DISSERTATION

TOPICS IN DESIGN-BASED AND BAYESIAN INFERENCE FOR SURVEYS

Submitted by

Daniel Hernandez-Stumpfhauser

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2012

Doctoral Committee:

Advisor: Jean Opsomer

F. Jay Breidt
Jennifer A. Hoeting
Sonia M. Kreidenweis

ABSTRACT

TOPICS IN DESIGN-BASED AND BAYESIAN INFERENCE FOR SURVEYS

We deal with two different topics in Statistics. The first topic in survey sampling deals with variance and variance estimation of estimators of model parameters in the design-based approach to analytical inference for survey data when sampling weights include post-sampling weight adjustments such as calibration. Under the design-based approach estimators of model parameters, if available in closed form, are written as functions of estimators of population totals and means. We examine properties of these estimators in particular their asymptotic variances and show how ignoring the post-sampling weight adjustments, i.e. treating sampling weights as inverses of inclusion probabilities, results in biased variance estimators. Two simple simulation studies for two common estimators, an estimator of a population ratio and an estimator of regression coefficients, are provided with the purpose of showing situations for which ignoring the post-sampling weight adjustments results in significant biased variance estimators.

For the second topic we consider Bayesian inference for directional data using the projected normal distribution. We show how the models can be estimated using Markov chain Monte Carlo methods after the introduction of suitable latent variables. The cases of random sample, regression, model comparison and Dirichlet process mixture models are covered and motivated by a very large dataset of daily departures of anglers. The number of parameters increases with sample size and thus the need of exploring alternatives. We explore mean field variational methods and identify a number of problems in the application of the method to these models, caused by the poor approximation of the variational distribution to the posterior distribution. We propose solutions to those problems by improving the mean field variational approximation through the use of the Laplace approximation for the regression case and through the use of novel Monte Carlo procedures for the mixture model case.

DEDICATION

to Carmen, Joaquincito and my family

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

Two different topics are covered in this dissertation. The first topic (Chapter 2) deals with the variance and variance estimation of estimators of model parameters in the design-based approach to analytic inference for survey data were estimators of model parameters make use of the sampling weights. For example, for simple linear regression the estimator of the slope is $\frac{\sum_{i\in s} w_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i\in s} w_i(x_i-\bar{x})^2}$ for covariate x and response y in sample s, and w_i representing survey weights. It is common for the weights w_i to include post-sampling adjustments such as calibration. However, in practice, variance estimation methods treat the sampling weights as inverses of inclusion probabilities, i.e. they ignore post-sampling weight adjustments, and doing so could lead to significantly biased variance estimators of estimators of model parameters depending on the nature of the post-sampling weight adjustments. Some formulas and simple simulation studies that show the bias effect in variance estimators when ignoring post-sampling weight adjustments are given in Chapter 1. Under a different approach of constructing efficient estimators of model parameters, similar formulas to the ones in Chapter 2 were given by Särndal, Swensson, and Wretman (1992, p.294-296) for the case of estimation of a population ratio and by Elvers et al. (1985) for an estimator of linear regression coefficients. Furthermore Goga and Ruiz-Gazen (2012) give variance formulas for the semiparametric case. Due to these facts the original aim of generalizing the results to the semiparametric context and to obtain estimators that fully account for the design and the post-sampling weight adjustments was not pursued.

The rest of the dissertation is devoted to the second topic: Bayesian methods for analysis of directional data based on the projected normal distribution. Directional data arise in

various ways and in many scientific disciplines. In Meteorology wind directions provide a natural source of circular data, (Breckling, 1989). The times of day at which thunderstorms occur and times of year at which heavy rain occur are other examples of circular data. We can also find circular data in Biology when studying bird navigation (Schmidt-Koenig, 1965 and Batschelet, 1981). Spherical data arise in Earth Sciences, Physics and Astronomy. Some examples are the epicentre of an earthquake and the distribution on the celestial sphere of sources of high-energy cosmic rays (Mardia and Edwards, 1982). For our application we have a large data set of daily departure of anglers, which we want to predict based on a set of spatial and temporal categorical variables. In Chapter 3 we present regression models for circular data and model selection based on the Deviance Information Criterion. We show how these models can be fit using Markov chain Monte Carlo methods after the introduction of suitable latent variables. Due to the very large size of the data set of daily departure of anglers (over 1,000,000 observations), we also show how they can be fit using variational/Laplace approximations which are fast and deterministic approximations to the posterior distribution. The mean field variational method (Ormerod and Wand, 2010) is based on the minimization of the Kullback-Leibler divergence between the posterior distribution and a variational distribution. The variational distribution is restricted to a manageable class of distributions and thus minimization of the Kullback-Leibler divergence is done over that class. Depending on the Bayesian model at hand and restrictions of the variational distribution, the mean field variational method can lead to poor inference. We identify a problem in the application of the mean field variational method and fix it by making use of the Laplace approximation.

In Chapter 4 we present Dirichlet process (DP) mixture models for directional data. The Dirichlet process (Ferguson, 1973) is a distribution over distributions. With probability one, distributions drawn from a DP are discrete. That is, a draw from a DP is a discrete distribution that places its probability mass on a countable (finite or infinite) subset of the underlying sample space. This discreteness is useful when modeling data with mixture

models. In the DP mixture, a discrete distribution is drawn from a DP. Mixture parameters are then drawn from this distribution and finally data are drawn conditional on the mixture parameters. These models can be fit using Markov chain Monte Carlo methods. Since our application is for the daily departure of anglers, we explore variational methods similar to the methods given by Blei and Jordan (2006). We identify a number of problems in the application of the variational method for DP mixture models for circular data and propose solutions to those problems by improving the variational approximation through the use of novel Monte Carlo procedures.

CHAPTER 2

ISSUES WITH VARIANCE ESTIMATION IN DESIGN-BASED APPROACHES TO ANALYTIC INFERENCE FOR SURVEY DATA

Summary

Design-based analytic inference for model parameters makes use of the survey weights in the construction of estimators of model parameters. In practice, the survey weights include post-sampling weight adjustments such as calibration. The major variance estimation methods in use today treat the survey weights as if they were inverses of inclusion probabilities. This accounts for the sampling design but not for any of the post-sampling adjustments, which can result in biased variance estimators. We present some theoretical results that allow us to identify situations for which ignoring the post-sampling weight adjustments indeed suffer from bias. Through simulation experiments, we illustrate the effect of the calibration on the variance estimators of two simple but common estimators: the estimator of a population ratio and the estimator of linear regression coefficients.

2.1 Introduction

An important use of survey data is for *analytic inference*, in which the target of inference is not the specific population from which the sample is drawn, but rather a statistical model. The statistical model can be thought of as the stochastic mechanism that generated the population being sampled. Then the sample is viewed as a set of data obtained by two random process operating in sequence: in a first step, the population is a realization from the statistical model, often referred to as the "superpopulation model," and in a second step, the sample is drawn from that particular population via a sampling design.

Two main approaches are in use for analytic inference: the *model-based approach to*

analytic inference, which relies primarily on model specification to capture the survey effects; and the *design-based approach to analytic inference*, which emphasizes the use of the *survey weights* in the construction of estimators. The survey weights account for the complex sampling design and often incorporate calibration and regression adjustments applied for reasons of consistency with other data sources or to improve the efficiency of estimators.

In the model-based approach for inference about superpopulation parameters $\boldsymbol{\theta}$, the effects of the sampling design and the survey data collection are incorporated as part of the model describing the data, by adding those effects to the distributional specification of the model itself. Once such a model is constructed and accepted as being a good representation of the data as observed in the sample, analysis proceeds using standard methods. Model-based approaches tend to be more efficient than the design-based approaches on which we will focus, but often require more information for implementation.

Under the design-based approach for inference about $\boldsymbol{\theta}$, one begins by defining census parameters $\boldsymbol{\theta}_N$, which are estimators of the model parameters that would be computed given a census of the complete population. Many census parameters are defined as the solution of population-level estimating equations, such as those obtained by setting the derivative of the population log-likelihood equal to zero. Once $\boldsymbol{\theta}_N$ are defined for the population, they are estimated from the sample data as $\hat{\boldsymbol{\theta}}_N$ using survey-weighted approaches, by treating them as functions of sample means and totals. It is usually assumed that the following central limit theorem (CLT) holds (Binder and Roberts, 2003):

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N\right) \xrightarrow{\mathcal{L}} \mathrm{N}\left(\mathbf{0}, \mathrm{V}_1\right).$$

For example, conditions were given by Hájek (1960) and Hájek (1964) for the asymptotic normality of the sample mean under simple random sampling without replacement and rejective sampling with varying probabilities. Other design-based CLT results for the sample mean were given by Rosén (1972) and Rosén (1997) for the case of probability proportional

to size without replacement (πps). Results for stratified multistage probability proportional to size with replacement designs were given by Krewski and Rao (1981).

Under suitable assumptions on the superpopulation, we also have a CLT (assuming $\sqrt{N}$-consistency of $\boldsymbol{\theta}_N$ for $\boldsymbol{\theta}$), written as

$$\sqrt{n/N}\sqrt{N}\left(\boldsymbol{\theta}_N - \boldsymbol{\theta}\right) \xrightarrow{\mathcal{L}} \left\{\lim_{n,N\to\infty}(n/N)\right\}^{1/2} \mathrm{N}\left(\mathbf{0}, \mathrm{V}_2\right).$$

Finally, from a standard argument (e.g., Rubin-Bleuer and Kratina, 2005) that the two are asymptotically independent, it follows that

$$\hat{\boldsymbol{\theta}}_N \text{ is asymptotically } \mathrm{N}\left(\boldsymbol{\theta}, (1/n)\left\{\mathrm{V}_1 + (n/N)\,\mathrm{V}_2\right\}\right),$$

and so inference relies on finding consistent variance estimators $\hat{\mathrm{V}}_1$ and $\hat{\mathrm{V}}_2$. If the sampling fraction n/N is negligible, then estimation of V_2 can be avoided since the latter variance component is asymptotically negligible.

The major variance estimation methods in use today for analytic inference treat the survey weights as if they were inverses of inclusion probabilities. Depending on the model under analysis as well as the nature of the post-sampling weight adjustments, this can result in biased variance estimation. The main goals of this article are two-fold. First, we aim to derive a set of general results that can be applied to modeling contexts in which the parameter estimators and the survey calibration estimators can be written as differentiable functions of survey totals. Second, rather than trying to provide an in-depth study of specific models and designs, we highlight two special cases and illustrate the effects of the calibration on the variance estimators through some simple simulation experiments. The original intent of this research was to generalize these results to the semiparametric context and to obtain estimators that fully account for the design and the calibration. However, a recent unpublished article by Goga and Ruiz-Gazen (2012) has covered a very similar topic, so that this original aim was abandoned.

The paper is organized as follows. In section 2.2 we present some results about the asymptotic variance of estimators in the design-based approach to analytic inference with proofs provided in the Supplement. In section 2.3 we consider some linear models with post-sampling weight adjustments (post-stratification and calibration) and a few simulation studies. In section 2.4 we present our conclusions.

2.2 Some results

Consider a population U consisting of N elements labeled $k = 1, 2, \ldots, N$. A sample s of size n is obtained from this population via a sampling design $p(s)$. Let $y, \boldsymbol{x}, \boldsymbol{z}$ denote variables and let $y_k, \boldsymbol{x}_k, \boldsymbol{z}_k$ be the values of those variables for the kth element in the population. We will use y to denote the model response variable, $\boldsymbol{x}$ will denote the model covariates and $\boldsymbol{z}$ will denote survey auxiliary variables. We will use t_j and $\bar{t}_j$ to denote the population total and population mean for some variable, labeled j, respectively and their "hat" versions, $\hat{t}_j$ and $\hat{\bar{t}}_j$, as the Horvitz-Thompson estimators of them. For example $\hat{t}_y = \sum_s \frac{y_k}{\pi_k}$ and $\hat{t}_x = \sum_s \frac{\boldsymbol{x}_k}{\pi_k}$, where π_k is the inclusion probability of element k.

In what follows, we will also use $\hat{\bar{h}}_i$ to denote a different estimator of a population mean that uses the auxiliary variables $\boldsymbol{z}_k$ (e.g. a regression estimator of a population mean) which will in general be a function of Horvitz-Thompson estimators. Finally, we will use $g\left(\hat{\bar{\boldsymbol{h}}}\right)$ to denote a function of $\hat{\bar{h}}_i$ estimators, which in the design-based approach to analytic inference will typically be a population-level estimator of a model parameter of interest. We will refer to this quantity as a "census parameter," as is commonly done in the survey literature.

Some of the results in this section are concerned about the asymptotic variance (denoted AVar) of estimators of population quantities. The asymptotic variance of an estimator will be defined here as the variance of the first order Taylor approximation around the target version of the estimator of interest. For example, let our estimator of interest be $\hat{R} = \hat{t}_y/\hat{t}_x$ with target $R = t_y/t_x$. To find the asymptotic variance of $\hat{R}$ we perform a first order Taylor approximation around the target R i.e. $\hat{R} \approx R + \frac{1}{t_x}\sum_s \frac{y_k - Rx_k}{\pi_k}$ and so

$\text{AVar}\left(\hat{R}\right) = \text{Var}\left(\frac{1}{t_x}\sum_s \frac{y_k - Rx_k}{\pi_k}\right)$. This is the most common approach to perform inference for nonlinear estimators for surveys, see for instance the classic text by Särndal, Swensson, and Wretman (1992).

The first of the following 3 results is about the order in probability of estimators which are functions of estimators of population means. The second and third results are about the asymptotic variance of these estimators for special cases of calibration and sampling designs. We will use these results in Section 2.3 to derive asymptotic formulas for two estimators, a ratio estimator and a regression estimator.

Result 1. *Let $g\left(\hat{\bar{\boldsymbol{h}}}\right)$ be a function of m estimators of population means $\hat{\bar{\boldsymbol{h}}} = \left(\hat{\bar{h}}_1, \hat{\bar{h}}_2, \ldots, \hat{\bar{h}}_m\right)$ and let g have continuous partial derivatives of order 2 at the population means $\bar{\boldsymbol{h}} = \left(\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_m\right)$. Let the estimators $\hat{\bar{h}}_i$ be continuous functions of J Horvitz-Thompson estimators of population means $\hat{\bar{h}}_i = \hat{\bar{h}}_i\left(\hat{\bar{\boldsymbol{t}}}\right)$, $\hat{\bar{\boldsymbol{t}}} = \left(\hat{\bar{t}}_1, \hat{\bar{t}}_2, \ldots, \hat{\bar{t}}_J\right)$ and let $\hat{\bar{h}}_i$ have continuous partial derivatives at the population means $\bar{\boldsymbol{t}} = (\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_J)$ and let n be the sample size. Assume that the sampling design is such that $\hat{\bar{t}}_j - \bar{t}_j = O_p\left(\frac{1}{\sqrt{n}}\right)$ for all population means $\bar{t}_j$. Then,*

$$g\left(\hat{\bar{\boldsymbol{h}}}\right) = g\left(\bar{\boldsymbol{h}}\right) + \sum_{i=1}^{m} \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i} \sum_{j=1}^{J} \frac{\partial \hat{\bar{h}}_i\left(\bar{\boldsymbol{t}}\right)}{\partial \bar{t}_j}\left(\hat{\bar{t}}_j - \bar{t}_j\right) + O_p\left(\frac{1}{n}\right).$$

Result 2. *Let the same conditions as in result 1 apply and let the $\hat{\bar{h}}_i$ estimators be regression estimators of means of $x's$ and y based on auxiliary information $\boldsymbol{z}$ i.e.*

$$\begin{aligned}
\hat{\bar{h}}_1 &= \hat{\bar{t}}_{ry} = \hat{\bar{t}}_{y\pi} + \sum_{l=1}^{L}\left(\bar{t}_{z_l} - \hat{\bar{t}}_{z_l\pi}\right)\hat{B}_{1l} \\
\hat{\bar{h}}_i &= \hat{\bar{t}}_{rx_i} = \hat{\bar{t}}_{x_i\pi} + \sum_{l=1}^{L}\left(\bar{t}_{z_l} - \hat{\bar{t}}_{z_l\pi}\right)\hat{B}_{il}
\end{aligned}$$

where $\hat{\bar{t}}_{y\pi}, \hat{\bar{t}}_{x_i\pi}, \hat{\bar{t}}_{z_l\pi}$ are the Horvitz-Thompson estimators of the population means $\bar{t}_y, \bar{t}_{x_i}$ and the known population means $\bar{t}_{z_l}$ respectively. The coefficients $\hat{B}_{i1}, \hat{B}_{i2}, \ldots, \hat{B}_{iL}$ are compo-

nents of the L-vector $\hat{\boldsymbol{B}}_i = \left(\hat{B}_{i1}, \hat{B}_{i2}, \ldots, \hat{B}_{iL}\right)^T = \left(\sum_s \boldsymbol{z}_k \boldsymbol{z}_k^T/\pi_k\right)^{-1} \sum_s \boldsymbol{z}_k x_{ik}/\pi_k$ *where* $\boldsymbol{z}_k = (z_{1k}, z_{2k}, \ldots, z_{Lk})^T$ *is the value of the auxiliary vector for the kth element in the population and* $\hat{\boldsymbol{B}}_1 = \left(\hat{B}_{11}, \hat{B}_{12}, \ldots, \hat{B}_{1L}\right)^T = \left(\sum_s \boldsymbol{z}_k \boldsymbol{z}_k^T/\pi_k\right)^{-1} \sum_s \boldsymbol{z}_k y_k/\pi_k$. *Under these conditions the asymptotic variance of* $g\left(\hat{\bar{\boldsymbol{h}}}\right)$ *is*

$$
\begin{aligned}
&\mathrm{AVar}\left(g\left(\hat{\bar{\boldsymbol{h}}}\right)\right) \\
&= \frac{1}{N^2}\mathrm{Var}\left(\sum_{k\in s}\left[\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_1}\frac{y_k}{\pi_k} + \sum_{i=2}^{m}\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i}\frac{x_{ik}}{\pi_k}\right]\right) + \\
&\quad \frac{1}{N^2}\mathrm{Var}\left(\sum_{k\in s}\left[\sum_{i=1}^{m}\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i}\frac{\boldsymbol{z}_k^T\boldsymbol{B}_i}{\pi_k}\right]\right) - \\
&\quad \frac{2}{N^2}\mathrm{Cov}\left(\sum_{k\in s}\frac{1}{\pi_k}\left[\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_1}y_k + \sum_{i=2}^{m}\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i}x_{ik}\right], \sum_{k\in s}\frac{1}{\pi_k}\left[\sum_{i=1}^{m}\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i}\boldsymbol{z}_k^T\boldsymbol{B}_i\right]\right)
\end{aligned}
$$

where the vectors $\boldsymbol{B}_1$ *and* $\boldsymbol{B}_2, \ldots, \boldsymbol{B}_m$ *are the population regression coefficients obtained by regressing the* z'*s on the* y *and* x'*s respectively and* N *is the population size.*

Result 3. *Let conditions in result 1 and result 2 apply and let* $\boldsymbol{z}_k = \left(\mathrm{I}_{\{k\in U_1\}}, \mathrm{I}_{\{k\in U_2\}}, \ldots, \mathrm{I}_{\{k\in U_L\}}\right)^T$ *for strata* $U_1, U_2, \ldots, U_L$*, the post-stratification case. Define* $r_k = \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial h_1}y_k + \sum_{i=2}^{m}\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial h_i}x_{ik}$, $\bar{r}_{U_l} = \sum_{k\in U_l} r_k$ *and* $\hat{N}_l = \sum_s \mathrm{I}_{\{k\in U_l\}}/\pi_k$. *Then the asymptotic variance of* $g\left(\hat{\bar{\boldsymbol{h}}}\right)$ *is*

$$
\begin{aligned}
\mathrm{AVar}\left(g\left(\hat{\bar{\boldsymbol{h}}}\right)\right) &= \frac{1}{N^2}\mathrm{Var}\left(\sum_{k\in s}\frac{1}{\pi_k}\left[r_k\right]\right) - \frac{1}{N^2}\mathrm{Var}\left(\sum_{l=1}^{L}\bar{r}_{U_l}\hat{N}_l\right) + \\
&\quad -\frac{2}{N^2}\sum_{l=1}^{L}\bar{r}_{U_l}\sum_{k=1}^{L}\sum_{i\in U_k}\sum_{j\in U_l}\left(r_i - \bar{r}_{U_k}\right)\frac{\pi_{ij}}{\pi_i\pi_j}.
\end{aligned}
$$

Furthermore, if the sampling design is such that $\frac{\pi_{ij}}{\pi_i\pi_j} = \begin{cases} a_l, & i = j \\ b_l, & i \neq j \end{cases}$ *for* $i, j \in U_l$ *and* $\frac{\pi_{ij}}{\pi_i\pi_j} = c$

for i and j in different stratum, then the asymptotic variance of $g\left(\hat{\hat{\boldsymbol{h}}}\right)$ is

$$\text{AVar}\left(g\left(\hat{\hat{\boldsymbol{h}}}\right)\right) = \frac{1}{N^2}\text{Var}\left(\sum_{k\in s}\frac{1}{\pi_k}\left[r_k\right]\right) - \frac{1}{N^2}\text{Var}\left(\sum_{l=1}^{L}\bar{r}_{U_l}\hat{N}_l\right).$$

In the next section, we consider two common models and the effect of calibration on the asymptotic variance of estimators. We derive asymptotic variance formulas for these two models and then run a few simple simulations to illustrate the calibration effects. Similar asymptotic variance formulas exist already in the literature, although the results are described there in less generality. In section 2.3.1 we consider the estimation of a ratio of two population totals, which is also done in Särndal et al. (1992, p.294-296). In section 2.3.2 we consider the estimation of regression coefficients and similar formulas can be found on Elvers et al. (1985).

2.3 Models

2.3.1 Ratio model

Consider the model

$$y_k = \beta x_k + \epsilon_k,\ \{\epsilon_k\}\,\text{iid N}\left(0, \sigma^2 x_k\right).$$

The census parameter $\beta_N = t_y/t_x = \sum_U y_k / \sum_U x_k$ is the maximum likelihood estimator of β under the stated superpopulation model. The sample s used to estimate β is obtained according to a sampling design with inclusion probabilities π_k. Auxiliary information on the population U is available to post-stratify the sample into L strata $U_1, U_2, \ldots, U_L$ with known sizes $N_1, N_2, \ldots, N_L$. This corresponds to the case where the $\boldsymbol{z}_k$ are post-stratum membership indicators. The survey weights to be used in estimation are $w_{ks} = \pi_k^{-1}\left(N_l/\hat{N}_l\right)$ for $k \in U_l$, with $\hat{N}_l = \sum_s I_{\{k\in U_l\}}/\pi_k$, $l = 1, 2, \ldots, L$. The census parameter β_N is a function of population totals and so the traditional approach to analytic inference substitutes population totals by estimates, $\hat{\beta}_N = \hat{t}_y/\hat{t}_x$, with $\hat{t}_y = \sum_s w_{ks}y_k$ and similarly for $\hat{t}_x$.

If we assume that the difference between β_N and the model parameter β is negligible, then the variance of the asymptotic distribution of $\hat{\beta}_N$ is obtained using Taylor methods, under standard design-based asymptotic assumptions. We will start by showing what would the asymptotic variance of $\hat{\beta}_N$ be if $w_k = \pi_k^{-1}$, i.e. if no post-sampling weight adjustments were used. In this case the linearized variance of $\hat{\beta}_N$ is obtained by finding the first order Taylor approximation to $\hat{\beta}_N$ in a neighborhood of the population totals t_y, t_x:

$$\text{AVar}_1\left(\hat{\beta}_N\right) = \frac{1}{t_x^2}\text{Var}\left(\hat{t}_e\right), \tag{1}$$

where $e_k = y_k - \beta_N x_k$ are the finite population-level model residuals. Notice that if the survey weights were now to include post-stratification adjustments $w_{kl} = \pi_k^{-1}\left(N_l/\hat{N}_l\right)$, $l = 1, 2, \ldots, L$, these would be sample dependent and would also be functions of estimators of population totals. Thus $\hat{\beta}_N$ is now a function of the total estimators $\hat{t}_{y1,\pi}, \hat{t}_{x1,\pi}, \hat{N}_{1,\pi}, \ldots,$ $\hat{t}_{yL,\pi}, \hat{t}_{xL,\pi}\hat{N}_{L,\pi}$

$$\hat{\beta}_N = \frac{\sum_l \frac{N_l}{\hat{N}_{l,\pi}}\hat{t}_{yl,\pi}}{\sum_l \frac{N_l}{\hat{N}_{l,\pi}}\hat{t}_{xl,\pi}},$$

where the $\hat{t}_{yl,\pi} = \sum_{i\in s\cap U_l} y_i/\pi_i$ are the Horvitz-Thompson estimators of the corresponding population totals, and similarly we have defined all other estimators of totals. We include the subscript π in the notation to distinguish $\hat{t}_{y,\pi}$ from $\hat{t}_y = \sum_s w_{ks}y_k$. Using result 2, the complete linearized variance of $\hat{\beta}_N$ obtained by making $\hat{\bar{h}}_1 = \frac{1}{N}\sum_s w_k y_k$, $\hat{\bar{h}}_2 = \frac{1}{N}\sum_s w_k y_k$ and $g\left(\hat{\bar{\boldsymbol{h}}}\right) = \frac{\hat{\bar{h}}_1}{\hat{\bar{h}}_2} = \hat{\beta}_N$ is readily shown to be:

$$\begin{aligned}
\text{AVar}_2\left(\hat{\beta}_N\right) &= \frac{1}{t_x^2}\text{Var}\left(\hat{t}_{e,\pi}\right) + \frac{1}{t_x^2}\text{Var}\left(\sum_{l=1}^{L}\bar{e}_{U_h}\hat{N}_{l,\pi}\right) - \frac{2}{t_x^2}\text{Cov}\left(\hat{t}_e, \sum_{l=1}^{L}\bar{e}_{U_l}\hat{N}_{l,\pi}\right) \\
&= \text{AVar}_1\left(\hat{\beta}_N\right) + \frac{1}{t_x^2}\text{Var}\left(\sum_{l=1}^{L}\bar{e}_{U_h}\hat{N}_{l,\pi}\right) - \frac{2}{t_x^2}\text{Cov}\left(\hat{t}_e, \sum_{l=1}^{L}\bar{e}_{U_l}\hat{N}_{l,\pi}\right),
\end{aligned} \tag{2}$$

where the $\bar{e}_{U_l} = \sum_{k\in U_l} e_k/N_l$ are the means of the population-level residuals in post-stratum l, and the e_k were defined previously. In general, the two extra terms in $\text{AVar}_2\left(\hat{\beta}_N\right)$ rel-

ative to $\text{AVar}_1\left(\hat{\beta}_N\right)$ are of the same order of magnitude as the first term and, depending on the sign of the covariance term, they can result in a larger or smaller overall linearized variance. Survey-specific software programs typically target $\text{AVar}_1\left(\hat{\beta}_N\right)$ and this could lead to bias. When the residuals e_k are uncorrelated with the $I_{\{k\in U_l\}}$, then $\bar{e}_{U_l} \approx 0$ and ignoring the second and third term in equation (2) is reasonable. Also, if the sampling design satisfies the conditions in result 3 the post-stratification is efficient i.e. $\text{AVar}_2\left(\hat{\beta}_N\right) < \text{AVar}_1\left(\hat{\beta}_N\right)$ and so ignoring the second and third term in equation (2) leads to conservative estimates. Some sampling designs that satisfy the conditions in result 3 are simple random sampling without replacement (SI), simple random sampling with replacement (SIR), Bernoulli sampling (BE), stratified Bernoulli sampling (STBE) and stratified simple random sampling with Replacement (STSIR) when strata and post-strata are the same and thus in those cases $\text{AVar}_2\left(\hat{\beta}_N\right) = \text{AVar}_1\left(\hat{\beta}_N\right) - \frac{1}{t_x^2}\text{Var}\left(\sum_{l=1}^{L}\bar{e}_{U_l}\hat{N}_{l,\pi}\right)$.

We performed a small simulation study to assess the difference between $\text{AVar}_2\left(\hat{\beta}_N\right)$ and $\text{AVar}_1\left(\hat{\beta}_N\right)$. We generated a population of size $N = 1000$ with the x_k generated from a $\text{Gamma}\,(1,1)$ distribution and we generated $y_k = \beta x_k + \varepsilon_k$ by generating the ε_k from a normal distribution with mean 0 and variance $\sigma^2 = 1$. Furthermore we generated normal observations ξ_k correlated with ε_k, i.e. $(\varepsilon_k, \xi_k) \overset{ind}{\sim} N\left((0,0), \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, and assigned the population units to two post-strata by making use of the indicator $z_k = \mathbf{I}\{\xi_k > 0\}$. Population unit k was assigned to post-stratum 1 if $z_k = 1$ and assigned to post-stratum 2 if $z_k = 0$. It can be shown that the correlation between ε_k and z_k is $\text{Corr}\,(\varepsilon_i, z_i) = \rho\sqrt{\frac{2}{\pi}}$, and this is how we controlled for the amount of correlation between the errors e_k and the post-stratum indicators $I_{\{k\in U_l\}}$, $l = 1, 2$.

We drew 10000 simple random samples of sizes $n = 100$ from this population, and estimated the population ratio β_N for each sample by the ratio of the estimated totals $\hat{\beta}_N = \sum_s w_{ks}y_k / \sum_s w_{ks}x_k$, with w_{ks} the post-stratified weights. Results in Table 1 show the relative bias, $\frac{1}{10000}\sum_{i=1}^{10000}(estimate - true)/true$, for estimators of variances $\text{AVar}_1\left(\hat{\beta}_N\right)$

Table 1: Post-stratification effect in variance estimation for the ratio model 2.3.1

$\text{Corr}(\varepsilon_i, z_i)$	-0.75	-0.5	0	0.5	0.75
$\frac{AVar_2}{AVar_1}$	0.426	0.771	0.999	0.758	0.420
$Rel.bias\left(\widehat{AVar}_{1,\pi}\right)$	1.270	0.265	-0.005	0.332	1.446
$Rel.bias\left(\widehat{AVar}_{2,\pi}\right)$	-0.038	-0.032	-0.014	0.004	0.020

and $\text{AVar}_2\left(\hat{\beta}_N\right)$. It shows that ignoring the calibration produces biased estimates of the variance, and accounting for calibration the variance estimates are nearly unbiased. The estimators of the variances were obtained by replacing all the population quantities in equations (1) and (2) by survey-weighted sample estimators. In both estimates of relative bias the "true" variance was taken as the sample variance of the 10000 estimates of β_N. Also, Table 1 shows the ratio $\text{AVar}_2/\text{AVar}_1$ to show how much smaller AVar_2 is, relative to AVar_1 when changing the value of $\text{Corr}(\varepsilon_i, z_i)$.

2.3.2 Linear model

Consider the model

$$y_k = x_k^T\boldsymbol{\beta} + \epsilon_k,\ \{\epsilon_k\}\ \text{iid N}\left(0, \sigma^2\right),$$

where x_k is a q-dimensional vector of covariates, for $k = 1, 2, \ldots, N$. The census parameter

$$\begin{aligned} B_N &= \left(X^TX\right)^{-1}X^TY \\ &= \left(\sum_U x_k x_k^T\right)^{-1}\left(\sum_U x_k y_k\right) \\ &= H^{-1}\boldsymbol{h}_0 \end{aligned}$$

is the maximum likelihood estimator of $\boldsymbol{\beta}$, where X is the design matrix and Y is the vector of responses. The $q \times q$ symmetric matrix $H = (h_{ij})$ defined here as $\left(\sum_U x_k x_k^T\right)$ has as elements the population totals $h_{ij} = \sum_U x_{ik}x_{jk}$, while the q-dimensional vector $\boldsymbol{h}_0 = (h_{10}, \ldots, h_{q0})$

has as elements the population totals $h_{i0} = \sum_U x_{ik} y_k$. The sample s used to estimate $\boldsymbol{\beta}$ is obtained according to a sampling design with inclusion probabilities π_k. Auxiliary information $z_k, k = 1, \ldots, N$ on the population U is available and calibrated weights are constructed in the following way (Särndal et al. 1992, p.232):

$$w_k = \frac{1}{\pi_k} \left(1 + \left(\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}_{z,\pi} \right)^T \hat{\bar{T}}^{-1} z_k \right),$$

where $\hat{\bar{T}} = \sum_s \frac{z_k z_k^T}{N \pi_k}$, and $\hat{\bar{\boldsymbol{t}}}_{z,\pi}$ is a p-dimensional vector of Horvitz-Thompson estimators of population means $\bar{\boldsymbol{t}}_z = \left(\bar{t}_{z_1}, \bar{t}_{z_2}, \ldots, \bar{t}_{z_p} \right)^T$. The census parameter B_N is a function of population totals and so the traditional design-based approach to analytic inference substitutes population totals by estimates,

$$\hat{B}_N = \hat{H}^{-1} \hat{\boldsymbol{h}}_0 \tag{3}$$

where $\hat{H}$ is a $q \times q$ symmetric matrix with elements $\hat{h}_{ij} = \sum_s x_{ik} x_{jk} w_k$ and $\hat{\boldsymbol{h}}_0$ is a q-dimensional vector with elements $\hat{h}_{i0} = \sum_s x_{ik} y_k w_k$.

If we assume that the difference between B_N and the model parameter $\boldsymbol{\beta}$ is negligible, then the variance of the asymptotic distribution of $\hat{B}_N$ is obtained using Taylor methods. We will again start by showing what would the asymptotic variance of $\hat{B}_N$ be if $w_k = \pi_k^{-1}$. From Särndal et al. (1992, p.194) the asymptotic variance of $\hat{B}_N$ is

$$\mathrm{AVar}_1 \left(\hat{B}_N \right) = H^{-1} \mathrm{Var} \left(\hat{\boldsymbol{t}}_1 \right) H^{-1}. \tag{4}$$

where $\hat{\boldsymbol{t}}_1$ is a q-dimensional vector with ith component equal to

$$\hat{t}_{1i} = \sum_{k \in s} \frac{1}{\pi_k} x_{ik} \left(y_k - x_k^T B_N \right) \tag{5}$$

and where x_{ik} is the value of the ith covariate for the kth population element. We now let $w_k = \frac{1}{\pi_k} \left(1 + \left(\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}_{z,\pi} \right)^T \hat{\bar{T}}^{-1} z_k \right)$ and define $C_{iy} = \bar{T}^{-1} \bar{\boldsymbol{t}}_{iy}$, $C_{ij} = \bar{T}^{-1} \bar{\boldsymbol{t}}_{ij}$ where $\bar{\boldsymbol{t}}_{iy}$ and $\bar{\boldsymbol{t}}_{ij}$ are p-dimensional vectors with elements $\bar{t}_{iy_r} = \frac{1}{N} \sum_{k \in U} z_{rk} x_{ik} y_k$ and $\bar{t}_{ij_r} = \frac{1}{N} \sum_{k \in U} z_{rk} x_{ik} x_{jk}$

respectively for $r = 1, 2, \ldots, p$. We now make use of result 2 to show that after some algebra the asymptotic variance of $\hat{B}_N$ is

$$\text{AVar}_2\left(\hat{B}_N\right) = H^{-1}\text{Var}\left(\hat{\boldsymbol{t}}_2\right)H^{-1} \tag{6}$$

where $\hat{\boldsymbol{t}}_2$ is a q-dimensional vector with ith component equal to

$$\hat{t}_{2i} = \sum_{k\in s}\frac{1}{\pi_k}\left[x_{ik}\left(y_k - x_k^T B_N\right) - z_k^T\left(C_{iy} - \sum_{j=1}^{q} C_{ij}B_j\right)\right] \tag{7}$$

where B_j is the jth component of B_N. The extra term in $\hat{\boldsymbol{t}}_2$ relative to $\hat{\boldsymbol{t}}_1$ is due to the calibration effect and not accounting for it could lead to bias in the estimation of the asymptotic variance of $\hat{B}_N$.

We performed a small simulation study to assess the difference between $\text{AVar}_2\left(\hat{B}_N\right)$ and $\text{AVar}_1\left(\hat{B}_N\right)$. Before going into details of the simulation we present the following result that was useful in generating the population values. The proof is given in Supplement 2.8.

Result 4. *For a SRS design and calibration weights* $w_k = \frac{1}{\pi_k}\left(1 + \left(\bar{\boldsymbol{t}}_z - \hat{\bar{\boldsymbol{t}}}_{z,\pi}\right)^T \hat{\bar{T}}^{-1} z_k\right)$, *define* $z_j^*, j = 1, 2, \ldots, p$ *as the* N*-dimensional vector with* k*th element equal to the value of the* j*th covariate for the* k*th population element (the* j*th column in the* Z *design matrix). Also define* $\eta_i = (x_{i1}e_1, \ldots, x_{iN}e_N)^T$, $e_k = y_k - x_k^T B_N$ *and* ξ_i *as the projection of* η_i *onto the span of* $\left\{z_1^*, \ldots, z_p^*\right\}$. *Then,*

$$\text{Var}\left(\hat{t}_{1i}\right) = \text{Var}\left(\hat{t}_{2i}\right) + N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}S_{\xi_i}^2$$

where $\hat{t}_{1i}$ *and* $\hat{t}_{2i}$ *are defined as in equations* (5) *and* (7) *respectively.*

This implies that for SRS $\text{AVar}_1\left(\hat{B}_i\right) \geq \text{AVar}_2\left(\hat{B}_i\right)$ and the calibration effect will depend on the amount of correlation between η_i and the z's.

We generated a population of size $N = 2000$ with ξ_1^* and ξ_2^* generated from independent normal $\text{N}(5, 1)$ and $\text{N}(3, 1)$ respectively. We then generated $x_1^* = 2 + 5\xi_1^* - 3\xi_2^* + \epsilon_1$ and

$x_2^* = 1 + 2\xi_1^* + 4\xi_2^* + \epsilon_2$ where ϵ_1 and ϵ_2 are independent normal with mean zero and variance σ_x^2 such that the $R^2 = 1 - SS_{err}/SS_{tot}$ between x_1^* and $\{1, \xi_1^*, \xi_2^*\}$, and R^2 between x_2^* and $\{1, \xi_1^*, \xi_2^*\}$ was approximately 0.75. This was done by setting $\sigma_x^2 = \frac{\left(1-R^2\right)SS_{reg}}{(n-2)R^2}$, where SS_{reg} can be easily estimated by making use of the mean of the distribution that generated x_1^* and the average of the x_1^* values. We then generated the y values by setting $y = 5 + 5x_1^* + x_2^* + \epsilon$, where the ϵ were draws from a $\mathrm{N}\left(0, \sigma_y^2\right)$ and σ_y^2 was such that R^2 between y and $\{1, x_1^*, x_2^*\}$ was close to 0.75. Finally we built different sets of calibration variables $\{1, z_1^*, z_2^*\}$ to show the effect that calibration can have on the variance of estimators. For example, in the first simulation we made z_1^* equal to η_1 + noise where η_1 is defined in result 4 and changed the value of R^2 between z_1^* and η_1 from 0.1 to 0.9 in increments of 0.1.

We drew 15000 samples of sizes $n = 400$ from these populations and for each one of these samples estimated $\mathrm{AVar}_1\left(\hat{B}_i\right)$ and $\mathrm{AVar}_2\left(\hat{B}_i\right)$ for $i = 0, 1, 2$ where $\hat{B}_0$ is the estimate of the intercept. The estimators of the variances were obtained by replacing all the population quantities in equations (4) and (6) by survey-weighted sample estimators and then we computed relative biases, $\frac{1}{15000}\sum_{i=1}^{15000}\left(estimate - true\right)/true$, taking as $true$ the sample variance of the 15000 estimates of $\hat{B}_i$ computed as in (3).

First simulation.

We made the calibration variables equal to $\{1, z_1^* = \eta_1 + \text{noise}, z_2^* = \xi_2^*\}$ and changed the values of $R^2_{z_1^*\eta_1} = 0.1, 0.2, \ldots, 0.9, 0.95$. Figure 1 shows plots of the asymptotic variance ratios $\mathrm{AVar}_2\left(\hat{B}_i\right)/\mathrm{AVar}_1\left(\hat{B}_i\right)$ and plots of relative biases of the variance estimators vs the value of $R^2_{z_1^*\eta_1}$. We note that as the correlation between z_1^* and η_1 gets bigger the calibration effect also increases but only affecting the variance of $\hat{B}_1$. This is because only η_1 is correlated with the z's while η_0 and η_2 are not. The variance estimator $\widehat{\mathrm{AVar}_1\left(\hat{B}_1\right)}$ is an estimator of $\mathrm{AVar}_1\left(\hat{B}_1\right)$ and hence significantly biased for the true variance when $\mathrm{AVar}_1\left(\hat{B}_1\right)$ is significantly different from $\mathrm{AVar}_2\left(\hat{B}_1\right)$.

Second simulation.

We made the calibration variables equal to $\{1, z_1^* = \eta_1 + \text{noise}, z_2^* = \eta_2 + \text{noise}\}$ such that

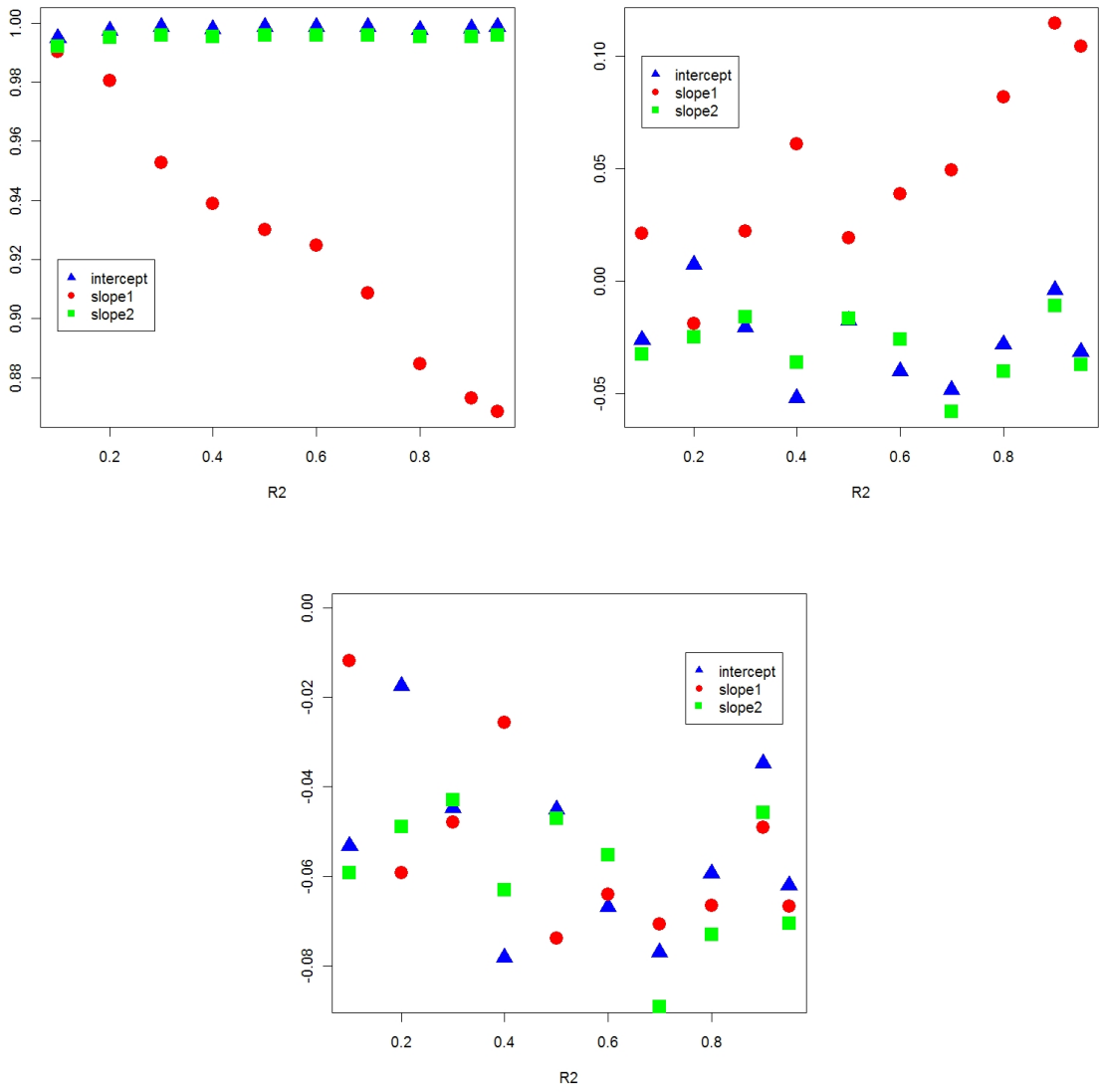

Figure 1: The top left plot shows $\text{AVar}_2\left(\hat{B}_i\right)/\text{AVar}_1\left(\hat{B}_i\right)$ for $i = 0, 1, 2$. The top right plot shows relative biases of $\widehat{\text{AVar}_1\left(\hat{B}_i\right)}$ and the bottom plot shows relative biases of $\widehat{\text{AVar}_2\left(\hat{B}_i\right)}$. All plots have as the horizontal axis the value of $R^2_{z_1^*\eta_1} = 0.1, 0.2, \ldots, 0.9, 0.95$.

$R^2_{z_1^*\eta_1} = R^2_{z_2^*\eta_2} = R^2 = 0.1, 0.2, \ldots, 0.9, 0.95$. Figure 2 shows plots of $\text{AVar}_2\left(\hat{B}_i\right)/\text{AVar}_1\left(\hat{B}_i\right)$ vs the value of R^2 and also relative biases of the variance estimators. We note that as the correlations of the η's with the $z^{*\prime}$s get bigger the calibration effect increases but only affecting the variance of $\hat{B}_1$ and $\hat{B}_2$. The variance of $\hat{B}_0$ is practically the same since $\eta_0 = (e_1, \ldots, e_N)$

is not correlated with the $z^{*\prime}$s.

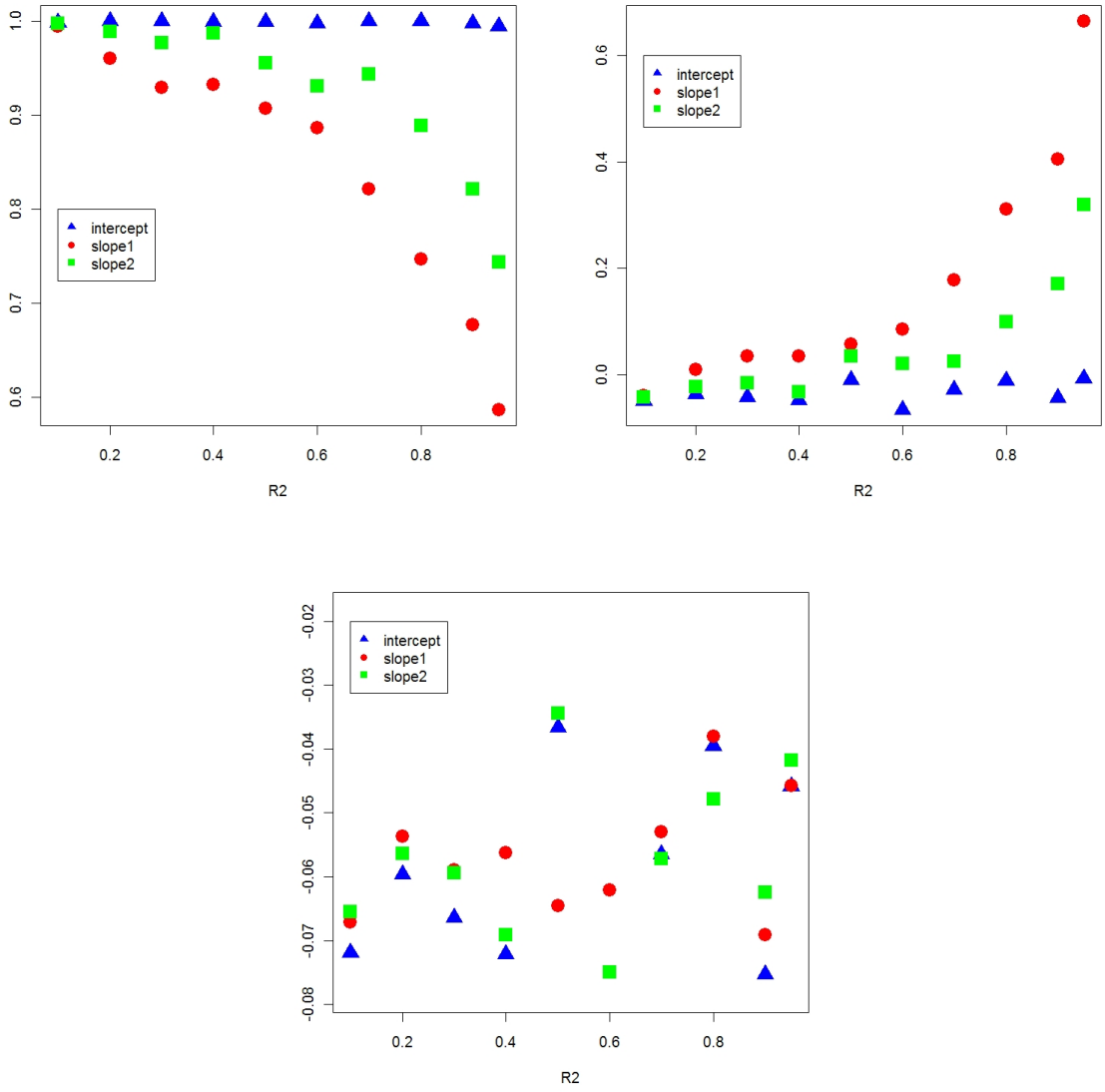

Figure 2: The top left plot shows $\text{AVar}_2\left(\hat{B}_i\right)/\text{AVar}_1\left(\hat{B}_i\right)$ for $i = 0, 1, 2$. The top right plot shows relative biases of $\widehat{\text{AVar}_1\left(\hat{B}_i\right)}$ and the bottom plot shows relative biases of $\widehat{\text{AVar}_2\left(\hat{B}_i\right)}$. All plots have as the horizontal axis the value of $R^2 = 0.1, 0.2, \ldots, 0.9, 0.95$.

Third simulation.

We made the calibration variables equal to $\{1, z_1^* = \eta_0 + \text{noise}, z_2^* = \eta_2 + \text{noise}\}$ such that $R^2_{z_1^*\eta_0} = R^2_{z_2^*\eta_2} = R^2 = 0.1, 0.2, \ldots, 0.9, 0.95$. Figure 3 shows plots of $\text{AVar}_2\left(\hat{B}_i\right)/\text{AVar}_1\left(\hat{B}_i\right)$

vs the value of R^2 and also relative biases of the variance estimators. We note that as the correlations of the η's with the $z^{*\prime}$s get bigger the calibration effect again increases but only affecting the variance of $\hat{B}_0$ and $\hat{B}_2$. The variance of $\hat{B}_1$ is practically the same since $\eta_1 = (x_{1,1}e_1, \ldots, x_{1,N}e_N)$ is not correlated with the $z^{*\prime}$s.

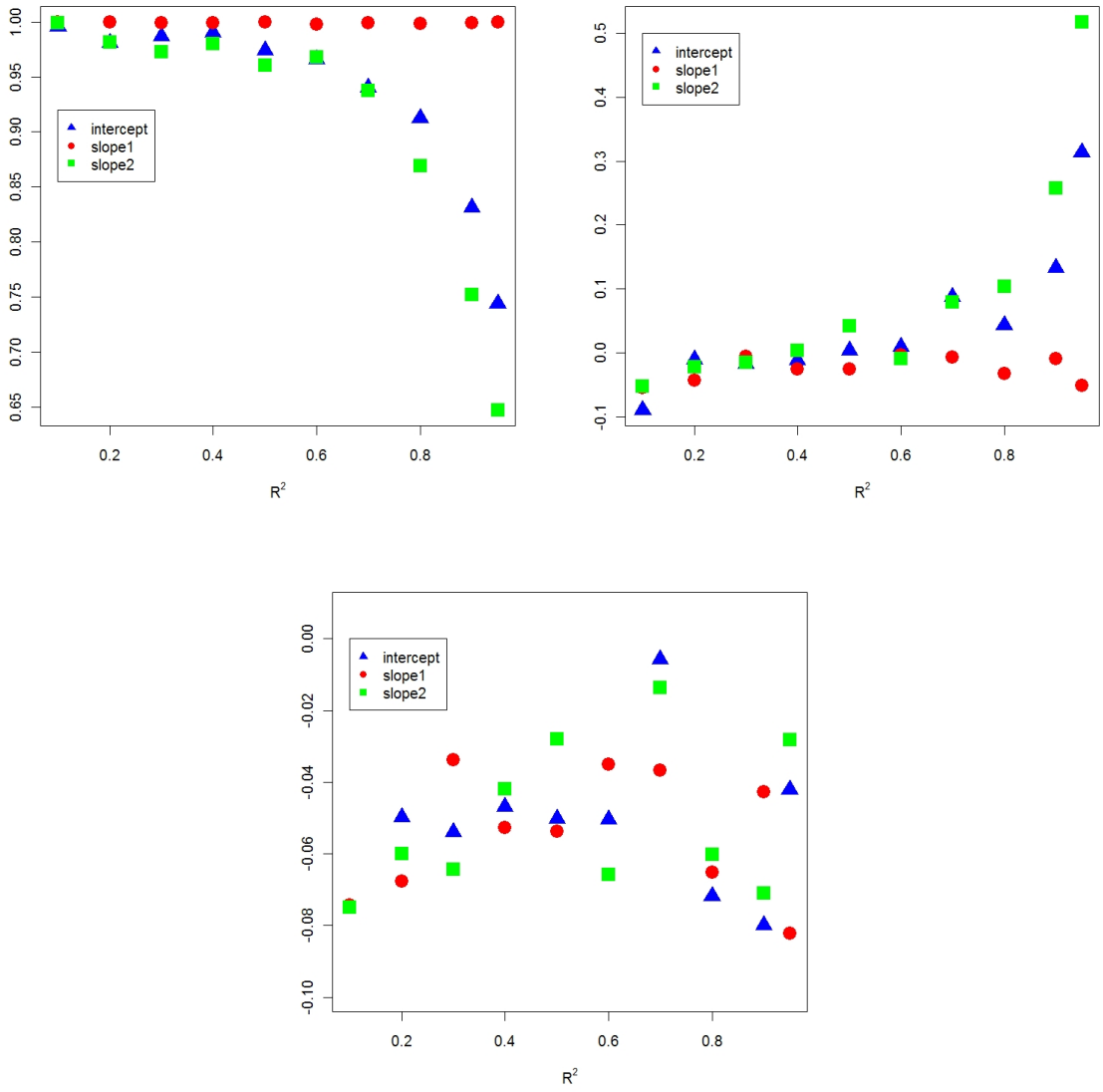

Figure 3: The top left plot shows $\text{AVar}_2\left(\hat{B}_i\right)/\text{AVar}_1\left(\hat{B}_i\right)$ for $i = 0, 1, 2$. The top right plot shows relative biases of $\widehat{\text{AVar}_1\left(\hat{B}_i\right)}$ and the bottom plot shows relative biases of $\widehat{\text{AVar}_2\left(\hat{B}_i\right)}$. All plots have as the horizontal axis the value of $R^2 = 0.1, 0.2, \ldots, 0.9, 0.95$.

2.4 Conclusions

We presented a simulation study for two common estimators, the ratio estimator and the regression estimator, for when sampling weights include calibration. For each of the estimators we explained situations for which ignoring the calibration leads to significantly biased variance estimators. For the ratio estimator and post-stratification, big bias effects occur for when the errors $e_k = y_k - \beta_N x_k$ are correlated with the post-strata indicators $I_{\{k \in U_l\}}$. For the regression estimator in Section 2.3.2, large bias effects occur for when the η_i (defined in result 4) are correlated with the calibration variables. There biases can be significant as shown by table 1 and by figures 1,2 and 3 and these biases are most often ignored in practice.

2.5 Supplement: Result 1

Corollary 5.1.5 in Fuller (1996, p.224-226) states:

Corollary 5.1.5. *Let* $\{\boldsymbol{X}_n\}$ *be a sequence of scalar random variables such that*

$$\boldsymbol{X}_n = a + O_p\left(r_n\right),$$

where $r_n \to 0$ *as* $n \to \infty$. *If* $g\left(x\right)$ *is a function with* s *continuous derivatives at* $x = a$, *then*

$$\begin{aligned} g\left(\boldsymbol{X}_n\right) &= g\left(a\right) + g^{(1)}\left(a\right)\left(\boldsymbol{X}_n - a\right) \\ &\quad + \cdots + \frac{1}{(s-1)!} g^{(s-1)}\left(a\right)\left(\boldsymbol{X}_n - a\right)^{s-1} + O_p\left(r_n^s\right), \end{aligned}$$

where $g^{(j)}\left(a\right)$ *is the* j*th derivative of* $g\left(x\right)$ *evaluated at* $x = a$.

The generalization to vector random variables is given after corollary 5.1.6 in Fuller (1996, p.225-226). Then, for $i = 1, 2, \ldots, m$ we have that

$$\hat{\bar{h}}_i\left(\hat{\bar{\boldsymbol{t}}}\right) = \bar{h}_i + O_p\left(\frac{1}{\sqrt{n}}\right),$$

and again by corollary 5.1.5 in Fuller (1996, p.225-226)

$$g\left(\hat{\bar{\boldsymbol{h}}}\right) = g\left(\bar{\boldsymbol{h}}\right) + \sum_{i=1}^{m} \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i} \sum_{j=1}^{J} \frac{\partial \hat{\bar{h}}_i\left(\bar{\boldsymbol{t}}\right)}{\partial \bar{t}_j} \left(\hat{\bar{t}}_j - \bar{t}_j\right) + O_p\left(\frac{1}{n}\right). \tag{8}$$

2.6 Supplement: Result 2

Using the notation in Särndal et al. (1992, p.235),

$$\frac{\partial \hat{\bar{h}}_i\left(\bar{\boldsymbol{t}}\right)}{\partial \bar{t}_i} \left(\hat{\bar{t}}_i - \bar{t}_i\right) \;=\; \frac{1}{N} \sum_{k \in s} \check{E}_{ik},$$

where $\check{E}_{1k} = \frac{E_{1k}}{\pi_k} = \frac{y_k - \boldsymbol{z}_k^T B_1}{\pi_k}$ and $\check{E}_{ik} = \frac{E_{ik}}{\pi_k} = \frac{x_{ik} - \boldsymbol{z}_k^T B_i}{\pi_k}$ and thus by substitution on equation (8)

$$g\left(\hat{\bar{\boldsymbol{h}}}\right) = g\left(\bar{\boldsymbol{h}}\right) + \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_1} \sum_{k \in s} \frac{\check{E}_{1k}}{N} + \sum_{i=2}^{m} \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i} \sum_{k \in s} \frac{\check{E}_{ik}}{N},$$

thus the asymptotic variance $\mathrm{AV}\left(g\left(\hat{\bar{\boldsymbol{h}}}\right)\right)$ is

$$\mathrm{AV}\left(g\left(\hat{\bar{\boldsymbol{h}}}\right)\right) = \mathrm{Var}\left(\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_1} \sum_{k \in s} \frac{\check{E}_{1k}}{N} + \sum_{i=2}^{m} \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i} \sum_{k \in s} \frac{\check{E}_{ik}}{N}\right),$$

and after rearanging terms

$$\mathrm{AV}\left(g\left(\hat{\bar{\boldsymbol{h}}}\right)\right) = \frac{1}{N^2} \mathrm{Var}\left(\sum_{k \in s}\left[\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_1} \frac{y_k}{\pi_k} + \sum_{i=2}^{m} \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i} \frac{x_{ik}}{\pi_k} - \sum_{i=1}^{m} \frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial \hat{\bar{h}}_i} \frac{\boldsymbol{z}_k^T B_i}{\pi_k}\right]\right).$$

2.7 Supplement: Result 3

For the post-stratification case the vector B_i in result 2 is the vector of averages in each stratum $B_1 = (\bar{y}_{U_1}, \ldots, \bar{y}_{U_L})^T$ and $B_i = (\bar{x}_{iU_1}, \ldots, \bar{x}_{iU_L})^T$. Letting $r_k = \frac{\partial g(\bar{\boldsymbol{h}})}{\partial \bar{h}_1} y_k + \sum_{i=2}^{m} \frac{\partial g(\bar{\boldsymbol{h}})}{\partial \bar{h}_i} x_{ik}$

and substituting in the second term of result 2, we obtain

$$\begin{aligned}\sum_{k\in s}\sum_{i=1}^{m}\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial\hat{\bar{h}}_i}\frac{\boldsymbol{z}_k^T\boldsymbol{B}_i}{\pi_k} &= \sum_{k\in s}\frac{1}{\pi_k}\sum_{l=1}^{L}I_{\{k\in U_l\}}\left(\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial\hat{\bar{h}}_1}\bar{y}_{U_l}+\sum_{i=2}^{m}\frac{\partial g\left(\bar{\boldsymbol{h}}\right)}{\partial\hat{\bar{h}}_i}\bar{x}_{iU_l}\right)\\ &= \sum_{l=1}^{L}\bar{r}_{U_l}\hat{N}_l.\end{aligned}$$

Now the third term in result 2, the covariance term, can be written as

$$\begin{aligned}\mathrm{Cov}\left(\sum_s\frac{1}{\pi_k}r_k,\sum_{l=1}^{L}\bar{r}_{U_l}\hat{N}_l\right) &= \sum_{l=1}^{L}\bar{r}_{U_l}\sum_{i\in U}\sum_{j\in U}\frac{r_i\triangle_{ij}}{\pi_i\pi_j}\\ &= \sum_{l=1}^{L}\bar{r}_{U_l}\left[\sum_{k=1}^{L}\sum_{i\in U}\sum_{j\in U}\frac{r_i\pi_{ij}}{\pi_i\pi_j}-\sum_{i\in U}\sum_{j\in U}r_i\right].\end{aligned}$$

For $i\in U_k$ we rewrite $\frac{r_i\pi_{ij}}{\pi_i\pi_j}$ as $\bar{r}_{U_k}\frac{\pi_{ij}}{\pi_i\pi_j}+\frac{\pi_{ij}}{\pi_i\pi_j}\left(r_i-\bar{r}_{U_k}\right)$ and substituting

$$\begin{aligned}&\mathrm{Cov}\left(\sum_s\frac{1}{\pi_k}r_k,\sum_{l=1}^{L}\bar{r}_{U_l}\hat{N}_l\right)\\ &=\sum_{l=1}^{L}\bar{r}_{U_l}\left[\sum_{k=1}^{L}\sum_{i\in U_k}\sum_{j\in U_l}\bar{r}_{U_k}\frac{\pi_{ij}}{\pi_i\pi_j}+\sum_{k=1}^{L}\sum_{i\in U_k}\sum_{j\in U_l}\left(r_i-\bar{r}_{U_k}\right)\frac{\pi_{ij}}{\pi_i\pi_j}-\sum_{i\in U}\sum_{j\in U_l}r_i\right]\\ &=\sum_{l=1}^{L}\bar{r}_{U_l}\left[\sum_{k=1}^{L}\bar{r}_{U_l}\bar{r}_{U_k}\sum_{i\in U_k}\sum_{j\in U_l}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)+\sum_{k=1}^{L}\sum_{i\in U_k}\sum_{j\in U_l}\left(r_i-\bar{r}_{U_k}\right)\frac{\pi_{ij}}{\pi_i\pi_j}\right]\\ &=\mathrm{Var}\left(\sum_{l=1}^{L}\bar{r}_{U_l}\hat{N}_l\right)+\sum_{l=1}^{L}\bar{r}_{U_l}\sum_{k=1}^{L}\sum_{i\in U_k}\sum_{j\in U_l}\left(r_i-\bar{r}_{U_k}\right)\frac{\pi_{ij}}{\pi_i\pi_j}.\end{aligned}$$

If $\frac{\pi_{ij}}{\pi_i \pi_j} = \begin{cases} a_l, & i = j \\ b_l, & i \neq j \end{cases}$ for $i, j \in U_l$ and $\frac{\pi_{ij}}{\pi_i \pi_j} = c$ for i and j in different stratum, the second term is zero. To see this, first suppose $k = l$, then

$$\begin{aligned} \sum_{i \in U_k} \sum_{j \in U_k} (r_i - \bar{r}_{U_k}) \frac{\pi_{ij}}{\pi_i \pi_j} &= \sum_{i \in U_k} (r_i - \bar{r}_{U_k}) \frac{1}{\pi_i} + \sum_{i \in U_k, j \in U_k, i \neq j} \sum (r_i - \bar{r}_{U_k}) \frac{\pi_{ij}}{\pi_i \pi_j} \\ &= \sum_{i \in U_k} (r_i - \bar{r}_{U_k}) a_k + b_k \left[\sum_{i \in U_k} (r_i - \bar{r}_{U_k}) N_{U_k} - \sum_{i \in U_k} (r_i - \bar{r}_{U_k}) \right] \\ &= 0 + 0, \end{aligned}$$

the terms in the bracket are the sum of all terms in the square minus the diagonal terms. Now suppose $k \neq l$

$$\begin{aligned} \sum_{i \in U_k} \sum_{j \in U_l} (r_i - \bar{r}_{U_k}) \frac{\pi_{ij}}{\pi_i \pi_j} &= \sum_{i \in U_k} (r_i - \bar{r}_{U_k}) \sum_{j \in U_l} \frac{\pi_{ij}}{\pi_i \pi_j} \\ &= 0, \end{aligned}$$

because $\frac{\pi_{ij}}{\pi_i \pi_j}$ is a constant.

2.8 Supplement: Result 4

Define $\gamma_1 = (x_{i1} y_1, \ldots, x_{iN} y_N)^T$ and $\gamma_2 = (x_{i1} \sum_{r=1}^q x_{r1} B_r, \ldots, x_{iN} \sum_{r=1}^q x_{rN} B_r)^T$. Then,

$$\begin{aligned} \gamma_1 - \gamma_2 &= \left[\gamma_1 - \text{proj}\left(\gamma_1; \left\{ z_1^*, \ldots, z_p^* \right\} \right) \right] - \left[\gamma_2 - \text{proj}\left(\gamma_2; \left\{ z_1^*, \ldots, z_p^* \right\} \right) \right] + \\ &\quad \left[\text{proj}\left(\gamma_1; \left\{ z_1^*, \ldots, z_p^* \right\} \right) - \text{proj}\left(\gamma_2; \left\{ z_1^*, \ldots, z_p^* \right\} \right) \right]. \end{aligned}$$

Define $\gamma_3 = \left[\gamma_1 - \text{proj}\left(\gamma_1; \left\{z_1^*, \ldots, z_p^*\right\}\right)\right] - \left[\gamma_2 - \text{proj}\left(\gamma_2; \left\{z_1^*, \ldots, z_p^*\right\}\right)\right]$ and define $\gamma_4 = \left[\text{proj}\left(\gamma_1; \left\{z_1^*, \ldots, z_p^*\right\}\right) - \text{proj}\left(\gamma_2; \left\{z_1^*, \ldots, z_p^*\right\}\right)\right]$. Thus,

$$\begin{aligned} ||\gamma_1 - \gamma_2||^2 &= ||\gamma_3 + \gamma_4||^2 \\ &= ||\gamma_3||^2 + ||\gamma_4||^2, \end{aligned}$$

the last equality due to the fact that γ_3 is orthogonal to γ_4. To see this notice that γ_4 is in the span of $\left\{z_1^*, \ldots, z_p^*\right\}$ while each bracket in γ_3 is orthogonal to every vector in the span of $\left\{z_1^*, \ldots, z_p^*\right\}$. Now notice that $||\gamma_1 - \gamma_2||^2 \propto \text{Var}\left(\hat{t}_{1i}\right)$, $||\gamma_3||^2 \propto \text{Var}\left(\hat{t}_{2i}\right)$ and $\gamma_4 = \text{proj}\left(\gamma_1 - \gamma_2; \left\{z_1^*, \ldots, z_p^*\right\}\right) = \xi_i$.

CHAPTER 3

HIERARCHICAL BAYESIAN SMALL AREA ESTIMATION FOR CIRCULAR DATA

Summary

We consider Bayesian regression models for circular data using the Projected Normal distribution. We show how they can be fit using Markov chain Monte Carlo methods after the introduction of suitable latent variables. We develop novel variational/Laplace approximation to the posterior distribution to dramatically speed up the computations. We apply these methods to a large dataset of daily departures of anglers, which we want to predict based on a set of spatial and temporal categorical covariates. We do model comparison based on the Deviance Information Criterion and make predictions using a composite estimation approach, balancing goodness-of-fit to the observations with prediction stability.

3.1 Introduction

Time-of-day observations are often modeled as coming from a circular distribution, in order to respect the special structure imposed by that type of data. In the application being considered here, we are interested in obtaining predictions of the daily distributions of the departures of recreational anglers along the coasts of the United States, as a function of the type of fishing trip, its location and time of year. The data are collected through the Marine Recreational Fisheries Statistics Survey (MRFSS), which is a national survey of recreational fishing activities in saltwater. Such activities constitute a multi-billion dollar industry in the United States. The survey is conducted by the U.S. National Marine Fisheries Service. Its major goal is to estimate recreational fish catch by species and size class, which are used

in fisheries stock assessments and in fisheries regulations, such as setting quotas on species, start and end dates for the fishing season, etc.

We begin by providing some background on MRFSS and explain why it was necessary to estimate the distributions of daily departures of anglers. MRFSS actually consists of two separate and complementary surveys. The Access Point Angler Intercept Survey (APAIS) collects data on catch at the fishing site. An on-site interviewer "intercepts" anglers as they leave the site. Data from this survey are used to estimate average catch per angler trip. The Coastal Household Telephone Survey (CHTS) is a separate survey that collects fishing activity data, through a stratified random-digit dialing (RDD) sample of households. The CHTS data are used to estimate total angler trips. Finally, the estimated total catch is obtained as the product:

$$\text{(estimated average catch per trip)} \times \text{(estimated total trips)}.$$

While the sampling design of the CHTS is straightforward, the APAIS consists of two or more stages of sampling with different sampling probabilities. The first stage consists of stratified unequal-probability selection of site-days in each fishing mode, with probabilities proportional to a known index of expected fishing pressure. Subsequent stages depend on the fishing mode. For boat-based modes, intermediate stages consist of selection of fishing boats and groups of anglers within boats, with each stage approximated by simple random sampling without replacement. The final stage of sampling in any mode consists of equal-probability selection of anglers observed leaving the site during the interviewer's on-site assignment within the selected day.

A serious problem with estimation in the original APAIS was that it ignored essentially all aspects of this design in the estimation, as noted by a US National Academy of Sciences panel (Sullivan et al. 2006). Valid design-based estimation for APAIS requires sampling weights accounting for the survey design. Consider two stages of selection. Let U_I denote

the set of all site-days in a domain of interest, and U_d denote the set of all anglers departing the site during that day. A sample $s_I \subset U_I$ of site-days is selected via a probability sampling design, with first-order inclusion probabilities $\pi_{Id} > 0$ for $d \in U_I$. Within each selected site-day, a sample $s_d \subset U_d$ of anglers is intercepted by the interviewer as they leave the site. Let $\pi_{a|d} > 0$ denote the probability that angler $a \in U_d$ is intercepted by the interviewer. Then, an unbiased estimator of the total $t_y = \sum_{d \in U_I} \sum_{a \in U_d} y_{da}$ for y_{da} = catch characteristic of angler a on site-day d is given by

$$\hat{t}_y = \sum_{d \in s_I} \sum_{a \in s_d} \frac{y_{da}}{\pi_{Id}\, \pi_{a|d}}.$$

While π_{Id} is based on known fishing pressure and can be readily calculated, the angler inclusion probability $\pi_{a|d}$ is unknown. Let Δ represent the segment of the day during which the interviewer visited the site on a given day. Let $F_{\Delta d}$ denote the fraction of anglers who departed the site during that time segment. The inverse of the angler inclusion probability is interpreted as a weight: the n_d anglers intercepted while departing the site during Δ_d represent all N_d anglers departing the site during Δ_d, and those N_d anglers in turn represent $(1/F_{\Delta d})$ anglers departing the site during the entire day. The weight is then

$$\pi_{a|d}^{-1} = (N_{\Delta d}/n_{\Delta d})(1/F_{\Delta d}).$$

The key problem preventing the use of this weight in APAIS estimation is that $F_{\Delta d}$ is unknown.

The data from the CHTS contain departure times for a random sample of angler trips, so that they can be used to estimate $F_{\Delta d}$. The ultimate goal of the current paper is therefore to obtain estimates of $F_{\Delta d}$ that can be used to obtain weights for the APAIS. In the process of doing so, we develop a flexible modeling approach that makes it possible to specify and fit regression models for circular data, based on the projected normal distribution. Finally, in order to predict the departure distribution for a specific wave, state and fishing mode

combination, composite estimation is used to combine the observed departure distribution with the model fit.

Directional data arise in various ways and in many scientific disciplines. In Meteorology wind directions provide a natural source of circular data, (Breckling, 1989). The times of day at which thunderstorms occur and times of year at which heavy rain occur are other examples of circular data. We can also find circular data in Biology when studying animal navigation for example bird navigation (Schmidt-Koenig, 1965 and Batschelet, 1981). Also in Medicine when analyzing deaths due to a disease at various times of year like month of onset of cases of lymphatic leukaemia in the UK, 1946-1960 (Lee, 1963). Circular data also occur in Psychology with studies of the mental maps which people use to represent their surroundings (Gordon, Jupp, and Byrne, 1989).

A circular observation can be regarded as a point on the unit circle or a unit vector in the plane. Given an initial orientation and direction of the circle, each circular observation can be specified by the angle between the initial direction and the point on the circle corresponding to the observation. The most basic distribution on the circle is the uniform distribution. It is often used as the "null model" in the construction of circular distributions. The most important family of distributions is the Von Mises distributions. The Von Mises distribution is unimodal and symmetrical about the mean direction. It shares many properties on the circle that the normal distribution satisfies on the line. In particular, it arises as maximum entropy distribution on the circle and it also has a maximum likelihood characterization (Mardia and Jupp, 2000, p.42). The two distributions are also related in the following way; if $\mathbf{x} = r\left(\cos\theta, \sin\theta\right)^T$ is bivariate normal with variance matrix $\sigma^2\mathbf{I}$, then the conditional distribution of θ given r is Von Mises (Mardia and Jupp, 2000, p.42).

Other useful distributions are wrapped distributions and projected distributions. Wrapped distributions consist of taking a distribution on the line and wrapping it around the circumference of the circle of unit radius. The most common wrapped distributions are Wrapped Poisson, Wrapped Normal and Wrapped Cauchy (Mardia and Jupp, 2000). Projected distri-

butions are obtained by radial projection of distributions on the plane. Projected normal distributions have been called off-set normal distributions by Mardia (1972), displaced normal by Kendall (1974) and angular normal by Watson (1983). The projected normal distributions were used by Presnell, Morrison, and Littell (1998) to introduce the Spherically Projected Multivariate Linear Model (SPMLM) for directional data. As noted by these authors, the underlying normal distributions make it convenient to specify regression models for circular data, which is also the reason we will be using this approach in the current paper. Recently, a Bayesian analysis for a random sample using the projected normal distribution was done by Nuñez-Antonio and Gutiérrez-Peña (2005). Our approach will also use a Bayesian model specification, and hence can be seen as a generalization of their work.

In this article, we will develop the circular data regression model having as effects mode, state, and wave and describe a Gibbs sampler estimation method. The existing Gibbs samplers are based on the introduction of a latent variable (a length variable) and sampling from its distribution is done via rejection methods. We introduce an extra latent variable that makes it easy and very fast to sample from the latent length distribution. However, due to large sample sizes, and the fact that we will do a model selection procedure, we also explore variational methods which are fast approximations to the posterior distribution based on deterministic algorithms. Finally, we develop a variational/Laplace approximation that seems to do an excellent job in approximating the posterior distribution. This approach is used to fit the SPMLM to the CHTS dataset.

The remainder of the paper is organized as follows. In Section 3.2, we provide basic background on projected normal distributions. In Sections 3.3 and 3.44, we describe the Gibbs sampler and model selection procedures, respectively. Section 3.5 discusses the variational and Laplace approximations. Finally in Section 3.6, the SPMLM fits are incorporated in a composite estimator to predict the fractions of anglers departing the sites.

3.2 Projected normal distribution regression model for departure times

Given the circular nature of the departure times (which can be viewed as angles on $(0, 2\pi]$), we will model them as random variables having a *projected bivariate normal distribution* $PN_2(\boldsymbol{\mu}, \mathbf{I}_2)$ (Presnell et al. 1998) and build a mixed effect model based on the following factors and interactions between them: state (s), wave (w) and mode (m). The dataset contains observations for 17 states, 6 waves, and 4 modes. Some levels for the state factor were removed because of lack of data within those state levels. Because a responding household could report on multiple trips, we also investigated the addition of a household factor (h). This factor has $215,003$ levels, which will require a separate approach to incorporate into the model. Hence, we will begin by describing the model without household effect.

After normalization to the unit circle, the distribution of the departure time T_{ijkt} of respondent t in given state i, wave j and mode k is denoted as

$$T_{ijkt} \overset{ind}{\sim} PN_2(\boldsymbol{\mu}_{ijk}, \boldsymbol{I}_2),$$

where $\boldsymbol{\mu}_{ijk} = \boldsymbol{\mu} + \boldsymbol{m}_k + \boldsymbol{s}_i + \boldsymbol{w}_j$, each term being a two dimensional vector, and $\boldsymbol{I}_2$ is the 2×2 identity matrix. For now, the specification for $\boldsymbol{\mu}_{ijk}$ corresponds to a model without any interaction between the factors.

In general, the angle Θ of a 2-dimensional unit random vector $\boldsymbol{U} = \boldsymbol{X}/\|\boldsymbol{X}\|$ has a projected bivariate normal distribution $PN_2(\boldsymbol{\mu}, \boldsymbol{I}_2)$ if the random variable $\boldsymbol{X}$ has a bivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{I}_2)$. The density of Θ can be written explicitly as Mardia and Jupp (2000, p.46):

$$f(\theta|\boldsymbol{\mu}) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2} \parallel \boldsymbol{\mu} \parallel^2\right\} \left[1 + \frac{\mathbf{u}^T \boldsymbol{\mu} \Phi\left(\mathbf{u}^T \boldsymbol{\mu}\right)}{\phi\left(\mathbf{u}^T \boldsymbol{\mu}\right)}\right] I_{(0,2\pi]}(\theta) \tag{9}$$

with $\mathbf{u} = (\cos\theta, \sin\theta)$, and where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal distribution and

density functions, respectively. The distribution $PN_2(\boldsymbol{\mu}, \boldsymbol{I}_2)$ is unimodal and rotationally symmetric about the mean direction vector $\boldsymbol{\eta}$, where $\boldsymbol{\eta} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\| = (\cos\omega, \sin\omega)$ and ω is called the mean direction.

In order to develop a model for the departure times, the projected normal distribution will be embedded in a hierarchical Bayesian framework that includes prior distributions for the factors. This will allow us to perform model selection, including the determination of whether fixed or random effects specifications are more appropriate for the different factors, and whether interactions between factors are needed. The goal of estimation is to obtain the posterior distribution of $\boldsymbol{\mu}_{ijk}$, which is the only parameter in the projected normal density (9). Once this distribution is obtained, the posterior distribution of any functional of $\boldsymbol{\mu}_{ijk}$, including the expected fraction of departures in a given time interval, can be obtained. We will return to this topic below.

3.3 Estimation

The approach is based on the introduction of suitable latent variables to define an augmented joint distribution with $\boldsymbol{\mu}_{ijk}$ (Nuñez-Antonio and Gutiérrez-Peña, 2005). Conjugate priors will be assumed in order to ensure that we can simulate from all full conditionals required for a Gibbs sampler. Suppose first that we could observe the values of $\boldsymbol{X}_{ijkt}$ for a sample of data and that the mean vector $\boldsymbol{\mu}_{ijk}$ is

$$\boldsymbol{\mu}_{ijk} = \boldsymbol{\mu} + \boldsymbol{m}_k + \boldsymbol{s}_i + \boldsymbol{w}_j, \tag{10}$$

where the mode effect is to be modeled as a "fixed" effect and the state and wave effects are "random" effects. We are interpreting these terms in the Bayesian context, implying that a fixed effect corresponds to having a predetermined vague prior and a random effect to having a prior with a variance parameter with its own prior distribution. The conjugate

priors corresponding to this model specification are

$$\boldsymbol{\mu} \sim N_2\left(\boldsymbol{\mu_0}, \sigma_0^2 \mathbf{I}_2\right)$$

$$\mathbf{m}_k \sim N_2\left(\mathbf{0}, \sigma_m^2 \mathbf{I}_2\right)$$

$$\mathbf{s}_i \mid \sigma_s^2 \sim N_2\left(\mathbf{0}, \sigma_s^2 \mathbf{I}_2\right)$$

$$\mathbf{w}_j \mid \sigma_w^2 \sim N_2\left(\mathbf{0}, \sigma_w^2 \mathbf{I}_2\right)$$

$$\sigma_s^2 \sim IG\left(\alpha_s, \beta_s\right) \propto \left(\sigma_s^2\right)^{-\alpha_s - 1} \exp\left\{-\frac{\beta_s}{\sigma_s^2}\right\}$$

$$\sigma_w^2 \sim IG\left(\alpha_w, \beta_w\right) \propto \left(\sigma_w^2\right)^{-\alpha_w - 1} \exp\left\{-\frac{\beta_w}{\sigma_w^2}\right\}. \tag{11}$$

It is straightforward to obtain the full conditional distributions for this model specification, after which Gibbs sampling can be used to obtain the posterior distributions of the model parameters.

In the application we are considering here, the $\boldsymbol{X}_{ijkt}$ are not observed. Let $\Theta_{ijkt} = T_{ijkt}\frac{2\pi}{24}$ represent the departure times normalized to the unit circle, and let

$$\boldsymbol{X}_{ijkt} = R_{ijkt}\boldsymbol{U}_{ijkt} = R_{ijkt}\left(\cos\Theta_{ijkt}, \sin\Theta_{ijkt}\right)^T.$$

For a given value of the random variable Θ_{ijkt}, it is possible to compute the corresponding value of $\boldsymbol{U}_{ijkt}$. However, the value of $\boldsymbol{X}_{ijkt}$ is unknown because of the unobservable component R_{ijkt}, which corresponds to the length of the vector $\boldsymbol{X}_{ijkt}$. Assuming

$$\boldsymbol{X}_{ijkt} \mid \boldsymbol{\mu}_{ijk} \overset{ind}{\sim} N_2\left(\boldsymbol{\mu}_{ijk}, \boldsymbol{I}_2\right), \tag{12}$$

it is clear that $\Theta_{ijkt} \mid \boldsymbol{\mu}_{ijk} \overset{ind}{\sim} PN_2\left(\boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right)$. The structure of the model suggests that we should treat the unobserved $R_{ijkt} = \|\boldsymbol{X}_{ijkt}\|$, $l = 1, 2, \ldots, n_{ijk}$ as latent variables. This was the approach followed by Nuñez-Antonio and Gutiérrez-Peña (2005), who obtained the

posterior distribution for an overall mean $\boldsymbol{\mu}$ via MCMC.

For a bivariate normal vector $\boldsymbol{X}$, consider the latent variable $R = \|\boldsymbol{X}\|$ defined on $(0, \infty)$. From equation (12), we can obtain its joint distribution with Θ by

$$f\left(\theta, r \mid \boldsymbol{\mu}_{ijk}\right) \;=\; (2\pi)^{-1} \exp\left\{-\frac{1}{2} \parallel \boldsymbol{\mu}_{ijk} \parallel^2\right\} \exp\left\{-\frac{1}{2}\left[r^2 - 2r\left(\mathbf{u}^T \boldsymbol{\mu}_{ijk}\right)\right]\right\} \mid \boldsymbol{J} \mid$$

where $\boldsymbol{J}$ is the Jacobian of the transformation $\mathbf{x} \rightarrow (\theta, r)$. Starting from this distribution, it is in principle again possible to obtain the full conditional distributions required for the Gibbs sampler, based on the conjugate priors described above.

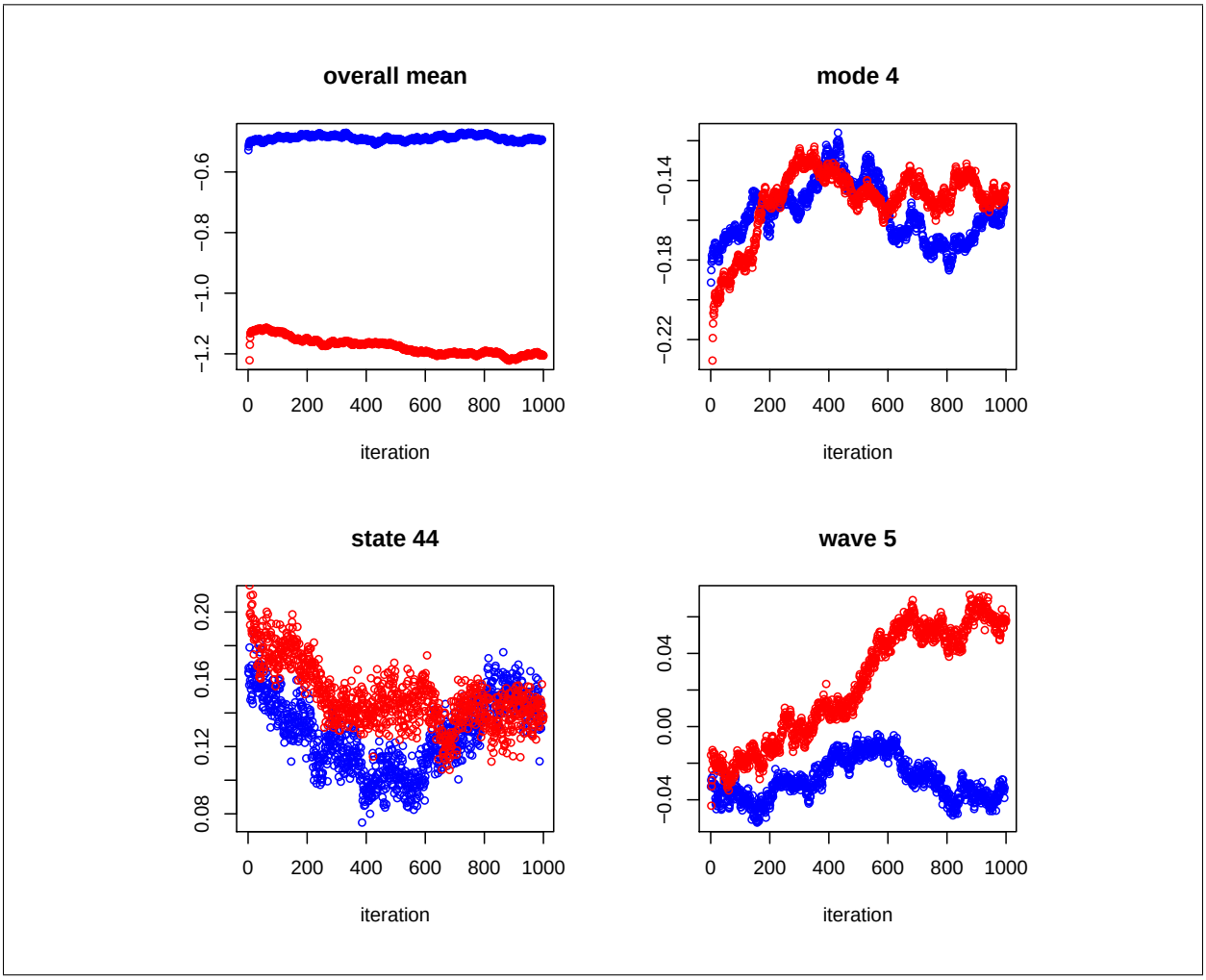

Figure 4: Output from the Gibbs sampler (before transformation) of the overall mean, the fixed effect for mode 4, and the random effects for state 44 and wave 5.

However, implementation of the Gibbs sampler for these data required that a number of issues be addressed. The most important problem was that, if the conjugate priors above were used, the sampling chain failed to converge even after 1000 iterations. Figure 4 shows examples of Gibbs sampler output applied to model (10) with a fixed mode effect and random state and wave effects, for the overall mean parameter and some specific mode, state and wave factors. Note that there are two traces in each plot, because each level of a factor is represented by two parameters.

This slow mixing is most likely due to large positive or negative posterior correlations between model parameters. For an illustration of this problem, we consider the simple random-intercept model as in Gilks et al. (1998, p.94-96),

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

with $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. We assume σ_α^2 and σ^2 known and a flat prior on μ. Let $\boldsymbol{y}$ denote the observed data. Gelfand et al. (1995) show that posterior correlations for this model are

$$\begin{aligned} \text{Corr}\left\{(\alpha_i, \mu) \mid \boldsymbol{y}\right\} &= -\left\{1 + \frac{m\sigma^2}{n\sigma_\alpha^2}\right\}^{-1/2} \\ \text{Corr}\left\{(\alpha_i, \alpha_{i'}) \mid \boldsymbol{y}\right\} &= \left\{1 + \frac{m\sigma^2}{n\sigma_\alpha^2}\right\}^{-1}. \end{aligned}$$

Hence for this simple model, large posterior correlations and poor mixing occur when σ^2/n is small relative to σ_α^2/m, or, speaking somewhat loosely, when the number of levels of the random intercept is small relative to the number of observations. Given the number of levels for the random effects in our data, it seems likely that we are suffering from the same issue here.

Remedies to the slow mixing problem involve reparameterizations of the parameters. Vines et al. (1996) propose a parameterization called *sweeping,* since the mean of the factors

are "swept" from the random and fixed effects effects and absorbed into μ. The sweeping parameterization applied to our model is:

$$\begin{aligned}
\boldsymbol{\mu}' &= \boldsymbol{\mu} + \overline{\mathbf{m}} + \overline{\mathbf{s}} + \overline{\mathbf{w}} \\
\mathbf{m}'_k &= \mathbf{m}_k - \overline{\mathbf{m}} \\
\mathbf{s}'_i &= \mathbf{s}_i - \overline{\mathbf{s}} \\
\mathbf{w}'_j &= \mathbf{w}_j - \overline{\mathbf{w}}.
\end{aligned}$$

Under this reparametrization, the hierarchical model for the departure times can be rewritten as

$$\begin{aligned}
\Theta_{ijkt} &\sim PN_2\left(\boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) \\
\boldsymbol{\mu}_{ijk} &= \boldsymbol{\mu}' + \mathbf{m}'_k + \mathbf{s}'_i + \mathbf{w}'_j \\
\boldsymbol{\mu}' &\sim N_2\left(\boldsymbol{\mu_0}, \sigma_0'^2 = \sigma_0^2 + \frac{\sigma_m^2}{K} + \frac{\sigma_s^2}{I} + \frac{\sigma_w^2}{J}\right) \\
\mathbf{m}'_{-K,c} &\sim N_{K-1}\left(\mathbf{0}, \sigma_m^2\left(\mathbf{I}_{K-1} - \frac{1}{K}\mathbf{J}_{K-1}\right)\right) \\
\mathbf{m}'_{K,c} &= -\sum_{k=1}^{K-1} \mathbf{m}'_{k,c} \\
\mathbf{s}'_{-I,c} \mid \sigma_s^2 &\sim N_{I-1}\left(\mathbf{0}, \sigma_s^2\left(\mathbf{I}_{I-1} - \frac{1}{I}\mathbf{J}_{I-1}\right)\right) \\
\mathbf{s}'_{I,c} &= -\sum_{i=1}^{I-1} \mathbf{s}'_{i,c} \\
\mathbf{w}'_{-J,c} \mid \sigma_w^2 &\sim N_{J-1}\left(\mathbf{0}, \sigma_w^2\left(\mathbf{I}_{J-1} - \frac{1}{J}\mathbf{J}_{J-1}\right)\right) \\
\mathbf{w}'_{J,c} &= -\sum_{j=1}^{J-1} \mathbf{w}'_{j,c} \\
\sigma_s^2 &\sim IG\left(\alpha_s, \beta_s\right) \\
\sigma_w^2 &\sim IG\left(\alpha_w, \beta_w\right)
\end{aligned}$$

where $\mathbf{m}'_{-K,c} = \left(\mathbf{m}'_{1,c}, \ldots, \mathbf{m}'_{K-1,c}\right)^T$, $\mathbf{m}'_{k,c}$ is the c^{th} component of $\mathbf{m}'_k$ and $c = 1, 2$, with

similar definitions for $\mathbf{s}'_{-I,c}$ and $\mathbf{w}'_{-J,c}$; N_p denotes a p-dimensional multivariate normal distribution; $\mathbf{I}_p$ and $\mathbf{J}_p$ are the $p \times p$ identity matrix and matrix of ones respectively. Note that the last components of each factor, $\boldsymbol{m}'_{K,c}, \boldsymbol{s}'_{I,c}$ and $\boldsymbol{w}'_{J,c}$, are fully determined once the other components are known.

To implement a Gibbs sampler for these new prior distributions, we need a new set of full conditional distributions. Let n represent the total sample size and

$$\bar{\mathbf{z}} = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{t=1}^{n_{ijk}} \left(\mathbf{x}_{ijkt} - \mathbf{m}'_k - \mathbf{s}'_i - \mathbf{w}'_j\right)/n.$$

For the mode factor $\boldsymbol{m}$, let $n_k^{(m)} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ijk}$ denote the total number of observations for level k. Let $\mathbf{V}_1^{(m)} = \text{diag}\left(n_1^{(m)}, \ldots, n_{K-1}^{(m)}\right)$ and $\mathbf{V}_2^{(m)} = n_K^{(m)}\mathbf{J} + \frac{1}{\sigma_m^2}\left(\mathbf{I} - \frac{1}{K}\mathbf{J}\right)^{-1} + \mathbf{V}_1^{(m)}$ and

$$\begin{aligned}
\mathbf{z}_{k,c}^{(m)} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n_{ijk}} \left(\mathbf{x}_{ijkt,c} - \boldsymbol{\mu}'_c - \mathbf{s}'_{i,c} - \mathbf{w}'_{j,c}\right) \qquad k = 1, \ldots, K \\
\boldsymbol{\mu}_c^{(m)} &= \left(\left(\mathbf{z}_{1,c}^{(m)} - \mathbf{z}_{K,c}^{(m)}\right)/n_1^{(m)}, \ldots, \left(\mathbf{z}_{K-1,c}^{(m)} - \mathbf{z}_{K,c}^{(m)}\right)/n_{K-1}^{(m)}\right)^T.
\end{aligned}$$

We similarly define $n_i^{(s)}$, $\mathbf{V}_1^{(s)}$, $\mathbf{V}_2^{(s)}$, $\mathbf{z}_{i,c}^{(s)}$, $\boldsymbol{\mu}_c^{(s)}$ and $n_j^{(w)}$, $\mathbf{V}_1^{(w)}$, $\mathbf{V}_2^{(w)}$, $\mathbf{z}_{j,c}^{(w)}$, $\boldsymbol{\mu}_c^{(w)}$ for the remaining two factors. The conditionals for the mean function components can be written explicitly as follows:

$$p\left(\boldsymbol{\mu}' \mid \cdot\right) = N_2\left(\frac{\sigma_0'^2}{1 + n\sigma_0'^2}\left(\frac{\boldsymbol{\mu}_0}{\sigma_0'^2} + n\bar{\mathbf{z}}\right), \frac{\sigma_0'^2}{1 + n\sigma_0'^2}\mathbf{I}_2\right)$$

$$p\left(\mathbf{m}'_{-K,c} \mid \cdot\right) = N_{K-1}\left(\left(\mathbf{V}_2^{(m)}\right)^{-1}\mathbf{V}_1^{(m)}\boldsymbol{\mu}_c^{(m)}, \left(\mathbf{V}_2^{(m)}\right)^{-1}\right)$$

$$p\left(\mathbf{s}'_{-I,c} \mid \cdot\right) = N_{I-1}\left(\left(\mathbf{V}_2^{(s)}\right)^{-1}\mathbf{V}_1^{(s)}\boldsymbol{\mu}_c^{(s)}, \left(\mathbf{V}_2^{(s)}\right)^{-1}\right)$$

$$p\left(\mathbf{w}'_{-J,c} \mid \cdot\right) = N_{J-1}\left(\left(\mathbf{V}_2^{(w)}\right)^{-1}\mathbf{V}_1^{(w)}\boldsymbol{\mu}_c^{(w)}, \left(\mathbf{V}_2^{(w)}\right)^{-1}\right).$$

The conditional distributions for the random effect variances are specified by

$$p\left(\sigma_s^2 \mid \cdot\right) \propto IG\left(\alpha_s + I - 1, \sum_{i=1}^{I} \frac{1}{2}\mathbf{s}_i'^T \mathbf{s}_i' + \beta_s\right) * \left(\frac{1}{\sigma_0'^2} \exp\left(-\frac{1}{\sigma_0'^2}\left[\frac{1}{2}\left(\boldsymbol{\mu}' - \boldsymbol{\mu}_0\right)^T\left(\boldsymbol{\mu}' - \boldsymbol{\mu}_0\right)\right]\right)\right)$$

$$p\left(\sigma_w^2 \mid \cdot\right) \propto IG\left(\alpha_w + J - 1, \sum_{j=1}^{J} \frac{1}{2}\mathbf{w}_j'^T \mathbf{w}_j' + \beta_w\right) * \left(\frac{1}{\sigma_0'^2} \exp\left(-\frac{1}{\sigma_0'^2}\left[\frac{1}{2}\left(\boldsymbol{\mu}' - \boldsymbol{\mu}_0\right)^T\left(\boldsymbol{\mu}' - \boldsymbol{\mu}_0\right)\right]\right)\right).$$

Finally, the conditional distribution of the latent length of the bivariate normal vector is

$$p\left(r_{ijkt} \mid \cdot\right) \propto r \exp\left[-\frac{1}{2}r^2 + b_{ijkt} r\right] \mathrm{I}_{(0,\infty)}\left(r\right), \tag{13}$$

with $b_{ijkt} = \mathbf{u}_{ijkt}^T \boldsymbol{\mu}_{ijk}$ and $\mathbf{u}_{ijkt}^T = \left(\cos\theta_{ijkt}, \sin\theta_{ijkt}\right)^T$.

The centered mean model components $\mathbf{m}'_{-K,c}$, $\mathbf{s}'_{-I,c}$ and $\mathbf{w}'_{-J,c}$ are correlated random vectors, so that they need to be generated as a block for each of the factors. Nevertheless, sampling from the corresponding multivariate normal distributions $p\left(\boldsymbol{\mu}' \mid \cdot\right)$, $p\left(\mathbf{m}'_{-K,c} \mid \cdot\right)$, $p\left(\mathbf{s}'_{-I,c} \mid \cdot\right)$, $p\left(\mathbf{w}'_{-J,c} \mid \cdot\right)$ is readily accomplished directly. Sampling from $p\left(\sigma_s^2 \mid \cdot\right)$ and $p\left(\sigma_w^2 \mid \cdot\right)$ is done via a Metropolis-Hastings algorithm with proposal distributions $IG\left(\alpha_s + I - 1, \sum_{i=1}^{I} \frac{1}{2}\mathbf{s}_i^T \mathbf{s}_i + \beta_s\right)$ and $IG\left(\alpha_w + J - 1, \sum_{j=1}^{J} \frac{1}{2}\mathbf{w}_j^T \mathbf{w}_j + \beta_w\right)$, respectively. Because the r_{ijkt} are latent, each iteration of the Gibbs sampler needs to draw n values from the conditional distribution (13). This density is concave and it can be shown that it belongs to an exponential family with canonical parameter b_{ijkt}, so that sampling from it is straightforward in principle, using algorithms such as Metropolis-Hastings or adaptive rejection sampling. However, the very large sample necessitates the use of an efficient algorithm to ensure that the Gibbs sampler can be run to convergence. In particular, we wanted to avoid having to explicitly iterate over the n random draws from the conditional distribution for each Gibbs

sampler realization.

We therefore implemented a method that allows us to sample directly from latent length distribution $p(r|b)$ after introducing a new convenient latent variable inside the Gibbs sampler. The idea comes from the slice sampler technique (Givens and Hoeting, 2005, pp.221-223), which is based on the introduction of suitable auxiliary variables. Defining $b = \mathbf{u}^T \boldsymbol{\mu}$, the distribution of the latent length is

$$p(r|\boldsymbol{\mu}, \theta) \propto r \exp\left(-\frac{1}{2}(r-b)^2\right). \tag{14}$$

We introduce the latent variable Y which has joint density with r given by

$$p(r, y|\boldsymbol{\mu}, \theta) \propto r \mathrm{I}_{\left(0, \exp\left\{-\frac{1}{2}(r-b)^2\right\}\right)}(y)\, \mathrm{I}_{(0,\infty)}(r).$$

Then, the full conditionals are

$$\begin{aligned} (Y|R=r, \boldsymbol{\mu}, \theta) &\sim \mathrm{U}\left(0, \exp\left\{-\frac{1}{2}(r-b)^2\right\}\right) \\ p(r|Y=y, \boldsymbol{\mu}, \theta) &\propto r \mathrm{I}_{\left(b+\max\left\{-b, -\sqrt{-2\ln y}\right\}, b+\sqrt{-2\ln y}\right)}(r), \end{aligned} \tag{15}$$

where this last one is very easy to sample from using the inverse cdf technique. Thus we draw $y \sim \mathrm{U}\left(0, \exp\left\{-\frac{1}{2}(r-b)^2\right\}\right)$ and independently we draw $u \sim \mathrm{U}(0,1)$. Finally we get a draw r by letting $r = \sqrt{(r_2^2 - r_1^2)\, u + r_1^2}$, where $r_1 = b + \max\left\{-b, -\sqrt{-2\ln y}\right\}$ and $r_2 = b + \sqrt{-2\ln y}$.

Sampling from $p(r_{ijkt} \mid \cdot)$ in this manner dramatically reduces the running time of the sampler since it only involves two uniform draws and finding the maximum of two numbers.

The model specification we have used so far contains a fixed effect for mode and random effects for wave and state. We will investigate alternative model specifications including interactions and chose between them, as further described in the next section. We now return to the issue of adding a respondent household factor in the model, to account for the

fact that a household can report on multiple trips. We consider here the following model

$$\boldsymbol{\mu}_{ijklt} = \boldsymbol{\mu} + \mathbf{m}_k + \mathbf{sw}_{ij} + \mathbf{h}_l \tag{16}$$

with $\boldsymbol{sw}_{ij}$ denoting a random interaction term for state and wave and $\boldsymbol{h}_l$ a bivariate random effect for household l. The sweeping adjustment discussed for the previous model specification was applied to the state-wave interaction term but not to the household term. The factor $\boldsymbol{h}$ has over 215,000 levels, so that it is simply not practical to apply the sweeping adjustment in this instance. However, based on the result by Gelfand et al. (1995) regarding the posterior correlations, we conjectured that the very large number of levels would result in negligible correlation between the levels of that factor. Figure 5 displays examples of the Gibbs sampler output applied to model (16). There do not appear to be undue convergence problems in this case.

3.4 Model selection

The previous section described the set-up of a Gibbs sampler that, when run to convergence, provides an approximation to the posterior distribution of all the model parameters. In order to find a suitable model for the distribution of the departure times, we wanted to investigate and compare alternative model specifications, including considering models with some of these effects removed, different effects treated as fixed and random, and interaction between these models.

A common measure of fit in the Bayesian literature is the *deviance* (Gelman et al. 2004, p.179-184). For a general Bayesian estimation problem, the deviance is defined as $D(y, \omega) = -2 \ln p(y \mid \omega)$ where y are the data, ω are the unknown parameters and $p(y \mid \omega)$ is the likelihood function. The deviance is proportional to the mean squared error if the model is normal with constant variance. The expected deviance $\mathrm{E}(D(y, \omega) \mid y)$ is a measure of how well the model fits and it can be estimated by the posterior mean deviance $\overline{D(y)} =$

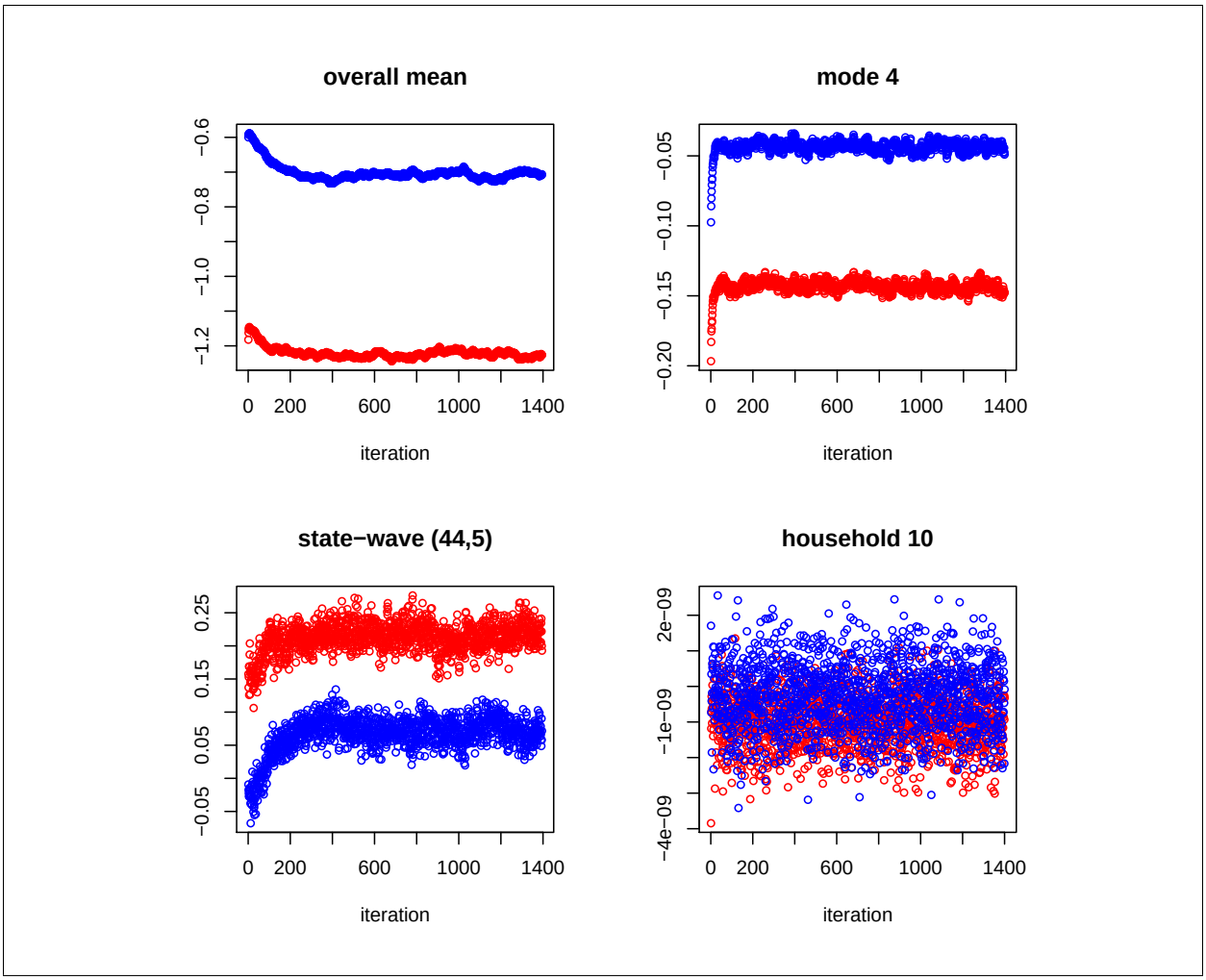

Figure 5: Output from the Gibbs sampler (after transformation) of the overall mean, the fixed effect representing mode 4, the interaction random effect representing state 44 and wave 5, and the random effect representing household 10.

$\frac{1}{B}\sum_{b=1}^{B} D(y,\omega^b)$, where the $\omega_b, b = 1,\ldots,B$ are random draws from the posterior distribution (as obtained from an MCMC chain at convergence). Let $\bar{\omega} = \sum_{b=1}^{B}\omega_b/B$. The difference between the posterior mean deviance and the deviance at $\bar{\omega}$,

$$p_D = \overline{D(y)} - D(y,\bar{\omega}),$$

is often interpreted as a measure of the effective number of parameters of a Bayesian model. More generally, p_D can be thought of as the number of "unconstrained" parameters in

the model, where a parameter counts as 1 if it is estimated without constraints or prior information, 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution, or an intermediate value if both the data and prior distributions are informative. For hierarchical models, the effective number of parameters strongly depends on the variance of the group-level parameters (Gelman et al. 2004, p.182).

A common model selection criterion in the Bayesian estimation context is the *deviance information criterion* (DIC):

$$\begin{aligned} DIC &= 2\overline{D(y)} - D(y, \bar{\omega}) \\ &= \overline{D(y)} + p_D. \end{aligned}$$

The DIC can be interpreted as a measure of goodness-of-fit, i.e. the estimated expected deviance, plus a "penalty" for model complexity in the form of the total number of effective parameters. When performing model selection, models with lower values of DIC are viewed as providing a more preferable tradeoff between fit and model complexity. We therefore used DIC to compare different model specifications for the departure time data.

Table 2 shows the DIC values obtained from different models applied to the departure time data. In interpreting the number of parameters, it should be noted that a level of a factor (e.g. $\boldsymbol{m}_k$) is represented by a pair of parameters. Hence, in a sweeping-reparametrized model with only a mode effect, there are 2 parameters for the overall mean and the 4-1=3 remaining free mode levels are represented in the projected normal model by 6 parameters, for a total of 8 possible parameters. The model with only a mode effect, with results in the first row of Table 2, resulted in a value of $p_D = 7.63$. The reduction of p_D relative to the fully unconstrained value of 8 indicates that the prior distributions resulting in only a small amount of "shrinking" for this factor. It is possible to similarly interpret the values of p_D for the other model specifications.

The results in Table 2 clearly show that models containing all three factors (mode, state,

Table 2: DIC results for different model specifications for the departure time data.

Fixed effects	Random effects	$D(\bar{\omega})$	$\overline{D(\omega)}$	p_D	DIC
Mode		2642698	2642706	7.631472	2642713
Mode	Wave	2631890	2631908	17.85838	2631926
Mode, Wave		2631890	2631907	17.2165	2631924
Mode	State	2628456	2628496	40.25385	2628536
Mode, State		2628456	2628495	39.48233	2628535
Mode	State, Wave	2618338	2618387	49.50114	2618437
Mode, Wave	State	2618337	2618387	49.07431	2618436
Mode, State	Wave	2618337	2618387	49.17294	2618436
Mode, State, Wave		2618337	2618386	49.06591	2618435
Mode, State×Wave		2615651	2615856	205.7165	2616062
Mode	State×Wave	2615669	2615848	178.9047	2616027
Mode	State×Wave, Household	2615668	2615847	179.3620	2616026

wave) consistently achieve lower DIC values than models that excluded any of those factors. While not shown here, models with mode as random effect performed worse than models with mode as fixed effect. In contrast, very similar DIC values were obtained with state and wave treated as either fixed or random. When we investigated models with interactions between the three factors, those with state-wave interactions scored better than any other arrangement of two-way interactions. As shown in Table 2, treating the state-wave interactions as a random effect gave better results than treating it as a fixed effect. Finally, the addition of a household random effect gave results that were virtually identical to those without the household effect. Based on all this, the following model was selected as the best-fitting model:

$$\boldsymbol{\mu}_{ijk} = \boldsymbol{\mu}' + \mathbf{m}'_k + \mathbf{sw}'_{ij}$$

where $\boldsymbol{sw}_{ij}$ denotes a random interaction effect between state and wave, with 99 total levels.

With the algorithm improvements discussed in Section 3.3, the Gibbs sampler still took a substantial time to run: about 2.15 seconds per iteration. Hence, we also explored an alternative approach to obtain the posterior distribution that would be more computationally efficient. In the next section we explore variational and Laplace approximations which are very fast approximations to the posterior distribution. We first consider the case of a random

sample of projected normals and then extend the method for the regression models considered so far.

3.5 Variational/Laplace approximation

Variational approximation methods involve approximations to marginal or posterior distributions in terms of an optimization problem (Ghahramani and Beal, 2001; Opper and Saad, 2001; Wainwright and Jordan, 2003). Mean-field methods are based on optimizing the Kullback-Leibler (KL) divergence with respect to a variational distribution (Ormerod and Wand, 2010). The KL divergence is defined as $\int q(w) \log\left\{\frac{q(w)}{p(w|y)}\right\} dw$ which is greater or equal than zero for all densities q and equal to zero if and only if $q(w) = p(w|y)$. Consider a model with parameters $\mathbf{W}$ and observations $\mathbf{y}$. Tractability is achieved by restricting q to a more manageable class of densities and then minimizing the KL divergence between q and p over that class. The restriction for the q density is that $q(w)$ factorizes into $\prod q_i(w_i)$ for some partition $\{w_1, \ldots, w_M\}$ of $\mathbf{w}$. It can be shown that the solutions satisfy (Ormerod and Wand, 2010):

$$q_i^*(w_i) \propto \exp\left\{\mathrm{E}_{-w_i} \log p(w_i|y, \mathbf{w}_{-i})\right\}, \tag{17}$$

where $\mathbf{w}_{-i}$ is $\mathbf{w}$ without w_i and E_{-w_i} denotes expectation with respect to the density $\prod_{j\neq i} q_j(w_j)$. Next we present a result that will allow us to compute some expectations needed for the algorithms in this section. The proof is given in Supplement 3.8.

Result 5. *Let* $r(\cos\theta, \sin\theta)^T = (X_1, X_2) \sim \mathrm{N}_2(\boldsymbol{\mu}, \mathbf{I}_2)$. *Define* $b = (\cos\theta, \sin\theta)^T \boldsymbol{\mu}$ *and* $C(b) = 1 + \sqrt{2\pi} b \exp\left\{\frac{1}{2}b^2\right\} \Phi(b)$. *Then the distribution of the latent length is* $p(r|\theta, \boldsymbol{\mu}) = \frac{1}{C(b)} r \exp\left\{-\frac{1}{2}r^2 + br\right\}$, *its moment generating function is*
$\frac{1}{C(b)}\left[1 + \sqrt{2\pi}(b+t)\exp\left\{\frac{1}{2}(b+t)^2\right\}\Phi(b+t)\right]$ *and the variance is bounded by 1,*
$0 < \mathrm{Var}(r|\theta, \boldsymbol{\mu}) < 1$.

We will first consider the case of a random sample of projected normals and then we will consider the regression case.

Algorithm 1 Iterative scheme for obtaining the parameters in the optimal densities for the random sample case.

Initialize $\mathrm{E}_{\mu}(\boldsymbol{\mu})$.
Cycle:

$$\begin{aligned} \mathrm{E}_{-\mu} &\leftarrow \sum \mathbf{u}_i \mathrm{E}(r_i) \\ \mathrm{E}_{\mu} &\leftarrow \frac{\boldsymbol{\mu}_0/\sigma_0^2 + \mathrm{E}_{-\mu}}{n + 1/\sigma_0^2}, \end{aligned}$$

where $\mathrm{E}(r_i) = b_i + \frac{\sqrt{2\pi}}{C(b_i)}\Phi(b_i)\exp\left\{\frac{1}{2}b_i^2\right\}$ and $C(b_i) = 1 + \sqrt{2\pi}b_i \exp\left(\frac{b_i^2}{2}\right)\Phi(b_i)$ and $b_i = \mathbf{u}_i^T \mathrm{E}_{\mu}$.

3.5.1 Random sample

Consider $\Theta_1, \Theta_2, \ldots, \Theta_n \overset{iid}{\sim} PN_2(\boldsymbol{\mu}, \mathbf{I}_2)$ and prior $p(\boldsymbol{\mu}) = N_2(\boldsymbol{\mu}_0, \sigma_0^2\mathbf{I}_2)$. We then approximate $p(\boldsymbol{\mu}, r_1, r_2, \ldots, r_n|\boldsymbol{\theta})$ by $q(\boldsymbol{\mu}, r_1, r_2, \ldots, r_n) = q_{\mu}(\boldsymbol{\mu}) q_{r_1}(r_1) q_{r_2}(r_2) \cdots q_{r_n}(r_n)$. From equation (17), the optimal densities take the form

$$\begin{aligned} q_{\mu}^*(\boldsymbol{\mu}) &\propto \exp\{\mathrm{E}_{-\mu} \log p(\boldsymbol{\mu}|\boldsymbol{\theta}, \mathbf{r})\} \\ &\propto N_2\left(\frac{\boldsymbol{\mu}_0/\sigma_0^2 + \mathrm{E}_{-\mu}(\sum r_i \mathbf{u}_i)}{n + (1/\sigma_0^2)}, \frac{1}{n + (1/\sigma_0^2)}\mathbf{I}_2\right) \\ q_{r_i}^*(r_i) &\propto \exp\{\mathrm{E}_{-r_i} \log p(r_i|\boldsymbol{\theta}, \boldsymbol{\mu}, \mathbf{r}_{-i})\} \\ &\propto r_i \exp\left(-\frac{1}{2}r_i^2 + r_i \mathbf{u}_i^T \mathrm{E}_{\mu}(\boldsymbol{\mu})\right), \end{aligned}$$

where $u_i^T = (\cos\theta_i, \sin\theta_i)$, $\mathrm{E}_{-\mu}(\sum r_i u_i) = \sum u_i \mathrm{E}(r_i) = \sum u_i \int r_i q_{r_i} dr_i$, and $\mathrm{E}_{\mu}(\boldsymbol{\mu}) = \int \boldsymbol{\mu} q_{\mu} d\boldsymbol{\mu}$. Each expectation is then repeatedly updated giving us the algorithm in Algorithm 1. The expectation $\mathrm{E}(r_i)$, shown in Algorithm 1, can be obtained from direct integration or from the moment generating function of the distribution of the latent length, see Supplement 3.8. Our factorization of q is such that we gain considerable tractability but this typically leads to low accuracy. The variance for $q_{\mu}(\boldsymbol{\mu})$ only depends on sample size and not on the data, which could lead to poor approximations especially when $\boldsymbol{\mu}$ is far away from the origin, Figure 6. However, it can be shown that for a random sample, the algorithm in Table 1 converges to the mode of the posterior distribution $p(\boldsymbol{\mu}|\boldsymbol{\theta})$ also called the maximum

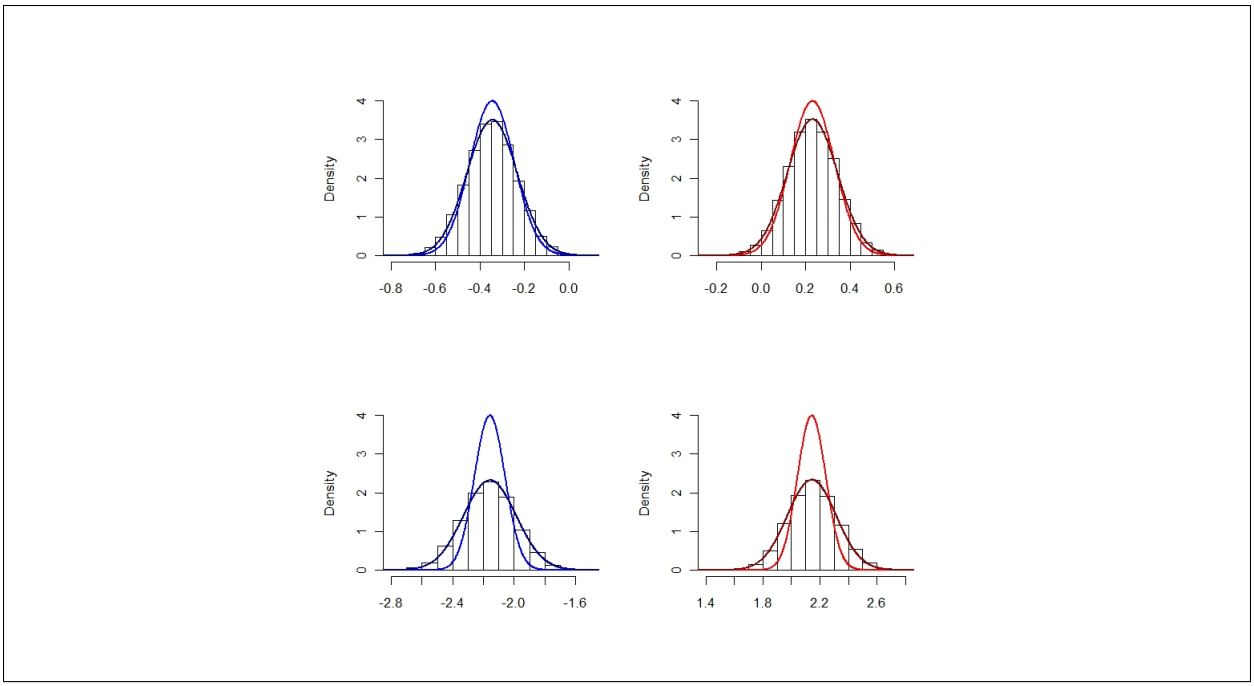

Figure 6: Histograms of the output from the Gibbs sampler with variational and Laplace approximations for a random sample of size 100 from projected normal with $||\boldsymbol{\mu}||^2 = 0.5$ (top) and for a random sample of size 100 with $||\boldsymbol{\mu}||^2 = 3$ (bottom). The tighter curves are the ones obtained by the variational approximation while the more accurate ones are obtained by the Laplace approximation.

a posteriori (MAP). The next result states this fact more explicitly. The proof is given in Supplement 3.9.

Result 6. *Let* $\theta_1, \ldots, \theta_n \overset{iid}{\sim} \mathrm{PN}_2(\boldsymbol{\mu}, \mathbf{I}_2)$ *and let* $p(\boldsymbol{\mu}) = \mathrm{N}_2(\boldsymbol{\mu}_0, \sigma^2\mathbf{I}_2)$. *Then the posterior distribution* $p(\boldsymbol{\mu}|\boldsymbol{\theta})$ *is unimodal and for any initialization of Algorithm 1, the algorithm converges to the posterior mode.*

To fix the problem of the variance of $q_\mu(\boldsymbol{\mu})$ being too tight we decided to compute a Laplace approximation to the posterior distribution of $\boldsymbol{\mu}$ using as mode E_μ obtained after convergence of Algorithm 1. In the Laplace approximation the goal is to find a multivariate Gaussian approximation which is centered on the mode of the posterior distribution of $\boldsymbol{\mu}$. The covariance matrix is taken as the inverse of minus the Hessian of the log posterior distribution evaluated at the mode (Bishop, 2006, pp.213-216). We will take E_μ (from the variational

method) as the mode of the Laplace approximation. The log posterior distribution for $\boldsymbol{\mu}$ is

$$\log p\left(\boldsymbol{\mu}|\boldsymbol{\theta}\right) = \log N_2\left(\boldsymbol{\mu}_0, \sigma_0^2\mathbf{I}_2\right) + \sum_{i=1}^{n} \log PN_2\left(\theta_i; \boldsymbol{\mu}, \mathbf{I}_2\right) + C$$

where C is a term that does not depend on $\boldsymbol{\mu}$. Thus, the Laplace approximation to the posterior distribution $p\left(\boldsymbol{\mu}|\boldsymbol{\theta}\right)$ is a bivariate normal distribution $\mathrm{N}_2\left(\mathrm{E}_\mu, \mathrm{V}\right)$ where

$$-\mathrm{V}^{-1} = \begin{pmatrix} \frac{\partial^2}{\partial\mu_1^2}\log p\left(\boldsymbol{\mu}|\boldsymbol{\theta}\right)|_{\mathrm{E}_\mu} & \frac{\partial^2}{\partial\mu_1\partial\mu_2}\log p\left(\boldsymbol{\mu}|\boldsymbol{\theta}\right)|_{\mathrm{E}_\mu} \\ \frac{\partial^2}{\partial\mu_1\partial\mu_2}\log p\left(\boldsymbol{\mu}|\boldsymbol{\theta}\right)|_{\mathrm{E}_\mu} & \frac{\partial^2}{\partial\mu_2^2}\log p\left(\boldsymbol{\mu}|\boldsymbol{\theta}\right)|_{\mathrm{E}_\mu} \end{pmatrix}$$

The second derivatives needed to compute the Hessian are given in Supplement 3.10. We can see from Figure 6 that $p\left(\boldsymbol{\mu}|\boldsymbol{\theta}\right)$ is very well approximated by the Laplace approximation. The next subsection will treat the regression case and find similar results.

3.5.2 Regression model

Consider $\Theta_{ijkt}{\sim}PN_2\left(\boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right)$, independent with $\mu_{ijk} = \mu + m_k + s_i + w_j$ and with normal priors with fixed variances on the overall and mode effects and random variances on the state and wave effects as specified in equation (11). Here we take the variational distribution as

$$q\left(\boldsymbol{\mu}, \mathbf{m}, \mathbf{s}, \mathbf{w}, \sigma_s^2, \sigma_w^2, r_1, r_2, \ldots, r_n\right) = q_\mu \prod q_{m_k} \prod q_{s_i} \prod q_{w_j} \prod q_{r_t} q_{\sigma_s^2} q_{\sigma_w^2}.$$

Making use of equation (17), the variational densities take the form

$$\begin{aligned} q_\mu^*\left(\boldsymbol{\mu}\right) &= N_2\left(\frac{\boldsymbol{\mu}_0/\sigma_0^2 + N\mathrm{E}\left(\bar{\mathbf{z}}\right)}{N + \left(1/\sigma_0^2\right)}, \frac{1}{N + \left(1/\sigma_0^2\right)}\mathbf{I}_2\right) \\ q_{m_k}^*\left(\boldsymbol{m}_k\right) &= N_2\left(\frac{\boldsymbol{m}_0/\sigma_m^2 + \mathrm{E}\left(\mathbf{z}_k^{(m)}\right)}{n_k^{(m)} + \left(1/\sigma_m^2\right)}, \frac{1}{n_k^{(m)} + \left(1/\sigma_m^2\right)}\mathbf{I}_2\right) \\ q_{s_i}^*\left(\boldsymbol{s}_i\right) &= N_2\left(\frac{\mathrm{E}\left(\mathbf{z}_i^{(s)}\right)}{n_i^{(s)} + \mathrm{E}\left(1/\sigma_s^2\right)}, \frac{1}{n_i^{(s)} + \mathrm{E}\left(1/\sigma_s^2\right)}\mathbf{I}_2\right) \end{aligned}$$

$$
\begin{aligned}
q^*_{w_j}(\boldsymbol{w}_j) &= N_2\left(\frac{\mathrm{E}\left(\mathbf{z}_j^{(w)}\right)}{n_j^{(w)}+\mathrm{E}\left(1/\sigma_w^2\right)}, \frac{1}{n_j^{(w)}+\mathrm{E}\left(1/\sigma_w^2\right)}\mathbf{I}_2\right) \\
q^*_{\sigma_s^2}\left(\sigma_s^2\right) &= IG\left(\alpha_s + I - 1, \mathrm{E}_s \sum_{i=1}^{I} \frac{1}{2}\mathbf{s}_i^T\mathbf{s}_i + \beta_s\right) \\
q^*_{\sigma_w^2}\left(\sigma_w^2\right) &= IG\left(\alpha_w + J - 1, \mathrm{E}_w \sum_{j=1}^{J} \frac{1}{2}\mathbf{w}_j^T\mathbf{w}_j + \beta_w\right) \\
q^*_{r_{ijkt}}(r_{ijkt}) &\propto r_{ijkt}\exp\left(-\frac{1}{2}r_{ijkl}^2 + r_{ijkt}\mathbf{u}_{ijkt}^T\mathrm{E}\left(\boldsymbol{\mu}_{ijk}\right)\right)
\end{aligned}
$$

where $\bar{\mathbf{z}}, \mathbf{z}_k^{(m)}, \mathbf{z}_i^{(s)}, \mathbf{z}_j^{(w)}$ are defined as previously and $\mathrm{E}_s \sum_{i=1}^{I} \frac{1}{2}\mathbf{s}_i^T\mathbf{s}_i = \sum_{i=1}^{I} \mathrm{Var}\left(\mathbf{s}_{i1}\right) + \mathrm{Var}\left(\mathbf{s}_{i2}\right) + \left(\mathrm{E}\left(\mathbf{s}_{i1}\right)\right)^2 + \left(\mathrm{E}\left(\mathbf{s}_{i2}\right)\right)^2$ and $\mathrm{E}_w \sum_{j=1}^{J} \frac{1}{2}\mathbf{w}_j^T\mathbf{w}_j$ defined similarly. All expectations are with respect to the variational density. Each expectation is then repeatedly updated giving us Algorithm 2. In analogy with the random sample case we will take $\left(\mathrm{E}\left(\boldsymbol{\mu}\right), \mathrm{E}\left(\mathbf{m}\right), \mathrm{E}\left(\mathbf{s}\right), \mathrm{E}\left(\mathbf{w}\right)\right)$, obtained after convergence of Algorithm 2, as the mode of the posterior distribution and get the Laplace approximation by finding the minus inverse of the Hessian matrix of the log posterior distribution evaluated at the mode. The log posterior distribution for $(\boldsymbol{\mu}, \mathbf{m}, \mathbf{s}, \mathbf{w})$

$\log p\left(\boldsymbol{\mu}, \mathbf{m}, \mathbf{s}, \mathbf{w}|\boldsymbol{\theta}\right)$ is

$$
\begin{aligned}
&\log N_2\left(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \sigma_0^2\mathbf{I}_2\right) + \sum_{k=1}^{K} \log N_2\left(\mathbf{m}_k; \mathbf{m}_0, \sigma_m^2\mathbf{I}_2\right) + \\
&(-I-\alpha_s)\log\left(\frac{1}{2}\sum_{i=1}^{I}\mathbf{s}_i^T\mathbf{s}_i + \beta_s\right) + (-J-\alpha_w)\log\left(\frac{1}{2}\sum_{j=1}^{J}\mathbf{w}_j^T\mathbf{w}_j + \beta_w\right) + \\
&\sum_{ijk}\sum_{t=1}^{n_{ijk}} \log PN_2\left(\theta_{ijkt}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) + C,
\end{aligned}
$$

where we have integrated out the variance components σ_s^2, σ_w^2. The second derivatives needed to compute the Hessian are given in Supplement 3.11. From Figure 7 we can see again that the posterior distributions are very well approximated by the Laplace approximation. We computed the DIC values for the same models in Table 2 using the Laplace approximation

Algorithm 2 Iterative scheme for obtaining the parameters in the optimal densities for the linear model case.

Initialize $\mathrm{E}(\boldsymbol{\mu}), \mathrm{E}(\mathbf{m}_k), \mathrm{E}(\mathbf{s}_i), \mathrm{E}(\mathbf{w}_j)$

Cycle:

$$
\begin{aligned}
b_{ijkt} &\leftarrow \mathbf{u}_{ijkt}^T \boldsymbol{\mu}_{ijk} \\
\mathrm{E}(r_{ijkt}) &\leftarrow b_{ijkt} + \frac{\sqrt{2\pi}}{c(b_{ijkt})} \Phi(b_{ijkt}) \exp\left\{\frac{1}{2} b_{ijkt}^2\right\} \\
\mathrm{E}\left(1/\sigma_s^2\right) &\leftarrow (\alpha_s + I) \Big/ \left(\mathrm{E}_s \sum_{i=1}^{I} \frac{1}{2} \mathbf{s}_i^T \mathbf{s}_i\right) \\
\mathrm{E}\left(1/\sigma_w^2\right) &\leftarrow (\alpha_w + J) \Big/ \left(\mathrm{E}_w \sum_{j=1}^{J} \frac{1}{2} \mathbf{w}_j^T \mathbf{w}_j\right) \\
\mathrm{E}(\bar{\mathbf{z}}) &\leftarrow \left(\frac{1}{n}\right) \sum_{ijkt} u_{ijkt} \mathrm{E}(r_{ijkt}) - \mathrm{E}(\mathbf{m}_k) - \mathrm{E}(\mathbf{s}_i) - \mathrm{E}(\mathbf{w}_j) \\
\mathrm{E}(\boldsymbol{\mu}) &\leftarrow \frac{\boldsymbol{\mu}_0/\sigma_0^2 + n\mathrm{E}(\bar{\mathbf{z}})}{n + (1/\sigma_0^2)} \\
\mathrm{E}\left(\mathbf{z}_k^{(m)}\right) &\leftarrow \sum_{ij} \sum_{l=1}^{n_{ijk}} u_{ijkl} \mathrm{E}(r_{ijkl}) - \mathrm{E}(\boldsymbol{\mu}) - \mathrm{E}(\mathbf{s}_i) - \mathrm{E}(\mathbf{w}_j) \\
\mathrm{E}(\mathbf{m}_k) &\leftarrow \frac{\boldsymbol{m}_0/\sigma_m^2 + \mathrm{E}\left(\mathbf{z}_k^{(m)}\right)}{n_k^{(m)} + (1/\sigma_m^2)} \\
\mathrm{E}\left(\mathbf{z}_i^{(s)}\right) &\leftarrow \sum_{kj} \sum_{t=1}^{n_{ijk}} u_{ijkt} \mathrm{E}(r_{ijkt}) - \mathrm{E}(\boldsymbol{\mu}) - \mathrm{E}(\mathbf{m}_k) - \mathrm{E}(\mathbf{w}_j) \\
\mathrm{E}(\mathbf{s}_i) &\leftarrow \frac{\mathrm{E}\left(\mathbf{z}_i^{(s)}\right)}{n_i^{(s)} + \mathrm{E}(1/\sigma_s^2)} \\
\mathrm{E}\left(\mathbf{z}_j^{(w)}\right) &\leftarrow \sum_{ik} \sum_{t=1}^{n_{ijk}} u_{ijkt} \mathrm{E}(r_{ijkt}) - \mathrm{E}(\boldsymbol{\mu}) - \mathrm{E}(\mathbf{m}_k) - \mathrm{E}(\mathbf{s}_i) \\
\mathrm{E}(\mathbf{w}_j) &\leftarrow \frac{\mathrm{E}\left(\mathbf{z}_j^{(w)}\right)}{n_j^{(w)} + \mathrm{E}(1/\sigma_w^2)}
\end{aligned}
$$

and obtained almost the same values.

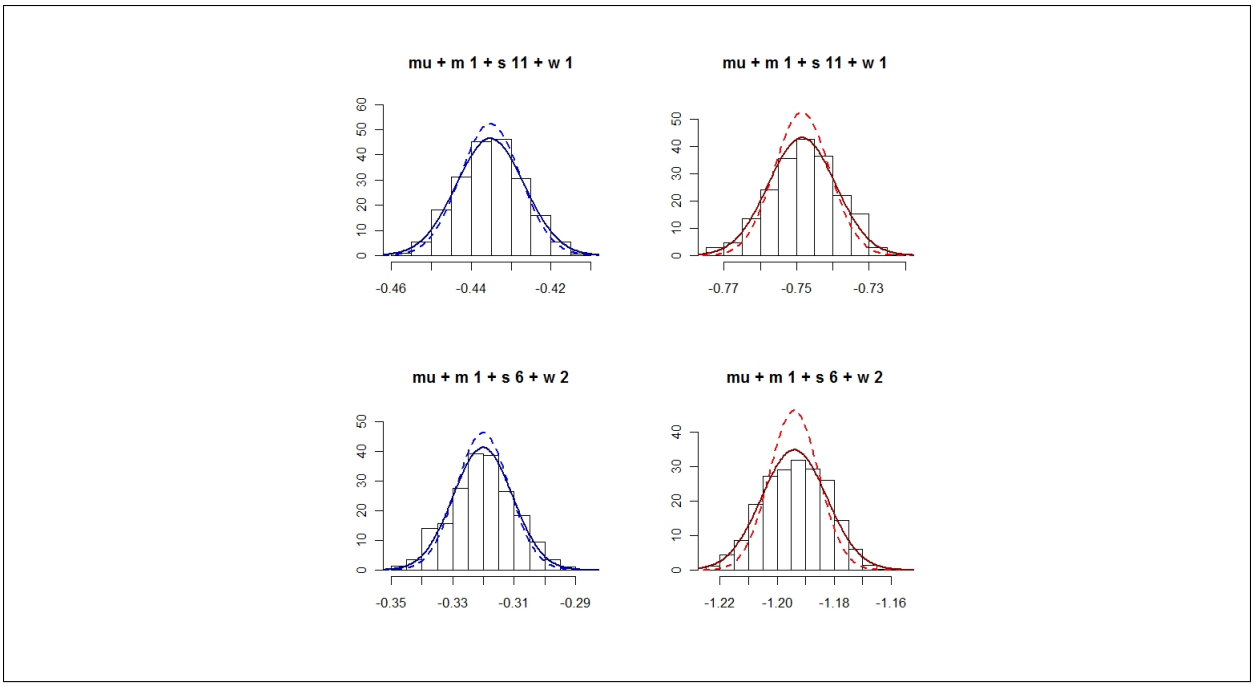

Figure 7: Histograms of the output from the Gibbs sampler with variational and Laplace approximations using the CHTS data for a model with mode as a fixed effect and state and wave as random effects. The dashed curves are the ones obtained by the variational approximation and the solid ones are obtained by the Laplace approximation. The first and second rows correspond to different combinations of the effects while the first and second columns correspond to the first and second components of $\boldsymbol{\mu}$, respectively.

3.6 Prediction of fractions of departures

As noted in Section 3.1, the goal of the paper is to estimate the population fraction of daily departures at a fishing site between the arrival and departure times of the interviewer, for a stratum U_{ijk} determined by mode, state and wave. The population fraction between two times, τ_1 and τ_2, is defined as

$$F_{N,ijk}(\tau_1, \tau_2) = \frac{\sum_{U_{ijk}} I_{[\tau_1,\tau_2)}(T_{ijkt})}{N_{ijk}}, \tag{18}$$

with $I_{[\tau_1,\tau_2)}(T) = 1$ when $\tau_1 \leq T < \tau_2$ and 0 otherwise. From the survey data, a direct estimator of this quantity is given by

$$\hat{F}^D_{ijk}(\tau_1, \tau_2) = \frac{\sum_{s_{ijk}} w_{ijkt} I_{[\tau_1,\tau_2)}(T_{ijkt})}{\sum_{s_{ijk}} w_{ijkt}}, \tag{19}$$

where $w_{ijkt} = 1/\pi_{ijkt}$ is the inverse of the inclusion probability for the tth element in stratum U_{ijk}. This estimator is asymptotically unbiased under repeated sampling from the target population, but is likely to be very variable in many strata because the sample size n_{ijk} is very small (or zero). For simplicity, we consider the case of simple random sampling with replacement (or without replacement with a vanishingly small sampling fraction). In that case, the variance of $\hat{F}^D_{ijk}(\tau_1, \tau_2)$ is given by

$$\mathrm{Var}(\hat{F}^D_{ijk}(\tau_1, \tau_2)) = \frac{1}{n_{ijk}} F_{N,ijk}(\tau_1, \tau_2)(1 - F_{N,ijk}(\tau_1, \tau_2)). \tag{20}$$

Under the assumption that the departure times follow a projected normal distribution, the population fraction $F_{N,ijk}(\tau_1, \tau_2)$ is expected to be very close to the probability $\Pr(\tau_1 \leq T < \tau_2 | \boldsymbol{\mu}_{ijk})$ under the projected normal distribution as long as N_{ijk} is sufficiently large, with $\boldsymbol{\mu}_{ijk}$ the "true" parameter value for the stratum. This is a non-random function of $\boldsymbol{\mu}_{ijk}$, so that a procedure that provides an estimate for $\boldsymbol{\mu}_{ijk}$ can be used to estimate $\Pr(\tau_1 \leq T < \tau_2 | \boldsymbol{\mu}_{ijk})$ and hence $F_{N,ijk}(\tau_1, \tau_2)$ as well. Hence, under the hierarchical Bayesian model described in the previous sections, we can obtain the posterior distribution of

$$F^M_{ijk}(\tau_1, \tau_2) = \Pr(\tau_1 \leq T < \tau_2 | \boldsymbol{\mu}_{ijk}) \tag{21}$$

given the sample data. However, obtaining this posterior distribution for each combination of state, wave, mode and time interval is far from trivial given the size of the dataset, because the integration of the projected normal density over the interval (τ_1, τ_2) needs to be performed at each draw from the posterior distribution $p(\boldsymbol{\mu}_{ijk}|\text{data})$. These draws are

obtained by either using iterations of Gibbs sampler after convergence or by drawing from the Variational/Laplace approximation distribution.

We therefore streamlined the computations by only computing the fractions for 24 one-hour intervals and by taking advantage of a number of results for the projected normal distribution. To simplify notation in what follows, let $F_{ijk}^{M}(\tau), \tau = 1, \ldots, 24$ denote the 1-hour fractions $F_{ijk}^{M}(\tau - 1, \tau)$. For each iteration b of the Gibbs sampler, we obtain a vector $\boldsymbol{\mu}_{ijk}^{b}$ for each one of the 368 combinations of indices ijk from which we compute the 24 fractions. Without simplifications, this implies that we would need to compute $368 \times 24 \times B$ integrals. To make this process more efficient, we first obtained the following two identities, derived in Supplement 3.12:

$$\begin{aligned}
\int_{\theta_1}^{\theta_2} f\left(\theta \mid \boldsymbol{\mu}_{ijk}\right) d\theta &= \Phi\left(-\rho_{ijk} \sin\left(\theta_1 - \omega_{ijk}\right)\right) \text{ if } \theta_2 - \theta_1 = \pi \\
\int_{\theta_1}^{\theta_2} f\left(\theta \mid \boldsymbol{\mu}_{ijk}\right) d\theta &= \Phi\left(-\rho_{ijk} \sin\left(\theta_1 - \omega_{ijk}\right)\right) \Phi\left(\rho_{ijk} \cos\left(\theta_1 - \omega_{ijk}\right)\right) \text{ if } \theta_2 - \theta_1 = \frac{\pi}{2},
\end{aligned}$$

where $f\left(\theta \mid \boldsymbol{\mu}_{ijk}\right)$ denotes the density of a $PN_2\left(\boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right)$ random variable normalized to the unit circle, and $\boldsymbol{\mu}_{ijk}^{T} = \rho_{ijk}\left(\cos\omega_{ijk}, \sin\omega_{ijk}\right)$. Then, at each iteration b the 24 fractions $F_{ijk}^{M,b}(\tau)$ are obtained by computing $F_{ijk}^{M,b}(1), \ldots, F_{ijk}^{M,b}(5)$ by numerical integration, followed by successive differencing using the second identity, i.e. we set $F_{ijk}^{M,b}(6) = \Phi\left(-\rho_{ijk}^{b} \sin\left(-\omega_{ijk}^{b}\right)\right)$ $\Phi\left(\rho_{ijk}^{b} \cos\left(-\omega_{ijk}^{b}\right)\right) - \sum_{t=1}^{5} F_{ijk}^{M,b}(t)$, $F_{ijk}^{M,b}(7) = \Phi\left(-\rho_{ijk}^{b} \sin\left(\frac{\pi}{12} - \omega_{ijk}^{b}\right)\right) \Phi\left(\rho_{ijk}^{b} \cos\left(\frac{\pi}{12} - \omega_{ijk}^{b}\right)\right) -$ $\sum_{\tau=2}^{6} F_{ijk}^{M,b}(\tau)$ and so on. In this manner, for each b and each ijk, we only need to compute 5 numerical integrals and 19 integrals using the formula above instead of 24 numerical integrals.

At the conclusion of this procedure, we obtain the posterior distribution of $F_{ijk}^{M}(\tau)$ for each one-hour fraction in each wave, state and mode combination. Figure 8 shows boxplots corresponding to the posterior distributions of the model-estimated fractions of departures for four state-wave-mode combinations in each 1-hour period, as well as histograms of the original data in those strata, which correspond to the estimator (19). The very narrow

boxplots reflect the fact that these estimated departure fractions are based on a very large sample size. It is also clear from these plots that the modeled distributions deviate substantially from the observed distributions, even when there appear to be many observations in a stratum.

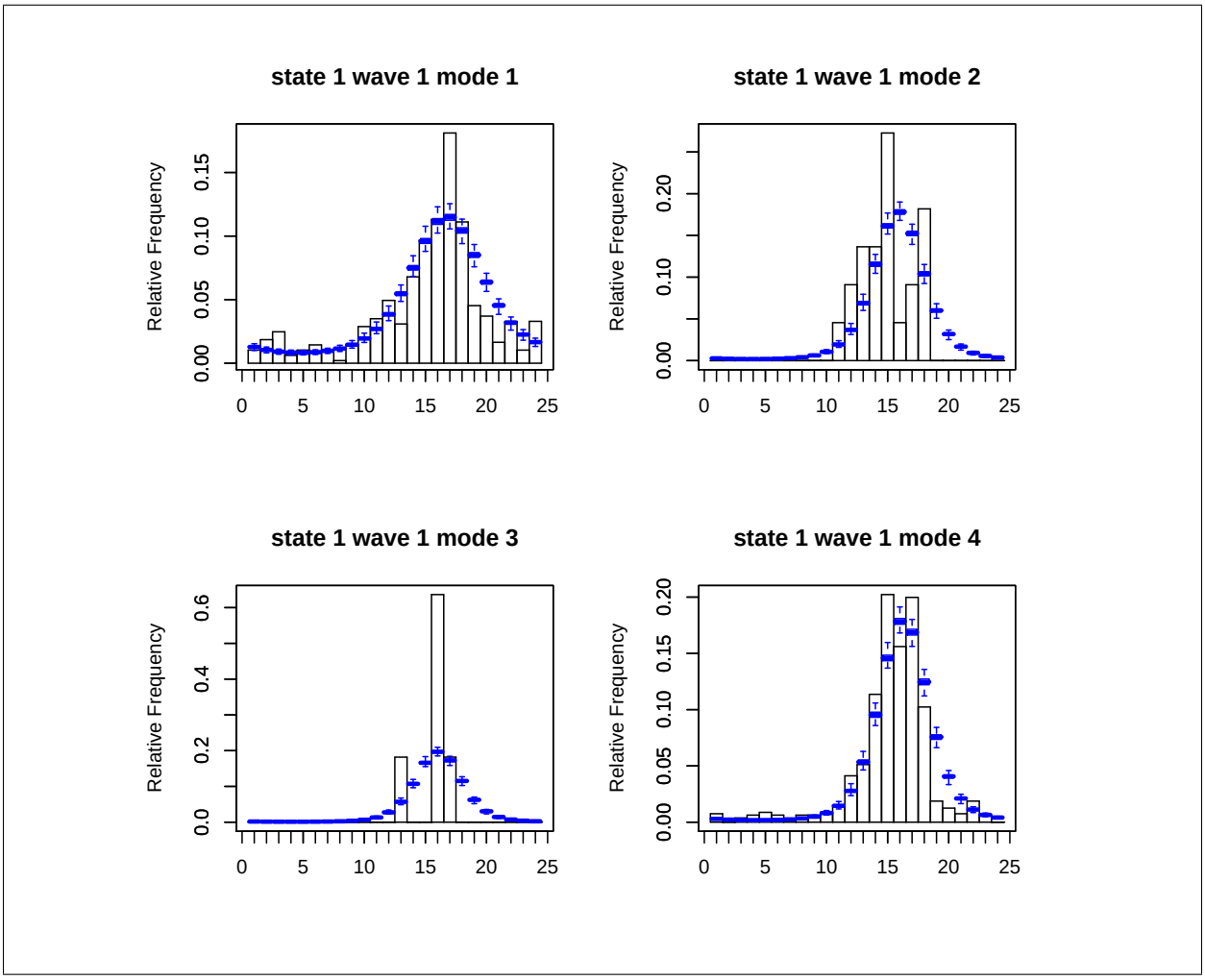

Figure 8: Histograms of departure time data and boxplots of the posterior distributions of the 1-hour fractions of departures. Four different combinations of state, wave, and mode are shown.

At this point, we have two possible estimators for the population fractions of departures in (18). The design-based estimator in (19) is unbiased but very variable in any stratum with small sample size, while the model-based estimator in (21) is very precise (low variance) but, being model-based, potentially biased if the model specification is incorrect. We now

consider the model above as a component of a small area estimation procedure. The goal of small area estimation is to combine a survey estimator that only uses data from a given "small area" (defined here as a mode-state-wave combination) with a model-based estimator that is based on data from the whole sample. In this case, the survey estimator is the direct estimator (19), and the model-based estimator is the posterior mean of (21), which we will denote by $\hat{F}^M_{ijk}(\tau)$. A large literature on small area estimation is available, with a range of different parametric and, more recently, nonparametric model specifications (see Rao (2003) for an overview). However, to the best of our knowledge, no small area estimation models for circular data exist. We therefore decided to apply *composite estimation* (Ghosh and Rao, 1994), which consists of taking a convex combination of both estimators. Composite estimation is generally applicable even in non-standard situations and is easy to implement. It also provides a simple way to trade off the bias and variance of the two estimators by adjusting the linear combination weights.

The composite estimator for the fraction of departures that occur in the interval $[\tau - 1, \tau), \tau = 1, \ldots, 24$ for state i, wave j and mode k is defined as

$$\hat{F}^C_{ijk}(\tau) = w_{ijk}\hat{F}^D_{ijk}(\tau) + (1 - w_{ijk})\,\hat{F}^M_{ijk}(\tau),$$

where $w_{ijk} \in [0, 1]$ is a weight further specified below. Note that, while the weight w_{ijk} can in principle depend on τ as well, we will use a single weight for all τ in a small area so that the composite estimator remains a valid fraction in the sense that $\sum_{\tau=1}^{24} \hat{F}^C_{ijk}(\tau) = 1$. The optimal weight w^{opt}_{ijk} in this context minimizes $\text{MSE}(\hat{F}^C_{ijk}(\tau))$, averaged over τ. Assuming that the direct estimator $\hat{F}^D_{ijk}(\tau)$ is unbiased and that the covariance $\text{Cov}\left(\hat{F}^D_{ijk}(\tau), \hat{F}^M_{ijk}(\tau)\right) = 0$,

$$w^{\text{opt}}_{ijk} = \frac{\text{MSE}\left(\hat{F}^M_{ijk}\right)}{\text{MSE}\left(\hat{F}^M_{ijk}\right) + \text{Var}\left(\hat{F}^D_{ijk}\right)}, \tag{22}$$

where we denote averaging over τ by removing it from the expressions. The assumption

that both estimators are uncorrelated is approximately met when the variance of the model-based estimator is negligible relative to that of the direct estimator. Given the very large sample size in this application, this is certainly reasonable here (see also Figure 8). Speaking somewhat loosely, the MSE of the direct estimator is equal to its variance, while that of the model-based estimator is equal to its bias due to model misspecification. The optimal weight w_{ijk}^{opt} therefore trades off the variance of the former and the bias of the latter. This optimal weight is unknown but will be estimated under a number of simplifying assumptions.

First, we assume that the magnitude of the bias of the model-based estimator averaged over time is approximately constant across small areas. Second, the variance of the direct estimator is assumed to be of the form C/n_{ijk}, which is reasonable given expression (20) above. Therefore, for any small area ijk, we estimate MSE $\left(\hat{F}_{ijk}^{M}\right)$ by

$$\widehat{MSE}(F^M) = \frac{1}{R}\sum_i\sum_j\sum_k\sum_\tau\left(\left(\hat{F}_{ijk}^{D}(\tau) - \hat{F}_{ijk}^{M}(\tau)\right)^2 - \frac{\hat{F}_{ijk}^{D}(\tau)\left(1-\hat{F}_{ijk}^{D}(\tau)\right)}{n_{ijk}}\right)$$

where R denotes the total number of cells over which this is computed, and the latter term inside the sum is the direct estimator of the variance of $\hat{F}_{ijk}^{D}(\tau)$.

The term Var $\left(\hat{F}_{ijk}^{D}\right)$ in (22) is replaced by the simplified "estimator" $\widehat{V}\left(\hat{F}_{ijk}^{D}\right) = 0.25/n_{ijk}$, which is the largest possible value for the variance of a proportion. We are using 0.25 in the numerator instead of the average of the $\hat{F}_{ijk}^{D}(\tau)\left(1-\hat{F}_{ijk}^{D}(\tau)\right)$ over τ, because especially in small areas with small sample sizes, the $\hat{F}_{ijk}^{D}(\tau)$ were 0 for many of the time intervals, resulting in very small estimates of Var $\left(\hat{F}_{ijk}^{D}\right)$ and hence skewing the composite estimator towards the direct estimator despite the small sample size. Using these estimators, the final weight for small area ijk is given by

$$w_{ijk} = \frac{\widehat{MSE}(F^M)}{\widehat{MSE}(F^M) + 0.25/n_{ijk}}.$$

In order to illustrate the effect of the above weighting procedure, Figure 9 shows the

direct, model-based and composite estimators for two small areas with different sample sizes. Clearly, when n_{ijk} is small, the weighting procedure will give a relatively small weight to the direct estimator and base the composite estimator primarily on the model-based estimator. In contrast, when the sample size is large, the composite estimator is very close to the direct estimator.

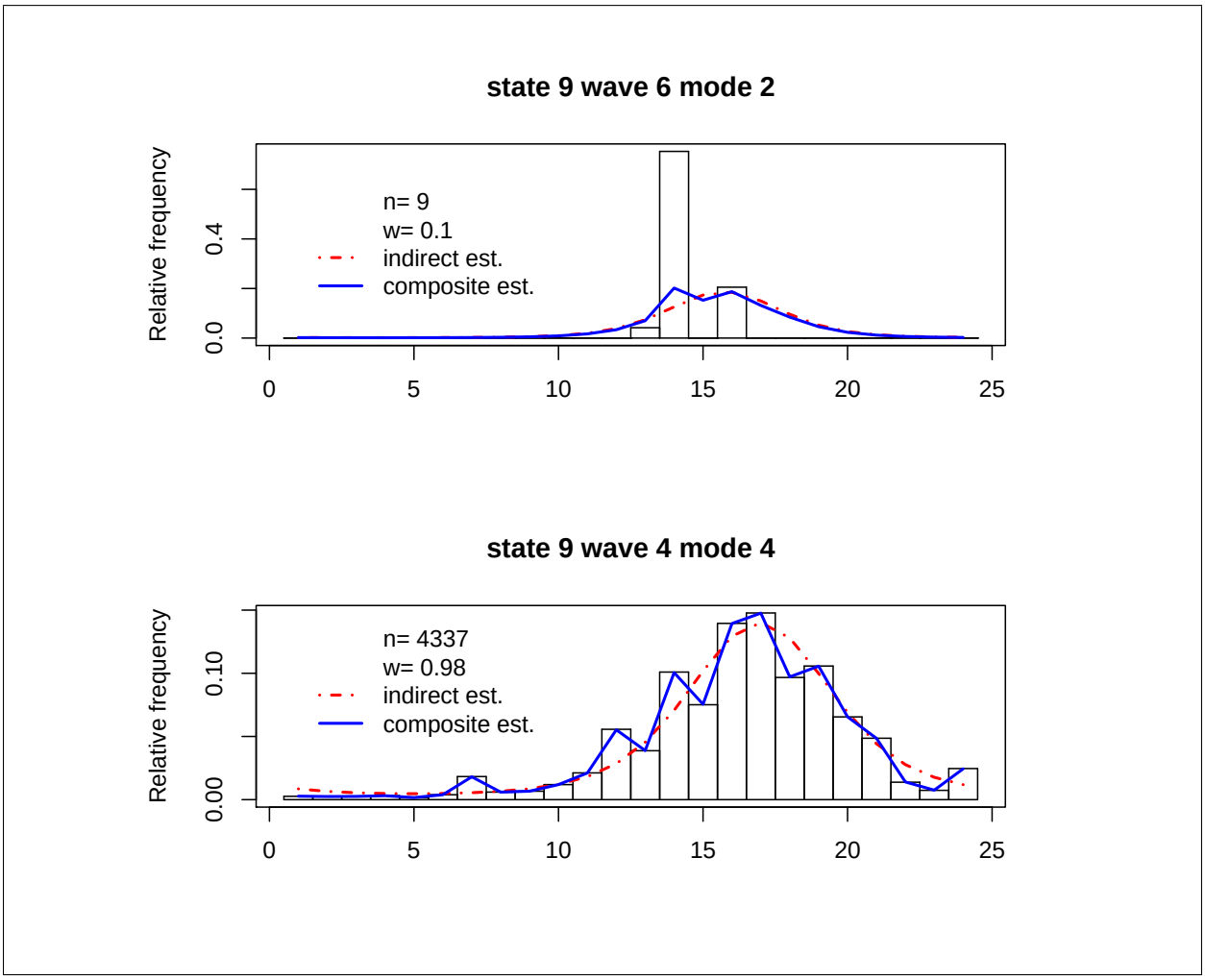

Figure 9: Examples of direct, model-based and composite estimators for the fractions of departures for two strata combinations. Top plot: stratum with small sample size; bottom plot: stratum with large sample size.

3.7 Conclusions

We described Bayesian inference for regression models when the response variable is circular using the projected normal distribution. By introducing another latent variable, we developed a new sampler that allowed us to draw from the latent length distribution in a simple manner speeding dramatically the Gibbs sampler compared to other Gibbs samplers that make use of rejection methods. We also presented the mean-field variational and Laplace approximations to the posterior distributions of the parameters in the regression models. The Laplace approximation explained in the paper (which makes use of the variational approximation to find the centers) worked exceptionally well, for our simulations and our data, and is dramatically faster than the Gibbs sampler. In the application to prediction of departure times, we made use of a convex combination of an indirect estimator (based on the Bayesian model) and a direct estimator (based solely on data in the small area). An obvious drawback of the current procedure is that the estimated weights to construct the composite estimators were obtained by somewhat ad hoc methods. While the resulting predictions were reasonable in terms of trading off goodness of fit and stability, a better approach would avoid the composite estimation altogether in favor of a more flexible model that can incorporate features of the observed data such as multi-modality and asymmetry. This is the topic of future work.

3.8 Supplement: proof of Result 5.

We begin by deriving the moment generating function. We define $\mathrm{M}_r(t) = \mathrm{E}(e^{tr}) = \frac{1}{C(b)} \int_0^\infty r \exp\left\{-\frac{1}{2}r^2 + r(b+t)\right\} dr$. Letting
$u = -\exp\left\{-\frac{1}{2}r^2 + r(b+t)\right\}$, we obtain $du = \left[r \exp\left\{-\frac{1}{2}r^2 + r(b+t)\right\}\right.$

$-(b+t)\exp\left\{-\frac{1}{2}r^2+r(b+t)\right\}\Big]dr$. Then,

$$\begin{aligned}
\mathrm{M}_r(t) &= \frac{1}{C(b)}\int_0^\infty\left[du+(b+t)\exp\left\{-\frac{1}{2}r^2+r(b+t)\right\}dr\right] \\
&= \frac{1}{C(b)}\left[u(r)|_0^\infty+(b+t)\exp\left\{\frac{1}{2}(b+t)^2\right\}\int_0^\infty\exp\left\{-\frac{1}{2}(r-(b+t))^2\right\}dr\right] \\
&= \frac{1}{C(b)}\left[1+\sqrt{2\pi}(b+t)\exp\left\{\frac{1}{2}(b+t)^2\right\}\Phi(b+t)\right]
\end{aligned}$$

where $C(b)=\int_0^\infty r\exp\left\{-\frac{1}{2}r^2+br\right\}dr$ is equal to the expression in the bracket above with $t=0$, i.e. $C(b)=1+\sqrt{2\pi}(b)\exp\left\{\frac{1}{2}(b)^2\right\}\Phi(b)$. To get the mean we take the derivative of $\mathrm{M}_r(t)$ with respect to t and evaluate at $t=0$.

$$\mathrm{E}(r)=\frac{d\mathrm{M}_r(t)}{dt}|_0=\frac{\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)}{C(b)}+b$$

We will now show that $0<\mathrm{Var}(r|\theta,\boldsymbol{\mu})<1$. After some algebra and making use of the moment generating function we obtain

$$\mathrm{E}\left(r^2|b\right)=b\mathrm{E}(r|b)+2$$

and

$$\begin{aligned}
\mathrm{Var}(r|b) &= \mathrm{E}\left(r^2|b\right)-(\mathrm{E}(r|b))^2 \\
&= 2-\frac{\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)}{C(b)}\mathrm{E}(r|b).
\end{aligned}$$

We will show that $\frac{\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)}{C(b)}\mathrm{E}(r|b)$ is greater than 1. Since

$$\frac{\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)}{C(b)}\mathrm{E}(r|b)-1=\frac{\frac{\left(\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\right)^2}{C(b)}-1}{C(b)},$$

we then need to show that $\frac{\left(\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\right)^2}{C(b)}>1$ or equivalently $\ln\left[\frac{\left(\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\right)^2}{C(b)}\right]>0$.

We will first prove it for the case of $b < 0$. The proof will make use of the following inequality that can be found in Abramowitz and Stegun (1964, p.298 f.7.1.13). For $x > 0$,

$$\frac{2}{x+\sqrt{x^2+4}} < \sqrt{2\pi}\exp\left\{\frac{1}{2}x^2\right\}\Phi^c(x) \leq \frac{2}{x+\sqrt{x^2+\frac{8}{\pi}}}$$

where $\Phi^c(x) = 1-\Phi(x) = \Phi(-x)$. From the above inequality we have that for $b < 0$

$$\frac{2}{|b|+\sqrt{b^2+4}} < \sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b) \leq \frac{2}{|b|+\sqrt{b^2+\frac{8}{\pi}}} \tag{23}$$

$$\frac{2b}{|b|+\sqrt{b^2+4}} > \sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\,b \geq \frac{2b}{|b|+\sqrt{b^2+\frac{8}{\pi}}} \tag{24}$$

$$1-\frac{2|b|}{|b|+\sqrt{b^2+4}} > 1+\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\,b \geq 1-\frac{2|b|}{|b|+\sqrt{b^2+\frac{8}{\pi}}} \tag{25}$$

$$\begin{aligned}
&\ln\left[\frac{\left(\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\right)^2}{C(b)}\right] \\
&= 2\ln\left(\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\right) - \ln\left(1+\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\,b\right) \\
&> 2\ln\left(\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(b)\right) - \ln\left(1-\frac{2|b|}{|b|+\sqrt{b^2+4}}\right) \\
&> 2\ln\left(\frac{2}{|b|+\sqrt{b^2+4}}\right) - \ln\left(1-\frac{2|b|}{|b|+\sqrt{b^2+4}}\right) \\
&= \ln 4 - \ln\left(|b|+\sqrt{b^2+4}\right) - \ln\left(\sqrt{b^2+4}-|b|\right) \\
&= \ln 4 - \ln\left[\left(\sqrt{b^2+4}-|b|\right)\left(\sqrt{b^2+4}+|b|\right)\right] \\
&= \ln 4 - \ln 4 = 0.
\end{aligned}$$

The inequalities on lines 2 and 3 are obtained by making use of (25) and (23) respectively.

Now for the case $b > 0$. We define $f(b) = \sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}$ just for convenience, and

$$\begin{aligned}
\frac{f(b)\,\Phi(b)}{C(b)}\mathrm{E}(r|b) - 1 &= \frac{\frac{(f(b)\Phi(b))^2}{C(b)} - 1}{C(b)} \\
&= \frac{(f(b)\,\Phi(b))^2 - 1 - f(b)\,\Phi(b)\,b}{(C(b))^2} \\
&= \frac{(f(b))^2(1-\Phi(-b))^2 - 1 - f(b)\,b\,(1-\Phi(-b))}{(C(b))^2} \\
&= \frac{(f(b))^2(1-2\Phi(-b)) - f(b)\,b + \left[(f(b)\,\Phi(-b))^2 - C(-b)\right]}{(C(b))^2} \\
&> \frac{(f(b))^2(1-2\Phi(-b)) - f(b)\,b}{(C(b))^2} \\
&= \frac{(f(b))\left[f(b)(1-2\Phi(-b)) - b\right]}{(C(b))^2}.
\end{aligned}$$

The inequality comes from the fact that $\left(\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}\Phi(-b)\right)^2 - C(-b) > 0$, as proved for the case $b < 0$. Finally, we show that $[f(b)(1-2\Phi(-b)) - b] \geq 0$ by showing that it is monotone with value equal to zero for when $b = 0$:

$$\begin{aligned}
\frac{d}{db}\left[\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}(1-2\Phi(-b)) - b\right] &= 1 + (1-2\Phi(-b))\sqrt{2\pi}\exp\left\{\frac{1}{2}b^2\right\}b \\
&> 0,
\end{aligned}$$

which completes the proof.

3.9 Supplement: proof of Result 6.

First we note that Algorithm 1 converges to a solution of the following fixed point equation

$$\boldsymbol{\mu}^\star = \frac{\boldsymbol{\mu}_0/\sigma_0^2 + \sum_{i=1}^n \mathbf{u}_i E(r_i|\boldsymbol{\mu}^\star, data)}{n + 1/\sigma_0^2},$$

where $E\left(r_i|\boldsymbol{\mu}^\star, data\right) = b_i + \frac{\sqrt{2\pi}}{c(b_i)}\Phi\left(b_i\right)\exp\left\{\frac{1}{2}b_i^2\right\}$ with $b_i = \mathbf{u}_i^T\boldsymbol{\mu}^\star$. Let μ_c^* be the cth component of $\boldsymbol{\mu}^*$ and similarly define $\mu_{0,c}$ and $u_{i,c}$.

$$\begin{aligned}
\mu_c^* &= \frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}\mathrm{E}\left(r_i|\boldsymbol{\mu}^*, data\right)}{n + 1/\sigma_0^2} \\
&= \int\left(\frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}r_i}{n + 1/\sigma_0^2}\right)p\left(\boldsymbol{r}|\boldsymbol{\mu}^*, data\right)d\boldsymbol{r} \\
&= \int\left(\frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}r_i}{n + 1/\sigma_0^2}\right)\frac{p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)}{p\left(\boldsymbol{\mu}^*|data\right)}d\boldsymbol{r}.
\end{aligned}$$

Multiplying by $p\left(\boldsymbol{\mu}^*|data\right)$ on both sides,

$$\mu_c^* p\left(\boldsymbol{\mu}^*|data\right) = \int\left(\frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}r_i}{n + 1/\sigma_0^2}\right)p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r}$$

$$\mu_c^*\int p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r} = \int\left(\frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}r_i}{n + 1/\sigma_0^2}\right)p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r}.$$

Subtracting $\mu_c^*\int p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r}$ on both sides,

$$0 = -\int\left(\mu_c^* - \frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}r_i}{n + 1/\sigma_0^2}\right)p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r}$$

and finally multiplying by $(n + 1/\sigma_0^2)$ on both sides, we obtain

$$\begin{aligned}
0 &= \int -\left(n + 1/\sigma_0^2\right)\left(\mu_c^* - \frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}r_i}{n + 1/\sigma_0^2}\right)p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r} \\
&= \int\frac{\partial}{\partial\mu_c}p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r} \\
&= \frac{\partial}{\partial\mu_c}\int p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)p\left(\boldsymbol{r}|data\right)d\boldsymbol{r} = \frac{\partial}{\partial\mu_c}p\left(\boldsymbol{\mu}^*|data\right).
\end{aligned}$$

The penultimate line results from the fact that $p\left(\boldsymbol{\mu}^*|\boldsymbol{r}, data\right)$ is a normal density. We just showed that Algorithm 1 converges to a critical point of $p\left(\boldsymbol{\mu}|data\right)$. Following the same steps but in the opposite direction we can show that every critical point of $p\left(\boldsymbol{\mu}|data\right)$ is a solution to the fixed point equation.

Next we will show that there is only one critical point of $p(\boldsymbol{\mu}|data)$, the mode. We will do this by showing that every critical point of $p(\boldsymbol{\mu}|data)$ is a local maximum. To do this we need to compute the Hessian matrix. For convenience we will define $f(\mu_c, \boldsymbol{r}) = \mu_c - \frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c} r_i}{n+1/\sigma_0^2}$. We will also notice that $f(\mu_c^*, \boldsymbol{r}) = \mu_c^* - \frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c} r_i}{n+1/\sigma_0^2} = \frac{\left[\sum_{i=1}^n u_{i,c}(\mathrm{E}(r_i|\boldsymbol{\mu}^*, data) - r_i)\right]^2}{n+1/\sigma_0^2}$. This last equality is because $\mu_c^* = \frac{\mu_{0,c}/\sigma_0^2 + \sum_{i=1}^n u_{i,c}\mathrm{E}(r_i|\boldsymbol{\mu}^*, data)}{n+1/\sigma_0^2}$. We compute

$$
\begin{aligned}
&\frac{\partial^2}{\partial \mu_c^2} p(\boldsymbol{\mu}^*|data) \\
&= \frac{\partial^2}{\partial \mu_c^2} p(\boldsymbol{\mu}|data)\,|_{\boldsymbol{\mu}^*} \\
&= \frac{\partial}{\partial \mu_c} \int -\left(n + 1/\sigma_0^2\right) f(\mu_c, \boldsymbol{r})\, p(\boldsymbol{\mu}|\boldsymbol{r}, data)\, p(\boldsymbol{r}|data)\, d\boldsymbol{r}|_{\boldsymbol{\mu}^*} \\
&= \left(n + 1/\sigma_0^2\right)^2 \int \left[f(\mu_c^*, \boldsymbol{r})\right]^2 p(\boldsymbol{\mu}^*|\boldsymbol{r}, data)\, p(\boldsymbol{r}|data)\, d\boldsymbol{r} - \left(n + 1/\sigma_0^2\right) p(\boldsymbol{\mu^*}|data) \\
&= \left(n + 1/\sigma_0^2\right) p(\boldsymbol{\mu^*}|data) \left[\left(n + 1/\sigma_0^2\right) \int \left[f(\mu_c^*, \boldsymbol{r})\right]^2 \frac{p(\boldsymbol{\mu}^*|\boldsymbol{r}, data)\, p(\boldsymbol{r}|data)}{p(\boldsymbol{\mu^*}|data)} d\boldsymbol{r} - 1\right] \\
&= \left(n + 1/\sigma_0^2\right) p(\boldsymbol{\mu^*}|data) \left[\left(n + 1/\sigma_0^2\right)^{-1} \int \left[f(\mu_c^*, \boldsymbol{r})\right]^2 p(\boldsymbol{r}|\boldsymbol{\mu}^*, data)\, d\boldsymbol{r} - 1\right],
\end{aligned}
$$

where the integral inside the bracket is

$$
\begin{aligned}
&\int \left[f(\mu_c^*, \boldsymbol{r})\right]^2 p(\boldsymbol{r}|\boldsymbol{\mu}^*, data)\, d\boldsymbol{r} \\
&= \int \sum_{i,j} u_{i,c} u_{j,c} \left[r_i - \mathrm{E}(r_i|\boldsymbol{\mu}^*, data)\right] \left[r_j - \mathrm{E}(r_j|\boldsymbol{\mu}^*, data)\right] p(\boldsymbol{r}|\boldsymbol{\mu}^*, data)\, d\boldsymbol{r} \\
&= \int \sum_{i=1}^n u_{i,c}^2 \left[r_i - \mathrm{E}(r_i|\boldsymbol{\mu}^*, data)\right] p(\boldsymbol{r}|\boldsymbol{\mu}^*, data)\, d\boldsymbol{r} \\
&= \sum_{i=1}^n u_{i,c}^2 \mathrm{Var}(r_i|\boldsymbol{\mu}^*, data) \\
&< \sum_{i=1}^n \mathrm{Var}(r_i|\boldsymbol{\mu}^*, data) \\
&< n.
\end{aligned}
$$

The inequality is due to Result 5. Thus,

$$\begin{aligned}\frac{\partial^2}{\partial\mu_c^2}p\left(\boldsymbol{\mu}^*|data\right) &< \left(n+1/\sigma_0^2\right)p\left(\boldsymbol{\mu}^*|data\right)\left[\frac{n}{n+1/\sigma_0^2}-1\right]\\ &< 0.\end{aligned}$$

Now we will compute $\frac{\partial^2}{\partial\mu_1\partial\mu_2}p\left(\boldsymbol{\mu}^*|data\right)$. Following similar steps as above:

$$\begin{aligned}\frac{\partial^2}{\partial\mu_1\partial\mu_2}p\left(\boldsymbol{\mu}^*|data\right) &= p\left(\boldsymbol{\mu}^*|data\right)\int\sum_{i=1}^{n}u_{i,1}u_{i,2}\left[\mathrm{E}\left(r_i|\boldsymbol{\mu}^*,data\right)-r_i\right]^2 p\left(\boldsymbol{r}|\boldsymbol{\mu}^*,data\right)d\boldsymbol{r}\\ &= p\left(\boldsymbol{\mu}^*|data\right)\sum_{i=1}^{n}u_{i,1}u_{i,2}\mathrm{Var}\left(r_i|\boldsymbol{\mu}^*,data\right).\end{aligned}$$

To simplify the notation we will write $\mathrm{Var}\left(r_i|\boldsymbol{\mu}^*,data\right)$ simply as $\mathrm{Var}\left(r_i\right)$. Finally, the determinant of the Hessian matrix evaluated at $\boldsymbol{\mu}^*$ is

$$\begin{aligned}\left[p\left(\boldsymbol{\mu}^*|data\right)\right]^2\Bigg[&\left(\sum_{i=1}^{n}u_{i,1}^2\mathrm{Var}\left(r_i\right)-\left(n+\frac{1}{\sigma_0^2}\right)\right)\left(\sum_{i=1}^{n}u_{i,2}^2\mathrm{Var}\left(r_i\right)-\left(n+\frac{1}{\sigma_0^2}\right)\right)\\ &-\left(\sum_{i=1}^{n}u_{i,1}u_{i,2}\mathrm{Var}\left(r_i\right)\right)^2\Bigg].\end{aligned}$$

The second bracket is equal to:

$$\begin{aligned}&\Bigg[\left(\sum_{i=1}^{n}u_{i,1}^2\mathrm{Var}\left(r_i\right)\right)\left(\sum_{i=1}^{n}u_{i,1}^2\mathrm{Var}\left(r_i\right)\right)-\left(\sum_{i=1}^{n}u_{i,1}u_{i,2}\mathrm{Var}\left(r_i\right)\right)^2\\ &+\left(n+\frac{1}{\sigma_0^2}\right)\left(\left(n+\frac{1}{\sigma_0^2}\right)-\sum_{i=1}^{n}\mathrm{Var}\left(r_i\right)\right)\Bigg],\end{aligned}$$

and by making use of Result 5 this last quantity is greater than:

$$\left(\sum_{i=1}^{n}u_{i,1}^2\mathrm{Var}\left(r_i\right)\right)\left(\sum_{i=1}^{n}u_{i,1}^2\mathrm{Var}\left(r_i\right)\right)-\left(\sum_{i=1}^{n}u_{i,1}u_{i,2}\mathrm{Var}\left(r_i\right)\right)^2,$$

which is greater than zero due to the Cauchy-Schwartz inequality. To see this define

$x_i = u_{i,1}\sqrt{\text{Var}\,(r_i)}$ and $y_i = u_{i,2}\sqrt{\text{Var}\,(r_i)}$. Then, the Cauchy-Schwartz inequality says $\left(\sum_{i=1}^n x_i y_i\right)^2 \leq \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right)$. We have shown that the determinant of the Hessian evaluated at the critical point is positive and also shown that the second partial derivative of $p\left(\boldsymbol{\mu}|data\right)$ with respect to μ_c evaluated at the critical point $\boldsymbol{\mu}^*$ is negative. These two facts together show that every critical point of $p\left(\boldsymbol{\mu}|data\right)$ is a maximum. Thus, there is only one critical point, the mode of $p\left(\boldsymbol{\mu}|data\right)$.

3.10 Supplement: derivatives to find the Hessian, random sample case.

$$\begin{aligned}
\frac{\partial^2}{\partial \mu_c^2}\log N_2\left(\boldsymbol{\mu}_0, \sigma_0^2\mathbf{I}_2\right) &= -\frac{1}{\sigma_0^2} \\
\frac{\partial^2}{\partial \mu_1 \partial \mu_2}\log N_2\left(\boldsymbol{\mu}_0, \sigma_0^2\mathbf{I}_2\right) &= 0 \\
\frac{\partial^2}{\partial \mu_c^2}\log PN_2\left(\theta_i; \boldsymbol{\mu}, \mathbf{I}_2\right) &= -1 + u_{i,c}^2 B_i \\
\frac{\partial^2}{\partial \mu_1 \partial \mu_2}\log PN_2\left(\theta_i; \boldsymbol{\mu}, \mathbf{I}_2\right) &= u_{i,1}u_{i,2}B_i,
\end{aligned}$$

where $B_i = 2 - \frac{\Phi(b_i)}{\varphi(b_i)}\left[1 + \frac{b_i\Phi(b_i)}{\varphi(b_i)}\right]^{-1}\left[\frac{\Phi(b_i)}{\varphi(b_i)}\left[1 + \frac{b_i\Phi(b_i)}{\varphi(b_i)}\right]^{-1} + b_i\right]$, $b_i = \mathbf{u}_i^T\boldsymbol{\mu}$, $\mathbf{u}_i^T = (\cos\theta_i, \sin\theta_i)$, and $u_{i,c}$ is the c^{th} component of u_i.

3.11 Supplement: derivatives to find the Hessian, regression case.

$$
\begin{aligned}
\frac{\partial^2}{\partial \mu_c \partial \mu_{c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= -\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl} \\
\frac{\partial^2}{\partial m_{k,c} \partial m_{k',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{k'k} \\
\frac{\partial^2}{\partial s_{i,c} \partial s_{i',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{i'i} \\
\frac{\partial^2}{\partial w_{j,c} \partial w_{j',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{j'j} \\
\frac{\partial^2}{\partial \mu_c \partial m_{k',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{k'k} \\
\frac{\partial^2}{\partial \mu_c \partial s_{i',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{i'i} \\
\frac{\partial^2}{\partial \mu_c \partial w_{j',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{j'j} \\
\frac{\partial^2}{\partial m_{k',c} \partial s_{i',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{k'k} \delta_{i'i} \\
\frac{\partial^2}{\partial m_{k',c} \partial w_{j',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{k'k} \delta_{j'j} \\
\frac{\partial^2}{\partial s_{i',c} \partial w_{j',c'}} \log PN_2\left(\theta_{ijkl}; \boldsymbol{\mu}_{ijk}, \mathbf{I}_2\right) &= \left(-\delta_{c,c'} + u_{ijkl,c} u_{ijkl,c'} B_{ijkl}\right) \delta_{i'i} \delta_{j'j} \\
\frac{\partial^2}{\partial s_{i,c} \partial s_{i',c'}} \log\left(\frac{1}{2}\sum_{j=1}^{I} \mathbf{s}_j^T \mathbf{s}_j + \beta_s\right) &= -I' \frac{\left(\frac{1}{2}\sum_{j=1}^{I} \mathbf{s}_j^T \mathbf{s}_j + \beta_s\right) \delta_{c,c'} \delta_{i'i} - s_{i,c} s_{i',c'}}{\left(\frac{1}{2}\sum_{j=1}^{I} \mathbf{s}_j^T \mathbf{s}_j + \beta_s\right)^2} \\
\frac{\partial^2}{\partial w_{j,c} \partial w_{j',c'}} \log\left(\frac{1}{2}\sum_{i=1}^{J} \mathbf{w}_i^T \mathbf{w}_i + \beta_w\right) &= -J' \frac{\left(\frac{1}{2}\sum_{i=1}^{J} \mathbf{w}_i^T \mathbf{w}_i + \beta_w\right) \delta_{c,c'} \delta_{j'j} - w_{j,c} w_{j',c'}}{\left(\frac{1}{2}\sum_{i=1}^{J} \mathbf{w}_i^T \mathbf{w}_i + \beta_w\right)^2},
\end{aligned}
$$

where $I' = I + \alpha_s$, $J' = J + \alpha_w$ and $\delta_{i,i'}$ is the kronecker delta which is equal to 1 if i equals i' and 0 otherwise. The values B_{ijkl} are defined as $B_{ijkl} = 2 - \frac{\Phi\left(b_{ijkl}\right)}{\varphi\left(b_{ijkl}\right)}\left[1 + \frac{b_{ijkl}\Phi\left(b_{ijkl}\right)}{\varphi\left(b_{ijkl}\right)}\right]^{-1}\left[\frac{\Phi\left(b_{ijkl}\right)}{\varphi\left(b_{ijkl}\right)}\left[1 + \frac{b_{ijkl}\Phi\left(b_{ijkl}\right)}{\varphi\left(b_{ijkl}\right)}\right]^{-1} + b_{ijkl}\right]$, where $b_{ijkl} = \mathbf{u}_{ijkl}^T \boldsymbol{\mu}_{ijk}$.

3.12 Supplement: projected normal identities.

Taking μ as $\mu = \rho(\cos\omega, \sin\omega)$, the projected normal density is

$$f(\theta \mid \mu) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\rho^2\right)\left[1 + \frac{\rho\cos(\theta-\omega)\,\Phi(\rho\cos(\theta-\omega))}{\varphi(\rho\cos(\theta-\omega))}\right].$$

Define $P(\rho;\theta_1,\theta_2)$ as

$$P(\rho;\theta_1,\theta_2) = \int_{\theta_1}^{\theta_2} f(\phi \mid \mu)\, d\phi,$$

so that

$$\begin{aligned}\frac{dP(\rho;\theta_1,\theta_2)}{d\rho} &= -\rho P(\rho;\theta_1,\theta_2) + \frac{1}{2\pi}\rho\exp\left(-\frac{1}{2}\rho^2\right)\int_{\theta_1}^{\theta_2}\cos^2(\phi-\omega)\,d\phi \\ &\quad + \frac{1}{2\pi}\exp\left(-\frac{1}{2}\rho^2\right)\int_{\theta_1}^{\theta_2}\cos(\phi-\omega)\frac{\Phi(\rho\cos(\phi-\omega))}{\varphi(\rho\cos(\phi-\omega))}d\phi \\ &\quad + \frac{1}{2\pi}\rho^2\exp\left(-\frac{1}{2}\rho^2\right)\int_{\theta_1}^{\theta_2}\cos^3(\phi-\omega)\frac{\Phi(\rho\cos(\phi-\omega))}{\varphi(\rho\cos(\phi-\omega))}d\phi.\end{aligned}$$

The last term in the sum can be written as

$$\begin{aligned}I_3 &= \frac{1}{2\pi}\rho^2\exp\left(-\frac{1}{2}\rho^2\right)\int_{\theta_1}^{\theta_2}\cos(\phi-\omega)\left(1-\sin^2(\phi-\omega)\right)\frac{\Phi(\rho\cos(\phi-\omega))}{\varphi(\rho\cos(\phi-\omega))}d\phi \\ &= \frac{\rho^2}{2\pi}\exp\left(-\frac{1}{2}\rho^2\right)\int_{\theta_1}^{\theta_2}\cos(\phi-\omega)\frac{\Phi(\rho\cos(\phi-\omega))}{\varphi(\rho\cos(\phi-\omega))}d\phi \\ &\quad - \frac{\rho^2}{2\pi}\exp\left(-\frac{1}{2}\rho^2\right)\int_{\theta_1}^{\theta_2}\cos(\phi-\omega)\sin^2(\phi-\omega)\frac{\Phi(\rho\cos(\phi-\omega))}{\varphi(\rho\cos(\phi-\omega))}d\phi.\end{aligned}$$

Taking $u = \sin(\phi-\omega)\frac{\Phi(\rho\cos(\phi-\omega))}{\varphi(\rho\cos(\phi-\omega))}$, the second term in I_3 becomes

$$\begin{aligned}I_{3,2} &= \frac{\exp\left(-\frac{\rho^2}{2}\right)}{2\pi}\int_{\theta_1}^{\theta_2} du + \frac{\rho}{2\pi}\exp\left(-\frac{\rho^2}{2}\right)\int_{\theta_1}^{\theta_2}\sin^2(\phi-\omega)\,d\phi \\ &\quad - \frac{\exp\left(-\frac{\rho^2}{2}\right)}{2\pi}\int_{\theta_1}^{\theta_2}\cos(\phi-\omega)\frac{\Phi(\rho\cos(\phi-\omega))}{\varphi(\rho\cos(\phi-\omega))}d\phi.\end{aligned}$$

Substituting all the corresponding integrals into $\frac{dP(\rho;\theta_1,\theta_2)}{d\rho}$, we obtain

$$\begin{aligned}\frac{dP\left(\rho;\theta_1,\theta_2\right)}{d\rho} &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}\rho^2\right)\left[\sin\left(\theta_2-\omega\right)\frac{\Phi\left(\rho\cos\left(\theta_2-\omega\right)\right)}{\varphi\left(\rho\cos\left(\theta_2-\omega\right)\right)}-\right.\\ &\qquad\left.\sin\left(\theta_1-\omega\right)\frac{\Phi\left(\rho\cos\left(\theta_1-\omega\right)\right)}{\varphi\left(\rho\cos\left(\theta_1-\omega\right)\right)}\right].\end{aligned}$$

This is a differential equation on ρ with initial value $P\left(0;\theta_1,\theta_2\right)=\frac{\theta_2-\theta_1}{2\pi}$

Proof of first identity:

Let $\theta_2-\theta_1=\pi$. This implies that $\sin\left(\theta_2-\omega\right)=-\sin\left(\theta_1-\omega\right)$ and $\cos\left(\theta_2-\omega\right)=-\cos\left(\theta_1-\omega\right)$. Hence,

$$\frac{dP\left(\rho;\theta_1,\theta_2\right)}{d\rho}=\sin\left(\theta_2-\omega\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\rho\sin\left(\theta_2-\omega\right)\right)^2\right)$$

so that

$$\begin{aligned}P\left(\rho;\theta_1,\theta_2\right) &= \int\frac{dP\left(\rho;\theta_1,\theta_2\right)}{d\rho}d\rho+const.\\ &= \int\sin\left(\theta_2-\omega\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\rho\sin\left(\theta_2-\omega\right)\right)^2\right)d\rho+const.\end{aligned}$$

Letting $z=\rho\sin\left(\theta_2-\omega\right)$,

$$\begin{aligned}P\left(\rho;\theta_1,\theta_2\right) &= \int\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dz+const.\\ &= \Phi\left(\rho\sin\left(\theta_2-\omega\right)\right)+const.\\ &= \Phi\left(-\rho\sin\left(\theta_1-\omega\right)\right).\end{aligned}$$

The value of the constant is obtained from the initial value $P\left(0;\theta_1,\theta_2\right)=\frac{\theta_2-\theta_1}{2\pi}=\frac{1}{2}$. Since $\Phi\left(0\right)=\frac{1}{2}$, the value of the constant is zero.

Proof of second identity:

Let $\theta_2 - \theta_1 = \frac{\pi}{2}$. This implies that $\sin(\theta_2 - \omega) = \cos(\theta_1 - \omega)$ and $\cos(\theta_2 - \omega) = -\sin(\theta_1 - \omega)$. Hence,

$$\begin{aligned}\frac{dP(\rho;\theta_1,\theta_2)}{d\rho} &= \cos(\theta_1-\omega)\,\varphi(\rho\cos(\theta_1-\omega))\,\Phi(-\rho\sin(\theta_1-\omega)) \\ &\quad -\sin(\theta_1-\omega)\,\varphi(\rho\sin(\theta_1-\omega))\,\Phi(\rho\cos(\theta_1-\omega)).\end{aligned}$$

Integrating the first term by parts, taking $u = \Phi(-\rho\sin(\theta_1 - \omega))$ and $dv = \cos(\theta_1 - \omega)\,\varphi(\rho\cos(\theta_1 - \omega))\,d\rho$, gives us

$$\begin{aligned}I_1 &\equiv \int \cos(\theta_1-\omega)\,\varphi(\rho\cos(\theta_1-\omega))\,\Phi(-\rho\sin(\theta_1-\omega))\,d\rho \\ &= \Phi(-\rho\sin(\theta_1-\omega))\,\Phi(\rho\cos(\theta_1-\omega)) + \\ &\quad + \int \sin(\theta_1-\omega)\,\varphi(\rho\sin(\theta_1-\omega))\,\Phi(\rho\cos(\theta_1-\omega))\,d\rho.\end{aligned}$$

Then,

$$\begin{aligned}P(\rho;\theta_1,\theta_2) &= \int \frac{dP(\rho;\theta_1,\theta_2)}{d\rho}d\rho + const. \\ &= \Phi(-\rho\sin(\theta_1-\omega))\,\Phi(\rho\cos(\theta_1-\omega)) + const. \\ &= \Phi(-\rho\sin(\theta_1-\omega))\,\Phi(\rho\cos(\theta_1-\omega)).\end{aligned}$$

The value of the constant is obtained from the initial value $P(0;\theta_1,\theta_2) = \frac{\theta_2-\theta_1}{2\pi} = \frac{1}{4}$. Since $\Phi(0)\,\Phi(0) = \frac{1}{4}$, the value of the constant is zero.

CHAPTER 4

DIRICHLET PROCESS MIXTURE MODELS FOR DIRECTIONAL DATA

Summary

We consider Dirichlet process mixture (DPM) models for directional data using the projected normal distribution. We show how they can be fit using Markov chain Monte Carlo (MCMC) methods after the introduction of suitable latent variables. A large dataset of daily departures of anglers makes the MCMC method infeasible and thus we explore mean field variational methods. We identify a number of problems in the application of the mean field variational method for DPM models for circular data, caused by the poor approximation of the variational approximation to the true posterior distribution. We propose solutions to those problems by improving the mean field variational approximation through the use of novel Monte Carlo procedures that are incorporated into the variational algorithm. The procedures are applied to the angler departure dataset, where the variational and MCMC methods are compared.

4.1 Introduction

In Hernandez-Stumpfhauser et al. (2011), a Bayesian hierarchical for directional data was used to obtain model predictions of the daily distributions of the departures of anglers from fishing sites. The current article will explore a more sophisticated approach that better handles some of the features often encountered in this type of data, including non-symmetric distributions and multi-modality.

We begin by briefly reviewing the data analysis context for both articles. We are interested in obtaining predictions of the daily distributions of the departures of recreational anglers along the coasts of the United States, as a function of the type of fishing trip, its location and time of the year. These predictions are an input into the estimation procedure

to estimate recreational fish catch by species and size class used by the U.S. National Marine Fisheries Service. The data for these estimates are collected by the Marine Recreational Fisheries Statistics Survey (MRFSS) which consists of two separate and complementary surveys: the Access Point Angler Intercept Survey (APAIS) which collects data on catch at the fishing site and the Coastal Household Telephone Survey (CHTS) which collects fishing activity data. The APAIS data are used to estimate average catch per angler trip and the CHTS data are used to estimate total angler trips. Roughly speaking, the estimated total catch is obtained as the product of the two previous estimates.

The APAIS consists of two or more stages of sampling with different sampling probabilities. A serious problem with the estimation in the original APAIS was that it did not incorporate crucial aspects of the sampling design in the estimation, as noted by the US National Academy of Sciences panel (Sullivan et al. 2006). An unbiased (and unfeasible) survey estimator of the total $t_y = \sum_{d \in U_I} \sum_{a \in U_d} y_{da}$ for y_{da} = catch characteristic of angler a on site-day d is given by

$$\hat{t}_y = \sum_{d \in s_I} \sum_{a \in s_d} \frac{y_{da}}{\pi_{Id}\, \pi_{a|d}},$$

where U_I denotes the set of all site-days in a domain of interest, U_d denotes the set of all anglers departing the site during that day, $\pi_{Id} > 0$ for $d \in U_I$ is the first-order inclusion probability of site-days and $\pi_{a|d} > 0$ is the probability that angler a on site day d is intercepted by the interviewer. All quantities are known except for $\pi_{a|d}$. Even if we assume that all anglers at the site are equally likely to be selected, so that $\pi_{a|d} = n_d/N_d$ with n_d the intercepted anglers at the site and N_d the total number of anglers, N_d is not observed because the interviewer is only at the site for a fraction of the 24-hour period.

In order to create APAIS estimates that continue to incorporate weighing to account for varying inclusion probabilities, $\hat{t}_y$ above was replaced by an estimator in which the unknown $\pi_{a|d}$ are replaced by model-based estimates. We rewrite $\pi_{a|d} = (n_d/N_{\Delta d})(F_{\Delta d})$, where Δ_d represents the "slice" of time during which the interviewer visited the site on a given day, $F_{\Delta d}$ denotes the fraction of anglers who departed the site during that time slice and $N_{\Delta d}$

represents the number of anglers departing the site during Δ_d. The unknown quantity $F_{\Delta d}$ will be estimated as a function of type of fishing trip (mode), its location (state) and time of the year (wave) is the quantity of interest in this paper.

The CHTS collects data on over one million angling trips from a random sample of anglers. These data contain trip departure times as well as site characteristics, so that they can be used to estimate $F_{\Delta d}$. Hernandez-Stumpfhauser et al. (2011) treated the departure times data as circular and developed a Bayesian modeling approach that made it possible to specify and fit regression models for circular data, based on the projected normal distribution (Presnell et al. 1998). After fitting the regression model to the departure times, they used composite estimation (Ghosh and Rao, 1994) to predict $F_{\Delta d}$ for each state-wave-mode "cell." While this resulted in reasonable predictions, the composite estimation was somewhat ad hoc and it was clear that the assumption of a single projected normal distribution for departures in each cell was not a good model for many cells.

Here in this paper we avoid the composite estimation procedure by considering Bayesian regression models for mixtures of projected normal distributions. The number of mixture components is unknown a priori and is to be inferred from the data, which will make it possible to fit the complicated data patterns present in the data. The clustering property of Dirichlet processes (DP) provides a nonparametric prior for the number of mixture components. The estimation of $F_{\Delta d}$ is then done by modeling the data in each combination of state, wave and mode as a sample from mixtures of projected normal distributions whose parameters follow some regression model. This allows for direct inference on uncertainty about density estimates, assessment of modality, and inference on the number of components.

We will first develop the model and describe a Gibbs sampler estimation method. However, due to large sample sizes, this approach will not be feasible for the application of interest. Hence, we also explore variational methods which are approximations to the posterior distribution based on deterministic algorithms. Finally we find the need to improve the variational approximations by making use of a sampling scheme that looks similar to a

Gibbs sampler.

The remainder of the paper is organized as follows. In section 4.2 we provide basic background on projected normal distributions and on DP mixture models. In sections 4.3 and 44. we describe the Gibbs sampler and variational algorithms for the case of a random sample from DP mixture of projected normal distributions, respectively. Section 4.5 discusses initialization of the variational algorithm, a major issue with the variational estimate, and improvements of the variational distribution. Section 4.6 presents two regression models and analysis of the CHTS data, and section 4.7 presents our conclusions.

4.2 Directional data and Dirichlet process mixture models

We first briefly review the projected normal distribution and introduce notation, ignoring the regression context for now. The directional data point θ_t, $t = 1, \ldots, n$ is drawn from

$$\theta_t | \boldsymbol{\mu} \sim \mathrm{PN}_p \left(\boldsymbol{\mu}, \mathbf{I}_p \right), \tag{26}$$

with mean vector $\boldsymbol{\mu}$. This means there is a multivariate normal random variable $\mathbf{X}_t | \boldsymbol{\mu} \sim \mathrm{N}_p \left(\boldsymbol{\mu}, \mathbf{I}_p \right)$ and θ_t is the polar representation of the projection of $\mathbf{X}_t$ onto the unit circle ($p = 2$) or the unit sphere ($p = 3$) for circular and spherical data respectively (Mardia and Jupp, 2000, p.46). Here we include the spherical case because it is a natural extension but we will primarily concerned with the circular case, since that is the case of interest in our data set. In the circular case θ is one-dimensional with support $\theta \in (0, 2\pi]$. The probability density function $\mathrm{PN}_2 \left(\boldsymbol{\mu}, \mathbf{I}_2 \right)$ can be written as

$$\mathrm{PN}_2 \left(\theta; \boldsymbol{\mu}, \mathbf{I}_2 \right) = \frac{1}{2\pi} \exp \left\{ -\frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} \right\} \left[1 + \frac{\mathbf{u}^T \boldsymbol{\mu} \Phi \left(\mathbf{u}^T \boldsymbol{\mu} \right)}{\varphi \left(\mathbf{u}^T \boldsymbol{\mu} \right)} \right]$$

(Mardia and Jupp, 2000, p.46). The vector $\mathbf{u}$ is equal to $(\cos\theta, \sin\theta)^T$ and the two-dimensional mean vector $\boldsymbol{\mu}$ is the only parameter in $\mathrm{PN}_2 \left(\boldsymbol{\mu}, \mathbf{I}_2 \right)$. The spherical case is

described in Supplement 4.7.

If the data points θ_t come from a mixture of projected normals, we allow for different mean vectors $\boldsymbol{\mu}_t$. When the means $\boldsymbol{\mu}_t$ come from some uncertain prior distribution $G(\cdot)$ on $\mathbb{R}^p$ and $G(\cdot)$ is modeled as a Dirichlet process, then the $\mathbf{X}_t$ come from a Dirichlet mixture of Normals (Escobar and West, 1995), and hence the θ_t come from a Dirichlet mixture of Projected Normals. The Dirichlet process $\mathrm{DP}(\alpha, G_0)$ is a distribution over distributions (Ferguson, 1973). It has two parameters, a scaling parameter $\alpha > 0$ and a base distribution G_0. The discreteness of the DP makes it suitable for the problem of placing priors on mixture components. A few examples of application areas in which DP models have been used are density estimation (Escobar and West, 1995), document modeling and genetics (Teh et al. 2004) and image analysis (Blei and Jordan, 2006).

In the remainder of this section, we give the definition of a Dirichlet process and two of its most common representations, the Polya urn and stick-breaking representations. The Gibbs sampler will be based on the Polya urn representation, while the variational approximations we will work with in later sections will be based on truncated stick-breaking representations.

Using the terminology of Ferguson (1973), let Ω be a set, and $\mathcal{F}$ a σ- field of subsets of Ω. Let G_0 be a finite, nonnull, nonnegative, finitely additive measure on $(\Omega, \mathcal{F})$. We say a random probability measure G on $(\Omega, \mathcal{F})$ is a Dirichlet process on $(\Omega, \mathcal{F})$ with parameter G_0, if for every $k = 1, 2, \ldots$ and measurable partition $B_1, B_2, \ldots, B_k$ of Ω, the joint distribution of the random probabilities $(G(B_1), \ldots, G(B_k))$ is Dirichlet with parameters $(G_0(B_1), \ldots, G_0(B_k))$. Suppose we draw a random measure G from a Dirichlet Process $\mathrm{DP}(\alpha, G_0)$, and independently draw n random variables $\boldsymbol{\mu}_t$ from G, $t = 1, 2, \ldots, n$. Then marginalizing out the random measure G, the joint distribution of $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n\}$ follows a *Polya urn* scheme (Blackwell and MacQueen, 1973),

$$\boldsymbol{\mu}_t | \boldsymbol{\mu}^{(t)} \sim \frac{\alpha}{\alpha + t - 1} G_0(\boldsymbol{\mu}_t) + \frac{1}{\alpha + t - 1} \sum_{j=1, j \neq t}^{n} \delta_{\boldsymbol{\mu}_j}(\boldsymbol{\mu}_t) \tag{27}$$

where $\delta_{\boldsymbol{\mu}_j}(\boldsymbol{\mu})$ denotes a unit point mass at $\boldsymbol{\mu} = \boldsymbol{\mu}_j$ and $\boldsymbol{\mu}^{(t)} = \left\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{t-1}, \boldsymbol{\mu}_{t+1} \ldots, \boldsymbol{\mu}_n\right\}$. Thus, a draw from $\boldsymbol{\mu}_t|\boldsymbol{\mu}^{(t)}$ is a draw from G_0 with probability $\frac{\alpha}{\alpha+t-1}$ and with probability $\frac{1}{\alpha+t-1}$ a uniform draw from $\left\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{t-1}, \boldsymbol{\mu}_{t+1} \ldots, \boldsymbol{\mu}_n\right\}$.

Sethuraman (1994) provides a characterization of the DP in terms of a stick-breaking construction. Consider two infinite collections of independent random variables $v_i \sim \text{Beta}(1, \alpha)$ and $\boldsymbol{\mu}_i \sim G_0$ for $i = 1, 2, \ldots$. Then,

$$\begin{aligned} \pi_i(\mathbf{v}) &= v_i \prod_{j=1}^{i-1}(1 - v_j) \\ G(\cdot) &= \sum_{i=1}^{\infty} \pi_i(\mathbf{v})\, \delta_{\boldsymbol{\mu}_i}(\cdot) \end{aligned}$$

The stick-breaking representation of the DP makes it clear that G is discrete with a support consisting of a countably infinite set of atoms, drawn independently from G_0. The proportions $\pi_i(\mathbf{v})$ are constructed by successively breaking a unit length stick. The "breaks" $\{v_1, v_2 \ldots\}$ are independent draws from $\text{Beta}(1, \alpha)$.

4.3 Estimation using Gibbs sampling

In this section, we assume that the data are a random sample from a DP mixture of projected normal distributions. Marginalizing out the random distribution $G(\cdot)$ the mean vectors follow a Polya urn scheme (27). If the base distribution is normally distributed, $G_0 = \text{N}_p(\boldsymbol{\mu}_0, \mathbf{I}_p)$, the full conditionals required for the Gibbs sampler are

$$\begin{aligned} \left(\boldsymbol{\mu}_t|\boldsymbol{\mu}^{(t)}, \mathbf{r}, \boldsymbol{\theta}\right) &\sim q_0 G_t(\boldsymbol{\mu}_t) + \sum_{j=1, j\neq t}^{n} q_j \delta_{\boldsymbol{\mu}_j}(\boldsymbol{\mu}_t) \\ p(r_t|\boldsymbol{\mu}, \boldsymbol{\theta}) &\propto r_t^{p-1} \exp\left(-\frac{1}{2} r_t^2 + \mathbf{u}^T \boldsymbol{\mu}_t r_t\right), \end{aligned} \tag{28}$$

where r_t denotes the length of the bivariate normal vector $\boldsymbol{x}_t$, $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{(t)}$ are defined as in Equations (26) and (27), $G_t(\boldsymbol{\mu}_t) = \text{N}_p\left(\frac{1}{2}(\mathbf{x}_t + \boldsymbol{\mu}_0), \frac{1}{2}\mathbf{I}_p\right)$, $q_0 \propto \alpha \text{N}_p(\mathbf{x}_t; \boldsymbol{\mu}_0, 2\mathbf{I}_p)$ and

$q_j \propto \mathrm{N}_p\left(\mathbf{x}_t; \boldsymbol{\mu}_j, \mathbf{I}_p\right)$ and $\mathbf{x}_t = r_t \mathbf{u}_t$. By $\mathrm{N}_p\left(\mathbf{x}_t; \boldsymbol{\mu}_j, \mathbf{W}_p\right)$, we denote the p-variate normal density with parameters $(\boldsymbol{\mu}_j, \mathbf{W}_p)$ evaluated at $\mathbf{x}_t$, and by $\mathrm{N}_p\left(\boldsymbol{\mu}_j, \mathbf{W}_p\right)$ we mean the p-variate normal distribution with parameters $\boldsymbol{\mu}_j, \mathbf{W}_p$. Thus, with probability proportional to q_0, $\left(\boldsymbol{\mu}_t | \boldsymbol{\mu}^{(t)}, \mathbf{r}, \boldsymbol{\theta}\right)$ is a draw from G_t and with probability proportional to q_j is equal to $\boldsymbol{\mu}_j$ for $j = 1, \ldots, t-1, t+1, \ldots, n$.

Draws from $p\left(r_t | \boldsymbol{\mu}, \boldsymbol{\theta}\right)$ can be obtained via Metropolis-Hastings, adaptive rejection sampling or by making use of the efficient sampler proposed in Hernandez-Stumpfhauser et al. 2011. This last one is the fastest and we briefly describe it next. Defining $b = \mathbf{u}^T \boldsymbol{\mu}$, the distribution of the latent length is

$$p\left(r | \boldsymbol{\mu}, \theta\right) \propto r^{p-1} \exp\left(-\frac{1}{2}\left(r-b\right)^2\right). \tag{29}$$

We introduce the latent variable Y which has joint density with r given by

$$p\left(r, y | \boldsymbol{\mu}, \theta\right) \propto r^{p-1} \mathrm{I}_{\left(0, \exp\left\{-\frac{1}{2}(r-b)^2\right\}\right)}\left(y\right) \mathrm{I}_{(0,\infty)}\left(r\right).$$

Then, the full conditionals are

$$\begin{aligned} \left(Y | R = r, \boldsymbol{\mu}, \theta\right) &\sim \mathrm{U}\left(0, \exp\left\{-\frac{1}{2}\left(r-b\right)^2\right\}\right) \\ p\left(r | Y = y, \boldsymbol{\mu}, \theta\right) &\propto r^{p-1} \mathrm{I}_{\left(b+\max\left\{-b, -\sqrt{-2\ln y}\right\}, b+\sqrt{-2\ln y}\right)}\left(r\right), \end{aligned} \tag{30}$$

where this last one is very easy to sample from using the inverse cdf technique. Thus we draw $y_t \sim \mathrm{U}\left(0, \exp\left\{-\frac{1}{2}\left(r_t - b_t\right)^2\right\}\right)$ and independently we draw $u_t \sim \mathrm{U}\left(0,1\right)$. Finally we get a draw r_t by letting $r_t = \sqrt[p]{\left(r_{2,t}^p - r_{1,t}^p\right) u_t + r_{1,t}^p}$, where $r_{1,t} = b_t + \max\left\{-b_t, -\sqrt{-2\ln y_t}\right\}$ and $r_{2,t} = b_t + \sqrt{-2\ln y_t}$.

We can also include prior distributions for the parameters in the Dirichlet Process $\mathrm{DP}\left(\alpha, G_0 = \mathrm{N}_p\left(\boldsymbol{\mu}_0, \mathbf{I}_p\right)\right)$. Suppose $\boldsymbol{\mu}_0 \sim \mathrm{N}_p\left(m, \sigma_\mu^2 \mathbf{I}_p\right)$ and following Escobar and West (1995), $\alpha \sim \mathrm{Gamma}(a, b)$. Given $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$, $\boldsymbol{\mu}_0$ is conditionally independent of $\boldsymbol{\theta}_t$ $(t = 1, \ldots, n)$

and depends only on the distinct values $\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_k^*$ of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$. The full conditional for $\boldsymbol{\mu}_0$ is Normal with mean $\left[k + \frac{1}{\sigma_\mu^2}\right]^{-1} \left[\frac{m}{\sigma_\mu^2} + \sum \boldsymbol{\mu}_j^*\right]$ and variance $\left[k + \frac{1}{\sigma_\mu^2}\right]^{-1} \mathbf{I}_p$. If $\alpha \sim$ Gamma(a, b) where b is the rate parameter, Escobar and West (1995) also show that the full conditional for α is the marginal distribution from a joint distribution for α and a continuous quantity η such that $p(\alpha, \eta|k) \propto p(\alpha)\, \alpha^{k-1} (\alpha + n)\, \eta^\alpha (1 - \eta)^{n-1}$ for $\alpha > 0$ and $0 < \eta < 1$. Making use of the auxiliary variable η, the conditional distribution of α is a mixture of two gamma densities

$$p(\alpha|\eta, k) = \pi_\eta \mathrm{G}(\alpha; a + k, b - \log \eta) + (1 - \pi_\eta)\, \mathrm{G}(\alpha; a + k - 1, b - \log \eta)$$

with weight π_η defined by $\pi_\eta / (1 - \pi_\eta) = (a + k - 1) / \{n (b - \log \eta)\}$. The distribution of η conditioned on α and k is a beta distribution

$$p(\eta|\alpha, k) = \mathrm{Beta}(\eta; \alpha + 1, n).$$

The Gibbs sampler based on the full conditionals described above can not change the $\boldsymbol{\mu}$ for more than one observation simultaneously, making the convergence to the posterior distribution slow (Neal, 2000). However, resampling $\boldsymbol{\mu}_j^*$ conditional on the configuration of all other parameters is straightforward (MacEachern and Muller, 1998). For a fixed j, the full conditional for $\boldsymbol{\mu}_j^*$ is the posterior distribution in the simple Bayesian model $\mathbf{x}_t \sim \mathrm{N}_p\left(\boldsymbol{\mu}_j^*, \mathbf{I}_p\right)$ and $\boldsymbol{\mu}_j^* \sim G_0 = N_p(\boldsymbol{\mu}_0, \mathbf{I}_p)$, or

$$\left(\boldsymbol{\mu}_j^*|\cdot\right) \sim \mathrm{N}_p\left(\left[\frac{1}{n_j + 1}\right]\left[\boldsymbol{\mu}_0 + \sum_{s=1}^{n_j} \mathbf{x}_s\right], \left[\frac{1}{n_j + 1}\right] \mathbf{I}_p\right),$$

where n_j are the number of observations associated to $\boldsymbol{\mu}_j^*$, $j = 1, \ldots, k$, $t = 1, \ldots, n$ and by $(\boldsymbol{\mu}^*|\cdot)$ we mean $\boldsymbol{\mu}^*$ conditioned on the value of all other parameters.

In our application, the quantity of interest is the predictive distribution, i.e. the distribution of a new observation given the data. Let $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n\}$, $\boldsymbol{\mu}_{n+1}$ be the (new) mean

of a new observation θ_{n+1} and $\boldsymbol{\theta}_n$ be the data. Conditioned on $\boldsymbol{\mu}$ and $\boldsymbol{\theta}_n$, the distribution of a new observation is independent of the data, i.e. $p\left(\theta_{n+1}|\boldsymbol{\mu},\boldsymbol{\theta}_n\right)=p\left(\theta_{n+1}|\boldsymbol{\mu}\right)$, which may be evaluated as $\int p\left(\theta_{n+1}|\boldsymbol{\mu}_{n+1}\right)dP\left(\boldsymbol{\mu}_{n+1}|\boldsymbol{\mu}\right)$, where

$$\begin{aligned} p\left(\theta_{n+1}|\boldsymbol{\mu}_{n+1}\right) &= \mathrm{PN}_p\left(\theta_{n+1};\boldsymbol{\mu}_{n+1},\mathbf{I}_p\right) \\ p\left(\boldsymbol{\mu}_{n+1}|\boldsymbol{\mu}\right) &= \alpha a_n G_0\left(\boldsymbol{\mu}_{n+1}\right)+a_n\sum_{t=1}^{n}\delta_{\boldsymbol{\mu}_t}\left(\boldsymbol{\mu}_{n+1}\right). \end{aligned} \tag{31}$$

Thus, we need to evaluate an integral of the form $\int \mathrm{PN}_p\left(\theta;\boldsymbol{\eta},\mathbf{I}_p\right)\mathrm{N}_p\left(\boldsymbol{\eta};\boldsymbol{\mu}_0,\mathbf{I}_p\right)d\boldsymbol{\eta}$ and this can be done relatively easily by making use of the joint distribution of r and θ, $p\left(r,\theta|\boldsymbol{\eta}\right)=r^{p-1}\mathrm{N}_p\left(r\mathbf{u};\boldsymbol{\eta},\mathbf{I}_p\right)$. Hence, the predictive distribution for a new observation requires computing an integral of the form $\int \mathrm{PN}_p\left(\theta;\boldsymbol{\eta},\mathbf{I}_p\right)\mathrm{N}_p\left(\boldsymbol{\eta};\boldsymbol{\mu}_0,\mathbf{I}_p\right)d\boldsymbol{\eta}=\int\int\left[r^{p-1}\mathrm{N}_p\left(r\mathbf{u};\boldsymbol{\eta},\mathbf{I}_p\right)\right.$ $\left.\mathrm{N}_p\left(\boldsymbol{\eta};\boldsymbol{\mu}_0,\mathbf{I}_p\right)\right]d\boldsymbol{\eta}dr$. Integrating first with respect to $\boldsymbol{\eta}$ gives a normal distribution with mean $\boldsymbol{\mu}_0$ and variance equal to $2\mathbf{I}_p$. Integrating secondly with respect to r gives a projected normal density $\mathrm{PN}_p\left(\theta;\boldsymbol{\mu}_0,2\mathbf{I}_p\right)$ which is equivalent to a projected normal density $\mathrm{PN}_p\left(\theta;\frac{\boldsymbol{\mu}_0}{\sqrt{2}},\mathbf{I}_p\right)$.

The predictive distribution can be approximated as follows

$$\begin{aligned} p\left(\theta_{n+1}|\boldsymbol{\theta}_n\right) &= \int p\left(\theta_{n+1}|\boldsymbol{\mu},\alpha,\boldsymbol{\mu}_0\right)dP\left(\boldsymbol{\mu},\alpha,\boldsymbol{\mu}_0|\boldsymbol{\theta}_n\right) \\ &\approx N^{-1}\sum_{i=1}^{N}p\left(\theta_{n+1}|\boldsymbol{\mu}(i),\alpha(i),\boldsymbol{\mu}_0(i)\right), \end{aligned}$$

where $p\left(\theta_{n+1}|\boldsymbol{\mu}(i),\alpha(i),\boldsymbol{\mu}_0(i)\right)=\alpha(i)a_n(i)\mathrm{PN}_p\left(\boldsymbol{\mu}_0(i),2\mathbf{I}_p\right)+a_n(i)\sum_{t=1}^{n}\mathrm{PN}_p\left(\boldsymbol{\mu}_t(i),\mathbf{I}_p\right)$, $a_n(i)=1/\left(\alpha(i)+n\right)$ and $\boldsymbol{\mu}(i),\alpha(i),\boldsymbol{\mu}_0(i)$ denotes the ith draw of the Gibbs sampler after convergence from a total of N draws.

In the above set-up, we fix the variance in the base distribution to be 1. For a single projected normal distribution, it is easy to see that $\mathrm{PN}_p\left(\boldsymbol{\mu},\sigma^2\mathbf{I}_p\right)$ is the same as the distribution $\mathrm{PN}_p\left(\boldsymbol{\mu}/\sigma,\mathbf{I}_p\right)$. While this is not true in the mixture context, the Gibbs sampler exhibited non-stationarity when we tried to include a random σ^2 in the base distribution. The choice $\sigma^2=1$ in the base distribution is somewhat arbitrary, and actually if we use

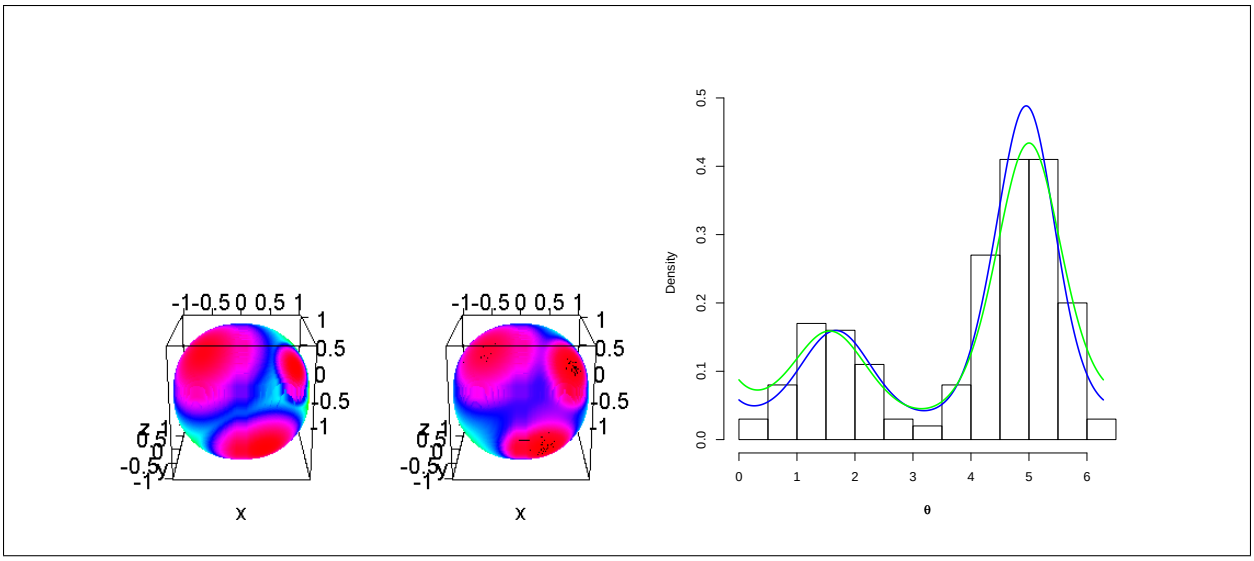

Figure 10: Simulated data and density estimates for spherical and circular cases. The sphere on the left represents the density of a mixture of 3 Projected Normal distributions from which a sample of size 300 was obtained. The sphere on the right represents the estimate of the density and the black dots are the data. The histogram is a random sample of size 200 from a mixture of two Projected Normal distributions. The light curve is the true density and the dark curve is the estimate.

the truncated stick-breaking representation, the variance in the base distribution seems to matter. The choice $\sigma^2 = 1$ led to reasonable fits, however, and hence we did not explore further fine-tuning of this parameter.

The Gibbs sampler based on all of the full conditionals described above will be referred as Algorithm III. Figure 10 shows predictive distributions obtained via the Gibbs sampler, for simulated spherical and circular data. The Gibbs sampler works very well and is easy to implement. A major problem with the Gibbs sampler is that if data sets are large, as in the application we consider in this article, it takes a long time to run or is even impossible to run. In the next section, we explore variational Bayes methods which are fast approximations to the posterior distribution. From now on we will focus on circular data ($p = 2$) although the case for spherical data ($p = 3$) would be treated in the same exact way.

4.4 Mean field variational Bayes approximation

Variational methods involve approximations to marginal or posterior distributions in terms of an optimization problem (Ghahramani and Beal, 2001; Opper and Saad, 2001; Wainwright and Jordan, 2003). Mean-field methods are based on minimizing the Kullback-Leibler (KL) divergence with respect to a variational distribution (Ormerod and Wand, 2010). The KL divergence between distributions $q(\cdot)$ and $p(\cdot|y)$ is defined as $\int q(w) \log\left\{\frac{q(w)}{p(w|y)}\right\} dw$, which is greater or equal to zero for all densities q and equal to zero if and only if $q(w) = p(w|y)$. Consider a model with parameters $\mathbf{w}$ and observations $\mathbf{y}$. Tractability is achieved by restricting $q(\cdot)$ to a more manageable class of densities and then minimizing the KL divergence between $q(\cdot)$ and $p(\cdot|y)$ over that class. The usual restriction for the $q(\cdot)$ density is that $q(w)$ factorizes into $\prod q_i(w_i)$ for some partition $\{w_1, \ldots, w_M\}$ of $\mathbf{w}$.

It can be shown that the solutions to the KL minimization satisfy (Ormerod and Wand, 2010):

$$q_i(w_i) \propto \exp\left\{\mathbf{E}_{-w_i} \log p(w_i|y, \mathbf{w}_{-i})\right\},$$

where $\mathbf{w}_{-i}$ is $\mathbf{w}$ without component w_i and $\mathbf{E}_{-w_i}$ denotes expectation with respect to the density $\prod_{j \neq i} q_j(w_j)$. Generally, the following iterative scheme is used to solve for q_i: initialize $q_1, q_2, \ldots, q_M$ and then cycle

$$\begin{aligned} q_1(w_1) &\propto \exp\left\{\mathbf{E}_{-w_1} \log p(w_1|y, \mathbf{w}_{-1})\right\} \\ &\vdots \\ q_M(w_1) &\propto \exp\left\{\mathbf{E}_{-w_M} \log p(w_M|y, \mathbf{w}_{-M})\right\} \end{aligned} \tag{32}$$

until the increase of a quantity called the *lower bound* $\underline{p}(y, q) = \exp\left[\mathbf{E}_q \log \frac{p(y,\mathbf{w})}{q(\mathbf{w})}\right]$ is negligible. The expectation in the lower bound is with respect to the variational distribution. Minimizing the KL divergence is equivalent to maximizing the lower bound, and working with the lower bound is easier than working with the KL divergence. Convergence to at least

local optima is guaranteed (Boyd and Vandenberghe, 2004). If conjugate priors are used, the q_i are part of recognizable families and updating q_i reduces to updating the parameters in such families.

The variational distribution proposed here is similar to the one proposed by Blei and Jordan (2006), which makes use of the stick-breaking representation of the Dirichlet process, briefly introduced in Section 4.2. A more complete description of the stick-breaking representation of the Dirichlet process mixture model for projected normal data is as follows:

1. Draw $V_i|\alpha \sim \text{Beta}\left(1, \alpha\right), \quad i = \{1, 2, \ldots\}$
2. Draw $\boldsymbol{\eta}_i|G_0 \sim G_0, \quad i = \{1, 2, \ldots\}$
3. For the tth data point:
   (a) Draw $Z_t|\left\{v_1, v_2, \ldots\right\} \sim \text{Mult}\left(\pi\left(\boldsymbol{v}\right)\right)$
   (b) Draw $X_t|z_t \sim \text{N}_2\left(\boldsymbol{\eta}_{z_t}, \mathbf{I}_2\right)$
   (c) Observe θ_t

where $G_0 = \text{N}_2\left(\boldsymbol{\mu}, \mathbf{I}_2\right)$, $\pi_i\left(\mathbf{v}\right) = v_i \prod_{j=1}^{i-1}\left(1 - v_j\right)$ and θ_t is the angle in the polar representation of $X_t = r_t\left(\cos\theta_t, \sin\theta_t\right)^T$.

As in Blei and Jordan (2006), we consider truncated stick-breaking representations as a family of variational distributions that approximates the distribution of the infinite-dimensional random measure G, where the random measure is expressed in terms of the infinite sets $\boldsymbol{V} = \{V_1, V_2, \ldots\}$ and $\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots\}$. This is done by selecting a value L and letting $q\left(v_L = 1\right) = 1$; this implies that the mixture proportions $\pi_i\left(\boldsymbol{v}\right)$ are equal to zero for $i > L$. The truncation level L is a variational parameter and can be set freely. The proposed factorized family of variational distributions we consider here is

$$q\left(\mathbf{v}, \boldsymbol{\eta}, \mathbf{r}, \mathbf{z}\right) = \prod_{i=1}^{L-1} q\left(v_i\right) \prod_{i=1}^{L} q\left(\boldsymbol{\eta}_i\right) \prod_{t=1}^{n} q\left(r_t\right) \prod_{t=1}^{n} q\left(z_t\right), \tag{33}$$

where we have written $q(v_i)$ as a shorthand for $q_{v_i}(v_i)$ and the same for all other parameters. We will continue with this shorthand notation whenever it does not lead to confusion as to which term in the variational approximation is being referred to.

To apply the iterative algorithm based on (32), we need to find the full conditionals $p(v_i|\cdot)$, $p(\boldsymbol{\eta}_i|\cdot)$, $p(r_t|\cdot)$, $p(z_t|\cdot)$, which are given by

$$\begin{aligned}
p(v_i|\cdot) &= \text{Beta}\left(v_i; \sum_{t=1}^{n} z_t^i + 1, \sum_{t=1}^{n}\sum_{j=1}^{\infty} z_t^{i+j} + \alpha\right) \\
p(\boldsymbol{\eta}_i|\cdot) &= \text{N}_2\left(\boldsymbol{\eta}_i; \left[\frac{1}{1+\sum z_t^i}\right]\left[\boldsymbol{\mu} + \sum_{t=1}^{n} z_t^i r_t \mathbf{u}_t\right], \left[\frac{1}{1+\sum z_t^i}\right]\mathbf{I}_2\right) \\
p(r_t|\cdot) &\propto r_t \exp\left\{-\frac{1}{2}r_t^2 + r_t \mathbf{u}_t^T \boldsymbol{\eta}_{z_t}\right\} \\
p\left(z_t^i = 1|\cdot\right) &\propto \left[v_i \prod_{j=1}^{i-1}(1-v_j)\right] \exp\left\{-\frac{1}{2}\left(r_t\mathbf{u}_t - \boldsymbol{\eta}_i\right)^T\left(r_t\mathbf{u}_t - \boldsymbol{\eta}_i\right)\right\}.
\end{aligned}$$

Applying the general iterative algorithm in (32), the variational distribution $q(v_i)$ must satisfy $q(v_i) \propto \exp\{\mathbf{E}_{-v_i} \log p(v_i|\cdot)\}$ and similarly for the variational distributions of all other parameters. Thus the variational distributions must satisfy

$$\begin{aligned}
q(v_i) &= \text{Beta}\left(v_i; \sum_{t=1}^{n} \mathbf{E}\left(z_t^i\right) + 1, \sum_{t=1}^{n}\sum_{j=i+1}^{L} \mathbf{E}\left(z_t^j\right) + \alpha\right) \qquad (34) \\
q(\boldsymbol{\eta}_i) &= \text{N}_2\left(\boldsymbol{\eta}_i; \left[\frac{1}{1+\sum \mathbf{E}\left(z_t^i\right)}\right]\left[\boldsymbol{\mu} + \sum_{t=1}^{n} \mathbf{E}\left(z_t^i\right)\mathbf{E}\left(r_t\right)\mathbf{u}_t\right], \left[\frac{1}{1+\sum \mathbf{E}\left(z_t^i\right)}\right]\mathbf{I}_2\right) \\
q(r_t) &\propto r_t \exp\left\{-\frac{1}{2}r_t^2 + r_t \mathbf{u}_t^T \sum_{i=1}^{L} \mathbf{E}\left(\boldsymbol{\eta}_i\right)\mathbf{E}\left(z_t^i\right)\right\} \\
q\left(z_t^i = 1\right) &\propto \exp\left\{\mathbf{E}\left(\log\left(v_i\right)\right) + \sum_{s=1}^{i-1}\mathbf{E}\left(\log\left(1-v_s\right)\right) - \frac{1}{2}\mathbf{E}\left(\|\boldsymbol{\eta}_i\|^2\right) + \mathbf{E}\left(r_t\right)\mathbf{u}_t^T\mathbf{E}\left(\boldsymbol{\eta}_i\right)\right\},
\end{aligned}$$

where all expectations are with respect to the variational distribution (33). The algorithm, which we will refer to as Algorithm 3, is shown on the next page.

An update of these expectations is equivalent to an update of the variational distributions $q(v_i), q(\boldsymbol{\eta}_i), q(r_t), q(z_t^i = 1)$. We keep updating these expectations until the increase of the

Algorithm 3 Iterative scheme for obtaining the parameters in the variational distributions in (34). The Ψ denotes the digamma function and Φ the cdf of standard normal.

Initialize $\mathbf{E}\left(z_t^i\right)$ and $\mathbf{E}\left(\boldsymbol{\eta}_i\right)$ for $i = 1, 2, \ldots, L$ and $t = 1, 2, \ldots, n$

$$
\begin{aligned}
\gamma_{i,1} &\leftarrow 1 + \sum_{t=1}^{n} \mathbf{E}\left(z_t^i\right) \\
\gamma_{i,2} &\leftarrow \sum_{t=1}^{n} \sum_{j=i+1}^{L} \mathbf{E}\left(z_t^j\right) + \alpha \\
\mathbf{E}\left(\log\left(v_i\right)\right) &\leftarrow \Psi\left(\gamma_{i,1}\right) - \Psi\left(\gamma_{i,1} + \gamma_{i,2}\right) \\
\mathbf{E}\left(\log\left(1 - v_s\right)\right) &\leftarrow \Psi\left(\gamma_{i,2}\right) - \Psi\left(\gamma_{i,1} + \gamma_{i,2}\right) \\
b_t &\leftarrow \mathbf{u}_t^T \sum_{i=1}^{L} \mathbf{E}\left(\boldsymbol{\eta}_i\right) \mathbf{E}\left(z_t^i\right) \\
\mathrm{C}\left(b_t\right) &\leftarrow 1 + \sqrt{2\pi} b_t \exp\left(b_t^2/2\right) \Phi\left(b_t\right) \\
\mathbf{E}\left(r_t\right) &\leftarrow \frac{\sqrt{2\pi}}{\mathrm{C}\left(b_t\right)} \exp\left(b_t^2/2\right) \Phi\left(b_t\right) + b_t \\
\mathbf{E}\left(\boldsymbol{\eta}_i\right) &\leftarrow \left[\frac{1}{1 + \sum \mathbf{E}\left(z_t^i\right)}\right] \left[\boldsymbol{\mu} + \sum_{t=1}^{n} \mathbf{E}\left(z_t^i\right) \mathbf{E}\left(r_t\right) \mathbf{u}_t\right] \\
\mathbf{E}\left(\boldsymbol{\eta}_{i1}^2\right) &\leftarrow \left[\frac{1}{1 + \sum \mathbf{E}\left(z_t^i\right)}\right] + \left[\mathbf{E}\left(\boldsymbol{\eta}_{i1}\right)\right]^2 \\
\mathbf{E}\left(\boldsymbol{\eta}_{i2}^2\right) &\leftarrow \left[\frac{1}{1 + \sum \mathbf{E}\left(z_t^i\right)}\right] + \left[\mathbf{E}\left(\boldsymbol{\eta}_{i2}\right)\right]^2 \\
\mathbf{E}\left(z_t^i\right) &\propto \exp\left\{\mathbf{E}\left(\log\left(v_i\right)\right) + \sum_{s=1}^{i-1} \mathbf{E}\left(\log\left(1 - v_s\right)\right) - \frac{1}{2}\mathbf{E}\left(\parallel \boldsymbol{\eta}_i \parallel^2\right) + \right. \\
&\qquad \left. \mathbf{E}\left(r_t\right) \mathbf{u}_t^T \mathbf{E}\left(\boldsymbol{\eta}_i\right)\right\}
\end{aligned}
$$

lower bound is negligible. The lower bound needs to be computed at the end of each iteration and its values should be monotonically increasing.

We now discuss the computation of the lower bound in more detail for the specific model we are considering here. The log lower bound $\log \underline{p}\left(\boldsymbol{\theta}, q\right) = \mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z})}{q(\mathbf{r}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z})}$ can be written

as follows

$$\begin{aligned}\mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z})}{q(\mathbf{r}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z})} &= \mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}|\boldsymbol{\eta}, \boldsymbol{z})\, p(\boldsymbol{z}|\boldsymbol{v})\, p(\boldsymbol{\eta})\, p(\boldsymbol{v})}{q(\mathbf{r})\, q(\boldsymbol{\eta})\, q(\boldsymbol{v})\, q(\boldsymbol{z})} \\ &= \mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}|\boldsymbol{\eta}, \boldsymbol{z})\, p(\boldsymbol{z}|\boldsymbol{v})}{q(\boldsymbol{z})} - \mathbf{E}_q \log q(\mathbf{r}) + \mathbf{E}_q \log \frac{p(\boldsymbol{\eta})}{q(\boldsymbol{\eta})} + \\ &\quad \mathbf{E}_q \log \frac{p(\boldsymbol{v})}{q(\boldsymbol{v})}.\end{aligned}$$

Using the fact that the joint distribution of $\mathbf{r}$ and $\boldsymbol{\theta}$ given all other parameters is $p(\mathbf{r}, \boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{z}) = \prod_{t=1}^{n} r_t \mathrm{N}_2\left(r_t \mathbf{u}_t; \sum_{i=1}^{L} z_t^i \boldsymbol{\eta}_i, \mathbf{I}_2\right)$, the first and second terms in $\log \underline{p}(\boldsymbol{\theta}, q)$ are

$$\begin{aligned}\mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}|\boldsymbol{\eta}, \boldsymbol{z})\, p(\boldsymbol{z}|\boldsymbol{v})}{q(\boldsymbol{z})} - \mathbf{E}_q \log q(\mathbf{r}) &= -n \log 2\pi + \sum_{t=1}^{n} \log \sum_{i=1}^{L} \exp\{S_{ti}\} \\ &\quad + \sum_{t=1}^{n} \log \mathrm{C}(b_t) - \sum_{t=1}^{n} \mathbf{E}_q(r_t)\, b_t\end{aligned}$$

where $S_{ti} = \mathbf{E}_q \log(v_i) + \sum_{s=1}^{i-1} \mathbf{E}_q(\log(1 - v_s)) - \frac{1}{2}\mathbf{E}_q(\| \boldsymbol{\eta}_i \|^2) + \mathbf{E}_q(r_t)\, \mathbf{u}_t^T \mathbf{E}_q(\boldsymbol{\eta}_i)$. The third and forth terms involve standard computations

$$\begin{aligned}\mathbf{E}_q \log \frac{p(\boldsymbol{\eta})}{q(\boldsymbol{\eta})} &= L - \sum_{i=1}^{L} \mathrm{var}_q(\boldsymbol{\eta}_i) + \sum_{i=1}^{L} \log \mathrm{var}_q(\boldsymbol{\eta}_i) - \frac{1}{2} \sum_{i=1}^{L} (\mathbf{E}_q(\boldsymbol{\eta}_l) - \boldsymbol{\mu})^T (\mathbf{E}_q(\boldsymbol{\eta}_l) - \boldsymbol{\mu}) \\ \mathbf{E}_q \log \frac{p(\boldsymbol{v})}{q(\boldsymbol{v})} &= (L-1) \log \alpha + \alpha \sum_{i=1}^{L-1} \mathbf{E}_q \log(1 - v_i) - \sum_{i=1}^{L-1} \log \Gamma(\gamma_{i,1} + \gamma_{i,2}) + \sum_{i=1}^{L-1} \log \Gamma(\gamma_{i,1}) \\ &\quad + \sum_{i=1}^{L-1} \log \Gamma(\gamma_{i,2}) - \sum_{i=1}^{L-1} (\gamma_{i,1} - 1)\, \mathbf{E}_q \log(v_i) - \sum_{i=1}^{L-1} \gamma_{i,2} \mathbf{E}_q \log(1 - v_i),\end{aligned}$$

where Γ denotes the gamma function and $\gamma_{i,1}, \gamma_{i,2}$ defined as above.

The parameters α and $\boldsymbol{\mu}$ play a strong role in the allocation of subjects to the mixture components. To allow for unknown α and $\boldsymbol{\mu}$ so that their values can be determined inside

the model estimation rather than fixed a priori, we choose the priors

$$\begin{aligned} p(\alpha) &= \text{Gamma}(\alpha; a, B) \\ p(\boldsymbol{\mu}) &= \text{N}_2\left(\boldsymbol{\mu}; 0, \sigma_0^2 \mathbf{I}_2\right), \end{aligned}$$

where B is the rate parameter and σ_0^2 is a large number. The full conditionals and variational distributions for α and $\boldsymbol{\mu}$ are then

$$\begin{aligned} p(\alpha|\cdot) &= \text{Gamma}\left(\alpha; a + L - 1, B - \sum_{i=1}^{L-1} \log(1 - v_i)\right) \\ q(\alpha) &= \text{Gamma}\left(\alpha; a + L - 1, B - \sum_{i=1}^{L-1} \mathbf{E}_q \log(1 - v_i)\right) \\ p(\boldsymbol{\mu}|\cdot) &= \text{N}_2\left(\boldsymbol{\mu}; \left[\frac{1}{L + \frac{1}{\sigma_0^2}}\right] \left[\sum_{i=1}^{L} \boldsymbol{\eta}_i\right], \left[\frac{1}{L + \frac{1}{\sigma_0^2}}\right] \mathbf{I}_2\right) \\ q(\boldsymbol{\mu}) &= \text{N}_2\left(\boldsymbol{\mu}; \left[\frac{1}{L + \frac{1}{\sigma_0^2}}\right] \left[\sum_{i=1}^{L} \mathbf{E}_q(\boldsymbol{\eta}_i)\right], \left[\frac{1}{L + \frac{1}{\sigma_0^2}}\right] \mathbf{I}_2\right). \end{aligned}$$

The parameters of the variational distributions $q(\alpha), q(\boldsymbol{\mu})$ can be easily updated by including them in Algorithm 3. The update of $\gamma_{i,2}, \mathbf{E}(\alpha), \mathbf{E}(\boldsymbol{\mu}), \mathbf{E}(\boldsymbol{\eta}_i)$ become

$$\begin{aligned} \mathbf{E}(\alpha) &\leftarrow \frac{a + L - 1}{B - \sum_{i=1}^{L-1} \mathbf{E} \log(1 - v_i)} \\ \gamma_{i,2} &\leftarrow \sum_{t=1}^{n} \sum_{j=i+1}^{L} \mathbf{E}\left(z_t^j\right) + \mathbf{E}(\alpha) \\ \mathbf{E}(\boldsymbol{\mu}) &\leftarrow \left[\frac{1}{L + \frac{1}{\sigma_0^2}}\right] \left[\sum_{i=1}^{L} \mathbf{E}(\boldsymbol{\eta}_i)\right] \\ \mathbf{E}(\boldsymbol{\eta}_i) &\leftarrow \left[\frac{1}{1 + \sum \mathbf{E}(z_t^i)}\right] \left[\mathbf{E}(\boldsymbol{\mu}) + \sum_{t=1}^{n} \mathbf{E}\left(z_t^i\right) \mathbf{E}(r_t)\, \mathbf{u}_t\right]. \end{aligned}$$

The variance in $q(\boldsymbol{\mu})$ has the undesirable property that it can be made arbitrarily small by making L large. Nevertheless, this is still preferable to picking a fixed value for $\boldsymbol{\mu}$. After

allowing α and $\boldsymbol{\mu}$ to be random, the log lower bound becomes

$$\begin{aligned}
\mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\mu}, \alpha)}{q(\mathbf{r}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\mu}, \alpha)} &= \mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}|\boldsymbol{\eta}, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{v}) p(\boldsymbol{\eta}|\boldsymbol{\mu}) p(\boldsymbol{v}|\alpha) p(\boldsymbol{\mu}) p(\alpha)}{q(\mathbf{r}) q(\boldsymbol{\eta}) q(\boldsymbol{v}) q(\boldsymbol{z}) q(\boldsymbol{\mu}) q(\alpha)} \\
&= \mathbf{E}_q \log \frac{p(\boldsymbol{\theta}, \mathbf{r}|\boldsymbol{\eta}, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{v})}{q(\boldsymbol{z})} - \mathbf{E}_q \log q(\mathbf{r}) + \mathbf{E}_q \log \frac{p(\boldsymbol{\eta}|\boldsymbol{\mu})}{q(\boldsymbol{\eta})} + \\
&\quad + \mathbf{E}_q \log \frac{p(\boldsymbol{v}|\alpha) p(\alpha)}{q(\boldsymbol{v}) q(\alpha)} + \mathbf{E}_q \log \frac{p(\boldsymbol{\mu})}{q(\boldsymbol{\mu})}.
\end{aligned}$$

The first and second terms are as before and the third, fourth and fifth terms become

$$\begin{aligned}
\mathbf{E}_q \log \frac{p(\boldsymbol{\eta}|\boldsymbol{\mu})}{q(\boldsymbol{\eta})} &= L - \sum_{i=1}^{L} \mathrm{var}_q(\boldsymbol{\eta}_i) + \sum_{i=1}^{L} \log \mathrm{var}_q(\boldsymbol{\eta}_i) \\
&\quad - \frac{1}{2} \sum_{i=1}^{L} (\mathbf{E}_q(\boldsymbol{\eta}_l) - \mathbf{E}_q \boldsymbol{\mu})^T (\mathbf{E}_q(\boldsymbol{\eta}_l) - \mathbf{E}_q \boldsymbol{\mu}) \\
\mathbf{E}_q \log \frac{p(\boldsymbol{v}|\alpha) p(\alpha)}{q(\boldsymbol{v}) q(\alpha)} &= a \log B - \log \Gamma(a) - \sum_{i=1}^{L-1} \log \Gamma(\gamma_{i,1} + \gamma_{i,2}) + \sum_{i=1}^{L-1} \log \Gamma(\gamma_{i,1}) + \\
&\quad + \sum_{i=1}^{L-1} \log \Gamma(\gamma_{i,2}) - \sum_{i=1}^{L-1} (\gamma_{i,1} - 1) \mathbf{E}_q \log(v_i) - \sum_{i=1}^{L-1} \gamma_{i,2} \mathbf{E}_q \log(1 - v_i) + \\
&\quad + \log \Gamma(a + L - 1) - (a + L - 1) \log \left(B - \sum_{i=1}^{L-1} \mathbf{E}_q \log(1 - v_i) \right) \\
\mathbf{E}_q \log \frac{p(\boldsymbol{\mu})}{q(\boldsymbol{\mu})} &= 1 + \log \frac{\mathrm{var}_q(\boldsymbol{\mu})}{\sigma_0^2} - \frac{1}{2\sigma_0^2} \mathbf{E}_q \left(\boldsymbol{\mu}^T \boldsymbol{\mu} \right).
\end{aligned}$$

Under the factorized variational approximation to the posterior distribution, the predictive distribution is approximated by a product of expectations,

$$\begin{aligned}
p(\theta_{n+1}|\boldsymbol{\theta}) &= \int \left(\sum_{i=1}^{\infty} \pi_i(\boldsymbol{v}) p(\theta_{n+1}|\boldsymbol{\eta}_i) \right) dP(\boldsymbol{\nu}, \boldsymbol{\eta}|\boldsymbol{\theta}) && (35) \\
&\approx \sum_{i=1}^{L} \mathbf{E}_q[\pi_i(\boldsymbol{\nu})] \mathbf{E}_q[p(\theta_{n+1}|\boldsymbol{\eta}_i)] && (36) \\
&= \sum_{i=1}^{L} \mathbf{E}_q[\pi_i(\boldsymbol{\nu})] \mathrm{PN}_2\left(\theta_{n+1}; \mathbf{E}_q(\boldsymbol{\eta}_i), \left[1 + \mathrm{var}_q \boldsymbol{\eta}_{i,1}\right] \mathbf{I}_2\right), && (37)
\end{aligned}$$

where $\mathbf{E}_q[\pi_i(\boldsymbol{\nu})] = \mathbf{E}_q(v_i) \prod_{s=1}^{i-1} \mathbf{E}_q(1 - v_s)$, $\mathrm{var_q}\boldsymbol{\eta}_{\mathrm{i,1}}$ is the variance parameter in $q(\boldsymbol{\eta}_i)$ and

the expectation $\mathbf{E}_q\left[p\left(\theta_{n+1}|\boldsymbol{\eta}_i\right)\right]$ which can be computed as the integral $\int\left[\mathrm{PN}_2\left(\theta_{n+1};\boldsymbol{\eta}_i,\mathbf{I}_2\right)\right.$ $\left.\mathrm{N}_2\left(\boldsymbol{\eta}_i;\mathbf{E}_q\left(\boldsymbol{\eta}_i\right),\left[\mathrm{var}_q\boldsymbol{\eta}_{i,1}\right]\mathbf{I}_2\right)\right]d\boldsymbol{\eta}_i$ equals a projected normal distribution $\mathrm{PN}_2\left(\theta_{n+1};\mathbf{E}_q\left(\boldsymbol{\eta}_i\right),\right.$ $\left.\left[1+\mathrm{var}_q\boldsymbol{\eta}_{i,1}\right]\mathbf{I}_2\right)$. This last integral is computed in the same way as we did for the predictive distribution when using the Gibbs sampler (31).

The variational algorithm is sensitive to initial values for the case of mixtures of normal distributions (Blei and Jordan, 2006) and even more sensitive to initial values in our case of mixtures of projected normal distributions. We observed that when the "circular variance" (defined as $1-\bar{R}$ where $\bar{R}$ is the length of the average vector $\frac{1}{n}\sum_{i=1}^{n}\left(\cos\theta_i,\sin\theta_i\right)^T$) is moderate to large, the variational distribution that maximizes the lower bound gives predictive distributions that are unimodal even when it is clear the presence of multiple modes, which defeats the purpose of having a mixture model. In the next section, we explain in more detail this major issue as well as some other computation aspects we had to deal with, and we propose solutions for them.

4.5 Initialization, fitting and improvement of the variational distribution

We applied the re-ordering step of Kurihara et al. (2006) in our implementation of the iterative algorithm. As noted by these authors, the algorithm operates in a space where component labels are distinguishable, which means that if we permute the labels, the total probability of the data changes. Since the average a priori mixture weights of the components are ordered by their size, the optimal labeling of the a posteriori variational components is also ordered according to cluster size. This is incorporated in the algorithm as a re-ordering step of components according to approximate size after each iteration of the algorithm. When we implemented this, the lower bound indeed increased significantly compared to the leaving the components unordered.

As noted in Section 4.4, the algorithm is also very sensitive to initial values. Next we show some of the initial values we chose and a big issue with one of the fits. We initialize

the variational distribution by making a histogram with number of bins equal to L. We treat the data in each one of the bins as being a random sample of size n_i from a projected normal distribution $\text{PN}_2\left(\boldsymbol{\eta}_i^*, \mathbf{I}_2\right)$ and we estimate $\boldsymbol{\eta}_i^*$ using its posterior mean for $i = 1, \ldots, L$. Hernandez-Stumpfhauser et al. (2011) show how to get the posterior mean for a projected normal random sample via variational methods and we also provide an algorithm on how to do this in Supplement 4.8.1. We then initialize the variational parameters as follows: $\mathbf{E}\left(z_t^i\right) \propto n_i$ if θ_t is in bin i and $\mathbf{E}\left(z_t^i\right) \propto m_t$ if θ_t is not in bin i for some number m_t. We set the means of the $\boldsymbol{\eta}s$ equal to the posterior means $\mathbf{E}\left(\boldsymbol{\eta}_i\right) \leftarrow \mathbf{E}\left(\boldsymbol{\eta}_i^* | \text{data in bin } i\right)$ and we let $\mathbf{E}\left(\alpha\right) = 1$ and $\mathbf{E}\left(\boldsymbol{\mu}\right) = \mathbf{0}$.

We choose different initializations of the algorithm by varying the m_t values and select the variational distribution that gives us the highest lower bound, as is recommended in the literature (Ormerod and Wand, 2010). A big issue arises when the circular variance of the data has a moderate to large value. The unimodal fit, obtained by initializing the algorithm with all the $\mathbf{E}\left(z_t^i\right)$ equal to each other and all the $\mathbf{E}\left(\boldsymbol{\eta}_i\right)$ equal to each other, seems to have the highest value of the lower bound even when it is clear from the histograms that the data exhibit multiple modes. To show this issue, in Figure 11, we generate data from a mixture of three projected normal distributions for three different values of the mean components $\boldsymbol{\mu}_1 = \rho\left(0, 1.5\right)^T, \boldsymbol{\mu}_2 = \rho\left(0, -1.7\right)^T, \boldsymbol{\mu}_3 = \rho\left(-2, 0\right)^T$ with proportions $\pi_1 = 0.5, \pi_2 = 0.2, \pi_3 = 0.3$ respectively. From left to right we changed the value of $\rho = 2, 1.5, 1$. The truncation level L was set constant for all fits at $L = 13$. We see that when $\rho = 2$ (small circular variance) the log lower bound for the multimodal fit (-344.14) is greater than the one for the unimodal fit (-432.49). When $\rho = 1.5$, the multimodal fit continues to dominate, but the difference in the log lower bounds is reduced (-408.78 vs. -439.69). At the other extreme, the log lower bound for the multimodal fit (-467.01) when $\rho = 1$ (larger circular variance) is less than the one for the unimodal fit (-451.06), even though we can clearly see the presence of 3 modes.

This problem is caused by the relatively poor approximation to the true posterior of the variational distribution. We investigated a number of approaches to improve the variational

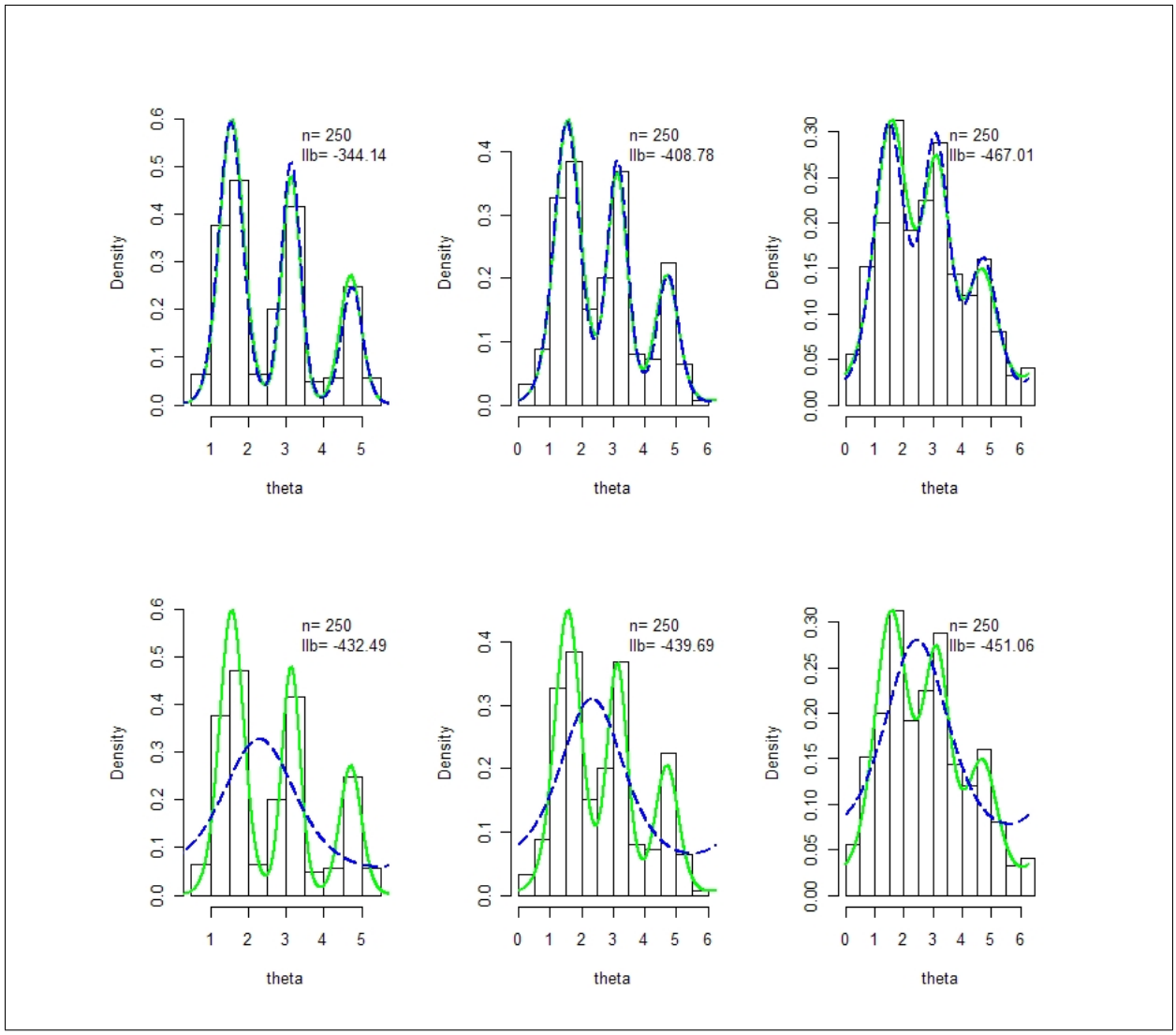

Figure 11: Histograms of simulated data and variational approximations to the predictive density for 3 different mixture models. Each data set is generated from a different mixture of 3 projected normal distributions with mean components $\boldsymbol{\mu}_1 = \rho(0, 1.5)^T, \boldsymbol{\mu}_2 = \rho(0, -1.7)^T, \boldsymbol{\mu}_3 = \rho(-2, 0)^T$ with proportions $\pi_1 = 0.5, \pi_2 = 0.2, \pi_3 = 0.3$ respectively. From left to right we changed the value of $\rho = 2, 1.5, 1$. Top and bottom rows display multimodal and unimodal fits respectively due to different initial values. Solid lines are the true densities and dashed lines are the predictive distributions via the variational approximation.

distribution, starting with alternative factorizations. This did not lead to tractable solutions, however, and was not further pursued. Another approach we investigated took advantage of fact that the predictive distributions (37) do not depend on $\boldsymbol{z}$ nor depend on $\mathbf{r}$. For this reason, after running Algorithm 3 we estimated marginal lower bounds $\log \underline{p}(\boldsymbol{\theta}, q) =$

$\mathbf{E}_q \log \frac{p(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{v})}{q(\boldsymbol{\eta},\boldsymbol{v})}$ via Montecarlo methods. We noticed that this approach led us to choose the multi-modal predictive distribution over the unimodal one in some cases but not all. We finally solved this poor approximation issue by using a set of sequential improvements that can be applied to arbitrary variational distributions. These improvements are derived in the following three results, which appear to be new.

Result 7. *Assume y and $(a,b)^T$ are continuous random vectors and $p(y,a,b)$ its joint density function. Let $q_0(a,b)$ be an arbitrary density function over $\{(a,b)\}$ and let $q_1(a,b) = p(a|y,b)\, q_0(b)$. Then, $\log \underline{p}(y;q_1) \geq \log \underline{p}(y;q_0)$.*

$$
\begin{aligned}
\log \underline{p}(y;q_1) &= \mathbf{E}_{q_1(a,b)} \log \frac{p(y,a,b)}{q_1(a,b)} \\
&= \mathbf{E}_{q_1(a,b)} \log \frac{p(y,a|b)\, p(b)}{p(a|y,b)\, q_0(b)} \\
&= \mathbf{E}_{q_1(a,b)} \left[\log \frac{p(y,a|b)}{p(a|y,b)} + \log \frac{p(b)}{q_0(b)} \right] \\
&= \mathbf{E}_{q_1(a,b)} \log \frac{p(y,a|b)}{p(a|y,b)} + \mathbf{E}_{q_1(a,b)} \log \frac{p(b)}{q_0(b)} \\
&= \int \left[\int \log \frac{p(y,a|b)}{p(a|y,b)} p(a|y,b)\, da \right] q_0(b)\, db + \mathbf{E}_{q_1(b)} \log \frac{p(b)}{q_0(b)} \\
&\geq \int \left[\int \log \frac{p(y,a|b)}{q_0(a|b)} q_0(a|b)\, da \right] q_0(b)\, db + \mathbf{E}_{q_0(b)} \log \frac{p(b)}{q_0(b)} \\
&= \mathbf{E}_{q_0(a,b)} \log \frac{p(y,a|b)}{q_0(a,b)} + \mathbf{E}_{q_0(a,b)} \log \frac{p(b)}{q_0(b)} \\
&= \mathbf{E}_{q_0(a,b)} \log \frac{p(y,a,b)}{q_0(a,b)} \\
&= \log \underline{p}(y;q_0)
\end{aligned}
$$

Result 8. *Assume y and $(a,b)^T$ are continuous random vectors and $p(y,a,b)$ its joint density function and let $q_0(a,b)$ be an arbitrary density function over $\{(a,b)\}$. Then, $\mathbf{E}_{q(a,b)} \log \frac{p(y,b)}{q(b)} \geq \mathbf{E}_{q(a,b)} \log \frac{p(y,a,b)}{q(a,b)}$.*

Let $q_1(a,b) = p(a|y,b)\,q(b)$. Using Result 7 we have that

$$\begin{aligned}
\mathbf{E}_{q(a,b)} \log \frac{p(y,a,b)}{q(a,b)} &\leq \mathbf{E}_{q_1(a,b)} \log \frac{p(y,a,b)}{q_1(a,b)} \\
&= \mathbf{E}_{q_1(a,b)} \log \frac{p(a|y,b)\,p(y,b)}{p(a|y,b)\,q(b)} \\
&= \mathbf{E}_{q_1(a,b)} \log \frac{p(y,b)}{q(b)} \\
&= \mathbf{E}_{q(a,b)} \log \frac{p(y,b)}{q(b)}
\end{aligned}$$

Result 9. *Assume y and $(a,b)^T$ are continuous random vectors and $p(y,a,b)$ its joint density function. Let $q_0(a,b)$ be an arbitrary density function. Let $q_1(a,b) = p(a|y,b)\,q_0(b)$ and let $q_2(a,b) = p(b|y,a)\,q_1(a)$. Then, $\log \underline{p}(y;q_2) \geq \log \underline{p}(y;q_1)$.*

$$\begin{aligned}
\log \underline{p}(y;q_2) &= \mathbf{E}_{q_2(a,b)} \log \frac{p(y,a,b)}{q_2(a,b)} \\
&= \mathbf{E}_{q_2(a,b)} \log \frac{p(b|y,a)\,p(y,a)}{p(b|y,a)\,q_1(a)} \\
&= \mathbf{E}_{q_2(a,b)} \log \frac{p(y,a)}{q_1(a)} \\
&= \int\int \log \frac{p(y,a)}{q_1(a)} p(b|y,a)\,q_1(a)\,dbda \\
&= \int \log \frac{p(y,a)}{q_1(a)} q_1(a)\,da \\
&= \int\int \log \frac{p(y,a)}{q_1(a)} q_1(b|a)\,q_1(a)\,dbda \\
&= \int\int \log \frac{p(y,a)}{q_1(a)} q_1(a,b)\,dbda \\
&= \mathbf{E}_{q_1(a,b)} \log \frac{p(y,a)}{q_1(a)} \\
&\geq \mathbf{E}_{q_1(a,b)} \log \frac{p(y,a,b)}{q_1(a,b)}
\end{aligned}$$

The last inequality follows from Result 8.

Our first improvement is based on Results 7 and 8. Letting $q_0(a,b)$ as the variational

fit with $b = \{\boldsymbol{z}, \boldsymbol{\mu}, \alpha\}$ and $a = \{\boldsymbol{v}, \boldsymbol{\eta}\}$, we consider the improvement $q_1(\boldsymbol{v}, \boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{\mu}, \alpha) = p(\boldsymbol{v}|\boldsymbol{z}, \alpha)\, p(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\mu})\, q_0(\boldsymbol{z}, \boldsymbol{\mu}, \alpha)$ based on Result 7. The distribution $p(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\mu})$ is not available in closed form but it can be very well approximated by a Laplace approximation (see Hernandez-Stumpfhauser et al. 2011). Notice that we could have included the latent lengths $\mathbf{r}$ in the vector b and there would have been no need of the Laplace approximation but by marginalizing over the latent lengths we improve the lower bound (Result 8). The approach is then to get a sample of size N from $q_1(\boldsymbol{v}, \boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{\mu}, \alpha)$. We do this as follows:

- Obtain $q(\boldsymbol{\mu})\, q(\alpha) \prod_{t=1}^{n} q(z_t)$ by making use of Algorithm 3
- Get a sample of size N from $q_0(\boldsymbol{z}, \boldsymbol{\mu}, \alpha) = q(\boldsymbol{\mu})\, q(\alpha) \prod_{t=1}^{n} q(z_t)$
- For each draw of q_0 we get a draw from

$$\begin{aligned} p(\boldsymbol{v}|\boldsymbol{z}, \alpha) &= \prod_{i=1}^{L-1} \text{Beta}\left(v_i; \sum_{t=1}^{n} z_t^i + 1, \sum_{t=1}^{n}\sum_{j=1}^{L} z_t^{i+j} + \alpha\right) \\ p(v_L = 1) &= 1 \\ p(\boldsymbol{\eta}|\boldsymbol{z}, \boldsymbol{\mu}) &= \prod_{i=1}^{L} \text{N}_2(\boldsymbol{\eta}_i; \mathbf{m}_i, \boldsymbol{W}_i), \end{aligned} \tag{38}$$

  where the $\mathbf{m}_i$ and the $\boldsymbol{W}_i$ are 2 by 1 vectors and 2 by 2 matrices that depend on $\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\mu}$ respectively.

Algorithms to obtain the $\mathbf{m}_i$ vectors and the W_i matrices are given in Supplement 4.8.2. After obtaining a sample of size N from q_1, the predictive distribution is estimated using the approximation

$$p(\theta_{n+1}|\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{s=1}^{N} \sum_{i=1}^{L} \pi_i\left(\boldsymbol{\nu}^{(s)}\right) \text{PN}_2\left(\theta_{n+1}; \boldsymbol{\eta}_i^{(s)}, \mathbf{I}_2\right)$$

where $\boldsymbol{\nu}^{(s)}, \boldsymbol{\eta}_i^{(s)}$ denote the *sth* draw from q_1.

The log lower bound $\log \underline{p}(\boldsymbol{\theta}, q_1) = \mathbf{E}_{q_1} \log \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z}, \alpha, \boldsymbol{\mu})}{q_1(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z}, \alpha, \boldsymbol{\mu})}$ is no longer available in closed

form but can be readily estimated via Monte Carlo methods using

$$\mathbf{E}_{q_1} \log \frac{p\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z}, \alpha, \boldsymbol{\mu}\right)}{q_1\left(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{z}, \alpha, \boldsymbol{\mu}\right)} \approx \frac{1}{N} \sum_{s=1}^{N} \log \frac{p\left(\boldsymbol{\theta}, \boldsymbol{\eta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{z}^{(s)}, \alpha^{(s)}, \boldsymbol{\mu}^{(s)}\right)}{q_1\left(\boldsymbol{\eta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{z}^{(s)}, \alpha^{(s)}, \boldsymbol{\mu}^{(s)}\right)}, \tag{39}$$

where $p\left(\boldsymbol{\theta}, \boldsymbol{\eta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{z}^{(s)}, \alpha^{(s)}, \boldsymbol{\mu}^{(s)}\right) = p\left(\boldsymbol{\theta}|\boldsymbol{\eta}^{(s)}, \boldsymbol{z}^{(s)}\right) p\left(\boldsymbol{z}^{(s)}|\boldsymbol{v}^{(s)}\right) p\left(\boldsymbol{\eta}^{(s)}|\boldsymbol{\mu}^{(s)}\right) p\left(\boldsymbol{v}^{(s)}|\alpha^{(s)}\right)$ $p\left(\alpha^{(s)}\right) p\left(\boldsymbol{\mu}^{(s)}\right)$ and $q_1\left(\boldsymbol{\eta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{z}^{(s)}, \alpha^{(s)}, \boldsymbol{\mu}^{(s)}\right) = p\left(\boldsymbol{v}^{(s)}|\boldsymbol{z}^{(s)}, \alpha^{(s)}\right) p\left(\boldsymbol{\eta}^{(s)}|\boldsymbol{\theta}, \boldsymbol{z}^{(s)}, \boldsymbol{\mu}^{(s)}\right)$ $q_0\left(\boldsymbol{z}^{(s)}\right) q_0\left(\boldsymbol{\mu}^{(s)}\right) q_0\left(\alpha^{(s)}\right)$. All those distributions are known in closed form.

Our second improvement is based on Result 9. Here we make $b = \{\boldsymbol{v}, \boldsymbol{\eta}, \mathbf{r}\}$ and $a = \{\boldsymbol{z}, \boldsymbol{\mu}, \alpha\}$ and we again take $q_0(a, b)$ as our variational fit. We then get a sample of size N from $q_2(a, b)$ as follows:

- Obtain $q_0\left(\boldsymbol{v}^{'}, \boldsymbol{\eta}^{'}, \mathbf{r}^{'}\right) = \prod_{i=1}^{L-1} q\left(v_i^{'}\right) \prod_{i=1}^{L} q\left(\boldsymbol{\eta}_i^{'}\right) \prod_{t=1}^{n} q\left(r_t^{'}\right)$ by making use of Algorithm 3
- Get a sample of size N from $q_0\left(\boldsymbol{v}^{'}, \boldsymbol{\eta}^{'}, \mathbf{r}^{'}\right)$
- For each sth draw $\left\{\boldsymbol{v}^{'}(s), \boldsymbol{\eta}^{'}(s), \mathbf{r}^{'}(s)\right\}$ get a sample of size N from $q_1(\boldsymbol{z}, \boldsymbol{\mu}, \alpha)$ by drawing from $p\left(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{v}^{'}(s), \boldsymbol{\eta}^{'}(s), \mathbf{r}^{'}(s)\right) p\left(\alpha|\boldsymbol{v}^{'}(s)\right) p\left(\boldsymbol{\mu}|\boldsymbol{\eta}^{'}(s)\right)$
- For each sth draw $\{\boldsymbol{z}(s), \boldsymbol{\mu}(s), \alpha(s)\}$ get a sample of size N from $q_2(\boldsymbol{v}, \boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{\mu}, \alpha)$ by drawing from

$$p\left(\boldsymbol{v}|\boldsymbol{z}(s), \alpha(s)\right) p\left(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{z}(s), \boldsymbol{\mu}(s)\right). \tag{40}$$

We again use the Laplace approximation to draw from $p\left(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{z}(s), \boldsymbol{\mu}(s)\right)$. Sampling from $q_0\left(\boldsymbol{v}^{'}, \boldsymbol{\eta}^{'}, \mathbf{r}^{'}\right)$ requires sampling from $q_0\left(\mathbf{r}^{'}\right)$ directly so we cannot use the sampling scheme in equation (30). Sampling directly from $q_0\left(\mathbf{r}^{'}\right)$ outside the Gibbs sampler can be done using rejection methods such as Metropolis-Hastings or adaptive rejection sampling. In Supplement 4.8.3, we show a simple and efficient way to sample approximately from $q_0\left(\mathbf{r}^{'}\right)$ by making use of the inverse cdf technique and finding such inverse via the Newton-Rhapson method.

To estimate the predictive distribution we do the same as what we did for the first improvement, equation (39). Notice that we no longer have q_1 in closed form which makes it a bit more difficult to estimate the log lower bound. We write

$$\begin{aligned}\mathbf{E}_{q_2}\log\frac{p\left(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{z},\alpha,\boldsymbol{\mu}\right)}{q_2\left(\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{z},\alpha,\boldsymbol{\mu}\right)} &= \mathbf{E}_{q_2}\log\frac{p\left(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{z},\alpha,\boldsymbol{\mu}\right)}{p\left(\boldsymbol{v}|\boldsymbol{z},\alpha\right)p\left(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\mu}\right)q_1\left(\boldsymbol{z},\boldsymbol{\mu},\alpha\right)} \\ &= \mathbf{E}_{q_2}\log\frac{p\left(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{z},\alpha,\boldsymbol{\mu}\right)}{p\left(\boldsymbol{v}|\boldsymbol{z},\alpha\right)p\left(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\mu}\right)} - \mathbf{E}_{q_1}\log q_1\left(\boldsymbol{z},\boldsymbol{\mu},\alpha\right)\end{aligned}$$

and to estimate the second term we approximate by

$$\begin{aligned}q_1\left(\boldsymbol{z},\boldsymbol{\mu},\alpha\right) &= \int p\left(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{v}^{'},\boldsymbol{\eta}^{'},\mathbf{r}^{'}\right)p\left(\alpha|\boldsymbol{v}^{'}\right)p\left(\boldsymbol{\mu}|\boldsymbol{\eta}^{'}\right)q_0\left(\boldsymbol{v}^{'},\boldsymbol{\eta}^{'},\mathbf{r}^{'}\right)d\boldsymbol{v}^{'}d\boldsymbol{\eta}^{'}d\mathbf{r}^{'} \\ &= \mathbf{E}_{q_0}p\left(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{v}^{'},\boldsymbol{\eta}^{'},\mathbf{r}^{'}\right)p\left(\alpha|\boldsymbol{v}^{'}\right)p\left(\boldsymbol{\mu}|\boldsymbol{\eta}^{'}\right) \\ &\approx \frac{1}{N_1}\sum_{s=1}^{N_1}p\left(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{v}^{'}\left(s\right),\boldsymbol{\eta}^{'}\left(s\right),\mathbf{r}^{'}\left(s\right)\right)p\left(\alpha|\boldsymbol{v}^{'}\left(s\right)\right)p\left(\boldsymbol{\mu}|\boldsymbol{\eta}^{'}\left(s\right)\right).\end{aligned}$$

Hence, the second term is approximated by

$$\begin{aligned}&\mathbf{E}_{q_1}\log q_1\left(\boldsymbol{z},\boldsymbol{\mu},\alpha\right) \\ &\approx \frac{1}{N_2}\sum_{l=1}^{N_2}\log\frac{1}{N_1}\sum_{s=1}^{N_1}p\left(\boldsymbol{z}\left(l\right)|\boldsymbol{\theta},\boldsymbol{v}^{'}\left(s\right),\boldsymbol{\eta}^{'}\left(s\right),\mathbf{r}^{'}\left(s\right)\right)p\left(\alpha\left(l\right)|\boldsymbol{v}^{'}\left(s\right)\right)p\left(\boldsymbol{\mu}\left(l\right)|\boldsymbol{\eta}^{'}\left(s\right)\right).\end{aligned}$$

To show the effect of the improvements, we consider the simulated data in Figure 11 for $\rho = 1$ and the unimodal and multimodal variational distributions that were obtained by choosing different initial values for the variational algorithm. The first column in Figure 12 shows how the log lower bound increases from -467.01 to -458.14 for the multimodal fit and from -451.06 to -451.02 for the unimodal fit after making use of (38) but still not enough to choose the multimodal fit over the unimodal fit. The second column shows how the log lower bound increases again from -458.14 to -445.67 for the multimodal fit and from -451.02 to -450.66 for the unimodal fit after making use of (40) and this second improvement is enough to choose the multimodal fit over the unimodal fit. Figure 12 also displays $10\%, 20\%, \ldots, 90\%$

point-wise credible intervals.

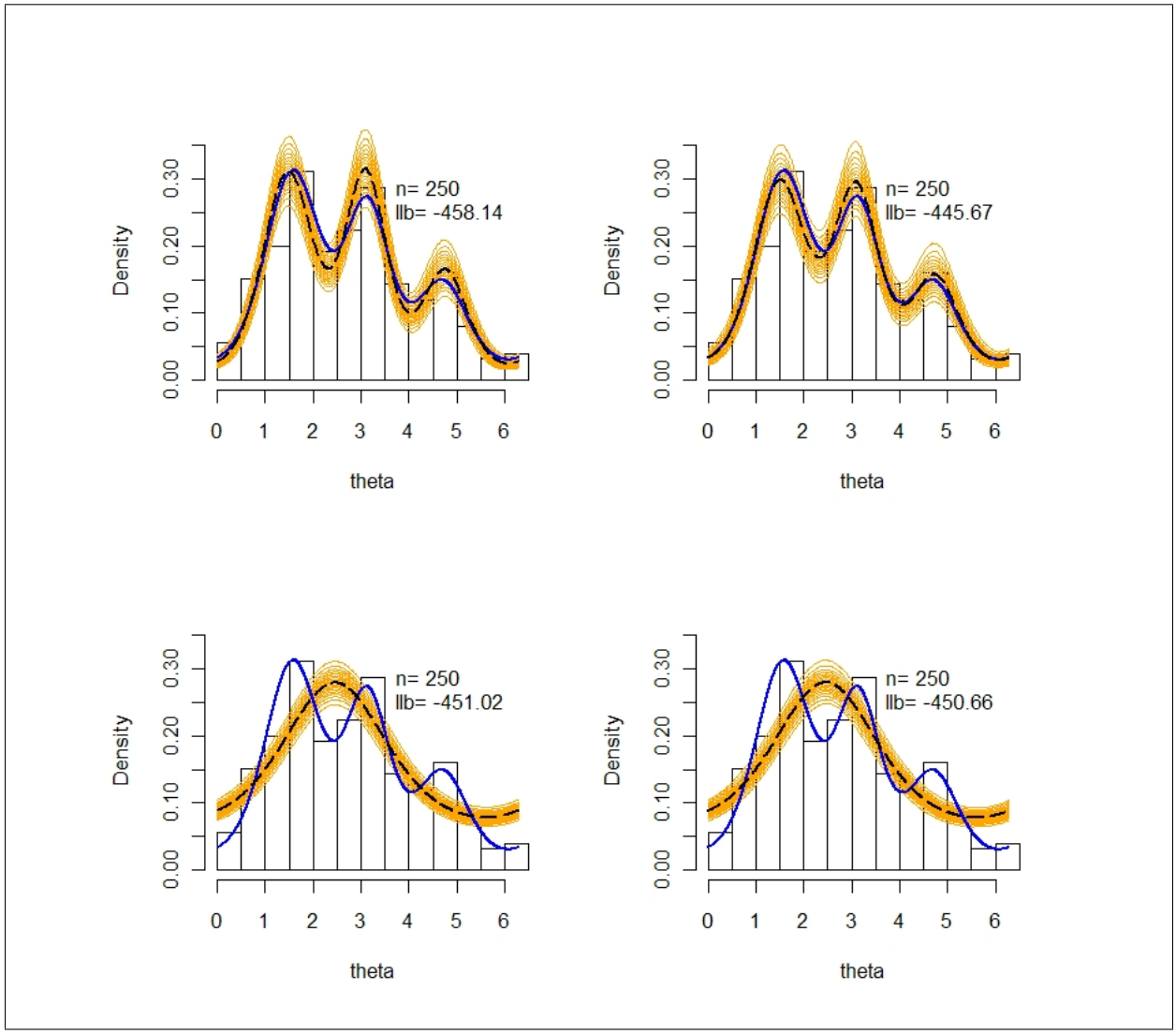

Figure 12: Improvements to the variational distribution in Figure 11. The data set is the same one as in Figure 11 with $\rho = 1$, (the third column in Figure 11). The first column corresponds to predictive distributions after applying (38) to both variational distributions in the third column of Figure 11. The second column displays predictive distributions after applying (40) to the same distributions. Solid lines are the true density while dashed lines are the predictive distributions obtained through the use of the variational approximation. Lines above and below the predictive distribution are $10\%, 20\%, \ldots, 90\%$ point-wise credible intervals.

Result 9 is a direct consequence of Result 7. While Result 7 is quite simple, it allows us to improve any approximation, including the approximations we have just proposed.

Notice that the Laplace approximation is not a necessary step. We could proceed as follows. Improve the variational distribution q_0 by using Result 7, i.e. $q_1(a,b) = p(a|\boldsymbol{\theta},b)\,q_0(b)$ with $b = \{\boldsymbol{z},\mathbf{r},\boldsymbol{\mu},\alpha\}$, $a = \{\boldsymbol{v},\boldsymbol{\eta}\}$. Then use Result 7 again to improve the approximation by taking $b = \{\boldsymbol{v},\boldsymbol{\eta},\mathbf{z},\boldsymbol{\mu},\alpha\}$ and $a = \{\mathbf{r}\}$ with the approximation being $q_2(a,b) = p(a|\boldsymbol{\theta},b)\,q_1(b)$. Then improve this last one by taking $b = \{\boldsymbol{v},\boldsymbol{\eta},\mathbf{r}\}$, $a = \{\mathbf{z},\boldsymbol{\mu},\alpha\}$ and so on. As long as all parameters are part of the a step, all the distributions of the parameters will be improved. In fact, we could improve one parameter at a time by having in each step only one parameter in a. Notice that this looks very much like a Gibbs sampler, with one of the differences being that a Gibbs sampler starts from some starting values while we start from a distribution. Another difference is that in a Gibbs sampler we collect a sample of size N after convergence of the chain while here at each step we get a sample of size N from an approximation which is improved at each step. If a model is such that it is known that the variational distribution is a good approximation then there is no need to use this sampling scheme. On the other hand, if the model is such that it is suspected that the variational distribution is a poor approximation, like in our case, then this sampling scheme could fix those issues.

4.6 Application

4.6.1 Background

The CHTS data contain over a million reported angling trips for 17 states, 6 waves (two-month time periods), and 4 modes of fishing (shore, private boat, charter boat, party boat). As noted in Section 1, our goal here is to get density estimates of departure times by state, wave and mode that can be used to obtain weights for the APAIS. In many state-wave-mode domains, the sample sizes are small (or even 0) and in many others the sample sizes are very big (up to 42000). In this section, we will present two types of Dirichlet Process mixture models that borrow strength across domains and apply those to the CHTS data. The subsections 4.6.2 and 4.6.3 will discuss the variational approximation to each one of the two models. For comparison purposes, we will also include computations needed for

the Gibbs sampler in subsection 4.6.2 and fit the data in the state of New Hampshire. The sample size for the state of New Hampshire is small and allows us to compare the Gibbs sampler to the variational approximations to both models. Finally in the last subsection we show results from applying the variational approximations to the whole data set (over a million observations).

4.6.2 Multiple populations model

The first model we present has a random distribution G_{ijk} for each combination of state, wave and mode. Each one of these random distributions is distributed as a Dirichlet process with a Projected Normal base distribution with mean of the form $\mathbf{m}_{ijk} = \mathbf{m}_0 + \mathbf{s}_i + \mathbf{w}_j + \mathbf{m}_k$. Here $\mathbf{m}_0, \mathbf{s}_i, \mathbf{w}_j, \mathbf{m}_k$ represent the overall, state, wave, and mode effects respectively. The hierarchical model is as follows:

$$\begin{aligned}
\theta_{ijkt}|\boldsymbol{\mu}_{ijkt} &\sim \mathrm{PN}_2\left(\boldsymbol{\mu}_{ijkt}, \mathbf{I}_2\right) \\
\boldsymbol{\mu}_{ijkt}|G_{ijk} &\sim G_{ijk} \\
G_{ijk} &\sim \mathrm{DP}\left(\alpha_0, G_0\right) \\
G_0\left(\cdot\right) &= \mathrm{N}_2\left(\cdot; \mathbf{m}_{ijk}, \mathbf{I}_2\right) \\
\mathbf{m}_{ijk} &= \mathbf{m}_0 + \mathbf{s}_i + \mathbf{w}_j + \mathbf{m}_k \\
\alpha_0 &\sim \mathrm{Gamma}\left(a, b\right)
\end{aligned}$$

where the effects $\mathbf{s}_i, \mathbf{w}_j, \mathbf{m}_k$ and the overall mean $\mathbf{m}_0$ are all distributed as bivariate normals with their own means and variances.

Next, we will show the full conditionals necessary for the Gibbs sampler in this model based on the Polya urn representation of the Dirichlet process. Conditioned on $\mathbf{m}_{ijk}$ and α_0,

the full conditionals are the same as in algorithm 1, or

$$\begin{aligned}\left(\boldsymbol{\mu}_{ijkt}|\boldsymbol{\mu}_{ijk}^{(t)},\mathbf{r},\mathbf{m}_{ijk},\alpha,\boldsymbol{\theta}\right) &\sim q_0 G_t\left(\boldsymbol{\mu}_{ijkt}\right)+\sum_{s=1,s\neq t}^{n_{ijk}} q_s\delta_{\boldsymbol{\mu}_{ijks}}\left(\boldsymbol{\mu}_{ijkt}\right)\\ p\left(r_{ijkt}|\boldsymbol{\mu}_{ijkt},\boldsymbol{\theta}\right) &\propto r_{ijkt}\exp\left(-\frac{1}{2}r_{ijkt}^2+\mathbf{u}_{ijkt}^T\boldsymbol{\mu}_{ijkt}r_{ijkt}\right),\end{aligned}$$

where n_{ijk} is the number of observations in cell ijk and as before, we define $\boldsymbol{\mu}_{ijk}^{(t)}=\{\boldsymbol{\mu}_{ijk,1},\ldots,\boldsymbol{\mu}_{ijk,t-1},\boldsymbol{\mu}_{ijk,t+1}\ldots,\boldsymbol{\mu}_{ijkn_{ijk}}\}$, $G_t\left(\boldsymbol{\mu}_{ijkt}\right)=\mathrm{N}_2\left(\frac{1}{2}\left(\mathbf{x}_{ijkt}+\mathbf{m}_{ijk}\right),\frac{1}{2}\mathbf{I}_2\right)$, $q_0\propto\alpha\mathrm{N}_2\left(\mathbf{x}_{ijkt};\mathbf{m}_{ijk},2\mathbf{I}_2\right)$ and $q_s\propto\mathrm{N}_2\left(\mathbf{x}_{ijkt};\boldsymbol{\mu}_{ijks},\mathbf{I}_2\right)$ and $\mathbf{x}_{ijkt}=r_{ijkt}\mathbf{u}_{ijkt}$. Thus with probability proportional to q_0, $\left(\boldsymbol{\mu}_{ijkt}|\boldsymbol{\mu}_{ijk}^{(t)},\mathbf{r},\mathbf{m}_{ijk},\alpha,\boldsymbol{\theta}\right)$ is a draw from G_t and with probability proportional to q_s is equal to $\boldsymbol{\mu}_{ijks}$ for $s=1,\ldots,t-1,t+1,\ldots,n_{ijk}$. Again as in Algorithm III, the full conditionals for the parameters in the base distribution depend only on the distinct values $\boldsymbol{\mu}_{ijk,1}^*,\ldots,\boldsymbol{\mu}_{ijk,l}^*$ of $\boldsymbol{\mu}_{ijk,1},\ldots,\boldsymbol{\mu}_{ijkn_{ijk}}$. The prior for the parameters in the base distribution are taken as bivariate normal, and under these priors it is straightforward to obtain the full conditionals for the parameters in the base distribution.

Slow mixing of the Gibbs sampler for normal linear models is common and several remedies exist that involve reparameterizations of the parameters. As with the model in Hernandez-Stumpfhauser et al. (2011), which uses the same data as in this paper, the chain fails to converge due to a slow mixing problem. We use the sweeping reparameterization proposed by Vines et al. (1996) again:

$$\begin{aligned}\mathbf{m}_0' &= \mathbf{m}_0+\overline{\mathbf{m}}+\overline{\mathbf{s}}+\overline{\mathbf{w}}\\ \mathbf{m}_k' &= \mathbf{m}_k-\overline{\mathbf{m}}\\ \mathbf{s}_i' &= \mathbf{s}_i-\overline{\mathbf{s}}\\ \mathbf{w}_j' &= \mathbf{w}_j-\overline{\mathbf{w}}\\ \mathbf{m}_{ijk} &= \mathbf{m}_0'+\mathbf{m}_k'+\mathbf{s}_i'+\mathbf{w}_j'.\end{aligned}$$

Under this reparameterization, the random effects are no longer priorly independent since

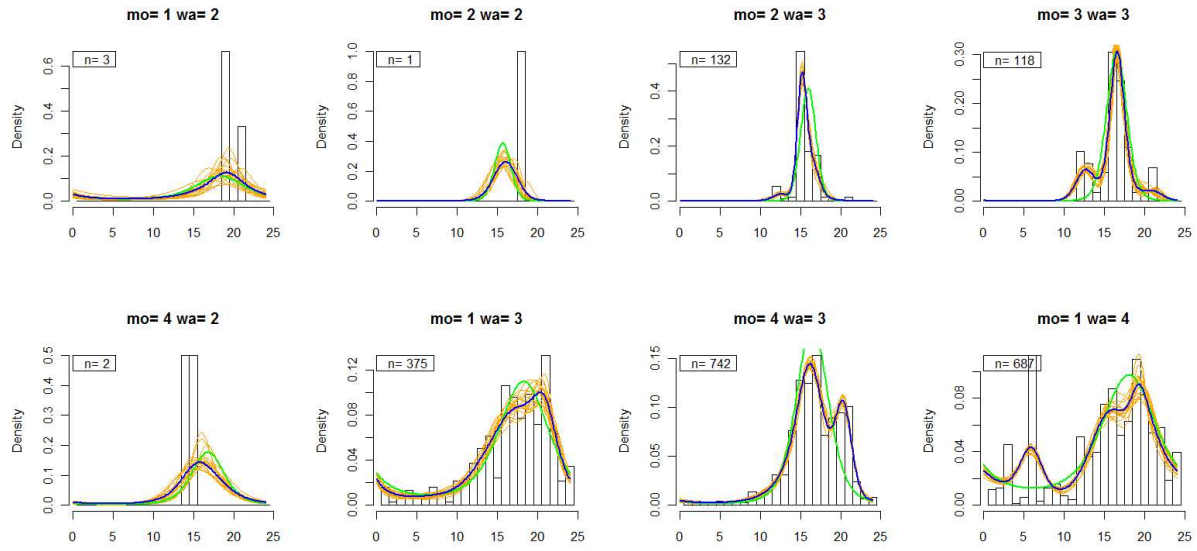

Figure 13: Results for the model in Section 4.6.2 for various waves and modes in the state of New Hampshire making use of the Gibbs sampler. The dark curve in each panel is the predictive distribution. Light curves are 20 Gibbs iterates sampled at random from all iterates, to show variability in the posterior distribution. The light unimodal curves are the posterior mean estimates of the mode effect.

they have to add to zero, for example $\sum_k \mathbf{m}'_k = 0$. The full conditionals for the parameters in the base distribution under the sweeping reparameterization are shown in Supplement 4.8.4.

To sample the precision parameter inside the Gibbs sampler we will use the method of Escobar and West (1995). In this model the full conditional for α may be expressed as a mixture of gamma distributions. The total number of mixture components equals the number of state-wave-mode combinations + 1. Derivations of the full conditional for the precision parameter are shown in Supplement 4.8.5.

Figures 13 and 14 show the predictive distributions under a mode effect model ($\mathbf{m}_{ijk} = \mathbf{m}_0 + \mathbf{m}_k$) for some of the combinations of mode and wave for the state of New Hampshire. The Gibbs sampler does a good job of fitting and adding multiple modes when data are rich, and reverting to the overall mode effect when data are sparse, as in Figure 13. Figure 14 displays the mode effect estimates and 20 curves from the sampler to show the

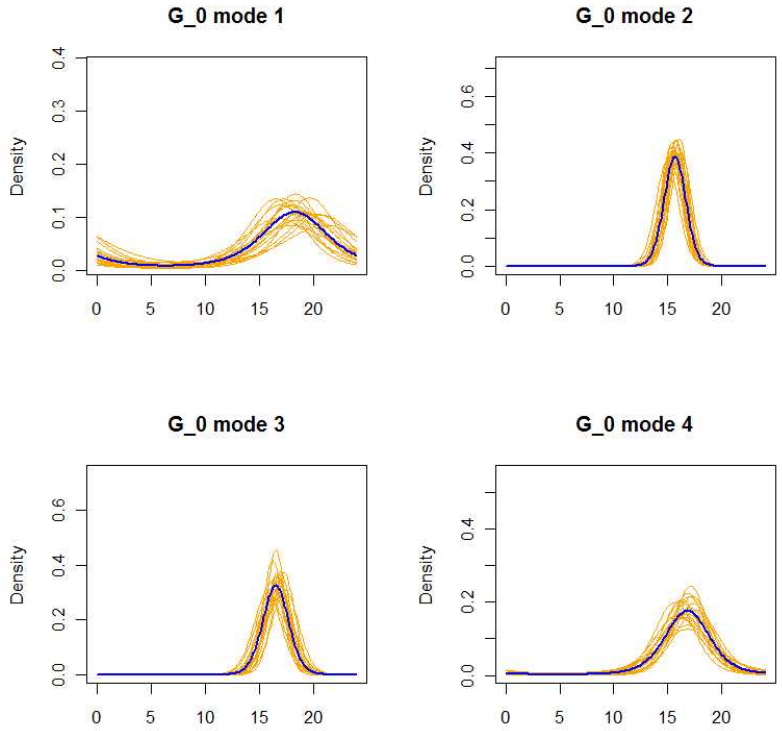

Figure 14: Results for the model in Section 4.6.2 for mode effects for the state of New Hampshire using the Gibbs sampler. The dark curve in each panel is the posterior mean estimate of the mode effect, computed by averaging all Gibbs iterates. Light curves are 20 Gibbs iterates sampled at random from all iterates, to show variability in the posterior distribution. Departures by anglers fishing from shore (mode=1) are more diffuse across time than departures by anglers fishing from boats (modes 2,3,4).

variability. Departures by anglers fishing from shore (mode=1) are more diffuse across time than departures by anglers fishing from boats (modes 2,3,4). The only disadvantage of the Gibbs sampler is its computational expense. To be able to use all of the CHTS data (all 17 states together), we make use of the variational approximation to the posterior distribution based on the stick-breaking representation of the process. We will first use the data from New Hampshire to compare the approximation to the Gibbs sampler and then in the last subsection we will use the full CHTS data.

The approximation is based on the stick-breaking representation of the Dirichlet process. For this model, the variational distributions look the same as before, equation (34), with the

difference that now we have a regression model for the parameters in the base distribution. The variational distributions for the parameters in the base distribution again have the undesirable property of having variances that can be made arbitrarily small by increasing the truncation levels on the number of mixture components. In this model there is a random distribution G_{ijk} for each combination of i, j, k and hence a potentially different truncation level L_{ijk} for each cell. The variational distributions are as follows:

$$
\begin{aligned}
q\left(v_{ijkl}\right) &= \text{Beta}\left(\sum_{t=1}^{n_{ijk}} \mathbf{E}\left(z_{ijkt}^{l}\right)+1, \sum_{t=1}^{n_{ijk}} \sum_{s=l+1}^{L_{ijk}} \mathbf{E}\left(z_{ijkt}^{s}\right)+\mathbf{E}\left(\alpha\right)\right) \\
q\left(\boldsymbol{\mu}_{ijkl}\right) &= \text{N}_2\left(\text{Var}\left(\boldsymbol{\mu}_{ijkl}\right)\left[\mathbf{E}\left(\mathbf{m}_{ijk}\right)+\sum_{t=1}^{n_{ijk}} \mathbf{E}\left(z_{ijkt}^{l}\right)\mathbf{E}\left(r_{ijkt}\right)\mathbf{u}_{ijkt}\right],\right. \\
&\qquad \left.\text{Var}\left(\boldsymbol{\mu}_{ijkl}\right)\mathbf{I}_2\right) \\
q\left(z_{ijkt}^{l}=1\right) &\propto \exp\left\{\mathbf{E}\left(\log\left(v_{ijkl}\right)\right)+\sum_{s=1}^{l-1}\mathbf{E}\left(\log\left(1-v_{ijks}\right)\right)\right. \\
&\qquad \left.-\frac{1}{2}\mathbf{E}\left(\parallel \boldsymbol{\mu}_{ijkl} \parallel^2\right)+\mathbf{E}\left(r_{ijkt}\right)\mathbf{u}_{ijkt}^{T}\mathbf{E}\left(\boldsymbol{\mu}_{ijkl}\right)\right\} \\
q\left(r_{ijkt}\right) &\propto r_{ijkt}\exp\left(-\frac{1}{2}r_{ijkt}^{2}+r_{ijkt}\mathbf{u}^{T}\sum_{l=1}^{L_{ijk}}\mathbf{E}\left(\boldsymbol{\mu}_{ijkl}\right)\mathbf{E}\left(z_{ijkt}^{l}\right)\right) \\
q\left(\mathbf{m}_0\right) &= \text{N}_2\left(\text{Var}\left(\mathbf{m}_0\right)\left[\sum_{ijk}\sum_{l=1}^{L_{ijk}}\mathbf{E}\left(\boldsymbol{\mu}_{ijkl}\right)-\mathbf{E}\left(\mathbf{m}_k\right)-\mathbf{E}\left(\mathbf{s}_i\right)-\mathbf{E}\left(\mathbf{w}_j\right)\right],\right. \\
&\qquad \left.\text{Var}\left(\mathbf{m}_0\right)\mathbf{I}_2\right) \\
q\left(\mathbf{m}_k\right) &= \text{N}_2\left(\text{Var}\left(\mathbf{m}_k\right)\left[\sum_{ij}\sum_{l=1}^{L_{ijk}}\mathbf{E}\left(\boldsymbol{\mu}_{ijkl}\right)-\mathbf{E}\left(\mathbf{m}_0\right)-\mathbf{E}\left(\mathbf{s}_i\right)-\mathbf{E}\left(\mathbf{w}_j\right)\right],\right. \\
&\qquad \left.\text{Var}\left(\mathbf{m}_k\right)\mathbf{I}_2\right) \\
q\left(\alpha\right) &= \text{Gamma}\left(a+\sum_{ijk}\left(L_{ijk}-1\right), b-\sum_{ijk}\sum_{l=1}^{L_{ijk}-1}\mathbf{E}_q\log\left(1-v_{ijkl}\right)\right)
\end{aligned}
$$

for $l = 1, 2, \ldots, L_{ijk}$, $t = 1, 2, \ldots, n_{ijk}$ and where $\text{Var}\left(\boldsymbol{\mu}_{ijkl}\right) = \left[1 + \sum_{t=1}^{n_{ijk}} \mathbf{E}\left(z_{ijkt}^{l}\right)\right]^{-1}$, $\text{Var}\left(\mathbf{m}_0\right) = \left[\sum_{ijk} L_{ijk} + \frac{1}{\sigma_0^2}\right]^{-1}$ and $\text{Var}\left(\mathbf{m}_k\right) = \left[\sum_{ij} L_{ijk} + \frac{1}{\sigma_m^2}\right]^{-1}$. The variational distributions for the state and wave effects are similar to that of the mode effect. The prior means for the effects were taken as 0 and the prior variances are $\sigma_0^2, \sigma_m^2, \sigma_s^2, \sigma_w^2$ for the overall, mode, state and wave effects respectively. All expectations are with respect to the variational distributions. The variational algorithm then updates these expectations one at a time in a similar way as in Algorithm 3. In the next subsection we show an alternative model in which the regression coefficients are at the level of the mixtures.

4.6.3 Single Dirichlet process prior on regression coefficients

For this second model we use indicators $m_{t,k}$ equal to 1 if mode of departure t equals k and 0 otherwise for $k = 1, 2, 3$. In the same way we define indicators $w_{t,j}$ equal to 1 if wave of departure t equals j and 0 otherwise for $j = 1, 2, 3, 4, 5$. Finally we define $s_{t,i}$ as 1 if state of departure t equals i for $i = 1, 2, \ldots, 16$. Let $x_t = \left(1, m_{t,1}, \ldots, m_{t,3}, w_{t,1}, \ldots, w_{t,5}, s_{t,1}, \ldots, s_{t,16}\right)^T$ be a vector of size q and define $\boldsymbol{\beta}_t = \begin{pmatrix} \beta_{t,1,0} & \cdots & \beta_{t,1,q} \\ \beta_{t,2,0} & \cdots & \beta_{t,2,q} \end{pmatrix}^T$ as a $q \times 2$ matrix with each column denoting the random effects of intercept and covariates for subject t. Now suppose departure times come from the following model

$$\begin{aligned} \theta_t|\boldsymbol{\beta}_t &\sim \text{PN}_2\left(\boldsymbol{\beta}_t^T x_t, \mathbf{I}_2\right) \\ \boldsymbol{\beta}_t|G &\sim G \\ G &\sim \text{DP}\left(\alpha_0, G_0\right) \\ G_0\left(\beta_1, \beta_2\right) &= \text{N}_q\left(\beta_1; \boldsymbol{\xi}_1, \mathbf{I}_q\right) \text{N}_q\left(\beta_2; \boldsymbol{\xi}_2, \mathbf{I}_q\right) \end{aligned}$$

where the means $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are normally distributed. In this model there is only one random distribution G and the base distribution G_0 is normally distributed.

Next we show the variational approximation to this model. We will start by showing

its stick-breaking representation and then we will show the corresponding variational distributions. Let $\boldsymbol{\beta}_{i,c}$ be the c^{th} column of $\boldsymbol{\beta}_i$, $c = 1, 2$ and all other parameters defined as before.

$$\begin{aligned}
v_i|\alpha &\sim \text{Beta}\,(1, \alpha) \\
\boldsymbol{\beta}_{i,c} &\sim \text{N}\,(\boldsymbol{\xi}_c, \mathbf{I}) \\
\pi_i &= v_i \prod_{j=1}^{i-1} (1 - v_j) \\
z_t &\sim \text{Mult}\,(\pi) \\
y_t &\sim \text{N}_2\left(\boldsymbol{\beta}_{z_t}^T x_t, \mathbf{I}_2\right) \\
y_t &= r_t\,(\cos\theta_t, \sin\theta_t)\,,
\end{aligned}$$

where $i = 1, 2, \ldots$ and the data consist of the angles $\theta_1, \theta_2, \ldots, \theta_n$. The variational approximation makes use of a truncated stick-breaking Dirichlet process. Denoting the trunctation

level as L, the variational distributions take the following forms:

$$\begin{aligned}
q(v_i) &= \text{Beta}\left(\sum_{t=1}^{n}\mathbf{E}\left(z_t^i\right)+1, \sum_{t=1}^{n}\sum_{j=i+1}^{L}\mathbf{E}\left(z_t^j\right)+\mathbf{E}(\alpha)\right) \\
q\left(\boldsymbol{\beta}_{i,c}\right) &= \text{N}\left(\left[\sum_{t=1}^{n}\mathbf{E}\left(z_t^i\right)x_t x_t^T+\mathbf{I}\right]^{-1}\left[\sum_{t=1}^{n}\mathbf{E}\left(z_t^i\right)\mathbf{E}(r_t)\,u_{t,c}x_t+\boldsymbol{\xi}_c\right],\right. \\
&\qquad \left.\left[\sum_{t=1}^{n}\mathbf{E}\left(z_t^i\right)x_t^T x_t+\mathbf{I}\right]^{-1}\right) \\
q\left(z_t^i=1\right) &\propto \exp\left(\mathbf{E}\log v_i+\sum_{j=1}^{i-1}\mathbf{E}\log(1-v_j)+\mathbf{E}(r_t)\,\mathbf{u}_t^T\mathbf{E}\left(\boldsymbol{\beta}_i^T\right)x_t-\right. \\
&\qquad \left.\frac{1}{2}\mathbf{E}\left[\left(\boldsymbol{\beta}_i^T x_t\right)^T\left(\boldsymbol{\beta}_i^T x_t\right)\right]\right) \\
q(r_t) &\propto r_t\exp\left\{-\frac{1}{2}r_t^2+r_t\mathbf{u}_t^T\left(\sum_{i=1}^{L}\mathbf{E}\left(z_t^i\right)\mathbf{E}\left(\boldsymbol{\beta}_i^T\right)\right)x_t\right\} \\
q(\alpha) &= \text{Gamma}\left(a+L-1, b-\sum_{i=1}^{L-1}\mathbf{E}\log(1-v_i)\right),
\end{aligned}$$

where $u_{t,1}=\cos\theta_t$ and $u_{t,2}=\sin\theta_t$ and $\mathbf{E}\left[\left(\boldsymbol{\beta}_i^T x_t\right)^T\left(\boldsymbol{\beta}_i^T x_t\right)\right]=\sum_{c=1}^{2}\text{tr}\left(x_t x_t^T\text{cov}\left(\boldsymbol{\beta}_{i,c}\right)\right)+\mathbf{E}\left(\boldsymbol{\beta}_{i,c}^T\right)x_t x_t^T\mathbf{E}\left(\boldsymbol{\beta}_{i,c}\right)$. We can easily have a variational distribution for $\boldsymbol{\xi}_c$ by making its prior a normal distribution. All expectations are with respect to the variational distributions. The variational algorithm consists of updating these expectations one at a time in a similar way as in Algorithm 3.

4.6.4 Comparison

Figures 15 and 16 show predictive distributions using the variational approximations for the models in Sections 4.6.2 and 4.6.3 using the data for New Hampshire with the purpose of comparing the results to the Gibbs sampler. The Gibbs sampler works very well, adding multiple modes when data are rich and reverting to the regression effect when data are sparse. The regression parameters are well estimated and we do not have the issue of sensitivity to

initial values as we do for the variational methods. Unfortunately the Gibbs sampler cannot handle large data sets, as previously mentioned. For the two variational approaches in Sections 4.6.2 and 4.6.3, the approximations of the form (33) that led to unimodal predictions (not shown) had the highest lower bound. We then applied the improvements introduced in Section 4.5. First, we obtained several variational approximations of the form (33) by choosing different initial values for the variational algorithms, and selected the one with the highest lower bound. Second, we improved the approximations by using Result 7, i.e. a sampling scheme similar to the one in (38). Finally, we compared the lower bounds among all improved approximations and chose the approximation with highest lower bound. The estimates in Figures 15 and 16 were obtained by using the approximations that had highest lower bound after making $b = \{\boldsymbol{z}\}$ in Result 7.

A drawback of the variational approximation in the multiple populations model is that the variational distributions for the parameters in the base distribution depend highly on the chosen truncation levels for each of the combinations of the factors. Hence, there is still the need to appropriately select these tuning parameters. In the model in Section 4.6.3, this problem is avoided by having the regression at the level of the mixtures. Unfortunately, the disadvantage of this last model is that we were once again not able to run it for the whole data set due to the sizes of some of the matrices involved in the variational algorithm.

For the reasons mentioned above, the analysis of the full dataset was ultimately only performed using the approach of Section 4.6.2. The variational distributions of the form (33) that gave the highest lower bound gave us again unimodal predictive distributions and so we had to look at variational distributions of the form (38) and chose the one with highest lower bound. For illustration, Figure 17 shows predictive distributions for some cells with a lot of data and some cells with sparse data.

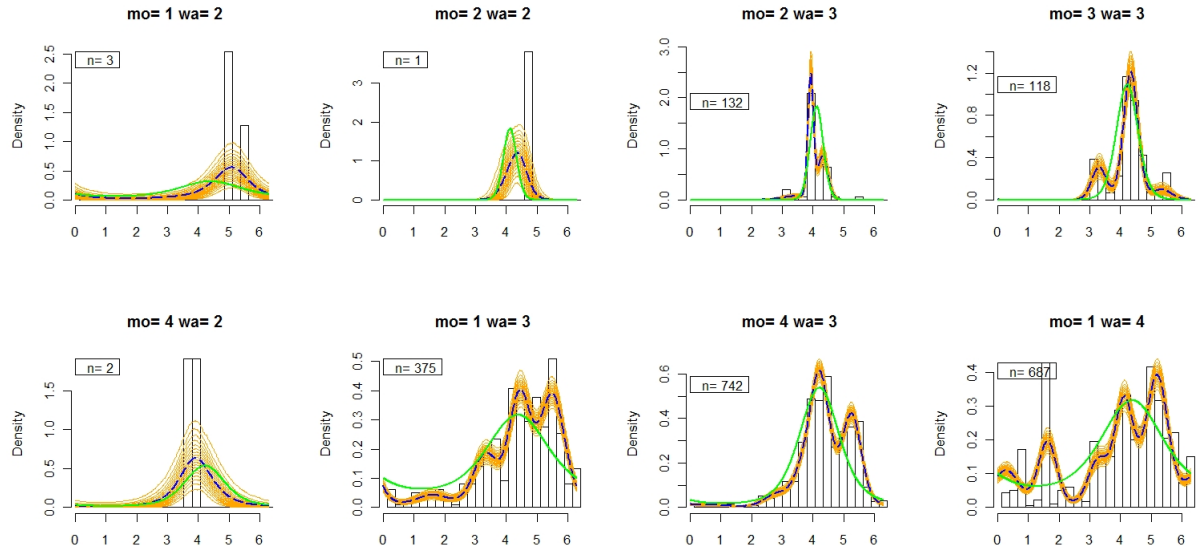

Figure 15: Results from the variational approximation to the *multiple populations* model for the state of New Hampshire. The dashed dark curve in each panel is the predictive distribution. The light solid curve is the posterior mean estimate of the mode effect. The light curves above and bellow the predictive distribution are $10\%, 20\%, \ldots, 90\%$ point-wise credible intervals.

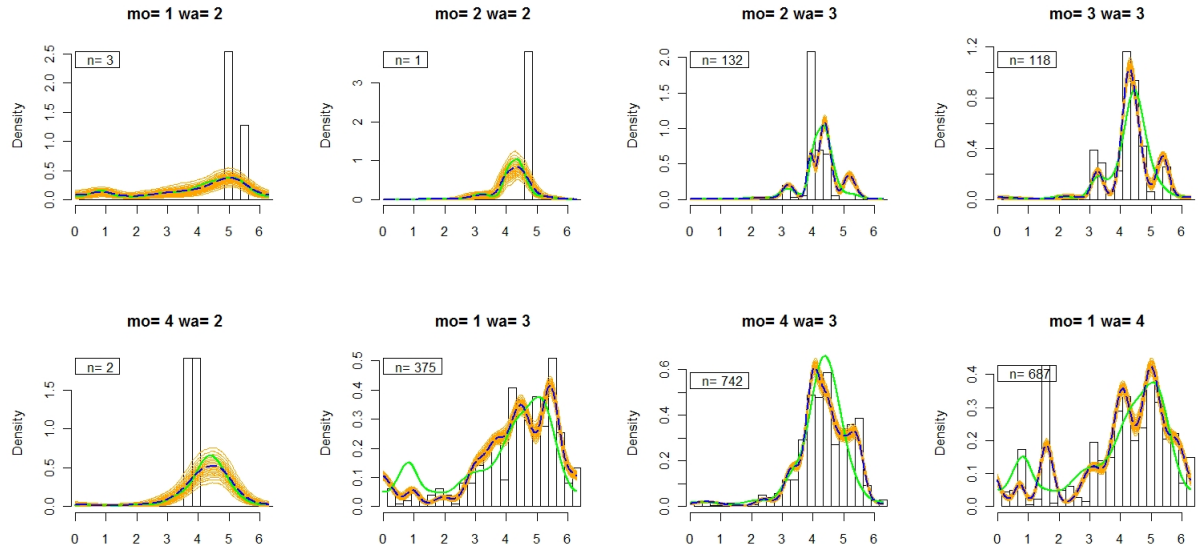

Figure 16: Results from the variational approximation to the *Dirichlet process prior on regression coefficients* model for the state of New Hampshire. The dark curve in each panel is the predictive distribution. The light solid curve is the posterior mean estimate of the mode effect.

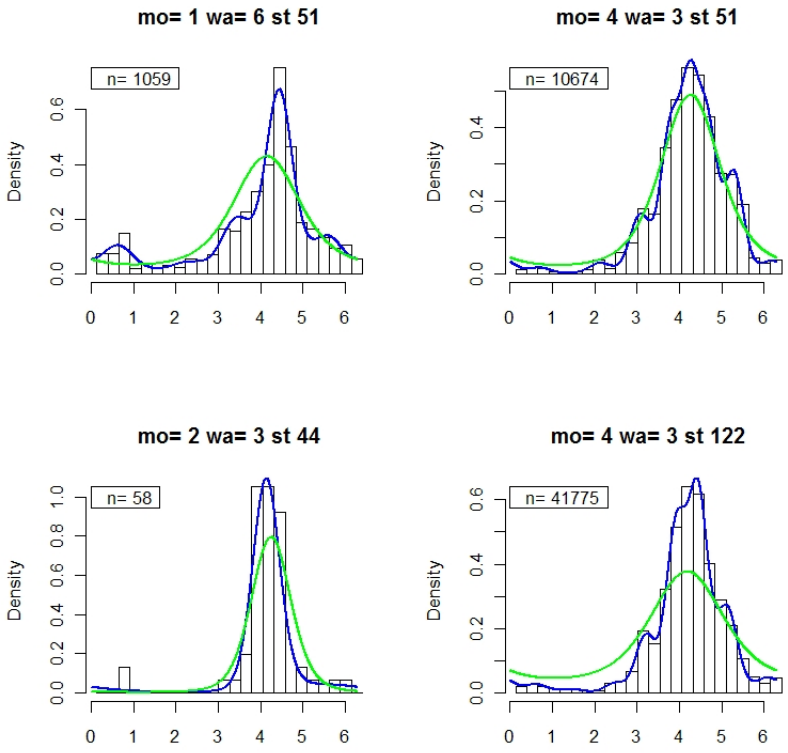

Figure 17: Predictive distributions for some of the combinations of state wave and mode obtained by using a variational approximation for the model described in Section 4.6.2 with a mode, state and wave effects model in the base distribution. Top left and top right are estimates for the state of Virginia for different combination of modes and waves. Bottom left and bottom right show estimates for the states of Rhode Island and Florida respectively. Mode 1 is shore, mode 2 is head boat and mode 4 is PR. Sample sizes for each of the combinations of state wave and mode are given in the plots. The dark curve in each panel is the predictive distribution. The light solid curve is the posterior mean estimate of the mode + state + wave effect.

4.7 Supplement: Projected normal density, spherical case

Let $\mathbf{X} = r\mathbf{u}^T = r\left(\cos\theta_1 \sin\theta_2, \sin\theta_1 \sin\theta_2, \cos\theta_1\right)^T \sim \mathrm{N}_3\left(\boldsymbol{\mu}, \mathrm{W}\right)$. The joint distribution of $\theta = (\theta_1, \theta_2)$ and r is:

$$\begin{aligned} p\left(\theta, r\right) &= r^2 \mathrm{N}_3\left(r\mathbf{u}; \boldsymbol{\mu}, \mathrm{W}\right) \\ &= r^2 \left(\frac{1}{2\pi}\right)^{\frac{3}{2}} \mid \mathrm{W} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(r\mathbf{u} - \boldsymbol{\mu}\right)^T \mathrm{W}^{-1}\left(r\mathbf{u} - \boldsymbol{\mu}\right)\right\} \\ &= r^2 \left(\frac{1}{2\pi}\right)^{\frac{3}{2}} \mid \mathrm{W} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}ar^2 + br + c\right\} \end{aligned}$$

where $a = \mathbf{u}^T \mathrm{W}^{-1}\mathbf{u}$, $b = \mathbf{u}^T \mathrm{W}^{-1}\boldsymbol{\mu}$, $c = -\frac{1}{2}\boldsymbol{\mu}^T \mathrm{W}^{-1}\boldsymbol{\mu}$. Integrating r out will give us the desired density $\mathrm{PN}_3(\boldsymbol{\mu}, \mathrm{W})$

$$\begin{aligned} p(\theta) &= \int_0^\infty p(\theta, r)\, dr \\ &= \left(\frac{1}{2\pi}\right)^{\frac{3}{2}} \mid \mathrm{W} \mid^{-\frac{1}{2}} \exp\{c\} \int_0^\infty r^2 \exp\left\{-\frac{1}{2}ar^2 + br\right\} dr. \end{aligned}$$

We first integrate by parts, with $u = r\exp\{br\}$ and $dv = r\exp\left\{-\frac{1}{2}ar^2\right\} dr$. Then $du = e^{br}(br+1)\, dr$ and $v = -\frac{1}{a}\exp\left\{-\frac{1}{2}ar^2\right\}$.

$$\begin{aligned} \int_0^\infty r^2 \exp\left\{-\frac{1}{2}ar^2 + br\right\} dr &= \int_0^\infty \frac{1}{a}(br+1)\exp\left\{-\frac{1}{2}ar^2 + br\right\} dr \\ &= \frac{1}{a}\left[b\int_0^\infty r\exp\left\{-\frac{1}{2}ar^2 + br\right\} dr + \right. \\ &\qquad \left. \int_0^\infty \exp\left\{-\frac{1}{2}ar^2 + br\right\} dr\right]. \end{aligned}$$

We now compute the second integral inside the bracket:

$$\begin{aligned} \int_0^\infty \exp\left\{-\frac{1}{2}ar^2 + br\right\} dr &= \exp\left\{\frac{b^2}{2a}\right\} \int_0^\infty \exp\left\{-\frac{a}{2}\left(r - \frac{b}{a}\right)^2\right\} dr \\ &= \exp\left\{\frac{b^2}{2a}\right\} \frac{1}{\sqrt{a}} \int_{-\frac{b}{\sqrt{a}}}^\infty \exp\left\{-\frac{1}{2}z^2\right\} dz \\ &= \exp\left\{\frac{b^2}{2a}\right\} \sqrt{\frac{2\pi}{a}}\Phi\left(\frac{b}{\sqrt{a}}\right) \\ &= \frac{1}{\sqrt{a}} \frac{\Phi\left(\frac{b}{\sqrt{a}}\right)}{\varphi\left(\frac{b}{\sqrt{a}}\right)}. \end{aligned}$$

To get the first integral inside the bracket we define $u = -\frac{1}{a}\exp\left\{-\frac{1}{2}ar^2 + br\right\}$, then

$$\begin{aligned}\int_0^\infty r\exp\left\{-\frac{1}{2}ar^2+br\right\}dr &= \int du + \frac{b}{a}\int_0^\infty \exp\left\{-\frac{1}{2}ar^2+br\right\}dr \\ &= u\left(r\right)|_0^\infty + \frac{b}{a}\frac{1}{\sqrt{a}}\frac{\Phi\left(\frac{b}{\sqrt{a}}\right)}{\varphi\left(\frac{b}{\sqrt{a}}\right)} \\ &= \frac{1}{a}\left[1+\frac{b}{\sqrt{a}}\frac{\Phi\left(\frac{b}{\sqrt{a}}\right)}{\varphi\left(\frac{b}{\sqrt{a}}\right)}\right].\end{aligned}$$

Finally,

$$\mathrm{PN}_3\left(\theta_1,\theta_2;\boldsymbol{\mu},\mathrm{W}\right) = \left(\frac{1}{2\pi a}\right)^{\frac{3}{2}} |\ \mathrm{W}\ |^{-\frac{1}{2}}\exp\left\{c\right\}\left(\left[1+\frac{b}{\sqrt{a}}\frac{\Phi\left(\frac{b}{\sqrt{a}}\right)}{\varphi\left(\frac{b}{\sqrt{a}}\right)}\right]\frac{b}{\sqrt{a}}+\frac{\Phi\left(\frac{b}{\sqrt{a}}\right)}{\varphi\left(\frac{b}{\sqrt{a}}\right)}\right).$$

4.8 Supplement: Computational derivations

4.8.1 Variational posterior mean for a projected normal random sample

To get the posterior means $\mathbf{E}\left(\boldsymbol{\eta}_i^*|\text{data in bin } i\right)$ we use the variational method explained in Hernandez-Stumpfhauser et al. (2011). The iterative algorithm is

$$\begin{aligned}\mathbf{E}\left(r_t\right) &\leftarrow \frac{\sqrt{2\pi}}{\mathrm{C}\left(b_t\right)}\exp\left(b_t^2/2\right)\Phi\left(b_t\right)+b_t \\ \mathbf{E}\left(\boldsymbol{\eta}_i^*|\text{data in bin } i\right) &\leftarrow \frac{1}{n_i}\sum_{t=1}^{n_i}\mathbf{u}_t\mathbf{E}\left(r_t\right)\end{aligned}$$

where $b_t = \mathbf{u}_t^T\mathbf{E}\left(\boldsymbol{\eta}_i^*|\text{data in bin } i\right)$ and $\mathrm{C}\left(b_t\right) \leftarrow 1+\sqrt{2\pi}b_t\exp\left(b_t^2/2\right)\Phi\left(b_t\right)$. This algorithm converges rapidly and it is not sensitive to initial values.

4.8.2 Laplace approximation

Each draw from $q_0(\mathbf{z})$ clusters the data, and then applying the method of Hernandez-Stumpfhauser et al. (2011),

$$\begin{aligned}\mathbf{E}(r_t) &\leftarrow \frac{\sqrt{2\pi}}{\mathrm{C}(b_t)}\exp\left(b_t^2/2\right)\Phi(b_t)+b_t \\ \mathbf{m}_i &\leftarrow \left[\frac{1}{n_i+1}\right]\left(\boldsymbol{\mu}+\sum_{t=1}^{n_i}\mathbf{u}_t\mathbf{E}(r_t)\right),\end{aligned}$$

where $b_t = \mathbf{u}_t^T\mathbf{m}_t$, $\mathrm{C}(b_t) \leftarrow 1+\sqrt{2\pi}b_t\exp(b_t^2/2)\Phi(b_t)$, and $\mathbf{u}_t$ are such that $z_t^i = 1$. The variance-covariance matrix W_i is found by taking the inverse of minus the Hessian of the log posterior distribution evaluated at $\mathbf{m}_i$ (Hernandez-Stumpfhauser et al. 2011). The log posterior distribution for $\boldsymbol{\eta}_i$ is

$$\log p(\boldsymbol{\eta}_i|\boldsymbol{\theta},\mathbf{z},\boldsymbol{\mu}) = \log \mathrm{N}_2(\boldsymbol{\eta}_i;\boldsymbol{\mu},\mathbf{I}_2)+\sum_{t=1}^{n_i}\log \mathrm{PN}_2(\theta_t;\boldsymbol{\eta}_i,\mathbf{I}_2)+C.$$

The second derivatives needed to compute the Hessian are
$\frac{\partial^2}{\partial\eta_c^2}\log \mathrm{N}_2(\boldsymbol{\eta}_i;\boldsymbol{\mu},\mathbf{I}_2) = -1$, $\frac{\partial^2}{\partial\eta_1\partial\eta_2}\log \mathrm{N}_2(\boldsymbol{\eta}_i;\boldsymbol{\mu},\mathbf{I}_2) = 0$,
$\frac{\partial^2}{\partial\eta_c^2}\log \mathrm{PN}_2(\theta_t;\boldsymbol{\eta}_i,\mathbf{I}_2) = -1+u_{t,c}^2B_{t,i}$, $\frac{\partial^2}{\partial\eta_1\partial\eta_2}\log \mathrm{PN}_2(\theta_t;\boldsymbol{\eta}_i,\mathbf{I}_2) = u_{t,1}u_{t,2}B_{t,i}$, where $B_{t,i} = 2-\frac{\Phi(b_{t,i})}{\varphi(b_{t,i})}\left[1+\frac{b_t\Phi(b_{t,i})}{\varphi(b_{t,i})}\right]^{-1}\left[\frac{\Phi(b_{t,i})}{\varphi(b_{t,i})}\left[1+\frac{b_t\Phi(b_{t,i})}{\varphi(b_{t,i})}\right]^{-1}+b_{t,i}\right]$, $b_{t,i} = \mathbf{u}_t^T\boldsymbol{\eta}_i$, $\mathbf{u}_t^T = (\cos\theta_t,\sin\theta_t)$, and $u_{t,c}$ is the c^{th} component of u_t.

4.8.3 Efficient algorithm to sample from $p(r|b)$

The cumulative distribution function (cdf) can be written as

$$F(r) = \frac{1}{C(b)}\left[1-\exp\left(-\frac{1}{2}r^2+br\right)+\sqrt{2\pi}b\exp\left(\frac{b^2}{2}\right)\left[\Phi(r-b)-\Phi(-b)\right]\right],$$

where $\Phi(\cdot)$ is the cdf of a standard normal random variable and b is the variational parameter in $q_0(r)$. We then sample from $q_0\left(\mathbf{r}'\right)$ by making use of the inverse cdf technique. We do

not have an explicit form for this inverse, but it is approximated by applying the Newton-Rhapson method using as initial value the mode of $q_0(r)$ which is equal to $\frac{b+\sqrt{b^2+4}}{2}$. That is, draw a uniform$(0,1)$, say u, and find r such that $F(r)-u=0$. We use the Newton-Rhapson method to find the root of $F(r)-u$.

$$r_{i+1} = r_i - \frac{F(r_i)-u}{F'(r_i)}$$

where $F'(r) = q_0(r)$ is the derivative of $F(r)$. Taking as initial value $r_0 = \frac{b+\sqrt{b^2+4}}{2}$ the algorithm is guaranteed to converge to the root of $F(r)-u$.

4.8.4 Full conditionals for the parameters in the base distribution in the multiple populations model

Let I, J, K be the number of levels for the state wave and mode factors and let $\mathbf{m}_{-K,c} = \left(\mathbf{m}'_{1,c}, \ldots, \mathbf{m}'_{K-1,c}\right)^T$ where $\mathbf{m}'_{k,c}$ is the c^{th} component of $\mathbf{m}'_k$ and $c = 1, 2$. Also, let all prior means of the factors be zero and prior variances of the overall, state, wave and mode factors be $\sigma_0^2, \sigma_s^2, \sigma_w^2, \sigma_m^2$ respectively. The full conditionals for the overall mean $\mathbf{m}'_0$ and mode effects $\mathbf{m}_{-K,c} = \left\{\mathbf{m}'_{1,c}, \ldots, \mathbf{m}'_{K-1,c}\right\}$ are shown next and full conditionals for all other factors would have equivalent forms:

$$\begin{aligned} p(\mathbf{m}'_0|\cdot) &= \mathrm{N}_2\left(\frac{\sigma_0'^2}{1+n^*\sigma_0'^2}\sum_{ijk}\sum_{t}^{n^*_{ijk}}\boldsymbol{\mu}^*_{ijkt} - \mathbf{m}'_k - \mathbf{s}'_i - \mathbf{w}'_j, \frac{\sigma_0'^2}{1+n^*\sigma_0'^2}\mathbf{I}_2\right) \\ p(\mathbf{m}_{-K,c}) &= \mathrm{N}_{K-1}\left((\mathbf{V}_2)^{-1}\mathbf{V}_1\boldsymbol{y}, (\mathbf{V}_2)^{-1}\right), \end{aligned}$$

where n^*_{ijk} is the number of distinct $\boldsymbol{\mu}'s$ for combination ijk of state, wave and mode. $n^* = \sum_{ijk} n^*_{ijk}$ is the total number of distinct $\boldsymbol{\mu}'s$, $\sigma_0'^2 = \sigma_0^2 + \frac{\sigma_s^2}{I} + \frac{\sigma_w^2}{J} + \frac{\sigma_m^2}{K}$, $n^*_k = \sum_{ij} n^*_{ijk}$, $\mathbf{V}_1 = \mathrm{diag}\left(n^*_1, \ldots, n^*_{K-1}\right)$, $\mathbf{V}_2 = n^*_K\mathbf{J} + \frac{1}{\sigma_m^2}(\mathbf{I}-\mathbf{J})^{-1} + \mathbf{V}_1$ where $\mathbf{I}, \mathbf{J}$ are the identity matrix and matrix of ones respectively. Finally $y_{k,c} = \sum_{ij}\sum_t^{n^*_{ijk}} \boldsymbol{\mu}^*_{ijkt,c} - \mathbf{m}'_{0,c} - \mathbf{s}'_{i,c} - \mathbf{w}'_{j,c}$ and

$\boldsymbol{y} = \left(\left(y_{1,c} - y_{K,c} \right) / n_1^*, \ldots, \left(y_{K-1,c} - y_{K,c} \right) / n_{K-1}^* \right)^T$.

4.8.5 Full conditional for the precision parameter in the multiple populations model

Let $S = IJK$ denote the total number of state-wave-mode combinations. From Escobar and West (1995), the conditional distribution of the number of distinct components n_{ijk}^* in cell ijk is $p\left(n_{ijk}^* | \alpha, n_{ijk}\right) = c_{n_{ijk}}\left(n_{ijk}^*\right) n_{ijk}! \alpha^{n_{ijk}^*} \frac{\Gamma(\alpha)}{\Gamma\left(\alpha + n_{ijk}\right)}$, $n_{ijk}^* = 1, 2, \ldots, n_{ijk}$ and $c_{n_{ijk}}\left(n_{ijk}^*\right) = p\left(n_{ijk}^* | \alpha = 1, n_{ijk}\right)$. Hence we deduce that

$$p\left(n_{1,1,1}^*, \ldots, n_S^* | \alpha, n_{1,1,1}, \ldots, n_S\right) = \prod_{ijk} p\left(n_{ijk}^* | \alpha, n_{ijk}\right).$$

Making use of the identity $\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)\beta(\alpha+1,n)}{\alpha\Gamma(n)}$ where $\beta(\cdot,\cdot)$ is the beta function, the full conditional for α can be written as:

$$\begin{aligned}
p(\alpha|\cdot) &\propto p(\alpha)\, p\left(n_{1,1,1}^*, \ldots, n_S^* | \alpha\right) \\
&\propto p(\alpha)\, \alpha^{\sum n_{ijk}^* - S} \prod_{ijk} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{ijk})} \\
&\propto p(\alpha)\, \alpha^{\sum n_{ijk}^* - S} \prod_{ijk} (\alpha + n_{ijk})\, \beta(\alpha + 1, n_{ijk}) \\
&\propto p(\alpha)\, \alpha^{\sum n_{ijk}^* - S} \prod_{ijk} (\alpha + n_{ijk}) \int_0^1 \xi^{\alpha} (1 - \xi)^{n_{ijk} - 1}\, d\xi.
\end{aligned}$$

This implies that $p(\alpha|\cdot)$ is the marginal distribution from a joint distribution for α and a continuous vector $\boldsymbol{\xi} = (\xi_{1,1,1}, \ldots, \xi_S)^T$ such that

$$p(\alpha|\cdot) \propto p(\alpha)\, \alpha^{\sum n_{ijk}^* - S} \left[\prod_{ijk} (\alpha + n_{ijk}) \right] \left[\prod_{ijk} \xi_{ijk}^{\alpha} (1 - \xi_{ijk})^{n_{ijk} - 1} \right].$$

If $p(\alpha) = \text{Gamma}(a, b)$ where b is the rate parameter, the distribution of α conditioning on $\boldsymbol{\xi}$ and all other parameters is:

$$\begin{aligned} p(\alpha|\cdot) &\propto \alpha^{\sum n^*_{ijk} - S + a - 1} \left[\prod_{ijk} (\alpha + n_{ijk})\right] \left[\exp\left(-\alpha\left(b - \sum \log(\xi_{ijk})\right)\right)\right] \\ &\propto \alpha^{\sum n^*_{ijk} - S + a - 1} \left[\alpha^S + c_1\alpha^{S-1} + \ldots + c_S\right] \left[\exp\left(-\alpha\left(b - \sum \log(\xi_{ijk})\right)\right)\right] \\ &\propto \left[\sum_{s=0}^{S} c_s \alpha^{\sum n^*_{ijk} + a - 1 - s}\right] \left[\exp\left(-\alpha\left(b - \sum \log(\xi_{ijk})\right)\right)\right], \end{aligned} \tag{41}$$

thus a mixture of gamma distributions with parameters $\Big(\sum_{ijk} n^*_{ijk} + a - s$, $b - \sum_{ijk} \log(\xi_{ijk})\Big)$. Finally the full conditional distribution of $\boldsymbol{\xi}$ is:

$$\begin{aligned} p(\boldsymbol{\xi}|\cdot) &\propto \prod_{ijk} \xi_{ijk}^{\alpha} (1 - \xi_{ijk})^{n_{ijk} - 1} \\ &\propto \prod_{ijk} \text{Beta}(\alpha + 1, n_{ijk}), \end{aligned} \tag{42}$$

thus $p(\boldsymbol{\xi}|\cdot)$ is the distribution of independent beta distributions with parameters $(\alpha + 1, n_{ijk})$. At each Gibbs iteration, the currently sampled values of n^*_{ijk} and α allow us to draw a new value of α by (a) first sampling $\boldsymbol{\xi}$ from the distribution in Equation (42), then (b) sampling the new α value from the mixture distribution Equation (41).

REFERENCES

Abramowitz, M. and I. A. Stegun (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Volume 55. U.S. Government Printing Office.

Batschelet, E. (1981). *Circular Statistics in Biology*. Academic Press.

Binder, D. and G. R. Roberts (2003). Design-based and model-based methods for estimating model parameters. In R. Chambers and C. Skinner (Eds.), *Analysis of Survey Data*. Wiley.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blackwell, D. and J. MacQueen (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics 1*, 353–355.

Blei, D. M. and M. I. Jordan (2006). Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis 1*, 121–144.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge.

Breckling, J. (1989). *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*. Springer-Verlag.

Elvers, E., C. Särndal, J. Wretman, and G. Örnberg (1985). Regression analysis and ratio analysis for domains: a randomization theory approach. *Canadian Journal of Statistics 13*, 185–199.

Escobar, M. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Fuller, W. A. (1996). *Introduction to Statistical Time Series* (2 ed.). New York, NY: John Wiley & Sons.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrizations for normal linear mixed models. *Biometrika 82*, 479–488.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Ghahramani, Z. and M. Beal (2001). Propagation algorithms for variational Bayesian learning. *Advances in Neural Information Processing Systems 13*, 507–513.

Ghosh, M. and J. N. K. Rao (1994). Small area estimation: an appraisal. *Statistical Science 9*, 55–93.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Givens, G. H. and J. A. Hoeting (2005). *Computational Statistics*. Wiley.

Goga, C. and A. Ruiz-Gazen (2012). Efficient estimation of nonlinear finite population parameters using nonparametrics. *JRSSB*. To appear.

Gordon, A. D., P. E. Jupp, and R. W. Byrne (1989). Variational inference for Dirichlet process mixtures. *British J. Math. Statist. Psych. 9*, 169–182.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5*, 361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics 35*, 1491–1523.

Hernandez-Stumpfhauser, D., J. D. Opsomer, and F. J. Breidt (2011). Hierarchical Bayesian small area estimation for circular data. Manuscript.

Kendall, D. G. (1974). Pole-seeking brownian motion and bird navigation. *Journal of the Royal Statistical Society 36*, 261–294.

Krewski, D. and J. Rao (1981). Inference from stratified samples: Properties of linearization, jackknife and balanced repeated replication methods. *Ann. Statist. 9*, 1010–1019.

Kurihara, K., M. Welling, and N. Vlassis (2006). Accelerated variational dirichlet process mixtures. In *NIPS. 2006*.

Lee, J. H. A. (1963). (correspondence). *British Med. J.*.

MacEachern, S. N. and P. Muller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*, 223–238.

Mardia, K. and R. Edwards (1982). Weighted distributions and rotating caps. *Biometrika 69*, 323–330.

Mardia, K. V. (1972). *Statistics of Directional Data*. Academic Press.

Mardia, K. V. and P. E. Jupp (2000). *Directional Statistics*. Chichester, UK: Wiley.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*, 249–265.

Nuñez-Antonio, G. and E. Gutiérrez-Peña (2005). A Bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics 32*(10), 995–1001.

Opper, M. and D. Saad (2001). *Advanced Mean Field Methods: Theory and Practice*. Cambridge.

Ormerod, J. T. and M. P. Wand (2010). Explaining variational approximations. *The American Statistician 64*, 140–153.

Presnell, B., S. P. Morrison, and R. C. Littell (1998). Projected multivariate linear models for directional data. *Journal of the American Statistical Association 93*(443), 1068–1077.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley-Interscience.

Rosén, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement, i and ii. *Annals of Mathematical Statistics 43*, 373–397; 748–776.

Rosén, B. (1997). On sampling with probability proportional to size. *J. Statist. Plann. Inference 62*, 159–191.

Rubin-Bleuer, S. and I. S. Kratina (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics 33*, 2789–2810.

Särndal, C. E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schmidt-Koenig, K. (1965). *Current problems in bird orientation. In D. Lehrman et al.(eds) Advances in the Study of Behaviour*. Academic Press.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Sullivan, P. J., F. J. Breidt, R. B. Ditton, B. A. Knuth, B. M. Leaman, V. M. O'Connell, G. R. Parsons, K. H. Pollock, S. J. Smith, and S. L. Stokes (2006). *Review of Recreational Fisheries Survey Methods*. Washington, DC: National Academies Press.

Teh, Y., M. I. Jordan, M. J. Beal, and D. M. Blei (2004). Hierarchical Dirichlet processes. *Technical Report 653, UC Berkeley Statistics*.

Vines, S. K., W. R. Gilks, and P. Wild (1996). Fitting Bayesian multiple random effects models. *Statistics and Computing 6*, 337–346.

Wainwright, M. and M. Jordan (2003). Graphical models, exponential families, and variational inference. Technical Report 649, U.C. Berkeley, Dept. of Statistics.

Watson, G. S. (1983). *Statistics on Spheres*. Wiley.