University of Alberta

ESTIMABILITY AND LIKELIHOOD INFERENCE FOR GENERAL HIERARCHICAL MODELS USING DATA CLONING

by

Khurram Nadeem

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Mathematical and Statistical Sciences Department

Fall, 2013

Edmonton, Alberta.

To my family and Adan, my nephew.

Abstract

Hierarchical models constitute one of the most useful classes of statistical models with applications in a broad range of disciplines including, among others, social sciences, epidemiology and environmental sciences. The widely used linear mixed effects models, their extension to generalized linear mixed models (GLMMs), and state-space models all arise as special cases of general hierarchical models. These models provide a powerful framework for modeling the effects of latent processes, called random effects, whose variability is only manifested through the observed data. However, maximum likelihood estimation for these models poses significant challenges because the likelihood function involves intractable integrals whose dimension depends on the random effects structure.

In this thesis, we use data cloning; a simple computational method that exploits advances in Bayesian computation, in particular the Markov Chain Monte Carlo (MCMC) method, to obtain maximum likelihood estimators of the parameters along with their asymptotic standard errors in general hierarchical models. We also suggest a frequentist method to obtain prediction intervals for random effects. Determining estimability of the parameters in a hierarchical model is a very difficult problem in general. This thesis also develops a simple data cloning based graphical test to not only check if the full set of parameters is estimable but also, and more importantly, if a specified function of the parameters is estimable. We exemplify our methodology by analyzing various GLMMs and state-space models. Using a focal population time series of song sparrow (*Melospiza melodia)* on Mandarte Island, British Columbia, Canada, we

show that data cloning can be efficiently employed to fit nonlinear non-Gaussian state-space models for conducting population viability analyses in the presence of observation error and missing values.

The quality of MCMC based Bayesian inference, and for that matter, that of data cloning based estimates, is crucially dependent on appropriate diagnosis of MCMC chains' convergence. This thesis also develops a diagnostic method for convergence of MCMC algorithms using a new empirical characteristic function (ECF) based nonparametric test for comparing *k*-multivariate distributions. We show that the ECF based convergence diagnostic is particularly useful in cases where the target distribution is multimodal.

Acknowledgements

I owe my gratitude to many people who played a positive role in this accomplishment. I am thankful to all who read my thesis, provided support and encouragement and have been there to listen to and endure with me.

I am especially grateful to my supervisor Dr. Subhash Lele for his sagacious advice and continuous support that made this thesis a great learning experience. Your passion and enthusiasm for research is contagious and inspiring. Thank you for giving me the opportunity to grow my own thinking in science and research.

I also extend my sincere gratitude to Dr. Prasad, Dr. Lewis and Dr. Schmuland for being part of my supervisory committee. I thank you all for dedicating your valuable time to reviewing and providing constructive criticism throughout these years. I am also deeply indebted to Dr. Bruce Smith for being the external examiner on the committee and for his careful review of this thesis.

Finally, a big thanks to my family for your love, encouragement and patience. I love you for being so supportive in realizing this dream. I also want to thank all my friends, colleagues and teachers in Pakistan and here in Canada. You all have been so instrumental throughout my career.

Contents

List of Tables

List of Figures

List of Acronyms

AGQ	Adaptive Gaussian Quadrature
AIC	Akaike Information Criterion
AMCMC	Adaptive Markov Chain Monte Carlo
CAR	Conditionally Autoregressive
CDF	Cumulative Distribution Function
CF	Characteristic Function
DC	Data Cloning
DCINLA	Data Cloned Integrated Nested Laplace Approximation
DCLR	Data Cloned Likelihood Ratio
ECF	Empirical Characteristic Function
ED	Energy Distance
EM	Expectation Maximization
FLA	Fully Exponential Laplace Approximation
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model

GQ	Gaussian Quadrature
INLA	Integrated Nested Laplace Approximation
IS	Importance Sampling
IWLS	Iteratively Reweighted Least Squares
LME	Linear Mixed Effects
LR	Likelihood Ratio
LRT	Likelihood Ratio Test
MCEM	Monte Carlo Expectation Maximization
MCMC	Markov Chain Monte Carlo
MCNR	Monte Carlo Newton-Raphson
MLE	Maximum Likelihood Estimator
PL	Profile Likelihood
PMCMC	Particle Markov Chain Monte Carlo
PPI	Population Prediction Interval
PQL	Penalized Quasi-Likelihood
PSRF	Potential Scale Reduction Factor
PVA	Population Viability Analysis
REML	Restricted Maximum Likelihood
SMC	Sequential Monte Carlo
SMR	Standardized Mortality Rate

Chapter 1

Introduction

A common thread to many statistical inference problems is the non-availability of data that should have ideally been observed to effectively model the phenomenon of interest. This constraint quite often leads to a hierarchical modeling framework involving two model components: a model for the unobserved data and a model linking the observed data to the unobserved data. Hierarchical models comprise one of the most useful classes of models in statistics such as Linear mixed effect (LME) models (Searle et al. 1992) and their extension to generalized linear mixed models (McCulloch et al. 2008), and state-space models (de Valpine and Hastings 2002). They have widespread use in various fields, for example, longitudinal data analysis (Diggle et al. 1994), epidemiology (Clayton and Kaldor 1987) and ecology and environmental sciences (Clark and Gelfand 2006; Royle and Dorazio 2008).

The general two-stage hierarchical modeling framework (Hobert 2000) is defined as follows. Let $(\mathbf{y}_1^T, \mathbf{u}_1^T)^T$, $(\mathbf{y}_2^T, \mathbf{u}_2^T)^T$, …, $(\mathbf{y}_n^T, \mathbf{u}_n^T)^T$ be *n* independent random vectors whose joint distribution depends on an unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. The observed data vector $\mathbf{y}_{(n)} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)^T$ is modeled conditionally on the unobservable random (also called *latent*) effects $(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$. A separate model is assumed for the marginal distribution of the random effects. Complete specification of the hierarchical model then consists of the following two models:

Hierarchy 1: $\mathbf{y}_i|\mathbf{u}_i \sim f_i(\mathbf{y}_i; \boldsymbol{\theta}_1|\mathbf{u}_i)$ (1.1 a)

Hierarchy 2: $\mathbf{u}_i \sim g_i(\mathbf{u}_i; \boldsymbol{\theta}_2)$, (1.1 b)

where f_i and g_i are some parametric densities with known forms. In general, f_i may depend on covariates associated with $\mathbf{y}_i$, but we have suppressed such a dependence in our

notation. The corresponding likelihood function is given as the marginal distribution of the observed data $\mathbf{y}_{(n)}$, viewed as a function of unknown parameters $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}; \mathbf{y}_{(n)}) = \prod_{i=1}^{n} \int f_i(\mathbf{y}_i; \boldsymbol{\theta}_1 | \mathbf{u}_i) g_i(\mathbf{u}_i; \boldsymbol{\theta}_2) d\mathbf{u}_i. \quad (1.2)$$

We notate $\widehat{\boldsymbol{\theta}}$ as the maximum likelihood estimator (MLE); the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}; \mathbf{y}_{(n)})$.

Generally, maximization of (1.2) is an intractable problem unless f_i and g_i have a conjugate relationship. A notable example is the normal linear model (Searle et al. 1992) for which the likelihood function exists in closed form, that is, integrals in (1.2) can be computed analytically. On the other hand, the most important instance of a non-conjugate relationship between the model components arises in the class of generalized linear mixed models (GLMMs) for which equation (1.2) does not exist in a closed form. This, together with a usually large dimension of random effects $\mathbf{u}_i$, poses significant challenges in numerical computation of the integrals involved. Resultantly, likelihood based inference in GLMMs is generally based on various approximate methods (Breslow and Clayton 1993; McCulloch 1997; McCulloch et al. 2008; Pinheiro and Chao 2006). The most commonly used approaches to analyze hierarchical models are therefore Bayesian, based on the Markov Chain Monte Carlo (MCMC) algorithm and noninformative priors (Gilks et al.1996; Spiegelhalter et al. 2004; Robert and Casella 2005). We present a review of both Bayesian and likelihood based methods for analyzing GLMMs in Chapter 2.

Recently Lele et al. (2007) introduced an alternative MCMC based method, called data cloning (DC), to obtain maximum likelihood estimates (MLEs) and their standard errors in state-space models, a particular class of hierarchical models (de Valpine and Hastings 2002). Data cloning is related to simulated annealing algorithm (Brooks and Morgan 1995) and is an adaptation of a computational maximum likelihood approach developed by Robert (1993). See also Doucet et al. (2002), Kuk (2003), and Jacquier et al. (2007) for methods similar to DC. The main advantage of DC is that it coaxes the Bayesian computational machinery to obtain frequentist inference, thereby inheriting all the computational advantages of the Bayesian approach at the same time avoiding the pitfalls of having the inference depend on the choice of the prior distribution.

In this work we extend the DC algorithm in several directions. We demonstrate that DC can be efficiently employed to obtain MLEs of model parameters and those of

the corresponding asymptotic standard errors in general hierarchical models. A common feature of hierarchical models arising in applied work is their complexity which, coupled with data limitations, casts doubts about the identifiability of the overall model (Gustafson 2009; Lele 2010). Mathematical determination of model estimability is a very intractable problem in general. In this thesis we develop a simple DC based diagnostic tool to establish model estimability in general hierarchical models.

Another widely used class of hierarchical models consists of state-space time series models (de Valpine and Hastings 2002). These models are especially useful in population dynamics modeling because population time series are often available only in the form of estimates that are subject to observation error. State-space models provide a flexible tool to incorporate such errors and missing observations in estimating the underlying population growth model. Estimation of the growth model then allows one to forecast future population trajectories to estimate the extinction risk of a study population – a key component of population viability analysis (Mills 2008). In this thesis we develop an efficient DC based methodology to conduct population viability analysis (PVA) within a wide class of nonlinear growth models in the presence of observation error. To exemplify our methodology, we reanalyze the viability of a population previously studied by Sæther et al. (2000). While Sæther et al. (2000) simply assumed that abundance counts were error free; we use DC method to fit the state-space theta-logistic model (Gilpin and Ayala 1973) to assess the presence of observation error. We then apply a DC based algorithm (Ponciano et al. 2009) to conduct information based model selection that strongly confirms the presence of observation error. The analyses also highlight the need for incorporating other key population processes into the PVA such as spatial distribution, dispersal and habitat attributes.

The main attraction of the DC approach is its use of the well-known Bayesian MCMC methodology. The quality of the resulting statistical inference is therefore dependent on appropriate diagnosis of the MCMC chains convergence. However, the posterior distributions induced by hierarchical likelihoods are often multimodal, leading to poor MCMC mixing and thereby complicating convergence assessment. Towards this end, we introduce a new diagnostic method for assessing convergence of MCMC algorithms based on our new empirical characteristic function (ECF) based nonparametric test for comparing *k*-multivariate distributions. We show that the new test is very sensitive in detecting shifts in different features of a multivariate density such as scale and multimodality and, therefore, plays a key role in assessing convergence to multimodal posteriors.

The rest of the thesis is organized as follows. Chapter 2 provides an overview of the recent and commonly used techniques for analyzing general hierarchical models. In Chapter 3, we develop a DC based algorithm to conduct likelihood based inference in hierarchical models, including maximum likelihood estimation, random effects' prediction and estimability diagnostics. Chapter 4 demonstrates applicability of the DC algorithm for analyzing general state-space models. We illustrate the methodology in the context of PVA where we fit population growth models in the presence of observation error and missing data. We further illustrate how information theoretic model selection can be performed using data cloning. In chapter 5, we formulate an MCMC convergence diagnostic procedure using our new ECF based test for comparing k multivariate populations that we also develop in the same chapter. Finally, in Chapter 6, conclusions and avenues for future research are presented.

Chapter 2

Overview

This chapter provides an overview of the existing approaches to analyzing general hierarchical models within both Bayesian and frequentist frameworks. In particular, we focus on estimation techniques for GLMMs which we describe in Sections 2.1 to 2.5. In Section 2.6 we briefly review more recent Monte Carlo based methods for estimation and prediction in hierarchical models, in particular, state-space models. Section 2.7 provides a summary of the chapter.

The common building block in both Bayesian and frequentist paradigms is the likelihood function of the model parameters evaluated at the observed data. The fundamental feature differentiating the frequentist and Bayesian approaches to statistical inference is that the former assumes the model parameter vector $\boldsymbol{\theta}$ to be a fixed point in the parameter space, while the latter incorporates *prior uncertainty* about $\boldsymbol{\theta}$ via a *prior probability distribution*, $\pi(\boldsymbol{\theta})$. A detailed commentary on various approaches to statistical inference is provided by Barnett (1999). Throughout this thesis, we refer to the frequentist methods as 'likelihood based methods'.

Generalized linear mixed models are a part of a larger class of hierarchical linear models called multilevel models (Raudenbush and Bryk 2002). Observations in these models are arranged in clusters in a hierarchical structure. For instance, in educational research, observations on students are usually nested in classroom, school and county levels (Bryk and Raudenbush 1988). Another example is that of longitudinal studies where repeated observations are nested within subjects (Diggle et al.1994). Because observations in a given cluster or group tend to be similar, they are often positively correlated. Ignoring this group-induced correlation structure in data results in underestimation of standard errors and thereby leads to spuriously significant group effects (Raudenbush and

Bryk 2002). GLMMs attempt to incorporate the correlation structure by introducing a separate probability model for cluster level effects, also called *latent* or random effects. These models are derived from the well-known generalized linear model (GLM) by incorporating random effect terms in the linear predictor. We proceed with a brief introduction to GLMs.

2.1 Generalized Linear Models

At the heart of statistical modeling are the regression models defined in terms of explanatory variables, $\mathbf{x}_i^T$. The simplest of these are the general linear regression models that assume a linear functional relation between the mean of a continuous response Y_i and the covariates $\mathbf{x}_i^T$. However, the observed response is often a nominal or a count variable that does not relate to $\mathbf{x}_i^T$ in a straightforward linear fashion. For instance, when Y_i is a binomial response, interest lies in modeling the probability of success, π_i, in terms of $\mathbf{x}_i^T$. Clearly, a linear regression model is implausible here as π_i is restricted in the interval (0,1). Generalized linear models form a flexible class of models that admit nonlinear regression relations by transforming the conditional mean response, $E(Y_i|\mathbf{x}_i^T)$, to a *linear predictor* $\mathbf{x}_i^T\boldsymbol{\beta}$ (Nelder and Wedderburn 1972). Formally, GLMs achieve this by assuming the response distribution to be a member of the exponential family of distributions (Darmois 1935; Pitman 1936) whose *natural* parameter is related to covariates $\mathbf{x}_i^T$ via a smooth invertible link function $g(.)$. That is, conditional on the realized vector $\mathbf{x}_i^T$, the distribution of Y_i is given as

$$f(y_i;\boldsymbol{\beta},\phi) = e^{\left\{\frac{\left(y_i\theta_i - b(\theta_i)\right)}{a(\phi)} + c(y_i,\phi)\right\}},$$

where $a(.)$, $b(.)$ and $c(.)$ are known functions, θ_i's are *natural* parameters, and ϕ is the dispersion parameter. We can show that $\mu_i = E(Y_i|\mathbf{x}_i^T) = b'(\theta_i)$ and $Var(Y_i|\mathbf{x}_i^T) = b''(\theta_i)a(\phi)$, where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$ respectively. The model is then specified in terms of the link function $g(.)$ such that $g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the *p*-dimensional vector of regression coefficients and η_i is the linear predictor. The link is said to be *canonical* when $\theta_i = \eta_i$. McCullagh and Nelder (1989) provide a comprehensive treatment of GLMs covering models for both nominal and count data. A detailed exposition of the applications of GLMs in categorical data analyses can be found in Agresti (2002).

Denoting $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and $\boldsymbol{\varphi} = (\boldsymbol{\beta}^T, \phi)^T$ to be the vectors of response values and model parameters, the GLM log-likelihood function is given as

$$l(\boldsymbol{\varphi}; \mathbf{y}) = \sum_i \left\{ \frac{\left(y_i \theta_i - b(\theta_i)\right)}{a(\phi)} + c(y_i, \phi) \right\}.$$

The MLE of $\boldsymbol{\beta}$, which we denote as $\widehat{\boldsymbol{\beta}}$, is then the solution of the score equations

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i) x_{ij}}{a(\phi) b''(\theta_i) g'(\mu_i)} \ , j = 1, 2, \ldots, p.$$

In general, these equations must be solved numerically as no closed form solution exists. The dispersion parameter ϕ can be estimated separately from regression residuals. There exist efficient algorithms for computing $\widehat{\boldsymbol{\beta}}$ based on iteratively reweighted least squares (IWLS) such as Gauss-Newton and Fisher's scoring methods (McCullagh and Nelder 1989; Lang 2004). Recently Wang (2007) has extended these algorithms to situations where model parameters are subject to constraints. Hypothesis testing for regression coefficients can be performed using Wald's or score tests (McCullagh and Nelder 1989). Likelihood ratio test (LRT) can be used to compare nested models. For further details on inferential tools in GLMs, the reader is referred to McCullagh and Nelder (1989).

2.2 Generalized Linear Mixed Models

These models are an extension over both GLMs and linear mixed effects (LME) models. The LME model attempts to relate the mean of a Gaussian response vector $\mathbf{y}_i = (y_{i1}, y_{i1}, \ldots, y_{in_i})^T$, in cluster *i*, to observed covariates and random effects $\mathbf{u}_i$ via the linear predictor $\boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i$, where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known design matrices of the fixed and random effects respectively. Random effects are introduced in the model to incorporate the correlation structure in the components of $\mathbf{y}_i$ induced by the hierarchical structure of the data. The LME model then takes the form $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ and $\mathbf{u}_i$ are independent, zero-mean and Gaussian random vectors with variance-covariance matrices $\sigma_\varepsilon^2 \mathbf{I}_{n_i}$ and $\boldsymbol{\Sigma}(\boldsymbol{\psi})$ respectively; $\mathbf{I}_{n_i}$ is an identity matrix and $\boldsymbol{\Sigma}$ depends on unknown parameters $\boldsymbol{\psi}$. The likelihood function exists in closed form and the MLE's can be efficiently computed using the mixed model equations of Henderson (1950). Henderson et al. (1959) further showed that the resulting estimators of model parameters and those of random effects are *best linear unbiased estimator* (BLUE) and *best linear unbiased predictor*

(BLUP) respectively. Robinson (1991) provides a detailed account of BLUP in LME models and their relevance to other statistical estimation problems.

A natural extension of the LME models is to allow y_{ij} to be a categorical variable with distribution in the exponential family. This extension results in a class of models known as generalized linear mixed models (GLMMs). A single-level GLMM is defined as follows. Conditional on $\mathbf{u}_i$, y_{ij} are *independent and identically distributed* (iid) so that the conditional joint distribution of $\mathbf{y}_i$ is given as

$$f(\mathbf{y}_i|\mathbf{u}_i) = \prod_{i=1}^{n_i} e^{\left\{\frac{\left(y_{ij}\theta_{ij} - b(\theta_{ij})\right)}{a(\phi)} + c(y_{ij},\phi)\right\}}$$
$$= e^{\left\{\frac{\left(\mathbf{y}_i^T\boldsymbol{\theta}_i - \mathbf{1}^T b(\boldsymbol{\theta}_i)\right)}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}_i,\phi)\right\}},$$

where $a(.)$, $b(.)$ and $c(.)$ are defined as before, θ_{ij}'s are the natural parameters such that $\boldsymbol{\mu}_i = E(\mathbf{y}_i|\mathbf{u}_i) = b'(\boldsymbol{\theta}_i)$. As in GLMs, the model is then further specified using a smooth invertible link function $g(.)$ such that $g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i$. The random effects $\mathbf{u}_i$ are assumed to be Gaussian with mean $\mathbf{0}$ and variance-covariance $\boldsymbol{\Sigma}(\boldsymbol{\psi})$. The link function is typically the canonical link, i.e. $\boldsymbol{\eta}_i = \boldsymbol{\theta}_i$. Thus, the joint density of $\mathbf{y}_i$ and $\mathbf{u}_i$ takes the form

$$f(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\varphi}) = (2\pi)^{-q/2}|\boldsymbol{\Sigma}(\boldsymbol{\psi})|^{-1/2} e^{\left\{\frac{\left(\mathbf{y}_i^T(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{u}_i) - \mathbf{1}^T b(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{u}_i)\right)}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}_i,\phi) + \frac{\mathbf{u}_i^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}_i}{2}\right\}}, \quad (2.1)$$

where $\boldsymbol{\varphi}$ is the vector of model parameters and q is the length of $\mathbf{u}_i$. The likelihood function for the observed data $\mathbf{y}_{(n)} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)^T$ is the following integrated likelihood:

$$L\left(\boldsymbol{\varphi}; \mathbf{y}_{(n)}\right) = \prod_{i=1}^{n} \int f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|\mathbf{u}_i)\, g(\mathbf{u}_i; \boldsymbol{\Sigma}(\boldsymbol{\psi}))\, d\mathbf{u}_i$$
$$= \prod_{i=1}^{n_i} \int f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|\mathbf{u}_i)\, g(\mathbf{u}_i; \boldsymbol{\Sigma}(\boldsymbol{\psi}))\, d\mathbf{u}_i, \quad (2.2)$$

where $g(\mathbf{u}_i; \boldsymbol{\Sigma}(\boldsymbol{\psi})$ is the marginal density of $\mathbf{u}_i$, i.e. a Normal distribution with mean $\mathbf{0}$ and variance-covariance $\boldsymbol{\Sigma}(\boldsymbol{\psi})$.

As indicated earlier, the above likelihood function exists in closed form for the LME model where response $\mathbf{y}_i$ is Gaussian with the *identity* link: $\boldsymbol{\mu}_i = \boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i$. However, in general, the integrals appearing in (2.2) have no closed form, resulting in an intractable maximum likelihood estimation problem. Direct maximization of (2.2) using numerical integration is infeasible as the integral dimension depends on q, the length of $\mathbf{u}_i$, which is generally large. Common estimation techniques are, therefore, based either on approximations of the likelihood function (Breslow and Clayton 1993), or on its Monte Carlo estimation (McCulloch 1997) or on numerical approximations (Pinheiro and

Chao 2006) of the integrals appearing in (2.2). This difficulty has led to widespread use of Bayesian estimation of GLMMs which circumvents high dimensional integration by invoking MCMC algorithms and noninformative priors (Gilks et al. 1996; Spigelhalter et al. 2004). We briefly review these inferential techniques in the sequel below.

2.3 Direct Maximization of the Likelihood

2.3.1 Expectation Maximization Algorithm

The expectation maximization (EM) algorithm was introduced by Dempster et al. (1977) in their seminal work on maximum likelihood estimation in the presence of missing observations. The algorithm can also be coaxed to compute MLEs in GLMMs by treating latent variables as missing data. The key aspect of the algorithm is to break down the optimization problem into a series of simpler maximization problems whose solution converges to the MLE. Assuming $\boldsymbol{\varphi}^{(i)}$ to be the estimate of $\boldsymbol{\varphi}$ at the i^{th} iteration, the EM algorithm seeks to find an update estimate $\boldsymbol{\varphi}^{(i+1)}$ such that $l(\boldsymbol{\varphi}^{(i+1)}) \geq l(\boldsymbol{\varphi}^{(i)})$, where $l(.)$ denotes the log-likelihood function. Naturally, the objective is to maximize the difference

$$l\left(\boldsymbol{\varphi}^{(i+1)}\right) - l\left(\boldsymbol{\varphi}^{(i)}\right) = logf\left(\mathbf{y}_{(n)}; \boldsymbol{\varphi}^{(i+1)}\right) - logf\left(\mathbf{y}_{(n)}; \boldsymbol{\varphi}^{(i)}\right).$$

We can show using Jensen's inequality (1906) that

$$\begin{aligned} l\left(\boldsymbol{\varphi}^{(i+1)}\right) - l\left(\boldsymbol{\varphi}^{(i)}\right) &\geq \int f\left(\mathbf{u}; \boldsymbol{\varphi}^{(i)} | \mathbf{y}_{(n)}\right) \log\left(\frac{f\left(\mathbf{y}_i; \boldsymbol{\varphi}^{(i+1)} | \mathbf{u}\right) g\left(\mathbf{u}; \boldsymbol{\varphi}^{(i+1)}\right)}{f\left(\mathbf{y}_i; \boldsymbol{\varphi}^{(i)} | \mathbf{u}\right) g\left(\mathbf{u}; \boldsymbol{\varphi}^{(i)}\right)}\right) d\mathbf{u} \\ &\equiv \Delta\left(\boldsymbol{\varphi}^{(i+1)} | \boldsymbol{\varphi}^{(i)}\right), \end{aligned}$$

which gives

$$l\left(\boldsymbol{\varphi}^{(i+1)}\right) \geq l\left(\boldsymbol{\varphi}^{(i)}\right) + \Delta\left(\boldsymbol{\varphi}^{(i+1)} | \boldsymbol{\varphi}^{(i)}\right) \equiv \varpi\left(\boldsymbol{\varphi}^{(i+1)} | \boldsymbol{\varphi}^{(i)}\right),$$

and where $\varpi(.)$ is concave and bounded above by $l(\boldsymbol{\varphi})$. Using these properties, Dempster et al. (1977) showed that the choice of $\boldsymbol{\varphi}^{(i+1)}$ producing maximum possible increment in $l(\boldsymbol{\varphi})$ while moving from iteration *i* to *i*+1 is given as

$$\begin{aligned} \boldsymbol{\varphi}^{(i+1)} &= \text{argmax}_{\boldsymbol{\varphi}}\left\{\varpi\left(\boldsymbol{\varphi}^{(i+1)} | \boldsymbol{\varphi}^{(i)}\right)\right\} \\ &= \text{argmax}_{\boldsymbol{\varphi}}\left[E_{\mathbf{u}; \boldsymbol{\varphi}^{(i)} | \mathbf{y}_{(n)}}\left\{\log f\left(\mathbf{y}_{(n)}, \mathbf{u}; \boldsymbol{\varphi}\right)\right\}\right], \end{aligned} \tag{2.3}$$

where $f\left(\mathbf{y}_{(n)}, \mathbf{u}; \boldsymbol{\varphi}\right)$ is the *complete* data likelihood, i.e. the joint distribution of observables and the random effects given in (2.1). Thus, the EM algorithm consists of iteratively applying the following two steps until convergence.

1. E-Step: Evaluate the expectation in (2.3) under the conditional distribution of the random effects, $f\left(\mathbf{u};\boldsymbol{\varphi}^{(i)}|\mathbf{y}_{(n)}\right)$.
2. M-Step: Maximize this expectation over $\boldsymbol{\varphi}$ to obtain $\boldsymbol{\varphi}^{(i+1)}$.

Under mild regularity conditions, Wu (1983) showed that the EM sequence of estimates $\left\{\boldsymbol{\varphi}^{(i+1)}\right\}_{i=1}^{\infty}$ converges to a local maximum of the likelihood surface. However, the E-Step of the EM algorithm is intractable in most GLMMs since the conditional density $f\left(\mathbf{u};\boldsymbol{\varphi}^{(i)}|\mathbf{y}_{(n)}\right)$ does not exist in closed form. This is evident from noticing that

$$f\left(\mathbf{u};\boldsymbol{\varphi}|\mathbf{y}_{(n)}\right)=\frac{f\left(\mathbf{y}_{(n)},\mathbf{u};\boldsymbol{\varphi}\right)}{\int f\left(\mathbf{y}_{(n)},\mathbf{u};\boldsymbol{\varphi}\right)d\mathbf{u}}, \tag{2.4}$$

where the integral in the denominator is what we wish to avoid. This leads us to a Monte Carlo version of the EM algorithm described below.

2.3.2 Monte Carlo Expectation Maximization

To avoid direct computation of the intractable expectation arising in the E-Step (2.3), Wei and Tanner (1990) suggested approximating it by a Monte Carlo estimate. McCulloch (1997) implemented this method by sampling from $f\left(\mathbf{u};\boldsymbol{\varphi}|\mathbf{y}_{(n)}\right)$ using MCMC, resulting in the so called Monte Carlo EM (MCEM) algorithm. To give a brief sketch of the algorithm, let us consider the expected complete log-data likelihood of a GLMM, i.e.

$$E\left(\log f\left(\mathbf{y}_{(n)},\mathbf{u};\boldsymbol{\varphi}\right)\right)=E\left(\log f\left(\mathbf{y}_{(n)};\boldsymbol{\beta},\phi|\mathbf{u}\right)\right)+E\left(\log g\left(\mathbf{u};\boldsymbol{\Sigma}(\boldsymbol{\psi})\right)\right).$$

So, given an MCMC sample of size *N* from the conditional distribution in equation (2.4), Monte Carlo estimates of the expectations appearing in the right hand side of the equation above are respectively given as

$$\hat{E}\left(\log f\left(\mathbf{y}_{(n)};\boldsymbol{\beta},\phi|\mathbf{u}\right)\right)=\frac{1}{N}\sum_{j=1}^{N}\log f\left(\mathbf{y}_{(n)};\boldsymbol{\beta},\phi|\mathbf{u}^{(j)}\right),$$

and

$$\hat{E}\left(\log g\left(\mathbf{u};\boldsymbol{\Sigma}(\boldsymbol{\psi})\right)\right)=\frac{1}{N}\sum_{j=1}^{N}\log g\left(\mathbf{u};\boldsymbol{\Sigma}(\boldsymbol{\psi})\right).$$

The implementation of the M-Step is then facilitated by the fact that $\log f\left(\mathbf{y}_{(n)};\boldsymbol{\beta},\phi|\mathbf{u}^{(j)}\right)$ is the standard GLM likelihood whose maximization is straightforward using the Fisher scoring algorithm. Similarly, the second empirical expectation is simple to maximize since $g\left(\mathbf{u};\boldsymbol{\Sigma}(\boldsymbol{\psi})\right)$ is a Gaussian density.

As an alternative version of MCEM, McCulloch (1997) also introduced a Monte Carlo version of a Newton-Raphson (Lang 2004) type algorithm (MCNR) for simultaneously solving the expected score equations

$$E\left[\frac{\partial \log f(\mathbf{y}_{(n)};\boldsymbol{\beta},\phi|\mathbf{u})}{\partial \boldsymbol{\theta}}\right] = 0, \tag{2.5 a}$$

and

$$E\left[\frac{\partial \log g(\mathbf{u};\boldsymbol{\Sigma}(\boldsymbol{\psi}))}{\partial \boldsymbol{\psi}}\right] = 0, \tag{2.5 b}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$. Here, the second equation (2.5 b) is easy to solve since random effects are assumed Gaussian, while (2.5 a) can be handled using a scoring approach similar to that of in a standard GLM. McCulloch (1997), therefore, arrived at the following iteration equation

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(i)} + E\left[\mathbf{X}^T\mathbf{W}\left(\boldsymbol{\theta}^{(m)},\mathbf{u}\right)\mathbf{X}|\mathbf{y}_{(n)}\right]^{-1} \times$$
$$\mathbf{X}^T\left(E\left[\mathbf{X}\mathbf{W}\left(\boldsymbol{\theta}^{(m)},\mathbf{u}\right)\frac{\partial \boldsymbol{\eta}_{(n)}}{\partial \boldsymbol{\mu}_{(n)}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}\left(\mathbf{y}_{(n)} - \boldsymbol{\mu}_{(n)}\left(\boldsymbol{\theta}^{(m)},\mathbf{u}\right)\right)|\mathbf{y}_{(n)}\right]\right), \tag{2.6}$$

where

$$\boldsymbol{\mu}_{(n)}\left(\boldsymbol{\theta}^{(m)},\mathbf{u}\right) = E\left(\mathbf{y}_{(n)}|\mathbf{u}\right),$$
$$\mathbf{W}^{-1}\left(\boldsymbol{\theta}^{(m)},\mathbf{u}\right) = diag\left\{\frac{\partial \eta_i}{\partial \mu_i}\right\}^2 Var\left(\mathbf{y}_{(n)}|\mathbf{u}\right),$$
$$\frac{\partial \boldsymbol{\eta}_{(n)}}{\partial \boldsymbol{\mu}_{(n)}} = \left\{\frac{\partial \eta_i}{\partial \mu_i}\right\},$$

and **X** is the overall design matrix associated with the fixed effects. This scoring approach then proceeds iteratively by solving (2.6) together with (2.5 b) and an equation for the dispersion parameter ϕ. The expectations appearing in equation (2.6) can be estimated using Monte Carlo simulation, i.e. by generating samples from the conditional density $f\left(\mathbf{u};\boldsymbol{\varphi}|\mathbf{y}_{(n)}\right)$ via the MCMC algorithm. The whole iterative procedure is run until reasonable stabilization of the estimates is achieved. A computationally attractive feature of the MCNR approach is that it automates the M-Step of the EM algorithm.

Although Monte Carlo estimation of the expectations is quite appealing, the usual EM assurance that $l(\boldsymbol{\varphi}^{(i+1)}) \geq l(\boldsymbol{\varphi}^{(i)})$, no longer holds. Convergence of both MCEM and MCNR is also tricky because of Monte Carlo error. This results in requiring very high Monte Carlo sample sizes, making the algorithms computationally intractable. There exist methods to address this limitation. For instance, Booth and Hobert (1999) employed a rejection sampling scheme (Geweke 1996) to develop a rule that automatically determines sufficient sample size by using estimates of the Monte Carlo error. Zipunnikov and Booth (2006) replaced the MCMC step by randomized spherical-radial integration (Genz and Monahan, 1997). Their MCEM algorithm substantially reduces the computational burden as it involves Monte Carlo simulation from standard distributions. Furthermore,

the E-Step takes the form of a standard GLM leading to a simple IWLS procedure in the M-Step.

Nevertheless, both MCEM and MCNR can run into convergence issues when the likelihood surface is multimodal, e.g. in case of variance component problems (Searl et al. 1992). The EM algorithm in such cases may converge to a local maximum while MCNR may not converge at all due to the non-concavity of the likelihood surface (McCulloch 1997). A method that is designed to address this problem is presented below.

2.3.3 Simulated Maximum Likelihood

The method is related to Monte Carlo integration of definite integrals using *importance sampling* (IS) (Hammersley and Handscomb 1964) described as follows. Suppose we wish to evaluate an integral of the form $I_f = \int f(x)dx$, for some $f(x) \geq 0$. The idea is to express the integral as

$$I_f = \int \frac{f(x)}{p(x)} p(x)dx$$
$$= E_{p(x)}\left(\frac{f(x)}{p(x)}\right),$$

where $p(x)$is an *easy to sample* density function known as the IS distribution. Then, generating a large random sample $\{x_j\}_{j=1}^{N}$ from $p(x)$, IS estimate of I_f is given as $I_f \approx \sum_{j=1}^{N} \frac{f(x_j)}{p(x_j)}$, where $f(x_j)/p(x_j)$ are called *importance weights*. For application of IS in Bayesian inference, we refer the reader to Gelman et al. (2003).

Now, recall from equation (2.2) that

$$L(\boldsymbol{\varphi}; \mathbf{y}_{(n)}) = \prod_{i=1}^{n_i} \int f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|\mathbf{u}_i)\, g(\mathbf{u}_i; \boldsymbol{\Sigma}(\boldsymbol{\psi}))\, d\mathbf{u}_i,$$

where, for observed data $\boldsymbol{y}_i$, the i^{th} integral can be written as the expectation

$$E_g\big(f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|\mathbf{u}_i)\big) = \int f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|\mathbf{u}_i)\, g(\mathbf{u}_i; \boldsymbol{\Sigma}(\boldsymbol{\psi}))d\mathbf{u}_i,$$

suggesting that a Monte Carlo estimate can be obtained by sampling from the marginal distribution $g(\mathbf{u}_i; \boldsymbol{\Sigma}(\boldsymbol{\psi}))$. However, since $\boldsymbol{\Sigma}(\boldsymbol{\psi})$ must be estimated in practice, an IS estimate of (2.2) can be obtained as follows (Geyer and Thompson 1992; Gelfand and Carlin 1993)

$$L(\boldsymbol{\varphi}; \mathbf{y}_{(n)}) \approx \prod_{i=1}^{n_i} \left\{ \frac{1}{N} \sum_{j=1}^{N} \frac{f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|\mathbf{u}^{(j)}) g(\mathbf{u}^{(j)}; \boldsymbol{\Sigma}(\boldsymbol{\psi}))}{p_{\mathbf{u}}(\mathbf{u}^{(j)})} \right\}, \tag{2.7}$$

where $\{\mathbf{u}^{(j)}\}_{j=1}^{N}$ are simulated from the IS distribution $p_{\mathbf{u}}(\mathbf{u})$. The simulated likelihood is then maximized where either a single simulated importance sample is used or $p_{\mathbf{u}}(\mathbf{u})$ is iteratively updated using multiple simulations.

The simulated maximum likelihood (SML) approach is attractive in that it provides an unbiased estimate of the likelihood function and the parameter estimates converge to MLEs as the Monte Carlo sample size increases. However, Monte Carlo error remains very high in practice unless a good initial guess for $\widehat{\boldsymbol{\varphi}}$ is available (Jank and Booth 2003; McCulloch 1997). McCulloch (1997), therefore, suggested a hybrid SML algorithm starting it with the initial estimates obtained either from MCEM or MCNR. His simulation study based on a logit-normal model showed superior performance of the hybrid approach. It helped overcome convergence problems in both MCEM and MCNR and yielded a reliable implementation of the SML approach.

2.4 Approximations of the Likelihood

2.4.1 Penalized Quasi-Likelihood

Computational difficulties in maximizing the GLMM likelihood has resulted in various simpler approximations to the likelihood function itself. One such approximation is the penalized quasi-likelihood (PQL) initially proposed as an approximate Bayes procedure for certain common GLMMs by Laird (1978) and Stiratelli (1984). The PQL procedure leads to an approximate version of the Henderson et al.'s (1959) mixed-model equations arising from maximizing the joint distribution $f(\mathbf{y}, \mathbf{u}; \boldsymbol{\varphi})$ with respect to both $\boldsymbol{\varphi}$ and random effects $\mathbf{u}$ (Schall 1991; Breslow and Clayton 1993; Wolfinger 1993). The method is more flexible than the full ML procedure as only the first two moments of the conditional density $f(\mathbf{y}|\mathbf{u})$ need to be specified in terms of the GLMM model parameters. That is, we only assume that the conditional mean of the response vector $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ given the random effects $\mathbf{u}$, satisfies $E(y_i|\mathbf{u}) = \mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\mathbf{u})$, where $g(.)$ is the link function as defined before; and that $Var(y_i|\mathbf{u}) = a_i(\phi)v(\mu_i)$; $v(.)$ is a known variance function and a_i is a known constant.

The above parameterization leads to the following quasi-likelihood (Breslow and Clayton 1993),

$$e^{ql(\boldsymbol{\varphi},\mathbf{u})} \propto |\boldsymbol{\Sigma}(\boldsymbol{\psi})|^{-1/2} \int e^{-\frac{1}{2\phi}\sum_{i=1}^{n} d_i - \frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}}\, d\mathbf{u}, \tag{2.8}$$

where

$$d_i = -2\int_{y_i}^{\mu_i} \frac{y_i-\omega}{a_i(\phi)v(\omega)}d\omega$$

is known as the quasi-deviance and is related to the GLM of $\mathbf{y}$ conditional on $\mathbf{u}$, with $a_i(\phi) = \frac{\phi}{w_i}$; w_i are known weights. It can be shown, therefore, that

$$d_i = 2\{L(y_i; y_i) - L(y_i; \mu_i)\},$$

where $L(y_i; \mu_i)$ is the likelihood of y_i given μ_i. The PQL procedure then proceeds as follows. We rewrite equation (2.8) as

$$e^{ql(\boldsymbol{\varphi},\mathbf{u})} \propto |\boldsymbol{\Sigma}(\boldsymbol{\psi})|^{-1/2}\int e^{-\kappa(\mathbf{u})}\,d\mathbf{u},$$

where

$$\kappa(\mathbf{u}) = \frac{1}{2}\left[\frac{1}{\phi}\sum_{i=1}^{n} d_i - \mathbf{u}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}\right].$$

The integral arising in the right hand side of the equation above is then amenable to applying Laplace approximation (Tierney and Kadane 1986), i.e.

$$\int e^{-\kappa(\mathbf{u})}d\mathbf{u} \approx c|\kappa''(\mathbf{u}_o)|^{-1}e^{-\kappa(\mathbf{u}_o)},$$

where c is a constant and $\mathbf{u}_o$ minimizes $\kappa(\mathbf{u})$ so that $\kappa'(\mathbf{u}_o) = 0$ and $\kappa''(\mathbf{u}_o) > 0$. Thus, ignoring the constant c, we get

$$ql(\boldsymbol{\varphi}, \mathbf{u}) \approx -\frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\psi})| - \frac{1}{2}|\kappa''(\mathbf{u}_o)| - \kappa(\mathbf{u}_o), \tag{2.9}$$

where $\mathbf{u}_o$ is the solution to

$$\kappa'(\mathbf{u}) = -\sum_{i=1}^{n}\frac{(y_i-\mu_i)\mathbf{z}_i}{\phi a_i v(\mu_i)g'(\mu_i)} + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u} = 0,$$

that minimizes $\kappa(\mathbf{u})$. Further differentiation with respect to $\mathbf{u}$ yields

$$\kappa''(\mathbf{u}) = \sum_{i=1}^{n}\frac{\mathbf{z}_i^T\mathbf{z}_i}{\phi a_i v(\mu_i)[g'(\mu_i)]^2} + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}) + \mathbf{R}, \tag{2.10}$$

where $\mathbf{R}$ is the remainder term that is shown to have expectation zero. Therefore, in probability as a function of n, we can assume it to be of lower order than the two leading terms (Breslow and Clayton 1993). Thus, dropping $\mathbf{R}$ from (2.10) leads to the approximation

$$ql(\boldsymbol{\varphi}, \mathbf{u}) \approx \mathbf{Z}^T\mathbf{W}\mathbf{Z} + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}), \tag{2.11}$$

where $\mathbf{Z}$ is the design matrix with rows $\mathbf{z}_i^T$, and $\mathbf{W} = diag\{w_i\}$, with GLM iterated weights $w_i = \left\{\phi a_i v(\mu_i)[g'(\mu_i)]^2\right\}^{-1}$ (e.g., McCullagh and Nelder 1989, Section 2.5). Now combining equations (2.9) and (2.11) yields,

$$ql(\boldsymbol{\varphi}, \mathbf{u}) \approx -\frac{1}{2}\left\{|\mathbf{I} + \mathbf{Z}^T\mathbf{W}\mathbf{Z}\boldsymbol{\Sigma}(\boldsymbol{\psi})| + \frac{1}{\phi}\sum_{i=1}^{n} d_i + \mathbf{u}_o^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}_o\right\}, \tag{2.12}$$

Breslow and Clayton (1993) further approximated equation (2.12) by assuming that the GLM iterative weights vary slowly as a function of the mean. Thus, as the first term in (2.12) depends on $\boldsymbol{\beta}$ only through $\mathbf{W}$, ignoring it leads to the approximation,

$$ql(\boldsymbol{\varphi}, \mathbf{u}) \approx -\frac{1}{2\phi}\sum_{i=1}^{n} d_i - \frac{1}{2}\mathbf{u}_o^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}_o. \tag{2.13}$$

This is the same as Green's (1987) PQL which he developed for analyzing semiparametric regression models. We now maximize (2.13) to obtain estimates $\left(\widehat{\boldsymbol{\beta}}, \hat{\mathbf{u}}\right) = \left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi}), \hat{\mathbf{u}}(\boldsymbol{\psi})\right)$, where $\hat{\mathbf{u}}(\boldsymbol{\psi}) = \mathbf{u}_o\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi})\right)$. So, differentiation with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ produces the following score equations,

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)\mathbf{x}_i}{\phi a_i v(\mu_i) g'(\mu_i)} = 0, \tag{2.14}$$

and

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)\boldsymbol{z}_i}{\phi a_i v(\mu_i) g'(\mu_i)} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u} = 0, \tag{2.15}$$

Breslow and Clayton (1993) developed an IWLS algorithm for solving these nonlinear equations by modifying a Fisher's scoring algorithm proposed earlier by Green (1987). Their approach is attractive as it leads to BLUP of $(\boldsymbol{\beta}, \mathbf{u})$ by exploiting the close correspondence of the above score equations to mixed model equations of Henderson et al. (1959). More recently, Jiang (2000) has developed a nonlinear Gauss-Seidel algorithm for solving (2.14) and (2.15) that converges globally for virtually all typical GLMM problems. This method is particularly useful when there exist a large number of random effects that slow down the BLUP computations. Notice that, in solving (2.14) and (2.15), we assumed that $\boldsymbol{\psi}$ is known and fixed. This, of course, needs to be estimated in practice. Breslow and Clayton (1993) proposed a profile likelihood method by substituting the maximized value of (2.13) in equation (2.12) leading to a profile quasi-likelihood function in $\boldsymbol{\psi}$. They further showed that the resulting estimator is similar in spirit to restricted maximum likelihood (REML) estimation in linear mixed models (Patterson and Thompson 1971).

McCulloch (1997) points out that the PQL method is based on maximization of a quasi-likelihood and, therefore, should be regarded as a new estimation procedure in its own right. Not surprisingly, the series of approximations involved in PQL induce substantial bias in the resulting estimates. In fact, PQL estimates are known to be inconsistent and the bias cannot be corrected for even after applying higher order Laplace approximation (Lin and Breslow 1996). McCulloch (1997) conducted a simulation study to evaluate performance of PQL estimates by analyzing a binary-logistic GLMM with nor-

mally distributed intercepts. The simulations confirmed that parameter estimates were heavily biased for both fixed effects and the variance component.

2.4.2 Laplace Approximation

Let us recall from equation (2.2) that the GLMM likelihood in the i^{th} data cluster is given by a high dimensional integral in random effects $\mathbf{u}_i$,i.e.

$$\begin{aligned} L_i &= \int f(\boldsymbol{y}_i; \boldsymbol{\beta}, \phi|\mathbf{u}_i)\, g(\mathbf{u}_i; \boldsymbol{\Sigma}(\boldsymbol{\psi}))d\mathbf{u}_i \\ &= (2\pi)^{-q/2}|\boldsymbol{\Sigma}(\boldsymbol{\psi})|^{-1/2} \int e^{\kappa(\boldsymbol{\varphi},\, \boldsymbol{y}_{i,}\, \boldsymbol{u}_i)} d\mathbf{u}_i, \end{aligned} \tag{2.16}$$

where, from equation (2.1),

$$\begin{aligned} \kappa\big(\boldsymbol{\varphi},\, \mathbf{y}_{i,}\, \mathbf{u}_i\big) &= \frac{\left(\mathbf{y}_i^T(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{u}_i) - \mathbf{1}^T b(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{u}_i)\right)}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}_i, \phi) + \mathbf{u}_i^T\boldsymbol{\Sigma}(\boldsymbol{\psi})\mathbf{u}_i/2 \\ &= \frac{\left(\mathbf{y}_i^T\boldsymbol{\eta}_i - \mathbf{1}^T b(\boldsymbol{\eta}_i)\right)}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}_i, \phi) + \mathbf{u}_i^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}_i/2. \end{aligned}$$

Rather than basing the inference on a quasi-likelihood function as in the PQL approach, Pinheiro and Bates (1995) considered approximating the integral in (2.16) using Laplace approximation to *integrate out* the random effects. The Laplace's method has recently gained widespread use in various statistical inference problems ranging from approximating the likelihood (Pinheiro and Bates 1995; and Pinheiro and Chao 2006) to approximating Bayesian posterior moments (Tierney and Kadane 1986) and marginal posterior distributions (Rue and Martino 2009). Here we summarize Pinheiro and Chao's (2006) approach to approximating the integral in (2.16). Further details of the Laplace approximation in the context of GLMMs can be found in McCulloch and Searle (2001), Demidenko (2004), Hedeker and Gibbons (2006) and in Lee et al. (2006).

We consider a second-order Taylor expansion of $\kappa\big(\boldsymbol{\varphi},\, \mathbf{y}_{i,}\, \mathbf{u}_i\big)$ around $\widehat{\mathbf{u}}_i$, the maximizer of $\kappa\big(\boldsymbol{\varphi},\, \mathbf{y}_{i,}\, \mathbf{u}_i\big)$, i.e.

$$\kappa\big(\boldsymbol{\varphi},\, \mathbf{y}_{i,}\, \mathbf{u}_i\big) \approx \kappa\big(\boldsymbol{\varphi},\, \mathbf{y}_{i,}, \widehat{\mathbf{u}}_i\big) + \frac{\partial\kappa}{\partial \boldsymbol{u}_i^{\mathrm{T}}}(\,\mathbf{u}_i - \widehat{\mathbf{u}}_i) + \frac{1}{2}(\,\mathbf{u}_i - \widehat{\mathbf{u}}_i)^T \frac{\partial^2\kappa}{\partial\, \boldsymbol{u}_i \partial \boldsymbol{u}_i^{\mathrm{T}}}(\,\mathbf{u}_i - \widehat{\mathbf{u}}_i),$$

where $\widehat{\mathbf{u}}_i$ is the solution to

$$\begin{aligned} \frac{\partial\kappa}{\partial \mathbf{u}_i^T} &= \mathbf{Z}_i^T[\,\mathbf{y}_i - b'(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i)]/a(\phi)\; - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}_i \\ &= \mathbf{Z}_i^T[\,\mathbf{y}_i - \boldsymbol{\mu}_i]/a(\phi)\; - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\mathbf{u}_i = 0, \end{aligned} \tag{2.17}$$

and

$$\begin{aligned} \frac{\partial^2\kappa}{\partial\, \mathbf{u}_i \partial \mathbf{u}_i^T} &= -\mathbf{Z}_i^T b''(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i)/a(\phi) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}) \\ &= -\left(\mathbf{Z}_i^T \frac{\mathbf{v}_i}{a^2(\phi)}\mathbf{Z}_i - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\right), \end{aligned} \tag{2.18}$$

where $\mathbf{V_i} = \text{diag}\{Var(y_{ij}|\mathbf{u}_i)\}$.

It is obvious from (2.18) that $\frac{\partial^2 \kappa}{\partial\, \mathbf{u}_i \partial \mathbf{u}_i^T}$ is a negative-definite matrix, showing that $\kappa(.)$ is a strictly concave function in $\mathbf{u}_i$. Thus, $\widehat{\mathbf{u}}_i$ is a unique point of maximum of $\kappa(.)$ obtained by solving equation (2.17). The solution can be obtained by developing a simple recursion relation using the Newton-Raphson algorithm. A more computationally efficient approach is also possible by translating the maximization problem into a least-squares problem. Further details can be found in Pinheiro and Chao (2006).

Having $\widehat{\mathbf{u}}_i$ as the solution to (2.17), the approximate version of $\kappa(.)$ is then given as,

$$\kappa(\boldsymbol{\varphi},\, \mathbf{y}_{i,}\, \mathbf{u}_i) \approx \frac{\left(\mathbf{y}_i^T \boldsymbol{\eta}_i - \mathbf{1}^T b(\boldsymbol{\eta}_i)\right)}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}_i, \phi) + \frac{\widehat{\mathbf{u}}_i^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}) \widehat{\mathbf{u}}_i}{2} - \frac{1}{2}(\,\mathbf{u}_i - \widehat{\mathbf{u}}_i)^{\mathrm{T}} \left(\mathbf{Z}_i^T \frac{\mathbf{V_i}}{a^2(\phi)} \mathbf{Z}_i - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\right)(\,\mathbf{u}_i - \widehat{\mathbf{u}}_i).$$

Substituting it into (2.16) yields the following approximate log-likelihood,

$$\log L_i \approx \log \tilde{L}_i = \log(2\pi)^{-q/2} + \log|\boldsymbol{\Sigma}(\boldsymbol{\psi})|^{-1/2} + \frac{\left(\mathbf{y}_i^T \boldsymbol{\eta}_i - \mathbf{1}^T b(\boldsymbol{\eta}_i)\right)}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}_i, \phi) + \frac{\widehat{\mathbf{u}}_i^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}) \widehat{\mathbf{u}}_i}{2} + \log\left(\int e^{-\frac{1}{2}(\,\mathbf{u}_i - \widehat{\mathbf{u}}_i)^T \left[\mathbf{Z}_i^T \frac{\mathbf{V_i}}{a^2(\phi)} \mathbf{Z}_i - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\right](\,\mathbf{u}_i - \widehat{\mathbf{u}}_i)} d\mathbf{u}_i\right)$$

$$= \log(2\pi)^{-q/2} + \log|\boldsymbol{\Sigma}(\boldsymbol{\psi})|^{-1/2} + \frac{\left(\mathbf{y}_i^T \boldsymbol{\eta}_i - \mathbf{1}^T b(\boldsymbol{\eta}_i)\right)}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}_i, \phi) + \frac{\widehat{\mathbf{u}}_i^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}) \widehat{\mathbf{u}}_i}{2} + \log\left((2\pi)^{q/2} \left|\mathbf{Z}_i^T \frac{\mathbf{V_i}}{a^2(\phi)} \mathbf{Z}_i - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\right|^{-1/2}\right)$$

$$= -\frac{1}{2}\log\left|\boldsymbol{\Sigma}(\boldsymbol{\psi}) \mathbf{Q}_i(\phi) + \mathbf{I}\right| + \frac{\left(\mathbf{y}_i^T \boldsymbol{\eta}_i - \mathbf{1}^T b(\boldsymbol{\eta}_i)\right)}{a(\phi)} + \frac{\widehat{\mathbf{u}}_i^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}) \widehat{\mathbf{u}}_i}{2} + \mathbf{1}^T c(\mathbf{y}_i, \phi), \quad (2.19)$$

where

$$\mathbf{Q}_i(\phi) = \mathbf{Z}_i^T \frac{\mathbf{V_i}}{a^2(\phi)} \mathbf{Z}_i.$$

Thus, the Laplace approximation to the full data log-likelihood (see equation 2.2) is given as

$$l(\boldsymbol{\varphi}; \boldsymbol{y}_{(n)}) \approx \sum_{i=}^{n} \log \tilde{L}_i.$$

Maximization of the approximated log-likelihood in parameters $\boldsymbol{\varphi}$ involves a version of Fisher's scoring algorithm leading to fast implementation of the method (McCulloch and Searle 2001; Pinheiro and Chao 2006).

This first-order Laplace approximation as obtained in (2.19) have been known to produce biased estimates in certain distributional settings (Breslow and Lin 1995; Noh

and Lee 2007), especially when number of observations in each cluster is small (Engel 1998). A recent simulation study by Joe (2008) shows that the bias tends to be higher for binary and ordinal responses than count responses. Although computationally intensive, higher order Laplace approximations have also been considered in the literature (Breslow and Lin 1995; Raudenbush et al. 2000; Noh and Lee 2007) that generally produce more accurate approximates.

2.4.3 Adaptive Gaussian Quadrature

Gaussian quadrature (GQ) rules are numerical integration methods to approximate definite integrals of a given function by using a weighted average of the function at specified integration points in the domain (Stoer and Bulirsch 2002). In particular, Gauss-Hermite quadrature uses a set of *Q quadrature points* and weights, $\{z_q, w_q\}_{q=1}^{Q}$, to approximate an integral of the form

$$\int_{-\infty}^{\infty} f(\mathrm{z})\phi(z)dz \approx \sum_{q=1}^{Q} w_q\, f(z_q),$$

where $\phi(.)$ denotes standard normal density with corresponding distribution function $\Phi(.)$. In order to exemplify how GQ can be used to approximate the likelihood in (2.2), we assume that we have a single random effect in cluster *i*, distributed as $u_i \sim N(0, \tau^2)$. The likelihood in (2.2) then simplifies to

$$L(\boldsymbol{\varphi}; \mathbf{y}_{(n)}) = \prod_{i=1}^{n} \int f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|u_i) g(u_i; \tau^2)\, du_i, \tag{2.20}$$

where

$$\begin{aligned} L_i &= \int f(\mathbf{y}_i; \boldsymbol{\beta}, \phi|u_i)\ g(u_i; \tau^2) du_i \\ &= \int \left\{ \prod_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\beta}, \phi|u_i) \right\} g(u_i; \tau^2) du_i \\ &= \frac{1}{\tau} \int e^{\sum_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\beta}, \phi|u_i)}\ \Phi\left(\frac{u_i}{\tau}\right) du_i \\ &\approx \sum_{q=1}^{Q} w_q \left[e^{\sum_{j=1}^{n_i} f\left(y_{ij}; \boldsymbol{\beta}, \phi|u_i = \frac{z_q}{\tau}\right)} \right]. \end{aligned} \tag{2.21}$$

Substituting (2.21) into equation (2.20) gives the approximated log-likelihood as

$$l(\boldsymbol{\varphi}; \mathbf{y}_{(n)}) \approx \sum_{i=1}^{n} \log L_i.$$

In general, accuracy of the approximation increases with increasing number of quadrature points *Q*. Quadrature points and weights, $\{z_q, w_q\}_{q=1}^{Q}$, are both selected with respect to random effects distribution, $N(0, \tau^2)$, using a specific quadrature rule. These rules are available for Gaussian and other common kernels from the tables of Abramowitz and Stegun (1965). Although, these rules are complex for multiple integrals, the GLMM

structure allows approximating (2.2) by successive applications of single one-dimensional rules as shown in the simple case above. However, GQ becomes computationally intensive when several random effect terms are present. This is because it requires a large number of quadrature points per dimension to obtain accurate approximations.

Pinheiro and Bates (1995) and Pinheiro and Chao (2006) noticed that equation (2.21) can be viewed as a deterministic version of the Monte Carlo integration where, rather than generating $u_i's$ from $N(0,\tau^2)$, we have used sample points z_q with corresponding weights w_q that are fixed a priori. Pinheiro and Chao (2006) developed this idea to approximate GLMM likelihoods. Their idea is based on modifying GQ to obtain a determinist equivalent of IS which they call adaptive Gaussian quadrature (AGQ). The key aspect of AGQ is, therefore, to generate $\{z_q, w_q\}_{q=1}^{Q}$ from an *importance distribution*, rather than from the marginal distribution of the random effects. Keeping in view the Laplace approximation discussed in the previous section, their suggestion for the importance distribution is to use the Gaussian density

$$N\left(\hat{\mathbf{u}}_i, \left[\mathbf{Z}_i^T \frac{\mathbf{V}_i}{a^2(\phi)} \mathbf{Z}_i - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\right]^{-1}\right).$$

Because IS tends to be much more accurate than simple Monte Carlo integration (Geweke 1989), the AGQ also produces an improved approximation over GQ. It can be shown that AGQ reduces to the Laplace approximation (2.19) when used with a single quadrature point.

Pinheiro and Chao (2006) also developed a fast algorithm for implementation of AGQ that scales up efficiently to multilevel GLMMs. Their simulation results showed that AGQ performs favorably against PQL and the Laplace approximation. Simulation studies by Rabe-Hesketh et al. (2002, 2005) indicate that AGQ performs well as compared to GQ when both cluster size and intra-cluster correlations are high in multilevel GLMMs. However, a recent study by Cagnone and Monari (2012) shows that AGQ can be computationally very intensive when response is ordinal with several categories.

Although AGQ tends to be more accurate than the Laplace approximation, the latter is far less computationally intensive, especially when there exist a large number of random effects. There are recent attempts to exploit the computational efficiency of Laplace approximation. For instance, Rizopoulos et al. (2009) employ *fully exponential Laplace approximation* (FLA) of Tierney *et al.* (1989) for the joint modeling of survival and longitudinal data, yielding an improved version of the existing Laplace method. Bian-

concini and Cagnone (2012) further extend this approach to analyze GLMMs. Their simulations study shows that the FLA approach compares favorably against AGQ and effectively handles the high dimensional latent structures without substantially increasing the computational burden.

2.5 Bayesian Inference

Bayesian philosophy of statistical inference is fundamentally different from the frequentist (or *classical*) approach. In Bayesian statistics, probability statements are treated as representing *degree of belief* in the occurrence of an event, rather than the usual long-run relative frequency of that event (Barnett 1999). This 'degree of belief' interpretation is further combined with the famous Bayes' theorem to update *prior* belief about unknown model quantities (parameters and latent effects) based on observed data. We describe the general framework as follows, starting with the Bayes' Theorem.

Theorem 1.1 (Bayes' Theorem): Let $\{A_j\}_{j=1}^{n}$ be a collection of *n* mutually exclusive and exhaustive events and *A* be another event defined on the sample space. Then the conditional (or *inverse*) probability of the observing an event A_i given that *A* has already occurred is given as

$$P(A_i|A) = \frac{P(A|A_i)P(A_i)}{\sum_{j=1}^{n} P(A|A_j)P(A_j)}\ , i = 1,2, \dots n.$$

We now assume that sample data are a realization of a random vector $\boldsymbol{Y}$ whose probability distribution, $f(\mathbf{y}|\boldsymbol{\theta})$, has a known form, except for fixed but unknown parameter vector $\boldsymbol{\theta}$. The degree of uncertainty about $\boldsymbol{\theta}$ before observing data $\mathbf{y}$ is quantified by a prior probability distribution, $\pi(\boldsymbol{\theta})$. That is, $\pi(\boldsymbol{\theta})$ provides an *objective* means of quantifying subjective, or *a priori*, information available about $\boldsymbol{\theta}$, such as expert opinion or knowledge from previous studies. Our information about $\boldsymbol{\theta}$ is further increased after observing the sample $\mathbf{y}$, resulting in the updated distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, called the *posterior probability distribution*. The posterior distribution can be obtained by applying the Bayes' theorem as follows,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Thus, $\pi(\boldsymbol{\theta}|\mathbf{y})$ embodies all the available information about $\boldsymbol{\theta}$ and, therefore, forms the basis of all inferences, including Bayesian point estimation, credible intervals and predictive distributions. For a comprehensive treatment of Bayesian inferential methods includ-

ing the use of various types of prior distributions, we refer the reader to Berger (1985) and Bernardo and Smith (2001).

In case of the general hierarchical model defined in (1.1), the posterior for the model parameters $\boldsymbol{\theta}$ is defined as

$$\pi\left(\boldsymbol{\theta}\middle|\mathbf{y}_{(\mathrm{n})}\right) = \frac{\{\int f\left(\mathbf{y}_{(\mathrm{n})};\boldsymbol{\theta}_1|\mathbf{u}\right)g(\mathbf{u};\boldsymbol{\theta}_2)d\mathbf{u}\}\pi(\boldsymbol{\theta})}{c(\mathbf{y}_{(\mathrm{n})})},$$
$$= \frac{L(\boldsymbol{\theta};\mathbf{y}_{(\mathrm{n})})\pi(\boldsymbol{\theta})}{c(\mathbf{y}_{(\mathrm{n})})}, \tag{2.22}$$

where

$$c\left(\mathbf{y}_{(\mathrm{n})}\right) = \int f\left(\mathbf{y}_{(\mathrm{n})};\boldsymbol{\theta}_1|\mathbf{u}\right)g(\mathbf{u};\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta})d\mathbf{u}d\boldsymbol{\theta}$$

is the normalizing constant. Except for some simple model formulations, the posterior for $\boldsymbol{\theta}$ does not exist in a closed form. Its numerical computation is also intractable because of the high-dimensional integration involved in computing the normalizing constant. Fortunately, one can generate random numbers from $\pi\left(\boldsymbol{\theta}\middle|\mathbf{y}_{(\mathrm{n})}\right)$ by appealing to the standard MCMC algorithms (Gilks et al.1996; Spiegelhalter et al. 2004), such as Gibbs sampling (Gelfand and Smith 1990) and Metropolis-Hastings (Metropolis et al. 1953; Hastings 1970) algorithms, without ever computing the normalizing constant in (22). These algorithms are designed to generate random samples from intractable target probability distributions.

In addition to obtaining the posterior for $\boldsymbol{\theta}$, we can also obtain marginal posterior distribution for the latent effects $\mathbf{u}$, i.e.

$$\pi\left(\mathbf{u}\middle|\mathbf{y}_{(\mathrm{n})}\right) = \frac{\int f\left(\mathbf{y}_{(\mathrm{n})};\boldsymbol{\theta}_1|\mathbf{u}\right)g(\mathbf{u};\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{c(\mathbf{y}_{(\mathrm{n})})}. \tag{2.23}$$

This posterior distribution can then be used to obtain Bayesian prediction intervals for the latent effects. Notice that, MCMC algorithms are used to generate random numbers from (2.22) and (2.23) simultaneously by sampling from the following joint posterior distribution

$$\pi\left(\boldsymbol{\theta},\mathbf{u}\middle|\mathbf{y}_{(\mathrm{n})}\right) = \frac{f\left(\mathbf{y}_{(\mathrm{n})};\boldsymbol{\theta}_1|\mathbf{u}\right)g(\mathbf{u};\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta})}{c(\mathbf{y}_{(\mathrm{n})})}. \tag{2.24}$$

2.5.1 Bayesian Analysis of GLMMs

Let us recall the GLMM defined in Section 2.2. Assuming $\pi(\boldsymbol{\varphi})$, to be prior distribution of model parameters, the marginal posterior distributions for $\boldsymbol{\varphi}$ and the random effects $\mathbf{u}_i$ are respectively given as

$$\pi\left(\boldsymbol{\varphi}\middle|\mathbf{y}_{(\mathrm{n})}\right)=\frac{\{\prod_{i=1}^{n}\int f(\mathbf{y}_i;\boldsymbol{\beta},\phi|\mathbf{u}_i)g(\mathbf{u}_i;\boldsymbol{\Sigma}(\boldsymbol{\psi}))d\mathbf{u}_i\}\pi(\boldsymbol{\varphi})}{\int\{\prod_{i=1}^{n}\int f(\mathbf{y}_i;\boldsymbol{\beta},\phi|\mathbf{u}_i)g(\mathbf{u}_i;\boldsymbol{\Sigma}(\boldsymbol{\psi}))\pi(\boldsymbol{\varphi})d\boldsymbol{\varphi}d\mathbf{u}_{\mathrm{i}}\}}, \tag{2.25}$$

$$\pi\left(\mathbf{u}_{\mathrm{i}}\middle|\mathbf{y}_{(\mathrm{n})}\right)=\frac{\prod_{i=1}^{n}\int f(\mathbf{y}_i;\boldsymbol{\beta},\phi|\mathbf{u}_i)g(\mathbf{u}_i;\boldsymbol{\Sigma}(\boldsymbol{\psi}))\pi(\boldsymbol{\varphi})d\boldsymbol{\varphi}}{\int\{\prod_{i=1}^{n}\int f(\mathbf{y}_i;\boldsymbol{\beta},\phi|\mathbf{u}_i)g(\mathbf{u}_i;\boldsymbol{\Sigma}(\boldsymbol{\psi}))\pi(\boldsymbol{\varphi})d\boldsymbol{\varphi}d\mathbf{u}_i\}}. \tag{2.26}$$

The posterior distributions in (2.25) and (2.26) are generally numerically intractable as the dimension of $\mathbf{u}_i$, q, is usually large. Therefore, these posteriors are estimated by drawing MCMC samples from the following joint posterior distribution

$$\pi\left(\boldsymbol{\varphi},\mathbf{u}_{\mathrm{i}}\middle|\mathbf{y}_{(\mathrm{n})}\right)=\frac{f(\mathbf{y}_i;\boldsymbol{\beta},\phi|\mathbf{u}_i)g(\mathbf{u}_i;\boldsymbol{\Sigma}(\boldsymbol{\psi}))\pi(\boldsymbol{\varphi})}{\int\{\prod_{i=1}^{n}\int f(\mathbf{y}_i;\boldsymbol{\beta},\phi|\mathbf{u}_i)g(\mathbf{u}_i;\boldsymbol{\Sigma}(\boldsymbol{\psi}))\pi(\boldsymbol{\varphi})d\boldsymbol{\varphi}d\mathbf{u}_{\mathrm{i}}\}}.$$

Zeger and Karim (1991, 1992) employed the Gibbs sampler to compute these posteriors. They suggested using *noninformative priors* with the understanding that the mean of the posterior distribution is a reasonable approximation to the maximum likelihood estimator. The resulting estimators have become increasingly popular because of the advent of the MCMC methodology and availability of computer software such as the WinBUGS (Spiegelhalter et al. 2004). Rather than numerical maximization of a noisy function, these Bayesian methods have the advantage of requiring only the computation of the means and variances of the posterior distribution. However, the use of noninformative prior has its own difficulties, both in the context of the convergence of the MCMC algorithm and the true meaning of noninformativeness of a prior. Difficulties with the use of noninformative priors are thoroughly reviewed by Press (2003, Chapter 5), Barnett (1999, Chapter 6), Cox (2006) and Lele and Dennis (2009).

Some researchers (e.g. Datta 1996) have tried to exploit the simplicity of the Bayesian computation to obtain valid frequentist answers to difficult problems by trying to construct priors, the probability-matching priors, such that the credible intervals obtained from the posterior distributions are, in fact, the same as the confidence intervals. However, calculation of the appropriate probability-matching priors for a given model is a difficult task. Recently, extending some of the techniques from the Physics literature (Bennett 1976), various researchers have suggested using path sampling and bridge sampling methods (Gelman and Meng 1998) for obtaining estimates of the likelihood ratios using MCMC method. These, in turn, can be used to obtain the maximum likelihood estimators for most hierarchical models including GLMM. A related method based on nonparametric estimation of the likelihood surface was developed by deValpine (2008). However, these methods also face the difficult task of numerically maximizing a noisy function (Spall 2003) that can be very tricky and difficult.

2.6 Sequential Monte Carlo

Sequential Monte Carlo (SMC) (Doucet et al. 2001) techniques are a set of simulation-based methods designed to compute prediction distributions of the latent effects, i.e.

$$g(\mathbf{u}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{u})g(\mathbf{u})}{\int f(\mathbf{y}|\mathbf{u})g(\mathbf{u})d\mathbf{u}}. \tag{2.27}$$

As we have seen previously, models $f(.)$ and $g(.)$ may depend on unknown parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. However, we suppress this dependence in the ongoing section to reduce notational burden. Although applicable in much more generality, the SMC methods can best be understood in the context of state-space time series models (deValpine 2002). In a standard state-space formulation, the unobserved states $\{\mathbf{u}_t; t \in \mathbb{N}\}$ are modeled as a *Markov process* with initial distribution $g(\mathbf{u}_0)$ and the transition distribution $g(\mathbf{u}_t|\mathbf{u}_{t-1})$. The observations $\{\mathbf{y}_t; t \in \mathbb{N}\}$ are conditionally independent given the latent process $\{\mathbf{u}_t; t \in \mathbb{N}\}$ with marginal distributions $f(\mathbf{y}_t|\mathbf{u}_t)$. Let us denote $\mathbf{u}_{0:t} = \{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_t\}$ and $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_t\}$ as the latent states and observations up to time *t*. Then, we wish to obtain *recursively in time* the prediction distribution (2.27), i.e.

$$g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t}) = \frac{f(\mathbf{y}_{1:t}|\mathbf{u}_{1:t})g(\mathbf{u}_{1:t})}{\int f(\mathbf{y}_{1:t}|\mathbf{u}_{1:t})g(\mathbf{u}_{1:t})d\mathbf{u}_{1:t}}. \tag{2.28}$$

In particular, our goal is to estimate the *filtering* distributions, $g(\mathbf{u}_t|\mathbf{y}_{1:t})$, and the following expectation recursively in time:

$$\mathcal{E}(\hbar) = E_{g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t})}[\hbar(\mathbf{u}_{1:t})] = \int \hbar(\mathbf{u}_{1:t})g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t})d\mathbf{u}_{1:t}. \tag{2.29}$$

for some function of interest $\hbar(.)$; for instance, conditional mean of (2.28).

The Markov assumption allows us to obtain a straightforward recursive formula for the joint predictive distribution (2.28), i.e.

$$g(\mathbf{u}_{1:t+1}|\mathbf{y}_{1:t+1}) = g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t})\frac{f(\mathbf{y}_{t+1}|\mathbf{u}_{t+1})g(\mathbf{u}_{t+1}|\mathbf{u}_t)}{f(\mathbf{y}_{t+1}|\mathbf{y}_{1:t})}. \tag{2.30}$$

Similarly, we also obtain a recursive relation for the filtering distribution $g(\mathbf{u}_t|\mathbf{y}_{1:t})$, i.e.

$$g(\mathbf{u}_t|\mathbf{y}_{1:t-1}) = \int g(\mathbf{u}_t|\mathbf{u}_{t-1})g(\mathbf{u}_{t-1}|\mathbf{y}_{1:t-1})d\,\mathbf{u}_{t-1}. \tag{2.31}$$

which can be updated as

$$g(\mathbf{u}_t|\mathbf{y}_{1:t}) = \frac{f(\mathbf{y}_t|\mathbf{u}_t)g(\mathbf{u}_t|\mathbf{y}_{1:t-1})}{\int f(\mathbf{y}_t|\mathbf{u}_t)g(\mathbf{u}_t|\mathbf{y}_{1:t-1})d\mathbf{u}_t}. \tag{2.32}$$

It is clear that the above predictive distributions cannot be computed in a straightforward fashion owing to the difficulty in computing the normalizing constants that involve intractable integrals in latent states $\mathbf{u}_t$. There, however, exist simpler cases where exact analytical solutions exist. A notable exception is that of the linear Gaussian state-space model for which the well known *Kalman filter* (Harvey 1993) provides efficient

estimates of the predictive distributions. However, the problem remains highly intractable for nonlinear non-Gaussian state-space models.

Importance sampling (see Section 2.3.3) plays a key role in SMC methodology. In the following we explain how IS can be used in a recursive manner to compute estimates of the predictive distributions defined above.

2.6.1 Sequential Importance Sampling

As we have seen earlier, the main difficulty in computing $g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t})$ is due to the normalizing constant appearing in (2.28). This normalizing constant is in fact the likelihood function for the observed data $\mathbf{y}_{1:t}$, i.e.

$$L(\boldsymbol{\theta}; \mathbf{y}_{1:t}) = \int f(\mathbf{y}_{1:t}|\mathbf{u}_{1:t})\, g(\mathbf{u}_{1:t}) d\mathbf{u}_{1:t}.$$

Recalling from Section 2.3.3, a Monte Carlo estimate of $L(\boldsymbol{\theta}; \mathbf{y}_{1:t})$ using IS is given as follows

$$\hat{L}(\boldsymbol{\theta}; \mathbf{y}_{1:t}) = \frac{1}{N}\sum_{j=1}^{N} f\left(\mathbf{y}_{1:t}|\mathbf{u}_{1:\mathrm{t}}^{(j)}\right) \omega_{1:\mathrm{t}}^{(j)}. \tag{2.33}$$

where $\omega_{1:\mathrm{t}}^{(j)} = g(\mathbf{u}_{1:\mathrm{t}}^{(j)})/p\left(\mathbf{u}_{1:\mathrm{t}}^{(j)}\right)$ are the importance weights and $p(.)$ is the IS distribution whose support includes the support of $g(.)$. For a given value of $\boldsymbol{\theta}$, the above expression provides a consistent estimate of $L(\boldsymbol{\theta}; \mathbf{y}_{1:t})$ that is independent of the dimension of the integrand (Geweke 1989). We can obtain similar estimates for the expectation defined in (2.29). However, this simple IS scheme is not suitable for the recursive estimation of the predictive distributions as described in (2.30-31). This is because we need to obtain all the data $\mathbf{y}_{1:t}$ before estimating $g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t})$. Thus, we need to recompute the importance weights over $\mathbf{u}_{1:t}$ whenever we observe a new data $\mathbf{y}_{t+1}$. Clearly, this scheme becomes computationally prohibitive with increasing time steps.

An alternative is to modify the above procedure as follows. Let us set $t = 1$ and assume $p(\mathbf{u}_1|\mathbf{y}_1)$ be the IS distribution corresponding to $g(\mathbf{u}_1|\mathbf{y}_1)$ yielding an IS estimate $\hat{g}(\mathbf{u}_1|\mathbf{y}_1)$. That is,

$$\hat{g}(\mathbf{u}_1|\mathbf{y}_1) = \frac{1}{N}\sum_{j=1}^{N} g\left(\mathbf{y}_1|\mathbf{u}_1^{(j)}\right) \omega_1^{(j)},$$

where $\{\mathbf{u}_1^{(j)}\}$ are *N* samples (or *particles*) sampled from $p(\mathbf{u}_1|\mathbf{y}_1)$, and $\{\omega_1^{(j)}\}$ are the assigned importance weights to account for the discrepancy between the two distributions. Now, for $t = 2$, we notice from (2.30) that

$$\begin{aligned} g(\mathbf{u}_{1:2}|\mathbf{y}_{1:2}) &= g(\mathbf{u}_1|\mathbf{y}_1)\frac{f(\mathbf{y}_2|\mathbf{u}_2)g(\mathbf{u}_2|\mathbf{u}_1)}{f(\mathbf{y}_2|\mathbf{y}_1)} \\ &\propto g(\mathbf{u}_1|\mathbf{y}_1)f(\mathbf{y}_2|\mathbf{u}_2)g(\mathbf{u}_2|\mathbf{u}_1). \end{aligned}$$

This suggests introducing an IS distribution $p(\mathbf{u}_2|\mathbf{y}_2, \mathbf{u}_1)$ in order to generate samples from the joint distribution $g(\mathbf{u}_1|\mathbf{y}_1)p(\mathbf{u}_2|\mathbf{y}_2, \mathbf{u}_1)$ with corresponding weights $\{\omega_{1:2}^{(j)}\}$. The trick here is to obtain the importance weights $\{\omega_2^{(j)}\}$ by *updating* the weights obtained at $t = 1$. This procedure is further augmented by a resampling step yielding samples approximately distributed as $g(\mathbf{u}_{1:2}|\mathbf{y}_{1:2})$. The procedure is repeated recursively until the last time step; say $t = T$. Further details and refinements of this simple SMC can be found in Doucet et al. (2001).

2.6.2 Estimation and Prediction

Thus far we have assumed that the parameter vector $\boldsymbol{\theta}$ is known. However, in practice $\boldsymbol{\theta}$ must be estimated from the observed data $\mathbf{y}_{1:t}$. Let us first consider the Bayesian estimation in the state-space modeling context. As described in Andrieu et al. (2010), the most commonly used choice is to sample from $\pi(\boldsymbol{\theta}, \mathbf{u}_{1:t}|\mathbf{y}_{1:t})$ using MCMC methods (see equation 2.24) by alternately updating the latent states $\mathbf{u}_{1:t}$ conditional on $\boldsymbol{\theta}$ and vice versa. Sampling from $\pi(\boldsymbol{\theta}|\mathbf{y}_{1:t})$ is usually feasible in general state-space models (Andrieu et al. 2010). On the other hand, apart from simpler cases such as linear Gaussian models, sampling from $g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t})$ is intractable as one needs to design efficient proposal densities. The problem is further aggravated when $g(\mathbf{u}_t|\mathbf{u}_{t-1})$ cannot be expressed analytically but its simulation is feasible. However, Andrieu et al. (2010) have recently introduced an SMC based algorithm to construct efficient, potentially multidimensional, proposal distributions to improve upon the standard MCMC methods for simulating from $g(\mathbf{u}_{1:t}|\mathbf{y}_{1:t})$. The resulting particle MCMC (PMCMC) algorithm also overcomes the limitations suffered by the stand-alone SMC algorithms. Another attractive feature of PMCMC is that it is generally applicable in a wide class of hierarchical models including the nonlinear mixed effects models.

Maximum likelihood estimation techniques involving SMC are now well developed (Kantas et al. 2009). It is obvious from (2.33) that for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, it is possible to compute an unbiased estimate of $L(\boldsymbol{\theta}; \mathbf{y}_{1:t})$, numerically yielding a corresponding plug-in estimate for the log-likelihood, i.e. $\hat{l}(\boldsymbol{\theta}; \mathbf{y}_{1:t}) = \log \hat{L}(\boldsymbol{\theta}; \mathbf{y}_{1:t})$. However, this latter estimate is biased but the standard techniques exist to correct for the bias (Pitt 2002). This bias corrected estimate then serves as a basic ingredient in more sophisticated optimization procedures such as gradient methods and the EM algorithm. For further details, we refer the reader to Kantas et al. 2009. A recent SMC based approach for likelihood based

inference in hierarchical models is developed by Johansen et al. (2008). They employ simulated annealing ideas (Brooks and Morgan 1995) to construct a sequence of artificial distributions whose limiting support is concentrated around the MLE. They then use SMC to sample from these artificial distributions.

2.7 Summary

Computation of the likelihood function arising in the context of GLMMs involves integration over the distribution of the random effects, which is generally high dimensional. Thus, exact maximization of the likelihood function is intractable, leading to various approximate methods for evaluating the high dimensional integrals. These can be broadly divided into two types: (i) Monte Carlo approximations such as MCEM and simulated maximum likelihood, and (ii) approximate numerical integration techniques such as Laplace approximation and AGQ. Although, PQL also falls in the latter category, it overly simplifies the maximization problem resulting in biased estimates of both fixed effects and variance components. On the other hand, AGQ have been shown to provide accurate results in various GLMM settings but it becomes computationally costly when response is ordinal or when there exist a large number of latent variables.

Although simple Monte Carlo approximations, such as simulated maximum likelihood, suffer in terms of convergence and accuracy when the problem involves high-dimensional integration, they can be adequately improved by adopting the SMC methodology. The PMCMC algorithm that integrates SMC with the standard MCMC routines, provides a promising framework for computing highly intractable predictive distributions.

Bayesian paradigm offers a different approach to likelihood based inference by assuming prior distributions to account for parameter uncertainty before any data have been collected. Noninformative priors are commonly used to avoid subjectivity arising from the choice of prior distributions. However, it is still debatable whether the resulting inference is invariant to the choice of noninformative priors. This problem is especially important when fitting complex hierarchical models with scarce data and no guarantees on model identifiability (Lele 2010).

Chapter 3

Analysis of GLMMs using Data Cloning[1]

In this chapter we develop data cloning (DC) algorithm for computing maximum likelihood estimates (MLEs) and their standard errors for general hierarchical models. Earlier, Lele et al. (2007) reviewed the difficulties associated with Bayesian and likelihood based inference in general hierarchical models and proposed data cloning as an alternative approach. We refer the reader to Doucet et al. (2002), Kuk (2003), and Jacquier et al. (2007) for methods similar to data cloning.

The organization of this chapter is as follows. We begin with a theoretical development of the DC algorithm in Section 3.1. In Section 3.2 we present an algorithm involving MCMC implementation of the DC approach for computing MLEs and their standard errors for general hierarchical models. Section 3.3 presents our algorithm for obtaining DC based prediction intervals for the latent effects. We exemplify the DC approach using various important subclasses of GLMMs in Section 3.4. In Section 3.5 we develop a DC based estimability diagnostic algorithm. The chapter concludes with a brief summary in Section 3.6.

[1] A version of this chapter has been published. Lele S R, Nadeem K, and Schmuland B. Journal of the American Statistical Association 2010, 105.492: 1617-1625.

3.1 Data Cloning Algorithm

The key idea behind data cloning is to formulate the Bayes' rule as an iterated map on the space of probability distributions. Let us recall from Chapter 1 the Bayesian formulation of the general hierarchical model:

Hierarchy 1: $\mathbf{y}_{(n)}|\mathbf{u} \sim f(\mathbf{y}_{(n)}|\boldsymbol{\theta}_1, \mathbf{u})$ (3.1 a)

Hierarchy 2: $\mathbf{u} \sim g(\mathbf{u}|\boldsymbol{\theta}_2)$, (3.1 b)

Prior Distribution: $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, (3.1 c)

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ and we assume that $\pi(\boldsymbol{\theta})$ takes positive values on the p-dimensional parameter space $\boldsymbol{\Theta}$. The marginal posterior distribution of $\boldsymbol{\theta}$ is then given as (see equation 2.22),

$$\pi_1(\boldsymbol{\theta}|\mathbf{y}_{(n)}) = \frac{L(\boldsymbol{\theta};\mathbf{y}_{(n)})\pi(\boldsymbol{\theta})}{c(\mathbf{y}_{(n)})},$$

where $L(\boldsymbol{\theta};\mathbf{y}_{(n)})$ is the integrated likelihood function and $c(\mathbf{y}_{(n)}) = \int L(\boldsymbol{\theta};\mathbf{y}_{(n)})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the normalizing constant. Now if we substitute this posterior distribution as prior back again, we obtain

$$\pi_2(\boldsymbol{\theta}|\mathbf{y}_{(n)}) = \frac{[L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^2\pi(\boldsymbol{\theta})}{c(\mathbf{y}_{(n)})\int\frac{[L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^2\pi(\boldsymbol{\theta})}{c(\mathbf{y}_{(n)})}d\boldsymbol{\theta}},$$

which reduces to

$$\pi_2(\boldsymbol{\theta}|\mathbf{y}_{(n)}) = \frac{[L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^2\pi(\boldsymbol{\theta})}{c(2;\mathbf{y}_{(n)})}.$$

Then, by induction, it follows that the posterior distribution corresponding to the prior $\pi_{K-1}(\boldsymbol{\theta}|\mathbf{y}_{(n)})$ is given as

$$\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)}) = \frac{[L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^K\pi(\boldsymbol{\theta})}{c(K;\mathbf{y}_{(n)})}. \quad (3.2)$$

where $c(K;\mathbf{y}_{(n)}) = \int[L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^K\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the normalizing constant. This posterior distribution can therefore be looked upon as an iterated map $\pi_1 = \Im(\pi, L)$, $\pi_2 = \Im(\pi_1, L)$, …, $\pi_K = \Im(\pi_{K-1}, L)$. Lele et al. (2007) showed that this iterated map has a fixed point: a probability distribution degenerate at the MLE which is independent of the initial prior distribution π. Here we present a sketch of their result whereas a more general result is obtained in the next section.

Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, i.e. $L(\hat{\boldsymbol{\theta}};\mathbf{y}_{(n)}) > L(\boldsymbol{\theta};\mathbf{y}_{(n)})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Since the initial prior distribution $\pi(\boldsymbol{\theta})$ is positive everywhere on the parameter space, it follows that as $K \to \infty$

$$\frac{\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})})}{\pi_K(\widehat{\boldsymbol{\theta}}|\mathbf{y}_{(\mathrm{n})})} = \frac{\pi(\boldsymbol{\theta})[L(\boldsymbol{\theta};\mathbf{y}_{(\mathrm{n})})]^K}{\pi(\widehat{\boldsymbol{\theta}})[L(\widehat{\boldsymbol{\theta}};\mathbf{y}_{(\mathrm{n})})]^K} \to 0 \text{ if } \boldsymbol{\theta} \neq \widehat{\boldsymbol{\theta}}$$

and

$$\frac{\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})})}{\pi_K(\widehat{\boldsymbol{\theta}}|\mathbf{y}_{(\mathrm{n})})} = \frac{\pi(\boldsymbol{\theta})[L(\boldsymbol{\theta};\mathbf{y}_{(\mathrm{n})})]^K}{\pi(\widehat{\boldsymbol{\theta}})[L(\widehat{\boldsymbol{\theta}};\mathbf{y}_{(\mathrm{n})})]^K} \to 1 \text{ if } \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}.$$

Thus, the fixed point for the aforementioned iterated map is a probability distribution that is degenerate at the MLE $\widehat{\boldsymbol{\theta}}$. Furthermore, the degenerate distribution is independent of the initial distribution $\pi(\boldsymbol{\theta})$.

In this thesis we further establish the following result whose proof is given in the next subsection. We denote the Fisher information matrix of $\widehat{\boldsymbol{\theta}}$ by $I(\widehat{\boldsymbol{\theta}})$.

Theorem 3.1: Let $\boldsymbol{\Theta}_K$ be a random variable on $\Re^p$, the p-dimensional Euclidean space, with density function $\pi_K(.)$ and define the standardized variable $\boldsymbol{\Psi}_K = \sqrt{K}I(\widehat{\boldsymbol{\theta}})\big(\Theta_K - \widehat{\boldsymbol{\theta}}\big)$, then under suitable regularity conditions $\boldsymbol{\Psi}_K \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_p)$, as $K \to \infty$.

Thus, for large K and regardless of the choice of the initial distribution π, the posterior distribution $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})}) = \big[L(\boldsymbol{\theta};\mathbf{y}_{(\mathrm{n})})\big]^K \pi(\boldsymbol{\theta})/c(K;\ \mathbf{y}_{(\mathrm{n})})$ is approximately Normal with mean $\widehat{\boldsymbol{\theta}}$ and variance-covariance matrix equal to $\frac{1}{K}I^{-1}(\widehat{\boldsymbol{\theta}})$. This suggests that we can compute the MLE and the associated standard errors by computing the mean and variance of $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})})$ for large K.

3.1.1 Proof of Convergence

In this section we present a proof of Theorem 3.1. The proof is similar to that of Walker (1969) where, under certain regularity conditions, he showed stochastic convergence of $\pi_1(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})})$ to a degenerate distribution (degenerate at the MLE) as the sample size $n \to \infty$. However, the proof given here involves deterministic convergences of a sequence of functions; not the probabilistic convergences used in Walker (1969).

Let us fix some notation first. Let $f(\mathbf{y}_{(\mathrm{n})};\boldsymbol{\theta})$ denote the joint probability density function of the data vector $\mathbf{y}_{(\mathrm{n})}$. We assume that this is a bounded function as a function of $\boldsymbol{\theta}$. Also define

$$\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})}) = f^K(\mathbf{y}_{(\mathrm{n})}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/c(K),$$

where $c(K) = \int f^K(\mathbf{y}_{(\mathrm{n})}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. We are suppressing the dependence of $c(K)$, on $\mathbf{y}_{(\mathrm{n})}$ for notational simplicity.

Assumption 3.1: The function $f(.)$, as a function of $\boldsymbol{\theta}$, has a local maximum at $\boldsymbol{\theta}_\infty$ and $f(\boldsymbol{\theta}_\infty) > 0$ and $\pi(\boldsymbol{\theta}_\infty) > 0$. The maximum likelihood estimator is, by definition, denoted by $\boldsymbol{\theta}_\infty$.

Assumption 3.2: The function $\pi(.)$ is continuous at $\boldsymbol{\theta}_\infty$. The function $f(.)$ has continuous second derivatives in a neighborhood of $\boldsymbol{\theta}_\infty$ and the Hessian matrix $\boldsymbol{D}^2 f(\boldsymbol{\theta}_\infty)$ is strictly negative definite.

Assumption 3.3: For any $\delta > 0$, we have $\gamma(\delta) \coloneqq \sup\{ f(\boldsymbol{\theta}) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| > \delta\} < f(\boldsymbol{\theta}_\infty)$, where $||\,.\,||$ denotes the Euclidean norm of a vector.

Definition 3.1 (Neighborhood): Let $\boldsymbol{\Sigma} = \{-\boldsymbol{D}^2 f(\boldsymbol{\theta}_\infty)\}^{-1/2}$ and for some $\delta > 0$ define $N(\delta) \coloneqq \sup\{ \boldsymbol{\theta} : \|\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty)\| < \delta\}$. Because $\boldsymbol{\Sigma}$ is positive definite, this defines a system of neighborhoods of $\boldsymbol{\theta}_\infty$.

Definition 3.2: Let $\boldsymbol{\Theta}_K$ be a random variable on $\Re^p$ with density function $\pi_K(.)$ and define the standardized variable $\boldsymbol{\Psi}_K = \sqrt{K}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Theta}_K - \widehat{\boldsymbol{\theta}}\right)$ that has density function $g_K(\boldsymbol{\theta}) = \frac{|\boldsymbol{\Sigma}|}{K^{\frac{p}{2}}} \pi_K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)$.

Lemma 3.1 (Fatou's lemma): Let $f_1, f_1, \ldots$ be a sequence of non-negative measurable functions defined on a measure space $(E, \mathcal{B}, \mu)$. If $f: E \longrightarrow [0, \infty]$ such that the following *almost everywhere* pointwise limit exists

$$f(e) = \liminf_{k\to\infty} f_k(e)\,, \qquad e \in E,$$

then f is measurable and

$$\int_E f d\mu \leq \liminf_{k\to\infty} \int_E f_k(e) d\mu.$$

Lemma 3.2 (Scheffe's lemma): Let $X, X_1, X_2, \ldots$ be a sequence of continuous random variables in a probability space $(\Omega, \mathcal{F}, P)$, whose probability density functions are $f, f_1, f_1, \ldots$, respectively. If $\lim_{k\to\infty} f_k(x) = f(x)$ exists *almost everywhere* for all $x \in \Re$, then X_k converges to X in distribution, i.e. $X_k \xrightarrow{D} X$.

We now assume, without loss of generality, that $f(\boldsymbol{\theta}_\infty) = 1$. This is simply a standardized likelihood function and computation of the posterior distribution $\pi_K\left(\boldsymbol{\theta}\middle|\mathbf{y}_{(\mathrm{n})}\right)$ is invariant to such standardizations. Thus, $\boldsymbol{\Sigma} = \{-\boldsymbol{D}^2 f(\boldsymbol{\theta}_\infty)\}^{-1/2}$ corresponds to the square root of the inverse of the Fisher information matrix since

$\boldsymbol{D}^2 f(\boldsymbol{\theta}_\infty) = \boldsymbol{D}^2 \log f(\boldsymbol{\theta}_\infty)$.

Lemma 3.3: Under Assumptions 3.1 and 3.2, $f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)$ converges to $e^{-\|\boldsymbol{\theta}\|^2/2}$ uniformly on bounded sets of $\boldsymbol{\theta}$ as $K \to \infty$.

Proof: Fix δ_0 so small that $\boldsymbol{D}^2 f(\boldsymbol{\theta})$ is continuous on the neighborhood $N(\delta_0)$. For every $\boldsymbol{\theta}$ in this neighborhood, Taylor's theorem dictates that there is some $\boldsymbol{\theta}^+$ on the line segment joining $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\infty$ so that

$$f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_\infty) + \boldsymbol{D}f(\boldsymbol{\theta}_\infty)(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty)^T\{\boldsymbol{D}^2 f(\boldsymbol{\theta}^+)\}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty)$$
$$= 1 - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty)^{\mathrm{T}}\{-\boldsymbol{D}^2 f(\boldsymbol{\theta}^+)\}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty). \quad (3.3)$$

For any $\boldsymbol{\theta} \in \Re^p$, when K is large, the vector$\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma\theta}$ is in $N(\delta_0)$ and we have

$$f\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma\theta}\right) = 1 - \frac{\boldsymbol{\theta}^T\boldsymbol{\Sigma}^T\{-\boldsymbol{D}^2 f(\boldsymbol{\theta}_K)\}\boldsymbol{\Sigma\theta}}{2K},$$

for some $\boldsymbol{\theta}_K$ on the line segment joining $\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma\theta}$ and $\boldsymbol{\theta}_\infty$.

For $\varepsilon > 0$, choose $\delta(\varepsilon) < \delta_0$ so small that for $\boldsymbol{\theta} \in N\big(\delta(\varepsilon)\big)$, we have $\boldsymbol{D}^2 f(\boldsymbol{\theta})$ negative definite and $\|\boldsymbol{\Sigma}^T\{-\boldsymbol{D}^2 f(\boldsymbol{\theta})\}\boldsymbol{\Sigma} - \mathbf{I}\| \le \varepsilon$. Now, for some $0 \le x$ and $y \le K$, the following inequalities hold

$$\left|\left(1 - \frac{x}{K}\right)^K - \left(1 - \frac{y}{K}\right)^K\right| \le |x - y|, \quad (3.4\text{ a})$$

and

$$\left|\left(1 - \frac{y}{K}\right)^K - \exp(-y)\right| \le \frac{y^2}{K}, \quad (3.4\text{ b})$$

Let us fix $M > 1$, $0 < \varepsilon < 1$ and let $K \ge \max\left\{\left(\frac{M}{\delta(\varepsilon)}\right)^2, M^2\right\}$. Then for $\|\boldsymbol{\theta}\| < M$ we have $\boldsymbol{\theta}_K \in N\big(\delta(\varepsilon)\big)$, so using (3.4) with $x = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^T\{-\boldsymbol{D}^2 f(\boldsymbol{\theta}_K)\}\boldsymbol{\Sigma\theta}$ and $y = \|\boldsymbol{\theta}\|^2/2$ we get

$$\left|f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma\theta}\right) - e^{\left(-\frac{\|\boldsymbol{\theta}\|^2}{2}\right)}\right| \le \frac{\varepsilon M^2}{2} + \frac{M^4}{4K}.$$

As ε is arbitrary, this gives the result.

Corollaries to Lemma 3.3:

(1) By the continuity of π at $\boldsymbol{\theta}_\infty$ and Lemma 3.3, $\pi\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma\theta}\right) f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma\theta}\right)$ converges to $\pi(\boldsymbol{\theta}_\infty)\exp(-\|\boldsymbol{\theta}\|^2/2)$ uniformly on bounded sets.

(2) Lemma 3.3 and Fatou's lemma give us $\pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{\mathrm{p}/2} \le \liminf_K c(K)K^{\mathrm{p}/2}$. In particular, there is a constant $C > 0$ so that $\frac{1}{c(K)} \le CK^{\mathrm{p}/2}$.

Lemma 3.4: Under Assumptions 3.1 and 3.2, the following three statements are equivalent.

(a) $\boldsymbol{\Psi}_K \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_p)$ (convergence in distribution to a Normal random variable).

(b) The density g_K converges pointwise to a multivariate standard normal density function. That is, $c(K)K^{p/2} \longrightarrow \pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{p/2}$.

(c) $\boldsymbol{\Theta}_K \xrightarrow{D} \delta_\infty$, where δ_∞ indicates a degenerate distribution at $\boldsymbol{\theta}_\infty$.

Proof: To show (a) ⇒ (b). The density $g_K(.)$ can be written as

$$g_K(\boldsymbol{\theta}) = \frac{|\boldsymbol{\Sigma}|}{c(K)K^{p/2}}\pi\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right).$$

Let B be a bounded Borel set with positive Lebesgue measure. From the convergence in (a), we have

$$\frac{1}{(2\pi)^{p/2}}\int_B e^{\left(-\frac{\|\boldsymbol{\theta}\|^2}{2}\right)}d\boldsymbol{\theta} = \lim_K \frac{|\boldsymbol{\Sigma}|}{c(K)K^{p/2}}\int_B \pi\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)d\boldsymbol{\theta}.$$

On the other hand, the uniform convergence from Lemma 3.3 gives

$$\lim_K \int_B \pi\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)d\boldsymbol{\theta} = \pi(\boldsymbol{\theta}_\infty)\int_B e^{\left(-\frac{\|\boldsymbol{\theta}\|^2}{2}\right)}d\boldsymbol{\theta}.$$

Hence we can conclude that $c(K)K^{p/2} \longrightarrow \pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{p/2}$ as K converges to infinity. Combined with convergence in Lemma 3.3, this gives

$$g_K(\boldsymbol{\theta}) \longrightarrow \frac{1}{(2\pi)^{p/2}}\, e^{\left(-\frac{\|\boldsymbol{\theta}\|^2}{2}\right)}.$$

Also, (b) ⇒ (a) follows from the Scheffe's theorem, and, (a) ⇒ (c) is obvious. Now, we show (c) ⇒ (b). Because $-\boldsymbol{D}^2 f$ and π are continuous at $\boldsymbol{\theta}_\infty$ and $\boldsymbol{\Sigma}$ is strictly positive definite, from (3.3) we see that for any ε >0 we can find δ >0 so that $\boldsymbol{\theta} \in N(\delta)$ implies

$$f(\boldsymbol{\theta}) < 1 - \frac{1}{2}(1-\varepsilon)(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)^T\boldsymbol{\Sigma}^{-2}(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty), \tag{3.5 a}$$

$$\pi(\boldsymbol{\theta}) \le (1+\varepsilon)\pi(\boldsymbol{\theta}_\infty). \tag{3.5 b}$$

Also, by assumption (c), we may assume that K is so large that $1-\varepsilon \le \int_{N(\delta)}\pi_K(\boldsymbol{\theta})d\boldsymbol{\theta}$. Multiplying this inequality by $c(K)K^{p/2}(1-\varepsilon)^{-1}$ and using (3.5) gives

$$c(K)K^{p/2} \le K^{p/2}(1-\varepsilon)^{-1}\int_{N(\delta)}\pi_K(\boldsymbol{\theta})f^K(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$\le K^{p/2}(1-\varepsilon)^{-1}(\pi(\boldsymbol{\theta}_\infty)+\varepsilon)\int_{N(\delta)}\left[1-\frac{1}{2}(1-\varepsilon)(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)^T\boldsymbol{\Sigma}^{-2}(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)\right]^K d\boldsymbol{\theta}$$

$$\le K^{p/2}(1-\varepsilon)^{-1}(\pi(\boldsymbol{\theta}_\infty)+\varepsilon)\int_{N(\delta)} e^{\left[-\frac{K}{2}(1-\varepsilon)(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)^T\boldsymbol{\Sigma}^{-2}(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)\right]^K}d\boldsymbol{\theta}$$

$$= (1-\varepsilon)^{-1}(\pi(\boldsymbol{\theta}_\infty)+\varepsilon)|\boldsymbol{\Sigma}|(2\pi)^{\frac{p}{2}}.$$

By letting $K \to \infty$ and then $\varepsilon \to 0$, we get

$$\limsup_K c(K)K^{p/2} \le \pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{p/2}.$$

The other half comes from the inequality in Corollary 2 of Lemma 3.3.

Corollary to Lemma 3.4: Under Assumptions 3.1, 3.2, and 3.3, $\boldsymbol{\Theta}_K \xrightarrow{D} \delta_\infty$.

Proof: Using Assumption 3.3 and the second corollary to Lemma 3.3, we see that for any $\delta > 0$,

$$\frac{1}{c(K)} \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_\infty\|>\delta} \pi_K(x) f^K(x) dx \leq C K^{p/2} \gamma^K(\delta) \to 0.$$

This implies that $\boldsymbol{\Theta}_K \xrightarrow{D} \delta_\infty$.

Hence, the main result of the convergence of the DC algorithm, that under Assumptions 3.1, 3.2, and 3.3, $\boldsymbol{\Psi}_K \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_p)$, follows immediately.

Remark 3.1:. The proof given in Jacquir et al. (2007) assumes only Assumptions 3.1 and 3.2 The counter example below shows that they are not sufficient for convergence; Assumption 3.3 is necessary. Let $\Theta = \Re$, $\pi(\theta) = \frac{1}{2.5}\min\left(1, \frac{1}{4\theta^2}\right)$. Let the likelihood function be $f(\theta) = 1 - \frac{\theta^2}{2}$ when $|\theta| \leq 1$ and $f(\theta) = 1 - \frac{1}{|\theta|^3}$ when $|\theta| > 1$. In this case, $\sqrt{K} c(K) \to \infty$ and we do not get the convergence to a Normal distribution.

3.2 MCMC Implementation

The key trick in operationalizing the result obtained in Theorem 3.1 is to pretend that several independent copies, or *clones*, of the originally observed data vector are available. Specifically, we suppose that the statistical experiment underlying the observed data is independently repeated *K* times and, purely by chance, each results in exactly the same dataset, $\mathbf{y}_{(\mathrm{n})}$. We denote the resulting data set, consisting of *K* independent clones of the original data, as $\mathbf{y}^{(K)} = \left(\mathbf{y}_{(\mathrm{n})}, \mathbf{y}_{(\mathrm{n})}, \ldots, \mathbf{y}_{(\mathrm{n})}\right)$. It immediately follows that the likelihood function corresponding to this *K-cloned* dataset is given as $\left[L\left(\boldsymbol{\theta}; \mathbf{y}_{(\mathrm{n})}\right)\right]^K$. We further note that (a) the location of the maximum of the *K-cloned* likelihood function is exactly same as that of $L\left(\boldsymbol{\theta}; \mathbf{y}_{(\mathrm{n})}\right)$, and (b) the Fisher information matrix based on this likelihood is equal to *K* times the Fisher information matrix corresponding to $L\left(\boldsymbol{\theta}; \mathbf{y}_{(\mathrm{n})}\right)$. Let us now consider a Bayesian formulation of the general hierarchical model in equation (3.1), where we replace the original likelihood $L\left(\boldsymbol{\theta}; \mathbf{y}_{(\mathrm{n})}\right)$ by the *K-cloned* likelihood $\left[L\left(\boldsymbol{\theta}; \mathbf{y}_{(\mathrm{n})}\right)\right]^K$. Assuming $\pi(\boldsymbol{\theta})$ as a *proper* prior distribution on the parameter space, the posterior distribution of $\boldsymbol{\theta}$, conditional on the *K*-cloned dataset, $\mathbf{y}^{(K)}$, is given as

$$\pi_K\left(\boldsymbol{\theta} \middle| \mathbf{y}_{(\mathrm{n})}\right) = \frac{\left[\int f\left(\mathbf{y}_{(\mathrm{n})}; \boldsymbol{\theta}_1 | \mathbf{u}\right) g(\mathbf{u}; \boldsymbol{\theta}_2) d\mathbf{u}\right]^K \pi(\boldsymbol{\theta})}{c\left(K; \mathbf{y}_{(\mathrm{n})}\right)}$$

$$= \frac{[L(\boldsymbol{\theta};\mathbf{y}_{(\mathrm{n})})]^K \pi(\boldsymbol{\theta})}{c(K;\,\mathbf{y}_{(\mathrm{n})})}, \tag{3.6}$$

where $c(K;\,\mathbf{y}_{(\mathrm{n})}) = \int [L(\boldsymbol{\theta};\mathbf{y}_{(\mathrm{n})})]^K \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the normalizing constant. Interestingly, this posterior distribution is exactly the same as the one given by (3.1) via the iterated map formulation $\pi_1 = \Im(\pi\,, L)$, $\pi_2 = \Im(\pi_1\,, L)$, …, $\pi_K = \Im(\pi_{K-1}\,, L)$.

Thus, with sufficiently large *K*, Theorem 3.1 ensures that the MLE, $\widehat{\boldsymbol{\theta}}$, and the corresponding asymptotic standard errors can be obtained by computing the mean and variance of the posterior distribution in (3.6). However, owing to the presence of high dimensional integrals in the likelihood function, these posterior quantities are highly intractable to compute numerically. A way out is to resort to their Monte Carlo approximation by generating random variates $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B$ from $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})})$ and use their mean and variance to obtain the MLE and its asymptotic variance. Fortunately, as outlined below, such generation of random variates is quite straightforward using the MCMC technique.

Recall that the *K*-cloned likelihood function $[L(\boldsymbol{\theta};\mathbf{y}_{(\mathrm{n})})]^K$ corresponds to a thought experiment where *K* experimenters happen to obtain exactly the same data set $\mathbf{y}_{(\mathrm{n})}$ independently. We conduct this thought experiment using computers. We create the *K*-cloned dataset, $\mathbf{y}^{(K)} = (\mathbf{y}_{(\mathrm{n})}, \mathbf{y}_{(\mathrm{n})}, \ldots, \mathbf{y}_{(\mathrm{n})})$, by repeating the observed data vector *K* times. We pretend as if these data were obtained from the *K* independent experiments and use the standard MCMC approach to generate random variates $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B$ from the posterior $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})})$. This can be easily implemented using the freely available software packages such as WinBUGS (Spiegelhalter al., 2004) and JAGS (Plummer, 2011a, 2011b). Thus, if *K* is large, the MLE of the parameter $\boldsymbol{\theta}$ is simply the mean of these random variates. Furthermore, if the parameter space is continuous, *K* times the variance (or, variance–covariance matrix for the multiparameter case) of these random variates is the estimated variance of the MLE, the inverse of the Fisher information, $I^{-1}(\widehat{\boldsymbol{\theta}})$, based on the original data.

In addition, inference for transformations of the model parameters becomes readily available. Let $\tau(\boldsymbol{\theta})$ be a transformation from $\mathfrak{R}^p$ to $\mathfrak{R}^q$, where $1 \le q \le p$. Then, by the invariance property of the MLE, $\tau(\widehat{\boldsymbol{\theta}})$ is the MLE of $\tau(\boldsymbol{\theta})$. Also, we compute the variates $\tau(\boldsymbol{\theta}_1), \tau(\boldsymbol{\theta}_2), \ldots, \tau(\boldsymbol{\theta}_B)$ from the MCMC-generated random variates $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B$. Then, *K* times the variance (or, variance–covariance matrix for the multiparameter case)

of these transformed random variates is the estimated variance of $\tau(\widehat{\boldsymbol{\theta}})$, i.e. the inverse of the Fisher information $I^{-1}(\boldsymbol{\tau}(\widehat{\boldsymbol{\theta}}))$.

Thus, quite remarkably, the DC procedure avoids: (i) analytical or numerical evaluation of the high-dimensional integral which is a major computational hurdle for maximum likelihood estimation for GLMM; (ii) numerical optimization of a function; and (iii) numerical computation of the curvature of the likelihood function. The number of clones to be used in the procedure is completely under the control of the analyst. It can be made as large as necessary to achieve the desired accuracy of the resultant estimates.

3.2.1 Determining Adequate Number of Clones

An important issue in implementing the DC algorithm is the *practical* convergence of the algorithm, i.e., determining adequate number of clones so that the posterior distribution $\pi_K\left(\boldsymbol{\theta}\middle|\mathbf{y}_{(\mathrm{n})}\right)$ becomes nearly degenerate. We achieve this by plotting a standardized version of the largest eigenvalues, λ_K, of the variance-covariance matrix of the *K*-cloned posterior distribution as a function of *K*. We notice that the largest eigenvalue, λ_K, converges to zero at the same rate as 1/*K*, as do the marginal posterior variances. The standardized largest eigenvalues are computed as $\lambda_K^S = \lambda_K/\lambda_1$, where λ_1 is the largest eigenvalue of the posterior variance with a single clone. Therefore, convergence of the algorithm can be monitored by comparing λ_K^S with the expected rate 1/*K*. We choose number of clones so that (a) λ_K^S consistently decreases at rate 1/*K* with increasing values of *K* and (b) the posterior also becomes sufficiently close to normal distribution. We know that when (b) holds, $(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}_K)^T\mathbf{V}_K^{-1}(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}_K)$ is a chi-square random variate, where $\overline{\boldsymbol{\theta}}_K$ and $\mathbf{V}_K$ are mean and variance (covariance) of the *K*-cloned posterior distribution respectively. In this case, the following statistics are expected to be close to zero (Johnson and Wichern 2007): (i) $\omega_K = \frac{1}{B}\sum_{j=1}^{B}\left(O_j^{(K)} - E_j\right)^2$ where $O_j^{(K)} = (\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}_K)^T\mathbf{V}_K^{-1}(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}_K)$, and E_j are the quantiles for χ_p^2 random variable, and (ii) $\tilde{r}_K^2 = 1 - \mathrm{corr}^2\left(O_j^{(K)},\ E_j\right)$. Thus, convergence is achieved at the value of *K* for which ω_K and $\tilde{r}_K^2$ are below some specified threshold such as 0.001. The detailed implementation steps of the DC algorithm are outlined in Algorithm 3.1.

Algorithm 3.1 Data Cloning Algorithm with a Fixed Prior

1. Let $(K_1, K_2, \ldots, K_J)$ be a set of increasing positive integers and assume a proper prior distribution $\pi(\boldsymbol{\theta})$. Set $i = 1$ and proceed to the next step.
2. Set $K = K_i$ and construct a *K*-cloned data set $\mathbf{y}^{(K)} = \left(\mathbf{y}_{(\mathrm{n})}, \mathbf{y}_{(\mathrm{n})}, \ldots, \mathbf{y}_{(\mathrm{n})}\right)$ by repeating the observed data vector *K* times.
3. Pretend as if these data were *K* independent realizations from the data generating mechanism of $\mathbf{y}_{(\mathrm{n})}$. Conditional on these data, generate random ates $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(B)}$ from the distribution $\pi_K\left(\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})}\right) = \left[L\left(\boldsymbol{\theta}; \mathbf{y}_{(\mathrm{n})}\right)\right]^K \pi(\boldsymbol{\theta}) / c(K; \mathbf{y}_{(\mathrm{n})})$ using any MCMC type algorithm.
4. Use random variates generated in Step-3 to compute the posterior mean $\overline{\boldsymbol{\theta}}_K$ and variance-covariance matrix $\mathbf{V}_K$.
5. Compute the largest eigenvalue, λ_K, of the variance-covariance matrix $\mathbf{V}_K$. Divide it by the largest eigenvalue when number of clones is K_1 to obtain the standardized largest eigenvalue, i.e. $\lambda_K^S = \lambda_K / \lambda_{K_1}$.
6. Compute $\lambda_K^E = K_1 / K$. This is the standardized rate at which λ_K^S is expected to drop with increasing clone size.
7. Compute $\omega_K = \frac{1}{B}\sum_{j=1}^{B}\left(O_j^{(K)} - E_j\right)^2$ and $\tilde{r}_K^2 = 1 - \mathrm{corr}^2\left(O_j^{(K)}, E_j\right)$, where $O_j^{(K)} = (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_K)^T \mathbf{V}_K^{-1} (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_K)$ and E_j are the quantiles for χ_p^2 random variable
8. Plot the quantities λ_K^S, λ_K^E, ω_K and $\tilde{r}_K^2$ verses the number of clones *K*.
9. Jump to the next step if the following two conditions are satisfied:
   (a) The standardized largest eigenvalue λ_K^S is dropping at the expected rate λ_K^E.
   (b) Both ω_K and $\tilde{r}_K^2$ are below a small threshold, say 0.01.
   Otherwise, set next $i = i + 1$ and go back to Step-2.
10. Compute $\widehat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}_K$ and $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\theta}}} = \sqrt{K}\mathbf{V}_K$. These are, respectively, MLE of $\boldsymbol{\theta}$ and that of the inverse of the Fisher information matrix $I^{-1}(\boldsymbol{\theta})$.

3.2.2 Choosing the Prior Distribution

Theorem 3.1 holds for any choice of a proper prior distribution $\pi(\boldsymbol{\theta})$. This, in principle, means that DC based likelihood estimates remain invariant to the choice of prior distributions. However, substantially noninformative (or *vague*) priors can cause poor MCMC

mixing leading to slow convergence of the DC algorithm. On the other hand, using strongly disinformative (informative but wrong) priors can also slow down the convergence as very high cloning size *K* might be required to fully eliminate the priors' influence. Our experience dictates that using moderately vague priors usually results in faster implementation of the DC algorithm.

Alternatively, we also suggest assuming a multivariate normal prior centered at the MLE's obtained from fitting a fixed-effects GLM model, i.e. ignoring the presence of latent effects. For instance, consider fitting the following logistic GLMM,

Hierarchy 1: $Y_i|p_i \sim Binomial(n_i, p_i)$, where

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i + u_i, \text{ and } X \text{ is a covariate.}$$

Hierarchy 2: $u_i \sim N(0, \sigma_u^2)$.

The multivariate normal prior can be constructed using the following two-step procedure.

Step-1: Use Algorithm 3.1 to fit the simpler model ignoring the random effects **u**. That is, we only fit the logistic GLM with $\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i$, assuming a moderately vague prior distribution $\pi(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = (\alpha, \beta)^{\mathrm{T}}$. This yields the MLE $\widehat{\boldsymbol{\eta}}_{glm}$ and the corresponding variance-covariance matrix $I^{-1}(\widehat{\boldsymbol{\eta}}_{glm})$.

Step-2. Construct the multivariate normal prior as follows.

$$\pi(\boldsymbol{\theta}) = \phi\left(\boldsymbol{\theta}; \begin{bmatrix} \widehat{\boldsymbol{\eta}}_{glm} \\ 0 \end{bmatrix}, \begin{bmatrix} I^{-1}(\widehat{\boldsymbol{\eta}}_{glm}) & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix}\right)$$

where $\boldsymbol{\theta} = (\boldsymbol{\eta}^T, \log \sigma_u^2)^T$ and $\phi(;\mu, \sigma^2)$ indicates the Normal density with mean μ and variance σ^2.

We emphasize that Algorithm 3.1 assumes the same prior distribution for each iteration of the DC algorithm. We can further improve upon it by iteratively incorporating the information contained in $\pi_{K_{i-1}}\left(\boldsymbol{\theta}\middle|\mathbf{y}_{(\mathrm{n})}\right)$ when computing the next posterior distribution $\pi_{K_i}\left(\boldsymbol{\theta}\middle|\mathbf{y}_{(\mathrm{n})}\right)$. We operationalize this idea in Algorithm 3.2. An advantage of Algorithm 3.2 is that, apart from speeding up the DC computations, it can also help circumvent MCMC convergence issues. The *dclone* package in R (Sólymos 2010) provides an efficient and user friendly implementation of the DC algorithm. The package is available from the package section of the Comprehensive R Archive Network site (< http://cran.r-project.org/>).

Algorithm 3.2 Data Cloning Algorithm with Prior Updating

The algorithm is exactly the same as Algorithm 3.1, except for the following Step.

9. Jump to the next step if the following two conditions are satisfied:

(a) The standardized largest eigenvalue λ_K^S is dropping at the expected rate λ_K^E.

(b) Both ω_K and $\tilde{r}_K^2$ are below a small threshold, say 0.01.

Otherwise, set $i = i + 1$ and $\pi(\boldsymbol{\theta}) = N\left(\overline{\boldsymbol{\theta}}_K, \sqrt{K}\mathbf{V}_K\right)$, and go back to Step-2.

3.3 Prediction of Random Effects

An important inferential component of many hierarchical models is prediction of random effects. One can use MCMC along with data cloning to obtain point prediction and prediction intervals for the random effects. The method is based on the results of Harris (1989) where it is shown that if one uses the bootstrap distribution of the parameters as the 'prior', the posterior distribution of the random effects is the best approximation, in Kullback–Leibler divergence, to the true distribution. We suggest replacing the bootstrap distribution by the Normal approximation obtained by data cloning. This may also be looked upon as the prior invariant component of the posterior distribution. Thus, prediction inference on random effects is obtained by using

$$\pi\left(\mathbf{u}|\mathbf{y}_{(\mathrm{n})}\right) = \frac{\int f\left(\mathbf{y}_{(\mathrm{n})}|\mathbf{u},\boldsymbol{\theta}_1\right) g(\mathbf{u}|\boldsymbol{\theta}_2)\phi\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}},I^{-1}(\hat{\boldsymbol{\theta}})\right)d\boldsymbol{\theta}}{c\left(\mathbf{y}_{(\mathrm{n})}\right)},$$

where $\phi(\theta;\mu,\sigma^2)$ indicates the Normal density with mean μ and variance σ^2. The MCMC algorithm can be used to obtain the draws from this distribution without actually conducting the integration. We simply obtain the random numbers from

$$\pi\left(\mathbf{u},\boldsymbol{\theta}|\mathbf{y}_{(\mathrm{n})}\right) = \frac{f\left(\mathbf{y}_{(\mathrm{n})}|\mathbf{u},\boldsymbol{\theta}_1\right) g(\mathbf{u}|\boldsymbol{\theta}_2)\phi\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}},I^{-1}(\hat{\boldsymbol{\theta}})\right)}{c\left(\mathbf{y}_{(\mathrm{n})}\right)}$$

and utilize the **u** component only.

3.4 Illustrative Examples

In the following we apply data cloning to obtain maximum likelihood estimates and associated asymptotic standard errors for three important subclasses of Generalized Linear Mixed Models with wide applications in medical statistics and epidemiology. The detailed description of the scientific problems, statistical models, and the data is available in

Breslow and Clayton (1993). The following descriptions are borrowed from Breslow and Clayton (1993, section 6).

3.4.1 Logistic–Normal Mixed Model

Crowder (1978, Table 3) presented data on the proportion of seeds that germinated on each of 21 plates arranged according to a 2 × 2 factorial layout by seed variety and type of root extract. He noted that the within-group variation exceeded that predicted by binomial sampling theory. A natural way to account for extraneous plate-to-plate variability in this situation is by means of the following GLMM:

Hierarchy 1: $Y_i|p_i \sim Binomial(n_i, p_i)$, where

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_{SEED} SEED + \alpha_{\text{extract}} EXTRACT + \alpha_{\text{interaction}} SEED * EXTRACT + b_i$$

Hierarchy 2: $b_i \sim N(0, \sigma_b^2)$.

Breslow and Clayton (1993) provide the exact ML estimates of the parameters along with their standard errors based on numerical integration. In Table 3.1, we provide the results based on the data-cloning algorithms with two different priors. The first set of DC estimates (Data Cloning 1) is obtained via Algorithm 3.1 using a noninformative prior. The second set (Data Cloning 2) is computed using Algorithm 3.2 starting with a multivariate normal prior based on the GLM estimates. We compare these estimates with those based on noninformative Bayes estimates. The DC based MLEs and their SEs are nearly identical to the exact ML estimators and are invariant to the choice of the priors. Figure 3.1-a gives the DC convergence diagnostics and Figure 3.1-b shows the DC based point predictions and prediction intervals for the probability of germination along with those based on noninformative priors. These match reasonably well with the ones obtained by using noninformative Bayes approach.

3.4.2 Longitudinal Data

Thall and Vail (1990, Table 2) presented data from a clinical trial of 59 epileptics who were randomized to a new drug (*Trt* = 1) or a placebo (*Trt* = 0) as an adjuvant to the standard chemotherapy. Baseline data available at entry into the trial included the number of epileptic seizures recorded in the preceding eight-week period and age in years. The logarithm of one fourth of number of baseline seizures (*Base*) and the logarithm of age (*AGE*) were treated as covariates in the analysis. A multivariate response variable

Table 3.1 Maximum likelihood estimates and standard errors (SEs) using data cloning under two different priors and comparison with the estimates and variances using the noninformative Bayesian analysis.

Parameters		Data Cloning 1		Data Cloning 2		Noninformative Bayes	
Example 1	$\alpha_{\circ}$	-0.5484	(0.1693)	-0.5491	(0.1623)	-0.5488	(0.2129)
	α_1	0.0970	(0.2758)	0.0993	(0.2771)	0.0515	(0.3462)
	α_2	1.3372	(0.2403)	1.3378	(0.2357)	1.3583	(0.3076)
	α_{12}	-0.8113	(0.3837)	-0.8133	(0.3879)	-0.8181	(0.4762)
	σ	0.2376	(0.1069)	0.2361	(0.1061)	0.3546	(0.1469)
Example 2	$\alpha_{\circ}$	-0.4381	(0.1693)	-0.4397	(0.1372)	-0.5581	(0.1496)
	α_1	0.6078	(0.0901)	0.6084	(0.1181)	0.6560	(0.0893)
	σ	1.2888	(0.2112)	1.2890	(0.1992)	1.4468	(0.2214)
	γ	0.1770	(0.0111)	0.1770	(0.0101)	0.1429	(0.0388)
Example 3	$\alpha_{\circ}$	-1.3934	(1.1965)	-1.4070	(1.2343)	-1.4165	(1.2537)
	α_{Base}	0.8782	(0.1318)	0.8822	(0.1180)	0.8824	(0.1293)
	α_{Trt}	-0.9493	(0.3827)	-0.9448	(0.3959)	-0.9739	(0.3889)
	α_{BT}	0.3501	(0.1913)	0.3473	(0.1975)	0.3632	(0.1980)
	α_{Age}	0.4852	(0.3519)	0.4872	(0.3715)	0.4883	(0.3700)
	α_V	-0.1019	(0.0861)	-0.1016	(0.0872)	-0.1026	(0.0877)
	σ_b	0.3590	(0.0430)	0.3593	(0.0412)	0.3622	(0.0428)
	σ_{b1}	0.4623	(0.0622)	0.4621	(0.0635)	0.4934	(0.0697)

consisted of the counts of seizures during the two weeks before each of four clinic visits (Visit, coded −3, −1, 1, and 3). Preliminary analysis indicated that the counts were substantially lower during the fourth visit and a binary variable ($V4$ = 1 for fourth visit, 0 otherwise) was constructed to model such effects. Breslow and Clayton (1993) use the following GLMM for modeling these data:

Hierarchy 1: $Y_{ijk}|\mu_{jk} \sim Poisson(\mu_{jk})$, where

$$\log(\mu_{jk}) = \alpha_0 + \alpha_{\mathrm{AGE}} AGE + \alpha_{\mathrm{BASE}} BASE + \alpha_{\mathrm{Trt}} Trt + \alpha_{\mathrm{BT}}(BASE * Trt) + \alpha_{\mathrm{V_4}} V_4 + b_j + b_{jk}$$

Hierarchy 2: $b_j \sim N(0, \sigma_b^2)$ and $b_{jk} \sim N(0, \sigma_{b1}^2)$.

In Table 3.1, we present the MLEs obtained using data cloning procedure. The results again do not depend on the choice of the priors. In Figure 3.1-c, we show the convergence diagnostic plots and Figure 3.1-d shows the DC based-point predictions and

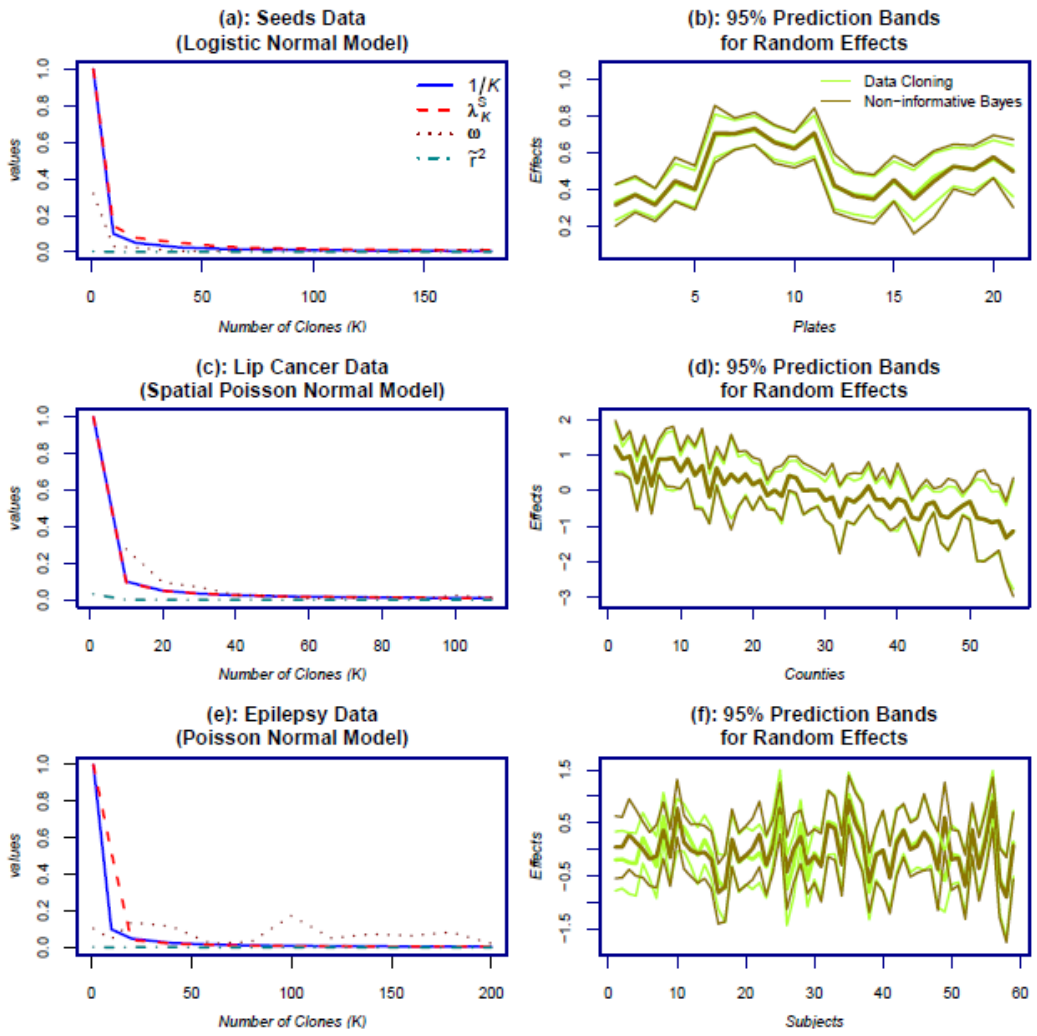

Figure 3.1 Data cloning convergence diagnostics and prediction of random effects for the three examples. The standardized eigenvalues converge to zero at the expected rate for all three cases. Data cloning based prediction intervals for random effects are quite similar to the ones obtained using noninformative priors.

prediction intervals for subject effects. These match reasonably well with the ones obtained using noninformative priors.

3.4.3 Spatial Smoothing of Disease Maps

One of the most common applications of GLMM is in the context of spatial smoothing of disease maps (Clayton and Kaldor 1987; Diggle, et al. 2002). We consider the data reported in Clayton and Kaldor (1987) on the number of lip cancer cases in the 56 counties of Scotland. Clayton and Kaldor (1987) proposed an empirical Bayes estimation of the county specific standardized mortality rates (SMRs) using several alternative assump-

tions about the distribution of the random effects. These data subsequently were analyzed by Breslow and Clayton (1993) using the PQL. In the following analysis, we use a proper, conditionally specified autoregression (CAR) model. A full discussion of these different analyses along with the Bayesian implementation is available inWinBUGS (Spiegelhalter et al. 2004, maps section). The model we use is as follows:

Hierarchy 1: $Y_i|\mu_i \sim Poisson(\mu_i)$.

Hierarchy 2: $\log(\mu_i) = \log e_i + \alpha_0 + \alpha_1 \frac{x_i}{10} + b_i$, where e_i = expected count and x_i = % of work force employed in agriculture, fishing, and forestry.

Hierarchy 3: $\mathbf{b} \sim \text{MVN}(\mathbf{0}, \mathbf{V})$ where $\mathbf{V} = \sigma^2(1 - \gamma\mathbf{C})^{-1}\mathbf{M}$, $M_{ij} = 1/e_i$, the inverse of the expected count in the i^{th} area, and $C_{ij} = e_i / e_j$. The spatial association parameter $\gamma \in (\gamma_{\min}, \gamma_{\max})$, where $\gamma_{\min}^{-1}$ and $\gamma_{\max}^{-1}$ are, max are the smallest and largest eigenvalues of $\mathbf{M}^{-1/2}\mathbf{C}\mathbf{M}^{1/2}$, respectively.

This ensures that the distribution of the random effects is a proper distribution. The maximum likelihood estimates and standard errors of the parameters are provided in Table 3.1. Convergence diagnostics are shown in Figure 3.1-e and predicted random effects and associated prediction intervals for counties are shown in Figure 3.1-f. They again match well with the ones based on noninformative priors.

3.5 Model Estimability

A desirable property in statistical model estimation is the ability to recover true parameter values given infinite amount of information available under the assumed model. However, the structure of the particular model at hand, or that of available data, may lead to *inestimability* of the model parameters. The key relevant concepts in this regard are that of structural identifiability of the model and its statistical estimability. Following Paulino and Pereira (1994), we state some basic ideas concerning model identifiability and estimability as follows.

A parametric statistical model is a probability triple $(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{P}})$ where $\boldsymbol{\mathcal{Y}}$ is the sample space with corresponding σ-algebra $\boldsymbol{\mathcal{A}}$, and $\boldsymbol{\mathcal{P}}$ is a family of probability measures defined on $(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{P}})$. The parametric structure of $\boldsymbol{\mathcal{P}}$ is specified as $\boldsymbol{\mathcal{P}} = \{P_{\boldsymbol{\theta}}: \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is a finite-dimensional parameter space. The following two definitions by Paulino and Pereira (1994) define the concepts of model identifiability.

Definition 3.3: Two points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of $\boldsymbol{\Theta}$ are called observationally equivalent (notated as $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$) if $P_{\boldsymbol{\theta}_1}(A) = P_{\boldsymbol{\theta}_2}(A)$, $\forall\, A \in \boldsymbol{\mathcal{A}}$.

In terms of the model likelihood function $L(\boldsymbol{\theta}; \mathbf{y}_{(n)})$, $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$ if and only if

$$L(\boldsymbol{\theta}_1; \mathbf{y}_{(n)}) = L(\boldsymbol{\theta}_2; \mathbf{y}_{(n)}), \qquad \forall\, \mathbf{y}_{(n)} \in \boldsymbol{\mathcal{Y}}.$$

Thus, the likelihood function is constant over all observationally equivalent points. Paulino and Pereira (1994) note that $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$ induces a partition of $\boldsymbol{\Theta}$ in equivalence classes, defined as $[\boldsymbol{\theta}_1] = \{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta} : P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}\}$. This partition is a quotient set of $\boldsymbol{\Theta}$ with respect to $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$, notated as $\boldsymbol{\Theta}/\sim$.

Definition 3.4 (Model Identifiability): i) The point $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ is said identifiable if $[\boldsymbol{\theta}_1] = \{\boldsymbol{\theta}_1\}$. ii) The parameter space $\boldsymbol{\Theta}$ (or the statistical model $P_{\boldsymbol{\theta}}$) is said structurally identifiable if $[\boldsymbol{\theta}] = \{\boldsymbol{\theta}\}$, $\forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta}$. That is, if $\boldsymbol{\Theta}/\sim$ is the finest possible partition, $\boldsymbol{\Theta}/\sim = \{\{\boldsymbol{\theta}\}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.

We now define the concept of parameter estimability under the regularity conditions stated in Assumption 3.1, 3.2 and 3.3.

Definition 3.5 (Parameter Estimability): The model $P_{\boldsymbol{\theta}}$ is said estimable if the set

$$N(\boldsymbol{\theta}) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : L(\boldsymbol{\theta}; \mathbf{y}_{(n)}) = L(\widehat{\boldsymbol{\theta}}; \mathbf{y}_{(n)})\},$$

is a single point set for all $\mathbf{y}_{(n)} \in \boldsymbol{\mathcal{Y}}$.

This definition essentially states that a model $P_{\boldsymbol{\theta}}$ is called estimable if existence and uniqueness of $\widehat{\boldsymbol{\theta}}$, the MLE, is ensured for all possible realizations from the sample space $\boldsymbol{\mathcal{Y}}$. Furthermore, the regularity conditions guarantee the consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}$. Thus, $P_{\boldsymbol{\theta}}$ is estimable if (i) $\widehat{\boldsymbol{\theta}}$ exists and is unique for all $\mathbf{y}_{(n)} \in \boldsymbol{\mathcal{Y}}$, and (ii) $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$.

Paulino and Pereira (1994), on the other hand, define parameter estimability as follows.

Definition 3.6: A function $\tau(\boldsymbol{\theta})$ is said estimable if it admits an unbiased estimator.

We, however, argue that model estimability is inherently linked to the existence of a consistent, rather than unbiased, estimator of $\boldsymbol{\theta}$. That is, for $P_{\boldsymbol{\theta}}$ to be estimable, the true value of $\boldsymbol{\theta}$ should be recoverable given infinite amount of information available under $P_{\boldsymbol{\theta}}$. On the contrary, unbiasedness of an estimator is not a sufficient condition for its consistency. We therefore employ Definition 3.5 as the definition of model estimability in this thesis.

The following theorem by Paulino and Pereira (1994) claims that estimability of model parameters $\boldsymbol{\theta}$ is a sufficient condition for model identifiability.

Theorem 3.2: If the model parameter vector $\boldsymbol{\theta}$ is estimable, $P_{\boldsymbol{\theta}}$ is identifiable.

It is important to note that model identifiability does not imply parameter estimability. Rather, model identifiability is a necessary condition for estimability.

Many hierarchical models have nonidentifiable parameters. For example, in the standard measurement error model $Y_i \mid \mu_i \sim N(\mu_i, \sigma^2)$ and $\mu_i \sim N(\mu, \tau^2)$ for $i = 1,2,...,n$, the parameters $(\mu, \sigma^2 + \tau^2)$are identifiable but parameters (μ, σ^2, τ^2) are not identifiable. It is known that (McCulloch and Searle 2001) for the Logistic–Normal model (Example 1, Section 4), if only one observation per stratum is available, the variance parameter σ^2 is confounded with the intercept parameter β_0. The analytical proof of this result, however, is difficult. In most practical applications, models are substantially more complex (Royle and Dorazio 2009; Clark and Gelfand 2006), making analytical proofs for identifiability of the parameters extremely difficult and are rarely attempted. Analysis is usually carried out as if the parameters are, in fact, identifiable (Lele 2010).

Common methods of assessing lack of parameter estimability include examining the rank of the Fisher information matrix or computing profile likelihoods after model estimation. Model inestimability may lead to rank deficiencies in the Fisher information matrix or extreme parameter correlations (Rodriguez-Fernandez et al 2006; Schittowiski 2007). However, if the likelihood attains its maximum at distinct modes, Fisher information matrix may be of full rank at each mode. On the other hand, profile likelihood exploration can fail to detect lack of estimability due to user defined bounds over which the likelihood is explored (Campbell and Lele 2013). Here, we present a simple DC based estimability diagnostic procedure that circumvents these limitations. We first state and prove the following theorem.

Theorem 3.3: Consider the set $N(\boldsymbol{\theta}) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}: L(\boldsymbol{\theta}; \mathbf{y}_{(n)}) = L(\widehat{\boldsymbol{\theta}}; \mathbf{y}_{(n)})\}$. Suppose this set is not a single point set, that is, the likelihood function is identical over the set $N(\boldsymbol{\theta})$. As $K \to \infty$, the posterior distribution converges to a distribution with density $\frac{\pi(\boldsymbol{\theta})}{\int_{N(\boldsymbol{\theta})} \pi(\boldsymbol{\theta})}$ for $\boldsymbol{\theta} \in N(\boldsymbol{\theta})$. If the set $N(\boldsymbol{\theta})$ is not a single point set, λ_K, the largest eigenvalue of the posterior variance matrix, does not converge to 0.

***Proof*:** Consider $\frac{\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)})}{\pi_K(\widehat{\boldsymbol{\theta}}|\mathbf{y}_{(n)})} = \frac{\pi(\boldsymbol{\theta}) f^K(\boldsymbol{\theta}|\mathbf{y}_{(n)})}{\pi(\widehat{\boldsymbol{\theta}}) f^K(\boldsymbol{\theta}|\mathbf{y}_{(n)})}$. It is obvious that for $\boldsymbol{\theta} \notin N(\boldsymbol{\theta})$,

$$\frac{\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)})}{\pi_K(\widehat{\boldsymbol{\theta}}|\mathbf{y}_{(n)})} = \frac{\pi(\boldsymbol{\theta}) f^K(\boldsymbol{\theta}|\mathbf{y}_{(n)})}{\pi(\widehat{\boldsymbol{\theta}}) f^K(\boldsymbol{\theta}|\mathbf{y}_{(n)})} \to 0.$$

It is also equally obvious that for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in N(\boldsymbol{\theta})$,

$$\frac{\pi_K(\boldsymbol{\theta}_1|\mathbf{y}_{(n)})}{\pi_K(\boldsymbol{\theta}_2|\mathbf{y}_{(n)})} = \frac{\pi(\boldsymbol{\theta}_1) f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_2) f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta}_2)} = \frac{\pi(\boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_2)}.$$

Hence the result follows.

Corollaries to Theorem 3.3:

1) Let $\tau(\boldsymbol{\theta})$ be a function of $\boldsymbol{\theta}$ such that it takes a unique value on the set $N(\boldsymbol{\theta})$. Then $\tau(\boldsymbol{\theta})$ is estimable.
2) Let $\pi_1(\boldsymbol{\theta})$ and $\pi_2(\boldsymbol{\theta})$ be two different prior distributions. Then, it follows that, as $K \to \infty$, the posterior distributions converge to $\frac{\pi_1(\boldsymbol{\theta})}{\int_{N(\boldsymbol{\theta})} \pi_1(\boldsymbol{\theta}) d\boldsymbol{\theta}}$ and $\frac{\pi_2(\boldsymbol{\theta})}{\int_{N(\boldsymbol{\theta})} \pi_2(\boldsymbol{\theta}) d\boldsymbol{\theta}}$ respectively. Hence the largest eigenvalue of the limiting posterior distribution depends on the choice of the prior distribution.

An immediate consequence of this result is that when the parameters are inestimable, as we increase the number of clones, the posterior distribution converges to a truncated prior distribution, truncated over the space of nonestimable parameter values. Consequently, the largest eigenvalue of the posterior variance matrix does not converge to zero. This result can be used to study lack of identifiability of the parameters in a statistical model as a whole. In practice, one may be interested in finding out whether certain functions of parameters are estimable. For example, in linear regression if the covariate matrix is singular, the regression parameters are nonestimable; however, the mean responses or differences in the treatment effects are estimable. Similarly in the applications of hierarchical models, a researcher might be interested in knowing if a specific parameter or a function of the parameters, $\tau(\boldsymbol{\theta})$, is estimable or not. The above result can be used to establish the estimability of $\tau(\boldsymbol{\theta})$. If the variance of the posterior distribution of the parameter of interest converges to zero, the parameter is estimable. Thus, data cloning not only alerts the researcher about nonestimability of the parameters in the model but also helps him/her in deciding if certain parameter(s) of interest are estimable or not. In the following, we illustrate the use of this technique in the context of hierarchical models.

3.5.1 Estimability Diagnostics

We start with a model where identifiability of various parameters is well established. Let $Y_i \mid \mu_i \sim N(\mu_i, \sigma^2)$ and $\mu_i \sim N(\mu, \tau^2)$ for $i = 1,2,...,n$. We generated a single realization from this model, used the MCMC algorithm to generate random variates from the posterior distribution for various number of clones and plotted λ_K^S as a function of *K*. We also plotted the posterior variance for various parameters that might be of interest. In Figure 3.2-a, it is clear that λ_K^S does not converge to zero as the number of clones is increased,

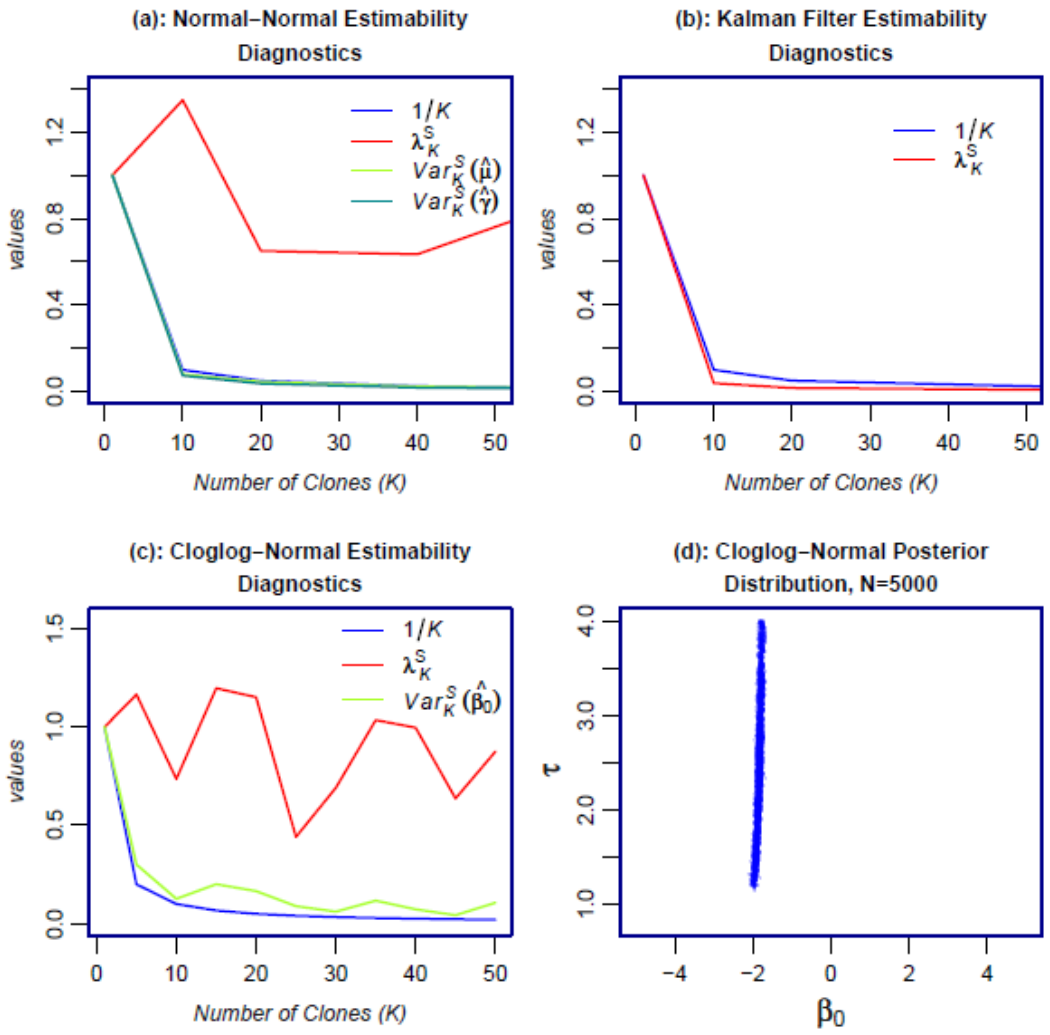

Figure 3.2 Estimability diagnostics using data cloning. In part (a), we consider Normal-Normal mixture. It is clear that λ_K^S does not converge to zero as K increases indicating non-estimability. However, the variance for μ does converge to zero indicating estimability. In part (b), we consider Kalman filter model. All parameters are estimable because λ_K^S does converge to zero as K increases. In part (c), we consider Binary-Normal mixture with complementary log-log link. It is clear that the model is non-estimable. Part (d) shows the posterior distribution is a truncated version of the prior distribution on a non-degenerate set supporting the non-estimability result further.

indicating non-estimability for the full model. On the other hand, the posterior variance for μ and $\gamma = \sigma^2 + \tau^2$ converges to zero as the number of clones increase, indicating their estimability. This shows that in the Normal–Normal model, μ and $\sigma^2 + \tau^2$ are estimable whereas σ^2 and τ^2 individually are not. Now we consider the classic Kalman filter model (Harvey 1993), $Y_i \mid \mu_i \sim N(\mu_i, \sigma^2)$ and $\mu_i \mid \mu_{i-1} \sim N(a + c\mu_{i-1}, \tau^2)$ for $i = 1,2,...,n$. The

Normal–Normal model above is a particular case of this model. However, introduction of correlation makes the parameters (a,c,σ^2,τ^2) identifiable as long as $c \neq 0$. In Figure 3.2-b, the plot of λ_K^S for the Kalman filter model clearly shows that the parameters are estimable.

Next we consider mixed Binary regression model. The analytical proof for the identifiability of various parameters in this model is difficult to establish (McCulloch and Searle 2001). Let $Y_i \mid p_i \sim Bernoulli(p_i)$, $p_i = 1 - \exp(-\exp(\beta_0 + \varepsilon_i))$ and $\varepsilon_i \sim N(0,\sigma^2)$ for $i = 1,2,...,n$. We considered n = 100 and the number of clones 1, 5, 10, . . . , 50. In Figure 3.2-c, we plot λ_K^S against K. It is obvious that the parameters in this model are nonestimable. To check this result, we also plot in Figure 3.2-d the posterior distribution based on 5000 observations and uniform priors. It is quite clear that the posterior distribution is nondegenerate even for such a large sample size and informative priors of *Uniform(*−5, 5*)* and *Uniform(*0.8, 4*)*. The marginal posterior distribution plot of β_0 as well as the data cloning plot for its variance as a function of the number of clones indicates that this parameter may be estimable. However, the rate at which the variance for β_0 converges to zero is not close to the theoretical rate of 1/*K* as was the case when the parameters are consistently estimable. Convergence may not necessarily indicate that the estimator is consistent for the true value. The posterior mean for β_0 was −1.82 (true value=−2) indicating possible inconsistency of this estimator.

3.5.2 Does Bayesian Learning Indicate Model Estimability?

The Bayesian perspective on identifiability is discussed in various articles (see, e.g., Gelfand and Sahu 1999 or Eberly and Carlin 2000). Both these articles note that sometimes the identifiability problems are subtly apparent in the convergence diagnostics for the MCMC or in the sensitivity of the posterior to the choice of the prior. They also discuss the concept of Bayesian learning when posterior distribution is changed due to the data. They seem to indicate that existence of Bayesian learning implies there are likely to be no problems with estimability. In the Binary–Normal example discussed above, the posterior distribution for the precision parameter $\tau = 1/\sigma^2$ was different than the prior distribution indicating some 'Bayesian learning' but clearly the parameter is nonestimable. Thus, some Bayesian learning is feasible even when the parameter is nonestimable. See also Lele (2010) for another example. This is concurrent with our result in Theorem 3.3 that the posterior distribution in the nonestimable parameter case is a truncated version of the

prior distribution, not necessarily the prior distribution itself. Similarly, we obtained good mixing and convergence (Gelman–Rubin's statistics of 1.06 and 1.12, respectively). These results also indicate that good mixing and convergence of the MCMC or evidence of Bayesian learning, although necessary, is not sufficient for estimability of the parameters. Convergence problems with MCMC and sensitivity to the choice of the prior can arise for various reasons. Aside from the possibility of nonestimability, they can also arise when the likelihood is relatively, but not exactly, flat or has multiple but unequal modes. These problems do not necessarily imply that the parameters are nonestimable. In data cloning, the information content of the sample is increased through cloning. By doing so, we eliminate the possibility of small information content affecting the convergence of MCMC and sensitivity to the choice of the prior. Thus, data-cloning-based test is clear and unambiguous. Of course, we consider this test as an additional tool to check for possible problems with the model and not a replacement of the checks proposed by Eberly and Carlin (2000) and others. Furthermore, in practice, published articles based on MCMC methodology seldom provide information on whether such checks were, in fact, conducted. Data-cloning methodology forces researchers to think about estimability issue carefully and to conduct such checks.

Hierarchical models are easy to construct and, thanks to MCMC, are easy to analyze. As a general principle, complexity of the model should not exceed the information content in the data (Lele 2010). Data cloning alerts the researcher to the potential pitfalls of the model such as nonestimability and points out any mismatch between the desired complexity of the model and what is feasible given the data.

3.6 Summary

In this chapter we refined the DC method introduced earlier by Lele et al. (2007). We obtained a general theoretical result that, for sufficiently large number of cloned data copies, the DC based posterior distribution is centered at the MLE with variance-covariance matrix equal to $1/K$ times the inverse of the Fisher information matrix. We implemented this result via a standard Bayesian formulation of the estimation problem, allowing the resulting DC algorithm to inherit the computational advantages offered by MCMC. We also developed a DC based algorithm for random effects' prediction. The illustrative examples in this chapter showed that DC method provides an efficient approach to analyzing GLMMs. We also obtained an important result related to the vexing problem of mod-

el identifiability. We showed that the ensuing DC based estimability diagnostic tool is very promising in resolving lack of estimability in complex hierarchical models.

Chapter 4

Population Viability Analysis: Incorporating Observation Error using State-Space Models[2]

Since the pioneering work of Shaffer (1981), population viability analysis (PVA) has become a key tool in wildlife management and conservation (Beissinger 2002). It is a procedure that uses population abundance data and population growth models to estimate the probability that a population will persist for a specified time into the future (Mills 2008). A typical PVA constitutes data collection, model formulation, model estimation and validation, and estimation of the extinction risk (Ralls et al. 2002). The last two decades have experienced a sea change in both the scale and complexity of PVA as population growth models have grown from modeling single population to spatially explicit metapopulations and beyond (Beissinger 2002). One of the major changes is the inclusion of environmental and demographic stochasticity (e.g. Dennis et al. 1991).

In this chapter we use data cloning to fit a specified population growth model to observed population time series in the presence of process variation and observation error (also called measurement error). Different population growth models have different extinction properties (Pascual et al. 1997; Henle et al. 2004) and hence the next important step is model selection. Estimation of the extinction risk is then based on computing vari-

[2] A version of this chapter has been published. Nadeem K, Lele S R. Oikos 2012, 121: 1656-1664.

ous extinction metrics by forecasting future population trajectories under the best fitting model.

Propagation of uncertainty in parameter estimation in forecasting future trajectories is an important issue. Not accounting for this uncertainty leads to inappropriate estimation of the extinction risk (Akçakaya and Raphael 1998; Ludwig 1999; Taylor et al. 2002). To incorporate the estimation uncertainty in forecasting, Dennis and Otten (2000) and Sæther et al. (2000) integrate over the bootstrap distribution of the parameter estimators. Here, we account for estimation uncertainty by integrating over the asymptotic normal distribution of the parameter estimates (see Section 3.3).

The chapter is organized as follows. In Section 4.1 we provide definitions of extinction metrics commonly used in PVA. In Section 4.2 we formulate the general state-space modeling setup and briefly review the currently used estimation methods. We present our model selection methodology for hierarchical models in Section 4.3. Section 4.4 presents a prediction algorithm for generating future population trajectories. We then develop algorithms for estimating the extinction metrics in Section 4.5. In Section 4.6 we analyze a focal population time series data to demonstrate the importance of incorporating observations error in PVAs. In Section 4.7 we extend our methodology to incorporate environmental covariates in PVA. The chapter concludes with a summary in Section 4.8.

4.1 Extinction Metrics used in PVA

Following is a detailed description of various extinction metrics used in the literature on PVA. Let N_t be the population abundance at t and $X_t = \ln(N_t)$. Throughout, $\mathbf{n} = (n_o, n_1, n_2, ..., n_q)$ denotes an abundance time series up to time q, and $\mathbf{x} = (x_o, x_1, x_2, ..., x_q)$ denotes corresponding log-abundances. Also, we represent random variables by capital letters and their realizations by small letters. For instance, x_t represents a realization of the random variable X_t. Vector valued random variables are denoted by bold-faced letters.

Throughout this chapter, we assume that N_t is a continuous random variable. That is, the process N_t serves as a continuous approximation to a discrete process for modeling growth rate in a population of interest.

***a) Population Prediction Interval* (PPI)** *(Saether et al. 2000)*

Let $G_{(r)}$ be the distribution function of the random variable X_{q+r}, the future log-population abundance at time *q*+*r*. The lower $(1-\alpha)$ prediction interval for the future log-population abundance X_{q+r} is then given as $[X_{\alpha}^{(r)},\infty)$, where $X_{\alpha}^{(r)} = G_{(r)}^{-1}(\alpha)$. This means future log-population at time $(q+r)$ will be somewhere in this interval with probability $(1-\alpha)$.

b) Conditional Time to Quasi-Extinction *(Dennis et al. 1991, Grimm and Wissel 2004; Morris and Doak 2002)*

Quasi-extinction (sometimes simply called extinction) is defined to occur when population size reaches some *threshold* level, denoted on the logarithmic scale as x_e. Conditional on the last observed log-population size $x_q > x_e$ and all the subsequent sample paths of the population process that reach the threshold x_e, conditional time to quasi-extinction (*T*) is defined as the first passage time of population abundance to reach x_e, i.e. $T := \min\{R > 0; X_{q+R} \le x_e\}$ where *R* is a random variable defined as the time a trajectory takes to reach x_e starting from x_q at time *q*. Probability distribution of this random variable is rarely known analytically. Monte Carlo estimation requires forecasting large number of future sample paths until extinction is observed. This is a computationally difficult task. Furthermore, in conservation planning, a short-term time horizon is often more useful. A practical version of time to extinction is defined as the first passage time to reach the threshold x_e conditional on population trajectories that reach x_e within time *t*. That is, $\tilde{T} := \{T > 0; X_{q+T} \le x_e \mid T \le t\}$. Some important metrics related to extinction times are mean, median and mode (also called *most likely*) time to extinction. We denote mean and median of $\tilde{T}$ by $\tilde{t}$ and $\tilde{\xi}_{(.5)}$ respectively.

c) Conditional Probability of Quasi-Extinction *(Staples et al. 2005)*

This is defined as the proportion of all sample paths hitting x_e within time *t* $\pi(n_e, t) := P\{X_{q+T} \le x_e, T \le t\}$. This probability is sometimes called *the extinction risk*.

The unconditional probability of extinction, usually denoted by π, is the probability that the population process will ever attain the extinction threshold (Dennis et al. 1991). Here, we only consider the conditional probability of extinction.

d) Probability of hitting a lower or upper threshold first

Let $x_v > x_e$ be an upper *policy-set population* threshold. Probability of reaching x_e before x_v is defined as $\pi^*_{[e,v]} := P\{X_{q+\omega} \le x_e\}$, where $\omega := \min(R > 0; X_{q+R} \in \{x_e, x_v\})$. The conditional version is defined as $\pi_{[e,v]} := P\{X_{q+\omega} \le x_e, 0 < \omega \le t\}$.

Similary the conditional probability of reaching x_v before x_e is defined as $\pi_{[v,e]} := P\{X_{q+\omega} \ge x_v, 0 < \omega \le t\}$, where $\omega := \min(R > 0; X_{q+R} \in \{x_e, x_v\})$.

e) Probability of recovering from Quasi-extinction

Given that a population trejectory has crossed a 'warning threshold' of x_s, it is of interest to know the probability that the population could recover within a specified time, before going extinct. This is given by $\pi_{[s,1]} := P\{0 < \omega' \le t - \omega, X_{q+t} > 0 \mid 0 < \omega < t\}$, where $\omega := \min(R > 0; X_{q+R} \in (0, x_s])$ and $\omega' := \min(R' > 0; X_{q+\omega+R'} > x_s)$.

4.1.1 Relationship between PPI's and Extinction Times

The population prediction intervals can be interpreted in terms of extinction times. Sæther et al. (2000), while discussing the concept of PPI, defined a relationship between PPI's and the conditional time to extinction, *T*. In the following we make this relationship precise.

Following Sæther et al. (2000), we consider a population to be *functionally extinct* when the surviving members of a sexually reproducing species are no longer able to reproduce. We therefore define $N_t = 1$ (equivalently $X_t = 0$) as the absorbing state. However, recalling that we are modeling population size as a continuous variable, the population is extinct by time *t* if $N_t \in (0,1]$, or equivalently $X_t \in (-\infty, 0]$. Therefore, the conditional probability of extinction, $\pi(1,t)$, can be written as:

$$\begin{aligned} P\{T \le t\} &= P\{X_{q+T} \le 0, T \le t\} \\ &= P\{X_{q+t} \le 0\}. \end{aligned} \tag{4.1}$$

Let us now define t_α to be the smallest time in which the lower (1- α) PPI contains the extinction barrier $x_e = 0$. Then, by the definition of PPIs, we have

$$P\{X_{q+t_\alpha} \geq 0\} = 1-\alpha \quad . \tag{4.2}$$

For which, $(4.1) \Rightarrow P\{T \geq t_\alpha\} = 1 - \alpha,$ (4.3)

or equivalently,

$$P\{T > t_\alpha\} = 1-\alpha. \tag{4.4}$$

Equations 4.2 and 4.3 together state that a $1-\alpha$ PPI obtained at time t_α also defines a corresponding lower $1-\alpha$ prediction interval, $[t_\alpha, \infty)$, for time to extinction, *T*.

Equation (4.4) says that extinction is predicted to occur after time t_α with probability $1-\alpha$. However, Sæther et al. (2000) interpreted it as saying that the probability of extinction after time t_α is α. This perhaps is due to a typographic error because, if we accept this interpretation, we have $P\{T > t_\alpha\} = \alpha. \Rightarrow P\{T \leq t_\alpha\} = 1-\alpha$ or $H_T(t_\alpha) = 1-\alpha$, where $H_T(.)$ is the cumulative distribution function of *T*. However, from Figure 4.2-a we have, $t_{.05} = 2$ and $t_{.1} = 4$, which according to Sæther et al.'s (2000) interpretation, respectively yield $H_T(2) = 0.95$ and $H_T(4) = 0.9$. This means that $H_T(4) < H_T(2)$. But this cannot be true since $H_T(t)$ is a monotone non-decreasing function in *t*; hence a contradiction.

4.2 Incorporating Observation Error: The State-Space Formulation

State-space formulation provides a powerful modeling tool for accommodating observation error and missing values in ecological analyses (McGowan et al. 2011). These models provide a flexible framework for estimating parameters of the population growth models in the presence of process variation and observation error (de Valpine and Hastings 2002; Clark and Bjørnstad 2004; Staples et al. 2004; Dennis et al. 2006; Lele 2006; Newman et al. 2006, Sæther et al. 2007). The basic nonlinear state-space model for population time series analysis (deValpine and Hastings 2002; Dennis et al. 2006; Lele 2006) is:

Process model: $X_t = m(X_{t-1}, \boldsymbol{\eta}) + E_t$ (4.5 a)

Observation model: $Y_t \sim f(y_t, X_t, \psi)$ (4.5 b)

Where $E_t \sim N(0, \sigma^2)$ represents process variation and $f(.)$ denotes observation error distribution that depends on an unknown parameter ψ. Different forms of the growth function, $m(X_{t-1})$, lead to different density-dependent growth models. For instance, $m(X_{t-1}) = X_{t-1} + a + bX_{t-1}$ corresponds to the stochastic Gompertz model (Gompertz 1825, Dennis and Taper 1994) and $m(X_{t-1}) = X_{t-1} + a + bN_{t-1}$ corresponds to the stochastic Ricker model (Ricker 1954). The parameter vector is $\boldsymbol{\varphi} = (\boldsymbol{\eta}^T, \sigma^2, \psi)^T$. In the rest of this chapter, we denote $g(x_t; m(X_{t-1}, \boldsymbol{\eta}), \sigma^2)$ as the process model density function, i.e. a Normal density with mean $m(X_{t-1}, \boldsymbol{\eta})$and variance σ^2.

Maximum-likelihood estimation in linear Gaussian state-space models can be conveniently obtained by using the Kalman filter (Harvey 1993, Schnute 1994). However, as we remarked in Chapter 2, likelihood based statistical inference for nonlinear non-Gaussian state-space models is extremely difficult. Evaluation of the likelihood function involves computationally intensive high dimensional numerical integration (Kitagawa 1987; deValpine 2002; deValpine and Hastings 2002). We refer the reader to Pedersen et al. (2011) for a review of methods of estimation for state-space population time series models.

We employ DC algorithm for the likelihood analysis of general nonlinear state-space population time series models. As we have seen in the previous chapter, data cloning can be used to obtain maximum likelihood estimates and associated standard errors in general hierarchical model. The prediction algorithm outlined in Section 4.4 can be used to predict unobserved states in a state space model. Handling missing data, higher order Markov models or spatial data is difficult with most existing methods of inference for hierarchical models. For example, Kitagawa's algorithm (deValpine and Hastings 2002) involves two or higher dimensional numerical integration if there are missing data or when the process model involves delayed density dependence, making it less practical for such cases. On the other hand, the Bayesian approach (Clark and Bjørnstad 2004) based on MCMC algorithms (Robert and Casella 2005) handles such situations without significant computational problems. Because data cloning uses the Bayesian computational machinery, it inherits all the computational advantages of the Bayesian approach at the same time avoiding the pitfalls of having the inference depend on the choice of the prior distribution.

4.3 Model Selection and Significance Testing

A key ingredient in hypothesis testing and model selection using information criteria (Burnham and Anderson 1998) is the likelihood ratio (LR) statistic. Likelihood ratios can also be used to compute the profile likelihood (PL) based confidence intervals that tend to have better statistical coverage properties than the Wald-type intervals (Meeker and Escobar 1995, Pawitan 2001). Profile likelihood calculations can be used to examine multimodality and likelihood ridges in the likelihood surface. Computation of LRs and PL is generally difficult in hierarchical models as one needs to calculate the maximized value of the likelihood surface. Recently, Ponciano et al. (2009) developed a DC based algorithm, called the DCLR algorithm, for computing LRs in hierarchical models. This algorithm is based on methods developed by Thompson and Guo (1991) and Thomson (1994) for Monte Carlo based estimation of LRs. Ponciano et al. (2009) implemented their algorithm in the context of state-space population dynamic models.

Here, in the context of state-space formulation (4.5), we further obtain Monte Carlo estimates of LR for the following two cases: (i) to compute LRs for comparing without observation models (4.5 a) versus with observation error models (4.5), and (ii) to compute LRs in case (i) when missing observations exist. We first provide a brief description of the DCLR algorithm as follows.

Let us assume, in general, that $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ be the parameters associated with the process model (4.5 a) and observation error model (4.5 b) respectively. Suppose we wish to compare two models defined by points $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\psi}^{(0)})$ and $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)})$ in the parameter space. This comparison can be performed by computing the likelihood ratio $L(\boldsymbol{\theta}^{(0)}, \boldsymbol{\psi}^{(0)})/L(\boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)})$. When the points $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)})$ and $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\psi}^{(0)})$ are MLEs under a full and a nested model respectively, this ratio can be used to conduct a LR test between these models. The DCLR algorithm consists of the following two-step procedure (Ponciano et al. 2009).

Step-1. Generate *m* random data samples $\mathbf{x}^{(1)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ from the conditional distribution

$$h\left(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)}\right) \propto f\left(\mathbf{y}|\mathbf{x}, \boldsymbol{\psi}^{(1)}\right) g\left(\mathbf{x}|\boldsymbol{\theta}^{(1)}\right)$$

of the latent states. These samples can be obtained via a straightforward MCMC algorithm using $g\left(\mathbf{x}|\boldsymbol{\theta}^{(1)}\right)$ as the prior distribution and $f\left(\mathbf{y}|\mathbf{x}, \boldsymbol{\psi}^{(1)}\right)$ as the likelihood in a usual Bayesian formulation.

Step-2. Estimate the desired LR using:

$$\frac{L(\boldsymbol{\theta}^{(0)},\boldsymbol{\psi}^{(0)})}{L(\boldsymbol{\theta}^{(1)},\boldsymbol{\psi}^{(1)})} \approx \frac{1}{m}\sum_{j=1}^{m}\frac{f(\mathbf{y}|\mathbf{x}^{(j)},\boldsymbol{\psi}^{(0)})g(\mathbf{x}^{(j)}|\boldsymbol{\theta}^{(0)})}{f(\mathbf{y}|\mathbf{x}^{(j)},\boldsymbol{\psi}^{(1)})g(\mathbf{x}^{(j)}|\boldsymbol{\theta}^{(1)})}.$$

We emphasize that the same algorithm can be readily adopted to compute LRs between two without observation error models (4.5 a) when missing observation exist. Furthermore, as implemented in Ponciano et al. (2009), it can be extended to the calculation of PLs.

4.3.1 Comparing Without versus With Observation Error Models

The above algorithm is only applicable when both the competing models are hierarchical models. Here we derive a Monte Carlo estimator of the LR between a without observation error model (4.5 a) and a with observation error state-space model (4.5) when no missing values are present. We begin by noticing that

$$\begin{aligned} L(\boldsymbol{\theta}^{(1)},\boldsymbol{\psi}^{(1)};\mathbf{y}) &= E_g\left(f(\mathbf{y}|\mathbf{x},\boldsymbol{\psi}^{(1)})\right) \\ &= \int f(\mathbf{y}|\mathbf{x},\boldsymbol{\psi}^{(1)})g(\mathbf{x}|\boldsymbol{\theta}^{(1)})d\mathbf{x}. \end{aligned}$$

This leads to the following Monte Carlo estimator

$$L(\boldsymbol{\theta}^{(1)},\boldsymbol{\psi}^{(1)};\mathbf{y}) \approx \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}|\mathbf{x}^{(j)},\boldsymbol{\psi}^{(1)}), \tag{4.6}$$

where the random trajectories $\mathbf{x}^{(1)},\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}$ are generated from the process model $g(\mathbf{x}|\boldsymbol{\theta}^{(1)})$. Also, $L(\boldsymbol{\theta}^{(0)};\mathbf{y}) = g(\mathbf{y}|\boldsymbol{\theta}^{(0)})$, so the Monte Carlo estimate of the LR is given as

$$\frac{L(\boldsymbol{\theta}^{(0)};\mathbf{y})}{L(\boldsymbol{\theta}^{(1)},\boldsymbol{\psi}^{(1)};\mathbf{y})} \approx g(\mathbf{y}|\boldsymbol{\theta}^{(0)})\left[\frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}|\mathbf{x}^{(j)},\boldsymbol{\psi}^{(1)})\right]^{-1}.$$

4.3.2 Comparing Without versus With Observation Error Models in the Presence of Missing Data

We now obtain a Monte Carlo estimator of the LR between a without observation error model (4.5 a) and a with observation error state-space model (4.5) when abundance time series contain missing observations. The derivation of the LR involves separate Monte Carlo estimates of $L(\boldsymbol{\theta}^{(0)},\boldsymbol{\psi}^{(0)})$ and $L(\boldsymbol{\theta}^{(1)},\boldsymbol{\psi}^{(1)})$. The estimator of $L(\boldsymbol{\theta}^{(1)},\boldsymbol{\psi}^{(1)})$ is given by (4.6). The form of this estimator remains the same regardless of the presence or absence of missing abundances. To see this, when missing values exist, we simply set

$f\left(\mathbf{y}|\mathbf{x}^{(j)}, \boldsymbol{\psi}^{(1)}\right) = \prod_{i \in S} f(y_i|x_i^{(j)}, \boldsymbol{\psi}^{(1)})$, where S is an index set of those years for which populations counts are available.

Let us now derive the estimate of $L\left(\boldsymbol{\theta}^{(0)}; \mathbf{y}\right)$ when missing abundances exist. For expositional simplicity, we consider a hypothetical time series of length six with two missing observations. However, the estimation procedure holds in general with arbitrary number of missing values. Let the time series be $y_1, y_2, \dot{y}_3, y_4, \dot{y}_5, y_6$, where $\dot{y}_3$ and $\dot{y}_5$ denote missing log-abundances. Then, exploiting the Markovian structure of (4.5 a), we have

$$\begin{aligned} L\left(\boldsymbol{\theta}^{(0)}; \mathbf{y}\right) &= \iint g(y_1)g(y_2|y_1)g(\dot{y}_3|y_2)g(y_4|\dot{y}_3)g(\dot{y}_5|y_4)g(y_6|\dot{y}_5)d\dot{y}_3 d\dot{y}_5 \\ &= g(y_1)g(y_2|y_1)\{\int g(y_4|\dot{y}_3)g(\dot{y}_3|y_2)d\dot{y}_3\}\{\int g(y_6|\dot{y}_5)g(\dot{y}_5|y_4)d\dot{y}_5\} \\ &= g(y_1)g(y_2|y_1)E_{g(\dot{y}_3|y_2)}[g(y_4|\dot{y}_3)|Y_2 = y_2]E_{g(\dot{y}_5|y_4)}[g(y_6|\dot{y}_5)|Y_4 = y_4] \\ &\approx g(y_1)g(y_2|y_1)\left\{\frac{1}{m}\sum_{j=1}^{m} g\left(y_4\middle|\dot{y}_3^{(j)}\right)\right\}\left\{\frac{1}{m}\sum_{j=1}^{m} g\left(y_6\middle|\dot{y}_5^{(j)}\right)\right\} \end{aligned} \quad (4.7)$$

where the random numbers $\dot{y}_3^{(1)}, \dot{y}_3^{(2)}, \dots, \dot{y}_3^{(m)}$ and $\dot{y}_5^{(1)}, \dot{y}_5^{(2)}, \dots, \dot{y}_5^{(m)}$ are generated from the conditional distribution $g\left(\dot{y}_3, \dot{y}_5|y_2, y_4, \boldsymbol{\theta}^{(0)}\right)$ via a simple Monte Carlo algorithm. The estimate of the likelihood ratio $L(\boldsymbol{\theta}^{(0)}; \mathbf{y})/L(\boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)}; \mathbf{y})$ is then simply obtained by taking ratio of (4.7) to (4.6).

4.4 Estimation Error and Prediction of Future Trajectories

Let $\hat{\boldsymbol{\varphi}} = (\hat{\boldsymbol{\eta}}^T, \hat{\sigma}^2, \hat{\psi})^T$ denote the MLEs. In the following, we describe how to predict future population states given the model and the MLEs of the parameters. Different extinction metrics are functions of the predicted future states and are computed using the predicted future trajectories.

If the true parameter values are known, prediction of the future states of the process is based on the conditional distribution of the future states given the observed population abundances. Let $\boldsymbol{X} = \left(X_0, X_1, X_2, .., X_q\right)$ and $\boldsymbol{Y} = \left(Y_0, Y_1, Y_2, .., Y_q\right)$ represent vectors of *unobserved* and estimated log population abundances respectively and let $\boldsymbol{X}^{(f)} = \left(X_{q+1}, X_{q+2}, .., X_{q+f}\right)$ be that of future abundances. If the growth process is a first order Markov process, the conditional distribution of $(\boldsymbol{X}, \boldsymbol{X}^{(f)})$ is

$$h\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}\middle|\boldsymbol{Y}\right) = \frac{g(X_0; \boldsymbol{\eta}, \sigma^2)\prod_{t=1}^{q+f} g(X_t; m(X_{t-1}; \boldsymbol{\eta}), \sigma^2)\prod_{t=0}^{q} f(Y_t; X_t, \psi)}{c(\boldsymbol{Y})}$$

where $g(.;\mu,\sigma^2)$ indicates the density function for $N(\mu,\sigma^2)$ and $c(\boldsymbol{Y})$is the normalizing constant. Substitution of known parameter values by their estimates leads to prediction intervals that have lower than nominal coverage. To account for the uncertainty in the parameter estimates, we can integrate over the bootstrap distribution (Harris 1989; Dennis and Otten 2000; Sæther et al. 2000). Here, as illustrated in Section 3.3, we integrate over the asymptotic Normal distribution (Hamilton 1986) of the parameter estimates to predict random effects in general hierarchical models. Simulation results in Torabi and Shokoohi (2012) show that actual coverage of such prediction intervals is close to nominal in many cases. Thus, for predicting future states, we use the following prediction distribution:

$$h\left(\boldsymbol{X},\boldsymbol{X}^{(f)}\middle|\boldsymbol{Y}\right)=\frac{\int g(X_0;\boldsymbol{\eta},\sigma^2)\prod_{t=1}^{q+f} g(X_t;m(X_{t-1};\boldsymbol{\eta}),\sigma^2)\prod_{t=0}^{q} f(Y_t;X_t,\psi)\rho(\widehat{\boldsymbol{\varphi}})d\widehat{\boldsymbol{\varphi}}}{c(\boldsymbol{Y})} \tag{4.8}$$

where $\rho(\widehat{\boldsymbol{\varphi}})$ denotes the asymptotic normal distribution of the estimators $\widehat{\boldsymbol{\varphi}}=(\widehat{\boldsymbol{\eta}}^T,\hat{\sigma}^2,\hat{\psi})$ We prefer DC algorithm because it gives MLE and the associated variance covariance matrix simultaneously. Although other computational algorithms (e.g. deValpine and Hastings 2002) may also be used to get the MLE, they require additional computation to estimate the variance-covariance matrix of the MLE.

The MCMC algorithm can be used to generate random numbers from the distribution in (4.8) without evaluating the high-dimensional integral. The full description of the algorithm is available in Section 4.5.1. Notice that missing data pose no special difficulties in the state-space formulation (4.5). Let *S* be the index set of those years for which populations counts are available. Then, in (4.8) we simply replace $\prod_{i=1}^{q} f(Y_t;X_t,\psi)$ by $\prod_{i\in S} f(Y_t;X_t,\psi)$. Otherwise, everything else including the computational effort remains exactly the same.

The growth models considered in this chapter are first-order density-dependent models. However, some populations exhibit higher-order density regulation, that is, population growth at time *t* is a function of lagged abundances $N_{t-l},N_{t-l+1},\dots,N_{t-2}$, $l\geq 2$. For instance, population fluctuations in many insect populations are manifested through delayed density dependence (see, for instance, Bjørnstad et al. 1998; Turchin et al. 1999). As alluded to in Section 4.2, the methodology developed in this chapter can flexibly handle the higher-order density regulation models. For example, for the second order Gompertz delayed density dependence model

$$m(X_{t-1},X_{t-2})=X_{t-1}+a+b_1X_{t-1}+b_2X_{t-2},$$

in (4.8) we only need to replace the joint distribution of hidden states,

$$g(X_0;\boldsymbol{\eta},\sigma^2)\prod_{t=1}^{q+f} g(X_t;m(X_{t-1};\boldsymbol{\eta}),\sigma^2)$$

by

$$g(X_0;\boldsymbol{\eta},\sigma^2)g(X_1;m(X_0;\boldsymbol{\eta}),\sigma^2)\prod_{t=1}^{q+f} g(X_t;m(X_{t-1},X_{t-2};\boldsymbol{\eta}),\sigma^2).$$

Again, the rest of the algorithm remains exactly the same.

4.5 Estimation of Extinction Metrics

4.5.1 Generating Random Number from $\boldsymbol{h}(\boldsymbol{X},\boldsymbol{X}^{(f)}|\boldsymbol{Y})$

Recall that if $\rho(\widehat{\boldsymbol{\varphi}})$is the prior distribution of model parameters, $\boldsymbol{\varphi}=(\boldsymbol{\eta}^T,\sigma^2,\psi)^T$, the joint posterior distribution of *unknowns* $(\boldsymbol{X},\boldsymbol{X}^{(f)},\boldsymbol{\varphi})$ given the data is

$$h(\boldsymbol{X},\boldsymbol{X}^{(f)}|\boldsymbol{Y})=\frac{g(X_0;\boldsymbol{\eta},\sigma^2)\prod_{t=1}^{q+f} g(X_t;m(X_{t-1};\boldsymbol{\eta}),\sigma^2)\prod_{t=0}^{q} f(Y_t;X_t,\psi)\rho(\widehat{\boldsymbol{\varphi}})}{C(\boldsymbol{Y})}.$$

MCMC algorithms (Robert and Casella 2005) are computational tools that allow one to generate random number from the marginal posterior distribution $\pi(\boldsymbol{X}^{(f)}|\boldsymbol{Y})$using only the numerator of the above equation which involves no integration. These algorithms are implemented in freely available software packages such as WinBUGS (Spiegelhalter et al., 2004) and JAGS (Plummer, 2011a, 2011b). Let us denote the MCMC-generated random numbers by $(\boldsymbol{X},\boldsymbol{X}^{(f)},\boldsymbol{\varphi})_k$, k = 1, 2,…, J. The random numbers from the marginal posterior distribution $\pi(\boldsymbol{X}^{(f)}|\boldsymbol{Y})$ are obtained by simply discarding the $(\boldsymbol{X},\boldsymbol{\varphi})$component of the random numbers $(\boldsymbol{X},\boldsymbol{X}^{(f)},\boldsymbol{\varphi})_k$, leaving$(\boldsymbol{X}^{(f)})_k$, k = 1, 2,…, J. These random numbers, i.e. $(\boldsymbol{X}^{(f)})_k$, k = 1, 2,…, J, then comprise J simulated future trajectories of the population process.

4.5.2 Computation of PPIs

We first recall the definition of a PPI. Assuming $G_{(r)}$ be the distribution function of the random variable X_{q+r}, the future population abundance at time q+r. The lower $(1-\alpha)$ prediction interval for the future log-population abundance X_{q+r} is then given as $[X_\alpha^{(r)},\infty)$, where $X_\alpha^{(r)}=G_{(r)}^{-1}(\alpha)$. The estimates, $\hat{G}_{(r)}$, of the true distribution functions $G_{(r)}$, r = 1,

2, … *f*, are available from the *J* future population trajectories simulated above. The PPIs plotted in Figure 4.2 are then simply computed from the estimated distribution functions $\hat{G}_{(r)}$.

4.5.3 Computation of other Extinction Metrics

4.5.3.1 Without Observation Error Models

We start by simulating random numbers $\hat{\boldsymbol{\varphi}}_1, \hat{\boldsymbol{\varphi}}_2, \ldots, \hat{\boldsymbol{\varphi}}_B$, with *B* sufficiently large, from $\rho(\hat{\boldsymbol{\varphi}})$, the asymptotic normal distribution of the MLEs. Then the following steps yield the extinction estimates.

Step 1. Set $m = 1$.

Step 2. Generate random future trajectories $\left(\boldsymbol{X}^{(f)}\right)_{m,k}$, $k = 1, 2 \ldots J$, from the posterior distribution $\pi\left(\boldsymbol{X}^{(f)} | X_q, \hat{\boldsymbol{\varphi}}_m\right)$ using a straightforward MCMC algorithm. Notice that for the present case, we do not have observation error and we condition on the last observed population abundance X_q.

Step 3. Use the trajectories obtained in *Step 2* to compute point estimates of the extinction metrics defined in Section 4.1. This yields, for instance, $\hat{\pi}(n_e, t)$ as an estimate of $\pi(n_e, t)$, which can be simply computed as the proportion of trajectories that reach the quasi-extinction threshold, n_e. The remaining metrics can be computed similarly.

Step 4. Set next $m = m + 1$. Stop if $m > B$, else go to *Step-2*.

Step 5. The above procedure (*Step 1-4*) produces random numbers $\hat{\pi}^{(1)}(n_e, t)$, $\hat{\pi}^{(2)}(n_e, t)$, ... , $\hat{\pi}^{(B)}(n_e, t)$ that provide an estimate of the bootstrap distribution of $\hat{\pi}(n_e, t)$. The point estimate of $\pi(n_e, t)$ is then simply given by computing an appropriate central measure of this sampling distribution. The lower and upper 2.5 percentiles of this distribution provide a 95% confidence interval for $\pi(n_e, t)$. Point estimates and the associated confidence intervals for the remaining extinction metrics can be computed in a similar fashion.

4.5.3.2 With Observation Error State-Space Models

The algorithm steps are outlined as follows.

Step 1. Simulate B time series data $\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}, \ldots, \boldsymbol{Y}_{(B)}$ under the estimated state-space model, each starting at the first observed population abundance Y_0 and is of length equal to that of the observed abundance time series.

Step 2. Fit the state-space model to each of the simulated time series above using data cloning. This generates the bootstrap parameter estimates $\widehat{\boldsymbol{\varphi}}_1, \widehat{\boldsymbol{\varphi}}_2, \ldots, \widehat{\boldsymbol{\varphi}}_B$ and corresponding variance-covariance estimates $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\varphi}}_1}, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\varphi}}_2}, \ldots, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\varphi}}_B}$, which together yield the prior distributions $\rho(\widehat{\boldsymbol{\varphi}}_1), \rho(\widehat{\boldsymbol{\varphi}}_2), \ldots, \rho(\widehat{\boldsymbol{\varphi}}_B)$.

The following steps use data $[\boldsymbol{Y}_{(1)}, \rho(\widehat{\boldsymbol{\varphi}}_1)], [\boldsymbol{Y}_{(2)}, \rho(\widehat{\boldsymbol{\varphi}}_2)], \ldots, [\boldsymbol{Y}_{(B)}, \rho(\widehat{\boldsymbol{\varphi}}_B)]$ generated above to produce the extinction estimates.

Step 3. Set $m = 1$.

Step 4. Generate random future trajectories $\left(\boldsymbol{X}^{(f)}\right)_{m,k}$, $k = 1, 2\ldots J$, from the posterior distribution $\pi\left(\boldsymbol{X}^{(f)}|\boldsymbol{Y}_{(m)}\right)$ using the algorithm described in Section 4.5.1 where $\rho(\widehat{\boldsymbol{\varphi}})$ is set equal to $\rho(\widehat{\boldsymbol{\varphi}}_m)$.

Step 5. Use the trajectories obtained in *Step 4* to compute point estimates of the extinction metrics listed in Section 4.1. This is similar to *step 3* in Section 4.5.3.1 above.

Step 6. Set next $m = m + 1$. Stop if $m > B$, else go to *Step-3*.

Step 7. Same as *Step 5* in Section 4.5.3.1 above.

4.6 Effect of observation error on PVA

We now illustrate our methodology using time series of song sparrow (*Melospiza melodia*) population abundances on Mandarte Island, British Columbia, Canada (Figure 4.1). The data were collected during 1975-1998 and is reported in Sæther et al. (2000). Population size is the number of territorial females alive each spring (around 30 April). For detailed field methods and population biology of the species, see Smith (1988), Hochachka *et al.* (1989), Smith and Arcese (1989) and Arcese *et al.* (1992). The original analysis by Sæther et al. (2000) considered theta-logistic and logistic growth models (Gilpin & Ayala 1973; Morris and Doak 2002). Based on the estimated parameters, they chose the logistic growth model to describe the data. Throughout their analysis, they assumed that observation error was not present by arguing that virtually all birds were banded and low shrub vegetation on the island allowed them to be enumerated accurately. In our analysis, we test the assumption of no observation error based on Akaike information criterion (AIC) comparisons of the various model fits.

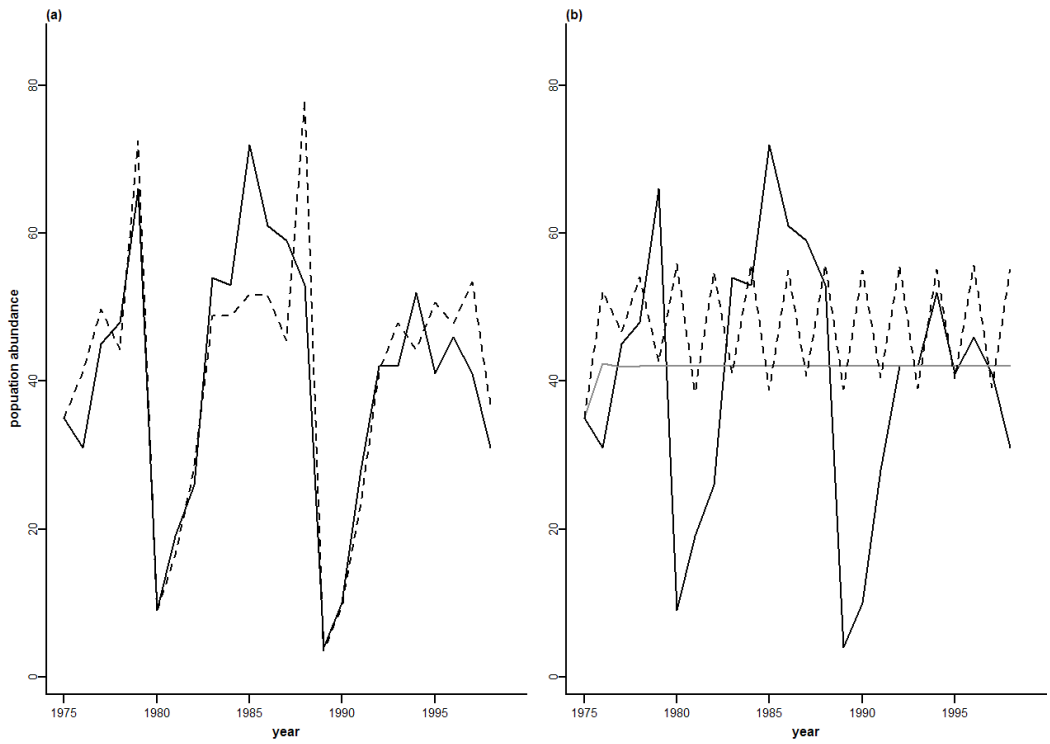

Figure 4.1. (a) Comparison of observed population counts (solid black line) of song sparrow from 1975-1998, on Mandarte Island, British Columbia, Canada with filtered population abundances, N_t, from the theta-logistic state-space model (dashed line) and, (b) smooth population trajectories obtained under the logistic model without observation error (solid gray line) and under theta-logistic model with observation error (dashed line). Theta-logistic model with observation error fits the data better than the logistic model without observation error.

Following Sæther et al. (2000), we also model song sparrow population data using the theta-logistic model. We first consider the model without observation error. The theta-logistic model (see equation 4.5 a) is defined by

$$m(X_{t-1}, \boldsymbol{\eta}) = X_{t-1} + r[1 - (N_{t-1}/K)^{\theta}],$$

where $\boldsymbol{\eta} = (r, K, \theta)^T$, r is the specific growth rate, K is the carrying capacity and θ represents the theta-logistic type of density dependence. The process variance is $\sigma^2 = \sigma_e^2 + \sigma_d^2/N_{t-1}$ where σ_e^2 and σ_d^2 denote the environmental and demographic variances respectively (Engen *et al*. 1998). Following Saether et al. (2000), instead of estimating demographic variance from the population data, we simply use the estimate of demographic variance ($\sigma_d^2 = 0.66$) reported in Sæther et al. (1998). This estimate is based on individual fluctuations in reproduction and survival of breeding females. Sæther

Table 4.1 Maximum likelihood estimates and standard errors of the model parameters. The likelihood function in each case is conditional on the first observed population count. The abbreviations *with* and *without* stand for with observation error model and without observation error model respectively.

Model	r	K	θ	σ	τ	ΔAIC_c
Theta-logistic (*with*)	0.4662 (0.1527)	51.1116 (3.8268)	5.3802 (2.3421)	0.1378 (0.1040)	0.2203 (0.0549)	0
Logistic (*without*)	1.1313 (0.3566)	41.5220 (5.0735)	-	0.6440 (0.1005)	-	28.31
Theta-logistic (*without*)	1.1826 (0.7673)	41.6922 (6.7235)	1.0286 (1.6660)	0.6446 (0.1003)	-	30.62
Logistic (*with*)	1.4274 (0.4376)	40.6571 (4.1627)	-	0.5080 (0.1648)	0.3732 (0.0939)	34.65

et al. (2000) assume that fluctuations in X_t are small and use least square method to estimate the parameters. Instead we compute maximum likelihood estimates using data cloning. The point estimates (Table 4.1) are comparable with those of Sæther et al. (2000). Under the no observation error model, estimate of the density regulation parameter θ is close to 1 suggesting that perhaps logistic model is sufficient to model the density dependent growth. We use AIC_c, the small sample bias-corrected version of AIC (Burnham and Anderson 2004), to compare logistic and theta-logistic models. The AIC_c difference (ΔAIC_c) between the theta-logistic and the logistic model is -2.307. Thus, assuming no observation error, the logistic growth model is sufficient to describe the song sparrow population process. This agrees with the conclusion in Sæther et al. (2000).

Next, instead of simply assuming no observation error, we test if it is present or not. We assume that observation errors are Lognormally distributed. That is, in (4.5), the observation error distribution $f(y_t, X_t, \psi)$ is Normal with mean zero and variance τ^2. The maximum likelihood estimates of the parameters $\boldsymbol{\varphi} = (\boldsymbol{\eta}^T, \sigma^2, \tau^2)^T$ are given in Table 4.1. Using the DCLR algorithm outlined in Section 4.3, we computed the ΔAIC_c values and ranked different models (Table 4.1). The ΔAIC_c values are calculated as the difference between the AICc value for a given model and the AIC_c value of the best-fitting model. The ΔAIC_c value for the best fitting model is, thus, 0. The theta-logistic model with observation error fits the data substantially better than both logistic model with observation error and the logistic model without observation error. Hence we conclude: (i) song sparrow population counts are subject to observation error, and, (ii) as compared to

the logistic growth function, theta-logistic provides a better functional form to describe the song sparrow population process.

Interestingly, had we fixed the logistic model as the correct functional form for population growth, we would have failed to detect the presence of observation error (Table 4.1) possibly leading to erroneous conclusions. This illustrates the importance of fitting and comparing various biologically plausible growth models both in the presence and absence of observation error. It has been pointed out to us (Mark Taper, personal communication) that for territorial species such as the song sparrows (Smith and Arcese 1989), the Hassel model (Hassell et al. 1976) or the generalized Beverton-Holt model (Smith and Slatkin 1973) might be biologically more plausible than the theta-logistic model. It will be worthwhile to try these additional models along with the meta-population approach.

The density regulation parameter θ, when observation error is taken into account (Table 4.1; Figure 4.6), is estimated to be much larger than 1 ($\hat{\theta} = 5.38$) indicating strong density regulation near the carrying capacity *K*. The estimated process variance under the theta-logistic model with observation error is smaller ($\hat{\sigma}^2 = 0.019$) than that for the logistic model without observation error ($\hat{\sigma}^2 = 0.441$). The large estimate of observation error variance ($\hat{\tau}^2 = 0.049$) under the theta-logistic model indicates that substantial component of variation in the population counts is probably due to observation error.

Dennis et al. (2006) discuss two methods for state prediction in state-space models: 1) A *filtered* value of the log-population abundance X_t defined as

$$E(X_t|Y_t = y_t, Y_{t-1} = y_{t-1}, .., Y_0 = y_0),$$

the mean of the log-population abundance given the previous and the current observations only and, 2) A *smoothed* value of X_t which is the mean value of X_t given *all* the observations $Y_0, Y_1, Y_2, .., Y_q$, including those that follow time *t*. The filtered and the smooth population trajectories for song sparrow data are shown in Figure 4.1 along with the actual observations. The observed data seem to fluctuate around a constant in a systematic fashion. The smoothed population trajectory predicted under no observation error logistic model is virtually constant (Figure 4.1-b) whereas the smoothed population trajectory of the best fitting theta-logistic model shows systematic cycles (Figure 4.1-b) and follows the observed fluctuations well. The filtered trajectory, N_t, from the theta-logistic state-space model also indicates that the model fits the data well (Figure 4.1-a).

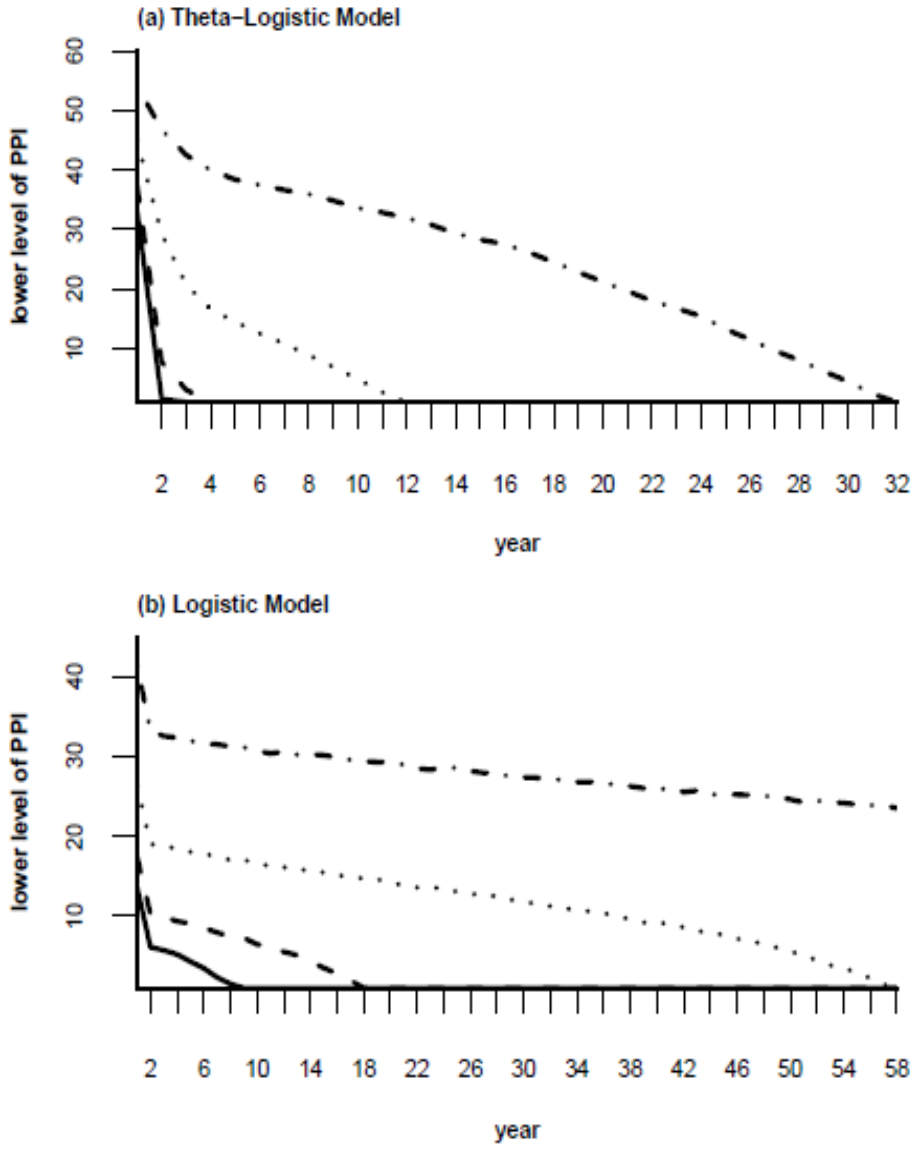

Figure 4.2 95% (solid line), 90% (dashed line), 75% (dotted line) and 50% (dotted-dashed line) lower bounds of prediction intervals for the future population abundance of song sparrows. PPI for theta-logistic model with observation error are mostly lower than for the logistic model without observation error.

Extinction properties under logistic and theta-logistic models are substantially different. Logistic model without observation error underestimates extinction risk substantially. Population prediction intervals (PPIs) for the future abundance of song sparrow population are wider under the theta-logistic model with observation error (Figure 4.2).

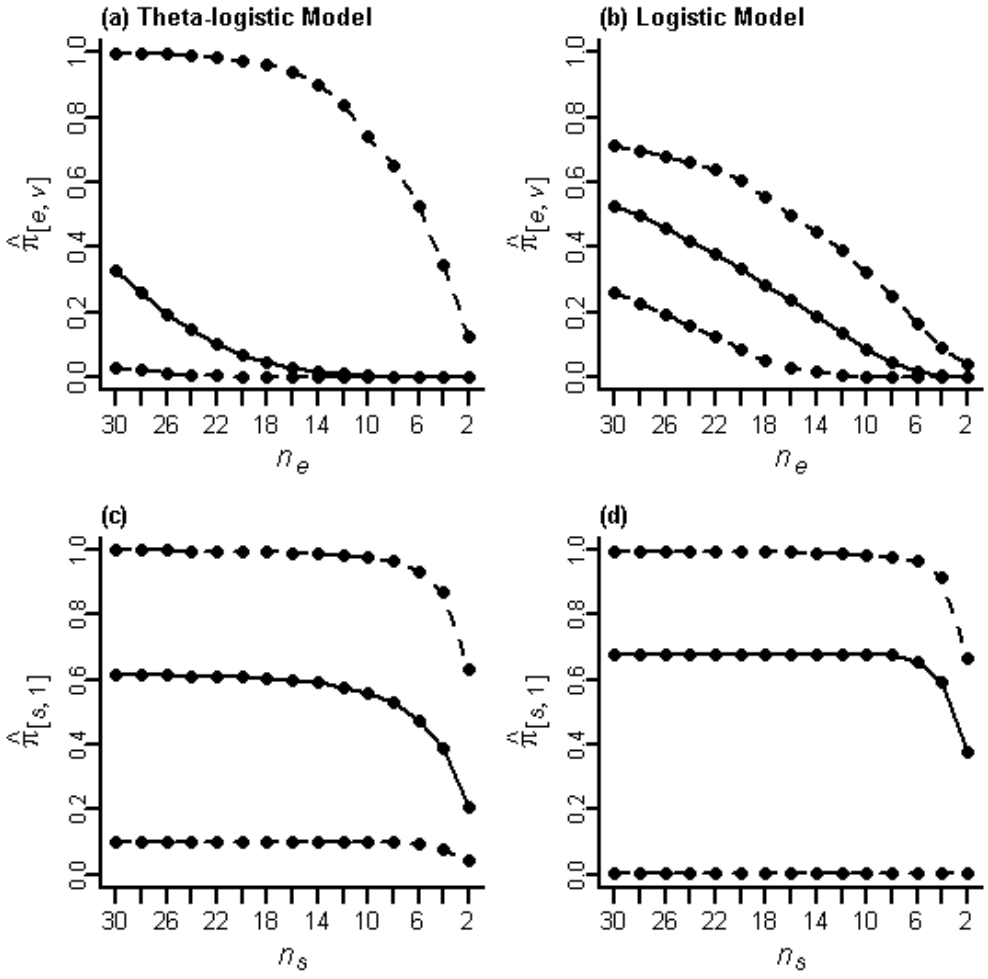

Figure 4.3 Profiles of probabilities of population going extinct before reaching a viable level, $\hat{\pi}_{[e,v]}$, and of probabilities of recovering from a lower threshold, $\hat{\pi}(s,1)$, under theta-logistic and logistic model for different thresholds along with 95% confidence intervals.

For instance, the 90% prediction interval obtained from the theta-logistic model, as compared to the one under the logistic model, predicts that the population will drop to much lower levels in the next fifteen years. In terms of time to extinction *T*, under the theta-logistic model, the song sparrow population is predicted to go extinct after just 4 years ($t_{0.1} = 4$) with probability 0.9 whereas under the logistic model it is after 18 years ($t_{0.1} = 18$). Estimates of $\pi_{[s,1]}$, the probability of recovering from a lower abundance level, are also different under these two models, especially at the lower warning threshold levels (Figure 4.3-c, d). The logistic model gives higher probabilities of surviving from

Table 4.2 Estimates of extinction metrics based on the predicted future trajectories. Numbers in parentheses are 95% confidence limits. Estimation is based on population forecast up to 100 time points into the future. The estimates correspond to $n_e = 3$, $n_v = 60$ and $n_s = 5$.

Extinction Metrics	Theta-logistic Model	Logistic Model
$\hat{\pi}(n_e,100)$	0.5769 (0.0034, 0.9563)	0.6466 (0.0556, .9996)
$\hat{\pi}_{[e,v]}$	0.0004 (0.00 , 0.2363)	0.0022 (0.00 , 0.0654)
$\hat{\pi}_{[v,e]}$	0.9988 (0.7516, 1.00)	0.9977 (0.9346, 1.00)
$\hat{\pi}_{[s,1]}$	0.4322 (0.0868, 0.9019)	0.6243 (0.0058, 0.9386)
$\hat{\hat{t}}$	36.3225 (14.9095, 51.5767)	43.4200 (12.1322, 51.6010)
$\hat{\hat{\xi}}_{(.5)}$	13.4200 (1.1925, 41.1521)	5.1902 (3.1533, 13.4767)

lower abundance levels. In contrast, the theta-logistic model predicts that the odds of song sparrow population recovering from a lower abundance level are relatively small.

At first, the extremely low estimates of $\pi_{[e,v]}$ under the theta-logistic model seem inconsistent with those of other extinction metrics (Figure 4.2-a, Figure 4.3-a). However, note that corresponding to the threshold abundance $n_e = 3$, estimates of the probability of quasi-extinction, $\pi(n_e,100)$, and $\pi_{[v,e]}$ are 0.577 and 0.999 respectively (Table 4.2). Thus, although the probability of quasi-extinction is large, the population is predicted to reach the carrying capacity before it crashes to extinction with high probability. This seemingly anomalous behavior is due to the large value of the density regulation parameter θ (Table 4.1; Figure 4.6) resulting in strong density regulation near the carrying capacity (Clark et al. 2010). The shape of the smoothed population trajectory (Figure 4.1-b) shows that the song sparrow population dynamics are intrinsically cyclical near its carrying capacity, $\hat{K} = 51.1$. Thus, when the population exceeds the carrying capacity, the strong regulatory mechanism kicks in and leads to a catastrophic decline of the population abundance. Uncertainty in the estimation of θ (Figure 4.6) further exacerbates this

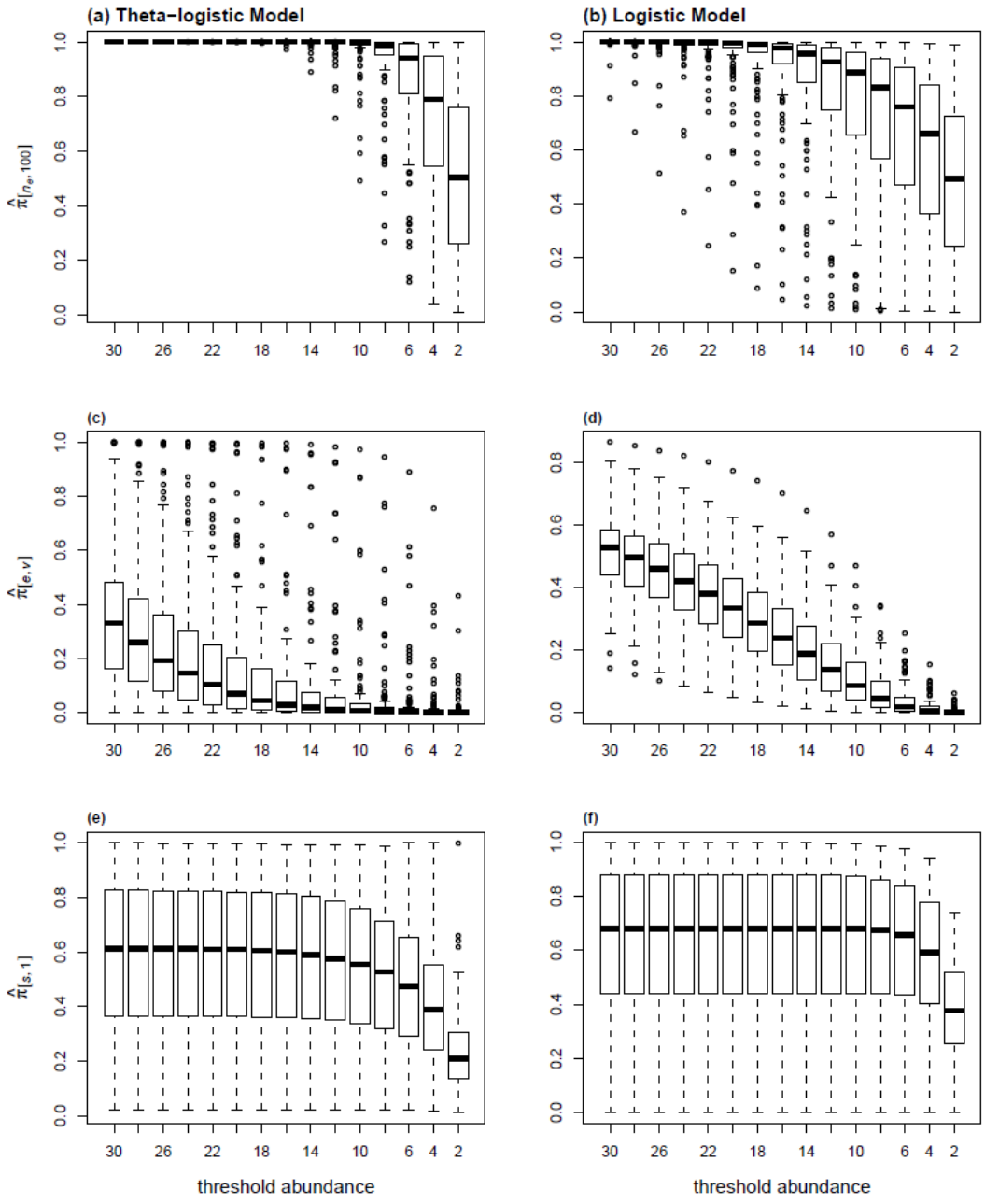

Figure 4.4 Approximate bootstrap distributions of the extinctions metrics. Figures (a), (c) and (e) in the left panel correspond to the theta-logistic model. Figures in the right panel correspond to the logistic model. The shape of these bootstrap distributions is also revealed by the corresponding extinction profiles plotted in Figure 4.5 and Figure 4.3.

phenomenon. On the other hand, the logistic model yields higher estimates of $\pi_{[e,v]}$ (Figure 4.3-b). Fixing the value of θ at 1 results in stable population dynamics.

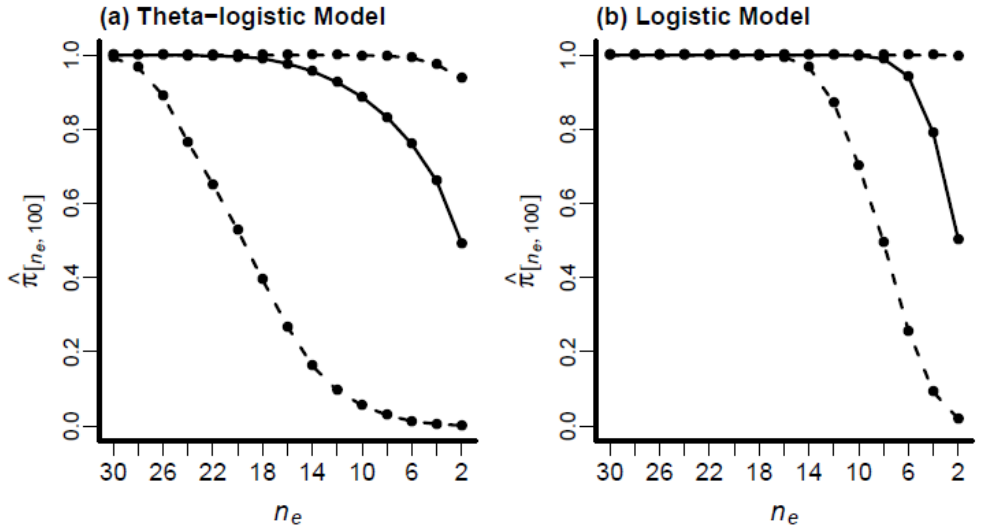

Figure 4.5 Profiles for probabilities of quasi-extinction, $\hat{\pi}(n_e,100)$, under theta-logistic and logistic model for different thresholds along with 95% confidence intervals.

The confidence bands shown in Figure 4.3 are based on the lower and upper 2.5 percentiles of the approximate bootstrap distribution of the corresponding estimated extinction metrics. In the case of $\hat{\pi}_{[s,1]}$ (Figure 4.3-c and d), sampling distribution has large variance resulting in wide confidence bands. This indicates that observed data lack information for the reliable estimation of $\pi_{[s,1]}$. However, these confidence intervals should be interpreted in conjunction with the shape of the corresponding sampling distributions. For instance, the distributions of $\hat{\pi}_{[s,1]}$ under the logistic model are left skewed (Figure 4.4-f). This shows that under the estimated model parameters, most of the future population trajectories that pass the warning thresholds would fail to recover to higher abundance levels. Despite the wide confidence bands, estimates of $\pi_{[s,1]}$ impart useful knowledge about the extinction risk. In comparison, the sampling distributions of $\hat{\pi}_{[s,1]}$ under the best fitting theta-logistic model are only slightly left skewed and therefore indicate a smaller chance of recovering from low abundance levels. Furthermore, confidence intervals for $\pi(n_e,100)$ are also wide (Table 4.2) but the corresponding sampling distributions are left skewed both under the theta-logistic model and the logistic model (Figure 4.4-a, b).

The warning threshold of five in Table 4.2 was chosen for two reasons: 1) Following the usual practice in PVA it is about 10% of the estimated carrying capacity and, 2) it is close to the smallest abundance level of four observed during the years, 1975-1998. We chose the quasi-extinction threshold, n_e, slightly smaller than n_s. Of course, one can plot $\pi(n_e,100)$ against various n_e values to get a better picture of the extinction risk. These plots for both logistic and theta-logistic models are shown in Figure 4.5.

We have implemented the R programs for model estimation, model selection and for the computation of extinction metrics in a user-friendly R package, PVAClone. The package is available to download from the packages section of the Comprehensive R Archive Network site (http://cran.r-project.org/).

4.6.1 Discussion

All population time series data contain observation error to some degree. It is well known that unaccounted for observation error leads to biased estimates of key model parameters (Freckleton et al. 2006; Barker and Sibly 2008). Barker and Sibly (2008) conducted a simulation study to investigate the effect of observation error in estimating the density regulation parameter theta of the theta-logistic model. Their results suggest that estimation of theta is subject to large bias especially when environmental perturbation is small. Our analysis of the song sparrow population counts also illustrates that incorporating observation error can result in substantially different estimates than when it is not incorporated. As we show for the song sparrow example, large changes in parameter estimates in turn lead to entirely different assessment of the extinction risk. These results highlight the fact that ignoring observation error could be potentially dangerous for conservation decisions. We therefore contend that the hypothesis of no observation error should always be rigorously tested against data.

Although the presence of density regulation parameter theta in the theta-logistic model provides a flexible description of density dependence, a recent simulation study by Clark et al. (2010) suggests that population abundance data generally lack information for reliable estimation of r and θ, especially when the observed population series is stationary, that is, fluctuating around its carrying capacity. This lack of information (Polansky et al. 2009, Clark et al. 2010), leads to multimodality and likelihood ridges in the likelihood surface where the best fitted growth rate curves are frequently biased toward concave fits ($\theta < 1$) but the likelihood ratio confidence regions include a wide range of models

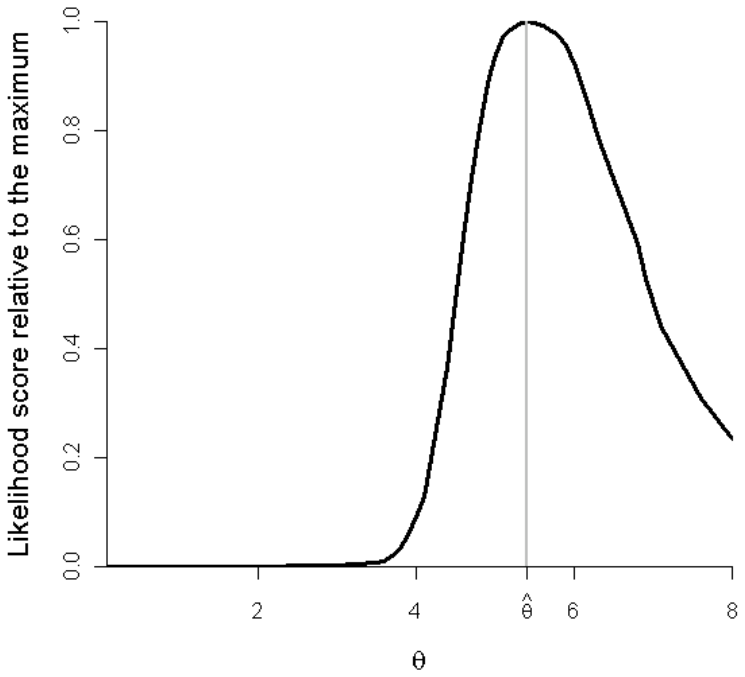

Figure 4.6 Profile likelihood (solid black line) for the density regulation parameter θ in the theta-logistic state-space model.

corresponding to more biologically plausible convex fits (θ >1). Clark et al. (2010), in light of these results, conclude that estimation of extinction risk from fitting theta-logistic model is prone to imprecision. However, their analysis also shows that for recovering populations, i.e. when the populations are fluctuating away from the stationary equilibrium, model fitting is not difficult. Barker and Sibly (2008) also observe the same phenomenon even when the observation error is present. However, we agree with the dictum that study of likelihood ridges and multimodality should be a part and parcel of any statistical analysis. To detect possible multimodalities, we plotted the profile likelihood for θ using data cloning (Ponciano et al. 2009). The profile likelihood (Figure 4.6) shows no signs of multimodality or ridges in the likelihood surface. In fact, the likelihood vanishes virtually to zero over the entire region of concavity of the growth rate curves (i.e. for θ < 1). Thus, large fluctuations in the song sparrow population away from its carrying capacity (Figure 4.1) have helped reduce uncertainty in estimating θ. If multimodality and likelihood ridges are present, bootstrap distribution, instead of the multivariate normal distribution used in this chapter, is probably a better way to incorporate uncertainty in parameter estimates in forecasting future states.

The extinction properties of the song sparrow population highlight a few important points. Traditionally, assessment of the extinction risk of a population is based on extinction metrics corresponding to a single quasi-extinction threshold (such as those given in Table 4.2). A single value for such a threshold is seldom available because it is determined by multitude of factors such as demographic variability and the genetic structure of the population (Morris and Doak 2002). It is therefore better to evaluate the extinction risk based on extinction metrics estimated for a range of quasi-extinction thresholds. The estimates for an extinction metric (such as $\pi_{[e,v]}$) and the associated confidence limits then can be plotted as a function of the threshold abundance levels (e.g. Figure 4.3-a,b). These plots, called *extinction profiles*, also reveal the shape of the distribution of extinction estimates (Figure 4.3, 4.5). This is a more useful measure of the precision of the estimate than the associated confidence interval alone.

Given several extinction metrics, which metrics should one use in practice? Perhaps probability of extinction, if known, best describes the extinction risk of a population. PPIs are a convenient alternative way of looking at the extinction risk of a population. Both these metrics, however, lack information about the ability of a population to recover from low abundance levels, as quantified by $\hat{\pi}_{[s,1]}$. This metric can be potentially useful for conservation planning as it can be used to rank populations in terms of their ability to recover from low abundance levels. We also notice that risk assessment based on a single extinction metric alone can be misleading. For instance, the estimates of $\pi_{[e,v]}$ for the song sparrow population under the theta-logistic model (Figure 4.3-a) seem to indicate that the population is highly likely to remain near the carrying capacity when, in fact, the probability of quasi-extinction, $\hat{\pi}(n_e, 100)$, is quite large (Table 4.2). We therefore recommend that the assessment of the extinction risk of a population should not be limited to a single extinction metric. Perhaps a better approach is to gain an overall picture of the extinction risk by obtaining extinction profiles corresponding to various extinction metrics.

The distribution of time to extinction is generally right skewed for most stochastic population growth models (see, Dennis et al. 1991 and Grimm and Wissel 2004). Median time to extinction is a better measure of a population's intrinsic ability to persist (Groom and Pascual 1998; Grimm and Wissel 2004) than mean time to extinction. The median extinction time (Table 4.2) is estimated to be higher (13.4 years) under the theta-logistic model than under the logistic model (5.2 years) . However, the population predic-

tion intervals (PPIs) indicate that extinction times are likely to be much shorter under the theta-logistic model (Figure 4.2). Following Sæther et al. (2000), we believe that PPIs are a better measure of extinction risk because they deal directly with extinction times. The conclusion based on the overall extinction profile is that extinction risk is predicted to be much greater under the theta-logistic model than the logistic model.

The song sparrow population on the Mandarte Island was, presumably, surveyed very accurately. Field methods indicate that virtually every adult bird was captured and identified with a combination of a numbered metal leg-band and plastic colored leg-bands (Arcese *et al.* 1992). This left us to wonder about the source of observation error detected in our analysis. Freckleton et al. (2006) discuss dispersal and other possible sources of observation error in census data. The average number of sparrows immigrating to Mandarte island during the study period was 1.6 whereas number of immigrants arriving after the 1980 and 1989 population crashes was 2 and 4 respectively (Smith et al. 2006). These immigrants were counted as part of the Mandarte island population and not accounted separately (Arcese et al. 1992), thus providing a possible source of observation error detected in the population counts. Arcese and Marr (2006) employed a balance equation model (Walters 1986) to study the effect of such immigration on the probability of extinction. Their results revealed that even a few immigrants at low population densities can result in demographic rescue of the population without altering the expected population abundance and provide a buffer against the effect of environmental stochasticity (see also, Stacey and Taper 1992). A better representation of the population process would, therefore, be achieved by explicitly incorporating a dispersal component into the theta-logistic model.

The song sparrow population dynamics provide a cautionary example of what might happen to PVA if one ignores key process components such as dispersal. We observe that contrary to a very high risk of local extinction, as predicted under the best fitting theta-logistic model (Figure 4.2-a), the song sparrow population at the Mandarte Island has not gone extinct after 1998 (Smith et al. 2006). Furthermore, despite the fact that the population did recover twice from very low abundances during the study period (Figure 4.1), the estimated recovery probabilities under the theta-logistic model are quite small (Figure 4.3-c). Clearly, the omission of dispersal process has substantially reduced the predictive power of the best fitting population growth model. Unfortunately we do not have enough information to decompose dispersal from observation error. Our conclusions are, therefore, two-fold: i) observation errors in PVA matter and ii) integrating these er-

rors in PVA is not always enough and can still lead to important biases in parameter estimates if other processes such as dispersal are ignored.

The song sparrow population at the Mandarte Island can be viewed as part of a large metapopulation consisting of other small neighbouring island populations (Smith et al. 1996). These small populations are constantly at risk of local extinction due to environmental variation. In fact two of the populations on smaller islands (within 7km of the Mandarte Island) did go extinct and remained uninhabited for the next two years (Smith et al. 1996). Considering that these islands share similar environmental conditions and that dispersal is an important factor in preventing local extinctions via demographic and genetic rescue (Arcese and Marr 2006), a more realistic approach is to quantify the extinction risk at the metapopulation scale.

4.7 Incorporating Environmental Covariates

Population abundance counts often accompany data on environmental time series processes such as climate conditions (Creel and Creel 2009; Luis et al. 2010; Hart and Gotelli 2011) or abundance of other species (Fryxell et al. 1998). Whenever available, these covariate processes should be added to the growth model to improve site and year-specific population forecasts (Dennis and Otten 2000). Let $\boldsymbol{w}_t$ be the *p*-dimensional vector representing the covariates processes. Then, conditional on the realized values of the covariates, the state-space model (4.5) takes the form

Process model: $$X_t = m(X_{t-1}; \boldsymbol{\eta}) + \boldsymbol{w}_t^T \boldsymbol{\beta} + E_t \quad (4.9\text{ a})$$

Observation model: $Y_t \sim f(y_t; X_t, \psi)$. (4.9 b)

where E_t is the normally distributed process noise as defined in (4.5). Also, the model parameter vector is given as $\boldsymbol{\varphi} = (\boldsymbol{\eta}^T, \boldsymbol{\beta}^T, \sigma^2, \psi)^T$.

Population viability analysis in the presence of covariates involves two additional steps: (i) testing the significance of the covariate effects, and, (ii) defining separate process models for those covariates whose effects are found significant. The second step is critical in generating future trajectories as covariate information is required to obtain one-step-ahead growth rate predictions. Significance of the covariate effects can be tested via LRT where the likelihood ratios can be computed using the algorithms outlined in Section 4.3.

We illustrate model estimation and model selection using population abundance estimates of San Joaquin kit foxes (*Vulpes macrotis mutica*) inhibiting the NPRS from

Table 4.3 Maximum likelihood estimates and standard errors of model parameters. The likelihood function in each case is conditional on the first observed population count.

Model	a	b	c	σ	τ	ΔAIC_c
H_3	-0.5605 (0.1461)	-0.0028 (0.0007)	0.0784 (0.0089)	0.0369 (0.0767)	0.2150 (0.0472)	0
H_1	0.7493 (0.3150)	-0.0047 (0.0019)	-	0.4352 (0.1181)	0.1321 (0.2198)	43.17
H_2	-0.8843 (0.1478)		0.0711 (0.0120)	0.0116 (0.0239)	0.2816 (0.0552)	418.74
H_0	0.0214 (0.0228)	-	-	0.02716 (0.0525)	0.5008 (0.1039)	∞

1983 to 1996. Earlier, ignoring the presence of observation error, Dennis and Otten (2000) conducted a PVA for the same population to illustrate their methodology of incorporating environmental covariates. They found that the annual growing season rainfall at lag two significantly improved the model fit. Because kit fox time series consisted of abundance estimates; we extend their approach by incorporating observation error using the state-space formulation (4.9). The process model in this case is the following augmented stochastic Ricker growth model: $X_t = X_{t-1} + a + bN_{t-1} + cR_{t-2} + E_t$, where R_t denotes the annual growing season rainfall (cm) recorded at time t. We assume that observation errors are Lognormally distributed. We emphasize that estimation of model parameters $\boldsymbol{\varphi} = (a, b, c, , \sigma^2, \tau^2)^T$ is conditional on the realized rainfall time series $\boldsymbol{R} = (R_{-2}, R_{-1}, .., R_{q-2})$.

Following Dennis and Otten (2000), we consider the following four biologically interesting hypotheses. $H_0: b = 0, c = 0$ (no density dependence, no rainfall effect); $H_1: b \neq 0, c = 0$ (density dependence, no rainfall effect); $H_2: b = 0, c \neq 0$ (no density dependence, rainfall effect); and, $H_3: b \neq 0, c \neq 0$ (density dependence, rainfall effect). These models are fitted separately to obtain AIC_c values for model comparison using the methodology outlined in Section 4.3. The resulting MLEs and the AIC_c values are reported in Table 4.3. It is clear that assuming the presence of observation error, the model defined under H_3 provides the most adequate description of the observed time series. Thus, both rainfall and population density are important in predicting future abundance of the San Joaquin kit fox population. We provide a comprehensive treatment of the hypothesis of no observation error elsewhere.

4.7.1 Prediction of Future Trajectories

Analogous to distribution (4.8), we now develop a prediction distribution to forecast future population trajectories while incorporating uncertainty in parameter estimation. For expositional simplicity, we assume that rainfall is the only covariate available in the context of previously analysed kit fox abundance time series. Since the growth rate function is now conditional on the rainfall, we need to define a separate model to account for the variability in future rainfall values. For instance, this could be an autoregressive time series model of order one (AR1). Let us denote the rainfall model as $k(\boldsymbol{R};\boldsymbol{\gamma})$. Also, recall that $\boldsymbol{X} = (X_0, X_1, X_2, .., X_q)$, $\boldsymbol{Y} = (Y_0, Y_1, Y_2, .., Y_q)$ and $\boldsymbol{X}^{(f)} = (X_{q+1}, X_{q+2}, .., X_{q+f})$ represent vectors of unobserved, estimated and future log population abundances respectively. We similarly define $\boldsymbol{R}^{(f)} = (R_{q-1}, R_q, R_{q+1}, .., R_{q+f-2})$ to be the vector of future rainfall values. Then, for a known parameter vector $\boldsymbol{\varphi} = (\boldsymbol{\eta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \sigma^2, \tau^2)^T$, the conditional prediction distribution for $(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)})$ is given as follows

$$\begin{aligned}\pi\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)} \middle| \boldsymbol{Y}, \boldsymbol{R}\right) &= \frac{\pi\left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)} | \boldsymbol{R}\right)}{c(\boldsymbol{Y};\boldsymbol{R})} \\ &= \frac{\pi\left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}^{(f)} | \boldsymbol{R}, \boldsymbol{R}^{(f)}\right) k\left(\boldsymbol{R}^{(f)} | \boldsymbol{R}\right)}{c(\boldsymbol{Y};\boldsymbol{R})} \\ &= \frac{f\left(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}, \boldsymbol{R}^{(f)}\right) g\left(\boldsymbol{X}^{(f)} | \boldsymbol{X}, \boldsymbol{R}, \boldsymbol{R}^{(f)}\right) g\left(\boldsymbol{X} | \boldsymbol{R}, \boldsymbol{R}^{(f)}\right) k\left(\boldsymbol{R}^{(f)} | \boldsymbol{R}\right)}{c(\boldsymbol{Y};\boldsymbol{R})},\end{aligned}$$

where $c(\boldsymbol{Y};\boldsymbol{R})$ is the normalizing constant. The densities appearing in the numerator above are evaluated at $\boldsymbol{\varphi}$. However, we have suppressed this dependence to simplify notation. We notice that the distribution of data vector $\boldsymbol{Y}$, conditional on the latent abundances $\boldsymbol{X}$, is independent of $(\boldsymbol{X}^{(f)}, \boldsymbol{R}, \boldsymbol{R}^{(f)})$. Furthermore, $g\left(\boldsymbol{X} \middle| \boldsymbol{R}, \boldsymbol{R}^{(f)}\right) = g(\boldsymbol{X}|\boldsymbol{R})$ and $g\left(\boldsymbol{X}^{(f)} \middle| \boldsymbol{X}, \boldsymbol{R}, \boldsymbol{R}^{(f)}\right) = g\left(\boldsymbol{X}^{(f)} \middle| \boldsymbol{X}, \boldsymbol{R}^{(f)}\right)$. So the above expression simplifies to

$$\pi\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)} \middle| \boldsymbol{Y}, \boldsymbol{R}\right) = \frac{f(\boldsymbol{Y}|\boldsymbol{X}) g(\boldsymbol{X}|\boldsymbol{R}) g\left(\boldsymbol{X}^{(f)} | \boldsymbol{X}, \boldsymbol{R}^{(f)}\right) k\left(\boldsymbol{R}^{(f)} | \boldsymbol{R}\right)}{c(\boldsymbol{Y};\boldsymbol{R})}. \tag{4.10}$$

In practice, one needs to estimate the model parameters $\boldsymbol{\varphi}$. Furthermore, as discussed in Section 4.4, we also need to incorporate the estimation uncertainty to ensure proper converge properties of the resulting prediction intervals. Again, we integrate over the asymptotic normal distribution $\rho(\widehat{\boldsymbol{\varphi}})$ of the MLE, $\widehat{\boldsymbol{\varphi}}$, to account for the uncertainty in parameter estimates. Therefore, the prediction distribution given by (4.10) becomes

$$\begin{aligned}\pi\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)} \middle| \boldsymbol{Y}, \boldsymbol{R}\right) &= \int \pi\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)}, \widehat{\boldsymbol{\varphi}} | \boldsymbol{Y}, \boldsymbol{R}\right) d\widehat{\boldsymbol{\varphi}} \\ &= \int \pi\left(\boldsymbol{X}, \boldsymbol{X}^{(f)} | \boldsymbol{Y}, \boldsymbol{R}, \boldsymbol{R}^{(f)}, \widehat{\boldsymbol{\varphi}}\right) k\left(\boldsymbol{R}^{(f)} | \boldsymbol{R}\right) \rho(\widehat{\boldsymbol{\varphi}}) d\widehat{\boldsymbol{\varphi}} \\ &= \int \frac{\pi\left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}^{(f)} | \boldsymbol{R}, \boldsymbol{R}^{(f)}, \widehat{\boldsymbol{\varphi}}\right) k\left(\boldsymbol{R}^{(f)} | \boldsymbol{R}\right) \rho(\widehat{\boldsymbol{\varphi}}) d\widehat{\boldsymbol{\varphi}}}{c(\boldsymbol{Y};\boldsymbol{R})}\end{aligned}$$

$$= \frac{\int f(\boldsymbol{Y}|\boldsymbol{X}) g(\boldsymbol{X}|\boldsymbol{R}) g(\boldsymbol{X}^{(f)}|\boldsymbol{X},\boldsymbol{R}^{(f)}) k(\boldsymbol{R}^{(f)}|\boldsymbol{R}) \rho(\hat{\boldsymbol{\varphi}}) d\boldsymbol{R}^{(f)} d\hat{\boldsymbol{\varphi}}}{c(\boldsymbol{Y};\boldsymbol{R})}$$

where the densities $f(.)$, $g(.)$, and $k(.)$ are all conditional on $\hat{\boldsymbol{\varphi}}$. Analogous to scheme outlined in Section 4.5.1, random numbers from this posterior distribution can be generated by adapting a straightforward MCMC algorithm. That is, we generate random numbers $\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)}, \boldsymbol{\varphi}\right)_k$, k = 1, 2,…, J, from the above conditional density $\pi\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)}, \boldsymbol{\varphi} | \boldsymbol{Y}, \boldsymbol{R}\right)$. The random numbers from the marginal posterior distribution $\boldsymbol{h}(\mathbf{X}^{(f)} \mid \mathbf{Y})$ are obtained by simply discarding the $\left(\boldsymbol{X}, \boldsymbol{R}^{(f)}, \boldsymbol{\varphi}\right)$ component of the random numbers $\left(\boldsymbol{X}, \boldsymbol{X}^{(f)}, \boldsymbol{R}^{(f)}, \boldsymbol{\varphi}\right)_k$, leaving $\left(\boldsymbol{X}^{(f)}\right)_k$, k = 1, 2,…, J. These random numbers, i.e. $\left(\boldsymbol{X}^{(f)}\right)_k$, k = 1, 2,…, J, then comprise J simulated future trajectories of the population process. We will provide elsewhere a detailed exposition of the methodology developed in the current section by revisiting the Dennis and Otten's (2000) PVA of the San Joaquin kit fox population.

4.8 Summary

State-space models provide a flexible framework to incorporate observation error when fitting stochastic population growth models. We demonstrated that DC is a powerful tool for computing MLEs of parameters in general state-space models with highly non-linear growth structure. In addition, the proposed estimation procedure elegantly handles the case when environmental covariates are available. We also showed that DC provides a unified computational framework for both model estimation and model selection. We also devised an efficient DC based algorithm to forecast future population trajectories while simultaneously accounting for observation error and estimation uncertainty. We illustrated the importance of incorporation of observation error in PVA by reanalyzing the population time series of song sparrow. Our analyses indicated that the extinction risks predicted by with and without observation error models are quite different. Further analysis of possible causes for observation error revealed that some component of the observation error might be due to unreported dispersal. A complete analysis of such data, thus, would require explicit spatial models and data on dispersal along with observation error. Our conclusions are, therefore, two-fold: 1) observation errors in PVA matter and 2) integrating these errors in PVA is not always enough and can still lead to important biases in parameter estimates if other processes such as dispersal are ignored.

Chapter 5

MCMC Convergence Assessment Using an Empirical Characteristic Function based Nonparametric Test[3]

Markov Chain Monte Carlo methods, such as the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) and the Gibbs sampler (Gelfand and Smith 1990; Geman and Genman 1984), are a set of algorithms designed to simulate random numbers from a target probability distribution. The underlying feature of these algorithms is to generate a Markov chain that *eventually* converges to the target distribution. Key applications of the MCMC methods arise in Bayesian inference where they are employed to obtain samples from intractable and, generally, multivariate posterior distributions to estimate quantities of interest such as posterior means and variances. The quality of these estimates, and hence the resulting inference, critically depends on convergence of the MCMC chain to the target posterior distribution. In most practical applications, some form of statistical diagnostics on the generated samples is used to assess convergence to a stationary distribution. The main objective of such diagnostic tools is, therefore, to determine the point at which the Markov chain has fully escaped its initial transient phase and has settled down to a steady-state behavior. The posterior quantities are computed

[3] A version of this chapter is in preparation for a peer reviewed publication. Nadeem K and Lele S R.

based on the samples obtained after discarding the random numbers generated during the *burn-in* period.

One class of approaches to assess convergence exploits the theoretical properties of the Markov transition kernel to determine the appropriate length of the burn-in period (Schervish and Carlin 1992; Rosenthal 1993; Polson 1996). However, this is mathematically laborious and problem-specific (Cowles and Carlin 1996). As a result, most of the existing methods take on an empirical approach: applying diagnostics directly on the output produced by the MCMC algorithms. Cowles and Carlin (1996) and Brooks and Roberts (1998) provide a detailed review of diagnostic methods associated with this latter approach. In this thesis, we consider only the empirical approach.

Most of the current diagnostic tools evaluate convergence of the Markov chain to a stationary distribution in terms of convergence of some functional of the chain. For example, the widely used method of Gelman and Rubin (1992) consists of generating m parallel chains to conduct a simple analysis of variance of some functional, $\theta(X_t)$, where X_t denotes the state of an arbitrary chain at time t. The method produces a variance ratio statistic of the form $\widehat{\mathrm{R}} = \left(\frac{\mathrm{d}+3}{\mathrm{d}+1}\right)\frac{\widehat{\mathrm{V}}}{\mathrm{W}}$ where $\hat{V}$ is an estimator of the variance of $\theta(X_t)$, σ^2, constructed as a weighted average of between and within chain variance estimators, while W is an estimator of σ^2that is based on the within chain variances. The term $\left(\frac{d+3}{d+1}\right)$ is a correction term associated with the approximations inherent in the estimator $\hat{V}$, which is distributed as χ^2_d. (see, for detail, Galman and Rubin (1992) and Brooks and Gelman (1998)). Gelman and Rubin (1992) conclude that convergence of each of the m chains is ensured when the value of $\hat{R}$ is close to 1. Brooks and Gelman (1998) further extend this method to monitor joint convergence of multidimensional MCMC chains. The other related variance ratio based methods include those from Brooks and Gelman (1998) and Brooks and Guidici (2000). Another example in a similar vein is that of the convergence diagnostic method by Giakoumatos et al. (1999). Their method employs subsampling methodology for time series (Politis and Romano 1994; Politis et al. 1997) to construct an estimator that is consistent for the functional θ of the stationary distribution of the Markov chain. Convergence is concluded if, as the sample size N increases, the range of the resulting $(1-\alpha)100\%$ confidence interval for θ drops at the expected rate $1/\sqrt{\mathrm{N}}$. The choice of the functional Giakoumatos et al. (1999) recommend is some large quantile (such as the 90$^{\text{th}}$ percentile) on the grounds that stabilization of the quantile estimates in the tail is a reliable indicator of convergence to the stationary distribution. For more ex-

amples of such functional based diagnostic methods, we refer the reader to Brooks and Roberts (1998) review paper.

Instead of an arbitrarily chosen functional, Robert et al. (1999) use a formal hypothesis test to compare distributions of MCMC output at two different time points. They use the two-sample Kolmogorov-Smirov test to detect convergence of univariate MCMC chains. For multidimensional chains, they compute p-value of the test for each parameter separately and use the minimum of the resulting p-values (henceforth named as min-p-value) to construct the *stopping rule*. Their diagnostic, however, has two major shortcomings. Firstly, as we show in Example1-2, min-p-value has significant downward bias in approximating the corresponding exact p-value. Thus, one may not be able to deduce convergence appropriately even when the sampler has fully converged. Secondly, because the diagnostic is based essentially on assessing convergence of the univariate marginals of the multidimensional chain, it does not take into account the multivariate features of the target distribution, such as the variance-covariance structure. In fact, convergence to the marginal distributions is only a necessary condition for convergence to the full target distribution. We rectify these limitations by replacing Robert et al.'s (1999) min-p-value based test with our new nonparametric test for comparing multivariate distributions.

The difficulty in using classical procedures based on empirical distribution function or rank statistics for multivariate observations has led researchers to using empirical characteristic functions (ECF). For recent examples of ECF based test procedures, see Fan (1997), Alba-Fernandéz et al.(2006), Hušková and Meintanis (2008) and references therein. For examples of tests based on other nonparametric techniques, see Li and Liu (2004), Székely and Rizzo (2004) and Liu and Modarres (2011). Here, we introduce a new ECF based nonparametric test to compare several multivariate distributions. We show that the new test, henceforth called the ECF test, has excellent power as compared to other tests for comparing multivariate distributions. It is particularly useful for comparing multimodal multivariate distributions that arise in many mixture models as well as in many practical situations where MCMC approaches are used. The test statistic is easy to compute and its null distribution is obtained easily using central limit theorem arguments. The main motivation of this chapter is to develop convergence diagnostics for MCMC algorithms when the target distribution is multimodal and multivariate. As we show, the ECF test is powerful in detecting differences in scale, skewness and multimodality, allowing the diagnostic to pronounce stationarity only when all the key features of the mul-

ti-dimensional distribution have converged. Hence, we evaluate the performance of our proposed diagnostic test by assessing the joint convergence of MCMC draws to known multivariate normal target distributions. We also illustrate the application of the diagnostic using an interesting MCMC based statistical modeling problem.

5.1 Comparing Multivariate Populations using ECF

In this section we present the construction of ECF test statistic for comparing two multivariate populations. The section also includes a simulation study to assess the level and power of the proposed test. The complete derivation of the test statistic and an approximation to its sampling distribution for the multivariate *k*-sample testing problem are given in Appendix A.

We first state the hypothesis of interest and its representation in terms of characteristic functions. Let $\mathrm{F}(.)$ denote the cumulative distribution function (CDF) of an absolutely continuous *p*-dimensional random vector X. We are interested in testing the equality of *k* distributions $\mathrm{F}^{(\mathrm{j})}(.)$, $j = 1, 2, \ldots, k$, i.e. we wish to test the following hypothesis:

$\mathrm{H_o}$: $\mathrm{F}^{(1)}(.) = \mathrm{F}^{(2)}(.) = \cdots = \mathrm{F}^{(\mathrm{k})}(.) = (\text{say } \mathrm{F}(.))$ vs.

H_a: $F^{(j)}(.) \neq F^{(l)}(.)$, for some $j \neq l$. (5.1)

Let $\varphi^{(j)}(\boldsymbol{t})$, $\boldsymbol{t} \in \mathbb{R}^p$, represent the characteristic function (CF) corresponding to $F^{(j)}(.)$, i.e.

$$\varphi^{(j)}(\boldsymbol{t}) = E\left[e^{i\boldsymbol{t}'\mathrm{X}^{(j)}}\right] = \int e^{i\boldsymbol{t}'\mathrm{X}^{(j)}} d\mathrm{X}^{(j)},\ i = \sqrt{-1},$$

then using the bijection between CDFs and CFs, the above null hypothesis can be equivalently written as

H_o: $\varphi^{(1)}(\boldsymbol{t}) = \varphi^{(2)}(\boldsymbol{t}) = \cdots = \varphi^{(k)}(\boldsymbol{t}) = \varphi(\boldsymbol{t}),\ \forall\ \boldsymbol{t} \in \mathbb{R}^p.$ (5.2)

Since $\varphi^{(j)}(\boldsymbol{t})$ is a complex function, using the Euler's formula $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, $\theta \in \mathbb{R}$, we can write $\varphi^{(j)}(\boldsymbol{t})$ in terms of its real and imaginary parts as follows:

$\varphi^{(j)}(\boldsymbol{t}) = E\left[\cos(\boldsymbol{t}'\mathrm{X}^{(j)})\right] + iE\left[\sin(\boldsymbol{t}'\mathrm{X}^{(j)})\right] \equiv \mu_{C_t}^{(j)} + i\mu_{S_t}^{(j)}.$

Because

$$\varphi^{(j)}(\boldsymbol{t}) = \varphi^{(l)}(\boldsymbol{t}), \Leftrightarrow\ \mu_{C_t}^{(j)} = \mu_{C_t}^{(l)}\ \text{ and }\ \mu_{S_t}^{(j)} = \mu_{S_t}^{(l)}, \forall\ \boldsymbol{t} \in \mathbb{R}^p, \quad (5.3)$$

this allows us to write H_o in terms of real and imaginary parts of $\varphi^{(j)}(\boldsymbol{t})$, $j = 1, 2, \ldots, k$. For this, let $\mathbf{B}_{-k}$ be a $k \times k$ contrast matrix with elements $b_{jj} = -(k-1)$ and $b_{jl} = 1$, $\forall\ j \neq l$. Also, define $\boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} = \mathbf{B}_{-k}\boldsymbol{\mu}_{C_t}$ and $\boldsymbol{\delta}_{\boldsymbol{t}}^{(S)} = \mathbf{B}_{-k}\boldsymbol{\mu}_{S_t}$ be vectors of linear contrasts,

where $\boldsymbol{\mu}_{C_t}$ and $\boldsymbol{\mu}_{S_t}$ notate the vectors of real and imaginary parts, respectively, of the *k* CFs. The following lemma, whose proof is given in Appendix A, represents H_o in terms of these contrast vectors.

Lemma A1: The null hypothesis defined in (5.1) and (5.2) can be equivalently stated as

$$H_o: \boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} = \boldsymbol{\delta}_{\boldsymbol{t}}^{(S)} = \mathbf{0}, \forall\, \boldsymbol{t} \in \mathbb{R}^p. \tag{5.4}$$

As in this section we are interested in the case with *k = 2* only, the above equation simplifies to

$$H_o: \mu_{C_t}^{(1)} - \mu_{C_t}^{(2)} = \mu_S^{(1)} - \mu_{S_t}^{(2)} = 0, \forall\, \boldsymbol{t} \in \mathbb{R}^p. \tag{5.5}$$

5.1.1 Empirical Characteristic Function

The ECF, defined as follows, is a consistent estimator of the characteristic function of a random vector **X**. Suppose we have a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ available from a *p*-dimensional distribution $F(.)$, then the ECF corresponding to $\varphi(\boldsymbol{t})$, the characteristic function of **X**, is defined as

$$\varphi_n(\boldsymbol{t}) = \frac{1}{n}\sum_{h=1}^{n} e^{i\boldsymbol{t}'\mathrm{X}_h} = \frac{1}{n}\sum_{h=1}^{n}\cos(\boldsymbol{t}'\mathrm{X}_h) + i\frac{1}{n}\sum_{h=1}^{n}\sin(\boldsymbol{t}'\mathrm{X}_h) \equiv \hat{\mu}_{C_t} + i\hat{\mu}_{S_t},$$

where $\hat{\mu}_{C_t}$ and $\hat{\mu}_{S_t}$ are consistent estimators of corresponding population quantities μ_{C_t} and μ_{S_t}, respectively. This leads us to define consistent estimators of the contrasts appearing in (5.5). Let $\beta_{\boldsymbol{t}} = \mu_{C_t}^{(1)} - \mu_{C_t}^{(2)}$ and $\gamma_{\boldsymbol{t}} = \mu_{S_t}^{(1)} - \mu_{S_t}^{(2)}$, then consistent estimators of $\beta_{\boldsymbol{t}}$ and $\gamma_{\boldsymbol{t}}$ are respectively given as $\hat{\beta}_{\boldsymbol{t}} = \hat{\mu}_{C_t}^{(1)} - \hat{\mu}_{C_t}^{(2)}$ and $\hat{\gamma}_{\boldsymbol{t}} = \hat{\mu}_{S_t}^{(1)} - \hat{\mu}_{S_t}^{(2)}$. The following results, which we prove in Appendix A for general *k*, state the asymptotic distributions of the statistics $\hat{\beta}_{\boldsymbol{t}}$ and $\hat{\gamma}_{\boldsymbol{t}}$.

Theorem A2: Assuming that H_o is true, for any $\boldsymbol{t} \neq \mathbf{0}$, as $n \to \infty$

$$\sqrt{n}\,\hat{\beta}_{\boldsymbol{t}} \xrightarrow{D} N\left(0, 2\sigma_{C_t}^2\right)$$

and

$$\sqrt{n}\hat{\gamma}_{\boldsymbol{t}} \xrightarrow{D} N\left(0, 2\sigma_{S_t}^2\right),$$

where $\sigma_{C_t}^2$ and $\sigma_{S_t}^2$ are variances related to the real and imaginary parts of the ECF respectively.

Theorem A3: Regardless of the truth or falsity of $\mathrm{H_o}$, we have

$$\lim_{\|\boldsymbol{t}\|\to\infty} \lim_{n\to\infty} \sqrt{n}\,\hat{\beta}_{\boldsymbol{t}} \xrightarrow{D} N(0,1)$$

and

$$\lim_{\|\boldsymbol{t}\|\to\infty}\lim_{n\to\infty}\sqrt{n}\hat{\gamma}_{\boldsymbol{t}} \xrightarrow{D} N(0,1).$$

where $\|\,.\,\|$ denotes the Euclidean norm of a vector.

5.1.2 The ECF Test Statistic

In order to motivate the construction of the test statistic, we first elucidate the asymptotic behavior of the statistics $\hat{\beta}_{\boldsymbol{t}}$ and $\hat{\gamma}_{\boldsymbol{t}}$ as described in the theorems stated above. When H_o is true, the means are exactly equal to zero for all values of $\boldsymbol{t}$. However, when H_o is false, the means oscillate away from zero until $\|\boldsymbol{t}\|$ becomes sufficiently large. In fact, as we point out in Appendix A, the means $\beta_{\boldsymbol{t}}$ and $\gamma_{\boldsymbol{t}}$ converge to zero exponentially fast as $\|\boldsymbol{t}\| \to \infty$. Thus, the asymptotic distribution of the statistics $\sqrt{n}\hat{\beta}_{\boldsymbol{t}}$ and $\sqrt{n}\hat{\gamma}_{\boldsymbol{t}}$ are centered away from zero only for small values of $\|\boldsymbol{t}\|$. This implies that, for small $\|\boldsymbol{t}\|$, we expect the observed absolute values of $\sqrt{n}\hat{\beta}_{\boldsymbol{t}}$ and $\sqrt{n}\hat{\gamma}_{\boldsymbol{t}}$ to be significantly larger than what are expected under the standard normal distribution. Our Remark-2 in Appendix A further explains this point.

Furthermore, the null hypothesis defined in (5.5) is essentially a union of hypotheses indexed by $\boldsymbol{t} \in \mathbb{R}^p$. Thus, the rejection of any one of these hypotheses leads to the rejection of H_o. In fact, as we show in Appendix A for general *k*, a necessary and sufficient condition for the falsity of H_o is stated as follows.

Theorem A4: $\mathrm{H_o}$, as stated in (5.5), is false if and only if $\max_{\mathrm{t}\in\mathrm{R^p}}\{|\beta_{\mathrm{t}}|\,,\ |\gamma_{\mathrm{t}}|\} > 0$.

Therefore, the above arguments suggest constructing a *union-intersection* type test in the following way.

Let $\mathbf{G}$ be a grid containing $\mathcal{L}$ vectors in $\mathbb{R}^p$. Construction of a suitable grid is explained in Appendix A. Notating the vectors in $\mathbf{G}$ as $\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_{\mathcal{L}}$, we obtain a collection of random variables $\left\{\left(\hat{\beta}_{\boldsymbol{t}_i}\,,\ \hat{\gamma}_{\boldsymbol{t}_i}\right), i = 1, 2, \dots, \mathcal{L}\right\}$ all marginally asymptotically distributed according to Theorems A2 and A3. As we show in the appendix, these $2\mathcal{L}$ random variables are asymptotically correlated; however, in order to approximate the sampling distribution of our test statistic, we proceed as if they are mutually asymptotically independent. Our simulations show that this simplification still results in a level $-\,\alpha$ test.

Let $\left(\mathbf{X}_j^{(1)}, \mathbf{X}_j^{(2)}, j = 1,2, \dots, n\right)$ denote n independent observations from populations $F_{X^{(1)}}$ and $F_{X^{(2)}}$ respectively. Here, for simplicity of exposition, we assume equal sample sizes $n_1 = n_2 = n$. The test is generally applicable for different sample sizes. The ECF test statistic is then defined as

$$T_n = \sqrt{n}[\max_{\boldsymbol{t}\in\mathbf{G}}\{|\hat{\alpha}_{\boldsymbol{t}}|\,,\ |\hat{\beta}_{\boldsymbol{t}}|\}],$$

with the asymptotic level $-\,\alpha$ rejection region given as

$$\mathrm{RJ}_{\alpha} = \left\{\left(\mathrm{X}_{\mathrm{j}}^{(1)}, \mathrm{X}_{\mathrm{j}}^{(2)}, \mathrm{j} = 1,2,\ldots,\mathrm{n}\right)\colon \mathrm{T_n} > \left|\mathrm{h}(\alpha, \tilde{\mathrm{L}})\right|\right\},$$

where $\tilde{\mathcal{L}} = 2\mathcal{L}$, α is the probability of observing at least one random number greater, in absolute value, than $\left|h(\alpha, \tilde{\mathcal{L}})\right|$ out of $\tilde{\mathcal{L}}$ numbers generated randomly under $N(0,1)$, and $-h\left(\alpha, \widetilde{\boldsymbol{\mathcal{L}}}\right) = \Phi^{-1}(\varphi)$, where $\Phi(.)$ is the standard normal distribution function and φ depends on both α and $\tilde{\mathcal{L}}$. Using the definition of α above, we can obtain the *critical value* $\left|h(\alpha, \widetilde{\boldsymbol{\mathcal{L}}})\right|$ as follows. We notice that $\alpha = 1 - (1 - 2\varphi)^{\tilde{\mathcal{L}}}$, which yields $\varphi = \frac{1-(1-\alpha)^{\frac{1}{\tilde{\mathcal{L}}}}}{2}$. So we can compute φ for a given α and $\tilde{\mathcal{L}}$, and then $h\left(\alpha, \tilde{\mathcal{L}}\right) = -\Phi^{-1}(\varphi)$. Alternatively, the *p-value* of the test can be computed as $p - value = 1 - [1 - 2\Phi(T_n)]^{\tilde{\mathcal{L}}}$.

5.1.3 Simulation Study

We consider testing the equality of two trimodal distributions; each a mixture of three trivariate normal distributions, using our ECF test. We compare the performance of the ECF test, with another multivariate *k*-sample test introduced by Székely and Rizzo (2004). Their test statistic is based on energy distances (or *e*-distances) as defined by Székely and Móri (2001). We refer the reader to Székely and Rizzo (2004) for further details. Henceforth, we abbreviate their test as the ED test. The reason we preferred this test for comparison is its implementation in a user friendly software package, *energy*, available from the Comprehensive R Archive Network site (< http://cran.r-project.org/>).

The null distribution we consider is defined as $F_{X^{(1)}} = \sum_{j=1}^{3} p_{1j}\, N(\boldsymbol{\mu}_j\,, \boldsymbol{\Sigma}_{\mathrm{j}})$, where the mixing probabilities p_{1j} and the parameters of the component trivariate normal distributions are

$$\boldsymbol{\mu}_1 = (0,0,0)',\ \boldsymbol{\mu}_2 = (7,7,7)', \boldsymbol{\mu}_3 = (16,16,16)'\,, p_{11} = 0.15, p_{12} = 0.30, p_{13} = 0.55,$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.4 & 0.5 \\ 0.4 & 1 & 0.6 \\ 0.5 & 0.6 & 1 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0.4 & 0.5 \\ 0.4 & 4 & 0.6 \\ 0.5 & 0.6 & 4 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_3 = \begin{bmatrix} 16 & 0.4 & 0.5 \\ 0.4 & 16 & 0.6 \\ 0.5 & 0.6 & 16 \end{bmatrix}.$$

We consider the following three alternative mixture distributions that differ from the null distribution either in terms of a shift in the location or shape of the local modes.

Location Shift (LC): $F_{X^{(2)}} = p_{11}N(\boldsymbol{\mu}_1\,,\ \boldsymbol{\Sigma}_1) + \ p_{12}N(\widetilde{\boldsymbol{\mu}}_2\,,\ \boldsymbol{\Sigma}_2) + p_{13}N(\boldsymbol{\mu}_3\,,\ \boldsymbol{\Sigma}_3)$

Scale Shift-1 (SC1): $F_{X^{(2)}} = \sum_{j=1}^{3} p_{2j}\, N(\boldsymbol{\mu}_j\,, \boldsymbol{\Sigma}_{\mathrm{j}})$

Scale Shift-2 (SC2): $F_{X^{(2)}} = p_{11}N(\boldsymbol{\mu}_1\,, \boldsymbol{\Sigma}_1) + \ p_{12}N\left(\boldsymbol{\mu}_2\,, \widetilde{\boldsymbol{\Sigma}}_2\right) + p_{13}N(\boldsymbol{\mu}_3\,, \boldsymbol{\Sigma}_3)$.

The parameters in the alternative models that differ from those in the null are given as

$$\tilde{\boldsymbol{\mu}}_2 = (-1,-1,-1)', p_{21} = 0, p_{22} = 0.50, p_{23} = 0.50,$$

and

$$\tilde{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 16 & -0.3 & -0.1 \\ -0.3 & 16 & -0.9 \\ -0.1 & -0.9 & 16 \end{bmatrix}.$$

Figure 5.1 summarizes the results of a Monte Carlo study conducted at a nominal level $\alpha = 0.05$ to assess the power of the ECF test against the alternatives described above. The percentage of rejection under the null hypothesis is generally less than the nominal level of 5% (Figure 5.1-a), i.e. it is a conservative test, showing that our approximation of the sampling distribution of T_n results in a test with smaller than the nominal level α. The ED test is also seen to have the correct level. The ED test clearly has smaller power as compared to the ECF test in rejecting the location and scale alternatives LC and SC1 (Figure 5.1-b,c). The ECF test also appears to be far more sensitive than the ED test in detecting the scale shift alternative SC2. The rejection percentage under ECF exceeds 80% at n = 300 whereas the ED test only has about 15% rejection rate at that sample size (Figure 5.1-d). Surprisingly, the power of the ED test remains very low in this case even at n = 500. This shows that our test is especially powerful in detecting shifts other than the location, such as shifts in scale, skewness and multimodality, which are generally difficult to detect in multivariate distributions. As we demonstrate in the next section, this feature of the ECF test is very promising in MCMC convergence diagnostics because the multivariate posterior distributions tend be multimodal in complex statistical modeling problems.

5.2 The ECF based MCMC Diagnostics

We now discuss the use of ECF test to assess convergence of an MCMC algorithm. We demonstrate the performance of the ECF diagnostic using Metropolis-Hastings MCMC output to sample from two known multivariate normal target distributions. We also apply the diagnostic to MCMC output for quadratic regression with errors in variables model. We compare the burn-in times decided under various diagnostic criteria.

The basic rationale of our diagnostic is as follows. Suppose we run r parallel MCMC chains with the *initial* multivariate output from the i^{th} chain denoted as ($\boldsymbol{X}_{i,1}$, $\boldsymbol{X}_{i,2}, \ldots, \boldsymbol{X}_{i,N}$), where $\boldsymbol{X}_{.,j}$ is a p-dimensional observation. We then split the whole initial output into k successive samples each containing s observations from i^{th} chain so that

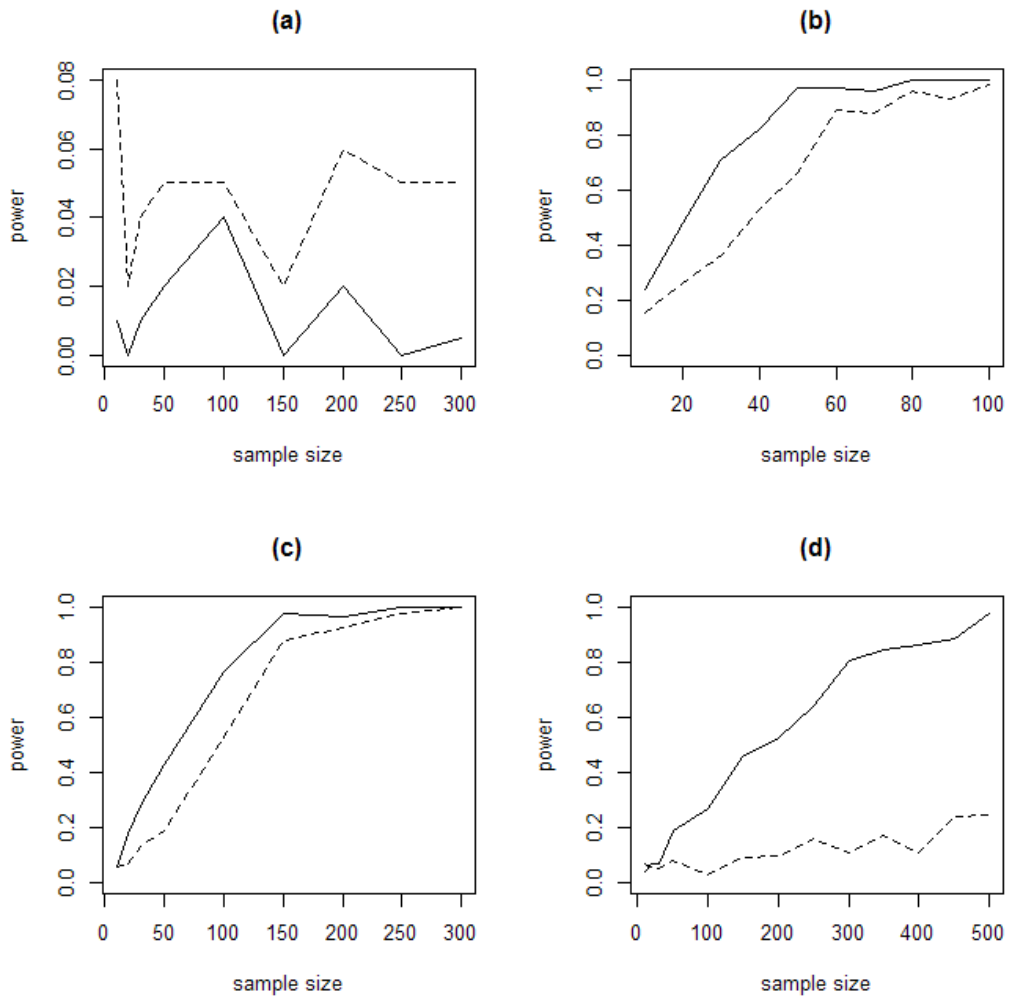

Figure 5.1 Rejection rate comparisons for the ECF (solid line) and ED (dashed line) tests under (a) the null distribution, (b) the location shift (LC), (c) the scale shift-1 (SC1), and (d) the scale shift-1 (SC2) alternatives.

$N = sk$. The number of observations in each sample is, therefore, $n = rs$. If the MCMC sampler has converged to the stationary distribution, all the k samples are drawn from that same distribution. Otherwise, if the sampler is still in the transient phase, at least one of these samples is observed from a different distribution. In order to assess if these k samples are realized from some common stationary distribution $F(.)$, our diagnostic employs the ECF test to formally test the convergence hypothesis:

$$H_o: F^{(1)}(.) = F^{(2)}(.) = \cdots = F^{(k)}(.). \tag{5.6}$$

The idea of applying a statistical test to consecutive *batches* of MCMC output as a measure of stationarity is not new. For one-dimensional case, Robert et al. (1999) used the Kolmogorov-Smirnov test to test the convergence hypothesis (5.6) for two halves (k=2) of a given MCMC output. For multi-dimensional chains, they proposed applying the Kolmogorov-Smirnov test to marginal components and using the minimum of component-wise p-values as the total test p-value (the min-p-value). This procedure, when applied to consecutive batches of MCMC output, results in a series of p-values. Assuming the validity of the convergence hypothesis, p-values are distributed as *Uniform*(0,1). Although min-p-value is not the exact p-value for the full hypothesis, they argue that the resulting p-value series can be considered as an approximate sample from *Uniform* (0,1) upon convergence. Their diagnostic is then based on a visual inspection of the min-p-value series: Convergence is deduced from the point on where p-values start behaving more like a sample from uniform distribution (see Figure 3.4 in Robert et al. 1999). They exemplify their diagnostic by analyzing a normal mixture hidden Markov chain model. Here, we replace their min-p-value series by the p-values generated from applying the ECF test. Furthermore, we formally test the uniformity of the resulting p-values.

We now give a precise formulation of our convergence algorithm. In order to compare our diagnostic with that of Robert et al.'s (1999), in rest of the chapter we consider k=2 samples within a given MCMC batch. Recall that $(X_{i,1}, X_{i,2}, \dots, X_{i,N})$ is the *initial* multivariate MCMC output available from the i^{th} chain. To fix ideas, we notate the combined initial output from the r chains as Block_1, which is split into m consecutive batches each containing $2n$ observations. All batches are further divided in two successive halves, each containing n observations. The whole splitting scheme is illustrated in Table 5.1. The rest of the algorithm is described as follows.

Step-1. Set h=1.

Step-2: Generate MCMC output to construct Block_h as described above.

Step-3. Apply the ECF test to test the convergence hypothesis (5.6) for each batch in $Block_h$, resulting in m p-values.

Step-4. Apply the ECF test to test the hypothesis that resulting m p-values represent a random sample from *Uniform*(0,1).

Step-5. Jump to *Step-7* if the null hypothesis in *Step-4* cannot be rejected, else continue running the sampler and go to the next step.

Step-6. Set $h = h + 1$ and repeat *Step-2* to *Step-5*.

Table 5.1 Construction of MCMC $Block_1$ for implementing the ECF convergence diagnostics: The block is split into *m* consecutive batches, each containing 2n observations. All batches are further divided in two successive halves, Sample-1 and Sample-2, containing n observations each. These n observations consist of s MCMC draws from each chain, i.e. n = rs. The table entries under columns 'Sample-i', i =1, 2, denote number of MCMC draws. Construction of the subsequent blocks is similar.

$Block_1$					
Chain	$Batch_1$		. . .	$Batch_m$	
	Sample-1	*Sample-2*		*Sample-1*	*Sample-2*
1	*s*	*s*		*s*	*s*
2	*s*	*s*		*s*	*s*
. . .	. . .	. . .	. . .	. . .	. . .
r	*s*	*s*		*s*	*s*
Total Draws (*n*=*rs*)	*rs*	*rs*		*rs*	*rs*

Step-7. Conclude convergence and compute, for each generated chain, *burn-in* = *m(h-1)N*, where *N* is the length of the chain in each block.

Thus, upon convergence, the last *mN* observations from each chain can be used to estimate the posterior quantities of interest. If convergence is deduced within the first block, as insurance, we suggest discarding the first 1000 observations as burn-in.

An important issue in testing convergence hypothesis as described above is the presence of autocorrelation in MCMC draws (see Goldman et al. 2008 for one treatment of this problem). As suggested by Robert et al. (1999), we reduce autocorrelation by *thinning* a given MCMC batch. For instance, we select every *qth* observation from a given batch where *q*=4 or 5 seems a good choice. In the rest of this chapter, we consider samples of size at least 500 MCMC draws in each batch after thinning. This sample size is based on the simulation results in the previous section showing that the ECF test enjoys high power at this size. The overall computational burden is manageable in most MCMC setups given the currently available computational resources.

We now exemplify our ECF diagnostic by sampling from three different target distributions. For comparison, we also implement diagnostics based on Gelman and Rubin's (1992) *potential scale reduction factor* (PSRF, also called the *shrink factor*), Robert et' al. (1999) min-p-value approach, and Giakoumatos et al.'s (1999) subsampling based algorithm.

5.2.1 Example 1

We use the trivariate normal target distribution with high correlations initially described in Cowles and Carlin's (1996) MCMC diagnostics review paper. Giakoumatos et al. (1999) also used the same to exemplify their subsampling diagnostic methodology. The target distribution has mean zero with large correlations of 0.90, 0.90 and 0.98; formally,

$$\begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}, \begin{bmatrix}1 & 4.5 & 9\\ 4.5 & 25 & 49\\ 9 & 49 & 100\end{bmatrix}\right).$$

We employed the standard Metropolis-Hasting random walk algorithm (Tierney 1994) to sample from this target distribution using five independent chains. The starting values for the chains were drawn from a multivariate normal distribution dispersed with respect to the target distribution. The proposal distribution was centered at the origin with a relatively weaker correlation structure than the target distribution. The resulting MCMC trace plot for the third parameter using first three chains appears to show good mixing and rapid overall convergence within the first 1000 iterations only (Figure 5.2-a). The trace plots of the other two parameters also show similar sample paths. Figure 5.2-b shows visual implementation of the ECF diagnostic described above where we use m = 50 batches per block. We also plot the corresponding Kolmogorov-Smirov min-p-values using the Robert et al.'s (1999) diagnostic (Figure 5.2-d).

Giakoumatos, et al.'s (1999) procedure involves calculation of successive confidence regions based on subsampling statistics (Politis and Romano 1994; Politis et al. 1997) whose *range* is proportional to $\frac{1}{\sqrt{N_j}}$ upon convergence. The subsampling statistic they preferred was the empirical 90$^{\text{th}}$ quantile arguing that stabilization of the estimated target distribution in the tails indicates satisfactory convergence of the MCMC chain. Let R_j denotes the range computed from a subsample of size N_j observations, then their diagnostic is based on a plot of coefficient of determination (*R.square*, Figure 5.2-c) computed from the linear model: $R_j = \beta \frac{1}{\sqrt{N_j}} + \varepsilon_j$, where ε_j are distributed with mean zero. Convergence is declared from point on where *R.square* > 0.999. For further details, we refer the reader to Giakoumatos, et al.'s (1999). Notice that this algorithm is essentially designed for a single MCMC chain. However, as we consider running multiple chains in this chapter, we plot minimum of the *R.square* computed from all the chains as convergence criterion.

The ECF and Robert et al.'s (1999) diagnostics generate a series of 50 p-values

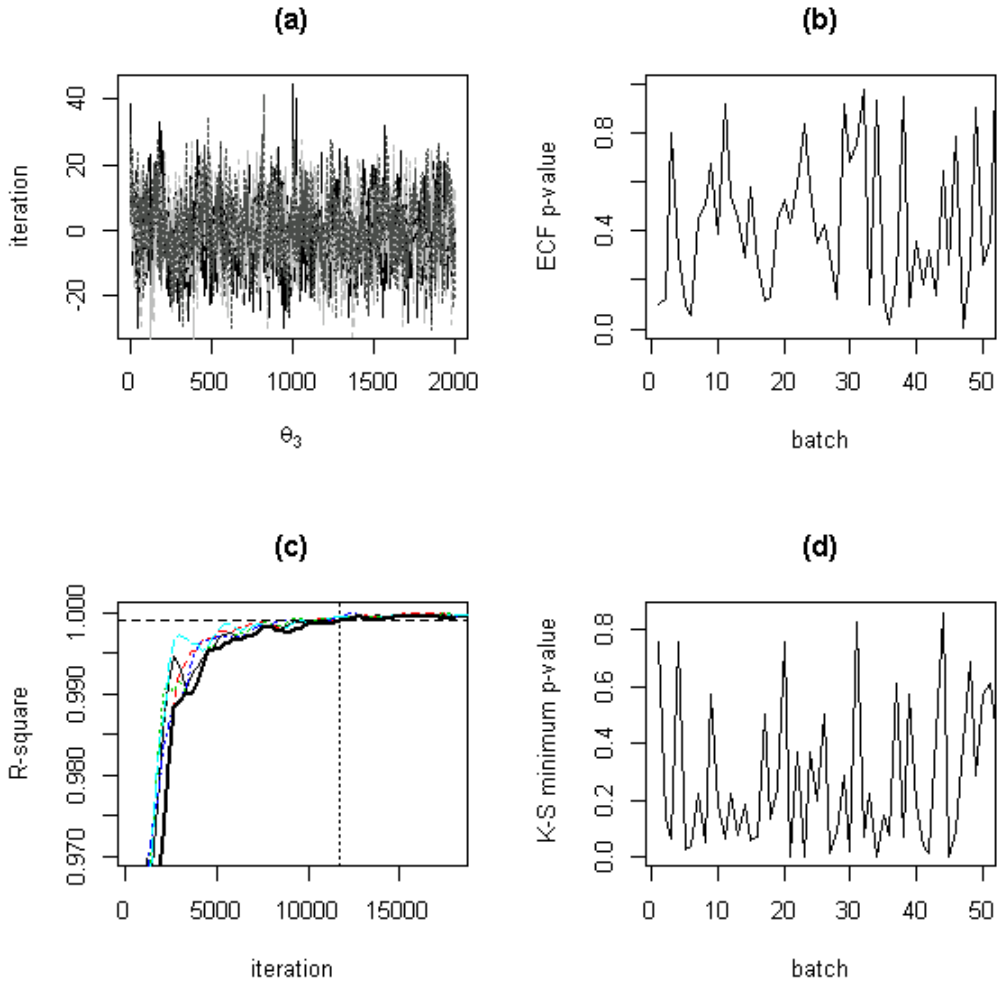

Figure 5.2 Convergence diagnostics for Example 1. (a) Trace plot of the first three chains for θ_3, and, (b) p-value series generated under the ECF diagnostic. (c) Coefficient of determination under the subsampling algorithm. Dotted lines: R-squared values for individual chains; dark solid line: the minimum R-squared values over the five chains; dashed line: threshold R.square = 0.999; vertical dotted line: burn-in=11700. (d) The min-p-value series generated under the Robert et al.'s (1999) diagnostic.

(Figure 5.2-b,d). We applied the ECF test to see if these p-value series represent samples from the *Uniform*(0,1) distribution. The resulting p-values for ECF and Robert et al.'s (1999) diagnostics are 1 and 0.018 respectively. Thus, of the two, only ECF diagnostic declares convergence within the first block. It is further evident from Figure 5.2-b that ECF based p-value series seem to emerge from *Uniform*(0,1), while min-p-value series shows a positively skewed distribution (Figure 5.2-d). In fact, min-p-value series has a

mean of 0.262; far less than the true mean 0.5. Since ECF declares convergence within the first block, we discard the first 1000 draws as burn-in.

The *R-square* values computed from Giakoumatos et al.'s (1999) subsampling algorithm are plotted in Figure 5.2-c. We see that minimum *R-square* value computed over the five chains exceeds the 0.999 threshold at about 11700 iterations which is therefore the burn-in size using the subsampling algorithm. We also computed PSRF over the first 10000 iterations. Both Gelman and Rubin's (1992) univariate and Brooks and Gelman's (1998) multivariate PSRFs were less than 1.001, confirming strong mixing as seen in Figure 5.2-a. Thus, both ECF and PSRF based diagnostics indicate rapid convergence of the sampler. This is further evident from MCMC based model parameter estimates after burning first 1000 observations: mean vector $(-.0080, .0018, .0040)^T$ and the variance-covariance $\begin{bmatrix} 1.08 & 4.82 & 9.470 \\ & 26.75 & 52.06 \\ & & 106.91 \end{bmatrix}$; these estimates are reasonably close to the corresponding true values.

5.2.2 Example 2

This example also appeared in both Cowles and Carlin's (1996) review paper and in Giakoumatos et al. (1999). The target distribution here is a bimodal density comprising a mixture of two multivariate normals with equal mixing proportion. The component normals share the following common covariance structure producing high correlations,

$\begin{bmatrix} 1 & 1.3 & 1.5 \\ 1.3 & 2.0 & 2.0 \\ 1.5 & 2.0 & 4.0 \end{bmatrix}$, with appreciably dispersed mean vectors $(0.0\,, 0.0,\ 0.0)^T$ and $(-6.0,\ -8.49, -12.0)^T$. The Metropolis-Hasting random walk algorithm (Tierney 1994) was used again to sample from the above distribution with five independent chains. The proposal density was centered at $(3.00\,, -4.245, -6.00)$ and was sufficiently dispersed with a strong correlation structure. The starting values were also drawn from a distribution sufficiently dispersed with respect to the target density. The resulting trace plot of the third parameter with the first three chains is depicted in Figure 5.3-a. The chains appear to converge quickly with frequent jumps between the modes. Similar pattern was observed for the remaining two parameters.

We implemented the ECF diagnostic with a block size of 100 batches. We further applied the ECF test on the generated p-value series (Figure 5.3-b) to see if they arise from the uniform model, yielding a p-value of 1.0. The corresponding p-value for the

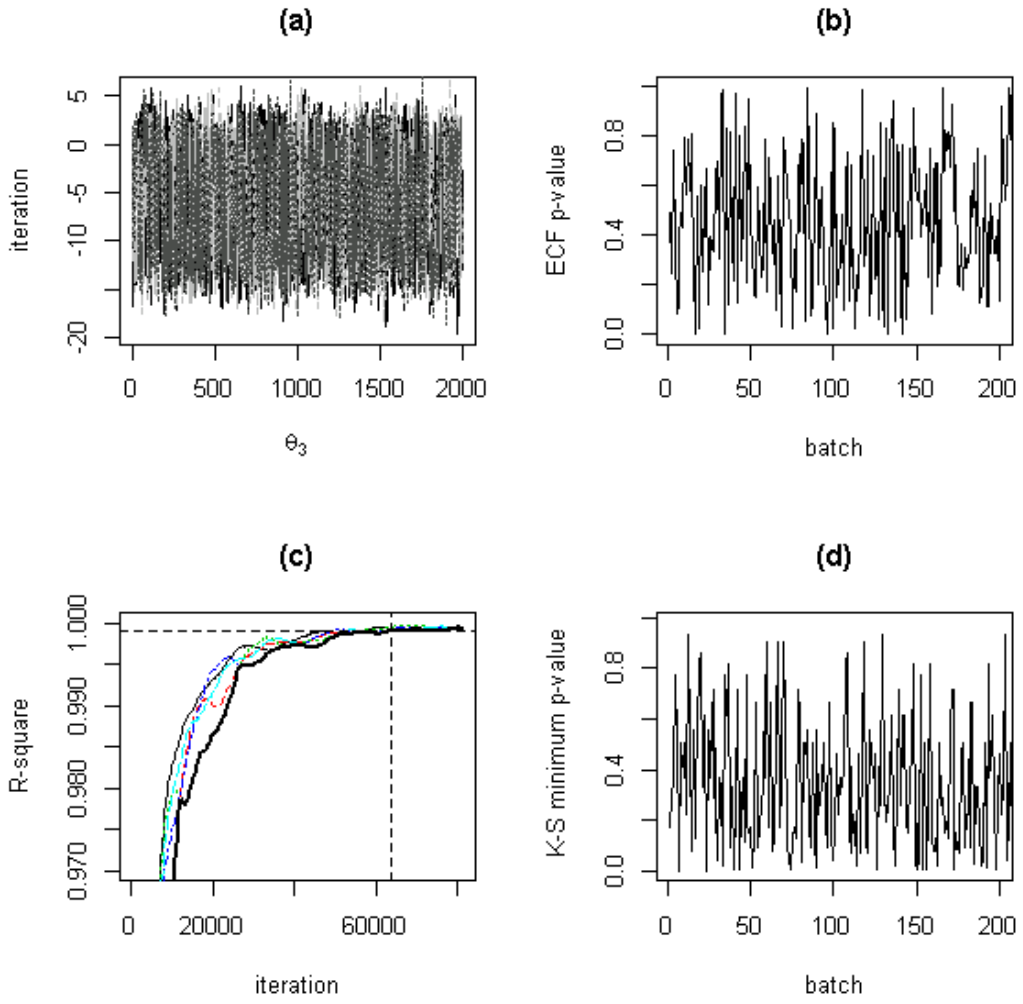

Figure 5.3 Convergence diagnostics for Example 2. (a) Trace plot of the first three chains for θ_3, and, (b) p-value series generated under the ECF diagnostic. (c) Coefficient of determination under the subsampling algorithm. Dotted lines: R-squared values for individual chains; dark solid line: the minimum R-squared values over the five chains; dashed line: threshold R.square = 0.999; vertical dotted line: burn-in=63500 (d) The min-p-value series generated under the Robert et al.'s (1999) diagnostic.

Robert et al.'s (1999) min-p-value series (Figure 5.3-d) was 0.177. However, with a block size of 200 batches, it dropped to 0 for min-p-value series but remained unchanged for the ECF diagnostic. Notice that mean of the min-p-value series (Figure 5.3-d) is 0.358, which confirms the downward bias in approximating the exact p-value of the Robert et al.'s (1999) diagnostic. Thus, the ECF diagnostic concludes convergence within the first block with a burn-in size of 1000 iterations. The Gelman and Rubin's (1992) diagnostic

also showed rapid convergence with the multivariate PSRF value of 1.01 computed from the first 1000 iterations.

The subsampling algorithm, however, showed more delayed convergence. It is clear from Figure 5.3-c that minimum R-square value exceeds 0.999 after a burn-in size of 63500 iterations. To see if a large burn-in size as determined by the subsampling algorithm is really necessary, we compared the empirical distributions based on the ECF and subsampling diagnostics. For this, we applied the ECF test on two samples consisting of 5000 MCMC draws each, obtained after burning the first 1000 draws for the ECF diagnostic, and 63500 draws for the subsampling diagnostic. The resulting p-value of 0.556 provides strong evidence that both samples arise from the same stationary distribution. Furthermore, the mean vector and covariance matrix of the ECF diagnostic based empirical distribution and those from target distribution (in parenthesis) are compared below. The estimate of the stationary distribution is based on 10000 iterations after burn-in=1000. The resulting MCMC estimates are reasonably close to the corresponding true parameter values, indicating that the sampler has explored the target distribution sufficiently well. This also shows that the subsampling diagnostic is very conservative in declaring convergence to the stationary distribution.

Mean Vector: $[-2.98(-3.99), -4.20(-4.25), -5.97(-6.00)]^T$

Covariance Matrix: $\begin{bmatrix} 9.72(10.00) & 13.64(14.00) & 18.89(19.50) \\ & 19.48(20.02) & 26.60(27.47) \\ & & 38.83(40.00) \end{bmatrix}$.

5.2.3 Example 3

Measurement error models play an important role in many scientific disciplines, such as epidemiology and environmental sciences, where predictor variables in a regression model often cannot be observed directly. For a comprehensive treatment of the related inferential methods, we refer the reader to Carroll et al. (2006). Here we consider Bayesian inference for the structural polynomial measurement error model (Huang and Huwang 2001). It is well known that the structural simple linear measurement error model is inestimable unless an estimate of the measurement error variance is available from additional data (Carroll et al. 2006). Interestingly, Huang and Huwang (2001) show that the polynomial measurement error model is fully identifiable provided the degree is greater than 1. They also derive consistent estimators of the model parameters.

The model is formalized as follows: $Y_i = \beta_0 + \beta_1 \xi_i + \beta_2 \xi_i^2 + ... + \beta_p \xi_i^p + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $\xi_i \sim N(\mu_\xi, \sigma_\xi^2)$ and $X_i \sim N(\xi_i, \sigma_\delta^2)$, while we only observe (Y_i, X_i), i = 1, 2, … , n. We simulated a random sample of size 50 observations under the quadratic model with the true parameter vector $(\beta_0, \beta_1, \beta_2, \mu_x, \sigma_\delta^2, \sigma_\varepsilon^2, \sigma_\xi^2)$ = (5, 2, 3, 0, 0.1, 1, 1). These parameters values are similar to those chosen by Huang and Huwang (2001). To draw samples from the posterior distribution of the parameter vector, we used the MCMC samplers implemented in JAGS 3.1.0 (Plummer 2003, 2011a) using the rjags package (Plummer 2011b) of the R computing software (Venables and Smith 2011). We generated 100 parallel chains each of length 10000 iterations. The reason we generate a large number of chains for this model is explored in the discussion section. All the parameters were given fairly noninformative proper priors except for μ_x which was given $N(\hat{\mu}_\xi, 0.01)$, where $\hat{\mu}_\xi = \overline{X}$, a consistent estimator of μ_x. Visual inspection of the MCMC trace plots of regressions coefficients and variance parameters (on natural logarithm scale) show reasonable mixing for $(\beta_0, \sigma_\varepsilon^2)$ only, while the other parameters show poor mixing (Figure 5.4, 5.5). However, the individual chains seem to stabilize after a few hundred initial iterations.

For this example we only focus on comparing the ECF diagnostic with Gelman and Rubin's (1992) PSRF. Agreeing to the visual impression, the univariate shrink factor shows convergence for $(\beta_0, \sigma_\varepsilon^2)$ only (Figures 5.6). While the shrink factor for other parameters initially drops and then stabilizes at values higher than 1.1, it seem to increase for σ_ξ^2 within the first 10000 values iterations. Furthermore, Brooks and Gelman's (1998) multivariate PSRF computed from the second half of the chain is 2.30, showing that the sampler is still running in the transient phase.

Next we applied the ECF diagnostic with m=30 batches per block and s = 10, resulting in sample size n = 1000 (see Table 5.1 for details). The thinning size was q=4 draws. The resulting p-value series up to *Block-4* is plotted in Figure 6-b. The hypothesis test for the uniformity of the p-values in *Block-1* resulted in a p-value of 0.058, while the test p-value was higher than 0.30 in remaining blocks. Therefore, based on the ECF diagnostic, we discard output from the first 30 batches as burn-in (2400 draws per chain). Notice that the posterior estimates of the parameters match well with those obtained by Huang and Huwang (2001) in their simulation study. Thus, the sampler appears to have

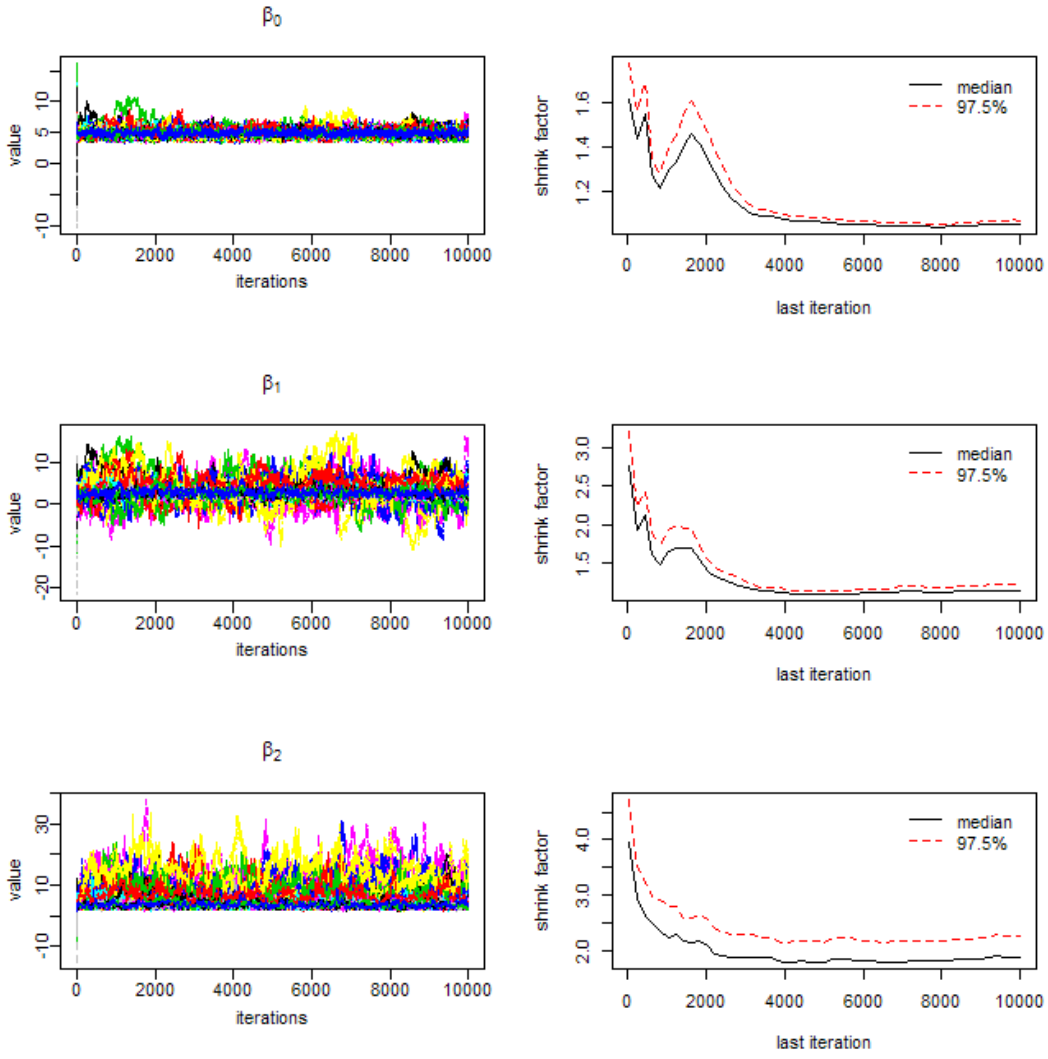

Figure 5.4. Left panel: Trace plots of the hundred parallel chains; Right panel: Corresponding Gelman and Rubin's (1992) PSRF values (shrink factors).

explored the posterior space sufficiently well after burn-in=2400.

5.3 Discussion

Ideally, we expect multiple chains to strongly mix as well as converge in terms of stabilization of the full empirical distribution, as we saw in Example 1-2. However, in practice most samplers produce chains that are only quasi-ergodic, i.e. they mix poorly and get confined to local modes for a computationally intractable amount of time (Murray 2010). This can happen, for instance, when the target density is highly multimodal, making it

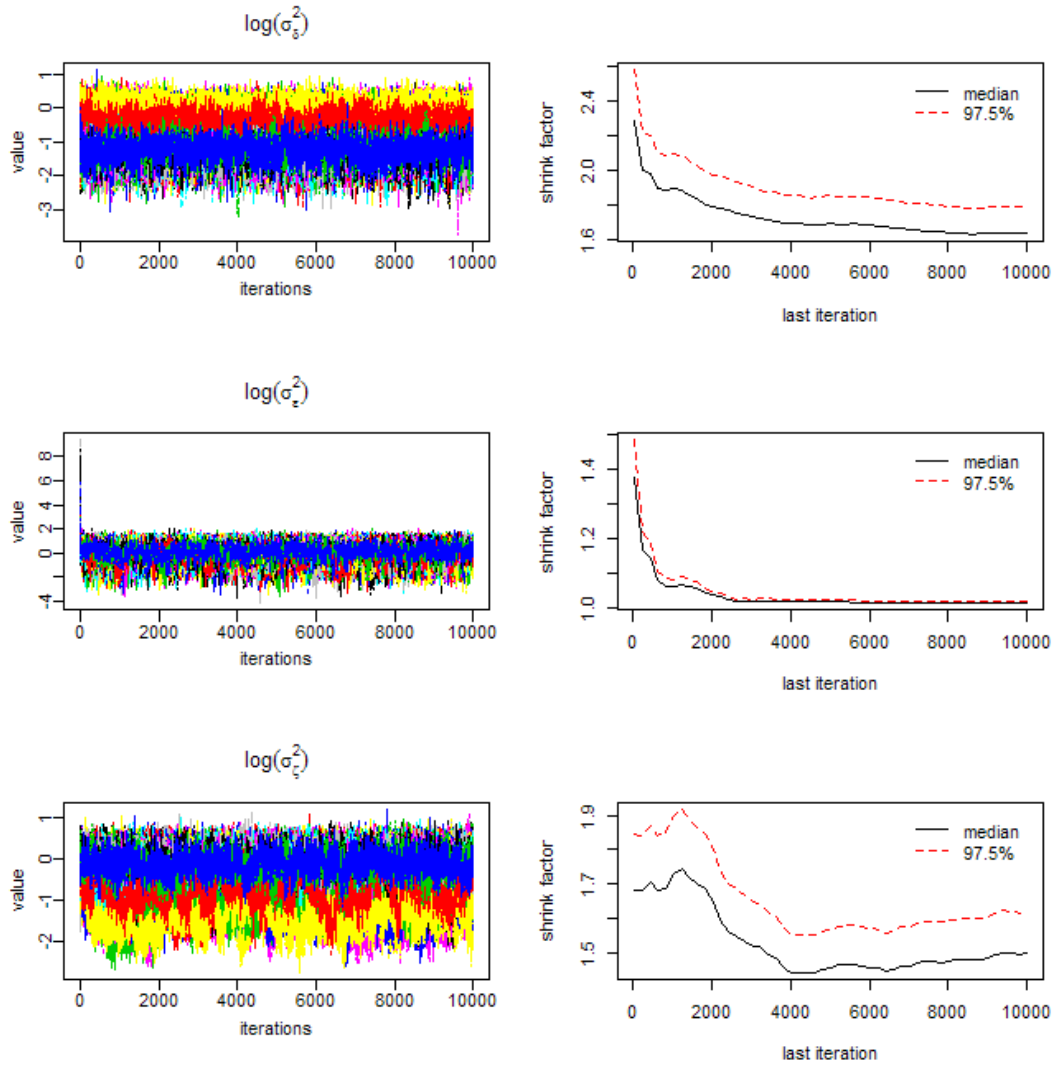

Figure 5.5. Left panel: Trace plots of the hundred parallel chains; Right panel: Corresponding Gelman and Rubin's (1992) PSRF values (shrink factors).

virtually impossible for a single quasi-ergodic chain to traverse the entire state-space. It is therefore useful to run many parallel chains so that the sampler can effectively explore various local modes. There indeed exists much debate on the efficacy of using a single long chain in assessing stationarity. For a detailed theoretical treatment of the subject, we refer the reader to Galman and Rubin (1992), Chauveau and Debolt (1997) and Mengeren et al. (1999). Despite the fact that individual chains might explore only local basins of attraction, collectively they can still provide useful information about the shape of the target density. This, however, blows up the PSRF even though the *ensemble* of chains exhibit stationary behavior. Example3 reflects this phenomenon where, although the

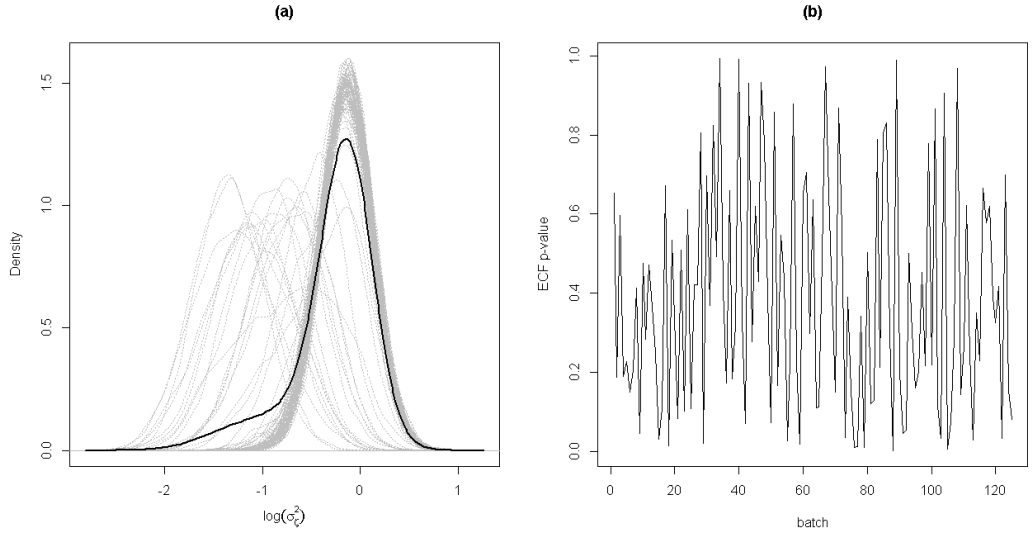

Figure 5.6. Convergence diagnosis for Example 3. (a) Density plots for $\log(\sigma_\zeta^2)$ after burn-in=2400 per chain; dotted lines: density plots based on the individual chains; solid line: density plot based on the 100 parallel chains. (b) p-value series generated under the ECF diagnostic.

chains seem to mix poorly, individually they explore a relatively small region of the state-space. This is evident from the marginal density plots computed from individual chains and that of their ensemble (Figure 5.6-a). Presence of various connected basins of attraction is apparent but the ensemble produces a unimodal distribution that has become stable after a few hundred initial iterations. This explains why ECF test pronounces convergence while PSRF requires further simulation because of insufficient mixing.

The stabilization of the full empirical posterior distribution, determined for instance in Example 3 using ECF, does not guarantee that the sampler has actually converged to the correct stationary distribution. Indeed, no diagnostic can achieve this because the stationary distribution will always be unknown to us in practice. Thus, designing efficient MCMC algorithms is crucial to have credence in the observed stationary distribution. This is currently an active area of research aiming at efficiently exploring high dimensional and, potentially, multimodal spaces in a computationally feasible manner. To get an overview of the recent developments, the reader is referred to Craiu et al.'s (2009) regional adaptive MCMC (AMCMC) and that of Andrieu et al.'s (2010) particle MCMC (PMCMC) methods.

Existing ECF tests are based on a weighted L_2 distance between the empirical characteristic functions where the weight is assigned through an even integrable weight

function w(t) (Hušková and Meintanis 2008; Meintanis and Swanepoel 2007). The choice of a suitable weight function depends on two aspects i) when some a priori knowledge exists about the form of the densities, w(t) is chosen to direct the power of the test towards frequencies where CF's differ maximally and, ii) to render a closed form of the test statistic suitable for computer calculation (Hušková and Meintanis 2008). These tests then rely on the permutation bootstrap approach for computing the test p-value. However, our ECF test avoids the arbitrariness introduced by w(t) and the test p-values are easy to calculate since the asymptotic CDF of the test statistic is exactly computable from the standard normal CDF. Furthermore, when a priori knowledge about the form of the densities does exist, it can be incorporated in our ECF test as well by constructing a grid **G** that is dense around the frequencies where CF's diverge maximally.

When the target distribution is high dimensional, computation of T_n can become computationally expensive because of the large size of the grid **G** (see Appendix A for the construction of **G**). A suitable approach in this case is to study the convergence of chains over a subset of parameters. We suggest monitoring convergence of various randomly selected subsets and continue sampling until convergence is achieved on all the subsets. Again, the stopping rule can be based on the resulting p-value series from all the subsets combined and then reduced to a single p-value for testing their uniformity.

5.4 Summary

In this chapter we introduced a diagnostic procedure for assessing the convergence of an MCMC generated multidimensional empirical distribution. We also introduced a new ECF based nonparametric test for the multivariate *k*-sample testing problem. The test is very sensitive in detecting shifts in different features of a multivariate density such as scale and multimodality and, therefore, forms the building block of our diagnostic algorithm.

Chapter 6

Conclusions and Future Research

This thesis sets out to develop a novel computational algorithm, data cloning, to conduct likelihood based analysis of one of the most useful classes of statistical models: the general hierarchical models. We also develop an MCMC convergence diagnostic procedure based on a new nonparametric test for comparing several multivariate populations. In this chapter, we review these contributions and discuss future direction.

The overview in Chapter 2 provides a glimpse of the existing approaches to analyzing GLMMs, an important class of hierarchical models. A common thread to the existing likelihood based methods is that they all employ some sort of approximations to the high dimensional multiple integral defining the likelihood function to avoid its explicit evaluation. Although some of these methods such as AGQ and FLA show good performance, they are more or less restricted to analyzing GLMMs only. On the other hand, Bayesian estimation of hierarchical models circumvents the aforementioned approximations by assuming prior distributions to integrate out the uncertainty concerning model parameters. The well developed MCMC methodology then allows sampling from the posterior distributions to conduct inference. A limitation of the Bayesian approach is that the inferences can be strongly dependent on the choice of prior distributions.

The DC algorithm developed in this thesis also uses Bayesian formulation and computational techniques. However, the resulting inferences are based on the classical frequentist paradigm and are invariant to the choice of prior distributions. Our methodology is applicable in most situations where the problem can be formulated as a Bayesian problem and where MCMC can be used to obtain random variates from the posterior distribution. Similar to the Bayesian methodology, data cloning avoids high-dimensional numerical integration and requires neither maximization nor differentiation of a function.

It is based only on the computation of the means and the variances. Application of the DC algorithm for estimating complex GLMMs and nonlinear state-space models in Chapter 3 and 4 respectively, shows that the method is viable for analysing a wide range of hierarchical models. Furthermore, as we described in Chapter 4, inference procedures such as model selection using information criteria, profile likelihood for inference in the presence of nuisance parameters etc. are also possible using data cloning.

One of the promising features of data cloning is the test for estimability of parameters in hierarchical models. Understanding estimability of the parameters is extremely important in practice, where models are complex and analytical results are sparse. Any valid scientific inference can only be based on identifiable parameters. Thus, checking for estimability is critical for good scientific practice. However crucial, analytical determination of identifiability is a very difficult problem in general. In this thesis we demonstrate that the DC algorithm provides a powerful numerical diagnostic tool to flag lack of model estimability in hierarchical models. However, there is further potential to improve upon the existing diagnostic. Theorem 3.3 shows that when some of the model parameters are nonidentifiable, the full posterior distribution is embedded in a lower dimensional manifold of the parameter space. This suggests that the existing manifold estimation techniques (Lin and Zha 2008) can be employed to devise a more general estimability diagnostic procedure. We plan to explore this possibility in a separate study. We also refer the reader to Campbell and Lele (2013) for an ANOVA based extension to the existing diagnostic approach.

Recently Rue and Martino (2009) introduced their integrated nested Laplace algorithm (INLA) for conducting approximate Bayesian inference in latent Gaussian models, an important subclass of hierarchical models that includes generalized linear mixed models, generalized additive models and state-space models as special cases. Their algorithm is based on Tierney and Kadane's (1986) Laplace approximation of marginal posterior distributions. The key feature of INLA is that it yields direct numerical approximations of the marginal posterior distributions, completely circumventing MCMC calculations. Baghishani et al. (2012) introduced an interesting further extension by developing a hybrid of INLA and the DC algorithm (DCINLA). This hybrid algorithm produces direct approximations to the DC based marginal posterior distributions, yielding MLEs and their standard errors. However, DCINLA does not produce the asymptotic variance-covariance matrix of the MLEs. Baghishani et al. (2012) illustrated the implementation of DCINLA by analyzing various standard GLMMs, producing estimates comparable with

those computed from other maximum likelihood estimation procedures. However, because data cloning induces a larger random effects structure than existing in the original data set, it is yet to be seen if DCINLA scales up to fitting complex models, such as the CAR model analyzed in Chapter 3. Extending DCINLA in this direction would be a useful contribution to likelihood based inference in general hierarchical models.

In Chapter 4, we present a flexible and computationally efficient methodology to include environmental covariates and delayed density dependence in PVAs while simultaneously accounting for observation error and parameter uncertainty. Analysis of song sparrow abundance counts implies a potential future extension to our approach: to incorporate observation error in spatially structured metapopulation models. These models have become a key tool for conservation and management of spatially fragmented populations (Dunning et al. 1995, Akçakaya 2000, Akçakaya et al. 2007). As accurate population estimates are seldom available for such fragmented populations, estimation of these models ignoring observation error can seriously limit their predictive strength. Recent applications of these models have overlooked the issue of observation error mainly due to the computational difficulties in handling complex hierarchical models (Dennis et al. 1998; Lele et al. 1998; Schtickzelle and Baguetti 2004; Jonzén et al. 2005). Extending the methods presented in this thesis to the spatially explicit PVA models would therefore be an important further contribution to the PVA tool kit.

As we pointed out at the onset of this thesis, quality of MCMC based Bayesian inference, and that of the DC algorithm for that matter, depends on satisfactory convergence of the MCMC chains. Controlling MCMC convergence in the hierarchical modeling context can be especially tricky as the target posterior distributions tend to be multimodal. Although, there exist a number of diagnostic tools for assessing convergence of univariate chains, only few procedures exist for controlling convergence to multivariate target distributions. Furthermore, these tests are mostly based on monitoring convergence of a functional of the multidimensional chain, overlooking multivariate features of the target distribution, such as the variance-covariance structure. In this thesis, we introduce a new diagnostic method that ensures that the empirical distribution of the multidimensional chain converges to a stationary distribution as a whole.

The diagnostic procedure we developed is based on a novel ECF based test for comparing k multivariate distributions. The simulation study in Chapter 5 shows that the test is quite powerful in detecting shifts in key features of multivariate distributions, such as skewness, variance-covariance and multimodality. Although not included in this the-

sis, we plan to evaluate the performance of our ECF test for k univariate distributions in a separate simulation study. We also plan to extend the test to assessing goodness-of-fit between two or more distributions. A further contribution would be to develop the test for assessing independence of a collection of random variables. We report these extensions elsewhere.

Appendix A

Derivation of the ECF Test

We shall assume throughout that the random vector **X** has absolutely continuous distribution function $F(\mathbf{x})$. Then, following the development in Section 5.1, we recall from (5.3) that $\varphi^{(j)}(\boldsymbol{t}) = \varphi^{(l)}(\boldsymbol{t}), \Leftrightarrow \mu_{C_t}^{(j)} = \mu_{C_t}^{(l)}$ and $\mu_{S_t}^{(j)} = \mu_{S_t}^{(l)}, \forall\, \boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}$, where $\mu_{C_t}^{(j)} = E\left[\cos(\boldsymbol{t}'\mathbf{X}^j)\right]$ and $\mu_{S_t}^{(j)} = E\left[\sin(\boldsymbol{t}'\mathbf{X}^j)\right]$. This allows us to write the null hypothesis in (5.2) in terms of real and imaginary parts of $\varphi^{(j)}(\boldsymbol{t}),\ j = 1, 2, \ldots, k$. Recall that $\mathbf{B}_{-k}$ is a $k \times k$ contrast matrix with elements $b_{jj} = -(k-1)$ and $b_{jl} = 1,\ \forall\, j \neq l$, so that $\boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} = \mathbf{B}_{-k}\boldsymbol{\mu}_{C_t}$ and $\boldsymbol{\delta}_{\boldsymbol{t}}^{(S)} = \mathbf{B}_{-k}\boldsymbol{\mu}_{S_t}$ define vectors of linear contrasts, where $\boldsymbol{\mu}_{C_t}$ and $\boldsymbol{\mu}_{S_t}$ notate the vectors of real and imaginary parts, respectively, of the k CFs. The following lemma represents H_o in terms of these contrast vectors.

Lemma A1: The null hypothesis defined in (5.1) and (5.2) can be equivalently stated as

$$H_o: \boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} = \boldsymbol{\delta}_{\boldsymbol{t}}^{(S)} = \mathbf{0}, \forall\, \boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}. \tag{A1}$$

Proof: It is trivial to show that, for any $\boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}$, the homogeneous system of linear equations defined by Eq. A1 is such that,

$$\boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} = \mathbf{0} \Leftrightarrow \mu_{C_t}^{(1)} = \mu_{C_t}^{(2)} = \cdots = \mu_{C_t}^{(k)} \tag{A2}$$

and

$$\boldsymbol{\delta}_{\boldsymbol{t}}^{(S)} = \mathbf{0} \Leftrightarrow \ \mu_{S_t}^{(1)} = \mu_{S_t}^{(2)} = \cdots = \mu_{S_t}^{(k)}. \tag{A3}$$

Then, it follows from (5.3) that Eqs. A2-A3 hold true for $\forall\, \boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}$ if and only if (5.2) holds for $\forall\, \boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}$. Hence the proof follows. ■

The following theorem states some properties of the characteristic function, proof of which can be found in Ushakov (1999) and Rosenthal (2006).

Theorem A1: For $\forall\, \boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}$:

(i) $\varphi(\boldsymbol{t})$ exists for any random vector **X**.

(ii) $|\varphi(\boldsymbol{t})| \leq |\varphi(\mathbf{0})| = 1$

(iii) $\overline{\varphi(\boldsymbol{t})} = \varphi(-\boldsymbol{t})$, where $\overline{\varphi(\boldsymbol{t})}$ is the complex conjugate of $\varphi(\boldsymbol{t})$.

(iv) $\varphi(\boldsymbol{t})$ is uniformly continuous in $\boldsymbol{t}$.

(v) If **X** has absolutely continuous distribution, $\lim_{\|\boldsymbol{t}\|\to\infty}|\varphi(\boldsymbol{t})| = 0$.

The following lemma states variance formulae of the random variables $\cos(\boldsymbol{t}'\mathrm{X})$ and $\sin(\boldsymbol{t}'\mathrm{X})$. We refer the reader to Fan (1997) for details.

Lemma A2: For any $\boldsymbol{t} \neq \mathbf{0}$

(i) $\sigma_{C_t}^2 \equiv \mathrm{Var}[\cos(\boldsymbol{t}'\mathbf{X})] = \frac{1}{2}\left[1 + \mu_{C_{2t}} - 2\mu_{C_t}^2\right]$, and, (A4)

(ii) $\sigma_{S_t}^2 \equiv \mathrm{Var}[\sin(\boldsymbol{t}'\mathbf{X})] = \frac{1}{2}\left[1 - \mu_{C_{2t}} - 2\mu_{S_t}^2\right]$. (A5)

In the following lemma we state a limiting result on Lemma A 2.

Lemma A3: As $\|\boldsymbol{t}\| \to \infty$, $\sigma_{C_t}^2 \to \frac{1}{2}$, $\sigma_{S_t}^2 \to \frac{1}{2}$.

Proof: From Theorem A1-v we have $\lim_{\|\boldsymbol{t}\|\to\infty}|\varphi(\boldsymbol{t})| = 0$. But $|\varphi(\boldsymbol{t})|^2 = \mu_{C_t}^2 + \mu_{S_t}^2 = 0$ $\Leftrightarrow \mu_{C_t} = \mu_{S_t} = 0$. Hence the proof follows immediately upon applying the limit in Eqs. A4-5. ■

Recall that the ECF for $F(.)$ is defined as

$\varphi_n(\boldsymbol{t}) = \frac{1}{n}\sum_{h=1}^{n} e^{i\boldsymbol{t}'\mathrm{X}_h} = \frac{1}{n}[\sum_{h=1}^{n}\cos(\boldsymbol{t}'\mathrm{X}_h) + i\sum_{h=1}^{n}\sin(\boldsymbol{t}'\mathrm{X}_h)] \equiv \hat{\mu}_{C_t} + i\hat{\mu}_{S_t}$,

where $\hat{\mu}_{C_t}$ and $\hat{\mu}_{S_t}$ are consistent estimators of μ_{C_t} and μ_{S_t}, respectively. We also define consistent estimators $\widehat{\boldsymbol{\delta}}_{\boldsymbol{t}}^{(C)} = \mathbf{B}_{-k}\widehat{\boldsymbol{\mu}}_{C_t}$ and $\widehat{\boldsymbol{\delta}}_{\boldsymbol{t}}^{(S)} = \mathbf{B}_{-k}\widehat{\boldsymbol{\mu}}_{S_t}$ of the contrast vectors appearing in Eq. A1. In the following theorem, assuming H_o is true, we state a result on the limiting marginal distribution of the components of $\widehat{\boldsymbol{\delta}}_{\boldsymbol{t}}^{(C)}$ and $\widehat{\boldsymbol{\delta}}_{\boldsymbol{t}}^{(S)}$. Let us first introduce some notation. We write

$\boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} \equiv (\beta_{\boldsymbol{t}}^{(1)}, \beta_{\boldsymbol{t}}^{(2)}, \dots, \beta_{\boldsymbol{t}}^{(k)})'$ and $\boldsymbol{\delta}_{\boldsymbol{t}}^{(S)} \equiv (\gamma_{\boldsymbol{t}}^{(1)}, \gamma_{\boldsymbol{t}}^{(2)}, \dots, \gamma_{\boldsymbol{t}}^{(k)})'$, where, all $\beta_{\boldsymbol{t}}^{(j)}$ and $\gamma_{\boldsymbol{t}}^{(j)}$ are linear combinations of the elements of $\boldsymbol{\mu}_{C_t}$ and $\boldsymbol{\mu}_{S_t}$, respectively, defined by the j^{th} row vector of **B**. For instance, $\beta_{\boldsymbol{t}}^{(1)}$ and $\gamma_{\boldsymbol{t}}^{(1)}$ are given as:

$$\beta_{\boldsymbol{t}}^{(1)} = -(k-1)\mu_{C_t}^{(1)} + \mu_{C_t}^{(2)} + \cdots + \mu_{C_t}^{(k)} \equiv \boldsymbol{g}_1'\boldsymbol{\mu}_{C_t} \tag{A6 a}$$

and

$$\gamma_{\boldsymbol{t}}^{(1)} = -(k-1)\mu_{S_t}^{(1)} + \mu_{S_t}^{(2)} + \cdots + \mu_{S_t}^{(k)} \equiv \boldsymbol{g}_1'\boldsymbol{\mu}_{S_t}. \tag{A6 b}$$

We also define $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ and $\hat{\gamma}_{\boldsymbol{t}}^{(j)}$ as the corresponding estimators. Before we state Theorem A2, we first present some results on the variance of $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ and $\hat{\gamma}_{\boldsymbol{t}}^{(j)}$ in the following lemmas.

Lemma A4:

(i) Assuming H_o is true, $\beta_{\boldsymbol{t}}^{(j)} = \gamma_{\boldsymbol{t}}^{(j)} = 0, \forall\, \boldsymbol{t} \in \mathbb{R}^p,\ j = 1, 2 \dots, k.$

(ii) Assuming H_o is false, $\lim_{\|\boldsymbol{t}\|\to\infty} \beta_{\boldsymbol{t}}^{(j)} = \lim_{\|\boldsymbol{t}\|\to\infty} \gamma_{\boldsymbol{t}}^{(j)} = 0\ ,\ j = 1, 2 \dots, k.$

(iii) For any $\boldsymbol{t} \in \mathbb{R}^p$, the statistics $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ and $\hat{\gamma}_{\boldsymbol{t}}^{(j)}$ are unbiased estimators of the contrasts $\beta_{\boldsymbol{t}}^{(j)}$ and $\gamma_{\boldsymbol{t}}^{(j)}$ respectively, $j = 1, 2 \dots, k.$

Proof:

(i) The proof follows immediately by using Eqs. A2-3.

(ii) We only show for $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ as the proof for $\hat{\gamma}_{\boldsymbol{t}}^{(j)}$ is analogous. The proof follows by noticing from Theorem A1-v that $\lim_{\|\boldsymbol{t}\|\to\infty}|\varphi(\boldsymbol{t})|^2 = \lim_{\|\boldsymbol{t}\|\to\infty}\mu_{C_t}^2 + \mu_{S_t}^2 = 0 \iff \mu_{C_t} = \mu_{S_t} = 0$, which implies $\lim_{\|\boldsymbol{t}\|\to\infty}\beta_{\boldsymbol{t}}^{(j)} = 0, j = 1,2,\dots,k$.

(iii) We only show for $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ as the proof for $\hat{\gamma}_{\boldsymbol{t}}^{(j)}$ is analogous. From the definition of ECF we have $\hat{\mu}_{C_t}^{(j)} = \frac{1}{\mathrm{n}}\sum_{h=1}^{n}\cos\left(\boldsymbol{t}'\mathrm{X}_h^{(j)}\right)$ where it is trivial to show that $E[\,\hat{\mu}_{C_t}^{(j)}] = \mu_{C_t}^{(j)}$, $\forall\, \boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}, \Rightarrow E[\widehat{\boldsymbol{\mu}}_{C_t}] = \boldsymbol{\mu}_{C_t}$. Thus, $E[\hat{\beta}_{\boldsymbol{t}}^{(j)}] = E[\boldsymbol{g}_j'\widehat{\boldsymbol{\mu}}_{C_t}] = \boldsymbol{g}_j'\boldsymbol{\mu}_{C_t} = \beta_{\boldsymbol{t}}^{(j)}$. ■

Lemma A5: Assuming H_o is true, for any $\boldsymbol{t} \in \mathbb{R}^p$, $\mathrm{Var}\left(\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}\right) = k(k-1)\sigma_{C_t}^2$ and $\mathrm{Var}\left(\sqrt{n}\hat{\gamma}_{\boldsymbol{t}}^{(j)}\right) = k(k-1)\sigma_{S_t}^2$, where $\sigma_{C_t}^2$ and $\sigma_{S_t}^2$ are defined in Eqs. A4-5, respectively.

Proof:

We begin by showing that $\sqrt{n}\widehat{\boldsymbol{\mu}}_{C_t}$ and $\sqrt{n}\widehat{\boldsymbol{\mu}}_{S_t}$ have diagonal variance-covariance matrices that are of the form $\mathbf{D}_{k,S_t} = \sigma_{S_t}^2\mathrm{I}_k$ and $\mathbf{D}_{k,C_t} = \sigma_{C_t}^2\mathrm{I}_k$, respectively, where I_k is an identity matrix. We only present the proof for $\mathbf{D}_{k,C_t}$, whereas $\mathbf{D}_{k,S_t}$ can be derived similarly.

The off-diagonal elements of $\mathbf{D}_{k,C_t}$ are all zero since the elements of $\widehat{\boldsymbol{\mu}}_{C_t}$are functions of independent random samples. The diagonal elements are given as:

$$d_{jj} = \mathrm{Var}\left(\sqrt{n}\hat{\mu}_{C_t}^{(j)}\right) = n\mathrm{Var}\left(\frac{1}{\mathrm{n}}\sum_{h=1}^{n}\cos(\boldsymbol{t}'\mathrm{X}_h^{(j)})\right) =$$

$$\frac{1}{\mathrm{n}}\sum_{h=1}^{n}\mathrm{Var}\left(\cos(\boldsymbol{t}'\mathrm{X}_h^{(j)})\right) = \sigma_{C_t}^2. \quad \text{(A7)}$$

Next, we derive the variance of $\hat{\beta}_{\boldsymbol{t}}^{(j)}$. Recalling from Eq. A6, we can write $\mathrm{Var}\left(\hat{\beta}_{\boldsymbol{t}}^{(j)}\right)$ as

$$\mathrm{Var}\left(\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}\right) = \mathrm{Var}\left(\boldsymbol{g}_j'\sqrt{n}\widehat{\boldsymbol{\mu}}_{C_t}\right) = \boldsymbol{g}_j'\mathrm{Var}\left(\sqrt{n}\widehat{\boldsymbol{\mu}}_{C_t}\right)\boldsymbol{g}_j = \boldsymbol{g}_j'\sigma_{C_t}^2\mathrm{I}_k\boldsymbol{g}_j = k(k-1)\sigma_{C_t}^2.$$

(A8)

The derivation of $\mathrm{Var}\left(\sqrt{n}\hat{\gamma}_{\boldsymbol{t}}^{(j)}\right)$ is also similar. Hence, the proof is complete. ■

Corollary A1: $\lim_{\|\boldsymbol{t}\|\to\infty} \mathrm{Var}\left(\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}\right) = \lim_{\|\boldsymbol{t}\|\to\infty} \mathrm{Var}\left(\sqrt{n}\hat{\gamma}_{\boldsymbol{t}}^{(j)}\right) = \binom{k}{2}$.

Proof: The proof follows immediately from applying the result in Lemma A3. ∎

Remark 1: When $k = 2$, $\boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} = \mathbf{0}$ in Eq. A1 results in two linear contrasts: $\beta_{\boldsymbol{t}}^{(1)} = -\mu_{C_{\boldsymbol{t}}}^{(1)} + \mu_{C_{\boldsymbol{t}}}^{(2)}$ and $\beta_{\boldsymbol{t}}^{(2)} = \mu_{C_{\boldsymbol{t}}}^{(1)} - \mu_{C_{\boldsymbol{t}}}^{(2)}$, one of which is clearly redundant. The same is true for contrasts associated with $\boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} = \mathbf{0}$. Throughout the Appendix, for $k = 2$, we define Eq. A1 in terms of $\beta_{\boldsymbol{t}}^{(1)}$ and $\gamma_{\boldsymbol{t}}^{(1)}$ only and simply denote them as $\beta_{\boldsymbol{t}}$ and $\gamma_{\boldsymbol{t}}$ respectively. Notice that all the results obtained herein remain valid for $k = 2$ as well. ∎

Lemma A6: Regardless of the truth or falsity of H_o,

$$\lim_{\|\boldsymbol{t}\|\to\infty} \mathrm{Var}\left(\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}\right) = \mathrm{Var}\left(\sqrt{n}\hat{\gamma}_{\boldsymbol{t}}^{(j)}\right) = \binom{k}{2}.$$

Proof: The proof is very similar to that of Lemma A5, except for the fact that Eq. A7 is derived as follows:

$$\begin{aligned} d_{jj} &= \mathrm{Var}\left(\sqrt{n}\hat{\mu}_{C_{\boldsymbol{t}}}^{(j)}\right) \\ &= n\mathrm{Var}\left(\frac{1}{\mathrm{n}}\sum_{h=1}^{n}\cos(\boldsymbol{t}'\mathrm{X}_h^{(j)})\right) \\ &= \frac{1}{n}\sum_{h=1}^{n}\mathrm{Var}\left(\cos\left(\boldsymbol{t}'\mathrm{X}_h^{(j)}\right)\right) \\ &= \tfrac{1}{n}\textstyle\sum_{h=1}^{n}\sigma_{j,C_{\boldsymbol{t}}}^2 = \sigma_{j,C_{\boldsymbol{t}}}^2 . \end{aligned} \qquad \text{(A9)}$$

However, it follows from Lemma A3 that , $\lim_{\|\boldsymbol{t}\|\to\infty} \sigma_{j,C_{\boldsymbol{t}}}^2 = \frac{1}{2}$. Thus, $\lim_{\|\boldsymbol{t}\|\to\infty} d_{jj} = \frac{1}{2}$, and a reconstruction of (A8) yields $\lim_{\|\boldsymbol{t}\|\to\infty} \mathrm{Var}\left(\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}\right) = \frac{k(k-1)}{2} = \binom{k}{2}$, completing the proof. ∎

Theorem A2: Assuming H_o is true, for any $\boldsymbol{t} \neq \mathbf{0}$, and for $\forall\, j$, as $n \to \infty$

$$\frac{\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N\left(0, 2\sigma_{C_{\boldsymbol{t}}}^2\right) \text{ and } \frac{\sqrt{n}\hat{\gamma}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N\left(0, 2\sigma_{S_{\boldsymbol{t}}}^2\right).$$

Proof:

We only present the proof for $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ as the steps remain similar for $\hat{\gamma}_{\boldsymbol{t}}^{(j)}$. We recall from Eq. A6 that $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ is a linear combination sample averages $\hat{\mu}_{C_{\boldsymbol{t}}}^{(1)}, \hat{\mu}_{C_{\boldsymbol{t}}}^{(2)}, \ldots, \hat{\mu}_{C_{\boldsymbol{t}}}^{(k)}$ where $\hat{\mu}_{C_{\boldsymbol{t}}}^{(j)} = \frac{1}{\mathrm{n}}\sum_{h=\mathbf{1}}^{n}\cos(\boldsymbol{t}'\mathrm{X}_h^{(j)})$ with $E[\,\hat{\mu}_{C_{\boldsymbol{t}}}^{(j)}] = \mu_{C_{\boldsymbol{t}}} < \infty$ and $\mathrm{Var}[\,\sqrt{n}\hat{\mu}_{C_{\boldsymbol{t}}}^{(j)}] = \sigma_{C_{\boldsymbol{t}}}^2 < \infty$, since $\cos(.)$ is a bounded function and $\left\{\hat{\mu}_{C_{\boldsymbol{t}}}^{(j)}\right\}_{j=1}^{k}$ is a sequence of independent and identically distributed (*iid*) random variables under H_o. Now by applying the strong central limit theorem on the sequence of *iid* random variables $\left\{\cos(\boldsymbol{t}'\mathrm{X}_h^{(j)})\right\}_{h=1}^{n}$, for any $\boldsymbol{t} \neq \mathbf{0}$, $\sqrt{n}\left(\hat{\mu}_{C_{\boldsymbol{t}}}^{(j)} - \mu_{C_{\boldsymbol{t}}}\right) \xrightarrow{D} N\left(0, \sigma_{C_{\boldsymbol{t}}}^2\right)$ as $n \to \infty$. Thus, as $n \to \infty$, $\sqrt{n}\left(\widehat{\boldsymbol{\mu}}_{C_{\boldsymbol{t}}} - \boldsymbol{\mu}_{C_{\boldsymbol{t}}}\right) \xrightarrow{D} N\left(\mathbf{0}, \sigma_{C_{\boldsymbol{t}}}^2 \mathrm{I}_k\right)$ where $\boldsymbol{\mu}_{C_{\boldsymbol{t}}} = (\mu_{C_{\boldsymbol{t}}}, \mu_{C_{\boldsymbol{t}}}, \ldots, \mu_{C_{\boldsymbol{t}}})'$. Notice that using Lemma A4-i, $\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}$ can be represented as $\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)} = \boldsymbol{g}_j'\sqrt{n}\left(\widehat{\boldsymbol{\mu}}_{C_{\boldsymbol{t}}} - \boldsymbol{\mu}_{C_{\boldsymbol{t}}}\right)$ with $E\left[\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}\right] = 0$, also $\mathrm{Var}\left(\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}\right) = k(k-1)\sigma_{C_{\boldsymbol{t}}}^2$ from Lemma A5. Hence, by using the properties of the multivariate normal distribution, we have $\lim_{n\to\infty} \sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)} \xrightarrow{D} N\left(0, k(k-1)\sigma_{C_{\boldsymbol{t}}}^2\right)$, or equivalently, $\lim_{n\to\infty} \frac{\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N\left(0, 2\sigma_{C_{\boldsymbol{t}}}^2\right)$. ∎

Corollary A2: $\lim_{\|\boldsymbol{t}\|\to\infty} \lim_{n\to\infty} \frac{\sqrt{n}\widehat{\beta}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N(0,1)$ and $\lim_{\|\boldsymbol{t}\|\to\infty} \lim_{n\to\infty} \frac{\sqrt{n}\widehat{\gamma}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N(0,1)$.

Proof: The proof follows immediately from applying the result in Lemma A3. ∎

Theorem A3: Regardless of the truth or falsity of H_o

$\lim_{\|\boldsymbol{t}\|\to\infty} \lim_{n\to\infty} \frac{\sqrt{n}\widehat{\beta}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N(0,1)$ and $\lim_{\|\boldsymbol{t}\|\to\infty} \lim_{n\to\infty} \frac{\sqrt{n}\widehat{\gamma}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N(0,1)$.

Proof:

Here also we only present the proof for $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ as the steps remain similar for $\hat{\gamma}_{\boldsymbol{t}}^{(j)}$. Assuming H_o, the proof is the same as for Corollary A2. To prove assuming the falsity of H_o, we proceed as follows.

Again, recall from Eq. A6 that $\hat{\beta}_{\boldsymbol{t}}^{(j)}$ is a linear combination of *independent* random variables $\hat{\mu}_{C_t}^{(1)}$, $\hat{\mu}_{C_t}^{(2)}$, …, $\hat{\mu}_{C_t}^{(k)}$ where $\hat{\mu}_{C_t}^{(j)} = \frac{1}{\mathrm{n}}\sum_{h=\mathbf{1}}^{n}\cos(\boldsymbol{t}'\mathrm{X}_h^{(j)})$ with $E[\,\hat{\mu}_{C_t}^{(j)}] = \mu_{C_t}^{(j)} < \infty$ and $\mathrm{Var}[\sqrt{n}\hat{\mu}_{C_t}^{(j)}] = \sigma_{j,C_t}^2 < \infty$, since $\cos(.)$ is a bounded function. Now by applying the strong central limit theorem on the sequence of *iid* random variables $\left\{\cos(\boldsymbol{t}'\mathrm{X}_h^{(j)})\right\}_{h=1}^{n}$, we have, for any $\boldsymbol{t} \neq \mathbf{0}$, $\sqrt{n}\left(\hat{\mu}_{C_t}^{(j)} - \mu_{C_t}^{(j)}\right) \xrightarrow{D} N\left(0, \sigma_{j,C_t}^2\right)$ as $n \to \infty$. Thus, as $n \to \infty$, $\sqrt{n}\left(\widehat{\boldsymbol{\mu}}_{C_t} - \boldsymbol{\mu}_{C_t}\right) \xrightarrow{D} N\left(\mathbf{0}, \mathbf{D}_{k,C}\right)$ where $\boldsymbol{\mu}_{C_t} = (\mu_{C_t}^{(1)}, \mu_{C_t}^{(2)}, \ldots, \mu_{C_t}^{(k)})'$ and $\mathbf{D}_{k,C_t}$ is a diagonal matrix as defined in the proof for Lemma A5. Notice that $\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}$ can be represented as

$\sqrt{n}\left(\hat{\beta}_{\boldsymbol{t}}^{(j)} - \beta_{\boldsymbol{t}}^{(j)}\right) = \boldsymbol{g}_j'\sqrt{n}\left(\widehat{\boldsymbol{\mu}}_{C_t} - \boldsymbol{\mu}_{C_t}\right)$ with $E\left[\sqrt{n}\left(\hat{\beta}_{\boldsymbol{t}}^{(j)} - \beta_{\boldsymbol{t}}^{(j)}\right)\right] = 0$ and

$\mathrm{Var}\left(\sqrt{n}\left(\hat{\beta}_{\boldsymbol{t}}^{(j)} - \beta_{\boldsymbol{t}}^{(j)}\right)\right) = \sigma_{j,C_t}^2$ from Eq. A9. Hence, by using the properties of the multivariate normal distribution, we have $\lim_{n\to\infty} \sqrt{n}\left(\hat{\beta}_{\boldsymbol{t}}^{(j)} - \beta_{\boldsymbol{t}}^{(j)}\right) \xrightarrow{D} N\left(0, \sigma_{\hat{\beta}_{\boldsymbol{t}}^{(j)}}^2\right)$. (A10)

Now by applying results in Lemma A4-ii and Lemma A6, we have

$\lim_{\|\boldsymbol{t}\|\to\infty}\lim_{n\to\infty} \sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)} \xrightarrow{D} N\left(0, \binom{k}{2}\right)$, or equivalently, $\lim_{\|\boldsymbol{t}\|\to\infty}\lim_{n\to\infty} \frac{\sqrt{n}\hat{\beta}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}} \xrightarrow{D} N(0,1)$. ■

In the following theorem we state a necessary and sufficient condition for the falsity of H_o as defined in Eq. A1.

Theorem A4: H_o, as stated in Eq. A1, is false if and only if $\max_{\boldsymbol{t}\in\mathbb{R}^p}\left\{\left\{\left|\beta_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}, \left\{\left|\gamma_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}\right\} > 0$.

Proof:

Let us first assume that H_o is false, then by *the continuity of the characteristic* function, there exists a region $\mathcal{R} \subset \mathbb{R}^p$ such that either

$$\mu_{C_t}^{(j)} \neq \mu_{C_t}^{(l)} \text{ or } \mu_{S_t}^{(j)} \neq \mu_{S_t}^{(l)} \tag{A11}$$

for some $j \neq l$ and all $\boldsymbol{t} \in \mathcal{R}$. Now assume that $\mu_{C_t}^{(j)} \neq \mu_{C_t}^{(l)}$ for some $j \neq l$ and all $\boldsymbol{t} \in \mathcal{R}$. Then it follows from Eq. A2 that $\boldsymbol{\delta}_{\boldsymbol{t}}^{(C)} \neq \mathbf{0} \Rightarrow \beta_{\boldsymbol{t}}^{(j)} \neq 0$ for some j, all $\boldsymbol{t} \in \mathcal{R}$. Since $\beta_{\boldsymbol{t}}^{(j)}$ and $\gamma_{\boldsymbol{t}}^{(j)}$ are continuous and bounded function in $\boldsymbol{t}$, we must either have $0 < \max_{\boldsymbol{t}\in\mathbb{R}^p} \left\{\left|\beta_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k} < \infty$ or $0 < \max_{\boldsymbol{t}\in\mathbb{R}^p} \left\{\left|\gamma_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k} < \infty$.

Let us now assume that $\max_{\boldsymbol{t}\in\mathbb{R}^p} \left\{\left\{\left|\beta_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}, \left\{\left|\gamma_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}\right\} > 0$. Specifically, suppose $\max_{\boldsymbol{t}\in\mathbb{R}^p} \left\{\left|\beta_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k} > 0$ which implies that $\beta_{\boldsymbol{t}}^{(j)} \neq 0$ for at least one j. Furthermore, as $\beta_{\boldsymbol{t}}^{(j)}$ is a continuous and bounded function in $\boldsymbol{t}$, there must exist a region $\mathcal{R} \subset \mathbb{R}^p$ such that $\beta_{\boldsymbol{t}}^{(j)} \neq 0$ for all $\boldsymbol{t} \in \mathcal{R}$. This necessitates that we must also have $\mu_{C_t}^{(j)} \neq \mu_{C_t}^{(l)}$ for some $j \neq l$, all $\boldsymbol{t} \in \mathcal{R}$, thereby proving the falsity of H_o. The proof is therefore complete. ■

The Test Statistic

The basic idea of the construction of our test is based on the results obtained in Theorems 3-4. First, recall from Lemma A4 that, for any $\boldsymbol{t} \in \mathbb{R}^p$, the statistics $\left\{\left\{\hat{\beta}_{\boldsymbol{t}}^{(j)}\right\}_{j=1}^{k}, \left\{\hat{\gamma}_{\boldsymbol{t}}^{(j)}\right\}_{j=1}^{k}\right\}$ have means $\left\{\left\{\beta_{\boldsymbol{t}}^{(j)}\right\}_{j=1}^{k}, \left\{\gamma_{\boldsymbol{t}}^{(j)}\right\}_{j=1}^{k}\right\}$, respectively, that vanish for large values of $\|\boldsymbol{t}\|$ when H_o is false. These parameters are exactly equal to zero for all $\boldsymbol{t} \in \mathbb{R}^{\boldsymbol{p}}$ when H_o holds true. Thus, the asymptotic distribution of these statistics are centered away from zero for small values of $\|\boldsymbol{t}\|$ when H_o is false. This implies that, for small $\|\boldsymbol{t}\|$, we expect their observed absolute values to be significantly larger from what are expected under the standard normal distribution (see *Remark 2* at the end). We use this basic rationale, together with the result in Theorem A4, to define our test statistic as follows.

Let $\mathbf{G}$ be a grid containing $\mathcal{L}$ vectors in $\mathbb{R}^p$ notated as $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{\mathcal{L}}$. Construction of a suitable grid is explained at end of this appendix. Using $\mathbf{G}$, we define a collection of statistics

$$\left\{ \left\{ \hat{\beta}_{\boldsymbol{t}}^{(j)} \right\}_{j=1}^{k} , \left\{ \hat{\gamma}_{\boldsymbol{t}}^{(j)} \right\}_{j=1}^{k} \right\}_{\boldsymbol{t}=\boldsymbol{t}_1}^{\boldsymbol{t}=\boldsymbol{t}_{\mathcal{L}}}. \tag{A12}$$

Notice that this collection consists of $\tilde{\mathcal{L}} = 2k\mathcal{L}$ random variables for $k > 2$. Notice that for $k = 2$, we have $\tilde{\mathcal{L}} = 2\mathcal{L}$ (see *Remark* 1). Let $\left(\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)}, \ldots, \mathbf{X}_i^{(k)};\ i = 1,2, \ldots, n\right)$ denote n independent observations from populations $F_{X^{(1)}}, F_{X^{(2)}}, \ldots, F_{X^{(k)}}$, respectively. For simplicity of exposition, we assume equal sample sizes $n_i = n,\ j = 1,2, \ldots, k$. Our ECF test statistic is then defined as $T_n = \frac{\sqrt{n}}{\sqrt{\binom{k}{2}}} \left[\max_{\boldsymbol{t} \in \mathbf{G}} \left\{ \left\{ \left| \hat{\beta}_{\boldsymbol{t}}^{(j)} \right| \right\}_{j=1}^{k} , \left\{ \left| \hat{\gamma}_{\boldsymbol{t}}^{(j)} \right| \right\}_{j=1}^{k} \right\} \right]$.

We notice from Theorems 2-3 that the statistics in collection A12 are all marginally asymptotically normal and, especially, are standard normal for large $\|\boldsymbol{t}\|$. It can be shown that these statistics are asymptotically correlated. However, in order to obtain a simple approximation to the distribution T_n, we proceed as if they are mutually asymptotically independent. Under this assumption, an asymptotic level $-\alpha$ rejection region given as

$$\mathrm{RJ}_\alpha = \left\{ \left(\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)}, \ldots, \mathbf{X}_i^{(k)};\ i = 1,2, \ldots, n \right) : T_n > \left| h(\alpha, \tilde{\mathcal{L}}) \right| \right\}, \tag{A13}$$

where the significance level α is the probability of observing at least one random number greater, in absolute value, than $\left| h(\alpha, \tilde{\mathcal{L}}) \right|$ out of $\tilde{\mathcal{L}}$ numbers generated randomly under $N(0,1)$, and $-h\left(\alpha, \tilde{\boldsymbol{\mathcal{L}}}\right) = \Phi^{-1}(\varphi)$, where $\Phi(.)$ is the standard normal distribution function and φ depends on both α and $\tilde{\mathcal{L}}$. Using the definition of α above, we can obtain the *critical value* $\left| h(\alpha, \tilde{\boldsymbol{\mathcal{L}}}) \right|$ as follows. We notice that $\alpha = 1 - (1 - 2\varphi)^{\tilde{\boldsymbol{\mathcal{L}}}}$, which yields $\varphi = \frac{1 - (1-\alpha)^{\frac{1}{\tilde{\mathcal{L}}}}}{2}$. So we can compute φ for a given α and $\tilde{\mathcal{L}}$, and then $h\left(\alpha, \tilde{\mathcal{L}}\right) = -\Phi^{-1}(\varphi)$.

Alternatively, the *p-value* of the test can be computed as $p-value = 1-[1-2\Phi(-T_n)]^{\tilde{L}}$. ■

The simulation study in Section 5.1.3 show that the ECF test given by A12 has correct level. We now show the consistency of our test as follows.

Theorem A5: The test defined in A12 is consistent against any alternative $H_a: F^{(j)}(.) \neq F^{(l)}(.)$, for some $j \neq l$.

Proof:

We first note from Theorem A4 that $\max_{\boldsymbol{t}\in\mathbb{R}^p}\left\{\left\{\left|\beta_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}, \left\{\left|\gamma_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}\right\} > 0$. In particular, because $\beta_{\boldsymbol{t}}^{(j)}$ and $\gamma_{\boldsymbol{t}}^{(j)}$ are continuous and bounded function of $\boldsymbol{t}$, there exists a region $\mathcal{R} \subset \mathbb{R}^p$ such that either $\beta_{\boldsymbol{t}}^{(j)} \neq 0$ or $\gamma_{\boldsymbol{t}}^{(j)} \neq 0$ for some j and all $\boldsymbol{t} \in \mathcal{R}$. We assume that the grid $\mathbf{G}$ is constructed such that $\mathbf{G}^* \equiv \mathbf{G} \cap \mathcal{R} \neq \{\}$. Furthermore, we suppose that $\mathbf{G}$ is such that either $\beta_{\boldsymbol{t}}^{(j)} \neq 0$ or $\gamma_{\boldsymbol{t}}^{(j)} \neq 0$ for some j and all $\boldsymbol{t} \in \mathbf{G}^*$. Let us also define

$\left|\beta_{\boldsymbol{t}^*}^{(j^*)}\right| = \max_{\boldsymbol{t}\in\mathbf{G}^*}\left\{\left|\beta_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}$ and $\left|\gamma_{\boldsymbol{t}^\circ}^{(j^\circ)}\right| = \max_{\boldsymbol{t}\in\mathbf{G}^*}\left\{\left|\gamma_{\boldsymbol{t}}^{(j)}\right|\right\}_{j=1}^{k}$, where the maxima occur respectively at some $\boldsymbol{t}^*$ and $\boldsymbol{t}^\circ$ contained in $\mathbf{G}^*$, for some j^* and j°. From here on we assume that $\left|\beta_{\boldsymbol{t}^*}^{(j^*)}\right| \geq \left|\gamma_{\boldsymbol{t}^\circ}^{(j^\circ)}\right|$.

Thus, we also have a statistic $\left|\hat{\beta}_{\boldsymbol{t}^*}^{(j^*)}\right|$ in the collection A12 such that

$\left|\hat{\beta}_{\boldsymbol{t}^*}^{(j^*)}\right| \xrightarrow{pr} \left|\beta_{\boldsymbol{t}^*}^{(j^*)}\right|$ and $\left(T_n - \frac{\sqrt{n}}{\sqrt{\binom{k}{2}}}\left|\hat{\beta}_{\boldsymbol{t}^*}^{(j^*)}\right|\right) \xrightarrow{pr} 0$. Also, by Eq. A10,

$\frac{\sqrt{n}}{\sqrt{\binom{k}{2}}}\hat{\beta}_{\boldsymbol{t}^*}^{(j^*)} \xrightarrow{D} N\left(\beta_{\boldsymbol{t}^*}^{(j^*)}, \sigma^2_{\beta_{\boldsymbol{t}^*}^{(j^*)}}\right) \Rightarrow T_n \xrightarrow{D} N\left(\beta_{\boldsymbol{t}^*}^{(j^*)}, \sigma^2_{\beta_{\boldsymbol{t}^*}^{(j^*)}}\right)$. Therefore, as $n \to \infty$

$$P_{H_a}\left(T_n > \left|h(\alpha, \tilde{M})\right|\right) = P_{H_a}\left(\frac{\sqrt{n}}{\sqrt{\binom{k}{2}}}\hat{\beta}_{\boldsymbol{t}^*}^{(j^*)} > h(\alpha, \tilde{M})\right) + P_{H_a}\left(\frac{\sqrt{n}}{\sqrt{\binom{k}{2}}}\hat{\beta}_{\boldsymbol{t}^*}^{(j^*)} < -h(\alpha, \tilde{M})\right)$$

$$= 1 - P_{H_a}\left(\frac{\sqrt{n}}{\sqrt{\binom{k}{2}}}\frac{\widehat{\beta}_{t^*}^{(j^*)} - \beta_{t^*}^{(j^*)})}{\sigma^2_{\beta_{t^*}^{(j^*)}}} \leq h(\alpha, \widetilde{M}) - \frac{\sqrt{n}\beta_{t^*}^{(j^*)}}{\sqrt{\binom{k}{2}}\sigma^2_{\beta_{t^*}^{(j^*)}}}\right) +$$

$$P_{H_a}\left(\frac{\sqrt{n}}{\sqrt{\binom{k}{2}}}\frac{\widehat{\beta}_{t^*}^{(j^*)} - \beta_{t^*}^{(j^*)})}{\sigma^2_{\beta_{t^*}^{(j^*)}}} \leq -h(\alpha, \widetilde{M}) - \frac{\sqrt{n}\beta_{t^*}^{(j^*)}}{\sqrt{\binom{k}{2}}\sigma^2_{\beta_{t^*}^{(j^*)}}}\right)$$

$$= 1 - \Phi\left(\lim_{n\to\infty} h(\alpha, \widetilde{M}) - \frac{\sqrt{n}\beta_{t^*}^{(j^*)}}{\sqrt{\binom{k}{2}}\sigma^2_{\beta_{t^*}^{(j^*)}}}\right) + \Phi\left(\lim_{n\to\infty} -h(\alpha, \widetilde{M}) - \frac{\sqrt{n}\beta_{t^*}^{(j^*)}}{\sqrt{\binom{k}{2}}\sigma^2_{\beta_{t^*}^{(j^*)}}}\right)$$

$= 1 - \Phi(-\infty) + \Phi(-\infty)$ =1 if $\beta_{t^*}^{(j^*)} > 0$, and,

$= 1 - \Phi(\infty) + \Phi(\infty)$ =1 if $\beta_{t^*}^{(j^*)} < 1$. ∎

We now state and prove the following theorem that we use in our concluding remark to further explain the construction our ECF test.

Theorem A6: Let $\mathcal{F}$ denote the class of all distribution functions, then

$\sup_{\mathcal{F}, \boldsymbol{t}\in\mathbb{R}^p} \mathrm{Var}[\cos(\boldsymbol{t}'\mathbf{X})] = \sup_{\mathcal{F}, \boldsymbol{t}\in\mathbb{R}^p} \mathrm{Var}[\sin(\boldsymbol{t}'\mathbf{X})] = 1$, where the random vector **X** has distribution function $F(.) \in \mathcal{F}$.

Proof:

Seaman et al. (1992) have shown that for any real-valued, measurable and essentially bounded function f, $\sup_{\mathcal{F}} \mathrm{Var}[f(\mathbf{X})] \leq (b-a)^2/4$, where $b = \operatorname{ess\,sup} f$ and $a = \operatorname{ess\,inf} f$. Hence the proof follows by noticing that $cos(\mathbf{X})$ and $sin(\mathbf{X})$ also belong to the class of real-valued, measurable and essentially bounded functions with $\operatorname{ess\,sup} cos = \operatorname{ess\,sup} sin = 1$ and $\operatorname{ess\,inf} cos = \operatorname{ess\,inf} sin = -1$. ∎

***Remark 2*:** The critical value $h(\alpha, \widetilde{M})$ in our ECF test A13 is determined based on the standard normal distribution, the asymptotic distribution of $\frac{\sqrt{n}\widehat{\beta}_{t}^{(j)}}{\sqrt{\binom{k}{2}}}$ for large **t**, regardless of the truth or falsity of H_0. However, under H_0, the asymptotic distribution is not stand-

ard normal, rather, it is $N\left(0, 2\sigma_{C_t}^2\right)$. From Theorem A6 , we notice that the variance $2\sigma_{C_t}^2$ has an upper bound of 2. For many continuous distributions this variance remains either close to 1 or less than 1 when for small values of $\|\boldsymbol{t}\|$. Therefore, we choose $N(0,1)$ as an approximation to the asymptotic distribution of $\frac{\sqrt{n}\widehat{\beta}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}}$ for small $\|\boldsymbol{t}\|$. This approximation works for $\frac{\sqrt{n}\widehat{\gamma}_{\boldsymbol{t}}^{(j)}}{\sqrt{\binom{k}{2}}}$ as well. Our simulation results in Section 5.1.3 show that the rejection region A13 still yields a level-alpha test. Furthermore, the test is *omnibus*, i.e. remains consistent against any alternative as shown in Theorem A5 and enjoys good statistical power as shown in Section 5.1.3.

Construction of the Grid

Here we explain the construction of a p-dimensional grid of vectors $\boldsymbol{t}$ such that $\|\boldsymbol{t}\|$ remains sufficiently small. Let $\mathbf{G} = \overbrace{\boldsymbol{g} \times \boldsymbol{g} \times \ldots \times \boldsymbol{g}}^{p}$ be the grid, where $\boldsymbol{g} = (g_1, g_2, \ldots, g_M)'$, and the scalars g_i are such that $g_1 < g_2 < \cdots < g_M$, $g_i \in \mathbb{R}^+$, $i = 1, 2, \ldots, M$. Notice that $\mathbf{G}$ contains $\mathcal{L} \equiv M^p$ vectors in $\mathbb{R}^p$. Furthermore, all these vectors lie within a p-dimensional sphere of radius $R = \sqrt{Mg_M^2}$, i.e. $\|\boldsymbol{t}\| < R, \ \forall\, \boldsymbol{t} \in \mathbf{G}$. A suitable value of g_M, for which $\|\boldsymbol{t}\|$ does not grow too large, can be chosen by plotting the real and imaginary parts of the k ECFs as a function of $\|\boldsymbol{t}\|$. That is, we choose g_M so that all the k estimated real and imaginary parts are substantially small beyond $\|\boldsymbol{t}\| = \sqrt{Mg_M^2}$. Although vectors in $\mathbf{G}$ lie in the positive orthant, the grid can be expanded in other orthants as well. We choose the positive orthant because (i) the characteristic function is *Hermitian*, and (ii) our simulations show that the test still enjoys good level and power properties with this choice.

Bibliography

Abramowitz, M. and Stegun, I. A. (1965), *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, New York: Dover publications.

Agresti, A. (2002), *Categorical Data Analysis*, New York: John Wiley & Sons.

Akçakaya, H. R. (2000), "Population viability analyses with demographically and spatially structured models," *Ecological Bulletins* ,48, 23-38.

Akçakaya, H. R., Mills, G. and Doncaster, C. P. (2007), "The role of metapopulations in conservation," in Key Topics in Conversation Biology, eds. D. Macdonald and J. Service, Oxford, UK: Blackwell Publishing, pp. 64-84.

Akçakaya, H. R. and Raphael, M. G. (1998), "Assessing human impact despite uncertainty: viability of the northern spotted owl metapopulation in the northwestern USA," *Biodiversity and Conservation*, 7, 875-894.

Alba-Fernandéz, V., Jiménez-Gamero, M. and Muñoz-García, J., (2006), "Goodness-of-fit tests based on the empirical characteristic function," *In COMPSTAT*, pp. 1059-1066.

Andrieu, C., Doucet, A. and Holenstein, R. (2010), "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B*, 72, 269-342.

Arcese, P. and Marr, A. B. (2006), "Population viability in the presence and absence of cowbirds, catastrophic mortality, and immigration," in *Conservation and Biology of Small Populations: the Song Sparrows of Mandarte Island*, eds. J. N. M. Smith et al., Oxford, UK: Oxford University Press, pp. 175-191.

Arcese, P., Smith, J. N., Hochachka, W. M., Rogers, C. M. and Ludwig, D. (1992), "Stability, regulation, and the determination of abundance in an insular Song Sparrow population," *Ecology* , 805-822.

Baghishani, H., Rue, H. and Mohammadzadeh, M. (2012), "On a hybrid data cloning method and its application in generalized linear mixed models," *Statistics and Computing*, 22, 597-613.

Barker, D. and Sibly, R. M. (2008), "The effects of environmental perturbation and measurement error on estimates of the shape parameter in the theta-logistic model of population regulation," *Ecological Modeling* , 219, 170-177.

Barnett, V. (1999), *Comparative Statistical Inference*, New York: John Wiley & Sons.

Beissinger, S. R. (2002), "Population viability analysis: past, present and future," in *Population Viability Analysis,* eds. S. R. Beissinger and D. R. Maccullough, Chicago: University of Chicago Press, pp. 5-17.

Bennett, C. H. (1976), "Efficient estimation of free energy differences from Monte Carlo data," *Journal of Compuational Physics*, 22, 245-268.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer.

Bernardo, J. M. and Smith, A. F. (2001), *Bayesian Theory*, New York: John Wiley & Sons.

Bianconcini, S. and Cagnone, S. (2012), "Estimation of generalized linear latent variable models via fully exponential Laplace approximation," *Journal of Multivariate Analysis*, 112, 183-193.

Bjørnstad, O. N., Begon, M., Stenseth, N. C., Falck, W., Sait, S. M. and Thompson, D. J. (1998), "Population dynamics of the Indian meal moth: demographic stochasticity and delayed regulatory mechanisms," *Journal of Animal Ecology* , 67, 110-126.

Booth, J. G. and Hobert, J. P. (1999), "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm," *Journal of the Royal Statistical Society: Series B*, 61, 265-285.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9-25.

Breslow, N. E. and Lin, X. (1995), "Bias correction in generalised linear mixed models with a single component of dispersion," *Biometrika* , 82, 81-91.

Brooks, S. P. and Gelman, A. (1998), "General methods for monitoring convergence of iterative simulations," *Journal Computational and Graphical Statistics*, 7, 434-455.

Brooks, S. P. and Giudici, P. (2000), "Markov chain Monte Carlo convergence assessment via two-way analysis of variance, *Journal Computational and Graphical Statistics*", 9, 266-285.

Brooks, S. P. and Morgan, B. J. T. (1995), "Optimization Using Simulated Annealing," *The Statistician*, 44, 241-257.

Brooks, S. P. and Roberts, G. O. (1998), "Assessing convergence of Markov chain Monte Carlo algorithms," *Statistics and Computing*, 8, 319-335.

Bryk, A. S. and Raudenbush, S. W. (1988), "Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model," *American Journal of Education*, 97, 65-108.

Burnham, K. P. and Anderson, D. R. (2004), "Multimodel inference understanding AIC and BIC in model selection," *Sociological Methods and Research*, 33, 261-304.

Cagnone, S. and Monari, P. (2012), "Latent variable models for ordinal data by using the adaptive quadrature approximation," *Computational Statistics*, 1-23.

Campbell, D. and Lele, S. R. (in press), "An ANOVA Test for Parameter Estimability using Data Cloning with Application to Statistical Inference for Dynamic Systems,".

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, London, UK: Chapman & Hall.

Chauveau, D. and Diebolt, J. (1997), "MCMC Convergence Diagnostic via the Central Limit Theorem," Techical Report 22, Université Paris-Est Marne-la-Vallée.

Clark, F., Brook, B. W., Delean, S., Reşit Akçakaya, H. and Bradshaw, C. J. (2010), "The theta-logistic is unreliable for modeling most census data," *Methods in Ecology and Evolution*, 1, 253-262.

Clark, J. S. and Bjørnstad, O. N. (2004), "Population time series: process variability, observation errors, missing values, lags, and hidden states," *Ecology* , 85, 3140-3150.

Clark, J. S. and Gelfand, A. eds. (2006), *Hierarchical Modeling for the Environmental Sciences: Statistical Methods and Applications*, New York: Oxford University Press.

Clayton, D. and Kaldor, J. (1987), "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping," *Biometrics* , 671-681.

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.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge UK: Cambridge University Press.

Craiu, R. V., Rosenthal, J. and Yang, C. (2009), "Learn from thy neighbor: Parallel-chain and regional adaptive MCMC," *Journal of the American Statistical Association*, 104, 1454-1466.

Creel, S. and Creel, M. (2009), "Density dependence and climate effects in Rocky Mountain elk: an application of regression with instrumental variables for population time series with sampling error," *Journal of Animal Ecology* , 78, 1291-1297.

Crowder, M. J. (1978), "Beta-binomial Anova for proportions," *Applied Statistics*, 27, 34-37.

Darmois, G. (1935), "Sur les lois de probabilites a estimation exhaustive," *C.R. Acad. sci. Paris* , 200, 1265-1266.

Datta, G. S. (1996), "On priors providing frequentist validity of Bayesian inference for multiple parametric functions," *Biometrika* , 83, 287-298.

de Valpine, P. (2008), "Improved estimation of normalizing constants from Markov chain Monte Carlo output," *Journal of Computational and Graphical Statistics*, 17, 333-351.

de Valpine, P. (2002), "Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state-space models," *Bulletin of Marine Science* , 70, 455-471.

de Valpine, P. and Hastings, A. (2002), "Fitting population models incorporating process noise and observation error," *Ecological Monographs*, 72, 57-76.

Demidenko, E. (2004), *Mixed Models: Theory and Applications*, New York: Wiley-Interscience.

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*, 39, 1-38.

Dennis, B. and Kemp, W. P. Taper, Mark L. (1998), "Joint density dependence," *Ecology*, 79, 426-441.

Dennis, B., Munholland, P. L. and Scott, J. M. (1991), "Estimation of growth and extinction parameters for endangered species," *Ecological Monographs* , 61, 115-143.

Dennis, B. and Otten, M. R. (2000), "Joint effects of density dependence and rainfall on abundance of San Joaquin kit fox," *The Journal of Wildlife Mangement*, 64, 388-400.

Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L. and Staples, D. F. (2006), "Estimating density dependence, process noise, and observation error," *Ecological Monographs* , 76, 323-341.

Dennis, B. and Taper, M. L. (1994), "Density dependence in time series observations of natural populations: estimation and testing," *Ecological Monographs* , 64, 205-224.

Diggle, P., Liang, K. and Zeger, S. (1994), *Analysis of Longitudinal Data*, Oxford. UK: Oxford University Press.

Diggle, P. J., Tawn, J. and Moyeed, R. (2002), "Model-based geostatistics," *Journal of the Royal Statistical Society: Series B*, 47, 299-350.

Doucet, A., De Freitas, N. and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, New York: Springer.

Doucet, A., Godsill, S. J. and Robert, C. P. (2002), "Marginal maximum a posteriori estimation using Markov chain Monte Carlo," *Statistics and Computing*, 12, 77-84.

Dunning Jr, J. B., Stewart, D. J., Danielson, B. J., Noon, B. R., Root, T. L., Lamberson, R. H. and Stevens, E. E. (1995), "Spatially explicit population models: current forms and future uses," *Ecological Applications* , 5, 3-11.

Eberly, L. E. and Carlin, B. P. (2000), "Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models," *Statistics in Medicine* , 19, 2279-2294.

Engel, B. (1998), "A simple illustration of the failure of PQL, IRREML and APHL as approximate ml methods for mixed models for binary data," *Biometrical Journal*, 40, 141-154.

Engen, S., Bakke, Ø. and Islam, A. (1998), "Demographic and environmental stochasticity-concepts and definitions," *Biometrics* , 54, 840-846.

Fan, Y. (1997), "Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function," *Journal of Multivariate Analysis*, 62, 36-63.

Freckleton, R. P., Watkinson, A. R., Green, R. E. and Sutherland, W. J. (2006), "Census error and the detection of density dependence," *Journal of Animal Ecology* , 75, 837-851.

Fryxell, J. M., Falls, J. B., Falls, E. A. and Brooks, R. J. (1998), "Long-term dynamics of small-mammal populations in Ontario," *Ecology* , 79, 213-225.

Gelfand, A. E. and Carlin, B. P. (1993), "Maximum-likelihood estimation for constrained-or missing-data models," *Canadian Journal of Statistics* , 21, 303-311.

Gelfand, A. E. and Sahu, S. K. (1999), "Identifiability, improper priors, and Gibbs sampling for generalized linear models," *Journal of the American Statistical Association*, 94, 247-253.

Gelfand, A. E. and Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003), *Bayesian Data Analysis*, London, UK: Chapman & Hall.

Gelman, A. and Meng, X. (1998), "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Statistical Sceince*, 13, 163-185.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Sceince*, 7, 457-472.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6, 721-741.

Geweke, J. (1996), *Handbook of Computational Economics*, Amsterdam: North-Holand, chap. 15.

Geweke, J. (1989), "Bayesian inference in econometric models using Monte Carlo integration," IEconometrica, 57, 1317-1339.

Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society: Series B*, 54, 657-699.

Giakoumatos, S., Vrontos, I., Dellaportas, P. and Politis, D. (1999), "A Markov chain Monte Carlo convergence diagnostic using subsampling," *Journal of Computational and Graphical Statistics*, 8, 431-451.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Gilpin, M. E. and Ayala, F. J. (1973), "Global models of growth and competition," *Proceedingsof the Natural Academy of Science* , 70, 3590-3593.

Goldman, E., Valiyeva, E. and Tsurumi, H. (2008), "Kolmogorov–Smirnov, Fluctuation, and Z g Tests for Convergence of Markov Chain Monte Carlo Draws," *Communications in Statistics-Simulations and Computation*, 37, 368-379.

Gompertz, B. (1825), "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies," *Philosophical Transactions of the Royal Statistical Society*, 115, 513-583.

Green, P. J. (1987), "Penalized likelihood for general semi-parametric regression models," *International Statistical Review*, 55, 245-259.

Grimm, V. and Wissel, C. (2004), "The intrinsic mean time to extinction: a unifying approach to analysing persistence and viability of populations," *Oikos* , 105, 501-511.

Groom, M. J. and Pascual, M. A. (1997), "The analysis of population persistence: an outlook on the practice of viability analysis," in *Conservation Biology for the Coming Decade*, eds. P. L. Fiedler and P. M. Kareiva, New York: Springer, pp. 4-27.

Gustafson, P. (2009), "What are the limits of posterior distributions arising from nonidentified models, and why should we care?" *Journal of the American Statistical Association*,104, 1682-1695.

Hamilton, J. D. (1986), "A standard error for the estimated state vector of a state-space model," *Journal of Econometrics* , 33, 387-397.

Hammersley, J. M. and Handscomb, D. C. (1964), *Monte Carlo Methods*, New York: John Wiley & Sons.

Harris, I. R. (1989), "Predictive fit for natural exponential families," *Biometrika* , 76, 675-684.

Hart, E. M. and Gotelli, N. J. (2011), "The effects of climate change on density-dependent population dynamics of aquatic invertebrates," *Oikos* , 120, 1227-1234.

Harvey, A. C. (1993), *Time Series Models*, Cambridge: The MIT Press.

Hassell, M. P., Lawton, J. and May, R. (1976), "Patterns of dynamical behaviour in single-species populations," *Journal of Animal Ecology*, 45, 471-486.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika* , 57, 97-109.

Hedeker, D. and Gibbson, R. D. (2006), *Longitudinal Data Analysis*, New Jersey: Wiley-Interscience.

Henderson, C. R. (1950), "Estimation of genetic parameters," *Annals of Mathematical Statistics*, 21, 309-310.

Henderson, C. R., Kempthorne, O., Searle, S. R. and Von Krosigk, C. (1959), "The estimation of environmental and genetic trends from records subject to culling," *Biometrics*, 15, 192-218.

Henle, K., Sarre, S. and Wiegand, K. (2004), "The role of density regulation in extinction processes and population viability analysis," *Biodiversity and Conservation*, 13, 9-52.

Hobert, J. P. (2000), "Hierarchical models: A current computational perspective," *Journal of the American Statistical Association*, 95, 1312-1316.

Hochachka, W. M., Smith, J. N. M. and Arcese, P., eds. (1989), *Song Sparrow*, London, UK: Academic Press.

Huang, Y. and Huwang, L. (2008), "On the polynomial structural relationship," *Canadian .Journal of Statistics*, 29, 495-512.

Hušková, M. and Meintanis, S. G. (2008), "Tests for the multivariate k-sample problem based on the empirical characteristic function," *Journal of Nonparametric Statistics*, 20, 263-277.

Jacquier, E., Johannes, M. and Polson, N. (2007), "MCMC maximum likelihood for latent state models," *Journal of Econometrics*, 137, 615-640.

Jank, W. and Booth, J. (2003), "Efficiency of Monte Carlo EM and simulated maximum likelihood in two-stage hierarchical models," *Journal of Computational and Graphical Statistics*, 12, 214-229.

Jensen, J. L. (1906), "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, 30, 175-193.

Jiang, J. (2000), "A nonlinear Gauss-Seidel algorithm for inference about GLMM," *Computational Statistics*, 15, 229-24.

Joe, H. (2008), "Accuracy of Laplace approximation for discrete response mixed models," *Computational Stattistics and Data Analysis*, 52, 5066-5074.

Johansen, A. M., Doucet, A. and Davy, M. (2008), "Particle methods for maximum likelihood estimation in latent variable models," *Statistics and Computing*, 18, 47-57.

Johnson, R. and Wichern, D. (2007), *Applied Multivariate Statistical Analysis*, New Jersey: Prentice Hall.

Jonzen, N., Pople, A. R., Grigg, G. C. and Possingham, H. P. (2005), "Of sheep and rain: large-scale population dynamics of the red kangaroo," *Journal of Animal Ecology* , 74, 22-30.

Kantas, N., Doucet, A., Singh, S. S. and Maciejowski, J. M., (2009), "An overview of sequential Monte Carlo methods for parameter estimation in general state-space models," *In: Proc. IFAC Symposium on System Identification (SYSID)*.

Karim, M. R. and Zeger, S. L. (1992), "Generalized linear models with random effects; salamander mating revisited," *Biometrics* , 48, 631-644.

Kitagawa, G. (1987), "Non-Gaussian State—Space Modeling of Nonstationary Time Series," *Journal of the American Statistical Association*, 82, 1032-1041.

Kuk, A. Y. (2003), "Automatic choice of driving values in Monte Carlo likelihood approximation via posterior simulations," *Statistics and Computing*, 13, 101-109.

Laird, N. (1978), "Empirical Bayes methods for two-way contingency tables," *Biometrika* , 65, 581-590.

Lange, K. (2004), *Optimization*, New York: Springer-Verlag.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006), *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*, London, UK: Chapman & Hall.

Lele, S. R. (2010), "Model complexity and information in the data: Could it be a house built on sand?" *Ecology* , 91, 3493-3496.

Lele, S. R. (2006), "Sampling variability and estimates of density dependence: a composite-likelihood approach," *Ecology* , 87, 189-202.

Lele, S. R. and Dennis, B. (2009), "Bayesian methods for hierarchical models: Are ecologists making a Faustian bargain," *Ecological Applications*, 19, 581-584.

Lele, S. R., Dennis, B. and Lutscher, F. (2007), "Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods," *Ecology Letters*, 10, 551-563.

Lele, S. R., Nadeem, K. and Schmuland, B. (2010), "Estimability and likelihood inference for generalized linear mixed models using data cloning," *Journal of the American Statistical Association*, 105, 1617-1625.

Lele, S., Taper, M. L. and Gage, S. (1998), "Statistical analysis of population dynamics in space and time using estimating functions," *Ecology* , 79, 1489-1502.

Li, J. and Liu, R. Y. (2004), "New nonparametric tests of multivariate locations and scales using data depth," *Statistical Science*, 19, 686-696.

Lin, T. and Zha, H. (2008), "Riemannian manifold learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30, 796-809.

Lin, X. and Breslow, N. E. (1996), "Bias correction in generalized linear mixed models with multiple components of dispersion," *Journal of the American Statistical Association*, 91, 1007-1016.

Liu, Z. and Modarres, R. (2011), "A triangle test for equality of distribution functions in high dimensions," *Journal of Nonparametric Statistics*, 23, 605-615.

Ludwig, D. (1999), "Is it meaningful to estimate a probability of extinction?" *Ecology*, 80, 298-310.

Luis, A. D., Douglass, R. J., Mills, J. N. and Bjørnstad, O. N. (2010), "The effect of seasonality, density and climate on the population dynamics of Montana deer mice, important reservoir hosts for Sin Nombre hantavirus," *Journal of .Animal Ecology* , 79, 462-470.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*. London, UK: Chapman & Hall.

McCulloch, C. E. (1997), "Maximum likelihood algorithms for generalized linear mixed models," *Journal of the American Statistical Association*, 92, 162-170.

McCulloch, C. E. and Searle, S. R. (2001), *Generalized, Linear and Mixed Models*, New York: John Wiley & Sons.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008), *Generalized linear mixed models*, Wiley Online Library.

McGowan, C. P., Runge, M. C. and Larson, M. A. (2011), "Incorporating parametric uncertainty into population viability analysis models," *Biological Conservation* , 144, 1400-1408.

Meeker, W. Q. and Escobar, L. A. (1995), "Teaching about approximate confidence regions based on maximum likelihood estimation," *The American Statistician*, 49, 48-53.

Meintanis, S. and Swanepoel, J. (2007), "Bootstrap goodness-of-fit tests with estimated parameters based on empirical transforms," *Statistics and Probability Letters*, 77, 1004-1013.

Mengersen, K. L., Robert, C. P. and Guihenneuc-Jouyaux, C. (1999), *"MCMC convergence diagnostics: a "reviewww (with discussion)","* in *Bayesian Statistics*, eds. J. Bernardo et al.,Oxford, UK: Oxford University Press, pp. 415-440.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), "Equation of state calculations by fast computing machines," *The Journal of chemical physics* , 21, 1087-1091.

Mills, L. S. (2008), *Conservation of Wildlife Populations: Demography, Genetics, and Management*, Oxford, UK: Blackwell Publishing.

Monahan, J. and Genz, A. (1997), "Spherical-radial integration rules for Bayesian computation," *Journal of the American Statistical Association*, 92, 664-674.

Morris, W. F. and Doak, D. F. (2002), *Quantitative Conservation Biology: Theory and Practice of Population Viability Analysis*, Massachusetts, USA: Sinauer Associates.

Murray, L., (2010), "Distributed Markov chain Monte Carlo," *In: Proceedings of Neural Information Processing Systems Workshop on Learning on Cores, Clusters and Clouds*.

Nadeem, K. and Lele, S. R. (2012), "Likelihood based population viability analysis in the presence of observation error," *Oikos* , 121, 1656-1664.

Nelder, J. A. and Wedderburn, R. W. (1972), "Generalized linear models," *Journal of the Royal Statistical Society: Series A*, 135, 370-384.

Newman, K., Buckland, S., Lindley, S., Thomas, L. and Fernández, C. (2006), "Hidden process models for animal population dynamics," *Ecological Applications* , 16, 74-86.

Noh, M. and Lee, Y. (2007), "REML estimation for binary data in GLMMs," *Journal of Multivariate Analysis*, 98, 896-915.

Pascual, M. A., Kareiva, P. and Hilborn, R. (1997), "The influence of model structure on conclusions about the viability and harvesting of Serengeti wildebeest," *Conservation Biology* , 11, 966-976.

Patterson, H. D. and Thompson, R. (1971), "Recovery of inter-block information when block sizes are unequal," *Biometrika* , 58, 545-554.

Paulino, C. D. M. and Pereira, C .A .B. (1994), "On identifiability of parametric statistical models," *Statistical Methods and Applications*, 3, 125-151.

Pawitan, Y. (2001), *In All Likelihood: Statistical Modeling and Inference using Likelihood*, Oxford UK: Oxford University Press.

Pedersen, M. W., Berg, C. W., Thygesen, U. H., Nielsen, A. and Madsen, H. (2011), "Estimation methods for nonlinear state-space models in ecology," *Ecological Modeling* , 222, 1394-1400.

Pinheiro, J. C. and Bates, D. M. (1995), "Approximations to the log-likelihood function in the nonlinear mixed-effects model," *Journal of Computational and Graphical Statistics*, 4, 12-35.

Pinheiro, J. C. and Chao, E. C. (2006), "Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models," *Journal of Computational and Graphical Statistics*, 15, 58-81.

Pitman, E., (1936), "Sufficient statistics and intrinsic accuracy," *In: Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 567-579.

Pitt, M. K. (2002), "Smooth particle filters for likelihood evaluation and maximisation," Working Paper, University of Warwick, Coventry.

Plummer, M. (a) (2011), "JAGS version 3.0.0 user manual," URL http://mcmc-jags.sourceforge.net.

Plummer, M. (b) (2011), "rjags: Bayesian graphical models using MCMC," *R package version 2: 0-4*, URL http://CRAN. R-project. org/package= rjags.

Plummer, M. (2003), "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," URL http://citeseer. ist. psu. edu/plummer03jags. html.

Polansky, L., De Valpine, P., Lloyd-Smith, J. O. and Getz, W. M. (2009), "Likelihood ridges and multimodality in population growth rate models," *Ecology* , 90, 2313-2320.

Politis, D. N. and Romano, J. P. (1994), "Large sample confidence regions based on subsamples under minimal assumptions," *The Annals of Statistics* , 22, 2031-2050.

Politis, D. N., Romano, J. P. and Wolf, M. (1997), "Subsampling for heteroskedastic time series," *Journal of Econometrics* , 81, 281-317.

Polson, N. G. (1996), *"Convergence of Markov chain Monte Carlo algorithms,"*, Oxford: Oxford University Press.

Ponciano, J. M., Taper, M. L., Dennis, B. and Lele, S. R. (2009), "Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning," *Ecology* , 90, 356-362.

Press, S. (2003), *Subjective and Objective Bayesian Statistics*, Hoboken, NJ: Wiley.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005), "Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects," *Journal of Econometrics*, 128, 301-323.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002), "Reliable estimation of generalized linear mixed models using adaptive quadrature," *The Stata Journal*, 2, 1-21.

Ralls, K., Beissinger, S. R. and Cochrane, F. J., eds. (2002), *Guidelines for Using Population Viability Analysis in Endangered-Species Management*, Chicago: University of Chicago Press.

Raudenbush, S. W. and Bryk, A. S. (2002), *Hierarchical Linear Models: Applications and Data Analyses Methods*, Thousand Oaks: CA Sage.

Raudenbush, S. W., Yang, M. and Yosef, M. (2000), "Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation," *Journal of Computational and Graphical Statistics*, 9, 141-157.

Ricker, W. E. (1954), "Stock and recruitment," *Journal of the Fisheries Board of Canada*, 11, 559-623.

Rizopoulos, D., Verbeke, G. and Lesaffre, E. (2009), "Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data," *Journal of the Royal Statistical Society: Series B*, 71, 637-654.

Robert, C. P. (1993), "Prior Feedback: Bayesian Tools for Maximum Likelihood Estimation," *Journal of Computational Statistics*, 8, 279-294.

Robert, C. P. and Casella, G. (2005), *Monte Carlo Statistical Methods*, New York : Springer.

Robert, C. P., Tobias. RydÉn and Titterington, D. (1999), "Convergence controls for MCMC algorithms, with applications to hidden Markov chains," *Journal of Statistical Computation and Simulation* , 64, 327-355.

Robinson, G. K. (1991), "That BLUP is a good thing: The estimation of random effects," *Statistical Science*, 6, 15-32.

Rodriguez-Fernandez, M., Mendes, P. and Banga, J. R. (2006), "A hybrid approach for efficient and robust parameter estimation in biochemical pathways," *BioSystems* , 83, 248-265.

Rosenthal, J. S. (2006), *A First Look at Rigorous Probability Theory*, Singapore: World Scientific Publishing Company Incorporated.

Rosenthal, J. S. (1993), "Rates of convergence for data augmentation on finite sample spaces," *The Annals of Applied Probability*, 3, 819-839.

Royle, J. A. and Dorazio, R. M. (2008), *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*, London, UK: Academic Press.

Rue, H., Martino, S. and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal StatisticalSociety: Series B* , 71, 319-392.

Sæther, B., Engen, S., Islam, A., McCleery, R. and Perrins, C. (1998), "Environmental stochasticity and extinction risk in a population of a small songbird, the great tit," *American Naturalist*, 151, 441-450.

Saether, B., Engen, S., Lande, R., Arcese, P. and Smith, J. (2000), "Estimating the time to extinction in an island population of song sparrows," *Proceedings of Royal Statistical Society B: Biological Sciences*, 267, 621.

Saether, B., Lillegård, M., Grøtan, V., Filli, F. and Engen, S. (2007), "Predicting fluctuations of reintroduced ibex populations: the importance of density dependence, environmental stochasticity and uncertain population estimates," *Journal of Animal Ecology*, 76, 326-336.

Schall, R. (1991), "Estimation in generalized linear models with random effects," *Biometrika*, 78, 719-727.

Schervish, M. J. and Carlin, B. P. (1992), "On the convergence of successive substitution sampling," *Journal of Computational and Graphical statistics*, 1, 111-127.

Schittkowski, K. (2007), "Experimental design tools for ordinary and algebraic differential equations," *Industrial and Engineering Chemistry Research*, 46, 9137-9147.

Schnute, J. T. (1994), "A general framework for developing sequential fisheries models," *Canadian Journal of Fisheries and Aquatic Sciences*, 51, 1676-1688.

Schtickzelle, N. and Baguette, M. (2004), "Metapopulation viability analysis of the bog fritillary butterfly using RAMAS/GIS," *Oikos* , 104, 277-290.

Seaman, J., Young, D. and Turner, D. (1992), "On the maximum variance of a bounded random variable," *International Journal of Mathematical Education in Science and Technology*, 23, 130-131.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992), *Variance Components*, New York: John Wiley and Sons.

Shaffer, M. L. (1981), "Minimum population sizes for species conservation," *Bioscience*, 31, 131-134.

Smith, J. N. M., ed. (1988), *Determinants of Lifetime Reproductive Success in the Song Sparrow*, Chicago: University of Chicago Press.

Smith, J. N. M. and Arcese, P. (1989), "How fit are floaters? Consequences of alternative territorial behaviors in a nonmigratory sparrow," *American Naturalist*, 133, 830-845.

Smith, J. N., Keller, L. F., Marr, A. B. and Arcese, P. (2006), *Conservation and Biology of Small Populations: The Song Sparrows of Mandarte Island*, Oxford University Press.

Smith, J. M. and Slatkin, M. (1973), "The stability of predator-prey systems," *Ecology*, 54, 384-391.

Smith, J. N., Taitt, M. J., Rogers, C. M., Arcese, P., Keller, L. F., Cassidy, A. L. and Hochachka, W. M. (1996), "A metapopulation approach to the population biology of the Song Sparrow Melospiza melodia," *Ibis* , 138, 120-128.

Sólymos, P. (2010), "dclone: Data Cloning in R," *The R Journal*, 2, 29-37.

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, New York: Wiley-Interscience.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D., eds. (2004), *WinBUGS Version 1.4 User Manual*, London: MRC Biostatistics Unit, Institute of Public Health.

Stacey, P. B. and Taper, M. (1992), "Environmental variation and the persistence of small populations," *Ecological Applications*, 2, 18-29.

Staples, D. F., Taper, M. L. and Dennis, B. (2004), "Estimating population trend and process variation for PVA in the presence of sampling error," *Ecology*, 85, 923-929.

Staples, D. F., Taper, M. L. and Shepard, B. B. (2005), "Risk-Based Viable Population Monitoring," *Conservation Biology*, 19, 1908-1916.

Stiratelli, R., Laird, N. and Ware, J. H. (1984), "Random-effects models for serial observations with binary response," *Biometrics*, 961-971.

Stoer, J. and Bulirsch, R. (2002), *Introduction to Numerical Analysis*, New York: Springer.

Székely, G. J. and Móri, T. (2001), "A characteristic measure of asymmetry and its application for testing diagonal symmetry," *Communications in Statistics – Theory and Methods*, 30, 1633-1639.

Székely, G. J. and Rizzo, M. L. (2004), "Testing for equal distributions in high dimension," *InterStat* , 5.

Taylor, B. L. (2002), "The reliability of using population viability analysis for risk classification of species," *Conservation Biology*, 9, 551-558.

Thall, P. F. and Vail, S. C. (1990), "Some covariance models for longitudinal count data with overdispersion," *Biometrics*, 46, 657-671.

Thompson, E. (1994), "Monte Carlo likelihood in genetic mapping," *Statistical Science*, 9, 355-366.

Thompson, E. and Guo, S. W. (1991), "Evaluation of likelihood ratios for complex genetic models," *Mathematical Medicine and Biology*, 8, 149-169.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *The Annals of Statistics*, 22, 1701-1728.

Tierney, L. and Kadane, J. B. (1986), "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, 81, 82-86.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989), "Fully exponential Laplace approximations to expectations and variances of nonpositive functions," *Journal of the American Statistical Association*, 84, 710-716.

Torabi, M. and Shokoohi, F. (2012), "Likelihood inference in small area estimation by combining time-series and cross-sectional data," *Journal of Multivariate Analysis*, 111, 213-221.

Turchin, P. (1999), "Population regulation: a synthetic view," *Oikos* , 153-159.

Ushakov, N. G. (1999), *Selected Topics in Characteristic Functions*, Utrecht: VSP.

Venables, W. and Smith, D. (2011), "The R development core team," *An Introduction to R: A programming environment for data analysis and graphics*, 2.13.2.

Walker, A. (1969), "On the asymptotic behaviour of posterior distributions," *Journal of the Royal Statistical Society: Series B*, 31, 80-88.

Walters, C. J. (1986), *Adaptive Management of Renewable Resources*, New York: McMillan.

Wang, Y. (2007), "Maximum likelihood computation based on the Fisher scoring and Gauss–Newton quadratic approximations," *Computational Statistics and Data Analysis*, 51, 3776-3787.

Wei, G. C. and Tanner, M. A. (1990), "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American Statistical Association*, 85, 699-704.

Wolfinger, R. (1993), "Laplace's approximation for nonlinear mixed models," *Biometrika* , 80, 791-795.

Wu, C. (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11, 95-103.

Zeger, S. L. and Karim, M. R. (1991), "Generalized linear models with random effects; a Gibbs sampling approach," *Journal of the American Statistical Association*, 86, 79-86.

Zipunnikov, V. V. and Booth, J. G. (2006), "Monte Carlo EM for Generalized Linear Mixed Models using Randomized Spherical Radial Integration," Working Paper, Cornell University.