ISSUES IN REVERSIBLE JUMP MARKOV CHAIN MONTE CARLO AND COMPOSITE EM ANALYSIS, APPLIED TO SPATIAL POISSON CLUSTER PROCESSES

by

John Mathias Castelloe

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Statistics in the
Graduate College of The
University of Iowa

December 1998

Thesis supervisor: Associate Professor Dale Zimmerman

ABSTRACT

In the analysis of spatial point patterns, it is generally assumed that the underlying spatial point process is "isotropic," i.e., that all characteristics are homogeneous with respect to direction. However, this is known in many applications not to be the case. For example, the distribution of plant seedling locations often exhibits directional asymmetry, or "anisotropy," due to factors such as prevailing wind direction and systematic migratory behavior of seed carriers. Failure to account for such directional inhomogeneity can result in erroneous inferences.

A special type of spatial point process is considered, namely the 2-dimensional Poisson cluster process with bivariate normal offspring dispersal (BVNPCP). Estimation of the parameters of a BVNPCP (the focus being the "cluster shape/scale parameter," the covariance matrix of the offspring dispersal distribution) is particularly challenging due to the substantial amount of latent data. The offspring relationships, number of parents and locations of parents are all unknown. Two approaches for testing for and estimating anisotropy are developed and applied to a collection of actual and simulated spatial point patterns.

The first approach considers the BVNPCP as a finite mixture model and combines EM algorithm parameter estimates, computed separately for different numbers of clusters, in a Bayesian model averaging type scheme. A "composite EM" estimator of the cluster shape/scale parameter is thus constructed, along with an estimated asymptotic variance computed from a combination of observed information matrices.

In the second approach, a reversible jump Markov chain Monte Carlo (RJMCMC) technique for 2-dimensional normal mixtures is developed. RJMCMC extends the traditional MCMC capabilities by providing for transitions between different parameter spaces, which are needed in our situation due to the unknown number of clusters. A new convergence assessment method, applicable to *any* RJMCMC situation in which distinct models can be identified, is designed and theoretically justified. Output analysis methods are also developed, including anisotropy testing/estimation, model checking and inference for number of clusters. The RJMCMC technique is flexible and has potential to apply to more complicated spatial point processes, and also other mixture-related problems.

Abstract approved: _____
Thesis supervisor

_____
Title and department

_____
Date

ISSUES IN REVERSIBLE JUMP MARKOV CHAIN MONTE CARLO AND COMPOSITE EM ANALYSIS, APPLIED TO SPATIAL POISSON CLUSTER PROCESSES

by

John Mathias Castelloe

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Statistics in the
Graduate College of The
University of Iowa

December 1998

Thesis supervisor: Associate Professor Dale Zimmerman

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

John Mathias Castelloe

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Statistics at the December 1998
graduation.

Thesis committee: _____
Thesis supervisor

_____
Member

_____
Member

_____
Member

_____
Member

To Raymond, Lillian, Helen, Edward, Mary, Ann, John, and the rest of my wonderfully supportive family, whose inspiration is always with me

ACKNOWLEDGMENTS

I would like to thank my advisor, Dale Zimmerman, for his valued guidance, patience, motivation, and reminders that sometimes it's best to "go back to first principles;" and for bringing this intriguing research topic to my attention. I am grateful to Kate Cowles for helping me learn the ways of the Bayesian force, and Russ Lenth for teaching me subtleties of statistical computing and experimental design. Many thanks also to my other committee members, Richard Dykstra, John Nason, and Ralph Russo, for helpful assistance and brainstorming along the way. I am deeply appreciative of my friends, who have inspired me to reach higher: Betsy, Johnny, Tammy, Chris, Erin, Audrey, Ken, Sam, Helen and many others. Heartfelt thanks to younger family members, for being there: Michael, Rebecca, Julie, Brian and Jennie. Thanks, mom, for the daily e-mails of encouragement. Finally, special thanks to my dad and grandpa, whose remarkable recovery and continuing dedication following victories over a heart attack and cancer have helped me believe that anything is possible.

ABSTRACT

In the analysis of spatial point patterns, it is generally assumed that the underlying spatial point process is "isotropic," i.e., that all characteristics are homogeneous with respect to direction. However, this is known in many applications not to be the case. For example, the distribution of plant seedling locations often exhibits directional asymmetry, or "anisotropy," due to factors such as prevailing wind direction and systematic migratory behavior of seed carriers. Failure to account for such directional inhomogeneity can result in erroneous inferences.

A special type of spatial point process is considered, namely the 2-dimensional Poisson cluster process with bivariate normal offspring dispersal (BVNPCP). In this process, "parent" events are assumed to be located uniformly in some region. Each parent event gives rise to a collection of "offspring" events, displaced according to a common bivariate normal distribution. The resulting point pattern is taken to be the collection all offspring events, with no information about parents recorded. If the covariance matrix (called the "cluster shape/scale parameter") of the bivariate normal distribution is a multiple of the identity matrix, then isotropy holds, with clusters having a circular shape. Otherwise, the process is anisotropic with elliptical clusters.

Estimation of the parameters of a BVNPCP is particularly challenging due to the substantial amount of latent data. The offspring relationships, number of parents and locations of parents are all unknown. In this thesis, two approaches for testing for and estimating anisotropy are developed and applied to a collection of

actual and simulated spatial point patterns. The cluster shape/scale parameter is re-parameterized in terms of anisotropy strength, anisotropy direction, and cluster size to allow for more transparent interpretation of results.

The first approach considers the BVNPCP as a finite mixture model and combines EM algorithm parameter estimates, computed separately for different numbers of clusters, in a Bayesian model averaging type scheme. A "composite EM" estimator of the cluster shape/scale parameter is thus constructed, along with an estimated asymptotic variance computed from a combination of observed information matrices.

In the second approach, a reversible jump Markov chain Monte Carlo (RJMCMC) technique for 2-dimensional normal mixtures is developed. RJMCMC extends the traditional MCMC capabilities by providing for transitions between different parameter spaces, which are needed in our situation due to the unknown number of clusters. A new convergence assessment method, applicable to *any* RJMCMC situation in which distinct models can be identified, is designed and theoretically justified. A "model" in our case is a given number of clusters, in other words, the number of components in a mixture. Output analysis methods are also developed, including anisotropy testing/estimation, model checking and inference for number of clusters. The RJMCMC technique is flexible and has potential to apply to more complicated spatial point processes, and also other mixture-related problems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION AND PRELIMINARIES

1.1 The Problem

Many naturally occurring phenomena give rise to data in the form of a set of event locations, or a *spatial point pattern*. Examples include the dispersal of trees in a forest and the locations of nuclei of a patch of biological cells. In the field of spatial point pattern analysis, a pattern is typically described as *random*, *regular*, or *aggregated*. Aggregated patterns can arise from either some sort of clustering mechanism or from environmental variation leading to high concentration of events in certain areas. Of particular interest in this thesis is a specific kind of process (defined formally in section 1.1.2) which generates aggregated patterns in the plane through a clustering mechanism.

As Ripley (1977) asserts in his influential paper on modeling spatial point patterns, in the theory of spatial point processes "one of the earliest and most intensively studied classes of models is the class of cluster processes" (p. 174), the most important subclass of which is the *Neyman-Scott process* (Neyman and Scott, 1958). A Neyman-Scott process (see section 1.1.2) consists of two stages: a set of *parent events* is distributed uniformly in a region, and *offspring events* are dispersed around the parents. The resulting *pattern* is taken to be the collection of all offspring locations, i.e., no information regarding parent events is recorded. The parameters of such a process describe the average number of parents per unit area, the distribution of the numbers of offspring per parent, and the dispersal of

offspring around their parents.

Despite the popularity of Neyman-Scott processes as models for point patterns, "little is known on the statistical estimation of [their] parameters in the planar case" (Stoyan, 1992, p. 67). Estimation is particularly challenging because there is a substantial amount of latent data underlying observed patterns: the number of parents, locations of parents, and relationships between offspring are all unknown. Furthermore, all methods known to the author assume that the dispersal of offspring around parents is radially symmetric (i.e., *isotropic*). The assumption of isotropy simplifies the mathematics considerably but is unrealistic for many situations.

The aim of this thesis is to develop tests of isotropy for a special type of Neyman-Scott process, and to estimate the parameters describing offspring dispersal, as well as some of the latent data, without assuming isotropy. Two main approaches are developed. The first (Chapter 3) combines results from several different EM algorithm runs, while the second (Chapters 4 – 6) involves a *reversible jump Markov chain Monte Carlo (RJMCMC)* scheme. Chapters 1 – 2 discuss necessary preliminaries. Results for real and simulated data sets are presented in Chapter 7, and a summary of new methods and scope for future research are given in Chapter 8.

1.1.1 Spatial Point Processes

Before presenting the specific model of interest, we define some basic terminology of spatial point processes and patterns. Diggle (1983) provides a good overview of spatial point pattern analysis. A *spatial point pattern* is a finite set of points (*events*) in a spatial domain A whose locations are modeled as random variables. A spatial point pattern is regarded as a partial realization of a *spatial point process* (SPP), which is a random mechanism for generating a countable set of events in A.

The region A is taken to be a window of observation (for our purposes, a "study region" in $\Re^2$) and not the entire domain of the process.

Let $N(B)$ denote the number of events in an arbitrary region $B \subset A$, $|B|$ the area of B, and $\mathrm{d}\mathbf{x}$ an infinitesimal region containing a point $\mathbf{x} \in A$. The simplest SPP to specify is a *homogeneous Poisson process* (HPP), in which the locations of events are independently and identically distributed according to the uniform distribution on A.

The *intensity function* $\lambda(\cdot)$ of a SPP is defined as follows:

$$\lambda(\mathbf{x}) = \lim_{|\mathrm{d}\mathbf{x}| \to 0} \left(\frac{E\left[N(\mathrm{d}\mathbf{x})\right]}{|\mathrm{d}\mathbf{x}|} \right).$$

The *second-order intensity function* $\lambda_2(\cdot)$ is defined similarly:

$$\lambda_2(\mathbf{x}, \mathbf{y}) = \lim_{|\mathrm{d}\mathbf{x}|, |\mathrm{d}\mathbf{y}| \to 0} \left(\frac{E\left[N(\mathrm{d}\mathbf{x}) N(\mathrm{d}\mathbf{y})\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right).$$

A minor variation of $\lambda_2(\cdot)$ is the *covariance density*

$$\zeta(\mathbf{x}, \mathbf{y}) = \lambda_2(\mathbf{x}, \mathbf{y}) - \lambda(\mathbf{x})\lambda(\mathbf{y}),$$

which can be interpreted as the covariance between event counts per unit area in two infinitesimal regions centered at $\mathbf{x}$ and $\mathbf{y}$.

A process is called *stationary* if there are no underlying environmental factors encouraging or discouraging the occurrence of events at particular locations, i.e., if all probabilistic statements about it in any region $B \subset A$ are invariant under arbitrary *translations* of B. For a stationary SPP, $\lambda(\mathbf{x}) \equiv \lambda$ (in which case λ is interpreted as the expected number of events per unit area), and $\lambda_2(\mathbf{x}, \mathbf{y}) \equiv \lambda_2(\mathbf{z})$ where $\mathbf{z} = \mathbf{x} - \mathbf{y}$. Stationarity is an important property of SPP's that must hold for many theoretical quantities to be well-defined; *we assume throughout that all SPP's discussed are stationary.*

Another property we will assume throughout is that of *orderliness.* A SPP is

orderly if multiple coincident events cannot occur, i.e.,

$$\lim_{|\mathrm{d}\mathbf{x}|\to 0}\left(\frac{P\left(N(\mathrm{d}\mathbf{x})>1\right)}{|\mathrm{d}\mathbf{x}|}\right)=0 \quad \forall \mathbf{x}\in A,$$

which implies that

$$\lim_{|\mathrm{d}\mathbf{x}|\to 0}\left(\frac{E\left[N(\mathrm{d}\mathbf{x})\right]}{P\left(N(\mathrm{d}\mathbf{x})=1\right)}\right)=1 \quad \forall \mathbf{x}\in A$$

(Diggle, 1983). We further assume, as in Diggle (1983), that

$$\lim_{|\mathrm{d}\mathbf{x}|\to 0}\left(\frac{E\left[N(\mathrm{d}\mathbf{x})N(\mathrm{d}\mathbf{y})\right]}{P\left(N(\mathrm{d}\mathbf{x})=N(\mathrm{d}\mathbf{y})=1\right)}\right)=1 \quad \forall \mathbf{x},\mathbf{y}\in A. \tag{1.1}$$

A SPP is *isotropic* if symmetry exists in every way with regard to direction, i.e., if all probability statements about it in any region $B \subset A$ are invariant under arbitrary *rotations* of B. If a SPP contains any violation of this condition, then it is defined as *anisotropic*. Note that under isotropy, $\lambda_2(\mathbf{z})$ further reduces to $\lambda_2(t)$ where $t = \mathbf{z}'\mathbf{z}$. For a HPP, $\lambda_2(\mathbf{x},\mathbf{y}) = \lambda^2$ and thus $\zeta(\mathbf{x},\mathbf{y}) = 0 \quad \forall \mathbf{x},\mathbf{y}\in A$.

Ripley's K-function (Ripley, 1976) is important for describing interaction between events at various ranges. It is defined as follows:

$$K(t)=\frac{1}{\lambda}E\left[\#\text{ of (other) events within } t \text{ of a randomly chosen event}\right].$$

For a HPP, $K(t) = \pi t^2$. Higher values indicate clustering, and lower values indicate regularity. Under isotropy, stationarity, orderliness and (1.1), $\lambda_2(t)$ and $K(t)$ have a simple relationship (Diggle, 1983, p. 48):

$$\lambda_2(t)=\lambda^2(2\pi t)^{-1}\frac{\mathrm{d}}{\mathrm{d}t}K(t).$$

Most methods in spatial point pattern analysis involve estimation of the intensity, second-order intensity, and/or K-function. Thus it is useful in model fitting to know the theoretical form of these quantities for different candidate point process models.

Finally, a spatial point *pattern* is represented as the event locations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where n is the observed number of events in A.

1.1.2 The Poisson Cluster Process with Bivariate Normal Displacement (BVNPCP)

Spatial cluster processes have been developed in several different forms (Neyman and Scott, 1958, 1972; Strauss, 1975; Kelly and Ripley, 1976; Ripley, 1977; Diggle, 1975, 1978, 1983). Neyman and Scott (1972) discuss several applications, including spatial distribution of larvae on an experimental field and clustering of galaxies. All instances involving analysis assume isotropy of the offspring dispersal distribution. We will study a process with offspring dispersal determined by a common bivariate normal distribution with arbitrary covariance matrix, thus allowing for anisotropy. First a more general cluster process, the *Poisson cluster process*, is defined (Neyman and Scott, 1958; Diggle, 1983):

Definition 1.1.1 *A* Poisson cluster process (PCP) *is given by the following 3 postulates:*

PCP1 *Parent events form a HPP in* $\Re^2$ *with intensity* ρ.

PCP2 *Each parent* j *produces a random number* S_j *of offspring, realized independently and identically for each parent according to a probability distribution* $\{p_s, s = 0, 1, \ldots\}$.

PCP3 *The positions of the offspring relative to their parents are independently and identically distributed in* $\Re^2$ *according to a common bivariate p.d.f.* $h(\cdot)$.

Note that a PCP is stationary, since the parent process is a HPP. Our model of interest is a special case of the PCP:

Definition 1.1.2 *A* bivariate normal Poisson cluster process (BVNPCP) *is a PCP with the following special distributions:*

1. *The cluster counts* $\{S_j\}$ *are distributed according to a common Poisson distribution with rate* ν.

2. *The offspring dispersal distribution* $h(\cdot)$ *is bivariate normal with mean zero and positive definite covariance matrix* $\mathbf{\Sigma}$.

A realization of a BVNPCP is taken to be the collection of all offspring events in a region $A \in \Re^2$. Perhaps an extra "Poisson" should be inserted into the name since there are two Poisson distributions involved (one of the parent process and the other of the cluster counts), but it is omitted for simplicity.

Of primary interest will be the estimation of $\mathbf{\Sigma}$. A BVNPCP with $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and σ^2 any positive constant, will produce circular clusters, whereas an arbitrary BVNPCP will produce elliptical clusters. The correspondence between components of $\mathbf{\Sigma}$ and the resulting cluster shape/scale is developed in section 1.1.3. While the model of elliptically-shaped clusters may not be general enough to convincingly represent a large class of natural processes, it is a substantial generalization of the commonly used class of BVNPCP's with radially symmetric offspring dispersal ($\mathbf{\Sigma} = \sigma^2\mathbf{I}$) (see, for example, Neyman and Scott (1972); Diggle (1983); Lawson (1995a)).

The bivariate normal component of the BVNPCP is (barely, perhaps) mathematically simple enough to make approaches such as maximum likelihood estimation and the EM algorithm analytically tractable. It seems a natural starting point, especially in light of the concept of geometric anistropy (see section 1.1.3). The RJMCMC approach (see Chapters 4 – 6), on the other hand, appears to offer the exciting possibility of generalization to a much larger class of offspring dispersal distributions. Hence the choice of the BVNPCP for analysis in this thesis: it is general

enough to make an important addition to the class of useful point process models, and simple enough to allow the comparison of different approaches for analysis.

1.1.3 Geometric Anisotropy

If a circle is "stretched," the result is an ellipse (the term "stretch" referring to the multiplication of the coordinates of any axis by a constant $c > 1$). For a spatial point process, a similar phenomenon can exist in the form of a force along an axis. For example, the path of the sun may influence the locations of one plant species relative to another through shading effects. A process which results from the introduction of such a axial force to an isotropic process, effectively changing the scale of a particular axis, is called *geometrically anisotropic.*

Geometric anisotropy is a popular model for anisotropy in the related field of geostatistics (Ecker and Gelfand, 1997; Zimmerman, 1994, section 13.5.3). The correlation between observations taken at different sites is taken to have elliptical contours, due to some directional force (for example, prevailing wind direction in Zimmerman (1994)). Although there are differences between the concept of geometric anisotropy in spatial point pattern analysis and that in geostatistics, a precedent seems to be in place for its use.

In spatial point process terminology, geometric anisotropy is defined as follows:

Definition 1.1.3 *A point process is* geometrically anisotropic *if the second-order intensity function has elliptical contours, i.e.,*

$$\lambda_2(\mathbf{x},\mathbf{y}) \equiv \lambda_2\left(\left[(\mathbf{x}-\mathbf{y})'\mathbf{M}^{-1}(\mathbf{x}-\mathbf{y})\right]^{1/2}\right) \qquad (1.2)$$

for some positive definite matrix $\mathbf{M}$ *with* $|\mathbf{M}| = 1$,

in which case we use the term geometric second-order intensity *and write*

$$\lambda_2(\mathbf{x},\mathbf{y}) \equiv \lambda^g_{2(M)}(t), \qquad \textit{where } t = \left[(\mathbf{x}-\mathbf{y})'\mathbf{M}^{-1}(\mathbf{x}-\mathbf{y})\right]^{1/2}.$$

The matrix $\mathbf{M}$ *induces a Mahalanobis (non-Euclidean) distance measure. It is assumed, for identifiability and without loss of generality, to have a determinant equal to unity.*

Note that under isotropy, (1.2) holds for $\mathbf{M} = \mathbf{I}$ (the identity matrix), and the contours are circular.

1.1.4 Geometric Anisotropy of the BVNPCP

The form of the second-order intensity function for a general PCP and a BVNPCP are established in Theorems 1.1.7 and 1.1.8. First two useful lemmas are presented, and the form of the (first-order) intensity function is derived.

Lemma 1.1.4 (Wald's equation) *Let* $X_1, X_2, \ldots$ *be i.i.d. random variables with finite mean. Let* N *be a non-negative integer-valued random variable independent of* $\{X_1, X_2, \ldots\}$ *and with finite mean. Then*

$$E\left(\sum_{i=1}^{N} X_i\right) = E(N)E\left(X_1\right)$$

Proof: See Grimmet and Stirzaker (1992, p. 396).

The following theorem is stated without proof in Diggle (1983, p. 55):

Theorem 1.1.5 *The intensity of a PCP is given by:* $\lambda = \rho\nu$.

Proof: Consider a PCP observed in a finite region A. Let

$$\begin{aligned} n_p &= \#(\text{parents in } A) \\ \mathbf{x}_{ij} &= j^{\text{th}} \text{ location of offspring of } i^{\text{th}} \text{ parent} \\ \mathbf{x} &= \text{location of arbitrary offspring from arbitrary parent} \\ B &= \text{finite region in } \Re^2 \end{aligned}$$

(Note: the dependence of these quantities on A is suppressed in the notation). Then:

$$\begin{aligned}
E[N(B)] &= \lim_{A\to\Re^2} E[N(B)] \\
&= \lim_{A\to\Re^2} E\left[\sum_{i=1}^{n_p}\sum_{j=1}^{S_i} 1_B(\mathbf{x}_{ij})\right] \\
&= \lim_{A\to\Re^2} \left\{E(n_p)E\left[\sum_{j=1}^{S} 1_B(\mathbf{x})\right]\right\} \\
&\quad \text{(by Lemma 1.1.4, since } n_p \text{ and } \left\{\textstyle\sum_{j=1}^{S_i} 1_B(\mathbf{x}_{ij})\right\} \text{ are independent,} \\
&\quad\quad \text{and } \left\{\textstyle\sum_{j=1}^{S_i} 1_B(\mathbf{x}_{ij})\right\} \text{ are i.i.d., } i = 1,\ldots,n_p) \\
&= \lim_{A\to\Re^2} \{E(n_p)E(S)P(\mathbf{x}\in B)\} \\
&\quad \text{(by Lemma 1.1.4, since } S \text{ and } 1_B(\mathbf{x}) \text{ are independent)} \\
&= \lim_{A\to\Re^2} \{\rho|A|\nu P(\mathbf{x}\in B)\} \\
&= \lim_{A\to\Re^2} \left\{\rho|A|\nu\frac{|B|}{|A|}\right\} \\
&\quad \text{(boundary effects possibly disturbing } P(\mathbf{x}\in B) \text{ become negligible} \\
&\quad\quad \text{for the fixed region } B \text{ as } A\to\Re^2;\ \mathbf{x} \text{ is an independent} \\
&\quad\quad \text{displacement of a uniformly distributed quantity, and thus} \\
&\quad\quad \text{marginally distributed uniformly)} \\
&= \rho\nu|B|.
\end{aligned}$$

Hence $\lambda = \rho\nu$. $\square$

Lemma 1.1.6 *Suppose $f(\cdot)$ is a p.d.f. defined on $\Re^2$ and continuous everywhere. Then, for any fixed $\mathbf{c}\in\Re^2$ and $D\subset\Re^2$, there exists $B<\infty$ such that*

$$\frac{1}{|D|}\int_D f(\mathbf{u}-\mathbf{c})\mathrm{d}\mathbf{u} \le B.$$

Proof: Since $f(\cdot)$ is a p.d.f., we have $\int_{\Re^2} f(\mathbf{u}-\mathbf{c})\mathrm{d}\mathbf{u} = 1$. Since $f(\cdot)$ is continuous, this clearly implies the existence of a constant $B<\infty$ such that $f(\mathbf{u}-\mathbf{c}) \le B \quad \forall\mathbf{u}\in$

$\Re^2$. Consequently,

$$\frac{1}{|D|}\int_D f(\mathbf{u}-\mathbf{c})\mathrm{d}\mathbf{u} \le \frac{1}{|D|}\int_D B\mathrm{d}\mathbf{u} = \frac{1}{|D|}|D|B = B. \quad \square$$

The following theorem is stated without proof in Ripley (1977, p. 174) and Diggle (1983, p. 55):

Theorem 1.1.7 *The second-order intensity function of a PCP with $h(\cdot)$ continuous is given by:*

$$\lambda_2(\mathbf{x},\mathbf{y}) = \lambda^2 + \rho E\left\{S(S-1)\right\}h_2(\mathbf{x}-\mathbf{y})$$

where

$$h_2(\mathbf{x}-\mathbf{y}) = \int_{\Re^2} h(\mathbf{x})h(\mathbf{x}-\mathbf{z})\mathrm{d}\mathbf{x},$$

the p.d.f. of the vector difference between two offspring from the same parent.

Proof: See Appendix A.1.

Geometric anisotropy of the BVNPCP is established by the following theorem:

Theorem 1.1.8 (Geometric anisotropy of BVNPCP) *The BVNPCP satisfies geometric anisotropy with $\mathbf{M} = |\mathbf{\Sigma}|^{-\frac{1}{2}}\mathbf{\Sigma}$ and*

$$\lambda^g_{2(M)}(t) = (\rho\nu)^2 + \rho\nu^2\left[\frac{1}{(2\pi)|2\mathbf{\Sigma}|^{1/2}}\exp\left(-\frac{t^2}{4|\mathbf{\Sigma}|^{\frac{1}{2}}}\right)\right].$$

Proof: Let $\mathbf{x}$ and $\mathbf{y}$ represent locations of two arbitrary (distinct) offspring from the same parent. Since $\mathbf{x},\mathbf{y} \sim N(\mathbf{0},\mathbf{\Sigma})$ and are independent, we have $\mathbf{x}-\mathbf{y} \sim N(\mathbf{0},2\mathbf{\Sigma})$. Also, since $S \sim$ Poiss(ν), we have $E[S(S-1)] = (\nu+\nu^2)-\nu = \nu^2$. Thus, using Theorems 1.1.5 and 1.1.7, we have

$$\begin{aligned}
\lambda_2(\mathbf{x},\mathbf{y}) &= \lambda^2 + \rho E[S(S-1)]h_2(\mathbf{x}-\mathbf{y}) \\
&= (\rho\nu)^2 + \rho\nu^2\left[\frac{1}{(2\pi)|2\mathbf{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{y})'(2\mathbf{\Sigma})^{-1}(\mathbf{x}-\mathbf{y})\right)\right] \\
&= (\rho\nu)^2 + \rho\nu^2\left[\frac{1}{(2\pi)|2\mathbf{\Sigma}|^{1/2}}\exp\left(-\frac{t^2}{4|\mathbf{\Sigma}|^{\frac{1}{2}}}\right)\right] \\
&\quad \text{where } t = \left[(\mathbf{x}-\mathbf{y})'\mathbf{M}^{-1}(\mathbf{x}-\mathbf{y})\right]^{1/2} \text{ and } \mathbf{M} = |\mathbf{\Sigma}|^{-\frac{1}{2}}\mathbf{\Sigma}.
\end{aligned}$$

Finally, $|\mathbf{M}| = 1$, thus satisfying the definition of $\lambda^g_{2(M)}(\cdot)$. $\square$

Note that for the BVNPCP, $\lambda^g_{2(M)}(t) > \lambda^2$ and decreases exponentially to λ^2 (the value for a HPP) in the limit as $t \to \infty$. In other words, the covariance density is strictly positive, being highest at $t = 0$ and decreasing exponentially to zero with the squared Mahalanobis distance between two locations.

So the anisotropy of a BVNPCP is completely determined by $\mathbf{\Sigma}$. The cluster shape/scale is governed by the elliptical contours of the $N(\mathbf{0}, \mathbf{\Sigma})$ density. In order to describe the cluster shape/scale in more useful terminology, we can re-parameterize $\mathbf{\Sigma}$ in terms of an *anisotropy parameterization*. First we define the usual parameterization, which we shall call the *regular parameterization*:

Definition 1.1.9 (regular parameterization of $\mathbf{\Sigma}$) *Consider a BVNPCP with cluster shape/scale parameter $\mathbf{\Sigma}$. Let $\mathbf{x} = (x_1, x_2)'$ be the location of an offspring relative to its parent. The* regular parameterization of $\mathbf{\Sigma}$, $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{22}, \sigma_{12})$, *is defined as follows:*

$$\mathbf{\Sigma} = \mathrm{Var}(\mathbf{x}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

where $\sigma_{11} = \mathrm{Var}(x_1)$, $\sigma_{22} = \mathrm{Var}(x_2)$, and $\sigma_{12} = \mathrm{Cov}(x_1, x_2)$.

Definition 1.1.10 (anisotropy parameterization of $\mathbf{\Sigma}$) *Consider a BVNPCP with cluster shape/scale parameter $\mathbf{\Sigma}$. Let $E_{\mathbf{\Sigma},\alpha}$ be the ellipse $\left\{\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} = \chi^2_2(1-\alpha)\right\}$, where $\mathbf{x} \in \Re^2$ and $\chi^2_2(1-\alpha)$ is the $(1-\alpha)^{\text{th}}$ quantile of the χ^2 distribution with 2 degrees of freedom. Let $I_{\mathbf{\Sigma},\alpha}$ be the interior of $E_{\mathbf{\Sigma},\alpha}$. Note that $E_{\mathbf{\Sigma},\alpha}$ describes the elliptical contours of $N(\mathbf{0}, \mathbf{\Sigma})$, and that $\mathbf{y} \sim N(\mathbf{0}, \mathbf{\Sigma}) \Rightarrow P(\mathbf{y} \in I_{\mathbf{\Sigma},\alpha}) = 1 - \alpha$* (Johnson and Wichern, 1992, Result 4.7).

The anisotropy parameterization of $\mathbf{\Sigma}$, (γ, ϕ, Ψ), *is defined as follows:*

1. *The* anisotropy strength, γ, *is defined as the ratio of the major semi-axis and minor semi-axis of $E_{\mathbf{\Sigma},\alpha}$. (Note that $\gamma \geq 1$).*

2. *The* anisotropy direction, ϕ, *is defined as the angle of inclination of the major axis of* $E_{\mathbf{\Sigma},\alpha}$, *which is the (smaller) angle between the major axis and the positive x-axis and lies in* $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

3. *The* cluster size, Ψ, *is defined as the square root of the determinant of* $\mathbf{\Sigma}$, *which is equal to* $\frac{Area(I_{\mathbf{\Sigma},\alpha})}{\pi\chi_2^2(1-\alpha)}$.

NOTE: (γ, ϕ, Ψ) *does* not *depend on* α, *which is used only to demonstrate the meaning of the magnitude of* Ψ. *This particular choice for* Ψ *is explained in the comments following the proof of Fact 1.1.11.*

For an isotropic BVNPCP with $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, note that $\gamma = 1$, $\Psi = \sigma^2$, and ϕ is not well-defined. The irrelevance of ϕ for isotropic BVNPCP's and the constraint $\gamma \geq 1$ render the anisotropy parameterization unsuitable for isotropy testing. However, it is still quite useful for estimation and descriptive purposes, especially for BVNPCP's with clear violations of isotropy.

The mathematical correspondence between $\mathbf{\Sigma}$ and the anisotropy parameterization is established by the following:

Fact 1.1.11 *Consider a BVNPCP with cluster shape/scale parameter* $\mathbf{\Sigma}$. *The two parameterizations* $(\sigma_{11}, \sigma_{22}, \sigma_{12})$ *and* (γ, ϕ, Ψ), *as defined in Definitions 1.1.9 and 1.1.10, are related as follows:*

$$\gamma = \left[\frac{\sigma_{11} + \sigma_{22} + \left[(\sigma_{11} - \sigma_{22})^2 + 4\sigma_{12}^2\right]^{1/2}}{\sigma_{11} + \sigma_{22} - \left[(\sigma_{11} - \sigma_{22})^2 + 4\sigma_{12}^2\right]^{1/2}}\right]^{1/2}$$

$$\phi = \begin{cases} 0, & \text{if } \sigma_{12} = 0 \text{ and } \sigma_{11} > \sigma_{22} \\ \frac{\pi}{2}, & \text{if } \sigma_{12} = 0 \text{ and } \sigma_{11} \leq \sigma_{22} \\ \arctan\left(\frac{-2\sigma_{12}}{\sigma_{22} - \sigma_{11} - [(\sigma_{11} - \sigma_{22})^2 + 4{\sigma_{12}}^2]^{1/2}}\right), & \text{if } \sigma_{12} \neq 0 \end{cases}$$

$$\Psi = \left[\sigma_{11}\sigma_{22} - \sigma_{12}^2\right]^{1/2}$$

and

$$\sigma_{11} = \Psi\left(\frac{1}{\gamma}\sin^2\phi + \gamma\cos^2\phi\right)$$

$$\sigma_{22} = \Psi\left(\frac{1}{\gamma}\cos^2\phi + \gamma\sin^2\phi\right)$$

$$\sigma_{12} = -\Psi\left(\frac{1}{\gamma} - \gamma\right)\sin\phi\cos\phi$$

Proof: First, $\Psi = \left[\sigma_{11}\sigma_{22} - {\sigma_{12}}^2\right]^{1/2}$ simply by definition. Write $\mathbf{\Sigma}^{-1}$ as

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} i_{11} & i_{12} \\ i_{12} & i_{22} \end{bmatrix} = \frac{1}{\sigma_{11}\sigma_{22} - {\sigma_{12}}^2}\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}. \tag{1.3}$$

Let $\mathbf{x} = (x_1, x_2)' \in \Re^2$ and consider the ellipse

$$E = \left\{\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} = 1\right\} = \left\{i_{11}x_1^2 + 2i_{12}x_1x_2 + i_{22}x_2^2 = 1\right\}. \tag{1.4}$$

Now define

$$\begin{aligned} a &= \text{major semi-axis of } E \\ b &= \text{minor semi-axis of } E \\ \phi &= \text{angle of inclination of major axis of } E \end{aligned}$$

Then E can also be written as (Batschelet, 1981, equation 13.3.2):

$$\begin{aligned} E = \Bigg\{&\left(\frac{\cos^2\phi}{a^2} + \frac{\sin^2\phi}{b^2}\right)x_1^2 + 2\cos\phi\sin\phi\left(\frac{1}{a^2} - \frac{1}{b^2}\right)x_1x_2 + \\ &\left(\frac{\sin^2\phi}{a^2} + \frac{\cos^2\phi}{b^2}\right)x_2^2 = 1\Bigg\}. \end{aligned} \tag{1.5}$$

The relationship between (a, b, ϕ) and (i_{11}, i_{22}, i_{12}) can be determined by equating the coefficients in 1.4 and 1.5. The mapping $(a, b, \phi) \mapsto (i_{11}, i_{22}, i_{12})$ is trivial, and the form of $(i_{11}, i_{22}, i_{12}) \mapsto (a, b, \phi)$ is derived by Batschelet (1981, section 13.5):

$$a = \frac{\sqrt{2}}{\left[i_{11} + i_{22} - \left[(i_{11} - i_{22})^2 + 4i_{12}^2\right]^{1/2}\right]^{1/2}}$$

$$b = \frac{\sqrt{2}}{\left[i_{11} + i_{22} + \left[(i_{11} - i_{22})^2 + 4i_{12}^2\right]^{1/2}\right]^{1/2}}$$

$$\phi = \begin{cases} 0, & \text{if } i_{12} = 0 \text{ and } i_{11} < i_{22} \\ \frac{\pi}{2}, & \text{if } i_{12} = 0 \text{ and } i_{11} \geq i_{22} \\ \arctan\left(\frac{-2i_{12}}{i_{11}-i_{22}-\left[(i_{11}-i_{22})^2+4i_{12}^2\right]^{1/2}}\right), & \text{if } i_{12} \neq 0 \end{cases}$$

Finally, the result is obtained by using the definition $\gamma = \frac{a}{b}$ and re-writing in terms of $(\sigma_{11}, \sigma_{22}, \sigma_{12})$ as determined by (1.3). $\square$

Note that (via simple substitution) the ellipse E can also be written in the following equivalent form in terms of (γ, ϕ, Ψ):

$$\begin{aligned} E &= \left\{ \left(\frac{1}{\gamma}\cos^2\phi + \gamma \sin^2\phi \right) x_1^2 + 2\cos\phi\sin\phi\left(\frac{1}{\gamma} - \gamma\right) x_1 x_2 + \right. \\ &\qquad \left. \left(\frac{1}{\gamma}\sin^2\phi + \gamma\cos^2\phi \right) x_2^2 = \Psi \right\} \\ &= \left\{ \mathbf{x}'\mathbf{M}^{-1}\mathbf{x} = \Psi \right\} \end{aligned}$$

$$\text{where } |\mathbf{M}| = 1 \text{ and } \mathbf{M} \text{ involves only } \gamma \text{ and } \phi,$$

thus justifying the representation chosen for Ψ in Definition 1.1.10.

1.1.5 Potential Applications in Ecology

Wright (1946) introduced the idea of a *genetic neighborhood* in population ecology as the "area from which the parents of central individuals may be treated as if drawn at random" (Crawford, 1984, p. 147). See Wright (1969) for a detailed explanation of the theory. The dispersal of pollen and seeds (combined) from parent plants is assumed to follow a bivariate normal distribution with covariance matrix $\mathbf{\Sigma} = \sigma_{nbd}^2\mathbf{I}$, where σ_{nbd} is estimated from measurements of pollen and seed dispersal *distances*, ignoring direction. The genetic neighborhood is then defined as the circle of radius $2\sigma_{nbd}$ (note that approximately 86.5% of observations from $N(\mathbf{0}, \sigma_{nbd}^2\mathbf{I})$ will lie in this circle). Many important ecological inferences are based on the concept of genetic neighborhoods. For example, genetic differentiation between two

populations is considered a function of the number of neighborhood diameters that separate them in space.

However, strong directionality of pollen and seed dispersal has been observed in nature. Directional migration patterns are quite common and are important in affecting gene flow, extinction and recolonization dynamics in natural populations. Crawford (1984, p. 157) expresses doubts about the neighborhood model:

> The basic model involves a number of assumptions that are unlikely to be true in nature. The most important are that dispersal distributions are normal, that these distributions have zero means and that they adequately reflect the form of gene dispersal between parents and offspring.

He also states that the most frequently encountered deviation from the assumed dispersal distribution is that of leptokurtosis, meaning an extended tail and excess of observations near the mean. Skewed dispersal distributions are often observed; directional behavior of pollinators and prevailing wind direction are cited as contributors to this effect.

Many adjustments for the effects of leptokurtosis have been proposed (Crawford, 1984), but all involve only adjustment of σ_{nbd}, leaving all other assumptions intact (most notably the assumption of radial symmetry). Methods to characterize the *directionality* of dispersal distributions could potentially produce an improved model for the genetic neighborhood.

1.2 Description of Datasets Used

One observed spatial point pattern and and a battery of twelve simulated patterns are analyzed in Chapter 7 using the techniques developed in Chapters 3 – 6. In this section we describe the patterns and the reasons for our choices.

1.2.1 Redwood Seedling Locations

Strauss (1975) studied the location of Redwood seedlings in an experimental plot. He states (p. 473) "it was felt that the seedlings would be scattered fairly randomly, except that a number of tight clusters would form around some of the redwood tree stumps present in the plot." Figure 1.1 shows the locations of the seedlings (with no information about the stumps). The diagonal line represents a discontinuity in the soil, below which very few Redwood stumps were found. Thus the clustering behavior is expected to be quite different in the two regions. The portion used for analysis in this thesis is indicated by a solid boundary and will henceforth be referred to as the "Redwood data." For convenience in analysis, the coordinate scale is chosen to produce a total area of 1.

Ripley (1977) extracts a square region (marked by dashed lines in Figure 1.1) mainly within the area of supposed clustering, citing computational convenience as his justification. He and other authors (Diggle, 1983; Lawson, 1993) have analyzed this square region as an isotropic PCP, leading to varied conclusions (see section 1.3).

The supposed clusters in the Redwood plot appear to exhibit strong directionality, suggesting a common northeast-southwest orientation. No reported analyses of this pattern account for or assess this directionality. Although there is no evidence suggesting that elliptical cluster shapes are reasonable, a visual inspection warrants the possibility. The Redwood data is thus analyzed as a BVNPCP.

1.2.2 Simulated Patterns

To enable a more thorough study of the performance of the methods developed, twelve simulated spatial point patterns are used for analysis. The number of such patterns is limited by constraints on computation time. Patterns are generated

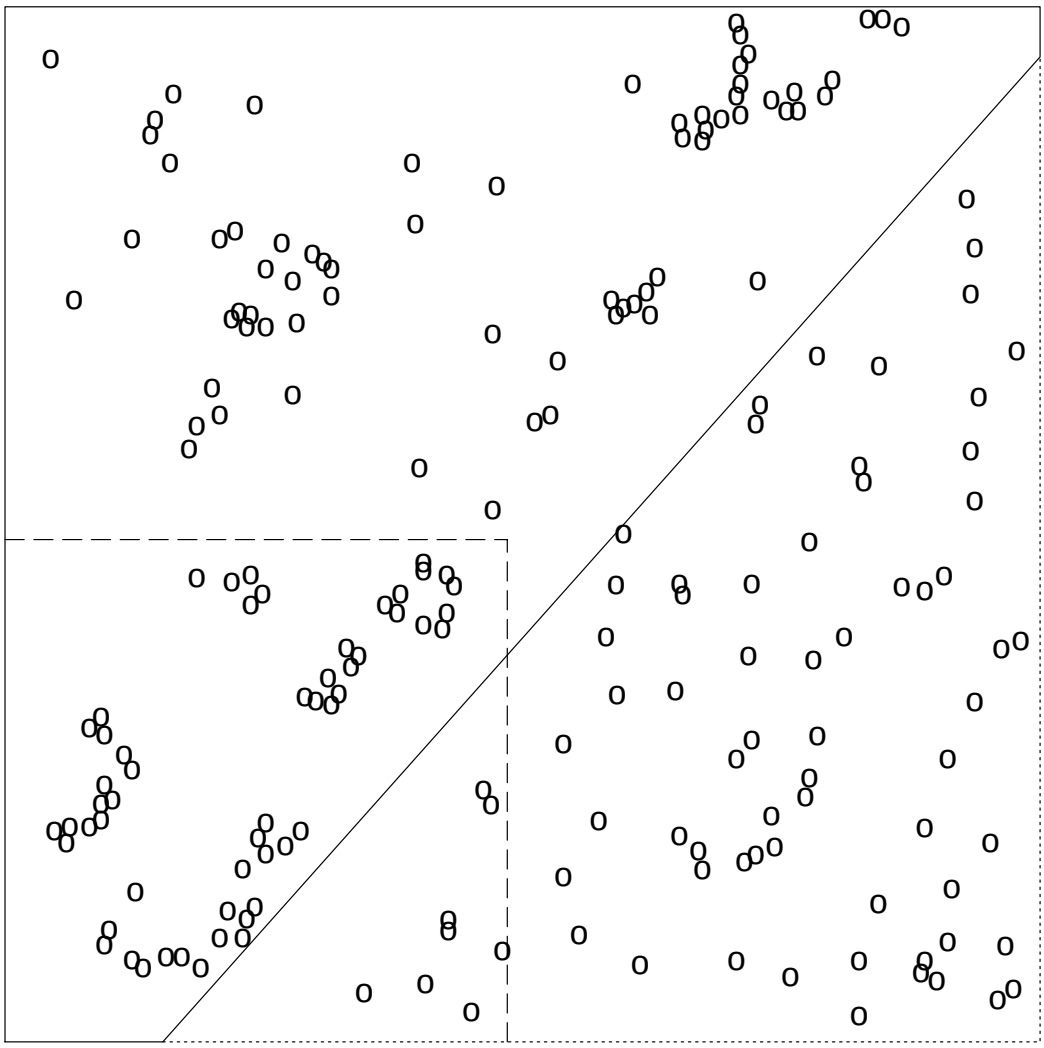


Figure 1.1: Location of Redwood seedlings in an experimental plot. Regions are marked according to use in this thesis (solid) and in other papers (dashed and dotted).

in the unit square according to a BVNPCP, conditional on number of clusters (k) and total number of offspring (n, set to 100), except that placement of parents is restricted to $[0.025, 0.975] \times [0.025, 0.975]$. (Note: terminology for such conditioning is given by Definition 3.1.1, and the equivalent specification given by Definition 3.1.3 is used for the actual simulation). The restriction on parent (cluster center) placement is used to reduce edge effects, since the robustness of methods to boundary effects is not studied.

Three factors are varied to generate the patterns: γ (anisotropy strength, set to 1, 1.5 or 3), k (number of clusters, set to 7 or 14) and Ψ (cluster size, set to .003 for $k = 7$ and .0015 for $k = 14$). Varying γ allows the study of different degrees of anisotropy, while varying k and Ψ allows the analysis of more and smaller clusters vs. fewer and larger clusters. The parameter ϕ (anisotropy direction) is set to $\frac{\pi}{6} = 30°$ for all anisotropic patterns. The use of a common value simplifies interpretation of results across patterns. Some methods we will develop later in the thesis analyze the variance difference $\sigma_{11} - \sigma_{22}$ and covariance σ_{12} separately. This particular choice of ϕ gives comparable (although not necessarily identical) importance to each in detecting departures from isotropy. Two replications of each factor combination are generated. To avoid selection bias, the first two such replications of each combination were accepted, regardless of the apparent adherence (or lack thereof) to model parameters.

The naming convention for the simulated BVNPCP realizations identify whether the underlying model is isotropic ("I") or anisotropic ("AI"), the value of γ in case of anisotropy ("1.5" or "3"), the value of k ("k7" or "k14"), and the replication ("a" or "b"). Table 1.1 shows the values of relevant quantities for the twelve simulated patterns, and Figures 1.2 – 1.7 show plots of the patterns.

Name	γ	k	Ψ	σ_{11}	σ_{22}	σ_{12}
I-k7-a	1	7	0.003	0.003	0.003	0
I-k7-b	1	7	0.003	0.003	0.003	0
I-k14-a	1	14	0.0015	0.0015	0.0015	0
I-k14-b	1	14	0.0015	0.0015	0.0015	0
AI-1.5-k7-a	1.5	7	0.003	0.003875	0.002625	0.001083
AI-1.5-k7-b	1.5	7	0.003	0.003875	0.002625	0.001083
AI-1.5-k14-a	1.5	14	0.0015	0.001938	0.001312	0.0005413
AI-1.5-k14-b	1.5	14	0.0015	0.001938	0.001312	0.0005413
AI-3-k7-a	3	7	0.003	0.007	0.003	0.003464
AI-3-k7-b	3	7	0.003	0.007	0.003	0.003464
AI-3-k14-a	3	14	0.0015	0.0035	0.0015	0.001732
AI-3-k14-b	3	14	0.0015	0.0035	0.0015	0.001732

Table 1.1: Simulated point patterns: values of BVNPCP parameters and realized latent data (to 4 significant digits). For all patterns, $n = 100$ and $\phi = \frac{\pi}{6}$.

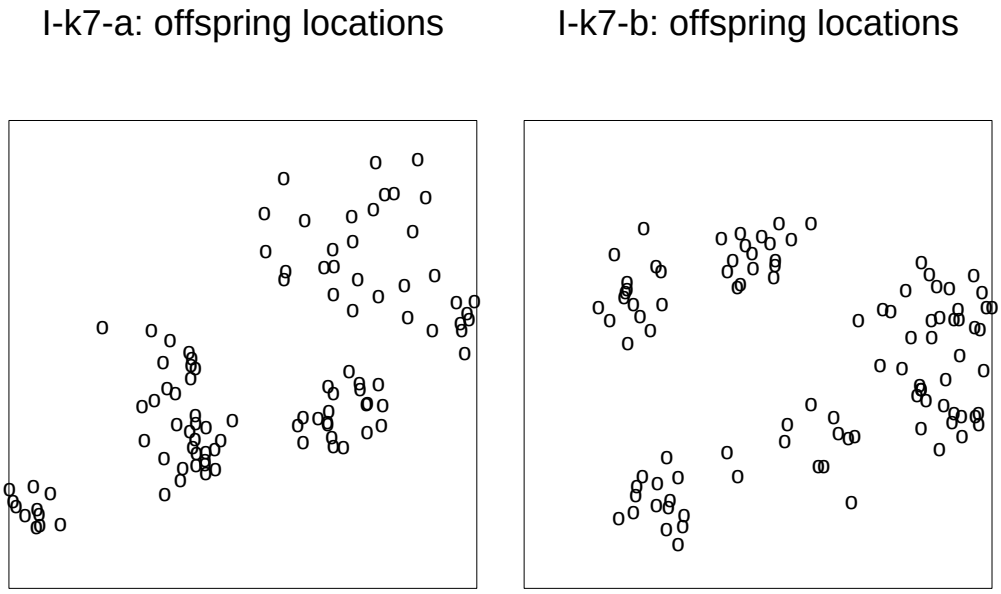


Figure 1.2: Simulated I-k7 patterns.

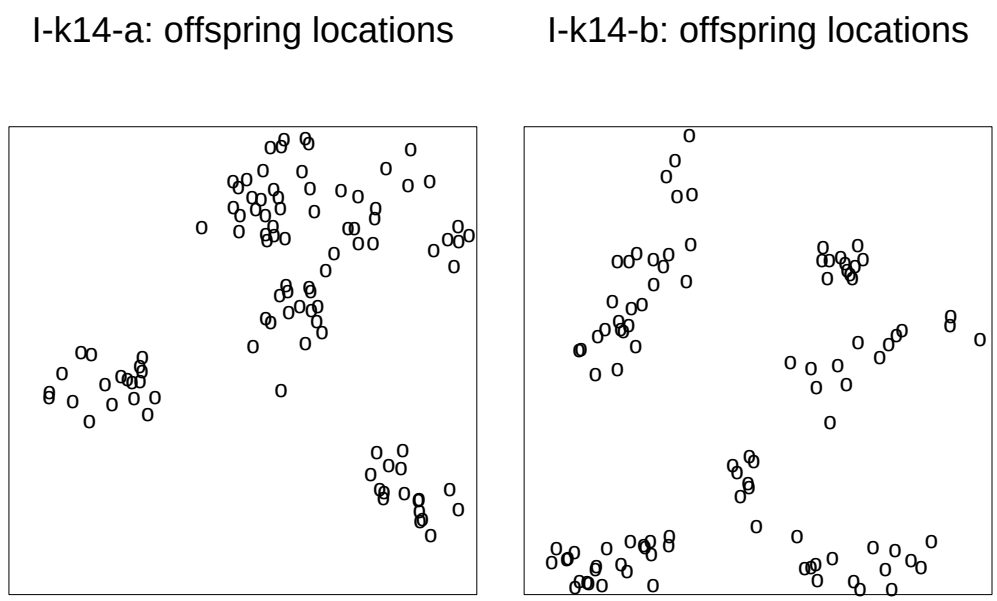


Figure 1.3: Simulated I-k14 patterns.

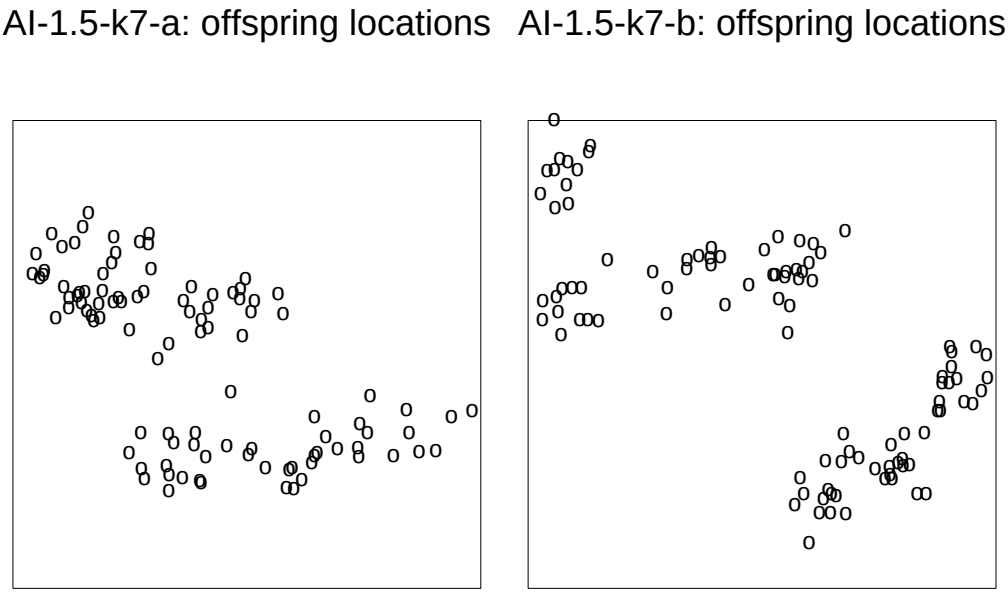


Figure 1.4: Simulated AI-1.5-k7 patterns.

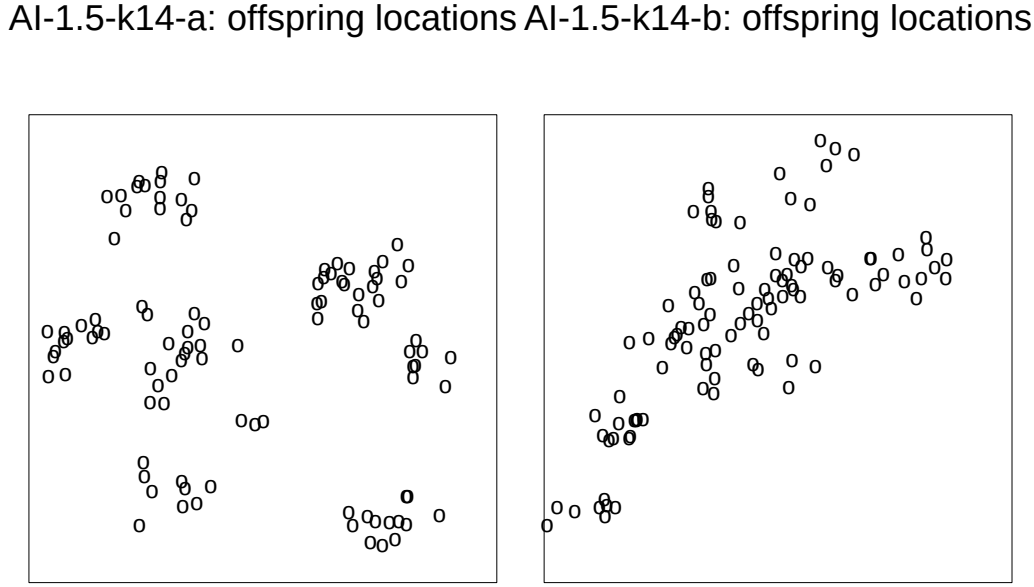


Figure 1.5: Simulated AI-1.5-k14 patterns.

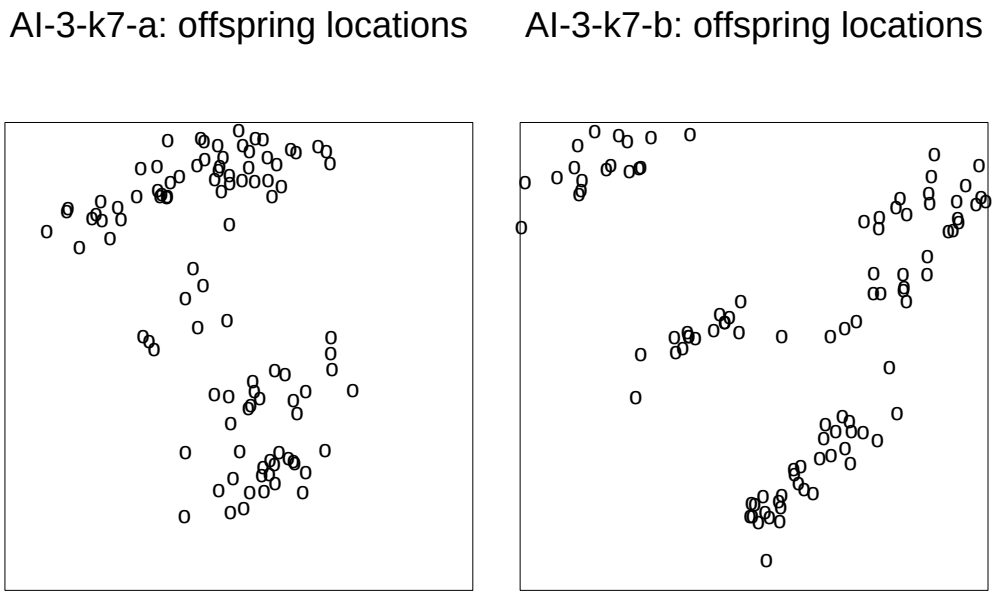


Figure 1.6: Simulated AI-3-k7 patterns.

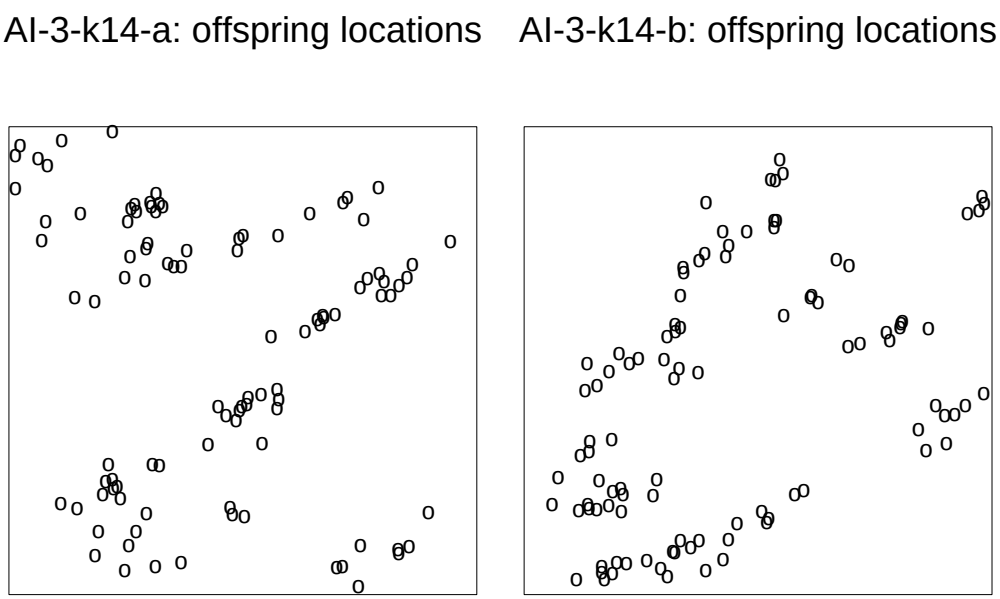


Figure 1.7: Simulated AI-3-k14 patterns.

1.3 Previous Approaches

To the author's knowledge, there are currently no other methods available to test isotropy of the offspring dispersal distribution of a PCP. Estimation even for isotropic PCP's has been a notoriously difficult problem with no clearly adequate solution, mostly because of the substantial amount of latent data (especially the unknown number of parents occurring in the region).

Furthermore, there is relatively little consideration of anisotropy at all in the spatial point pattern literature. There are several useful directional extensions of commonly used descriptive statistics of spatial point processes. However, in relatively few cases are the theoretical values known for commonly used models, and none have known distributions or standard error computations. Ohser and Stoyan (1981) define a version of the K-function which tallies the occurrence of other events within a *sector* whose origin is at an event. Stoyan and Stoyan (1994) derive an edge-corrected (i.e., including a correction for boundary effects) estimator for a generalized version which counts the occurrence of events in an arbitrarily shaped region around an event. Stoyan (1991) develops an estimator $\widehat{\lambda}_2(r, \phi)$, a kernel estimator of the second-order intensity as a function of distance r and direction ϕ. König and Schmidt (1992) construct an estimator of the distribution of direction between one event and another arbitrary event located with a given distance range from the first. Mugglestone and Renshaw (1996a,b) use spectral analysis to characterize anisotropic structure in a spatial pattern. Applications of these descriptive methods have been found mostly in the stereology and microscopy literature (see Stoyan and Beneš, 1991; Beneš et al., 1989; Carvajal-Gonzalez et al., 1989; König

and Ohser, 1988; Cruz-Orive et al., 1985).

Strauss (1975) developed a model for fixed-range interactions in spatial point processes (in which events are allowed to encourage or discourage the occurrence of other events within a certain fixed radius). He fit this model to the Redwood data (using the same region as in this thesis) and concluded there was substantial evidence of clustering. The interpretation of his "clustering tendency" parameter in relation to quantities considered in this thesis is unclear.

1.3.1 Least Squares Estimation for Isotropic PCP's

A popular method of parameter estimation for spatial point patterns is that of *least-squares estimation* (Diggle, 1983, Chapter 5). The essential idea is as follows: First a measure of some property of the point process (usually a function of distance t) is chosen. Examples include Ripley's $K(t)$, the distribution function of nearest-neighbor distances ($F(t)$), the distribution function of point-to-nearest-event distances ($G(t)$), and scaled versions of $\lambda_2(t)$ (Baudin, 1981; Stoyan, 1992). Say a measure $M(t)$ is chosen. The theoretical value $M(t;\theta)$ for different values of the unknown parameters (θ) is compared to an estimate $\widehat{M}(t)$ from the data, for various θ and t. A discrepancy function $D(\theta)$ is defined as

$$D(\theta) = \int_0^{t_0} \left[\left\{ \widehat{M}(t) \right\}^c - \{M(t;\theta)\}^c \right]^2 \mathrm{d}\mathbf{t}$$

with "tuning constants" t_0 and c. Then θ is estimated as the value $\hat{\theta}$ which minimizes $D(\theta)$. Diggle (1983) uses a quasi-Newton optimization procedure. A drawback of the method is the difficulty in choosing proper values for the tuning constants.

Ripley (1977) carries out a similar procedure with $K(t)$ for Strauss's fixed-range interaction model (using the square region from the lower-lefthand corner of the Redwood plot in Figure 1.1, scaled to give an area of 1) and, instead of using least

squares, tries a range of different parameter values and concludes that none provides an adequate fit. Diggle (1983) models the same region as a BVNPCP with $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. He uses $K(t)$ with $t_0 = 0.25$ and $c = 0.25$ to yield $(\hat{\rho}, \hat{\sigma}) = (25.6, 0.042)$, implying (surprisingly) the estimated presence of 25.6 clusters for only 62 observations.

1.3.2 MCMC Analysis

Granville and Smith (1995) consider a variant of the BVNPCP in which the cluster count distribution is geometric rather than Poisson. The cluster shape/scale parameter $\mathbf{\Sigma}$ is kept arbitrary. They develop a dimension-changing MCMC sampler based on a spatial birth-and-death process (Geyer and Møller, 1994) capable of modeling the number of clusters. However, they do not perform any convergence assessment. A similar method is suggested, although not developed, by Baddeley and van Lieshout (1993) for general PCP's. Lawson (1995a) develops a similar MCMC sampler for two classes of isotropic PCP's, simply mentioning that convergence assessment method is "based on Q-Q plots of the marginal distributions [of the parameters]." Lawson (1995b) models a variant of the isotropic BVNPCP in which the cluster shape/scale parameter is allowed to vary as a function of cluster center location, using a similar MCMC sampler and convergence assessment method. In all of these cases, output analysis is essentially restricted to the display of the posterior density estimate and reported modal values, and no details are given to allow one to reproduce the sampling algorithm. It is not clear whether any of the chains have satisfactory mixing properties, as rigorous convergence assessment is not performed.

Lawson (1993) mentions in a brief discussion contribution that he models the same Redwood data set used in Ripley (1977) and Diggle (1983) as a PCP with

a "'nearest parent' approximation to $h(\cdot)$" (not defined), using a Gibbs sampler (not explained) and Q-Q plots to assess convergence. He reports modal estimates of $k = 16$ clusters and $\sigma^2 = 0.00037$ (which is not defined and may or may not correspond to Diggle's σ).

These MCMC methods, most of which are based on the spatial birth-and-death process, can be considered precursors to the more versatile reversible jump Markov chain Monte Carlo technique (discussed in Chapters 4 – 6) developed by Green (1995).

CHAPTER 2
ATTEMPTS AT MAXIMUM LIKELIHOOD ESTIMATION

Consider the BVNPCP observed in a region $A \in \Re^2$. The parameters of this process can be represented as (see Definitions 1.1.1, 1.1.2 and 1.1.9)

$$\Phi = \{\rho, \nu, \mathbf{\Sigma}\} \equiv \{\rho, \nu, \sigma_{11}, \sigma_{22}, \sigma_{12}\}.$$

The observed data are

$$\mathbf{Y} = \{\text{locations of offspring in } A\} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)', \quad \text{where } \mathbf{y}_i = (y_{i1}, y_{i2})'. \tag{2.1}$$

The latent data can be expressed as follows:

$$k = \#(\text{parents in } A) \tag{2.2}$$

$$\boldsymbol{\mu} = \{\text{parent locations}\} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)', \quad \text{where } \boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})' \tag{2.3}$$

$$\mathbf{Z} = \{\text{“allocations”}\} = \begin{bmatrix} z_{11} & \cdots & z_{1k} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nk} \end{bmatrix}$$

$$\text{where } z_{ji} = \begin{cases} 1, & \text{if offspring } j \text{ belongs to parent } i \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

A CAUTIONARY NOTE: The above definitions are not entirely self-consistent because it is possible for parents in A to produce offspring outside A, and for parents outside A to produce offspring inside A. The author has not found a satisfactory remedy to account for this, and thus we will proceed as though the entire BVNPCP occurs within A. Estimation of intensity is not of concern in the thesis. As long as the study region is chosen carefully so that there is not an excess of clusters occurring on the boundary, this simplifying assumption is expected to have little

impact on the results. The Redwood pattern appears to meet this condition, and for the simulated patterns, parents were generated in an interior region (encompassing about 90% of the area). Further research is needed to adequately account for cases in which a significant number of clusters occur on the boundary of the study region.

The latent data component $\mathbf{Z}$ can also be referred to as "cluster memberships" or "parentage identifiers." The cluster counts are represented as the column sums of $\mathbf{Z}$:

$$\mathbf{s} = \{\text{cluster counts}\} = (S_1, \ldots, S_k)', \quad \text{where } S_i = \textstyle\sum_{j=1}^{n} z_{ji}.$$

Note that the "sample size" (n, the total number of offspring), is random. However, we proceed as is standard in statistical inference for spatial point processes and condition on the observed sample size (see e.g. Ripley (1977, 1981, 1988); Diggle (1983); Baddeley and Møller (1989)).

For the rest of this chapter (and also for use throughout the thesis), define the generic notation $p(\cdot)$ to denote a likelihood, p.d.f. or p.m.f., the meaning in each case being defined by the context. Also define the notation $a \sim b$ to denote that a is distributed according to the b distribution. Let the observed-data likelihood be represented as $p(\mathbf{Y}|\Phi, n)$. It is possible to express this likelihood in closed form by writing the complete-data likelihood $p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\mu}, k|\Phi, n)$ and integrating over the latent data. We can write the complete-data likelihood as:

$$p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\mu}, k|\Phi, n) = p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\mu}, k, \Phi, n)p(\mathbf{Z}|\boldsymbol{\mu}, \mathbf{s}, k, \Phi, n)p(\boldsymbol{\mu}, \mathbf{s}|k, \Phi, n)p(k|\Phi, n). \tag{2.5}$$

At this point a useful lemma is presented:

Lemma 2.1 *Let $X_1, \ldots, X_m$ be random variables, with $\mathbf{X} = (X_1, \ldots, X_m)'$. The following two statements are equivalent:*

$$X_1, \ldots, X_m \text{ are independent and identically distributed as } \text{Poiss}(\lambda) \tag{2.6}$$

and

$$\sum_{i=1}^{m} X_i \sim \text{Poiss}(m\lambda) \quad \textit{and} \quad \mathbf{X} \left| \left\{ \sum_{i=1}^{m} X_i = n \right\} \sim \text{Mult}\left(n, \frac{1}{m}\mathbf{1}\right) \right. \tag{2.7}$$

Proof:

(2.6) $\Rightarrow$ (2.7):

Assume (2.6). Then by Theorem 5.1 of Taylor and Karlin (1994), we have

$$\sum_{i=1}^{m} X_i \sim \text{Poiss}(m\lambda).$$

Also,

$$p\left(\mathbf{X} \left| \sum_{i=1}^{m} X_i = n \right.\right) = \frac{p(\mathbf{X}, n)}{p(n)} = \frac{\prod_{i=1}^{m} \frac{\lambda^{X_i} \exp(-\lambda)}{X_i!}}{\left[\frac{(m\lambda)^n \exp(-m\lambda)}{n!}\right]} = \binom{n}{X_1 \cdots X_m} \left(\frac{1}{m}\right)^n,$$

and thus (2.7) holds.

(2.7) $\Rightarrow$ (2.6):

Assume (2.7). Then

$$\begin{aligned} p(\mathbf{X}) &= p\left(\mathbf{X}, \sum_{i=1}^{m} X_i\right) \\ &= p\left(\mathbf{X} \left| \sum_{i=1}^{m} X_i = n \right.\right) P\left(\sum_{i=1}^{m} X_i = n\right) \\ &= \binom{n}{X_1 \cdots X_m} \left(\frac{1}{m}\right)^n \frac{(m\lambda)^n \exp(-m\lambda)}{n!} \\ &= \prod_{i=1}^{m} \frac{\lambda^{X_i} \exp(-\lambda)}{X_i!} \end{aligned}$$

Thus $X_1, \ldots, X_m$ are independent and identically distributed as Poiss(λ). $\square$

Each factor on the right-hand-side of (2.5) is derived in what follows. First note that

$$p(k|\Phi, n) = \frac{p(n|k, \Phi)p(k|\Phi)}{p(n|\Phi)}. \tag{2.8}$$

Now $n|\{k, \Phi\} \sim \text{Poiss}(k\nu)$ (by Lemma 2.1) and $k|\Phi \sim \text{Poiss}(\rho|A|)$ (Diggle, 1983, section 4.2), and so we have

$$\begin{aligned} p(n|k, \Phi) &= \frac{(k\nu)^n \exp(-k\nu)}{n!} \\ p(k|\Phi) &= \frac{(\rho|A|)^k \exp(-\rho|A|)}{k!}. \end{aligned}$$

The denominator of (2.8) is computed as:

$$\begin{aligned}
p(n|\Phi) &= P\left(\sum_{i=1}^{k} S_i = n \,\middle|\, \Phi\right) \\
&= \sum_{q=0}^{\infty} P\left(\sum_{i=1}^{q} S_i = n \,\middle|\, k=q, \Phi\right) P\left(k=q|\Phi\right) \\
&= \left(\frac{\nu^n \exp\left\{-\left[1-\exp(-\nu)\right]\rho|A|\right\}}{n!}\right) \cdot \\
&\quad \sum_{q=0}^{\infty} q^n \left[\frac{\left[\rho|A|\exp(-\nu)\right]^q \exp\left\{-\rho|A|\exp(-\nu)\right\}}{q!}\right] \\
&= \left(\frac{\nu^n \exp\left\{-\left[1-\exp(-\nu)\right]\rho|A|\right\}}{n!}\right) E\left(X^n\right) \\
&\quad \text{where } X \sim \text{Poiss}\left(\rho|A|\exp(-\nu)\right) \\
&= \left(\frac{\nu^n \exp\left\{-\left[1-\exp(-\nu)\right]\rho|A|\right\}}{n!}\right) \sum_{j=1}^{n} a_{n,j}\left[\rho|A|\exp(-\nu)\right]^j \\
&\quad \text{where } a_{n,j} = \begin{cases} 1, & \text{if } j=1 \text{ or } j=n \\ j(a_{n-1,j}) + a_{n-1,j-1}, & \text{otherwise,} \end{cases}
\end{aligned}$$

where the last equality follows from an induction argument (shown in Appendix A.2).

Next observe that $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)|\{k, \Phi, n\}$ are independent and distributed uniformly on A, and $\mathbf{s}|\{k, \Phi, n\} \sim \text{Mult}\left(n, \frac{1}{k}\mathbf{1}\right)$ (by Lemma 2.1), and $\boldsymbol{\mu}$ and $\mathbf{s}$ are independent, so that

$$p(\boldsymbol{\mu}, \mathbf{s}|k, \Phi, n) = \frac{1}{|A|^k}\binom{n}{S_1 \cdots S_k}\frac{1}{k^n}. \tag{2.9}$$

Given the offspring counts, all possible allocations satisfying the offspring counts are clearly equally likely (marginally, not taking into account the offspring locations). Denote the set of all possible allocations as $\Omega(\mathbf{s})$. The cardinality of $\Omega(\mathbf{s})$ is given by

$$\#(\Omega(\mathbf{s})) = \binom{n}{S_1 \cdots S_k}, \tag{2.10}$$

and so

$$p(\mathbf{Z}|\boldsymbol{\mu}, \mathbf{s}, k, \Phi, n) = \frac{1}{\binom{n}{S_1 \cdots S_k}}. \tag{2.11}$$

Finally, since $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ are independent, we have

$$p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\mu}, k, \Phi, n) = \prod_{i=1}^{k} \prod_{j=1}^{n} [f(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})]^{Z_{ji}} \tag{2.12}$$

where $f(\mathbf{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ denotes the density of $N(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Now that the form of the complete-data likelihood has been determined, it can be integrated to produce the observed-data likelihood:

$$\begin{aligned} p(y|\Phi, n) &= \int \cdots \int p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\mu}, k|\Phi, n) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\mathbf{Z} \, \mathrm{d}k \\ &= \sum_{k=0}^{\infty} \sum_{\mathbf{s} \in \Lambda_n(k)} \sum_{\mathbf{Z} \in \Omega(\mathbf{s})} \iint_A \cdots \iint_A p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\mu}, k|\Phi, n) \, \mathrm{d}\boldsymbol{\mu} \end{aligned} \tag{2.13}$$

$$\text{where } \Lambda_n(k) = \text{all possible values of } \mathbf{s} \text{ given } k \text{ and } n \tag{2.14}$$

and $\Omega(\mathbf{s})$ is as defined in (2.10).

The integral over $\boldsymbol{\mu}$ in (2.13) can be reduced (shown in Appendix A.3) to a product of terms of the form $cP(\mathbf{X} \in A)$, where c is a simple algebraic expression and $\mathbf{X}$ has a bivariate normal distribution with easily computable parameters. Thus, the integral can be calculated numerically using readily available techniques. For example, if A is a square region, then the integral is calculated by the function `pmvnorm` in S-Plus version 4.5 for Windows (Mathsoft, Inc.).

The summation over k in (2.13) can be justifiably truncated, for example at n (otherwise it would not make sense to model the data as a cluster process in the first place). This reduces the observed-data likelihood to the summation of a finite number of computable terms, which can thus be maximized (in principle, at least) by an optimization procedure such as the Nelder-Mead simplex method (see Nelder and Mead (1965), Olsson and Nelson (1975) and Press, Flannery, Teukolsky, and Vetterling (1988, section 10.4)).

The summations over $\mathbf{s}$ and $\mathbf{Z}$, however, pose serious problems for even moderately-sized data sets. The number of terms grows astronomically with n. The

cardinality of $\Lambda_n(k)$ is difficult to calculate, but it can be shown by a simple combinatorial argument that the number of ways to choose a collection of non-zero counts is $\binom{n-1}{k-1}$, and so $\#\left(\Lambda_n(k)\right) > \binom{n-1}{k-1}$. Expressions describing the exact number of such terms are unwieldy, but it will suffice to demonstrate that for $k = 2$ and n odd, we have

$$\sum_{\mathbf{s}\in\Lambda_n(2)} \sum_{\mathbf{Z}\in\Omega(\mathbf{s})} 1 = 2^{n-1}.$$

This result follows from the fact that there are 2^n ways to allocate n offspring to 2 ordered clusters, and each such possibility has a redundant duplicate (for n odd, at least) since order should not be counted. Thus, it would appear that computation of even one likelihood value (using a truncation of k) for a moderately-sized data set is not possible anytime in this millenium.

Obviously some allocations are highly unlikely and could be justifiably discarded. However, determining which allocations to discard would require a separate analysis altogether. If such an effort is to be undertaken, there are more suitable methods to consider, most notably the EM algorithm and Markov chain Monte Carlo.

For illustrative purposes, we implement a Nelder-Mead simplex (NMS) algorithm to find a local maximum of the observed-data likelihood (2.13) for a very simple pattern, shown in Figure 2.1. This pattern is a realization of a BVNPCP on the unit square conditional on the number of parent events, parent locations, and cluster counts. The true values of the parameters and other quantities used to create the pattern are: $n = 14$, $k = 2$, $S_1 = 7$, $S_2 = 7$, $\gamma = 3$, $\Psi = .0015$, $\phi = \frac{\pi}{4}$, $\boldsymbol{\mu}_1 = (.33, .67)'$, and $\boldsymbol{\mu}_2 = (.67, .33)'$.

Following the guidelines of Olsson and Nelson (1975) for bounded parameters, we transform $\Phi = \{\rho, \nu, \sigma_{11}, \sigma_{22}, \sigma_{12}\}$ to $\{\log\rho, \log\nu, \log\sigma_{11}, \log\sigma_{22}, 2z(\rho_{12})\}$, where

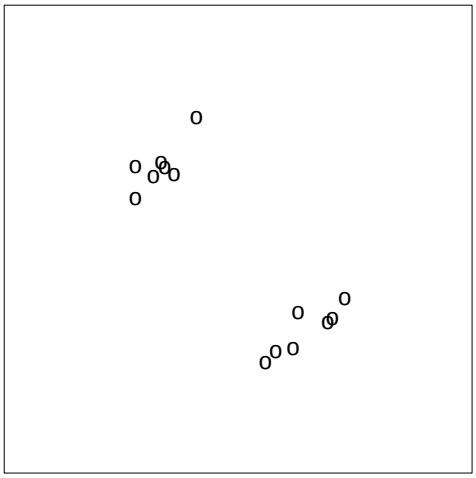


Figure 2.1: Small test pattern to demonstrate Nelder-Mead simplex MLE procedure.

ρ_{12} is the correlation and $z(\rho_{12})$ is defined in Definition 3.5.1, for use in the actual algorithm. The starting simplex (shown in Table 2.1, in a more directly interpretable parameterization) is chosen to be very close to the true value of the parameter vector (to represent a "best case" scenario). Only the values $\{1, 2, 3, 4\}$ are used for k in each likelihood computation. The clusters in the test pattern (Figure 2.1) were intentionally located far from the boundary so that the term $\prod_{i=1}^{k} (P(\mathbf{x}_i \in A))^{\oplus}$ in (2.13) (see equation (A.4) in Appendix A.3) is extremely close to 1 and need not be computed at each iteration.

The NMS algorithm was run until the relative difference between likelihood values at successive iterations was less than 0.0001 (i.e., with a fractional tolerance of 0.0001). The simplex converged in 84 iterations and required 18.27 hours of computation time. Virtually all of the compuatation time was spent in calculation of the likelihood. Table 2.2 shows the resulting parameter estimates, along with

ρ	ν	σ_{11}	σ_{22}	ρ_{12}
3	8	0.003	0.002	0.75
2	8	0.003	0.002	0.75
3	7	0.003	0.002	0.75
3	8	0.002	0.002	0.75
3	8	0.003	0.003	0.75
3	8	0.003	0.002	0.85

Table 2.1: Starting simplex used in NMS algorithm for small test pattern.

true values and also values computed separately using the true allocations $\mathbf{Z}$ and the usual sample correlation coefficient and sample variance. The source code was written in C++, and compiled and run using the same type of computer as discussed at the end of section 7.1.

	ρ	ν	σ_{11}	σ_{22}	ρ_{12}
NMS	2.29651	7.65753	0.003154	0.002622	0.827185
Truth	(k =2)	($S_1 = S_2$ =7)	0.0025	0.0025	0.8
Estimates given $\mathbf{Z}$			0.003150	0.002464	0.824253

Table 2.2: Parameter estimates from NMS algorithm implemented for small test pattern, along with true values and estimates computed using knowledge of $\mathbf{Z}$.

The NMS estimates of σ_{11}, σ_{22} and ρ_{12} are very close to the estimates obtained with knowledge of $\mathbf{Z}$. This is not too surprising since the pattern has clear structure,

and the NMS starting values are close to the true values. Experimentation with other starting simplex values suggests that the algorithm converges to many different local maxima, and many more iterations of the algorithm are usually required.

CHAPTER 3
COMPOSITE EM ANALYSIS

3.1 Mixture Model Specification of the BVNPCP

Define a conditional version of the bivariate normal Poisson cluster process (see Definition 1.1.2) as follows:

Definition 3.1.1 (BVNPCP(A,k,n)) *A* BVNPCP(A,k,n) *with parameter* $\mathbf{\Sigma}$ *is defined as a BVNPCP with cluster shape/scale parameter* $\mathbf{\Sigma}$ *occurring entirely within a region* A*, conditional on the realized values of the number of clusters* (k) *and total number of offspring* (n)*.*

Theorem 3.1.2 *The BVNPCP*(A,k,n) *is completely determined by the following three postulates:*

C1 *The* k *parent events are independently distributed uniformly on* A*, i.e.,*

$$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \sim U(A) \text{ and are independent.}$$

C2 *Each parent* j *produces a random number* S_j *of offspring, where*

$$S_1, \ldots, S_k \sim \text{Mult}\left(n, \frac{1}{k}\mathbf{1}\right).$$

C3 *The positions of the offspring relative to their parents are independently and identically distributed as* $N(\mathbf{0}, \mathbf{\Sigma})$*, conditional on being confined to* A*.*

Proof: **C1** follows from Definition 1.1.2 and the definition of a HPP (see postulate **PP2** in section 4.2 of Diggle (1983)). **C3** follows from Definition 1.1.2, and **C2** follows from Lemma 2.1. □

Note that the parameters ρ and ν of the BVNPCP become irrelevant (and can therefore be treated as absent) in the BVNPCP(A, k, n). Next define the *mixture model specification* of the BVNPCP(A, k, n) (terminology which is justified by Theorem 3.1.4) as follows:

Definition 3.1.3 (mixture model specification of BVNPCP(A, k, n)) *The* mixture model specification of the BVNPCP(A, k, n) *is defined by the following three postulates:*

MM1 *k parent events (also called* components*) are independently distributed uniformly on A, with locations given by*

$$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \overset{\text{i.i.d.}}{\sim} U(A).$$

MM2 *Let $\mathbf{Z}$ be defined as in* (2.4) *and $\mathbf{z}_j$ denote the j^{th} row of $\mathbf{Z}$. Define the notation "$\mathbf{z}_j = q$" to represent*

$$\mathbf{z}_{ji} = \begin{cases} 1, & \text{if } i = q \\ 0, & \text{otherwise.} \end{cases}$$

Allocations are determined independently as

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \overset{\text{i.i.d.}}{\sim} \text{Mult}\left(1, \frac{1}{k}\mathbf{1}\right),$$

i.e.,

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \text{ are independent with } P(\mathbf{z}_j = q) = \frac{1}{k} \quad \forall q \in \{1, \ldots, k\}.$$

MM3 *The positions of the offspring relative to their parents (parentage being determined by $\mathbf{Z}$) are independently and identically distributed as $N(\mathbf{0}, \boldsymbol{\Sigma})$, conditional on being confined to A.*

Theorem 3.1.4 *The BVNPCP(A, k, n) and the mixture model specification of the BVNPCP(A, k, n) (Definitions* 3.1.1 *and* 3.1.3*) are equivalent.*

Proof: Let A, k, and n be given. Both specifications possess the same parameter ($\boldsymbol{\Sigma}$) with the same meaning. The random quantities $\boldsymbol{\mu}$, $\mathbf{Z}$ and $\mathbf{Y}$ completely determine either specification, so it will suffice to prove that their joint distribution is equivalent for the two specifications. By definition, $\mathbf{Z}$ and $\boldsymbol{\mu}$ are independent (of each other and of $\boldsymbol{\Sigma}$), and the distributions of $\boldsymbol{\mu}$ and $\mathbf{Y}|\{\boldsymbol{\mu}, \mathbf{Z}, \boldsymbol{\Sigma}\}$ the same (the latter given by (2.12)), for both specifications. Thus all that remains is to establish agreement on the distribution of $\mathbf{Z}$. Using notation developed in Definition 3.1.3, we have for the mixture model specification

$$\begin{aligned} p(\mathbf{Z}) &= \prod_{j=1}^{n} P(\mathbf{z}_j = q_j) = \frac{1}{k^n} \\ &\quad \text{for any } (q_1, \ldots, q_n) \text{ satisfying } q_j \in \{1, \ldots, k\} \text{ for each } j \in \{1, \ldots, k\}, \\ &\quad \text{i.e., for any } \mathbf{Z}. \end{aligned}$$

For the BVNPCP(A, k, n) we have

$$\begin{aligned} p(\mathbf{Z}) &= p(\mathbf{Z}, S_1, \ldots, S_k) \\ &= p(\mathbf{Z}|S_1, \ldots, S_k)p(S_1, \ldots, S_k) \\ &= \frac{1}{\binom{n}{S_1 \cdots S_k}} \binom{n}{S_1 \cdots S_k} \frac{1}{k^n} \\ &= \frac{1}{k^n}, \end{aligned}$$

and thus the distributions of $\mathbf{Z}$ are equivalent, completing the proof. □

Thus the term "BVNPCP(A, k, n)" will be used, and terminology for mixture models (e.g. "components" and "allocations") will be used when appropriate.

NOTE: Observe that the offspring dispersal distribution of the BVNPCP(A, k, n) is technically a truncated bivariate normal, with the truncation depending on $\boldsymbol{\mu}$. Attempts to account for this would render the model intractable for the types of analyses to be developed. Thus, proceeding as explained in the cautionary note

on page 27, we ignore the truncation and model a common bivariate normal offspring dispersal distribution. As discussed there, the effect of this simplification is expected to be minor, in light of careful choice of data sets used for analysis.

Now we again turn our attention to the problem of estimating $\mathbf{\Sigma}$ for a BVNPCP observed on a region A. Likelihood forms relevant for analysis are derived in section 3.2. Sections 3.3 and 3.4 describe estimation of $\mathbf{\Sigma}$ for a BVNPCP(A, k, n) using the EM algorithm. In section 3.5, a new technique is developed to combine $\mathbf{\Sigma}$ estimates from several different BVNPCP(A, k, n)'s to arrive at a *composite EM* estimate of $\mathbf{\Sigma}$, along with an associated variance estimate. First, the likelihoods associated with the technique are developed.

3.2 Likelihoods Associated with Mixture Models

Consider the BVNPCP(A, k, n) developed in section 3.1. We proceed as is standard in the analysis of mixture models and treat $\boldsymbol{\mu}$ as an unknown parameter instead of a random quantity. The parameters to be estimated are thus $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is treated as a nuisance parameter. Sometimes a mixture model analysis also involves estimation of *mixing proportions*, but in our case the mixing proportions are determined by the BVNPCP(A, k, n) model assumptions to be equal (with value $\frac{1}{k}$, a result of the common Poisson distribution of cluster counts). The notation to follow represents conditioning on k but suppresses dependence on n and A. The distribution of the observed data $\mathbf{Y}$ given the parameters and latent data is called the *classification likelihood* and is given by

$$p\left(\mathbf{Y}\,|k, \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Z}\right) \;=\; \prod_{j=1}^{n} f\left(\mathbf{y}_j | \boldsymbol{\mu}_{\mathbf{z}_j}, \mathbf{\Sigma}\right)$$

$$= \prod_{j=1}^{n} \prod_{i=1}^{k} [f(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})]^{z_{ji}} \tag{3.1}$$

where $f(\cdot|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ denotes the density of $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

The distribution of the observed *and* latent data ($\mathbf{Y}$ and $\mathbf{Z}$) given the parameters is called the *complete-data likelihood* and is given as

$$\begin{aligned} p(\mathbf{Y}, \mathbf{Z} | k, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= p(\mathbf{Y} | k, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z})\, p(\mathbf{Z} | k, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{k^n} \prod_{j=1}^{n} \prod_{i=1}^{k} [f(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})]^{z_{ji}} \end{aligned} \tag{3.2}$$

where $f(\cdot|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ denotes the density of $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

Finally, the marginal distribution of the observed data only given the parameters is called the *mixture likelihood*, or *observed-data likelihood*, and is given by

$$\begin{aligned} p(\mathbf{Y} | k, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{j=1}^{n} p(\mathbf{y}_j | k, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \prod_{j=1}^{n} \left[\int p(\mathbf{y}_j, \mathbf{z}_j | k, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\mathbf{z}_j \right] \\ &= \prod_{j=1}^{n} \left[\int p(\mathbf{y}_j | k, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\mathbf{z}_j | k)\, \mathrm{d}\mathbf{z}_j \right] \\ &= \prod_{j=1}^{n} \left[\sum_{i=1}^{k} p(\mathbf{y}_j | k, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{z}_j = i)\, P(\mathbf{z}_j = i | k) \right] \\ &= \prod_{j=1}^{n} \left[\frac{1}{k} \sum_{i=1}^{k} f(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \right] \\ &= \frac{1}{k^n} \prod_{j=1}^{n} \sum_{i=1}^{k} f(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \end{aligned} \tag{3.3}$$

where $f(\cdot|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ denotes the density of $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

3.3 An EM Algorithm Clustering Approach for Fixed Number of Clusters

The EM algorithm (Dempster, Laird, and Rubin, 1977) is a useful approach to maximum likelihood estimation in the presence of missing (latent) data. Unfortunately, the EM algorithm for normal mixture models (McLachlan and Basford, 1988) cannot be used directly to estimate the parameters of a BVNPCP because the dimension of the parameter space varies with the unknown realized value of k (number of clusters). However, the parameters of a BVNPCP(A, k, n) *can* be estimated. Then parameter estimates and associated variance estimates for different k can be combined, using estimated probabilities that each k is the truth, to form a composite EM estimate and associated variance estimate of the parameters of the underlying BVNPCP. In this section we describe the EM algorithm approach for fixed k. The asymptotic covariance of the estimators is derived in section 3.4, and the technique to combine estimates is developed in section 3.5.

Let the notation $\mathbf{L}(\boldsymbol{\theta}; \mathbf{X})$ denote the log-likelihood function for parameters $\boldsymbol{\theta}$ and data (observed and/or latent) $\mathbf{X}$. The EM algorithm is a technique to utilize the complete-data likelihood to find a solution to the equation

$$\frac{\partial}{\partial(\boldsymbol{\mu}, \boldsymbol{\Sigma})}\mathbf{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{Y}, k) = 0, \tag{3.4}$$

thus obtaining a local maximum of the observed-data (mixture) likelihood (under mild regularity conditions, which are satisfied in our case: see McLachlan and Basford (1988, section 1.6), and for a more general discussion of convergence properties, Wu (1983)). The approach is usually applied when the form of the observed-data likelihood is intractable; in our case, it is easy to express but difficult to maximize by other techniques. McLachlan and Basford (1988, section 1.6) discuss drawbacks of other methods such as Newton-Raphson. The solution to (3.4) is not necessarily

the maximum likelihood estimate (MLE). However, in the case of normal mixture models with common covariance matrix (our situation), the MLE exists and is strongly consistent (i.e., converges to the true value almost surely with increasing sample size) (McLachlan and Basford, 1988, section 2.2). Fraley and Raftery (1998) promote a method (using agglomerative hierarchical clustering, which we describe later in this section) to produce good starting values for the EM algorithm which improve the chances of the solution to (3.4) providing a global maximum. Perhaps a more reliable strategy would be to try many different starting values and compare the mixture likelihood values (since its closed form is available in our case) of the solutions, but construction of a good battery of starting values is a difficult endeavor. We choose the approach of Fraley and Raftery (1998). Regardless, Lehmann (1983, chapter 6) establishes that many desirable properties (such as asymptotic efficiency) hold for any solution to (3.4) under mild regularity conditions, which are not rigorously verified in this thesis but are suspected to hold. Fraley and Raftery (1998) and other authors treat the solution to (3.4) as an MLE, and we follow this precedent.

The EM algorithm consists of two steps, the *E-step* and *M-step*. First the algorithm is initialized with starting values for the latent data (in our case, $\mathbf{Z}$). In the M-step, the complete-data log-likelihood $\mathbf{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{Y}, \mathbf{Z}, k)$ is maximized over the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ conditional on the current values of the latent data. In the E-step, the expectation of $\mathbf{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{Y}, \mathbf{Z}, k)$ over the latent data is computed, conditional on the current values of the parameters. The M-step and E-step are alternated repeatedly until the value of $\mathbf{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{Y}, \mathbf{Z}, k)$ (evaluated at the estimated parameter values and latent data values) converges, as determined by relative differences between iterations.

Starting values for the EM algorithm are obtained as suggested in Fraley and

Raftery (1998) via an agglomerative hierarchical clustering technique, a classification analysis method (see Gordon (1981, section 3.3.1)). The goal is to determine an optimal set of allocations $\mathbf{Z}$ ("optimal" in the sense that the classification likelihood is maximized over a restricted but specially chosen subspace of the latent data domain). Estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which are functions of $\mathbf{Z}$ are constructed to maximize the classification likelihood conditional on $\mathbf{Z}$. For computational reasons, not all possible allocations can be considered. The process is started by treating each data point as a separate cluster. Then two clusters are merged into one, the particular clusters to merge being chosen to maximize the classification likelihood (3.1) over the estimated parameters. This stepwise process is continued until there are k clusters. The allocations at that point are taken as the estimate $\widehat{\mathbf{Z}}$. Estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on $\widehat{\mathbf{Z}}$ could be used, but their properties are not as desirable as those obtained from the EM algorithm (for example, they are not necessarily consistent, as stated in McLachlan and Basford (1988, section 1.12)).

Derivation of $\widehat{\boldsymbol{\Sigma}}(\mathbf{Z}_0)$ and $\widehat{\boldsymbol{\mu}}(\mathbf{Z}_0)$, maximizers of the classification likelihood for fixed $\mathbf{Z}_0$, is now shown. First define the sample mean of a cluster i as

$$\bar{\mathbf{y}}_i = \frac{1}{\sum_{j=1}^{n} z_{ji}} \sum_{j=1}^{n} z_{ji}\mathbf{y}_j. \tag{3.5}$$

Then the classification log-likelihood is

$$\begin{aligned}
&\mathbf{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}; \mathbf{Y}, k\right) \\
&= \log \prod_{i=1}^{k} \prod_{j=1}^{n} \left[f(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\right]^{z_{ji}} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{n} z_{ji} \log f(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{n} z_{ji} \left[-\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \left(-n\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right) \\
&= \left(-n\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\mathrm{tr}\left[\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)\right] \\
&= \left(-n\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\mathrm{tr}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\right] \\
&= \left(-n\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\mathrm{tr}\left[\sum_{i=1}^{k}\left\{\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right)' + \right.\right. \\
&\qquad \left.\left.\sum_{j=1}^{n} z_{ji}\left(\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i\right)\left(\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i\right)'\right\}\boldsymbol{\Sigma}^{-1}\right] \\
&= \left(-n\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\mathrm{tr}\left[\left\{\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right)'\right\}\boldsymbol{\Sigma}^{-1}\right] - \\
&\qquad \frac{1}{2}\sum_{i=1}^{k}\mathrm{tr}\left[\left(\sum_{j=1}^{n} z_{ji}\right)\left(\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i\right)\left(\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\right] \\
&\qquad\quad \left(\text{since } \sum_{i=1}^{k} z_{ji} = 1 \quad \forall j\right) \\
&= \left(-n\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}|\right) - \frac{n}{2}\mathrm{tr}\left[\left\{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right)'\right\}\boldsymbol{\Sigma}^{-1}\right] - \\
&\qquad \frac{1}{2}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} z_{ji}\right)\left(\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i\right). \qquad (3.6)
\end{aligned}
$$

The maximum of the classification likelihood for a fixed $\mathbf{Z}_0$ can then be computed most conveniently as

$$\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \mathbf{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}_0; \mathbf{Y}, k\right) = \max_{\boldsymbol{\Sigma}}\left[\max_{\boldsymbol{\mu}} \mathbf{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}_0; \mathbf{Y}, k\right)\right].$$

From (3.6) it is clear that $\mathbf{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}_0; \mathbf{Y}, k\right)$ is maximized over $\boldsymbol{\mu}$ uniquely by

$$\widehat{\boldsymbol{\mu}}(\mathbf{Z}_0) = \left(\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_k\right)' \quad \text{computed at } \mathbf{Z}_0. \qquad (3.7)$$

By Lemma 3.2.2 of Anderson (1984), we have that

$$\mathbf{L}\left(\widehat{\boldsymbol{\mu}}(\mathbf{Z}_0), \boldsymbol{\Sigma}, \mathbf{Z}_0; \mathbf{Y}, k\right) = \left(-n\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}|\right) -$$

$$\frac{n}{2}\mathrm{tr}\left[\left\{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)'\right\}\boldsymbol{\Sigma}^{-1}\right]$$

is maximized over $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{\Sigma}}(\mathbf{Z}_0) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)' \quad (3.8)$$

computed at $\mathbf{Z}_0$.

Thus we have

$$\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \mathbf{L}\left(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{Z}_0;\mathbf{Y},k\right) = -n\log(2\pi) - \frac{n}{2}\log\left|\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)'\right| - n,$$

and so the agglomerative hierarchical clustering algorithm chooses a cluster merge at each stage to minimize $\left|\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)'\right|$. This result is often referred to as the *determinant criterion* and is due to Friedman and Rubin (1967). Fraley (1999) develops efficient techniques to perform the clustering, which are implemented in the MCLUST/EMCLUST software (Fraley, 1998).

The allocations $\widehat{\mathbf{Z}}_0$ given by the agglomerative hierarchical clustering algorithm are then used as the starting values for the EM algorithm, which involves maximization (over $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) and integration (over the conditional distribution of $\mathbf{Z}$ given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) of the *complete-data* log-likelihood (see (3.2)). However, note that (3.2) is simply a constant multiple of the classification likelihood (3.1) for fixed k and n. Even though the E-step produces estimates $\widehat{\mathbf{Z}}$ that are not integer-valued, (3.5) and (3.6) are still valid when $z_{ji} \in \{0,1\}$ is replaced with $\widehat{z}_{ji} \in (0,1)$. Thus the M-step of the EM algorithm is given by (3.7) and (3.8). As far as the E-step is concerned, since $\mathbf{L}\left(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{Z};\mathbf{Y},k\right)$ is a linear function in $\mathbf{Z}$ (see (3.6)), its conditional expectation over $\mathbf{Z}$ given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is determined easily from the conditional distribution of $\mathbf{Z}$ given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which is derived below.

First note that

$$p\left(\mathbf{Z}\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right) = \frac{p\left(\mathbf{Y}\,|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right)p\left(\mathbf{Z}\,|\boldsymbol{\mu},\boldsymbol{\Sigma},k\right)}{p\left(\mathbf{Y}\,|\boldsymbol{\mu},\boldsymbol{\Sigma},k\right)}$$

$$\propto \quad p\left(\mathbf{Y}\,|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right)$$
$$\text{since } p\left(\mathbf{Z}\,|\boldsymbol{\mu},\boldsymbol{\Sigma},k\right)=\frac{1}{k^n}.$$

Also,

$$\begin{aligned} p\left(\mathbf{Y}\,|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right) &= \prod_{j=1}^{n}\prod_{i=1}^{k}\left[f(\mathbf{y}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma})\right]^{z_{ji}} \\ &= \prod_{j=1}^{n}\left[\prod_{i=1}^{k}\mathrm{I}(\mathbf{z}_j=i)\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\right\}\right], \\ &\qquad \text{(where I}(\cdot)\text{ is the indicator function)} \end{aligned}$$

and so $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are independent (by factorization), and for each $j\in\{1,\ldots,n\}$

$$P\left(\mathbf{z}_j=i\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right) \quad\propto\quad \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\right\}$$
$$\text{for } i\in\{1,\ldots,k\}.$$

Thus

$$P\left(\mathbf{z}_j=i\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\right\}}{\sum_{q=1}^{k}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_q\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_q\right)\right\}} \qquad (3.9)$$
$$\text{independently for each } j\in\{1,\ldots,n\}.$$

Note that $\sum_{i=1}^{k}P\left(\mathbf{z}_j=i\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right)=1 \quad \forall j$. Since

$$E\left(z_{ji}\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right)=P\left(\mathbf{z}_j=i\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right),$$

the E-step is solved.

Putting it altogether, we have the form of the normal mixture model EM algorithm for fixed k, as described in Fraley and Raftery (1998):

Algorithm 3.3.1 (EM Algorithm for Fixed k) *Estimates for the parameters $(\boldsymbol{\Sigma},\boldsymbol{\mu})$ of a BVNPCP(A,k,n) are computed as follows, for a given* tolerance ϵ*:*

Step 1: Agglomerative Hierarchical Clustering *Perform agglomerative hierarchical clustering, choosing the clusters to merge at each stage by minimizing*

$$\left|\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n}z_{ji}\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)'\right|,$$

to yield an initial allocation estimate $\widehat{\mathbf{Z}}(0)$.

Step 2: M-step *Given* $\widehat{\mathbf{Z}}(t) = \{\widehat{z}_{ji}(t)\}$, *compute*

$$\begin{aligned} \widehat{\boldsymbol{\mu}}(t+1) &= (\widehat{\boldsymbol{\mu}}(t+1)_1, \ldots, \widehat{\boldsymbol{\mu}}(t+1)_k) \\ &\quad \textit{where } \widehat{\boldsymbol{\mu}}(t+1)_i = \frac{1}{\sum_{j=1}^{n} \widehat{z}_{ji}(t)} \sum_{j=1}^{n} \widehat{z}_{ji}(t)\mathbf{y}_j \end{aligned}$$

and

$$\widehat{\boldsymbol{\Sigma}}(t+1) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} \widehat{z}_{ji}(t) \left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right) \left(\mathbf{y}_j - \bar{\mathbf{y}}_i\right)' \ .$$

Step 3: Convergence Check *If* $t > 0$ *and*

$$\frac{\left|\mathbf{L}\left(\widehat{\boldsymbol{\mu}}(t+1), \widehat{\boldsymbol{\Sigma}}(t+1); \mathbf{Y}, \widehat{\mathbf{Z}}(t), k\right) - \mathbf{L}\left(\widehat{\boldsymbol{\mu}}(t), \widehat{\boldsymbol{\Sigma}}(t); \mathbf{Y}, \widehat{\mathbf{Z}}(t), k\right)\right|}{\left|\mathbf{L}\left(\widehat{\boldsymbol{\mu}}(t+1), \widehat{\boldsymbol{\Sigma}}(t+1); \mathbf{Y}, \widehat{\mathbf{Z}}(t), k\right)\right|} < \epsilon,$$

then go to Step 5. Otherwise proceed to Step 4.

Step 4: E-step *Given* $(\widehat{\boldsymbol{\mu}}(t+1), \widehat{\boldsymbol{\Sigma}}(t+1))$, *compute* $\widehat{\mathbf{Z}}(t+1)$ *according to*

$$\widehat{z}_{ji}(t+1) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}(t+1)_i\right)' \left[\widehat{\boldsymbol{\Sigma}}(t+1)\right]^{-1} \left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}(t+1)_i\right)\right\}}{\sum_{q=1}^{k} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}(t+1)_q\right)' \left[\widehat{\boldsymbol{\Sigma}}(t+1)\right]^{-1} \left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}(t+1)_q\right)\right\}},$$

increment t *by 1, and go to Step 2.*

Step 5: Termination *Set the final estimates to* $\widehat{\boldsymbol{\Sigma}}^{(k)} = \widehat{\boldsymbol{\Sigma}}(t+1)$ *and* $\widehat{\boldsymbol{\mu}}^{(k)} = \widehat{\boldsymbol{\mu}}(t+1)$. *Estimates of the expected allocations can be taken from the last iteration of the E-step to give* $\widehat{\mathbf{Z}}^{(k)} = \widehat{\mathbf{Z}}(t)$.

3.4 Computation of Approximate Variance of Parameter Estimates for Fixed k

As mentioned in section 3.3, estimates $(\widehat{\boldsymbol{\Sigma}}^{(k)}, \widehat{\boldsymbol{\mu}}^{(k)})$ from Algorithm 3.3.1 are not guaranteed to be MLE's of the observed-data likelihood. However, Theorem 4.1 of Chapter 6 in Lehmann (1983) can be used to obtain the asymptotic distribution of $(\widehat{\boldsymbol{\Sigma}}^{(k)}, \widehat{\boldsymbol{\mu}}^{(k)})$, since they are solutions to (3.4). The regularity conditions of Lehmann's theorem are not rigorously verified in this thesis, but are suspected to

hold. However, since many authors (for example, McLachlan and Basford (1988, sections 1.9 and 2.4)) advocate the use of the asymptotic distribution implied by the theorem for computation of approximate variance in our situation, and for lack of a better alternative, that is how we proceed.

Before stating the asymptotic distribution, some terminology regarding information matrices is introduced. For simplicity of notation, let

$$\boldsymbol{\theta} = \{\boldsymbol{\sigma}, \boldsymbol{\mu}\} = (\sigma_{11}, \sigma_{22}, \sigma_{12}, \mu_{11}, \mu_{12}, \ldots, \mu_{k1}, \mu_{k2})'$$

and

$$\widehat{\boldsymbol{\theta}}^{(k)} = \left\{\widehat{\boldsymbol{\sigma}}^{(k)}, \widehat{\boldsymbol{\mu}}^{(k)}\right\} = \left(\hat{\sigma}_{11}^{(k)}, \hat{\sigma}_{22}^{(k)}, \hat{\sigma}_{12}^{(k)}, \hat{\mu}_{11}^{(k)}, \hat{\mu}_{12}^{(k)}, \ldots, \hat{\mu}_{k1}^{(k)}, \hat{\mu}_{k2}^{(k)}\right)'$$

Definition 3.4.1 (Observed Information Matrix) *The observed information matrix for the BVNPCP(A, k, n) is given by*

$$\mathrm{I}_{\mathrm{o}}\left(\boldsymbol{\theta} \middle| \mathbf{Y}, k\right) = \left\{\frac{-\partial^2 \mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, k\right)}{\partial \boldsymbol{\theta}^2}\right\},$$

the negative of the Hessian matrix of the observed-data (mixture) likelihood.

Definition 3.4.2 (Complete Information Matrix) *The complete information matrix for the BVNPCP(A, k, n) is given by*

$$\mathrm{I}_{\mathrm{c}}\left(\boldsymbol{\theta} \middle| \mathbf{Y}, \mathbf{Z}, k\right) = \left\{\frac{-\partial^2 \mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k\right)}{\partial \boldsymbol{\theta}^2}\right\},$$

the negated Hessian matrix of the complete-data likelihood.

Definition 3.4.3 (Missing Information Matrix) *The missing information matrix for the BVNPCP(A, k, n) is given by*

$$\mathrm{I}_{\mathrm{m}}\left(\boldsymbol{\theta} \middle| \mathbf{Y}, k\right) = E_{\mathbf{Z}}\left\{\frac{-\partial^2 \log p\left(\mathbf{Z} \middle| \boldsymbol{\theta}, \mathbf{Y}, k\right)}{\partial \boldsymbol{\theta}^2} \middle| \boldsymbol{\theta}, \mathbf{Y}, k\right\}.$$

Definition 3.4.4 (Fisher Information Matrix) *The Fisher (or expected) information matrix for the BVNPCP(A, k, n) is given by*

$$\mathrm{I}\left(\boldsymbol{\theta} | k\right) = E_{\mathbf{Y}}\left[\mathrm{I}_{\mathrm{o}}\left(\boldsymbol{\theta} \middle| \mathbf{Y}, k\right)\right].$$

If the regularity conditions are assumed to hold, then Theorem 4.1 of Chapter 6

in Lehmann (1983) yields

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}) \xrightarrow{D} N\left(\mathbf{0}, [\mathrm{I}\,(\boldsymbol{\theta}|k)]^{-1}\right),$$

where "$\xrightarrow{D}$" denotes convergence in distribution.

The Fisher information $\mathrm{I}\,(\boldsymbol{\theta}|k)$ is intractable to work with in our situation, so as suggested by McLachlan and Basford (1988, sections 1.9 and 2.4), we use the observed information $\mathrm{I_o}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k)$ calculated at the EM estimate $\widehat{\boldsymbol{\theta}}^{(k)}$. Thus we use the approximation

$$\widehat{\boldsymbol{\theta}}^{(k)} \overset{\bullet}{\sim} N\left(\boldsymbol{\theta}, [\mathrm{I_o}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k)]^{-1}\big|_{\widehat{\boldsymbol{\theta}}^{(k)}}\right), \tag{3.10}$$

where "$\overset{\bullet}{\sim}$" denotes "is approximately distributed as."

McLachlan and Basford (1988), however, do not compute the observed information matrix $\mathrm{I_o}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k)$, directly or indirectly. They instead use an approximation, the accuracy of which is unknown. Dasgupta and Raftery (1998) analyze a similar model and suggest the use of an approach such as the supplemented EM algorithm (Meng and Rubin, 1991) (which also approximates the observed information matrix) to obtain variance estimates.

However, $\mathrm{I_o}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k)$ *can* be computed in closed form, and that is how we will obtain variance estimates. Derivatives of the observed-data likelihood are difficult to calculate, but two results due to Louis (1982), which we state in the Lemma below, allow us to work with the complete-data likelihood:

Lemma 3.4.5 (Louis) *For arbitrary* $(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}, k)$, *if* $\mathrm{I_o}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k)$, $\mathrm{I_c}\,(\boldsymbol{\theta}|\,\mathbf{Y}, \mathbf{Z}, k)$ *and* $\mathrm{I_m}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k)$ *exist then*

$$\mathrm{I_m}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k) = \mathrm{Var}_{\mathbf{Z}}\left\{\frac{\partial \mathbf{L}\,(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k)}{\partial \boldsymbol{\theta}}\,\middle|\,\boldsymbol{\theta}, \mathbf{Y}, k\right\} \tag{3.11}$$

and

$$\mathrm{I_o}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k) = E_{\mathbf{Z}}\left[\mathrm{I_c}\,(\boldsymbol{\theta}|\,\mathbf{Y}, \mathbf{Z}, k)|\,\boldsymbol{\theta}, \mathbf{Y}, k\right] - \mathrm{I_m}\,(\boldsymbol{\theta}|\,\mathbf{Y}, k). \tag{3.12}$$

Proof: See Louis (1982).

The result (3.12) is often called the *Missing Information Principle*. Due to the construction of Algorithm 3.3.1, we also have

$$E_{\mathbf{Z}}\left\{\left.\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right|\boldsymbol{\theta},\mathbf{Y},k\right\}\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}}=\left\{\left.\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right|\boldsymbol{\theta},\mathbf{Y},k\right\}\Big|_{\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\mathbf{Z}}^{(k)}}=\mathbf{0},$$

and so

$$\begin{aligned}
&\mathrm{Var}_{\mathbf{Z}}\left\{\left.\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right|\boldsymbol{\theta},\mathbf{Y},k\right\}\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}}\\
&= E_{\mathbf{Z}}\left[\left\{\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right\}'\Big|\boldsymbol{\theta},\mathbf{Y},k\right]\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}}-\\
&\quad\left[E_{\mathbf{Z}}\left\{\left.\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right|\boldsymbol{\theta},\mathbf{Y},k\right\}\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}}\right]\left[E_{\mathbf{Z}}\left\{\left.\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right|\boldsymbol{\theta},\mathbf{Y},k\right\}\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}}\right]'\\
&= E_{\mathbf{Z}}\left[\left\{\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right\}'\Big|\boldsymbol{\theta},\mathbf{Y},k\right]\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}} \qquad (3.13)
\end{aligned}$$

Using Lemma 3.4.5 and (3.13), we thus approximate the asymptotic distribution of $\widehat{\boldsymbol{\theta}}^{(k)}$ as

$$\widehat{\boldsymbol{\theta}}^{(k)} \overset{\bullet}{\sim} N\left(\boldsymbol{\theta}, \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}^{(k)})\right) \qquad (3.14)$$

where

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}^{(k)}) &= \left[E_{\mathbf{Z}}\left\{\left.\frac{-\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}^2}\right|\boldsymbol{\theta},\mathbf{Y},k\right\}\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}} - \qquad (3.15)\right.\\
&\qquad\left.E_{\mathbf{Z}}\left[\left\{\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\boldsymbol{\theta}}\right\}'\Big|\boldsymbol{\theta},\mathbf{Y},k\right]\Big|_{\widehat{\boldsymbol{\theta}}^{(k)}}\right]^{-1}.
\end{aligned}$$

Detailed expressions for (3.15) for the BVNPCP(A,k,n) are derived in Appendix A.4.

3.5 Composite EM Analysis of the BVNPCP

The results of sections 3.3 and 3.4 give us estimates of $\boldsymbol{\Sigma}$ for a BVNPCP(A,k,n), along with their approximate variances. In this section, we use these quantities from BVNPCP(A,k,n)'s for a range of k's to construct an overall, or *composite EM*, estimate of $\boldsymbol{\Sigma}$ (along with approximate variance matrix) that accounts for the uncertainty in estimation of k. The approximate asymptotic distribution of the composite EM estimate is derived and used to construct isotropy tests and compute

confidence regions for various parameterizations and components of $\mathbf{\Sigma}$.

3.5.1 Derivation of the Estimator

Adopting a Bayesian perspective, we can consider each BVNPCP(A, k, n) as a candidate model for the observed spatial point pattern $\mathbf{Y}$, the models being indexed by k. Instead of attempting to choose one "correct" model, we will implement a *Bayesian model averaging* scheme. The first step in such a scheme is obtaining estimated model probabilities

$$\left\{\widehat{P}(\text{number of clusters } = k|\mathbf{Y})\right\} = \{\widehat{p}^{(k)}\}$$

for a reasonable range $\{k_{\text{lo}}, \ldots, k_{\text{hi}}\}$ of possible k (note the suppression of dependence on $\mathbf{Y}$ in the notation).

Fraley and Raftery (1998, section 2.4) promote the use of the *Bayesian Information Criterion (BIC)* (Schwarz, 1978) to assess model probabilities for our situation. They state "although the regularity conditions for BIC do not hold for mixture models, there is considerable theoretical and practical support for its use in this context," citing Leroux (1992); Roeder and Wasserman (1997); Dasgupta and Raftery (1998); Campbell et al. (1998); Mukerjee et al. (1998). The BIC for a particular k is defined as

$$\begin{aligned} BIC_k^{\text{EM}} &= 2\mathbf{L}\left(\widehat{\boldsymbol{\theta}}^{(k)}; \mathbf{Y}, k\right) - (\#\text{parameters})\log(n) && (3.16) \\ &= 2\mathbf{L}\left(\widehat{\boldsymbol{\theta}}^{(k)}; \mathbf{Y}, k\right) - (2k+3)\log(n), && (3.17) \end{aligned}$$

since we have 2 parameters for each cluster mean $\boldsymbol{\mu}_i$ and 3 for $\boldsymbol{\sigma}$. (Note: there is variation in the literature regarding the use of $\log(n)$ vs. $\log(rn)$, where r is the dimension of an observation $\mathbf{y}_i$, but we choose $\log(n)$ as in Fraley and Raftery (1998)). The BIC can be used to calculate approximate Bayes factors (see Kass and Raftery (1995)), and also posterior model probabilities for given prior probabilities

of each candidate model. We assign equal prior probabilities to models in a range $\{k_{\rm lo}, \ldots, k_{\rm hi}\}$ and compute estimated model probabilities, like in Raftery (1993, equation 11), as

$$\widehat{p}^{(k)} = \frac{\exp\left(\frac{1}{2}BIC_k^{\rm EM}\right)}{\sum_{q=k_{\rm lo}}^{k_{\rm hi}} \exp\left(\frac{1}{2}BIC_q^{\rm EM}\right)} \tag{3.18}$$

Next we obtain parameter estimates and variance approximations from BVN-PCP(A, k, n)'s for $k \in \{k_{\rm lo}, \ldots, k_{\rm hi}\}$ using the methods of sections 3.3 and 3.4. Instead of considering $\boldsymbol{\theta}$ as a parameter, we now follow the Bayesian paradigm and consider it a random vector $\boldsymbol{\theta}^\star$ with a distribution of its own. Using Result 8(iii) of Berger (1985, section 4.7.8), we have

$$\boldsymbol{\theta}^\star \overset{\bullet}{\sim} N\left(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\rm Var}(\widehat{\boldsymbol{\theta}}^{(k)})\right), \tag{3.19}$$

where $\widehat{\boldsymbol{\theta}}^{(k)}$is the EM estimate and $\widehat{\rm Var}(\widehat{\boldsymbol{\theta}}^{(k)})$ is given by (3.15).

For investigation of isotropy we are only interested in estimation of $\boldsymbol{\sigma}$, and so we extract the appropriate subvector $\boldsymbol{\sigma}^\star$ from $\boldsymbol{\theta}^\star$, subvector $\widehat{\boldsymbol{\sigma}}^{(k)}$ from $\widehat{\boldsymbol{\theta}}^{(k)}$, and submatrix $\widehat{\rm Var}(\widehat{\boldsymbol{\sigma}}^{(k)})$ from $\widehat{\rm Var}(\widehat{\boldsymbol{\theta}}^{(k)})$ in (3.19) to obtain

$$\boldsymbol{\sigma}^\star \overset{\bullet}{\sim} N\left(\widehat{\boldsymbol{\sigma}}^{(k)}, \widehat{\rm Var}(\widehat{\boldsymbol{\sigma}}^{(k)})\right). \tag{3.20}$$

A generalization of equations 14 and 15 of Raftery (1993) from scalar to vector form gives estimates of $E(\boldsymbol{\sigma}^\star|\mathbf{Y})$ and ${\rm Var}(\boldsymbol{\sigma}^\star|\mathbf{Y})$ as

$$\widehat{E(\boldsymbol{\sigma}^\star|\mathbf{Y})} = \sum_{k=k_{\rm lo}}^{k_{\rm hi}} \widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)} \tag{3.21}$$

and

$$\begin{aligned}\widehat{{\rm Var}(\boldsymbol{\sigma}^\star|\mathbf{Y})} \;=\; & \sum_{k=k_{\rm lo}}^{k_{\rm hi}} \left\{\widehat{\rm Var}(\widehat{\boldsymbol{\sigma}}^{(k)}) + \left[\widehat{\boldsymbol{\sigma}}^{(k)}\right]\left[\widehat{\boldsymbol{\sigma}}^{(k)}\right]'\right\}\widehat{p}^{(k)} \;-\; \\ & \left[\sum_{k=k_{\rm lo}}^{k_{\rm hi}} \widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)}\right]\left[\sum_{k=k_{\rm lo}}^{k_{\rm hi}} \widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)}\right]'.\end{aligned} \tag{3.22}$$

This suggests an estimator

$$\widehat{\boldsymbol{\sigma}} = \widehat{E(\boldsymbol{\sigma}^\star|\mathbf{Y})} \tag{3.23}$$

with approximate variance given by

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\sigma}}) = \widehat{\mathrm{Var}(\boldsymbol{\sigma}^\star|\mathbf{Y})}. \tag{3.24}$$

Although (3.21) suggests that the asymptotic distribution of $\widehat{\boldsymbol{\sigma}}$ is a *mixture* of normal distributions, we will approximate the distribution as multivariate normal in constructing confidence regions and test statistics. This is somewhat reasonable in cases where a small collection of nearby k's are dominant in estimated model probabilities, but certainly less reasonable in situations with many and/or disparate supported values of k. In our analyses, the former situation is more common, but caution is nevertheless advised.

So we will take $\widehat{\boldsymbol{\sigma}}$ of (3.23) as our composite EM estimator of $\boldsymbol{\sigma}$ and approximate its asymptotic distribution from (3.14) and (3.21) – (3.24) as

$$\widehat{\boldsymbol{\sigma}} \overset{\bullet}{\sim} N\left(\boldsymbol{\sigma}, \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\sigma}})\right). \tag{3.25}$$

Note that $\widehat{\mathrm{Var}(\boldsymbol{\sigma}^\star|\mathbf{Y})}$ in (3.22) can be re-written as

$$\left\{\sum_{k=k_{\mathrm{lo}}}^{k_{\mathrm{hi}}} \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\sigma}}^{(k)})\widehat{p}^{(k)}\right\} \quad + \tag{3.26}$$

$$\left\{\left[\sum_{k=k_{\mathrm{lo}}}^{k_{\mathrm{hi}}} \left(\widehat{\boldsymbol{\sigma}}^{(k)}\right)\left(\widehat{\boldsymbol{\sigma}}^{(k)}\right)' \widehat{p}^{(k)}\right] - \left[\sum_{k=k_{\mathrm{lo}}}^{k_{\mathrm{hi}}} \widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)}\right]\left[\sum_{k=k_{\mathrm{lo}}}^{k_{\mathrm{hi}}} \widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)}\right]'\right\} \tag{3.27}$$

and that

$$\mathrm{diag}\left\{\left[\sum_{k=k_{\mathrm{lo}}}^{k_{\mathrm{hi}}} \left(\widehat{\boldsymbol{\sigma}}^{(k)}\right)\left(\widehat{\boldsymbol{\sigma}}^{(k)}\right)' \widehat{p}^{(k)}\right] - \left[\sum_{k=k_{\mathrm{lo}}}^{k_{\mathrm{hi}}} \widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)}\right]\left[\sum_{k=k_{\mathrm{lo}}}^{k_{\mathrm{hi}}} \widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)}\right]'\right\} \geq \mathbf{0}$$

since (3.27) is positive semidefinite, with equality holding only in the trivial case that all $\widehat{\boldsymbol{\sigma}}^{(k)}\widehat{p}^{(k)}$ are equal. The fact that (3.27) is positive semidefinite is evident from the consideration of $\{\widehat{\boldsymbol{\sigma}}^{(k)}\}$ as observations from some (arbitrary) distribution and (3.27) as a weighted sample variance estimate for this distribution.

Therefore the variances on the diagonal of $\widehat{\mathrm{Var}(\boldsymbol{\sigma}^\star|\mathbf{Y})}$ are at least as large as

pooled variance estimates using the weighted average

$$\sum_{k=k_{\rm lo}}^{k_{\rm hi}} \widehat{\rm Var}(\widehat{\boldsymbol{\sigma}}^{(k)})\widehat{p}^{(k)},$$

and so the composite EM method *inflates* variance estimates which would be used in a naive combination of separate analyses by k. This is appropriate since the uncertainty in estimation of k should be accounted for.

In our analyses we found that the diagonal elements of (3.27) were sometimes smaller and sometimes larger than those of (3.26).

3.5.2 Applications for Anisotropy Estimation and Testing

The composite EM estimator developed in section 3.5.1 can be used to produce a 3-dimensional confidence region for $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{22}, \sigma_{12})'$, utilizing approximate normality. It can also be used to obtain more interpretable confidence regions and intervals in terms of the components of and other parameterizations of $\boldsymbol{\sigma}$.

Let $f(\boldsymbol{\sigma})$ be any function of interest. Then the multivariate Δ-method yields

$$f(\widehat{\boldsymbol{\sigma}}) \overset{\bullet}{\sim} N\left(f(\boldsymbol{\sigma}), J\widehat{\rm Var}(\widehat{\boldsymbol{\sigma}})J'\right), \tag{3.28}$$

where $\widehat{\boldsymbol{\sigma}}$ and $\widehat{\rm Var}(\widehat{\boldsymbol{\sigma}})$ are the composite EM estimate and its approximate variance, and

$$J = \left\{\frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}}\right\}\Big|_{\widehat{\boldsymbol{\sigma}}} = \left[\frac{\partial f(\boldsymbol{\sigma})}{\partial \sigma_{11}}, \frac{\partial f(\boldsymbol{\sigma})}{\partial \sigma_{22}}, \frac{\partial f(\boldsymbol{\sigma})}{\partial \sigma_{12}}\right]\Big|_{\widehat{\boldsymbol{\sigma}}} \tag{3.29}$$

Useful choices of $f(\boldsymbol{\sigma})$ are discussed below.

To facilitate comparisons with results from Markov chain Monte Carlo (in which careful choice of parameterization is vital for some methods), we focus on "normalized" versions of components of $\boldsymbol{\sigma}$ (meaning parameterizations that are likely to achieve the best approximation to normality). First we define a useful transformation of the correlation coefficient which improves approximate normality:

Definition 3.5.1 (Fisher's z-transformation) *Define the theoretical correlation coefficient as*

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}. \tag{3.30}$$

Then the Fisher's z-transformation of ρ_{12} *is defined as*

$$z(\rho_{12}) = \frac{1}{2}\log\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right).$$

Componentwise confidence intervals are constructed for the following choices of $f(\boldsymbol{\sigma})$:

$$\log\sigma_{11}, \quad \log\sigma_{22}, \quad z(\rho_{12}), \quad \log\gamma, \quad \log\Psi, \quad \phi, \tag{3.31}$$

the first 3 involving the regular parameterization (see Definition 1.1.9), and the last 3 involving the anisotropy parameterization (see Definition 1.1.10). Detailed expressions of the Jacobians J (3.29) for these parameters are shown in Appendix A.5.

Note that the null value of $\log\gamma$ in the case of isotropy is 0, which is on the boundary of the parameter space. Thus an isotropy test utilizing an estimate of $\log\gamma$ is subject to criticism. However, a confidence interval for $\log\gamma$ is still meaningful, especially in cases where there is clear anisotropy, the strength of which one wishes to assess. There does not appear to be an adequate adjustment to the definition of γ, or its estimation, that will render it inarguably acceptable for isotropy testing. Perhaps there should not be, as the test of isotropy is really a 2-degree-of-freedom test, comparing the situations ($\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$) and ($\boldsymbol{\Sigma}$ arbitrary).

A more appropriate test of isotropy is the simultaneous assessment of the difference in variances and the covariance, the normalized version of which is

$$\boldsymbol{\sigma}^c = (\log\sigma_{11} - \log\sigma_{22},\ z(\rho_{12}))'. \tag{3.32}$$

A well-defined isotropy test is a test of $\boldsymbol{\sigma}^c = \mathbf{0}$. A 2-dimensional confidence region for $\boldsymbol{\sigma}^c$ can be plotted, with its deviation from the null value $\mathbf{0}$ suggesting the nature and extent of anisotropy. The components of the Jacobian J (3.29) for $\boldsymbol{\sigma}^c$ are given

in Appendix A.5.

By (3.28), an approximate $100(1-\alpha)\%$ confidence interval for a scalar function $f(\boldsymbol{\sigma})$ is computed as

$$f(\widehat{\boldsymbol{\sigma}}) \pm z_{\left(1-\frac{\alpha}{2}\right)} \left[J\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\sigma}})J'\right]^{\frac{1}{2}}, \tag{3.33}$$

where $\widehat{\boldsymbol{\sigma}}$ is the composite EM parameter estimate, $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\sigma}})$ is its approximate variance, J is as defined in (3.29), and $z_{\left(1-\frac{\alpha}{2}\right)}$ is the $(1-\frac{\alpha}{2})^{\text{th}}$ quantile of the standard normal distribution.

Similarly, an approximate $100(1-\alpha)\%$ confidence region for $\boldsymbol{\sigma}^c$ is

$$[\widehat{\boldsymbol{\sigma}}^c]' \left[\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\sigma}}^c)\right]^{-1} [\widehat{\boldsymbol{\sigma}}^c] \leq \chi^2_2(1-\alpha), \tag{3.34}$$

where $\boldsymbol{\sigma}^c = (\log\hat{\sigma}_{11} - \log\hat{\sigma}_{22},\ z(\rho_{12}))'$ is the composite EM parameter estimate, $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\sigma}}^c)$ is its approximated variance according to (3.28), and $\chi^2_2(1-\alpha)$ is the $(1-\alpha)^{\text{th}}$ quantile of the Chi-square distribution with 2 degrees of freedom. An isotropy test can be conducted by computing the left-hand side of (3.34), say X_{obs}, and obtaining a p-value of $P(X > X_{obs})$ where $X \sim \chi^2_2$.

CHAPTER 4
RJMCMC ALGORITHM DESIGN

4.1 Motivation

While the composite EM technique of Chapter 3 offers an appealing method to account for the unknown number of clusters k in estimation of $\mathbf{\Sigma}$ for a BVNPCP, it is subject to criticism on the grounds of the questionable accuracy of the BIC used to estimate model probabilities. An alternative modeling strategy is to adopt a *fully* Bayesian perspective by specifying the BVNPCP as a Bayesian hierarchical model including all unknowns, parameters and latent data alike, and assigning a distribution to each unknown quantity. In particular, k can be allowed to vary in the hierarchical model, permitting uncertainty of its true value to be inherently accounted for in the estimation of quantities of interest (in our case, $\mathbf{\Sigma}$).

As in Chapter 3, we will think of the BVNPCP in terms of a mixture model. Markov chain Monte Carlo (MCMC) methods (Hastings, 1970; Metropolis et al., 1953; Geman and Geman, 1984; Gelfand and Smith, 1990) have been successfully applied to problems in finite mixture analysis. Diebolt and Robert (1994); Lavine and West (1992); Bensmail, Celeux, Raftery, and Robert (1995) and other authors have developed Gibbs sampling approaches to analyze univariate and multivariate normal mixture models for *fixed* k. In such approaches, k is treated as a model indicator, and any of a number of available model selection techniques utilizing marginal likelihood estimates are applied to in an attempt to choose the "best" k.

Examples of such methods applied recently include the Laplace-Metropolis estimator (Raftery, 1995), importance-sampling-based estimators (Newton and Raftery, 1994), the Schwarz BIC criterion (Schwarz, 1978) and Approximate Weight of Evidence (Banfield and Raftery, 1993).

Typical MCMC methods (e.g. the Gibbs sampler and Metropolis-Hastings algorithm) apply only to situations in which the dimension of the parameter space is fixed. However, we wish to model k as a parameter, in which case the dimension of the parameter space is *not* fixed (e.g., the dimensionality of $\boldsymbol{\mu}$ varies with k). Stephens (1997) develops a generalization of the Metropolis-Hastings algorithm based on the Markov spatial birth-and-death process (Geyer and Møller, 1994) to allow for varying parameter space dimension, applying it to multivariate normal mixtures. Another method to handle varying parameter space dimensions is *jump diffusion* (Grenander and Miller, 1994; Phillips and Smith, 1995). Carlin and Chib (1995) design a sampler consisting of several parallel chains, each traversing its own parameter space, and take the output on a given sweep to be the state of one of the parallel chains. This approach appears to work well but requires a large amount of additional analytical effort and computer time for problems such as ours; it is probably not feasible for more than a handful of k values. A more flexible technique applicable to Bayesian hierarchical models with varying dimension, which can be considered a generalization of the methods of Stephens (1997) and Phillips and Smith (1995), is *reversible jump Markov chain Monte Carlo (RJMCMC)*, developed by Green (1995). RJMCMC is essentially a random sweep Metropolis-Hastings method adapted for general state spaces. Richardson and Green (1997) develop an application of RJMCMC to univariate normal mixture models. We construct a version applicable to bivariate normal mixture models, suitable for modeling a

BVNPCP.

Construction of the algorithm itself is presented in this chapter. A primary drawback of dimension-changing MCMC methods has been the lack of suitable convergence assessment techniques. In Chapter 5 we propose a new convergence assessment method applicable to MCMC situations which are indexed by models (k). Output analysis procedures are discussed in Chapter 6.

4.2 A Bayesian Hierarchical Model Specification of the BVNPCP

A fully Bayesian specification of the BVNPCP is achieved as a *Bayesian hierarchical model (BHM)*, in which *prior distributions* are assigned to all unknown quantities (sometimes generically called "parameters," including both latent data and quantities that would traditonally be called parameters in a frequentist analysis). Some of these distributions are defined in terms of fixed *hyperparameters*. First we present our definition of the BVNPCP as a Bayesian hierarchical model (*BVNPCP-BHM*(A, n)):

Definition 4.2.1 (Bayesian hierarchical model specification of a BVNPCP) *A BVNPCP-BHM*(A, n) *for a study region* A *and observed total offspring count* n *is defined as follows:*

1. *The positions of the offspring relative to the locations* $\boldsymbol{\mu}$ *of their parents (parentage being determined by* $\mathbf{Z}$*) are independently and identically distributed as* $N(\mathbf{0}, \boldsymbol{\Sigma})$*, conditional on being confined to* A*, i.e.,*

$$\mathbf{y}_j \,|\{k, \boldsymbol{\mu}, \mathbf{Z}, \boldsymbol{\Sigma}\} \;\sim\; N\left(\boldsymbol{\mu}_{\mathbf{z}_j}, \boldsymbol{\Sigma}\right) \text{ and are independent, } , \forall j \in \{1, \ldots, n\},$$
$$\text{conditional on } \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \in A.$$

2. *Let the allocations (*$\mathbf{Z}$*) be defined as in* (2.4) *and* $\mathbf{z}_j$ *denote the* j^{th} *row of* $\mathbf{Z}$*.*

Define the notation "$\mathbf{z}_j = q$" to represent

$$\mathbf{z}_{ji} = \begin{cases} 1, & \text{if } i = q \\ 0, & \text{otherwise.} \end{cases}$$

Allocations are determined independently as

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \overset{\text{i.i.d.}}{\sim} \text{Mult}\left(1, \frac{1}{k}\mathbf{1}\right),$$

i.e.,

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \text{ are independent with } P(\mathbf{z}_j = q) = \frac{1}{k} \quad \forall q \in \{1, \ldots, k\}.$$

3. *Parent event locations ($\boldsymbol{\mu}$, also called* cluster centers*) are independently distributed uniformly on A, i.e.,*

$$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \overset{\text{i.i.d.}}{\sim} U(A) \text{ and are independent.}$$

4. *The number of parents (k) is uniformly distributed on the set of integers $\{k_{\text{lo}}, \ldots, k_{\text{hi}}\}$, i.e.,*

$$k \sim U\{k_{\text{lo}}, \ldots, k_{\text{hi}}\}$$

where k_{lo} and k_{hi} are fixed hyperparameters.

5. *The cluster shape/scale parameter ($\boldsymbol{\Sigma}$, common to all clusters) is distributed (independent of all other quantities) according to an Inverse Wishart distribution, according to*

$$\boldsymbol{\Sigma}^{-1} \sim W_2(m, V),$$

where m is the (fixed hyperparameter) degrees of freedom parameter and V the (fixed hyperparameter) covariance matrix parameter. (Generally, we will take $m = 2$ and $V^{-1} = m\sigma^2\mathbf{I}$ for reasonable σ^2, implying isotropy in the most uninformative way possible).

The Inverse Wishart distribution is a typical choice for the prior distribution of a covariance matrix (see e.g. Stephens (1997), who also studies bivariate normal

mixtures) and is a multivariate generalization of the Inverse Gamma distribution, which is very commonly used for scalar variances (e.g., Diebolt and Robert, 1994). The Inverse Wishart family is a conjugate family of prior distributions, thus making Gibbs sampling feasible for updating $\mathbf{\Sigma}$ (as we will see in section 4.4).

A realization of a BVNPCP-BHM(A, n) conditional on the realized values of k and $\mathbf{\Sigma}$ is clearly equivalent to the BVNPCP(A, k, n) and mixture model specifications (see Definitions 3.1.1 and 3.1.3). Prior distributions of k and $\mathbf{\Sigma}$ are chosen to be as uninformative as possible (while still being proper) so as to have minimal effect on inference. We could have used a Poiss(ρ) prior for k (except disallowing $k = 0$) to match the BVNPCP assumption, but this is not practical unless either (1) there is evidence suggesting likely values of k, or (2) the same process is observed in more than one study region. The quantities k and ρ would not be separately identifiable for only one realized pattern in a region A. In fact, in the case of only one realized pattern (as considered in this thesis), inference for k can be considered equivalent to inference for ρ. The parameter ν of the BVNPCP becomes irrelevant due to conditioning on total number of offspring n.

A useful representation of a Bayesian hierarchical model is a *Directed Acyclic Graph (DAG)*, which is shown for our BVNPCP-BHM(A, n) model in Figure 4.1. We follow the same conventions as Spiegelhalter et al. (1995) and Richardson and Green (1997) by enclosing unknown quantities in circles and fixed or observed quantities in boxes. Each such enclosed quantity is called a *node*. An arc represents a direct probabilistic dependency and points from a *parent* (not to be confused with the term "parent event" in BVNPCP terminology) to a *child* (i.e., children nodes are stochastically dependent on their parent node(s)). Lauritzen, Dawid, Larsen, and Leimer (1990) establish that the joint distribution of all random quantities is

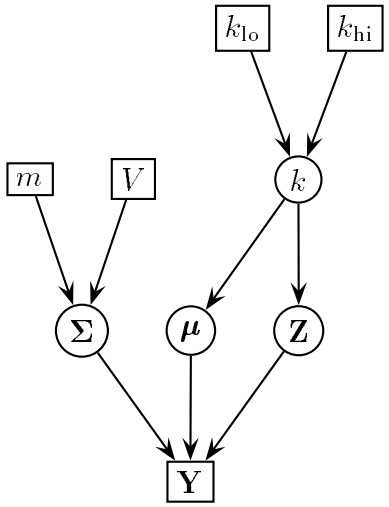


Figure 4.1: Directed Acyclic Graph (DAG) for a BVNPCP-BHM.

fully specified in terms of the conditional distribution of each node given its parents. Also, for any node η, once the values of its parent nodes are given, no other nodes besides the descendants of η are informative concerning η.

For simplicity of notation, let the unknown quantities be represented as $\boldsymbol{\theta}$ and the (fixed) hyperparameters as ξ. Then we have

$$\text{Unknown quantities:} \quad \boldsymbol{\theta} = (k, \boldsymbol{\mu}, \mathbf{Z}, \boldsymbol{\Sigma})$$

$$\text{Fixed hyperparameters:} \quad \xi = (k_{\text{lo}}, k_{\text{hi}}, m, V)$$

$$\text{Observed data:} \quad \mathbf{Y}$$

The *joint prior distribution* refers to the distribution of $\boldsymbol{\theta}$ before the data $\mathbf{Y}$ are observed and is typically written as "$p(\boldsymbol{\theta})$." The notation "$p(\boldsymbol{\theta}|\xi)$" might be more appropriate, but dependence on ξ is implied and suppressed in notation. Also note in particular that since the distribution of $\boldsymbol{\Sigma}^{-1}$ is more convenient to specify than that of $\boldsymbol{\Sigma}$, the quantities $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ may be used interchangeably in notation

when the meaning is not ambiguous. For the BVNPCP-BHM(A, n), the joint prior distribution can be determined as follows:

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\theta}|\xi) \\ &= p(k, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \mathbf{Z}|\xi) \\ &= p(\boldsymbol{\mu}|k, \boldsymbol{\Sigma}, \mathbf{Z}, \xi)p(\mathbf{Z}|k, \boldsymbol{\Sigma}, \xi)p(\boldsymbol{\Sigma}^{-1}|k, \xi)p(k|\xi) \\ &= p(\boldsymbol{\mu}|k)p(\mathbf{Z}|k)p(\boldsymbol{\Sigma}^{-1}|m, V)p(k|k_{\text{hi}}, k_{\text{lo}}) \end{aligned} \tag{4.1}$$

where, using Definition 4.2.1,

$$p(k|k_{\text{hi}}, k_{\text{lo}}) = \frac{1}{k_{\text{hi}} - k_{\text{lo}} + 1}$$
$$p(\boldsymbol{\mu}|k) = \frac{1}{|A|^k}$$
$$p(\mathbf{Z}|k) = \frac{1}{k^n}$$

and

$$p(\boldsymbol{\Sigma}^{-1}|m, V) = C^{-1}|V|^{-\frac{m}{2}} \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{m-3}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(V^{-1}\boldsymbol{\Sigma}^{-1}\right)\right\}$$
$$\text{where} \quad m \geq 2 \quad \text{and} \quad C = 2^m \pi^{\frac{1}{2}} \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{m-1}{2}\right).$$

The likelihood for a Bayesian hierarchical model is defined as the distribution of the observed data given all other quantities, and is given by Definition 4.2.1 as:

$$\begin{aligned} p(\mathbf{Y}|\boldsymbol{\theta}) &= p(\mathbf{Y}|\boldsymbol{\mu}, \mathbf{Z}, \boldsymbol{\Sigma}) \\ &= \prod_{j=1}^{n} \left[\frac{1}{2\pi}\left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_{\mathbf{z}_j}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_{\mathbf{z}_j}\right)\right\}\right] \\ &= \frac{1}{(2\pi)^n}|\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\left(\mathbf{y}_j - \boldsymbol{\mu}_{\mathbf{z}_j}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_{\mathbf{z}_j}\right)\right\} \end{aligned} \tag{4.2}$$

Note that this is equivalent to the classification likelihood defined in (3.1).

The aim of MCMC is to construct a Markov chain $\left\{\boldsymbol{\theta}^{(t)}\right\}$ whose limiting distribution is $p(\boldsymbol{\theta}|\mathbf{Y})$, the *posterior distribution* of unknown quantities given the observed data, thus allowing us to obtain a (dependent) sample (approximately) from that distribution. (Note that dependence on ξ is again suppressed in notation).

The chain is started at an initial value $\boldsymbol{\theta}^0$ (chosen at will). For a given state $\boldsymbol{\theta}^s$, a new state $\boldsymbol{\theta}^{s+1}$ is produced by "updating" some component of $\boldsymbol{\theta}^s$ (a "component" being defined as any subset of $\boldsymbol{\theta}$, ranging from one scalar parameter to the entire set of parameters). The value of $\boldsymbol{\theta}^s$ is saved at pre-determined stages (e.g., perhaps after each member of a exhaustive set of components is updated in turn) to create a sequence $\left\{\boldsymbol{\theta}^{(t)}\right\}$, each member of which is referred to as a *sweep*.

The updating must be performed according to certain criteria to ensure the proper limiting distribution. The two most common types of updating schemes which satisfy these criteria are *Gibbs sampling* and the *Metropolis-Hastings algorithm*, the former actually being a special case of the latter (see Brooks (1998) for a review). If the *full conditional distribution* of a component $\boldsymbol{\theta}_c$ given all other quantities of the model ($\{\boldsymbol{\theta}_{(c)}, \mathbf{Y}, \xi\}$) can be determined and easily sampled from, then a Gibbs step can be implemented, in which $\boldsymbol{\theta}_c$ is updated by randomly generating a new value from $p(\boldsymbol{\theta}_c|\boldsymbol{\theta}_{(c)}, \mathbf{Y}, \xi)$. (Note: the notation $\boldsymbol{\theta}_{(c)}$ represents all quantities in $\boldsymbol{\theta}$ except those in $\boldsymbol{\theta}_c$). Otherwise, if updating $\boldsymbol{\theta}_c$ will not alter the parameter space of $\boldsymbol{\theta}$, then a traditional Metropolis-Hastings step can be implemented. In this type of update, a proposed new value $\boldsymbol{\theta}_c^\star$ is simulated from *any* distribution having the correct support, but accepted only with a computable probability (computed using the forms of the proposal distribution, prior distribution and likelihood). If it is not accepted, then the chain retains its current value of $\boldsymbol{\theta}_c$. For the BVNPCP-BHM(A, n), we will see that a Gibbs update works nicely for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{Z}$. However, neither a Gibbs nor any other traditional Metropolis-Hastings update will work for k, since a change in k alters the dimension of $\boldsymbol{\theta}$. Therefore a new mechanism is required to handle transitions from one parameter space into another. RJMCMC,

explained in the next section, provides such a mechanism. We will use it to design new types of Markov chain transitions ("moves") which update k along with selected other components of $\boldsymbol{\theta}$.

4.3 RJMCMC Methodology

Green (1995) introduces a new mechanism, *reversible jump Markov chain Monte Carlo (RJMCMC)*, for Markov chain updates which allow transition between parameter spaces of differing dimension ("dimension-changing moves"). He establishes a Markov transition kernel $\kappa(\boldsymbol{\theta}, \mathrm{d}\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ and $\mathrm{d}\boldsymbol{\theta}$ may belong to different parameter spaces with different dimensions, which is aperiodic, irreducible, and satisfies *detailed balance*:

$$\int_A \int_B P\left(\mathrm{d}\boldsymbol{\theta}^a | \mathbf{Y}\right) \kappa\left(\boldsymbol{\theta}^a, \mathrm{d}\boldsymbol{\theta}^b\right) = \int_B \int_A P\left(\mathrm{d}\boldsymbol{\theta}^b | \mathbf{Y}\right) \kappa\left(\boldsymbol{\theta}^b, \mathrm{d}\boldsymbol{\theta}^a\right) \tag{4.3}$$

for any Borel sets A, B in the combined parameter space $\Theta_a \cup \Theta_b$, where $\boldsymbol{\theta}^a \in \Theta_a$ and $\boldsymbol{\theta}^b \in \Theta_b$.

Detailed balance essentially means that the equilibrium probability of moving to A and then B is equal to that of moving to B and then A. These conditions (aperiodicity, irreducibility and detailed balance) are more than enough to ensure ergodicity and the correct limiting distribution $p(\boldsymbol{\theta}|\mathbf{Y})$ of a chain implementing transitions according to $\kappa(\cdot, \cdot)$. Thus after a suitable "burn-in period" (to be assessed more rigorously in Chapter 5), we can treat samples from a Markov chain with transitions given by $\kappa(\cdot, \cdot)$ as dependent observations approximately from $p(\boldsymbol{\theta}|\mathbf{Y})$. (Note: the Gibbs sampler and Metropolis-Hastings algorithm are actually special cases of $\kappa(\cdot, \cdot)$ for which the parameter space is fixed). See Green (1995) for details.

We now describe the mechanism for dimension-changing moves designed by

Green (1995), dissecting the scheme into more components to improve clarity. Consider a pair of moves M_a and M_b that provide transitions between parameter spaces Θ_a and Θ_b, possibly of different dimension. Let $\boldsymbol{\theta}^a \in \Theta_a$ and $\boldsymbol{\theta}^b \in \Theta_b$. The notation $\boldsymbol{\theta} \xrightarrow{M_a} \boldsymbol{\theta}^\star$ represents an implementation of M_a which updates the state of the Markov chain from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^\star$, where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\star$ *could* be identical. The reversible jump mechanism is essentially a method to determine acceptance probabilities $\alpha(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b)$ and $\alpha(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a; M_b, M_a)$ such that moves are implemented according to

$$\boldsymbol{\theta}^a \xrightarrow{M_b} \begin{cases} \boldsymbol{\theta}^b, & \text{with probability } \alpha(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b) \\ \boldsymbol{\theta}^a, & \text{with probability } 1 - \alpha(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b) \end{cases} \tag{4.4}$$

and

$$\boldsymbol{\theta}^b \xrightarrow{M_a} \begin{cases} \boldsymbol{\theta}^a, & \text{with probability } \alpha(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a; M_b, M_a) \\ \boldsymbol{\theta}^b, & \text{with probability } 1 - \alpha(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a; M_b, M_a). \end{cases} \tag{4.5}$$

The acceptance probabilities are determined as follows. From $\boldsymbol{\theta}^a$, a move of type M_b is started by proposing a new state $\boldsymbol{\theta}^b$ according to:

$$c\,(M_b; \boldsymbol{\theta}^a) = \text{probability of choosing this particular move type } M_b \text{ when at } \boldsymbol{\theta}^a, \tag{4.6}$$

$$d\,(D_{\boldsymbol{\theta}^a}) = \text{probabilities of discrete random variables } D_{\boldsymbol{\theta}^a} \text{ (if any) generated as part of the move attaining their realized values (otherwise, 1),} \tag{4.7}$$

$$q\,(U_{\boldsymbol{\theta}^a}) = \text{density of continuous random variables } U_{\boldsymbol{\theta}^a} \text{ (if any) generated as part of the move (otherwise, 1),} \tag{4.8}$$

$$\left|\frac{\partial\,(T_{\boldsymbol{\theta}^b})}{\partial\,(T_{\boldsymbol{\theta}^a})}\right| = \text{Jacobian of deterministic components (if any) of the mapping } (\boldsymbol{\theta}^a, D_{\boldsymbol{\theta}^a}, U_{\boldsymbol{\theta}^a}) \mapsto \left(\boldsymbol{\theta}^b, D_{\boldsymbol{\theta}^b}, U_{\boldsymbol{\theta}^b}\right), \text{ represented more succinctly as } T_{\boldsymbol{\theta}^a} \mapsto T_{\boldsymbol{\theta}^b}, \text{ where } (T_{\boldsymbol{\theta}^a}, T_{\boldsymbol{\theta}^b}) \text{ are terms in } (\boldsymbol{\theta}^a, D_{\boldsymbol{\theta}^a}, U_{\boldsymbol{\theta}^a}) \text{ and } (\boldsymbol{\theta}^b, D_{\boldsymbol{\theta}^b}, U_{\boldsymbol{\theta}^b}) \text{ involved in any}$$

$$\text{besides trivial identity mappings (otherwise, 1),} \tag{4.9}$$

where $c\left(M_a; \boldsymbol{\theta}^b\right)$, $d\left(D_{\boldsymbol{\theta}^b}\right)$ and $q\left(U_{\boldsymbol{\theta}^b}\right)$ are defined analogously for a "reverse" move M_a which proposes a transition from $\boldsymbol{\theta}^b$ to $\boldsymbol{\theta}^a$, and

$$\dim\left(T_{\boldsymbol{\theta}^a}\right) = \dim\left(T_{\boldsymbol{\theta}^b}\right), \tag{4.10}$$

where (4.10) is referred to as the *dimension-matching condition.*

Then the acceptance probabilities are given by:

$$\alpha(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b) = \min\left\{1, R\left(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b\right)\right\} \tag{4.11}$$

where

$$\begin{aligned} R\left(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b\right) \quad = \quad & \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^b)}{p(\mathbf{Y}|\boldsymbol{\theta}^a)}\right]\left[\frac{p(\boldsymbol{\theta}^b)}{p(\boldsymbol{\theta}^a)}\right]\left[\frac{c\left(M_a; \boldsymbol{\theta}^b\right)}{c\left(M_b; \boldsymbol{\theta}^a\right)}\right] \cdot \\ & \left[\frac{d\left(D_{\boldsymbol{\theta}^b}\right)}{d\left(D_{\boldsymbol{\theta}^a}\right)}\right]\left[\frac{q\left(U_{\boldsymbol{\theta}^b}\right)}{q\left(U_{\boldsymbol{\theta}^a}\right)}\right]\left|\frac{\partial\left(T_{\boldsymbol{\theta}^b}\right)}{\partial\left(T_{\boldsymbol{\theta}^a}\right)}\right|, \end{aligned} \tag{4.12}$$

and

$$\alpha(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a; M_b, M_a) = \min\left\{1, R\left(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a; M_b, M_a\right)\right\} \tag{4.13}$$

where

$$R\left(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a; M_b, M_a\right) = \frac{1}{R\left(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b\right)}. \tag{4.14}$$

(If the denominator of an acceptance probability is calculated as zero, then the convention is to set the acceptance probability to zero, since the move would not be possible).

So, computation of the acceptance probability for a jump $\boldsymbol{\theta}^a \xrightarrow{M_b} \boldsymbol{\theta}^b$ requires a reconstruction of how the *reverse* jump $\boldsymbol{\theta}^b \xrightarrow{M_a} \boldsymbol{\theta}^a$ would have occurred.

If a pair of moves (M_a/M_b) is designed according to (4.4) – (4.14), then detailed balance holds, and a "reversible jump" between parameter spaces Θ_a and Θ_b is established.

4.4 Algorithm Design for the BVNPCP-BHM

The RJMCMC mechanism allows a considerable amount of flexibility in the design of state updates (moves) in a MCMC strategy. The challenge is to construct a collection of move types so that (a) dimension-changing moves yield acceptance probabilities in a moderate range (where "moderate" is not well-defined for RJMCMC) and (b) the entire collection of moves exhaustively updates (or at least attempts to) all components of $\boldsymbol{\theta}$. Acceptance probabilities that are too low will tend to produce poor "mixing" properties (i.e., the chain will converge very slowly to its limiting distribution, and dependence between successive sweeps will likely be very high). On the other hand, acceptance probabilities that are too high will tend to correspond to only minor state changes, and may cause the chain to get "stuck" (i.e., fail to traverse all areas or modes of the combined parameter space). Thus, to put it succinctly, dimension-changing moves must be bold but sensible.

The RJMCMC strategy we develop for a BVNPCP-BHM(A, n) is roughly based on the move types used in Richardson and Green (1997), but adapted for bivariate data, and modified to overcome a flaw apparently missed by Richardson and Green (1997). An essentially unlimited amount of fine-tuning is possible for the dimension-changing moves; we implement details as suggested by limited experimentation and do not claim our choices to be optimal. We suspect that fine-tuning is unlikely to lead to significant improvement. Richardson and Green (1997, p. 741) state that "it is rarely worth fine-tuning the proposal distribution, especially if doing so prevents simple and explicit random variate generation." We do not exclude the possibility that the addition of clever new move types may significantly improve the design, but this is left for future research. See section 4.5 for a discussion of the unexpected result of attempts to improve a move by adding a Gibbs update.

Let the notation $p(\boldsymbol{\theta}_c|\cdots)$ represent the full conditional distribution of a component $\boldsymbol{\theta}_c$ of $\boldsymbol{\theta}$ given the values of all other components, i.e. $p(\boldsymbol{\theta}_c|\boldsymbol{\theta}_{(c)})$. Our RJMCMC strategy for the BVNPCP-BHM(A, n) consists of the following collection of move types:

$M_{\boldsymbol{\mu}}$ (Update $\boldsymbol{\mu}$): update $\boldsymbol{\mu}$ via a Gibbs step, by generating a new value from $p(\boldsymbol{\mu}|\cdots)$

$M_{\boldsymbol{\Sigma}}$ (Update $\boldsymbol{\Sigma}$): update $\boldsymbol{\Sigma}$ via a Gibbs step, by generating a new value from $p(\boldsymbol{\Sigma}|\cdots)$

$M_{\mathbf{Z}}$ (Update $\mathbf{Z}$): update $\mathbf{Z}$ via a Gibbs step, by generating a new value from $p(\mathbf{Z}|\cdots)$

M_S/M_C (Split/Combine): attempt to either split a cluster in two or combine two "neighboring" clusters into one

M_B/M_D (Birth/Death): attempt to either generate a new cluster center at a random location, or delete an existing cluster center

Note that (M_S/M_C) is a "reversible jump" pair of dimension-changing moves, as is (M_B/M_D).

The birth/death move pair, although perhaps somewhat redundant and inefficient in the presence of split/combine, is included because of the possibility of split/combine moves being insufficient to explore certain regions of the parameter space. Perhaps the addition or deletion of a cluster center in a certain area would improve the situation, but is discouraged by the limited capability of the split/combine mechanism. At some point in the chain, a birth or death move may attempt such a maneuver.

Details of the move types, and finally the overall algorithm, are given in the next several subsections. First we define some notation and new terminology that will be used.

For any symbol "a", the quantities $(\boldsymbol{\theta}^a, k^a, \boldsymbol{\mu}^a, \boldsymbol{\Sigma}^a, \mathbf{Z}^a)$ denote the *current* values of $(\boldsymbol{\theta}, k, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z})$ at a given state "a" in the Markov chain (where "a" is not an index of time or sweeps, but rather is set to a value suggestive of its contextual meaning). The absence of a symbol (when not ambiguous) may also suggest a state, e.g., we could discuss a transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^\star$.

The notation $\bar{\mathbf{y}}_i$ retains the same meaning as given in (3.5). Since $z_{ji} \in \{0, 1\}$ always holds in our RJMCMC method, we also define

$$n_i = \sum_{j=1}^{n} z_{ji} \tag{4.15}$$

to indicate the number of offspring allocated to cluster i.

In combining clusters, acceptance probabilities are reasonable only for attempts to combine nearby clusters. Thus, for the M_C move, a definition of "adjacency" is needed. Since we are modeling (in general) a geometrically anisotropic process, the usual Euclidean distance is inappropriate. Instead, an alternative measure of adjacency is defined:

Definition 4.4.1 ($NN_{\boldsymbol{\Sigma}}$) *The $\boldsymbol{\Sigma}-Nearest-Neighbor$ ($NN_{\boldsymbol{\Sigma}}$) of a cluster i is defined as*

$$NN_{\boldsymbol{\Sigma}}(i) = \arg\min_{q \neq i} \left[\left(\boldsymbol{\mu}_q - \boldsymbol{\mu}_i \right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_q - \boldsymbol{\mu}_i \right) \right],$$

in other words, the cluster with the closest center to its own in terms of the Mahalanobis distance induced by $\boldsymbol{\Sigma}$.

In deriving full conditional distributions and ascertaining which moves can be performed in parallel, it is helpful to consult a *Conditional Independence Graph (CIG)* of the model, which is shown in Figure 4.2.

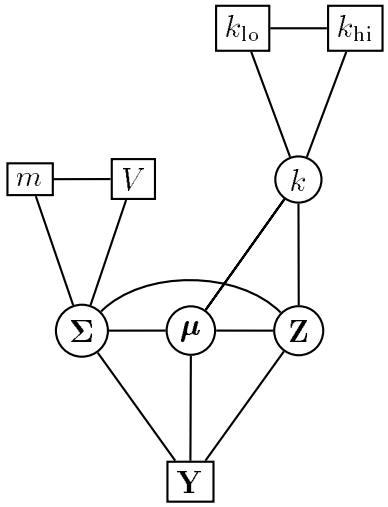


Figure 4.2: Conditional Independence Graph (CIG) for a BVNPCP-BHM.

The CIG is formed by "moralizing" the DAG, i.e. connecting common parents of children nodes, and dropping the arrows (Lauritzen and Spiegelhalter, 1988). The conditional distribution of a node in the CIG given all other quantities can be reduced to its distribution given nodes it is directly connected to. (In other words, nodes not connected in the CIG are conditionally independent given all other nodes). We see from Figure 4.2 that there are no conditional independencies among $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$ and $\mathbf{Z}$, and thus we cannot implement any Gibbs steps in parallel.

In general, the derivations of full conditional distributions and other densities are given by Definition 4.2.1, 4.1, 4.2, and Figures 4.1 and 4.2. Detailed reasons are given only for nonstandard steps in the derivations.

4.4.1 $M_{\boldsymbol{\mu}}$ Details

The full conditional distribution of $\boldsymbol{\mu}$ given all other quantities can be determined as follows:

$$\begin{aligned}
p(\boldsymbol{\mu}|\cdots) &= p(\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z},\boldsymbol{\Sigma},k) \\
&= \frac{p(\mathbf{Y}|\boldsymbol{\mu},\mathbf{Z},\boldsymbol{\Sigma},k)p(\boldsymbol{\mu}|\mathbf{Z},\boldsymbol{\Sigma},k)}{p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Sigma},k)} \\
&= \frac{p(\mathbf{Y}|\boldsymbol{\mu},\mathbf{Z},\boldsymbol{\Sigma})p(\boldsymbol{\mu}|k)}{p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Sigma},k)} \\
&\propto p(\mathbf{Y}|\boldsymbol{\mu},\mathbf{Z},\boldsymbol{\Sigma})p(\boldsymbol{\mu}|k) \\
&\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n}z_{ji}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\right\} \\
&= \exp\left\{-\frac{n}{2}\mathrm{tr}\left[\left\{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n}z_{ji}\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)\left(\mathbf{y}_j-\bar{\mathbf{y}}_i\right)'\right\}\boldsymbol{\Sigma}^{-1}\right] - \right. \\
&\qquad \left.\frac{1}{2}\sum_{i=1}^{k}\left(\sum_{j=1}^{n}z_{ji}\right)\left(\bar{\mathbf{y}}_i-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{y}}_i-\boldsymbol{\mu}_i\right)\right\} \\
&\qquad \text{(see (3.6) for details of the previous step)} \\
&\propto \prod_{i=1}^{k}\exp\left\{-\frac{1}{2}n_i\left(\bar{\mathbf{y}}_i-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{y}}_i-\boldsymbol{\mu}_i\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}\left(\boldsymbol{\mu}_i-\bar{\mathbf{y}}_i\right)'\left[\frac{1}{n_i}\boldsymbol{\Sigma}\right]^{-1}\left(\boldsymbol{\mu}_i-\bar{\mathbf{y}}_i\right)\right\},
\end{aligned}$$

implying that $(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k)$ are independently distributed with $\boldsymbol{\mu}_i|\cdots \sim N\left(\bar{\mathbf{y}}_i,\frac{1}{n_i}\boldsymbol{\Sigma}\right)$.

Thus the update $\boldsymbol{\theta} \xrightarrow{M_{\mu}} \boldsymbol{\theta}^{\star}$ can be performed with a Gibbs step as follows:

$$(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{Z},k) \xrightarrow{M_{\mu}} (\boldsymbol{\mu}^{\star},\boldsymbol{\Sigma},\mathbf{Z},k)$$

where $(\boldsymbol{\mu}_1^{\star},\ldots,\boldsymbol{\mu}_k^{\star})$ are generated independently according to

$$\boldsymbol{\mu}_i^{\star} \sim N\left(\bar{\mathbf{y}}_i,\frac{1}{n_i}\boldsymbol{\Sigma}\right), \quad i\in\{1,\ldots,k\}. \tag{4.16}$$

4.4.2 $M_{\boldsymbol{\Sigma}}$ Details

The full conditional distribution of $\boldsymbol{\Sigma}^{-1}$ given all other quantities can be determined as follows:

$$\begin{aligned}
p(\boldsymbol{\Sigma}^{-1}|\cdots) &= p(\boldsymbol{\Sigma}^{-1}|\mathbf{Y},\mathbf{Z},\boldsymbol{\mu},m,V) \\
&= \frac{p(\mathbf{Y}|\boldsymbol{\Sigma},\mathbf{Z},\boldsymbol{\mu},m,V)p(\boldsymbol{\Sigma}^{-1}|\mathbf{Z},\boldsymbol{\mu},m,V)}{p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\mu},m,V)} \\
&= \frac{p(\mathbf{Y}|\boldsymbol{\Sigma},\mathbf{Z},\boldsymbol{\mu})p(\boldsymbol{\Sigma}^{-1}|m,V)}{p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\mu})} \\
&\propto p(\mathbf{Y}|\boldsymbol{\Sigma},\mathbf{Z},\boldsymbol{\mu})p(\boldsymbol{\Sigma}^{-1}|m,V) \\
&\propto \left[\left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\right\}\right]\cdot \\
&\qquad \left[\left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{m-3}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(V^{-1}\boldsymbol{\Sigma}^{-1}\right)\right\}\right] \\
&= \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{n+m-3}{2}}\exp\left\{-\frac{1}{2}\left[\mathrm{tr}\left(V^{-1}\boldsymbol{\Sigma}^{-1}\right)+\right.\right. \\
&\qquad \left.\left.\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\right]\right\} \\
&= \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{n+m-3}{2}}\exp\left\{-\frac{1}{2}\left[\mathrm{tr}\left(V^{-1}\boldsymbol{\Sigma}^{-1}\right)+\right.\right. \\
&\qquad \left.\left.\mathrm{tr}\left(\left\{\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\right\}\boldsymbol{\Sigma}^{-1}\right)\right]\right\} \\
&= \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{n+m-3}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\left(V^{-1}+\right.\right.\right. \\
&\qquad \left.\left.\left.\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\right)\boldsymbol{\Sigma}^{-1}\right]\right\},
\end{aligned}$$

implying that

$$\boldsymbol{\Sigma}^{-1}|\cdots \sim W_2\left(m+n,\left[V^{-1}+\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\right]^{-1}\right).$$

Thus the update $\boldsymbol{\theta} \xrightarrow{M_{\boldsymbol{\Sigma}}} \boldsymbol{\theta}^{\star}$ can be performed with a Gibbs step as follows:

$$(\boldsymbol{\Sigma},\boldsymbol{\mu},\mathbf{Z},k) \xrightarrow{M_{\boldsymbol{\Sigma}}} (\boldsymbol{\Sigma}^{\star},\boldsymbol{\mu},\mathbf{Z},k)$$

where $\boldsymbol{\Sigma}^\star$ is generated according to

$$\left[\boldsymbol{\Sigma}^{-1}\right]^\star \sim W_2\left(m+n, \left[V^{-1}+\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\right]^{-1}\right). \tag{4.17}$$

Note that the update involves essentially a weighted contribution of prior and observed data, with weight $\frac{n}{n+m}$ given to the data. Thus, with $m=2$ and $n \geq 100$ (as in the applications in this thesis), the prior has very little influence provided V is chosen reasonably.

4.4.3 $M_{\mathbf{Z}}$ Details

The full conditional distribution of $\mathbf{Z}$ given all other quantities can be determined as follows:

$$\begin{aligned}
p(\mathbf{Z}|\cdots) &= p(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Sigma},\boldsymbol{\mu},k) \\
&= \frac{p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Sigma},\boldsymbol{\mu},k)p(\mathbf{Z}|\boldsymbol{\Sigma},\boldsymbol{\mu},k)}{p(\mathbf{Y}|\boldsymbol{\Sigma},\boldsymbol{\mu},k)} \\
&= \frac{p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Sigma},\boldsymbol{\mu})p(\mathbf{Z}|k)}{p(\mathbf{Y}|\boldsymbol{\Sigma},\boldsymbol{\mu})} \\
&\propto p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Sigma},\boldsymbol{\mu})p(\mathbf{Z}|k) \\
&\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_{\mathbf{z}_j}\right)\right\} \\
&= \prod_{j=1}^{n}\left[\prod_{i=1}^{k}\mathrm{I}(\mathbf{z}_j=i)\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\right\}\right],
\end{aligned}$$

and so $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are independent (by factorization), and for each $j \in \{1,\ldots,n\}$

$$P\left(\mathbf{z}_j=i\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\right\}$$
$$\text{for } i \in \{1,\ldots,k\}.$$

Thus

$$P\left(\mathbf{z}_j=i\,|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma},k\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\right\}}{\sum_{q=1}^{k}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_q\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_q\right)\right\}}$$

independently for each $j \in \{1, \ldots, n\}$,

duplicating the result (3.9).

Thus the update $\boldsymbol{\theta} \xrightarrow{M_{\mathbf{Z}}} \boldsymbol{\theta}^\star$ can be performed with a Gibbs step as follows:

$$(\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k) \xrightarrow{M_{\mathbf{Z}}} (\mathbf{Z}^\star, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k)$$

where $(\mathbf{z}_1^\star, \ldots, \mathbf{z}_n^\star)$ are generated independently according to

$$P\left(\mathbf{z}_j^\star = i \,|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_i\right)\right\}}{\sum_{q=1}^{k} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_q\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_q\right)\right\}}. \tag{4.18}$$

4.4.4 (M_S/M_C) Details

The split/combine move pair (M_S/M_C) is designed as a "reversible jump" satisfying (4.4) – (4.14). We first describe the mechanisms and then derive the corresponding acceptance probabilities, starting with the simpler combine mechanism.

The combine move attempts a transition from $\boldsymbol{\theta}^S$ to $\boldsymbol{\theta}^C$ through the following sequence of steps:

Algorithm 4.4.2 (COMBINE) *Implement the move* $\boldsymbol{\theta}^S \xrightarrow{M_C} \boldsymbol{\theta}^C$ *as follows:*

1. *Initialize* $\boldsymbol{\theta}^C$ *to* $\boldsymbol{\theta}^S$.

2. *Choose a cluster, say* i_1, *from the uniform distribution on the integers* $\{1, \ldots, k^S\}$, *i.e.,*

$$i_1 \sim U\{1, \ldots, k^S\}$$

3. *Determine* $i_2 = NN_{\boldsymbol{\Sigma}^C}(i_1)$ *according to* $\boldsymbol{\theta}^C$, *i.e., identify the* $\boldsymbol{\Sigma}$*-Nearest-Neighbor* i_2 *of* i_1 *in the current state.*

4. *Combine clusters* i_1 *and* i_2 *into cluster* $i^\star$ *by averaging the cluster centers and re-allocating offspring from* i_1 *and* i_2 *to* $i^\star$, *as follows:*

   (a) *Set* k^C *to* $k^S - 1$.

(b) *Set* $i^\star$ *to* $\min(i_1, i_2)$.

(c) *Set* $\boldsymbol{\mu}_{i^\star}^C = \frac{1}{2}\left(\boldsymbol{\mu}_{i_1}^S + \boldsymbol{\mu}_{i_2}^S\right)$.

(d) *Set* $\boldsymbol{\mu}_i^C$ *to* $\boldsymbol{\mu}_{i+1}^S$ *for* $i \in \{\max(i_1, i_2), \ldots, k^C\}$ *and then discard* $\boldsymbol{\mu}_{k^C+1}$.

(e) *For all* j *such that* $\mathbf{z}_j^C = \max(i_1, i_2)$, *set* $\mathbf{z}_j^C = i^\star$.

(f) *For all* j *such that* $\mathbf{z}_j^C \in \{\max(i_1, i_2) + 1, \ldots, k^C + 1\}$, *set* $\mathbf{z}_j^C$ *to* $\mathbf{z}_j^C - 1$.

5. *Update* $\boldsymbol{\theta}^S$ *according to*

$$\boldsymbol{\theta}^S \xrightarrow{M_C} \begin{cases} \boldsymbol{\theta}^C, & \text{with probability } \alpha(\boldsymbol{\theta}^S, \boldsymbol{\theta}^C; M_S, M_C) \\ \boldsymbol{\theta}^S, & \text{with probability } 1 - \alpha(\boldsymbol{\theta}^S, \boldsymbol{\theta}^C; M_S, M_C), \end{cases}$$

where $\alpha(\boldsymbol{\theta}^S, \boldsymbol{\theta}^C; M_S, M_C)$ *is determined by* (4.29).

The split move attempts a transition from $\boldsymbol{\theta}^C$ to $\boldsymbol{\theta}^S$ through the following sequence of steps:

Algorithm 4.4.3 (SPLIT) *Implement the move* $\boldsymbol{\theta}^C \xrightarrow{M_S} \boldsymbol{\theta}^S$ *as follows:*

1. *Initialize* $\boldsymbol{\theta}^S$ *to* $\boldsymbol{\theta}^C$.

2. *Choose a cluster* $i^\star$ *from the uniform distribution on the integers* $\{1, \ldots, k^C\}$, *i.e.,*

$$i^\star \sim U\{1, \ldots, k^C\}$$

3. *Split cluster* $i^\star$ *into clusters* i_1 *and* i_2 *as follows:*

(a) *Set* k^S *to* $k^C + 1$.

(b) *Set* $i_1 = i^\star$ *and* $i_2 = k^S$.

(c) *Sample* $\mathbf{u}$ *from* $N\left(\mathbf{0}, \boldsymbol{\Sigma}^C\right)$.

(d) *Set* $\boldsymbol{\mu}_{i_1}^S = \boldsymbol{\mu}_{i^\star}^C - \mathbf{u}$ *and set* $\boldsymbol{\mu}_{i_2}^S = \boldsymbol{\mu}_{i^\star}^C + \mathbf{u}$.

(e) *For all j such that $\mathbf{z}_j^S = i^\star$, sample $\mathbf{z}_j^S$ from $\{i_1, i_2\}$, analogously to (4.18), independently according to*

$$P\left(\mathbf{z}_j^S = i_1\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_{i_1}^S\right)'\left[\boldsymbol{\Sigma}^S\right] - 1\left(\mathbf{y}_j - \boldsymbol{\mu}_{i_1}^S\right)\right\}}{\sum_{q\in\{i_1,i_2\}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_q^S\right)'\left[\boldsymbol{\Sigma}^S\right]^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_q^S\right)\right\}}.$$

4. *Determine $NN_1 = NN_{\boldsymbol{\Sigma}^S}(i_1)$ according to $\boldsymbol{\theta}^S$, i.e., identify the $\boldsymbol{\Sigma}$-Nearest-Neighbor NN_1 of i_1 in the proposed new state.*

5. *Determine $NN_2 = NN_{\boldsymbol{\Sigma}^S}(i_2)$ according to $\boldsymbol{\theta}^S$, i.e., identify the $\boldsymbol{\Sigma}$-Nearest-Neighbor NN_2 of i_2 in the proposed new state.*

6. *If $NN_1 \neq i_2$ and $NN_2 \neq i_1$, then preserve the current state, i.e., update $\boldsymbol{\theta}^C$ according to $\boldsymbol{\theta}^C \xrightarrow{M_S} \boldsymbol{\theta}^C$ (since the reverse move from $\boldsymbol{\theta}^S$ to $\boldsymbol{\theta}^C$ would be impossible).*

7. *If $NN_1 = i_2$ or $NN_2 = i_1$, then update $\boldsymbol{\theta}^C$ according to*

$$\boldsymbol{\theta}^C \xrightarrow{M_S} \begin{cases} \boldsymbol{\theta}^S, & \text{with probability } \alpha(\boldsymbol{\theta}^C, \boldsymbol{\theta}^S; M_C, M_S) \\ \boldsymbol{\theta}^C, & \text{with probability } 1 - \alpha(\boldsymbol{\theta}^C, \boldsymbol{\theta}^S; M_C, M_S), \end{cases}$$

where $\alpha(\boldsymbol{\theta}^C, \boldsymbol{\theta}^S; M_C, M_S)$ is determined by (4.27).

We will specify later that a choice of (split/combine) $\equiv (M_S/M_C)$ (meaning that one of split or combine will be attempted) will be made with probability $\frac{3}{16}$ at an arbitrary state $\boldsymbol{\theta}$ of the chain. If (M_S/M_C) is chosen, then the M_S and M_C move types are attempted with equal probability, provided $k > k_{\text{lo}}$ for M_C and $k < k_{\text{hi}}$ for M_S. Encountering an inappropriate k is interpreted as "not choosing the move", so that this information belongs in $c(\cdot;\cdot)$. In other words,

$$c\left(M_S; \boldsymbol{\theta}^C\right) = \begin{cases} \frac{3}{40}, & \text{if } k^C < k_{\text{hi}} \\ 0, & \text{otherwise} \end{cases} \tag{4.19}$$

and

$$c\left(M_C;\boldsymbol{\theta}^S\right)=\begin{cases}\frac{3}{40}, & \text{if } k^S>k_{\text{lo}}\\ 0, & \text{otherwise.}\end{cases} \tag{4.20}$$

Now we consider discrete random variables generated as part of the moves. In the combine move, it is important to realize that (i_1,i_2) are actually chosen as a pair: the same pair *could* result if either is chosen first. So for M_C we have $D_{\boldsymbol{\theta}}s=(i_1,i_2)$. If $(NN_{\boldsymbol{\Sigma}}s(i_1)=i_2$ and $NN_{\boldsymbol{\Sigma}}s(i_2)=i_1)$, then either i_1 *or* i_2 could be initially chosen in the combine mechanism, resulting in the same pair (i_1,i_2) to combine and hence the same move. The probability of either being chosen is $\frac{2}{k^S}$. If $(NN_{\boldsymbol{\Sigma}}s(i_1)=i_2$ and $NN_{\boldsymbol{\Sigma}}s(i_2)\neq i_1)$ or $(NN_{\boldsymbol{\Sigma}}s(i_1)\neq i_2$ and $NN_{\boldsymbol{\Sigma}}s(i_2)=i_1)$, then only one cluster choice would create the pair (i_1,i_2). If $(NN_{\boldsymbol{\Sigma}}s(i_1)\neq i_2$ and $NN_{\boldsymbol{\Sigma}}s(i_2)\neq i_1)$, then (i_1,i_2) would not be chosen in M_C. Thus we have

$$d\left(D_{\boldsymbol{\theta}}s\right)=\frac{1}{k^S}\left[\mathrm{I}\left(NN_{\boldsymbol{\Sigma}}s(i_1)=i_2\right)+\mathrm{I}\left(NN_{\boldsymbol{\Sigma}}s(i_2)=i_1\right)\right]. \tag{4.21}$$

In the split move, the discrete quantities generated are the cluster to split ($i^\star$, with probability $\frac{1}{k^C}$) and the new allocations for offspring belonging to that cluster. New values are assigned to $\mathbf{z}_j^S$ for j such that $\mathbf{z}_j^C=i^\star$, and the 2 possible values for each are $\{i_1,i_2\}=\{i^\star,k^C+1\}$. Hence

$$D_{\boldsymbol{\theta}^C}=\{i^\star\}\cup\left\{\mathbf{z}_j^S \text{ for } j \text{ such that } \mathbf{z}_j^C=i^\star\right\},$$

and

$$d\left(D_{\boldsymbol{\theta}^C}\right)=\frac{1}{k^C}\prod_{j:\mathbf{z}_j^C=i^\star}\frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_{z_j^S}^S\right)'\left[\boldsymbol{\Sigma}^S\right]^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_{z_j^S}^S\right)\right\}}{\sum_{q\in\{i^\star,k^C+1\}}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j-\boldsymbol{\mu}_q^S\right)'\left[\boldsymbol{\Sigma}^S\right]^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_q^S\right)\right\}}. \tag{4.22}$$

There are no continuous quantities generated for the combine move, so

$$q\left(U_{\boldsymbol{\theta}}s\right)=1. \tag{4.23}$$

For the split move, $U_{\boldsymbol{\theta}^C}=\mathbf{u}$ and

$$q\left(U_{\boldsymbol{\theta}^C}\right)=\frac{1}{2\pi}\left|\boldsymbol{\Sigma}^C\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{u}'\left[\boldsymbol{\Sigma}^C\right]^{-1}\mathbf{u}\right\}. \tag{4.24}$$

The only terms involved in non-trivial deterministic mappings are $T_{\boldsymbol{\theta}^C}=$

$\{\boldsymbol{\mu}^C_{i^*}, \mathbf{u}\}$ and $T_{\boldsymbol{\theta}^S} = \{\boldsymbol{\mu}^S_{i_1}, \boldsymbol{\mu}^S_{i_2}\}$. Expressed in terms of scalars, this mapping is

$$\begin{aligned}
\mu^S_{i_1,1} &= \mu^C_{i^*,1} - u_1 \\
\mu^S_{i_1,2} &= \mu^C_{i^*,2} - u_2 \\
\mu^S_{i_2,1} &= \mu^C_{i^*,1} + u_1 \\
\mu^S_{i_2,2} &= \mu^C_{i^*,2} + u_2
\end{aligned}$$

Note that the dimension-matching condition (4.10) is satisfied. The Jacobian of the mapping is calculated as

$$\begin{aligned}
\left|\frac{\partial\left(T_{\boldsymbol{\theta}^S}\right)}{\partial\left(T_{\boldsymbol{\theta}^C}\right)}\right| &= \begin{vmatrix}
\frac{\partial \mu^S_{i_1,1}}{\partial \mu^C_{i^*,1}} & \frac{\partial \mu^S_{i_1,1}}{\partial \mu^C_{i^*,2}} & \frac{\partial \mu^S_{i_1,1}}{\partial \mu^C_{u_1}} & \frac{\partial \mu^S_{i_1,1}}{\partial \mu^C_{u_2}} \\
\frac{\partial \mu^S_{i_1,2}}{\partial \mu^C_{i^*,1}} & \frac{\partial \mu^S_{i_1,2}}{\partial \mu^C_{i^*,2}} & \frac{\partial \mu^S_{i_1,2}}{\partial \mu^C_{u_1}} & \frac{\partial \mu^S_{i_1,2}}{\partial \mu^C_{u_2}} \\
\frac{\partial \mu^S_{i_2,1}}{\partial \mu^C_{i^*,1}} & \frac{\partial \mu^S_{i_2,1}}{\partial \mu^C_{i^*,2}} & \frac{\partial \mu^S_{i_2,1}}{\partial \mu^C_{u_1}} & \frac{\partial \mu^S_{i_2,1}}{\partial \mu^C_{u_2}} \\
\frac{\partial \mu^S_{i_2,2}}{\partial \mu^C_{i^*,1}} & \frac{\partial \mu^S_{i_2,2}}{\partial \mu^C_{i^*,2}} & \frac{\partial \mu^S_{i_2,2}}{\partial \mu^C_{u_1}} & \frac{\partial \mu^S_{i_2,2}}{\partial \mu^C_{u_2}}
\end{vmatrix} \\
&= \begin{vmatrix}
1 & 0 & -1 & 0 \\
0 & 1 & 0 & -1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1
\end{vmatrix} \\
&= 4. \qquad (4.25)
\end{aligned}$$

Finally, the likelihood and prior ratios are determined simply by plugging in the values of $\boldsymbol{\theta}^C$ and $\boldsymbol{\theta}^S$:

$$\begin{aligned}
\left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^S)}{p(\mathbf{Y}|\boldsymbol{\theta}^C)}\right]\left[\frac{p(\boldsymbol{\theta}^S)}{p(\boldsymbol{\theta}^C)}\right] &= \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^S)}{p(\mathbf{Y}|\boldsymbol{\theta}^C)}\right]\left[\frac{p(\boldsymbol{\mu}^S|k^S)p(\mathbf{Z}^S|k^S)p(k^S|k_{\mathrm{hi}},k_{\mathrm{lo}})}{p(\boldsymbol{\mu}^C|k^C)p(\mathbf{Z}^C|k^C)p(k^C|k_{\mathrm{hi}},k_{\mathrm{lo}})}\right] \\
&= \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^S)}{p(\mathbf{Y}|\boldsymbol{\theta}^C)}\right]|A|\left(\frac{k^C}{k^C+1}\right)^n. \qquad (4.26)
\end{aligned}$$

Now we have all the factors required for calculation of the acceptance probabilities. These are computed according to (4.11)–(4.14), using values from (4.19)–(4.26), as follows:

$$\alpha(\boldsymbol{\theta}^C, \boldsymbol{\theta}^S; M_C, M_S) = \min\left\{1, R\left(\boldsymbol{\theta}^C, \boldsymbol{\theta}^S; M_C, M_S\right)\right\} \qquad (4.27)$$

where

$$R\left(\boldsymbol{\theta}^C, \boldsymbol{\theta}^S; M_C, M_S\right) = \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^S)}{p(\mathbf{Y}|\boldsymbol{\theta}^C)}\right]\left[\frac{p(\boldsymbol{\theta}^S)}{p(\boldsymbol{\theta}^C)}\right]\left[\frac{c\left(M_C; \boldsymbol{\theta}^S\right)}{c\left(M_S; \boldsymbol{\theta}^C\right)}\right] \cdot \left[\frac{d\left(D_{\boldsymbol{\theta}^S}\right)}{d\left(D_{\boldsymbol{\theta}^C}\right)}\right]\left[\frac{q\left(U_{\boldsymbol{\theta}^S}\right)}{q\left(U_{\boldsymbol{\theta}^C}\right)}\right]\left|\frac{\partial\left(T_{\boldsymbol{\theta}^S}\right)}{\partial\left(T_{\boldsymbol{\theta}^C}\right)}\right|, \tag{4.28}$$

and

$$\alpha(\boldsymbol{\theta}^S, \boldsymbol{\theta}^C; M_S, M_C) = \min\left\{1, R\left(\boldsymbol{\theta}^S, \boldsymbol{\theta}^C; M_S, M_C\right)\right\} \tag{4.29}$$

where

$$R\left(\boldsymbol{\theta}^S, \boldsymbol{\theta}^C; M_S, M_C\right) = \frac{1}{R\left(\boldsymbol{\theta}^C, \boldsymbol{\theta}^S; M_C, M_S\right)}. \tag{4.30}$$

Note that the strategy for generation of $\mathbf{u}$ in the split mechanism could be implemented differently. In Algorithm 4.4.3, the 2 new clusters are displaced in opposite directions from the original cluster on the scale of the variation of offspring about parents. This scale could be increased or decreased (i.e., generating the displacement $\mathbf{u}$ from $N(\mathbf{0}, q\boldsymbol{\Sigma}^C)$ for some constant q) in pilot runs to ascertain values yielding better acceptance rates for a given data set (although this would be somewhat time consuming, and perhaps not worth the effort). Richardson and Green (1997) pursue a different strategy (in one dimension), generating u from a Beta distribution and displacing by a multiple of u (depending on current estimates of variances and mixing proportions). However, a consequence of this strategy is that a combine move may not be reversible, since the hypothetical reverse split move might need to generate a u outside $[0, 1]$ to accomplish the required displacement, a feat impossible for the Beta distribution. Richardson and Green (1997) do not appear to account for this possibility. Analogous moves (correcting for the hypothetical reverse split problem, which can be accomplished easily through adjustment of $d(\cdot)$) could certainly be implemented in our 2-dimensional scheme.

4.4.5 (M_B/M_D) Details

The birth/death move pair (M_B/M_D) is also designed as a "reversible jump" satisfying (4.4) – (4.14). We first describe the mechanisms and then derive the corresponding acceptance probabilities.

The death move attempts a transition from $\boldsymbol{\theta}^B$ to $\boldsymbol{\theta}^D$ through the following sequence of steps:

Algorithm 4.4.4 (DEATH) *Implement the move* $\boldsymbol{\theta}^B \xrightarrow{M_D} \boldsymbol{\theta}^D$ *as follows:*

1. *Initialize* $\boldsymbol{\theta}^D$ *to* $\boldsymbol{\theta}^B$.

2. *Choose a cluster, say* $i^\star$, *from the uniform distribution on the integers* $\{1, \ldots, k^B\}$, *i.e.,*

$$i^\star \sim U\{1, \ldots, k^B\}$$

3. *Delete cluster* $i^\star$ *and re-allocate its offspring to other clusters, as follows:*

   (a) *Set* k^D *to* $k^B - 1$ *and discard* $\boldsymbol{\mu}_{i^\star}$.

   (b) *Re-label remaining cluster-centers as* $1, \ldots, k^D$ *and re-label* $\mathbf{Z}^D$ *accordingly (except for* j *such that* $\mathbf{z}_j^B = i^\star$, *which are handled below).*

   (c) *For all* j *such that* $\mathbf{z}_j^B = i^\star$, *sample* $\mathbf{z}_j^D$ *from* $\{1, \ldots, k^D\}$, *analogously to* (4.18), *independently according to*

$$P\left(\mathbf{z}_j^D = i\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_i^D\right)' \left[\boldsymbol{\Sigma}^D\right]^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_i^D\right)\right\}}{\sum_{q=1}^{k^D} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_q^D\right)' \left[\boldsymbol{\Sigma}^D\right]^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}_q^D\right)\right\}}.$$

4. *Update* $\boldsymbol{\theta}^B$ *according to*

$$\boldsymbol{\theta}^B \xrightarrow{M_D} \begin{cases} \boldsymbol{\theta}^D, & \text{with probability } \alpha(\boldsymbol{\theta}^B, \boldsymbol{\theta}^D; M_B, M_D) \\ \boldsymbol{\theta}^B, & \text{with probability } 1 - \alpha(\boldsymbol{\theta}^B, \boldsymbol{\theta}^D; M_B, M_D), \end{cases}$$

where $\alpha(\boldsymbol{\theta}^B, \boldsymbol{\theta}^D; M_B, M_D)$ *is determined by* (4.41).

The birth move attempts a transition from $\boldsymbol{\theta}^D$ to $\boldsymbol{\theta}^B$ through the following sequence of steps:

Algorithm 4.4.5 (BIRTH) *Implement the move* $\boldsymbol{\theta}^D \xrightarrow{M_B} \boldsymbol{\theta}^B$ *as follows:*

1. *Initialize* $\boldsymbol{\theta}^B$ *to* $\boldsymbol{\theta}^D$.

2. *Create a new cluster* $i^\star$ *and give all offspring a chance to switch to this cluster, as follows:*

   (a) *Set* k^B *to* $k^D + 1$.

   (b) *Set* $i^\star = k^B$.

   (c) *Sample* $\boldsymbol{\mu}^B_{i^\star}$ *from the uniform distribution on* A, $U(A)$.

   (d) *Update* $\mathbf{Z}^B$ *so that offspring can either stay in their current clusters or switch to* $i^\star$*: for all* j, *sample* $\mathbf{z}^B_j$ *from* $\{\mathbf{z}^D_j, i^\star\}$, *analogously to* (4.18), *independently according to*

$$P\left(\mathbf{z}^B_j = i^\star\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}^B_{i^\star}\right)'\left[\boldsymbol{\Sigma}^B\right]^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}^B_{i^\star}\right)\right\}}{\sum_{q\in\{\mathbf{z}^D_j, i^\star\}}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}^B_{q}\right)'\left[\boldsymbol{\Sigma}^B\right]^{-1}\left(\mathbf{y}_j - \boldsymbol{\mu}^B_{q}\right)\right\}}.$$

3. *Update* $\boldsymbol{\theta}^D$ *according to*

$$\boldsymbol{\theta}^D \xrightarrow{M_B} \begin{cases} \boldsymbol{\theta}^B, & \text{with probability } \alpha(\boldsymbol{\theta}^D, \boldsymbol{\theta}^B; M_D, M_B) \\ \boldsymbol{\theta}^D, & \text{with probability } 1 - \alpha(\boldsymbol{\theta}^D, \boldsymbol{\theta}^B; M_D, M_B), \end{cases}$$

   where $\alpha(\boldsymbol{\theta}^D, \boldsymbol{\theta}^B; M_D, M_B)$ *is determined by* (4.39).

A choice of (M_B/M_D) (meaning that one of birth or death will be attempted) will be made with probability $\frac{1}{16}$ at an arbitrary state $\boldsymbol{\theta}$ of the chain. If (M_B/M_D) is chosen, then the M_B and M_D move types are attempted with equal probability, provided $k > k_{\text{lo}}$ for M_D and $k < k_{\text{hi}}$ for M_B. As with (M_S/M_C), encountering an inappropriate k is interpreted as "not choosing the move", so that this information

belongs in $c(\cdot;\cdot)$. So we have

$$c\left(M_B;\boldsymbol{\theta}^D\right) = \begin{cases} \frac{1}{40}, & \text{if } k^D < k_{\text{hi}} \\ 0, & \text{otherwise} \end{cases} \tag{4.31}$$

and

$$c\left(M_D;\boldsymbol{\theta}^B\right) = \begin{cases} \frac{1}{40}, & \text{if } k^B > k_{\text{lo}} \\ 0, & \text{otherwise.} \end{cases} \tag{4.32}$$

For the death move M_D we choose a cluster $i^\star$ to delete and then re-allocate its offspring. Hence

$$D_{\boldsymbol{\theta}^B} = \{i^\star\} \cup \{\mathbf{z}_j^D \text{ for } j \text{ such that } \mathbf{z}_j^B = i^\star\}$$

$$d\left(D_{\boldsymbol{\theta}^B}\right) = \frac{1}{k^B} \prod_{j:\mathbf{z}_j^B = i^\star} \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_{z_j^D}^D\right)' \left[\boldsymbol{\Sigma}^D\right]^{-1} \left(\mathbf{y}_j - \boldsymbol{\mu}_{z_j^D}^D\right)\right\}}{\sum_{q=1}^{k^B-1} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_q^D\right)' \left[\boldsymbol{\Sigma}^D\right]^{-1} \left(\mathbf{y}_j - \boldsymbol{\mu}_q^D\right)\right\}}. \tag{4.33}$$

In the birth move M_B, the only discrete quantities generated are the new allocations for each offspring. The 2 possible values for the j^{th} offspring are $\mathbf{z}_j^D$ and $i^\star$. Hence $D_{\boldsymbol{\theta}^D} = \mathbf{Z}^B$ and

$$d\left(D_{\boldsymbol{\theta}^D}\right) = \prod_{j=1}^{n} \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_{z_j^B}^B\right)' \left[\boldsymbol{\Sigma}^B\right]^{-1} \left(\mathbf{y}_j - \boldsymbol{\mu}_{z_j^B}^B\right)\right\}}{\sum_{q \in \{\mathbf{z}_j^D, i^\star\}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_j - \boldsymbol{\mu}_q^B\right)' \left[\boldsymbol{\Sigma}^B\right]^{-1} \left(\mathbf{y}_j - \boldsymbol{\mu}_q^B\right)\right\}}. \tag{4.34}$$

There are no continuous quantities generated for the death move, so

$$q\left(U_{\boldsymbol{\theta}^B}\right) = 1. \tag{4.35}$$

For the birth move, $U_{\boldsymbol{\theta}^D} = \boldsymbol{\mu}_{i^\star}$ and

$$q\left(U_{\boldsymbol{\theta}^D}\right) = \frac{1}{|A|}. \tag{4.36}$$

There are no non-trivial deterministic mappings, and therefore

$$\left|\frac{\partial\left(T_{\boldsymbol{\theta}^B}\right)}{\partial\left(T_{\boldsymbol{\theta}^D}\right)}\right| = 1. \tag{4.37}$$

Finally, the likelihood and prior ratios are determined simply by plugging in the values of $\boldsymbol{\theta}^D$ and $\boldsymbol{\theta}^B$:

$$\left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^B)}{p(\mathbf{Y}|\boldsymbol{\theta}^D)}\right]\left[\frac{p(\boldsymbol{\theta}^B)}{p(\boldsymbol{\theta}^D)}\right] = \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^B)}{p(\mathbf{Y}|\boldsymbol{\theta}^D)}\right]\left[\frac{p(\boldsymbol{\mu}^B|k^B)p(\mathbf{Z}^B|k^B)p(k^B|k_{\text{hi}},k_{\text{lo}})}{p(\boldsymbol{\mu}^D|k^D)p(\mathbf{Z}^D|k^D)p(k^D|k_{\text{hi}},k_{\text{lo}})}\right]$$

$$= \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^B)}{p(\mathbf{Y}|\boldsymbol{\theta}^D)}\right] |A| \left(\frac{k^D}{k^D+1}\right)^n. \tag{4.38}$$

Now we have all the factors required for calculation of the acceptance probabilities. These are computed according to (4.11)–(4.14), using values from (4.31)–(4.38), as follows:

$$\alpha(\boldsymbol{\theta}^D, \boldsymbol{\theta}^B; M_D, M_B) = \min\left\{1, R\left(\boldsymbol{\theta}^D, \boldsymbol{\theta}^B; M_D, M_B\right)\right\} \tag{4.39}$$

where

$$R\left(\boldsymbol{\theta}^D, \boldsymbol{\theta}^B; M_D, M_B\right) = \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^B)}{p(\mathbf{Y}|\boldsymbol{\theta}^D)}\right] \left[\frac{p(\boldsymbol{\theta}^B)}{p(\boldsymbol{\theta}^D)}\right] \left[\frac{c\left(M_D; \boldsymbol{\theta}^B\right)}{c\left(M_B; \boldsymbol{\theta}^D\right)}\right] \cdot \left[\frac{d\left(D_{\boldsymbol{\theta}^B}\right)}{d\left(D_{\boldsymbol{\theta}^D}\right)}\right] \left[\frac{q\left(U_{\boldsymbol{\theta}^B}\right)}{q\left(U_{\boldsymbol{\theta}^D}\right)}\right] \left|\frac{\partial\left(T_{\boldsymbol{\theta}^B}\right)}{\partial\left(T_{\boldsymbol{\theta}^D}\right)}\right|, \tag{4.40}$$

and

$$\alpha(\boldsymbol{\theta}^B, \boldsymbol{\theta}^D; M_B, M_D) = \min\left\{1, R\left(\boldsymbol{\theta}^B, \boldsymbol{\theta}^D; M_B, M_D\right)\right\} \tag{4.41}$$

where

$$R\left(\boldsymbol{\theta}^B, \boldsymbol{\theta}^D; M_B, M_D\right) = \frac{1}{R\left(\boldsymbol{\theta}^D, \boldsymbol{\theta}^B; M_D, M_B\right)}. \tag{4.42}$$

Richardson and Green (1997), in one dimension, implement birth/death of *empty* clusters only. This makes more sense in their situation, because they model their mixing proportions instead of enforcing an equality constraint. An empty cluster i_0 in their model can be given a "weight" of zero (i.e., $P(\mathbf{z}_j = i_0) = 0$). We tried an experimental empty cluster birth/death move type, the result being that births were hardly ever accepted, and there were usually no empty clusters to choose from for a death.

4.4.6 The Overall Form of the Algorithm

Define the move type "$M_{(SCBD)}$" as

$$M_{(SCBD)} = \begin{cases} M_S, & \text{with probability } \frac{3}{8} \\ M_C, & \text{with probability } \frac{3}{8} \\ M_B, & \text{with probability } \frac{1}{8} \\ M_D, & \text{with probability } \frac{1}{8} \end{cases}$$

Our Markov chain consists of a sequence of *states* of $\boldsymbol{\theta}$ resulting from individual updates. Not all of these states are saved. The value of the chain is saved at regular intervals, the index of which we call a *sweep*. The value of the chain at sweep t is denoted $\boldsymbol{\theta}^{(t)}$. Also the generic notation $\boldsymbol{\theta}_\bullet$ is used to denote the value of $\boldsymbol{\theta}$ at a particular state. (If two or more instances of $\boldsymbol{\theta}_\bullet$ appear together in the same equation, they are not necessarily equal).

Using terminology from sections 4.4.1 – 4.4.5, our RJMCMC algorithm for the BVNPCP-BHM(A, n) is represented as follows:

Algorithm 4.4.6 (RJMCMC for BVNPCP-BHM) *For the BVNPCP-BHM(A, n), implement RJMCMC as follows:*

1. *Specify values for all fixed hyperparameters* $\xi = (k_{\text{lo}}, k_{\text{hi}}, m, V)$.

2. *Choose an initial value* $\boldsymbol{\theta}_0$ *for the chain, in any manner desired (possibly using the observed data), and set*

$$\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}_0.$$

3. *For* $t \in \{1, \ldots, T\}$, *perform the following sequence of moves:*

$$\boldsymbol{\theta}^{(t-1)} \overset{M_{(SCBD)}}{\longrightarrow} \boldsymbol{\theta}_\bullet \overset{M_Z}{\longrightarrow} \boldsymbol{\theta}_\bullet \overset{M_\mu}{\longrightarrow} \boldsymbol{\theta}_\bullet \overset{M_\Sigma}{\longrightarrow} \boldsymbol{\theta}^{(t)}.$$

Due to limitations on storage space, we save the value at every 10^{th} sweep of each chain. The order of the sequence of move types (and the choice of random vs.

systematic scanning, and the choice of frequency of each move type in the sequence) could certainly be changed without affecting the limiting distribution. The *rate* of convergence, i.e. the mixing properties, may depend on the strategy used. We did not experiment with different orderings. We chose to perform dimension-changing moves first and save the state directly after the Gibbs steps, suspecting that the chain may require an update of $(\boldsymbol{\mu}, \mathbf{Z}, \boldsymbol{\Sigma})$ to "get comfortable" immediately after jumping into a new dimension, possibly producing more sensible output. The $\boldsymbol{\Sigma}$ update is performed last, essentially because our inference focuses on this parameter. These reasons are rather ad-hoc, and we suggest future research to assess the effect of such choices.

An important consequence of our particular construction of the dimension-changing moves is that cluster labels (i.e. $\{1, \ldots, k\}$) are not informative. We have imposed no ordering restriction on the cluster centers $\boldsymbol{\mu}$. Thus, for example, $\boldsymbol{\mu}_1^{(100)}$ and $\boldsymbol{\mu}_1^{(500)}$ have no meaningful connection, even if $k^{(100)} = k^{(500)}$. Also, even though $\mathbf{z}_j^{(t)}$ refers to the allocation of the same offspring j on all sweeps, its value is only meaningful in terms of the labels of $\boldsymbol{\mu}^{(t)}$ on the given sweep. This limits our selection of applicable convergence assessment and output analysis methods. However, we offer a reasonable solution to the inability to analyze convergence of *all* parameters in Chapter 5, and we shall see in Chapter 6 that a considerable number of output analysis methods are still applicable. Another ramification of this "label-switching problem" (as it is often called) is that estimation of cluster-specific features is not possible. This is not a concern in our situation, because we are primarily interested in modeling $\boldsymbol{\Sigma}$, which we take to be common for all clusters.

Stephens (1997, Chapter 3) develops two alternative algorithms to attempt to deal with the label-switching problem. His approach is essentially to estimate

the cluster labels in subsets of the MCMC output with identical k. Richardson and Green (1997), in one dimension, impose ordering restrictions on the cluster centers and design moves so that the ordering cannot be disturbed. There are several drawbacks of this strategy, however. Robert (1997) warns that it "may create traps for the resulting Gibbs sampler," citing Diebolt and Robert (1994), and "slow convergence down." Gilks (1997) asserts that the ordering restriction "worsens mixing in the MCMC algorithm" and is not necessary for valid Bayesian inference. Nobile (1997) points out that if strong prior information is available, then its use may run contrary to the ordering constraint.

As mentioned by several authors in the discussion of Richardson and Green (1997), a sensible ordering restriction does not appear to be feasible in more than one dimension. Many authors agree with Nobile (1997) on "deferring to the post-processing stage the decision on whether and which constraints to impose." Since the method of Stephens (1997) can only approximate labels, and we are not concerned with estimating cluster-specific features, we have chosen to restrict our use of methods to those that are invariant to label-switching.

4.5 The Effect of Incorporating Gibbs Updates into Dimension-changing Moves

Especially if the acceptance rate of a dimension-changing move is very low, it may be advisable to attempt to "improve" the move. One possible option is to add to the move a final stage which updates parameters that do not change as part of the existing move, by generating new values from their full conditional distribution using the new values of parameters which *do* change as part of the existing move, presumably making the collection of new parameter values more "in synch." It

seems that this would make a move more "intelligent" by encouraging it to produce a more realistic set of parameter values and consequently increase the acceptance probability. For example, if a split move in our RJMCMC algorithm is attempted with only 2 existing clusters, then the supposed cluster shape/scale might change dramatically. Would the split move benefit from an added Gibbs update of $\mathbf{\Sigma}$ given the new parent locations and cluster memberships?

Consider a dimension-changing reversible move pair (M_a/M_b) in an arbitrary RJMCMC sampler. Suppose the parameter vector can be partitioned (disjointly) into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where all parameters whose values potentially change in (M_a/M_b) are contained in $\boldsymbol{\theta}_1$.

For simplicity of notation, the "reversible jump device" part of (4.12) is represented using a function $G(\cdot)$:

$$\frac{G(\boldsymbol{\theta}^a|\boldsymbol{\theta}^b; M_a)}{G(\boldsymbol{\theta}^b|\boldsymbol{\theta}^a; M_b)} = \left[\frac{c\left(M_a; \boldsymbol{\theta}^b\right)}{c\left(M_b; \boldsymbol{\theta}^a\right)}\right] \left[\frac{d\left(D_{\boldsymbol{\theta}^b}\right)}{d\left(D_{\boldsymbol{\theta}^a}\right)}\right] \left[\frac{q\left(U_{\boldsymbol{\theta}^b}\right)}{q\left(U_{\boldsymbol{\theta}^a}\right)}\right] \left|\frac{\partial\left(T_{\boldsymbol{\theta}^b}\right)}{\partial\left(T_{\boldsymbol{\theta}^a}\right)}\right|.$$

Since $\boldsymbol{\theta}_1$ is not updated as part of the moves,

$$\frac{G(\boldsymbol{\theta}^a|\boldsymbol{\theta}^b; M_a)}{G(\boldsymbol{\theta}^b|\boldsymbol{\theta}^a; M_b)} \equiv \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a)}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b)}.$$

Also, note that the posterior ratio can be used in (4.12) instead of the likelihood and prior ratios:

$$\left[\frac{p(\boldsymbol{\theta}^b|\mathbf{Y})}{p(\boldsymbol{\theta}^a|\mathbf{Y})}\right] = \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta}^b)}{p(\mathbf{Y}|\boldsymbol{\theta}^a)}\right] \left[\frac{p(\boldsymbol{\theta}^b)}{p(\boldsymbol{\theta}^a)}\right].$$

The (un-truncated) acceptance probability for M_b is then

$$\begin{aligned} R\left(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b\right) &= \frac{p(\boldsymbol{\theta}^b|\mathbf{Y})}{p(\boldsymbol{\theta}^a|\mathbf{Y})} \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a)}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b)} \\ &= \frac{p(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b|\mathbf{Y})}{p(\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a|\mathbf{Y})} \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a)}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b)} \\ &= \frac{p(\boldsymbol{\theta}_2^b|\boldsymbol{\theta}_1^b, \mathbf{Y}) p(\boldsymbol{\theta}_1^b|\mathbf{Y})}{p(\boldsymbol{\theta}_2^a|\boldsymbol{\theta}_1^a, \mathbf{Y}) p(\boldsymbol{\theta}_1^a|\mathbf{Y})} \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a)}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b)} \\ &\quad \text{(where } \boldsymbol{\theta}_2^b = \boldsymbol{\theta}_2^a\text{)}. \end{aligned}$$

Consider an alternative move type (M_a'/M_b'), which functions as follows:

- M_b': Given the current state $\boldsymbol{\theta}^a$, generate $\boldsymbol{\theta}_1^b$ exactly as in M_b, and then sample $\boldsymbol{\theta}_2^b$ from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^b, \mathbf{Y})$.

- M_a': Given the current state $\boldsymbol{\theta}^b$, generate $\boldsymbol{\theta}_1^a$ exactly as in M_a, and then sample $\boldsymbol{\theta}_2^a$ from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^a, \mathbf{Y})$.

This is a valid move pair for RJMCMC. The only change from (M_a/M_b) is the insertion of terms $p(\boldsymbol{\theta}_2^b|\boldsymbol{\theta}_1^a, \mathbf{Y})$ into $q\left(U_{\boldsymbol{\theta}^a}\right)$ and/or $d\left(D_{\boldsymbol{\theta}^a}\right)$, and $p(\boldsymbol{\theta}_2^a|\boldsymbol{\theta}_1^b, \mathbf{Y})$ into $q\left(U_{\boldsymbol{\theta}^b}\right)$ and/or $d\left(D_{\boldsymbol{\theta}^b}\right)$ (see (4.8) and (4.7)). The new (un-truncated) acceptance probability is

$$\begin{aligned}
R\left(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a', M_b'\right) &= \frac{p(\boldsymbol{\theta}^b|\mathbf{Y})}{p(\boldsymbol{\theta}^a|\mathbf{Y})} \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a')}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b')} \\
&= \frac{p(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b|\mathbf{Y})}{p(\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a|\mathbf{Y})} \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a')}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b')} \\
&= \frac{p(\boldsymbol{\theta}_2^b|\boldsymbol{\theta}_1^b, \mathbf{Y}) p(\boldsymbol{\theta}_1^b|\mathbf{Y})}{p(\boldsymbol{\theta}_2^a|\boldsymbol{\theta}_1^a, \mathbf{Y}) p(\boldsymbol{\theta}_1^a|\mathbf{Y})} \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a')}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b')} \\
&= \frac{p(\boldsymbol{\theta}_2^b|\boldsymbol{\theta}_1^b, \mathbf{Y}) p(\boldsymbol{\theta}_1^b|\mathbf{Y})}{p(\boldsymbol{\theta}_2^a|\boldsymbol{\theta}_1^a, \mathbf{Y}) p(\boldsymbol{\theta}_1^a|\mathbf{Y})} \left[\frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a)}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b)} \frac{p(\boldsymbol{\theta}_2^b|\boldsymbol{\theta}_1^b, \mathbf{Y})}{p(\boldsymbol{\theta}_2^a|\boldsymbol{\theta}_1^a, \mathbf{Y})}\right] \\
&= \frac{p(\boldsymbol{\theta}_1^b|\mathbf{Y})}{p(\boldsymbol{\theta}_1^a|\mathbf{Y})} \frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a)}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b)} \\
&\qquad \text{(where } \boldsymbol{\theta}_2^b \neq \boldsymbol{\theta}_2^a \text{ necessarily).}
\end{aligned}$$

Thus the component $\frac{p(\boldsymbol{\theta}_2^b|\boldsymbol{\theta}_1^b, \mathbf{Y})}{p(\boldsymbol{\theta}_2^a|\boldsymbol{\theta}_1^a, \mathbf{Y})}$ of the posterior ratio has been *eliminated* from the acceptance probability! This is not the only change from $R\left(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b; M_a, M_b\right)$, however: the value of $\frac{G(\boldsymbol{\theta}_1^a|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b; M_a)}{G(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^a, \boldsymbol{\theta}_2^a; M_b)}$ may be different because $\boldsymbol{\theta}_2^b = \boldsymbol{\theta}_2^a$ in (M_a/M_b) but not necessarily in (M_a'/M_b'). It is difficult to imagine cases where this difference would be beneficial, but we do not deny the possibility. Regardless, it is apparent that this type of move modification does not yield the anticipated effect.

CHAPTER 5
RJMCMC CONVERGENCE ASSESSMENT

Before conducting inference using output from a Markov chain Monte Carlo sampler, the output should be analyzed to determine a point at which the sampler has "converged" to the proper limiting distribution. There are two distinct aspects of convergence to consider:

1. Are the samples being generated from the correct distribution?

2. Has the entire parameter space been traversed?

It is difficult to rigorously verify either condition; a general strategy which we will follow is to run several chains started at over-dispersed values. If at some point all chains are generating samples from approximately the same distribution, then this distribution is presumed to be the correct one (a justifiable assumption when the Markov chain is designed properly). If the starting values are appropriately over-dispersed, then it is also likely that the parameter space has been thoroughly traversed as well.

5.1 Choice of Parameters to Monitor

In MCMC convergence assessment it is recommended that, if feasible, all parameters are monitored, and if not, then at least one representative parameter of each "type" is monitored. The output of the RJMCMC sampler (Algorithm 4.4.6) consists of $k^{(t)}$, $\boldsymbol{\Sigma}^{(t)}$, $\boldsymbol{\mu}^{(t)}$, and $\mathbf{Z}^{(t)}$ for each sweep t. Firstly, k and $\boldsymbol{\Sigma} \equiv (\sigma_{11}, \sigma_{22}, \sigma_{12})$ can be monitored easily, as these parameters retain the same meaning from sweep

to sweep. As mentioned at the end of section 4.4.6, label-switching inhibits the possibility of monitoring individual components of $\boldsymbol{\mu}$ and $\mathbf{Z}$. However, we have devised an approach to monitor a combination of $\boldsymbol{\mu}$ and $\mathbf{Z}$ which *is* identifiable. A certain number of offspring are "marked," and the parent locations of these offspring are tracked from sweep to sweep. We choose to monitor 3 offspring, chosen as:

1. an event near the center of a clearly defined cluster,

2. an event located between 2 clusters that are potential competitors for ownership of this event, and

3. an isolated event that could potentially be the sole member of a cluster, or an outlier in another cluster.

The purpose of these particular choices is to attempt to monitor parent locations that are expected to fluctuate across sweeps in different ways. This approach essentially boils down to monitoring $\boldsymbol{\mu}_{\mathbf{z}_{j_1}}$, $\boldsymbol{\mu}_{\mathbf{z}_{j_2}}$ and $\boldsymbol{\mu}_{\mathbf{z}_{j_3}}$ (6 scalar parameters in all) for 3 chosen offspring j_1, j_2 and j_3. These quantities retain the same meaning from sweep to sweep, and they represent instances of both parent locations and offspring allocations. We emphasize that the choice of offspring to track can be made after the sampler is run, since we are only using the usual sampler output.

The point patterns analyzed in this thesis are shown in Figures C.1 – C.2, with the offspring whose parent locations are to be monitored in convergence assessment marked as "1", "2" and "3." Detailed descriptions of the implementation of the RJMCMC algorithm for the Redwood data and simulated patterns are postponed until Chapter 7, but some figures displaying results of such RJMCMC runs are referred to in this chapter for explanatory purposes.

5.2 Initial Assessment

A sensible first step for any convergence assessment technique is inspection of trace plots for each scalar parameter chosen to monitor. A collection of such trace plots is displayed in Figure D.1 for the first $2,000$ sweeps of a RJMCMC sampler run for the Redwood data, with every 10^{th} sweep shown, and in Figure D.2 for all $200,000$ sweeps of this run with every 1000^{th} sweep shown. It is not possible to ascertain "convergence" from such plots, but they can be helpful in revealing any major problems. The trace plots in Figure D.1 show that at least the $\boldsymbol{\Sigma}$ parameters appear to explore different regions of the parameter space over time, without returning. This indicates that sufficient mixing has not yet occurred, and the sampler should probably be run longer. Note that k occasionally stays at one value for long periods of time, and the values of $\boldsymbol{\Sigma}$ components appear to change along with k. In contrast, the trace plots in Figure D.2 appear to be well-behaved in the sense that variation is more homogeneous over time; thus there is no indication of trouble. Note the occasional spikes in the trace plots of the tracked parent locations: these represent instances of the offspring being allocated to an unusual cluster. As long as these spikes occur somewhat regularly over time, they are not indicators of convergence trouble.

Since we do not monitor allocations $\mathbf{Z}$ in their pure form, it is informative to check allocations for at least a handful of sweeps. Figure D.3 shows allocations at the last occurrence of $k = 7, 10, 12$ and 15, for the same RJMCMC run. Many authors, particularly in the hidden Markov chain literature, have proposed techniques for monitoring allocations (Gruet et al., 1998; Robert et al., 1998; Robert and Titterington, 1998; Robert and Mengersen, 1997). Most involve constructing grayscale plots of observation number vs. sweep number with the darkness of the

plotted points representing allocations. The grayscale patterns across sweeps then suggest whether allocations are remaining stable or fluctuating wildly for each offspring. Such allocation plots become useless in the presence of a significant amount of label-switching, however, and so are not useful for our model. Even if offspring tend to stay with the same clusters over time, the labels may change as an artifact of the dimension-changing mechanisms.

Another useful feature to monitor as an initial assessment is autocorrelation functions (ACF's) of the parameters at different lags. The ACF estimates the correlation between $\theta^{(t)}$ and $\theta^{(t+g)}$ for a given parameter θ and lag g. High ACF's for a parameter indicate slow mixing, which is not in itself a sign of lack of convergence, but does provide a warning that convergence is likely to be slow. A chain with high ACF's will take a long time to traverse the entire parameter space. High ACF's also warn that it will be inappropriate to estimate variances with the usual sample variance estimator. ACF's for normalized versions of parameters (except ϕ) in the RJMCMC run on the Redwood data, using every 10^{th} sweep for the last $100,000$ sweeps, are shown in Figure D.4. Normalized versions are used in anticipation of their use to construct confidence intervals and tests, methods for which we must carefully deal with autocorrelation. Since we saved only every 10^{th} value from the MCMC output, "lag-1" could technically be considered lag-10. Note the extremely high ACF of k, which is not surprising given that dimension changes do not occur very often. The ACF's of $\log \sigma_{11}, \log \sigma_{22}$, and $\log \Psi$ follow suit, since the cluster size tends to vary predictably with k. The parameters describing cluster *shape* $(z(\rho_{12}), \log \gamma)$, however, have lower ACF's, suggesting that shape estimates may not vary as much with k. All tracked $\boldsymbol{\mu}$ parameters except $\boldsymbol{\mu}_{j_3 2}$ have extremely low ACF; the high ACF for $\boldsymbol{\mu}_{j_3 2}$ is likely due to extended periods of time in different

clusters.

Finally, an assessment of acceptance rates for dimension-changing moves is useful in targeting any move types which may be inefficient. As mentioned in the beginning of section 4.4, reasonable ranges of acceptance rates for dimension-changing moves have not yet been established, but one could at least compare acceptance rates for different move types. Such comparisons should not be taken too seriously, however, as some moves may have lower acceptance rates but provide for transitions not covered by other moves (as we suspect is the case for our birth/death move, although for the relatively small data sets used in this thesis we cannot evaluate this supposition).

5.3 Previous Related Approaches

Virtually none of the existing MCMC convergence assessment techniques apply to RJMCMC due to the transitions between different parameter spaces. A thorough review of MCMC convergence assessment techniques is provided by Cowles and Carlin (1996) and Mengersen, Robert, and Guihenneuc-Jouyaux (1998). Most are univariate, considering only one parameter at a time. Currently, the two most popular types are those developed by Geweke (1991) and Gelman and Rubin (1992). Geweke (1991) proposes comparing (univariate) sample means of a parameter computed from different parts of a chain, using variance estimates adjusted for autocorrelation. Gelman and Rubin (1992) propose an analysis of variance (ANOVA) type approach in which several chains are run, and the ratio of a pooled variance estimate and a within-chain variance estimate, similar to the comparison between total mean-square and error mean square in a one-way ANOVA with "chain" being the factor, is calculated. The idea is that if the two variances are comparable, then

the chains are probably realizations from a common distribution, presumably the correct limiting distribution. This method depends on the absence of other significant factors, but for our BVNPCP-BHM, k could be considered a factor in this paradigm, since parameters are expected to vary considerably with k. Neither of these two popular methods (nor any others that the author is aware of) are sufficient to detect lack of convergence *within* k. Convergence within k really should be assessed also, since k is essentially a model indicator, and some models may be less well-behaved than others.

Extensions to Geweke's technique do not appear to be feasible, since output from a RJMCMC sampler for a given k consists of a series of uninterrupted sequences separated by visits to other values of k, and thus an autocorrelation would need to be assessed in each of these sequences. It is an extension of Gelman and Rubin's method, both from univariate to multivariate *and* 1-way-ANOVA to 2-way-ANOVA, that we develop in the next section. First we discuss some other extensions of their technique which are relevant.

5.3.1 Brooks and Gelman's Multivariate Potential Scale Reduction Factor (MPSRF)

Brooks and Gelman (1996) introduce several different versions of Gelman and Rubin's convergence diagnostic and suggest monitoring both numerator and denominator, not just a ratio. One of the versions is multivariate in the sense of providing an upper bound of an analogous convergence diagnostic computed for a set of scalar parameters.

We will focus on their multivariate convergence diagnostic, but first derive the univariate analogue. It requires running $C > 1$ chains of a MCMC sampler (with

T sweeps each, say) with over-dispersed starting values. A number m of successive (overlapping) "batches" of increasing length (multiples of a base batch length b) of the output are analyzed from each chain. Let $\theta_c^{(qb+1)}, \ldots, \theta_c^{(2qb)}$, denote the q^{th} batch of length qb, from chain c for a scalar parameter θ, where $c \in \{1, \ldots, C\}$. Successive batches for $q = 1, \ldots, \frac{T}{b}$ are used. Brooks and Gelman (1996) propose monitoring $\widehat{V}^{(q)}(\theta)$, $W^{(q)}(\theta)$ and $\frac{\widehat{V}^{(q)}(\theta)}{W^{(q)}(\theta)}$ (which they call the *potential scale reduction factor*, or PSRF), defined below, computed for each batch.

Defining $\bar{\theta}_c^{(\cdot)}$ and $\bar{\theta}_\cdot^{(\cdot)}$ as

$$\bar{\theta}_c^{(\cdot)} = \frac{1}{qb} \sum_{t=qb+1}^{2qb} \theta_c^{(t)} \quad \text{and} \quad \bar{\theta}_\cdot^{(\cdot)} = \frac{1}{qbC} \sum_{c=1}^{C} \sum_{t=qb+1}^{2qb} \theta_c^{(t)},$$

the quantities of interest are defined as follows:

$$\widehat{V}^{(q)}(\theta) = \frac{qb-1}{qb} W^{(q)}(\theta) + \left(1 + \frac{1}{C}\right) B(\theta)/(qb)$$

and

$$W^{(q)}(\theta) = \frac{1}{C(qb-1)} \sum_{c=1}^{C} \sum_{t=qb+1}^{2qb} \left(\theta_c^{(t)} - \bar{\theta}_c^{(\cdot)}\right)^2$$

where

$$B(\theta)/(qb) = \frac{1}{C-1} \sum_{c=1}^{C} \left(\bar{\theta}_c^{(\cdot)} - \bar{\theta}_\cdot^{(\cdot)}\right)^2.$$

The value of $\widehat{V}^{(q)}(\theta)$ should be larger than $W^{(q)}(\theta)$ for small q, since the starting values are over-dispersed; they may approach a common value as q increases, indicating that the variation is homogeneous across chains. It may happen that the numerator and denominator happen to fluctuate together but yield a ratio close to 1, so Brooks and Gelman (1996) recommend monitoring these individually in addition to the ratio. They mention that, provided the starting values are appropriately over-dispersed, the settling of $\frac{\widehat{V}^{(q)}(\theta)}{W^{(q)}(\theta)}$ to a neighborhood of 1, *and* of $\widehat{V}^{(q)}(\theta)$ and $W^{(q)}(\theta)$ approximately to a common value for $q \geq q_0$, are generally adequate reasons to justify inferences based on posterior means and variances of the collection of samples $\{\theta^{(q_0 b+1)}, \theta^{(q_0 b+2)}, \ldots\}$. This situation often suggests additionally that the

chains are following the same distribution, but they warn that only approximate equivalence of the first 2 moments across chains has been established. It is difficult to determine how close to 1 is "close enough": they cite a cutoff of 1.2 as a rule of thumb in one of their examples.

The multivariate version for a vector $\boldsymbol{\theta}$ of parameters is defined analogously, estimating posterior variance-covariance matrices instead of scalar variances:

Defining $\bar{\boldsymbol{\theta}}_c^{(\cdot)}$ and $\bar{\boldsymbol{\theta}}_\cdot^{(\cdot)}$ as

$$\bar{\boldsymbol{\theta}}_c^{(\cdot)} = \frac{1}{qb}\sum_{t=qb+1}^{2qb}\boldsymbol{\theta}_c^{(t)} \quad \text{and} \quad \bar{\boldsymbol{\theta}}_\cdot^{(\cdot)} = \frac{1}{qbC}\sum_{c=1}^{C}\sum_{t=qb+1}^{2qb}\boldsymbol{\theta}_c^{(t)},$$

the multivariate convergence diagnostics are given by

$$\widehat{V}^{(q)}(\boldsymbol{\theta}) = \frac{qb-1}{qb}W^{(q)}(\boldsymbol{\theta}) + \left(1+\frac{1}{C}\right)B(\boldsymbol{\theta})/(qb)$$

and

$$W^{(q)}(\boldsymbol{\theta}) = \frac{1}{C(qb-1)}\sum_{c=1}^{C}\sum_{t=qb+1}^{2qb}\left(\boldsymbol{\theta}_c^{(t)}-\bar{\boldsymbol{\theta}}_c^{(\cdot)}\right)\left(\boldsymbol{\theta}_c^{(t)}-\bar{\boldsymbol{\theta}}_c^{(\cdot)}\right)'$$

where

$$B(\boldsymbol{\theta})/(qb) = \frac{1}{C-1}\sum_{c=1}^{C}\left(\bar{\boldsymbol{\theta}}_c^{(\cdot)}-\bar{\boldsymbol{\theta}}_\cdot^{(\cdot)}\right)\left(\bar{\boldsymbol{\theta}}_c^{(\cdot)}-\bar{\boldsymbol{\theta}}_\cdot^{(\cdot)}\right)'.$$

The multivariate PSRF (MPSRF) is then defined as a maximum root statistic-type measure of distance between $\widehat{V}^{(q)}(\boldsymbol{\theta})$ and $W^{(q)}(\boldsymbol{\theta})$:

$$MPSRF(\boldsymbol{\theta}) = \max_{a\in\Re^p}\frac{a'\widehat{V}^{(q)}(\boldsymbol{\theta})a}{a'W^{(q)}(\boldsymbol{\theta})a},$$

where p is the dimension of θ. They proceed to prove that $MPSRF(\boldsymbol{\theta})$ can be represented in terms of the maximum eigenvalue of $\left[W^{(q)}(\boldsymbol{\theta})\right]^{-1}\widehat{V}^{(q)}(\boldsymbol{\theta})$, and that it provides an upper bound on the collection of univariate PSRF's, $\frac{\widehat{V}^{(q)}(\theta_i)}{W^{(q)}(\theta_i)}$, where θ_i is the i^{th} scalar component of $\boldsymbol{\theta}$. We now present these results in generic notation.

Lemma 5.3.1 *For two non-singular, positive definite and symmetric $p\times p$ matrices M and N,*

$$\max_{a\in\Re^p}\frac{a'Ma}{a'Na} = \lambda,$$

where λ is the largest eigenvalue of $N^{-1}M$.

Proof: See Mardia, Kent, and Bibby (1979, Theorem A.9.2). □

Lemma 5.3.2 *Let M and N be two non-singular, positive definite and symmetric $p \times p$ matrices, and denote the diagonal elements as $\{m_1, \ldots, m_p\}$ and $\{n_1, \ldots, n_p\}$, respectively. Then*

$$\max_{a \in \Re^p} \frac{a'Ma}{a'Na} \geq \max_{i \in \{1,\ldots,p\}} \frac{m_i}{n_i}.$$

Proof: Let $\mathbf{i}_j$ denote a $p \times 1$ vector of zeroes with the j^{th} entry replaced by 1. Then

$$\max_{a \in \Re^p} \frac{a'Ma}{a'Na} \geq \max_{j \in \{1,\ldots,p\}} \frac{\mathbf{i}_j' M \mathbf{i}_j}{\mathbf{i}_j' N \mathbf{i}_j} = \max_{i \in \{1,\ldots,p\}} \frac{m_i}{n_i}. \quad \square$$

Note that the collection $\{v_1, \ldots, v_p\}$ of diagonal elements of the multivariate version of $\widehat{V}$ are equivalent to the univariate versions, and that the same holds for the diagonal elements $\{w_1, \ldots, w_p\}$ of W. Thus Lemma 5.3.2 establishes that Brooks and Gelman's MPSRF is an upper bound of the unvariate PSRF's. They suggest monitoring this MPSRF, and also $f\left(\widehat{V}^{(q)}(\boldsymbol{\theta})\right)$ and $f\left(W^{(q)}(\boldsymbol{\theta})\right)$ for some real-valued function $f(\cdot)$, such as the determinant.

5.3.2 Brooks and Giudici's Proposed RJMCMC Diagnostic

Brooks and Giudici (1998) introduce the first proposed method, a univariate one, specifically designed for RJMCMC convergence assessment. The basic idea is to compute various decompositions of the estimated variance of a collection of samples of a scalar parameter from C different chains. Two factors determine the decompositions: "model" (the indicator of the different parameter spaces) and "chain." The scalar parameter chosen must have the same meaning across all models. They claim that the decompositions correspond to three pairs of variance estimates, with each member of a pair estimating the same quantity. Thus they propose following

the method of Brooks and Gelman (1996) by monitoring each of these 3 pairs and the 3 ratios they produce.

Brooks and Giudici do not specify how batches should be chosen for analysis. For simplicity of notation, we will consider calculations for one batch only. Suppose $C > 1$ chains of a RJMCMC sampler are run. Let θ be a scalar parameter in the chain (with equivalent interpretation across models), T denote the batch size, and M denote the total number of different models (different parameter spaces) visited by any chain for this batch. Define θ^r_{cm} as the r^{th} value of θ occurring in chain c and model m. Also define R_{cm} as the number of times model m occurs in chain c and $R_{\cdot m}$ as the number of times model m occurs across chains. Note that $R_{c\cdot} = T$ and the total number of sweeps in the batch over all chains is CT. Brooks and Giudici (1998) define the following quantities (note: the subscripts on the left-hand side are parts of the names, and do not correspond to values of indices on the right-hand side):

$$\begin{aligned}
\widehat{V}(\theta) &= \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^r_{cm}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2 \\
W_c(\theta) &= \frac{1}{C}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\frac{\left(\theta^r_{cm}-\bar{\theta}^{\cdot}_{c\cdot}\right)^2}{T-1} \\
W_m(\theta) &= \frac{1}{M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\frac{\left(\theta^r_{cm}-\bar{\theta}^{\cdot}_{\cdot m}\right)^2}{R_{\cdot m}-1} \\
W_mW_c(\theta) &= \frac{1}{CM}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\frac{\left(\theta^r_{cm}-\bar{\theta}^{\cdot}_{cm}\right)^2}{R_{cm}-1} \\
B_m(\theta) &= \sum_{m=1}^{M}\frac{\left(\bar{\theta}^{\cdot}_{\cdot m}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2}{M-1} \qquad (5.1) \\
B_mW_c(\theta) &= \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}\right)^2}{C(M-1)} \qquad (5.2)
\end{aligned}$$

where

$$\begin{aligned}
\bar{\theta}^{\cdot}_{cm} &= \frac{1}{R_{cm}} \sum_{r=1}^{R_{cm}} \theta^r_{cm} \\
\bar{\theta}^{\cdot}_{c\cdot} &= \frac{1}{T} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \theta^r_{cm} \\
\bar{\theta}^{\cdot}_{\cdot m} &= \frac{1}{R_{\cdot m}} \sum_{c=1}^{C} \sum_{r=1}^{R_{cm}} \theta^r_{cm} \\
\bar{\theta}^{\cdot}_{\cdot\cdot} &= \frac{1}{CT} \sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \theta^r_{cm}.
\end{aligned}$$

(Note: we have corrected two obvious typographical errors in the definitions of W_m and W_c).

Brooks and Giudici claim the following:

1. Both $\widehat{V}(\theta)$ and $W_c(\theta)$ should well approximate the true variation of θ under the stationary distribution of the Markov chain (and this comparison is essentially the original Gelman and Rubin comparison)

2. Both $W_m(\theta)$ and $W_m W_c$ should well approximate the true mean within-model variance

3. Both $B_m(\theta)$ and $B_m W_c$ should well approximate the true between-model variance.

It is true that, in the case of equal R_{cm} counts, these 6 quantities correspond to the descriptions they attach using ANOVA terminology. However, in the case of unequal R_{cm} counts, the meanings of the quantities are unclear. In general, the R_{cm} counts will be dramatically different, as some models are less likely than others and hence visited infrequently. Brooks and Giudici encounter this situation in their own example: the second and third comparisons break down when *one* of the chains visits a rare model *once* late in the sequence. It is easy to see why this occurs: the

comparisons are based on *unweighted* sample variances of means, allowing imprecise sample means from rare models to heavily influence their values. While it may be useful in some situations to have such diagnostics to detect rare model visits, we do not feel that this satisfies the definition of a *convergence* diagnostic. It is perfectly fine for some models to be more unlikely than others. We reconsider Brooks and Giudici's apparent initial motives and develop a strategy from scratch by considering appropriate two-way unbalanced ANOVA models.

5.4 A New Multivariate Strategy for RJMCMC

In this section we design a convergence diagnostic especially for RJMCMC situations in which different parameter spaces ("models") are indexed by some parameter in the chain. Our convergence diagnostic detects the following:

1. variation between chains (i.e., the target of the original Gelman and Rubin diagnostic: variation that is not homogeneous across chains),

2. an interaction between models and chains (i.e., between-model variation that differs from one chain to another), and

3. significant differences in the frequencies of model visits from one chain to another.

Any one of these three conditions would indicate that the chains are not living in the same stationary distribution, and hence that convergence has not occurred.

5.4.1 Forms of Variation Estimators

Suppose we have a RJMCMC sampler which produces output of a parameter vector Θ, with some $k \in \Theta$ indexing "model" and $\boldsymbol{\theta} \subset \Theta$ a vector of parameters

which retain the same meaning across models ($k \notin \boldsymbol{\theta}$). Let the output of $\boldsymbol{\theta}$ be represented as $\left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots\right)$. Suppose $C > 1$ chains of this sampler are run for the same number of sweeps. For simplicity of notation, we will consider output from one *batch* of size qb only, i.e., $\left(\boldsymbol{\theta}_1^{(qb+1)}, \ldots, \boldsymbol{\theta}_1^{(2qb)}\right), \ldots, \left(\boldsymbol{\theta}_C^{(qb+1)}, \ldots, \boldsymbol{\theta}_C^{(2qb)}\right)$ for some q and base batch size b. We now represent this collection in a more convenient notation (as in section 5.3.2), which we describe completely below.

Let

$$\boldsymbol{\theta} = \text{vector of parameters retaining same interpretation across models} \tag{5.3}$$

$$\theta = \text{arbitrary scalar component of } \boldsymbol{\theta} \tag{5.4}$$

$$C = \text{number of chains} \tag{5.5}$$

$$T = \text{batch size (this many sweeps per chain)} \tag{5.6}$$

$$M = \text{number of distinct models visited by any chain} \tag{5.7}$$

$$\boldsymbol{\theta}_{cm}^{r} = \text{value of } \boldsymbol{\theta} \text{ for } r^{\text{th}} \text{ occurrence of model } m \text{ in chain } c \tag{5.8}$$

$$R_{cm} = \text{number of times model } m \text{ occurred in chain } c \tag{5.9}$$

$$R_{\cdot m} = \sum_{c=1}^{C} R_{cm} \tag{5.10}$$

$$\bar{\boldsymbol{\theta}}_{cm}^{\cdot} = \frac{1}{R_{cm}} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}_{cm}^{r} \tag{5.11}$$

$$\bar{\boldsymbol{\theta}}_{c\cdot}^{\cdot} = \frac{1}{T} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}_{cm}^{r} \tag{5.12}$$

$$\bar{\boldsymbol{\theta}}_{\cdot m}^{\cdot} = \frac{1}{R_{\cdot m}} \sum_{c=1}^{C} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}_{cm}^{r} \tag{5.13}$$

$$\bar{\boldsymbol{\theta}}_{\cdot\cdot}^{\cdot} = \frac{1}{CT} \sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}_{cm}^{r}. \tag{5.14}$$

Our convergence diagnostic is based on the following estimates of variation: (note: the subscripts on the left-hand side are parts of the names, and do not correspond to values of indices on the right-hand side):

$$\widehat{V}(\theta) = \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta_{cm}^{r}-\bar{\theta}_{\cdot\cdot}^{\cdot}\right)^2 \tag{5.15}$$

$$W_c(\theta) = \frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta_{cm}^{r}-\bar{\theta}_{c\cdot}^{\cdot}\right)^2 \tag{5.16}$$

$$W_m(\theta) = \frac{1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta_{cm}^{r}-\bar{\theta}_{\cdot m}^{\cdot}\right)^2 \tag{5.17}$$

$$W_mW_c(\theta) = \frac{1}{C(T-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta_{cm}^{r}-\bar{\theta}_{cm}^{\cdot}\right)^2 \tag{5.18}$$

Note that these quantities may be interpreted as total variation ($\widehat{V}$), variation within chains (W_c), variation within models (W_m), and variation within models and chains (W_mW_c). The first of two comparisons we will use involves $\widehat{V}$ and W_c, which are defined in the same way as Brooks and Giudici (1998), and correspond (except for minor differences in multiplicative factors) to the original Gelman and Rubin diagnostic. The second involves W_m and W_mW_c, which are defined differently so as to correspond meaningfully to elements of appropriate ANOVA models. We establish these correpondences, for both pairs of variation estimates, in the next section.

5.4.2 Interpretation from an ANOVA Perspective

The output from the RJMCMC sampler can be considered as a collection of observations from a factorial design, in which the factors are "chain" and/or "model." An analysis of variance (ANOVA) can be used to assess the significance of factors and interactions. The primary exception to the usual assumptions of

ANOVA approaches is that the samples are not independent. However, we shall see that certain quantities constructed from ANOVA features are still useful in suggesting and interpreting our convergence diagnostics. Consider the three ANOVA models defined in Tables 5.1 – 5.3.

ANOVA 1

$$\boldsymbol{\theta}^r_{cm} = \mu + \alpha_c + e^r_{cm(1)}$$

where: $\alpha_c \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{ch}})$

$e^r_{cm(1)} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{er(ch)}})$

Source	df	SS
chain	$C-1$	$T\sum_{c=1}^{C}\left(\bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$
error(chain)	$C(T-1)$	$\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{c\cdot}\right)^2$
total	$CT-1$	$\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$

Table 5.1: ANOVA 1: One-way ANOVA with factor chain (random), balanced.

We represent model as a "fixed" factor and chain as "random," which is certainly debatable. However, basically the same conclusions are reached regardless of how the factors are treated (differing only in description of effects and minor coefficient changes). For example, if model were treated as random, effects would be described in terms of σ^2_{mo}, not the individual effects $\{\beta_m\}$. If chain were treated as random, effects would be described in terms of $\{\alpha_c\}$ instead of σ^2_{ch}.

Winer (1971, pp. 212 and 403) establishes the expressions for degrees of freedom entries. All terms which have the same notation in the three ANOVA's (e.g., $\mu, \alpha_c, \beta_m, \sigma^2_{\text{ch}}$) are *equivalent*. The error terms $(e^r_{cm(1)}, e^r_{cm(2)}, e^r_{cm(3)})$ are labeled

ANOVA 2

$\boldsymbol{\theta}^r_{cm} = \mu + \beta_m + e^r_{cm(2)}$

where: $\sum_{m=1}^{M} \beta_m = 0$

$e^r_{cm(2)} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{er(mo)}})$

"σ^2_{mo}" $= \frac{1}{M-1} \sum_{m=1}^{M} {\beta_m}^2$

Source	df	SS
model	$M-1$	$\sum_{m=1}^{M} R_{\cdot m} \left(\bar{\theta}^{\cdot}_{\cdot m} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$
error(model)	$CT-M)$	$\sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{\cdot m}\right)^2$
total	$CT-1$	$\sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$

Table 5.2: ANOVA 2: One-way ANOVA with factor model (fixed), unbalanced.

differently because they are in general *not* equivalent for the three models. In comparing entries in the ANOVA's with (5.15) – (5.18), it is clear that

$$\widehat{V} \equiv \text{MS}_{\text{tot}} \text{ for ANOVA 1,} \tag{5.19}$$

$$W_c \equiv \text{MS}_{\text{er(ch)}} \text{ for ANOVA 1,} \tag{5.20}$$

$$W_m \equiv \text{MS}_{\text{er(mo)}} \text{ for ANOVA 2, } and \tag{5.21}$$

$$W_m W_c \equiv \text{MS}_{\text{er(ch}\times\text{mo)}} \text{ for ANOVA 3,} \tag{5.22}$$

where "MS" denotes mean-square. We can of course *not* claim that an ANOVA model is a realistic description of the output from parallel chains of a RJMCMC sampler, since the assumptions of independence and normality in general do not hold. However, the effects of dependence are likely to be at least approximately cancelled out since we are focusing on ratios of mean-squares. The convergence diagnostics of Gelman and Rubin (1992); Brooks and Gelman (1996); and Brooks

ANOVA 3

$$\boldsymbol{\theta}^r_{cm} = \mu + \alpha_c + \beta_m + (\alpha\beta)_{cm} + e^r_{cm(3)}$$

where:

$$\alpha_c \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{ch}})$$

$$\textstyle\sum_{m=1}^{M} \beta_m = 0$$

$$(\alpha\beta)_{cm} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{ch}\times\text{mo}})$$

$$e^r_{cm(3)} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{er(ch}\times\text{mo)}})$$

$$\text{“}\sigma^2_{\text{mo}}\text{”} = \tfrac{1}{M-1}\textstyle\sum_{m=1}^{M} {\beta_m}^2$$

Source	df	SS
chain	$C-1$	$T\sum_{c=1}^{C}\left(\bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$
model	$M-1$	$\sum_{m=1}^{M} R_{\cdot m}\left(\bar{\theta}^{\cdot}_{\cdot m} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$
chain×model	$(C-1)(M-1)$	$\sum_{c=1}^{C}\sum_{m=1}^{M} R_{cm}\left(\bar{\theta}^{\cdot}_{cm} - \bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot m} + \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$
error(chain×model)	$C(T-M)$	$\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{cm}\right)^2$
total	$CT-1$	$\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2$

Table 5.3: ANOVA 3: Two-way ANOVA with factors model (fixed), chain (random) and chain×model interaction (random, unrestricted), balanced across chain only.

and Giudici (1998) all make this same implicit assumption. Furthermore, we will not rely on approximate normality for inferences. Thus we will proceed by considering the sampler output as occurring approximately according to an ANOVA model (not specifying yet which one(s), but using Tables 5.1 – 5.3 as appropriate).

Derivations of expected mean-squares for the three ANOVA models (shown in Appendix A.6) reveal that the expected values of (5.15) – (5.18) under ANOVA assumptions are given as follows:

$$E\widehat{V} = \sigma^2_{\text{er(ch)}} + \left[\frac{(C-1)T}{CT-1}\right]\sigma^2_{\text{ch}} \tag{5.23}$$

$$EW_c = \sigma^2_{\text{er(ch)}} \tag{5.24}$$

$$EW_m = \sigma^2_{\text{er(ch}\times\text{mo)}} \quad + \tag{5.25}$$

$$\left[\frac{(C-1)T}{CT-M} + \frac{1}{C^2(CT-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{(CR_{cm}-R_{\cdot m})^2}{R_{\cdot m}}\right]\sigma^2_{\text{ch}} + \tag{5.26}$$

$$\left[\frac{2}{C^2(CT-M)T}\sum_{m=1}^{M}\left\{\sum_{c=1}^{C}(CR_{cm}-R_{\cdot m})^2\right\}\beta^2_m\right] + \tag{5.27}$$

$$\left[\frac{CT}{CT-M} + \frac{-1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R^2_{cm}}{R_{\cdot m}} \quad + \right.$$

$$\left.\frac{2}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}(CR_{cm}-R_{\cdot m})\frac{R^2_{cm}}{R_{\cdot m}}\right]\sigma^2_{\text{ch}\times\text{mo}} \tag{5.28}$$

$$EW_mW_c = \sigma^2_{\text{er(ch}\times\text{mo)}} \tag{5.29}$$

If the set of within-chain model frequencies is equivalent for all chains (i.e., $R_{cm} = \frac{R_{\cdot m}}{C}\ \forall c, m$), then (5.25) – (5.28) simplifies to

$$EW_m = \sigma^2_{\text{er(ch}\times\text{mo)}} + \left[\frac{(C-1)T}{CT-M}\right]\sigma^2_{\text{ch}} + \left[\frac{(C-1)T}{CT-M}\right]\sigma^2_{\text{ch}\times\text{mo}}. \tag{5.30}$$

For large T (and any $\{R_{cm}\}$),

$$E\widehat{V} \approx \sigma^2_{\text{er(ch)}} + \left[\frac{C-1}{C}\right]\sigma^2_{\text{ch}} \tag{5.31}$$

and

$$EW_m \approx \sigma^2_{\text{er(ch}\times\text{mo)}} + \left[\frac{C-1}{C}\right]\sigma^2_{\text{ch}} + \tag{5.32}$$

$$\left[\frac{1}{C^2(CT-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{(CR_{cm}-R_{\cdot m})^2}{R_{\cdot m}}\right]\sigma^2_{\text{ch}} + \tag{5.33}$$

$$\left[\frac{2}{C^3T^2-C^2MT}\sum_{m=1}^{M}\left\{\sum_{c=1}^{C}(CR_{cm}-R_{\cdot m})^2\right\}\beta^2_m\right] + \tag{5.34}$$

$$\left[1 + \left(\frac{-1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R^2_{cm}}{R_{\cdot m}}\right) + \right.$$

$$\left(\frac{2}{CT} \sum_{c=1}^{C} \sum_{m=1}^{M} \left(C R_{cm} - R_{\cdot m} \right) \frac{R_{cm}^2}{R_{\cdot m}} \right) \Bigg] \sigma^2_{\text{ch}\times\text{mo}}. \tag{5.35}$$

If additionally the set of within-chain model frequencies is equivalent for all chains ($R_{cm} = \frac{R_{\cdot m}}{C}\ \forall c, m$), then (5.32) – (5.35) simplifies to (for large T):

$$EW_m \ \approx \ \sigma^2_{\text{er(ch}\times\text{mo)}} + \left[\frac{C-1}{C} \right] \sigma^2_{\text{ch}} + \left[\frac{C-1}{C} \right] \sigma^2_{\text{ch}\times\text{mo}}. \tag{5.36}$$

Notice that (5.26) and (5.27), in the presence of chain and model effects, respectively, increase as the model frequencies across chains, R_{cm}, deviate more from the frequencies $\frac{R_{\cdot m}}{C}$ that would occur if the set of within-chain model frequencies were equivalent for all chains.

The expression (5.28) can be characterized as follows. Let

$$X = X_1 + X_2 + X_3$$

where

$$\begin{aligned} X_1 &= \frac{CT}{CT - M} \\ X_2 &= \frac{-1}{CT - M} \sum_{c=1}^{C} \sum_{m=1}^{M} \frac{R_{cm}^2}{R_{\cdot m}} \\ X_3 &= \frac{2}{CT} \sum_{c=1}^{C} \sum_{m=1}^{M} \left(C R_{cm} - R_{\cdot m} \right) \frac{R_{cm}^2}{R_{\cdot m}}. \end{aligned}$$

The ranges of X_2 and X_3 can be determined from consideration of two extreme cases,

$$A. \qquad R_{cm} = \frac{R_{\cdot m}}{C}\ \forall c, \quad \text{and}$$

$$B. \qquad R_{cm} = \begin{cases} R_{\cdot m}, & \text{for } c = c'_m \\ 0, & \text{for } c \neq c'_m \end{cases} \qquad \text{for some } \{c'_1, \ldots, c'_M\},$$

to be

$$\frac{-CT}{CT - M} \leq X_2 \leq \frac{-T}{CT - M}$$

and

$$0 \leq X_3 \leq \frac{2(C-1)}{CT} \sum_{m=1}^{M} R_{\cdot m}^2 \quad .$$

Also, X is strictly positive because (i) if case A holds, then $X = \frac{(C-1)T}{CT-M}$, and (ii) if case A does not hold, then X_3 is strictly positive. In general, X increases (although not necessarily monotonically) as the set of within-cell model frequencies becomes less homogeneous across chains.

Thus we can conclude the following about the ratios $\frac{E\widehat{V}}{EW_c}$ and $\frac{EW_m}{EW_mW_c}$:

1. $\frac{E\widehat{V}}{EW_c} \geq 1$, with $\frac{E\widehat{V}}{EW_c} = 1$ indicating the absence of a chain effect. The greater $\frac{E\widehat{V}}{EW_c}$, the stronger the chain effect, with each term in the numerator and denominator stabilizing as $T \to \infty$ and thus preserving the validity of the magnitude as $T \to \infty$.

2. $\frac{EW_m}{EW_mW_c} \geq 1$, with $\frac{EW_m}{EW_mW_c} = 1$ indicating:

   (a) the absence of a chain effect, *and*

   (b) the absence of a chain×model interaction, *and*

   (c) either (i) no model effect or (ii) equality of the set of within-chain model frequencies across chains, or both.

   The greater the violation of any combination of these three criteria (2a)–(2c), the larger $\frac{EW_m}{EW_mW_c}$ becomes. The relative weights of the three criteria as $T \to \infty$ (i.e., the sensitivity of the ratio to violations of each of the three criteria) are not yet fully understood. We can at least reason by (5.36) that when the set of within-chain model frequencies are somewhat homogeneous across chains (i.e., $CR_{cm} \approx R_{\cdot m}\ \forall c, m$), then the ratio has approximately equal sensitivity to (2a) and (2b), and so either a significant chain effect or chain×model interaction should be detected.

   These properties clearly suggest the design of a convergence diagnostic based

on the two ratios $\frac{\widehat{V}}{W_c}$ and $\frac{W_m}{W_m W_c}$. We suggest the use of both ratios, because it may help to narrow down the cause of any violations of convergence. In the next section, we show the exact form of the diagnostic technique we propose.

Expressions analogous to (5.1) and (5.2), represented with the proper degrees of freedom terms in the denominators, yield expected mean-squares that do not appear to be useful for comparison purposes. Further research is needed to determine what a ratio based on analogues of (5.1) and (5.2) would actually detect.

5.4.3 The Convergence Assessment Strategy

Define the following multivariate versions of (5.15) – (5.18):

$$\widehat{V}(\boldsymbol{\theta}) = \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{\cdot\cdot}^{\cdot}\right)\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{\cdot\cdot}^{\cdot}\right)' \tag{5.37}$$

$$W_c(\boldsymbol{\theta}) = \frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{c\cdot}^{\cdot}\right)\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{c\cdot}^{\cdot}\right)' \tag{5.38}$$

$$W_m(\boldsymbol{\theta}) = \frac{1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{\cdot m}^{\cdot}\right)\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{\cdot m}^{\cdot}\right)' \tag{5.39}$$

$$W_m W_c(\boldsymbol{\theta}) = \frac{1}{C(T-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{cm}^{\cdot}\right)\left(\boldsymbol{\theta}_{cm}^{r}-\bar{\boldsymbol{\theta}}_{cm}^{\cdot}\right)' \tag{5.40}$$

Define the following set of potential scale reduction factors, for a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, using (5.15)–(5.18) and (5.37)–(5.40):

$$PSRF_1(\theta_i) = \frac{\widehat{V}(\theta_i)}{W_c(\theta_i)} \tag{5.41}$$

$$PSRF_2(\theta_i) = \frac{W_m(\theta_i)}{W_m W_c(\theta_i)} \tag{5.42}$$

$$MPSRF_1(\boldsymbol{\theta}) = \text{maximum eigenvalue of } [W_c(\boldsymbol{\theta})]^{-1}\,\widehat{V}(\boldsymbol{\theta}) \tag{5.43}$$

$$MPSRF_2(\boldsymbol{\theta}) = \text{maximum eigenvalue of } [W_m W_c(\boldsymbol{\theta})]^{-1}\,W_m(\boldsymbol{\theta}). \tag{5.44}$$

By Lemmas 5.3.1 and 5.3.2, we have that

$$MPSRF_1(\boldsymbol{\theta}) \geq \max_i PSRF_1(\theta_i) \text{ and } MPSRF_2(\boldsymbol{\theta}) \geq \max_i PSRF_2(\theta_i). \tag{5.45}$$

Our convergence assessment technique consists of the following steps:

Algorithm 5.4.1 (RJMCMC Convergence Assessment) *Implement the following procedure as a convergence assessment technique for RJMCMC applied to a model with parameters* Θ*, using* (5.15)–(5.18), (5.37)–(5.40) *and* (5.41)–(5.44)*:*

1. *Identify a parameter* $k \in \Theta$ *which is an indicator of "model" and select a parameter vector* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \subset \Theta$ *consisting of quantities which have the same interpretation across* k *(but with* $k \notin \boldsymbol{\theta}$*).*

2. *Simulate* $C > 1$ *chains of equal length* T *via RJMCMC, with over-dispersed starting values.*

3. *Choose a base batch size* b *(Brooks and Gelman (1996) recommend, for example,* $b \approx \frac{T}{20}$*).*

4. *Let the notation* $S^{(q)}(\cdot)$ *represent a statistic* S *computed for the* q^{th} *batch* $(\boldsymbol{\theta}_1^{(qb+1)}, \ldots, \boldsymbol{\theta}_1^{(2qb)})$, $\ldots$, $(\boldsymbol{\theta}_C^{(qb+1)}, \ldots, \boldsymbol{\theta}_C^{(2qb)})$. *For batches* $q = 1, \ldots, \frac{T}{b}$*, do the following:*

   (a) *Plot* ${MPSRF_1}^{(q)}(\boldsymbol{\theta})$ *vs.* q *and* ${MPSRF_2}^{(q)}(\boldsymbol{\theta})$ *vs.* q *(separately or together).*

   (b) *Plot the maximum eigenvalues of* $\widehat{V}^{(q)}(\boldsymbol{\theta})$ *and* ${W_c}^{(q)}(\boldsymbol{\theta})$ *together vs.* q*.*

   (c) *Plot the maximum eigenvalues of* ${W_m}^{(q)}(\boldsymbol{\theta})$ *and* ${W_m W_c}^{(q)}(\boldsymbol{\theta})$ *together vs.* q*.*

   (d) *Optionally plot* ${PSRF_1}^{(q)}(\theta_i)$ *vs.* q *and* ${PSRF_2}^{(q)}(\theta_i)$ *vs.* q*.*

   (e) *Optionally plot the numerator and denominator of* ${PSRF_1}^{(q)}(\theta_i)$ *together vs.* q*.*

(f) *Optionally plot the numerator and denominator of* $PSRF_2{}^{(q)}(\theta_i)$ *together vs.* q.

5. *Determine* q_0 *such that for* $q \geq q_0$ *the plots in Step 4a have settled close to 1,* and *the plots in Step 4b have settled approximately to a common value,* and *the plots in Step 4c have settled approximately to a common value.*

6. *Discard the first* $q_0 b$ *sweeps from each chain, and then pool the remaining ones together to use for inference.*

We prefer the maximum eigenvalue to the determinant for monitoring individual matrices, since it is on the same scale as the univariate variance estimates and hence can conveniently be displayed in the same plot. The method can be performed on more than one parameter vector $\boldsymbol{\theta}$. It may be useful to use a collection of related sets of scalar parameters in order to target which sets are mixing faster than others. The purpose of the $MPSRF$ is to provide a safe (conservative) alternative to the monitoring of a large number of scalar parameters individually. However, the individual scalar parameters can still be monitored (Steps 4d–4f), providing more detailed information.

For our BVNPCP-BHM(A, n) model, we monitor two collections of parameters, $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, which are defined in section 5.1. The associated plots for the Redwood data and all simulated patterns, for chains of length $20,000$ (runs of length $200,000$ where every 10^{th} sweep is saved) are displayed in Figures E.1 – E.13. Very fast convergence is implied in each case. Further research is needed to study the performance of this convergence assessment technique, since we have only applied it to a small collection of similar datasets for only one type of model.

CHAPTER 6
RJMCMC OUTPUT ANALYSIS

In this chapter we present the details of methods used for analysis of a post-convergent RJMCMC sample (convergence being determined by Algorithm 5.4.1). Results for these methods as applied to our real and simulated data sets, along with those from the composite EM technique, are discussed in Chapter 7.

6.1 Notation

Suppose we have run a RJMCMC sampler for the BVNPCP-BHM(A, n) and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is some p-dimensional subset of parameters of the model (as opposed to Chapter 4, where we took $\boldsymbol{\theta}$ to represent all unknowns). Assume convergence assessment has been implemented and a collection of values from a *total* of T post-convergent sweeps from C chains ($\frac{T}{C}$ sweeps from each chain) is to be used for inference. Some methods need to differentiate between output from different chains, while others need to differentiate between output with different k, while still others analyze all output collectively. So, we will use the following sets of notation as appropriate, sometimes interchangeably provided the meaning is clear:

$$\begin{aligned}
\boldsymbol{\theta}^{(\cdot)} &= \text{collection of } T \text{ values of } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \text{ consisting of} \\
&\qquad \tfrac{T}{C} \equiv T_{\text{ch}} \text{ post-convergent sweeps from each of } C \text{ chains} \\
&\equiv \left(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}\right) \\
&\equiv \left(\left(k^{(1)}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}, \mathbf{Z}^{(1)}\right), \ldots, \left(k^{(T)}, \boldsymbol{\mu}^{(T)}, \boldsymbol{\Sigma}^{(T)}, \mathbf{Z}^{(T)}\right)\right)
\end{aligned}$$

$$\equiv \left(\boldsymbol{\theta}^{(1;1)}, \ldots, \boldsymbol{\theta}^{(T_{\rm ch};1)}, \ \ldots \ , \boldsymbol{\theta}^{(1;C)}, \ldots, \boldsymbol{\theta}^{(T_{\rm ch};C)}\right)$$

and

$$\begin{aligned}
\boldsymbol{\theta}^{(\cdot;c)} &= \left(\boldsymbol{\theta}^{(1;c)}, \ldots, \boldsymbol{\theta}^{(T_{\rm ch};c)}\right) \\
\bar{\boldsymbol{\theta}}^{(\cdot;c)} &= \frac{1}{T_{\rm ch}} \sum_{t=1}^{T_{\rm ch}} \boldsymbol{\theta}^{(t;c)} \\
\bar{\boldsymbol{\theta}}^{(\cdot)} &= \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\theta}^{(t)},
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\theta}^{(\cdot|k)} &= \text{ set of values at the } T_k \text{ sweeps for which } k^{(t)} = k \\
&\equiv \left(\boldsymbol{\theta}^{(1|k)}, \ldots, \boldsymbol{\theta}^{(T_k|k)}\right).
\end{aligned}$$

6.2 Preliminaries: Tools for Analysis

6.2.1 Autocorrelation Function

The *lag-g autocorrelation* of two scalar components θ_i and θ_j (i could equal j) from chain c is defined as the correlation between $\theta_i^{(t)}$ and $\theta_j^{(t+g)}$. It is estimated by the *autocorrelation function (ACF)*:

$$ACF_g\left(\theta_i^{(\cdot;c)}, \theta_j^{(\cdot;c)}\right) = \frac{\sum_{t=1}^{T_{\rm ch}-g} \left(\theta_i^{(t;c)} - \bar{\theta}_i^{(\cdot;c)}\right)\left(\theta_j^{(t+g;c)} - \bar{\theta}_j^{(\cdot;c)}\right)}{\sum_{t=1}^{T_{\rm ch}} \left(\theta_i^{(t;c)} - \bar{\theta}_i^{(\cdot;c)}\right)\left(\theta_j^{(t;c)} - \bar{\theta}_j^{(\cdot;c)}\right)}. \tag{6.1}$$

Note that in general $ACF_g\left(\theta_i^{(\cdot;c)}, \theta_j^{(\cdot;c)}\right) \neq ACF_g\left(\theta_j^{(\cdot;c)}, \theta_i^{(\cdot;c)}\right)$.

Define $ACF_g\left(\theta_i^{(\cdot;c)}\right)$ as $ACF_g\left(\theta_i^{(\cdot;c)}, \theta_i^{(\cdot;c)}\right)$ and

$ACF_g\left(\boldsymbol{\theta}^{(\cdot;c)}\right) = p \times p$ matrix with $(i,j)^{\rm th}$ entry $ACF_g\left(\theta_i^{(\cdot;c)}, \theta_j^{(\cdot;c)}\right)$. (6.2)

6.2.2 Batch Sampling

One of the methods we will use to construct approximate confidence intervals and regions for RJMCMC parameters requires estimation of posterior variances.

Let Φ^2 be the true posterior variance of a parameter θ_i. The usual sample variance estimator

$$\frac{1}{T-1}\sum_{t=1}^{T}\left(\theta_i^{(t)}-\bar{\theta}_i^{(\cdot)}\right)^2$$

will tend to underestimate Φ^2 if there is significant positive autocorrelation in the chain (which there typically will be). One possible remedy is to separate $\theta_i^{(\cdot)}$ into batches (consecutive within each chain) and then compute the mean of each batch and the sample variance of the batch means. These means should exhibit less autocorrelation than the samples themselves.

Suppose we separate $\theta_i^{(\cdot)}$ into m batches (consecutive within each chain), each of size b, with the sample means of the batches denoted

$$\bar{\theta}_i^{(\cdot)(1)},\ \ldots\ ,\bar{\theta}_i^{(\cdot)(m)}.$$

Then if these batch means are relatively uncorrelated,

$$\widehat{\Phi}^2_{\mathrm{BS}}\left(\theta_i^{(\cdot)}\right)=\frac{b}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_i^{(\cdot)(j)}-\bar{\theta}_i^{(\cdot)}\right)^2$$

is a better estimate of Φ^2 (Roberts, 1995, p. 50). Usually the strongest correlation between batch means is the lag-1 autocorrelation. Ripley (1987, p. 155) suggests choosing b large enough so that the ACF_1 of batch means is below 0.05. Since we wish to use samples from different chains, we compute the ACF_1 for a given batch size separately in each chain, and choose b large enough so that the ACF_1 of size-b batch means is below 0.05 for all chains. For each candidate b we compute $m=\lfloor\frac{T_{\mathrm{ch}}}{b}\rfloor C$ and $r=$ integer remainder of $\frac{T_{\mathrm{ch}}}{b}$, where "$\lfloor\cdot\rfloor$" denotes "greatest integer less than or equal to", and then use $\frac{m}{C}$ batches of size b from each chain, ignoring the first r sweeps of each chain, to compute the C ACF_1's.

We start with $b=10$ and increment by 10 (instead of 1, to save computation time) until we encounter one (b_0, say) which achieves the $ACF_1=0.05$ cutoff for all C chains. Then we see if we can increase b_0 (to b_1, say) and still maintain the same

number of batches (m_1, say). (Since we only try multiples of 10 for b, the remainder r described above may be large enough at b_0 so that we can increase the batch size and still maintain the same number of batches, thus using as much of the output as possible). Finally, we then combine all batch means for all chains and estimate Φ^2 by

$$\widehat{\mathrm{Var}}_{\mathrm{BS}}\left(\theta^{(\cdot)}\right) = \widehat{\Phi}^2_{\mathrm{BS}}\left(\theta^{(\cdot)}\right) = \frac{b_1}{m_1 - 1}\sum_{j=1}^{m_1}\left(\bar{\theta}_i^{(\cdot)(j)} - \bar{\theta}_i^{(\cdot)}\right)^2. \tag{6.3}$$

The number of batches, m_1, should be large enough for this estimate to have reasonable accuracy. In our analyses, we make sure that $m_1 \geq 12$; we are able to find a suitable b_1 in each case, and in most cases m_1 is much larger than 12.

An analogous vector (multivariate) version of the batch sampling variance estimate can also be calculated. The choice of batch size is not as straightforward, however. There may be (cross)-autocorrelation between different scalar components of $\boldsymbol{\theta}$, and so the matrix form of ACF_1, (6.2), must be checked. Due to random fluctuations in the ACF, it is very difficult to find a batch size for which all $C \times p^2 ACF_1$'s fall below the suggested cutoff. We experimented with independently generated sequences (theoretical $ACF_g = 0 \ \forall \ g$) and encountered a surprisingly large amount of variation in batch mean ACF_1's. Thus we follow a different strategy than in the scalar (univariate) case and choose a batch size b_0 such that each cross-autocorrelation estimate in each chain has fallen below the cutoff of 0.05 at *some* point in the past, i.e. for some $b \leq b_0$. As before, we increase b_0 if possible to b_1 to obtain the largest possible batch size for the same number of batches (m_1) corresponding to b_0. Let $\boldsymbol{\Phi}$ be the true posterior variance-covariance matrix of a parameter vector $\boldsymbol{\theta}$. We then estimate $\boldsymbol{\Phi}$ by

$$\widehat{\mathrm{Var}}_{\mathrm{BS}}\left(\boldsymbol{\theta}^{(\cdot)}\right) = \widehat{\boldsymbol{\Phi}}_{\mathrm{BS}}\left(\boldsymbol{\theta}^{(\cdot)}\right) = \frac{b_1}{m_1 - 1}\sum_{j=1}^{m_1}\left(\bar{\boldsymbol{\theta}}_i^{(\cdot)(j)} - \bar{\boldsymbol{\theta}}^{(\cdot)}\right)\left(\bar{\boldsymbol{\theta}}_i^{(\cdot)(j)} - \bar{\boldsymbol{\theta}}^{(\cdot)}\right)'. \tag{6.4}$$

6.2.3 Circular Data Methods

When the anisotropy parameterization of $\mathbf{\Sigma}$ is used (see Definition 1.1.10), the parameter ϕ must receive special treatment since it is a circular (or, directional or angular) parameter. Actually, ϕ is an *axial* parameter since it takes on values in an interval of length π rather than 2π. For now, consider a circular random variable η which takes on values in $[0, 2\pi)$. We will discuss special accommodations for an axial variable at the end of this section. For η, values near 0 should be considered "close" to those near 2π. Fisher (1993) and Mardia (1972) provide a wealth of methods for the analysis of circular data, and we present a review of relevant techniques here.

Let $\eta_1, \ldots, \eta_n$ be an independent random sample from some circular distribution defined on $[0, 2\pi)$. The analogue of a mean for linear data is referred to as the *circular mean*, or *mean direction*. Suppose the true mean direction is ω. The *sample circular mean* $\bar{\eta}$ is defined as

$$\bar{\eta} = \widehat{\omega}(\boldsymbol{\eta}) = \begin{cases} \arctan\left(\frac{\mathcal{S}}{\mathcal{C}}\right), & \text{if } \mathcal{S} > 0 \text{ and } \mathcal{C} > 0 \\ \arctan\left(\frac{\mathcal{S}}{\mathcal{C}}\right) + \pi, & \text{if } \mathcal{C} < 0 \\ \arctan\left(\frac{\mathcal{S}}{\mathcal{C}}\right) + 2\pi, & \text{if } \mathcal{S} < 0 \text{ and } \mathcal{C} > 0 \end{cases} \tag{6.5}$$

where

$$\mathcal{C} = \sum_{i=1}^{n} \cos(\eta_i) \quad \text{and} \quad \mathcal{S} = \sum_{i=1}^{n} \sin(\eta_i)$$

The p^{th} centered sample trigonometric moment is defined as

$$\mathcal{M}_p = \frac{1}{n}\sum_{j=1}^{n} \cos[p(\eta_j - \bar{\eta})] + i\frac{1}{n}\sum_{j=1}^{n} \sin[p(\eta_j - \bar{\eta})].$$

The first two are

$$\bar{\mathcal{R}} = \mathcal{M}_1 = \frac{1}{n}\sqrt{\mathcal{C}^2 + \mathcal{S}^2} = \frac{1}{n}\sum_{i=1}^{n} \cos(\eta_i - \bar{\eta})$$

(also called the *mean resultant length*) and

$$\widehat{\aleph}_2 = \mathcal{M}_2 = \frac{1}{n}\sum_{i=1}^{n} \cos[2(\eta_i - \bar{\eta})].$$

A commonly used measure of spread is the *sample circular dispersion*,

$$\widehat{\delta}(\boldsymbol{\eta}) = \frac{1 - \widehat{\aleph}_2}{2\bar{\mathcal{R}}^2}.$$

A nonparametric confidence interval for the mean direction ω is (see Fisher, 1993, p. 76)

$$\bar{\eta} \pm \arcsin\left(z_{\frac{\alpha}{2}} \left(\frac{\widehat{\delta}}{n} \right)^{\frac{1}{2}} \right). \tag{6.6}$$

Note: if the computed value of $z_{\frac{\alpha}{2}} \left(\frac{\widehat{\delta}}{n} \right)^{\frac{1}{2}}$ is greater than 1, this confidence interval is ill-defined, but at least covers $\left[\bar{\eta} - \frac{\pi}{2}, \bar{\eta} + \frac{\pi}{2} \right]$. We are not interested in the variance of the posterior *mean* of ϕ, but rather the variance of the posterior distribution of ϕ. Thus we will use confidence intervals of the form $\bar{\eta} \pm \arcsin\left(z_{\frac{\alpha}{2}} \left(\widehat{\delta} \right)^{\frac{1}{2}} \right)$.

An analogue of the linear correlation coefficient for circular data is the *sample circular correlation coefficient*. For a paired sample of directions $(\eta_1, \zeta_1), \ldots, (\eta_n, \zeta_n)$, it is defined as (Fisher and Lee, 1983)

$$\widehat{\rho}_{\mathrm{T}}(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \frac{\sum_{1 \le i < j \le n} \sin(\eta_i - \eta_j) \sin(\zeta_i - \zeta_j)}{\left[\left(\sum_{1 \le i < j \le n} \sin^2(\eta_i - \eta_j) \right) \left(\sum_{1 \le i < j \le n} \sin^2(\zeta_i - \zeta_j) \right) \right]^{\frac{1}{2}}}.$$

The value of $\widehat{\rho}_{\mathrm{T}}$ lies in $[-1, 1]$, with $\widehat{\rho}_{\mathrm{T}} = 1 \;\Rightarrow\; \eta = (\zeta + d_0) \pmod{2\pi}$ for some d_0 and $\widehat{\rho}_{\mathrm{T}} = -1 \;\Rightarrow\; \eta = (-\zeta + d_0) \pmod{2\pi}$ for some d_0.

Also, a circular analogue of the ACF_g is given by

$$\begin{aligned} \widehat{\rho}_{\mathrm{T},g}(\boldsymbol{\eta}) &= \widehat{\rho}_{\mathrm{T}} \text{ calculated from } (\eta_1, \eta_{g+1}), \ldots, (\eta_{n-g}, \eta_n) \\ &= \frac{\sum_{1 \le i < j \le n-g} \sin(\eta_i - \eta_j) \sin(\eta_{i+g} - \eta_{j+g})}{\left[\left(\sum_{1 \le i < j \le n-g} \sin^2(\eta_i - \eta_j) \right) \left(\sum_{1 \le i < j \le n-g} \sin^2(\eta_{i+g} - \eta_{j+g}) \right) \right]^{\frac{1}{2}}}. \end{aligned}$$

Batch sampling can be implemented for circular data as well, to estimate the circular dispersion δ. Suppose that $\eta \in \boldsymbol{\theta}$ is a circular parameter sampled in RJMCMC. A batch size b_0 is chosen large enough so that $\widehat{\rho}_{\mathrm{T},1}\left(\eta^{(\cdot;c)}\right) \le 0.05$ for

$c = 1, \ldots, C$ (and increased to b_1 to keep the same number of batches $m_1 \geq 12$, as before). Then the estimate is calculated from the sample circular means of the batches:

$$\widehat{\delta}_{\mathrm{BS}}(\eta^{(\cdot)}) = b_1 \widehat{\delta}\left(\widehat{\omega}(\eta^{(\cdot)(1)}), \ldots, \widehat{\omega}(\eta^{(\cdot)(m_1)})\right).$$

If a circular variable η is confined to $[a, b]$ where $|a-b| = \frac{2\pi}{p}$, it is called *p-axial*. For example, ϕ as defined in Definition 1.1.10 is 2-axial with $a = -\frac{\pi}{2}$ and $b = \frac{\pi}{2}$. Analysis of p-axial data is performed by first transforming to 1-axial ("vectorial") data:

$$\eta \longrightarrow \eta^\star = [p(\eta - a)] \pmod{2\pi}, \tag{6.7}$$

performing all analyses on the vectorial data, and then back-transforming the results (e.g., confidence interval endpoints) back to p-axial form via

$$\eta^\star \longrightarrow \eta = \frac{\eta^\star}{p} + a \tag{6.8}$$

Fisher (see 1993, p. 37).

6.2.4 Posterior Density Estimates

Perhaps the most useful and descriptive display of RJMCMC output is via posterior density estimates. For the parameter k, these take the form of simple histograms. For a sample $\theta^{(\cdot)}$ of a continuous linear parameter (perhaps bounded), we calculate a nonparametric density estimate according to the `density` function in S-Plus 4.5 for Windows (Mathsoft, Inc.), using default options. This employs a Gaussian window of width $\mathcal{W} = \frac{\mathrm{range}(\theta^{(\cdot)})}{\log_2(T)+1}$. The density estimate is evaluated at 50 equally spaced points in the range $[\min(\theta^{(\cdot)}) - 0.75\mathcal{W}\ ,\ \ \max(\theta^{(\cdot)}) + 0.75\mathcal{W}]$.

A nonparametric density estimate can also be computed for samples of the anisotropy direction $\phi^{(\cdot)}$. For a sample of circular data $(\eta_1, \ldots, \eta_n) \in [0, 2\pi)$, a

quartic kernel is used (Fisher, 1993, p. 26):

$$w(x) = \begin{cases} 0.9375(1-x^2)^2, & \text{if } -1 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

The bandwidth h_0 for $n > 15$ is chosen as

$$h_0 = \sqrt{7} n^{-\frac{1}{5}} \widehat{\kappa}^{-\frac{1}{2}}$$

where

$$\widehat{\kappa} = \begin{cases} 2\bar{\mathcal{R}} + \bar{\mathcal{R}}^3 + \frac{5}{6}\bar{\mathcal{R}}^5, & \text{if } \bar{\mathcal{R}} < 0.53 \\ -0.4 + 1.39\bar{\mathcal{R}} + \frac{0.43}{1-\bar{\mathcal{R}}}, & \text{if } 0.53 \le \bar{\mathcal{R}} < 0.85 \\ \frac{1}{\bar{\mathcal{R}}^3 - 4\bar{\mathcal{R}}^2 + 3\bar{\mathcal{R}}}, & \text{if } \bar{\mathcal{R}} \ge 0.85. \end{cases}$$

Then the nonparametric density estimate at a point x is

$$\widehat{d}(x) = \frac{1}{nh_0} \sum_{i=1}^{n} w\left(\frac{\min(|x-\eta_i|, 2\pi - |x-\eta_i|)}{h_0} \right). \tag{6.9}$$

We choose to evaluate (6.9) at 128 equally spaced points in $[0, 2\pi)$. The density estimate at x for $\phi^{(\cdot)}$ is computed as $\widehat{d}\left(2\left[x + \frac{\pi}{2}\right]\right)$.

6.3 Assessment of Model Adequacy

Before proceeding to conduct inference using post-convergent RJMCMC output, it is wise to perform some type of "model adequacy" check to see if the data conforms to the BVNPCP-BHM assumptions. Since k really indexes different models possessing different parameter sets, and we use a vague prior for k, it makes more sense to assess model adequacy separately for each k. Many methods are available to perform model-checking using MCMC output. We can apply these methods to subsets of RJMCMC output separated by k. The label-switching issue (see section 4.4.6) is not a deterrent for any model-checking approaches, since we will consider each sweep of the chain as a separate instance of the model. We explore two different paradigms for model adequacy assessment: (a) use of discrepancy

measures with posterior predictive densities, and (b) cross-validation.

6.3.1 Posterior Predictive Densities and Discrepancy Measures

For this section let $\mathbf{Y}^{\text{obs}}$ denote the *observed* values of offspring locations and $\mathbf{Y}^{\star}$ denote a *replication* of $\mathbf{Y}$ with the same sample size, n. The *posterior predictive density* $p(\mathbf{Y}^{\star}|\mathbf{Y}^{\text{obs}})$ describes the marginal distribution of the locations of a new set of offspring conditional on the observed offspring locations. A *discrepancy measure* $D(\mathbf{Y};\boldsymbol{\theta})$ measures disagreement between the data $\mathbf{Y}$ and model with parameters $\boldsymbol{\theta}$; it may reduce to $D(\mathbf{Y})$, which measures deviation of $\mathbf{Y}$ from model assumptions inherent for all $\boldsymbol{\theta}$.

If the distribution of $D(\mathbf{Y};\boldsymbol{\theta})$ under the model assumptions is known, then a quick assessment can be implemented by computing $D\left(\mathbf{Y}^{\text{obs}};\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right)$, where $\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}$ is the mode of $p\left(\mathbf{Y}^{\text{obs}}\left|\boldsymbol{\theta}^{(\cdot|k)}\right.\right)$, i.e., the maximizer of the model likelihood over all posterior samples. This choice of $\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}$ seems sensible since it represents the "best" model for this k. A *Bayesian p-value* can be computed as

$$p_D = P\left(D\left(\mathbf{Y}^{\star};\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right) \geq D\left(\mathbf{Y}^{\text{obs}};\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right)\right). \tag{6.10}$$

If p_D is close to 0, then the discrepancy between $\mathbf{Y}^{\text{obs}}$ and $\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}$ is excessive; if p_D is close to 1, then the discrepancy is significantly less than would be expected under natural sampling variability. Either extreme indicates a poor fit to the model.

Another method to obtain a Bayesian p-value, Monte Carlo-style, is to compute $D\left(\mathbf{Y}^{\text{obs}};\boldsymbol{\theta}^{(t|k)}\right)$ and $D\left(\mathbf{Y}^{\star(t)};\boldsymbol{\theta}^{(t|k)}\right)$ for $t = 1,\ldots,T_k$, where $\mathbf{Y}^{\star(t)}$ is a sample from $p\left(\mathbf{Y}^{\star}\left|\boldsymbol{\theta}^{(t|k)}\right.\right)$. A Bayesian p-value with similar interpretation as (6.10), except that the discrepancy measured is between $\mathbf{Y}^{\text{obs}}$ and $\boldsymbol{\theta}^{(\cdot)}$ overall, is

$$p_D = \frac{1}{T_k}\sum_{t=1}^{T_k} \mathrm{I}\left[D\left(\mathbf{Y}^{\star(t)};\boldsymbol{\theta}^{(t|k)}\right) \geq D\left(\mathbf{Y}^{\text{obs}};\boldsymbol{\theta}^{(t|k)}\right)\right], \tag{6.11}$$

the proportion of times that data randomly generated under model assumptions displays greater discrepancy with $\boldsymbol{\theta}^{(t|k)}$ than does the observed data.

As far as choices of forms of $D(\cdot;\cdot)$ are concerned, Gelman, Carlin, Stern, and Rubin (1995, p. 172) recommend using more than one and trying to "reflect aspects of the model that are relevant to scientific purposes to which the inference will be applied." Our approach uses two forms of $D(\cdot;\cdot)$: one a goodness-of-fit statistic for bivariate normality, $D_{\mathrm{CR}}(\cdot;\cdot)$, and the other a measure $D_{\boldsymbol{\Sigma}}(\cdot;\cdot)$ of the discrepancy between $\widehat{\boldsymbol{\Sigma}}$ estimated from the data (given $\mathbf{Z}^{(t|k)}$, but not $\boldsymbol{\mu}^{(t|k)}$ or $\boldsymbol{\Sigma}^{(t|k)}$) and $\boldsymbol{\Sigma}^{(t|k)}$.

The first, $D_{\mathrm{CR}}\left(\mathbf{Y};\boldsymbol{\theta}^{(t|k)}\right)$, is based on the bivariate normality goodness-of-fit technique given in Johnson and Wichern (1992, pp. 158–164). Basically it estimates $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on $\mathbf{Y}$ and $\mathbf{Z}^{(t|k)}$ only, and then uses these estimates to construct normal-theory based approximate $100(1-\alpha)\%$ confidence regions (perhaps more appropriately called *prediction regions*) for individual $\mathbf{y}_j$'s. Then the number of $\mathbf{y}_j$'s falling within their confidence region is counted and compared to the expected count, $(1-\alpha)n$. Using $\mathbf{Z}^{(t|k)}$ only, compute for each $j \in \{1,\ldots,n\}$:

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_i^{(t|k)} &= \frac{1}{n_i^{(t|k)}} \sum_{j=1}^{n} z_{ji}^{(t|k)} \mathbf{y}_j, \quad i = 1,\ldots,k \\
\widehat{\boldsymbol{\Sigma}}^{(t|k)} &= \frac{1}{n-1} \sum_{j=1}^{n} \left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}_{\mathbf{z}_j^{(t|k)}}^{(t|k)}\right)\left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}_{\mathbf{z}_j^{(t|k)}}^{(t|k)}\right)' \qquad (6.12) \\
\left[d_j^2\right]^{(t|k)} &= \left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}_{\mathbf{z}_j^{(t|k)}}^{(t|k)}\right)' \left[\widehat{\boldsymbol{\Sigma}}^{(t|k)}\right]^{-1} \left(\mathbf{y}_j - \widehat{\boldsymbol{\mu}}_{\mathbf{z}_j^{(t|k)}}^{(t|k)}\right)
\end{aligned}$$

where

$$n_i = \sum_{j=1}^{n} z_{ji}.$$

Under bivariate normality and correct allocations $\mathbf{Z}$, $d_j^2 \overset{\bullet}{\sim} \chi_2^2$. Choose a confidence level α (we use $\alpha = 0.5$, as suggested by Johnson and Wichern (1992)) and compute

$\chi^2_2(1-\alpha)$, the $(1-\alpha)^{\text{th}}$ quantile of χ^2_2. Then define

$$D_{\text{CR}}\left(\mathbf{Y};\boldsymbol{\theta}^{(t|k)}\right) = \frac{1}{n}\sum_{j=1}^{n} \text{I}\left(\left[d_j^2\right]^{(t|k)} \geq \chi^2_2(1-\alpha)\right). \tag{6.13}$$

Note that $n D_{\text{CR}}\left(\mathbf{Y};\boldsymbol{\theta}^{(t|k)}\right) \overset{\bullet}{\sim} \text{Binomial}(n,\alpha)$, so that an approximate p-value for this discrepancy measure is

$$\begin{aligned} p_D &= P\left(D_{\text{CR}}\left(\mathbf{Y}^\star;\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right) \geq D_{\text{CR}}\left(\mathbf{Y}^{\text{obs}};\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right)\right) \\ &= P\left(\frac{D_{\text{CR}}\left(\mathbf{Y}^\star;\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right)-\alpha}{\sqrt{\frac{\alpha(1-\alpha)}{n}}} \geq \frac{D_{\text{CR}}\left(\mathbf{Y}^{\text{obs}};\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right)-\alpha}{\sqrt{\frac{\alpha(1-\alpha)}{n}}}\right) \\ &= P\left(z \geq \frac{D_{\text{CR}}\left(\mathbf{Y}^{\text{obs}};\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right)-\alpha}{\sqrt{\frac{\alpha(1-\alpha)}{n}}}\right) \qquad \text{where } z \sim N(0,1). \end{aligned}$$

Additionally, a Chi-square plot of $\left[d_j^2\right]^{(t|k)}$ vs. $\chi^2_1\left(\frac{\left[d_j^2\right]^{(t|k)}-0.5}{100}\right)$ could be displayed, with deviations from a straight line indicating various types of violation of bivariate normality (see Johnson and Wichern, 1992, p. 161).

We also implement the Monte Carlo approach by simulating a dataset $\mathbf{Y}^{\star(t)}$ for each $\boldsymbol{\theta}^{(t|k)}$: for $j = 1,\ldots,n$ generate $\mathbf{y}_j^{\star(t)} \sim N\left(\boldsymbol{\mu}_{\mathbf{z}_j^{(t|k)}}^{(t|k)}, \boldsymbol{\Sigma}^{(t|k)}\right)$. Then we compute the Bayesian p-value as in (6.11).

The second discrepancy measure we use, $D_{\boldsymbol{\Sigma}}\left(\mathbf{Y};\boldsymbol{\theta}^{(t|k)}\right)$, utilizes the asymptotic distribution of an estimator $\widehat{\boldsymbol{\Sigma}}^{(t|k)}$ to judge its "distance" from the true value $\boldsymbol{\Sigma}^{(t|k)}$. It is computed as follows.

Compute $\widehat{\boldsymbol{\Sigma}}^{(t|k)}$ as in (6.12), and let

$$\boldsymbol{\Sigma}^{(t|k)} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \quad \widehat{\boldsymbol{\Sigma}}^{(t|k)} = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix},$$

$$\boldsymbol{\sigma}^{(t|k)} = \begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{bmatrix}, \quad \widehat{\boldsymbol{\sigma}}^{(t|k)} = \begin{bmatrix} s_{11} \\ s_{22} \\ s_{12} \end{bmatrix},$$

and

$$\Gamma^{(t|k)} = \begin{bmatrix} 2{\sigma_{11}}^2 & 2{\sigma_{12}}^2 & 2\sigma_{11}\sigma_{12} \\ 2{\sigma_{12}}^2 & 2{\sigma_{22}}^2 & 2\sigma_{12}\sigma_{22} \\ 2\sigma_{11}\sigma_{12} & 2\sigma_{12}\sigma_{22} & \sigma_{11}\sigma_{22} + {\sigma_{12}}^2 \end{bmatrix}.$$

It can be shown that $\frac{1}{n}\Gamma^{(t|k)}$ is the asymptotic variance of $\widehat{\boldsymbol{\sigma}}^{(t|k)}$ under bivariate normality and correct allocations $\mathbf{Z}$, in the sense that

$$\sqrt{n}\left(\widehat{\boldsymbol{\sigma}}^{(t|k)} - \boldsymbol{\sigma}^{(t|k)}\right) \xrightarrow{D} N(\mathbf{0}, \Gamma).$$

Define the discrepancy measure as

$$D_{\boldsymbol{\Sigma}}\left(\mathbf{Y}; \boldsymbol{\theta}^{(t|k)}\right) = \left(\widehat{\boldsymbol{\sigma}}^{(t|k)} - \boldsymbol{\sigma}^{(t|k)}\right)' \left[\frac{1}{n}\Gamma\right]^{-1} \left(\widehat{\boldsymbol{\sigma}}^{(t|k)} - \boldsymbol{\sigma}^{(t|k)}\right). \tag{6.14}$$

Under bivariate normality, $D_{\boldsymbol{\Sigma}}\left(\mathbf{Y}; \boldsymbol{\theta}^{(t|k)}\right) \overset{\bullet}{\sim} \chi^2_3$. But since this asymptotic approximation is not as accurate as that for $D_{\mathrm{CR}}(\cdot;\cdot)$, we only implement a Monte Carlo scheme (completely analogous to that for $D_{\mathrm{CR}}(\cdot;\cdot)$) to compute a Bayesian p-value p_D for $D_{\boldsymbol{\Sigma}}\left(\mathbf{Y}; \boldsymbol{\theta}^{(t|k)}\right)$ for each k.

6.3.2 Cross-validation Methods

Let $\mathbf{y}_j^{\mathrm{obs}}$ denote the j^{th} observed offspring location, $\mathbf{y}^\star$ a replication for an arbitrary offspring (i.e., with possibly different allocation), and $\mathbf{y}_{(j)}^{\mathrm{obs}}$ the collection of all observed offspring locations except the j^{th}. Gelfand, Dey, and Chang (1992) propose the use of the *cross-validation predictive density* $p\left(\mathbf{y}^\star \middle| \mathbf{y}_{(j)}^{\mathrm{obs}}, k\right)$ to assess the fit of the data to the model indexed by k. This density suggests which values of $\mathbf{y}^\star$ are likely when the model is fitted using all data except $\mathbf{y}_j^{\mathrm{obs}}$. Each observed $\mathbf{y}_j^{\mathrm{obs}}$ can be compared to an estimate of the corresponding $p\left(\cdot \middle| \mathbf{y}_{(j)}^{\mathrm{obs}}, k\right)$ density to determine how well it supports the model.

We use two applications of the cross-validation predictive density. First, the *conditional predictive ordinate* for $\mathbf{y}_j^{\mathrm{obs}}$ under model k can be estimated for each j

and each k considered. The theoretical value is defined as

$$CPO_{j|k} = p\left(\mathbf{y}_j^{\text{obs}} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right). \tag{6.15}$$

Second, a measure of how likely it is to obtain a $p\left(\mathbf{y}^\star \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right)$ value smaller than $CPO_{j|k}$ (i.e., how likely it is for a new observation $\mathbf{y}^\star$ to support the model less than $\mathbf{y}_j^{\text{obs}}$ does) is defined as (using the same "d_3" name as given in Gelfand, Dey, and Chang (1992)):

$$d_{3_{j|k}} = P\left(p\left(\mathbf{y}^\star \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right) \leq p\left(\mathbf{y}_j^{\text{obs}} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right)\right), \tag{6.16}$$

$$\text{where } \mathbf{y}^\star \sim p\left(\cdot \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right). \tag{6.17}$$

Estimates of (6.15) and (6.16) can be used to detect observations which are "outliers" in the sense of not being supported by the model. Also, summaries of (6.15) and (6.16) over j can be used as a measure of overall fit of the data to the model. We describe such applications later; first we construct the estimators.

Estimators of $CPO_{j|k}$ and $d_{3_{j|k}}$ can be based on RJMCMC output, separated by k. These utilize the *cross-validation likelihood*, defined as "$p\left(\mathbf{y}_j^{\text{obs}} \left| \mathbf{y}_{(j)}^{\text{obs}}, \boldsymbol{\theta} \right.\right)$" where $\boldsymbol{\theta}$ is the set of all model parameters. In our BVNPCP-BHM, the definition of $\boldsymbol{\theta}$ actually depends on $\mathbf{Y}$, since $\boldsymbol{\theta}$ for $\mathbf{Y}$ includes $\mathbf{z}_j$, but $\boldsymbol{\theta}$ for $\mathbf{y}_{(j)}^{\text{obs}}$ does not. Also, the $\mathbf{y}_j$'s are independent, and so $\mathbf{y}_j^{\text{obs}}$ does not depend on $\mathbf{y}_{(j)}^{\text{obs}}$. Hence our unfixed parameter space necessitates a modified definition and notation for the cross-validation likelihood, represented and computed as follows:

$$\begin{aligned}
p\left(\mathbf{y}_j^{\text{obs}} \left| \boldsymbol{\theta}_{(j)} \right.\right) &= p\left(\mathbf{y}_j^{\text{obs}} \left| \boldsymbol{\Sigma}, \boldsymbol{\mu}, \mathbf{z}_{(j)}, k \right.\right) \\
&= p\left(\mathbf{y}_j^{\text{obs}} \left| \boldsymbol{\Sigma}, \boldsymbol{\mu}, k \right.\right) \\
&= \int p\left(\mathbf{y}_j^{\text{obs}}, \mathbf{z}_j \left| \boldsymbol{\Sigma}, \boldsymbol{\mu}, k \right.\right) \mathrm{d}\mathbf{z}_j \\
&= \int p\left(\mathbf{y}_j^{\text{obs}} \left| \boldsymbol{\Sigma}, \boldsymbol{\mu}, \mathbf{z}_j \right.\right) p\left(\mathbf{z}_j \left| k \right.\right) \mathrm{d}\mathbf{z}_j
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} p\left(\mathbf{y}_j^{\text{obs}} | \boldsymbol{\Sigma}, \boldsymbol{\mu}, \mathbf{z}_j = i\right) P\left(\mathbf{z}_j = i | k\right) \\
&= \frac{1}{k} \sum_{i=1}^{k} p\left(\mathbf{y}_j^{\text{obs}} | \boldsymbol{\Sigma}, \boldsymbol{\mu}, \mathbf{z}_j = i\right) \\
&= \frac{1}{k} \sum_{i=1}^{k} f\left(\mathbf{y}_j^{\text{obs}} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right) \qquad (6.18)
\end{aligned}
$$

where $f(\cdot|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ denotes the density of $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Note that (6.18) is equivalent to the mixture likelihood (3.3). The estimation of $CPO_{j|k}$ based on RJMCMC requires a sample $\left\{\boldsymbol{\theta}_{(j)}^{(1|k)}, \ldots, \boldsymbol{\theta}_{(j)}^{(T_k|k)}\right\}$, which is tempting to obtain by taking $\left\{\boldsymbol{\theta}^{(1|k)}, \ldots, \boldsymbol{\theta}^{(T_k|k)}\right\}$ and removing $\mathbf{z}_j$ from each. However, $\left\{\boldsymbol{\theta}^{(1|k)}, \ldots, \boldsymbol{\theta}^{(T_k|k)}\right\}$ is a Monte Carlo sample from $p\left(\boldsymbol{\theta} \left| \mathbf{Y}^{\text{obs}}, k\right.\right)$. The proper technique in our cross-validation setting is to use a sample from $p\left(\boldsymbol{\theta} \left| \mathbf{y}_{(j)}^{\text{obs}}, k\right.\right)$. Fortunately it is not necessary to re-run the RJMCMC sampler without $\mathbf{y}_j^{\text{obs}}$; the required sample can be obtained via weighted bootstrap resampling of the RJMCMC output (also called sampling/importance resampling: see Rubin (1988)):

For $t = 1, \ldots, T_k$, compute

$$w_{jt} = \frac{1}{p\left(\mathbf{y}_j^{\text{obs}} \left| \boldsymbol{\theta}_{(j)}^{(t|k)}\right.\right)}.$$

Then normalize the weights to yield

$$w_{jt}^{\star} = \frac{w_{jt}}{\sum_{t=1}^{T_k} w_{jt}}.$$

Sample, with replacement, T_k values from $\left\{\boldsymbol{\theta}^{(1|k)}, \ldots, \boldsymbol{\theta}^{(T_k|k)}\right\}$ with probabilities $\left\{w_{j1}^{\star}, \ldots, w_{jT_k}^{\star}\right\}$ to yield $\left\{\boldsymbol{\theta}^{\star(1|k)}, \ldots, \boldsymbol{\theta}^{\star(T_k|k)}\right\}$. Then discard $\mathbf{z}_j$ from each to produce $\left\{\boldsymbol{\theta}_{(j)}^{\star(1|k)}, \ldots, \boldsymbol{\theta}_{(j)}^{\star(T_k|k)}\right\}$, a cross-validation sample from $p\left(\boldsymbol{\theta} \left| \mathbf{y}_{(j)}^{\text{obs}}, k\right.\right)$.

Using the cross-validation sample $\boldsymbol{\theta}_{(j)}^{\star(\cdot|k)}$, $CPO_{j|k}$ is estimated, using (6.18), by

$$\widehat{CPO}_{j|k} = \left[\frac{1}{T_k} \sum_{t=1}^{T_k} \frac{1}{p\left(\mathbf{y}_j^{\text{obs}} \left| \boldsymbol{\theta}_{(j)}^{\star(\cdot|k)}\right.\right)}\right]^{-1}, \qquad (6.19)$$

the harmonic mean of the cross-validation likelihood values. Gelfand (1995) uses the form (6.19), except with $\boldsymbol{\theta}_{(j)}^{(\cdot|k)}$ in place of $\boldsymbol{\theta}_{(j)}^{\star(\cdot|k)}$, since he does not implement importance resampling (which is strange, because then the method is not really cross-validatory). For large data sets, this is unlikely to make a difference. However, to be on the safe side, we incorporate the importance resampling.

Estimation of $d_{3_{j|k}}$ requires an additional sampling step, the generation of samples from the cross-validation predictive density $p\left(\mathbf{y}^{\star} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right)$. This can be accomplished using our cross-validation sample $\boldsymbol{\theta}_{(j)}^{\star(\cdot|k)}$ as follows.

Note that

$$\begin{aligned} p\left(\mathbf{y}^{\star}, \boldsymbol{\theta}_{(j)} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right) &= p\left(\mathbf{y}^{\star} \left| \mathbf{y}_{(j)}^{\text{obs}}, \boldsymbol{\theta}_{(j)}, k \right.\right) p\left(\boldsymbol{\theta}_{(j)} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right) \\ &= p\left(\mathbf{y}^{\star} \left| \boldsymbol{\theta}_{(j)}, k \right.\right) p\left(\boldsymbol{\theta}_{(j)} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right). \end{aligned} \tag{6.20}$$

Therefore, if we can simulate $\left(\mathbf{y}^{\star}, \boldsymbol{\theta}^{\star}{}_{(j)}\right)$ jointly from (6.20), then $\mathbf{y}^{\star}$ will marginally be a sample from $p\left(\mathbf{y}^{\star} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right)$. We already have $\boldsymbol{\theta}_{(j)}^{\star(\cdot|k)}$ from $p\left(\boldsymbol{\theta} \left| \mathbf{y}_{(j)}^{\text{obs}}, k \right.\right)$, so we can obtain the required sample by generating $\mathbf{y}^{\star(t|k)}$ from $p\left(\mathbf{y}^{\star} \left| \boldsymbol{\theta}_{(j)}^{\star(t|k)}, k \right.\right)$, using (6.18):

1. Sample z from $U\{1, \ldots, k\}$.
2. Sample $\mathbf{y}^{\star(t|k)}$ from $N\left(\boldsymbol{\mu}_z^{\star(t|k)}, \boldsymbol{\Sigma}^{\star(t|k)}\right)$.

Then we have an estimate of $d_{3_{j|k}}$:

$$\widehat{d}_{3_{j|k}} = \frac{1}{T_k} \sum_{t=1}^{T_k} \mathrm{I}\left[p\left(\mathbf{y}^{\star(t|k)} \left| \boldsymbol{\theta}_{(j)}^{\star(t|k)} \right.\right) \leq p\left(\mathbf{y}_j^{\text{obs}} \left| \boldsymbol{\theta}_{(j)}^{\star(t|k)} \right.\right)\right], \tag{6.21}$$

where the densities $p(\cdot|\cdot)$ are computed via (6.18).

If the collection of $\widehat{d}_{3_{j|k}}$ for $j = 1, \ldots, n$ is "roughly centered around 0.5 without many extreme values," this indicates a good model fit (Gelfand, Dey, and Chang, 1992). If $d_{3_{j|k}}$ is small, then $\mathbf{y}_j^{\text{obs}}$ does not support model k. On the other extreme, an excess of large $d_{3_{j|k}}$ suggest that variation predicted by model k is not supported

by the data.

Plots of $\widehat{CPO}_{j|k}$ vs. j, constructed separately for each k, can identify which data points $\mathbf{y}_j^{\text{obs}}$ support or fail to support the model for each k. The sum of $\log \widehat{CPO}_{j|k}$ can be used as a measure to compare model fits (see section 6.4). However, the irrelevance of the magnitude of $\widehat{CPO}_{j|k}$ to model *adequacy* renders it useless for validating model assumptions in general. The statistic $\widehat{d}_{3_{j|k}}$ fills this role; histograms or boxplots of $\widehat{d}_{3_{1|k}}, \ldots, \widehat{d}_{3_{n|k}}$ for each k can be used to assess the overall fit of the BVNPCP-BHM conditional on each k considered, according to the criteria discussed in the previous paragraph.

6.4 Model Comparison (Inference for k)

Inference for k, the number of clusters, is possible through a variety of techniques. The aim of this thesis is not to choose a particular model (i.e., a BVNPCP(A, k, n) for a particular k). Inference for $\boldsymbol{\Sigma}$ will involve contributions from all candidate models visited by the Markov chains, accounting for the uncertainty of k implicitly as part of RJMCMC rather than as a secondary Bayesian model averaging-type procedure as in composite EM analysis.

However, there are a wealth of model comparison methods, few of which appear to have been applied to RJMCMC, and so we feel that an investigation of model comparison possibilities for RJMCMC is in order. Furthermore, it will be interesting to see how well conclusions from methods which analyze output separately for each k tally with the marginal distribution of k from the RJMCMC sampler (i.e., the "visit frequencies" as shown in the posterior histograms of k).

6.4.1 RJMCMC Model Visit Frequencies

The number of clusters, k, can actually be estimated quite directly from RJMCMC, using samples from the marginal posterior distribution of k:

$$\widehat{p}(k|\mathbf{Y}) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{I}\left(k^{(t)} = k\right). \tag{6.22}$$

The variance of each $\widehat{p}(k|\mathbf{Y})$ can be estimated via batch sampling of the indicator function $\mathrm{I}\left(k^{(t)} = k\right)$ (an idea apparently implemented by Carlin and Chib (1995) in their non-RJMCMC dimension-changing sampler):

$$\widehat{\mathrm{Var}}\left(\widehat{p}(k|\mathbf{Y})\right) = \frac{1}{T}\left[\frac{b_1}{m_1 - 1}\sum_{j=1}^{m_1}\left(\widehat{p}_j(k|\mathbf{Y}) - \widehat{p}(k|\mathbf{Y})\right)^2\right] \tag{6.23}$$

where $\widehat{p}_j(k|\mathbf{Y})$ = mean of $\mathrm{I}(k^{(\cdot)} = k)$ for j^{th} batch

and b_1, m_1 = batch size and number of batches used.

By the Central Limit Theorem (quite appropriate here since T is in general very large),

$$\widehat{p}(k|\mathbf{Y}) \overset{\bullet}{\sim} N\left(p(k|\mathbf{Y}), \widehat{\mathrm{Var}}\left(\widehat{p}(k|\mathbf{Y})\right)\right). \tag{6.24}$$

6.4.2 Use of Model Adequacy / Checking Criteria

As suggested by Gelfand, Dey, and Chang (1992), $\widehat{CPO}_{j|k}$ and $\widehat{d}_{3_{j|k}}$ can be used for model comparison via determination of which models appear to be more adequate than others as an explanation for the data. Such comparisons are rather ad-hoc and are difficult to interpret on a meaningful scale, but nevertheless useful at least from a descriptive point of view.

Gelfand, Dey, and Chang (1992) suggest, as one possible strategy, favoring values of k with higher $\sum_{j=1}^{n} \log \widehat{CPO}_{j|k}$ or higher $\sum_{j=1}^{n}\left(\widehat{d}_{3_{j|k}} - 0.5\right)^2$. The exponentiated difference

$$\exp\left(\sum_{j=1}^{n} \log \widehat{CPO}_{j|k_1} - \sum_{j=1}^{n} \log \widehat{CPO}_{j|k_2}\right)$$

may also be used as a surrogate for the Bayes factor, called the "pseudo-Bayes factor" (see Gelfand, 1995, p. 150) in comparing candidate models k_1 and k_2. An alternative is to display adjacent boxplots of $\widehat{CPO}_{j|k}$ or $\widehat{d}_{3_{j|k}}$ for different k, favoring k for which $\widehat{CPO}_{j|k}$ values appear higher and $\widehat{d}_{3_{j|k}}$ values are more concentrated around 0.5. This alternative method may be more robust to outliers than use of a scalar summary. Still another possibility is to plot $\widehat{CPO}_{j|k_1}$ vs. $\widehat{CPO}_{j|k_2}$ in a scatterplot matrix covering all pairs of k values considered.

6.4.3 Bayes Factor Approximations

Perhaps the most popular tool for model comparison in any Bayesian framework is the Bayes factor (see Kass and Raftery (1995) for a review). The Bayes factor for comparing two models k_1 and k_2 is defined as the ratio of *marginal likelihoods* for the two models,

$$B_{12} = \frac{p\left(\mathbf{Y}\,|k_1\right)}{p\left(\mathbf{Y}\,|k_2\right)}.$$

Its name is suitable because B_{12} is the factor by which the prior odds of k_1 over k_2 must be multiplied to obtain the posterior odds:

$$\frac{p\left(k_1\,|\mathbf{Y}\right)}{p\left(k_2\,|\mathbf{Y}\right)} = \frac{\left[\frac{p(\mathbf{Y}|k_1)p(k_1)}{p(\mathbf{Y})}\right]}{\left[\frac{p(\mathbf{Y}|k_2)p(k_2)}{p(\mathbf{Y})}\right]} = \left[\frac{p(k_1)}{p(k_2)}\right]\left[\frac{p\left(\mathbf{Y}\,|k_1\right)}{p\left(\mathbf{Y}\,|k_2\right)}\right].$$

If the candidate models are taken to be equally likely *a priori* (which they are, in our case), then the Bayes factor completely determines the posterior odds of each pair of candidate models, and thus the posterior distribution of k. Inference for k then focuses on estimation of constant multiples of the marginal likelihoods, i.e., computation of $\widehat{cp(\mathbf{Y}|k)}$ for some constant c. Many varieties of such estimators are available. We concentrate on those that are invariant to label-switching (see section 4.4.6). A popular method due to Chib (1995) is unfortunately unavailable to us due to the label-switching problem. Another, the Laplace-Metropolis estimator

(see Raftery, 1995, section 10.4.1) appears to be intractable due to the inability to estimate posterior variance matrices involving $\boldsymbol{\mu}$ and/or $\mathbf{Z}$ (again, label-switching being the culprit). There are possible alternatives, e.g., use of an observed information matrix, but it is not clear how such asymptotic variance estimates should be computed, or how accurate they would be, given that MLE's are not produced by RJMCMC. Further research is needed to explore the feasibility of a Laplace-Metropolis approach for our model.

Fortunately, all other commonly used marginal likelihood estimators for MCMC *are* available to us, and so we can concentrate on these. Many of these methods involve computation of a likelihood using posterior samples. Although the "likelihood" for our BVNPCP-BHM is the classification likelihood as specified by (4.2) and (3.1), there is no reason we cannot use other likelihood forms as well. Raftery (1995, section 10.5) supports the use of the mixture likelihood (given by (3.3)) for model comparison using MCMC for mixtures.

In the MCMC framework, the likelihood can be treated simply as a function $L(\boldsymbol{\theta}, \mathbf{Y})$ of the model parameters and data. The marginal likelihood estimators we will use rely on the availability of samples of the function $L(\boldsymbol{\theta}, \mathbf{Y})$ from the posterior distribution of $\boldsymbol{\theta}$. If $L(\boldsymbol{\theta}, \mathbf{Y}) \equiv p(\mathbf{Y} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}, k)$, the BVNPCP-BHM (and classification) likelihood, posterior samples are clearly available directly from RJMCMC, which implements the Monte Carlo integration

$$\int \cdots \int p(\mathbf{Y} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}, k) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}, k \,|\mathbf{Y}) \mathrm{d}\boldsymbol{\mu}\ \mathrm{d}\boldsymbol{\Sigma}\ \mathrm{d}\mathbf{Z}\ \mathrm{d}k.$$

If $L(\boldsymbol{\theta}, \mathbf{Y}) \equiv p(\mathbf{Y} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, k)$, the mixture likelihood, using posterior samples directly from RJMCMC is analogous to performing the Monte Carlo integration

$$\int \cdots \int p(\mathbf{Y} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, k) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}, k \,|\mathbf{Y}) \mathrm{d}\boldsymbol{\mu}\ \mathrm{d}\boldsymbol{\Sigma}\ \mathrm{d}\mathbf{Z}\ \mathrm{d}k.$$

A more direct Monte Carlo integration would be

$$\int \cdots \int p\left(\mathbf{Y} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, k\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, k \,|\mathbf{Y}\right) \mathrm{d}\boldsymbol{\mu} \; \mathrm{d}\boldsymbol{\Sigma} \; \mathrm{d}k,$$

but since samples of $(\boldsymbol{\Sigma}, \boldsymbol{\mu}, k)$ from $p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}, k \,|\mathbf{Y}\right)$ are *marginally* from $p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, k \,|\mathbf{Y}\right)$, RJMCMC does indeed produce valid samples of the mixture likelihood $p\left(\mathbf{Y} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, k\right)$ as well.

Two different varieties of marginal likelihood estimators are used. The first involves penalized likelihoods, three forms of which we use: the Bayesian Information Criterion (BIC, Schwarz (1978)), Approximate Weight of Evidence (AWE, Banfield and Raftery (1993)) and Akaike Information Criterion (AIC, Akaike (1973)). These are all estimators of $2 \log p(\mathbf{Y}|k) + c$. The penalized likelihoods are typically evaluated at the MLE, but since the MLE is unavailable in RJMCMC, either the maximum (e.g., Raftery (1995, p. 178)) or average (e.g., Carlin and Louis (1996, p. 231)) of the posterior likelihood samples can be used. Since our study of model comparison procedures for RJMCMC is rather exploratory, we try both approaches.

It seems that use of the mixture likelihood is more appropriate than the classification likelihood (except for the AWE, which is specifically designed to utilize the classification likelihood), especially since specification of the dimension of $\mathbf{Z}$ is unclear. Raftery (1995, section 10.5) supports this approach in the mixture model context, although he cautions that the BIC is not known to be valid for mixture models. However, Fraley and Raftery (1998) cite examples (mentioned in section 3.5.1 of this thesis) supporting the use of the BIC for mixture models.

The penalties employed by the estimators include a specification of the number of scalar parameters, which for the mixture likelihood is $2k + 3$ (two scalar coordinates for each $\boldsymbol{\mu}_i$, plus σ_{11}, σ_{22} and σ_{12}). The two forms of BIC computed

from RJMCMC output are

$$BIC_k^{\max} = 2 \max_t \log p\left(\mathbf{Y} \left| \boldsymbol{\mu}^{(t|k)}, \boldsymbol{\Sigma}^{(t|k)} \right.\right) - (2k+3)\log n \tag{6.25}$$

and

$$BIC_k^{\text{mean}} = 2\frac{1}{T_k}\sum_{t=1}^{T_k} \log p\left(\mathbf{Y} \left| \boldsymbol{\mu}^{(t|k)}, \boldsymbol{\Sigma}^{(t|k)} \right.\right) - (2k+3)\log n. \tag{6.26}$$

The two forms of AIC are

$$AIC_k^{\max} = 2 \max_t \log p\left(\mathbf{Y} \left| \boldsymbol{\mu}^{(t|k)}, \boldsymbol{\Sigma}^{(t|k)} \right.\right) - 2(2k+3) \tag{6.27}$$

and

$$AIC_k^{\text{mean}} = 2\frac{1}{T_k}\sum_{t=1}^{T_k} \log p\left(\mathbf{Y} \left| \boldsymbol{\mu}^{(t|k)}, \boldsymbol{\Sigma}^{(t|k)} \right.\right) - 2(2k+3). \tag{6.28}$$

Finally, the two forms of AWE are

$$AWE_k^{\max} = 2 \max_t \log p\left(\mathbf{Y} \left| \boldsymbol{\mu}^{(t|k)}, \boldsymbol{\Sigma}^{(t|k)}, \mathbf{Z}^{(t|k)} \right.\right) - 2\left(2k+3+\frac{3}{2}\right)\log n \tag{6.29}$$

and

$$AWE_k^{\text{mean}} = 2\frac{1}{T_k}\sum_{t=1}^{T_k} \log p\left(\mathbf{Y} \left| \boldsymbol{\mu}^{(t|k)}, \boldsymbol{\Sigma}^{(t|k)}, \mathbf{Z}^{(t|k)} \right.\right) - 2\left(2k+3+\frac{3}{2}\right)\log n. \tag{6.30}$$

Because we assume equal prior model probabilities (equal $p(k)$), posterior model probability estimates for a set of candidate models $\{k_{\min}, \ldots, k_{\max}\}$ can constructed via:

$$\widehat{p}(k|\mathbf{Y}) = \frac{\exp\left(\frac{1}{2}[2\log \widehat{p\,(\mathbf{Y}\,|k\,)} + c]\right)}{\sum_{q=k_{\min}}^{k_{\max}} \exp\left(\frac{1}{2}[2\log \widehat{p\,(\mathbf{Y}\,|q\,)} + c]\right)} \tag{6.31}$$

where $\left\{[2\log \widehat{p\,(\mathbf{Y}\,|k\,)} + c]\right\}$ is any of $\{BIC_k^{\max}\}$, $\{BIC_k^{\text{mean}}\}$, $\{AIC_k^{\max}\}$, $\{AIC_k^{\text{mean}}\}$, $\{AWE_k^{\max}\}$, or $\{AWE_k^{\text{mean}}\}$. This follows because, assuming $\{k_{\min}, \ldots, k_{\max}\}$ covers the set of feasible models,

$$\begin{aligned} p\,(k\,|\mathbf{Y}\,) &= \frac{p\,(\mathbf{Y}\,|k\,)p(k)}{p(\mathbf{Y}\,)} \\ &\approx \frac{p\,(\mathbf{Y}\,|k\,)p(k)}{\sum_{q=k_{\min}}^{k_{\max}} p\,(\mathbf{Y}\,|q\,)p(q)} \\ &= \frac{\exp\left(\frac{1}{2}\left[2\log p\,(\mathbf{Y}\,|k\,) + c\right]\right)}{\sum_{q=k_{\min}}^{k_{\max}} \exp\left(\frac{1}{2}\left[2\log p\,(\mathbf{Y}\,|q\,) + c\right]\right)}. \end{aligned}$$

The second type of marginal likelihood estimator used is the importance sampling estimator based on the mixture likelihood, which has general form

$$\widehat{p}(\mathbf{Y}\,|k) = \frac{\sum_{t=1}^{T_k} \frac{p\left(\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\,|k\right)p\left(\mathbf{Y}\,\middle|\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\right)}{p^{\star}\left(\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\,|k\right)}}{\sum_{t=1}^{T_k} \frac{p\left(\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\,|k\right)}{p^{\star}\left(\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\,|k\right)}}, \tag{6.32}$$

where $p^{\star}(\cdot)$ is an importance sampling density (Newton and Raftery, 1994). If $p^{\star}(\cdot)$ is chosen to be equal to the posterior $p\left(\mathbf{Y}\,\middle|\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\right)$, then (6.32) simplifies to (using the name given by Newton and Raftery (1994))

$$\widehat{p}_2(\mathbf{Y}\,|k) = \left[\frac{1}{T_k}\sum_{t=1}^{T_k}\frac{1}{p\left(\mathbf{Y}\,\middle|\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\right)}\right]^{-1}, \tag{6.33}$$

the harmonic mean of the sample of mixture likelihood values, which converges almost surely to the correct value but does not, in general, satisfy a Gaussian central limit theorem (Raftery, 1995, p. 169).

A more robust version of (6.32) is constructed using for $p^{\star}(\cdot)$ a mixture of prior and posterior densities,

$$p^{\star}\left(\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\,|k\right) = \delta p\left(\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\,|k\right) + (1-\delta)p\left(\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\,|\mathbf{Y},k\right),$$

where $0 < \delta < 1$, preferably small. To avoid the necessity of simulating from the prior, Newton and Raftery (1994) suggest using all T_k values from the posterior sample and "imagining that a further $\left[\frac{\delta T_k}{1-\delta}\right]$ values of $[\boldsymbol{\mu},\boldsymbol{\Sigma}]$ are drawn from the prior, all with likelihoods $[p\,(\mathbf{Y}\,|\boldsymbol{\mu},\boldsymbol{\Sigma},k\,)]$ equal to their expected value $[p\,(\mathbf{Y}\,|k\,)]$," leading to

$$\widehat{p}_4(\mathbf{Y}\,|k) = \frac{\sum_{t=1}^{T_k} \frac{p\left(\mathbf{Y}\,\middle|\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\right)}{\delta\widehat{p}_4(\mathbf{Y}\,|k)+(1-\delta)p\left(\mathbf{Y}\,\middle|\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\right)} + \left(\frac{\delta}{1-\delta}\right)T_k}{\sum_{t=1}^{T_k} \frac{1}{\delta\widehat{p}_4(\mathbf{Y}\,|k)+(1-\delta)p\left(\mathbf{Y}\,\middle|\boldsymbol{\mu}^{(t|k)},\boldsymbol{\Sigma}^{(t|k)}\right)} + \frac{\left(\frac{\delta}{1-\delta}\right)T_k}{\widehat{p}_4(\mathbf{Y}\,|k)}}, \tag{6.34}$$

which can be solved iteratively. The "expected value" of $p\,(\mathbf{Y}\,|\boldsymbol{\mu},\boldsymbol{\Sigma},k\,)$ means its expectation under the prior distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given k,

$$\begin{aligned}\int\cdots\int p\,(\mathbf{Y}\,|\boldsymbol{\mu},\boldsymbol{\Sigma},k\,)p\,(\boldsymbol{\mu},\boldsymbol{\Sigma}\,|k\,)\mathrm{d}\boldsymbol{\mu}\;\mathrm{d}\boldsymbol{\Sigma} &= \int\cdots\int p\,(\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|k\,)\mathrm{d}\boldsymbol{\mu}\;\mathrm{d}\boldsymbol{\Sigma}\\ &= p\,(\mathbf{Y}\,|k\,).\end{aligned}$$

We use $\delta = 0.1$ and encounter no convergence problems; the solution to (6.34) is obtained in every case within a handful of iterations. Unlike $\widehat{p}_2(\mathbf{Y}|k)$, $\widehat{p}_4(\mathbf{Y}|k)$ *does* satisfy a Gaussian central limit theorem, in addition to being strongly consistent (Raftery, 1995).

Both $\widehat{p}_2(\mathbf{Y}|k)$ and $\widehat{p}_4(\mathbf{Y}|k)$ can produce posterior model probability estimates for $\{k_{\min}, \ldots, k_{\max}\}$, assuming equal priors, via

$$\widehat{p}(k|\mathbf{Y}) = \frac{\widehat{p}_i(\mathbf{Y}|k)}{\sum_{q=k_{\min}}^{k_{\max}} \widehat{p}_i(\mathbf{Y}|q)} \tag{6.35}$$

for $i = 2$ or 4.

6.5 Estimation of $\boldsymbol{\Sigma}$ and Isotropy Testing

Two very different approaches to performing inference on the cluster shape/scale parameter $\boldsymbol{\Sigma}$ using RJMCMC output are implemented: highest posterior density interval (HPDI) calculation and batch sampling-based variance estimation. For both approaches we use posterior samples $\boldsymbol{\theta}^{(\cdot)}$ across all k, thus implicitly accounting for uncertainty in the number of clusters. We also work with normalized versions of the "regular" and "anisotropy" parameterizations of $\boldsymbol{\Sigma}$, as discussed in Definitions 1.1.9 and 1.1.10 and section 3.5.2.

6.5.1 HPD Intervals and Tests

For this section we define the notation $\theta(j)$ to denote the j^{th} order statistic of a sample $\theta^{(\cdot)} = \{\theta^{(1)}, \ldots, \theta^{(T)}\}$ of either a linear or circular parameter. A $100(1-\alpha)\%$ HPD interval for a linear parameter θ of a model analyzed via MCMC is defined as the shortest interval containing at least $100(1-\alpha)\%$ of the posterior samples, which is given for a unimodal sample as:

$$[\theta(t^\star), \theta\,(t^\star + \lfloor (1-\alpha)T - \epsilon \rfloor)] \tag{6.36}$$

where $t^\star$ is such that

$$\theta\left(t^\star + \lfloor (1-\alpha)T - \epsilon \rfloor\right) - \theta(t^\star) = \min_{1 \le t \le T - \lfloor (1-\alpha)T - \epsilon \rfloor} \left\{\theta\left(t + \lfloor (1-\alpha)T - \epsilon \rfloor\right) - \theta(t)\right\}$$

and "$\lfloor \cdot \rfloor$" denotes "greatest integer less than or equal to" and $\epsilon > 0$ is on the order of machine precision. There are analogous definitions for multimodal samples, but they are not needed in our case because all of our posterior samples turn out to be unimodal. Chen and Shao (1998) prove that the coverage probability of (6.36) converges almost surely to the correct value. The HPD interval is considered superior to the commonly used equal-tail *Bayesian credible interval*, especially if the posterior sample is not symmetric (Chen and Shao, 1998). It can be used to construct confidence intervals for $\log \sigma_{11}$, $\log \sigma_{22}$, $z(\rho_{12})$, $\log \gamma$, $\log \Psi$ and $\log \sigma_{11} - \log \sigma_{22}$. Note in particular that an HPD interval for $\log \gamma$ cannot possibly contain 0, the null value for isotropy. A valid isotropy test can be conducted, however, using two HPD intervals for

$$\boldsymbol{\sigma}^c \equiv (\log \sigma_{11} - \log \sigma_{22},\ z(\rho_{12}))$$

and a Bonferroni correction. As discussed in section 3.5.2, a test of $\mathrm{H}_0 : \boldsymbol{\sigma}^c = 0$ vs. $\mathrm{H}_1 : \boldsymbol{\sigma}^c \neq 0$ is a test of isotropy for the BVNPCP. We compute the two achieved significance levels

$$\begin{aligned} ASL_{\text{VarDiff}} &= 1 - (\text{confidence level of largest HPD interval for } \log \sigma_{11} - \log \sigma_{22} \\ &\qquad \text{which excludes } 0) \end{aligned}$$

and

$$\begin{aligned} ASL_{\text{Cov}} &= 1 - (\text{confidence level of largest HPD interval for } z(\rho_{12}) \\ &\qquad \text{which excludes } 0) \end{aligned}$$

and then compute the Bonferroni-corrected p-value of the isotropy test as

$$p = \min\left\{1,\ 2\min\left(ASL_{\text{VarDiff}}, ASL_{\text{Cov}}\right)\right\}. \tag{6.37}$$

The standard formula (6.36) for HPD intervals does not apply for circular

parameters (e.g., ϕ). To the author's knowledge, no methods have been previously proposed to obtain HPD intervals for circular parameters. It seems perfectly reasonable, however, to define a new interval length measure appropriate for a circular parameter $\eta \in [a, b)$ and construct a HPD interval in the same fashion, except allowing the interval the possibility of wrapping around b and resuming at a.

Define the distance measure "$\ominus$" as

$$\eta(i) \ominus \eta(j) = \begin{cases} \eta(i) - \eta(j), & \text{if } i \geq j \\ (\eta(i) - a) + (b - \eta(j)), & \text{if } i < j \end{cases}$$

and interval notation "$\odot$" as

$$\eta(i) \odot \eta(j) = \begin{cases} (\eta(i), \eta(j)), & \text{if } i \geq j \\ (a, \eta(i)) \cup (\eta(j), b), & \text{if } i < j. \end{cases} \tag{6.38}$$

Also define the operator "$\oslash$" as

$$j \oslash T = \begin{cases} j, & \text{if } j \leq T \\ j \pmod{T}, & \text{if } j > T, \end{cases}$$

which is equivalent to the "mod" operator except that $j \oslash j = j$.

Then a $100(1-\alpha)\%$ HPD interval for η is

$$\eta(t^\star) \ \odot \ \eta\left((t^\star + \lfloor (1-\alpha)T - \epsilon \rfloor) \oslash T\right) \tag{6.39}$$

where $t^\star$ is such that

$$\eta\left((t^\star + \lfloor (1-\alpha)T - \epsilon \rfloor) \oslash T\right) \ \ominus \ \eta(t^\star)$$

$$= \min_{1 \leq t \leq T} \left\{ \eta\left((t + \lfloor (1-\alpha)T - \epsilon \rfloor) \oslash T\right) \ \ominus \ \eta(t) \right\}.$$

A $100(1-\alpha)\%$ HPD interval can be constructed for ϕ using $\left\{\phi^{(1)}, \ldots, \phi^{(T)}\right\}$ with $a = -\frac{\pi}{2}$ and $b = \frac{\pi}{2}$. Caution is advised, however, since there are no established results regarding the asymptotic coverage probability.

6.5.2 Batch Sampling-Based Confidence Regions and Tests

Another approach to estimating a vector or scalar component of $\boldsymbol{\Sigma}$ is to assume approximate normality of the posterior sample of the component and estimate its variance via batch sampling. We caution from the outset that approximate normality has not been rigorously assessed for posterior samples of $\boldsymbol{\sigma}^c$, $\log \sigma_{11}$, $\log \sigma_{22}$, $z(\rho_{12})$, $\log \gamma$ or $\log \Psi$, but we implement the batch sampling approach mostly for investigative purposes. Since we perform analyses on several simulated data sets whose true underlying model parameters are known, we can carry out a (albeit small-scale) study of its performance. We use the normalized parameterizations of $\boldsymbol{\Sigma}$ to make the approximate normality assumption more reasonable. Most of the nonparametric posterior density estimates calculated for scalar parameters in our analyses (the primary exception being $\log \gamma$, which is skewed right for isotropic models) seem to resemble normal curves, at least suggesting our investigation is worthwhile.

For a scalar linear parameter θ (e.g., $\log \sigma_{11}$, $\log \sigma_{22}$, $z(\rho_{12})$, $\log \gamma$ or $\log \Psi$), the posterior variance can be estimated via batch sampling (see section 6.2.2). A (very, perhaps) approximate $100(1-\alpha)\%$ confidence interval for θ is then

$$\bar{\theta}^{(\cdot)} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathrm{Var}}_{\mathrm{BS}}(\theta)}. \tag{6.40}$$

A two-dimensional approximate $100(1-\alpha)\%$ confidence region for $\boldsymbol{\sigma}^c$ can be constructed using a multivariate batch sampling variance estimate:

$$\left[\bar{\boldsymbol{\sigma}}^{c(\cdot)}\right]' \left[\widehat{\mathrm{Var}}_{\mathrm{BS}}\left(\boldsymbol{\sigma}^{c(\cdot)}\right)\right]^{-1} \left[\bar{\boldsymbol{\sigma}}^{c(\cdot)}\right] \leq \chi^2_2(1-\alpha). \tag{6.41}$$

The deviation of this elliptical region from the null value $\mathbf{0}$ suggests the nature and extent of anisotropy. The p-value for an isotropy test of $\mathrm{H}_0 : \boldsymbol{\sigma}^c = 0$ vs.

$H_1 : \boldsymbol{\sigma}^c \neq 0$ is given by

$$P\left(X \geq \left[\bar{\boldsymbol{\sigma}}^{c(\cdot)}\right]' \left[\widehat{\mathrm{Var}}_{\mathrm{BS}}\left(\boldsymbol{\sigma}^{c(\cdot)}\right)\right]^{-1} \left[\bar{\boldsymbol{\sigma}}^{c(\cdot)}\right]\right), \quad \text{where } X \sim \chi^2_2. \tag{6.42}$$

Finally, a nonparametric (not assuming approximate normality, but nevertheless quite approximate) $100(1-\alpha)\%$ confidence interval can be obtained for ϕ, using the special version of batch sampling for circular dispersion (see section 6.2.2). First $\phi^{(\cdot)}$ is converted to $\phi^{\star(\cdot)}$ via (6.7). Then the sample circular mean and batch sampling circular dispersion estimate of $\phi^{\star(\cdot)}$ are used to produce the confidence interval (using the "$\odot$" notation defined in (6.38)):

$$\begin{aligned}&\left(\left[\bar{\phi}^{\star(\cdot)} - \arcsin\left(z_{\frac{\alpha}{2}}\sqrt{\widehat{\delta}_{\mathrm{BS}}\left(\phi^{\star(\cdot)}\right)}\right)\right] \pmod{2\pi}\right) \\ &\quad \odot \quad \left(\left[\bar{\phi}^{\star(\cdot)} + \arcsin\left(z_{\frac{\alpha}{2}}\sqrt{\widehat{\delta}_{\mathrm{BS}}\left(\phi^{\star(\cdot)}\right)}\right)\right] \pmod{2\pi}\right).\end{aligned} \tag{6.43}$$

Then the endpoints ($c^\star_{\mathrm{lo}}$ and $c^\star_{\mathrm{hi}}$, say) are back-transformed via (6.8) to give a $100(1-\alpha)\%$ confidence interval $c_{\mathrm{lo}} \odot c_{\mathrm{hi}}$ for ϕ. Recall that this confidence interval is ill-defined if $z_{\frac{\alpha}{2}}\sqrt{\widehat{\delta}_{\mathrm{BS}}\left(\phi^{\star(\cdot)}\right)} > 1$, in which case we report that at least $\bar{\phi}^{\star(\cdot)} \pm \frac{\pi}{2}$, back-transformed appropriately, is covered.

6.5.3 Comments on the Two Approaches

The HPD interval approach makes no assumptions about the form of the posterior distribution and is more theoretically sound, and hence our preferred approach. A drawback of this method is its inability to directly construct multi-dimensional confidence regions (although componentwise confidence intervals can certainly be combined to yield cube-shaped regions, but these may be unnecessarily large and lead to overly conservative multi-dimensional tests).

The batch sampling approach was included primarily for this reason, as it can produce multi-dimensional ellipsoidal confidence regions and tests incorporating covariances in posterior samples. However, the distributional approximations

required are certainly suspect. Furthermore, there appears to be a serious problem with *over*-estimation of variance in batch sampling for RJMCMC (see Chapter 7 for examples). We have not determined the exact cause, but we have observed strong negative correlations between batch means, likely the result of different k's dominating different batches. This is a good topic for future research.

CHAPTER 7
IMPLEMENTATION AND RESULTS OF ANALYSES, WITH COMPARISONS OF METHODS

7.1 Implementation of RJMCMC Algorithm

For the Redwood data and each of the 12 simulated patterns, 3 chains of the RJMCMC sampler for the BVNPCP-BHM(A, n) (see Definition 4.2.1 and Algorithm 4.4.6) were run for 200,000 sweeps apiece. Some information was recorded using all sweeps (see section 7.2), but due to limitations on storage space, only every 10^{th} sweep was saved. Originally, only 100,000 sweeps were run (also saving every 10^{th} sweep), and this was found to be insufficient for batch sampling (the ACF cutoff of 0.05 could not be achieved for some methods). Due to the high degree of autocorrelation, we feel that simulation of longer chains, discarding some sweeps, is preferable to simulation of shorter chains, saving all sweeps.

Hyperparameter values are displayed in Table 7.1. Values for V (specified in terms of $\frac{1}{m}V^{-1}$, the inverse of the prior mean of $\mathbf{\Sigma}^{-1}$) all represent isotropic processes; they are chosen to be consistent with the cluster size implied by the true $\mathbf{\Sigma}$ model values for the simulated patterns, and according to casual visual inspection for the Redwood data. As discussed in section 4.4.2, setting $m = 2$ makes the prior for $\mathbf{\Sigma}$ as uninformative as possible. We expect the prior to have very little impact on the behavior of the algorithm. Rigorous sensitivity analysis would require a large number of simulations and is saved for future research.

Starting values used for k and $\mathbf{\Sigma}$ are shown in Tables 7.2 and 7.3. For chain 1,

Pattern(s)	k_{lo}	k_{hi}	m	$\left(\frac{1}{m}V^{-1}\right)_{11}$	$\left(\frac{1}{m}V^{-1}\right)_{22}$	$\left(\frac{1}{m}V^{-1}\right)_{12}$
Redwood	1	30	2	0.0025	0.0025	0
I-k7-$*$	1	30	2	0.003	0.003	0
I-k14-$*$	1	30	2	0.0015	0.0015	0

Table 7.1: Hyperparameter values used for prior specifications in RJMCMC.

$\boldsymbol{\mu}_1$ is initially located at the center of the region. For chains 2 and 3, starting values are set to a random sample (selected without replacement) of offspring locations. Instead of selecting initial values for $\mathbf{Z}$, we implement M_Z (then $M_{\boldsymbol{\mu}}$ and $M_{\boldsymbol{\Sigma}}$) before starting step 3 of Algorithm 4.4.6.

Chain	k	σ_{11}	σ_{22}	σ_{12}
1	1	0.1	0.1	0
2	15	0.003	0.003	0
3	30	0.001	0.001	0

Table 7.2: Starting values for k and $\boldsymbol{\Sigma}$ used in RJMCMC, Redwood data.

The starting values were chosen to be over-dispersed, as required by our convergence assessment technique. An upper limit of $k_{\text{hi}} = 30$ seems reasonable for all of the patterns. The result of the analysis performed by Diggle (1983) (see section 1.3.1 and Figure 1.1) on a similar Redwood pattern, however, implies the

Chain	k	σ_{11}	σ_{22}	σ_{12}
1	1	0.1	0.1	0
2	15	0.002	0.002	0
3	30	0.001	0.001	0

Table 7.3: Starting values for k and $\mathbf{\Sigma}$ used in RJMCMC, simulated patterns.

presence of approximately 67 clusters in our Redwood data set. Once this was realized, we ran a chain with $k_{\text{hi}} = 100$ and a starting k of 90. The value of k fell below 10 within 2,000 sweeps and remained below 20 for the remainder of the 200,000 sweeps.

The source code for the RJMCMC sampler was written in C++ using matrix and random number libraries authored by Davies (1997) and compiled with the HP-UX CC compiler, version A.10.36, to generate ANSI style code. Simulations were run on a Hewlett Packard workstation model B132L running HP-UX version 10.20, which has a 132 MHz PA-7300CL CPU and 128 Mb RAM. Simulation run times (for each chain) range from 12.2 to 21.46 hours with a mean of 16.2 hours. (Some longer run times resulted from shared usage of computers).

7.2 RJMCMC Algorithm Performance and Convergence Assessment

Table B.1 shows acceptance rates for dimension-changing moves, calculated for each data set from all 600,000 sweeps from the 3 chains. Acceptance rates for split/combine range from 0.008 to 0.0522, while those for birth/death are considerably lower, ranging from 0.002 to 0.016. Richardson and Green (1997) report

acceptance rates between 0.04 and 0.18 in their one-dimensional normal mixture RJMCMC sampler. It seems that higher acceptance rates would certainly be desirable; however, there are currently no established standards for dimension-changing MCMC samplers.

Table B.2 displays occurrence rates for move-disqualifying conditions (also using all 600,000 sweeps for each data set). We define a "move-disqualifying condition" as a situation which immediately sets the acceptance probability to 0. The three types of such occurrences in our sampler are:

1. a $NN_{\mathbf{\Sigma}}$ violation in the split move, resulting from an attempt to generate 2 new cluster centers, neither of which is the other's $\mathbf{\Sigma}$-Nearest-Neighbor (see Definition 4.4.1),

2. a split or birth attempt when $k = k_{\text{hi}}$, and

3. a combine or death attempt when $k = k_{\text{lo}}$.

Since we use $k_{\text{lo}} = 1$, the third type is uninteresting. The occurrence of $NN_{\mathbf{\Sigma}}$ violations (on average, in 12% of split attempts) does not seem excessive. In very few cases (and, always at the very beginning of each chain, and only for chains starting at $k = k_{\text{hi}}$) was a split or birth move attempted at $k = k_{\text{hi}}$, suggesting that our choice of k_{hi} is reasonable. (Note: these occurrences do not imply that the move would have been accepted if k_{hi} were higher; the acceptance probability in such cases is set to 0 before any other components are calculated).

We observed trace plots for each RJMCMC simulation (not shown, except for chain 2 for the Redwood pattern in Figures D.1 – D.2) to make sure that no anomalies occurred. In all cases, the value of k settled to its eventual neighborhood within several hundred sweeps.

The convergence assessment method of Algorithm 5.4.1 was applied to the saved sweeps (20,000 per chain) in each pattern, for the parameters discussed in section 5.1, with a base batch size $b = 500$ (yielding 20 total batches). Henceforth a "sweep" will refer to a saved state (for example, "sweep 100" corresponds to the 1000^{th} sweep of the original chain). Relevant plots are shown in Figures E.1 – E.13. Convergence appears to be attained remarkably quickly in each case. In the numerator-and-denominator trace plots (right-hand side of each page), the two lines in each pair are practically indistinguishable and stabilize (with the possible exception of Figures E.4(b) and E.13(b)) to a common value by the 15^{th} batch (most far sooner). MPSRF's are *never* higher than 1.2 (even for the first batch, which analyzes sweeps 500–1000), and they stay below 1.01 past the 15^{th} batch in each case.

Although we could justify a diagnosis of convergence at the 15^{th} batch (or even sooner), we conservatively chose the 20^{th} batch and declared the last 10,000 sweeps of each chain suitable for inference. Thus, in terms of section 6.1, we have $C = 3$, $T_{\text{ch}} = 10,000$ and $T = 30,000$.

Programs for convergence assessment were written in S-Plus and implemented in S-Plus version 4.5 for Windows. The run time for each data set was only a few minutes.

7.3 BVNPCP-BHM Model Adequacy Assessment

The methods discussed in section 6.3 were applied to all 13 data sets to ascertain whether model assumptions are supported by the data. For each pattern, only values of k occurring at a frequency of at least 0.001 (i.e., at least 30 times) are considered. Since all simulated patterns were generated from only one kind of

model, the BVNPCP-BHM, it is difficult to assess the performance of these methods. The most informative evaluation is obtained by observing results for correct vs. incorrect values of k for a given data set. A study of the behavior of the methods for patterns which deviate from model assumptions would clearly be useful, and is encouraged in future research.

Chi-square plots for each pattern and k (not displayed) constructed at the mode show no signs of trouble, except for occasional outliers for some values of k (which do not correspond consistently to correct vs. incorrect k). Figures H.1 – H.2 display p-values for the 3 discrepancy measures used. P-values from $D_{\mathrm{CR}}\left(\mathbf{Y}^{\mathrm{obs}};\widetilde{\boldsymbol{\theta}}^{(\cdot|k)}\right)$, analyzed at each mode only, fluctuate wildly and have no discernible relationship to k, thus casting doubt on its effectiveness. Those from $D_{\mathrm{CR}}\left(\mathbf{Y}^{\mathrm{obs}};\boldsymbol{\theta}^{(t|k)}\right)$ (with Monte Carlo simulation) are much more stable, indicating trouble in simulated patterns only for $k = 7 - 9$ for AI-3-k7-b (which is strange, since these are equal or close to the actual k values). For the Redwood pattern, p-values are very high for low k, implying that for these k, clusters are much smaller than expected under the model. Finally, p-values for $D_{\boldsymbol{\Sigma}}\left(\mathbf{Y}^{\mathrm{obs}};\boldsymbol{\theta}^{(t|k)}\right)$ are extremely stable, rarely deviating far from 0.5. We realize that this discrepancy measure is probably a poor choice: although it is based on proper intentions of evaluating the conformity of posterior $\boldsymbol{\Sigma}$ samples to the data, the computation of the $\widehat{\boldsymbol{\Sigma}}$ estimate used in $D_{\boldsymbol{\Sigma}}\left(\mathbf{Y}^{\mathrm{obs}};\boldsymbol{\theta}^{(t|k)}\right)$ is similar to the generation of $\boldsymbol{\Sigma}$ from its full conditional distribution in the RJMCMC sampler (the primary difference being use of $\boldsymbol{\mu}$ values vs. sample cluster means), and thus is bound to be well-behaved regardless of model appropriateness.

Box plots of $\widehat{CPO}_{j|k}$ (see section 6.3.2) values are shown in Figures H.3 – H.4. For each pattern, distribution of $\widehat{CPO}_{j|k}$ varies mostly in the upper tail and consistently shows higher values for higher k.

The $\widehat{d}_{3_{j|k}}$ statistic is more useful than $\widehat{CPO}_{j|k}$ for determining model *adequacy*. Box plots of $\widehat{d}_{3_{j|k}}$ for different k are shown in Figures H.7 – H.8. Histograms (shown in Figure H.9 for the Redwood data) provide a more thorough assessment. For the Redwood pattern, histograms for each k show a mode of $\widehat{d}_{3_{j|k}}$ around 0.7 and an excess of very low values, implying an excess of offspring very close to cluster centers and an excess of isolated offspring with distant parent cluster centers, respectively. This is consistent with a leptokurtic distribution (see section 1.1.5), which is commonly observed in pollen and seed dispersal. Box plots and histograms of $\widehat{d}_{3_{j|k}}$ for the simulated patterns look reasonable in most cases (and very similar across k for each pattern), except for AI-3-k14-b, which exhibits behavior similar to the Redwood pattern.

Programs to compute $\widehat{CPO}_{j|k}$ and $\widehat{d}_{3_{j|k}}$ were written in C++ and implemented similarly to those for the RJMCMC algorithm. Graphical displays are achieved with S-Plus programs. The computations of $\widehat{d}_{3_{j|k}}$, $D_{\mathrm{CR}}\left(\mathbf{Y}^{\mathrm{obs}};\boldsymbol{\theta}^{(t|k)}\right)$ and $D_{\boldsymbol{\Sigma}}\left(\mathbf{Y}^{\mathrm{obs}};\boldsymbol{\theta}^{(t|k)}\right)$ are computer intensive due to a large amount of required Monte Carlo simulation, but can be performed simultaneously. Run times were not recorded, but seemed to average about 6 hours per data set.

7.4 Inference for k: RJMCMC and Composite EM

For the purpose of model comparison (in our case, equivalent to inference for k), the same k are used as in section 7.3. As mentioned in section 6.4.2, model adequacy criteria can also be used for comparing models. The discrepancy measures (Figures H.1 – H.2) do not seem to favor any values of k over others, except for the Redwood pattern, where smaller k correspond to a worse fit (reflecting tighter

clustering than expected under model assumptions). The box plots for $\widehat{CPO}_{j|k}$, and especially plots of $\sum_{j=1}^{n} \log \widehat{CPO}_{j|k}$ (Figures H.5 – H.6), favor higher k for most patterns. We suspect that $\sum_{j=1}^{n} \log \widehat{CPO}_{j|k}$ may be too sensitive to extremely low $\widehat{CPO}_{j|k}$ values, which do occur in our analyses. Wherever box plots or histograms of $\widehat{d}_{3_{j|k}}$ indicate questionable model fits, they generally do so for all k and thus do not appear to be effective for model comparison.

The posterior density estimates of k from RJMCMC are shown as histograms in Figures F.1 – F.2, and as interval estimates of model probabilities (95% confidence interval for each k: see section 6.4.1) in Figures G.1 – G.6. Table M.1 (far right column) shows, for each data set, the minimum number of batches used (minimum over k) in computing the variance estimates.

Estimated model probabilities from composite EM (see section 3.5.1) are shown in Figures G.1 – G.2. For each pattern, we computed $\widehat{p}(k|\mathbf{Y})$ from the composite EM estimate for all $(k_{\min} - 4, \ldots, k_{\max} + 4)$, where $(k_{\min}, \ldots, k_{\max})$ are k visited by RJMCMC with frequency ≥ 0.001. (Note: $k = 1$ was not used). Then we continued to try more values of k until the estimated probabilities for the lowest four (unless $k = 2$ had been included) and highest four values were negligible. For two patterns, the Redwoods and AI-3-k14-b, values of k higher than those visited by RJMCMC were given non-negligible probabilities. To be on the safe side, we calculated BIC_k^{EM} for $k = 2, \ldots, 80$ for each of these patterns and observed that no additional values of k were supported.

Figures G.3 – G.6 show estimated model probabilities computed from various marginal likelihood estimates using RJMCMC output (see section 6.4.3). The label "RHarmMn" refers to $\widehat{p}_4(\mathbf{Y}|k)$, while "HarmMean" refers to $\widehat{p}_2(\mathbf{Y}|k)$.

Note that for some patterns (e.g., AI-1.5-k7-b and AI-3-k7-a), there is one

clearly dominant k, but for others, there is more variability.

In general, $\{\widehat{p}(k|\mathbf{Y})\}$ from RJMCMC visit frequencies (VF) seems to agree most closely with those from $BIC_k^{\max}$, second most with BIC_k^{mean} (which tends to favor slightly lower k), somewhat with AWE_k^{mean} and AIC_k^{mean} (which favor moderatly *higher* k), very little with BIC_k^{EM}, $AWE_k^{\max}$ and $AIC_k^{\max}$ (which favor much higher k), and least of all with $\widehat{p}_2(\mathbf{Y}|k)$ and $\widehat{p}_4(\mathbf{Y}|k)$ (which almost always favor the highest k's). The RJMCMC visit frequencies almost always place highest estimated model probabilities on k *below* the true value in simulated patterns, and most other methods do this most of the time. A probable explanation for this is that methods can be easily "fooled" into treating overlapping clusters as single clusters, but there is nothing to encourage *over*estimation of k.

As would be expected, model probability estimates using likelihood *averages* favor lower k's than corresponding versions using likelihood *maximums*, since the penalty carries more weight in the presence of smaller likelihoods. Since AIC has a lower penalty than BIC, it makes sense that it would prefer higher k.

Estimated model probabilities from composite EM are remarkably similar to those from $AIC_k^{\max}$, and also quite similar to those from $AWE_k^{\max}$. We suspect that this is due to the combination of 2 opposite forces:

1. composite EM uses a true (local, at least) maximum likelihood estimate, while $AIC_k^{\max}$ and $AWE_k^{\max}$ only use estimates from the largest sample likelihoods observed in RJMCMC output, and

2. BIC enforces a larger penalty than AIC, and seemingly AWE also. We anticipate that model probability estimates from composite EM would agree

more with $BIC_k^{\max}$ as the number of sweeps in the RJMCMC sampler approaches ∞, since higher and higher sample mixture likelihood values will be encountered, perhaps occasionally approaching the MLE.

The importance sampling marginal likelihood estimates ($\widehat{p}_2(\mathbf{Y}|k)$ and $\widehat{p}_4(\mathbf{Y}|k)$) appear to suffer the same fate as $\sum_{j=1}^{n} \log \widehat{CPO}_{j|k}$, being sensitive to extremely low values going into the harmonic mean (or, in the case of $\widehat{p}_4(\mathbf{Y}|k)$, a slightly robustified harmonic mean). A handful of extremely small mixture likelihood values tend to occur in most models (as ascertained by observing computed values from several data sets), producing significant impacts on the harmonic means. We suspect that this is the cause of higher k's being consistently favored by these methods: for higher k, cluster centers are more numerous and thus more likely to be able to accommodate isolated offspring, preventing them from contributing extremely small values to the mixture likelihood.

Overall, the different methods considered to estimate the number of clusters exhibit quite different behavior. A similar conclusion is reached by Raftery (1995, section 10.5), who compares the Laplace-Metropolis estimator, $\widehat{p}_2(\mathbf{Y}|k)$, another importance sampling marginal likelihood estimator using a specially constructed importance density, BIC and AWE for a one-dimensional normal mixture. He states that the "Laplace-Metropolis estimator is the only one that seems to be in the right ballpark." Most striking perhaps is the fact that composite EM places high probability on values of k that are not supported at all in RJMCMC, for the Redwoods and AI-3-k14-b. For both patterns, the Markov chain started at $k = 30$ certainly has a chance to explore the same possibilities, but quickly moves to lower k values. RJMCMC samplers consistently spend most of their time visiting k below its true value in the 12 simulated patterns, leading to suspicion that perhaps RJMCMC for

the BVNPCP-BHM systematically underestimates k in general.

It seems that, surprisingly, the behavior of these various model-comparison methods is not very predictable from visual assessment of how many clusters there appear to be, or how well-separated they are.

All marginal likelihood estimators from RJMCMC ($\widehat{p}_2(\mathbf{Y}|k)$, $\widehat{p}_4(\mathbf{Y}|k)$, $BIC_k^{\max}$, BIC_k^{mean}, $AWE_k^{\max}$, AWE_k^{mean}, $AIC_k^{\max}$, AIC_k^{mean}) are computed in C++ programs, and corresponding model probabilities are computed and displayed in S-Plus. Composite EM estimates for each k, and corresponding BIC_k^{EM}, are obtained from MCLUST/EMCLUST (Fraley, 1998), a suite of S-Plus and Fortran routines for EM analysis of mixture models. Estimated model probabilities from composite EM are computed and displayed in S-Plus. Run times are very short (several seconds) for these methods.

7.5 Inference for Σ: RJMCMC and Composite EM

Estimates of various components and parameterizations for $\boldsymbol{\Sigma}$ can be constructed both from RJMCMC output (section 6.5) and composite EM estimates (section 3.5.2), and also used for tests of isotropy. Output from all post-convergent sweeps is used in RJMCMC methods. Information on batch sizes used in RJMCMC batch sampling methods is displayed in Table M.1. Figures I.1 – I.2 show 90%, 95% and 99% confidence regions for $\boldsymbol{\sigma}^c$ (see (3.32)) using batch sampling ("BS") and composite EM methods, and pairs of 95% HPD intervals for the two scalar components of $\boldsymbol{\sigma}^c$ (with a joint 90% confidence level). Also shown are associated p-values of isotropy tests. The point of isotropy ($\boldsymbol{\sigma}^c = 0$) is indicated in each plot, and the true value is indicated in the plot for each simulated data set. Figures J.1 – J.12

display posterior density estimates from RJMCMC (see section 6.2.4) and 95% confidence intervals from batch sampling and composite EM methods, for the scalar parameterizations $\log \sigma_{11}$, $\log \sigma_{22}$, $z(\rho_{12})$, $\log \gamma$, $\log \Psi$ and ϕ (see section 3.5.2). The true model values are shown for each simulated pattern.

We emphasize that for RJMCMC, the most informative and theoretically sound output analysis methods are posterior density estimation and HPD intervals. Normal approximations used to construct confidence intervals and test statistics for *both* batch sampling methods *and* composite EM analysis are certainly questionable. However, posterior means from RJMCMC and point estimates from composite EM (easily located on plots as the centers of confidence regions and intervals) are still valid, as are variance estimates in composite EM.

One could argue that a presentation of RJMCMC posterior density estimates, HPD intervals and corresponding isotropy p-values, and composite EM estimates and associated variances would constitute a sufficient summary of results for analysis of $\mathbf{\Sigma}$. However, since we have a collection of simulated data sets with known model values, we proceed to implement and discuss inferential results using the distributional approximations. Although a set of only 12 simulated patterns is nowhere near enough for a proper performance study, we can at least get a general idea of the behavior of the methods considered.

In confidence intervals for ϕ, those from batch sampling which are ill-defined are represented with ellipsis ($\cdots$) markings (see section 6.2.3). For I-k14-b, the composite EM interval for ϕ covers the entire range and is not shown.

We are hesitant to trust confidence regions and test results from composite EM for the Redwoods and AI-3-k14-b since highly separated values of k contribute to the estimates, casting serious doubt on the validity of the normal approximation

(see section 3.5.1). However, for lack of a better alternative, and in the interest of comparing the performance of composite EM to that of other methods, we proceed as usual for these data sets.

The confidence regions and intervals from composite EM and HPD agree fairly well in most cases, and those from batch sampling are occasionally similar but usually much wider. Isotropy / anisotropy is diagnosed correctly by all methods for all simulated patterns, except batch sampling for AI-1.5-k7-a and AI-1.5-k14-b, in which ridiculously large confidence regions are produced. As mentioned in section 6.5.3, batch sampling appears to frequently over-estimate variances. Unreasonably large batch sampling variance estimates tend to occur most in patterns whose RJMCMC samplers produced high ACF's (investigated for each pattern and showed for the Redwoods in Figure D.4) and high variability in k (these two phenomena typically occurring together).

For the Redwood pattern, the isotropy test is strongly rejected by batch sampling ($p = 0.000291$) and HPD intervals ($p = 0$) but borderline for composite EM ($p = 0.0572$). This represents a notable exception to the trend of HPD intervals and composite EM regions being similar; the variance estimates in composite EM are much larger due to strong contributions from very different k's.

True values for $\boldsymbol{\sigma}^c$ are contained in 90% confidence regions in all cases *except*:

1. for batch sampling: AI-1.5-k7-b

2. for composite EM: AI-1.5-k7-a (99% CR contains true $\boldsymbol{\sigma}^c$), AI-1.5-k7-b, and AI-3-k7-a (95% CR contains true $\boldsymbol{\sigma}^c$)

3. for joint HPD intervals: AI-1.5-k7-a (although very close), AI-1.5-k7-b, AI-1.5-k14-b, AI-3-k7-a (very close), and AI-3-k14-a.

(Note: only 95% HPD intervals were computed; so, information for other confidence levels for the HPD method is not shown).

Table 7.4 shows the number of times true values were included in scalar confidence intervals. There is no "true value" for ϕ in isotropic patterns. The 4 instances of the failure of HPD intervals to include the true value of $\log\gamma$ occur in the 4 isotropic patterns: it is of course not possible for a HPD interval for $\log\gamma$ to contain 0.

Parameter	Batch Sampling	Composite EM	HPD Interval
$\log\sigma_{11}$	11	8	6
$\log\sigma_{22}$	12	9	11
$z(\rho_{12})$	11	10	11
$\log\gamma$	12	11	8
$\log\Psi$	11	9	7
ϕ	7 (of 8)	6 (of 8)	4 (of 8)

Table 7.4: Coverage of true value achieved by 95% confidence intervals for simulated patterns. Entry is number out of 12 (or 8, in the case of ϕ) patterns in which true value is contained.

The batch sampling intervals almost always include the truth but are unnecessarily large. Intervals from composite EM are consistently lower for "size" parameters ($\log\sigma_{11}, \log\sigma_{22}, \log\Psi$) and $\log\gamma$ than corresponding HPD intervals. This is probably due to the fact that composite EM favors higher k's than RJMCMC, and thus smaller clusters (presumably with more variable shape as well). Composite

EM intervals are usually bigger than corresponding HPD intervals for

$$\log \sigma_{11}, \log \Psi, \text{ and } (\phi \text{ for "k14" patterns})$$

and smaller for

$$\log \sigma_{22}, z(\rho_{12}), \text{ and } (\phi \text{ for "k7" patterns}).$$

For the Redwood data, confidence intervals from composite EM are huge compared to HPD intervals, and often larger than those from batch sampling, likely due to strong contributions from a wide range of k (7-9,15,20-24) compared to RJMCMC (7-14).

A brief overall assessment of point estimates $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ from RJMCMC posterior means and composite EM analysis is provided in Figures K.1 – K.2, which display bivariate normal contours of the offspring dispersal distribution (drawn to scale with the boundary) characterized by each $\widehat{\mathbf{\Sigma}}$ and by each true $\mathbf{\Sigma}$ (for simulated patterns). It is apparent from these plots that RJMCMC consistently produces larger cluster size estimates than composite EM. Despite the tendency of RJMCMC to favor low values of k, its cluster shape/scale point estimates appear quite reasonable (very close to the truth for 6 simulated patterns, larger size for 5, and smaller size for 1). The composite EM estimates look even better (except for AI-3-k14-b, in which BIC_k^{EM} is mysteriously fond of $k = 21$).

Analogous plots constructed separately for k = 7,10,12 and 15 are shown in Figure K.3 for the Redwood data. The relationship between k and estimated cluster size is quite apparent. Posterior density estimates for various scalar components of $\mathbf{\Sigma}$, computed separately for each of these k and overlaid with the all-k estimates, are displayed in Figure L.1.

Source code was written in C++ for all batch sampling and HPD interval calculations. A suite of S-Plus functions for circular data analysis (Davies, 1996) is

used to compute and display circular density estimates. Composite EM parameter estimates and variances are computed in S-plus programs written to operate on output from MCLUST/EMCLUST (Fraley, 1998). S-Plus programs were written to compute all other intervals and tests and to display all graphs. Run times are longest for computation of composite EM variance estimates (several hours if $k \geq 20$ are included), but would be drastically decreased by conversion to C++. All other run times are reasonable (seconds or minutes).

CHAPTER 8
CONCLUSION

8.1 Summary of New Methods

The composite EM and RJMCMC approaches are, to the author's knowledge, the first fully developed methods for inference for model parameters of any spatial cluster process and/or anisotropic point process.

The aim of the EM algorithm applied to mixture models is usually to estimate the number of components and cluster centers. Our approach is somewhat the opposite in that we treat these as nuisance parameters and focus on estimation of the common "cluster shape/scale parameter" $\mathbf{\Sigma}$ (although we still have several methods to assess the number of clusters). Composite EM (Chapter 3) is apparently the first method in the field of mixture analysis to:

1. combine estimates of $\mathbf{\Sigma}$ for different numbers of components into a composite estimate $\widehat{\boldsymbol{\sigma}} = (\hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{12})$,

2. compute asymptotic variance estimates directly from the observed information matrix (for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$) without relying on approximations to the form of the observed information matrix, or

3. develop an overall estimate of the variance-covariance matrix of the $\widehat{\boldsymbol{\sigma}}$ accounting for uncertainty in k.

Even if the normal approximation to the distribution of $\widehat{\boldsymbol{\sigma}}$ is inappropriate, point estimates and variance estimates are still reasonable.

The RJMCMC algorithm developed for the BVNPCP-BHM(A, n) (Chapter 4) is the first RJMCMC algorithm capable of modeling mixtures of more than one dimension (ours being two-dimensional).

The convergence assessment technique developed in Chapter 5 is the first convergence assessment method with a solid theoretical foundation for RJMCMC, and the first multivariate technique (i.e., capable of analyzing convergence of parameter vectors) for any dimension-changing MCMC sampler. It is applicable to *any* RJMCMC sampler with a parameter which retains the same meaning across models.

In RJMCMC output analysis (Chapter 6), we develop (to the author's knowledge) the first usage of batch sampling or HPD intervals to estimate a circular parameter from any MCMC method. It is perhaps also the first established method to estimate circular parameters from a MCMC sampler without treating them as linear.

8.2 Scope for Future Research

Perhaps the most appropriate next step would be to simulate a much larger number of point patterns to enable a more thorough study of the performance of methods developed in this thesis. The large number of sweeps used in RJMCMC in this thesis was necessitated only by batch sampling; a much smaller number of sweeps would suffice for all other methods. Inclusion of patterns which deviate in different ways from model assumptions would help to determine the robustness of the methods. A study of the effect of smaller or larger samples sizes (total number of offspring) would also add significantly to understanding of their behavior and success.

In the composite EM technique, the regularity conditions guaranteeing the

correct asymptotic distribution of $\widehat{\boldsymbol{\sigma}}^{(k)}$ from the EM algorithm should be checked. We suspect that they hold (e.g., third-order partial derivatives are known to satisfy the appropriate criteria for multivariate normal distributions, a result which is likely to extend to mixtures of multivariate normals). A study of the appropriateness (or lack thereof) of the normal approximation to the asymptotic distribution of the composite EM estimate $\widehat{\boldsymbol{\sigma}}$ is also clearly in order. Although the agglomerative clustering method is considered to generate good starting values for the EM algorithm, it would be prudent to try a battery of different starting values to check whether a larger local maximum can be obtained. Assessment of the accuracy of BIC_k^{EM} in estimating model probabilities is another priority. Several alternative EM-type algorithms (e.g. classification EM and stochastic EM) exist and could be used to develop similar composite methods combining information from separate analyses by k (Celeux, Chauveau, and Diebolt, 1996; Diebolt and Ip, 1995; Celeux and Govaert, 1995).

For the RJMCMC algorithm, incorporation of some kind of label estimation method or ordering restriction would allow a greater variety of output analysis options. The RJMCMC sampler should be compared with fixed-k MCMC samplers to see if it provides a beneficial "tunneling" effect (i.e., ability to move between distant high-probability regions of the parameter space that are separated by valleys of low density, via dimension-changing jumps). Richardson and Green (1997, p. 751) report that in previous work in mixture estimation, fixed-k samplers have been plagued by slow mixing. They carry out a small experiment for a simple univariate mixture model which demonstrates that output from a RJMCMC sampler collected for a particular value of k mixes faster than output from a corresponding fixed-k sampler. It would also be interesting to see how inferences using RJMCMC output

separated by k compare to those from the fixed-k samplers. Different orderings of move types can be attempted, to see if any particular orderings affect mixing or output analysis results. There are numerous possibilities for modification of existing move types and invention of new ones. A study of the sensitivity of the results to specification of the prior distribution should be performed.

The validity of the RJMCMC convergence assessment technique in the absence of certain ANOVA assumptions (especially independence of samples) should be assessed. Since convergence seemed to occur very quickly for all examples in this thesis, a study of the sensitivity of the diagnostics to different types of violations of convergence would help to define the effectiveness of the technique. It may be possible to construct additional diagnostics (e.g., using different ratios of mean-squares) to add to the ability to detect convergence failure.

The model adequacy criteria used in RJMCMC were for the most part inconclusive in assessing the patterns studied in this thesis. A study of the sensitivity of these methods to various deviations from model assumptions would help to determine their potential. Better discrepancy measures could certainly be developed.

For model comparison, it would be desirable to seek an acceptable way to use a Laplace-Metropolis estimator even in the presence of label-switching; for many authors, it is the Bayes factor approximation of choice. Better importance sampling-based marginal likelihood estimators could also be constructed by the use of different importance-sampling densities.

The mystery of frequent over-estimation of variance by batch sampling methods should definitely be analyzed in more detail; we suspect that the cause is *negative* autocorrelation at higher lags. The validity of the normal approximation relied on to

construct confidence regions and test statistics should also be assessed. Perhaps *bivariate* HPD regions (using a 2-dimensional kernel density estimate computed from the posterior samples) could be explored; resulting tests and confidence regions would likely be less conservative.

Analysis of a larger number of simulated point patterns would be helpful in determining the coverage probabilities of confidence regions produced by the different methods. The asymptotic properties of HPD intervals for circular parameters can also be researched.

There are many possible alterations/extensions to the BVNPCP model considered in this thesis that are bound to yield more widely applicable methods. An attempt should be made to account for the effects of the boundary of the study region, either by adjusting the model specification or incorporating edge-corrections into estimators. Unequal mixing proportions (i.e., possibly different expected numbers of offspring per cluster) can easily be modeled; perhaps isolated offspring would more easily be accommodated. Alternatively (or in addition), the model can be expanded to allow for "noise" (events not belonging to any clusters), as in Fraley and Raftery (1998). If more information is available *a priori* (e.g., probabilistic assertions about relationships between certain offspring, certain regions more likely to contain parent events, etc.), then a more informative prior could be used. For example, genetic data recorded for seedlings can produce prior probabilities of each pair of seedlings descending from a common parent. Genetic and spatial data are usually analyzed separately in the study of population dynamics; the combination of these two types of data offers the prospect of improved ecological inferences. The location of a set of potential parent trees in a region (perhaps only an unknown fraction of which can actually produce seedlings) can provide a mixture of discrete and

uniform distributions for a prior on $\boldsymbol{\mu}$. Types of dispersal distributions other than bivariate normal can be modeled, dramatically improving the applicability of the methods. For example, a dispersal distribution can be based on Gaussian plumes, which are typically used to model the flow of particles from a smokestack in the presence of a prevailing wind (see Thompson and Greenkorn, 1988; Pasquill, 1974).

Composite EM has a somewhat limited capacity to incorporate these kinds of extensions due to heavy reliance on closed-form solutions for maximization, asymptotic variances, etc., which may be rendered intractable by overly complex models. However, RJMCMC is a very flexible technique which requires only the quantities used in Metropolis Hastings moves (traditional or reversible jump) to be known analytically. As better understanding of convergence assessment methods and validity of various output analysis techniques increases, RJMCMC may become a more powerful and widely usable tool for analysis of quite complicated variable-dimension models.

APPENDIX A
SELECTED PROOFS AND DERIVATIONS

A.1 Proof of Theorem 1.1.7

Proof: Consider the PCP observed in a finite region A. Assume that $h(\cdot)$ is continuous.

Let

$(\mathrm{a}_1, \mathrm{b}_1), \ldots, (\mathrm{a}_{n_d}, \mathrm{b}_{n_d})$ be all ordered pairs of events in A from different parents

and

$(\mathrm{c}_1, \mathrm{d}_1), \ldots, (\mathrm{c}_{n_s}, \mathrm{d}_{n_s})$ be all ordered pairs of events in A from the same parent

where

$$n_d = \#(\text{ordered pairs of events in } A \text{ from different parents})$$

and

$$n_s = \#(\text{ordered pairs of events in } A \text{ from the same parent}).$$

Note that

$$n_s = \sum_{i=1}^{n_p} S_i(S_i - 1)$$

where

$$n_p = \#(\text{parents}).$$

Also define

$$N_i(\mathrm{d}\mathbf{x}) = \#(\text{offspring from parent } i \text{ in region } \mathrm{d}\mathbf{x}).$$

As in the proof of Theorem 1.1.5, the notation suppresses dependence on A.

Consider two fixed locations $\mathbf{x}, \mathbf{y} \in A$ (well in the interior of A so that boundary effects are negligible). Then

$$N(\mathrm{d}\mathbf{x})N(\mathrm{d}\mathbf{y}) = \sum_{j=1}^{n_d} 1_{\mathrm{d}\mathbf{x}}(\mathbf{a}_j)1_{\mathrm{d}\mathbf{y}}(\mathbf{b}_j) + \sum_{j=1}^{n_s} 1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j)1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j)$$

and so

$\lambda_2(\mathbf{x}, \mathbf{y})$

$$= \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} \frac{E\left[N(\mathrm{d}\mathbf{x})N(\mathrm{d}\mathbf{y})\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|}$$

$$= \lim_{A\to\Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} \frac{E\left[\sum_{j=1}^{n_d} 1_{\mathrm{d}\mathbf{x}}(\mathbf{a}_j)1_{\mathrm{d}\mathbf{y}}(\mathbf{b}_j) + \sum_{j=1}^{n_s} 1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j)1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j)\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right\}$$

$$= \lim_{A\to\Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} \frac{E\left[\sum_{j=1}^{n_d} 1_{\mathrm{d}\mathbf{x}}(\mathbf{a}_j)1_{\mathrm{d}\mathbf{y}}(\mathbf{b}_j)\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right\} \tag{A.1}$$

$$+ \lim_{A\to\Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} \frac{E\left[\sum_{j=1}^{n_s} 1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j)1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j)\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right\} \tag{A.2}$$

We can simplify (A.1) as follows:

$$\lim_{A\to\Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} \frac{E\left[\sum_{j=1}^{n_d} 1_{\mathrm{d}\mathbf{x}}(\mathbf{a}_j)1_{\mathrm{d}\mathbf{y}}(\mathbf{b}_j)\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right\}$$

$$= \lim_{A\to\Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} \frac{E\left[\sum_{i=1}^{n_p} \sum_{j=1,j\neq i}^{n_p} N_i(\mathrm{d}\mathbf{x})N_j(\mathrm{d}\mathbf{y})\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right\}$$

$$= \lim_{A\to\Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} \frac{E\left[n_p\left(n_p-1\right)E\left\{N_i(\mathrm{d}\mathbf{x})N_j(\mathrm{d}\mathbf{y})\right\}\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right\}$$

(by Lemma 1.1.4: n_p and $\{N_i(\mathrm{d}\mathbf{x})N_j(\mathrm{d}\mathbf{y})\}$ are independent,

and $\{N_i(\mathrm{d}\mathbf{x})N_j(\mathrm{d}\mathbf{y})\}$ are i.i.d. (for $i \neq j$))

$$= \lim_{A\to\Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}|\to 0} E\left[n_p\left(n_p-1\right)\right] \left(\frac{E\left[N_i(\mathrm{d}\mathbf{x})\right]}{|\mathrm{d}\mathbf{x}|}\right)\left(\frac{E\left[N_j(\mathrm{d}\mathbf{y})\right]}{|\mathrm{d}\mathbf{y}|}\right) \right\}$$

(by independence of N_i and N_j)

$$= \lim_{A \to \Re^2} \left\{ (\rho|A|)^2 \left[\frac{ES}{|A|} \right] \left[\frac{ES}{|A|} \right] \right\}$$

(by stationarity, holding when $A \to \Re^2$)

$$= \lim_{A \to \Re^2} \left\{ \rho^2 \nu^2 \right\}$$

$$= \lambda^2$$

(by Theorem 1.1.5)

We can simplify (A.2) as follows:

$$\lim_{A \to \Re^2} \left\{ \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}| \to 0} \frac{E\left[\sum_{j=1}^{n_s} 1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j) 1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j)\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right\}$$

$$= \lim_{A \to \Re^2} \left\{ [E n_s] \left[\lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}| \to 0} \frac{E\left[1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j) 1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j)\right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right] \right\}$$

(by Lemma 1.1.4: n_s and $\{1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j) 1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j)\}$ are independent,

and $\{1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j) 1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j)\}$ are i.i.d.)

$$= \lim_{A \to \Re^2} \left\{ E\left[\sum_{i=1}^{n_p} S_i(S_i - 1) \right] \left[\lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}| \to 0} \frac{E\left[E\left\{ 1_{\mathrm{d}\mathbf{x}}(\mathbf{c}_j) 1_{\mathrm{d}\mathbf{y}}(\mathbf{d}_j) \;\middle|\; \mathbf{p} \right\} \right]}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \right] \right\}$$

(where $\mathbf{p}$ is the location of the common parent of $\mathbf{c}_j$ and $\mathbf{d}_j$)

$$= \lim_{A \to \Re^2} \left\{ \left[E\left\{ n_p \right\} \right] \left[E\left\{ S(S-1) \right\} \right] \cdot \right.$$
$$\left. \left[\lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}| \to 0} \frac{1}{|\mathrm{d}\mathbf{x}||\mathrm{d}\mathbf{y}|} \int_A P\left\{ \mathbf{c}_j \in \mathrm{d}\mathbf{x}, \mathbf{d}_j \in \mathrm{d}\mathbf{y} \;\middle|\; \mathbf{p} \right\} \frac{1}{|A|} \mathrm{d}\mathbf{p} \right] \right\}$$

(by Lemma 1.1.4: n_p and $\{S_i(S_i-1)\}$ are independent,

and $\{S_i(S_i-1)\}$ are i.i.d.;

and since $\mathbf{p}$ is uniformly distributed on A, having p.d.f. $\frac{1}{|A|}$)

$$= \lim_{A \to \Re^2} \left\{ \left[\rho |A| E\left\{ S(S-1) \right\} \right] \cdot \right.$$
$$\left. \left[\frac{1}{|A|} \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}| \to 0} \int_A \frac{P\left\{ \mathbf{c}_j \in \mathrm{d}\mathbf{x} \;\middle|\; \mathbf{p} \right\}}{|\mathrm{d}\mathbf{x}|} \frac{P\left\{ \mathbf{d}_j \in \mathrm{d}\mathbf{y} \;\middle|\; \mathbf{p} \right\}}{|\mathrm{d}\mathbf{y}|} \mathrm{d}\mathbf{p} \right] \right\}$$

(since the locations of $\mathbf{c}_j$ and $\mathbf{d}_j$ are indpendent given $\mathbf{p}$)

$$= \lim_{A \to \Re^2} \left\{ \left[\rho E\left\{ S(S-1) \right\} \right] \cdot \right.$$
$$\left. \left[\lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}| \to 0} \int_A \left\{ \frac{1}{|\mathrm{d}\mathbf{x}|} \int_{\mathrm{d}\mathbf{x}} h\left(\mathbf{u}-\mathbf{p}\right) \mathrm{d}\mathbf{u} \right\} \left\{ \frac{1}{|\mathrm{d}\mathbf{y}|} \int_{\mathrm{d}\mathbf{y}} h\left(\mathbf{u}-\mathbf{p}\right) \mathrm{d}\mathbf{u} \right\} \mathrm{d}\mathbf{p} \right] \right\}$$

$$= \rho E\left\{ S(S-1) \right\} \cdot$$
$$\lim_{A \to \Re^2} \left\{ \int_A \lim_{|\mathrm{d}\mathbf{x}|,|\mathrm{d}\mathbf{y}| \to 0} \left\{ \frac{\int_{\mathrm{d}\mathbf{x}} h\left(\mathbf{u}-\mathbf{p}\right) \mathrm{d}\mathbf{u}}{|\mathrm{d}\mathbf{x}|} \frac{\int_{\mathrm{d}\mathbf{y}} h\left(\mathbf{u}-\mathbf{p}\right) \mathrm{d}\mathbf{u}}{|\mathrm{d}\mathbf{y}|} \right\} \mathrm{d}\mathbf{p} \right\}$$

(interchanging limit and integral, using Lemma 1.1.6 and

the Bounded Convergence Theorem: see Chung (1974, p. 42))

$$= \rho E\left\{S(S-1)\right\} \lim_{A \to \Re^2} \left\{ \int_A h\left(\mathbf{x}-\mathbf{p}\right) h\left(\mathbf{y}-\mathbf{p}\right) \mathrm{d}\mathbf{p} \right\}$$

(by the Fundamental Theorem of Calculus:

see Khuri (1993, Theorem 6.4.8))

$$= \rho E\left\{S(S-1)\right\} h_2(\mathbf{x}-\mathbf{y})$$

(by the definition of h_2)

Finally, putting it all together, we have

$$\lambda_2(\mathbf{x},\mathbf{y}) = (A.1) + (A.2) = \lambda^2 + \rho E\left\{S(S-1)\right\} h_2(\mathbf{x}-\mathbf{y}). \quad \square$$

A.2 Derivation of $E(X^n)$ for $X \sim \text{Poiss}(\lambda)$

Suppose $X \sim \text{Poiss}(\lambda)$ and n is a positive integer. Let $\Psi(t)$ denote the moment generating function of Poiss(λ),

$$\Psi(t) = \exp\left\{\lambda\left[\exp(t)-1\right]\right\},$$

and $\Psi^{(m)}(t_0)$ denote the m^{th} derivative of $\Psi(t)$ with respect to t evaluated at t_0,

$$\Psi^{(m)}(t) = \left.\frac{\partial^m \Psi(t)}{\partial t^m}\right|_{t_0}.$$

Then $E(X^n) = \Psi^{(n)}(0)$.

We prove the following lemma by induction:

Lemma A.2.1 *For any positive integer n,*

$$\Psi^{(n)}(t) = \sum_{j=1}^{n} a_{n,j} \lambda^j \exp\left\{\lambda\left[\exp(t)-1\right] + jt\right\},$$

where

$$a_{n,j} = \begin{cases} 1, & \text{if } j = 1 \text{ or } j = n \\ j(a_{n-1,j}) + a_{n-1,j-1}, & \text{otherwise.} \end{cases} \tag{A.3}$$

Proof: First note that $\Psi^{(1)}(t) = \lambda \exp\{\lambda[\exp(t) - 1]\}$, and so the lemma holds for $n = 1$. Suppose that the lemma holds for $n = m$, where $m \geq 1$. We must show that it holds for $n = m + 1$, i.e.,

$$\Psi^{(m+1)}(t) = \sum_{j=1}^{m+1} a_{m+1,j} \lambda^j \exp\{\lambda[\exp(t) - 1] + jt\},$$

where $\{a_{m+1,j}\}$ is given by (A.3):

$$\begin{aligned}
\Psi^{(m+1)}(t) &= \sum_{j=1}^{m} \left(a_{m,j} \lambda^j \exp\{\lambda[\exp(t) - 1] + jt\}\right)(\lambda \exp(t) + j) \\
&= \sum_{j=1}^{m} \left(a_{m,j} \lambda^{j+1} \exp\{\lambda[\exp(t) - 1] + (j+1)t\} \; + \right. \\
&\qquad \left. j a_{m,j} \lambda^j \exp\{\lambda[\exp(t) - 1] + jt\}\right) \\
&= a_{m,1} \lambda \exp\{\lambda[\exp(t) - 1] + t\} \quad + \\
&\qquad \sum_{j=2}^{m} (a_{m,j-1} + j a_{m,j}) \lambda^j \exp\{\lambda[\exp(t) - 1] + jt\} \quad + \\
&\qquad a_{m,m} \lambda^{m+1} \exp\{\lambda[\exp(t) - 1] + (m+1)t\} \\
&= a_{m+1,1} \lambda \exp\{\lambda[\exp(t) - 1] + t\} \quad + \\
&\qquad \sum_{j=2}^{m} a_{m+1,j} \lambda^j \exp\{\lambda[\exp(t) - 1] + jt\} \quad + \\
&\qquad a_{m+1,m+1} \lambda^{m+1} \exp\{\lambda[\exp(t) - 1] + (m+1)t\} \\
&= \sum_{j=1}^{m+1} a_{m+1,j} \lambda^j \exp\{\lambda[\exp(t) - 1] + jt\}.
\end{aligned}$$

Thus the lemma is satisfied for any integer $n \geq 1$, and the proof is complete. □

We can determine $E(X^n)$ from Lemma A.2.1 with $t = 0$:

$$E(X^n) = \Psi^{(n)}(0)$$

$$= \sum_{j=1}^{n} a_{n,j}\lambda^j \exp\{\lambda[\exp(0)-1]+j(0)\}$$
$$= \sum_{j=1}^{n} a_{n,j}\lambda^j, \quad \text{where } \{a_{n,j}\} \text{ are defined by (A.3).}$$

A.3 Simplification of Integral in Observed-data Likelihood (2.13)

Consider the observed-data likelihood of the BVNPCP observed in a region $A \in \Re^2$, as given by (2.13). First, we have

$$\iint_A \cdots \iint_A p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\mu}, k|\Phi, n)\, \mathrm{d}\boldsymbol{\mu}$$
$$= p(k|\Phi,n)p(\boldsymbol{\mu}, \mathbf{s}|k,\Phi,n)p(\mathbf{Z}|\boldsymbol{\mu},\mathbf{s},k,\Phi,n)\iint_A \cdots \iint_A p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\mu},k,\Phi,n)\, \mathrm{d}\boldsymbol{\mu}$$

since $p(k|\Phi,n)p(\boldsymbol{\mu},\mathbf{s}|k,\Phi,n)p(\mathbf{Z}|\boldsymbol{\mu},\mathbf{s},k,\Phi,n)$ is constant in $\boldsymbol{\mu}$ (see (2.5),(2.8),(2.9) and (2.11)). Define the notation

$$[X_i]^{\oplus} = \begin{cases} X_i, & \text{if } S_i > 0 \\ 1, & \text{otherwise} \end{cases} \quad \text{(for any expression } X_i \text{ depending on } i)$$

and

$$[X_i]^{\circledcirc} = \begin{cases} X_i, & \text{if } S_i > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{(for any expression } X_i \text{ depending on } i)$$

and an alternative indexing scheme for $\mathbf{Y}$:

$$\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ik} \equiv (y_{i1;1}, y_{i1;2})', \ldots, (y_{ik;1}, y_{ik;2})'$$
$$\equiv \text{locations of offspring from parent } i.$$

The remaining integral can be re-written as:

$$\iint_A \cdots \iint_A p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\mu},k,\Phi,n)\, \mathrm{d}\boldsymbol{\mu}$$
$$= \iint_A \cdots \iint_A \prod_{i=1}^{k}\prod_{j=1}^{n} [h(\mathbf{y}_j - \boldsymbol{\mu}_i)]^{z_{ji}}\, \mathrm{d}\boldsymbol{\mu}$$

$$
\begin{aligned}
&= \iint_A \cdots \iint_A \prod_{i=1}^{k} \left[\prod_{j=1}^{S_i} h(\mathbf{y}_{ij} - \boldsymbol{\mu}_i) \right]^{\oplus} \, \mathrm{d}\boldsymbol{\mu} \\
&= \iint_A \cdots \iint_A \prod_{i=1}^{k} \left[\prod_{j=1}^{S_i} \left\{ \frac{1}{\sqrt{2\pi(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}} \exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \cdot \right. \right. \right. \\
&\quad \left. \left. \left. \left\{ \sigma_{22} \left(y_{ij;1} - \mu_{i1} \right)^2 + \sigma_{11} \left(y_{ij;2} - \mu_{i2} \right)^2 - 2\sigma_{12} \left(y_{ij;1} - \mu_{i1} \right) \left(y_{ij;2} - \mu_{i2} \right) \right\} \right] \right\} \right]^{\oplus} \, \mathrm{d}\boldsymbol{\mu} \\
&= \iint_A \cdots \iint_A \prod_{i=1}^{k} \left[\left\{ \frac{1}{\left(\sqrt{2\pi(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \right)^{S_i}} \exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \cdot \right. \right. \right. \\
&\quad \left\{ \sigma_{22} \sum_{j=1}^{S_i} \left(y_{ij;1} - \mu_{i1} \right)^2 + \sigma_{11} \sum_{j=1}^{S_i} \left(y_{ij;2} - \mu_{i2} \right)^2 - \right. \\
&\quad \left. \left. \left. 2\sigma_{12} \sum_{j=1}^{S_i} \left(y_{ij;1} - \mu_{i1} \right) \left(y_{ij;2} - \mu_{i2} \right) \right\} \right] \right\} \right]^{\oplus} \, \mathrm{d}\boldsymbol{\mu} \\
&= \prod_{i=1}^{k} \left(\frac{1}{\left(\sqrt{2\pi(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \right)^{S_i}} \iint_A \exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \cdot \right. \right. \\
&\quad \left\{ \sigma_{22} \left(\sum_{j=1}^{S_i} y_{ij;1}^2 - 2\mu_{i1} \sum_{j=1}^{S_i} y_{ij;1} + S_i \mu_{i1}^2 \right) + \right. \\
&\quad \sigma_{11} \left(\sum_{j=1}^{S_i} y_{ij;2}^2 - 2\mu_{i2} \sum_{j=1}^{S_i} y_{ij;2} + S_i \mu_{i2}^2 \right) - \\
&\quad \left. \left. 2\sigma_{12} \left(\sum_{j=1}^{S_i} y_{ij;1} y_{ij;2} - \mu_{i1} \sum_{j=1}^{S_i} y_{ij;2} - \mu_{i2} \sum_{j=1}^{S_i} y_{ij;1} + S_i \mu_{i1} \mu_{i2} \right) \right\} \right] \, \mathrm{d}\boldsymbol{\mu} \Bigg)^{\oplus} \\
&= \frac{1}{\left(\sqrt{2\pi(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \right)^{n}} \prod_{i=1}^{k} \left(\iint_A \exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \cdot \right. \right. \\
&\quad \left\{ \sigma_{22} \left(S_i \left[\mu_{i1} - \frac{1}{S_i} \sum_{j=1}^{S_i} y_{ij;1} \right]^2 + \sum_{j=1}^{S_i} y_{ij;1}^2 - \frac{1}{S_i} \left[\sum_{j=1}^{S_i} y_{ij;1} \right]^2 \right) + \right. \\
&\quad \sigma_{11} \left(S_i \left[\mu_{i2} - \frac{1}{S_i} \sum_{j=1}^{S_i} y_{ij;2} \right]^2 + \sum_{j=1}^{S_i} y_{ij;2}^2 - \frac{1}{S_i} \left[\sum_{j=1}^{S_i} y_{ij;2} \right]^2 \right) - \\
\end{aligned}
$$

$$
\begin{aligned}
& 2\sigma_{12}\left(S_i\left[\mu_{i1}-\frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;1}\right]\left[\mu_{i2}-\frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;2}\right]+\sum_{j=1}^{S_i}y_{ij;1}y_{ij;2}-\right. \\
& \left.\left.\frac{1}{S_i}\left[\sum_{j=1}^{S_i}y_{ij;1}\right]\left[\sum_{j=1}^{S_i}y_{ij;2}\right]\right)\right\}\Bigg]\ \mathrm{d}\boldsymbol{\mu}\Bigg)^{\oplus} \\
= \quad & \frac{1}{\left(\sqrt{2\pi(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\right)^n}\prod_{i=1}^{k}\left(\exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\cdot\right.\right. \\
& \left\{\sigma_{22}\left(\sum_{j=1}^{S_i}y_{ij;1}^2-\frac{1}{S_i}\left[\sum_{j=1}^{S_i}y_{ij;1}\right]^2\right)+\sigma_{11}\left(\sum_{j=1}^{S_i}y_{ij;2}^2-\frac{1}{S_i}\left[\sum_{j=1}^{S_i}y_{ij;2}\right]^2\right)-\right. \\
& \left.\left.2\sigma_{12}\left(\sum_{j=1}^{S_i}y_{ij;1}y_{ij;2}-\frac{1}{S_i}\left[\sum_{j=1}^{S_i}y_{ij;1}\right]\left[\sum_{j=1}^{S_i}y_{ij;2}\right]\right)\right\}\right]\cdot \\
& \left[\frac{2\pi}{S_i}\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\right]\iint_A\frac{1}{2\pi\sqrt{\left(\frac{\sigma_{11}}{S_i}\right)\left(\frac{\sigma_{22}}{S_i}\right)-\left(\frac{\sigma_{12}}{S_i}\right)^2}}\exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\cdot\right. \\
& \left\{\sigma_{22}\left(S_i\left[\mu_{i1}-\frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;1}\right]^2\right)+\sigma_{11}\left(S_i\left[\mu_{i2}-\frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;2}\right]^2\right)-\right. \\
& \left.\left.2\sigma_{12}\left(S_i\left[\mu_{i1}-\frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;1}\right]\left[\mu_{i2}-\frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;2}\right]\right)\right\}\right]\ \mathrm{d}\boldsymbol{\mu}\Bigg)^{\oplus} \\
= \quad & \left[\frac{1}{\left(\sqrt{2\pi(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\right)^n}\right]\exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\cdot\right. \\
& \left.\left\{\sigma_{22}\sum_{j=1}^{n}y_{j1}^2+\sigma_{11}\sum_{j=1}^{n}y_{j2}^2-2\sigma_{12}\sum_{j=1}^{n}y_{j1}y_{j2}\right\}\right]\left[\prod_{i=1}^{k}\left(\frac{2\pi}{S_i}\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\right)^{\oplus}\right]\cdot \\
& \exp\left[\frac{-1}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\sum_{i=1}^{k}\left(\frac{1}{S_i}\left\{\sigma_{22}\left[\sum_{j=1}^{S_i}y_{ij;1}\right]^2+\sigma_{11}\left[\sum_{j=1}^{S_i}y_{ij;2}\right]^2-\right.\right.\right. \\
& \left.\left.\left.2\sigma_{12}\left[\sum_{j=1}^{S_i}y_{ij;1}\right]\left[\sum_{j=1}^{S_i}y_{ij;2}\right]\right\}\right)^{\circledcirc}\right]\left[\prod_{i=1}^{k}\left(P(\mathbf{x}_i\in A)\right)^{\oplus}\right] \qquad \text{(A.4)}
\end{aligned}
$$

where

$$
\mathbf{x}_i\ \sim\ N\left(\left(\frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;1},\ \frac{1}{S_i}\sum_{j=1}^{S_i}y_{ij;2}\right)',\frac{1}{S_i}\boldsymbol{\Sigma}\right).
$$

A.4 Derivation of Asymptotic Variance for the BVNPCP(A, k, n) in Composite EM

In this section we derive expressions for

$$E_{\mathbf{Z}}\left\{ \frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k\right)}{\partial \boldsymbol{\theta}^2} \middle| \boldsymbol{\theta}, \mathbf{Y}, k \right\}$$

and

$$E_{\mathbf{Z}}\left[\left\{ \frac{\partial \mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k\right)}{\partial \boldsymbol{\theta}} \right\} \left\{ \frac{\partial \mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k\right)}{\partial \boldsymbol{\theta}} \right\}' \middle| \boldsymbol{\theta}, \mathbf{Y}, k \right]$$

for use in (3.15) for the BVNPCP(A, k, n) . The resulting variance estimate $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}^{(k)})$ can then be obtained by plugging in the EM estimates $\widehat{\boldsymbol{\theta}}^{(k)} = \{\widehat{\boldsymbol{\Sigma}}^{(k)}, \widehat{\boldsymbol{\mu}}^{(k)}\}$ for $\boldsymbol{\theta}$ and $\widehat{\mathbf{Z}}^{(k)}$ for $\{\widetilde{z}_{ji}\}$ (see Algorithm 3.3.1 and the definition of $\widetilde{z}_{ji}$ below), and inverting the matrix as given in (3.15).

The following notation is used:

- Number of components in mixture model: k
- Parameters of mixture model:

$$\boldsymbol{\theta} \;=\; \{\boldsymbol{\Sigma}, \boldsymbol{\mu}\} = \{\sigma_{11}, \sigma_{22}, \sigma_{12}, \mu_{11}, \mu_{12}, \ldots, \mu_{k1}, \mu_{k2}\}$$

- Observed data (offspring locations):

$$\mathbf{Y} \;=\; \{y_{11}, y_{12}, \ldots, y_{n1}, y_{n2}\}$$

- Latent data (allocations):

$$\mathbf{Z} \;=\; \{z_{11}, \ldots, z_{n1}, \;\ldots,\; z_{1k}, \ldots, z_{nk}\}$$

- Conditional expectations:

$$\widetilde{z}_{ji} = E\left[z_{ji} \,|\boldsymbol{\theta}, \mathbf{Y}, k \right].$$

The complete-data log-likelihood $\mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k\right)$ is given by (3.1),(3.2) and (3.6)

as:

$$
\begin{aligned}
&\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)\\
&= -n\log k - n\log(2\pi) - \frac{n}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_j-\boldsymbol{\mu}_i\right)\\
&= -n\log k - n\log(2\pi) - \frac{n}{2}\log(\sigma_{11}\sigma_{22}-\sigma_{12}^2) -\\
&\quad \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left[\frac{\sigma_{22}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}(y_{j1}-\mu_{i1})^2 + \frac{\sigma_{11}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}(y_{j2}-\mu_{i2})^2 -\right.\\
&\quad \left.\frac{2\sigma_{12}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}(y_{j1}-\mu_{i1})(y_{j2}-\mu_{i2})\right]
\end{aligned}
$$

Because of symmetry, we need not give all expressions in full form. In what follows, the shorthand notation "1" $\leftrightarrow$ "2" is used to prescribe that all occurrences of the index "1" should be replaced by "2", and vice-versa. In other words: σ_{11} becomes σ_{22}, σ_{22} becomes σ_{11}, y_{j1} becomes y_{j2}, y_{j2} becomes y_{j1}, μ_{j1} becomes μ_{j2}, and μ_{j2} becomes μ_{j1}.

First-order partial derivatives:

$$
\begin{aligned}
&\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\\
&= \frac{-n\sigma_{22}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} + \frac{\sigma_{22}^2}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j1}-\mu_{i1}\right)^2 +\\
&\quad \frac{\sigma_{12}^2}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j2}-\mu_{i2}\right)^2 -\\
&\quad \frac{\sigma_{12}\sigma_{22}}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)
\end{aligned}
$$

$$
\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}} = \frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}} \quad \text{with "1"} \leftrightarrow \text{"2"}
$$

$$\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \sigma_{12}}$$

$$= \frac{n\sigma_{12}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2} - \frac{\sigma_{22}\sigma_{12}}{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j1}-\mu_{i1}\right)^2 -$$

$$\frac{\sigma_{11}\sigma_{12}}{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j2}-\mu_{i2}\right)^2 +$$

$$\frac{\sigma_{11}\sigma_{22}+\sigma_{12}^2}{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)$$

$$\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i1}}$$

$$= \frac{\sigma_{22}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j1}-\mu_{i1}\right) - \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ji}\left(y_{j2}-\mu_{i2}\right)$$

$$\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i2}} = \frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i1}} \quad \text{with "1"} \leftrightarrow \text{"2"}$$

Expressions for $E_{\mathbf{Z}}\left\{\left.\frac{\partial^2 \mathbf{L}(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k)}{\partial \boldsymbol{\theta}^2}\right| \boldsymbol{\theta},\mathbf{Y},k\right\}$:

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \sigma_{11}\partial \sigma_{11}}\right| \boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \frac{n\sigma_{22}^2}{2\left|\boldsymbol{\Sigma}\right|^2} - \frac{\sigma_{22}^3}{\left|\boldsymbol{\Sigma}\right|^3}\sum_{i=1}^{k}\sum_{j=1}^{n} \widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2 - \frac{\sigma_{12}^2\sigma_{22}}{\left|\boldsymbol{\Sigma}\right|^3}\sum_{i=1}^{k}\sum_{j=1}^{n} \widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2 +$$

$$\frac{2\sigma_{12}\sigma_{22}^2}{\left|\boldsymbol{\Sigma}\right|^3}\sum_{i=1}^{k}\sum_{j=1}^{n} \widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \sigma_{22}\partial \sigma_{22}}\right| \boldsymbol{\theta},\mathbf{Y},k\right] = E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \sigma_{11}\partial \sigma_{11}}\right| \boldsymbol{\theta},\mathbf{Y},k\right]$$
$$\text{with "1"} \leftrightarrow \text{"2"}$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}\partial\sigma_{22}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \frac{n\sigma_{12}^2}{2|\boldsymbol{\Sigma}|^2} - \frac{\sigma_{12}^2\sigma_{22}}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2 - \frac{\sigma_{12}^2\sigma_{11}}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2 +$$
$$\frac{\sigma_{12}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}\partial\sigma_{12}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \frac{-n\sigma_{22}\sigma_{12}}{|\boldsymbol{\Sigma}|^2} + \frac{2\sigma_{22}^2\sigma_{12}}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2 +$$
$$\frac{\sigma_{12}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2 -$$
$$\frac{\sigma_{22}(\sigma_{11}\sigma_{22}+3\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}\partial\sigma_{12}}\right|\boldsymbol{\theta},\mathbf{Y},k\right] = E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}\partial\sigma_{12}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$
with "1" $\leftrightarrow$ "2"

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}\partial\sigma_{12}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \frac{n(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^2} - \frac{\sigma_{22}(\sigma_{11}\sigma_{22}+3\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2 -$$
$$\frac{\sigma_{11}(\sigma_{11}\sigma_{22}+3\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2 +$$
$$\frac{2\sigma_{12}(3\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}\partial\mu_{i1}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$
$$= \frac{-\sigma_{22}^2}{\left|\boldsymbol{\Sigma}\right|^2}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)+\frac{\sigma_{12}\sigma_{22}}{\left|\boldsymbol{\Sigma}\right|^2}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}\partial\mu_{i2}}\right|\boldsymbol{\theta},\mathbf{Y},k\right] = E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}\partial\mu_{i1}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$
with "1" $\leftrightarrow$ "2"

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}\partial\mu_{i2}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$
$$= \frac{-\sigma_{12}^2}{\left|\boldsymbol{\Sigma}\right|^2}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)+\frac{\sigma_{12}\sigma_{22}}{\left|\boldsymbol{\Sigma}\right|^2}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}\partial\mu_{i1}}\right|\boldsymbol{\theta},\mathbf{Y},k\right] = E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}\partial\mu_{i2}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$
with "1" $\leftrightarrow$ "2"

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}\partial\mu_{i1}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$
$$= \frac{2\sigma_{22}\sigma_{12}}{\left|\boldsymbol{\Sigma}\right|^2}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)-\frac{\sigma_{11}\sigma_{22}+\sigma_{12}^2}{\left|\boldsymbol{\Sigma}\right|^2}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)$$

$$E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}\partial\mu_{i2}}\right|\boldsymbol{\theta},\mathbf{Y},k\right] = E_{\mathbf{Z}}\left[\left.\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}\partial\mu_{i1}}\right|\boldsymbol{\theta},\mathbf{Y},k\right]$$
with "1" $\leftrightarrow$ "2"

$$E_{\mathbf{Z}}\left[\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}\partial\mu_{i1}}\middle|\boldsymbol{\theta},\mathbf{Y},k\right] = \frac{-\sigma_{22}}{|\boldsymbol{\Sigma}|}\sum_{j=1}^{n}\widetilde{z}_{ji}$$

$$E_{\mathbf{Z}}\left[\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i2}\partial\mu_{i2}}\middle|\boldsymbol{\theta},\mathbf{Y},k\right] = \frac{-\sigma_{11}}{|\boldsymbol{\Sigma}|}\sum_{j=1}^{n}\widetilde{z}_{ji}$$

$$E_{\mathbf{Z}}\left[\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}\partial\mu_{i2}}\middle|\boldsymbol{\theta},\mathbf{Y},k\right] = \frac{\sigma_{12}}{|\boldsymbol{\Sigma}|}\sum_{j=1}^{n}\widetilde{z}_{ji}$$

$$E_{\mathbf{Z}}\left[\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}\partial\mu_{i'1}}\middle|\boldsymbol{\theta},\mathbf{Y},k\right] = E_{\mathbf{Z}}\left[\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i2}\partial\mu_{i'2}}\middle|\boldsymbol{\theta},\mathbf{Y},k\right]$$
$$= E_{\mathbf{Z}}\left[\frac{\partial^2 \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}\partial\mu_{i'2}}\middle|\boldsymbol{\theta},\mathbf{Y},k\right] = 0 \quad \text{for } i \neq i'$$

Expressions for $E_{\mathbf{Z}}\left[\left\{\frac{\partial \mathbf{L}(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k)}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{\partial \mathbf{L}(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k)}{\partial\boldsymbol{\theta}}\right\}'\middle|\boldsymbol{\theta},\mathbf{Y},k\right]$:

At this stage we require conditional expectations of cross-products of z_{ji} for various j and i. First, since $z_{ji} \in \{0,1\}$, we have

$$E\left(z_{ji}^2|\boldsymbol{\theta},\mathbf{Y},k\right) = E\left(z_{ji}|\boldsymbol{\theta},\mathbf{Y},k\right) = \widetilde{z}_{ji}\ .$$

Allocations for different offspring are independent, so that

$$\begin{aligned} E\left(z_{ji}z_{j'i'}|\boldsymbol{\theta},\mathbf{Y},k\right) &= E\left(z_{ji}|\boldsymbol{\theta},\mathbf{Y},k\right)E\left(z_{j'i'}|\boldsymbol{\theta},\mathbf{Y},k\right) \\ &= \widetilde{z}_{ji}\widetilde{z}_{j'i'} \quad \text{for any } i,i' \text{ and } j' \neq j\ . \end{aligned}$$

For each offspring j, only one z_{ji} can be 1, so that

$$E\left(z_{ji}z_{ji'}|\boldsymbol{\theta},\mathbf{Y},k\right) = 0 \quad \text{for any } i' \neq i\ .$$

So, for example,

$$E\left[\left(a\sum_{i=1}^{k}\sum_{j=1}^{n}c_{ji}z_{ji}\right)\left(b\sum_{i=1}^{k}\sum_{j=1}^{n}d_{ji}z_{ji}\right)\middle|\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \quad ab\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}c_{ji}d_{ji}+\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}c_{ji}d_{j'i'}\right].$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)^{2}\middle|\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$\begin{aligned}
= \quad & \frac{-n\sigma_{22}^{3}}{2\left|\boldsymbol{\Sigma}\right|^{3}}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^{2}+\frac{n^{2}\sigma_{22}^{2}}{4\left|\boldsymbol{\Sigma}\right|^{2}}-\frac{n\sigma_{22}\sigma_{12}^{2}}{2\left|\boldsymbol{\Sigma}\right|^{3}}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^{2}+\\
& \frac{n\sigma_{12}\sigma_{22}^{2}}{\left|\boldsymbol{\Sigma}\right|^{3}}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)+\\
& \frac{\sigma_{22}^{4}}{4\left|\boldsymbol{\Sigma}\right|^{4}}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^{4}+\right.\\
& \qquad\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^{2}\left(y_{j'1}-\mu_{i'1}\right)^{2}\right]+\\
& \frac{\sigma_{12}^{2}\sigma_{22}^{2}}{2\left|\boldsymbol{\Sigma}\right|^{4}}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^{2}\left(y_{j2}-\mu_{i2}\right)^{2}+\right.\\
& \qquad\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^{2}\left(y_{j'2}-\mu_{i'2}\right)^{2}\right]-\\
& \frac{\sigma_{12}\sigma_{22}^{3}}{\left|\boldsymbol{\Sigma}\right|^{4}}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^{3}\left(y_{j2}-\mu_{i2}\right)+\right.\\
& \qquad\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^{2}\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]+\\
& \frac{\sigma_{12}^{4}}{4\left|\boldsymbol{\Sigma}\right|^{4}}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^{4}+\right.\\
& \qquad\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^{2}\left(y_{j'2}-\mu_{i'2}\right)^{2}\right]-
\end{aligned}$$

$$\frac{\sigma_{12}^3\sigma_{22}}{|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^3+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^2\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]+$$

$$\frac{\sigma_{12}^2\sigma_{22}^2}{|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}}\right)^2\middle|\boldsymbol{\theta},\mathbf{Y},k\right] = E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)^2\middle|\boldsymbol{\theta},\mathbf{Y},k\right]$$

with "1" $\leftrightarrow$ "2"

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}}\right)^2\middle|\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \frac{n^2\sigma_{12}^2}{|\boldsymbol{\Sigma}|^2}-\frac{2n\sigma_{12}^2\sigma_{22}}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2-\frac{2n\sigma_{11}\sigma_{12}^2}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2+$$

$$\frac{2n\sigma_{12}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)+$$

$$\frac{\sigma_{22}^2\sigma_{12}^2}{|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^4+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'1}-\mu_{i'1}\right)^2\right]+$$

$$\frac{2\sigma_{11}\sigma_{12}^2\sigma_{22}}{|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left. \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n} \widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'2}-\mu_{i'2}\right)^2 \right] -$$

$$\frac{2\sigma_{12}\sigma_{22}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{\left|\boldsymbol{\Sigma}\right|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^3\left(y_{j2}-\mu_{i2}\right)+\right.$$

$$\left. \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n} \widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right) \right] +$$

$$\frac{\sigma_{11}^2\sigma_{12}^2}{\left|\boldsymbol{\Sigma}\right|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^4+\right.$$

$$\left. \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n} \widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^2\left(y_{j'2}-\mu_{i'2}\right)^2 \right] -$$

$$\frac{2\sigma_{11}\sigma_{12}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{\left|\boldsymbol{\Sigma}\right|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^3+\right.$$

$$\left. \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n} \widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^2\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right) \right] +$$

$$\frac{(\sigma_{11}\sigma_{22}+\sigma_{12}^2)^2}{\left|\boldsymbol{\Sigma}\right|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left. \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n} \widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right) \right]$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \frac{n^2\sigma_{11}\sigma_{22}}{4\left|\boldsymbol{\Sigma}\right|^2} - \frac{n\sigma_{11}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{4\left|\boldsymbol{\Sigma}\right|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2 -$$

$$\frac{n\sigma_{22}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{4\left|\boldsymbol{\Sigma}\right|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2 +$$

$$\frac{n\sigma_{11}\sigma_{12}\sigma_{22}}{|\mathbf{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)+$$

$$\frac{\sigma_{11}^2\sigma_{22}^2+\sigma_{12}^4}{4|\mathbf{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'2}-\mu_{i'2}\right)^2\right]+$$

$$\frac{\sigma_{12}^2\sigma_{22}^2}{4|\mathbf{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^4+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'1}-\mu_{i'1}\right)^2\right]-$$

$$\frac{\sigma_{12}\sigma_{22}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{2|\mathbf{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^3\left(y_{j2}-\mu_{i2}\right)+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]+$$

$$\frac{\sigma_{11}^2\sigma_{12}^2}{4|\mathbf{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^4+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^2\left(y_{j'2}-\mu_{i'2}\right)^2\right]-$$

$$\frac{\sigma_{11}\sigma_{12}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{2|\mathbf{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^3+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^2\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]+$$

$$\frac{\sigma_{11}\sigma_{12}^2\sigma_{22}}{|\mathbf{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left. \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$\begin{aligned}
= \quad & \frac{-n^2\sigma_{12}\sigma_{22}}{2|\boldsymbol{\Sigma}|^2}+\frac{n\sigma_{12}\sigma_{22}^2}{|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2+ \\
& \frac{n\sigma_{12}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{2|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2- \\
& \frac{n\sigma_{22}(\sigma_{11}\sigma_{22}+3\sigma_{12}^2)}{2|\boldsymbol{\Sigma}|^3}\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)- \\
& \frac{\sigma_{12}\sigma_{22}^3}{2|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^4+\right. \\
& \left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'1}-\mu_{i'1}\right)^2\right]- \\
& \frac{\sigma_{12}\sigma_{22}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{2|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)^2+\right. \\
& \left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'2}-\mu_{i'2}\right)^2\right]+ \\
& \frac{\sigma_{22}^2(\sigma_{11}\sigma_{22}+3\sigma_{12}^2)}{2|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^3\left(y_{j2}-\mu_{i2}\right)+\right. \\
& \left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]- \\
& \frac{\sigma_{11}\sigma_{12}^3}{2|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^4+\right.
\end{aligned}$$

$$\left. \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n} \widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^2\left(y_{j'2}-\mu_{i'2}\right)^2 \right] +$$

$$\frac{\sigma_{12}^2(3\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{2|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^3+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j2}-\mu_{i2}\right)^2\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right] -$$

$$\frac{\sigma_{12}\sigma_{22}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^4}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left.\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{i'=1}^{k}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i'}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \quad E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \quad \text{with "1"} \leftrightarrow \text{"2"}$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \quad \frac{-n\sigma_{22}^2}{2|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)+\frac{\sigma_{22}^3}{2|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^3+\right.$$

$$\left.\sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i'1}\right)^2\right] +$$

$$\frac{\sigma_{12}^2\sigma_{22}}{2|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left.\sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'2}-\mu_{i'2}\right)^2\right] -$$

$$\frac{\sigma_{12}\sigma_{22}^2}{|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)+\right.$$

$$\left.\sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]+$$

$$\frac{n\sigma_{12}\sigma_{22}}{2|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)-\frac{\sigma_{12}\sigma_{22}^2}{2|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)+\right.$$

$$\left.\sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)^2\right]-$$

$$\frac{\sigma_{12}^3}{2|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^3+\right.$$

$$\left.\sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'2}-\mu_{i'2}\right)^2\right]+$$

$$\frac{\sigma_{12}^2\sigma_{22}}{|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left.\sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{22}}\right)\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i2}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \quad E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \quad \text{with "1"} \leftrightarrow \text{"2"}$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{11}}\right)\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i2}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \quad \frac{-n\sigma_{11}\sigma_{22}}{2|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)+\frac{\sigma_{11}\sigma_{22}^2}{2|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right)+\right.$$

$$\left. \sum_{i'=1}^{k} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \widetilde{z}_{ji} \widetilde{z}_{j'i} \left(y_{j2} - \mu_{i2}\right) \left(y_{j'1} - \mu_{i'1}\right)^2 \right] +$$

$$\frac{\sigma_{11}\sigma_{12}^2}{2|\mathbf{\Sigma}|^3} \left[\sum_{j=1}^{n} \widetilde{z}_{ji} \left(y_{j2} - \mu_{i2}\right)^3 + \right.$$

$$\left. \sum_{i'=1}^{k} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \widetilde{z}_{ji} \widetilde{z}_{j'i} \left(y_{j2} - \mu_{i2}\right) \left(y_{j'2} - \mu_{i'2}\right)^2 \right] -$$

$$\frac{\sigma_{11}\sigma_{12}\sigma_{22}}{|\mathbf{\Sigma}|^3} \left[\sum_{j=1}^{n} \widetilde{z}_{ji} \left(y_{j1} - \mu_{i1}\right) \left(y_{j2} - \mu_{i2}\right)^2 + \right.$$

$$\left. \sum_{i'=1}^{k} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \widetilde{z}_{ji} \widetilde{z}_{j'i} \left(y_{j2} - \mu_{i2}\right) \left(y_{j'1} - \mu_{i'1}\right) \left(y_{j'2} - \mu_{i'2}\right) \right] +$$

$$\frac{n\sigma_{12}\sigma_{22}}{2|\mathbf{\Sigma}|^2} \sum_{j=1}^{n} \widetilde{z}_{ji} \left(y_{j1} - \mu_{i1}\right) - \frac{\sigma_{12}\sigma_{22}^2}{2|\mathbf{\Sigma}|^3} \left[\sum_{j=1}^{n} \widetilde{z}_{ji} \left(y_{j1} - \mu_{i1}\right)^3 + \right.$$

$$\left. \sum_{i'=1}^{k} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \widetilde{z}_{ji} \widetilde{z}_{j'i} \left(y_{j1} - \mu_{i1}\right) \left(y_{j'1} - \mu_{i'1}\right)^2 \right] -$$

$$\frac{\sigma_{12}^3}{2|\mathbf{\Sigma}|^3} \left[\sum_{j=1}^{n} \widetilde{z}_{ji} \left(y_{j1} - \mu_{i1}\right) \left(y_{j2} - \mu_{i2}\right)^2 + \right.$$

$$\left. \sum_{i'=1}^{k} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \widetilde{z}_{ji} \widetilde{z}_{j'i} \left(y_{j1} - \mu_{i1}\right) \left(y_{j'2} - \mu_{i'2}\right)^2 \right] +$$

$$\frac{\sigma_{12}^2\sigma_{22}}{|\mathbf{\Sigma}|^3} \left[\sum_{j=1}^{n} \widetilde{z}_{ji} \left(y_{j1} - \mu_{i1}\right)^2 \left(y_{j2} - \mu_{i2}\right) + \right.$$

$$\left. \sum_{i'=1}^{k} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \widetilde{z}_{ji} \widetilde{z}_{j'i} \left(y_{j1} - \mu_{i1}\right) \left(y_{j'1} - \mu_{i'1}\right) \left(y_{j'2} - \mu_{i'2}\right) \right]$$

$$E_{\mathbf{Z}} \left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k\right)}{\partial \sigma_{22}} \right) \left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}, k\right)}{\partial \mu_{i1}} \right) \middle| \boldsymbol{\theta}, \mathbf{Y}, k \right]$$

$$= \quad E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \sigma_{11}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i2}}\right)\middle|\boldsymbol{\theta},\mathbf{Y},k\right] \quad \text{with "1"} \leftrightarrow \text{"2"}$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \sigma_{12}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i1}}\right)\middle|\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$\begin{aligned}
= \quad & \frac{n\sigma_{12}\sigma_{22}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right) - \frac{\sigma_{12}\sigma_{22}^2}{|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^3 + \right. \\
& \left. \sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i'1}\right)^2\right] - \\
& \frac{\sigma_{11}\sigma_{12}\sigma_{22}}{|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^2 + \right. \\
& \left. \sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'2}-\mu_{i'2}\right)^2\right] + \\
& \frac{\sigma_{22}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right) + \right. \\
& \left. \sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right] - \\
& \frac{n\sigma_{12}^2}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right) + \frac{\sigma_{12}^2\sigma_{22}}{|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2\left(y_{j2}-\mu_{i2}\right) + \right. \\
& \left. \sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)^2\right] + \\
& \frac{\sigma_{11}\sigma_{12}^2}{|\boldsymbol{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^3 + \right. \\
& \left. \sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'2}-\mu_{i'2}\right)^2\right] -
\end{aligned}$$

$$\frac{\sigma_{12}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)}{|\mathbf{\Sigma}|^3}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)^2+\right.$$

$$\left.\sum_{i'=1}^{k}\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)\left(y_{j'2}-\mu_{i'2}\right)\right]$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}}\right)\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i2}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \quad E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\sigma_{12}}\right)\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \quad \text{with "1"} \leftrightarrow \text{"2"}$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}}\right)^2\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$= \quad \frac{\sigma_{22}^2}{|\mathbf{\Sigma}|^2}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)^2+\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i1}\right)\right]-$$

$$\frac{2\sigma_{12}\sigma_{22}}{|\mathbf{\Sigma}|^2}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j1}-\mu_{i1}\right)\left(y_{j2}-\mu_{i2}\right)+\right.$$

$$\left.\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'2}-\mu_{i2}\right)\right]+$$

$$\frac{\sigma_{12}^2}{|\mathbf{\Sigma}|^2}\left[\sum_{j=1}^{n}\widetilde{z}_{ji}\left(y_{j2}-\mu_{i2}\right)^2+\sum_{j=1}^{n}\sum_{\substack{j'=1\\j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'2}-\mu_{i2}\right)\right]$$

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i2}}\right)^2\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \quad = \quad E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}}\right)^2\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

with "1" $\leftrightarrow$ "2"

$$E_{\mathbf{Z}}\left[\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i1}}\right)\left(\frac{\partial\mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial\mu_{i2}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right]$$

$$
\begin{aligned}
= \quad & \frac{\sigma_{11}\sigma_{22}+\sigma_{12}^2}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'2}-\mu_{i2}\right)- \\
& \frac{\sigma_{12}\sigma_{22}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i1}\right)- \\
& \frac{\sigma_{11}\sigma_{12}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'2}-\mu_{i2}\right)
\end{aligned}
$$

$$
\begin{aligned}
& E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i1}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i'1}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \\
= \quad & \frac{\sigma_{22}^2}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i'1}\right)- \\
& \frac{\sigma_{12}\sigma_{22}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'2}-\mu_{i'2}\right)- \\
& \frac{\sigma_{12}\sigma_{22}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)+ \\
& \frac{\sigma_{12}^2}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'2}-\mu_{i'2}\right)
\end{aligned}
$$

$$
\begin{aligned}
& E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i2}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i'2}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \\
= \quad & E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i1}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i'1}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \quad \text{with ``1'' } \leftrightarrow \text{ ``2''}
\end{aligned}
$$

$$
\begin{aligned}
& E_{\mathbf{Z}}\left[\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i1}}\right)\left(\frac{\partial \mathbf{L}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{Z},k\right)}{\partial \mu_{i'2}}\right)\middle|\,\boldsymbol{\theta},\mathbf{Y},k\right] \\
= \quad & \frac{\sigma_{11}\sigma_{22}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'2}-\mu_{i'2}\right)-
\end{aligned}
$$

$$\frac{\sigma_{12}\sigma_{22}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j1}-\mu_{i1}\right)\left(y_{j'1}-\mu_{i'1}\right)-$$

$$\frac{\sigma_{11}\sigma_{12}}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'2}-\mu_{i'2}\right)+$$

$$\frac{{\sigma_{12}}^2}{|\boldsymbol{\Sigma}|^2}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\widetilde{z}_{ji}\widetilde{z}_{j'i}\left(y_{j2}-\mu_{i2}\right)\left(y_{j'1}-\mu_{i'1}\right)$$

A.5 Jacobians for Confidence Intervals / Regions in Composite EM

In this section we give expressions for the Jacobian ((3.29), before estimates $\widehat{\boldsymbol{\sigma}}$ are plugged in):

$$J=\left\{\frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}}\right\}=\left[\frac{\partial f(\boldsymbol{\sigma})}{\partial \sigma_{11}},\frac{\partial f(\boldsymbol{\sigma})}{\partial \sigma_{22}},\frac{\partial f(\boldsymbol{\sigma})}{\partial \sigma_{12}}\right]$$

for use in (3.28) for the BVNPCP(A,k,n) , for the following choices of $f(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma}=(\sigma_{11},\sigma_{22},\sigma_{12})'$:

1. $\boldsymbol{\sigma}^c=(\log\sigma_{11}-\log\sigma_{22},\ z(\rho_{12}))'$ (see (3.32))
2. $\log\sigma_{11}$ (see Definition 1.1.9)
3. $\log\sigma_{22}$ (see Definition 1.1.9)
4. $z(\rho_{12})$ (see Definition 3.5.1)
5. $\log\gamma$ (see Definition 1.1.10 and Fact 1.1.11)
6. $\log\Psi$ (see Definition 1.1.10 and Fact 1.1.11)
7. ϕ (see Definition 1.1.10 and Fact 1.1.11).

$$\frac{\partial \boldsymbol{\sigma}^c}{\partial \boldsymbol{\sigma}} = \begin{bmatrix} \frac{1}{\sigma_{11}} & \frac{-1}{\sigma_{22}} & 0 \\ \frac{-\sigma_{12}\sqrt{\sigma_{22}}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\sqrt{\sigma_{11}}} & \frac{-\sigma_{12}\sqrt{\sigma_{11}}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\sqrt{\sigma_{22}}} & \frac{\sqrt{\sigma_{11}\sigma_{22}}}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} \end{bmatrix}$$

$$\frac{\partial \log \sigma_{11}}{\partial \boldsymbol{\sigma}} = \left[\frac{1}{\sigma_{11}},\ 0,\ 0\right]$$

$$\frac{\partial \log \sigma_{22}}{\partial \boldsymbol{\sigma}} = \left[0,\ \frac{1}{\sigma_{22}},\ 0\right]$$

$$\frac{\partial z(\rho_{12})}{\partial \boldsymbol{\sigma}} = \frac{1}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\left[-\sigma_{12}\left(\frac{\sigma_{22}}{\sigma_{11}}\right)^{\frac{1}{2}},\ -\sigma_{12}\left(\frac{\sigma_{11}}{\sigma_{22}}\right)^{\frac{1}{2}},\ 2\left(\sigma_{11}\sigma_{22}\right)^{\frac{1}{2}}\right]$$

$$\frac{\partial \log \gamma}{\partial \boldsymbol{\sigma}} = \frac{1}{2\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)\left[\left(\sigma_{11}-\sigma_{22}\right)^2+4\sigma_{12}^2\right]^{\frac{1}{2}}}\begin{bmatrix} \sigma_{22}(\sigma_{11}-\sigma_{22})-2\sigma_{12}^2 \\ \sigma_{11}(\sigma_{22}-\sigma_{11})-2\sigma_{12}^2 \\ 2\sigma_{12}(\sigma_{11}+\sigma_{22}) \end{bmatrix}'$$

$$\frac{\partial \log \Psi}{\partial \boldsymbol{\sigma}} = \frac{1}{2\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)}\left[\sigma_{22},\ \sigma_{11},\ -2\sigma_{12}\right]$$

$$\frac{\partial \phi}{\partial \boldsymbol{\sigma}} = \frac{1}{(\sigma_{22}-\sigma_{11})^2-(\sigma_{22}-\sigma_{11})\left[\left(\sigma_{22}-\sigma_{11}\right)^2+4\sigma_{12}^2\right]^{\frac{1}{2}}+4\sigma_{12}^2} \cdot \begin{bmatrix} -\sigma_{12}\left\{1+(\sigma_{11}-\sigma_{22})\left[\left(\sigma_{11}-\sigma_{22}\right)^2+4\sigma_{12}^2\right]^{-\frac{1}{2}}\right\} \\ \sigma_{12}\left\{1+(\sigma_{11}-\sigma_{22})\left[\left(\sigma_{11}-\sigma_{22}\right)^2+4\sigma_{12}^2\right]^{-\frac{1}{2}}\right\} \\ (\sigma_{11}-\sigma_{22})+\left[\left(\sigma_{11}-\sigma_{22}\right)^2+4\sigma_{12}^2\right]^{\frac{1}{2}}-4\sigma_{12}^2\left[\left(\sigma_{11}-\sigma_{22}\right)^2+4\sigma_{12}^2\right]^{-\frac{1}{2}} \end{bmatrix}'$$

A.6 Derivation of Expectations of Variation Estimates Used in Convergence Assessment

First, we (trivially) have (5.24) by Winer (1971, p. 163) and (5.29) by Winer (1971, p. 325).

Proof of (5.23): Consider ANOVA 1 in Table 5.1. From Winer (1971, p. 165), we have that

$$E\mathrm{MS}_{\mathrm{ch}} = T\sigma^2_{\mathrm{ch}} + \sigma^2_{\mathrm{er(ch)}}$$

and so

$$\begin{aligned} E\widehat{V} &= E\mathrm{MS}_{\mathrm{tot}} \\ &= \frac{1}{CT-1} E\mathrm{SS}_{\mathrm{tot}} \\ &= \frac{1}{CT-1} E\left[\mathrm{SS}_{\mathrm{ch}} + \mathrm{SS}_{\mathrm{er(ch)}}\right] \\ &= \frac{1}{CT-1}\left[(C-1)E\mathrm{MS}_{\mathrm{ch}} + C(T-1)E\mathrm{MS}_{\mathrm{er(ch)}}\right] \\ &= \sigma^2_{\mathrm{er(ch)}} + \frac{(C-1)T}{CT-1}\sigma^2_{\mathrm{ch}} \quad \square \end{aligned}$$

Proof of (5.25): Consider ANOVA 2 and ANOVA 3 in Tables 5.2 – 5.3. The quantity $\mathrm{SS}_{\mathrm{er(mo)}}$ in ANOVA 2 can be re-written as follows:

$$\begin{aligned} \mathrm{SS}_{\mathrm{er(mo)}} &= \sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}} \left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{\cdot m}\right)^2 \\ &= \sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}} \left[\left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{cm}\right) + \left(\bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right) + \left(\bar{\theta}^{\cdot}_{cm} - \bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot m} + \bar{\theta}^{\cdot}_{\cdot\cdot}\right)\right]^2 \\ &= \sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}} \left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{cm}\right)^2 + \sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}} \left(\bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2 + \\ &\quad \sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}} \left(\bar{\theta}^{\cdot}_{cm} - \bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot m} + \bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2 + \\ &\quad 2\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}} \left(\theta^r_{cm} - \bar{\theta}^{\cdot}_{cm}\right)\left(\bar{\theta}^{\cdot}_{c\cdot} - \bar{\theta}^{\cdot}_{\cdot\cdot}\right) + \end{aligned}$$

$$2\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^{r}_{cm}-\bar{\theta}^{\cdot}_{cm}\right)\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right)+$$
$$2\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right).$$

Each term can then be simplified:

$$\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^{r}_{cm}-\bar{\theta}^{\cdot}_{cm}\right)^{2}=\mathrm{SS}_{\mathrm{er(ch\times mo)}},$$
$$\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)^{2}=\mathrm{SS}_{\mathrm{ch}},$$
$$\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right)^{2}=\mathrm{SS}_{\mathrm{ch\times mo}},$$
$$\begin{aligned}
&2\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^{r}_{cm}-\bar{\theta}^{\cdot}_{cm}\right)\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\\
&\quad=\quad 2\sum_{c=1}^{C}\sum_{m=1}^{M}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\left[\sum_{r=1}^{R_{cm}}\left(\theta^{r}_{cm}-\bar{\theta}^{\cdot}_{cm}\right)\right]\\
&\quad=\quad 2\sum_{c=1}^{C}\sum_{m=1}^{M}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)[0]\\
&\quad=\quad 0,
\end{aligned}$$
$$\begin{aligned}
&2\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\theta^{r}_{cm}-\bar{\theta}^{\cdot}_{cm}\right)\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\\
&\quad=\quad 2\sum_{c=1}^{C}\sum_{m=1}^{M}\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\left[\sum_{r=1}^{R_{cm}}\left(\theta^{r}_{cm}-\bar{\theta}^{\cdot}_{cm}\right)\right]\\
&\quad=\quad 2\sum_{c=1}^{C}\sum_{m=1}^{M}\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right)[0]\\
&\quad=\quad 0,
\end{aligned}$$

and finally, using the notation

$$\theta^{\cdot}_{cm}=\sum_{r=1}^{R_{cm}}\theta^{r}_{cm},\ \theta^{\cdot}_{c\cdot}=\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\theta^{r}_{cm},\ \theta^{\cdot}_{\cdot m}=\sum_{c=1}^{C}\sum_{r=1}^{R_{cm}}\theta^{r}_{cm},\ \text{ and }\theta^{\cdot}_{\cdot\cdot}=\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\theta^{r}_{cm},$$
$$2\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right)$$

$$
\begin{aligned}
&= 2\sum_{c=1}^{C}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\left[\sum_{m=1}^{M}\left\{R_{cm}\left(\bar{\theta}^{\cdot}_{cm}-\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot m}+\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\right\}\right]\\
&= 2\sum_{c=1}^{C}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\left[\sum_{m=1}^{M}\left(\theta^{\cdot}_{cm}-\frac{1}{M}\theta^{\cdot}_{c\cdot}-\frac{1}{C}\theta^{\cdot}_{\cdot m}+\frac{1}{CM}\theta^{\cdot}_{\cdot\cdot}\right)\right]\\
&= 2\sum_{c=1}^{C}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)\left[\theta^{\cdot}_{c\cdot}-\theta^{\cdot}_{c\cdot}-\frac{1}{C}\theta^{\cdot}_{\cdot\cdot}+\frac{1}{C}\theta^{\cdot}_{\cdot\cdot}\right]\\
&= 0.
\end{aligned}
$$

Therefore,

$$\mathrm{SS}_{\mathrm{er(mo)}} = \mathrm{SS}_{\mathrm{er(ch\times mo)}} + \mathrm{SS}_{\mathrm{ch}} + \mathrm{SS}_{\mathrm{ch\times mo}}.$$

Now we derive the expected mean-squares for chain and chain×model, using terminology from ANOVA 3 (Table 5.3). Several steps in the derivations use the following ANOVA assumptions (and not always with explicit reference):

$$\begin{aligned}
&\sum_{m=1}^{M}\beta_m = 0,\ E\alpha_c = 0,\ E(\alpha\beta)_{cm} = 0,\ Ee^{r}_{cm} = 0,\ \text{ and}\\
&\text{all } \{\alpha_c\}, \{(\alpha\beta)_{cm}\}, \{e^{r}_{cm}\} \text{ are mutually independent.}
\end{aligned} \tag{A.5}$$

For simplicity of notation, let $e^{r}_{cm} \equiv e^{r}_{cm(3)}$.

First, $E\mathrm{MS}_{\mathrm{ch}}$:

$$
\begin{aligned}
E\mathrm{MS}_{\mathrm{ch}} &= \frac{1}{C-1}E\left\{T\sum_{c=1}^{C}\left(\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right)^2\right\}\\
&= \frac{1}{C-1}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left\{\bar{\theta}^{\cdot}_{c\cdot}-\bar{\theta}^{\cdot}_{\cdot\cdot}\right\}^2\\
&= \frac{1}{C-1}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left\{\frac{1}{T}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}\left(\mu+\alpha_c+\beta_{m'}+(\alpha\beta)_{cm'}+e^{r'}_{cm'}\right)-\right.\\
&\quad \left.\frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}\left(\mu+\alpha_{c'}+\beta_{m'}+(\alpha\beta)_{c'm'}+e^{r'}_{c'm'}\right)\right\}^2\\
&= \frac{1}{C-1}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left\{\left(\mu+\alpha_c+\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}\beta_{m'}+\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}(\alpha\beta)_{cm'}+\right.\right.\\
&\quad \left.\left.\frac{1}{T}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}e^{r'}_{cm'}\right)-
\end{aligned}
$$

$$\left(\mu + \frac{1}{C}\sum_{c'=1}^{C}\alpha_{c'} + \frac{1}{CT}\sum_{m'=1}^{M}R_{\cdot m'}\beta_{m'} + \frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}(\alpha\beta)_{c'm'} + \frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}e^{r'}_{c'm'}\right)\Bigg\}^2$$

$$= \frac{1}{C-1}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\Bigg\{\left(\alpha_c - \frac{1}{C}\sum_{c'=1}^{C}\alpha_{c'}\right) + \left(\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}\beta_{m'} - \frac{1}{CT}\sum_{m'=1}^{M}R_{\cdot m'}\beta_{m'}\right) + \left(\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}(\alpha\beta)_{cm'} - \frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}(\alpha\beta)_{c'm'}\right) + \left(\frac{1}{T}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}e^{r'}_{cm'} - \frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}e^{r'}_{c'm'}\right)\Bigg\}^2$$

$$= \frac{1}{C-1}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\Bigg\{\left(\alpha_c - \frac{1}{C}\sum_{c'=1}^{C}\alpha_{c'}\right)^2 + \left(\frac{1}{CT}\sum_{m'=1}^{M}\left[CR_{cm'} - R_{\cdot m'}\right]\beta_{m'}\right)^2 + \left(\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}(\alpha\beta)_{cm'} - \frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}(\alpha\beta)_{c'm'}\right)^2 + \left(\frac{1}{T}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}e^{r'}_{cm'} - \frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}e^{r'}_{c'm'}\right)^2\Bigg\}$$

(all sums of cross-products are 0, by (A.5))

$$= \frac{1}{C-1}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\Big\{(\alpha_c - \bar{\alpha}_\cdot)^2 + \left(\frac{1}{C^2T^2}\sum_{m'=1}^{M}\left[CR_{cm'} - R_{\cdot m'}\right]^2\beta^2_{m'}\right) + \left(\frac{1}{T^2}\sum_{m'=1}^{M}R^2_{cm'}(\alpha\beta)^2_{cm'} - \frac{2}{CT^2}\sum_{m'=1}^{M}R^2_{cm'}(\alpha\beta)^2_{cm'} +$$

$$
\begin{aligned}
& \frac{1}{C^2T^2}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}^2(\alpha\beta)_{c'm'}^2\Bigg) \;+\; \left(\bar{e}_{c\cdot}^{\cdot}-\bar{e}_{\cdot\cdot}^{\cdot}\right)^2\Bigg\} \\
= \;& \frac{1}{C-1}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\Bigg\{\left(\alpha_c-\bar{\alpha}_{\cdot}\right)^2+\frac{1}{C^2T^2}\sum_{m'=1}^{M}\left[CR_{cm'}-R_{\cdot m'}\right]^2\beta_{m'}^2 \;+ \\
& \frac{1}{CT^2}\sum_{m'=1}^{M}\left[CR_{cm'}^2-2R_{cm'}^2\right](\alpha\beta)_{cm'}^2 \;+ \\
& \frac{1}{C^2T^2}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}^2(\alpha\beta)_{c'm'}^2 \;+\; \left(\bar{e}_{c\cdot}^{\cdot}-\bar{e}_{\cdot\cdot}^{\cdot}\right)^2\Bigg\} \\
= \;& \frac{1}{C-1}\Bigg\{T(C-1)\sigma_{\text{ch}}^2+\frac{T}{C^2T^2}\sum_{m'=1}^{M}\left[\sum_{c=1}^{C}\left(CR_{cm'}-R_{\cdot m'}\right)^2\right]\beta_{m'}^2 \;+ \\
& T\frac{C-2}{CT^2}\sum_{m'=1}^{M}\sum_{c=1}^{C}R_{cm'}^2E(\alpha\beta)_{cm'}^2+\frac{1}{CT}\sum_{m'=1}^{M}\sum_{c'=1}^{C}R_{c'm'}^2E(\alpha\beta)_{c'm'}^2 \;+ \\
& T(C-1)\frac{\sigma_{\text{er(ch}\times\text{mo)}}^2}{T}\Bigg\} \\
= \;& T\sigma_{\text{ch}}^2+\frac{1}{(C-1)C^2T}\sum_{m=1}^{M}\left[\sum_{c=1}^{C}\left(CR_{cm}-R_{\cdot m}\right)^2\right]\beta_m^2 \;+ \\
& \left[\frac{1}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\right]\sigma_{\text{ch}\times\text{mo}}^2 \;+\; \sigma_{\text{er(ch}\times\text{mo)}}^2 \quad .
\end{aligned}
$$

Next, $E\text{MS}_{\text{ch}\times\text{mo}}$:

$$
\begin{aligned}
E\text{MS}_{\text{ch}\times\text{mo}} &= \frac{1}{(C-1)(M-1)}E\left\{\sum_{c=1}^{C}\sum R_{cm}\left(\bar{\theta}_{cm}^{\cdot}-\bar{\theta}_{c\cdot}^{\cdot}-\bar{\theta}_{\cdot m}^{\cdot}+\bar{\theta}_{\cdot\cdot}^{\cdot}\right)^2\right\} \\
= \;& \frac{1}{(C-1)(M-1)}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left\{\bar{\theta}_{cm}^{\cdot}-\bar{\theta}_{c\cdot}^{\cdot}-\bar{\theta}_{\cdot m}^{\cdot}+\bar{\theta}_{\cdot\cdot}^{\cdot}\right\}^2 \\
= \;& \frac{1}{(C-1)(M-1)}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\Bigg\{\frac{1}{R_{cm}}\sum_{r'=1}^{R_{cm}}\left(\mu+\alpha_c+\beta_m+(\alpha\beta)_{cm}+e_{cm}^{r'}\right)- \\
& \frac{1}{T}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}\left(\mu+\alpha_c+\beta_{m'}+(\alpha\beta)_{cm'}+e_{cm'}^{r'}\right)- \\
& \frac{1}{R_{\cdot m}}\sum_{c'=1}^{C}\sum_{r'=1}^{R_{cm}}\left(\mu+\alpha_{c'}+\beta_m+(\alpha\beta)_{c'm}+e_{c'm}^{r'}\right)+
\end{aligned}
$$

$$\frac{1}{IT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}\sum_{r'=1}^{R_{cm}}\left(\mu+\alpha_{c'}+\beta_{m'}+(\alpha\beta)_{c'm'}+e^{r'}_{c'm'}\right)\Bigg\}^2$$

$$= \frac{1}{(C-1)(M-1)}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\{(\mu+\alpha_c+\beta_m+(\alpha\beta)_{cm}+\bar{e}_{cm}^{\cdot})-$$

$$\left(\mu+\alpha_c+\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}\beta_{m'}+\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}(\alpha\beta)_{cm'}+\bar{e}_{c\cdot}^{\cdot}\right)-$$

$$\left(\mu+\frac{1}{R_{\cdot m}}\sum_{c'=1}^{C}R_{c'm}\alpha_{c'}+\beta_m+\frac{1}{R_{\cdot m}}\sum_{c'=1}^{C}R_{c'm}(\alpha\beta)_{c'm}+\bar{e}_{\cdot m}^{\cdot}\right)+$$

$$\left(\mu+\frac{1}{C}\sum_{c'=1}^{C}\alpha_{c'}+\frac{1}{CT}\sum_{m'=1}^{M}R_{\cdot m'}\beta_{m'}+\frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}(\alpha\beta)_{c'm'}+\bar{e}_{\cdot\cdot}^{\cdot}\right)\Bigg\}^2$$

$$= \frac{1}{(C-1)(M-1)}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\Bigg\{\left(\frac{1}{C}\sum_{c'=1}^{C}\alpha_{c'}-\frac{1}{R_{\cdot m}}\sum_{c'=1}^{C}R_{c'm}\alpha_{c'}\right)+$$

$$\left(\frac{1}{CT}\sum_{m'=1}^{M}R_{\cdot m'}\beta_{m'}-\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}\beta_{m'}\right)+$$

$$\left((\alpha\beta)_{cm}-\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}(\alpha\beta)_{cm'}-\frac{1}{R_{\cdot m}}\sum_{c'=1}^{C}R_{c'm}(\alpha\beta)_{c'm}+\right.$$

$$\left.\frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}(\alpha\beta)_{c'm'}\right)+(\bar{e}_{cm}^{\cdot}-\bar{e}_{c\cdot}^{\cdot}-\bar{e}_{\cdot m}^{\cdot}+\bar{e}_{\cdot\cdot}^{\cdot})\Bigg\}^2$$

$$= \frac{1}{(C-1)(M-1)}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\Bigg\{\left(\frac{1}{R_{\cdot m}}\sum_{c'=1}^{C}R_{c'm}\alpha_{c'}-\frac{1}{C}\sum_{c'=1}^{C}\alpha_{c'}\right)^2+$$

$$\left(\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}\beta_{m'}-\frac{1}{CT}\sum_{m'=1}^{M}R_{\cdot m'}\beta_{m'}\right)^2+$$

$$\left((\alpha\beta)_{cm}-\frac{1}{T}\sum_{m'=1}^{M}R_{cm'}(\alpha\beta)_{cm'}-\frac{1}{R_{\cdot m}}\sum_{c'=1}^{C}R_{c'm}(\alpha\beta)_{c'm}+\right.$$

$$\left.\frac{1}{CT}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}(\alpha\beta)_{c'm'}\right)^2+(\bar{e}_{cm}^{\cdot}-\bar{e}_{c\cdot}^{\cdot}-\bar{e}_{\cdot m}^{\cdot}+\bar{e}_{\cdot\cdot}^{\cdot})^2\Bigg\}$$

(all sums of cross-products are 0, by (A.5))

At this point we simplify the last portion:

$$
\begin{aligned}
& E\left(\bar{\bar{e}}_{cm} - \bar{\bar{e}}_{c\cdot} - \bar{\bar{e}}_{\cdot m} + \bar{\bar{e}}_{\cdot\cdot}\right)^2 \\
&= E\left[\bar{\bar{e}}_{cm} - \frac{1}{M}\sum_{m'=1}^{M}\bar{\bar{e}}_{cm'} - \frac{1}{C}\sum_{c'=1}^{C}\bar{\bar{e}}_{c'm} + \frac{1}{CM}\sum_{c'=1}^{C}\sum_{m'=1}^{M}\bar{\bar{e}}_{c'm'}\right]^2 \\
&= E\left[(\bar{\bar{e}}_{cm})^2 + \frac{1}{M^2}\sum_{m'=1}^{M}(\bar{\bar{e}}_{cm'})^2 + \frac{1}{C^2}\sum_{c'=1}^{C}(\bar{\bar{e}}_{c'm})^2 + \frac{1}{C^2M^2}\sum_{c'=1}^{C}\sum_{m'=1}^{M}(\bar{\bar{e}}_{c'm'})^2 - \right. \\
&\quad \frac{2}{M}(\bar{\bar{e}}_{cm})^2 - \frac{2}{C}(\bar{\bar{e}}_{cm})^2 + \frac{2}{CM}(\bar{\bar{e}}_{cm})^2 + \frac{2}{CM}(\bar{\bar{e}}_{cm})^2 - \frac{2}{CM^2}\sum_{m'=1}^{M}(\bar{\bar{e}}_{cm'})^2 - \\
&\quad \left.\frac{2}{C^2M}\sum_{c'=1}^{C}(\bar{\bar{e}}_{c'm})^2\right] \\
&= E\left[(\bar{\bar{e}}_{cm})^2 + \frac{1}{M}(\bar{\bar{e}}_{cm})^2 + \frac{1}{C}\bar{\bar{e}}_{cm})^2 + \frac{1}{CM}(\bar{\bar{e}}_{cm})^2 - \right. \\
&\quad \frac{2}{M}(\bar{\bar{e}}_{cm})^2 - \frac{2}{C}(\bar{\bar{e}}_{cm})^2 + \frac{2}{CM}(\bar{\bar{e}}_{cm})^2 + \frac{2}{CM}(\bar{\bar{e}}_{cm})^2 - \frac{2}{CM}(\bar{\bar{e}}_{cm})^2 - \\
&\quad \left.\frac{2}{CM}(\bar{\bar{e}}_{cm})^2\right] \\
&= \frac{1}{C^2M^2}\left[(CM)^2 + C^2M + CM^2 + CM - 2C^2M - 2CM^2 + \right. \\
&\quad \left. 2CM + 2CM - 2CM - 2CM\right] E(\bar{\bar{e}}_{cm})^2 \\
&= \frac{1}{C^2M^2}\left[(CM)^2 - C^2M - CM^2 + CM\right]\frac{\sigma^2_{\text{er(ch}\times\text{mo)}}}{R_{cm}} \\
&= \frac{(C-1)(M-1)}{R_{cm}CM}\sigma^2_{\text{er(ch}\times\text{mo)}}
\end{aligned}
$$

Now, resuming again with $E\text{MS}_{\text{ch}\times\text{mo}}$:

$$
\begin{aligned}
& E\text{MS}_{\text{ch}\times\text{mo}} \\
&= \frac{1}{(C-1)(M-1)}E\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\left\{\left(\frac{1}{R^2_{\cdot m}C^2}\sum_{c'=1}^{C}\left(CR_{c'm} - R_{\cdot m}\right)^2\alpha^2_{c'}\right) + \right. \\
&\quad \left(\frac{1}{C^2T^2}\sum_{m'=1}^{M}\left(CR_{cm'} - R_{\cdot m'}\right)^2\beta^2_{m'}\right) + \\
&\quad \left.\left((\alpha\beta)^2_{cm} + \frac{1}{T^2}\sum_{m'=1}^{M}R^2_{cm'}(\alpha\beta)^2_{cm'} + \frac{1}{R^2_{\cdot m}}\sum_{c'=1}^{C}R^2_{c'm}(\alpha\beta)_{c'm} + \right.\right.
\end{aligned}
$$

$$
\begin{aligned}
& \frac{1}{C^2T^2}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}^2(\alpha\beta)_{c'm'}^2-\frac{2}{T}R_{cm}(\alpha\beta)_{cm}^2-\frac{2}{R_{\cdot m}}R_{cm}(\alpha\beta)_{cm}^2+ \\
& \frac{2}{CT}R_{cm}(\alpha\beta)_{cm}^2+\frac{2}{TR_{\cdot m}}R_{cm}^2(\alpha\beta)_{cm}^2-\frac{2}{CT^2}\sum_{m'=1}^{M}R_{cm'}^2(\alpha\beta)_{cm'}^2- \\
& \left.\frac{2}{R_{\cdot m}CT}\sum_{c'=1}^{C}R_{c'm}^2(\alpha\beta)_{c'm}\right)+\left(\frac{(C-1)(M-1)}{R_{cm}CM}\sigma^2_{\mathrm{er(ch\times mo)}}\right)\Bigg\} \\
= \quad & \frac{1}{(C-1)(M-1)}\Bigg\{\left(\sum_{m=1}^{M}R_{\cdot m}\left[\frac{1}{R_{\cdot m}^2C^2}\sum_{c'=1}^{C}\left(CR_{c'm}-R_{\cdot m}\right)^2\sigma^2_{\mathrm{ch}}\right]\right)+ \\
& \left(T\sum_{c=1}^{C}\left[\frac{1}{C^2T^2}\sum_{m'=1}^{M}\left(CR_{cm'}-R_{\cdot m'}\right)^2\beta_{m'}^2\right]\right)+ \\
& \Bigg(CT\sigma^2_{\mathrm{ch\times mo}}+T\sum_{c=1}^{C}\left[\frac{1}{T^2}\sum_{m'=1}^{M}R_{cm'}^2\sigma^2_{\mathrm{ch\times mo}}\right]+ \\
& \sum_{m=1}^{M}R_{\cdot m}\left[\frac{1}{R_{\cdot m}^2}\sum_{c'=1}^{C}R_{c'm}^2\sigma^2_{\mathrm{ch\times mo}}\right]+CT\left[\frac{1}{C^2T^2}\sum_{c'=1}^{C}\sum_{m'=1}^{M}R_{c'm'}^2\sigma^2_{\mathrm{ch\times mo}}\right]- \\
& \sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}\left[\frac{2}{T}R_{cm}\sigma^2_{\mathrm{ch\times mo}}\right]-\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}\left[\frac{2}{R_{\cdot m}}R_{cm}\sigma^2_{\mathrm{ch\times mo}}\right]+ \\
& \sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}\left[\frac{2}{CT}R_{cm}\sigma^2_{\mathrm{ch\times mo}}\right]+\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}\left[\frac{2}{TR_{\cdot m}}R_{cm}^2\sigma^2_{\mathrm{ch\times mo}}\right]- \\
& T\sum_{c=1}^{C}\left[\frac{2}{CT^2}\sum_{m'=1}^{M}R_{cm'}^2\sigma^2_{\mathrm{ch\times mo}}\right]-\sum_{m=1}^{M}R_{\cdot m}\left[\frac{2}{R_{\cdot m}CT}\sum_{c'=1}^{C}R_{c'm}^2\sigma^2_{\mathrm{ch\times mo}}\right]\Bigg)+ \\
& \left(\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}\left[\frac{(C-1)(M-1)}{R_{cm}CM}\sigma^2_{\mathrm{er(ch\times mo)}}\right]\right)\Bigg\} \\
= \quad & \frac{1}{(C-1)(M-1)}\Bigg\{\left(\frac{1}{C^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\left(CR_{cm}-R_{\cdot m}\right)^2}{R_{\cdot m}}\sigma^2_{\mathrm{ch}}\right)+ \\
& \left(\frac{1}{C^2T}\sum_{c=1}^{C}\sum_{m=1}^{M}\left(CR_{cm}-R_{\cdot m}\right)^2\beta_m^2\right)+ \\
& \Bigg(CT\sigma^2_{\mathrm{ch\times mo}}+\frac{1}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\sigma^2_{\mathrm{ch\times mo}}+
\end{aligned}
$$

$$
\begin{aligned}
& \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^2}{R_{\cdot m}}\sigma^2_{\text{ch}\times\text{mo}} + \frac{1}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\sigma^2_{\text{ch}\times\text{mo}} - \\
& \frac{2}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\sigma^2_{\text{ch}\times\text{mo}} - 2\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^2}{R_{\cdot m}}\sigma^2_{\text{ch}\times\text{mo}} + \\
& \frac{2}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\sigma^2_{\text{ch}\times\text{mo}} + \frac{2}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^3}{R_{\cdot m}}\sigma^2_{\text{ch}\times\text{mo}} - \\
& \frac{2}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\sigma^2_{\text{ch}\times\text{mo}} - \frac{2}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\sigma^2_{\text{ch}\times\text{mo}}\Bigg) + \\
& \left((C-1)(M-1)\sigma^2_{\text{er(ch}\times\text{mo)}}\right)\Big\} \\
= \quad & \left[\frac{1}{C^2(C-1)(M-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{(CR_{cm}-R_{\cdot m})^2}{R_{\cdot m}}\right]\sigma^2_{\text{ch}} + \\
& \frac{1}{C^2T(C-1)(M-1)}\sum_{m=1}^{M}\left[\sum_{c=1}^{C}(CR_{cm}-R_{\cdot m})^2\right]\beta_m^2 + \\
& \frac{1}{(C-1)(M-1)}\left[CT + \frac{1}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2 + \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^2}{R_{\cdot m}} + \frac{1}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2 - \right. \\
& \frac{2}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2 - 2\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^2}{R_{\cdot m}} + \frac{2}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2 + \frac{2}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^3}{R_{\cdot m}} - \\
& \left.\frac{4}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\right]\sigma^2_{\text{ch}\times\text{mo}} + \sigma^2_{\text{er(ch}\times\text{mo)}} \\
= \quad & \left[\frac{1}{C^2(C-1)(M-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{(CR_{cm}-R_{\cdot m})^2}{R_{\cdot m}}\right]\sigma^2_{\text{ch}} + \\
& \frac{1}{C^2T(C-1)(M-1)}\sum_{m=1}^{M}\left[\sum_{c=1}^{C}(CR_{cm}-R_{\cdot m})^2\right]\beta_m^2 + \\
& \frac{1}{(C-1)(M-1)}\left[CT - \frac{C+1}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2 - \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^2}{R_{\cdot m}} + \right. \\
& \left.\frac{2}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^3}{R_{\cdot m}}\right]\sigma^2_{\text{ch}\times\text{mo}} + \sigma^2_{\text{er(ch}\times\text{mo)}}
\end{aligned}
$$

Finally, putting it all together,

$$
\begin{aligned}
EW_m &= E\mathrm{MS}_{\mathrm{er(mo)}} \\
&= \frac{1}{CT-M} E\left[\mathrm{SS}_{\mathrm{er(ch\times mo)}} + \mathrm{SS}_{\mathrm{ch}} + \mathrm{SS}_{\mathrm{ch\times mo}}\right] \\
&= \frac{1}{CT-M}\left[C(T-M)E\mathrm{MS}_{\mathrm{er(ch\times mo)}} + (C-1)E\mathrm{MS}_{\mathrm{ch}} + \right. \\
&\quad \left.(C-1)(M-1)E\mathrm{MS}_{\mathrm{ch\times mo}}\right] \\
&= \frac{C(T-M)}{CT-M}\sigma^2_{\mathrm{er(ch\times mo)}} + \\
&\quad \frac{C-1}{CT-M}\left\{T\sigma^2_{\mathrm{ch}} + \frac{1}{(C-1)C^2T}\sum_{m=1}^{M}\left[\sum_{c=1}^{C}\left(CR_{cm}-R_{\cdot m}\right)^2\right]\beta_m^2 + \right. \\
&\quad \left.\left[\frac{1}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2\right]\sigma^2_{\mathrm{ch\times mo}} + \sigma^2_{\mathrm{er(ch\times mo)}}\right\} + \\
&\quad \frac{(C-1)(M-1)}{CT-M}\left\{\left[\frac{1}{C^2(C-1)(M-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\left(CR_{cm}-R_{\cdot m}\right)^2}{R_{\cdot m}}\right]\sigma^2_{\mathrm{ch}} + \right. \\
&\quad \frac{1}{C^2T(C-1)(M-1)}\sum_{m=1}^{M}\left[\sum_{c=1}^{C}\left(CR_{cm}-R_{\cdot m}\right)^2\right]\beta_m^2 + \\
&\quad \frac{1}{(C-1)(M-1)}\left[CT - \frac{C+1}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}R_{cm}^2 - \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^2}{R_{\cdot m}} + \right. \\
&\quad \left.\left.\frac{2}{T}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^3}{R_{\cdot m}}\right]\sigma^2_{\mathrm{ch\times mo}} + \sigma^2_{\mathrm{er(ch\times mo)}}\right\} \\
&= \left[\frac{C(T-M)+C-1+(C-1)(M-1)}{CT-M}\right]\sigma^2_{\mathrm{er(ch\times mo)}} + \\
&\quad \frac{1}{CT-M}\left[(C-1)T + \frac{1}{C^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\left(CR_{cm}-R_{\cdot m}\right)^2}{R_{\cdot m}}\right]\sigma^2_{\mathrm{ch}} + \\
&\quad \frac{2}{(CT-M)C^2T}\sum_{m=1}^{M}\left[\sum_{c=1}^{C}\left(CR_{cm}-R_{\cdot m}\right)^2\right]\beta_m^2 + \\
&\quad \frac{1}{CT-M}\left[CT + \sum_{c=1}^{C}\sum_{m=1}^{M}\left(\frac{C-1}{CT} - \frac{C+1}{CT} - \frac{1}{R_{\cdot m}} + \frac{2R_{cm}}{TR_{\cdot m}}\right)R_{cm}^2\right]\sigma^2_{\mathrm{ch\times mo}} \\
&= \sigma^2_{\mathrm{er(ch\times mo)}} + \left[\frac{(C-1)T}{CT-M} + \frac{1}{C^2(CT-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\left(CR_{cm}-R_{\cdot m}\right)^2}{R_{\cdot m}}\right]\sigma^2_{\mathrm{ch}} +
\end{aligned}
$$

$$\frac{2}{C^3T - C^2MT}\sum_{m=1}^{M}\left[\sum_{c=1}^{C}\left(CR_{cm} - R_{\cdot m}\right)^2\right]\beta_m^2 +$$
$$\left[\frac{CT}{CT-M} + \frac{-1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{R_{cm}^2}{R_{\cdot m}} +\right.$$
$$\left.\frac{2}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}\left(CR_{cm} - R_{\cdot m}\right)\frac{R_{cm}^2}{R_{\cdot m}}\right]\sigma^2_{\mathrm{ch}\times\mathrm{mo}}\ . \quad \square$$

APPENDIX B
RJMCMC ALGORITHM PERFORMANCE

Pattern	Split	Combine	Birth	Death
Redwoods	0.051881 11669 (224919)	0.052235 11754 (225020)	0.016041 1202 (74933)	0.015028 1129 (75128)
I-k7-a	0.045025 10124 (224853)	0.045321 10208 (225239)	0.010568 793 (75037)	0.009763 731 (74871)
I-k7-b	0.024376 5494 (225390)	0.024257 5448 (224598)	0.005143 386 (75053)	0.00611 458 (74959)
I-k14-a	0.036347 8184 (225163)	0.036521 8205 (224666)	0.007409 551 (74366)	0.007321 555 (75805)
I-k14-b	0.025911 5850 (225777)	0.025978 5841 (224844)	0.007473 560 (74940)	0.007805 581 (74439)
AI-1.5-k7-a	0.01712 3847 (224713)	0.01695 3819 (225311)	0.00367 274 (74654)	0.004408 332 (75322)
AI-1.5-k7-b	0.008435 1892 (224292)	0.008272 1869 (225953)	0.002047 153 (74728)	0.002679 201 (75027)
AI-1.5-k14-a	0.041356 9313 (225190)	0.040739 9178 (225290)	0.00797 595 (74651)	0.010071 754 (74869)
AI-1.5-k14-b	0.036242 8142 (224656)	0.037254 8387 (225130)	0.00972 730 (75102)	0.00667 501 (75112)
AI-3-k7-a	0.013192 2969 (225056)	0.013185 2962 (224649)	0.003597 269 (74782)	0.004026 304 (75513)
AI-3-k7-b	0.023627 5312 (224828)	0.023558 5295 (224761)	0.004029 302 (74960)	0.004599 347 (75451)
AI-3-k14-a	0.038546 8688 (225392)	0.038566 8662 (224603)	0.008633 650 (75289)	0.009289 694 (74716)
AI-3-k14-b	0.045239 10167 (224740)	0.045145 10176 (225407)	0.010964 821 (74879)	0.01091 818 (74974)

Table B.1: Acceptance rates for dimension-changing moves, for all sweeps in all chains. Entries are: acceptance rate, #successes, (#total).

Pattern	NN_{Σ} Violation in M_S	$k = k_{\text{hi}}$ in M_S/M_B	$k = k_{\text{lo}}$ in M_C/M_D
Redwoods	0.172258 38744 (224919)	3.3e-005 10 (299852)	7e-006 2 (300148)
I-k7-a	0.155804 35033 (224853)	2e-005 6 (299890)	3e-005 9 (300110)
I-k7-b	0.083331 18782 (225390)	0 0 (300443)	3e-006 1 (299557)
I-k14-a	0.143243 32253 (225163)	1e-005 3 (299529)	7e-006 2 (300471)
I-k14-b	0.074268 16768 (225777)	2e-005 6 (300717)	2.3e-005 7 (299283)
AI-1.5-k7-a	0.168722 37914 (224713)	7e-006 2 (299367)	7.7e-005 23 (300633)
AI-1.5-k7-b	0.053511 12002 (224292)	0 0 (299020)	0.000136 41 (300980)
AI-1.5-k14-a	0.071482 16097 (225190)	3e-006 1 (299841)	6.7e-005 20 (300159)
AI-1.5-k14-b	0.202906 45584 (224656)	0 0 (299758)	0.000183 55 (300242)
AI-3-k7-a	0.158165 35596 (225056)	1.3e-005 4 (299838)	1e-005 3 (300162)
AI-3-k7-b	0.089428 20106 (224828)	3e-006 1 (299788)	9e-005 27 (300212)
AI-3-k14-a	0.070739 15944 (225392)	0 0 (300681)	0 0 (299319)
AI-3-k14-b	0.104592 23506 (224740)	7e-006 2 (299619)	1e-005 3 (300381)

Table B.2: Occurrence rates of move-disqualifying conditions, for all sweeps in all chains. Entries are: occurrence rate, #occurrences, (#total).

APPENDIX C
POINT PATTERNS, SHOWING TRACKED OFFSPRING

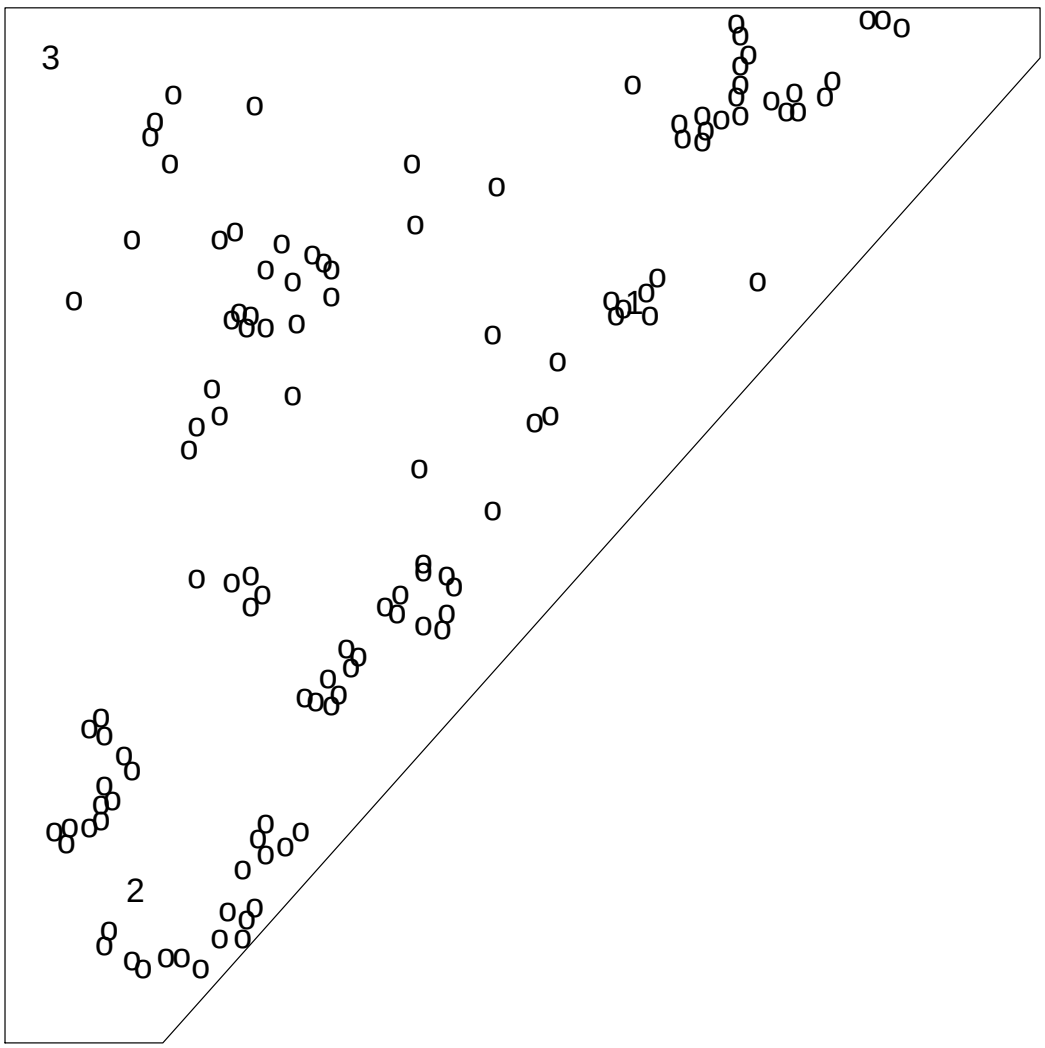
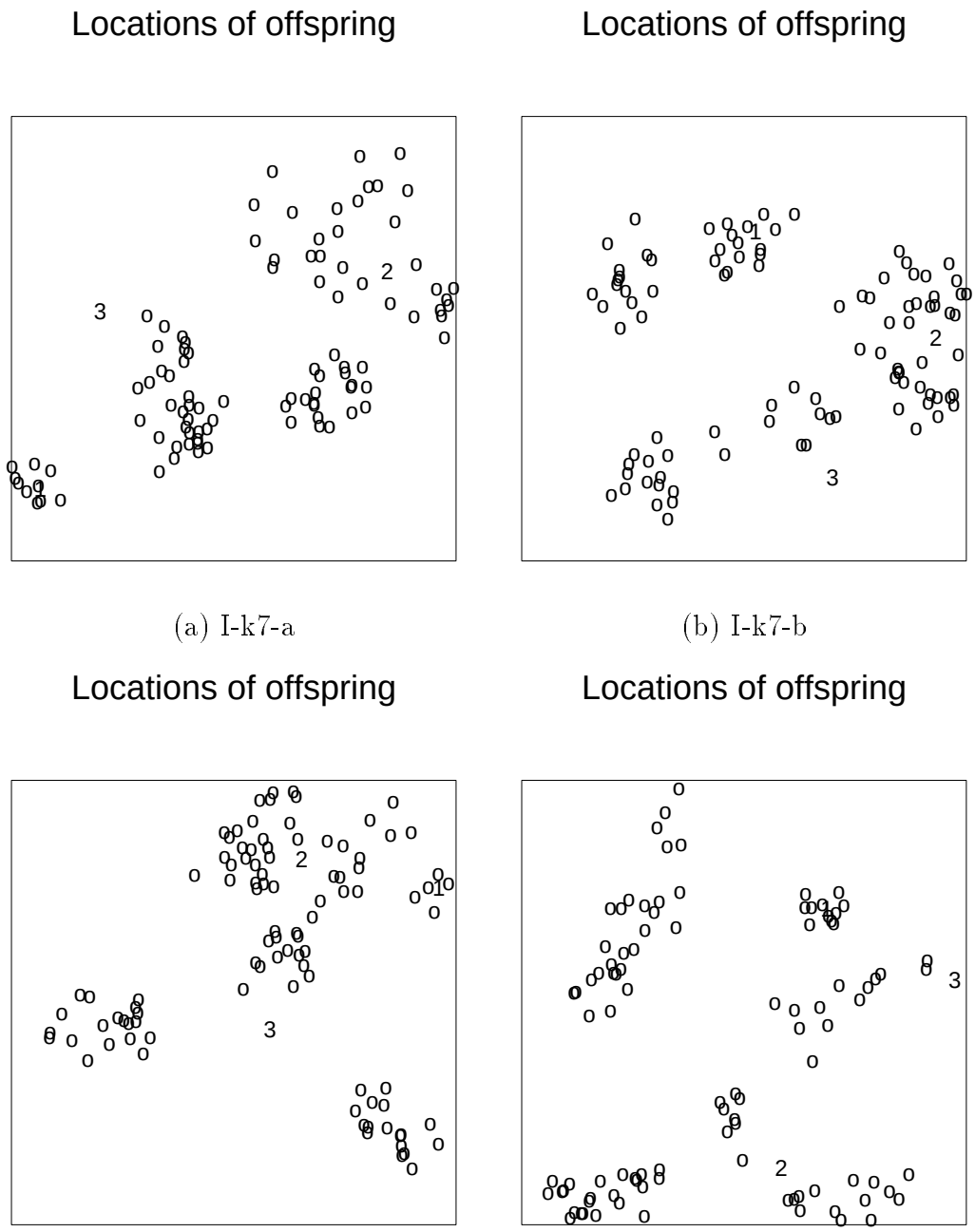


Figure C.1: Locations of Redwood seedlings, with tracked offspring marked.



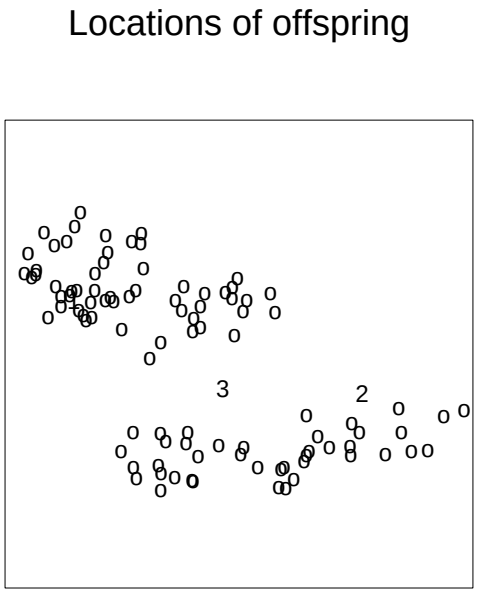
(a) I-k7-a

(b) I-k7-b

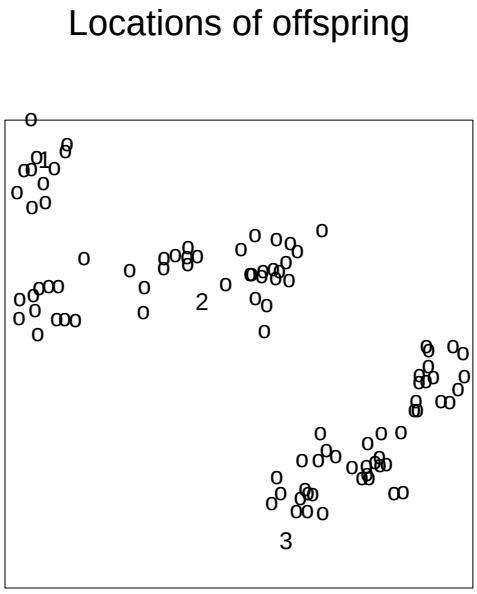
(c) I-k14-a

(d) I-k14-b

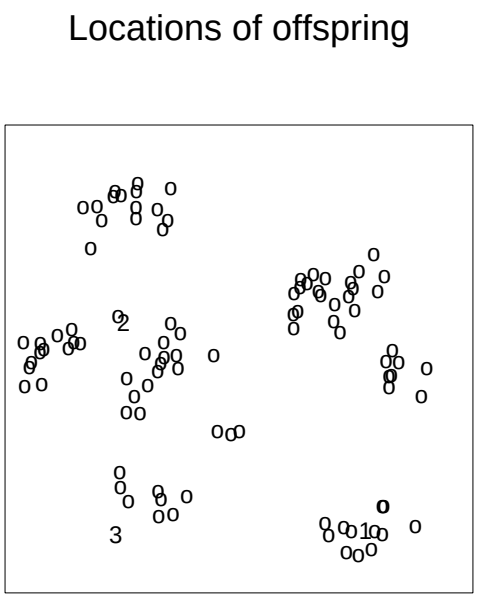
Figure C.2: Simulated point patterns, with tracked offspring marked.



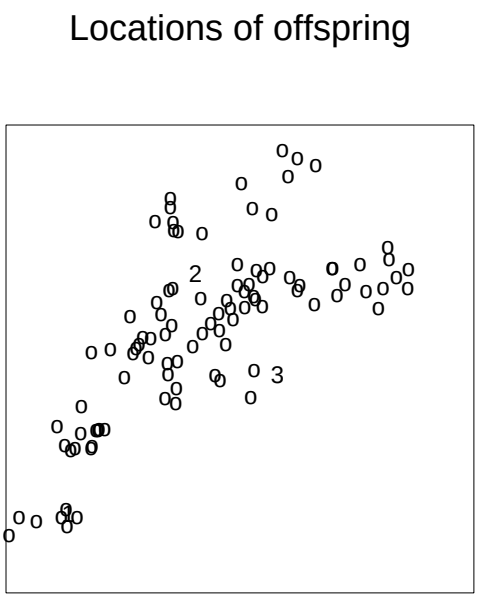
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

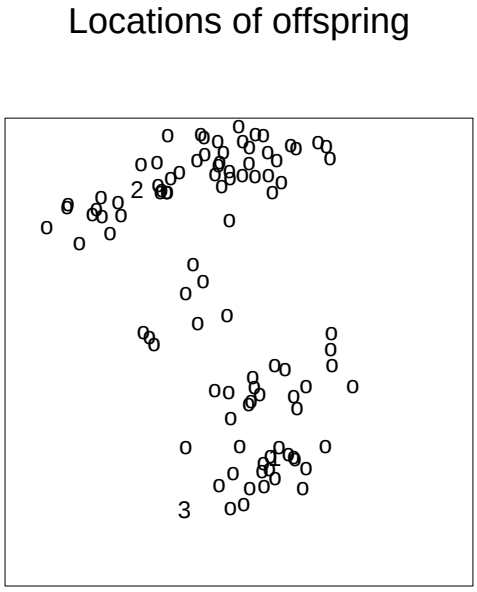


(g) AI-1.5-k14-a

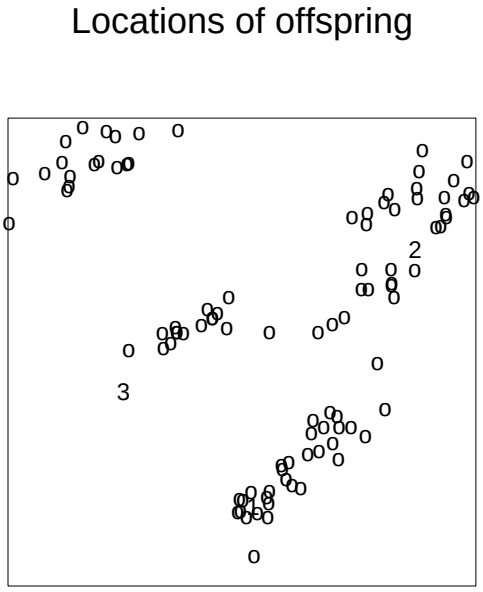


(h) AI-1.5-k14-b

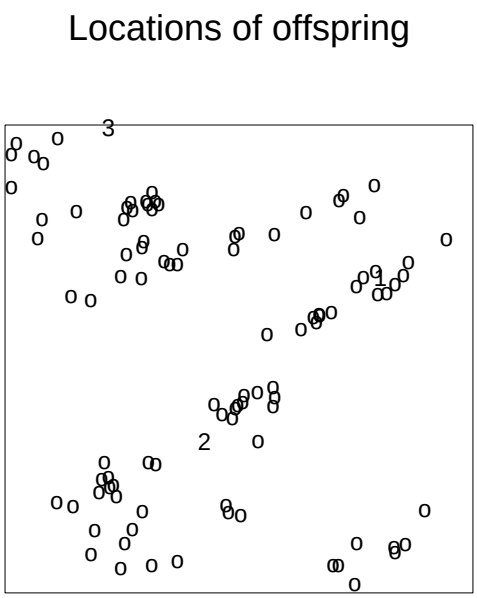
Figure C.2 (continued).



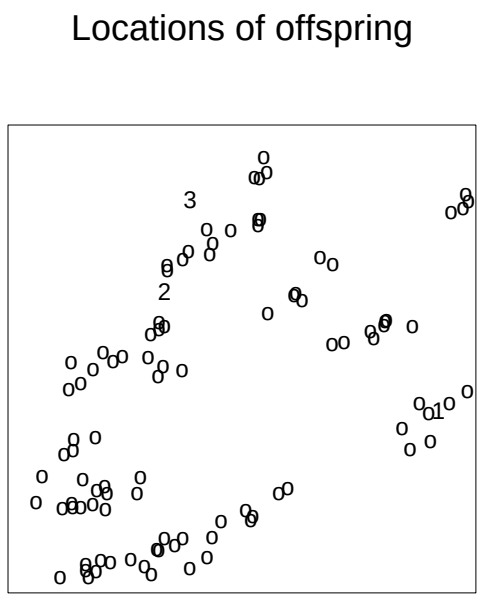
(i) AI-3-k7-a



(j) AI-3-k7-b



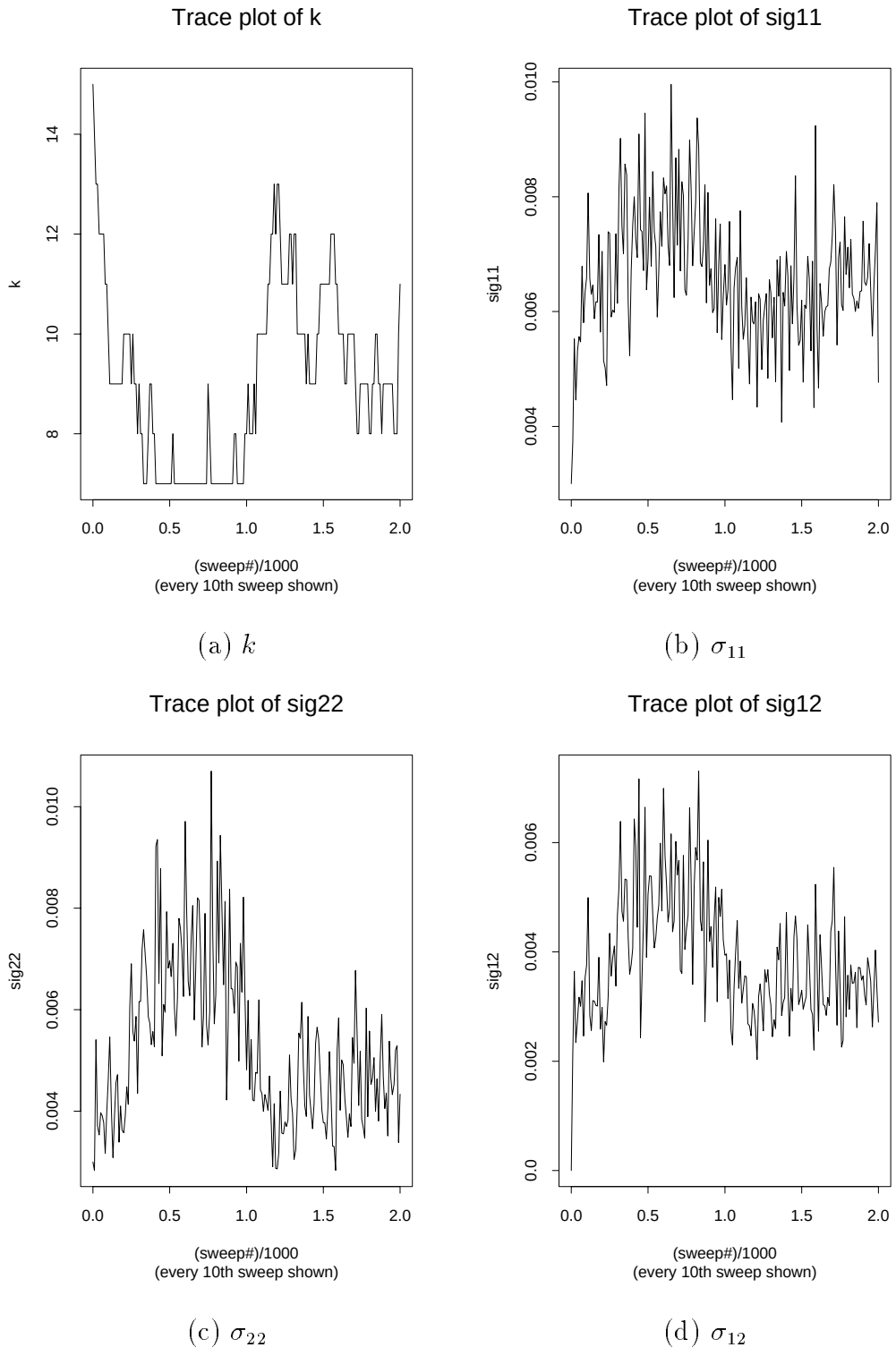
(k) AI-3-k14-a



(l) AI-3-k14-b

Figure C.2 (continued).

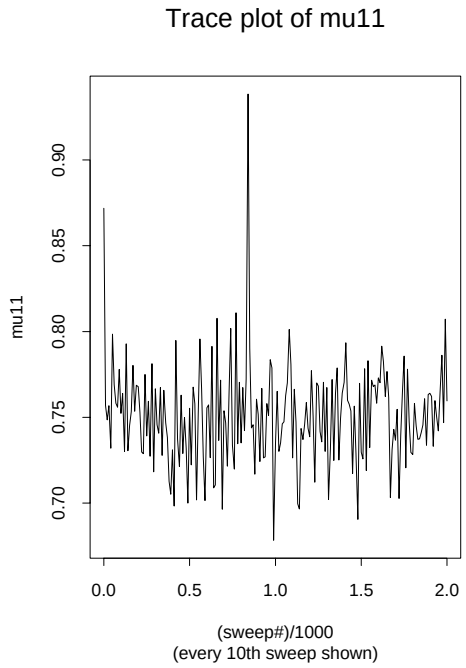
APPENDIX D
SAMPLE TRACE PLOTS, CLUSTER MEMBERSHIPS, AND ACF'S FOR REDWOOD DATA



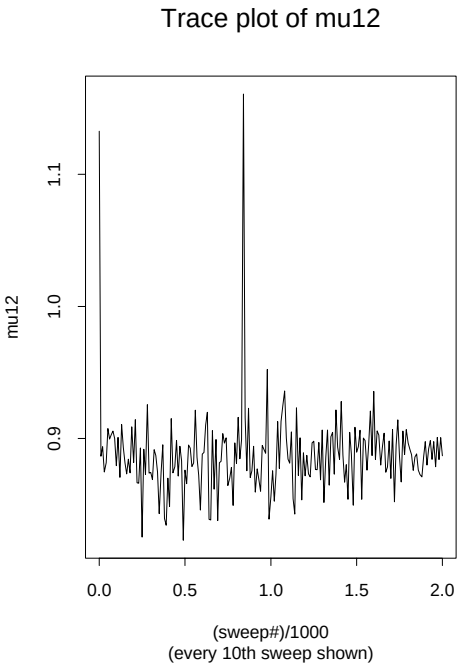
(a) k (b) σ_{11}

(c) σ_{22} (d) σ_{12}

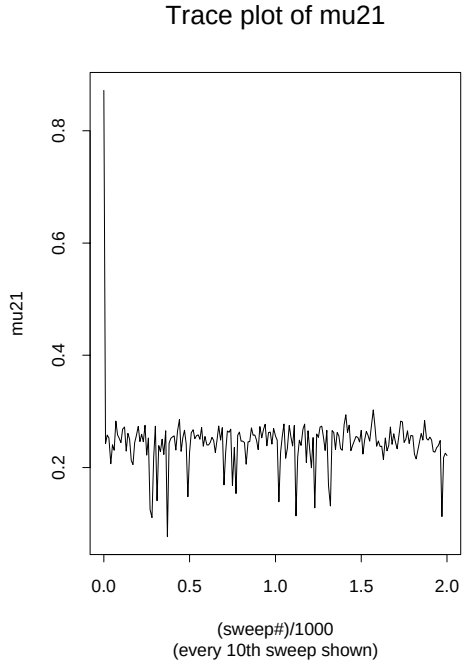
Figure D.1: Trace plots of monitored parameters for a 2,000-sweep RJMCMC run, Redwood data.



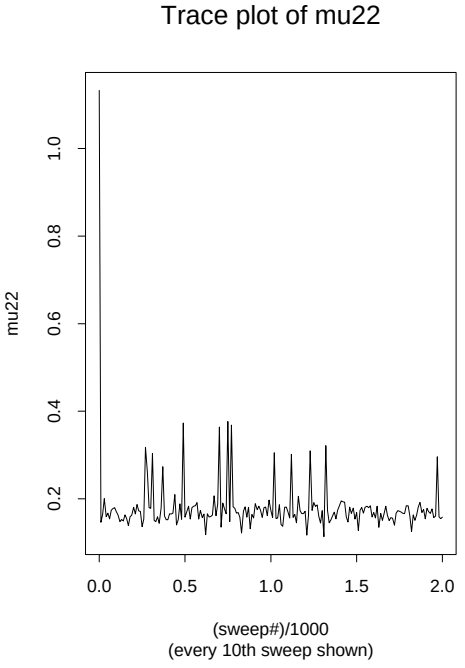
(e) $\boldsymbol{\mu}_{j_1,1}$



(f) $\boldsymbol{\mu}_{j_1,2}$

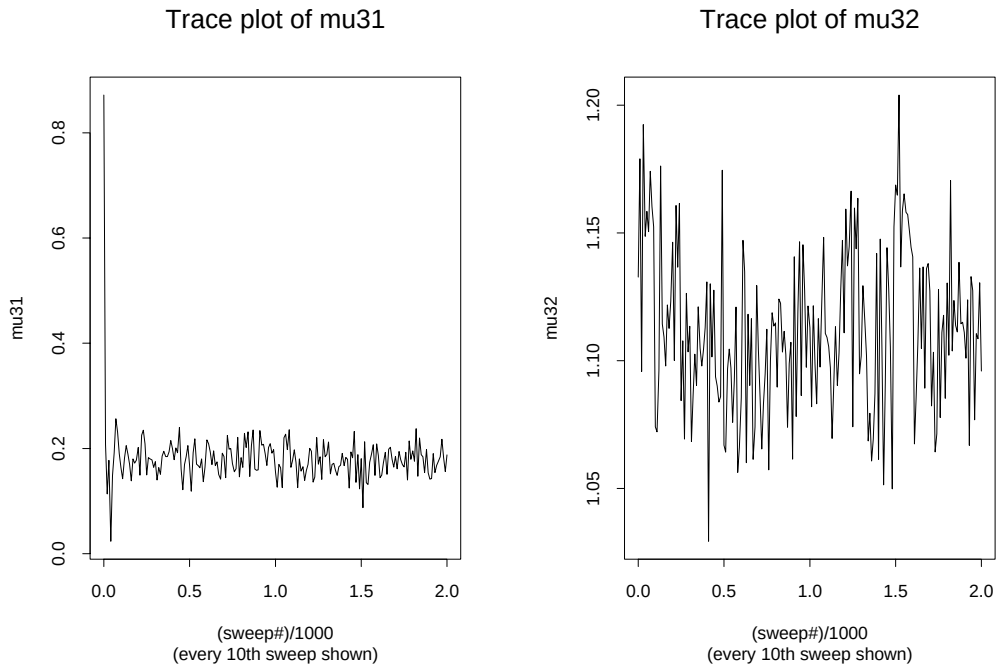


(g) $\boldsymbol{\mu}_{j_2,1}$



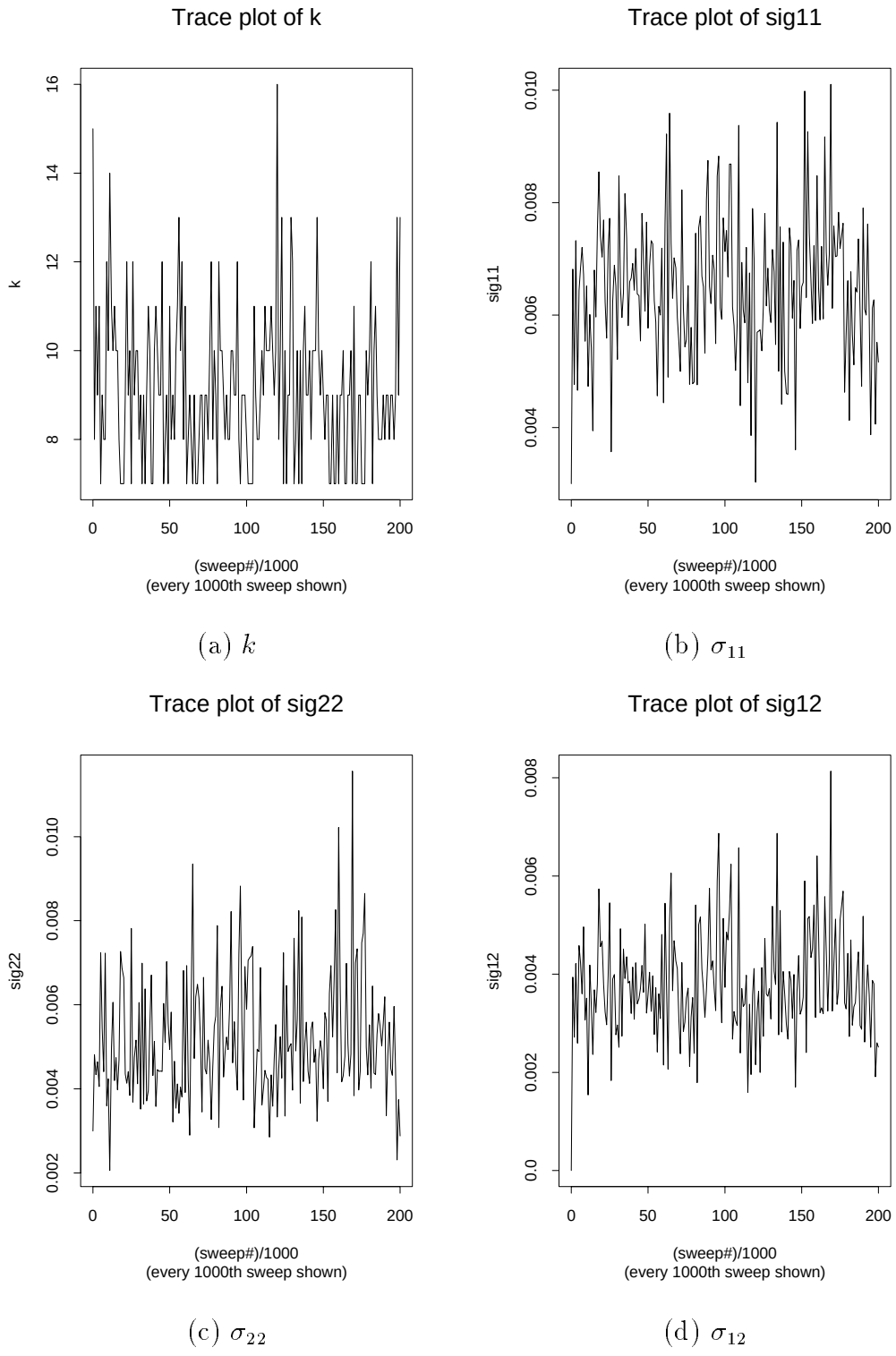
(h) $\boldsymbol{\mu}_{j_2,2}$

Figure D.1 (continued).



(i) $\boldsymbol{\mu}_{j_3,1}$ (j) $\boldsymbol{\mu}_{j_3,2}$

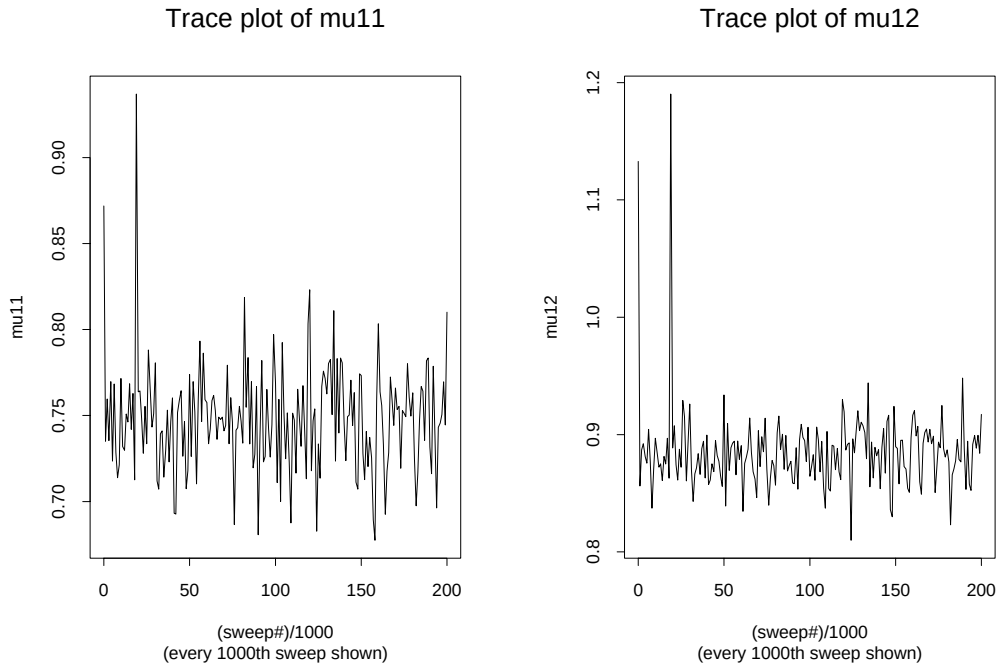
Figure D.1 (continued).



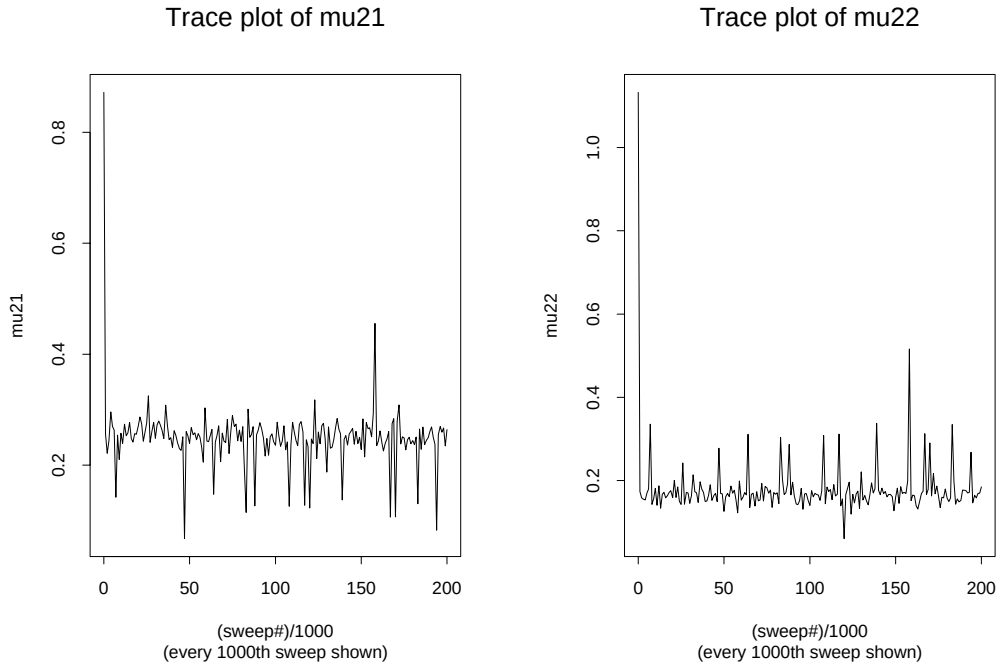
(a) k (b) σ_{11}

(c) σ_{22} (d) σ_{12}

Figure D.2: Trace plots of monitored parameters for a 200,000-sweep RJMCMC run, Redwood data.

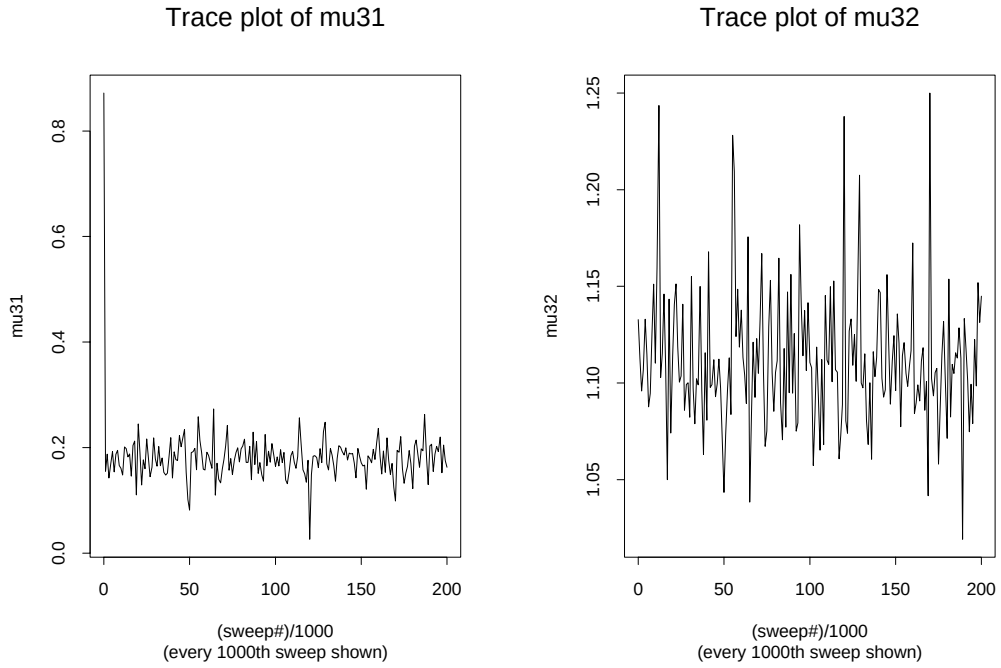


(e) $\boldsymbol{\mu}_{j_1,1}$ (f) $\boldsymbol{\mu}_{j_1,2}$



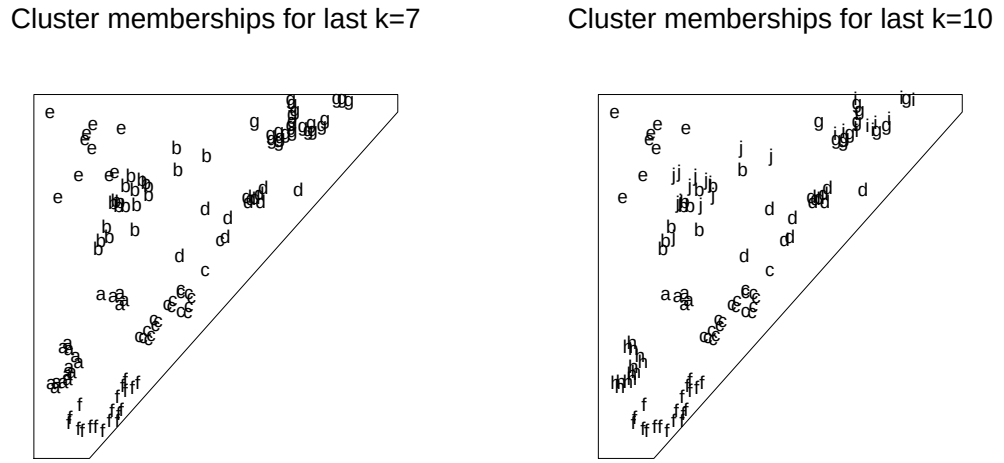
(g) $\boldsymbol{\mu}_{j_2,1}$ (h) $\boldsymbol{\mu}_{j_2,2}$

Figure D.2 (continued).

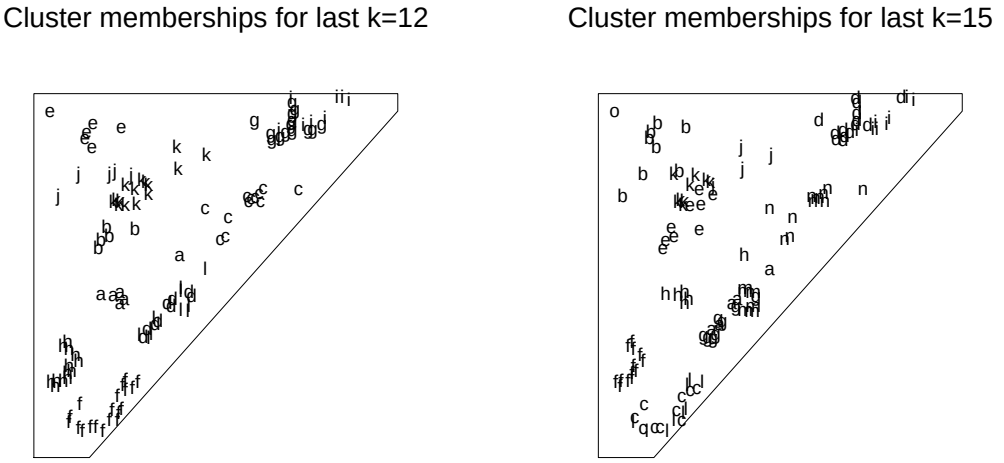


(i) $\boldsymbol{\mu}_{j_3,1}$ (j) $\boldsymbol{\mu}_{j_3,2}$

Figure D.2 (continued).

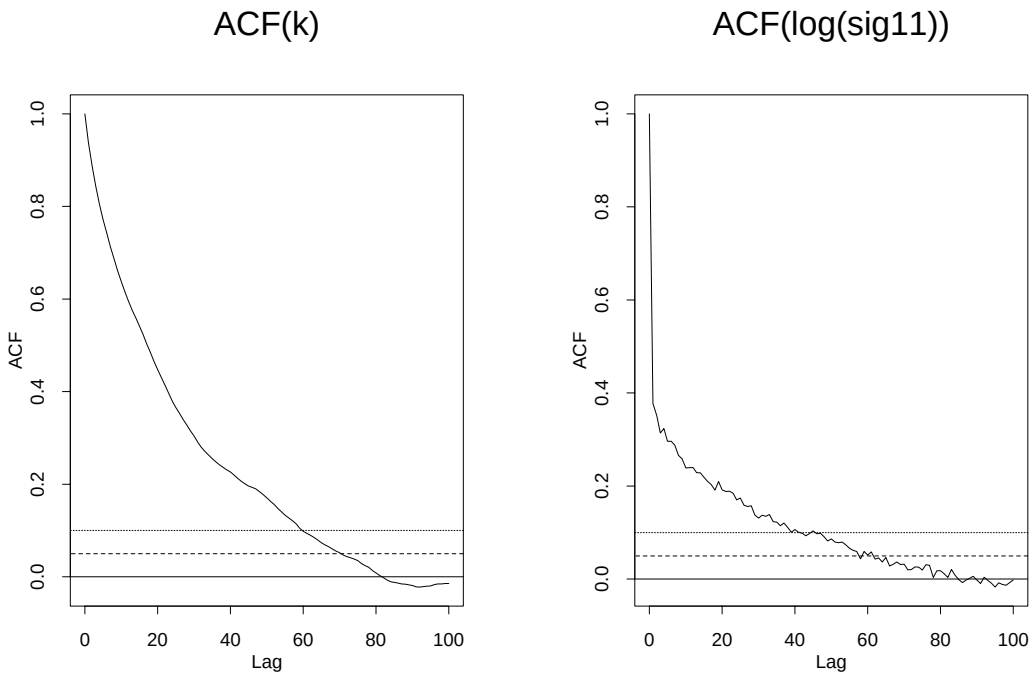


(a) $k = 7$ (b) $k = 10$

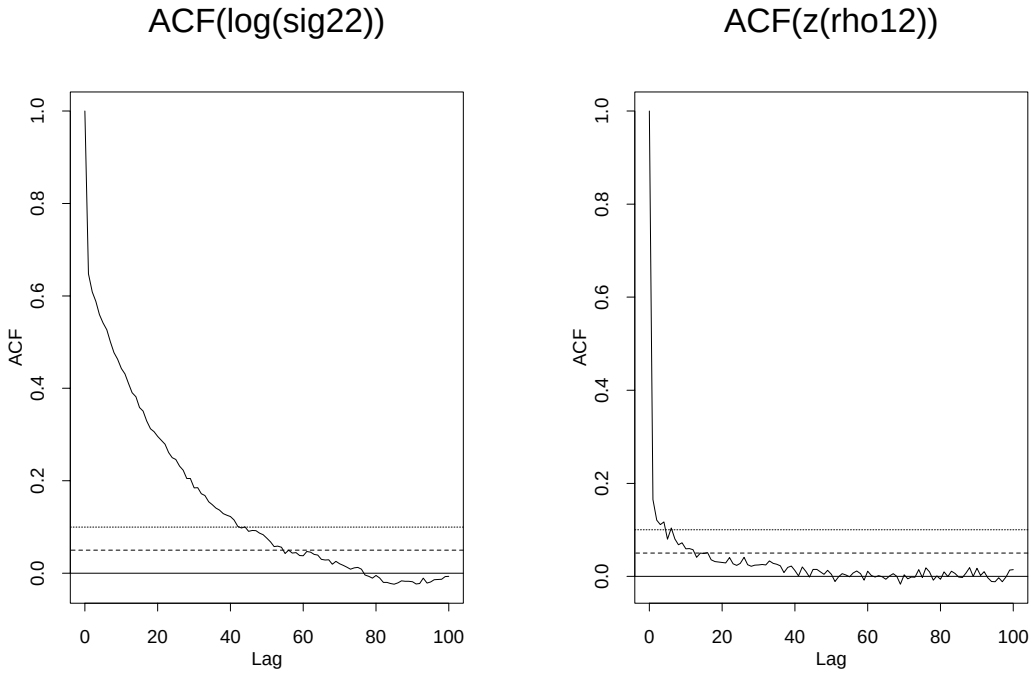


(c) $k = 12$ (d) $k = 15$

Figure D.3: Sample cluster memberships, Redwood data.

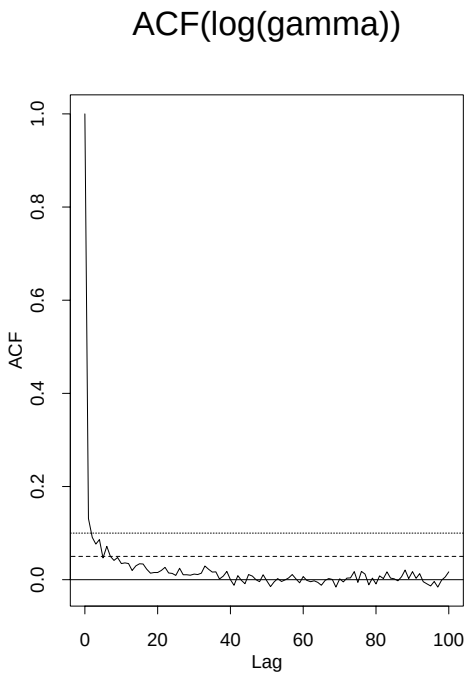


(a) k (b) $\log \sigma_{11}$

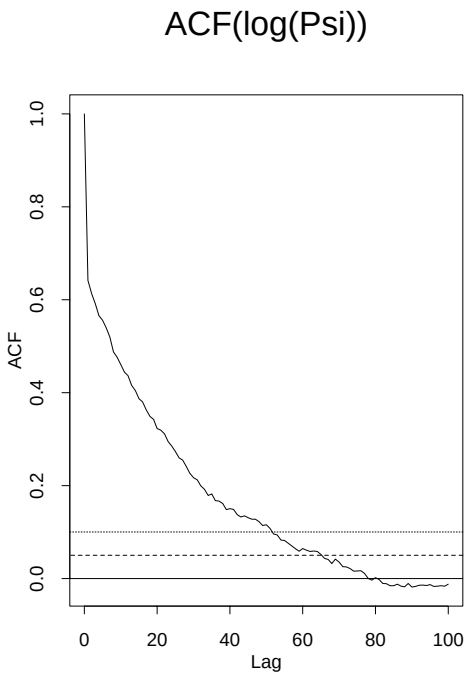


(c) $\log \sigma_{22}$ (d) $z(\rho_{12})$

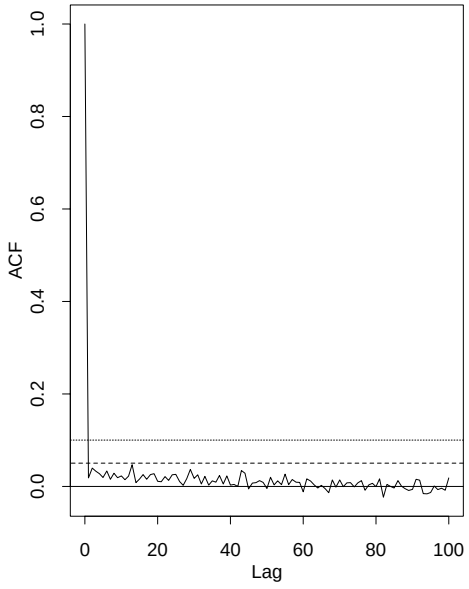
Figure D.4: Autocorrelation functions of normalized versions of monitored parameters in latter half of sweeps (every 10^{th} value used), Redwood data.



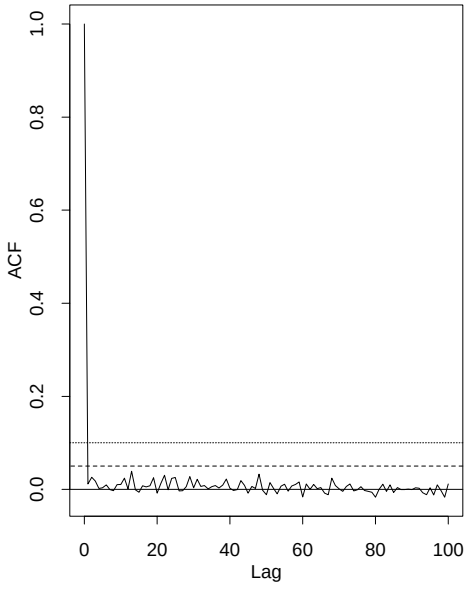
(e) $\log \gamma$



(f) $\log \Psi$

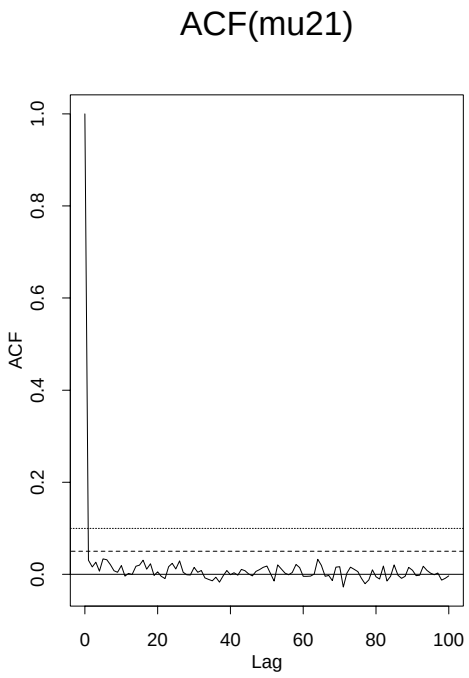


(g) $\boldsymbol{\mu}_{j_1,1}$

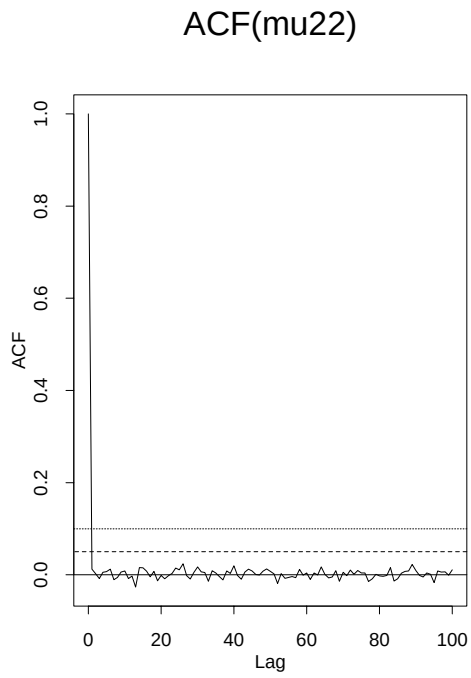


(h) $\boldsymbol{\mu}_{j_1,2}$

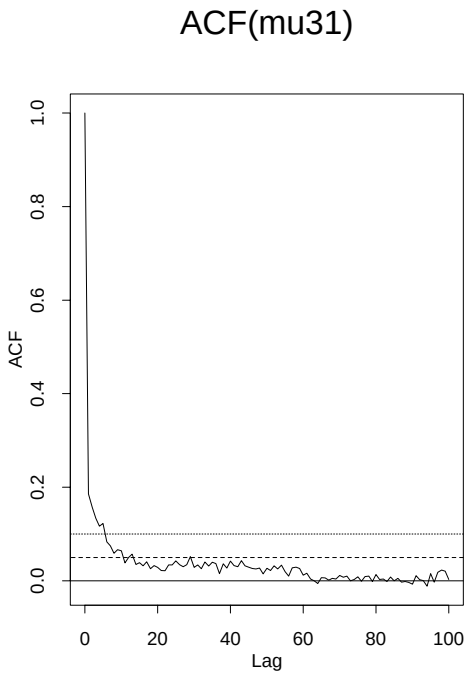
Figure D.4 (continued).



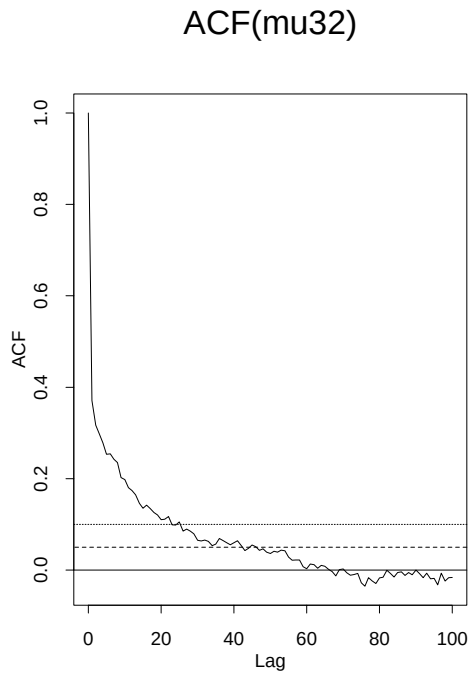
(i) $\boldsymbol{\mu}_{j_2,1}$



(j) $\boldsymbol{\mu}_{j_2,2}$



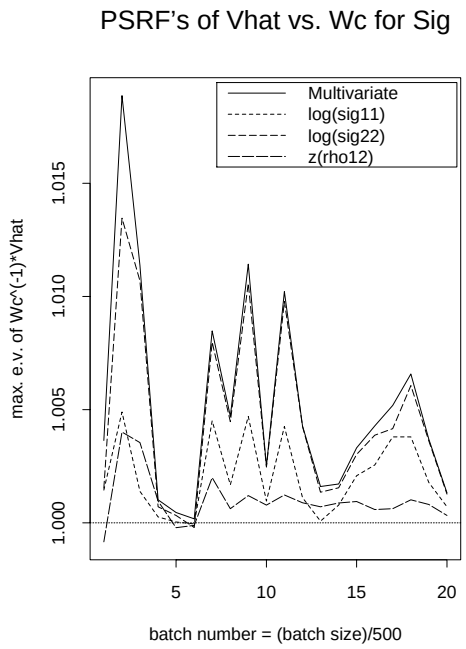
(k) $\boldsymbol{\mu}_{j_3,1}$



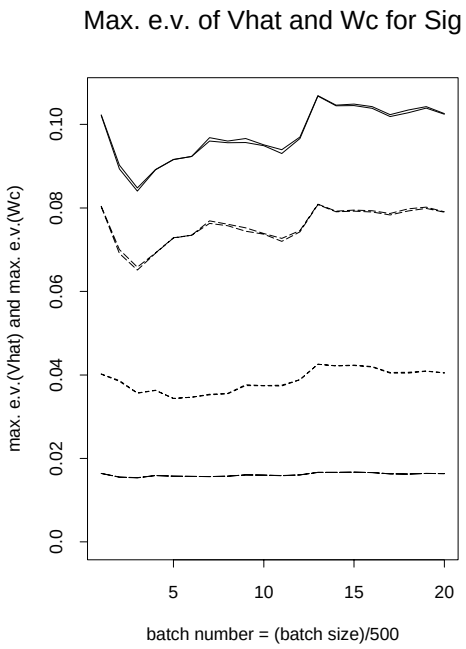
(l) $\boldsymbol{\mu}_{j_3,2}$

Figure D.4 (continued).

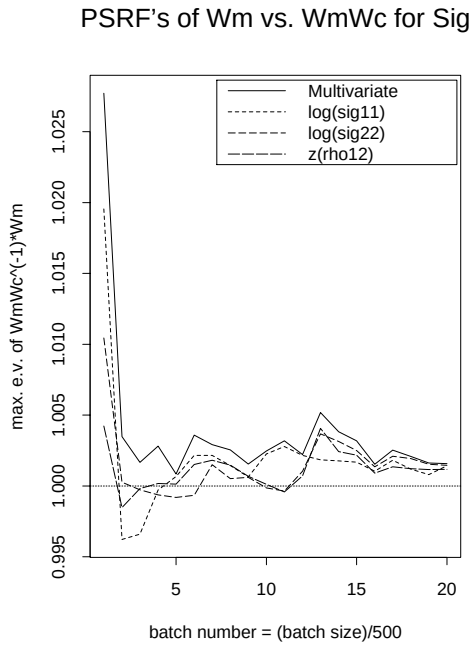
APPENDIX E
CONVERGENCE ASSESSMENT PLOTS



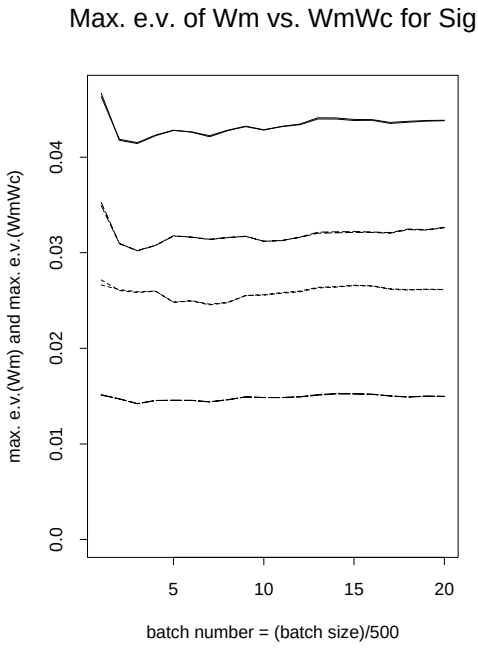
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

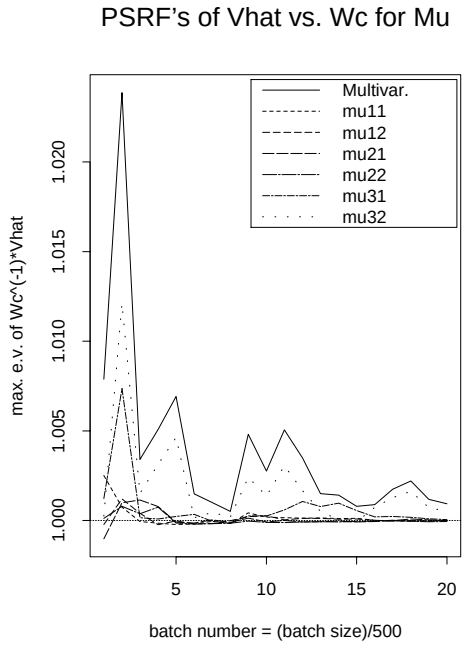


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. W_mW_c

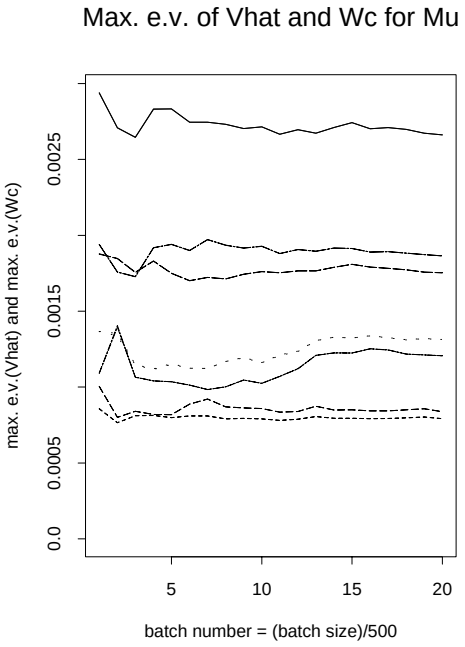


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. W_mW_c

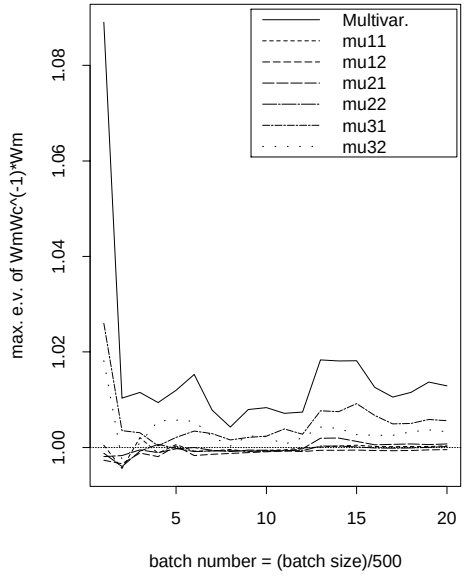
Figure E.1: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, Redwood data.



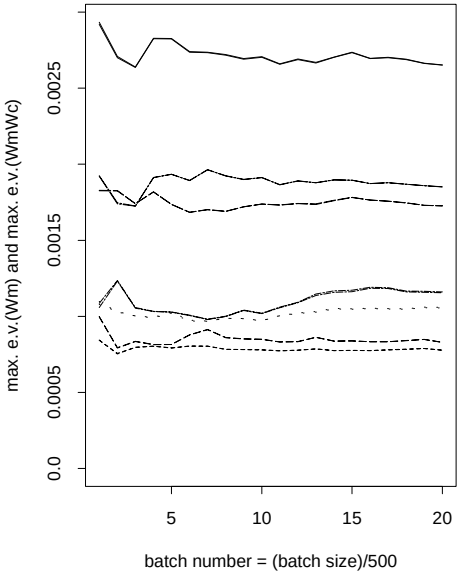
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

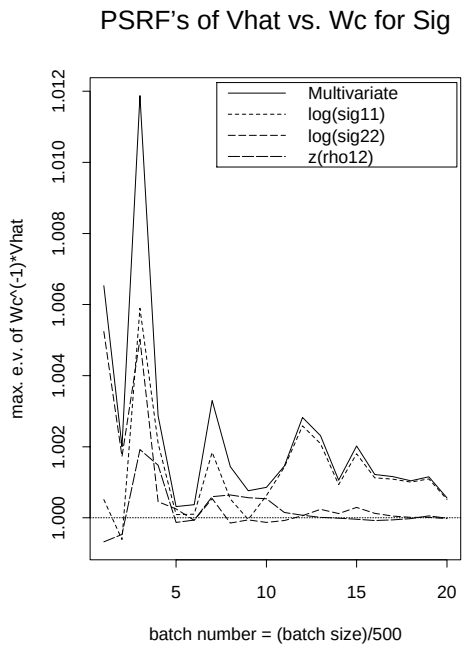


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c

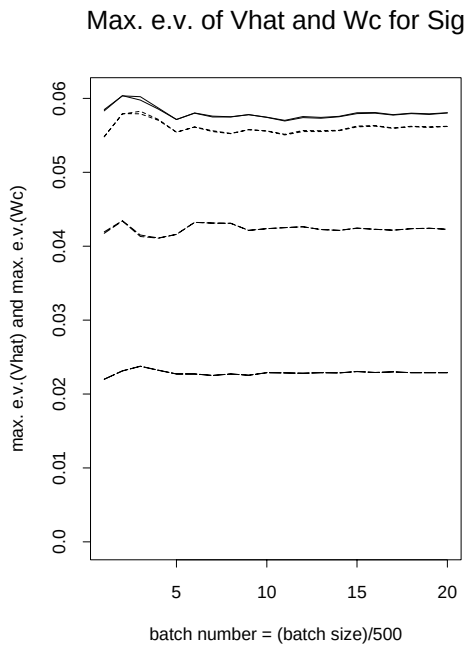


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

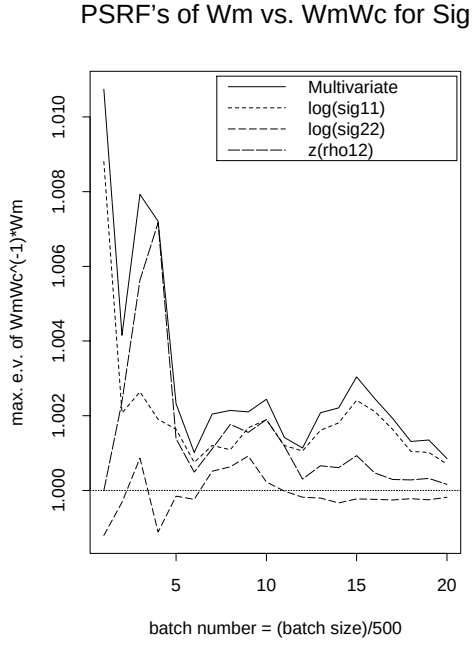
Figure E.1 (continued).



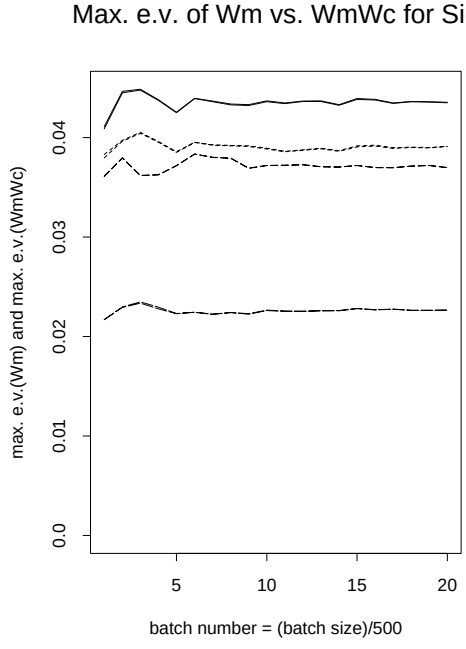
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

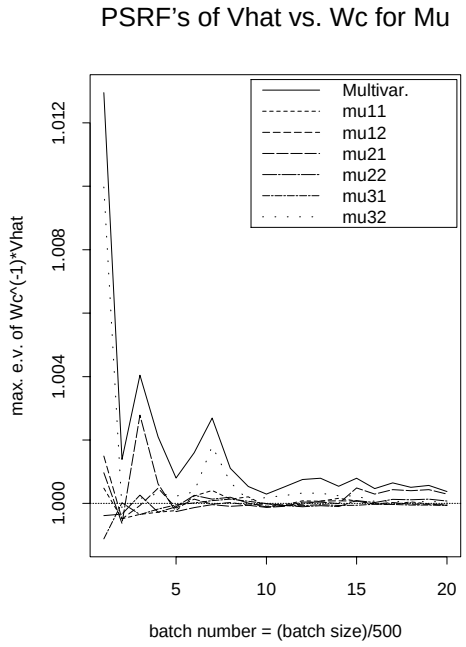


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

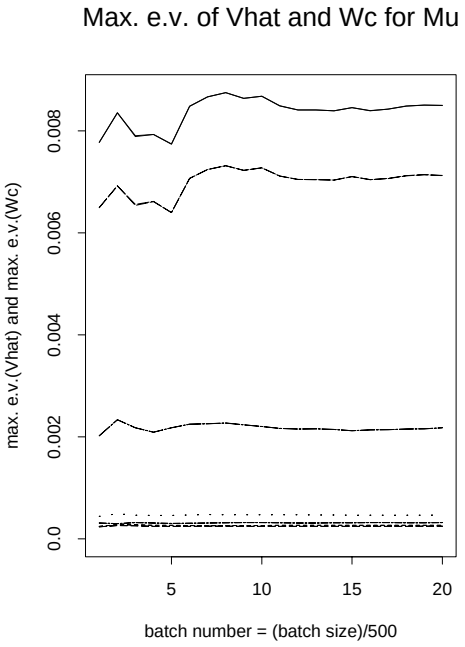


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

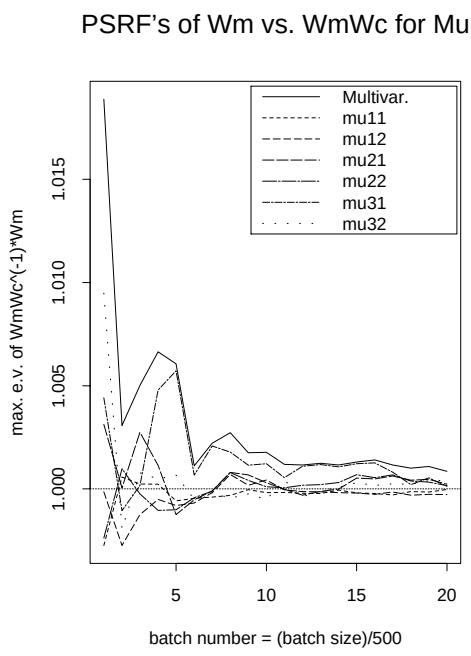
Figure E.2: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, I-k7-a.



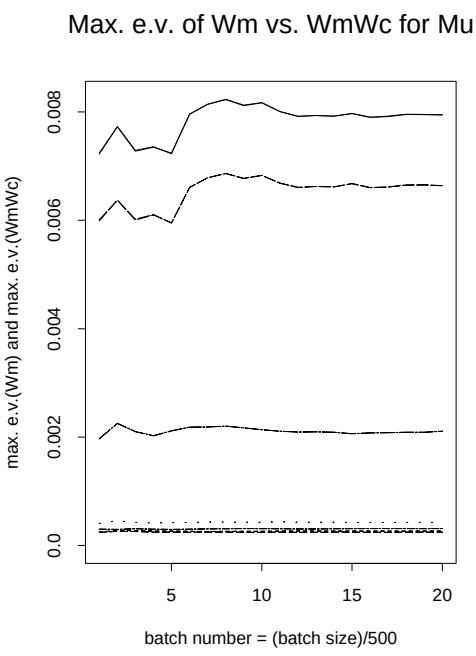
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

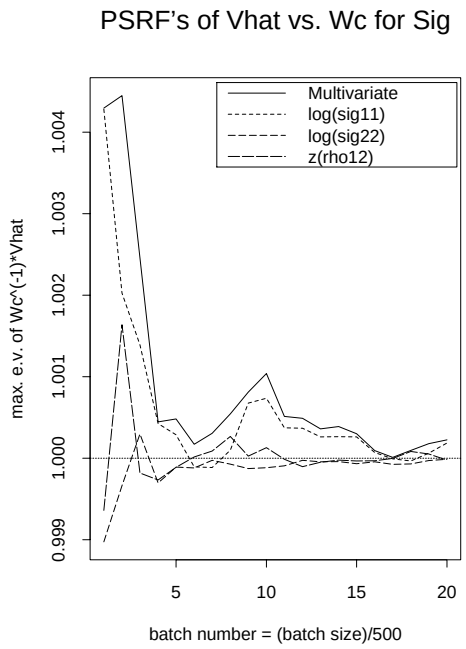


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

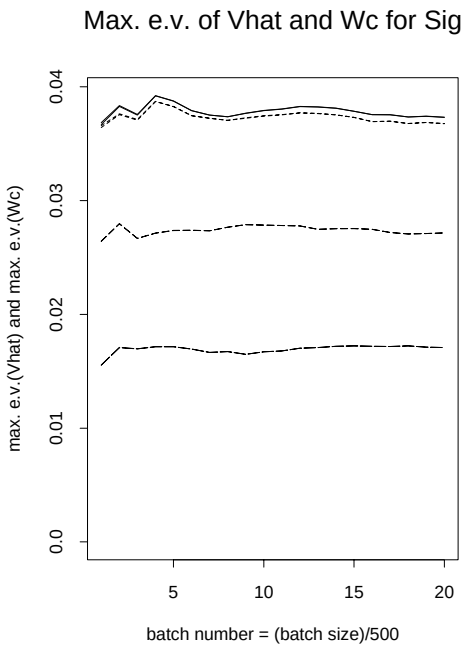


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

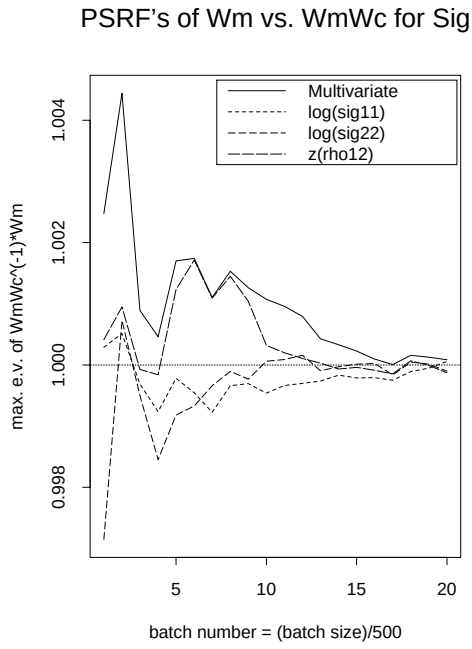
Figure E.2 (continued).



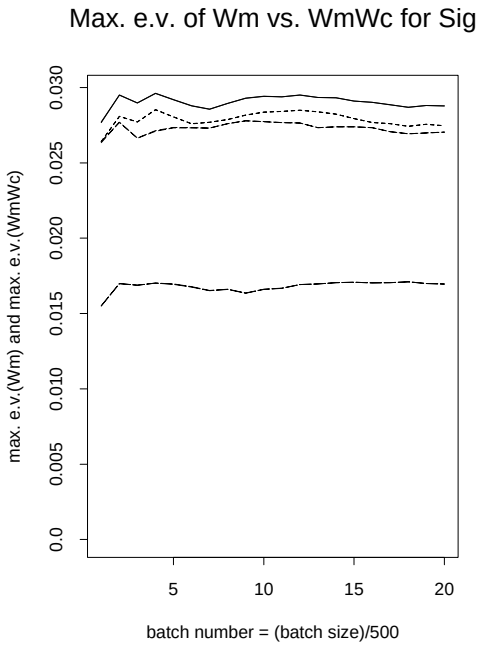
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

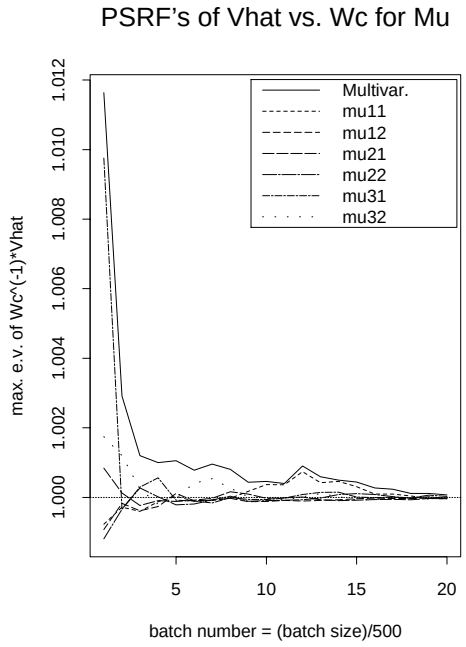


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

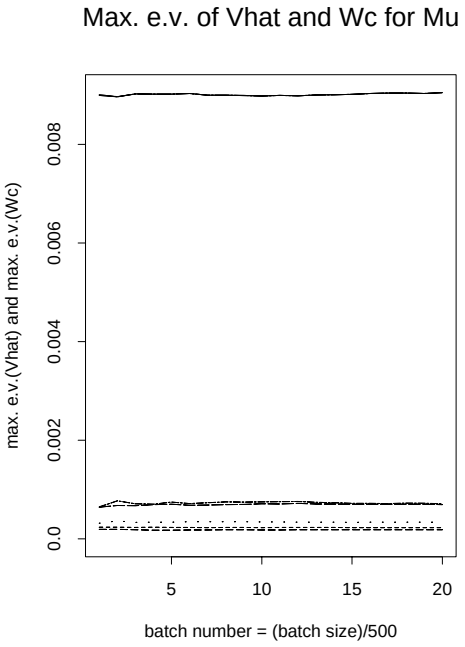


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

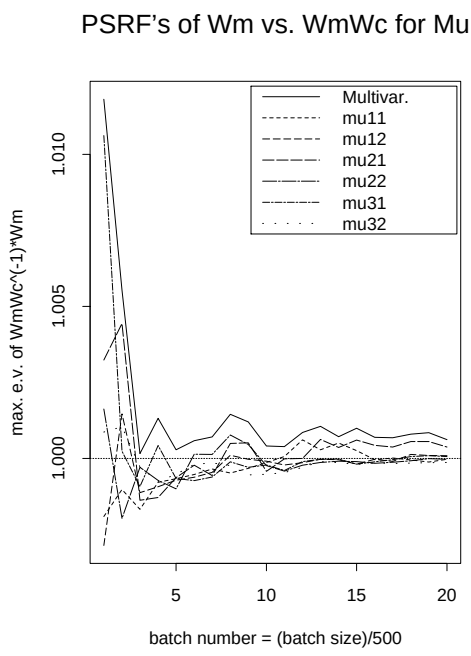
Figure E.3: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, I-k7-b.



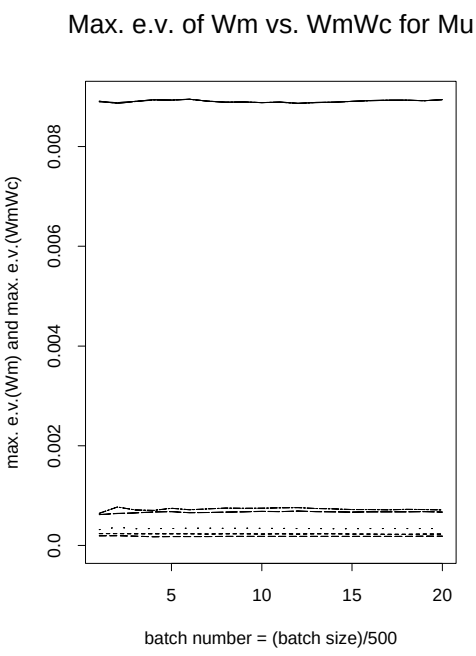
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

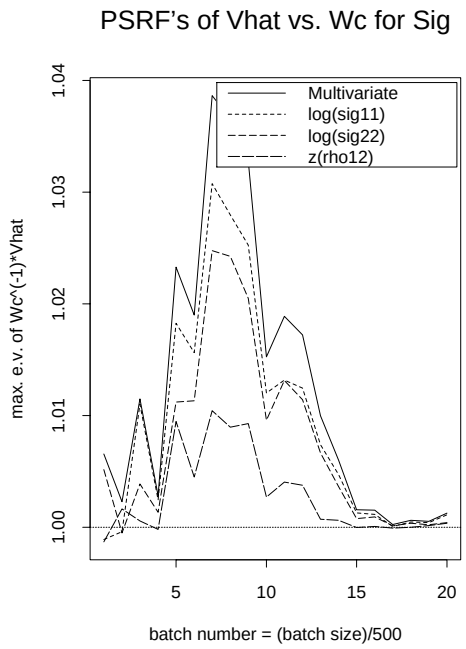


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c

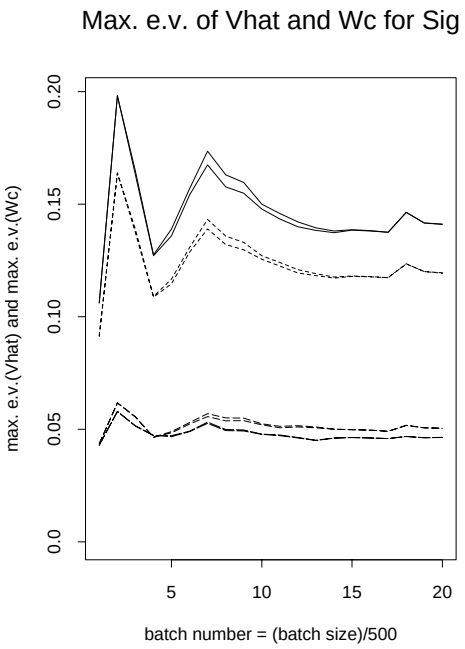


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

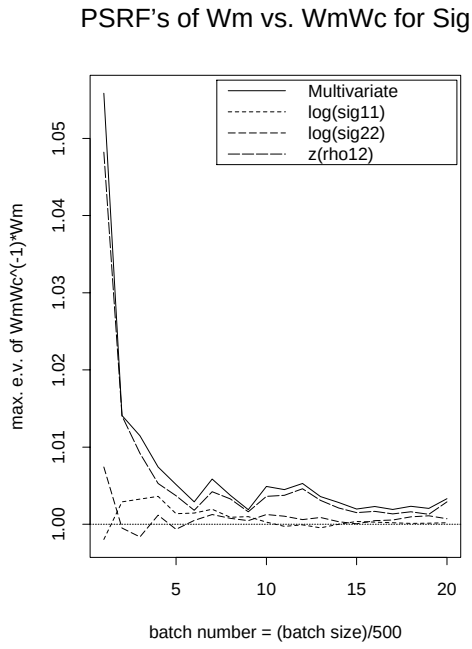
Figure E.3 (continued).



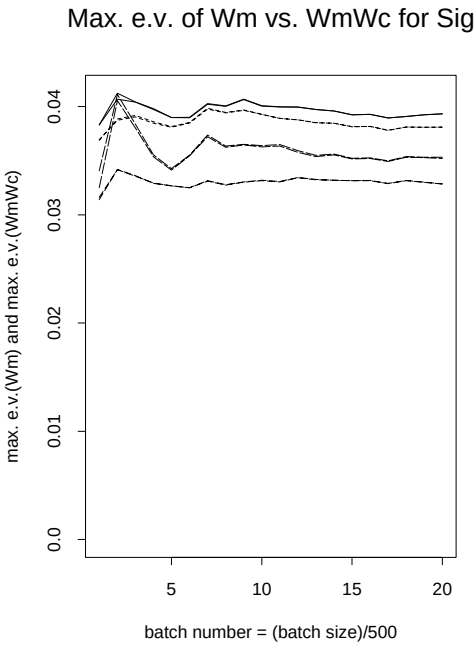
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

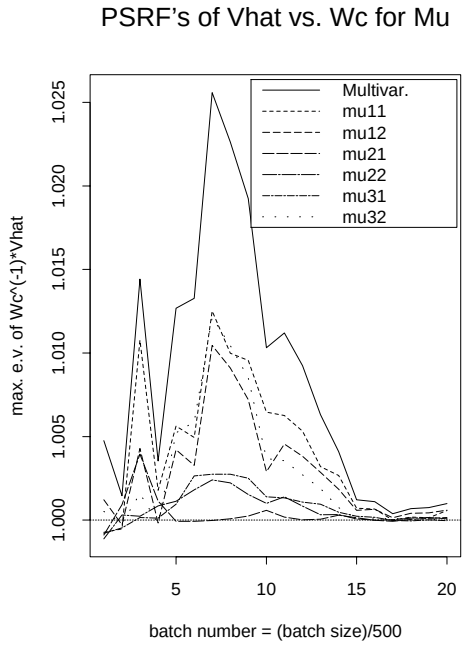


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

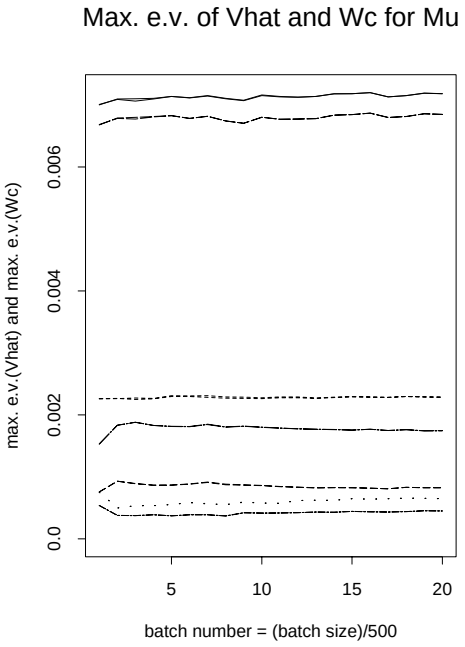


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

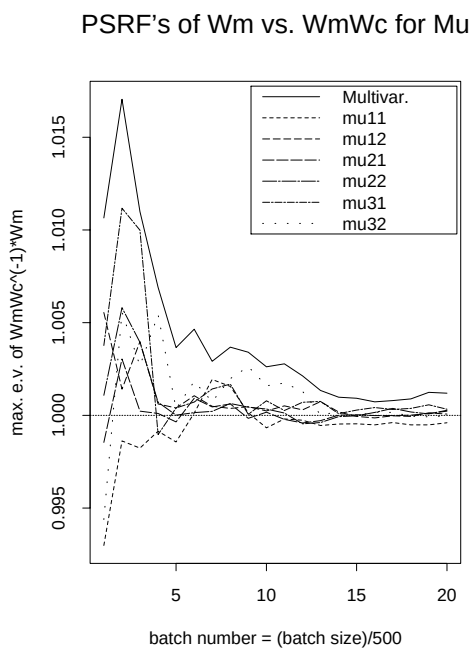
Figure E.4: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, I-k14-a.



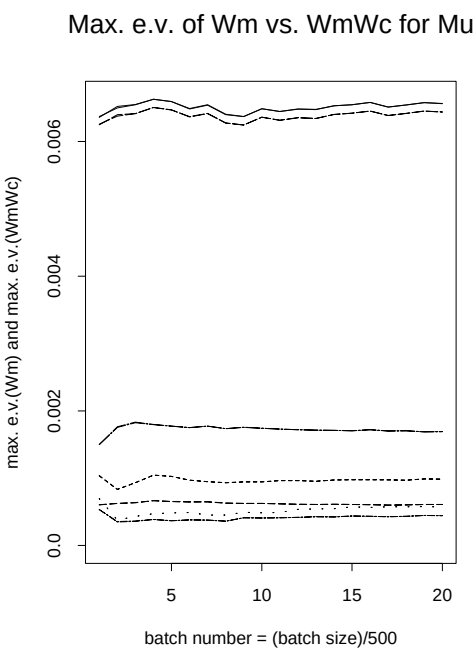
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

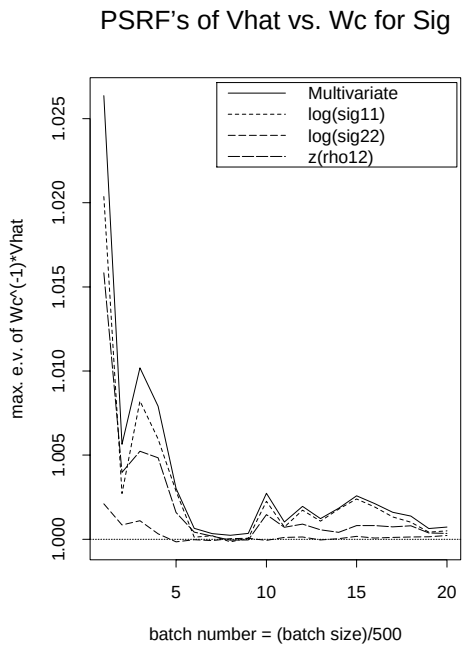


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c

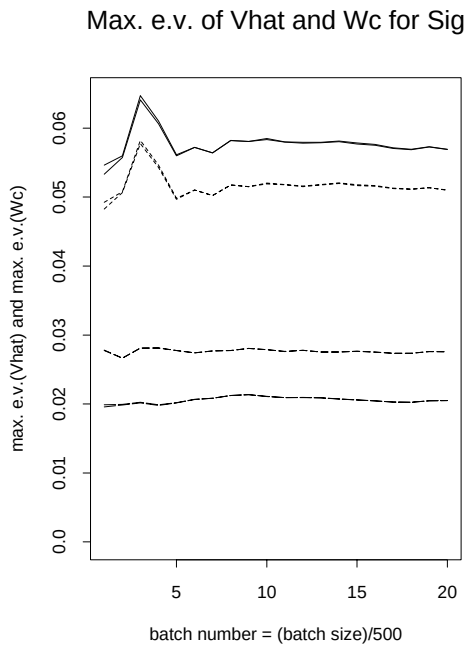


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

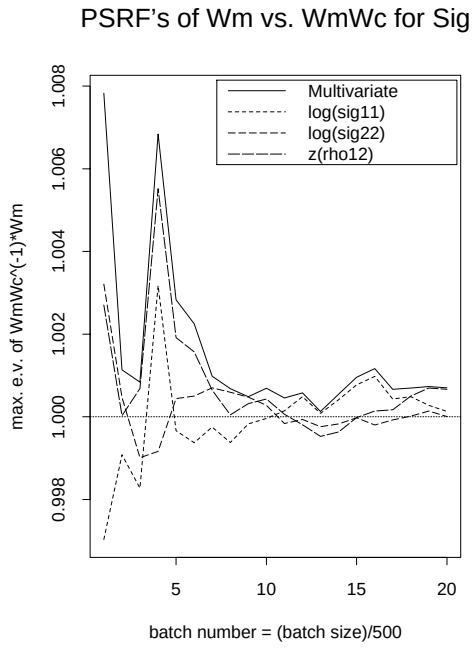
Figure E.4 (continued).



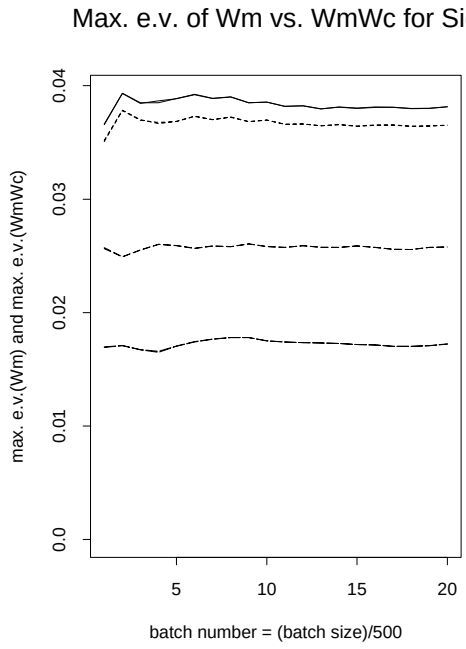
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

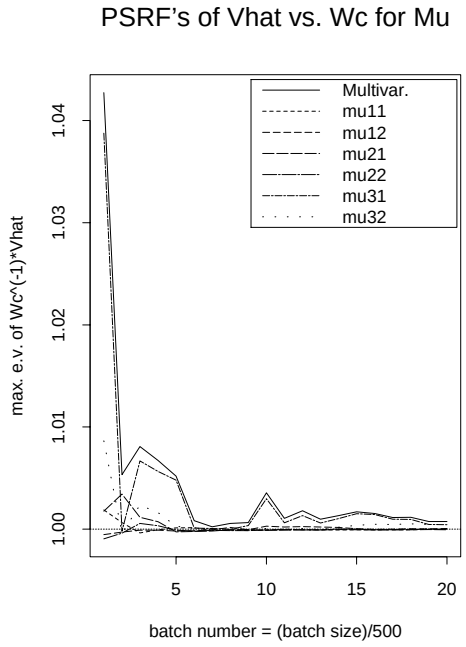


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

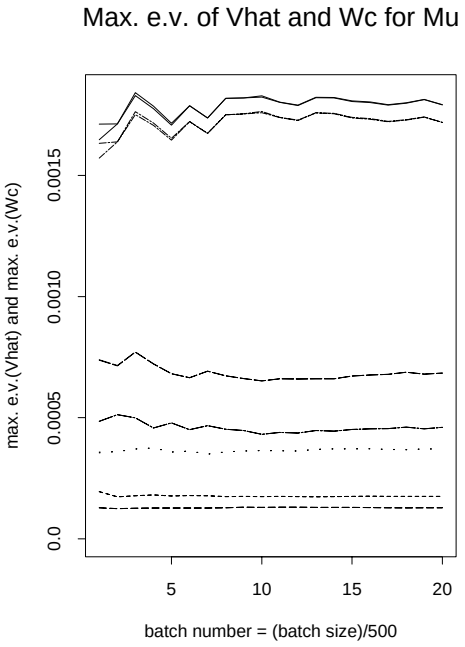


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

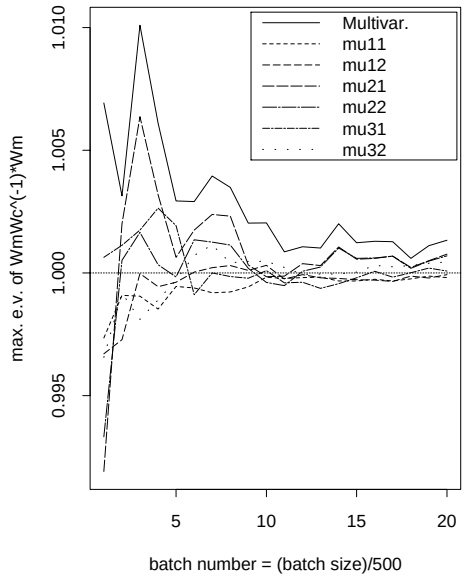
Figure E.5: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, I-k14-b.



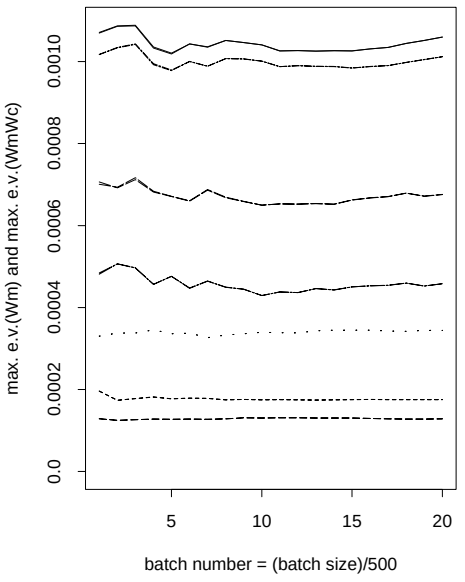
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

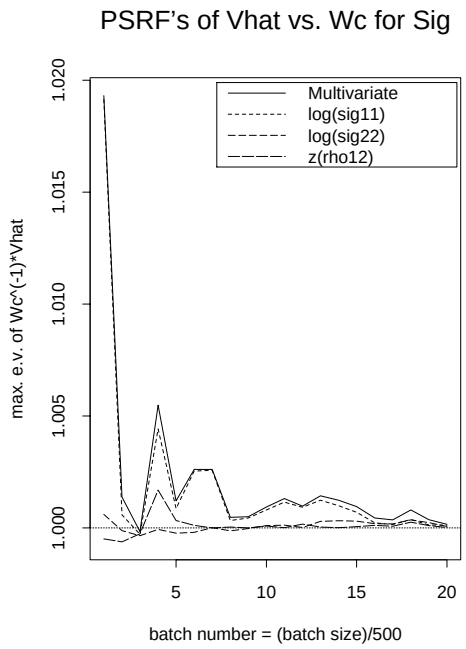


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c

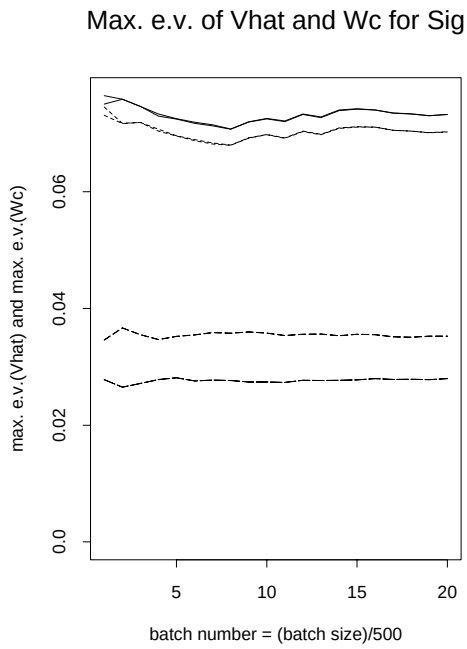


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

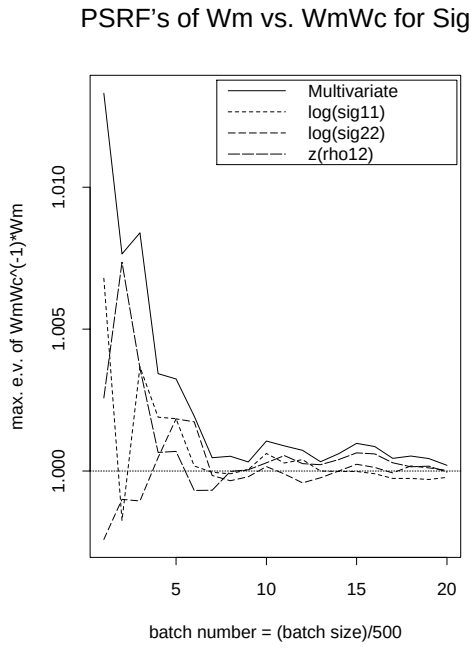
Figure E.5 (continued).



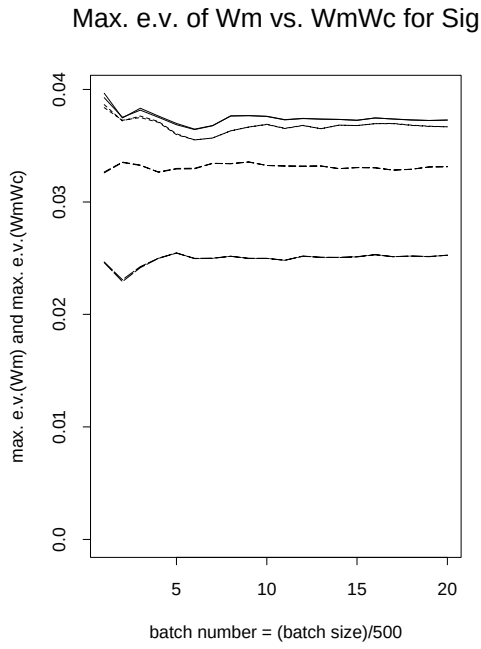
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

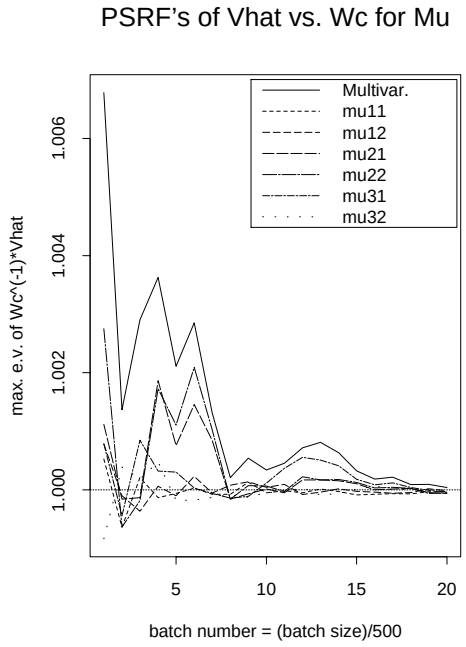


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

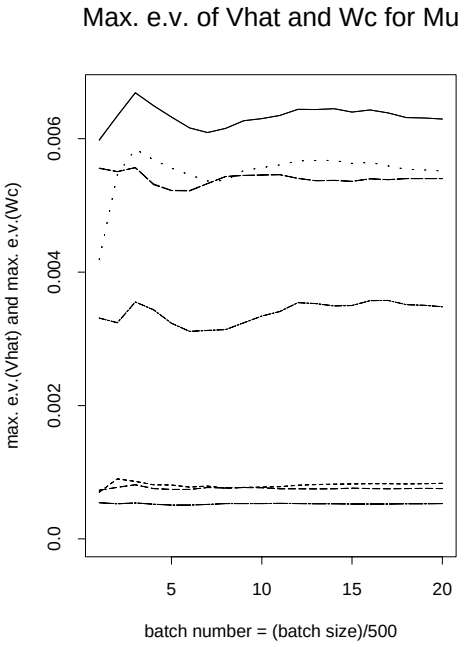


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

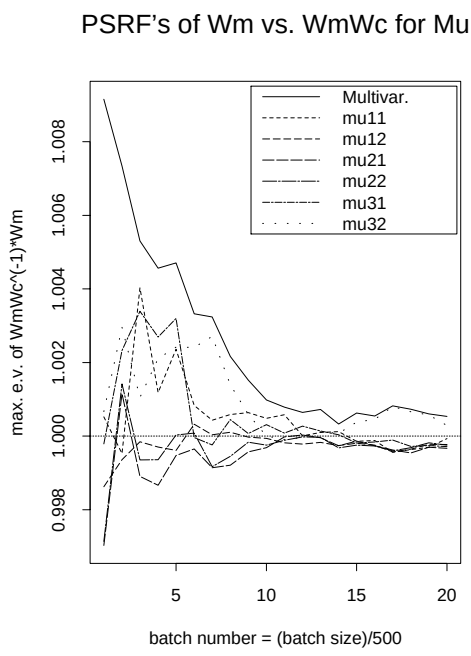
Figure E.6: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-1.5-k7-a.



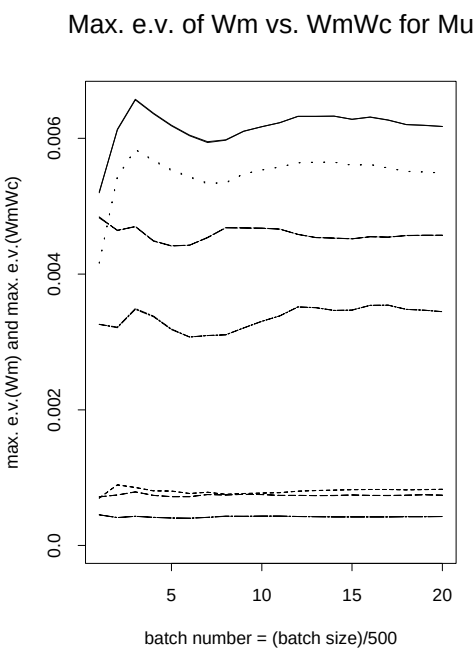
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

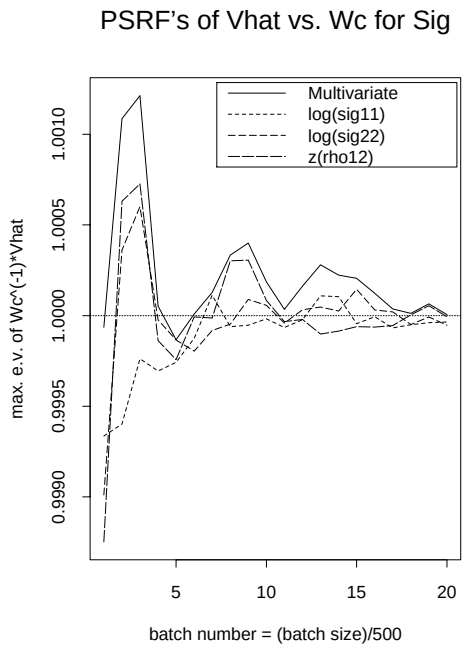


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c

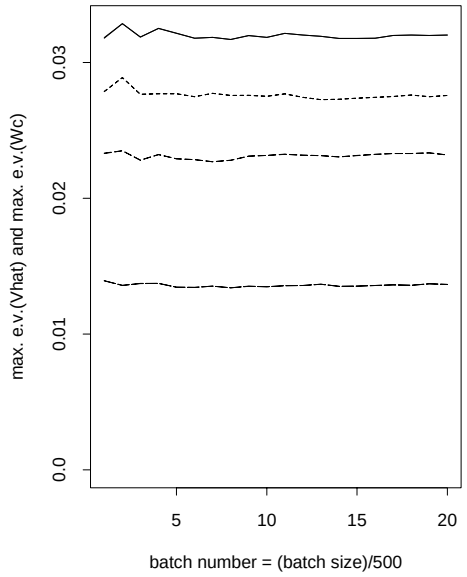


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

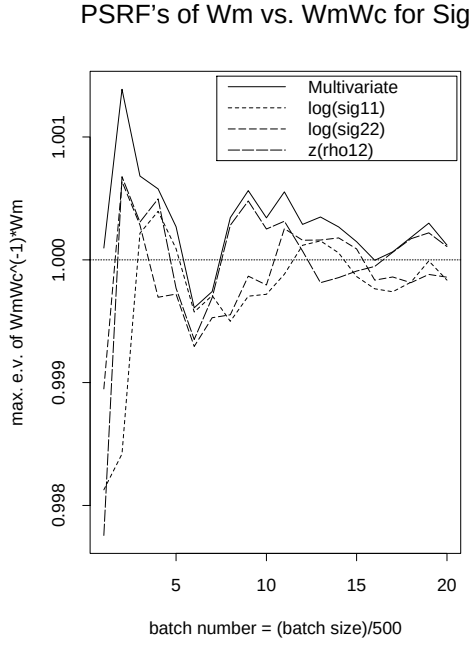
Figure E.6 (continued).



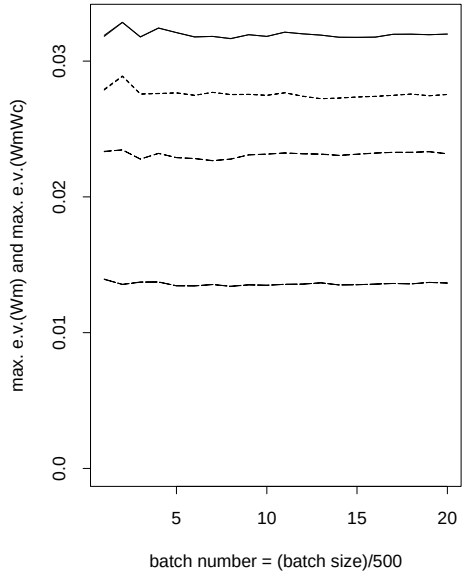
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

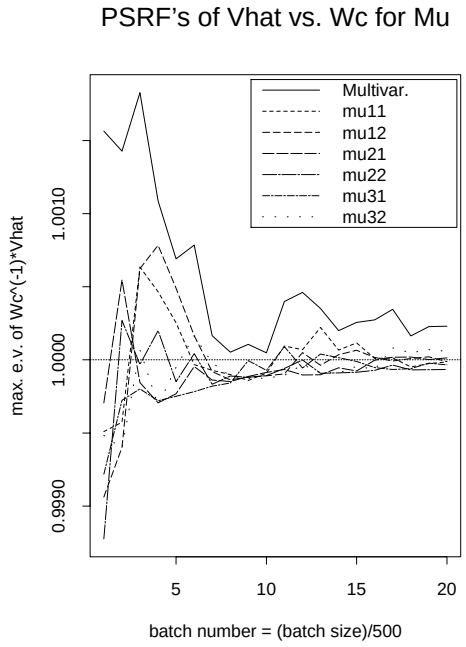


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. W_mW_c

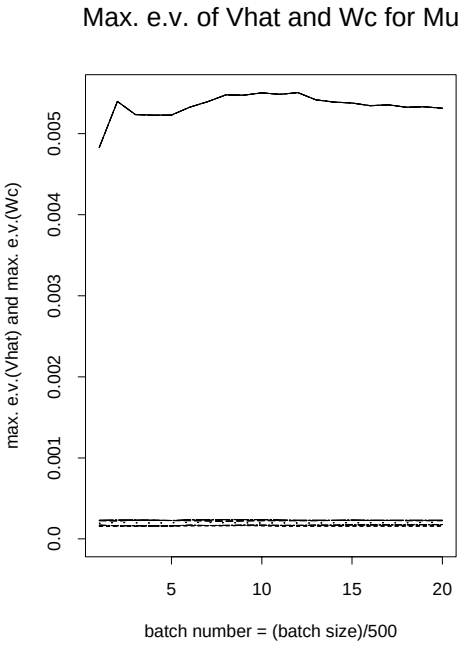


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. W_mW_c

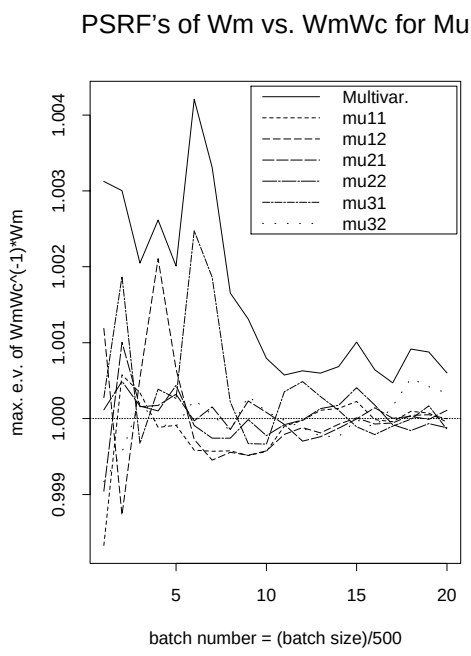
Figure E.7: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-1.5-k7-b.



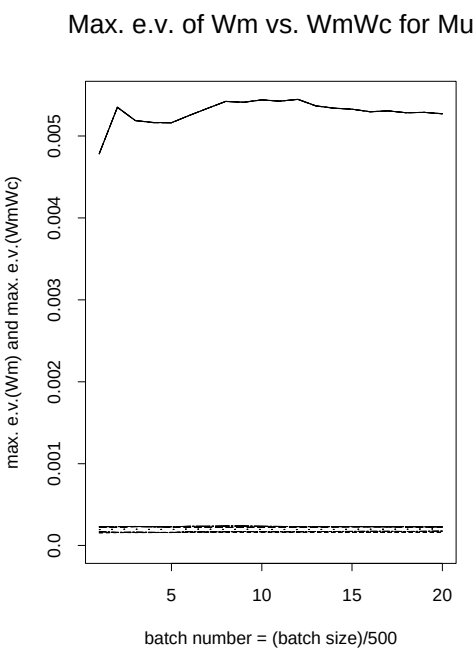
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

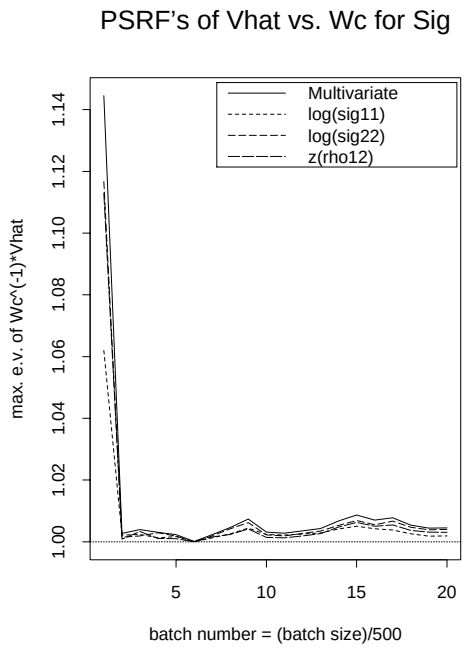


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

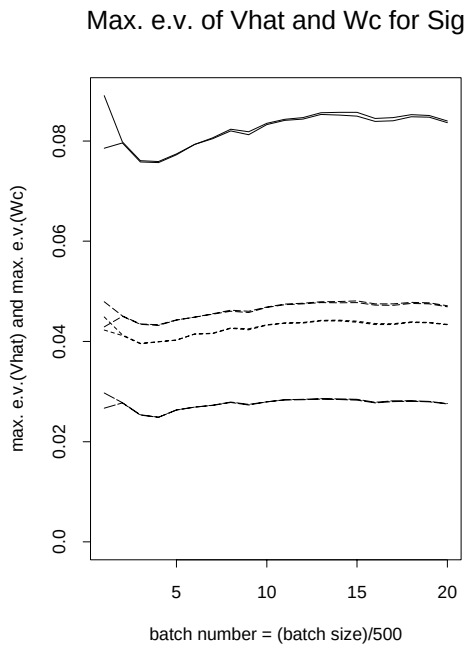


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

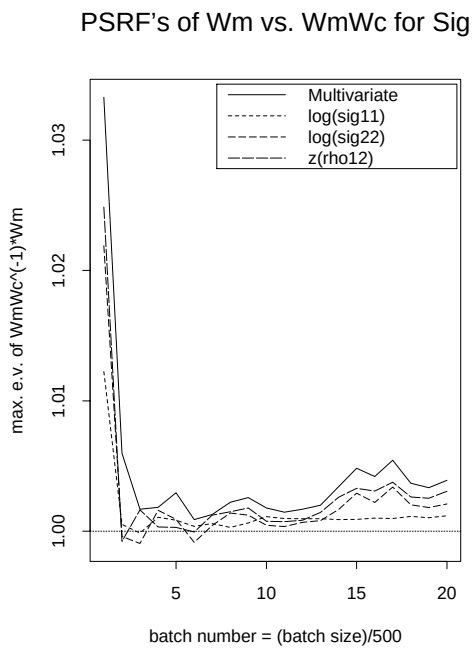
Figure E.7 (continued).



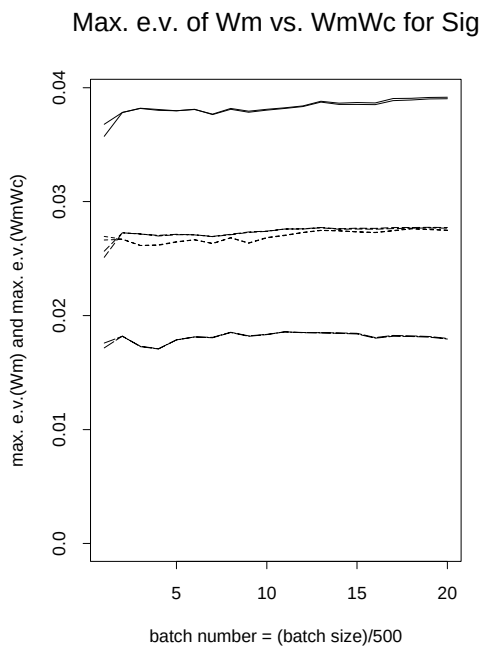
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

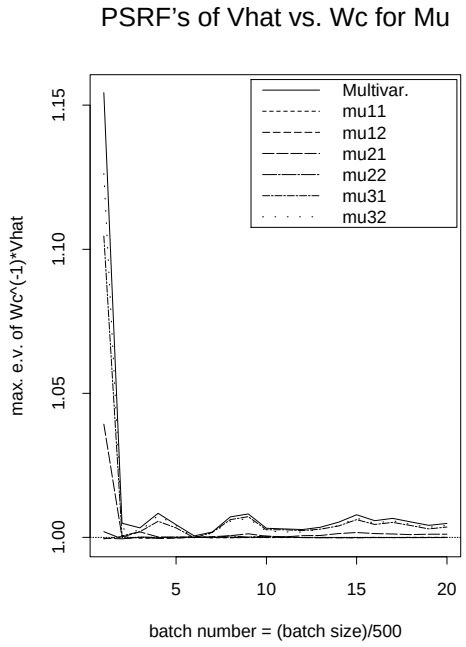


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

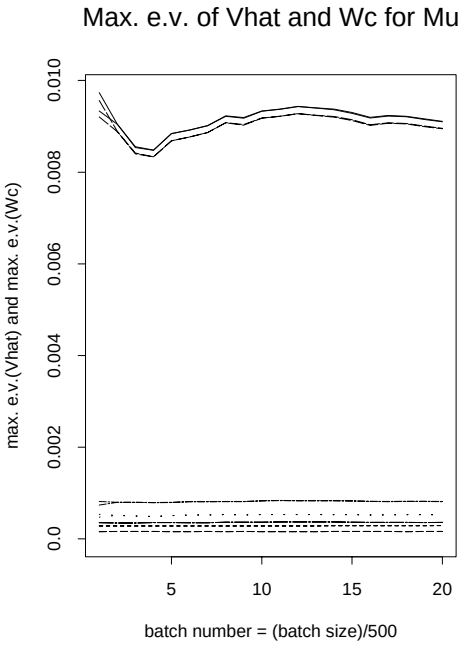


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

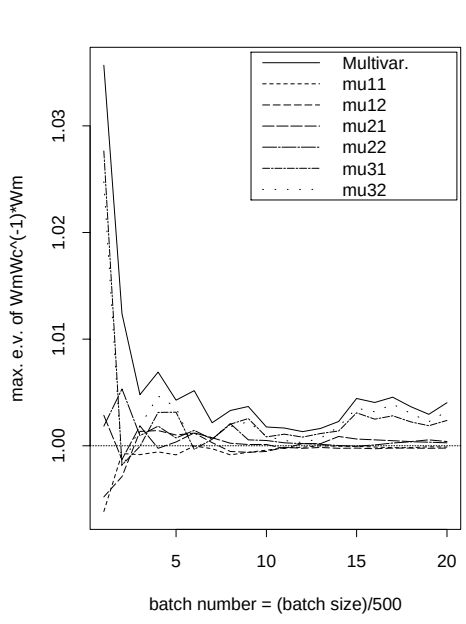
Figure E.8: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-1.5-k14-a.



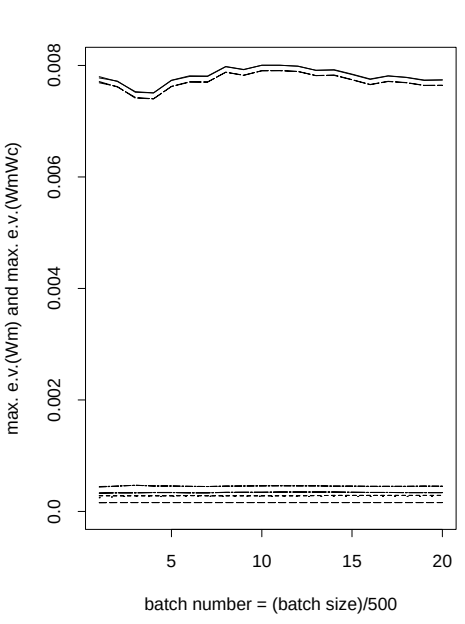
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

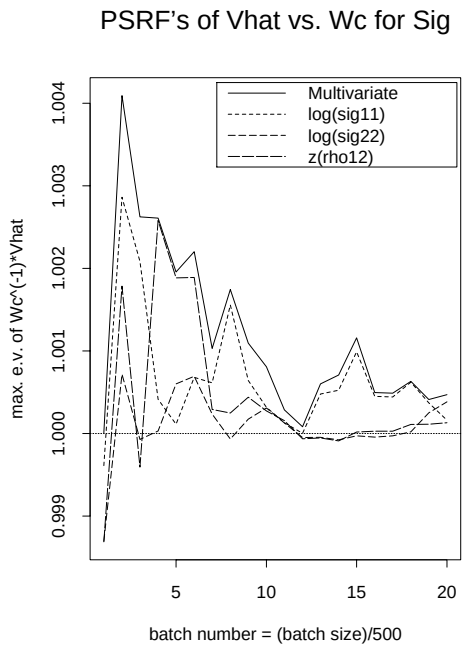


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c

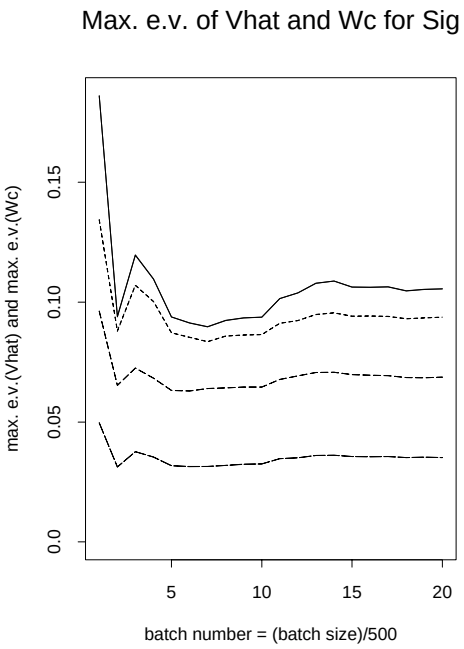


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

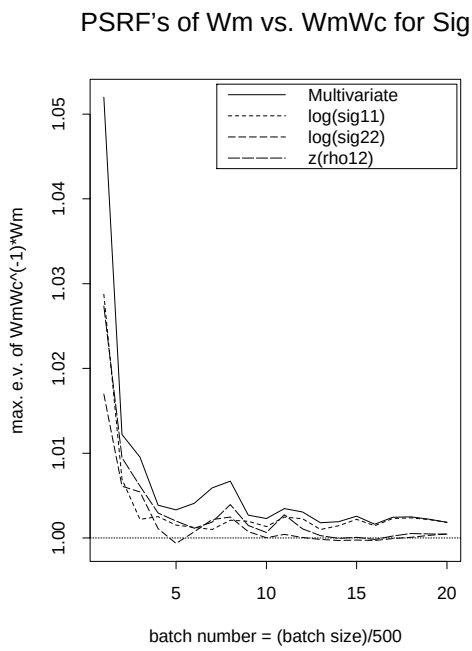
Figure E.8 (continued).



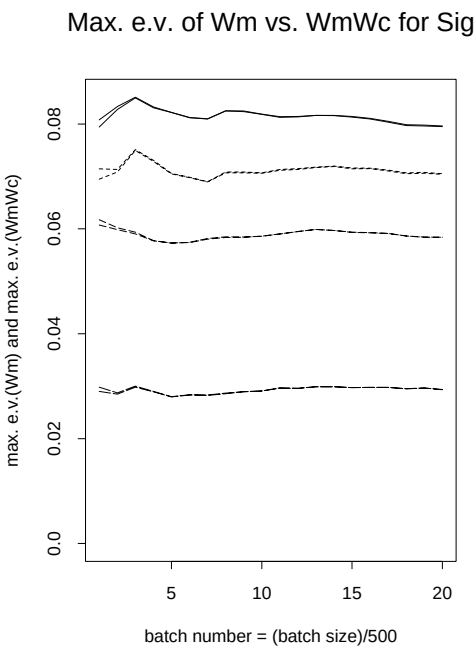
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

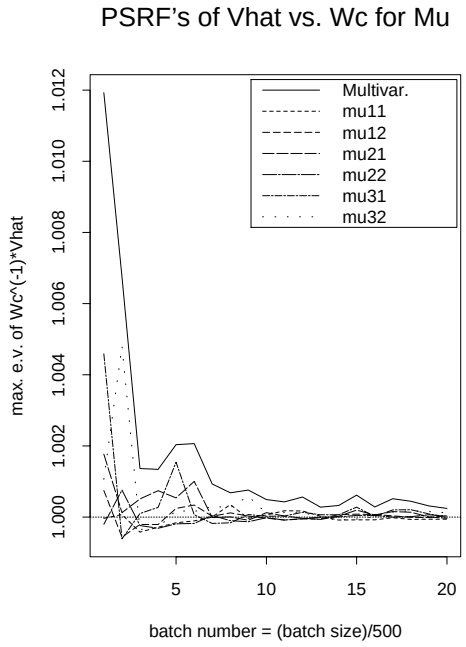


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

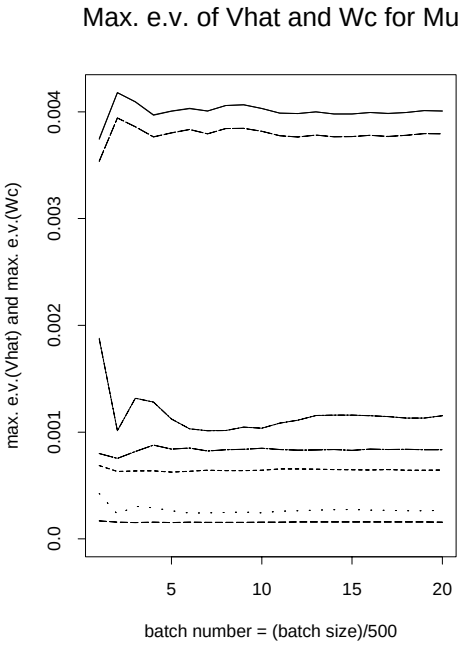


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

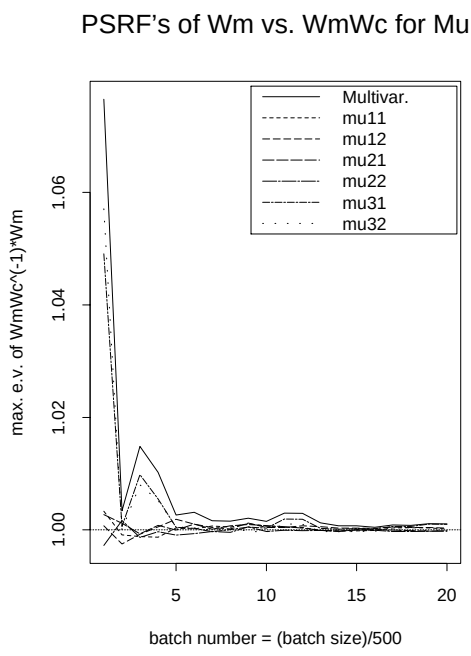
Figure E.9: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-1.5-k14-b.



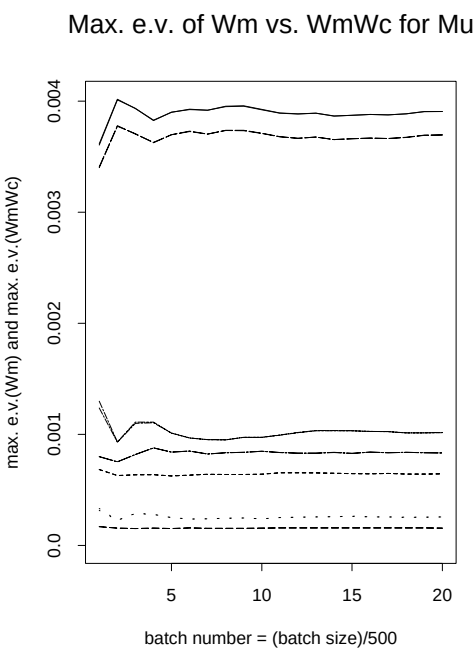
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

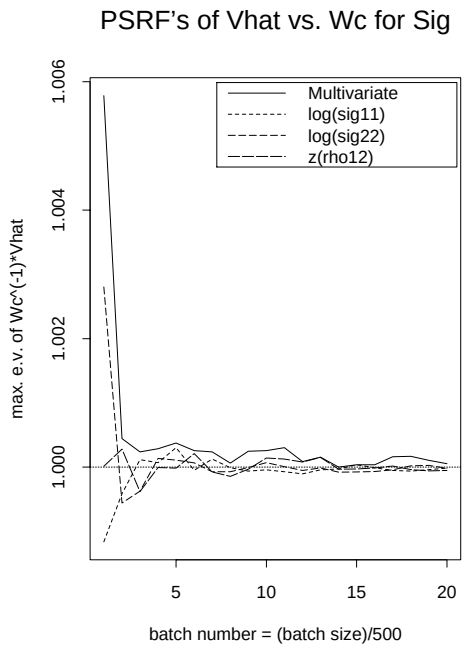


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

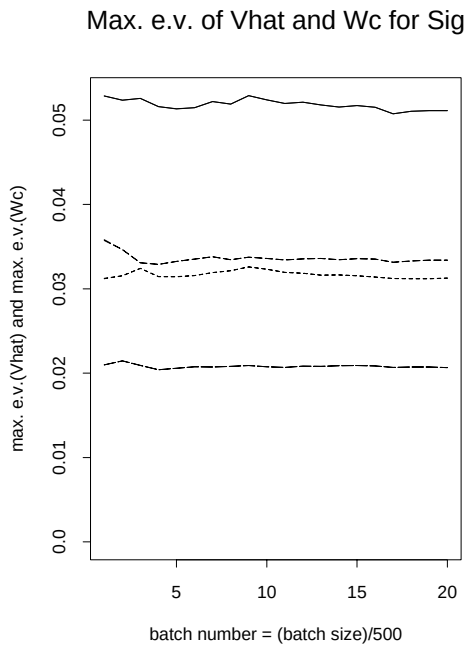


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

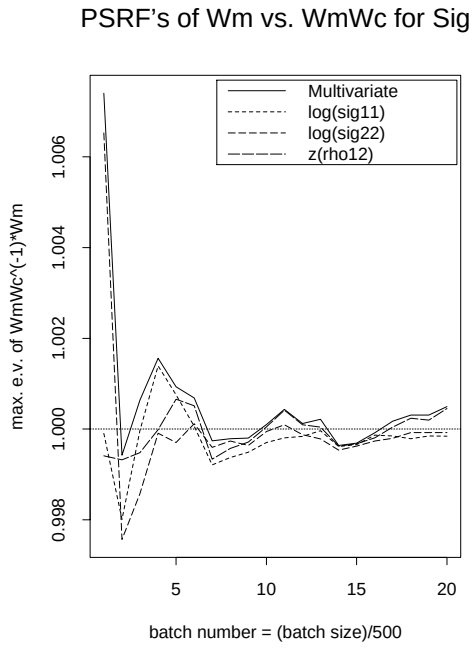
Figure E.9 (continued).



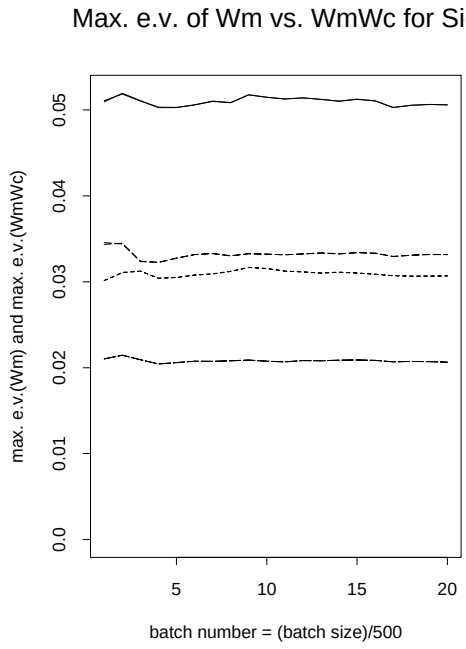
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

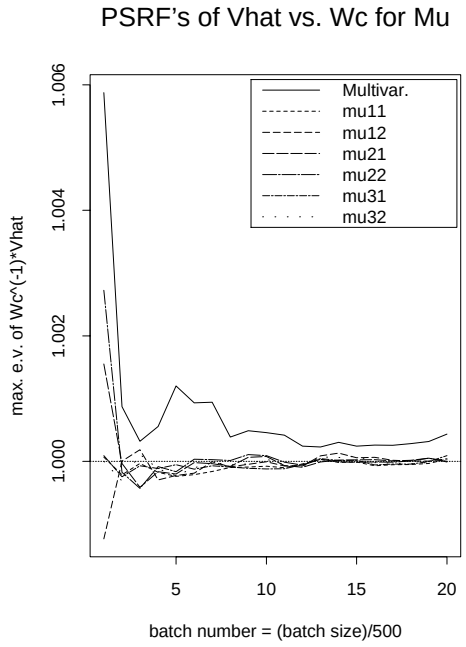


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

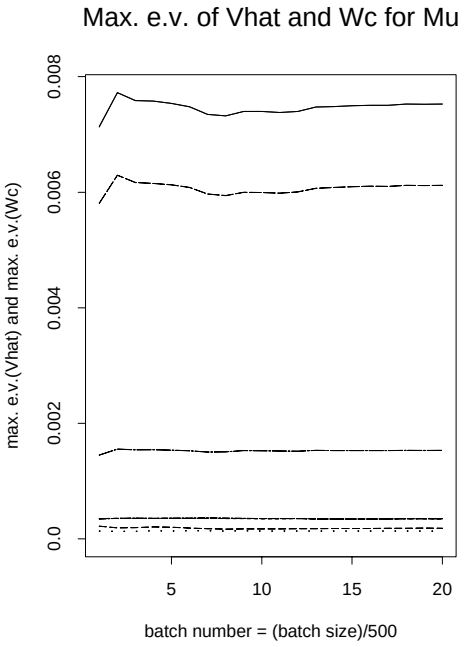


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

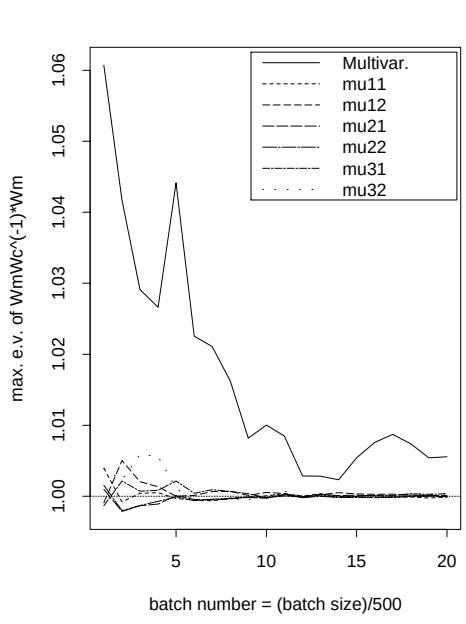
Figure E.10: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-3-k7-a.



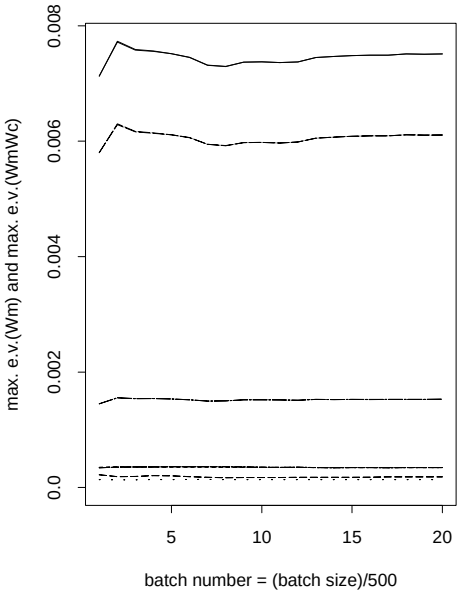
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

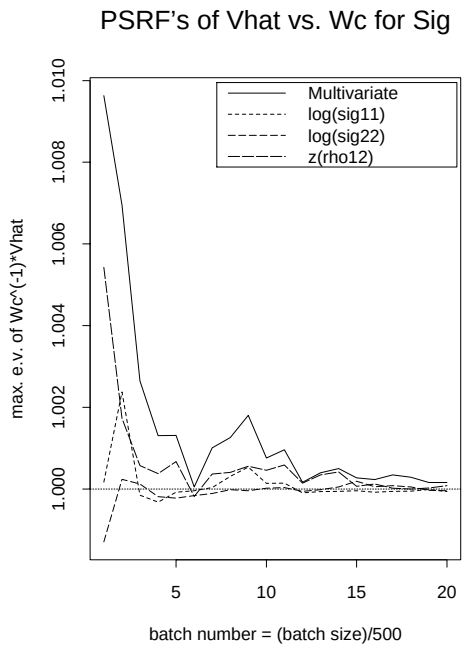


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c

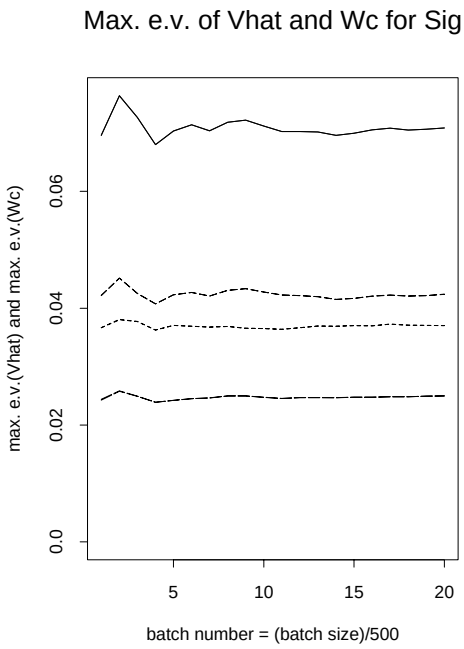


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

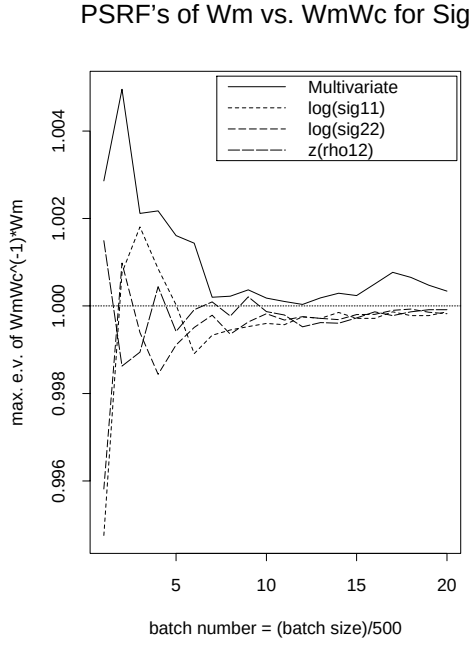
Figure E.10 (continued).



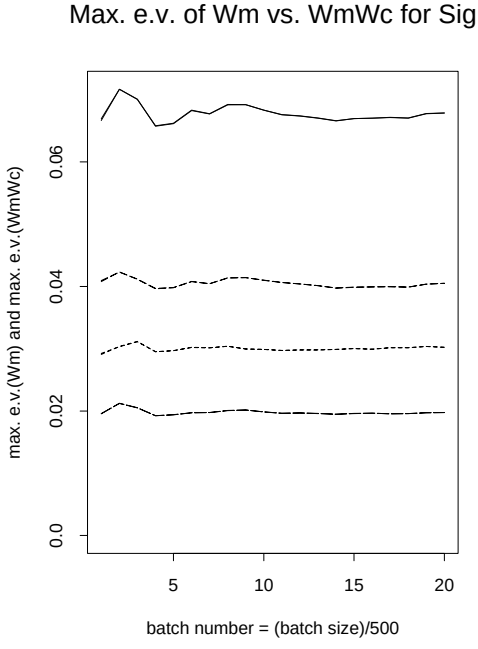
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

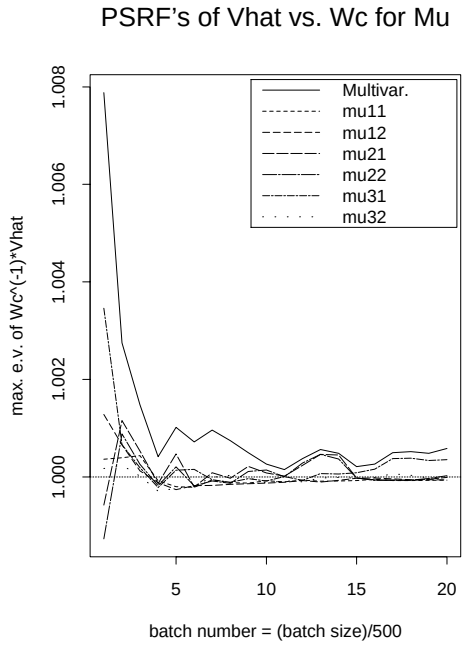


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

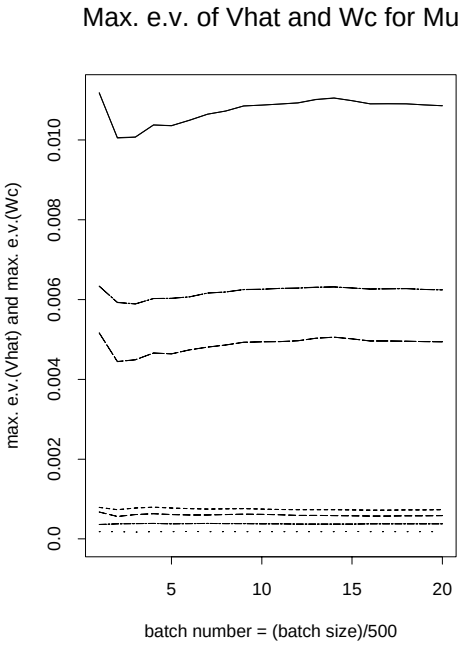


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

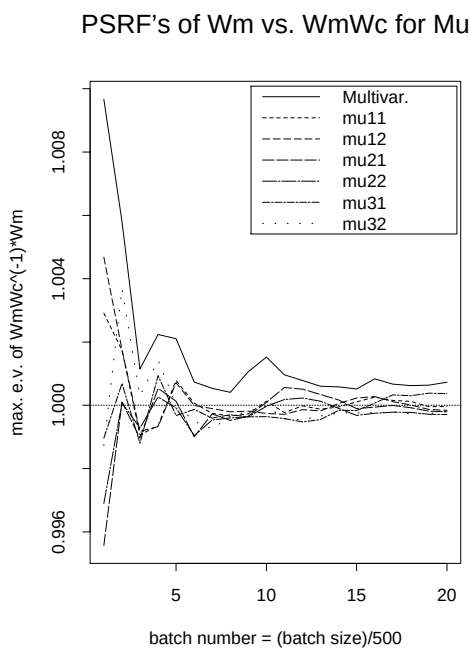
Figure E.11: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-3-k7-b.



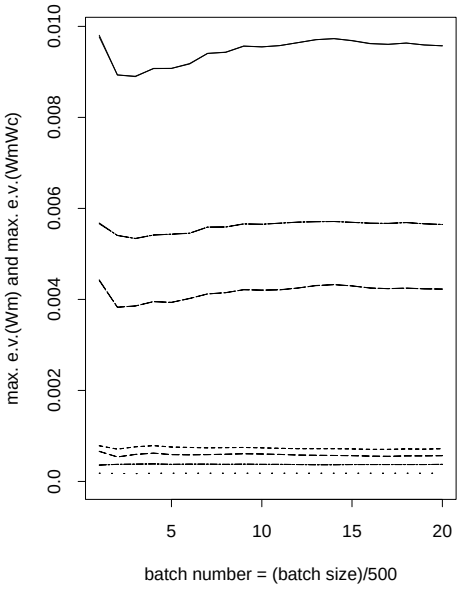
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

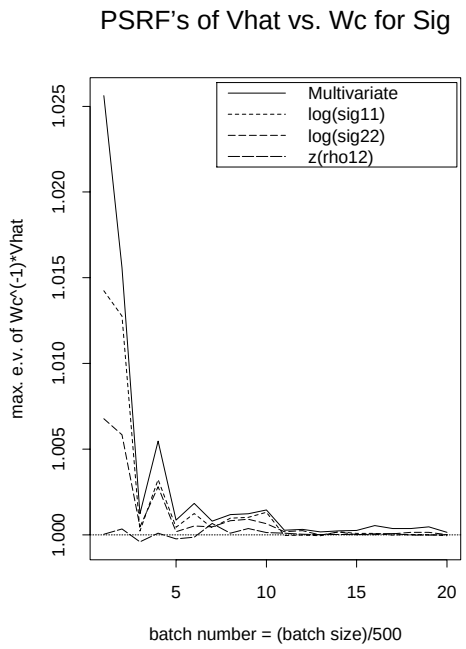


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

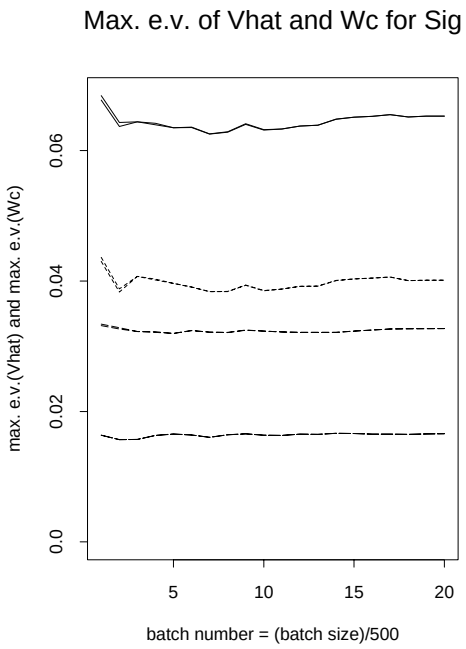


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

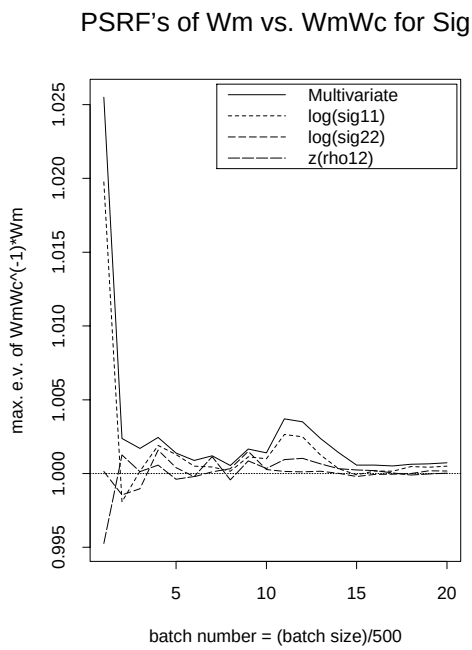
Figure E.11 (continued).



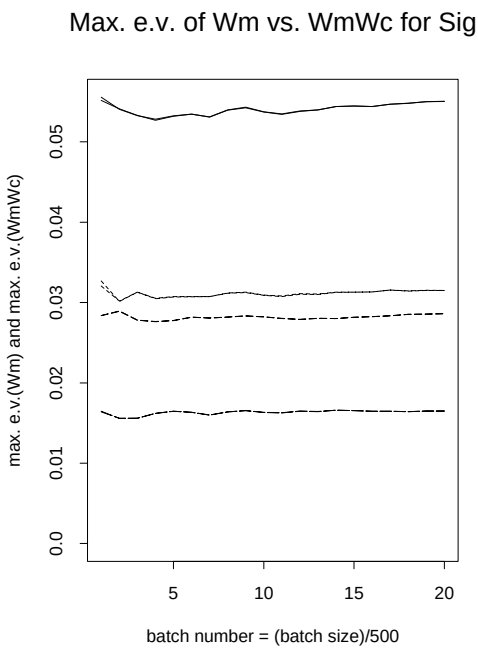
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

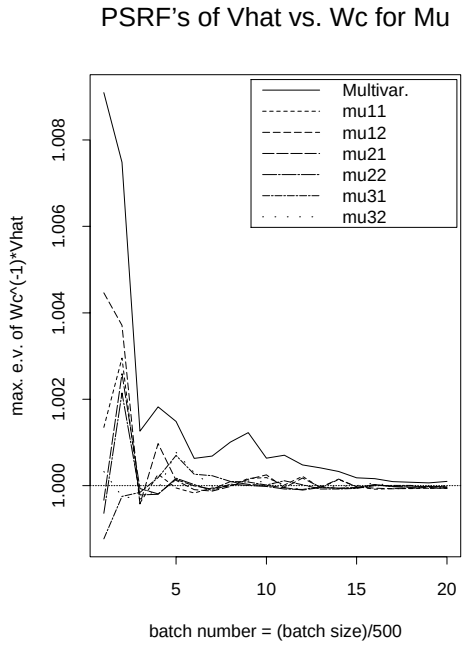


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

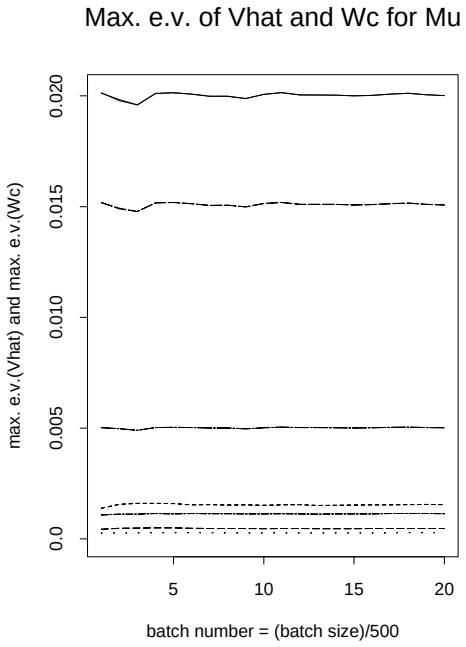


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

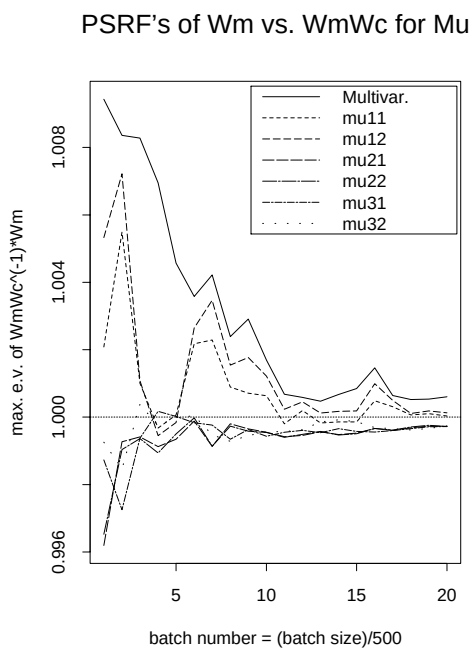
Figure E.12: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-3-k14-a.



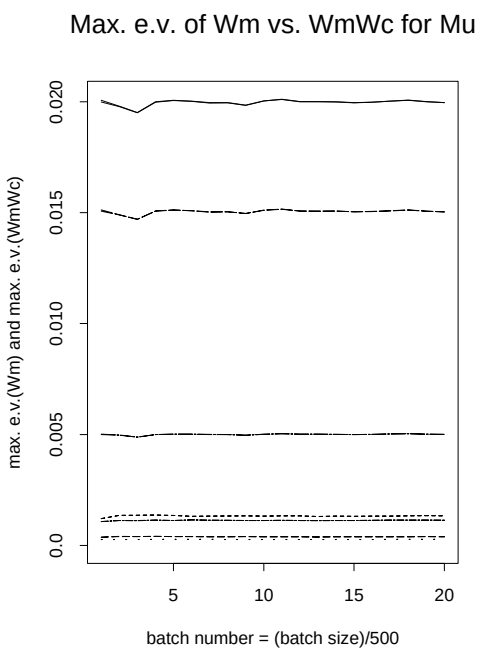
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c

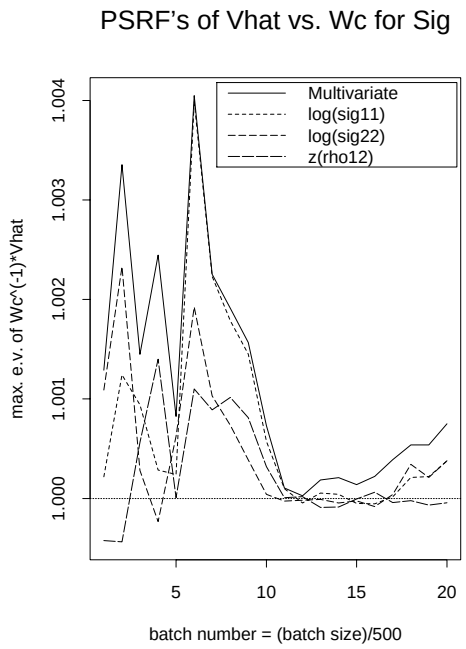


(g) PSRF: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

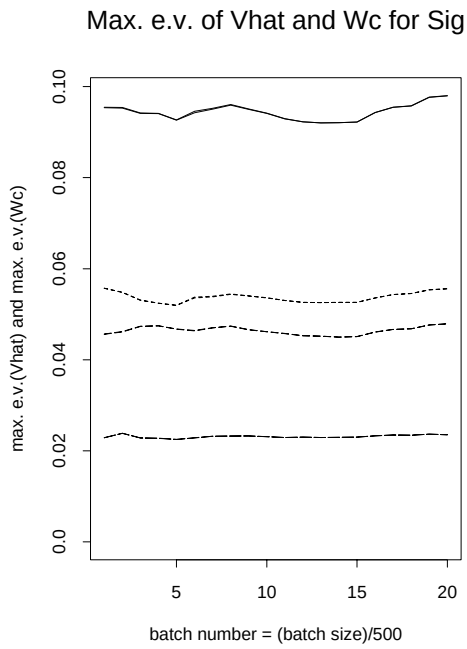


(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. $W_m W_c$

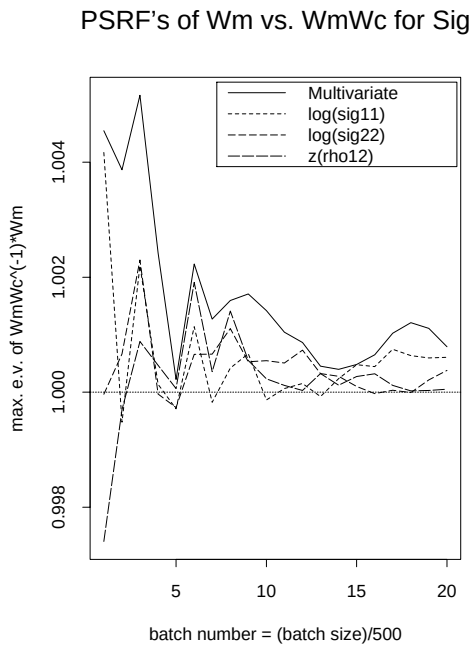
Figure E.12 (continued).



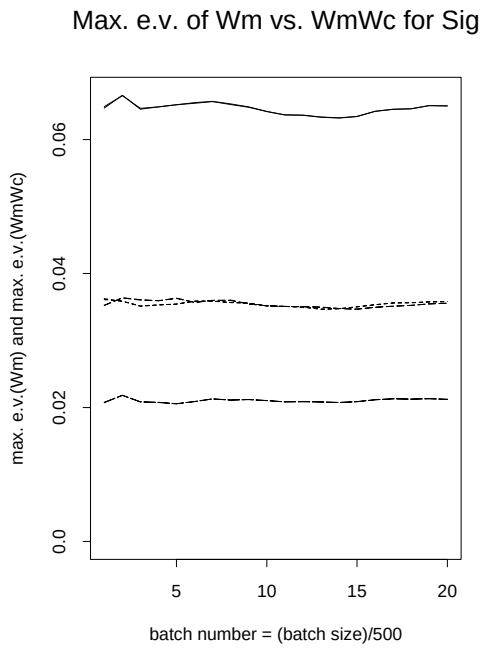
(a) PSRF: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c



(b) Max. e.v.: $\boldsymbol{\Sigma}$, $\widehat{V}$ vs. W_c

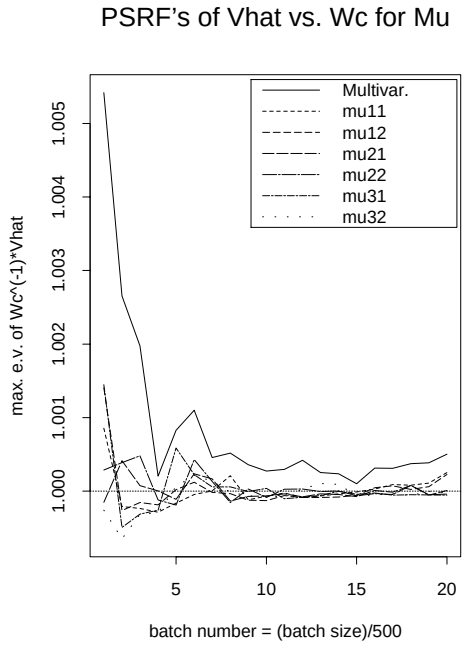


(c) PSRF: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

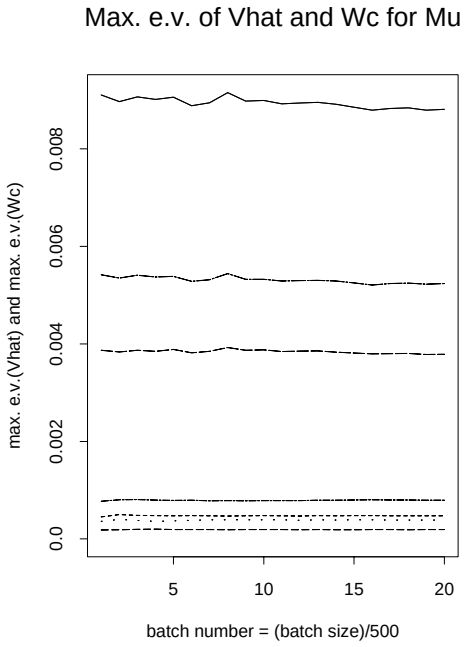


(d) Max. e.v.: $\boldsymbol{\Sigma}$, W_m vs. $W_m W_c$

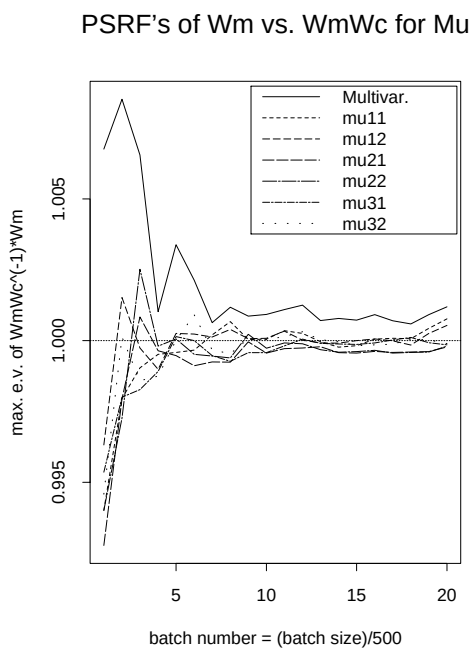
Figure E.13: Potential scale reduction factor and maximum eigenvalue plots for $(\log \sigma_{11}, \log \sigma_{22}, z(\rho_{12}))$ and $(\mu_{j_1 1}, \mu_{j_1 2}, \mu_{j_2 1}, \mu_{j_2 2}, \mu_{j_3 1}, \mu_{j_3 2})$, AI-3-k14-b.



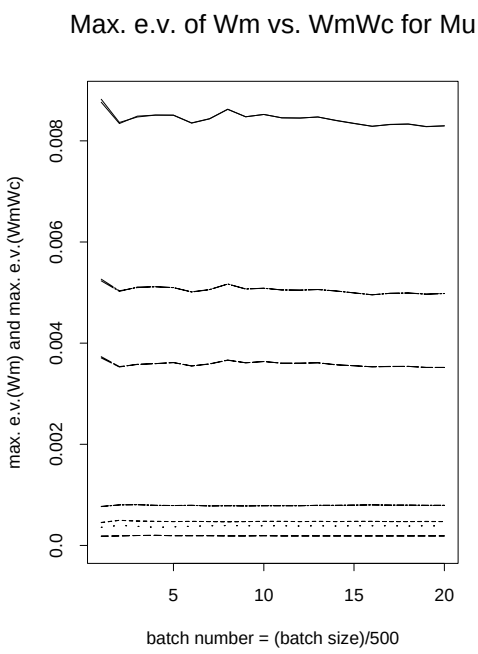
(e) PSRF: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(f) Max. e.v.: $\boldsymbol{\mu}$, $\widehat{V}$ vs. W_c



(g) PSRF: $\boldsymbol{\mu}$, W_m vs. W_mW_c



(h) Max. e.v.: $\boldsymbol{\mu}$, W_m vs. W_mW_c

Figure E.13 (continued).

APPENDIX F
HISTOGRAMS OF K = NUMBER OF CLUSTERS

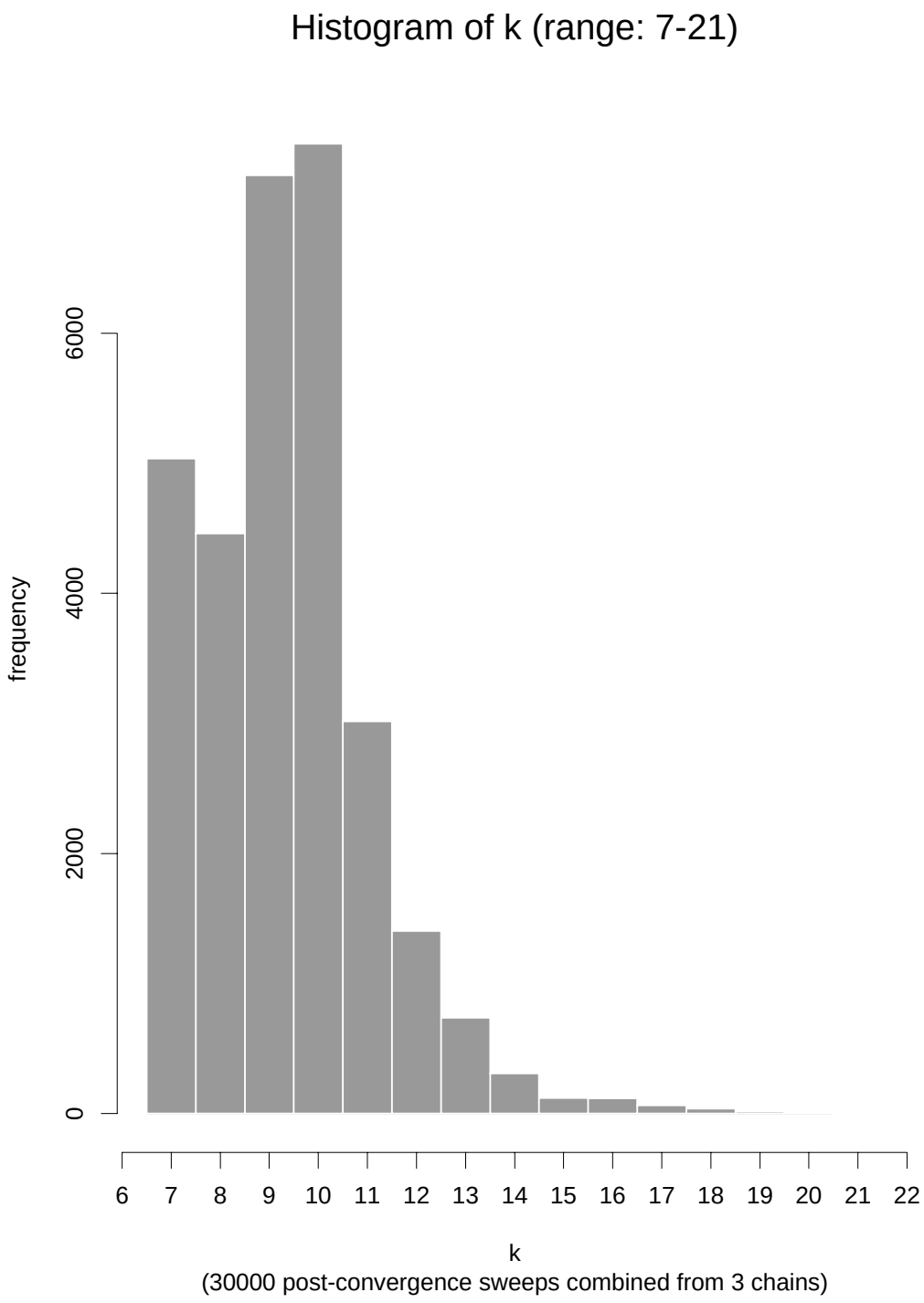
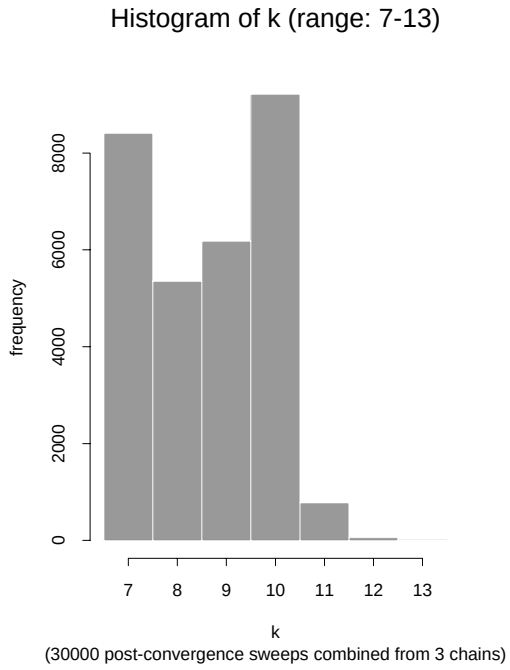
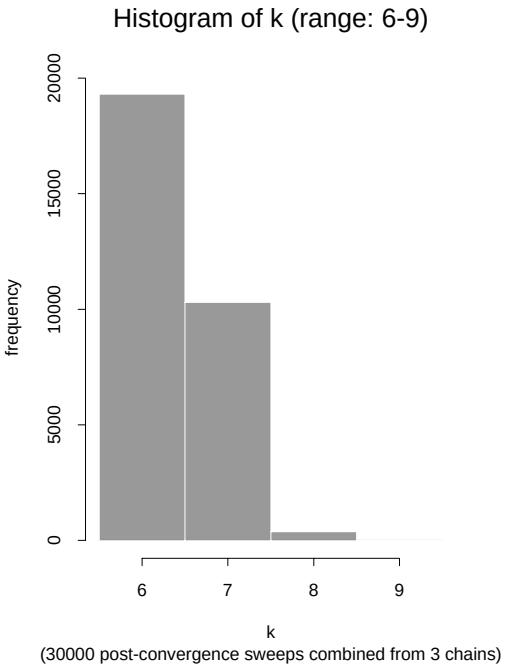


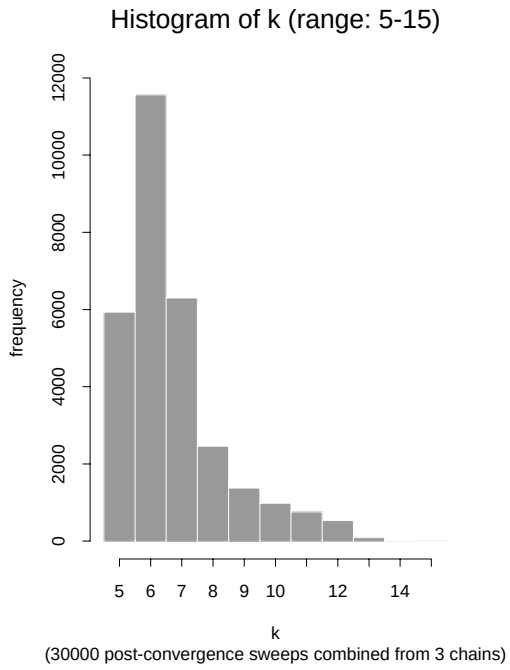
Figure F.1: Histogram of k in post-convergent RJMCMC sweeps, Redwood data.



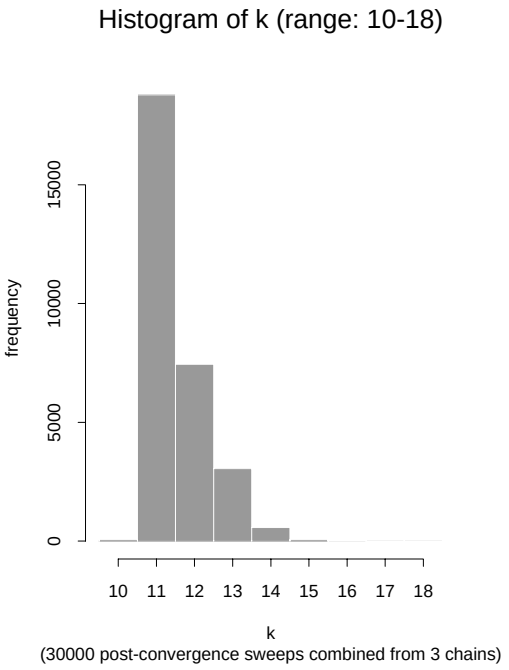
(a) I-k7-a



(b) I-k7-b

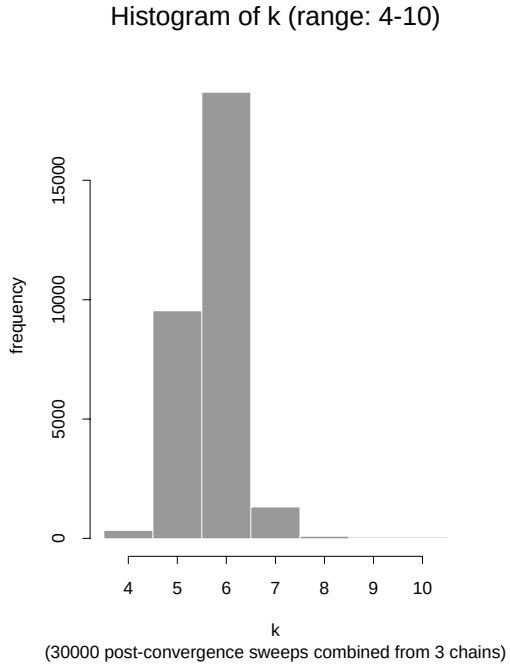


(c) I-k14-a

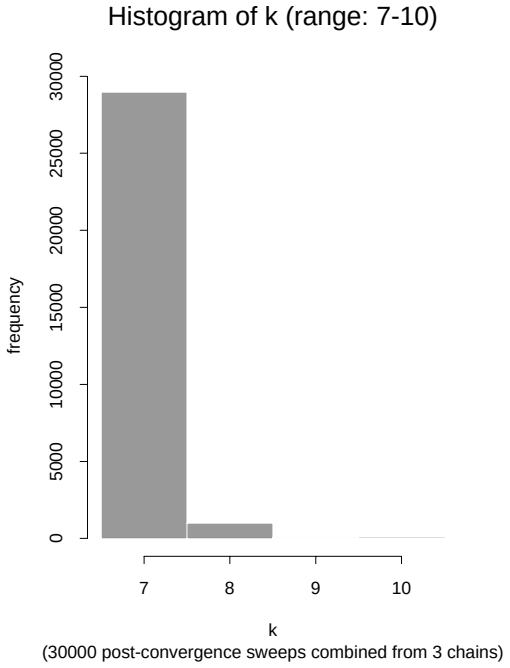


(d) I-k14-b

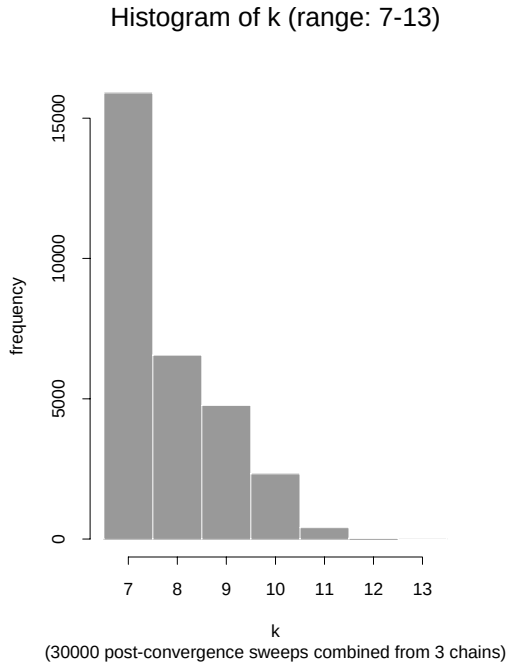
Figure F.2: Histogram of k in post-convergent RJMCMC sweeps, simulated patterns.



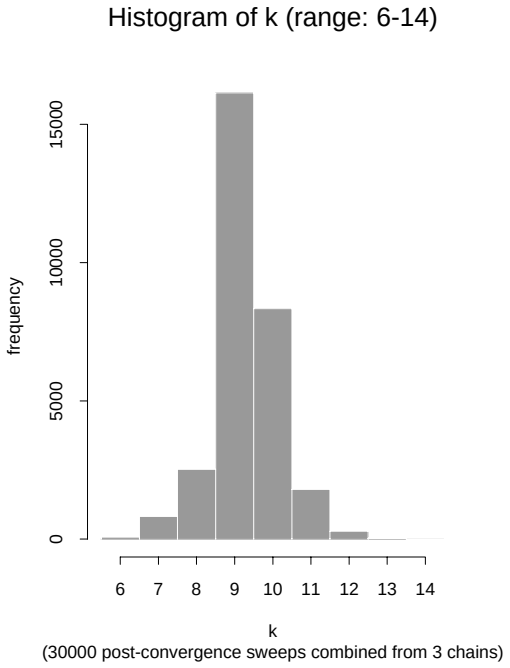
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

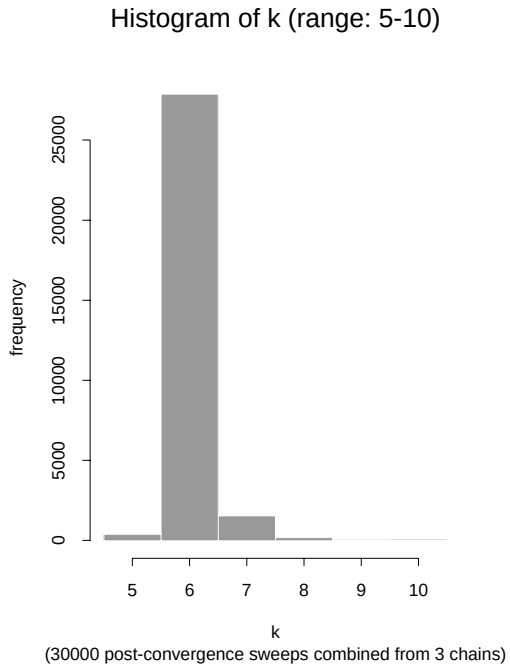


(g) AI-1.5-k14-a

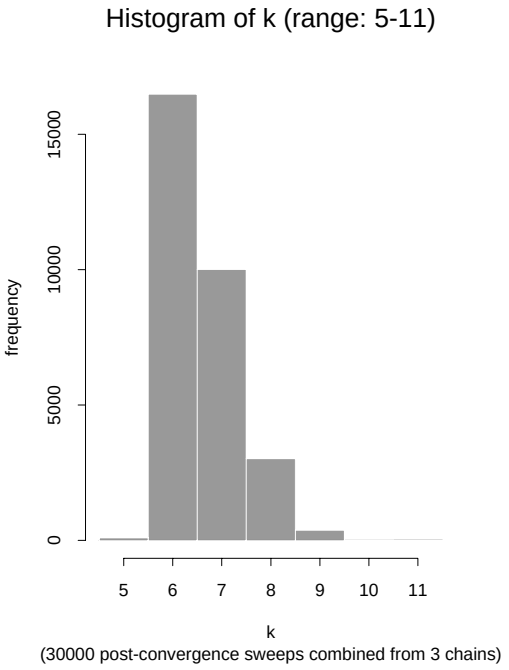


(h) AI-1.5-k14-b

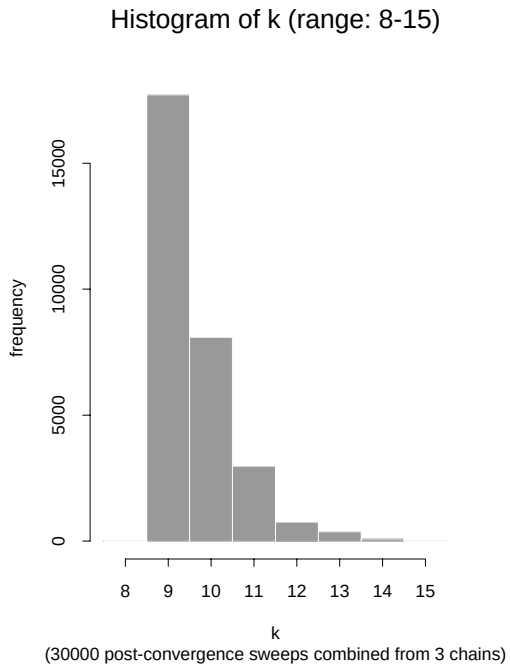
Figure F.2 (continued).



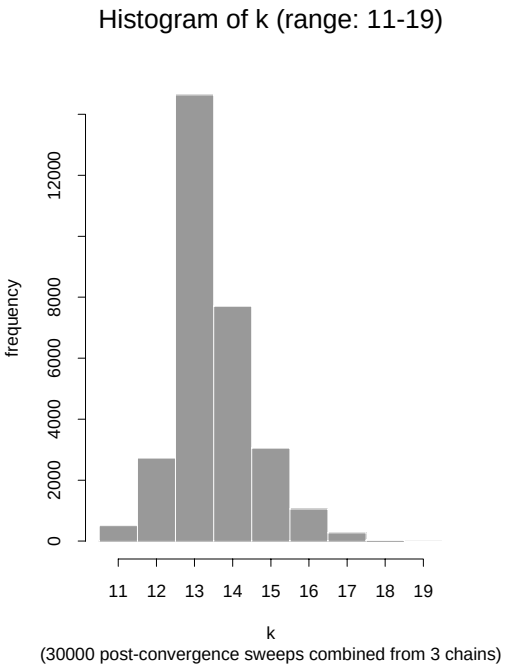
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure F.2 (continued).

APPENDIX G
P(K) ESTIMATES USING DIFFERENT METHODS

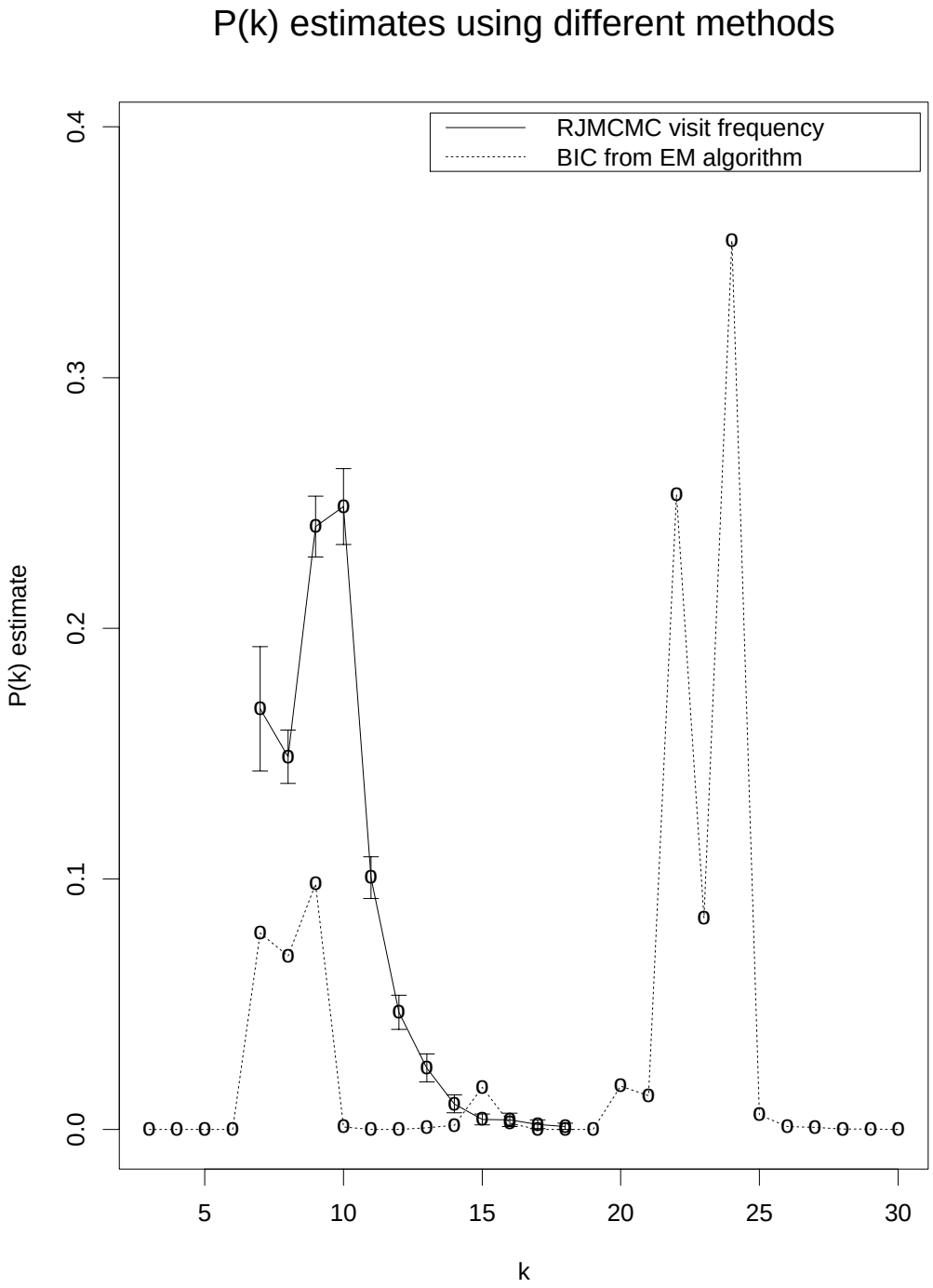
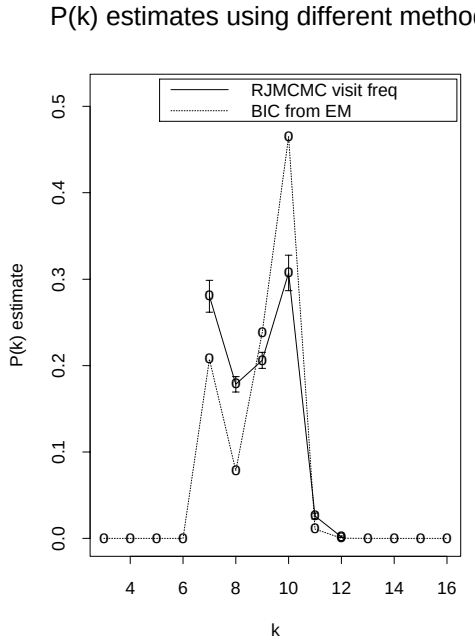
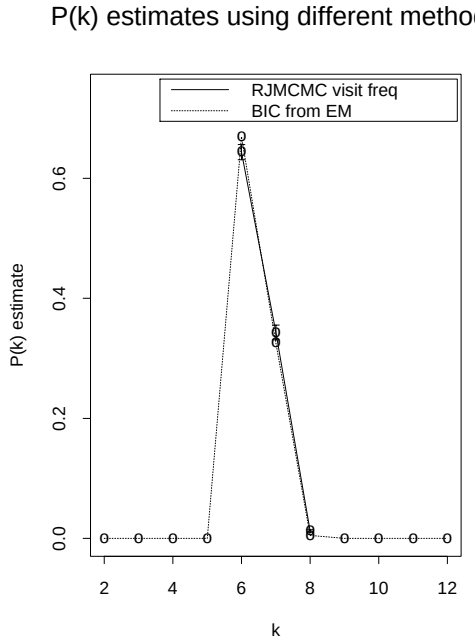


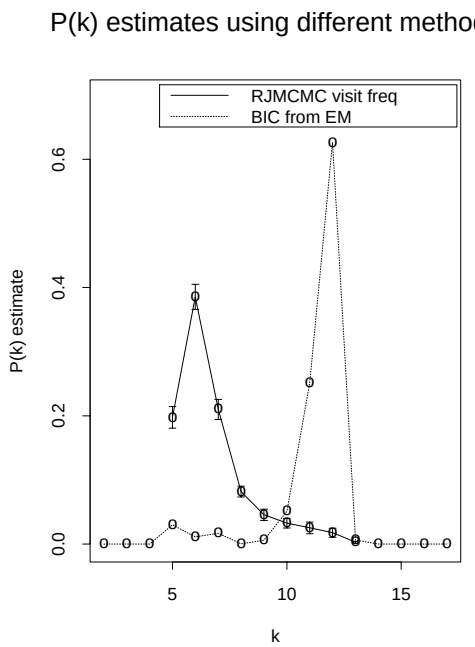
Figure G.1: $P(k)$ estimates using visit frequency from RJMCMC vs. BIC from EM, Redwood data.



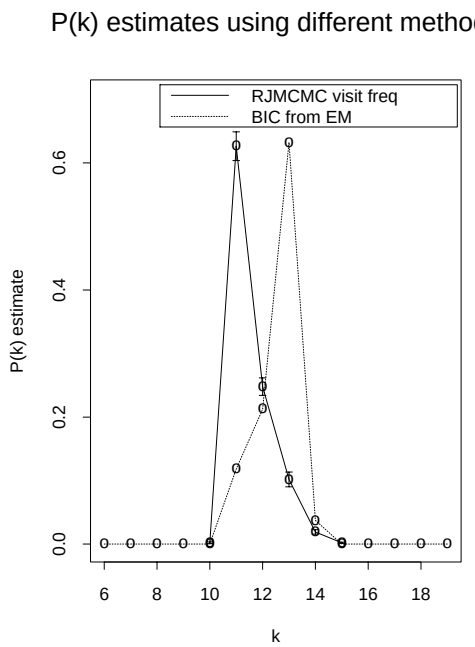
(a) I-k7-a



(b) I-k7-b

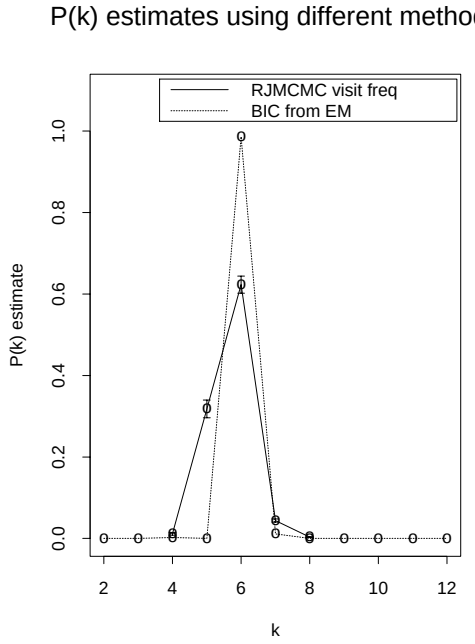


(c) I-k14-a

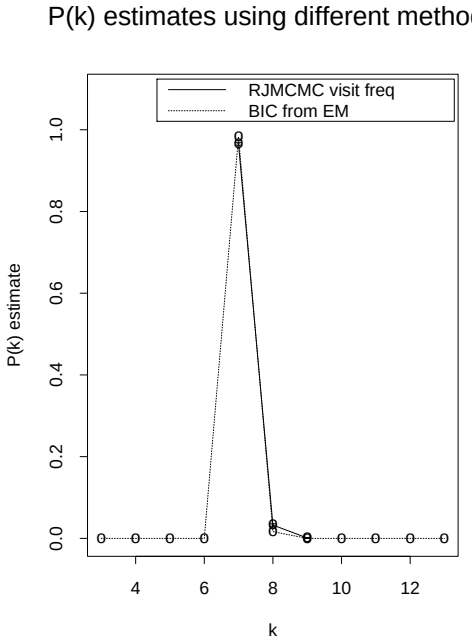


(d) I-k14-b

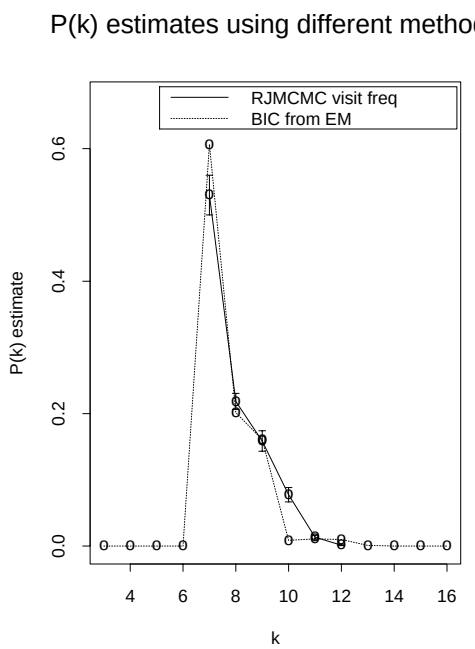
Figure G.2: $P(k)$ estimates using visit frequency from RJMCMC vs. BIC from EM, simulated patterns.



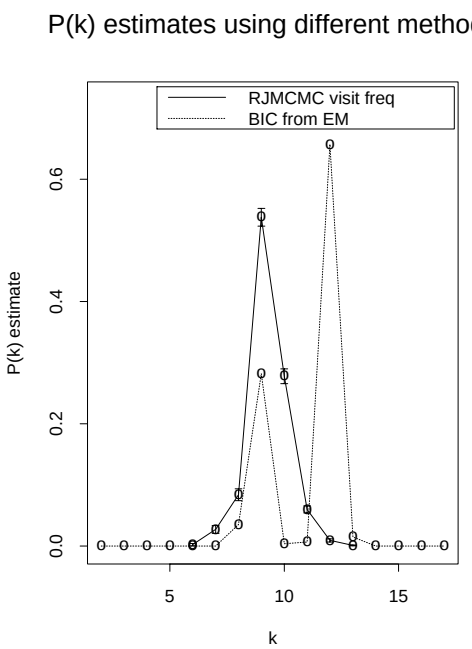
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

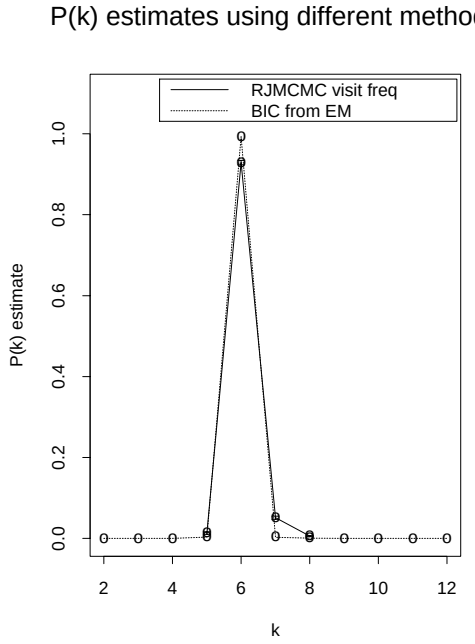


(g) AI-1.5-k14-a

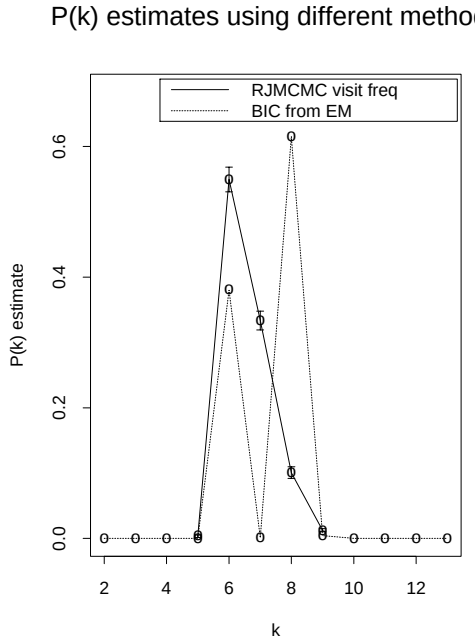


(h) AI-1.5-k14-b

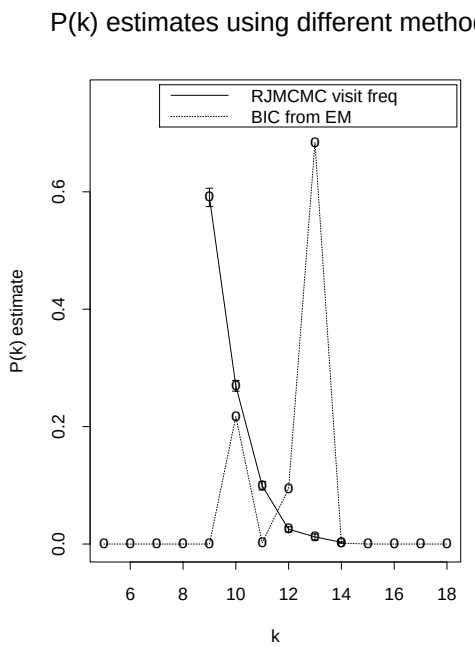
Figure G.2 (continued).



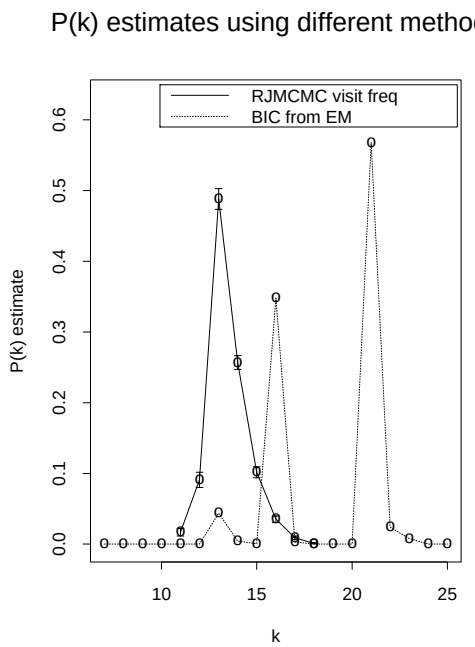
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure G.2 (continued).

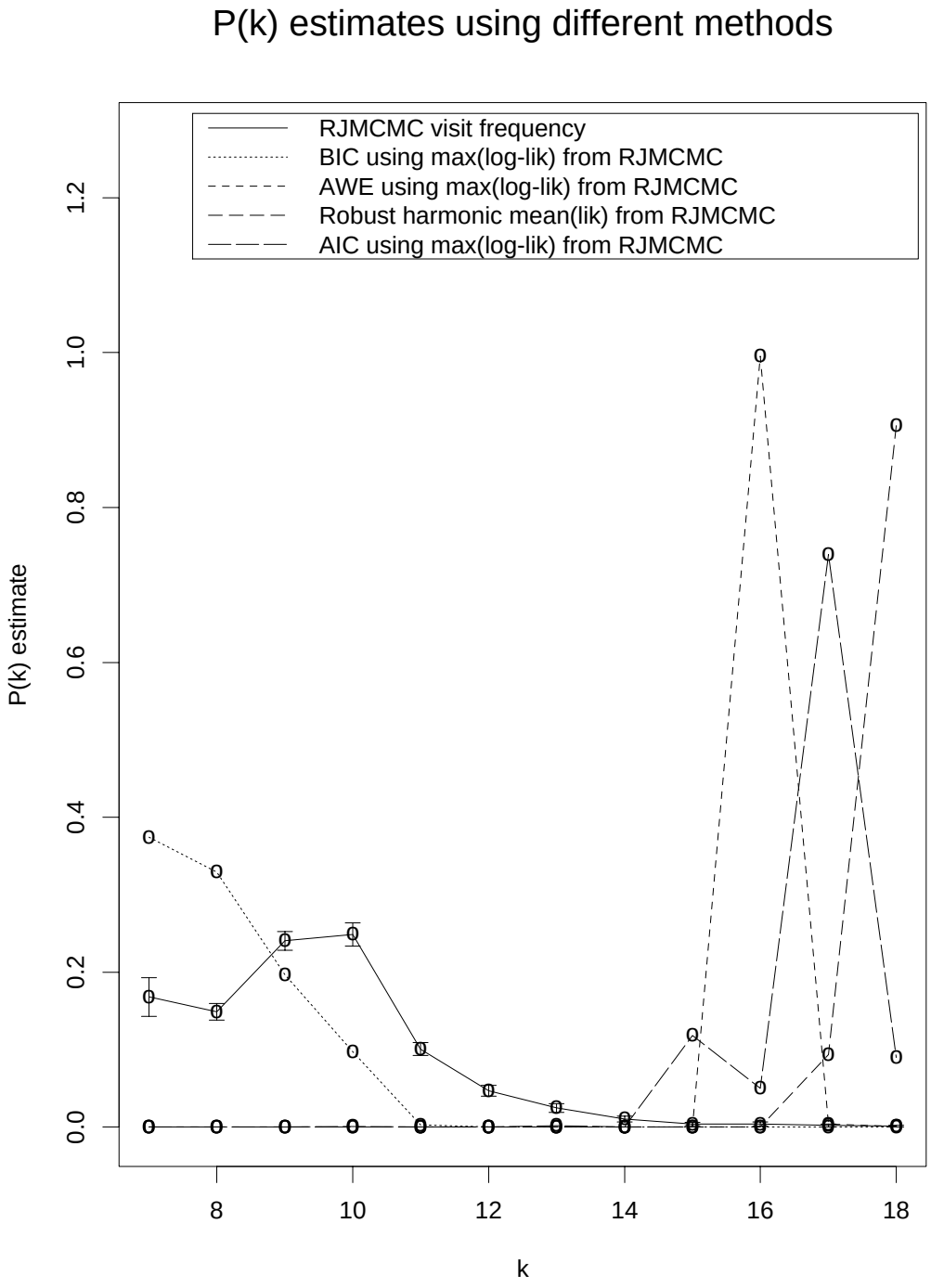
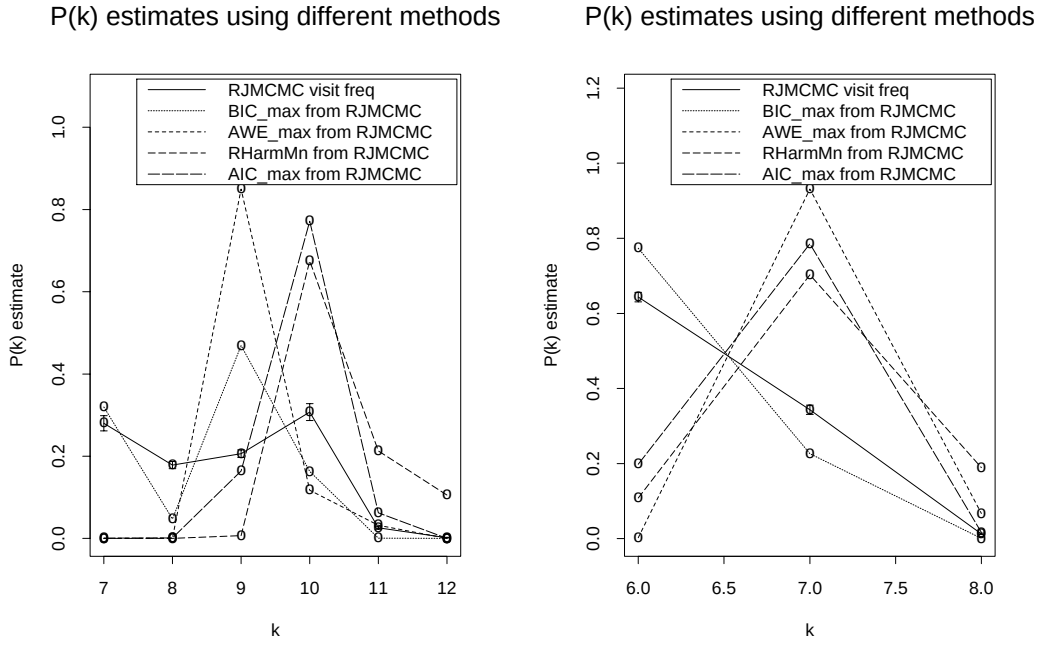
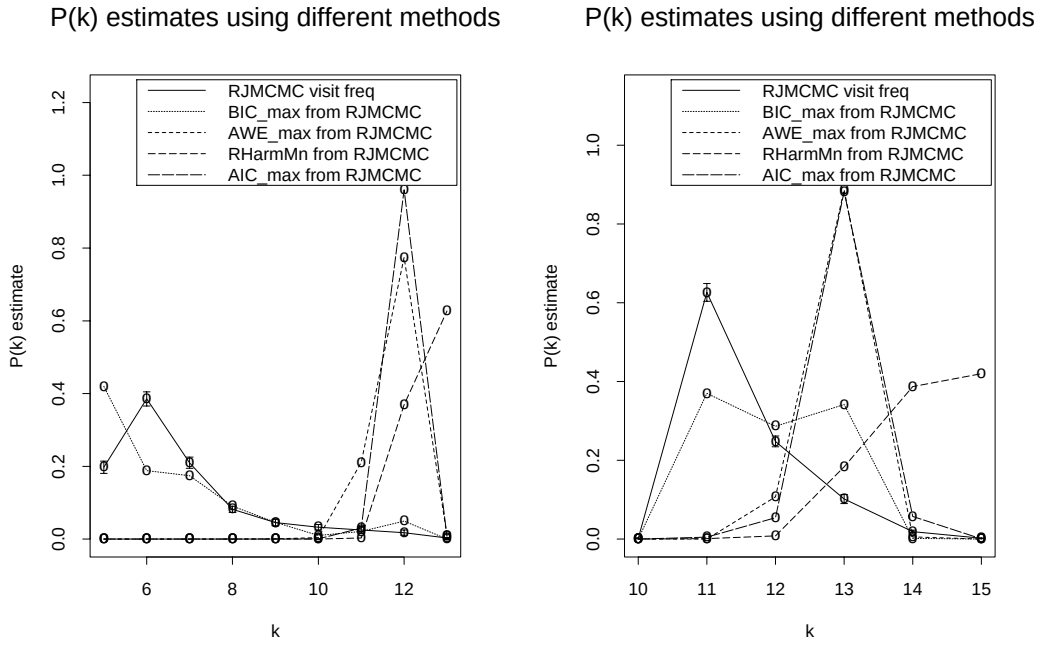


Figure G.3: $P(k)$ estimates using visit frequency from RJMCMC vs. penalized max marginal likelihoods and robust harmonic marginal likelihood mean, Redwood data.



(a) I-k7-a

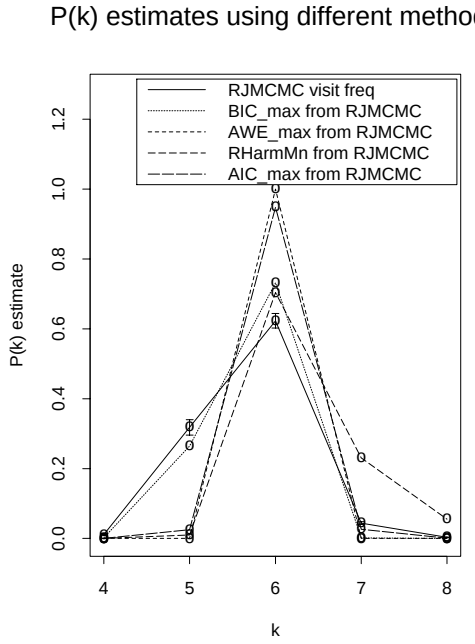
(b) I-k7-b



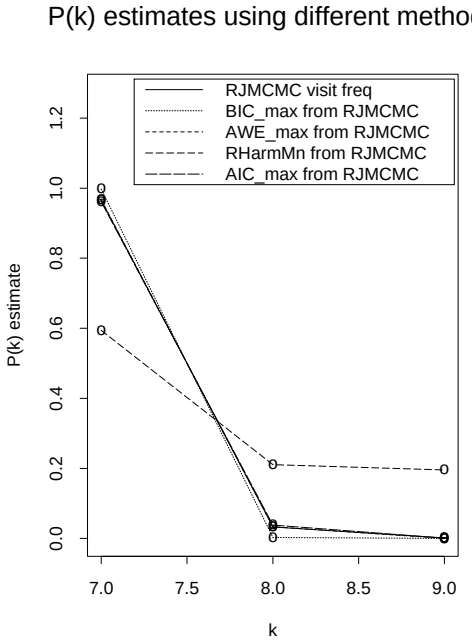
(c) I-k14-a

(d) I-k14-b

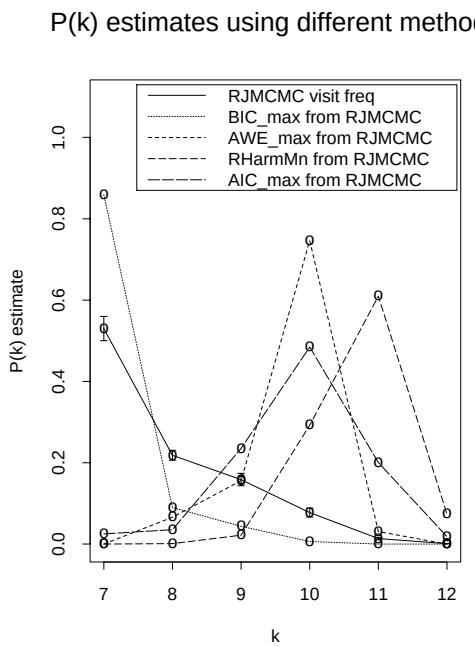
Figure G.4: $P(k)$ estimates using visit frequency from RJMCMC vs. penalized max marginal likelihoods and robust harmonic marginal likelihood mean, simulated patterns.



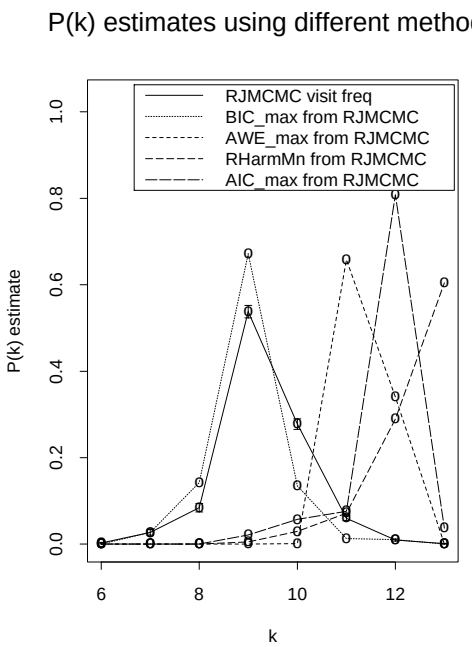
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

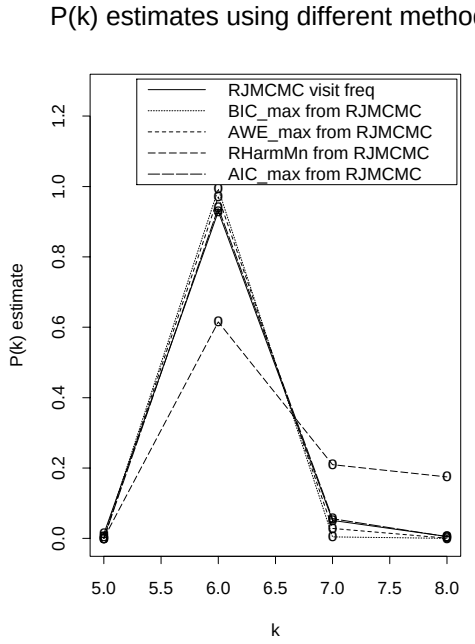


(g) AI-1.5-k14-a

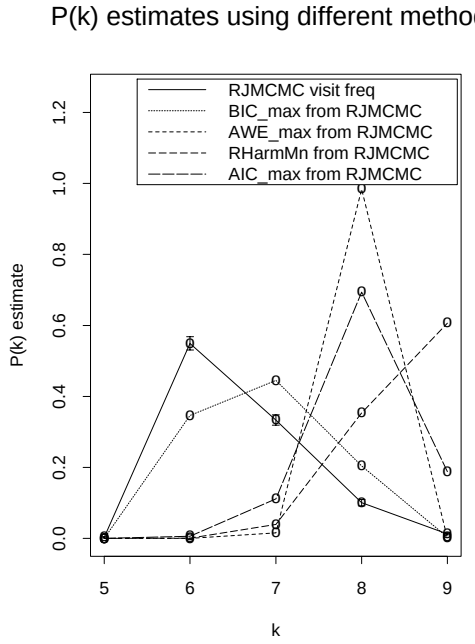


(h) AI-1.5-k14-b

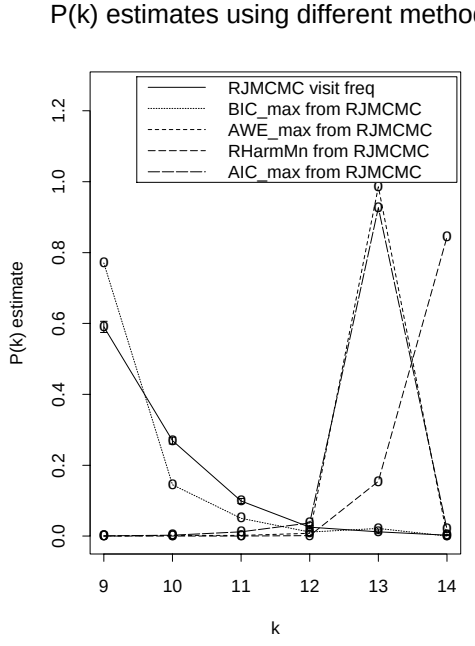
Figure G.4 (continued).



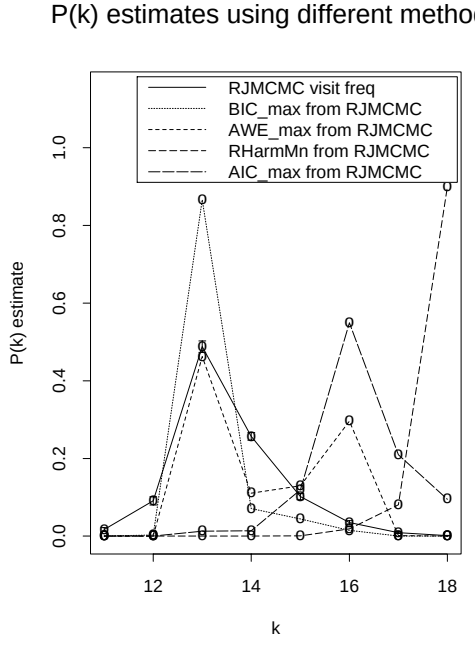
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure G.4 (continued).

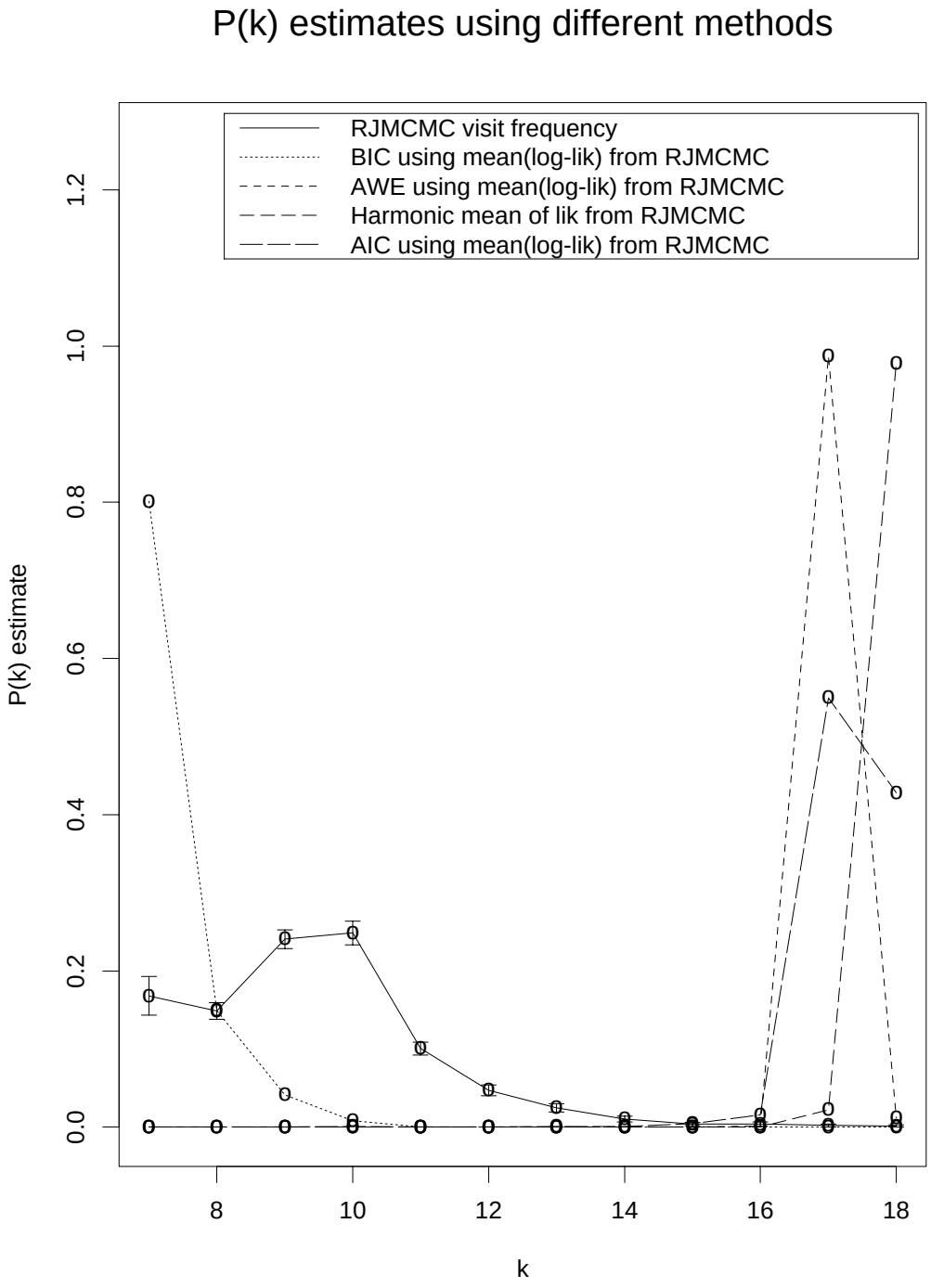
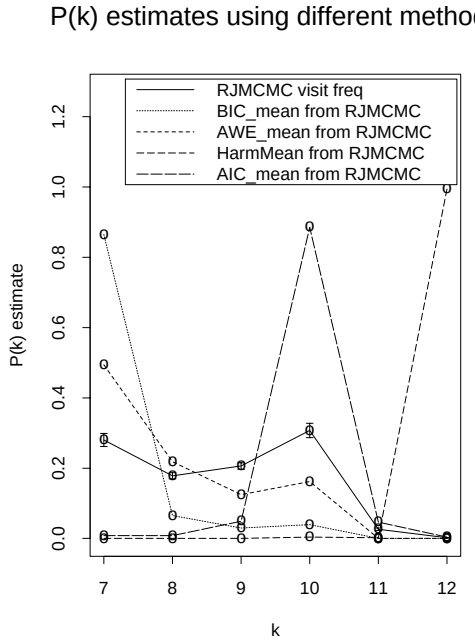
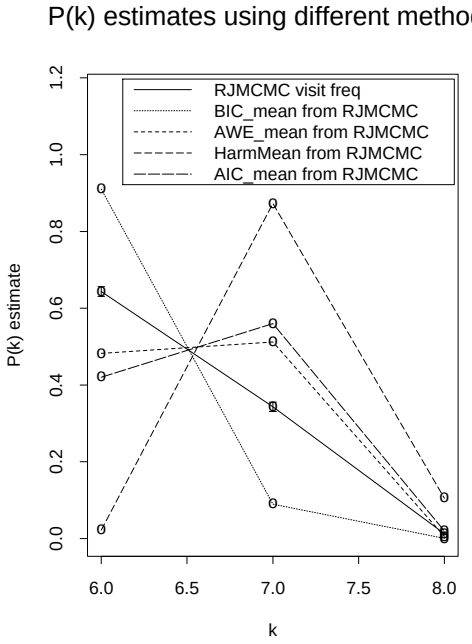


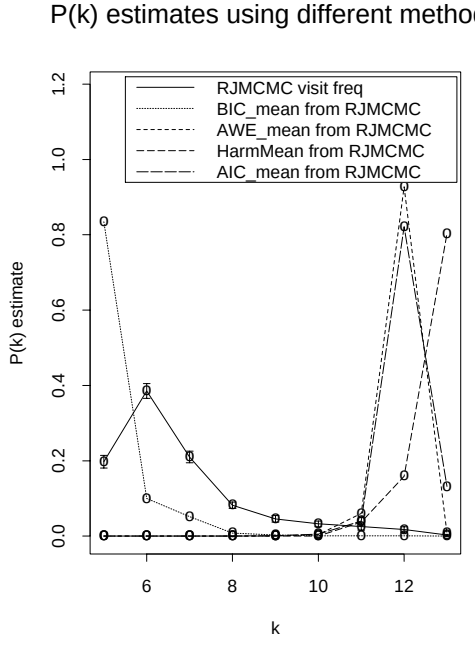
Figure G.5: $P(k)$ estimates using visit frequency from RJMCMC vs. penalized mean marginal likelihoods and harmonic marginal likelihood mean, Redwood data.



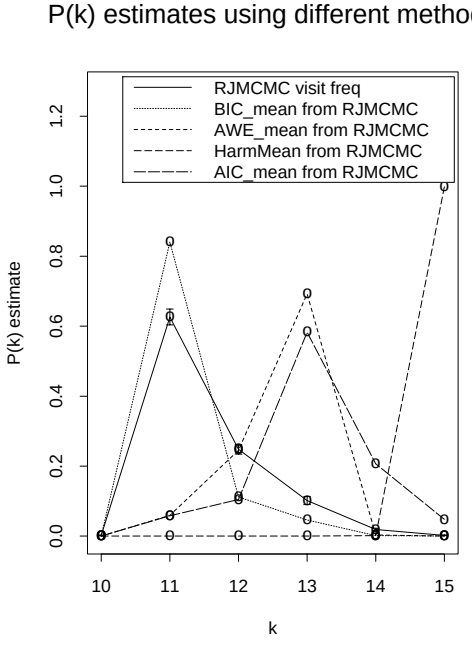
(a) I-k7-a



(b) I-k7-b

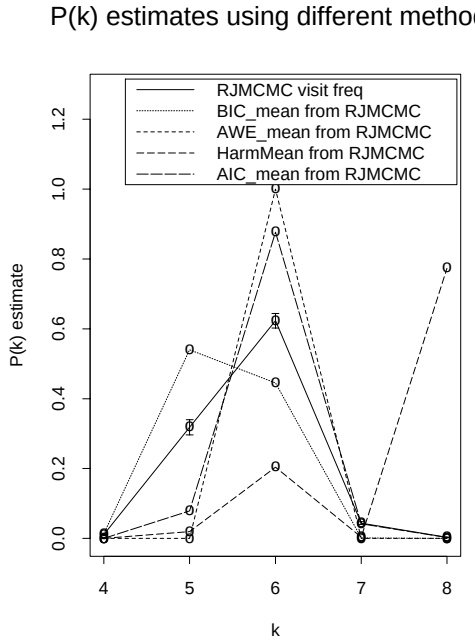


(c) I-k14-a

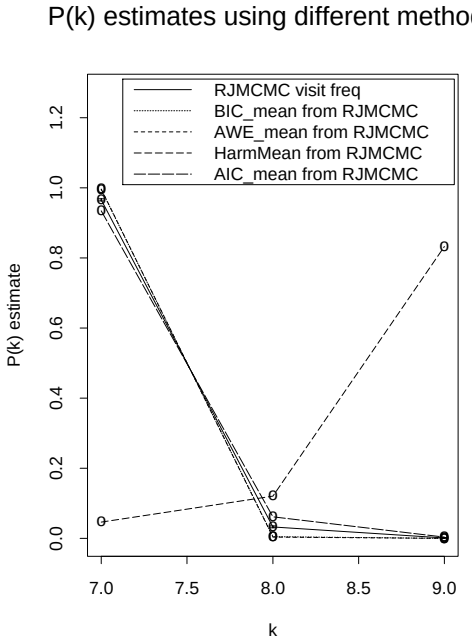


(d) I-k14-b

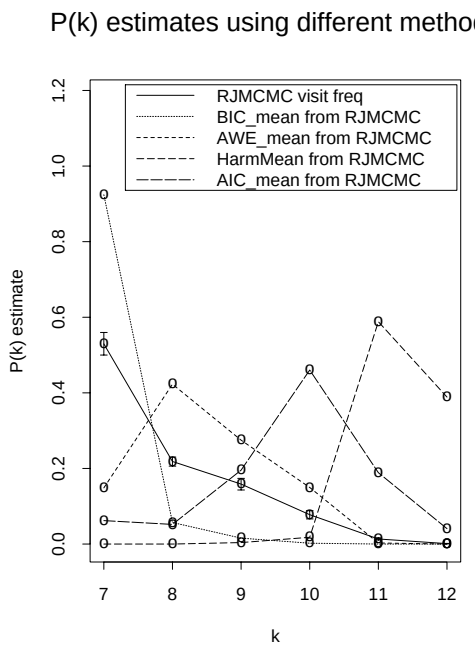
Figure G.6: $P(k)$ estimates using visit frequency from RJMCMC vs. penalized mean marginal likelihoods and harmonic marginal likelihood mean, simulated patterns.



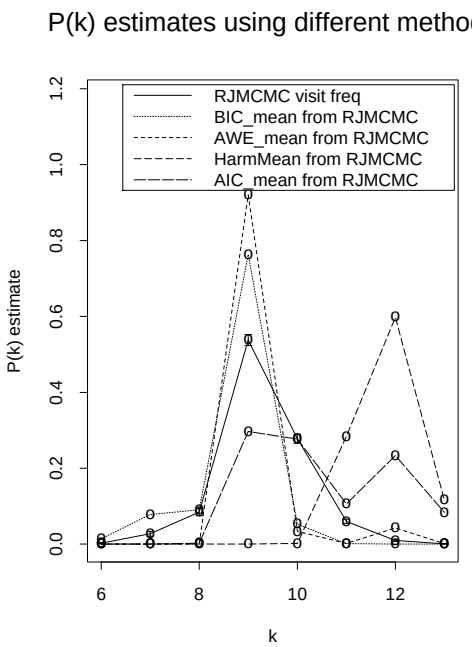
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

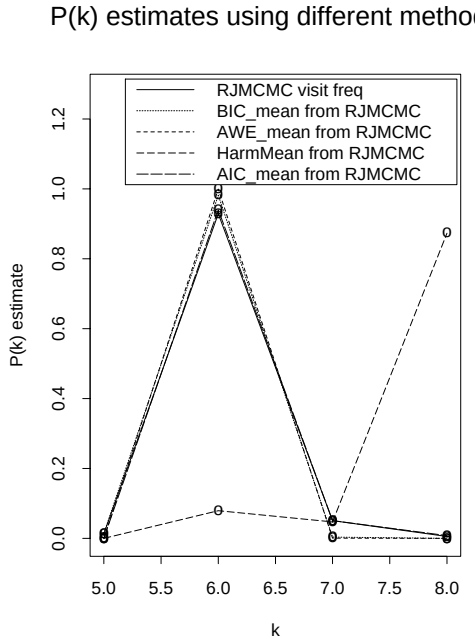


(g) AI-1.5-k14-a

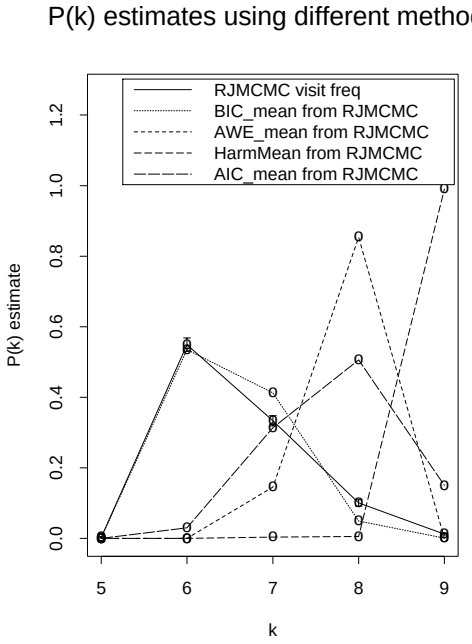


(h) AI-1.5-k14-b

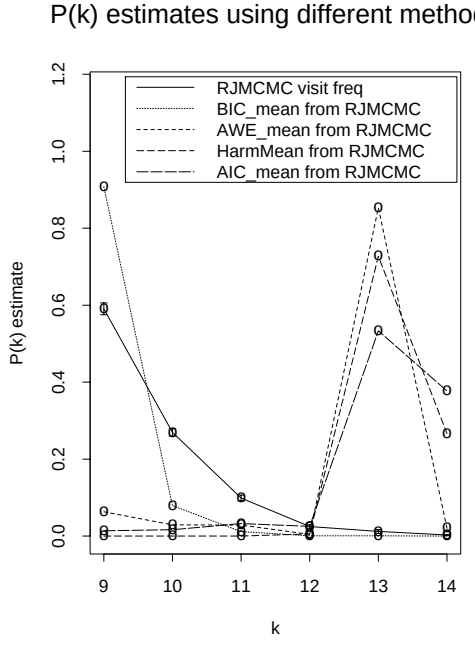
Figure G.6 (continued).



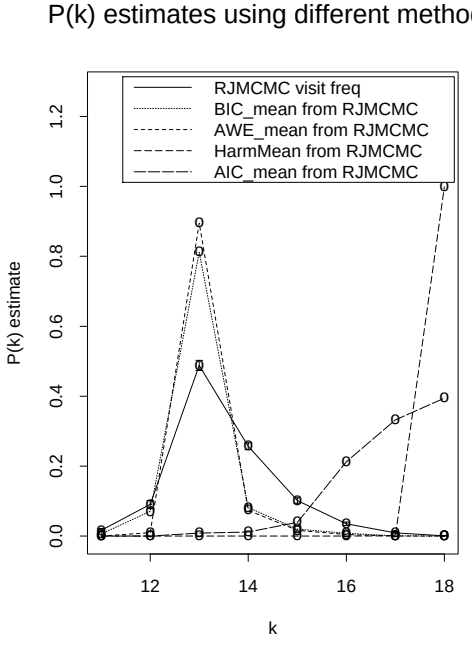
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure G.6 (continued).

APPENDIX H
MODEL ADEQUACY AND COMPARISON CRITERIA

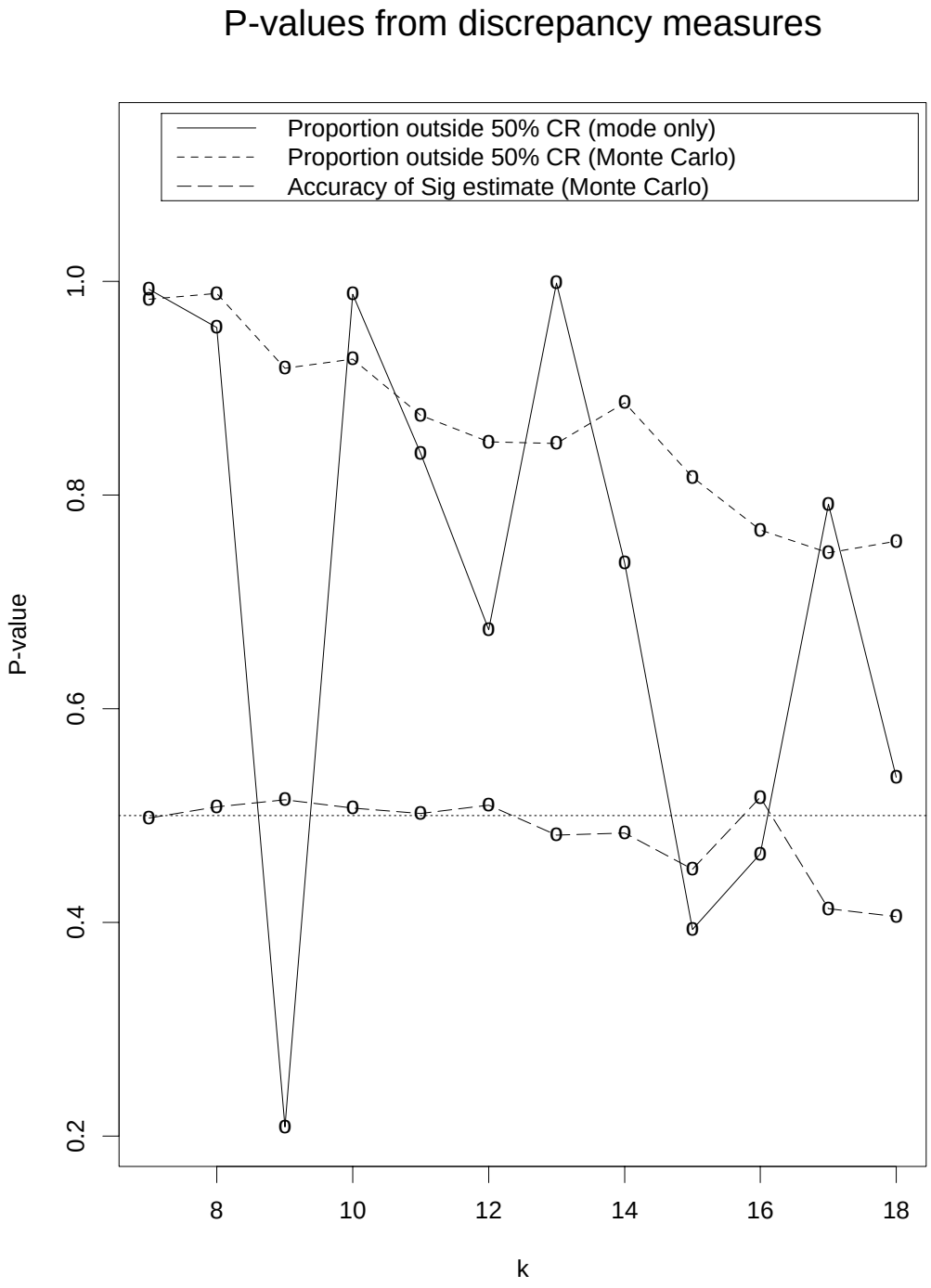
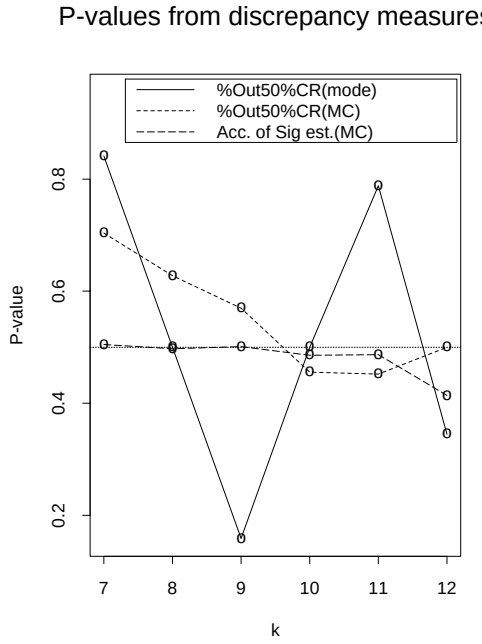
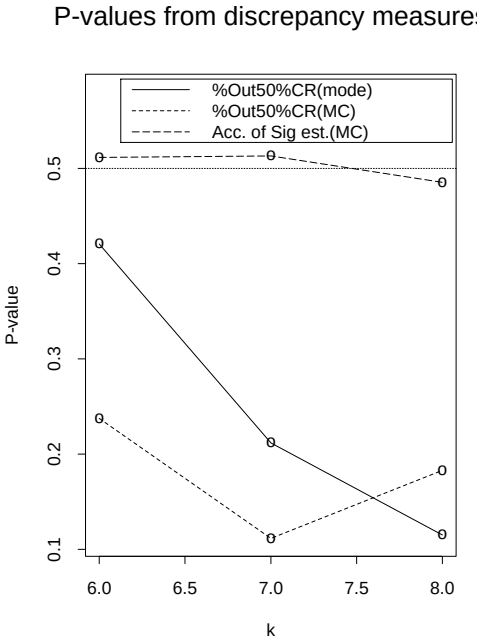


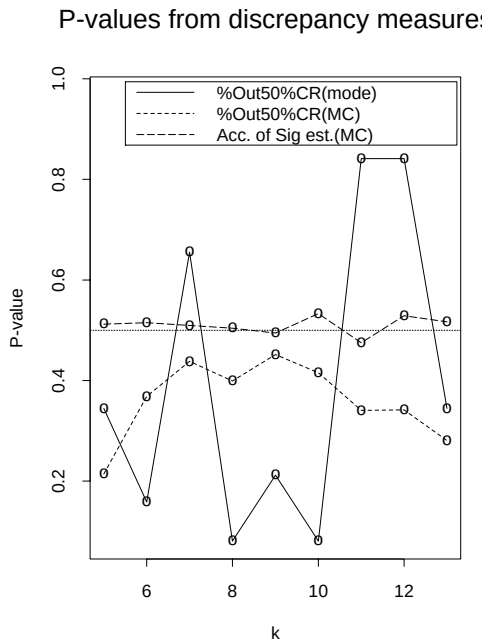
Figure H.1: P-values from posterior predictive distribution-based discrepancy measures, Redwood data.



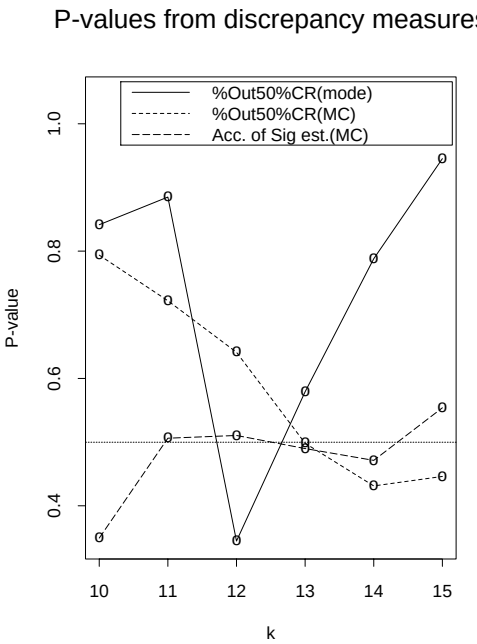
(a) I-k7-a



(b) I-k7-b

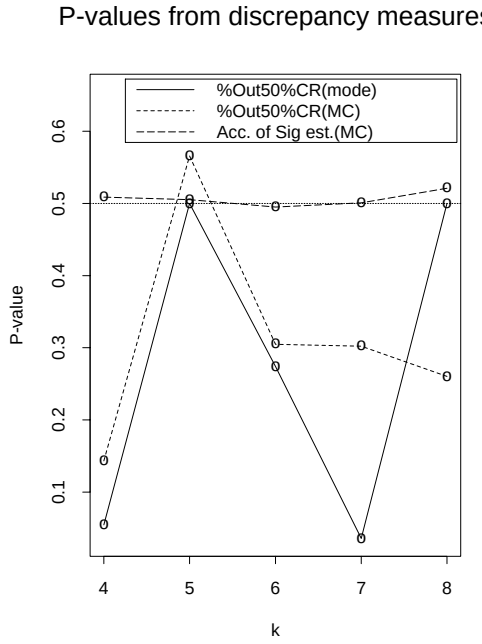


(c) I-k14-a

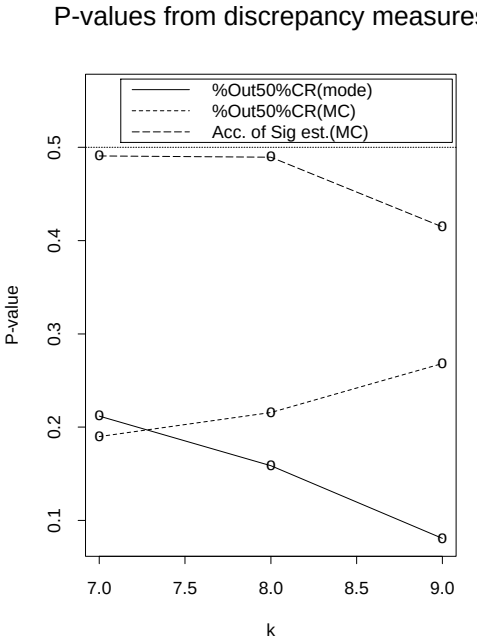


(d) I-k14-b

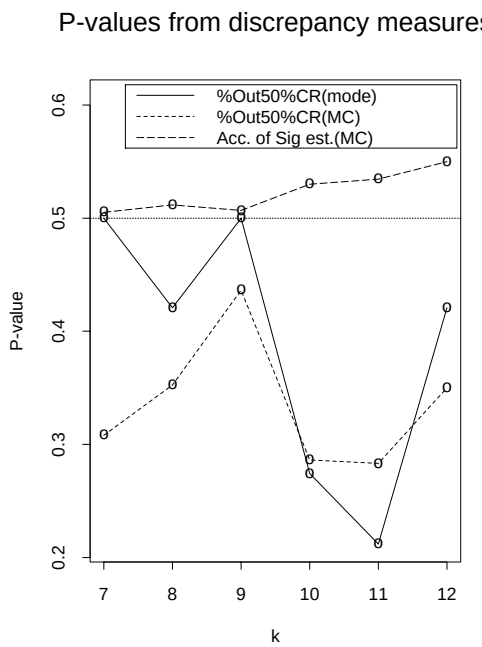
Figure H.2: P-values from posterior predictive distribution-based discrepancy measures, simulated patterns.



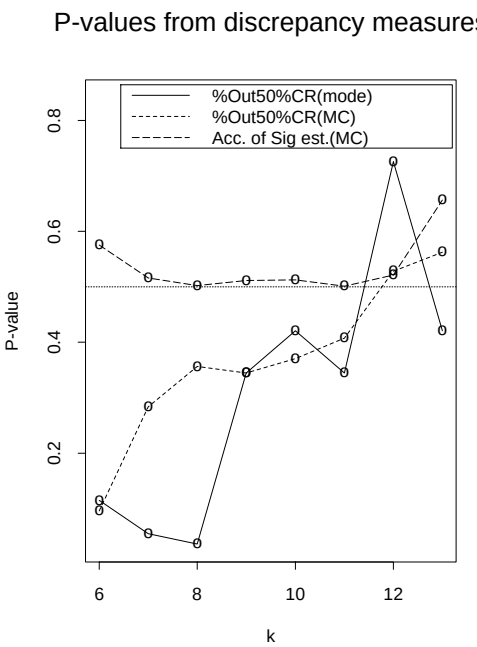
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

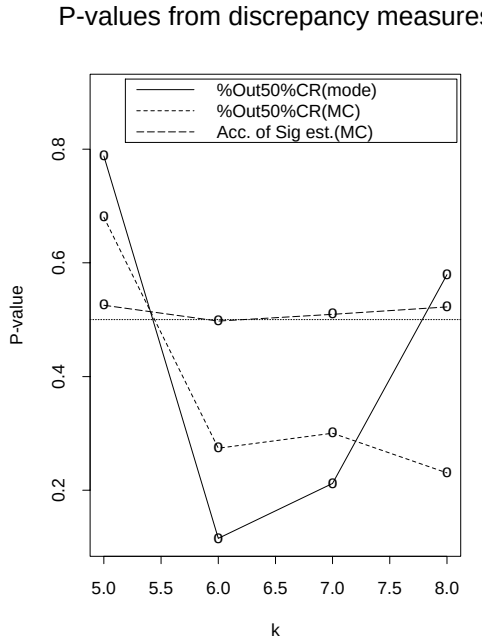


(g) AI-1.5-k14-a

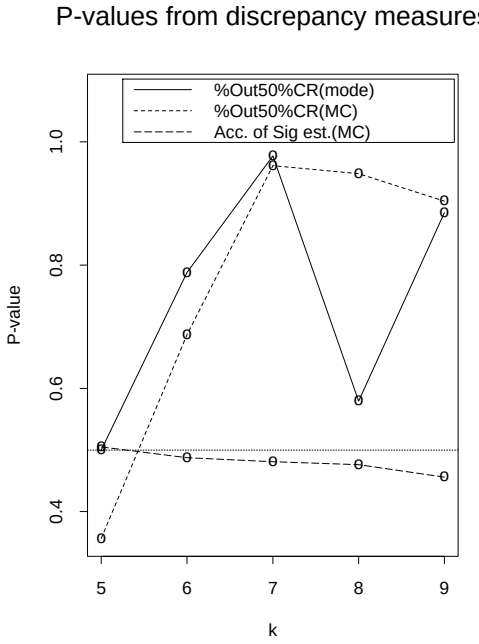


(h) AI-1.5-k14-b

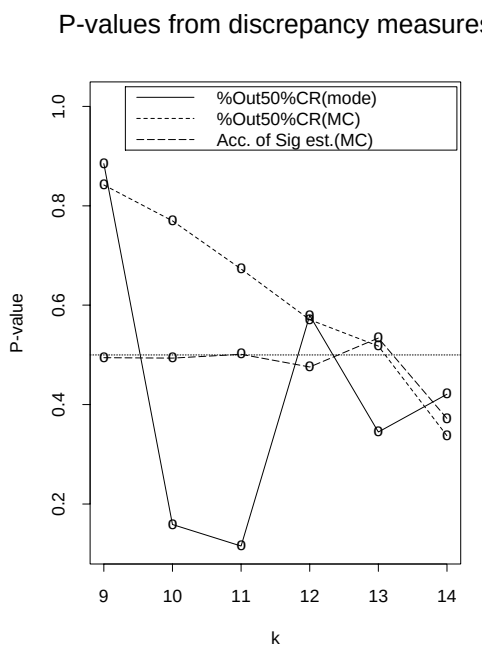
Figure H.2 (continued).



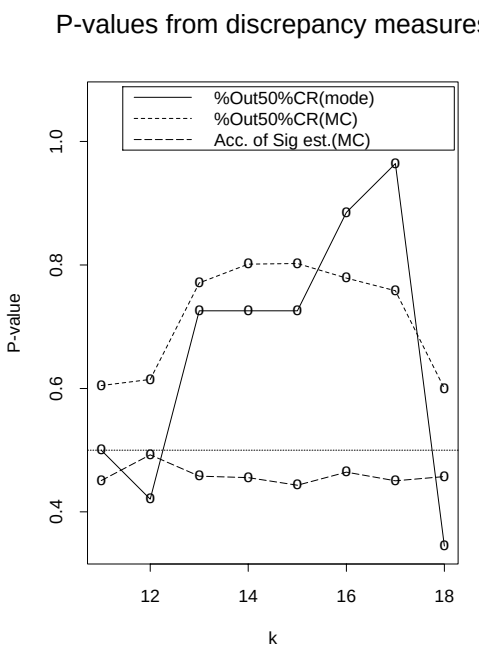
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure H.2 (continued).

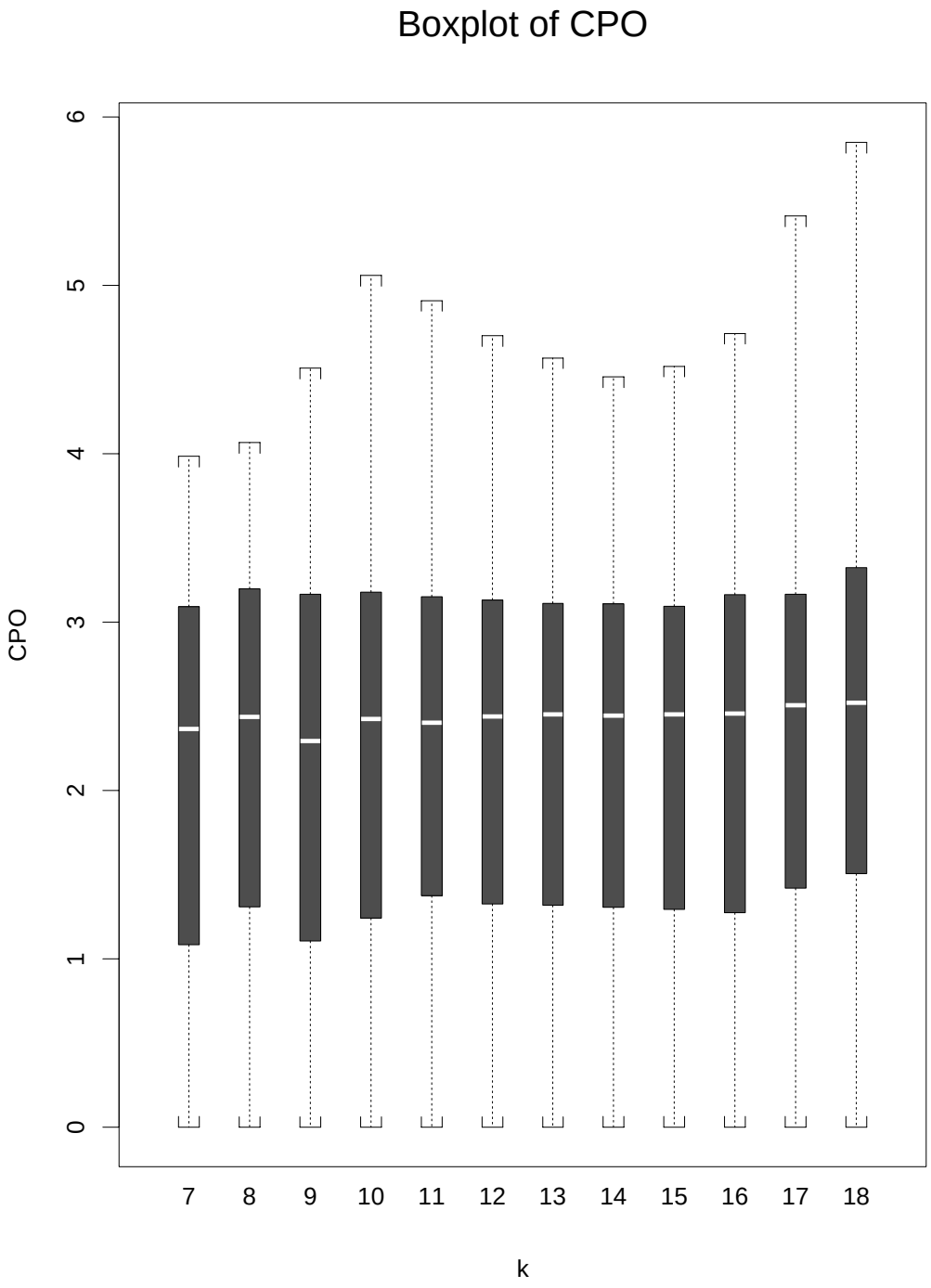
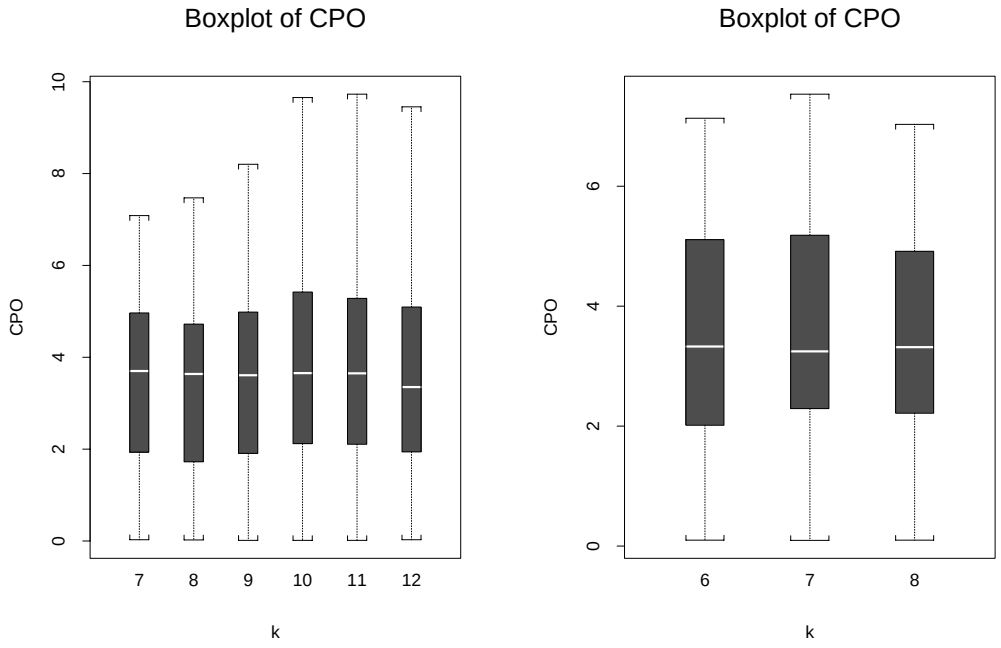
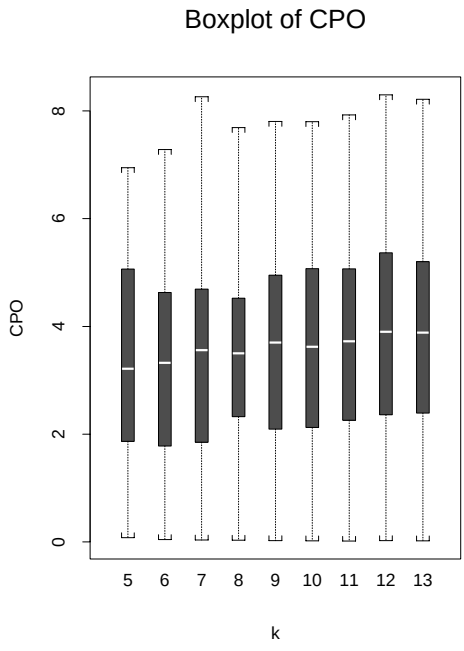


Figure H.3: Boxplots of $\widehat{CPO}_{j|k}$ values for different k, Redwood data.

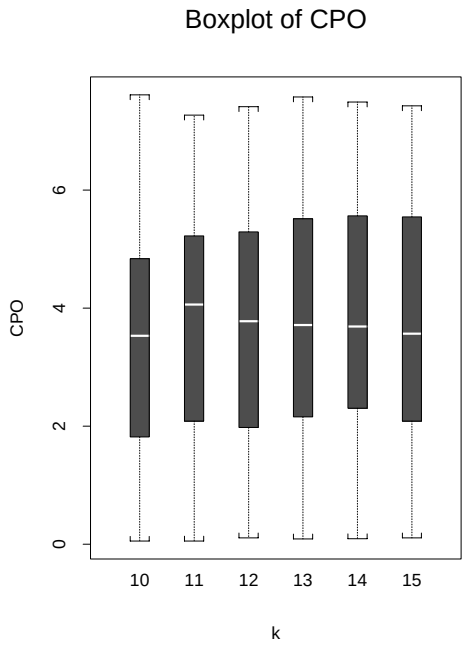


(a) I-k7-a

(b) I-k7-b

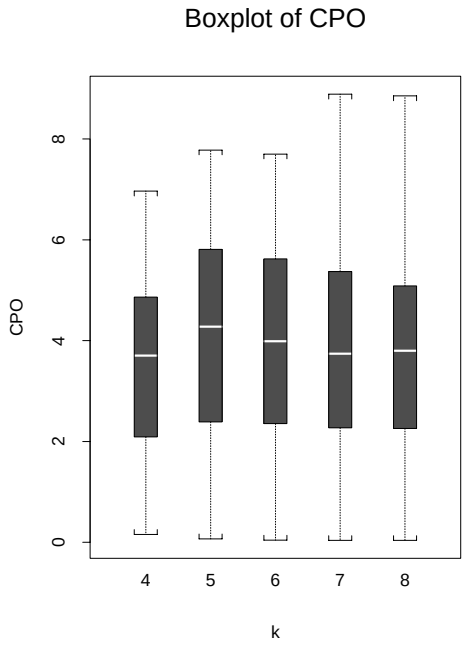


(c) I-k14-a

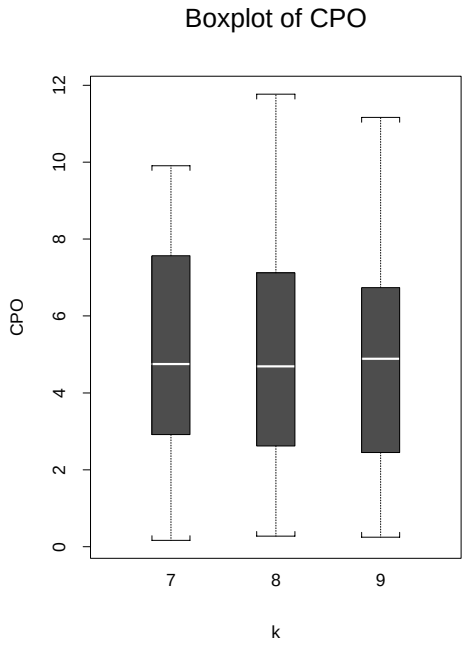


(d) I-k14-b

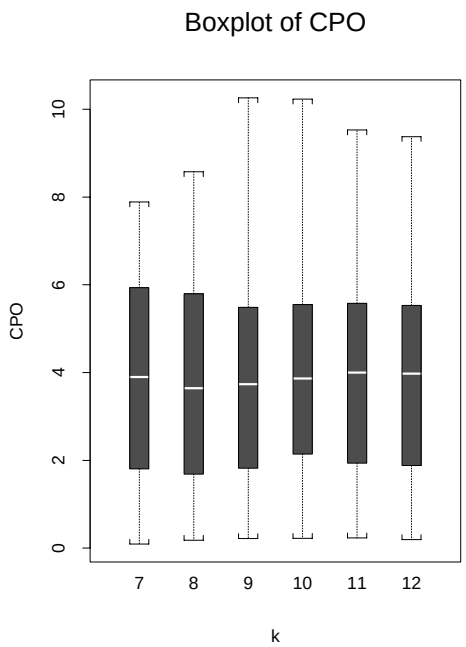
Figure H.4: Boxplots of $\widehat{CPO}_{j|k}$ values for different k, simulated patterns.



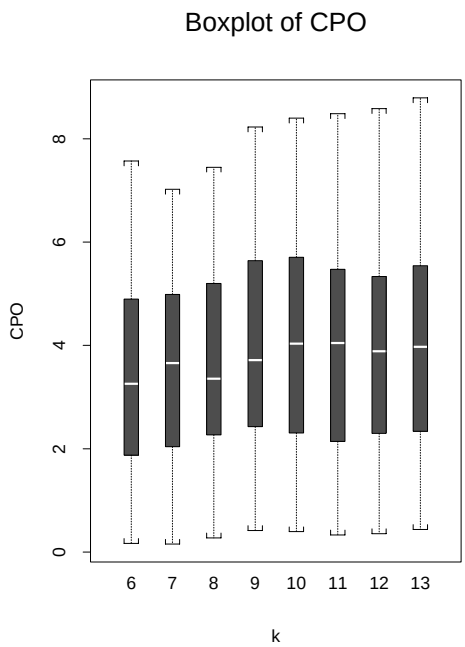
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

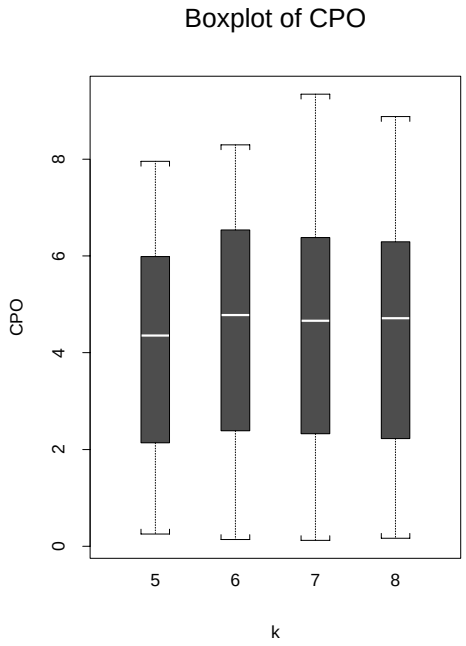


(g) AI-1.5-k14-a

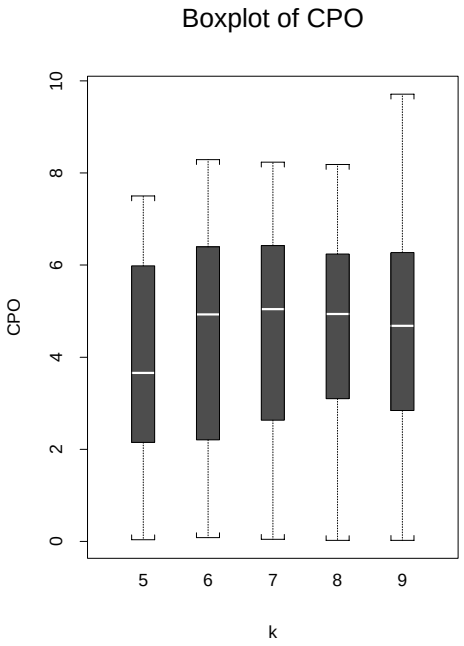


(h) AI-1.5-k14-b

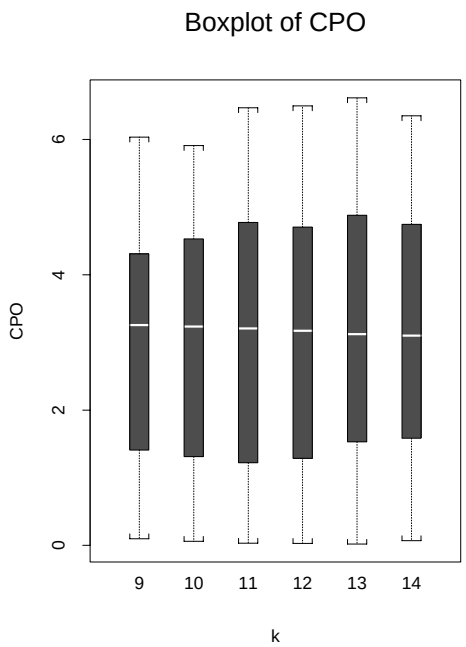
Figure H.4 (continued).



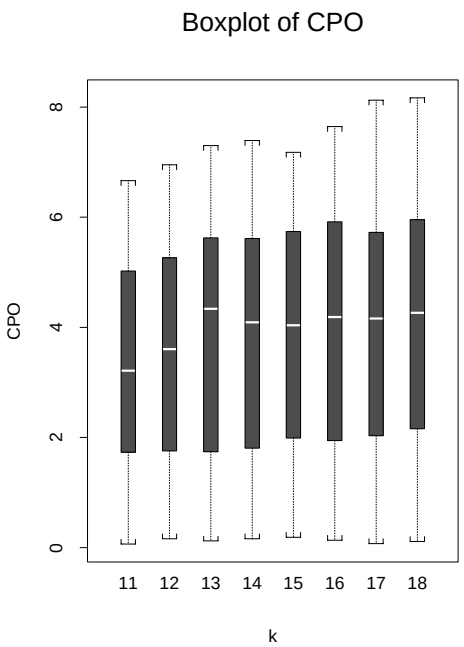
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure H.4 (continued).

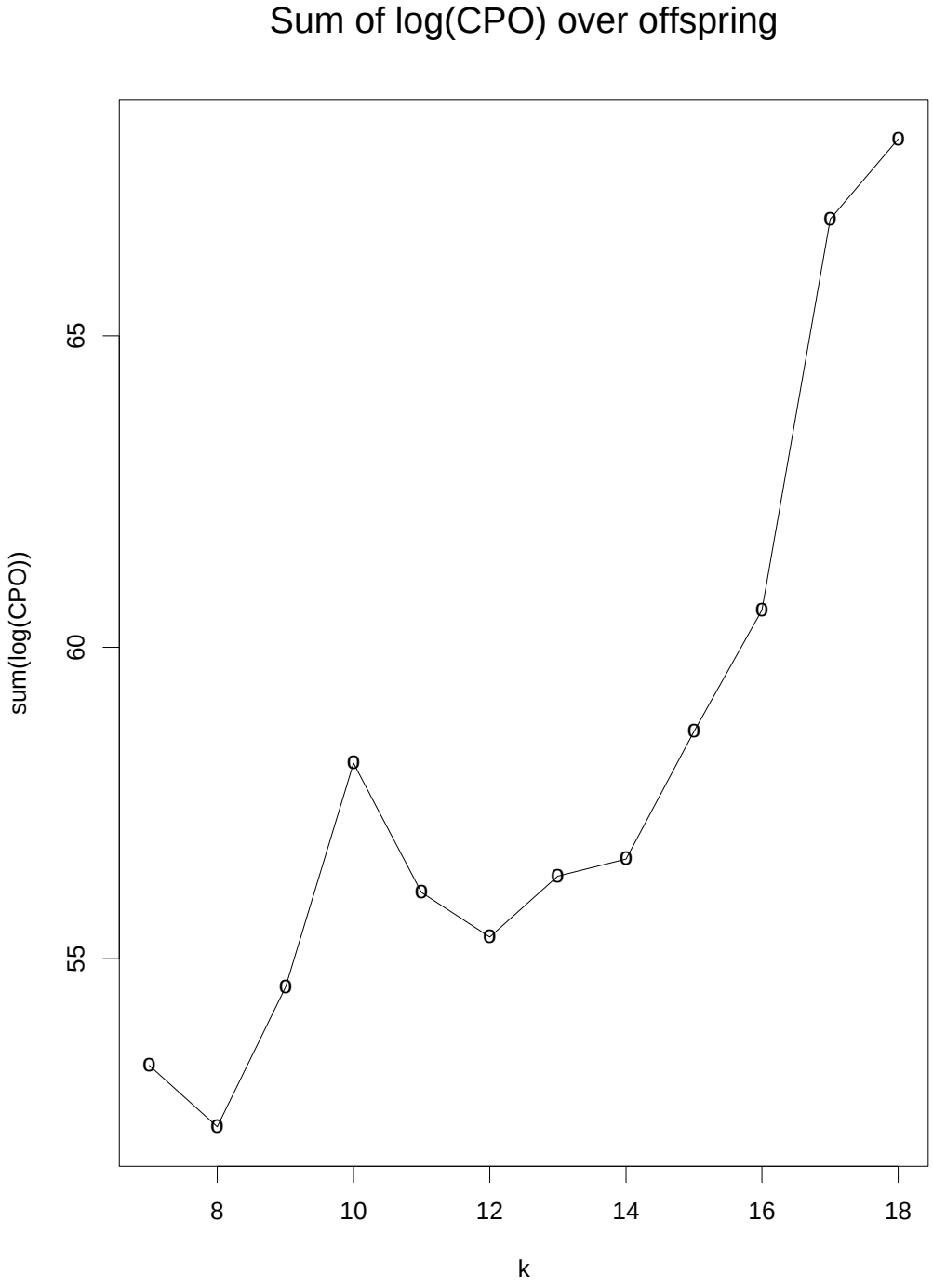
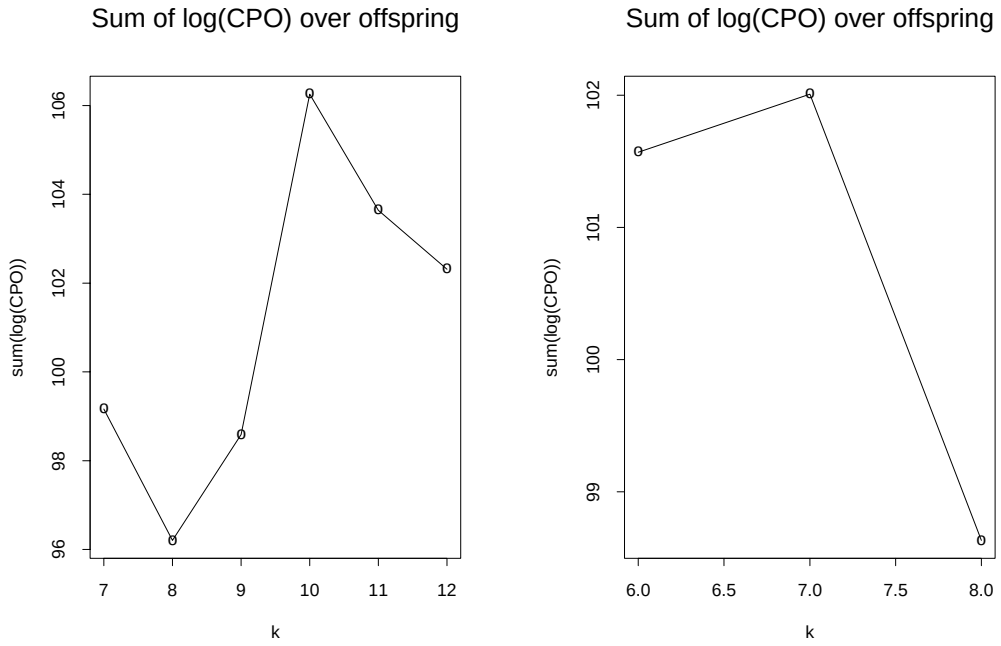
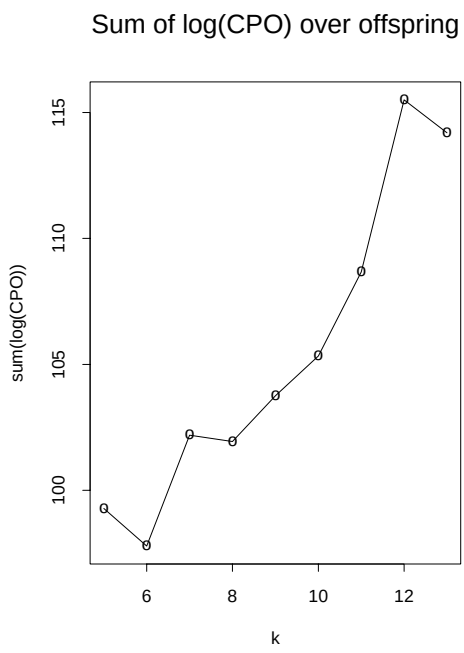


Figure H.5: Sum of log $\widehat{CPO}_{j|k}$ for different k, Redwood data.

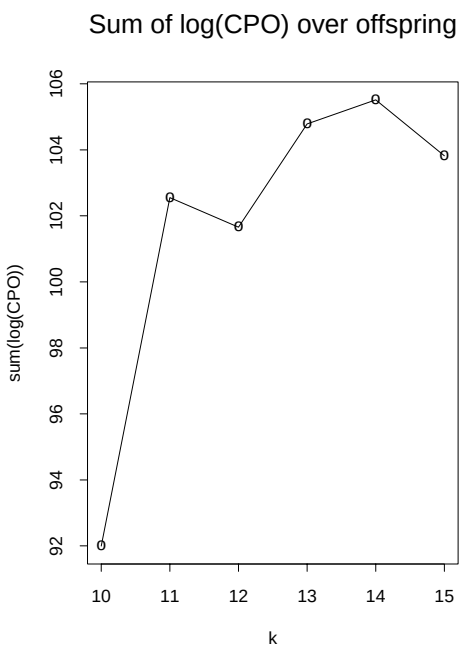


(a) I-k7-a

(b) I-k7-b

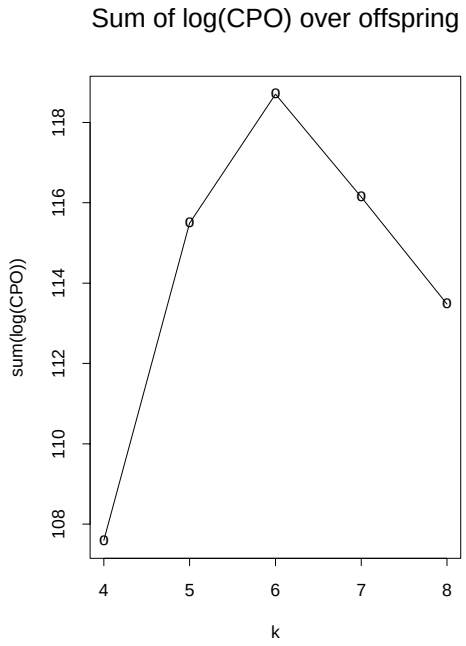


(c) I-k14-a

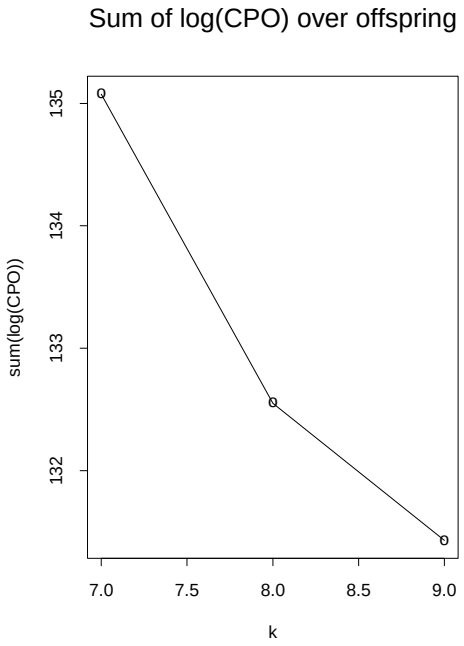


(d) I-k14-b

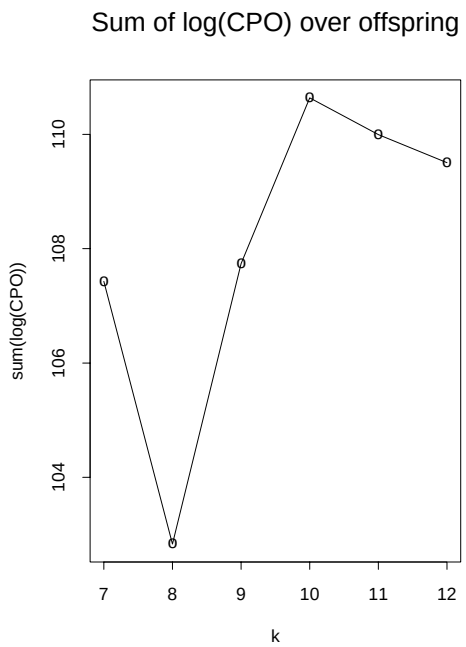
Figure H.6: Sum of $\log \widehat{CPO}_{j|k}$ for different k, simulated patterns.



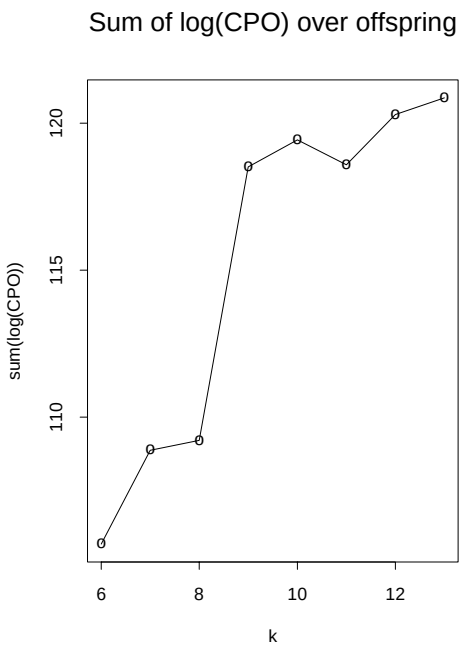
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

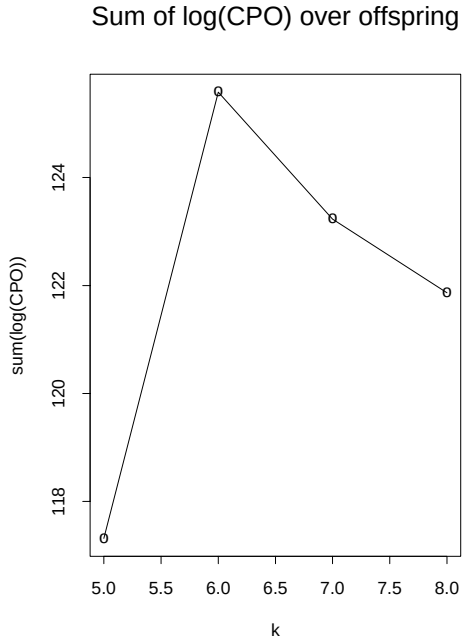


(g) AI-1.5-k14-a

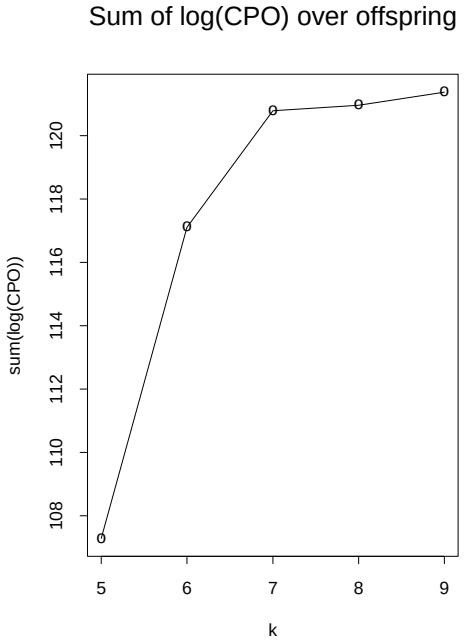


(h) AI-1.5-k14-b

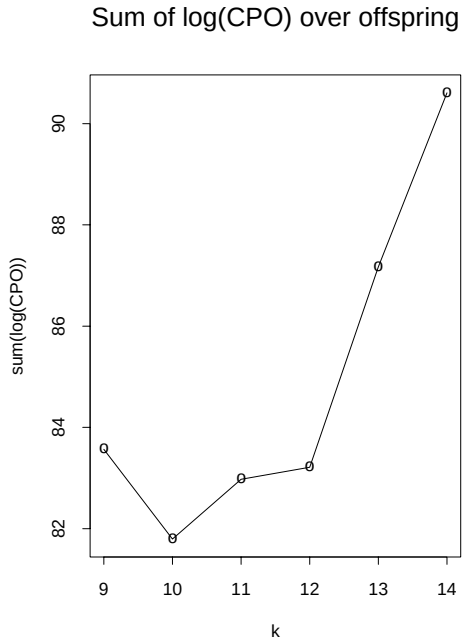
Figure H.6 (continued).



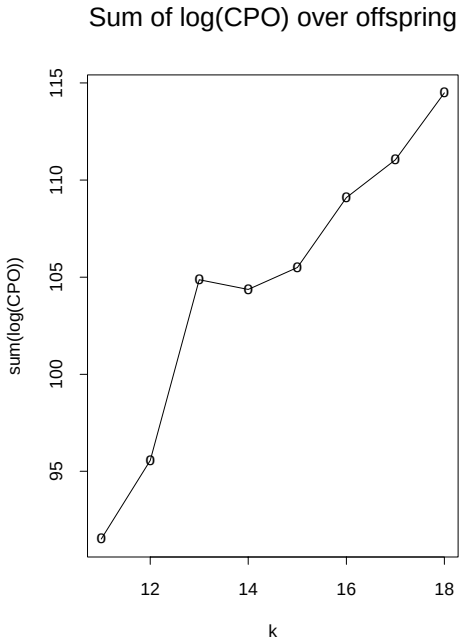
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure H.6 (continued).

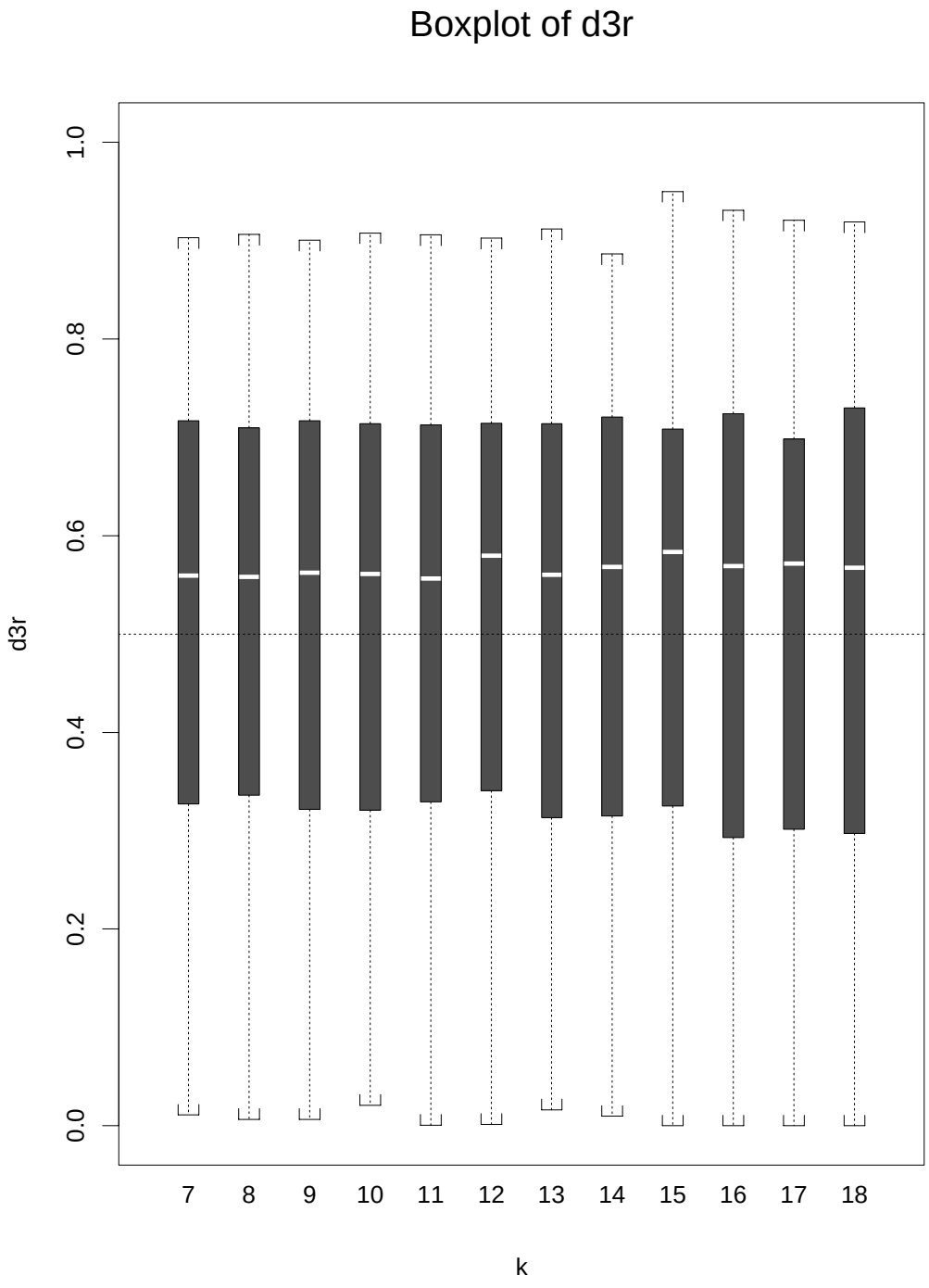
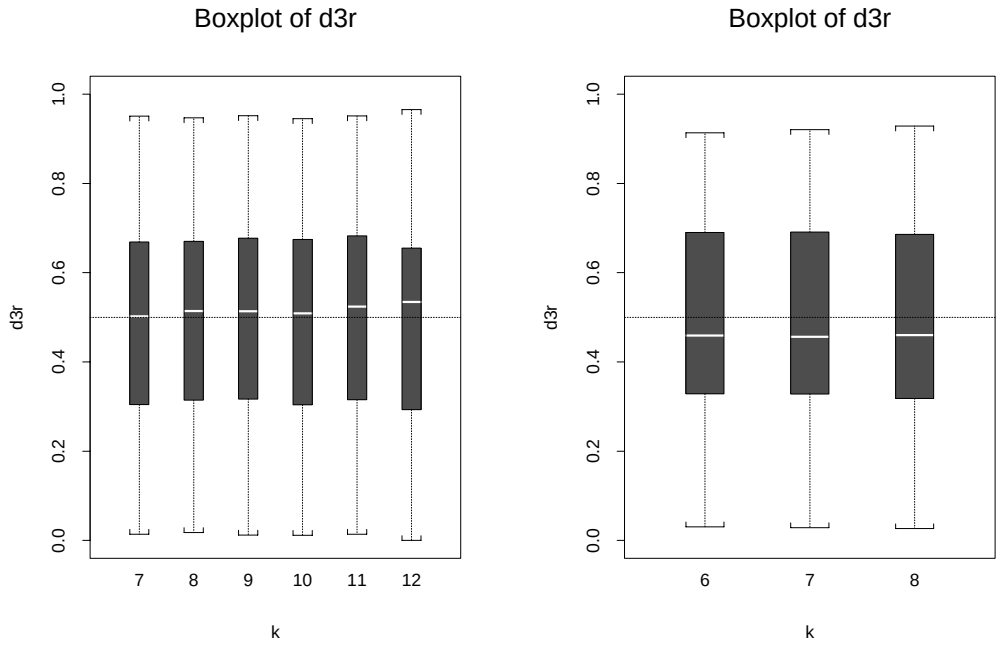
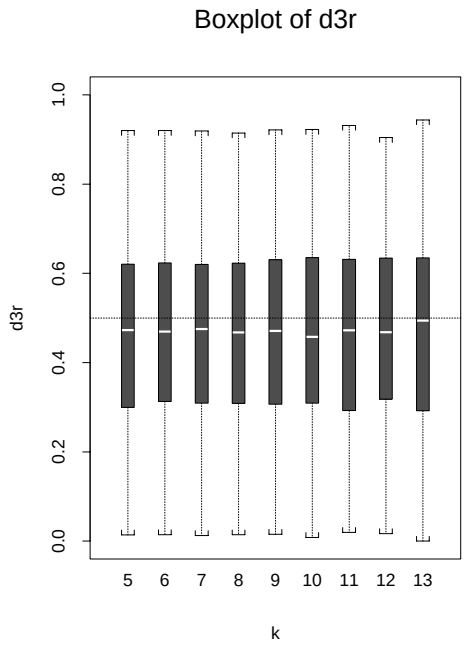


Figure H.7: Boxplots of $\widehat{d}_{3_{j|k}}$ values for different k, Redwood data.

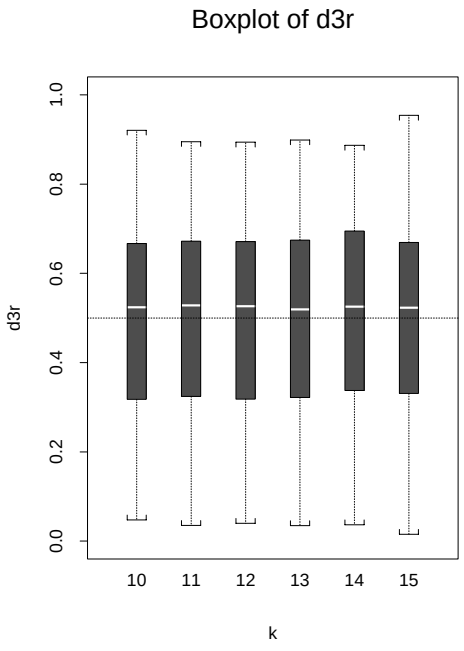


(a) I-k7-a

(b) I-k7-b

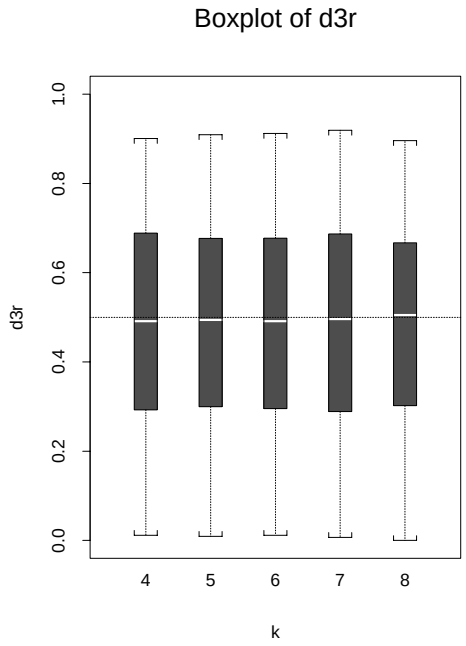


(c) I-k14-a

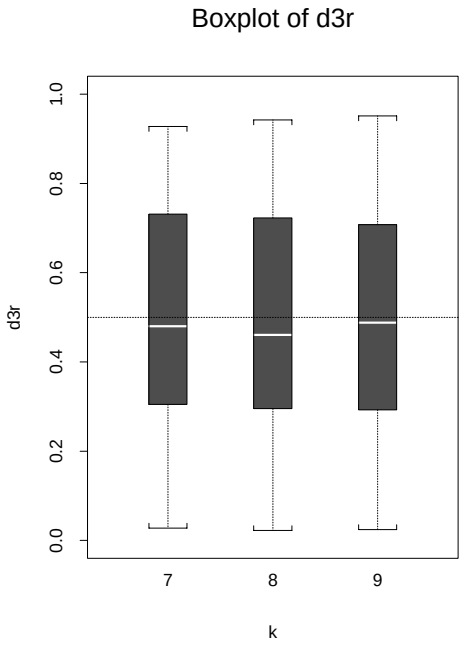


(d) I-k14-b

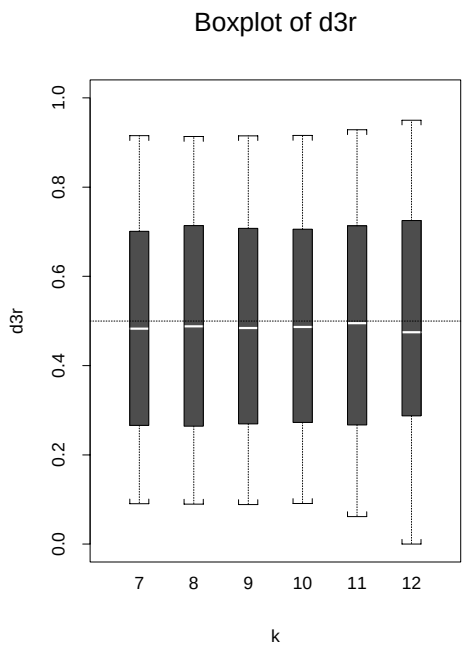
Figure H.8: Boxplots of $\widehat{d}_{3_{j|k}}$ values for different k, simulated patterns.



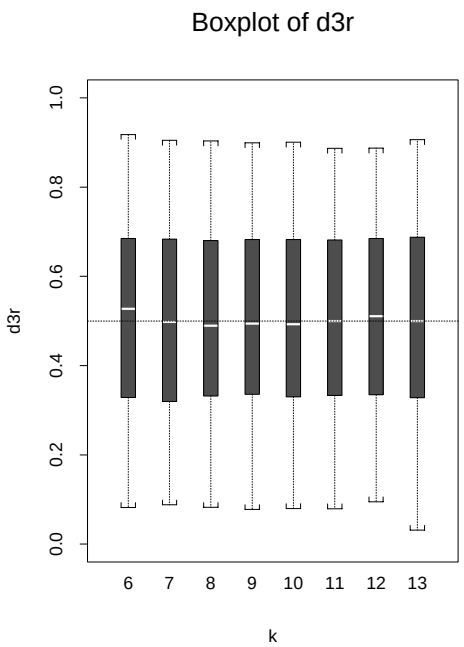
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

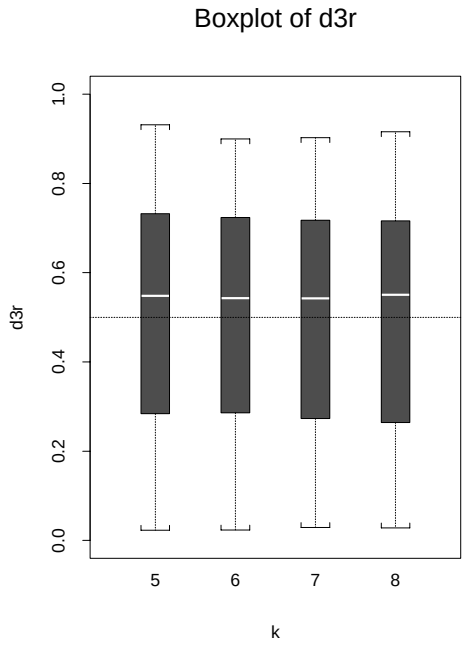


(g) AI-1.5-k14-a

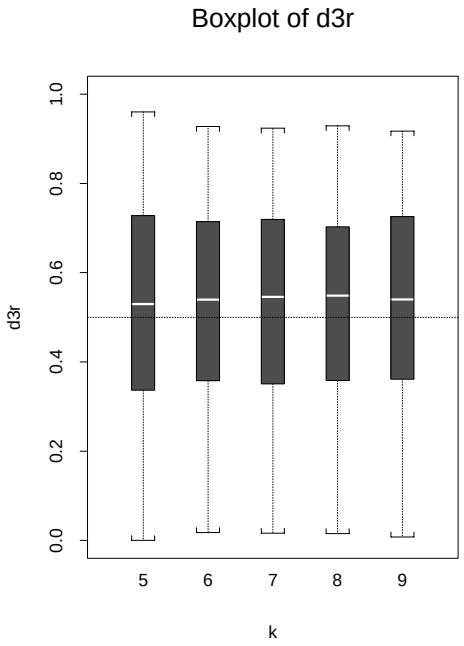


(h) AI-1.5-k14-b

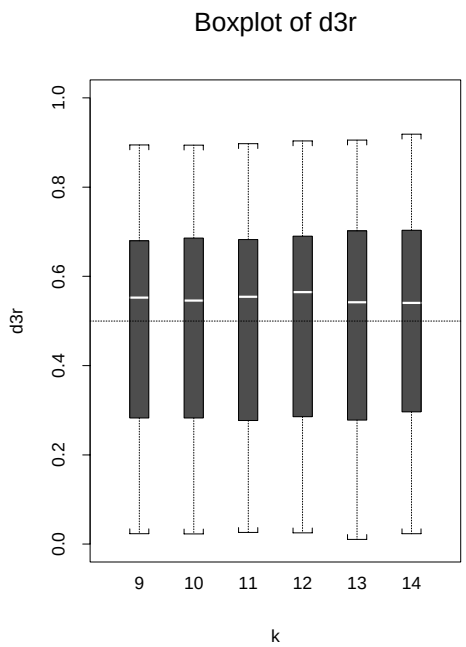
Figure H.8 (continued).



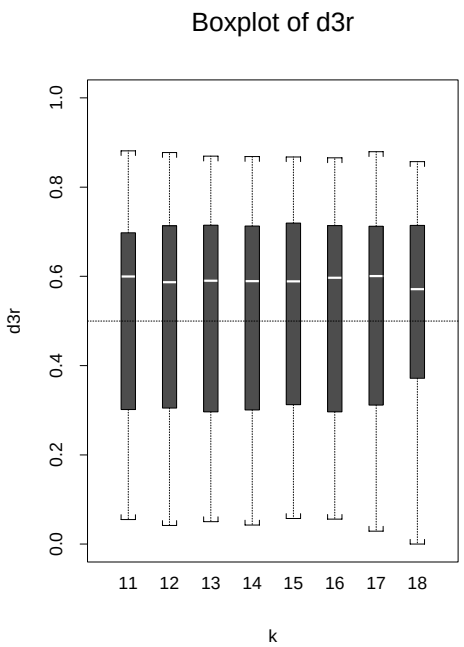
(i) AI-3-k7-a



(j) AI-3-k7-b

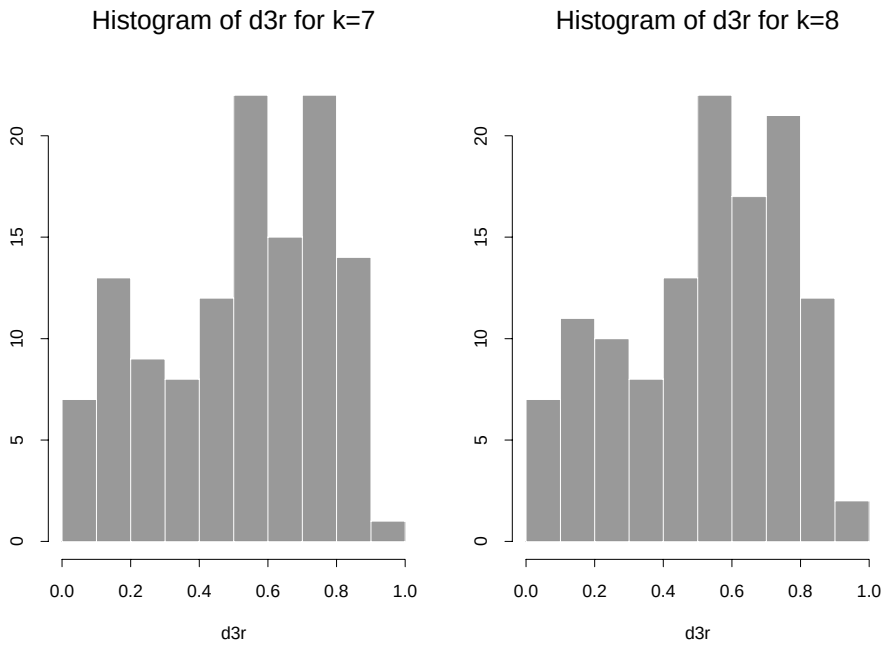


(k) AI-3-k14-a

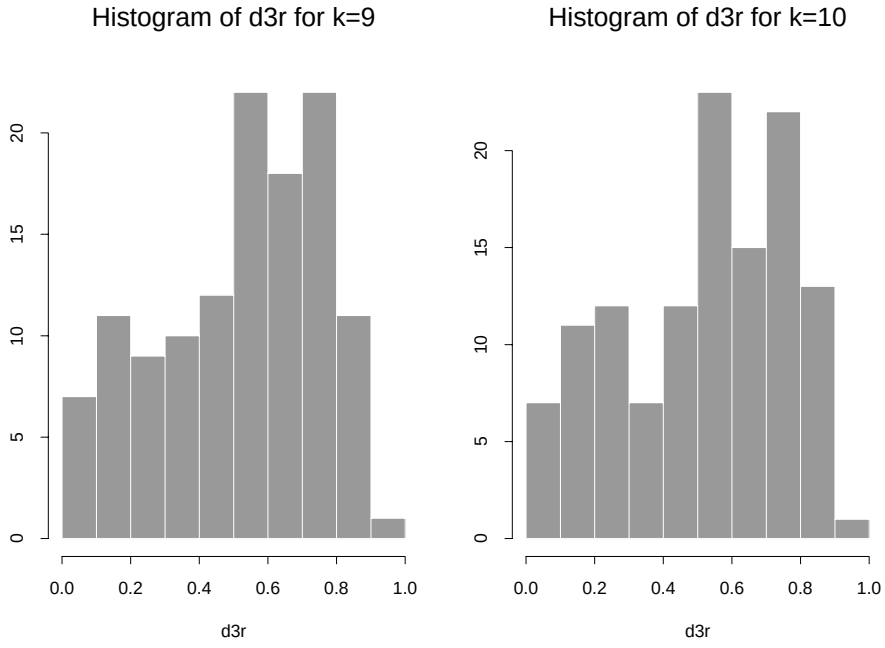


(l) AI-3-k14-b

Figure H.8 (continued).

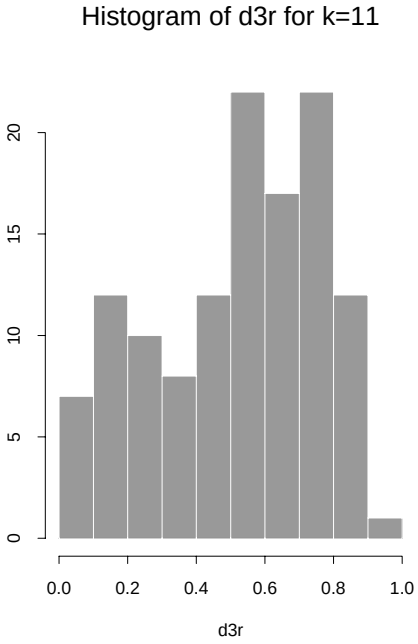


(a) $k = 7$ (b) $k = 8$

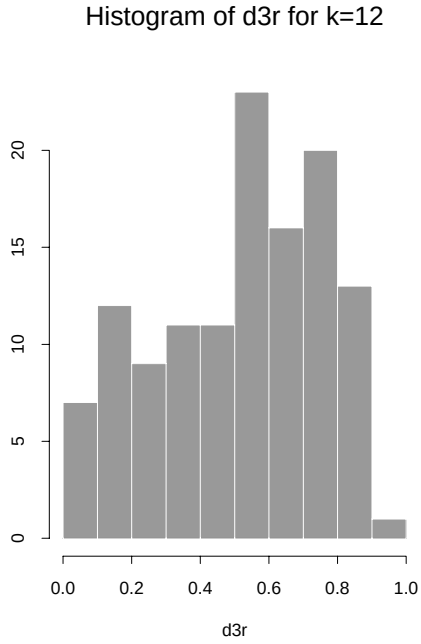


(c) $k = 9$ (d) $k = 10$

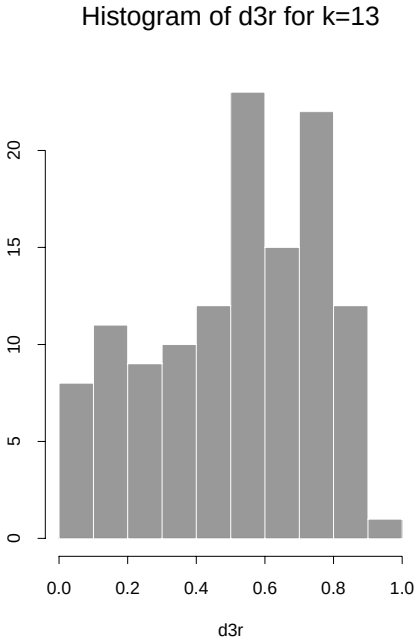
Figure H.9: Histograms of $\widehat{d}_{3_{j|k}}$ by k, Redwood data.



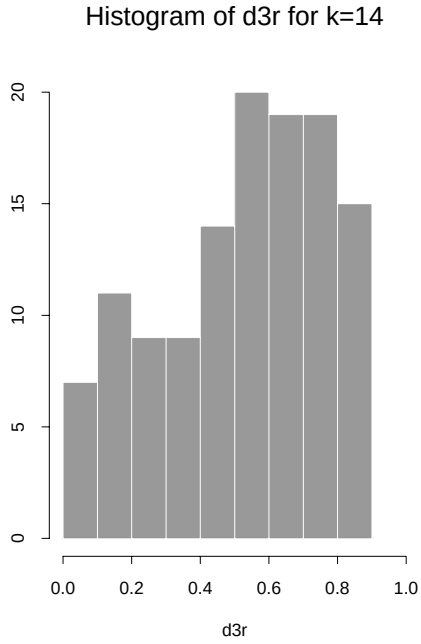
(e) $k = 11$



(f) $k = 12$

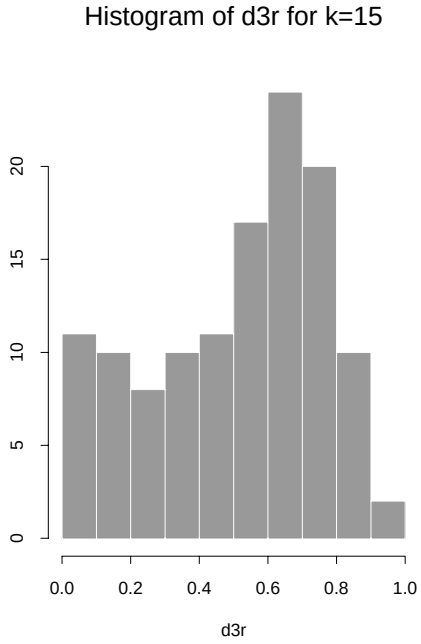


(g) $k = 13$

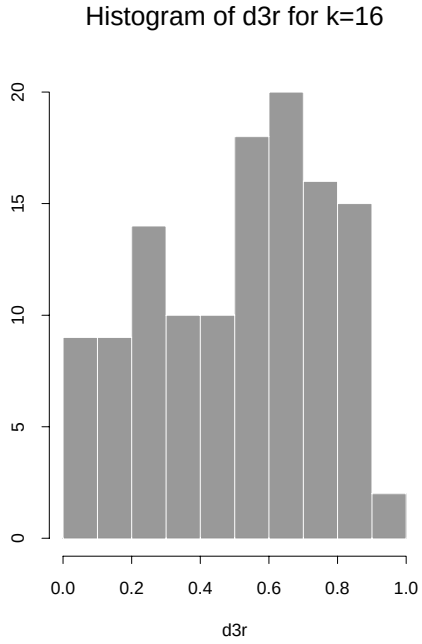


(h) $k = 14$

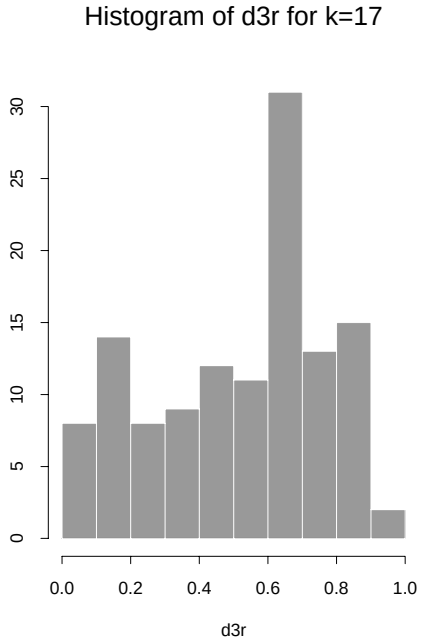
Figure H.9 (continued).



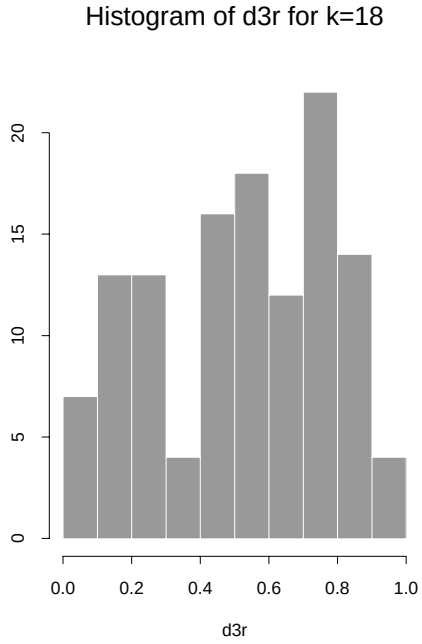
(i) $k = 15$



(j) $k = 16$



(k) $k = 17$



(l) $k = 18$

Figure H.9 (continued).

APPENDIX I
CONFIDENCE REGIONS AND TESTS FOR ISOTROPY / ANISOTROPY

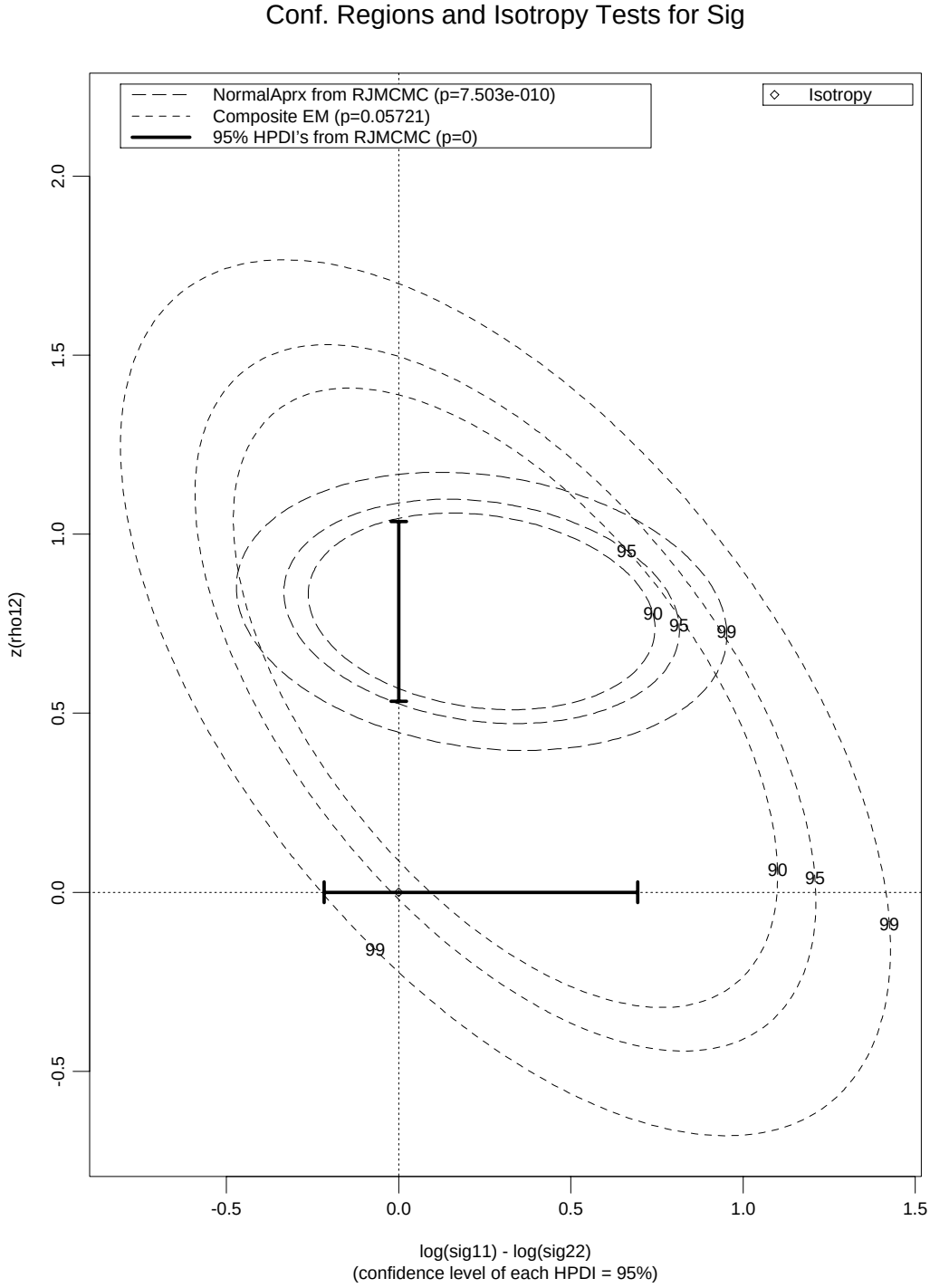
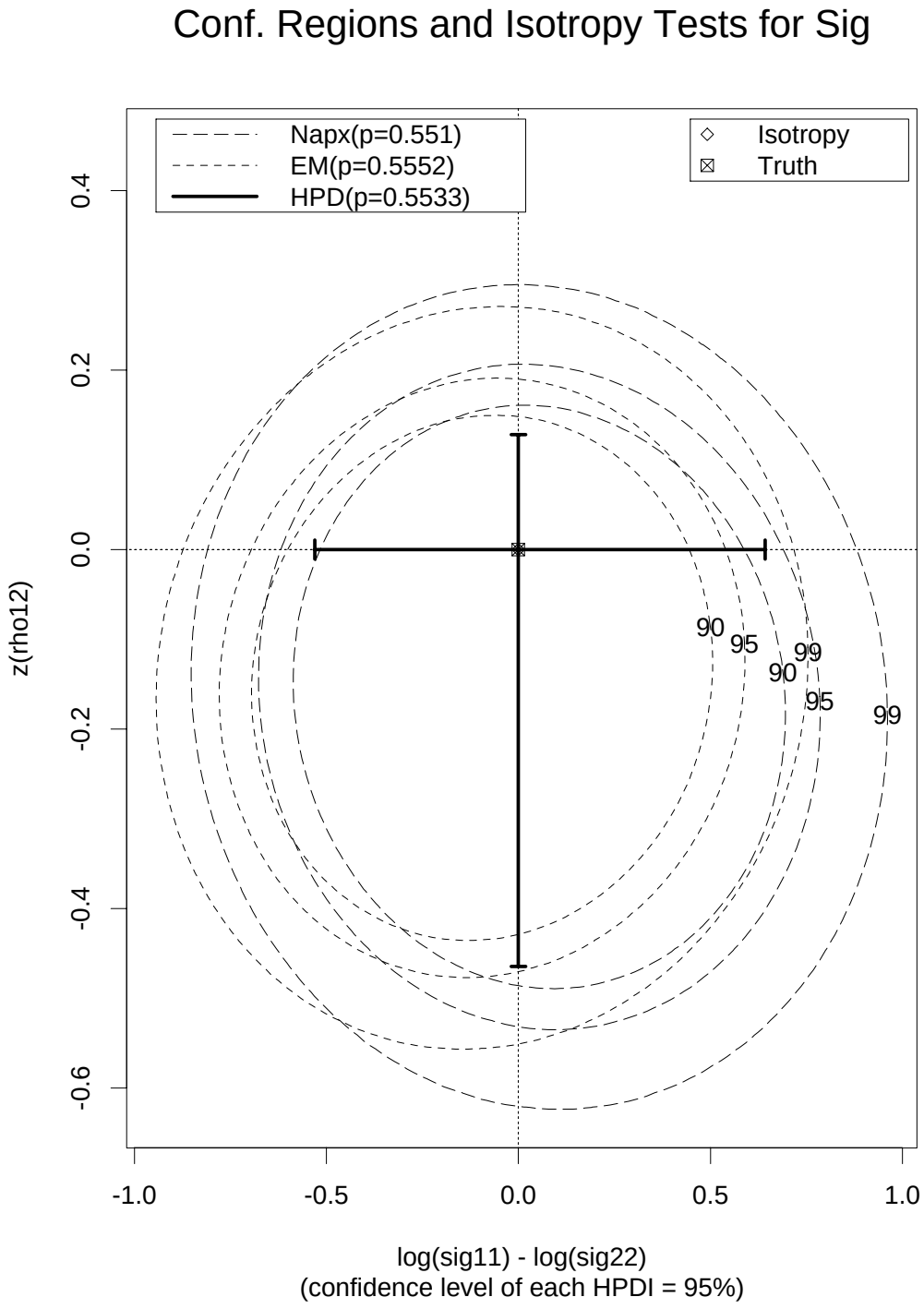
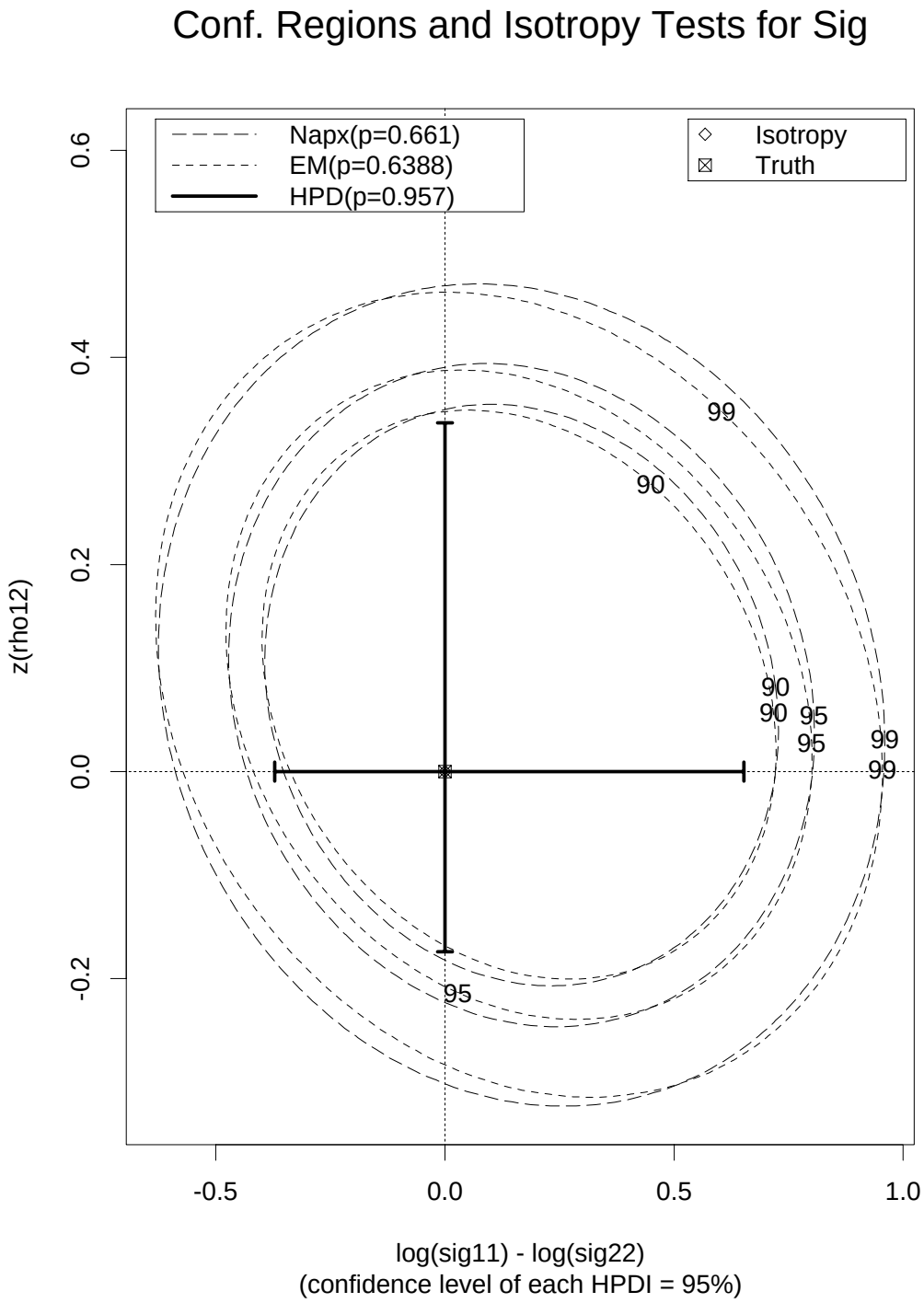


Figure I.1: Confidence regions and tests for isotropy/anisotropy using composite EM and HPDR and normal approximation from RJMCMC, Redwood data.



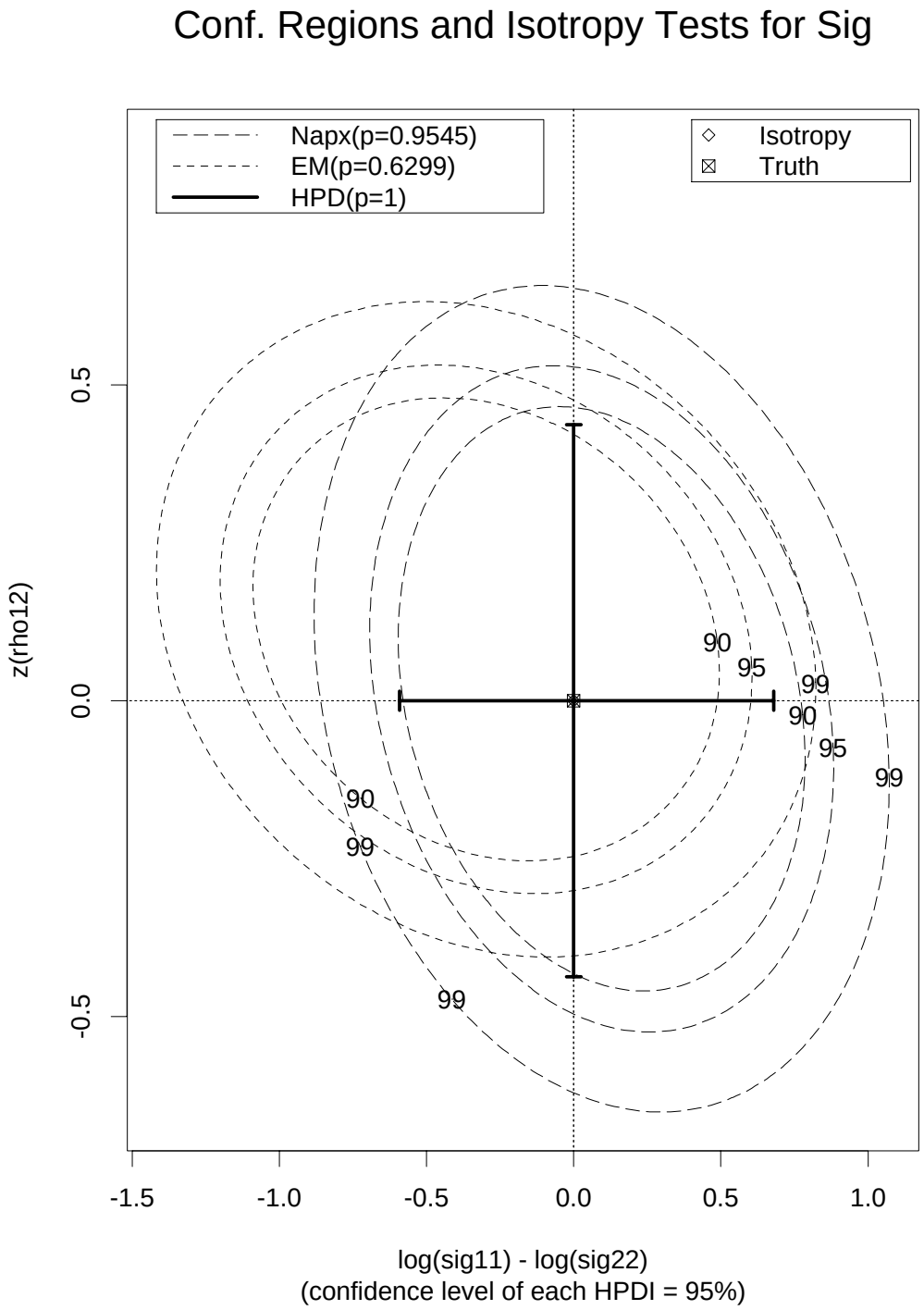
(a) I-k7-a

Figure I.2: Confidence regions and tests for isotropy/anisotropy using composite EM and HPDR and normal approximation from RJMCMC, simulated patterns.



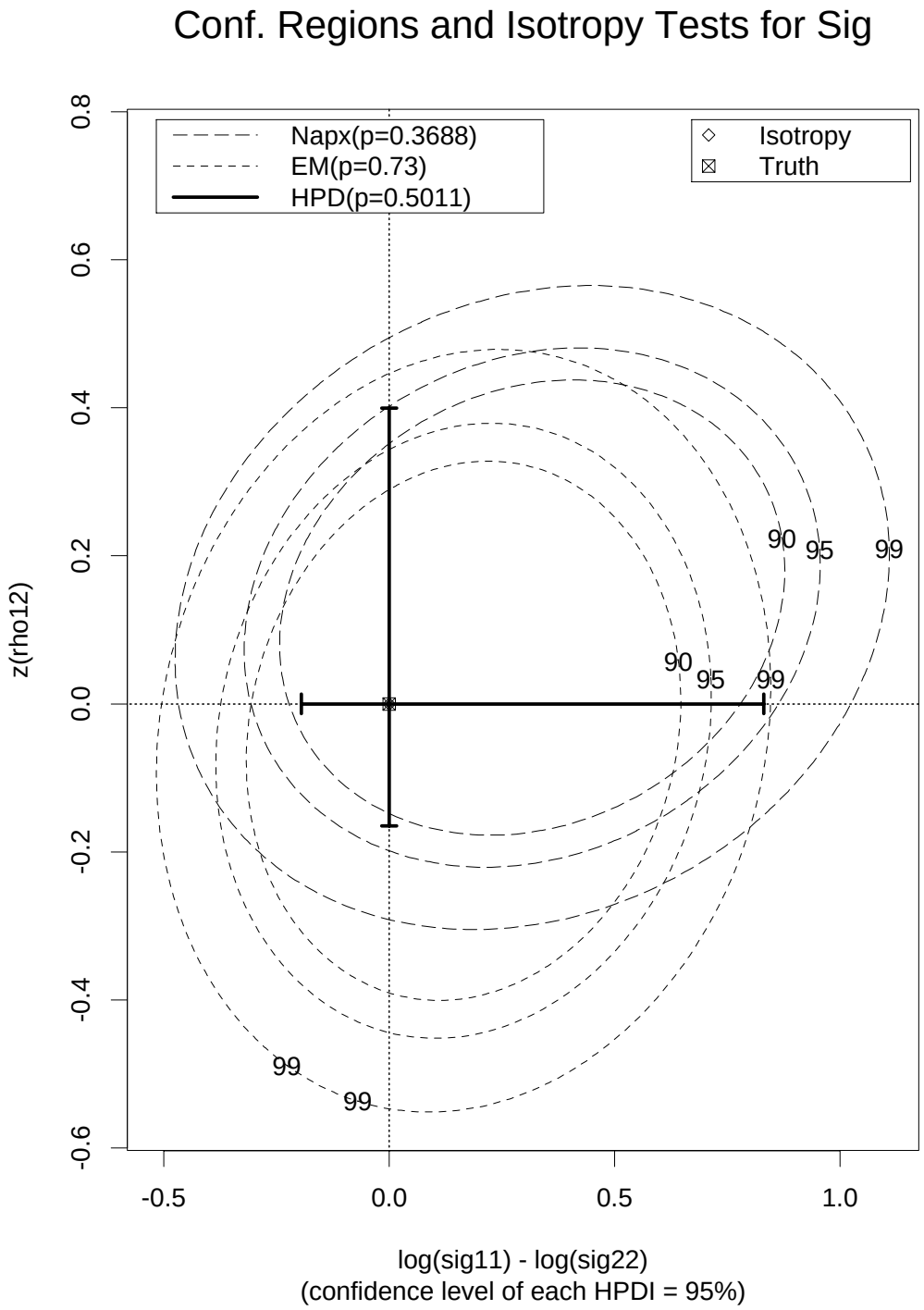
(b) I-k7-b

Figure I.2 (continued).



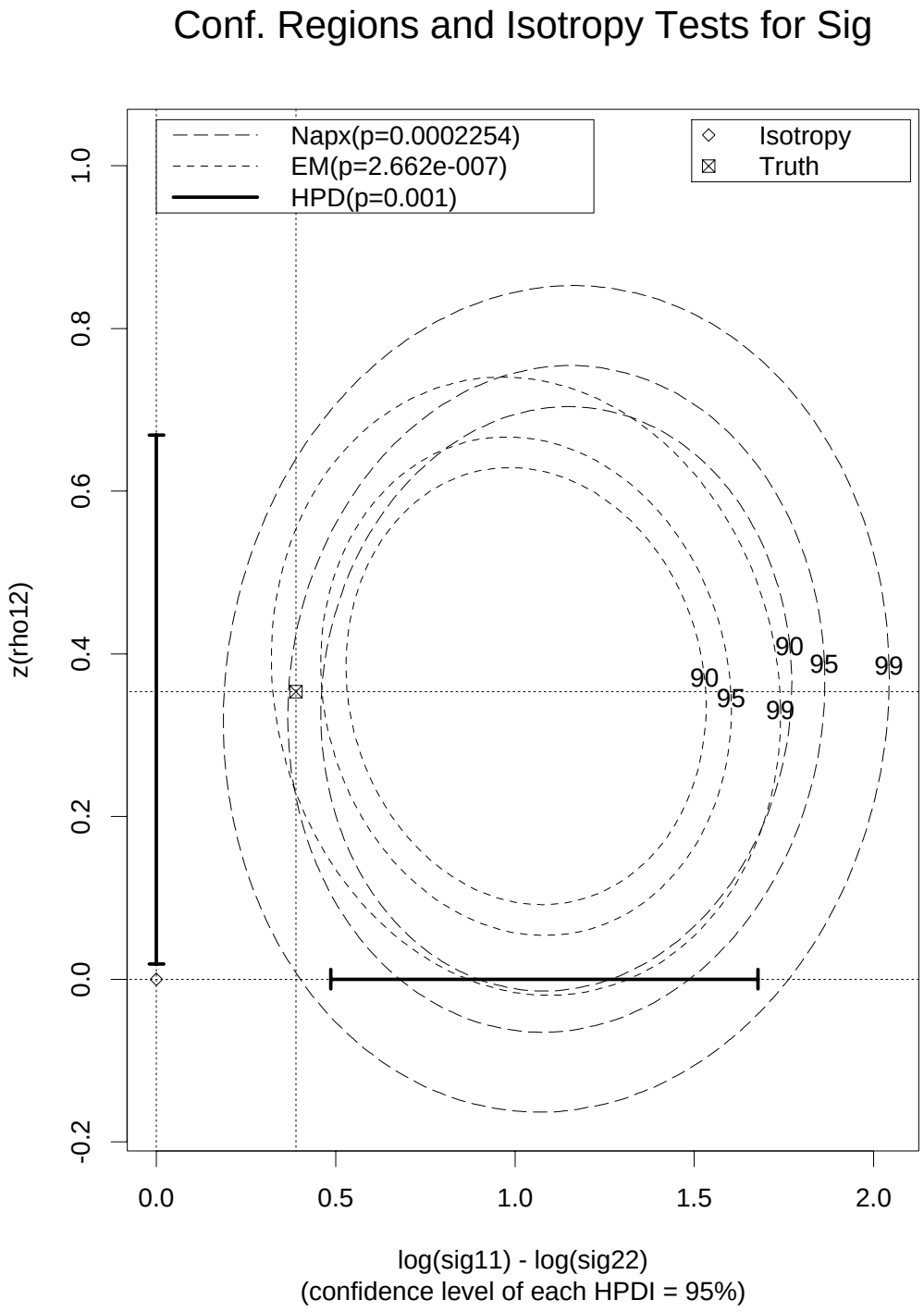
(c) I-k14-a

Figure I.2 (continued).



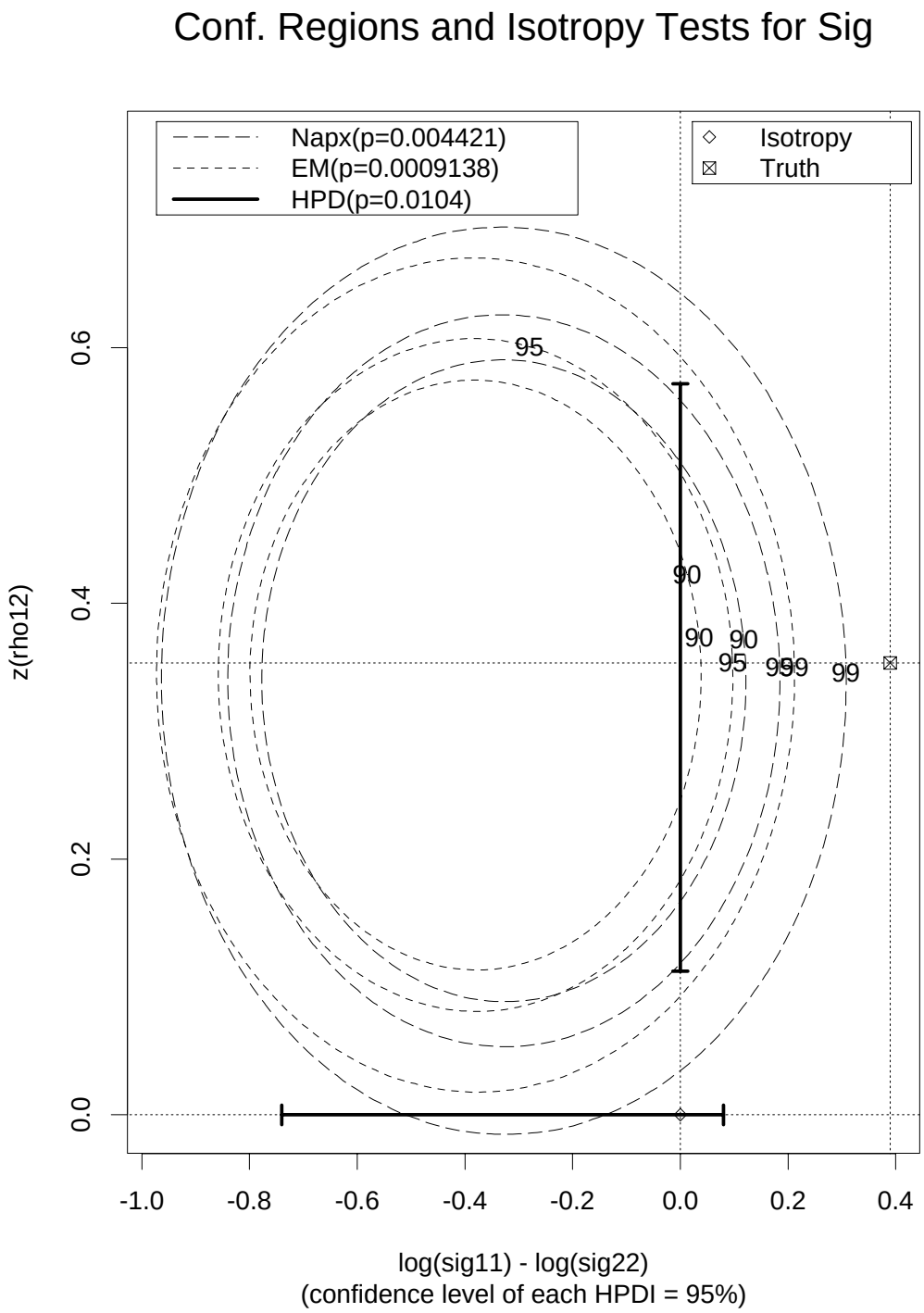
(d) I-k14-b

Figure I.2 (continued).



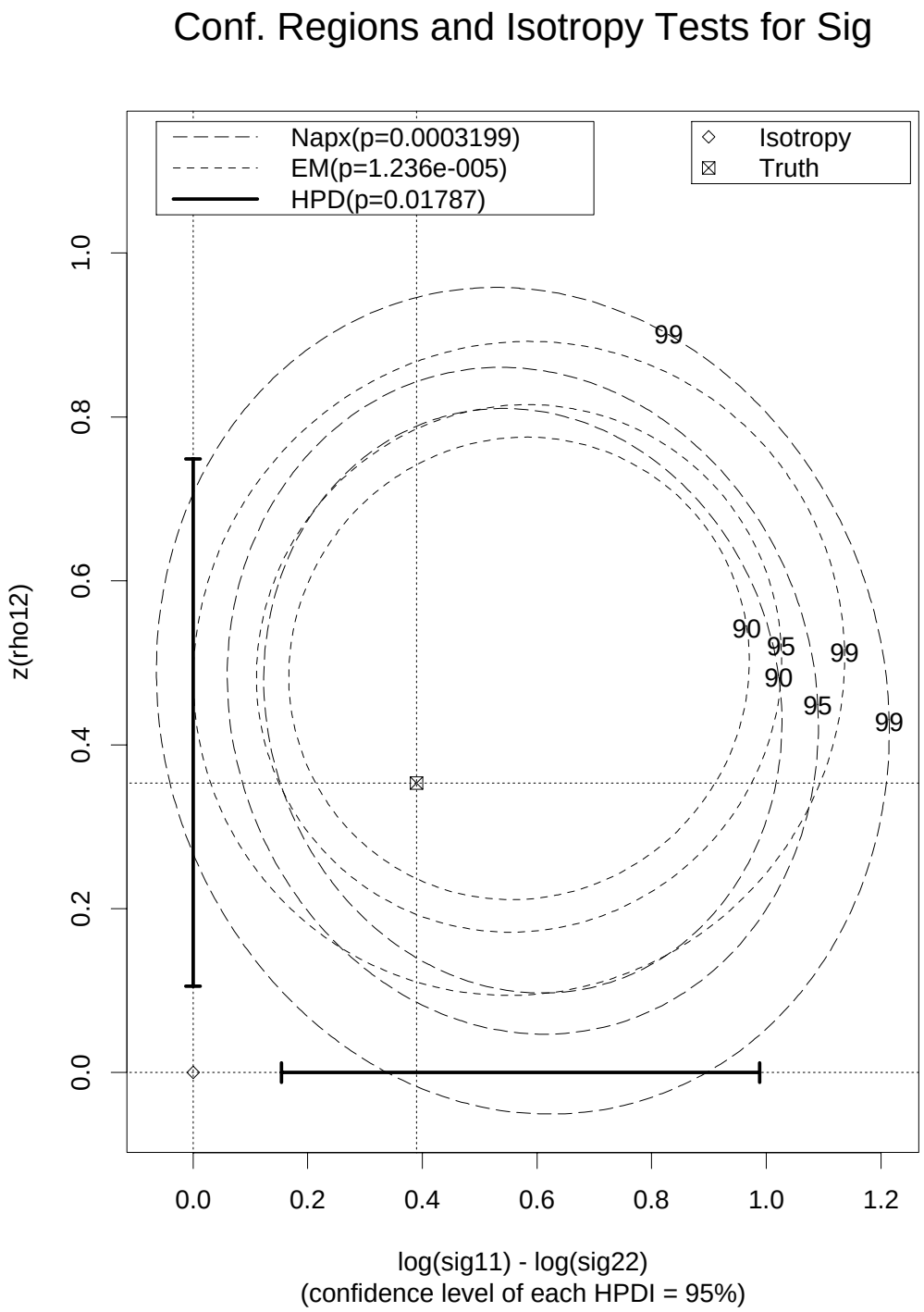
(e) AI-1.5-k7-a

Figure I.2 (continued).



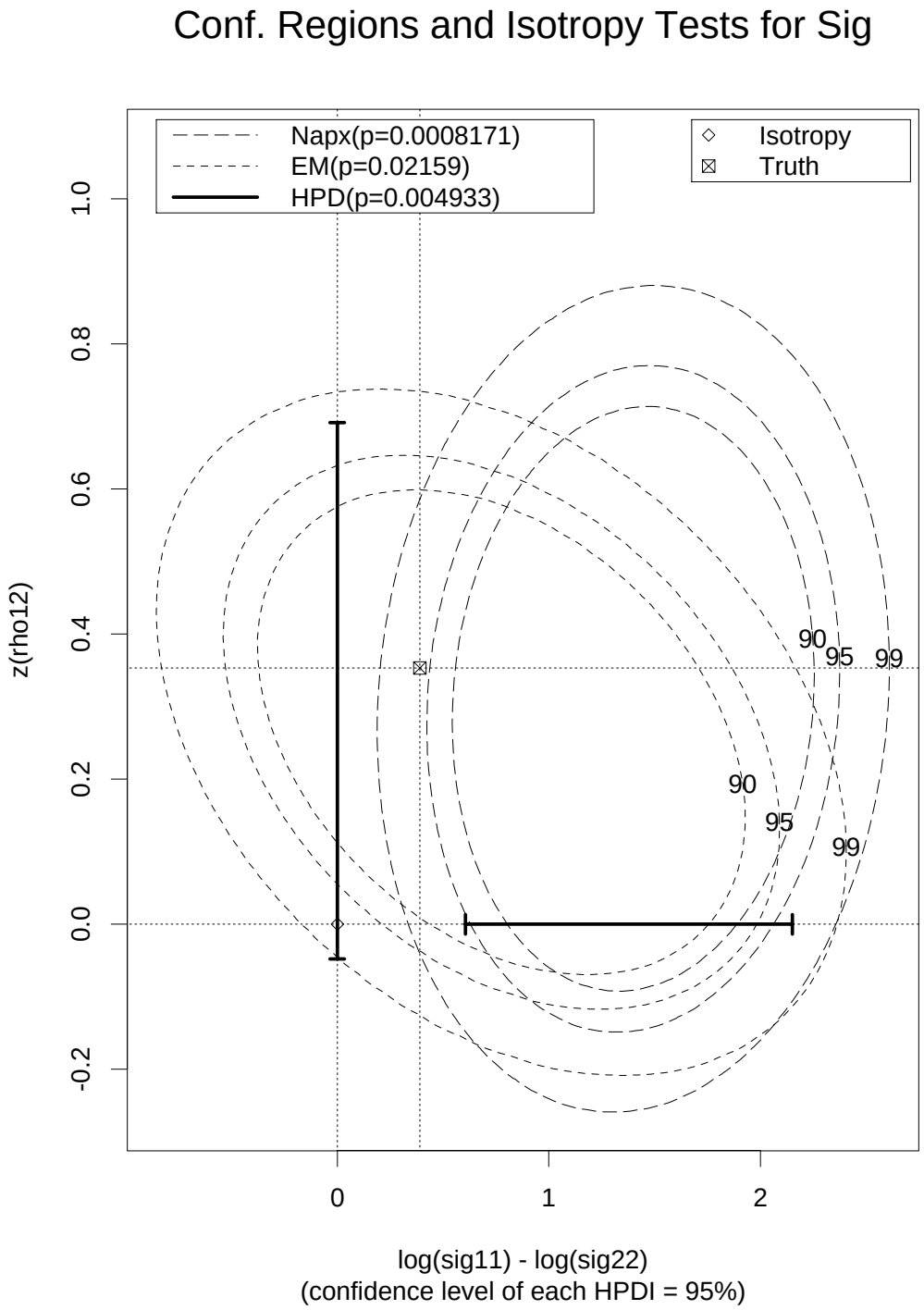
(f) AI-1.5-k7-b

Figure I.2 (continued).



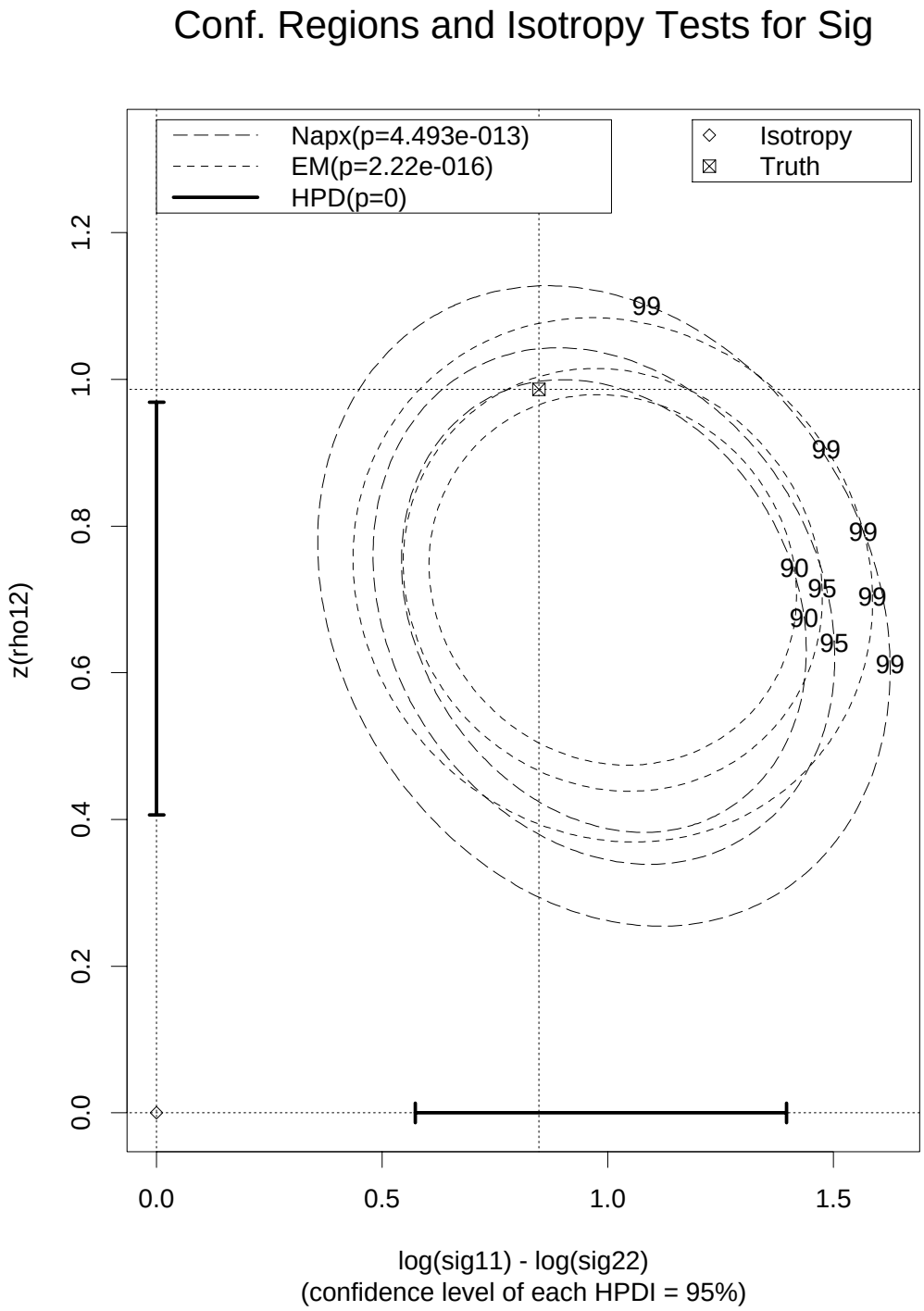
(g) AI-1.5-k14-a

Figure I.2 (continued).



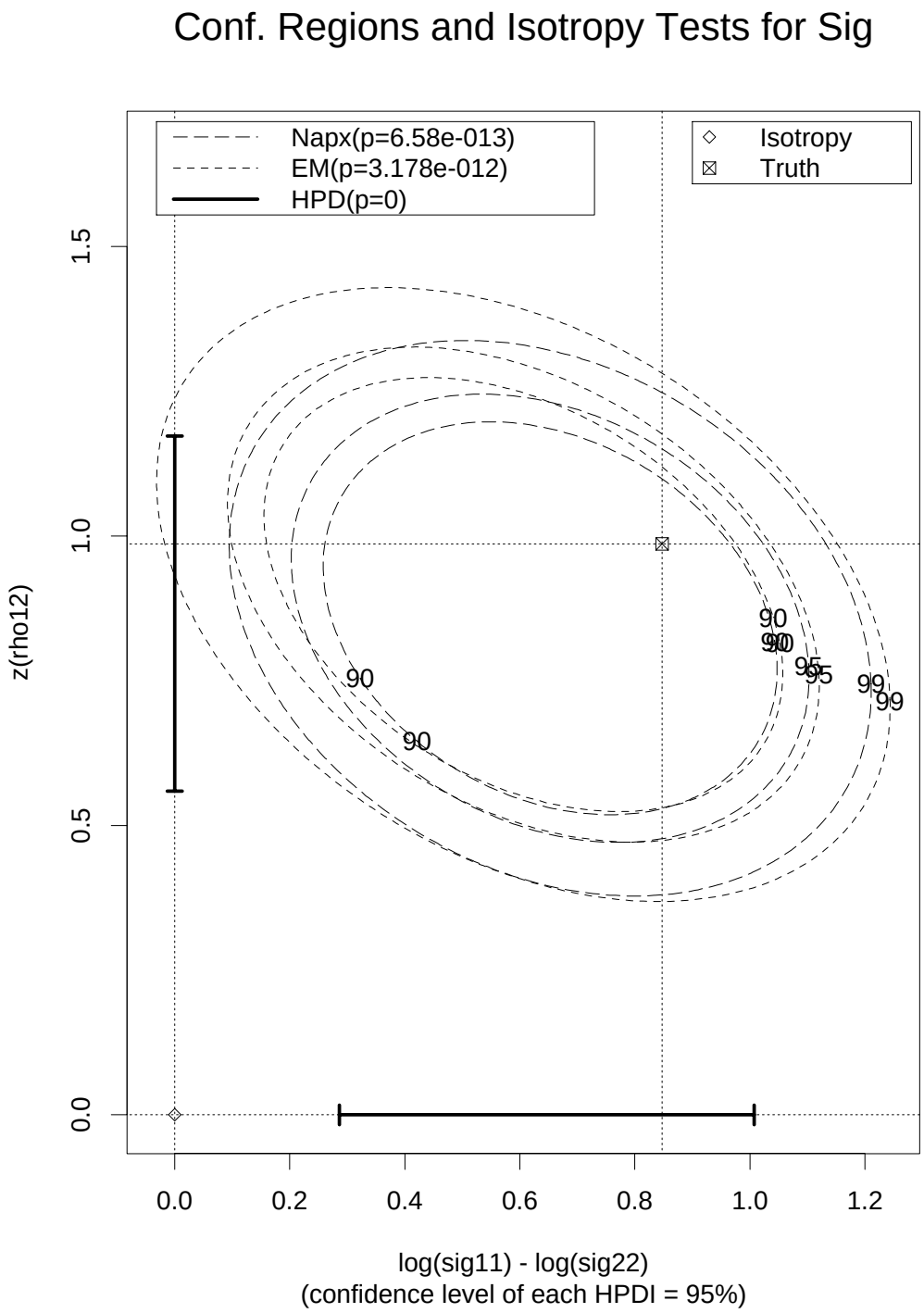
(h) AI-1.5-k14-b

Figure I.2 (continued).



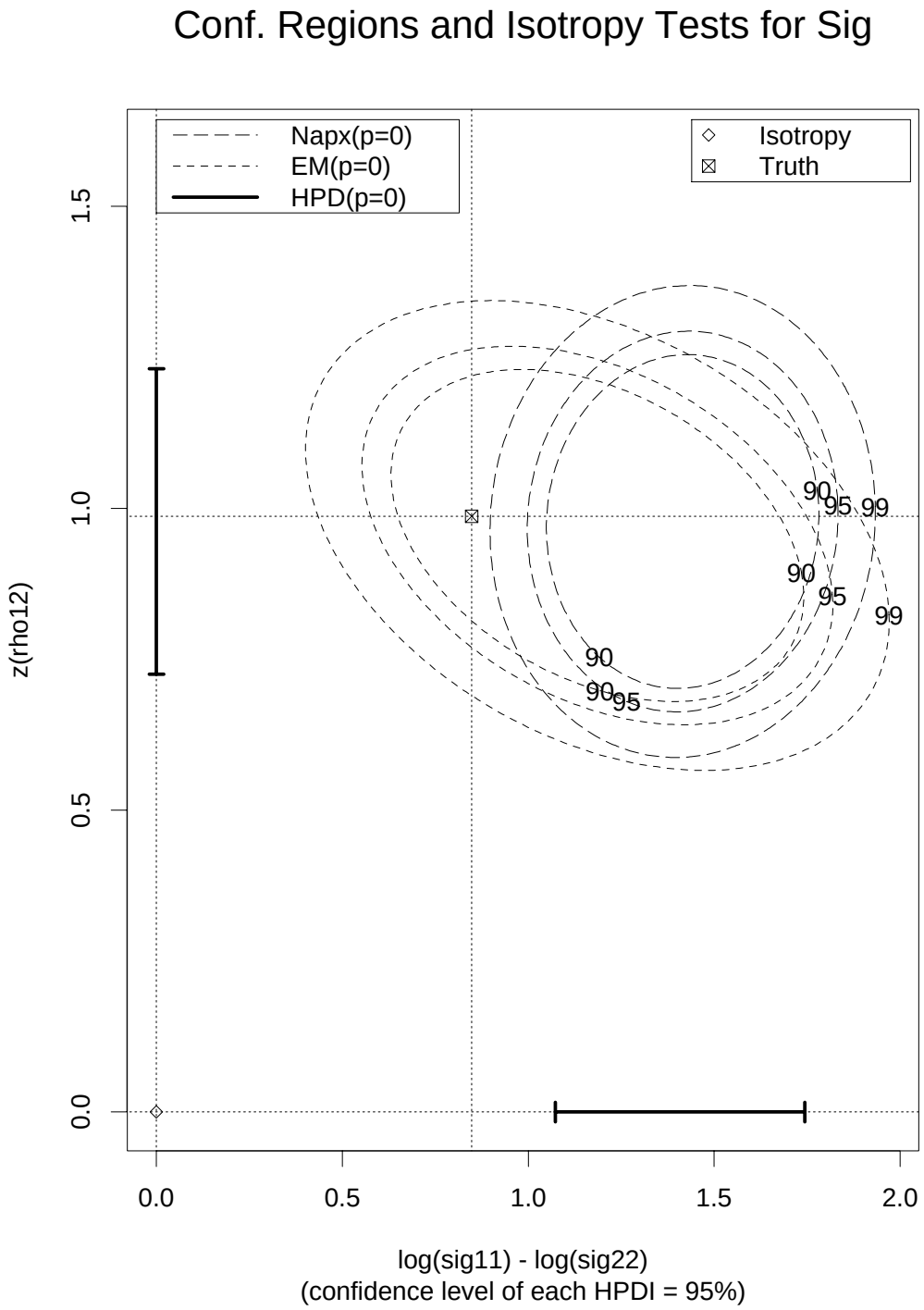
(i) AI-3-k7-a

Figure I.2 (continued).



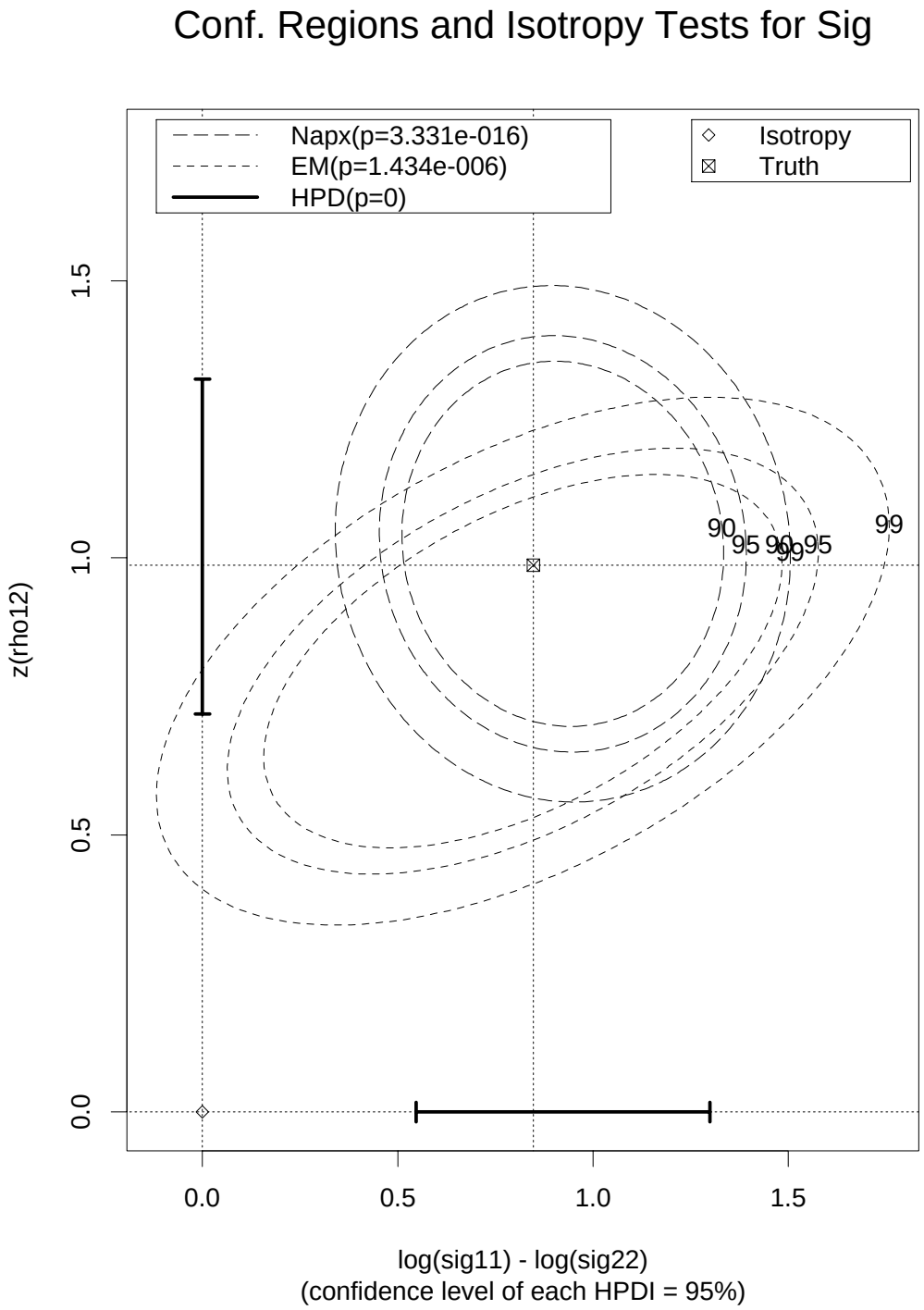
(j) AI-3-k7-b

Figure I.2 (continued).



(k) AI-3-k14-a

Figure I.2 (continued).



(l) AI-3-k14-b

Figure I.2 (continued).

APPENDIX J
POSTERIOR DENSITY ESTIMATES AND COMPONENTWISE CONFIDENCE INTERVALS FOR SIGMA PARAMETERS

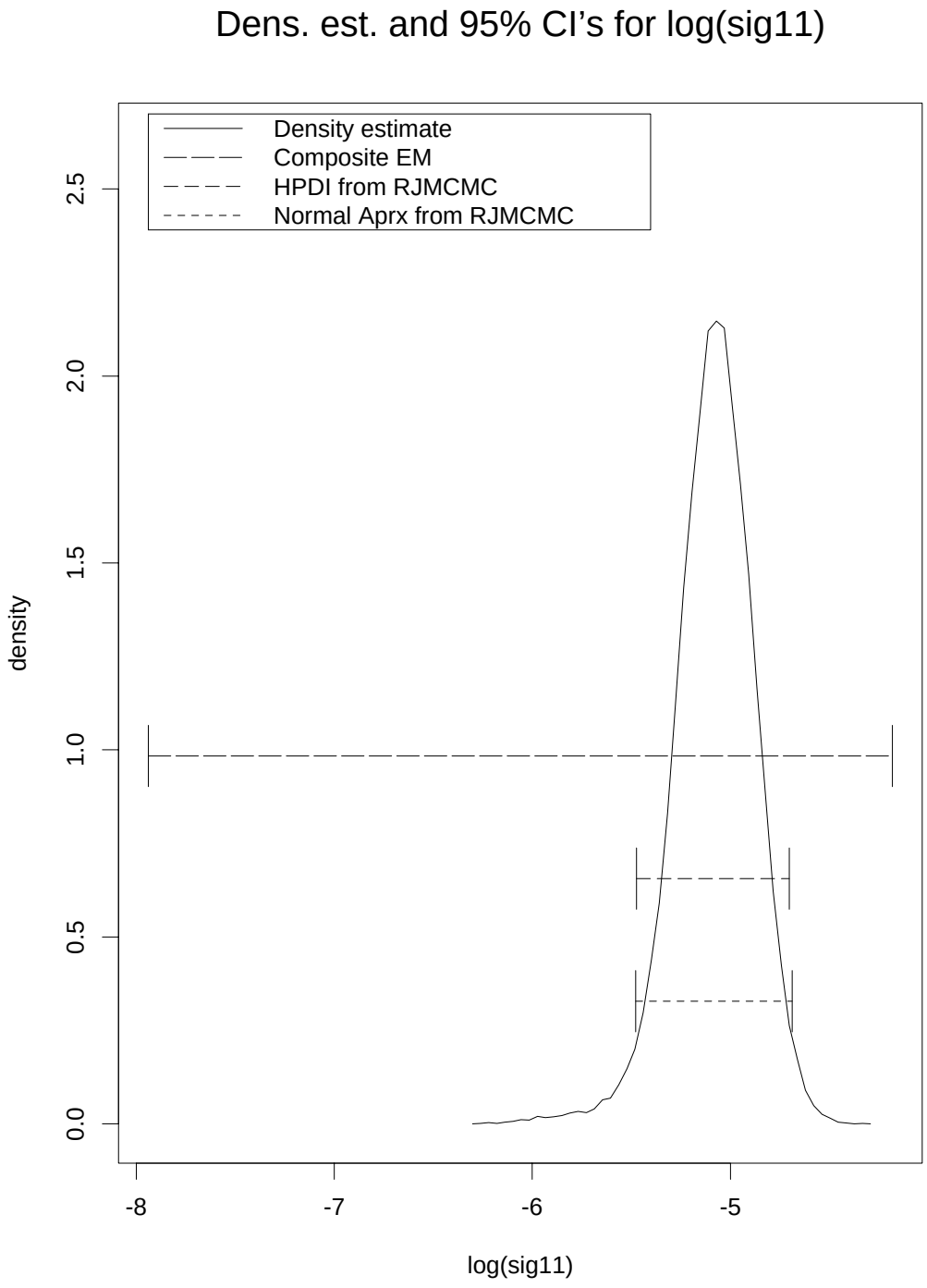
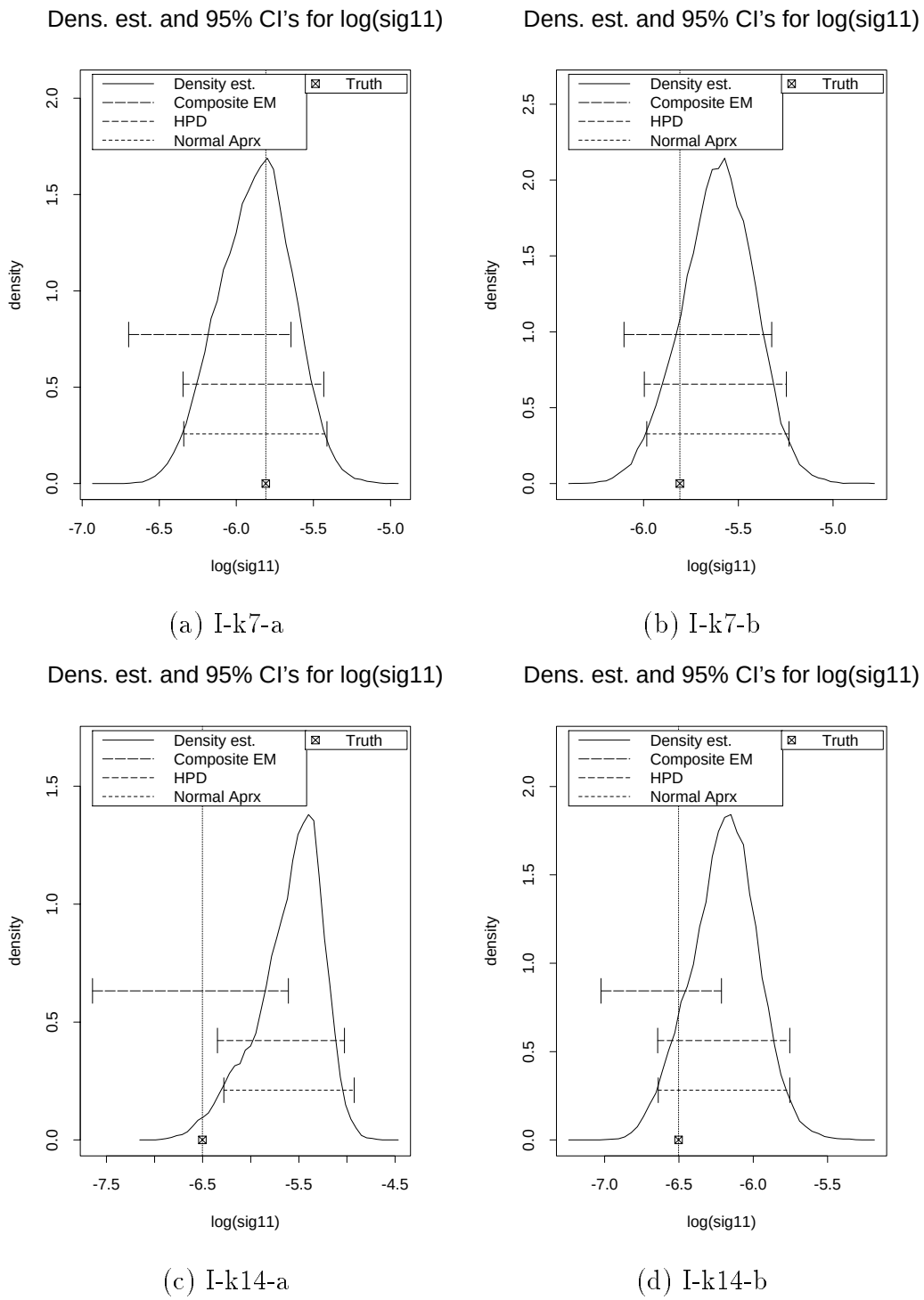


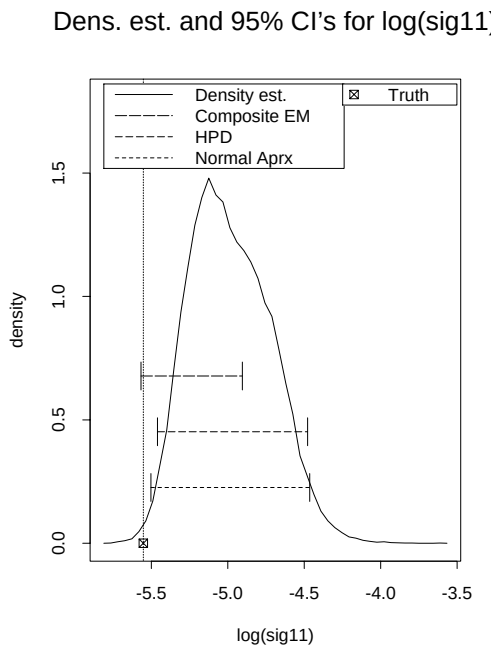
Figure J.1: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log(\sigma_{11})$, using composite EM, HPDR and normal approximation, Redwood data.



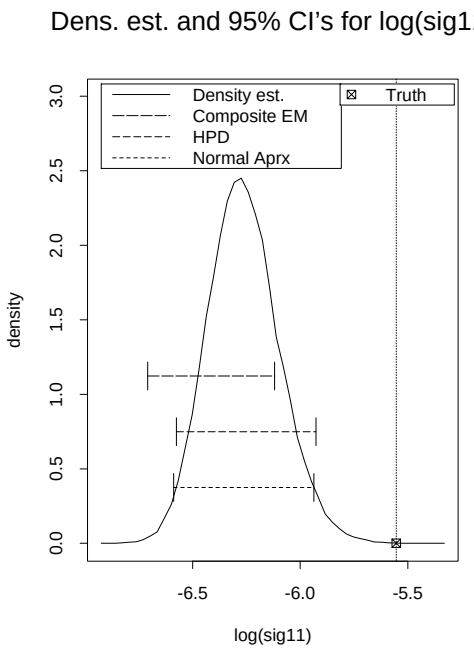
(a) I-k7-a (b) I-k7-b

(c) I-k14-a (d) I-k14-b

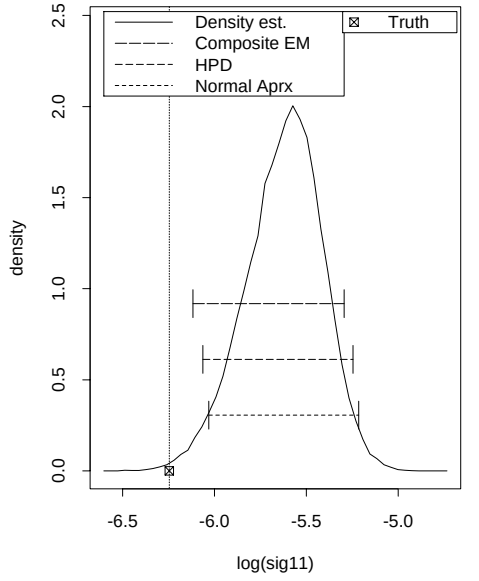
Figure J.2: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log(\sigma_{11})$, using composite EM, HPDR and normal approximation, simulated patterns.



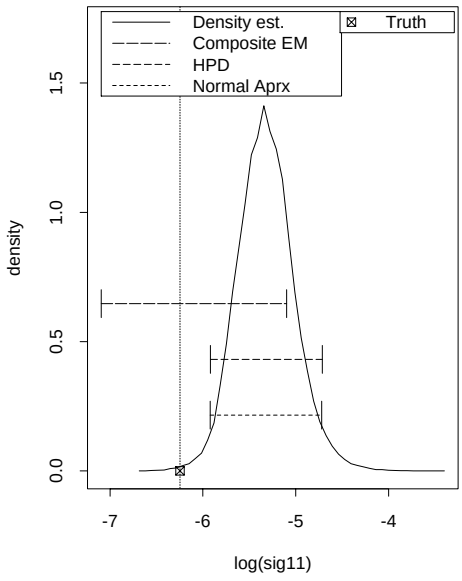
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

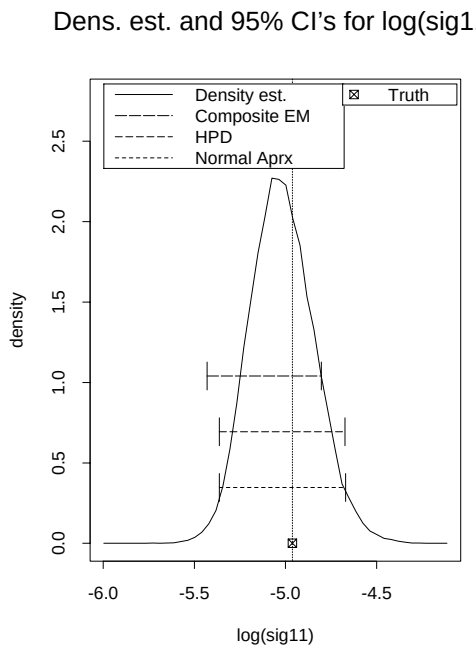


(g) AI-1.5-k14-a

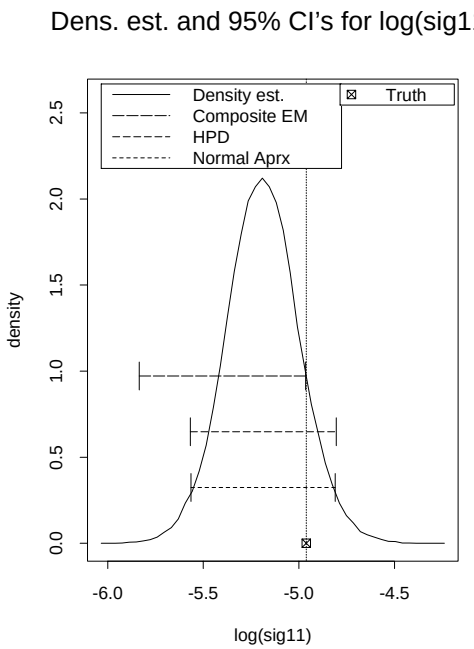


(h) AI-1.5-k14-b

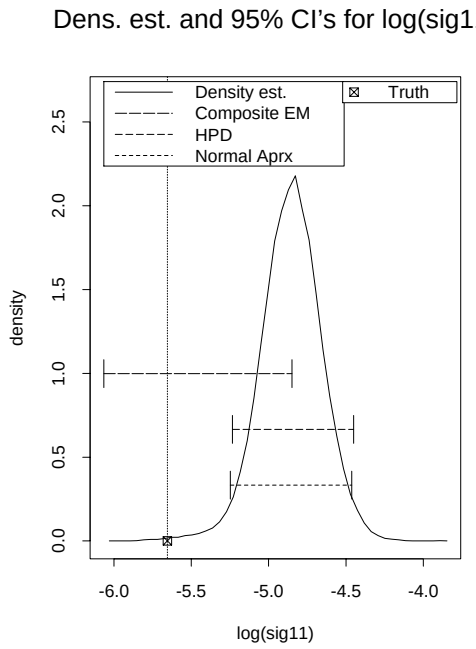
Figure J.2 (continued).



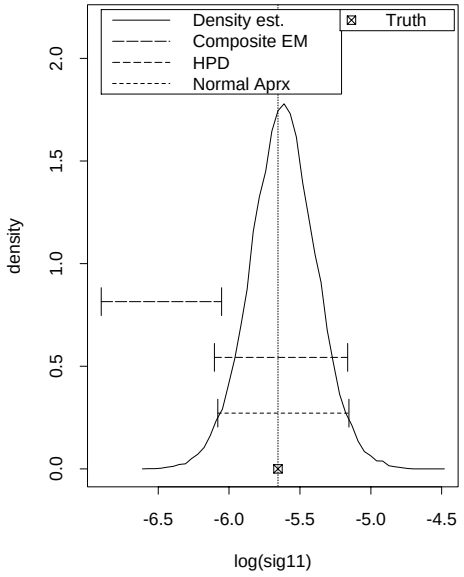
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure J.2 (continued).

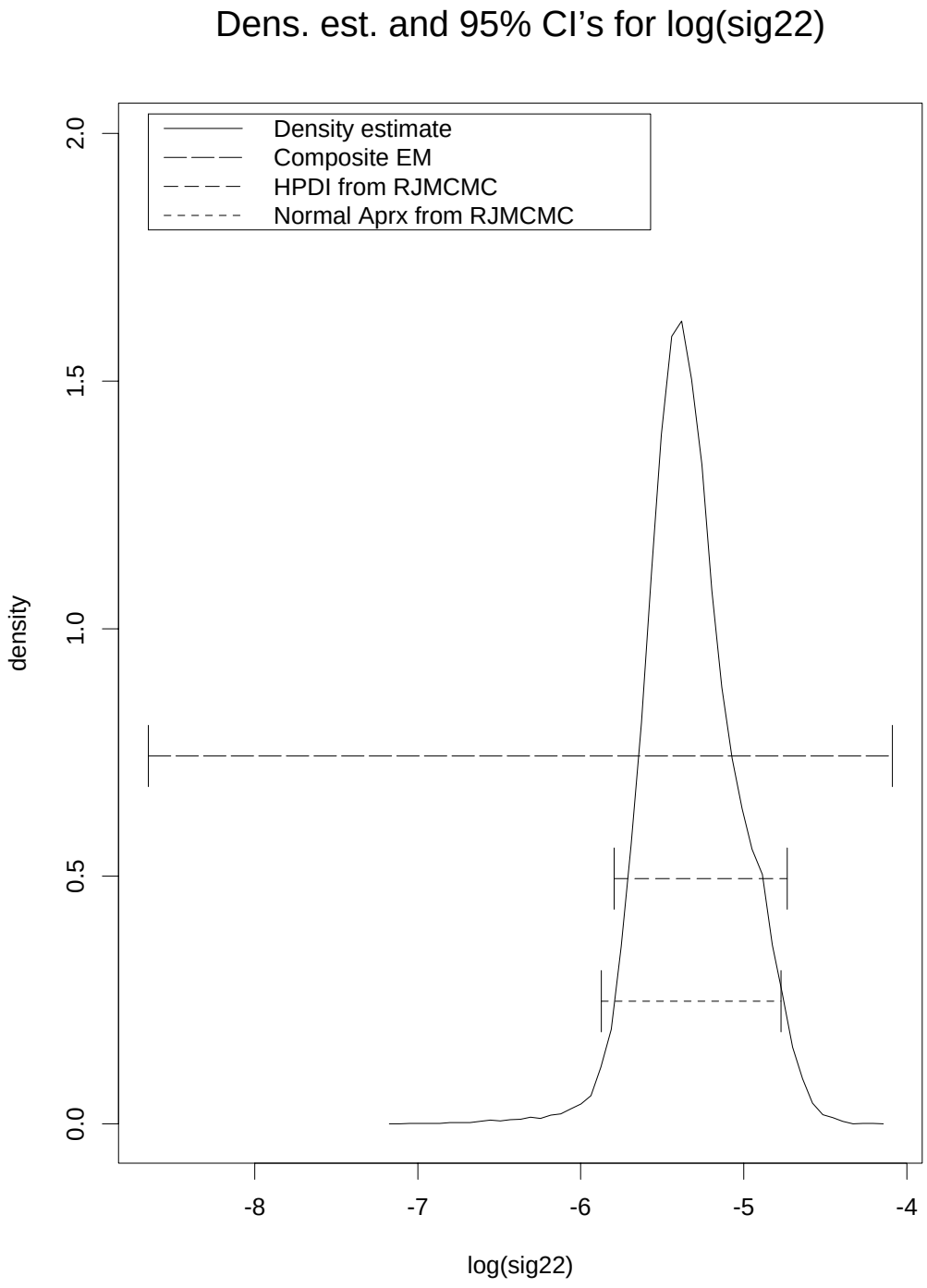
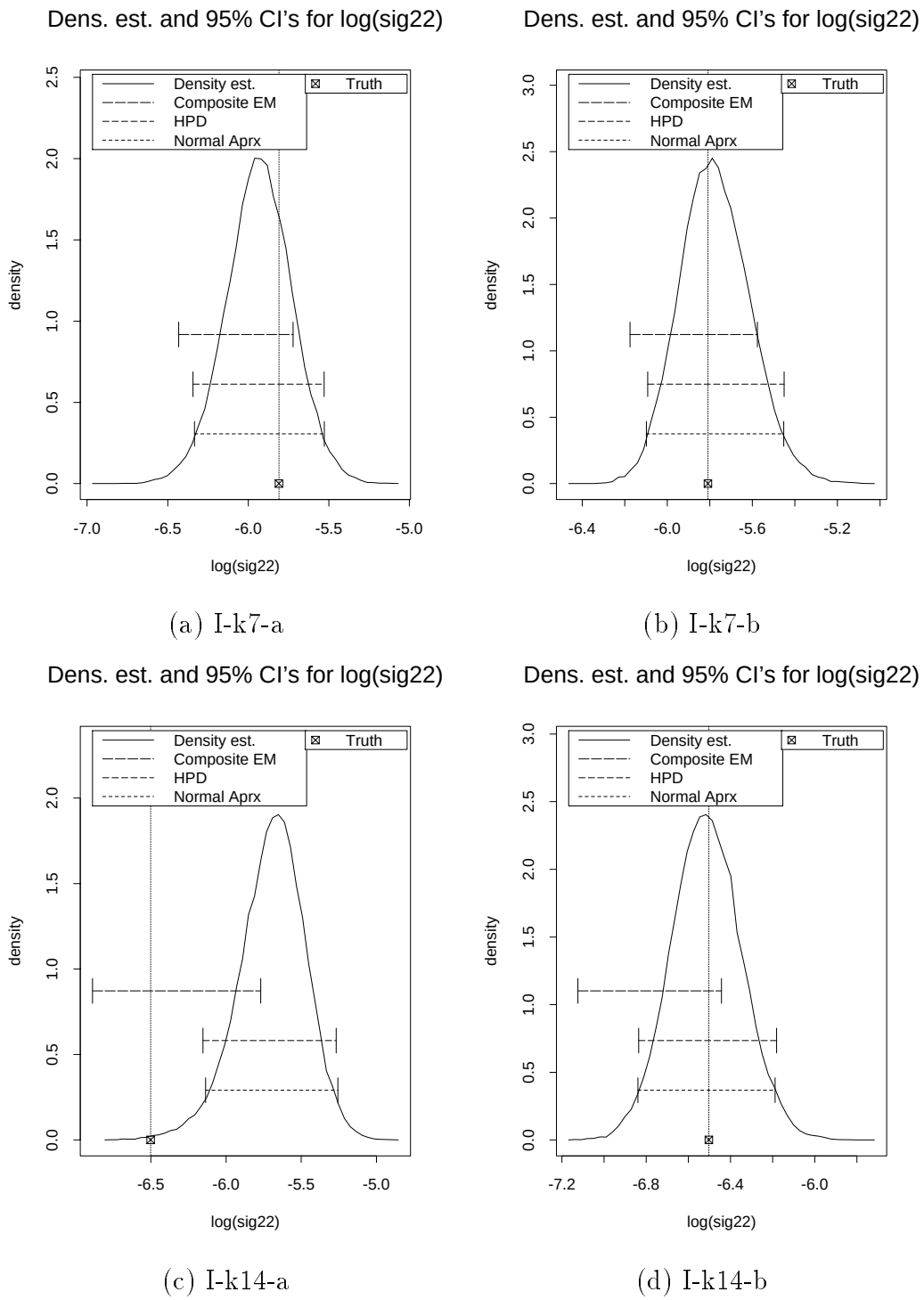


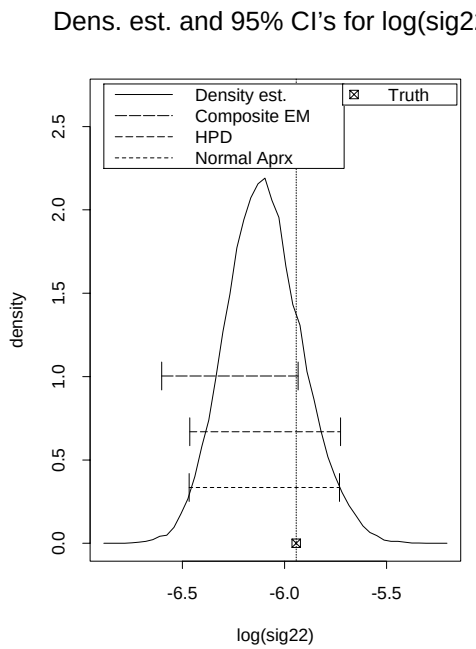
Figure J.3: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log(\sigma_{22})$, using composite EM, HPDR and normal approximation, Redwood data.



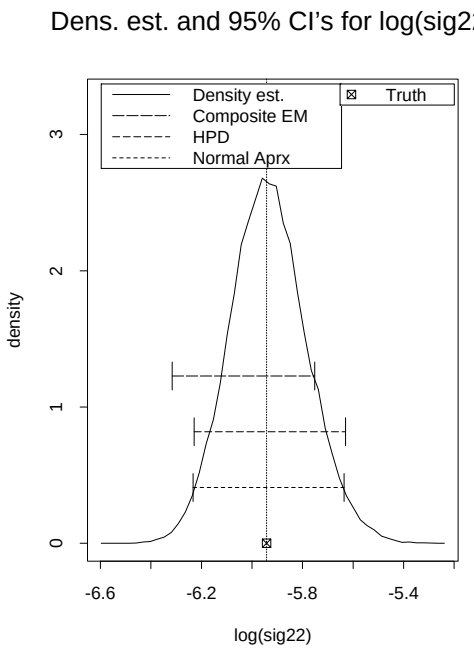
(a) I-k7-a (b) I-k7-b

(c) I-k14-a (d) I-k14-b

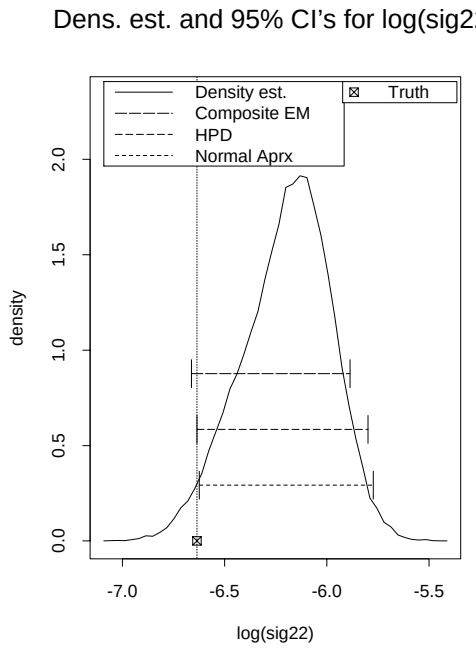
Figure J.4: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log(\sigma_{22})$, using composite EM, HPDR and normal approximation, simulated patterns.



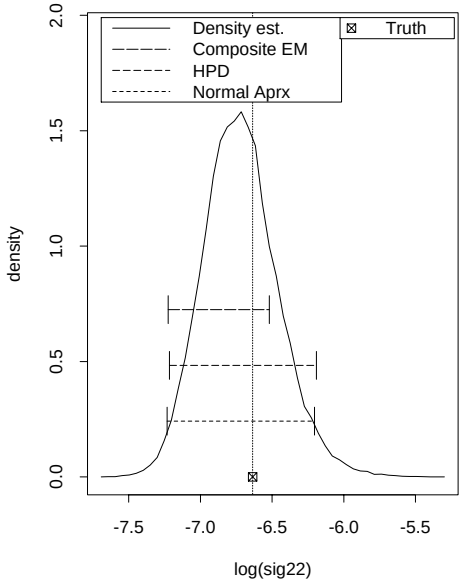
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

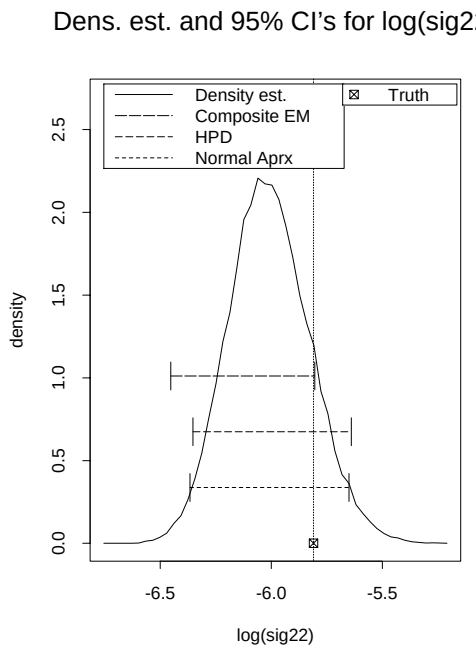


(g) AI-1.5-k14-a

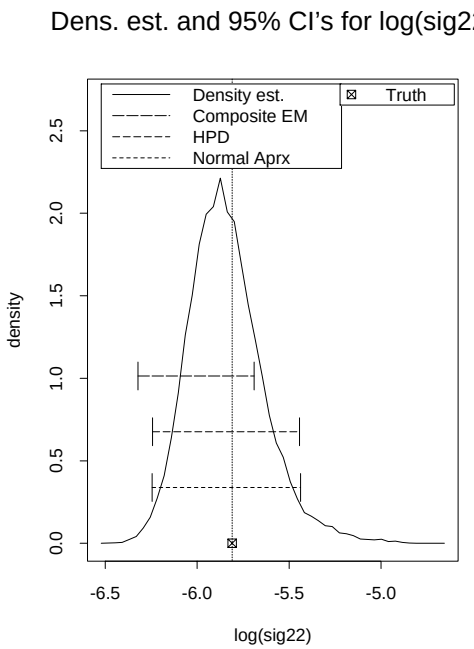


(h) AI-1.5-k14-b

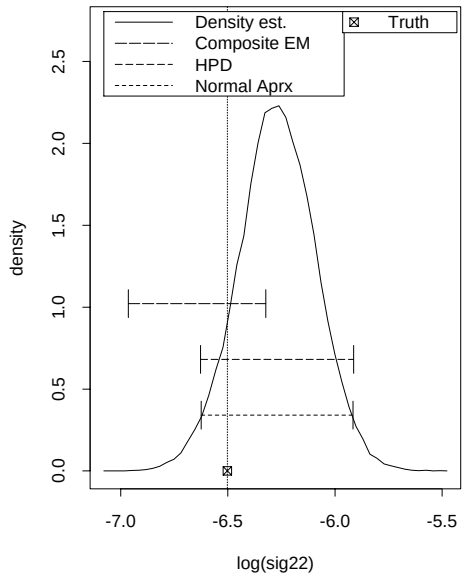
Figure J.4 (continued).



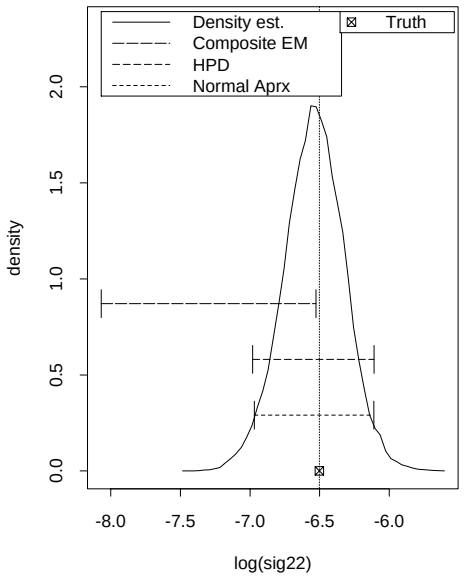
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure J.4 (continued).

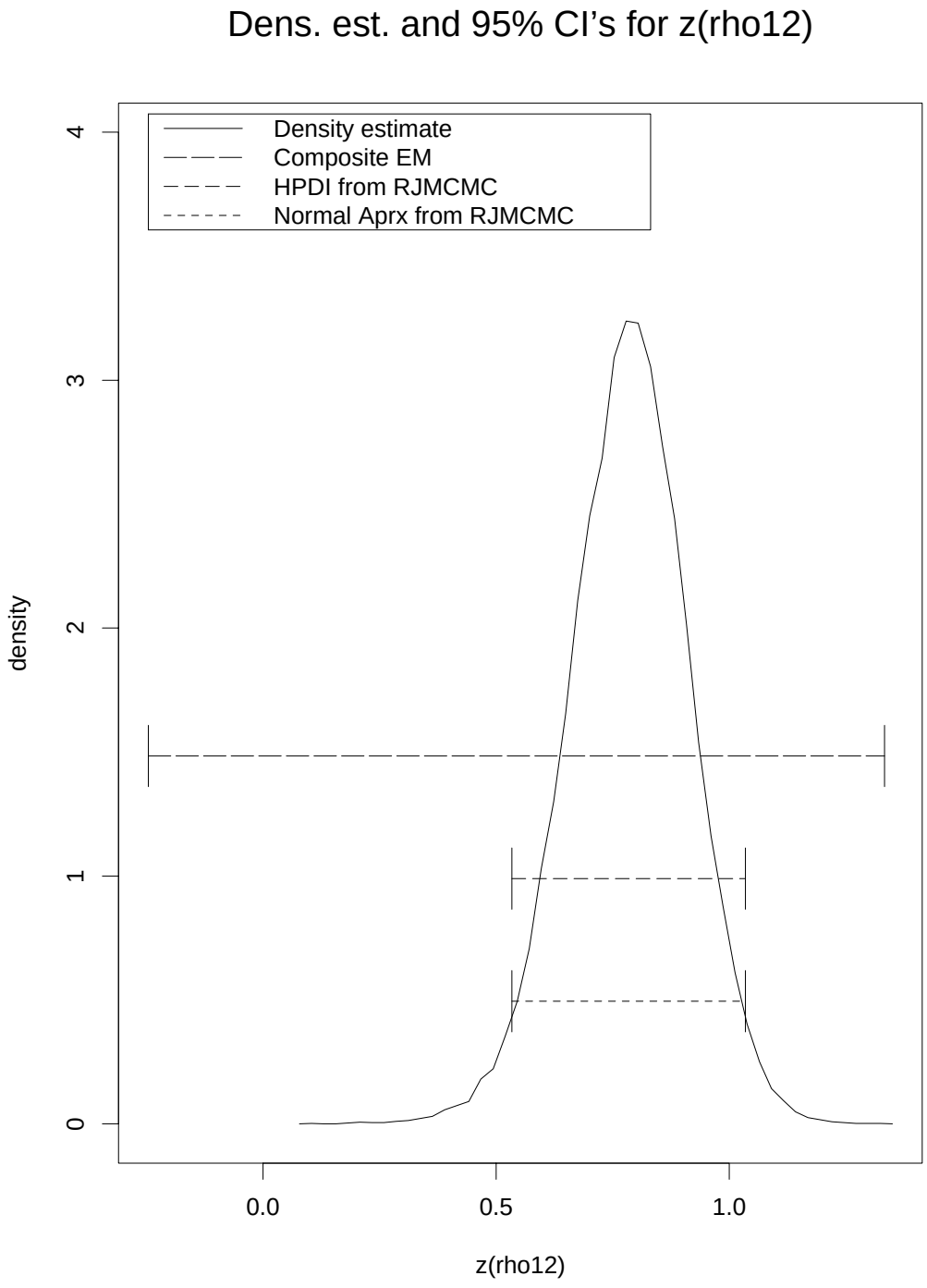
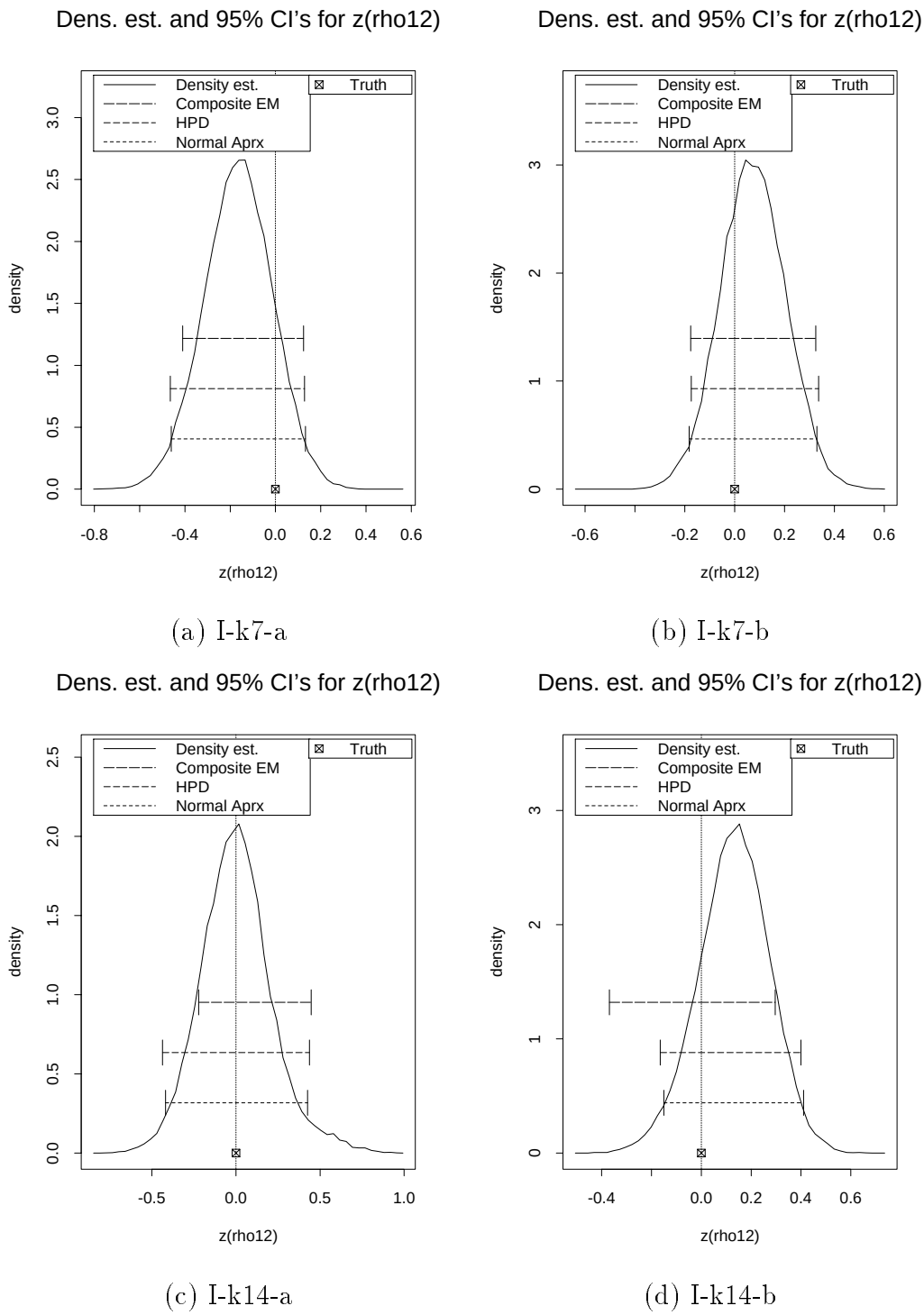


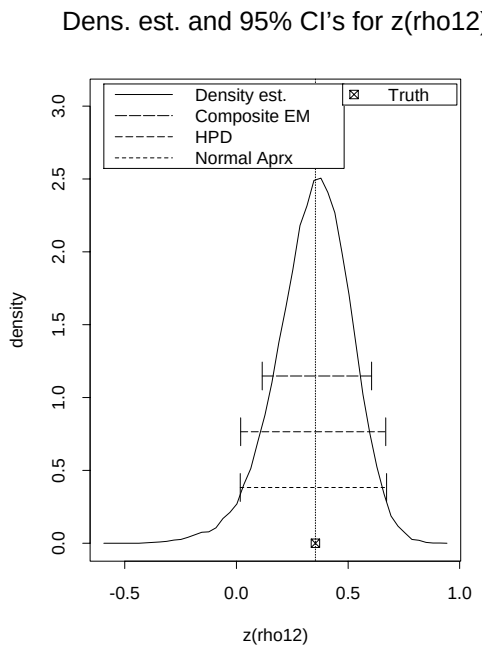
Figure J.5: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $z(\rho_{12})$, using composite EM, HPDR and normal approximation, Redwood data.



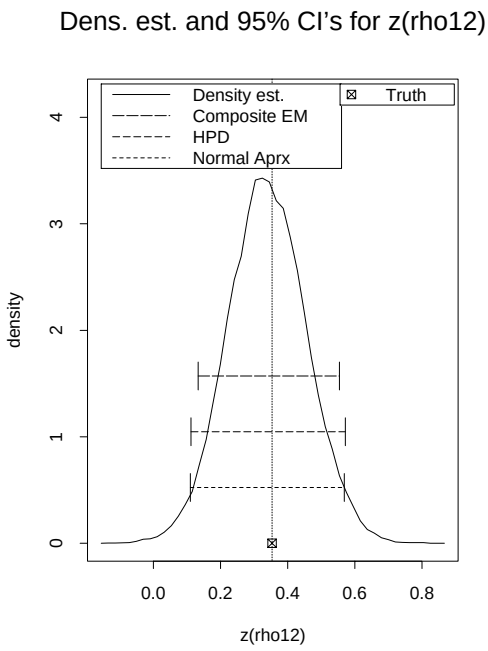
(a) I-k7-a (b) I-k7-b

(c) I-k14-a (d) I-k14-b

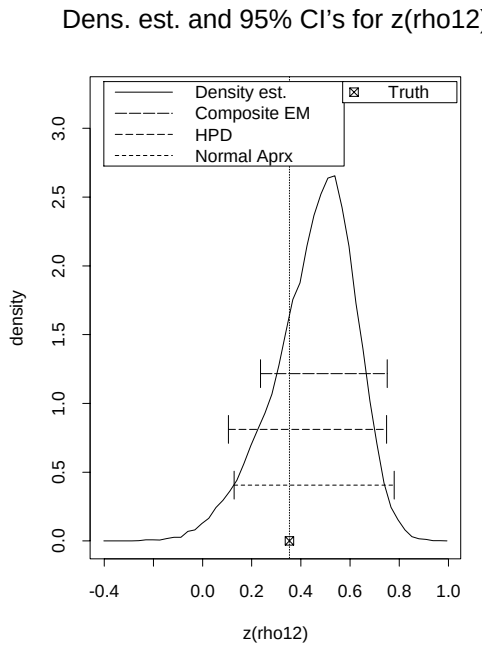
Figure J.6: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $z(\rho_{12})$, using composite EM, HPDR and normal approximation, simulated patterns.



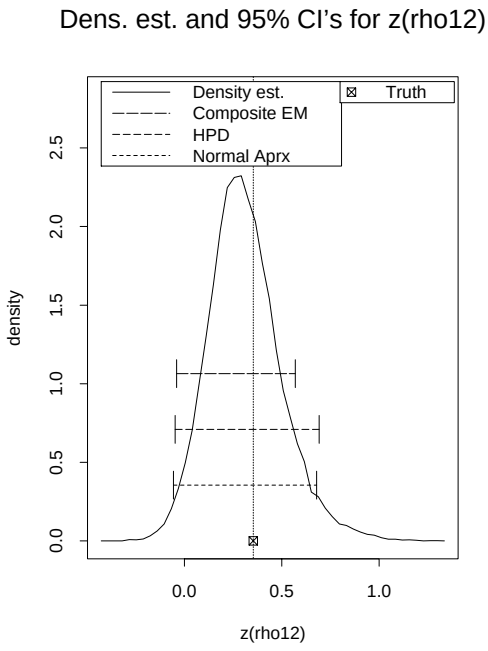
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

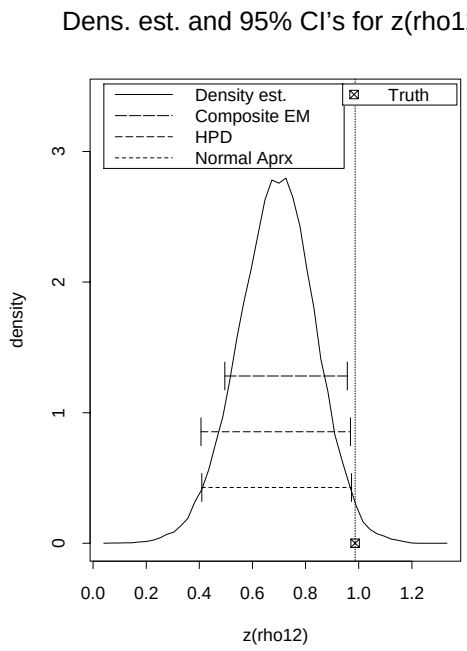


(g) AI-1.5-k14-a

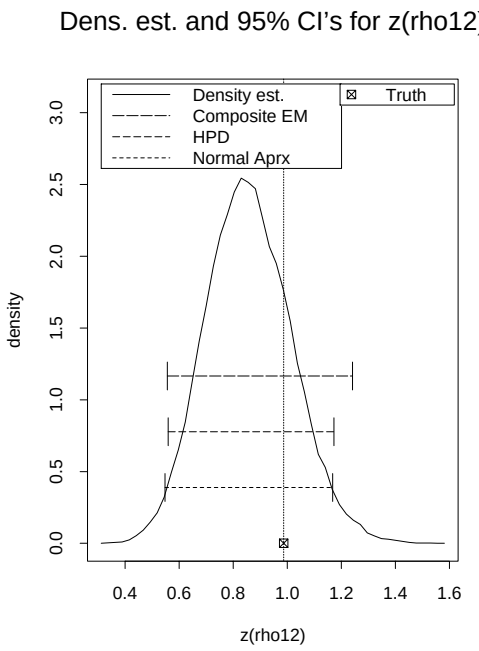


(h) AI-1.5-k14-b

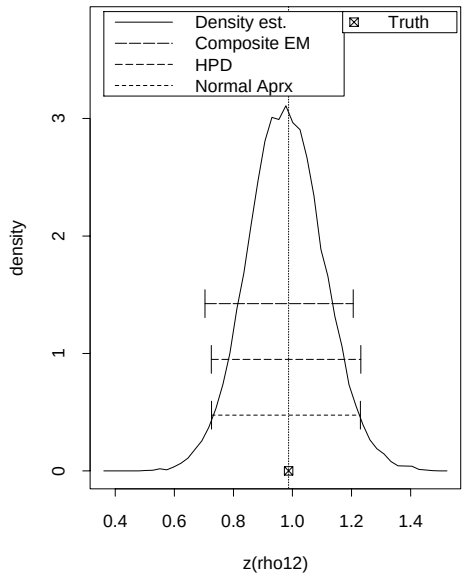
Figure J.6 (continued).



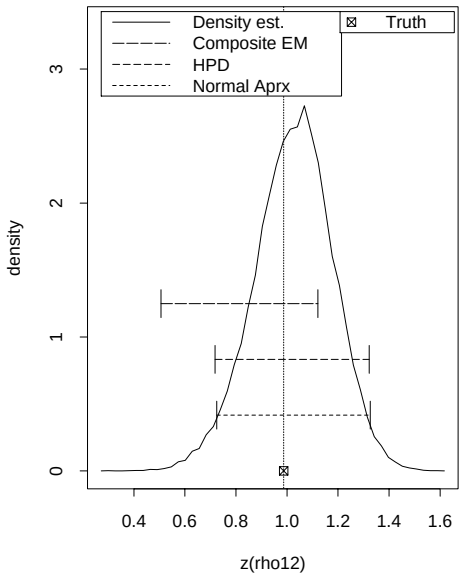
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure J.6 (continued).

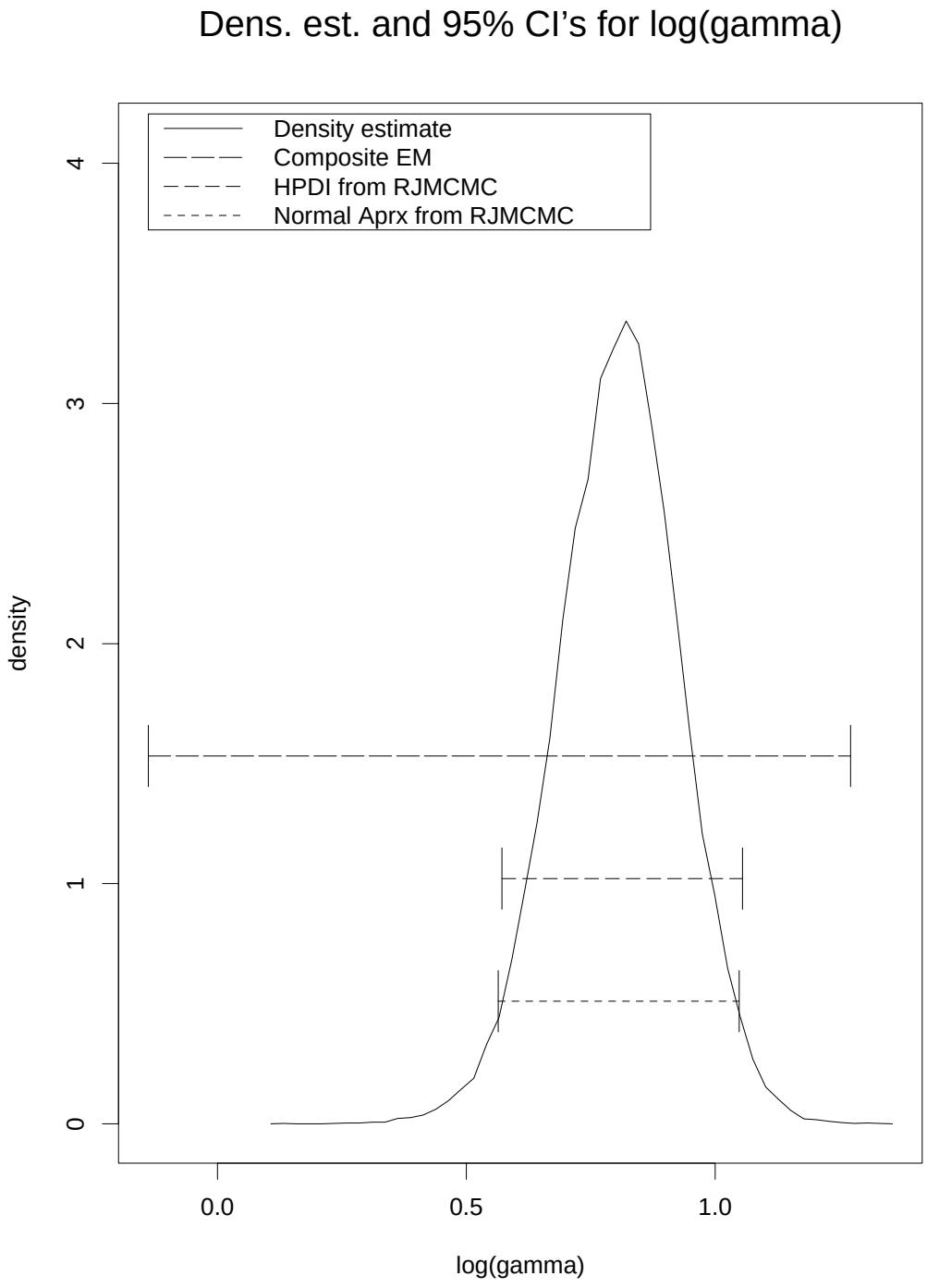
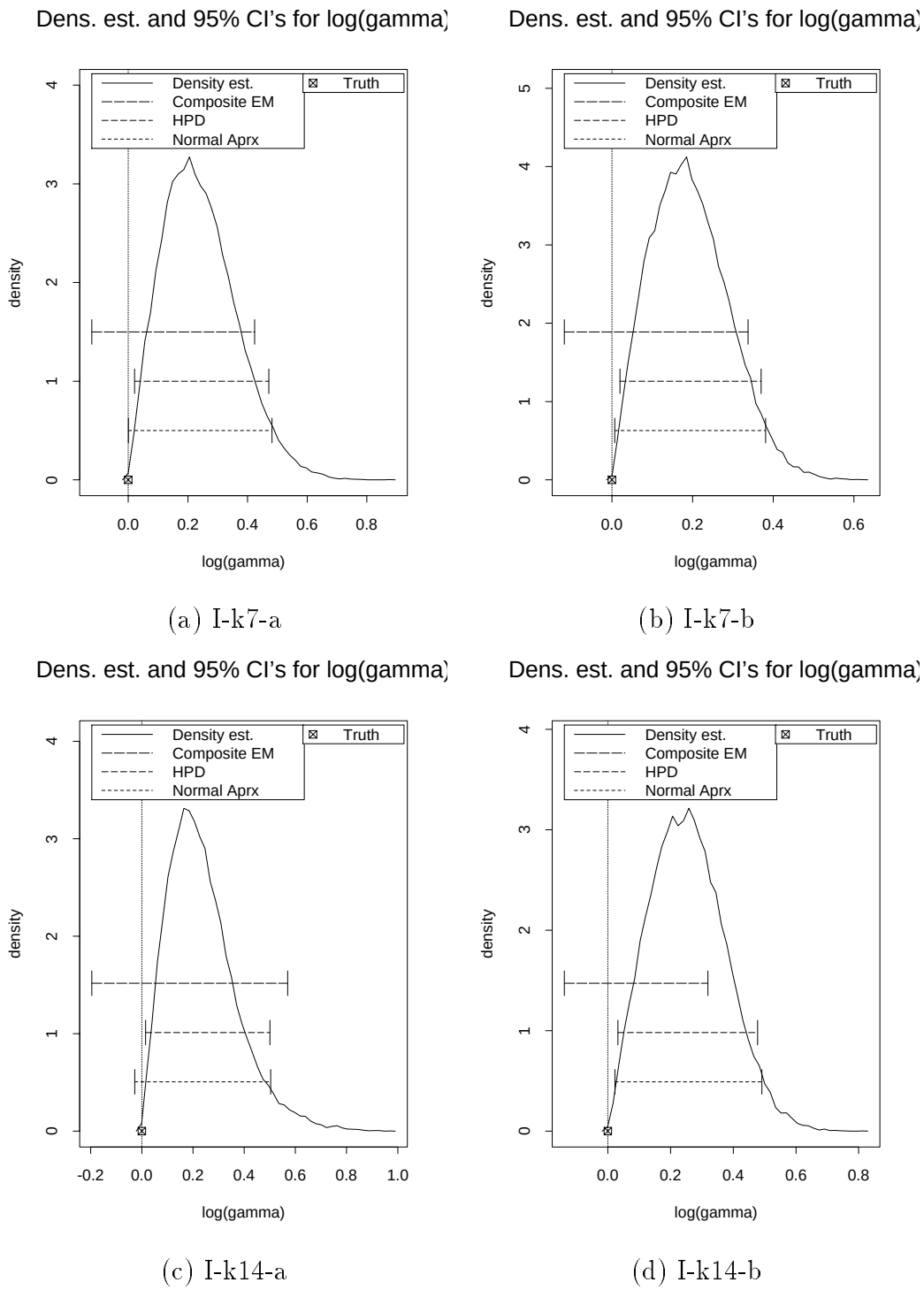


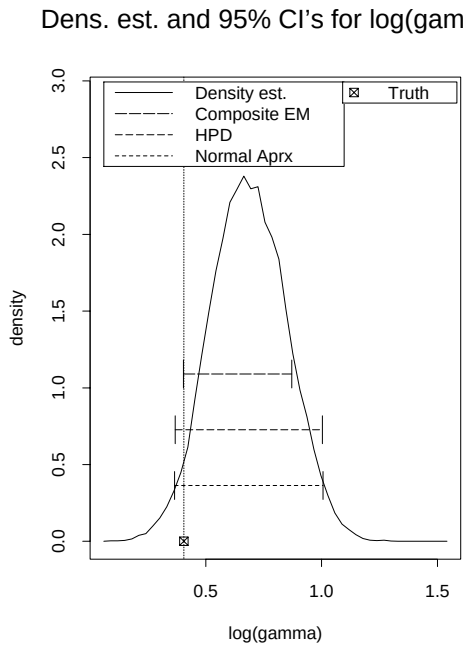
Figure J.7: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log \gamma$, using composite EM, HPDR and normal approximation, Redwood data.



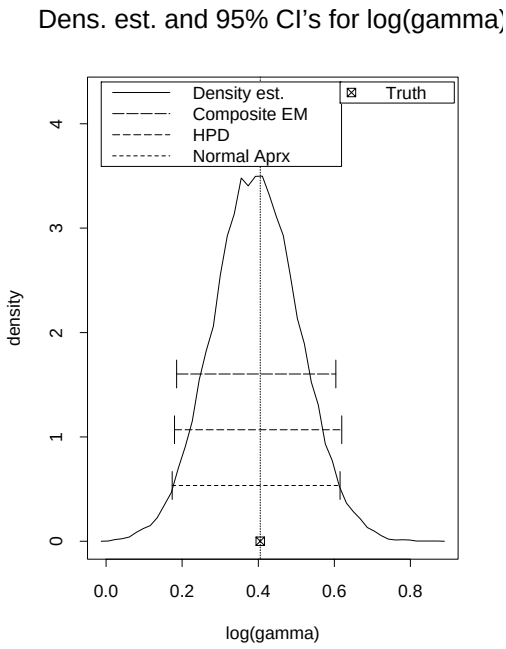
(a) I-k7-a (b) I-k7-b

(c) I-k14-a (d) I-k14-b

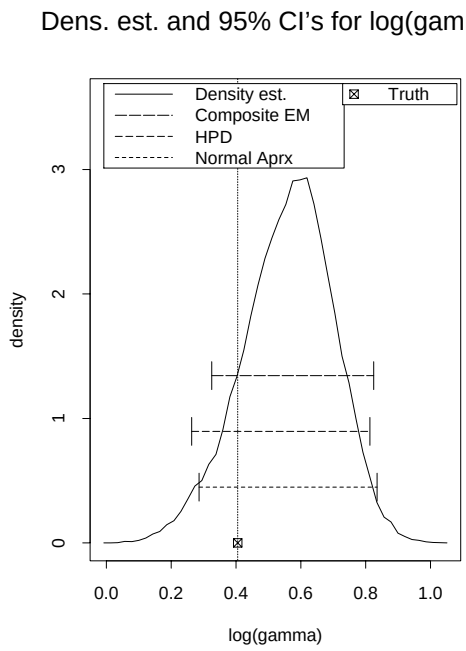
Figure J.8: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log \gamma$, using composite EM, HPDR and normal approximation, simulated patterns.



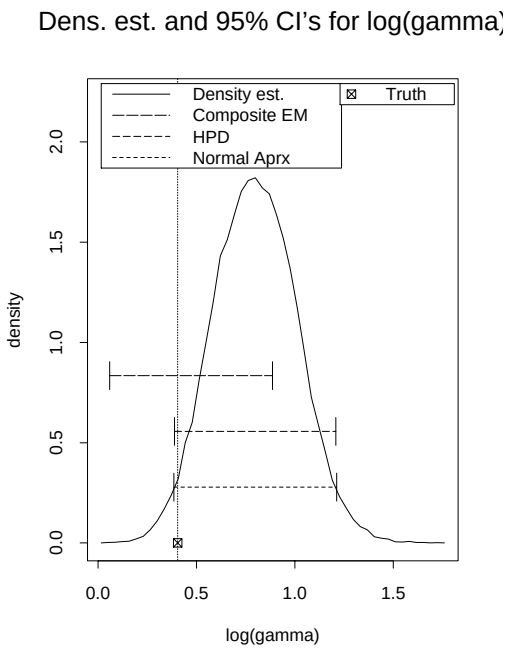
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

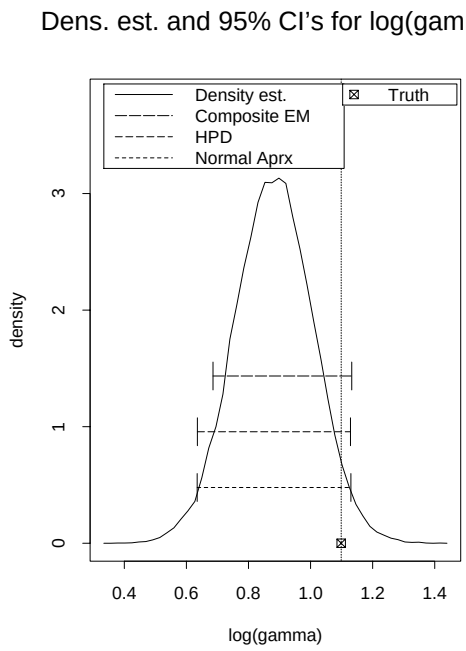


(g) AI-1.5-k14-a

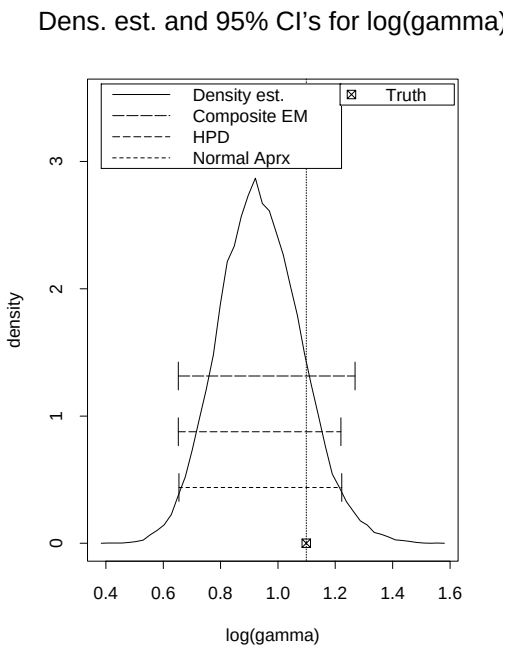


(h) AI-1.5-k14-b

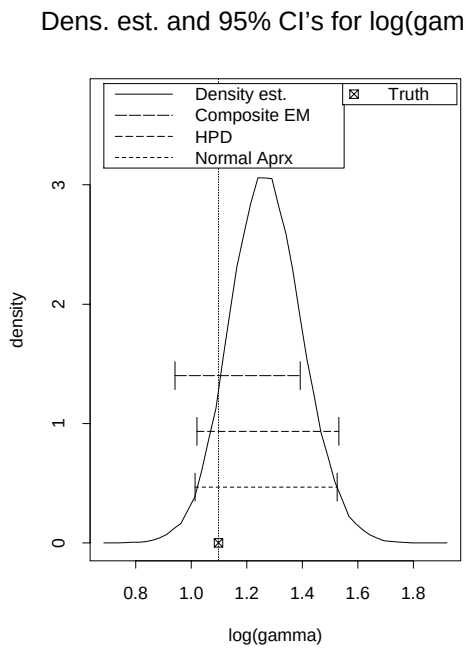
Figure J.8 (continued).



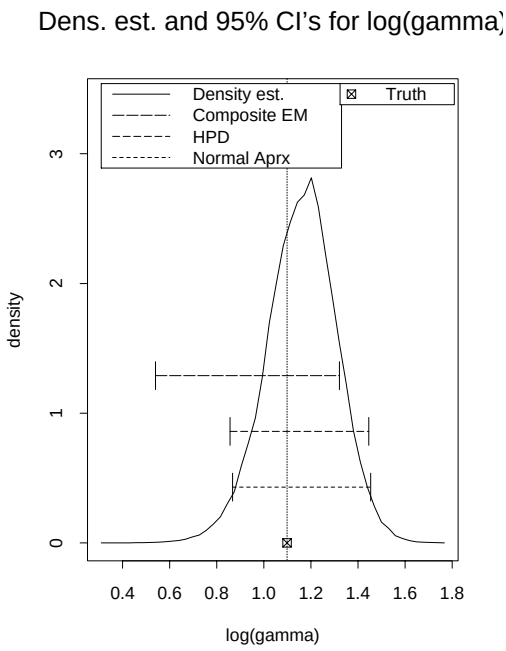
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure J.8 (continued).

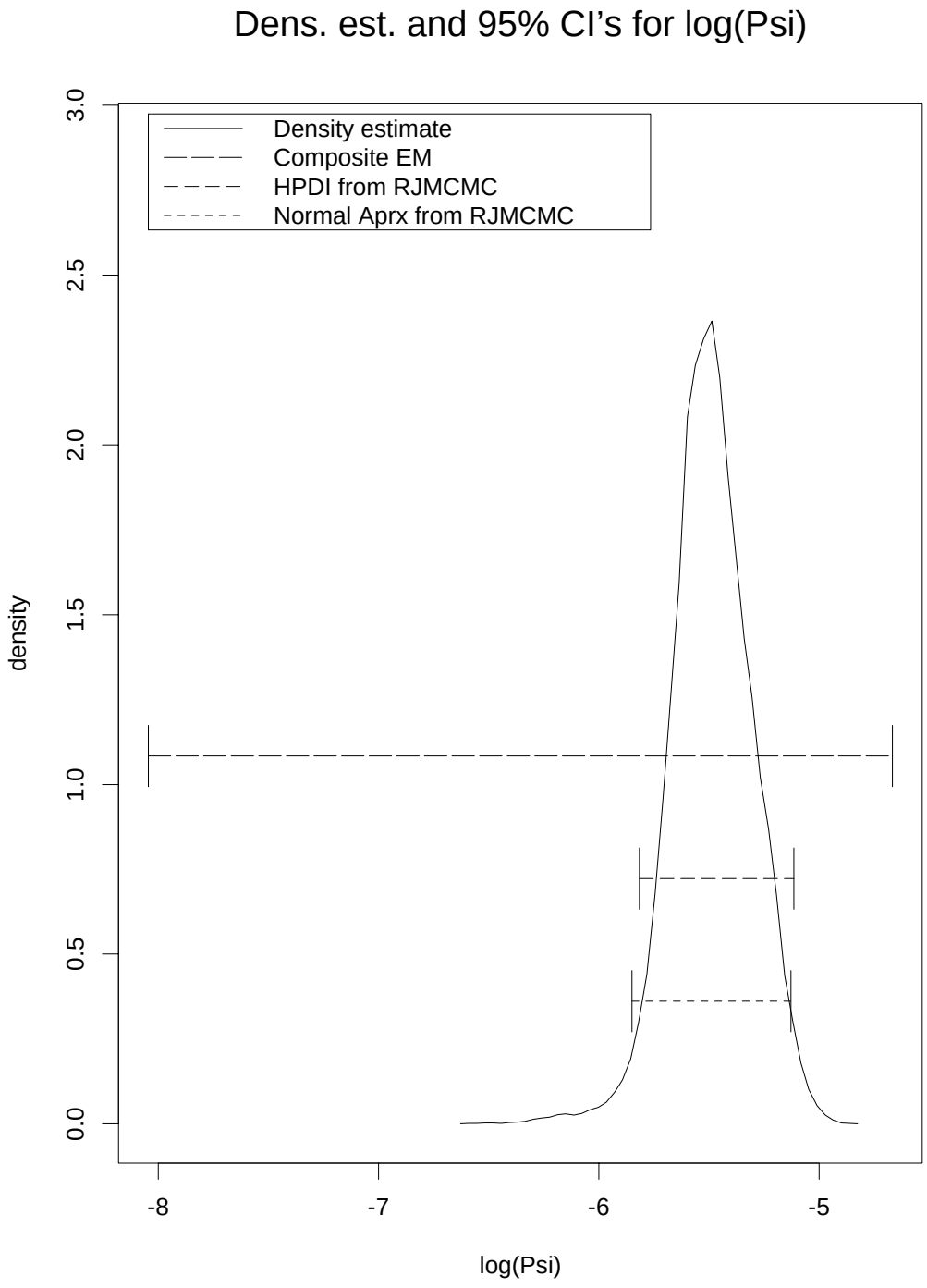
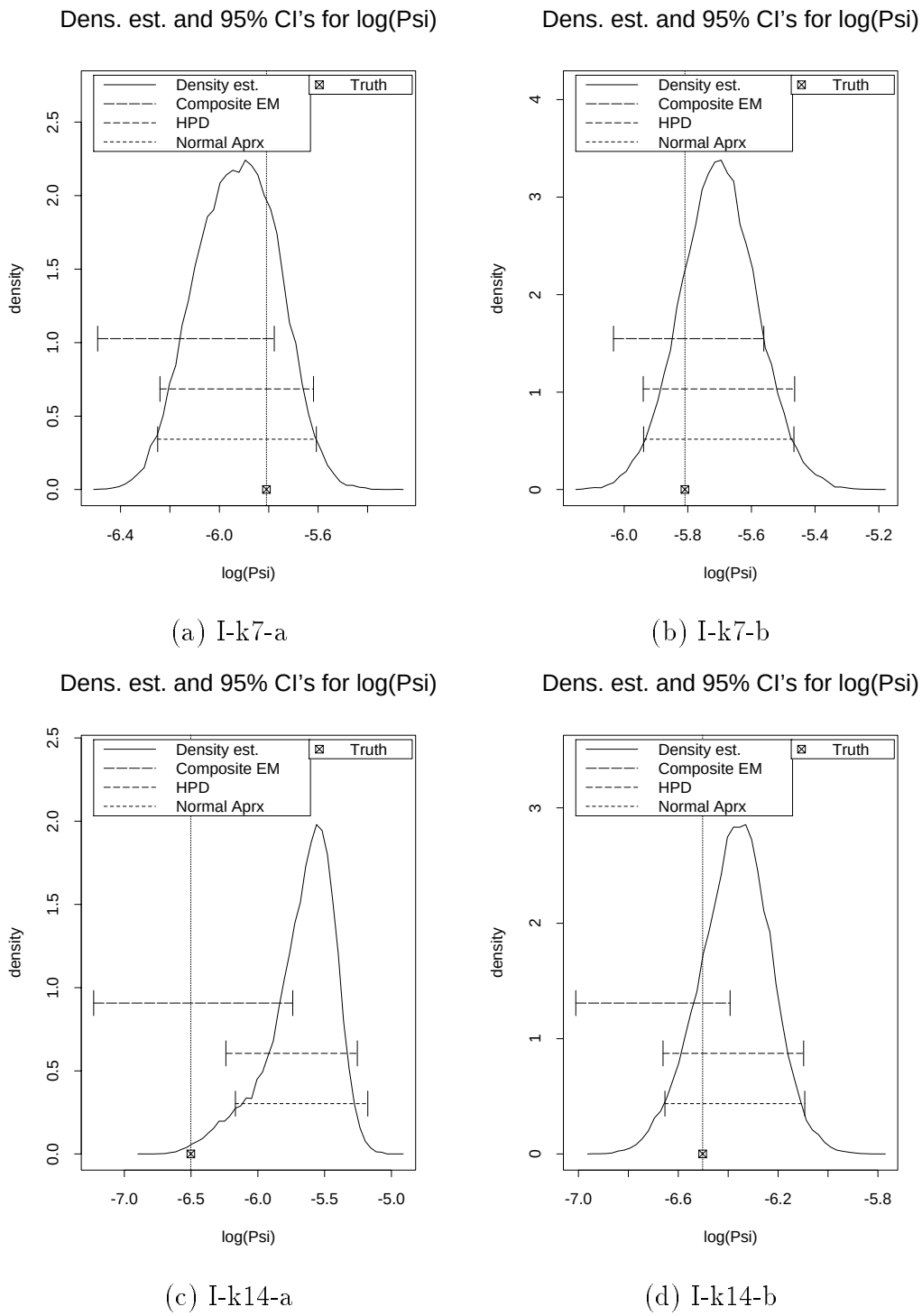


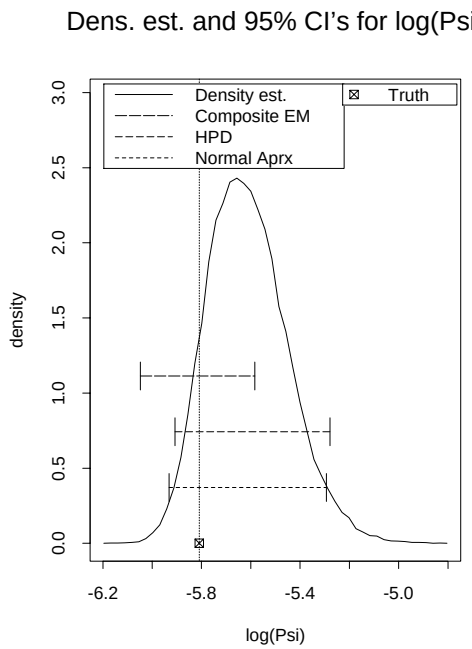
Figure J.9: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log \Psi$, using composite EM, HPDR and normal approximation, Redwood data.



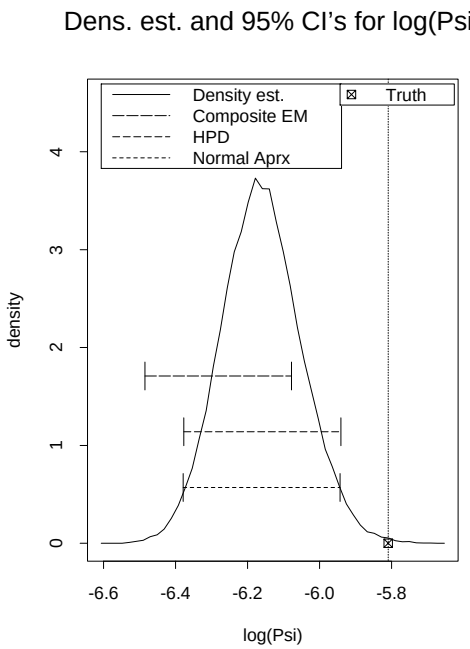
(a) I-k7-a (b) I-k7-b

(c) I-k14-a (d) I-k14-b

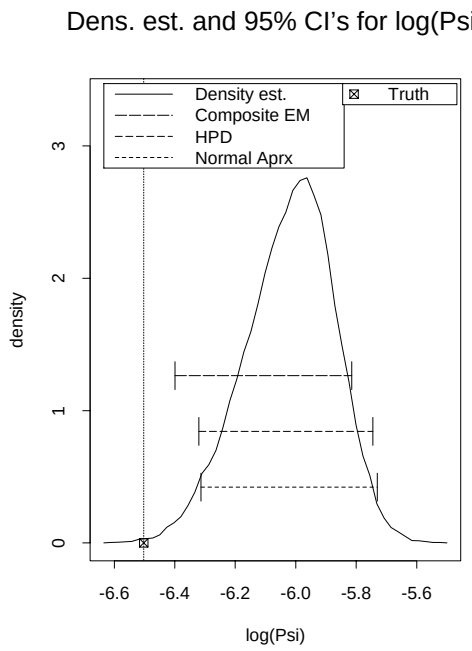
Figure J.10: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for $\log \Psi$, using composite EM, HPDR and normal approximation, simulated patterns.



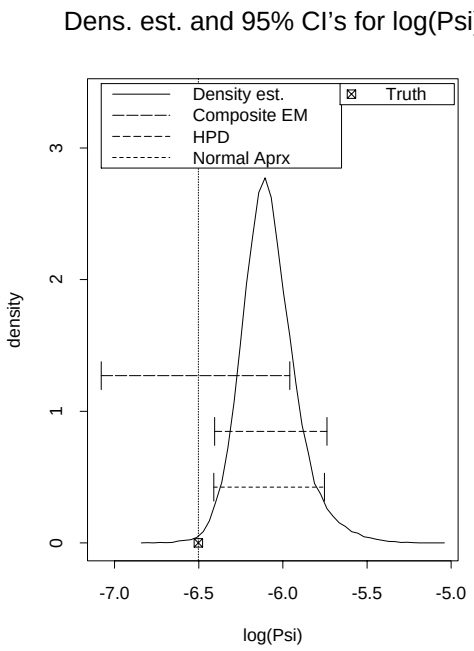
(e) AI-1.5-k7-a



(f) AI-1.5-k7-b

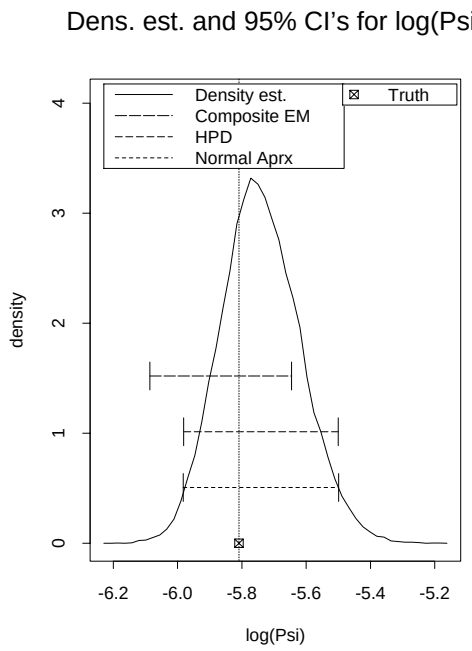


(g) AI-1.5-k14-a

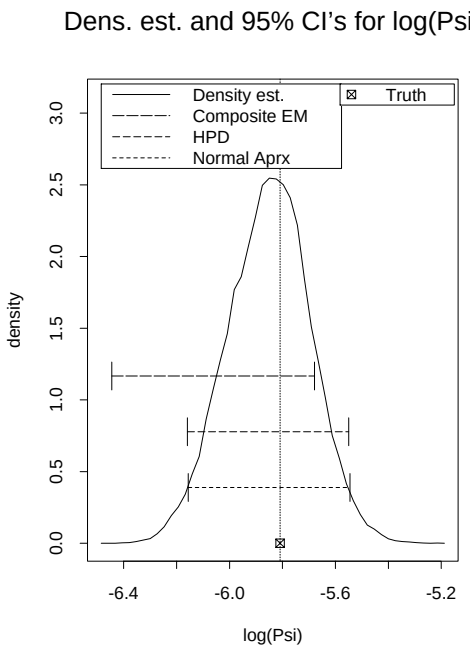


(h) AI-1.5-k14-b

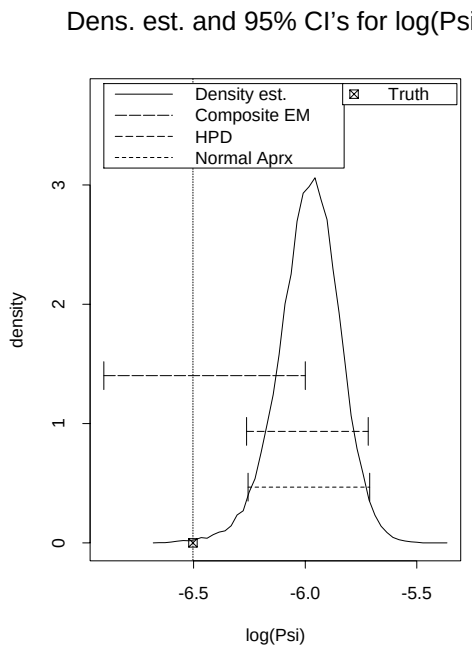
Figure J.10 (continued).



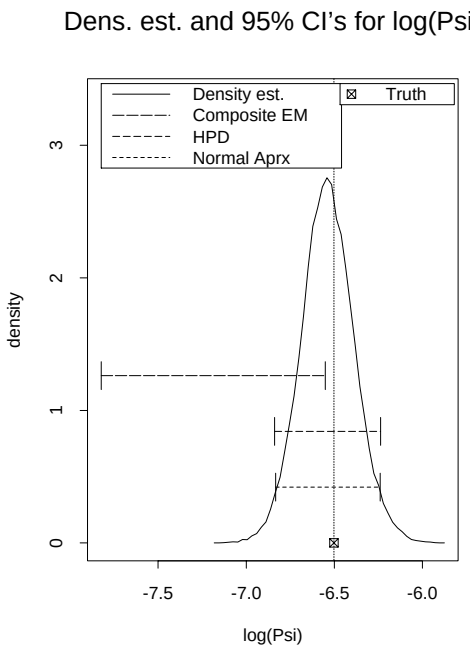
(i) AI-3-k7-a



(j) AI-3-k7-b



(k) AI-3-k14-a



(l) AI-3-k14-b

Figure J.10 (continued).

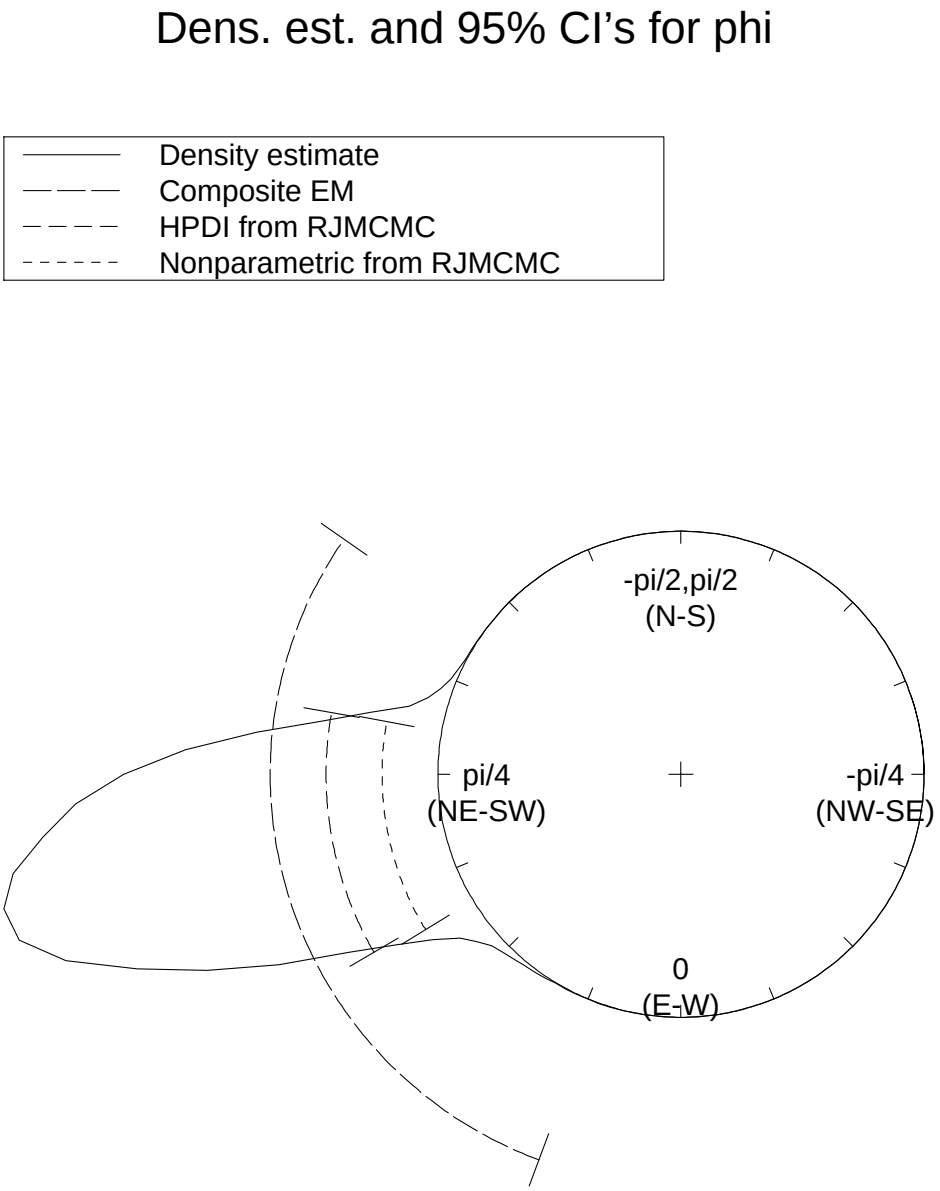
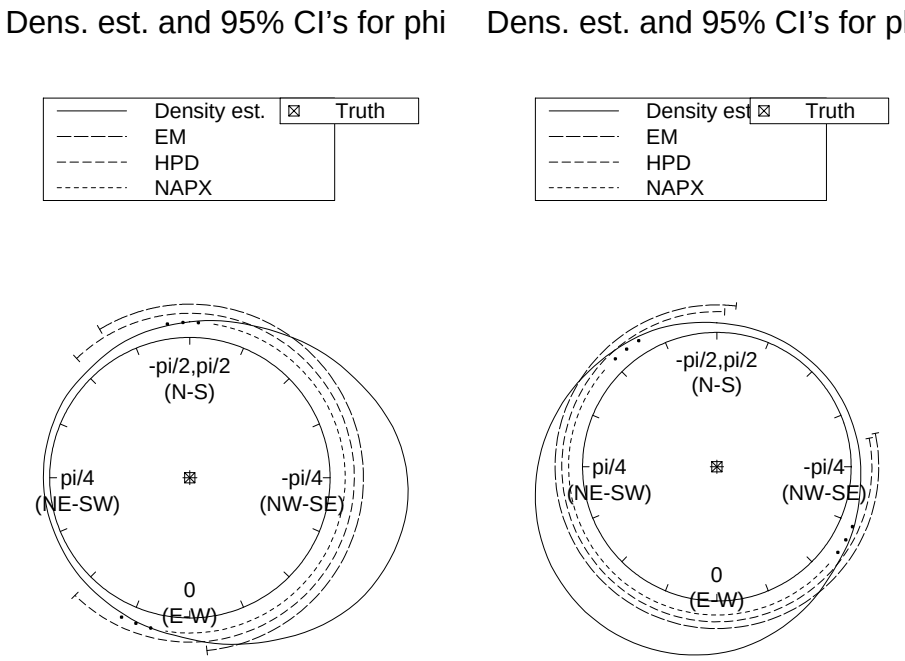
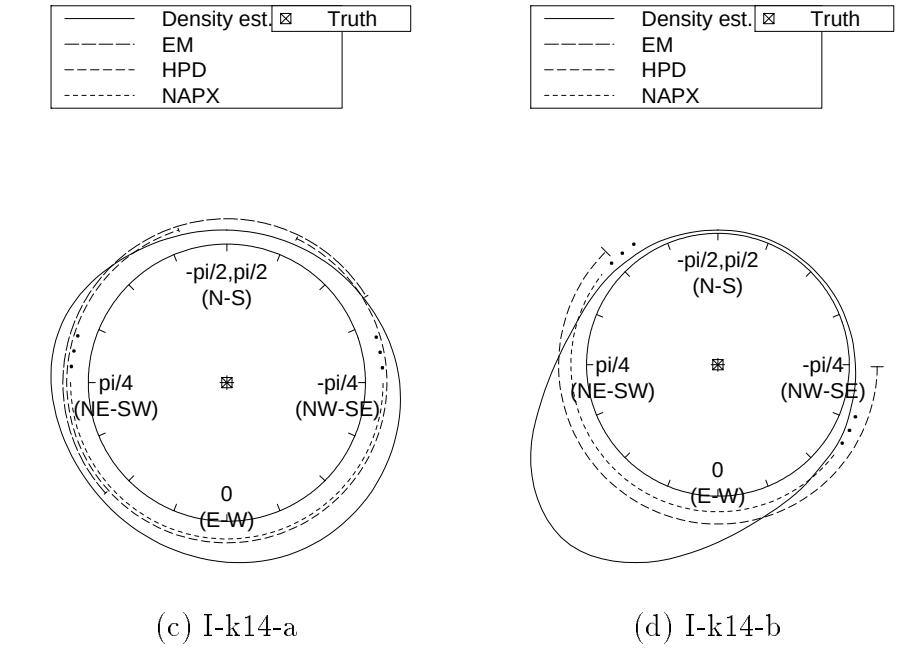


Figure J.11: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for ϕ, using composite EM, HPDR and normal approximation, Redwood data.

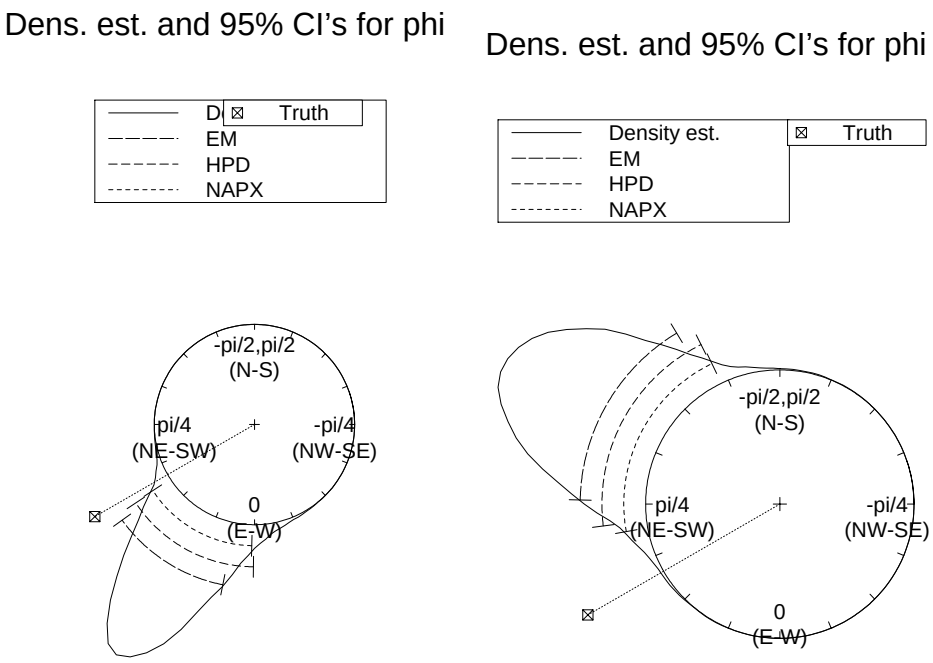


(a) I-k7-a (b) I-k7-b



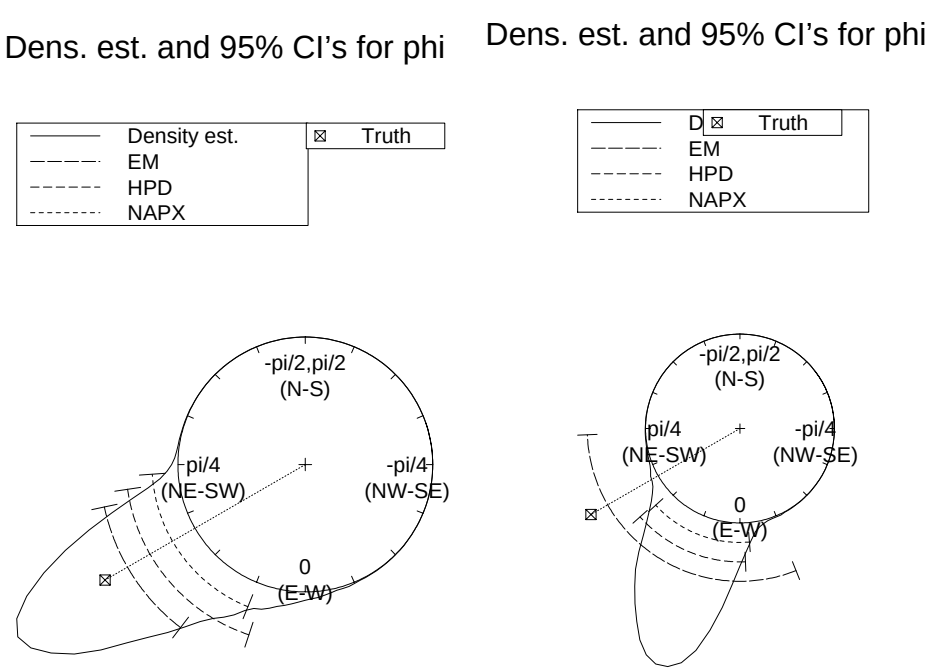
(c) I-k14-a (d) I-k14-b

Figure J.12: Non-parametric Gaussian posterior density estimate and 95% confidence intervals for ϕ, using composite EM, HPDR and normal approximation, simulated patterns.



(e) AI-1.5-k7-a

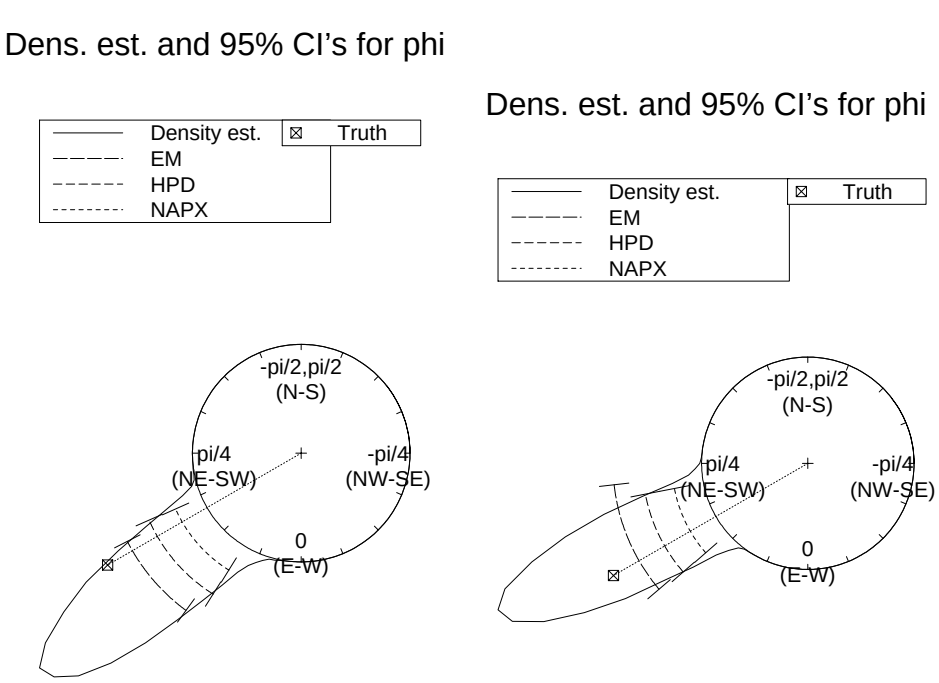
(f) AI-1.5-k7-b



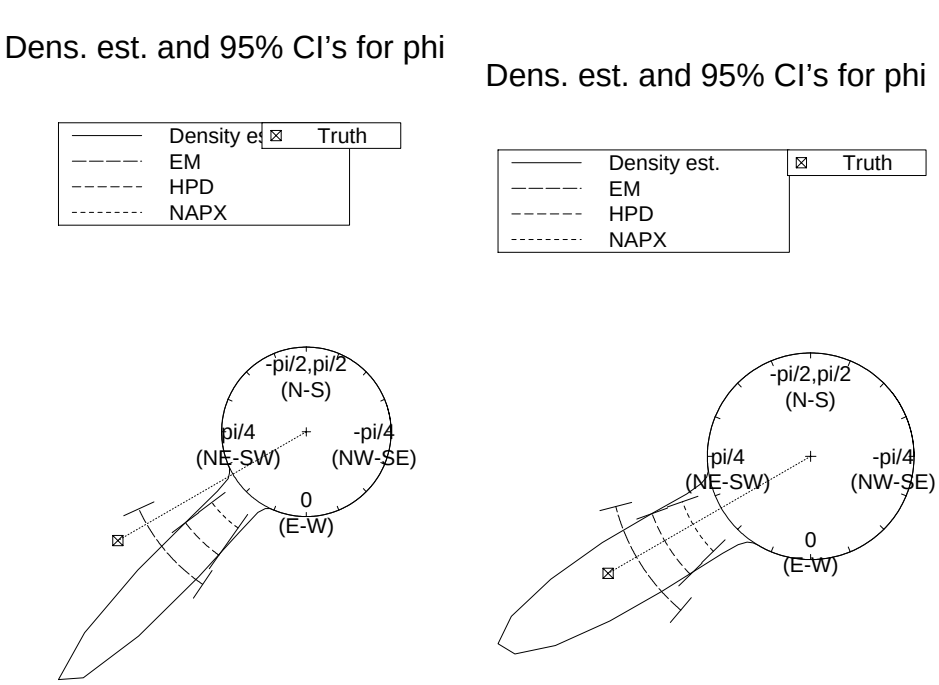
(g) AI-1.5-k14-a

(h) AI-1.5-k14-b

Figure J.12 (continued).



(i) AI-3-k7-a (j) AI-3-k7-b



(k) AI-3-k14-a (l) AI-3-k14-b

Figure J.12 (continued).

APPENDIX K
BIVARIATE NORMAL CONTOURS OF ESTIMATED OFFSPRING DISPERSAL DISTRIBUTION

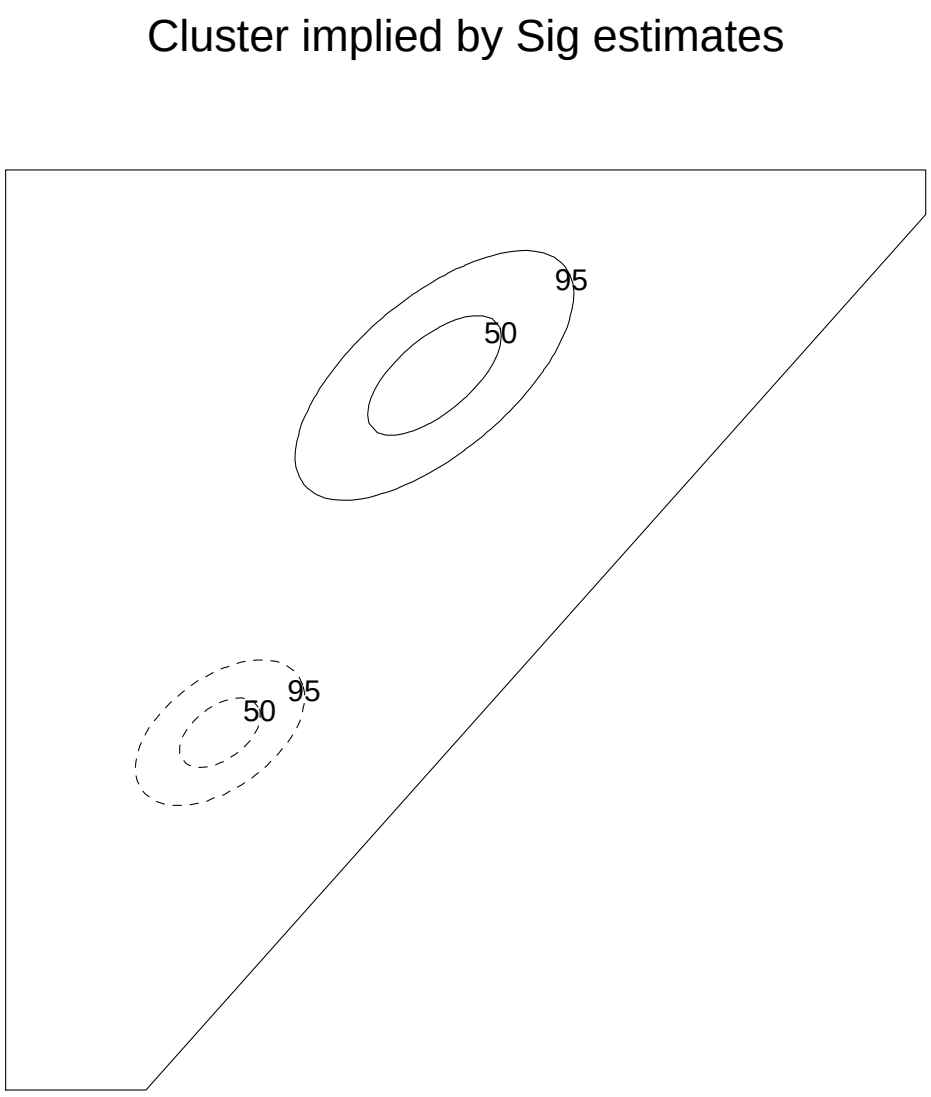
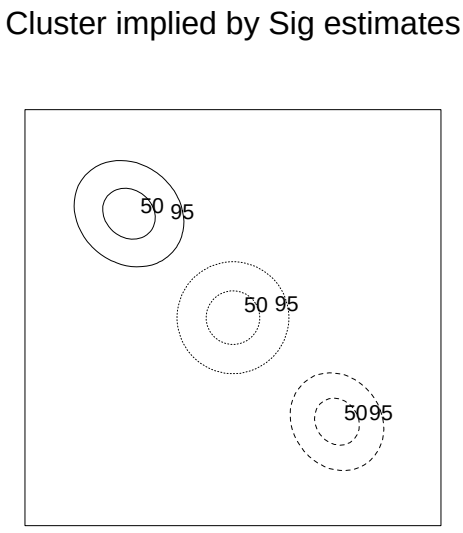
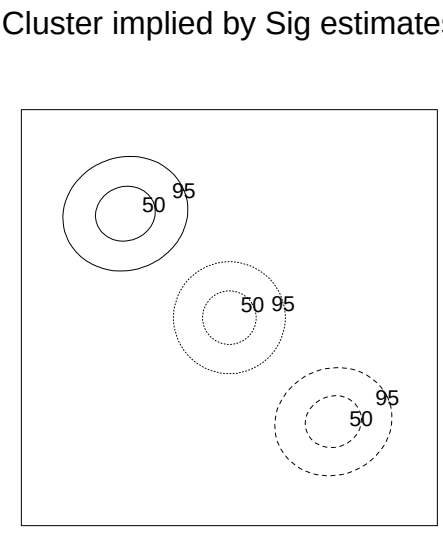


Figure K.1: Bivariate normal contours of estimated offspring dispersal distribution, using RJMCMC posterior means and composite EM, shown to scale, Redwood data.



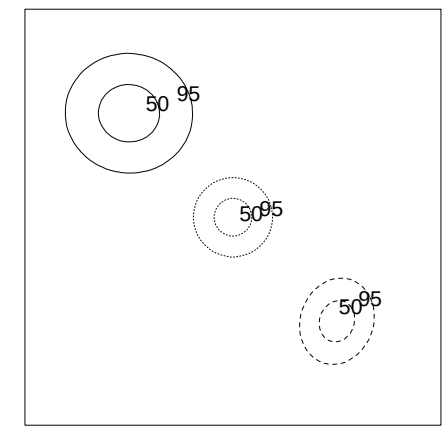
RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

(a) I-k7-a



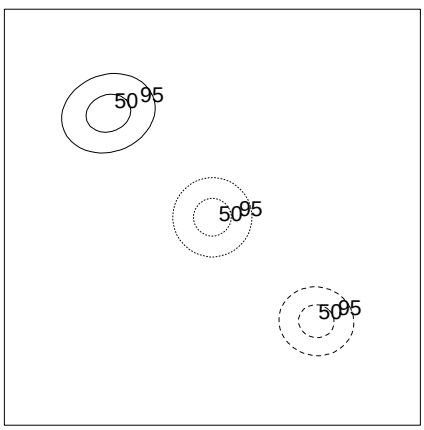
RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

(b) I-k7-b



RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

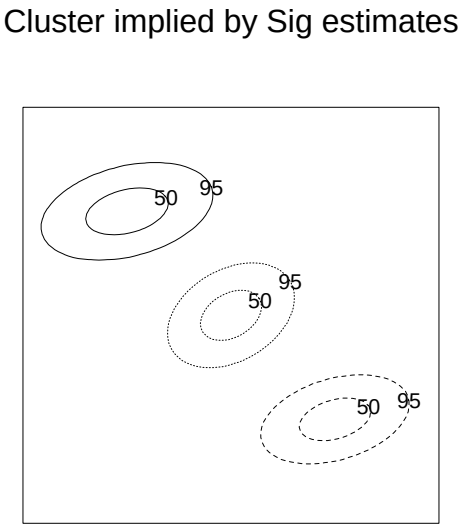
(c) I-k14-a



RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

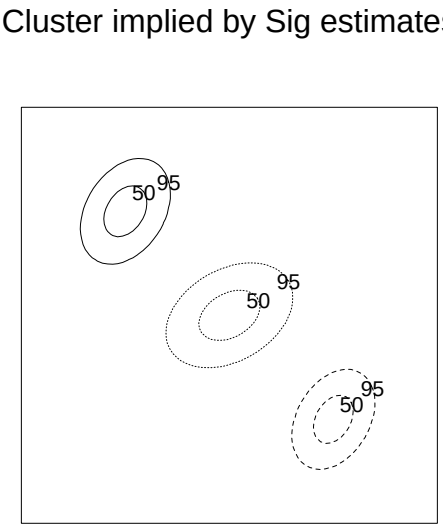
(d) I-k14-b

Figure K.2: Bivariate normal contours of estimated offspring dispersal distribution, using RJMCMC posterior means and composite EM, shown to scale, simulated patterns.



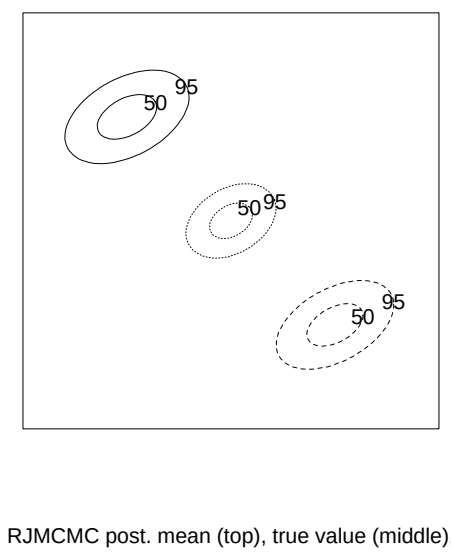
RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

(e) AI-1.5-k7-a



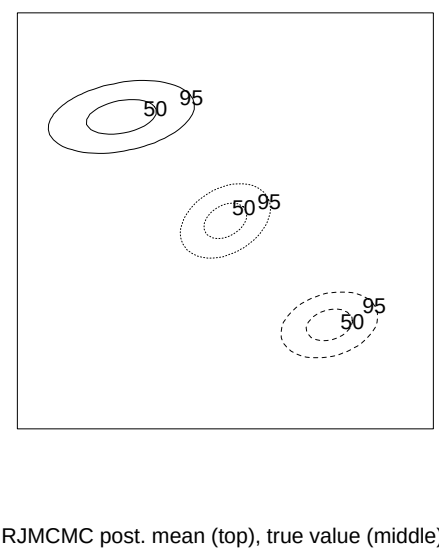
RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

(f) AI-1.5-k7-b



RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

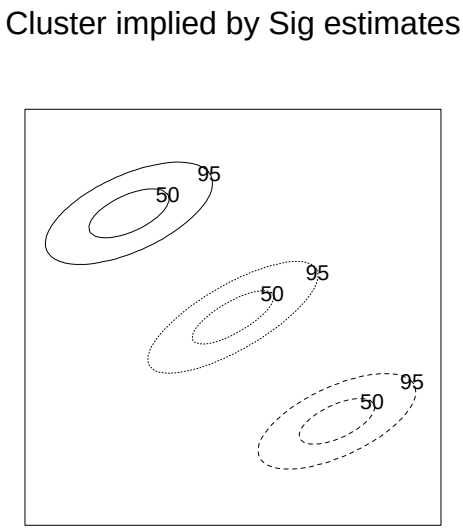
(g) AI-1.5-k14-a



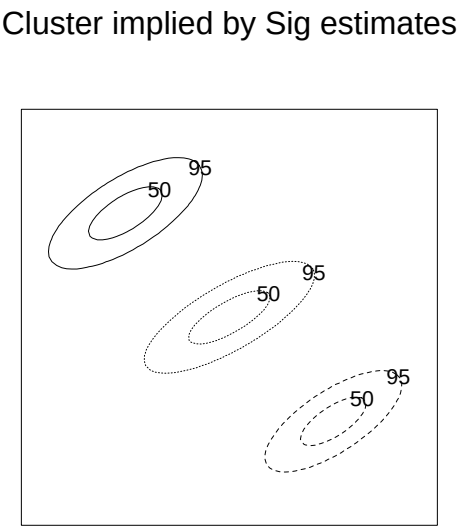
RJMCMC post. mean (top), true value (middle), and composite EM (bottom)

(h) AI-1.5-k14-b

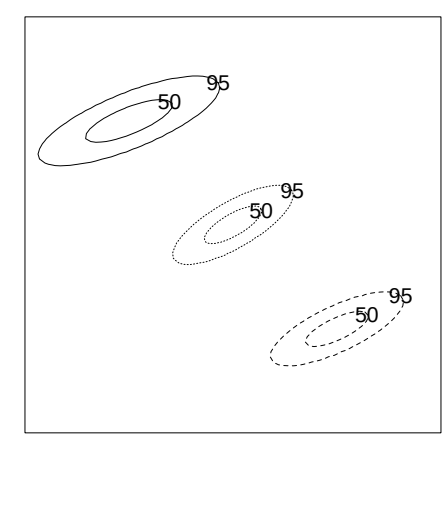
Figure K.2 (continued).



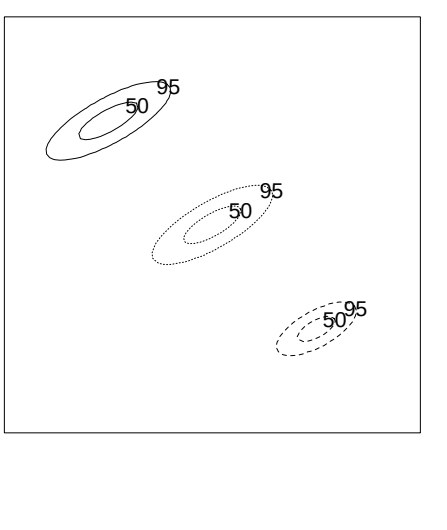
(i) AI-3-k7-a



(j) AI-3-k7-b

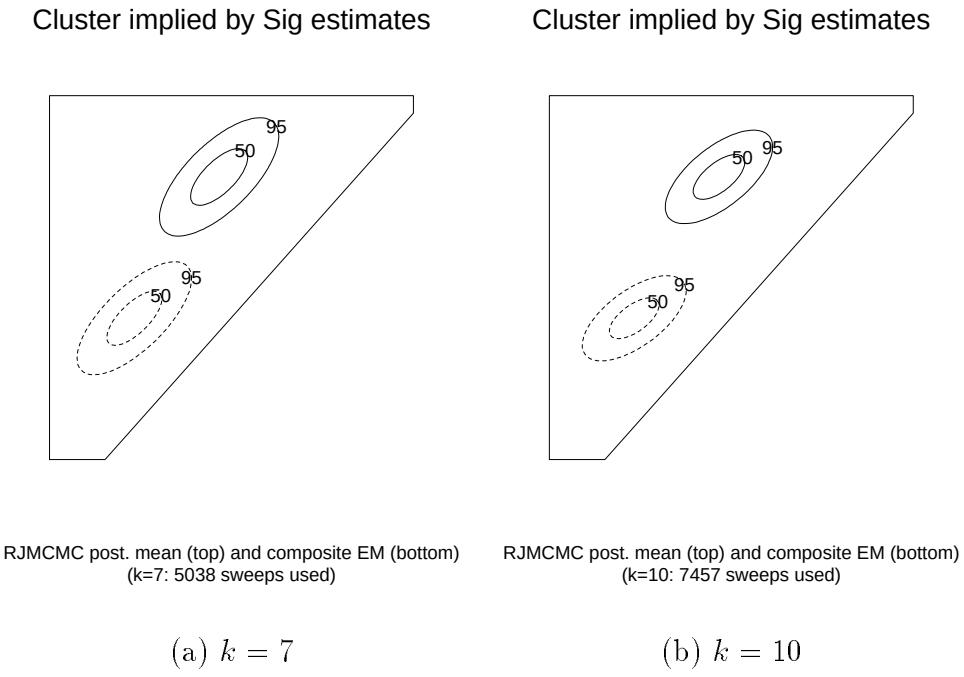


(k) AI-3-k14-a

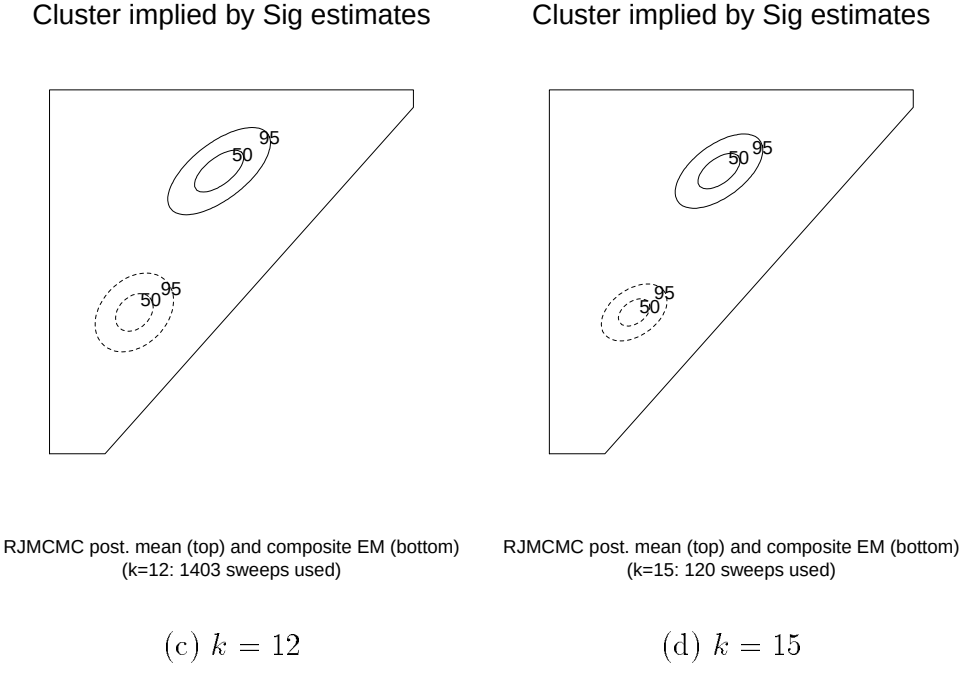


(l) AI-3-k14-b

Figure K.2 (continued).



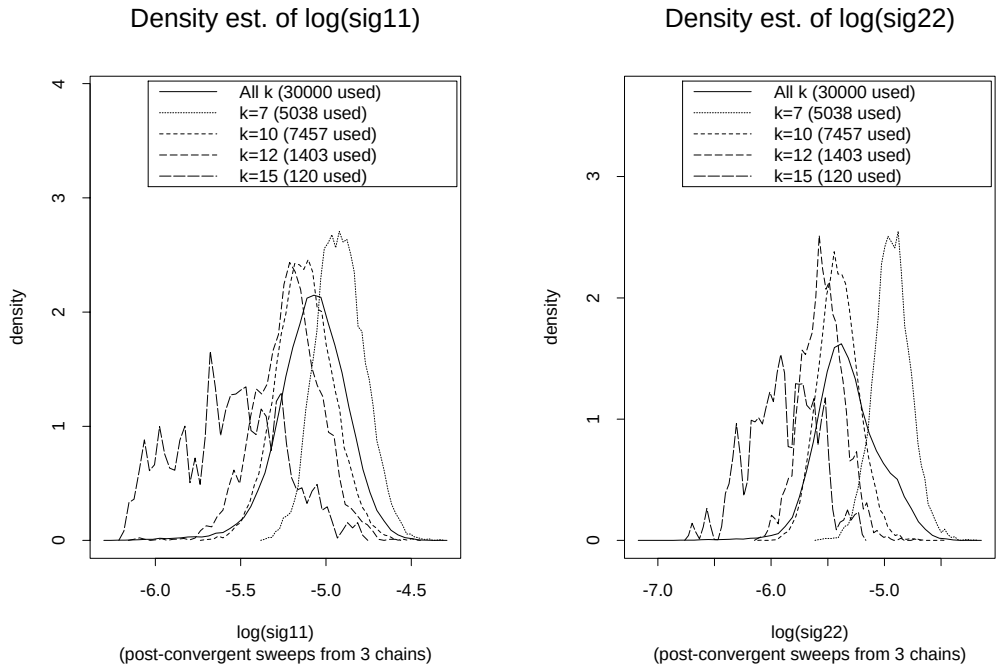
(a) $k = 7$ (b) $k = 10$



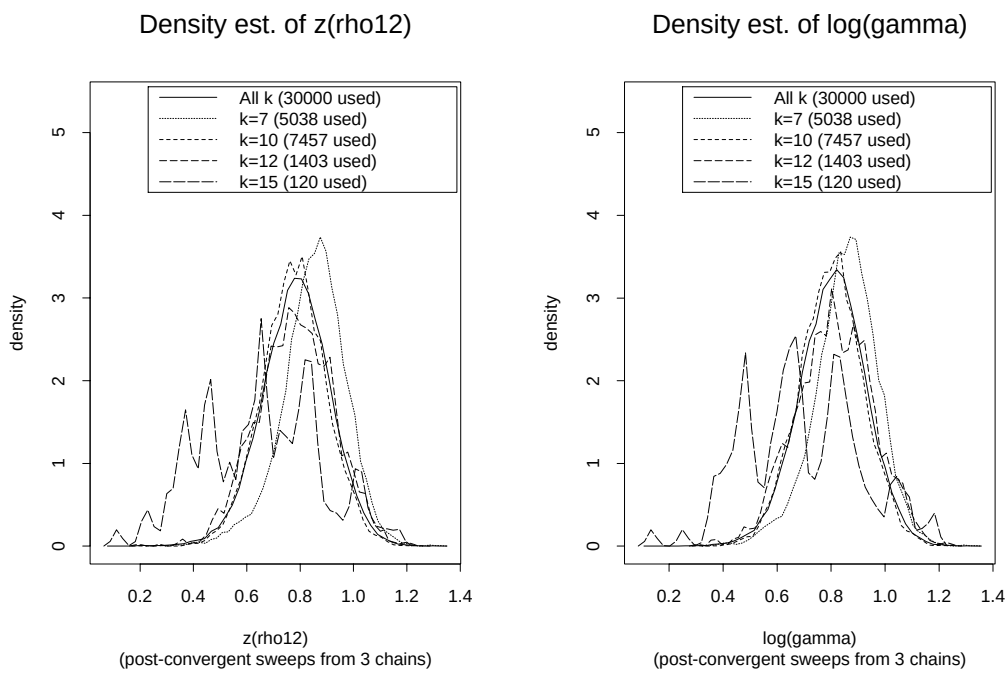
(c) $k = 12$ (d) $k = 15$

Figure K.3: Bivariate normal contours of estimated offspring dispersal distribution, by k, Redwood data.

APPENDIX L
POSTERIOR DENSITY ESTIMATES, BY K, REDWOOD DATA

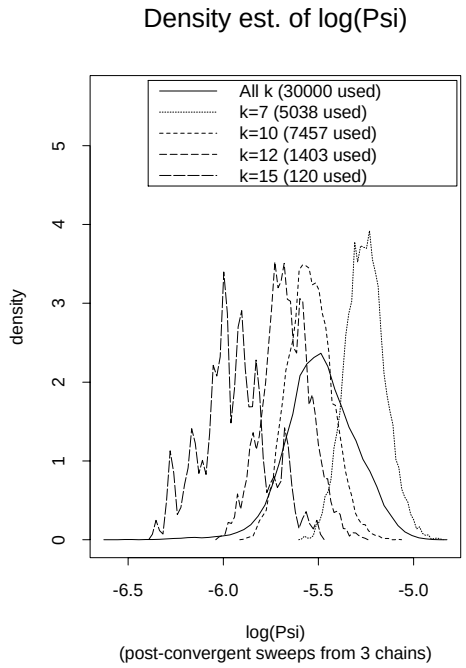


(a) $\log \sigma_{11}$ (b) $\log \sigma_{22}$

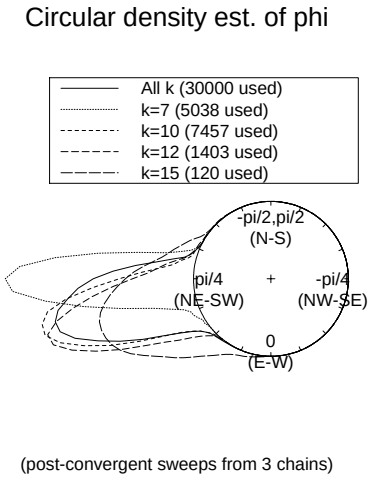


(c) $z(\rho_{12})$ (d) $\log \gamma$

Figure L.1: RJMCMC posterior density estimates, by k, Redwood data.



(e) $\log \Psi$



(f) ϕ

Figure L.1 (continued).

APPENDIX M
TABLES OF RJMCMC DETAILED RESULTS

Pattern	$\log \sigma_{11}$	$\log \sigma_{22}$	$z(\rho_{12})$	$\log \gamma$	$\log \Psi$	ϕ	$\boldsymbol{\sigma}^c$	$\widehat{p}(k\|\mathbf{Y})$
Redwoods	60 (500)	60 (500)	105 (285)	114 (263)	60 (500)	54 (555)	111 (270)	48 (625)
I-k7-a	249 (120)	498 (60)	498 (60)	498 (60)	249 (120)	102 (294)	375 (80)	114 (263)
I-k7-b	750 (40)	750 (40)	3000 (10)	1500 (20)	750 (40)	1500 (20)	999 (30)	600 (50)
I-k14-a	90 (333)	54 (555)	129 (232)	165 (181)	81 (370)	135 (222)	150 (200)	36 (833)
I-k14-b	300 (100)	300 (100)	249 (120)	300 (100)	300 (100)	498 (60)	249 (120)	198 (151)
AI-1.5-k7-a	174 (172)	270 (111)	111 (270)	174 (172)	198 (151)	1500 (20)	123 (243)	114 (263)
AI-1.5-k7-b	3000 (10)	3000 (10)	1500 (20)	3000 (10)	999 (30)	3000 (10)	1500 (20)	426 (70)
AI-1.5-k14-a	165 (181)	81 (370)	81 (370)	165 (181)	198 (151)	81 (370)	186 (161)	81 (370)
AI-1.5-k14-b	54 (555)	249 (120)	129 (232)	90 (333)	174 (172)	333 (90)	129 (232)	150 (200)
AI-3-k7-a	999 (30)	3000 (10)	3000 (10)	3000 (10)	3000 (10)	1500 (20)	3000 (10)	249 (120)
AI-3-k7-b	186 (161)	999 (30)	426 (70)	426 (70)	186 (161)	600 (50)	375 (80)	426 (70)
AI-3-k14-a	300 (100)	198 (151)	498 (60)	300 (100)	249 (120)	249 (120)	300 (100)	198 (151)
AI-3-k14-b	498 (60)	81 (370)	333 (90)	498 (60)	87 (344)	999 (30)	426 (70)	69 (434)

Table M.1: Batch sampling details. Entries are: #batches, (batch size) used. Entries for $\widehat{p}(k|\mathbf{Y})$ correspond to minimum #batches over k.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov and Csaki (1973), pp. 267–281.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.). John Wiley and Sons, New York.

Baddeley, A. and Møller, J. (1989). Nearest-neighbor Markov point processes and random sets. *International Statistical Review* **57**, 89–121.

Baddeley, A. J. and van Lieshout, M. N. M. (1993). Stochastic geometry models in high-level vision. *Journal of Applied Statistics* **20**, 231–256.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Batschelet, E. (1981). *Circular Statistics in Biology.* Academic Press, New York.

Baudin, M. (1981). Likelihood and nearest-neighbor distance properties of multidimensional Poisson cluster processes. *Journal of Applied Probability* **18**, 879–888.

Beneš, V., Fendrych, F., and Suchánek, V. (1989). On some quantitative methods for evaluation of anisotropic structures. *Acta Stereologica* **8**, 695–700.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1995). Inference in model-based cluster analysis. Technical Report 285, Department of Statistics, University of Washington.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer-Verlag, New York.

Berger, J. O. et al., eds. (1992). *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, April 15-20, 1991.* Oxford University Press, New York.

Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., eds. (1996). *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting, June 5-9, 1994.* Oxford University Press, New York.

Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., eds. (1998). *Bayesian Statistics 6*. Oxford University Press, New York (in press).

Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., eds. (1988). *Bayesian Statistics 3: Proceedings of the Third Valencia International Meeting, June 1-5, 1987*. Oxford University Press, Cambridge, MA.

Bollen, K. A. and Long, S. J., eds. (1993). *Testing Structural Equation Models*. Sage Publications, Newbury Park, CA.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician* **47**, 69–100.

Brooks, S. P. and Gelman, A. (1996). General methods for monitoring convergence of iterative simulations. Department of Mathematics, University of Bristol and Department of Statistics, Columbia University (in press).

Brooks, S. P. and Giudici, P. (1998). Convergence assessment for reversible jump MCMC simulations. In Bernardo, Berger, Dawid, and Smith (1998), pp. (to appear).

Campbell, J. G., Fraley, C., Murtagh, F., and Raftery, A. E. (1998). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, (to appear).

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, Methodological* **57**, 473–484.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York.

Carvajal-Gonzalez, S., König, D., Downs, A. M., Nguyen, Q., Vassy, J., and Rigaut, J. P. (1989). Analysis of histological architecture by point process modelling and spatial statistics applied to three-dimensional images from laser scanning confocal microscopy. *Acta Stereologica* **8**, 407–412.

Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation* **55**, 287–314.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* **28**, 781–793.

Chen, M. and Shao, Q. (1998). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* **7**, (to appear).

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chung, K. L. (1974). *A Course in Probability Theory* (2nd ed.). Academic Press, San Diego.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* **91**, 883–904.

Crawford, T. J. (1984). What is a population? In Sharrocks (1984), pp. 135–173.

Cruz-Orive, L. M., Hoppeler, O. M., and Weibel, E. R. (1985). Stereological analysis of anisotropic structures using directional statistics. *Applied Statistics* **34**, 14–32.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **93**, 294–302.

Davies, R. (1997). NEWMAT/NEWRAN Matrix and Random Number Generator C++ Libraries. New Zealand.

Davies, S. (1996). Software for Circular Data Analysis. CSIRO, Australia.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–22.

Diebolt, J. and Ip, E. H. S. (1995). Stochastic EM: method and application. In Gilks, Richardson, and Spiegelhalter (1996), Chapter 15, pp. 259–274.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B, Methodological* **56**, 363–375.

Diggle, P. J. (1975). Robust density estimation using distance methods. *Biometrika* **62**, 39–48.

Diggle, P. J. (1978). On parameter estimation for spatial point processes. *Journal of the Royal Statistical Society, Series B, Methodological* **40**, 178–181.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press, New York.

Ecker, M. D. and Gelfand, A. E. (1997). Modeling and inference for geometrically anisotropic spatial data (in press).

Fisher, N. I. (1993). *Statistical Analysis of Circular Data.* Cambridge University Press, Cambridge.

Fisher, N. I. and Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika* **70**, 327–332.

Fraley, C. (1998). MCLUST/EMCLUST Software. Department of Statistics, University of Washington.

Fraley, C. (1999). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing* **20**, 270–281. (to appear).

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. Technical Report 329, Department of Statistics, University of Washington.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association* **62**, 1159–1178.

Gelfand, A. E. (1995). Model determination using sampling-based methods. In Gilks, Richardson, and Spiegelhalter (1996), Chapter 9, pp. 145–162.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In Berger et al. (1992), pp. 147–159.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, J. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall, New York.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulations using multiple sequences. *Statistcal Science* (7), 457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, 721–741.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Berger et al. (1992), pp. 169–188.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandanavian Journal of Statistics* **21**, 359–373.

Gilks, W. R. (1997). Discussion contribution. In Richardson and Green (1997), pp. 770–771.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York.

Gordon, A. D. (1981). *Classification*. Chapman and Hall, New York.

Granville, V. and Smith, R. L. (1995). Clustering and Neyman-Scott process parameter simulation via Gibbs sampling. Statistical Laboratory, University of Cambridge (in press).

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Grenander, U. and Miller, M. I. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B, Methodological* **56**, 549–603.

Grimmet, G. R. and Stirzaker, D. R. (1992). *Probability and Random Processes* (2nd ed.). Oxford University Press Inc., New York.

Gruet, M., Philippe, A., and Robert, C. P. (1998). MCMC control spreadsheets for exponential mixture estimation. INRA, Jouy-en-Josas (in press).

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis* (3rd ed.). Prentice Hall, Englewood Cliff, NJ.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kelly, F. P. and Ripley, B. D. (1976). A note on Strauss's model for clustering. *Biometrika* **63**, 357–360.

Khuri, A. I. (1993). *Advanced Calculus with Applications in Statistics*. Wiley, New York.

König, D. and Ohser, J. (1988). On the estimation of second-order and further characteristics of random planar points structures. *Acta Stereologica* **7**, 1–16.

König, D. and Schmidt, V. (1992). Directional distributions for multi-dimensional random point processes. *Communications in Statistics – Stochastic Models* **8**, 617–636.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks* **20**, 491–505.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **50**, 157–224.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *The Canadian Journal of Statistics* **20**, 451–461.

Lawson, A. (1993). Discussion contribution on The Gibbs sampler and other Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 61–62.

Lawson, A. B. (1995a). Markov chain Monte Carlo methods for spatial cluster processes. In Meyer and Rosenberger (1996), pp. 314–319.

Lawson, A. B. (1995b). MCMC methods for putative pollution source problems in environmental epidemiology. *Statistics in Medicine* **14**, 2473–2485.

Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley and Sons, New York.

Leroux, M. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics* **20**, 1350–1360.

Lewis, P. A. W., ed. (1972). *Stochastic Point Processes: Statistical Analysis, Theory, and Applications.* Wiley-Interscience, New York.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* **44**, 226–233.

Mardia, K. V. (1972). *Statistics of Directional Data.* Academic Press, New York.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, Inc., London.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

Meng, X. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1998). MCMC convergence diagnostics: a "reviewww". CREST-INSEE, Paris (in press).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.

Meyer, M. M. and Rosenberger, J. L., eds. (1996). *Proceedings of the 27th Symposium on the Interface, Pittsburgh, PA, June 21-24, 1995*, Volume 27 of *Statistics and manufacturing with subthemes in environmental statistics, graphics and imaging : computing science and statistics*. Interface Foundation of North America, Fairfax, VA.

Mugglestone, M. A. and Renshaw, E. (1996a). The exploratory analysis of biviariate spatial point patterns using cross-spectra. *EnvironMetrics* **7**, 361–377.

Mugglestone, M. A. and Renshaw, E. (1996b). A practical guide to the spectral analysis of spatial point processes. *Computational Statistics and Data Analysis* **21**, 43–65.

Mukerjee, S., Feigelson, E. D., Babu, G. J., Murtagh, F., Fraley, C., and Raftery, A. (1998). Three types of Gamma ray bursts. Technical report, Department of Astronomy and Astrophysics, Pennsylvania State University.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7**, 308–313.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B, Methodological* **56**, 3–48.

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society, Series B, Methodological* **20**, 1–43.

Neyman, J. and Scott, E. L. (1972). Processes of clustering and applications. In Lewis (1972), pp. 646–681.

Nobile, A. (1997). Discussion contribution. In Richardson and Green (1997), pp. 771.

Ohser, J. and Stoyan, D. (1981). On the second-order and orientation analysis of planar stationary point processes. *Biometrical Journal* **23**, 523–533.

Olsson, D. M. and Nelson, L. S. (1975). The Nelder-Mead simplex procedure for function minimization. *Technometrics* **7**, 45–51.

Pasquill, F. (1974). *Atmospheric Diffusion: The Dispersion of Windborne Material from Industrial and other Sources* (2nd ed.). Ellis Horwood Limited, Chichester.

Patil, G. P., Pielou, E. C., and Waters, W. E., eds. (1971). *Statistical Ecology*, Volume 1. Pennsylvania State University Press, University Park.

Petrov, B. N. and Csaki, F., eds. (1973). *2nd International Symposium on Information Theory*. Akademiai Kiado, Budapest.

Phillips, D. B. and Smith, A. F. M. (1995). Bayesian model comparison via jump diffusions. In Gilks, Richardson, and Spiegelhalter (1996), Chapter 13, pp. 215–240.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In Bollen and Long (1993), Chapter 7, pp. 163–180.

Raftery, A. E. (1995). Hypothesis testing and model selection. In Gilks, Richardson, and Spiegelhalter (1996), Chapter 10, pp. 163–188.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B, Methodological* **59**, 731–792.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability* **13**, 255–266.

Ripley, B. D. (1977). Modelling spatial patterns,. *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 172–212.

Ripley, B. D. (1981). *Spatial Statistics*. Wiley, New York.

Ripley, B. D. (1987). *Stochastic Simulation*. John Wiley and Sons, New York.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

Robert, C. P. (1997). Discussion contribution. In Richardson and Green (1997), pp. 758–764.

Robert, C. P. and Mengersen, K. L. (1997). Reparameterisation issues in mixture modelling and their bearing on MCMC algorithms. CREST, INSEE, Paris (in press).

Robert, C. P., Ryden, T., and Titterington, D. M. (1998). Convergence controls for MCMC algorithms, with applications to hidden Markov chains. CREST-INSEE, Paris (in press).

Robert, C. P. and Titterington, D. M. (1998). Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, (to appear).

Roberts, G. O. (1995). Markov chain concepts related to sampling algorithms. In Gilks, Richardson, and Spiegelhalter (1996), Chapter 3, pp. 145–162.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**, 894–902.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In Bernardo, DeGroot, Lindley, and Smith (1988), pp. 395–402.

Schwarz, C. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sharrocks, B. A., ed. (1984). *Evolutionary Ecology: the 23rd Symposium of the British Ecological Society*. Blackwell Scientific, Oxford.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., and Inskip, H. (1995). Hepatitis B: a case study in MCMC methods. In Gilks, Richardson, and Spiegelhalter (1996), Chapter 2, pp. 21–44.

Stanford, J. L. and Vardeman, S. B., eds. (1994). *Statistical Methods for Physical Sciences*, Volume 28 of *Methods of Experimental Physics*. Academic Press, San Diego.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. Ph. D. thesis, University of Oxford.

Stoyan, D. (1991). Describing the anisotropy of marked planar point processes. *Statistics* **22**, 449–462.

Stoyan, D. (1992). Statistical estimation of model parameters of planar Neyman-Scott cluster processes. *Metrika* **39**, 67–74.

Stoyan, D. and Beneš, V. (1991). Anisotropy analysis for particle systems. *Journal of Microscopy* **164**(2), 159–168.

Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields*. John Wiley and Sons, New York.

Strauss, D. J. (1975). A model for clustering. *Biometrika* **62**, 467–475.

Taylor, H. M. and Karlin, S. (1994). *An Introduction to Stochastic Modeling* (Revised ed.). Academic Press, San Diego.

Thompson, V. L. and Greenkorn, R. A. (1988). Non-gaussian dispersion in model smokestack plumes. *AIChE Journal* **34**, 223–228.

Winer, B. J. (1971). *Statistical Principles in Experimental Design* (2nd ed.). McGraw-Hill Book Company, Inc., New York.

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics* **31**, 39–59.

Wright, S. (1969). *The Theory of Gene Frequencies*, Volume 2 of *Evolution and the Genetics of Populations*. University of Chicago Press, Chicago.

Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *Annals of Statistics* **11**, 95–103.

Zimmerman, D. L. (1994). Statistical analysis of spatial data. In Stanford and Vardeman (1994), Chapter 13, pp. 375–402.