2014

Accuracy assessment in multivariate Bayesian forecasting linear and nonlinear models

Clark Joachim Kogan

ACCURACY ASSESSMENT IN MULTIVARIATE BAYESIAN FORECASTING LINEAR AND NONLINEAR MODELS

BY

CLARK JOACHIM KOGAN

B.A. Physics, University of Montana, Missoula, Montana, 2008

Dissertation

presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mathematics, Traditional Mathematical Sciences Research

The University of Montana

Missoula, MT

May 2014

Approved by:

Sandy Ross, Associate Dean of The Graduate School

Graduate School

Dr. Leonid V. Kalachev, Chair

Department of Mathematical Sciences, The University of Montana, Missoula, MT

Dr. Hans P.A. Van Dongen, Co-chair

Sleep and Performance Research Center, Washington State University, Spokane, WA

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Dr. Brian Steele
Department of Mathematical Sciences, The University of Montana,
Missoula, MT

Dr. Jesse Johnson
Department of Mathematical Sciences, The University of Montana,
Missoula, MT

Dr. Jon Graham
Department of Mathematical Sciences, The University of Montana,
Missoula, MT

June 6, 2014

Abstract

Reduced alertness and high levels of cognitive fatigue due to sleep loss bring forth substantial risks in today's 24/7 society. Biomathematical models can be used to help mitigate such risks by predicting quantitative levels of fatigue under sleep loss. These models help manage risk by providing information on the timing at which high levels of fatigue will occur; countermeasures can then be taken to reduce accident risk at such critical times.

Many quantitative models exist to predict cognitive performance based on homeostatic and circadian processes (Mallis et al., 2004). These models have typically been fitted to group average data. Due to large individual variation, group-average predictions are often inaccurate for a given individual. However, since individual differences are trait-like, between subjects variation can be captured by individualizing model parameters using the technique of Bayesian forecasting. In many cases the amount of data collected, and consequently, the prediction accuracy, will be limited by factors such as cost and availability. However; prediction accuracy may still be improved by including information from alternative, correlated performance measures in a multivariate Bayesian forecasting framework.

When collecting data from two performance measures, we consider methods of sampling that obtain a desired average level of prediction accuracy for minimal data collection cost. We assess the prediction accuracy using the Bayesian mean squared error (MSE) and derive this measure for a general Bayesian linear model. To understand how the accuracy depends on the number of measurements from primary and secondary tasks in the simplest case, we apply the equation to specify the accuracy for the bivariate Bayesian linear model of subject means. For this simple model, we further assume that observations from each performance measure have a fixed cost per data point, and use this assumption to determine the number of measurements of each variable needed to minimize the cost while still obtaining no less than the desired level of accuracy.

To aid the extension of the findings from the linear case to state of the art nonlinear biomathematical fatigue models, we focus on obtaining our extended measure of accuracy for the nonlinear case. Computing this accuracy analytically is often infeasible without reliance on model approximations. Model simulations can be used to compute this accuracy; however, such simulations can be time consuming, especially for models that lack analytic solutions and require that a system of differential equations be solved to produce model dynamics. Much of this computational burden in assessing estimator accuracy, however, is produced by using the Bayesian MMSE estimator, and could be reduced by taking advantage of the quicker to compute Bayesian MAP estimator. We show how for a nonlinear biomathematical model that the accuracy assessment using repeated simulation with the

MAP estimator yields a reasonable estimate of the accuracy obtained using the MMSE estimator. Still, however, for any given case, determination of whether the MMSE accuracy can be approximated with the MAP accuracy requires these time consuming simulations. We begin to analytically identify classes of models where the MMSE accuracy can be approximated by the MAP accuracy. We consider a class of quadratic Bayesian models, and show by analytic approximation that for this class, the MMSE has twice the accuracy of the MAP.

Contents

Chapter 1

Introduction

1.1 Background

Use of models

Reduced alertness and high levels of cognitive fatigue due to sleep loss bring forth substantial risks in today's 24/7 society. Biomathematical models can be used to help mitigate such risks by predicting quantitative levels of fatigue under sleep loss (Van Dongen et al, 2007). These models help manage risk by providing information on the timing at which high levels of fatigue will occur; countermeasures can then be taken to reduce accident risk at such critical times.

Truck driver application

For example, biomathematical models can be used to make predictions of truck driver performance. Driving performance is often assessed using the variance in lane position, referred to more simply as lane variability (Forsman et al, 2012). Lane variability at intervals can be computed from lane position, which in turn can be estimated using lane-tracking cameras and complex video signal processing software. Once predictions of driver lane variability have been constructed, they may be used to inform a driver alert system, or help trucking companies or individual drivers make decisions concerning driver schedules.

Model biology

Performance predictions can be made by modeling the neurobiology underlying temporal changes in cognitive fatigue. The brain's drive for sleep and wakefulness is regulated largely by homeostatic and circadian processes (Van Dongen & Dinges, 2005). The homeostatic process, which is responsible for balancing the time spent awake and asleep, can be thought of as a pressure for sleep that increases during wakefulness and decreases during sleep. The circadian process, on the other hand, is responsible for managing the drive for wakefulness over the course of the day. This process has been directly linked to the suprachiasmatic nuclei, a specialized region located in the hypothalamus of the brain which is responsible for keeping track of the time of day. State of the art fatigue models use our understanding of these two processes to predict performance throughout the day.

Individual differences

Many quantitative models exist to predict cognitive performance based on homeostatic and circadian processes (Mallis et al, 2004, McCauley et al, 2013). These models have typically been fitted to group average data and have successfully been used to predict group-average performance in operational settings. There is also demand for these models in operative environments where fewer individuals are being considered (i.e. commercial trucking.) However, for a given individual the group-average prediction can be quite inaccurate. The reason for this inaccuracy is that the larger part of the variation in performance due to sleep loss is due to trait-like individual differences. Therefore, we cannot rely on the group-average model to accurately predict the absolute level of performance impairment for a given individual. However, since the individual differences are trait-like, between subjects variation can be captured by individualizing model parameters.

Group average advantage

Individual parameter estimates may be obtained by fitting a given model to subject-specific performance data, as opposed to fitting the model with group-average data. However, group-average data has certain advantages. Group data will not only help to average out noise, but likely also be collected over a greater diversity of schedules. This schedule diversity aids the fitting process by enabling unique identification of parameters that can only be separated by observing performance under more extreme schedules of sleep restriction. Therefore, in tailoring a model to an individual by fitting only subject-specific performance data, we let

go of the noise reduction and schedule variety which otherwise would be gained from the group.

Bayesian forecasting

To both retain the strength of group data and capture trait-like fluctuations away from the group-average, we take on the assumption that the individual is a member of a specific population. This assumption allows us to use data from this population to specify prior distributions for the individual parameters. Bayesian forecasting can then be used to combine the individual and group data to produce more robust estimates of subject specific performance. Performance predictions will then naturally take on population mean values in the absence of individual-specific data, and will converge to the best representation of the individual at hand as more individual-specific data is obtained.

Secondary task data

With an unlimited amount of individual-specific data, performance prediction accuracy can be maximized. Data, however, may be limited by factors such as cost and availability. For instance, estimates of lane variability may, at times, be unavailable, as lane tracking cameras are known to be unreliable in darkness, and when snow or sand are covering lane markers (Forsman et al, 2012). Alternatively, each data point collected from the primary task may be costly, and a second, more cost effective task measure may also be available for streaming. In consideration of this range of data-limited scenarios, we pose the question: is it possible to use other information sources, specifically secondary performance measures, to improve the accuracy of those predictions on a primary measure?

PERCLOS

Lane variability is not the only manner in which driving performance may be assessed. For instance, infrared cameras may be used to monitor eyelid closure. Slow eyelid closure ("droop"), is a more effective measure of drowsiness than fast blinking. PERCLOS is one such measure, which measures the proportion of time during a minute for which the eyelid covers more than 80 percent of the pupil. This measurement of driving performance has been found to correlate with lane variability. During abnormal driving conditions, it may be possible to use PERCLOS to obtain more accurate predictions of lane variability.

Framework for including secondary task

Secondary measures of cognitive performance can be considered under the same homeostatic/circadian biological modeling framework. However, because the individual differences are not the same from one task to the next, we cannot reasonably use the same model parameters for a secondary task. Instead, we can consider task specific model parameters, which have some degree of between subjects' correlation with the primary parameters. A multivariate Bayesian forecasting approach can then be used to combine population data from the primary and secondary measures with new individual data on these measures. Information collected about the performance on a secondary task measure can then be transferred to knowledge about the primary task.

Given this multivariate modeling framework, the question remains whether the secondary task will result in a significant improvement in performance predictions on the primary task, and if so, how an optimal balance of data from primary and secondary tasks can be constructed when there is some cost to collection.

1.2 Mathematical framework

Mathematical Framework

In this section, we develop the mathematical framework for obtaining individual parameter estimates and response forecasts (i.e., performance predictions) via Bayesian forecasting. Furthermore, we construct notation and definitions for assessing the accuracy of these parameter estimates and response forecasts. In later chapters, such a framework will be helpful in showing how secondary tasks influence the accuracy of estimates and forecasts for a primary task.

Model specification

In this thesis, we consider Bayesian models of subject-specific performance, formulated as

$$\boldsymbol{y}_i = \xi\left(\boldsymbol{\phi}_i, \boldsymbol{x}_i\right), \tag{1.2.1}$$

where ξ represents the neurobiological performance model which may be either linear or nonlinear in the parameters, $\boldsymbol{\phi}_i$ represents a subject-specific parameter vector, and $\boldsymbol{x}_i$ is a subject-specific covariate vector. The subject-specific vectors are not subscripted with a subject-specific index as might be expected, as we are assuming that the characteristics of

the population have already been estimated, and we may now focus on making predictions for just a single individual. We assume $\boldsymbol{\phi}_i$ to be a normal random variable

$$\boldsymbol{\phi}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.2.2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ represent population characteristics that have been estimated in the population stage. We use the letter f to represent probability distributions, and subscript it with the random variable for the distribution that is represented. For instance, $f_{\boldsymbol{\phi}_i}$ will be used to represent the prior distribution for $\boldsymbol{\phi}_i$.

Bayesian estimators

Bayes rule is used to construct the posterior distribution by combining the prior, conditional distribution and normalizing constant as

$$f_{\boldsymbol{\phi}_i|\boldsymbol{y}_i} = \frac{f_{\boldsymbol{y}_i|\boldsymbol{\phi}_i} f_{\boldsymbol{\phi}_i}}{\int f_{\boldsymbol{y}_i|\boldsymbol{\phi}_i} f_{\boldsymbol{\phi}_i} d\boldsymbol{\phi}_i}. \tag{1.2.3}$$

From this distribution, we may construct estimates of $\boldsymbol{\phi}_i$. Two estimators which are most commonly used in Bayesian inference are the maximum a posteriori estimator (MAP) and the minimum mean squared error estimator (MMSE). These estimators represent the maximum and mean of the posterior distribution, and can be constructed for both parameters and unobserved responses. The MMSE estimate is that which minimizes the mean squared error conditional on the observed data, whereas the MAP does not have this requirement.

We denote the MMSE estimator of $\boldsymbol{\phi}_i$ with

$$\hat{\boldsymbol{\phi}}_i^{\,E} \equiv \text{MMSE}\left[\boldsymbol{\phi_i}\right] = E_{\boldsymbol{\phi_i},\boldsymbol{y}_i}\left[\boldsymbol{\phi_i}\right] = \int_{-\infty}^{\infty} \boldsymbol{\phi_i} f(\boldsymbol{\phi_i}|\boldsymbol{y}_i) d\boldsymbol{\phi_i} \tag{1.2.4}$$

and the MAP estimator with

$$\hat{\boldsymbol{\phi}}^{\,A} \equiv \text{MAP}\left[\boldsymbol{\phi_i}\right] = \underset{\boldsymbol{\phi_i}}{\text{argmax}}\left[f(\boldsymbol{\phi_i}|\boldsymbol{y}_i)\right]. \tag{1.2.5}$$

We denote the MMSE predictor of the unobserved response $\boldsymbol{y}_i^*$ with

$$\hat{\boldsymbol{y}}_i^E \equiv \text{MMSE}\left[\boldsymbol{y}_i^*\right] = E_{\boldsymbol{y}_i^*,\boldsymbol{y}_i}\left[\boldsymbol{y}_i^*\right] = \int_{-\infty}^{\infty} \boldsymbol{y}_i^* f(\boldsymbol{y}_i^*|\boldsymbol{y}_i) d\boldsymbol{y}_i^* \tag{1.2.6}$$

and the MAP predictor of $\boldsymbol{y}_i^*$ with

$$\hat{\boldsymbol{y}}_i^A \equiv \text{MAP}\left[\boldsymbol{y}_i^*\right] = \underset{\boldsymbol{y}_i^*}{\text{argmax}}\left[f(\boldsymbol{y}_i^*|\boldsymbol{y}_i)\right]. \tag{1.2.7}$$

Accuracy assessment

To assess the accuracy of estimators for the parameter and estimators of an unobserved response (i.e., predictors), we use the *mean squared error* loss function. We define the *posterior expected squared parameter loss* for a single parameter $\hat{\phi}_i$ as

$$\mathcal{R}\left[\hat{\phi}_i\right] \equiv E_{\phi_i|\boldsymbol{y}_i}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] \tag{1.2.8}$$

where $\hat{\phi}_i$ is an estimator of the model parameter ϕ_i and depends on the observed data y. For a parameter vector, this extends to

$$\mathcal{R}\left[\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}\right] \equiv \operatorname{tr}\left\{E_{\boldsymbol{\phi_i}|\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{\phi}}_{\boldsymbol{i}} - \boldsymbol{\phi_i}\right)\left(\hat{\boldsymbol{\phi}}_{\boldsymbol{i}} - \boldsymbol{\phi_i}\right)'\right]\right\}. \tag{1.2.9}$$

Furthermore, we define the *posterior expected squared prediction loss* of $\hat{y}_i$ as

$$\mathcal{R}\left[\hat{y}_i\right] \equiv E_{y_i^*|\boldsymbol{y}}\left[\left(\hat{y}_i - y_i^*\right)^2\right], \tag{1.2.10}$$

where $\hat{y}_i$ is a predictor of y_i^* and depends on $\boldsymbol{y}$. For a vector of predictions, we have

$$\mathcal{R}\left[\hat{\boldsymbol{y}}_i\right] \equiv \operatorname{tr}\left\{E_{\boldsymbol{y}_i^*|\boldsymbol{y}}\left[\left(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*\right)\left(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*\right)'\right]\right\}. \tag{1.2.11}$$

As a result of this dependence on $\boldsymbol{y}_i$, in the nonlinear case, we anticipate that both $\mathcal{R}\left[\hat{\boldsymbol{\phi}}_i\right]$ and $\mathcal{R}\left[\hat{\boldsymbol{y}}_i\right]$ will also depend on $\boldsymbol{y_i}$, and will therefore vary from one dataset to the next. To get an overall measure of the accuracy for a particular Bayesian model, we may instead consider the average value of $\mathcal{R}\left[\hat{\boldsymbol{\phi}}_i\right]$ and $\mathcal{R}\left[\hat{\boldsymbol{y}}_i\right]$ after integrating out the effects of y. This quantity is termed the *Bayesian mean squared parameter error* for $\hat{\boldsymbol{\phi}}_i$ and the *Bayesian mean squared prediction error* for $\hat{\boldsymbol{y}}_i$, and we differentiate them from $\mathcal{R}$ by using the tilde,

$$\tilde{\mathcal{R}}\left[\hat{\boldsymbol{\phi}}_i\right] \equiv \int_{-\infty}^{\infty} \mathcal{R}\left[\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}\right] f\left(\boldsymbol{y}_i\right) d\boldsymbol{\phi}_i, \tag{1.2.12}$$

$$\tilde{\mathcal{R}}\left[\hat{\boldsymbol{y}}_i\right] \equiv \int_{-\infty}^{\infty} \mathcal{R}\left[\hat{\boldsymbol{y}}_i\right] f\left(\boldsymbol{y}_i\right) d\boldsymbol{y}_i. \tag{1.2.13}$$

To compare the accuracy of the parameter estimators and predictors for a particular observed data point, we define the *posterior expected squared parameter loss ratio* of $\hat{\boldsymbol{\phi}}_i^A$ to $\hat{\boldsymbol{\phi}}_i^E$ as

$$\mathcal{P}\left[\hat{\boldsymbol{\phi}}_i^A, \hat{\boldsymbol{\phi}}_i^E\right] \equiv \frac{\mathcal{R}\left[\hat{\boldsymbol{\phi}}_i^A\right]}{\mathcal{R}\left[\hat{\boldsymbol{\phi}}_i^E\right]}, \tag{1.2.14}$$

and the *posterior expected squared prediction loss ratio* of $\hat{\boldsymbol{y}}_a$ to $\hat{\boldsymbol{y}}_e$ as

$$\mathcal{P}\left[\hat{\boldsymbol{y}}_i^A, \hat{\boldsymbol{y}}_i^E\right] \equiv \frac{\mathcal{R}\left[\hat{\boldsymbol{y}}_i^A\right]}{\mathcal{R}\left[\hat{\boldsymbol{y}}_i^E\right]}. \tag{1.2.15}$$

Substituting the $\tilde{\mathcal{R}}$ for $\mathcal{R}$ allows us to construct the *Bayesian mean squared parameter error ratio* of $\hat{\boldsymbol{\phi}}_i^A$ to $\hat{\boldsymbol{\phi}}_i^E$ as

$$\tilde{\mathcal{P}}\left[\hat{\boldsymbol{\phi}}_i^A, \hat{\boldsymbol{\phi}}_i^E\right] \equiv \frac{\tilde{\mathcal{R}}\left[\hat{\boldsymbol{\phi}}_i^A\right]}{\tilde{\mathcal{R}}\left[\hat{\boldsymbol{\phi}}_i^E\right]}, \tag{1.2.16}$$

and the *Bayesian mean squared prediction error ratio* of $\hat{\boldsymbol{y}}^A$ to $\hat{\boldsymbol{y}}^E$ as

$$\tilde{\mathcal{P}}\left[\hat{\boldsymbol{y}}_i^A, \hat{\boldsymbol{y}}_i^E\right] \equiv \frac{\tilde{\mathcal{R}}\left[\hat{\boldsymbol{y}}_i^A\right]}{\tilde{\mathcal{R}}\left[\hat{\boldsymbol{y}}_i^E\right]}, \tag{1.2.17}$$

which are both data independent measures which we can use to compare estimator accuracy of the MAP and MMSE. In this thesis, we formulate a modeling framework which includes correlated random effects for subject by task.

1.3 Literature review

Kay (1993) formulates the Bayesian linear model as

$$\boldsymbol{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{w} \tag{1.3.1}$$

where

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}_\theta, \boldsymbol{C}_{\boldsymbol{\theta}}\right) \tag{1.3.2}$$

and

$$\boldsymbol{w} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{C}_w\right). \tag{1.3.3}$$

For this model, Kay (1993) determines the posterior variance to be

$$\boldsymbol{C}_{\theta|x} = \left(\boldsymbol{C}_\theta^{-1} + \boldsymbol{H}'\boldsymbol{C}_w^{-1}\boldsymbol{H}\right)^{-1}. \tag{1.3.4}$$

Furthermore, Kay (1993) determines the MMSE estimator to be

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu}_\theta + \left(\boldsymbol{C}_\theta^{-1} + \boldsymbol{H}'\boldsymbol{C}_w^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}'\boldsymbol{C}_w^{-1}\left(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_\theta\right), \tag{1.3.5}$$

and the Bayesian MSE for the i^{th} parameter to be

$$\text{Bmse}(\hat{\theta}_i) = \left[\boldsymbol{C}_{\theta|x}\right]_{\text{ii}}. \tag{1.3.6}$$

In Section 3.1, we give an analogous derivation to obtain the posterior variance and Bayesian MSE for the Bayesian linear model formulated assuming $\boldsymbol{C}_w$ to be a diagonal matrix.

A linear model for multiple task variables may be considered under the linear Bayesian modeling framework considered by Kay (1993) when we further specify how different task variables will be denoted. For instance, Hall & Clutter (2004) specify a multivariate modeling framework to simultaneously model and make predictions of multiple measures of timber volume. These authors consider a multivariate nonlinear mixed effects model with 3 levels of grouping and r response variables. They let y_{ijkl} denote the l^{th} response variable at the k^{th} measurement time, on the j^{th} second-level group, and i^{th} first-level group, f_ℓ denote the model function, ϕ_{ijkl} denote a random effects vector for a particular response variable a the bottom level, and $\boldsymbol{v}_{\text{ijkl}}$ denote the covariate for the r^{th} response variable at the bottom level. Stacking the model equations for the r responses, they obtain,

$$\boldsymbol{y}_{\text{ijk}} = \boldsymbol{f}\left(\boldsymbol{\phi}_{\text{ijk}}, \boldsymbol{v}_{\text{ijk}}\right) + \epsilon_{\text{ijk}}, \tag{1.3.7}$$

where $\boldsymbol{y}_{\text{ijk}} = (y_{\text{ijk1}}, ..., y_{\text{ijkr}})'$ and $\boldsymbol{f}$, $\boldsymbol{\phi}_{\text{ijk}}$, $\boldsymbol{v}_{\text{ijk}}$, and $\boldsymbol{\epsilon}_{\text{ijk}}$ are defined in a similar manner. By defining a similar multivariate modeling framework, we may take advantage of the results of Kay (1993) to determine the Bayesian MSE where multiple tasks are concerned.

Including secondary variables in the manner of Hall & Clutter (2004), however, is not the only way in which they may be included. For instance, Chandler et al. (2013) estimated the ability to predict individual differences in cognitive impairment due to sleep loss using data from subject-specific measures on secondary tasks and information on the timing and duration of sleep. Sleep schedule information was captured using the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model, a biomathematical model that represents the temporal dynamics of cognitive performance impairment. The ability to predict individual performance was assessed by fitting a Generalized Linear Model (GLM) that combined Readiness Screening Tools (RSTs) with the SAFTE model to predict individual performance on the Psychomotor Vigilance Task (PVT). Individual performance measures from the Flight Fit cognitive test battery and from the PMI Fit screener were considered along with the Stanford Sleepiness Scale for inclusion in the GLM. The measures were each individually assessed using a series of Repeated Measures Analyses of Variance (ANOVAs) and a series of Hierarchical Linear Models (HLMs), and those measures which showed both significant variation over time, and significant correlation with PVT were included in the

final GLM. The final model utilized by Chandler was

$$\text{PVT} = -0.126\text{FAST} + 0.029\text{rawRT} + 0.03\text{daRT} + -0.242\text{shiftAcc} + 0.211\text{SV} \tag{1.3.8}$$

where PVT represents lapses on the PVT, FAST represents predicted performance effectiveness from the SAFTE model, rawRT represents reaction time on the PVT, daRT represents divided attention reaction time, shiftAcc represents attention shifting accuracy, and SV represents saccadic velocity. It was determined that the inclusion of subject-specific measures increased the model explanatory power form 13.8% to 35.7%.

Additionally, covariates may be included using the Kalman filter (Kay, 2013). The Kalman filter is the sequential Bayesian MMSE estimator for a discrete time signal embedded in noise. The estimator is derived by assuming discrete-time evolution of the state of the system modeled by a linear state transition model embedded in process noise. Process noise is noise which becomes recursively embedded in a state transition model as a system updates. The observation model is then embedded in white gaussian noise to create the observations. The Kalman filter is a sequential estimator, estimating the state at the n^{th} time point using only the estimate at the $n-1^{\text{th}}$ time point along with observations at time n. Kay (1993) specifies the Kalman filter model as

$$\begin{aligned} \boldsymbol{s}[n] &= \boldsymbol{A}\boldsymbol{s}[n-1] + \boldsymbol{B}\boldsymbol{u}[n], n \geq 0 \\ \boldsymbol{x}[n] &= \boldsymbol{h}'[n]\boldsymbol{s}[n] + w[n] \end{aligned} \tag{1.3.9}$$

where

$$\boldsymbol{u}[n] \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}), \tag{1.3.10}$$

$$\boldsymbol{s}[-1] \sim \mathcal{N}\left(\boldsymbol{\mu}_s, \boldsymbol{C}_s\right), \tag{1.3.11}$$

and

$$w[n] \sim \mathcal{N}\left(0, \sigma_n^2\right). \tag{1.3.12}$$

Secondary tasks may be used to help predict primary tasks in such a framework by letting $x[n]$ represent a vector of secondary tasks at time n, and $s[n]$ represent the primary task at time n. As we will see in the discussion section, with certain prior assumptions the Kalman filter can be viewed as the Bayesian MMSE estimator for a special case of the general linear Bayesian model. For this class of problems, the benefit of solving them under a Kalman filtering framework is the ability to make estimates sequentially in time in the most computationally efficient manner.

Chapter 2

Forecasting for Univariate Linear Models

A demonstration of how to utilize multiple tasks to increase the accuracy of predictions for a single task requires knowledge of how to assess prediction accuracy. In this chapter we demonstrate how to obtain the prediction accuracy for a single parameter linear Bayesian model by deriving the Bayesian MSE for both the MMSE and MAP estimators. This chapter is then used as a guiding framework to follow as the accuracy of the general linear Bayesian model obtained in Chapter 3.

The results of this chapter give the accuracy of the MAP and MMSE estimators, as measured using the Bayesian MSE. For the model considered, the MMSE and MAP estimators are equivalent, and therefore, so are their corresponding marginal mean squared parameter and prediction errors. The Bayesian MSE is first obtained by deriving the parameter posterior distribution and using it to obtain the response posterior distribution. We continue by showing that the parameter and response MMSE and MAP estimators are equivalent, and the posterior posterior expected squared loss can be taken as the variance of the corresponding posterior distribution. Finally, since this result is independent of the data, we find the Bayesian MSE to be the same as the posterior expected squared loss for this case.

2.1 Univariate linear Bayesian model of subject means

Model formulation

Suppose that y_{ik} is the k^{th} experimental observation for individual i measured at time t_k from performance outcome y. We first consider the model

$$\boldsymbol{y}_i = b_i + \boldsymbol{\epsilon}_i, \tag{2.1.1}$$

where $\boldsymbol{y}_i$ is a vector of m_i observed responses for a particular individual. Furthermore, b_i is a random effect used to model the mean response for a particular individual, and is assumed to arise from a normal distribution with mean μ and variance δ^2, both independent of the individual,

$$b_i \sim \mathcal{N}\left(\mu, \delta^2\right). \tag{2.1.2}$$

Lastly, $\boldsymbol{\epsilon}_i$ is a vector of additive measurement errors, which are independently and identically normally distributed with mean zero and variance σ^2,

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}\right). \tag{2.1.3}$$

Posterior parameter distribution

We can explicitly write the prior and conditional distributions as

$$f\left(b_i\right) = \frac{1}{\sqrt{2\pi}\delta} \exp\left[-\frac{\left(b_i - \mu\right)^2}{2\delta^2}\right] \tag{2.1.4}$$

and

$$f\left(\boldsymbol{y}_i | b_i\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m_i} \exp\left[-\sum_{k=1}^{m_i} \frac{\left(y_{\text{ik}} - b_i\right)^2}{2\sigma^2}\right]. \tag{2.1.5}$$

The posterior distribution for b_i can be obtained using Bayes Theorem:

$$f\left(b_i | \boldsymbol{y}_i\right) = \frac{f\left(\boldsymbol{y}_i | b_i\right) f\left(b_i\right)}{\int f\left(\boldsymbol{y}_i | b_i\right) f\left(b_i\right) db_i} = c_0 f\left(\boldsymbol{y} \,|b_i\,\right) f\left(b_i\right), \tag{2.1.6}$$

where c_0 does not depend on b_i. Including the exact distributional forms allows us to specify

$$f\left(b_i | \boldsymbol{y}_i\right) = c_0 \exp\left[-\frac{1}{2}\left[\sum_{k=1}^{m_i} \frac{\left(y_{\text{ik}} - b_i\right)^2}{\sigma^2} + \frac{\left(b_i - \mu\right)^2}{\delta^2}\right]\right]. \tag{2.1.7}$$

To show that the distribution is normal and determine the mean and variance, we expand the squares and collect the coefficients of the powers of b_i. We determine that

$$\begin{aligned} f\left(b_i|\boldsymbol{y}_i\right) &= c_0 \exp\left[-\frac{1}{2}\left[\frac{\sum_{k=1}^{m_i} y_{\mathrm{ik}}^2}{\sigma^2} - \frac{2b_i\sum_{k=1}^{m_i} y_{\mathrm{ik}}}{\sigma^2} + \frac{m_i b_i^2}{\sigma^2} + \frac{b_i^2}{\delta^2} - \frac{2b_i\mu}{\delta^2} + \frac{\mu^2}{\delta^2}\right]\right] \\ &= c_0 \exp\left[-\frac{1}{2}\left[b_i^2\left[\frac{m_i}{\sigma^2}+\frac{1}{\delta^2}\right] - 2b_i\left[\frac{\sum_{k=1}^{m_i} y_{\mathrm{ik}}}{\sigma^2}+\frac{\mu}{\delta^2}\right] + \left[\frac{\sum_{k=1}^{m_i} y_{\mathrm{ik}}^2}{\sigma^2}+\frac{\mu^2}{\delta^2}\right]\right]\right]. \end{aligned} \tag{2.1.8}$$

We define the coefficients,

$$c_1 = \frac{m_i}{\sigma^2} + \frac{1}{\delta^2}, \tag{2.1.9}$$

$$c_2 = \frac{\sum_{k=1}^{m_i} y_{\mathrm{ik}}}{\sigma^2} + \frac{\mu}{\delta^2}, \tag{2.1.10}$$

$$c_3 = \frac{\sum_{k=1}^{m_i} y_{\mathrm{ik}}^2}{\sigma^2} + \frac{\mu^2}{\delta^2}, \tag{2.1.11}$$

factor out the coefficient from b_i, and complete the square to get the expression into the form of a normal distribution, from which we can determine the mean and variance. This is done as

$$\begin{aligned} f\left(b_i|\boldsymbol{y}_i\right) &= c_0 \exp\left[-\frac{1}{2}\left[b_i^2 c_1 - 2b_i c_2 + c_3\right]\right] \\ &= c_0 \exp\left[-\frac{1}{2}\frac{\left[b_i^2 - \frac{2b_i c_2}{c_1} + \frac{c_3}{c_1}\right]}{c_1^{-1}}\right] \\ &= c_0 \exp\left[-\frac{1}{2}\frac{\left[b_i^2 - \frac{2b_i c_2}{c_1} + \left(\frac{c_2}{c_1}\right)^2\right] + \frac{c_3}{c_1} - \left(\frac{c_2}{c_1}\right)^2}{c_1^{-1}}\right] \\ &= c_0 \exp\left[-\frac{1}{2}\frac{\left[b_i - \frac{c_2}{c_1}\right]^2}{c_1^{-1}}\right] \exp\left[-\frac{c_4 c_1}{2}\right], \end{aligned} \tag{2.1.12}$$

where

$$c_4 = \frac{c_3}{c_1} - \left(\frac{c_2}{c_1}\right)^2. \tag{2.1.13}$$

The last exponential on the right does not depend on b_i and we include it in the constant c_0 without changing the notation for c_0. Substituting in the coefficients c_1, c_2, c_3 allows us to determine that

$$f\left(b_i|\boldsymbol{y}_i\right) = c_0 \exp\left[-\frac{1}{2}\left[\frac{\left(b_i - \left[\frac{\sum_{k=1}^{m} y_{\mathrm{ik}}}{\sigma^2} + \frac{\mu}{\delta^2}\right] / \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right]\right)^2}{\left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right]^{-1}}\right]\right], \tag{2.1.14}$$

from which it follows that (Kay, 1993, p. 319)

$$b_i|\boldsymbol{y}_i \sim \mathcal{N}\left(\left[\frac{\sum_{k=1}^{m_i} y_{\text{ik}}}{\sigma^2} + \frac{\mu}{\delta^2}\right] / \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right], \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right]^{-1}\right). \tag{2.1.15}$$

As we will consider further in Chapter 4, the normality of the posterior distribution does not generalize to the nonlinear case.

Posterior response distribution

From the posterior distribution for b, we can move to the posterior response distribution for responses $\boldsymbol{y}^*$ that have yet to be observed. As $E[\epsilon] = 0$, the mean of the posterior response distribution will be the same as that of the posterior for b. The variance is computed by adding the error variance to the posterior variance. The posterior response distribution is thus,

$$\boldsymbol{y}_i^*|\boldsymbol{y}_i \sim \mathcal{N}\left(\left[\frac{\sum_{k=1}^{m_i} y_{\text{ik}}}{\sigma^2} + \frac{\mu}{\delta^2}\right] / \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right], \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right]^{-1} + \sigma^2\right). \tag{2.1.16}$$

Parameter MMSE

As it is our eventual aim to estimate b and, subsequently, the unobserved responses $\boldsymbol{y}^*$, let us consider two estimation options and assess the accuracy of the estimates. As determined in Kay (1993, p. 319), the MMSE is found to be simply

$$\hat{b}_i^E = E_{b_i|\boldsymbol{y}_i}[b_i] = \left[\frac{\sum_{k=1}^{m_i} y_{\text{ik}}}{\sigma^2} + \frac{\mu}{\delta^2}\right] / \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right]. \tag{2.1.17}$$

Parameter MAP

The MAP estimator is defined as

$$\hat{b}_i^A = \underset{b_i}{\text{argmax}}\left[f\left(b_i|\boldsymbol{y}_i\right)\right]. \tag{2.1.18}$$

Since for this model, the posterior is normal and, thus, achieves a maximal value at the mean, we find that the MAP and MMSE are equivalent (Kay, 1993, p. 358),

$$\hat{b}_i \equiv \hat{b}_i^A = \hat{b}_i^E. \tag{2.1.19}$$

As will be considered in Chapter 4, this property does not, in general, extend to the nonlinear case.

Response MMSE

As for b_i, the MMSE estimate of unobserved response values $\boldsymbol{y}_i^*$ can be found from its relevant posterior, yielding

$$\hat{\boldsymbol{y}_i^E} = E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\boldsymbol{y}_i^*\right] = \left[\frac{\sum_{k=1}^{m_i} y_{\text{ik}}}{\sigma^2} + \frac{\mu}{\delta^2}\right] / \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right]. \tag{2.1.20}$$

Since in this case, the mean of the parameter posterior is the same as that of the posterior response distribution, we find that

$$\hat{\boldsymbol{y}_i^E} = \hat{b}_i. \tag{2.1.21}$$

This also will not hold in the nonlinear case.

Response MAP

Again as for b_i, the MAP estimate for the response will be

$$\hat{\boldsymbol{y}_i^A} = \underset{b_i}{\text{argmax}}\left[f\left(\boldsymbol{y}_i^*|\boldsymbol{y}_i\right)\right], \tag{2.1.22}$$

and due to the normality, it is equivalent to the MMSE, or more concisely,

$$\hat{\boldsymbol{y}}_i \equiv \hat{\boldsymbol{y}_i^A} = \hat{\boldsymbol{y}_i^E}. \tag{2.1.23}$$

Parameter MMSE and MAP accuracy

We next consider the accuracy of the estimators, as measured by the posterior expected squared loss. Using the variance computing formula, we find that

$$\begin{aligned}\mathcal{R}\left[\hat{b_i}\right] &= E_{b_i|\boldsymbol{y}_i}\left[\left(\hat{b}_i - b_i\right)^2\right] \\ &= \text{Var}_{b_i|\boldsymbol{y}_i}\left[\hat{b_i} - b_i\right] + E_{b_i|\boldsymbol{y}_i}\left[\hat{b_i} - b_i\right]^2.\end{aligned} \tag{2.1.24}$$

Furthermore, we note that since $b_i|\boldsymbol{y}_i$ is normally distributed as given in (2.1.15), it will have mean $\hat{b_i} \equiv \left[\frac{\sum_{k=1}^{m_i} y_{\text{ik}}}{\sigma^2} + \frac{\mu}{\delta^2}\right] / \left[\frac{m_i}{\sigma^2} + \frac{1}{\delta^2}\right]$. Therefore, $E_{b_i|\boldsymbol{y}_i}\left[\hat{b_i} - b_i\right] = 0$. This allows us to determine that

$$\mathcal{R}\left[\hat{b_i}\right] = \text{Var}_{b_i|\boldsymbol{y}_i}\left[\hat{b_i} - b_i\right]. \tag{2.1.25}$$

Finally, as noted in Kay (1993, p. 320), since $\hat{b_i}$ is fixed for $\boldsymbol{y}_i$ given,

$$\mathcal{R}\left[\hat{b_i}\right] = \text{Var}_{b_i|\boldsymbol{y}_i}\left[b_i\right]. \tag{2.1.26}$$

Therefore the posterior expected squared parameter loss is equivalent to the variance of the posterior distribution. This result holds true for the nonlinear case.

Parameter MMSE and MAP accuracy over data

Sometimes, our interest is not in the evaluation of an estimator for a specific dataset, but instead in the evaluation of an estimator over all datasets generated by a certain modeling process. To accomplish this, we compute the Bayesian MSE. This can be computed using iterated expectations as

$$\begin{aligned}\tilde{\mathcal{R}}\left[\hat{b_i}\right] &= E_{b_i,\boldsymbol{y}_i}\left[\left(\hat{b_i}-b_i\right)^2\right] \\ &= E_{\boldsymbol{y}_i}\left[E_{b_i|\boldsymbol{y}_i}\left[\left(\hat{b_i}-b_i\right)^2\right]\right]. \\ &= E_{\boldsymbol{y}_i}\left[\mathrm{Var}_{b_i|\boldsymbol{y}_i}\left[b_i\right]\right]\end{aligned} \tag{2.1.27}$$

Finally, as shown in Kay (1993, p. 320), since for this model $\mathrm{Var}_{b_i|\boldsymbol{y}_i}\left[b_i\right]$ is independent of $\boldsymbol{y}_i$,

$$\tilde{\mathcal{R}}\left[\hat{b_i}\right] = \mathrm{Var}_{b_i|\boldsymbol{y}_i}\left[b_i\right]. \tag{2.1.28}$$

Thus, the Bayesian MSE for the MMSE and MAP estimators is the same as the posterior expected squared loss. In summary, this comes back to the following ideas: the posterior is normal, and so the MAP is equal to the MMSE. The MMSE estimator is unbiased, so the posterior expected squared loss of the MMSE estimator is just the posterior variance. Finally, the posterior variance does not depend on the data values, but rather, just the number of data points, and so taking the expectation over $\boldsymbol{y}$ has no effect. This result does not hold in the general nonlinear case, as the variance can depend on the data.

Response MMSE and MAP accuracy

We repeat the same procedure to determine the accuracy of the predictors. Using the variance computing formula, we find that

$$\begin{aligned}\mathcal{R}\left[\hat{\boldsymbol{y}}_i\right] &= E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[(\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i^*)^2\right] \\ &= \mathrm{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i^*\right]+\left(E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i^*\right]\right)^2 \\ &= \mathrm{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\boldsymbol{y}_i^*\right]+\left(E_{b_i|\boldsymbol{y}_i}\left[\hat{b_i}-b_i-\boldsymbol{\epsilon}_i\right]\right)^2 \\ &= \mathrm{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\boldsymbol{y}_i^*\right].\end{aligned} \tag{2.1.29}$$

Therefore the posterior expected squared prediction loss for the MAP and MMSE estimators is equivalent to the variance of the posterior response distribution.

Response MMSE and MAP accuracy over data

Taking the expectation over $\boldsymbol{y}_i$, we find that

$$\begin{aligned}\tilde{\mathcal{R}}\left[\hat{\boldsymbol{y}}_i\right] &= E_{\boldsymbol{y}_i^*,\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i^*\right)^2\right] \\ &= E_{\boldsymbol{y}_i}\left[E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i^*\right)^2\right]\right] \\ &= E_{\boldsymbol{y}_i}\left[\mathrm{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\boldsymbol{y}_i^*\right]\right].\end{aligned} \tag{2.1.30}$$

Since for this model $\mathrm{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\boldsymbol{y}_i^*\right]$ is independent of $\boldsymbol{y}_i$, we find that

$$E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i^*\right)^2\right] = \mathrm{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[\boldsymbol{y}_i^*\right]. \tag{2.1.31}$$

Thus, the Bayesian MSE of the MMSE and MAP response estimators is the same as the posterior expected squared loss for any given dataset.

Summary

For Bayesian forecasting scenarios which can appropriately be explained by the simple model from this section, (2.1.16) and (2.1.31) can be used to solve for the number of data points required to obtain a certain level of accuracy. Specifically,

$$m = \sigma^2\left(\tilde{\mathcal{R}}^{-1}-\frac{1}{\delta^2}\right), \tag{2.1.32}$$

where $\tilde{\mathcal{R}}^{-1}$ represents the desired level of risk as measured by the Bayesian MSE.

In this chapter, we have demonstrated how to assess the accuracy of the predictions and determine the number of data points required to obtain a fixed prediction accuracy for a single task for the univariate linear Bayesian model of subject means. This chapter will be used as a guiding framework to follow for future chapters.

Chapter 3

Forecasting for Multivariate Linear Models

In the previous chapter, we demonstrated how to assess prediction accuracy for the response MMSE and MAP for a univariate linear Bayesian model with pre-specified priors, given performance data for a single individual on a single task. In this Chapter, we include secondary tasks and reassess the prediction accuracy on the primary task with such secondary tasks included. In Section 3.1 we introduce the general linear Bayesian model and demonstrate how to assess the accuracy of predictions for multiple task variables. This model differs from that formulated previously not only in allowing for multiple task variables, but also in that it includes fixed effects and arbitrary design matrices, which for simplicity were not yet included in the last chapter. In Section 3.2, we consider a special case of this model which includes only fixed subject means and random subject-specific, task-specific means, and solve for the accuracy assessment in a manner that reduces the dimensionality of the required matrix inversion. In Section 3.3 we examine how task variance components (i.e. the task specific between subject variances, the between subjects correlation, and the error variance) influence prediction accuracy, in order to enable informed decisions when selecting particular task variables. We also consider how prediction accuracy depends on primary and secondary task sample size in Section 3.3 and, in Section 3.4, how to use this information to minimize the cost of data collection for a desired level of prediction accuracy.

3.1 General linear Bayesian model

Here we formulate the general linear mixed model and demonstrate how to assess the accuracy of predictions, following the framework of the last section. With reference to the nonlinear cases in chapters to follow, we focus on both MMSE and MAP estimators. Our results give the Bayesian MSE of the MMSE and MAP, which we find to be equivalent again.

Let y_{ijk} be the k^{th} observation for an individual i measured on performance task j, where performance tasks range from $j = 1, .., s$, and nested measurement times range from $k = 1, .., m_{\mathrm{ij}}$. The total number of measurements for the individual is $m_i = \sum_{j=1}^{s} m_{\mathrm{ij}}$. Let us consider the subject-specific model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{3.1.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of effects which are estimated in the population stage of estimation and do not vary over individual. The remaining terms are subject-specific. $\boldsymbol{X}_i$ is the design matrix specifying the relationship between $\boldsymbol{\beta}$ and the response vector, $\boldsymbol{b}_i$ is a $q \times 1$ parameter vector, $\boldsymbol{Z}_i$ is the design matrix specifying a linear relationship between the Bayesian parameters and the response vector, and $\boldsymbol{\epsilon}_i$ represents the measurement error. The individual response vector, $\boldsymbol{y}_i$, is of length m_i and is ordered such that time is varied first, and task second. To illustrate, for $(m_i, m_{\mathrm{i1}}, m_{\mathrm{i2}}) = (2, 2, 3)$ this would result in the response vector

$$\boldsymbol{y}_i = \begin{pmatrix} y_{\mathrm{i11}} \\ y_{\mathrm{i12}} \\ y_{\mathrm{i21}} \\ y_{\mathrm{i22}} \\ y_{\mathrm{i23}} \end{pmatrix}. \tag{3.1.2}$$

For model (3.1.1) we make two distributional assumptions. First, we assume a normal distribution on the random effects,

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \tag{3.1.3}$$

where $\boldsymbol{D}$ is assumed positive definite. This assumption is made without loss of generality, since an indefinite model can always be reformulated as a positive-definite model with lower dimension (Pinheiro & Bates, 2009, p. 58). This distributional assumption allows us to

write the probability distribution for $\boldsymbol{b}_i$ as

$$f\left(\boldsymbol{b}_i\right) = \frac{\exp\left[-\frac{1}{2}\boldsymbol{b}_i{}'\boldsymbol{D}^{-1}\boldsymbol{b}_i\right]}{(2\pi)^{q/2}|\boldsymbol{D}|^{\frac{1}{2}}}. \tag{3.1.4}$$

Secondly, we assume a normal distribution on the errors,

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_i), \tag{3.1.5}$$

where $\boldsymbol{\Lambda_i}$ is also assumed positive definite. The distribution of the responses given the random effects is normal:

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Lambda}_i), \tag{3.1.6}$$

which implies

$$f\left(\boldsymbol{y}_i|\boldsymbol{b}_i\right) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\right)'\boldsymbol{\Lambda}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\right)\right]}{(2\pi)^{m_{\mathrm{ij}}/2}|\boldsymbol{\Lambda}_i|^{\frac{1}{2}}} \tag{3.1.7}$$

where $\boldsymbol{\Lambda}_i$ is the symmetric, positive definite correlation matrix for the errors.

The posterior distribution for $\boldsymbol{b}_i$ can be obtained using Bayes Theorem, from which we know that

$$f\left(\boldsymbol{b}_i|\boldsymbol{y}_i\right) = c_0 \cdot f\left(\boldsymbol{y}_i|\boldsymbol{b}_i\right) f\left(\boldsymbol{b}_i\right), \tag{3.1.8}$$

where c_0 is the normalization constant which does not depend on $\boldsymbol{b}_i$. Including the exact distributional forms allows us to specify

$$f\left(\boldsymbol{b}_i|\boldsymbol{y}_i\right) = c_0 \exp\left[-\frac{1}{2}\left[\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\right)'\boldsymbol{\Lambda}_i{}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\right) + \boldsymbol{b}_i{}'\boldsymbol{D}^{-1}\boldsymbol{b}_i\right]\right]. \tag{3.1.9}$$

which can be expanded to yield

$$f\left(\boldsymbol{b}_i|\boldsymbol{y}_i\right) = c_0 \exp\left[-\frac{1}{2}\left[\begin{array}{l}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)'\boldsymbol{\Lambda}_i{}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right) - \boldsymbol{b}_i{}'\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i{}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right) \\ -\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{b}_i{}'\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{b}_i{}'\boldsymbol{D}_i{}^{-1}\boldsymbol{b}_i\end{array}\right]\right]. \tag{3.1.10}$$

Noting that $\boldsymbol{b}_i{}'\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i{}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)$ is a scalar allows us to determine that it is equal to its transpose

$$\boldsymbol{b}_i{}'\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i{}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right) = \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_i, \tag{3.1.11}$$

and therefore,

$$f\left(\boldsymbol{b}_i|\boldsymbol{y}_i\right) = c_0 \exp\left[-\frac{1}{2}\left[\begin{array}{l}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)'\boldsymbol{\Lambda}_i{}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right) - 2\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_i \\ + \boldsymbol{b}_i{}'\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{b}_i{}'\boldsymbol{D}_i{}^{-1}\boldsymbol{b}_i\end{array}\right]\right]. \tag{3.1.12}$$

The term $(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' \boldsymbol{\Lambda}_i{}^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ is constant w.r.t. $\boldsymbol{b}_i$ and can therefore be absorbed by the constant c_0, so that

$$f(\boldsymbol{b}_i|\boldsymbol{y}_i) = c_0 \exp\left[-\frac{1}{2}\left[\boldsymbol{b}_i'(\boldsymbol{Z}_i'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i + \boldsymbol{D}^{-1})\boldsymbol{b}_i - 2(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_i\right]\right]. \quad (3.1.13)$$

Using results from Appendix B and noting that since the inverse of a symmetric matrix is itself symmetric, $\boldsymbol{\Lambda}_i{}^{-1}$ is symmetric, we find (as noted in Kay (1993, p. 328)) that the posterior distribution is normal,

$$f(\boldsymbol{b}_i|\boldsymbol{y}_i) \sim \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_i, \tilde{\boldsymbol{D}}_i\right), \quad (3.1.14)$$

where

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\boldsymbol{D}}_i\boldsymbol{Z}_i'\boldsymbol{\Lambda}_i{}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \quad (3.1.15)$$

and

$$\tilde{\boldsymbol{D}}_i = \left(\boldsymbol{Z}_i'\boldsymbol{\Lambda}_i{}^{-1}\boldsymbol{Z}_i + \boldsymbol{D}^{-1}\right)^{-1}. \quad (3.1.16)$$

In (3.1.14) we found the distribution of $\boldsymbol{b}_i$. However, it is often the case that we are interested only in those elements of $\boldsymbol{b}_i$ that will be used in making predictions for certain tasks. We will assume that the data have been organized so that we wish to predict only tasks $1...a, a \leq s$. To partition the model, we define a matrix $\boldsymbol{P}$ to extract the first q_a out of q rows of a matrix. We let

$$\boldsymbol{P} = \left(\begin{array}{cc} \boldsymbol{I}_{q_a} & \boldsymbol{0}_{q_a \times (q-q_a)} \end{array}\right), \quad (3.1.17)$$

where q_a is the number of random effects used to model the first a tasks, and $\boldsymbol{0}_{q\times(q-q_a)}$ is a $q_a \times (q - q_a)$ matrix of zeros. Then the parameters of interest for making predictions are $\boldsymbol{b}_{\text{ia}} = \mathbf{P}\mathbf{b}_i$. Using results from Appendix B, we find that the marginal posterior for $\boldsymbol{b}_{\text{ia}}$ is normal:

$$\boldsymbol{b}_{\text{ia}} \sim \mathcal{N}\left(\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i, \boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}'\right), \quad (3.1.18)$$

where the mean and variance are directly extracted from the first q_a rows of $\tilde{\boldsymbol{\beta}}_i$, and from the first q_a rows and q_a columns of $\tilde{\boldsymbol{D}}_i$.

From the posterior distribution for $\boldsymbol{b}_{\text{ia}}$, we can move to the posterior response distribution for responses $\boldsymbol{y}_i^*$ that have yet to be observed. We let $\boldsymbol{X}_i^*, \boldsymbol{Z}_i^*$ be the design matrices for $\boldsymbol{y}_i^*$. Then (3.1.1) becomes:

$$\boldsymbol{y}_i^* = \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{b}_{\text{ia}} + \boldsymbol{\epsilon}_i^*, \quad (3.1.19)$$

which allows us to determine the expected value and variance as follows:

$$\begin{aligned} E\left[\boldsymbol{y}_i^*\right] &= E\left[X_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{b}_{\text{ia}} + \boldsymbol{\epsilon}_i^*\right] \\ &= E\left[\boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta}\right] + E\left[\boldsymbol{Z}_i^*\boldsymbol{b}_{\text{ia}}\right] \\ &= \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^* E\left[\boldsymbol{b}_{\text{ia}}\right] \\ &= \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i, \end{aligned} \tag{3.1.20}$$

$$\begin{aligned} \text{Var}\left[\boldsymbol{y}_i^*\right] &= \text{Var}\left[\boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{b}_{\text{ia}} + \boldsymbol{\epsilon}_i^*\right] \\ &= \text{Var}\left[\boldsymbol{Z}_i^*\boldsymbol{b}_{\text{ia}}\right] + \text{Var}\left[\boldsymbol{\epsilon}_i^*\right] \\ &= \boldsymbol{Z}_i^*\text{Var}\left[\boldsymbol{b}_{\text{ia}}\right]\left(\boldsymbol{Z}_i^*\right)' + \boldsymbol{\Lambda}_i^* \\ &= \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}'(\boldsymbol{Z}_i^*)' + \boldsymbol{\Lambda}_i^*. \end{aligned} \tag{3.1.21}$$

Thus, the posterior distribution for the unobserved responses is

$$f\left(\boldsymbol{y}_i^*|\boldsymbol{y}_i\right) \sim \mathcal{N}\left(\boldsymbol{X}_i^*\boldsymbol{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i, \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}'(\boldsymbol{Z}_i^*)' + \boldsymbol{\Lambda}_i^*\right). \tag{3.1.22}$$

Before moving to the estimators, we obtain the joint distribution of $(\boldsymbol{y}_i^*, \boldsymbol{b}_{\text{ia}}|\boldsymbol{y}_i)$. The motivation is that in the non-linear case (to be discussed later), we find that this distribution is easy to maximize, and we want to obtain the linear case equivalence to $f\left(\boldsymbol{y}_i^*|\boldsymbol{y}_i\right)$. We first note that since we have assumed the errors and Bayesian parameters to be normal and independent of one-another, we know that

$$\begin{pmatrix} \boldsymbol{\epsilon}_i^* \\ \boldsymbol{b}_{\text{ia}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \boldsymbol{P}\tilde{\boldsymbol{\beta}}_i \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Lambda}_i^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}\tilde{\boldsymbol{\beta}}_i \end{pmatrix}\right). \tag{3.1.23}$$

Using results from Appendix B, we can construct the vector of interest through the linear transformation

$$\begin{pmatrix} \boldsymbol{y}_i^* \\ \boldsymbol{b}_{\text{ia}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{m_{\text{ia}}} & \boldsymbol{Z}_i^* \\ \mathbf{0}_{a\times m_{\text{ia}}} & \boldsymbol{I}_a \end{pmatrix}\begin{pmatrix} \boldsymbol{\epsilon}_i^* \\ \boldsymbol{b}_{\text{ia}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{X}_i^*\boldsymbol{\beta} \\ \mathbf{0}_{a\times a} \end{pmatrix}. \tag{3.1.24}$$

We then know that

$$\begin{pmatrix} \boldsymbol{y}_i^* \\ \boldsymbol{b}_{\text{ia}} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.1.25}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{I}_{m_{\text{ia}}} & \boldsymbol{Z}_i^* \\ \mathbf{0}_{a\times m_{\text{ia}}} & \boldsymbol{I}_a \end{pmatrix}\begin{pmatrix} \boldsymbol{\Lambda}_i^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}\tilde{\boldsymbol{\beta}}_i \end{pmatrix}\begin{pmatrix} \boldsymbol{I}_{m_{\text{ia}}} & \boldsymbol{Z}_i^* \\ \mathbf{0}_{a\times m_{\text{ia}}} & \boldsymbol{I}_a \end{pmatrix}. \tag{3.1.26}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{I}_{m_{\text{ia}}} & \boldsymbol{Z}_i^* \\ \mathbf{0}_{a\times m_{\text{ia}}} & \boldsymbol{I}_a \end{pmatrix}\begin{pmatrix} \boldsymbol{\Lambda}_i^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}\tilde{\boldsymbol{\beta}}_i \end{pmatrix}\begin{pmatrix} \boldsymbol{I}_{m_{\text{ia}}} & \boldsymbol{Z}_i^* \\ \mathbf{0}_{a\times m_{\text{ia}}} & \boldsymbol{I}_a \end{pmatrix}'. \tag{3.1.27}$$

As it is our eventual aim to estimate the parameters or new responses, let us consider two estimation options and assess the accuracy of the estimates. The Bayesian MMSE parameter estimate minimizes the squared error loss function, and is found to be simply

$$\hat{\boldsymbol{b}_{\text{ia}}} \equiv \hat{\boldsymbol{b}_{\text{ia}}^{E}} = \boldsymbol{P}\tilde{\boldsymbol{\beta}}_i. \tag{3.1.28}$$

As the distribution is normal, the MAP and MMSE are equal (Kay, 1993, p. 358),

$$\hat{\boldsymbol{b}_{\text{ia}}^{A}} = \hat{\boldsymbol{b}_{\text{ia}}^{E}}. \tag{3.1.29}$$

Note also that the MMSE for estimating the full random effects vector would be

$$\hat{\boldsymbol{b}_i^{A}} = \tilde{\boldsymbol{\beta}}_i \tag{3.1.30}$$

and

$$\hat{\boldsymbol{b}_i} \equiv \hat{\boldsymbol{b}_i^{A}} = \hat{\boldsymbol{b}_i^{E}}. \tag{3.1.31}$$

Therefore,

$$\hat{\boldsymbol{b}_{\text{ia}}} \equiv \boldsymbol{P}\tilde{\boldsymbol{\beta}}_i = \boldsymbol{P}\hat{\boldsymbol{b}_i}, \tag{3.1.32}$$

so the estimate of $\hat{\boldsymbol{b}_{\text{ia}}}$ is just the extraction of the first a rows of $\hat{\boldsymbol{b}_i}$.
As for $\boldsymbol{b}_{\text{ia}}$, the MMSE estimate of unobserved response values, $\boldsymbol{y}_i^*$, can be found from its the posterior of $\boldsymbol{y}_i^*$,

$$\hat{\boldsymbol{y}_i} \equiv \hat{\boldsymbol{y}_i^{E}} = \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i. \tag{3.1.33}$$

Since the posterior response distribution is normal,

$$\hat{\boldsymbol{y}_i} \equiv \hat{\boldsymbol{y}_i^{A}} = \hat{\boldsymbol{y}_i^{E}}. \tag{3.1.34}$$

Note that

$$\begin{aligned} \xi\left(\boldsymbol{\beta}, \hat{\boldsymbol{b}_{\text{ia}}}, \boldsymbol{x}_i\right) &= \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\hat{\boldsymbol{b}_{\text{ia}}} \\ &= \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i \\ &= \hat{\boldsymbol{y}_i}, \end{aligned} \tag{3.1.35}$$

which tells us that the estimator for the posterior response is equal to the model function evaluated at the estimate for the random effects. Concerning the joint MAP of $(\boldsymbol{y}_i^*, \boldsymbol{b}_{\text{ia}})$, we can determine this by maximizing $f\left(\boldsymbol{y}_i^*, \boldsymbol{b}_{\text{ia}} | \boldsymbol{y}_i\right)$. Since this distribution is normal, it will yield a maximum value at

$$\begin{pmatrix} \boldsymbol{I}_{m_{\text{ia}}} & \boldsymbol{Z}_i^* \\ \mathbf{0}_{a\times m_{\text{ia}}} & \boldsymbol{I}_a \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \boldsymbol{P}\boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} \\ \mathbf{0}_{a\times a} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i \\ \boldsymbol{P}\tilde{\boldsymbol{\beta}}_i \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{y}_i} \\ \hat{\boldsymbol{b}_{\text{ia}}} \end{pmatrix}. \tag{3.1.36}$$

Therefore, maximizing $f\left(\boldsymbol{y}_i^*, \boldsymbol{b}_{\mathrm{ia}} | \boldsymbol{y}_i\right)$ will yield the same estimates as separately maximizing $f\left(\boldsymbol{y}_i^* | \boldsymbol{y}_i\right)$ and $f\left(\boldsymbol{b}_{\mathrm{ia}} | \boldsymbol{y}_i\right)$.

We next consider the accuracy of the estimator of $\boldsymbol{b}_{\mathrm{ia}}$, as measured by the expected squared loss. We define

$$E_x[g(x)] \equiv \int_{-\infty}^{\infty} g(x) f(x) dx, \tag{3.1.37}$$

where $f(x)$ is the probability distribution function for x, and

$$\mathrm{Var}_x[g(x)] \equiv E_x\left[(g(x) - E[g(x)])^2\right]. \tag{3.1.38}$$

The posterior expected squared loss is

$$\mathcal{R}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}}\right] = \mathrm{tr}\left\{E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)'\right]\right\}. \tag{3.1.39}$$

Using the standard formula for computing the variance, we have that

$$\mathcal{R}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}}\right] = \mathrm{tr}\left\{\mathrm{Var}_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right] + E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right] E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right]'\right\}. \tag{3.1.40}$$

Noting that

$$E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right] = E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}}\right] - E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\boldsymbol{b}_{\mathrm{ia}}\right] = \hat{\boldsymbol{b}_{\mathrm{ia}}} - E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\boldsymbol{b}_{\mathrm{ia}}\right] = \hat{\boldsymbol{b}_{\mathrm{ia}}} - \hat{\boldsymbol{b}_{\mathrm{ia}}} = 0, \tag{3.1.41}$$

we find that

$$\begin{aligned} \mathcal{R}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}}\right] &= \mathrm{tr}\left\{\mathrm{Var}_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right]\right\} \\ &= \mathrm{tr}\left\{\mathrm{Var}_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\boldsymbol{b}_{\mathrm{ia}}\right]\right\}, \end{aligned} \tag{3.1.42}$$

since $\hat{\boldsymbol{b}_{\mathrm{ia}}}$ is fixed for $\boldsymbol{y}_i$ given. Taking the expectation over all datasets and using iterated expectations we find that:

$$\begin{aligned} \tilde{\mathcal{R}}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}}\right] &= \mathrm{tr}\left\{E_{\boldsymbol{b}_{\mathrm{ia}},\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)'\right]\right\} \\ &= \mathrm{tr}\left\{E_{\boldsymbol{y}_i}\left[E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)'\right]\right]\right\} \\ &= E_{\boldsymbol{y}_i}\left[\mathrm{tr}\left\{E_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)\left(\hat{\boldsymbol{b}_{\mathrm{ia}}} - \boldsymbol{b}_{\mathrm{ia}}\right)'\right]\right\}\right] \\ &= E_{\boldsymbol{y}_i}\left[\mathcal{R}\left[\hat{\boldsymbol{b}_{\mathrm{ia}}}\right]\right] \\ &= E_{\boldsymbol{y}_i}\left[\mathrm{tr}\left\{\mathrm{Var}_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\boldsymbol{b}_{\mathrm{ia}}\right]\right\}\right] \\ &= \mathrm{tr}\left\{E_{\boldsymbol{y}_i}\left[\mathrm{Var}_{\boldsymbol{b}_{\mathrm{ia}}|\boldsymbol{y}_i}\left[\boldsymbol{b}_{\mathrm{ia}}\right]\right]\right\}. \end{aligned} \tag{3.1.43}$$

Applying (3.1.42) and noting that by (3.1.18) $\text{Var}_{\boldsymbol{b}_{\text{ia}}|\boldsymbol{y}_i}[\boldsymbol{b}_{\text{ia}}]$ is independent of $\boldsymbol{y}$ we find that (as given in Kay (1993, p. 391))

$$\begin{aligned}\tilde{\mathcal{R}}\left[\hat{\boldsymbol{b}_{\text{ia}}}\right] &= \text{tr}\left\{\text{Var}_{\boldsymbol{b}_{\text{ia}}|\boldsymbol{y}_i}[\boldsymbol{b}_{\text{ia}}]\right\} \\ &= \mathcal{R}\left[\hat{\boldsymbol{b}_{\text{ia}}}\right].\end{aligned} \tag{3.1.44}$$

Thus, the Bayesian MSE of $\hat{\boldsymbol{b}_{\text{ia}}}$ is the same as the posterior expected squared loss.

We repeat the same procedure to determine the accuracy of the predictors of $\boldsymbol{y}_i^*$. Using the variance computing formula, we find that

$$\begin{aligned}\mathcal{R}[\hat{\boldsymbol{y}}_i] &= \text{tr}\left\{E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*)(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*)'\right]\right\} \\ &= \text{tr}\left\{\text{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*] + E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*]\, E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*]'\right\}.\end{aligned} \tag{3.1.45}$$

The expectation on the right is zero, since $E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\boldsymbol{y}_i^*] = \boldsymbol{X}_i^*\mathbf{P}\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i = \hat{\boldsymbol{y}}_i$. As such,

$$\mathcal{R}[\hat{\boldsymbol{y}}_i] = \text{tr}\left\{\text{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\boldsymbol{y}_i^*]\right\}. \tag{3.1.46}$$

Therefore the posterior expected squared loss is equivalent to the trace of the variance of the posterior response distribution. Taking the expectation over $\boldsymbol{y}_i$ and using (3.1.46), we find that

$$\begin{aligned}\tilde{\mathcal{R}}[\hat{\boldsymbol{y}}_i] = \text{tr}\left\{E_{\boldsymbol{y}_i^*,\boldsymbol{y}_i}\left[(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*)(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*)'\right]\right\} &= \text{tr}\left\{E_{\boldsymbol{y}_i}\left[E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*)^2\right]\right]\right\} \\ &= E_{\boldsymbol{y}_i}\left[\text{tr}\left\{E_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}\left[(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i^*)^2\right]\right\}\right] \\ &= E_{\boldsymbol{y}_i}\left[\mathcal{R}[\hat{\boldsymbol{y}}_i]\right] \\ &= E_{\boldsymbol{y}_i}\left[\text{tr}\left\{\text{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\boldsymbol{y}_i^*]\right\}\right].\end{aligned} \tag{3.1.47}$$

Since for this model, $\text{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\boldsymbol{y}_i^*]$ is independent of $\boldsymbol{y}_i$, we find that

$$\tilde{\mathcal{R}}[\hat{\boldsymbol{y}}_i] = \text{tr}\left\{\text{Var}_{\boldsymbol{y}_i^*|\boldsymbol{y}_i}[\boldsymbol{y}_i^*]\right\}. \tag{3.1.48}$$

Thus, the Bayesian MSE $\hat{\boldsymbol{y}}_i$ is the same as the MSE for any given dataset.

3.2 Multivariate Bayesian model of subject means

In the last section we analytically determined the Bayesian MSE of both the MMSE and MAP estimators over datasets for the general linear Bayesian model and found that these expressions require matrix inversion. In certain special cases, the dimensions of the matrix

that must be inverted can be reduced in size. In this chapter, we consider such a special case, and reformulate the Bayesian MSE of both the MMSE and MAP estimators with the reduced size matrix inversion.

The model of subject means can be defined as

$$y_{\mathrm{ijk}} = \beta_j + b_{\mathrm{ij}} + \epsilon_{\mathrm{ijk}}, \tag{3.2.1}$$

where β_j is a fixed effect to differentiate tasks, b_{ij} is a random effect for individual i and task j, and ϵ_{ijk} is the additive measurement error, where k indexes time.

To apply the results of Section 3.1 to this model, we specify it as a subset of the General Linear Multivariate Mixed Model (3.1.1), where we make certain stipulations on the design and error variance matrices.

We make those stipulations using the direct sum. Such notation is used to describe the regular structure of the error covariance and design matrices so as to keep algebraic computations as simple as possible.

The direct sum is defined as a diagonalization of matrices,

$$\bigoplus_{j=1}^{n} \boldsymbol{A}_i = \begin{pmatrix} \boldsymbol{A}_1 & \mathbf{0} & \dots \\ \mathbf{0} & \boldsymbol{A}_2 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}, \tag{3.2.2}$$

where each of the $\boldsymbol{A}_i's$ represent a unique matrix of arbitrary dimension. We also use the notation $\boldsymbol{I}_s$ to represent the $s \times s$ identity matrix, $\mathbf{1}_{m_{\mathrm{ij}}}$ to represent the $m_j \times 1$ vector of ones,

$$\mathbf{1}_{m_{\mathrm{ij}}} = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix}. \tag{3.2.3}$$

The model of subject means is constructed by specifying in (3.1.1) the fixed effects design matrix, random effects design matrix and covariance matrix as:

$$\boldsymbol{X}_i = \bigoplus_{j=1}^{s} \mathbf{1}_{m_{\mathrm{ij}}}, \tag{3.2.4}$$

$$\boldsymbol{Z}_i = \bigoplus_{j=1}^{s} \mathbf{1}_{m_{\mathrm{ij}}}, \tag{3.2.5}$$

and

$$\boldsymbol{\Lambda}_i = \bigoplus_{j=1}^{s} \sigma_j^2 \boldsymbol{I}_{m_{\mathrm{ij}}}. \tag{3.2.6}$$

To illustrate the case where $s = 2, m_{\mathrm{i}1} = 2, m_{\mathrm{i}2} = 3$, the fixed effects, random effects and error covariance design matrices are constructed as

$$\boldsymbol{X_i} = \bigoplus_{j=1}^{2} \mathbf{1}_{m_{\mathrm{ij}}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \tag{3.2.7}$$

$$\boldsymbol{Z_i} = \bigoplus_{j=1}^{2} \mathbf{1}_{m_{\mathrm{ij}}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \tag{3.2.8}$$

and

$$\boldsymbol{\Lambda}_i = \oplus_{j=1}^{2} \sigma_j^2 \boldsymbol{I}_{m_{\mathrm{ij}}} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_2^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_2^2 \end{pmatrix}. \tag{3.2.9}$$

The full model (3.1.1) then becomes

$$\begin{aligned} \begin{pmatrix} y_{\mathrm{i}11} \\ y_{\mathrm{i}12} \\ y_{\mathrm{i}21} \\ y_{\mathrm{i}22} \\ y_{\mathrm{i}23} \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_{\mathrm{i}1} \\ b_{\mathrm{i}2} \end{pmatrix} + \begin{pmatrix} \epsilon_{\mathrm{i}11} \\ \epsilon_{\mathrm{i}12} \\ \epsilon_{\mathrm{i}21} \\ \epsilon_{\mathrm{i}22} \\ \epsilon_{\mathrm{i}23} \end{pmatrix} \\ &= \begin{pmatrix} \beta_1 \\ \beta_1 \\ \beta_2 \\ \beta_2 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} b_{\mathrm{i}1} \\ b_{\mathrm{i}1} \\ b_{\mathrm{i}2} \\ b_{\mathrm{i}2} \\ b_{\mathrm{i}2} \end{pmatrix} + \begin{pmatrix} \epsilon_{\mathrm{i}11} \\ \epsilon_{\mathrm{i}12} \\ \epsilon_{\mathrm{i}21} \\ \epsilon_{\mathrm{i}22} \\ \epsilon_{\mathrm{i}23} \end{pmatrix}. \end{aligned} \tag{3.2.10}$$

As can be seen from the rows of (3.2.10), model (3.1.1) simplifies to (3.2.1), the usual form of the Subject Means Model. From (3.1.18) and (3.1.16), we know the posterior distribution is normal with covariance matrix

$$\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P'} = \boldsymbol{P}(\boldsymbol{Z_i'}\boldsymbol{\Lambda}_i^{-1}\boldsymbol{Z}_i + \boldsymbol{D}^{-1})^{-1}\boldsymbol{P'}. \tag{3.2.11}$$

Applying (3.2.4) and (3.2.6), we can write

$$\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i^{-1}\boldsymbol{Z}_i = \left(\overset{s}{\underset{j=1}{\oplus}}\mathbf{1}_{m_{\mathrm{ij}}}\right)' \left(\overset{s}{\underset{j=1}{\oplus}}\sigma_j^2\boldsymbol{I}_{m_{\mathrm{ij}}}\right)^{-1}\left(\overset{s}{\underset{j=1}{\oplus}}\mathbf{1}_{m_{\mathrm{ij}}}\right). \tag{3.2.12}$$

Using properties from Appendix A yields:

$$\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i^{-1}\boldsymbol{Z}_i = \left(\overset{s}{\underset{j=1}{\oplus}}\mathbf{1}_{m_{\mathrm{ij}}}\right)' \left(\overset{s}{\underset{j=1}{\oplus}}\frac{1}{\sigma_j^2}\boldsymbol{I}_{m_{\mathrm{ij}}}\right)\left(\overset{s}{\underset{j=1}{\oplus}}\mathbf{1}_{m_{\mathrm{ij}}}\right) \tag{3.2.13}$$

Using property (A.0.1) allows us to combine the direct sums to find that

$$\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i^{-1}\boldsymbol{Z}_i = \overset{s}{\underset{j=1}{\oplus}}\left[\mathbf{1}_{m_{\mathrm{ij}}}'\frac{1}{\sigma_j^2}\boldsymbol{I}_{m_{\mathrm{ij}}}\mathbf{1}_{m_{\mathrm{ij}}}\right] = \overset{s}{\underset{j=1}{\oplus}}\left[\frac{1}{\sigma_j^2}\mathbf{1}_{m_{\mathrm{ij}}}'\mathbf{1}_{m_{\mathrm{ij}}}\right] = \overset{s}{\underset{j=1}{\oplus}}\frac{m_{\mathrm{ij}}}{\sigma_j^2}. \tag{3.2.14}$$

Therefore, we find the posterior variance to be

$$\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}' = \boldsymbol{P}(\overset{s}{\underset{j=1}{\oplus}}\frac{m_{\mathrm{ij}}}{\sigma_j^2} + \boldsymbol{D}^{-1})^{-1}\boldsymbol{P}'. \tag{3.2.15}$$

Furthermore, the posterior mean is

$$\begin{aligned}\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i &= \boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{Z}_i'{}^{'}\boldsymbol{\Lambda}_i{}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)\\ &= \boldsymbol{P}(\overset{s}{\underset{j=1}{\oplus}}\frac{m_{\mathrm{ij}}}{\sigma_j^2} + \boldsymbol{D}^{-1})^{-1}\left(\overset{s}{\underset{j=1}{\oplus}}\mathbf{1}_{m_{\mathrm{ij}}}\right)'\left(\overset{s}{\underset{j=1}{\oplus}}\frac{1}{\sigma_j^2}\boldsymbol{I}_{m_{\mathrm{ij}}}\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right).\end{aligned} \tag{3.2.16}$$

Using results from Appendix B, we find that

$$\boldsymbol{P}\tilde{\boldsymbol{\beta}}_i = \boldsymbol{P}(\overset{s}{\underset{j=1}{\oplus}}\frac{m_{\mathrm{ij}}}{\sigma_j^2} + \boldsymbol{D}^{-1})^{-1}\left(\overset{s}{\underset{j=1}{\oplus}}\frac{1}{\sigma_j^2}\mathbf{1}_{m_{\mathrm{ij}}}\right)'\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right). \tag{3.2.17}$$

In the previous section, determination of the Bayesian MSE of the MMSE and the MAP required matrix inversion of $\boldsymbol{\Lambda}_i$, which had dimensions $s \times m_{\mathrm{ij}}$. By analytically computing the matrix $\boldsymbol{Z}_i{}'\boldsymbol{\Lambda}_i^{-1}\boldsymbol{Z}_i$, we have reduced the maximum size of the matrix inversions required to s.

3.3 Bivariate Bayesian model of subject means

In Section 3.1, we constructed a general formula for assessing the accuracy of the MMSE and MAP estimators for a general linear Bayesian model. In Section 3.2, we showed that further specifying the model to include only fixed subject means and random subject-specific, task-specific means allowed us to reduce the dimensionality of the matrix inversion required

to assess the estimator accuracy. In this section, we aim additionally to understand the relationship between the task variance components and the prediction accuracy.

Comprehending how variance components influence prediction accuracy is difficult for the multivariate linear Bayesian model with an arbitrary number of task variables. In addition, analytic minimization of the data collection cost for this model requires optimization of a matrix system of arbitrary size. To simplify analytical calculations and understanding of results, we therefore choose to further specify a bivariate version of the multivariate Bayesian model of subject means before considering the effect of the task variance components and minimizing the data collection cost.

We restrict the multivariate Bayesian model of subject means in equation (3.2.1) by mandating that the task subscript j can take on values $j = 1, 2$, where $j = 1$ represents the primary task for which we are interested in enhancing predictions, and $j = 2$ represents a secondary task. We now write the primary task extraction matrix as

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 \end{pmatrix}. \tag{3.3.1}$$

Furthermore, we may specify the between-subjects covariance matrix as

$$\boldsymbol{D} = \begin{pmatrix} \delta_1^2 & \rho\delta_1\delta_2 \\ \rho\delta_1\delta_2 & \delta_2^2 \end{pmatrix}, \tag{3.3.2}$$

where ρ represents the correlation between tasks 1 and 2, and δ_j^2 represents the task-specific variance. Finally, we may specify the the error variance matrix as

$$\boldsymbol{\Lambda} = \bigoplus_{j=1}^{2} \frac{m_{\mathrm{ij}}}{\sigma_j^2} = \begin{pmatrix} \frac{m_{\mathrm{i1}}}{\sigma_1^2} & 0 \\ 0 & \frac{m_{\mathrm{i2}}}{\sigma_2^2} \end{pmatrix}, \tag{3.3.3}$$

where m_{i1} is the number of observations for an individual on task 1, m_{i2} is the number of observations for this individual on task 2, σ_1^2 is the error variance associated with task 1, and σ_2^2 is the error variance associated with task 2. An example of the bivariate specification is given in (3.2.10).

Using (3.2.15) and (3.3.3), we determine the the posterior variance of $\boldsymbol{b}$ for the bivariate linear model of subject means to be

$$\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P'} = \boldsymbol{P}\Big(\bigoplus_{j=1}^{2} \frac{m_{\mathrm{ij}}}{\sigma_j^2} + \boldsymbol{D}^{-1}\Big)^{-1}\boldsymbol{P'}. \tag{3.3.4}$$

The inverse of the prior covariance matrix is

$$\boldsymbol{D}^{-1} = \begin{pmatrix} \frac{1}{\delta_1^2(1-\rho^2)} & \frac{-\rho}{\delta_1\delta_2(1-\rho^2)} \\ \frac{-\rho}{\delta_1\delta_2(1-\rho^2)} & \frac{1}{\delta_2^2(1-\rho^2)} \end{pmatrix}. \tag{3.3.5}$$

The posterior variance of $\boldsymbol{b}_{\mathrm{i}1}$ (3.2.15) simplifies to

$$\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}' = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{m_{\mathrm{i}1}}{\sigma_1^2} + \frac{1}{\delta_1^2(1-\rho^2)} & \frac{-\rho}{\delta_1\delta_2(1-\rho^2)} \\ \frac{-\rho}{\delta_1\delta_2(1-\rho^2)} & \frac{m_{\mathrm{i}2}}{\sigma_2^2} + \frac{1}{\delta_2^2(1-\rho^2)} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{3.3.6}$$

Taking the inverse and multiplying by the matrices $\boldsymbol{P}$ and $\boldsymbol{P}'$ (defined in (3.1.17)) from left and right, respectively, to extract the upper left element, we find that the posterior variance is

$$\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}' = \frac{\frac{m_{\mathrm{i}2}}{\sigma_2^2} + \frac{1}{\delta_2^2(1-\rho^2)}}{\left(\frac{m_{\mathrm{i}1}}{\sigma_1^2} + \frac{1}{\delta_1^2(1-\rho^2)}\right)\left(\frac{m_{\mathrm{i}2}}{\sigma_2^2} + \frac{1}{\delta_2^2(1-\rho^2)}\right) - \frac{\rho^2}{\delta_1^2\delta_2^2(1-\rho^2)^2}}, \tag{3.3.7}$$

for $0 \leq \rho < 1$. In the limit as $\rho \to 1$, this simplifies to

$$\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}' = \frac{1}{m_{\mathrm{i}2}\frac{\delta_2^2}{\delta_1^2\sigma_2^2} + \frac{m_{\mathrm{i}1}}{\sigma_1^2} + \frac{1}{\delta_1^2}}. \tag{3.3.8}$$

In this form it is difficult to compare the accuracy gain from the two tasks. We can reformulate (3.3.7) by working with the precision instead of the variance, which yields certain intuitive properties. We define the precision η of a random variable $\boldsymbol{z}$ as

$$\eta(\boldsymbol{z}) = \mathrm{Var}[\boldsymbol{z}]^{-1}. \tag{3.3.9}$$

It follows that the posterior precision of $\boldsymbol{b}_{\mathrm{i}1}$ is

$$\eta\left(\boldsymbol{b}_{\mathrm{i}1}|\,\boldsymbol{y}\right) = \left(\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}'\right)^{-1}. \tag{3.3.10}$$

Since $\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}'$ is a scalar, the precision is simply the reciprocal of the variance:

$$\eta\left(\boldsymbol{b}_{\mathrm{i}1}|\,\boldsymbol{y}\right) = \frac{1}{\boldsymbol{P}\tilde{\boldsymbol{D}}_i\boldsymbol{P}'}. \tag{3.3.11}$$

We rearrange (3.3.7) (see Appendix C) and show that the posterior precision can be written as

$$\eta = \frac{m_1}{\sigma_1^2} + \frac{1}{\delta_1^2} + \lambda_i, \tag{3.3.12}$$

where

$$\lambda_i = \frac{m_{\mathrm{i}2}\frac{\rho^2}{\delta_1^2(1-\rho^2)}}{m_{\mathrm{i}2} + \frac{\sigma_2^2}{\delta_2^2(1-\rho^2)}}. \tag{3.3.13}$$

This formulation is more suitable for interpretation as it readily simplifies to the posterior precision of a single task (see (2.1.15)) when $\rho = 0$. It is not difficult to show that the Fisher

Information for the mean of a normal distribution from which m observations are taken is $\frac{m}{\sigma^2}$, which is the first term in (3.3.12). Next, the precision of the marginal prior

$$\boldsymbol{b}_{\mathrm{i}1} \sim \mathcal{N}\left(0, \delta_1^2\right) \tag{3.3.14}$$

is $\frac{1}{\delta_1^2}$, which is the second term in (3.3.12). Finally, the term λ is a component of the posterior precision that depends on the number of observations from the second task.

As all three terms represent sources of information about $\boldsymbol{b}_{\mathrm{i}1}$, and all terms are included in the posterior precision in the same manner, we will hitherto refer to each of these terms as sources of information.

As such, we may loosly interpret $\frac{m_{\mathrm{i}1}}{\sigma_1^2}$ to be the information obtained about $\boldsymbol{b}_{\mathrm{i}1}$ from the primary task data, $\frac{1}{\delta_1^2}$ to be the information obtained about $\boldsymbol{b}_{\mathrm{i}1}$ from its marginal prior, and λ to be the information obtained about $\boldsymbol{b}_{\mathrm{i}1}$ from the secondary task data. The total information gain is the sum of information gained from the primary task, the secondary task, and the prior.

To obtain a better understanding of the information gain from the secondary task, we continue by looking at the univariate effects of the sample size and prior covariance terms on λ.

To show how the information on $\boldsymbol{b}_{\mathrm{i}1}$ from $\boldsymbol{y}_{\mathrm{i}2}$ is affected by the secondary task sample size, we define the constants

$$\lambda_{\max} = \frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)} \tag{3.3.15}$$

and

$$\tilde{m}_h = \frac{\sigma_2^2}{\delta_2^2\left(1-\rho^2\right)}, \tag{3.3.16}$$

so that

$$\lambda\left(m_{\mathrm{i}2}\right) = \frac{m_{\mathrm{i}2}\lambda_{\max}}{m_{\mathrm{i}2}+\tilde{m}_h}. \tag{3.3.17}$$

We find that the information gained through the secondary task is a nonlinear growth function of the number of data points observed from the task.

The information obtained from the secondary task increases at an approximately constant rate of $\frac{\lambda_{\max}}{\tilde{m}_h} = \rho^2\sigma_2^2\frac{\delta_2^2}{\delta_1^2}$ per measurement for small values of $m_{\mathrm{i}2}$, it reaches half of its maximum when $m_{\mathrm{i}2} = \tilde{m}_{1/2}$ and at larger values of $m_{\mathrm{i}2}$ tends asymptotically to a maximum information content value of $\lambda_{\max}$. To better understand $\lambda_{\max}$, we can decompose it as follows:

$$\lambda_{\max} = \frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)} = \frac{1}{\delta_1^2\left(1-\rho^2\right)} - \frac{1}{\delta_1^2}. \tag{3.3.18}$$

We then note that $\frac{1}{\delta_1^2(1-\rho^2)}$ is the precision of

$$\boldsymbol{b}_{\mathrm{i}1}|\boldsymbol{b}_{\mathrm{i}2} \sim \mathcal{N}\left(\rho\frac{\delta_1}{\delta_2}\boldsymbol{b}_{\mathrm{i}2}, \delta_1^2\left(1-\rho^2\right)\right), \tag{3.3.19}$$

and δ_1^2 is the precision of the marginal prior.
Therefore, we find that the information added by the secondary task is

$$\lambda_{\max} = \eta\left(\boldsymbol{b}_{\mathrm{i}1}|\boldsymbol{b}_{\mathrm{i}2}\right) - \eta\left(\boldsymbol{b}_{\mathrm{i}1}\right). \tag{3.3.20}$$

Now that we have assessed how the information gained from the secondary task depends on the number of data points, we move to considering how this information is influenced by the correlation ρ between the primary and secondary tasks. Note that the term ρ always appears as ρ^2 in λ_i. We may interpret ρ^2 as the proportion of the prior variance in the marginal prior of $\boldsymbol{b}_{\mathrm{i}1}$ that can be explained by $\boldsymbol{b}_{\mathrm{i}2}$. To see this, note that the variance from (3.3.19) is the variance in $\boldsymbol{b}_{\mathrm{i}1}$ unexplained by $\boldsymbol{b}_{\mathrm{i}2}$, and the prior variance can be found in (3.3.14). The variance that can be explained with $\boldsymbol{b}_{\mathrm{i}2}$ can then be computed as

$$\begin{aligned}\text{ExplainedVariance} &= \text{TotalVariance} - \text{UnexplainedVariance}\\ &= \delta_1^2 - \left(1-\rho^2\right)\delta_1^2\\ &= \rho^2\delta_1^2,\end{aligned} \tag{3.3.21}$$

and the proportion of the variance that can be explained with $\boldsymbol{b}_{\mathrm{i}2}$ is indeed

$$\frac{\text{ExplainedVariance}}{\text{TotalPriorVariance}} = \frac{\rho^2\delta_1^2}{\delta_1^2} = \rho^2. \tag{3.3.22}$$

We can rewrite (3.3.13) as

$$\lambda\left(\rho^2\right) = \alpha_{\min}\frac{\rho^2}{(1-\rho^2) + \frac{\alpha_{\min}}{\alpha_{\max}}}, \tag{3.3.23}$$

where

$$\alpha_{\max} = \frac{\delta_2^2}{\delta_1^2}\frac{m_{\mathrm{i}2}}{\sigma_2^2}, \tag{3.3.24}$$

and

$$\alpha_{\min} = \frac{1}{\delta_1^2}. \tag{3.3.25}$$

Thus, we see that the information is a rational function of the squared correlations. We can use this function to assess how different levels of correlation between the primary and secondary task would affect the improvement in the accuracy of the primary task parameter estimate.

Next, let us consider how σ_2^2 affects the information obtained from the secondary task. We can rewrite (3.3.13) as

$$\lambda\left(\sigma_2^2\right) = \frac{\lambda_{\max}}{1 + \frac{\sigma_2^2}{\gamma_1}}, \tag{3.3.26}$$

where

$$\gamma_1 = \delta_2^2\left(1 - \rho^2\right) m_{\mathrm{i}2}. \tag{3.3.27}$$

We find that as the error variance gets small, the information converges to $\lambda_{\max}$, and when the error variance gets large, the information converges to 0.

Next, let us consider how δ_2^2 affects the information obtained from the secondary task:

$$\lambda\left(\delta_2^2\right) = \lambda_{\max}\frac{\delta_2^2}{\tilde{\delta}_{1/2} + \delta_2^2}, \tag{3.3.28}$$

where

$$\tilde{\delta}_{1/2} = \frac{\sigma_2^2}{\left(1 - \rho^2\right) m_{\mathrm{i}2}}. \tag{3.3.29}$$

We find that for small values of the prior variance for the secondary task, the information obtained is approximately linearly proportional to the prior variance, with coefficient

$$\frac{\lambda_{\max}}{\tilde{\delta}_{1/2}} = \frac{\rho^2 m_{\mathrm{i}2}}{\delta_1^2\sigma_2^2}. \tag{3.3.30}$$

As the prior variance for the secondary task goes to zero, the information added by the secondary task goes to zero, and as the prior variance for the secondary task goes to infinity, the information added goes to $\lambda_{\max}$.

Next, let us consider how δ_1^2 affects the information obtained from the secondary task,

$$\lambda\left(\delta_1^2\right) = \frac{\gamma_2}{\delta_1^2}, \tag{3.3.31}$$

where

$$\gamma_2 = \frac{m_{\mathrm{i}2}\rho^2}{\left(1 - \rho^2\right)\left(m_{\mathrm{i}2} + \frac{\sigma_2^2}{\delta_2^2\left(1-\rho^2\right)}\right)}. \tag{3.3.32}$$

We find that for non-zero $m_{\mathrm{i}2}, \rho$, as the prior variance for $\boldsymbol{b}_{\mathrm{i}1}$ gets small, the information about the primary task from the secondary task goes to ∞. As the prior variance gets large, the information about the primary task from the secondary task goes to zero.

Figure 3.1 summarizes the effect of each of the parameters on the information gained from the secondary task.

We showed that for the Bayesian bivariate model of subject means, the precision can be written as the sum of terms related to various sources of information. This property

motivates us to suggest that the precision is an easier and more intuitive measure of accuracy than the MSE. Using the precision as a measure of accuracy, we obtained an analytic form for the precision of the Bayesian bivariate model of subject means. We continued by describing the influence of the number of data points from the secondary task, a result which we will build on in Section 3.4. To inform the choice of which tasks will result in maximal information gain for a primary task, we also described the influence of the various task variance components. In scenarios where a group means model is relevant, these can be used to quickly choose the task which will be best suited to give increase accuracy.

3.4 Cost minimization for the bivariate linear Bayesian model of subject means

In Section 3.3 we detailed how the accuracy of the parameter MMSE and MAP estimators for the primary task depend on the sample sizes for the bivariate Bayesian model of subject means. We now continue by using this result to minimize the cost of data collection for a simple example. We consider two tasks which are modeled by the Bayesian bivariate model of subject means, and assume that each measurement from task 1 costs c_1 to collect and each measurement from task 2 costs c_2 to collect, resulting in the total cost of

$$c_t = c_1 m_{\mathrm{i}1} + c_2 m_{\mathrm{i}2}. \tag{3.4.1}$$

We then aim to answer the question: how many observations should we measure on the primary task variable and how many observations should we measure on the secondary task variable to minimize the total cost, assuming we pursue a fixed precision η_i on the primary task, where combining equations (3.3.12) and (3.3.17) yields

$$\eta_i = \frac{m_{\mathrm{i}1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{m_{\mathrm{i}2}\lambda_{\max}}{m_{\mathrm{i}2} + \tilde{m}_h}. \tag{3.4.2}$$

We can solve for $m_{\mathrm{i}1}$ as follows:

$$m_{\mathrm{i}1} = \sigma_1^2 \left(\eta_i - \frac{1}{\delta_1^2} - \frac{m_{\mathrm{i}2}\lambda_{\max}}{m_{\mathrm{i}2} + \tilde{m}_h} \right). \tag{3.4.3}$$

Substituting into the cost equation, the total cost is

$$c_t = c_1\sigma_1^2 \left(\eta_i - \frac{1}{\delta_1^2} - \frac{m_{\mathrm{i}2}\lambda_{\max}}{m_{\mathrm{i}2} + \tilde{m}_h} \right) + c_2 m_{\mathrm{i}2}. \tag{3.4.4}$$

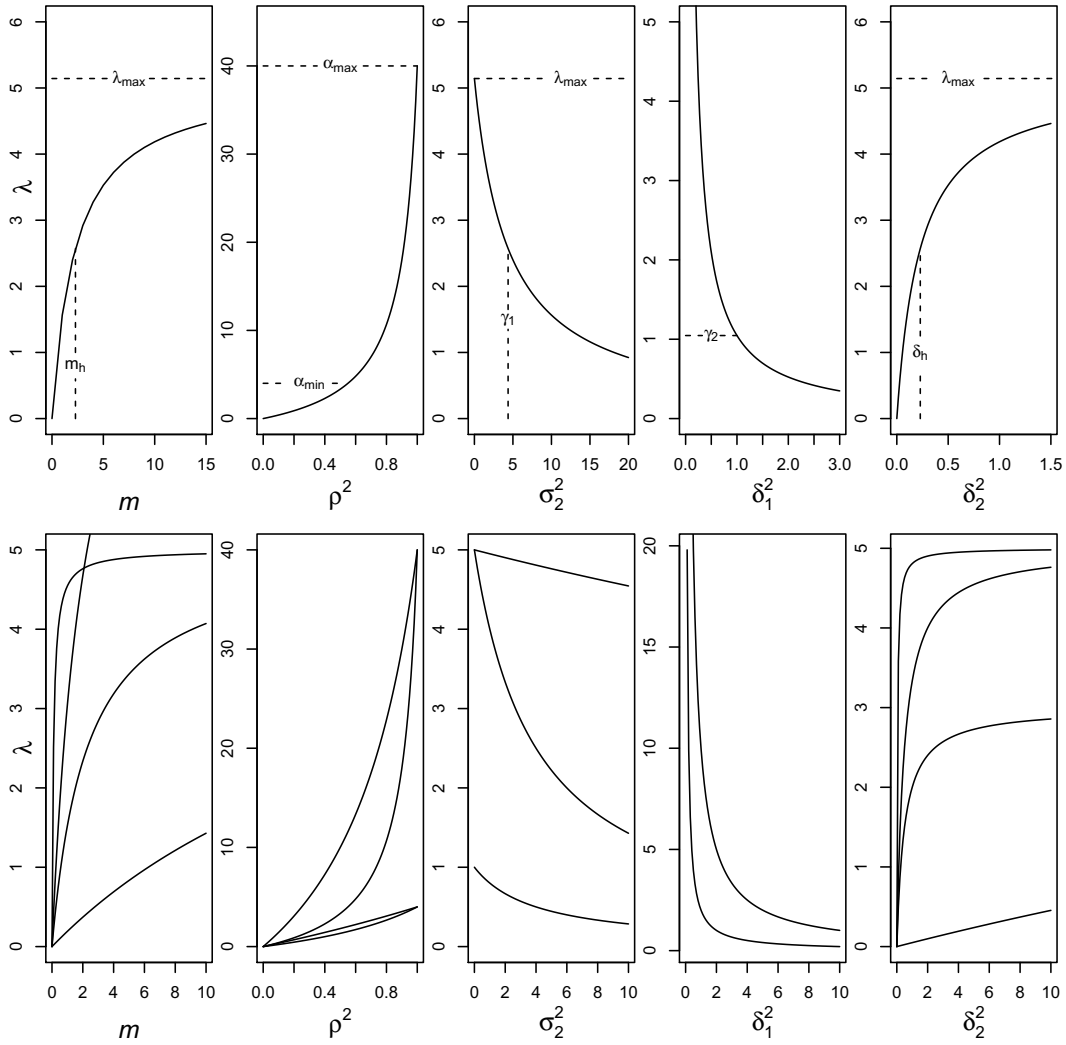

Figure 3.1: The first row shows an example of the effect of the number of datapoints from a secondary task (m), the squared correlation between the primary and secondary tasks (ρ^2), the error variance of the secondary task (σ_2^2), the between-subjects variance of the primary task (δ_1^2), and the between-subjects variance of the secondary task (δ_2^2) on the additional information obtained from the secondary task (λ). The fixed parameter values are $m_2 = 10, \rho = 0.75, \delta_1^2 = 0.25, \delta_2^2 = 1.0, \sigma_1^2 = 1.0, \sigma_2^2 = 1.0$. All parameters except the ones represented on the x-axis are held at their fixed values. The second row shows a range of effects each of the primary and secondary task variables may have on λ. On the bottom row from left to right, the first plot shows λ vs. m, where where m_h=2.28,0.1,25,2.28 and λ_{max}=5,5,5,10, the second plot shows λ vs. ρ^2, where a_{min}=4,20,4,20 and a_{max}=40,40,4,4, the third plot shows λ vs. δ_1^2, where λ_{max}=5,5,1, and γ_1=100,4,4, the fourth plot shows, γ_2=2,10, the fifth plot shows, λ_{max}=5,5,5,3, and δ_h=0.04,0.5,100,0.5.

To minimize the total cost, we take the derivative of the total cost with respect to $m_{\mathrm{i}2}$ and set it equal to zero,

$$\frac{\partial c_t}{\partial m_{\mathrm{i}2}} = c_1\sigma_1^2\left(\frac{m_{\mathrm{i}2}\lambda_{\max}}{\left(m_{\mathrm{i}2}+\tilde{m}_h\right)^2} - \frac{\lambda_{\max}}{m_{\mathrm{i}2}+\tilde{m}_h}\right) + c_2 = 0. \tag{3.4.5}$$

We then solve for $m_{\mathrm{i}2}$ to find the number of observations of the secondary task needed to minimize the cost:

$$\hat{m}_{\mathrm{i}2} = \pm\sigma_1\sqrt{\frac{c_1}{c_2}\lambda_{\max}\tilde{m}_h} \; - \tilde{m}_h. \tag{3.4.6}$$

There are two solutions; the only possible solution is the one for which

$$\begin{aligned}
\hat{m}_{\mathrm{i}2} &= \sigma_1\sqrt{\frac{c_1}{c_2}\lambda_{\max}\tilde{m}_h} \; - \tilde{m}_h \\
&= \sigma_1\sqrt{\frac{c_1}{c_2}\frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)}\frac{\sigma_2^2}{\delta_2^2\left(1-\rho^2\right)}} - \frac{\sigma_2^2}{\delta_2^2\left(1-\rho^2\right)} \\
&= \frac{\sigma_2}{\delta_1\delta_2^2\left(1-\rho^2\right)}\left(\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho - \delta_1\sigma_2\right).
\end{aligned} \tag{3.4.7}$$

The corresponding number of measurements for the first task is

$$\begin{aligned}
\hat{m}_{\mathrm{i}1} &= \sigma_1^2\left(\eta_i - \frac{1}{\delta_1^2} - \frac{\hat{m}_{\mathrm{i}2}\lambda_{\max}}{\hat{m}_{\mathrm{i}2}+\tilde{m}_h}\right) \\
&= \sigma_1^2\left(\eta_i - \frac{1}{\delta_1^2} - \frac{\frac{\sigma_2}{\delta_1\delta_2^2\left(1-\rho^2\right)}\left(\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho - \delta_1\sigma_2\right)\frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)}}{\frac{\sigma_2}{\delta_1\delta_2^2\left(1-\rho^2\right)}\left(\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho - \delta_1\sigma_2\right) + \frac{\sigma_2^2}{\delta_2^2\left(1-\rho^2\right)}}\right) \\
&= \sigma_1^2\left(\eta_i - \frac{1}{\delta_1^2} - \frac{\left(\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho - \delta_1\sigma_2\right)\frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)}}{\left(\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho - \delta_1\sigma_2\right) + \delta_1\sigma_2}\right) \\
&= \sigma_1^2\left(\eta_i - \frac{1}{\delta_1^2} - \frac{\left(\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho - \delta_1\sigma_2\right)\frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)}}{\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho}\right) \\
&= \sigma_1^2\left(\eta_i - \frac{1}{\delta_1^2} - \frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)} + \frac{\delta_1\sigma_2\frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)}}{\sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho}\right) \\
&= \left(\eta_i - \frac{1}{\delta_1^2\left(1-\rho^2\right)}\right)\sigma_1^2 + \frac{\sqrt{c_2}\rho\sigma_1\sigma_2}{\sqrt{c_1}\delta_1\delta_2\left(1-\rho^2\right)}.
\end{aligned} \tag{3.4.8}$$

The optimal sample size on the secondary task will only be greater than zero if

$$\delta_1\sigma_2 < \sqrt{\frac{c_1}{c_2}}\delta_2\sigma_1\rho, \tag{3.4.9}$$

which can be rearranged to read

$$\frac{c_2}{c_1} < \frac{\delta_2^2/\sigma_2^2}{\delta_1^2/\sigma_1^2}\rho^2. \tag{3.4.10}$$

The optimal solution for the sample size on the primary task will only be greater than zero if

$$\left(\eta_i - \frac{1}{\delta_1^2\left(1-\rho^2\right)}\right)\sigma_1^2 + \frac{\sqrt{c_2}\rho\sigma_1\sigma_2}{\sqrt{c_1}\delta_1\delta_2\left(1-\rho^2\right)} > 0, \tag{3.4.11}$$

which can be rearranged to read

$$\eta_i > \frac{1}{\delta_1^2\left(1-\rho^2\right)}\left(1 - \rho\sqrt{\frac{c_2}{c_1}}\frac{\delta_1/\sigma_1}{\delta_2/\sigma_2}\right). \tag{3.4.12}$$

Inequality (3.4.10) tells us whether any measurements from a secondary task will be required to obtain the fixed accuracy η_i for minimal cost. We find that whether measurements from the secondary task will be useful is not dependent on the level of accuracy that we require. If this inequality is not met, then no measurements should be collected from the primary task, and a total of

$$\hat{m}_{\mathrm{i}1} = \sigma_1^2\eta_i - \frac{\sigma_1^2}{\delta_1^2} \tag{3.4.13}$$

measurements should be collected from the primary task. Inequality (3.4.12) tells us when data from the primary task will be needed in obtaining the fixed accuracy η_i for minimal cost. If this inequality is not met, then no measurements should be collected from the primary task, and a total of

$$\hat{m}_{\mathrm{i}2} = \tilde{m}_h\frac{\left(\eta_i - \frac{1}{\delta_1^2}\right)}{\lambda_{\max} - \eta_i - \frac{1}{\delta_1^2}} \tag{3.4.14}$$

measurements should be collected from the secondary task.

We note that for a given scenario, the optimal sample sizes $\hat{m}_{\mathrm{i}1}$ and $\hat{m}_{\mathrm{i}2}$ will likely not be integer values. The suggestion we give for such scenarios is to consider values of $\hat{m}_{\mathrm{i}2}$ that are both rounded down and rounded up. For $\hat{m}_{\mathrm{i}2}$ rounded down, round $\hat{m}_{\mathrm{i}1}$ up and add to $\hat{m}_{\mathrm{i}1}$ until the desired level of accuracy is met. For $\hat{m}_{\mathrm{i}2}$ rounded up, again round $\hat{m}_{\mathrm{i}1}$ up, but this time remove from $\hat{m}_{\mathrm{i}1}$ until the desired accuracy is just met. Use the cost equation to determine the cost of these two scenarios, and choose the scenario that minimizes the cost.

Chapter 4

Forecasting Accuracy for Nonlinear Models

In this thesis, we have been investigating the individualization of biomathematical models of performance over multiple task variables. We consider a multivariate Bayesian framework for combining population information with data on a new individual to make parameter estimates and performance predictions. In the previous chapter, we demonstrated how to assess the accuracy of predictions in a general linear Bayesian model for multiple task variables and showed how to determine which tasks to collect data from to obtain a fixed accuracy with minimum cost. In order to demonstrate how to utilize multiple tasks to increase the prediction accuracy for a model that is not necessarily linear, we next consider the problem of assessing the accuracy of predictions in the nonlinear case.

In Section 4.1, we formulate the general nonlinear Bayesian model, and consider complications that arise in this case as we apply Bayes rule to determine the posterior distribution. We show that in general the accuracy in the nonlinear case must be assessed numerically, and detail a way to do this more quickly via repeated simulation using the MAP estimator as opposed to the MMSE. In Section 4.2, we apply the suggested accuracy assessment procedure to a nonlinear model describing performance dynamics over 88 hours of total sleep deprivation and find that the procedure quickly obtains a reasonable approximation of the Bayesian mean squared prediction error. Finally, in Section 4.3 we consider analytically separating classes of nonlinear models by whether or not the accuracy of the MMSE estimator may be reasonably assessed with the MAP estimator. We illustrate this separation

by considering a class of quadratic Bayesian models.

4.1 Nonlinear multivariate Bayesian model

We now extend the general Bayesian linear model (3.1.1) to the general Bayesian nonlinear model,

$$\boldsymbol{y}_i = \xi\left(\boldsymbol{\phi}_i,\boldsymbol{x}_i\right) + \boldsymbol{\epsilon}_i, \tag{4.1.1}$$

where the model parameters $\boldsymbol{\phi}_i$ are linear functions of the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{b}_i$,

$$\boldsymbol{\phi}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i. \tag{4.1.2}$$

Furthermore, $\boldsymbol{x}_i$ is the covariate vector for subject i and ξ is assumed nonlinear with respect to the model parameters $\boldsymbol{\phi}_i$. As before, we assume that the random effects are normally distributed,

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \tag{4.1.3}$$

which implies

$$f\left(\boldsymbol{b}_i\right) = \frac{\exp\left[-\frac{1}{2}\boldsymbol{b}_i{}'\boldsymbol{D}^{-1}\boldsymbol{b}_i\right]}{(2\pi)^{q/2}|\boldsymbol{D}|^{\frac{1}{2}}}. \tag{4.1.4}$$

We also assume that the errors are normally distributed,

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\boldsymbol{0},\boldsymbol{\Lambda}_i\right), \tag{4.1.5}$$

from which we find

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim \mathcal{N}(\xi(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,\boldsymbol{x}_i), \boldsymbol{\Lambda}_i), \tag{4.1.6}$$

which implies

$$f\left(\boldsymbol{y}_i|\boldsymbol{b}_i\right) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{y}_i - \xi\left(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,\boldsymbol{x}_i\right)\right)'\boldsymbol{\Lambda}_i^{-1}\left(\boldsymbol{y}_i - \xi\left(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,\boldsymbol{x}_i\right)\right)\right]}{(2\pi)^{\,m_{\mathrm{ij}}/2}|\boldsymbol{\Lambda}_i|^{\frac{1}{2}}}. \tag{4.1.7}$$

Following of Bayes Theorem, the prior and conditional distributions are multiplied to specify the posterior up to a constant c_0,

$$f(\boldsymbol{b}|\boldsymbol{y}) = c_0 \exp\left[-\frac{1}{2}\left[\left(\boldsymbol{y}_i - \xi\left(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,\boldsymbol{x}_i\right)\right)'\boldsymbol{\Lambda}_i^{-1}\left(\boldsymbol{y}_i - \xi\left(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,\boldsymbol{x}_i\right)\right) + \boldsymbol{b}_i{}'\boldsymbol{D}_i^{-1}\boldsymbol{b}_i\right]\right]. \tag{4.1.8}$$

In the linear case we were able to show that the exponential term in the posterior (3.1.9) was quadratic in $\boldsymbol{b}_i$, and hence, normal. In the nonlinear case, the exponential term will

generally not be quadratic, and hence the posterior will not be normal. Specific nonlinear cases may result in distributional forms that exhibit normal conjugacy; however, without further specification of the model, we are not able in general to make such determination. Therefore, we are generally unable to analytically compute the normalizing constant required to specify the posterior distribution and obtain the MMSE estimator (Kay, 1994, p. 317).

Analytical derivation of the MAP also depends on the specific nonlinear function, and in general will not have a closed form. In all cases, however, we can obtain the MAP and MMSE estimates numerically.

The MMSE estimate can be obtained by first using the Metropolis algorithm to obtain samples from $f(\boldsymbol{b}_i|\boldsymbol{y}_i)$, and then averaging the samples. Subsequently, a sample from the posterior response may be constructed by simulating measurement errors $\boldsymbol{\epsilon}$ from (4.1.5) and evaluating (4.1.1) for each sample of $\boldsymbol{b}_i$. The response MMSE can then be obtained by taking the mean of $\boldsymbol{y}_i^*$ over samples. For more information on

We may compute the MAP estimate of $\boldsymbol{b}$ as

$$\hat{\boldsymbol{b}_i} = \underset{\boldsymbol{b}_i}{\operatorname{argmin}} \left[(\boldsymbol{y}_i - \xi(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \boldsymbol{x}_i))' \boldsymbol{\Lambda}_i^{-1} (\boldsymbol{y}_i - \xi(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \boldsymbol{x}_i)) + \boldsymbol{b}_i' \boldsymbol{D}^{-1} \boldsymbol{b}_i \right], \quad (4.1.9)$$

which is straight forward to compute with the Gauss-Newton algorithm. Note that $\hat{\boldsymbol{b}_i}$ is the maximum of the joint posterior for all $\boldsymbol{b}_i$ and should be differentiated from $\hat{\boldsymbol{b}_{\text{ia}}}$, the maximum of the marginal posterior for only the parameters associated with the primary task. As a result of the normality of $\boldsymbol{b}|\boldsymbol{y}$, the MAP estimate of $\boldsymbol{b}_{\text{ia}}$ may be easily obtained in the linear case by extracting the first a parameters of $\hat{\boldsymbol{b}_i}$. In the nonlinear case, $\hat{\boldsymbol{b}_{\text{ia}}}$ cannot necessarily be extracted from $\hat{\boldsymbol{b}_i}$. In general, obtaining $\hat{\boldsymbol{b}_{\text{ia}}}$ requires numerical integration, in which case we may as well obtain the more accurate MMSE estimate. Concerning the response, we can obtain the joint MAP of $(\boldsymbol{y}_i^*|\boldsymbol{b}_i)$ by first using the definition of conditional probability (B.0.3) to show that

$$f(\boldsymbol{y}_i^*, \boldsymbol{b}_i|\boldsymbol{y}_i) = f(\boldsymbol{y}_i^*|\boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{y}_i). \quad (4.1.10)$$

In the linear case, both $f(\boldsymbol{y}_i^*|\boldsymbol{b}_i)$ and $f(\boldsymbol{b}_i|\boldsymbol{y}_i)$ are normal, and so by normal conjugacy, $f(\boldsymbol{y}_i^*, \boldsymbol{b}_i|\boldsymbol{y}_i)$ will be jointly normal. As a result, the MAP of $\boldsymbol{y}_i^*$ can be obtained as a partition of the MAP of $(\boldsymbol{y}_i^*, \boldsymbol{b}_i)$. In the nonlinear case, we again cannot construct the MAP of $\boldsymbol{y}_i^*$ without the use of a numerical integration technique. However; we can obtain the joint MAP of $(\boldsymbol{y}_i^*, \boldsymbol{b}_i)$ without numerical integration. Obtaining this MAP requires that the partial derivatives with respect to $\boldsymbol{y}_i^*$ and $\boldsymbol{b}_i$ will be zero at the maxima. We take the derivative of

$-2\log f\left(\boldsymbol{y}_i^*, \boldsymbol{b}_i|\boldsymbol{y}_i\right)$ as

$$\begin{aligned}
\frac{\partial\left[-2\log f\left(\boldsymbol{y}_i^*, \boldsymbol{b}_i|\boldsymbol{y}_i\right)\right]}{\partial \boldsymbol{y}_i^*} &= \frac{\partial\left[-2\log\left[f\left(\boldsymbol{y}_i^*|\boldsymbol{b}_i\right) f\left(\boldsymbol{b}_i|\boldsymbol{y}_i\right)\right]\right]}{\partial \boldsymbol{y}_i^*} \\
&= \frac{\partial\left[-2\log f\left(\boldsymbol{y}_i^*|\boldsymbol{b}_i\right)\right]}{\partial \boldsymbol{y}_i^*} - \frac{\partial\left[-2\log f\left(\boldsymbol{b}_i|\boldsymbol{y}_i\right)\right]}{\partial \boldsymbol{y}_i^*} \\
&= \frac{\partial\left[-2\log f\left(\boldsymbol{y}_i^*|\boldsymbol{b}_i\right)\right]}{\partial \boldsymbol{y}_i^*} \\
&= -2\frac{\partial}{\partial \boldsymbol{y}_i^*}\begin{bmatrix} -\frac{1}{2}\left(\boldsymbol{y}_i^* - \xi\left(\boldsymbol{A}_i^*\boldsymbol{\beta} + \boldsymbol{B}_i^*\boldsymbol{b}_i, \boldsymbol{x}_i^*\right)\right)' \\ \boldsymbol{\Lambda}_i^{-1}\left(\boldsymbol{y}_i - \xi\left(\boldsymbol{A}_i^*\boldsymbol{\beta} + \boldsymbol{B}_i^*\boldsymbol{b}_i, \boldsymbol{x}_i^*\right)\right) - \log\left[(2\pi)^{m_{\mathrm{ij}}/2}|\boldsymbol{\Lambda}_i^*|^{\frac{1}{2}}\right] \end{bmatrix} \\
&= \frac{\partial}{\partial \boldsymbol{y}_i^*}\left[\left(\boldsymbol{y}_i^* - \xi\left(\boldsymbol{A}_i^*\boldsymbol{\beta} + \boldsymbol{B}_i^*\boldsymbol{b}_i, \boldsymbol{x}_i^*\right)\right)'\boldsymbol{\Lambda}_i^{-1}\left(\boldsymbol{y}_i^* - \xi\left(\boldsymbol{A}_i^*\boldsymbol{\beta} + \boldsymbol{B}_i^*\boldsymbol{b}_i, \boldsymbol{x}_i^*\right)\right)\right] \\
&= \left(\boldsymbol{y}_i^* - \xi\left(\boldsymbol{A}_i^*\boldsymbol{\beta} + \boldsymbol{B}_i^*\boldsymbol{b}_i, \boldsymbol{x}_i^*\right)\right)'\boldsymbol{\Lambda}_i^{*-1}.
\end{aligned} \tag{4.1.11}$$

Setting the derivative of the joint posterior to zero and solving for $\boldsymbol{y}_i$, we obtain the MAP estimate as

$$\hat{\boldsymbol{y}_i} = \xi\left(\boldsymbol{A}_i^*\boldsymbol{\beta} + \boldsymbol{B}_i^*\hat{\boldsymbol{b}_i}, \boldsymbol{x}_i^*\right), \tag{4.1.12}$$

where $\hat{\boldsymbol{b}_i}$ is the MAP of $\boldsymbol{b}_i$. Stated plainly, the joint MAP of the parameters and response may be obtained by applying the model function to the parameter MAP. Note again that this is not the MAP of $\boldsymbol{y}_i^*$ alone, which would require numerical integration.

For the nonlinear Bayesian model, we propose that the *accuracy* of the MAP and MMSE may be assessed by repeatedly simulating data from the model. For this simulation, we would first choose values for $\boldsymbol{D}$, $\boldsymbol{\beta}$, and each $\boldsymbol{x}_i$, $\boldsymbol{A}_i$, $\boldsymbol{B}_i$, and $\boldsymbol{\Lambda}_i$. Furthermore, we would also specify the prediction scenario by specifying $\boldsymbol{x}_i^*$, $\boldsymbol{A}_i^*$, $\boldsymbol{B}_i^*$, and $\boldsymbol{\Lambda}_i^*$. For simplicity, we would usually assume the same design for each individual $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \ldots$, $\boldsymbol{A}_1 = \boldsymbol{A}_2 = \ldots$, $\boldsymbol{B}_1 = \boldsymbol{B}_2 = \ldots$, and $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2 = \ldots$, and similarily assume the same design over individuals for the predictions scenario. We would then repeatedly simulate individual parameter vectors $\boldsymbol{\phi}_i$ from (4.1.3). Applying the model function (4.1.1) to the simulated model parameters and simulating errors $\boldsymbol{\epsilon}_i$ from (4.1.5) for each individual, we would produce simulated responses $\boldsymbol{y}_i$. With these responses, we would make use of the Gauss-Newton algorithm to obtain the joint parameter MAP, and use the model function (this time using $\boldsymbol{x}_i^*$, $\boldsymbol{A}_i^*$, $\boldsymbol{B}_i^*$) to construct the responses at the parameter MAP. Furthermore, we would use the Metropolis algorithm to obtain a sample from the posterior distribution, and apply the model function (using $\boldsymbol{x}_i^*$, $\boldsymbol{A}_i^*$, $\boldsymbol{B}_i^*$) to each sample and again simulate measurement error (using $\boldsymbol{\Lambda}_i^*$) to obtain the

posterior response distribution. As mentioned, taking the mean of the posterior response distribution yields MMSE estimates. Repeating this procedure results in a distribution of MAP and MMSE estimates. We can then take the sample MSE to estimate the Bayesian MSE for each of the estimators.

Unlike the general linear Bayesian model of Section 3.1, the non-normality of the posterior in the case of the nonlinear Bayesian model may result in unequal MAP and MMSE estimators. Furthermore, the accuracies of these estimators may be different. In fact, in Section 4.3, we consider a special case of the quadratic Bayesian model for which we show, via analytic approximation, that both the estimators and their accuracies differ.

For most nonlinear modeling scenarios, however, it is difficult to find a simple approximation which makes the posterior distribution analytically computable. For a single parameter model, when the model is nearly linear over the entire parameter region where the likelihood is moderate, we may approximate the model as linear around a single parameter value. Such approximation moves us to the realm of the general Bayesian linear model, where the MAP and MMSE and their accuracies are equal. Note that if we make this linear approximation when it is not justified, we will be led to believe that the MAP and MMSE and their accuracies are equal, when indeed, they may be quite different. Still, to reduce the time required for the accuracy simulations by an order of magnitude, we hypothesize that in many modeling scenarios, the accuracy of the MMSE may be assessed by repeated simulation of the MAP estimator. In the next section, we present a numerical example to motivate the approximation of the MMSE with the MAP.

4.2 Nonlinear accuracy assessment simulation

In the last section, we reviewed differences in assessing the accuracy of predictions for linear and nonlinear models, brought up computational time issues with the assessment in nonlinear models, and proposed an alternative method for quicker estimation. In this section, we consider an application where the assessment of the response MMSE accuracy using the response MAP estimator is quicker than such assessment using the response MMSE estimator.

The application is based on a total sleep deprivation study described in Van Dongen et al., (2003) where subjects' cognitive performance was measured on the psychomotor vigilance task (PVT) every two hours over 88 hours of total sleep deprivation. Van Dongen et al. (2007) used subjects' performance on the PVT to estimate model parameters for an individualizable version of the waking portion of the model of Borbély (1982). The original

version contains a sum of sinusoids which, for the sake of simplicity and without noteably changing the model, we replace with a single sinusoid. The modified model is written as

$$y_{\mathrm{ij}} = \xi_i \exp\left(-\rho_i \left(t_{\mathrm{ij}} - t_{\mathrm{i0}}\right)\right) + \gamma_i \sin\left(2\pi\left(\frac{t_{\mathrm{ij}} - \phi_i}{\tau}\right)\right) + \kappa_i + \epsilon_{\mathrm{ij}}, \tag{4.2.1}$$

where τ is the circadian period, ρ_i is the subject-specific homeostatic build-up rate constant, γ_i is the subject-specific circadian amplitude, κ_i is the subject specific basal performance level, ξ_i is the initial homeostatic state, ϕ_i is the initial circadian phase, t_{ij} represents the j^{th} measurement time, t_{i0} represents the initial time, and ϵ_{ij} represents the j^{th} measurement error. The random effects terms are assumed to follow normal and lognormal distributions,

$$\begin{aligned} \rho_i &\sim \mathrm{Log}-\mathcal{N}\left(\ln\left(\rho_0\right), \psi^2\right), \\ \gamma_i &\sim \mathrm{Log}-\mathcal{N}\left(\ln\left(\gamma_0\right), \omega^2\right), \\ \kappa_i &\sim \mathcal{N}\left(\kappa_0, \chi^2\right). \end{aligned} \tag{4.2.2}$$

Also, the error is assumed to be normal,

$$\epsilon_{\mathrm{ij}} \sim \mathcal{N}\left(0, \sigma^2\right). \tag{4.2.3}$$

Van Dongen et al. (2007) estimated the population parameters as:

$$\rho_0 = 0.0350, \gamma_0 = 4.30, \kappa_0 = 29.7, \xi_0 = -28.0, \phi_0 = 0.600, \psi^2 = 1.15, \omega^2 = 0.294, \chi^2 = 36.2, \sigma^2 = 77.6. \tag{4.2.4}$$

Here we simulate performance on the PVT for 5000 hypothetical individuals every two hours over the 88 hours of wakefulness. For each individual we obtain the MAP and MMSE estimates using the first 22 data points. The joint MAP estimate for the parameters ρ, γ, and κ is obtained using a Newton-type algorithm to maximize the posterior distribution. The MMSE estimates are determined by first obtaining a sample from the parameter posterior density with the Metropolis algorithm. A sample from the response posterior is then obtained by evaluating the model at the simulated parameter values and simulating the errors to produce the response vector. Simulated and estimated response values are shown for a single individual in Figure 4.1 (a).

Calculation of all 5000 response MAP estimates took 6 minutes, whereas the corresponding calculations for the MMSE estimates took 433 minutes, or 76 times as long. The MSE was calculated using the remaining the last 22 data points for each simulated individual. A histogram of the MSE around the MMSE estimates and a similar histogram for the MAP

estimates are shown in Figure 4.1 (b). The estimated marginal MSE for the MAP estimate was 95.8±0.5, whereas the estimated marginal MSE for the MMSE estimate was 94.1±0.5, where the term after the ± represents the standard error in the mean. Therefore, we estimate that in this case, there was a 1.8 percent error in estimating the accuracy of the MMSE accrued by using the accuracy of the MAP. Deeming this to be a reasonably small estimation error, we conclude that for this modeling scenario, the MAP estimator accuracy can reasonably be used to approximate the MMSE estimator accuracy.

This scenario, however, represents only a particular case for which a large number of observations were used to obtain the posterior distribution, and we suggest this to be the reason why the Bayesian MSE for the MAP and MMSE were approximately equal. In general, as the amount of data collected increases the width of the posterior distribution will decrease. As the width decreases, a linear approximation of the model to construct the posterior distribution becomes increasingly accurate. Such a linear approximation will result in a normal posterior distribution for which the MMSE and MAP will be approximately equal.

Much of the consideration in this thesis concerns including information from a secondary task. As we saw when considering the bivariate linear Bayesian model of subject means, the secondary task only has a significantly large effect on the prediction accuracy when the number of data points from the primary task is small. When considering the nonlinear case, as more data are collected on the primary task, additional data collected on the secondary task contribute progressively less to improving parameter estimates for the primary task.

A substantial need for the inclusion of a secondary task will only occur when data from the primary task is sparse. To consider whether the Bayesian MSE of the MMSE could reasonably be assessed by the MAP under such low data scenarios, further simulations of the model of Borbély (1982) with less data points per individual than that used in Section 4.2 were conducted. Specifically, we assumed two measurements were collected on the primary task at 0 and 2 hours after the start of the total sleep deprivation period, and two measurements were collected on the secondary task at 24 and 26 hours. We used this data to make predictions at 48 and 50 hours after the start of the sleep deprivation period, and compared the predictions to the simulated responses. The estimated Bayesian MSE of the MMSE estimator was 109 ± 7 and the estimated Bayesian MSE of the MAP estimator was 149 ± 10. When the simulation was repeated with the secondary task measurements removed, the estimated Bayesian MSE of the MMSE estimator was 101 ± 7 and the Bayesian MSE of the MAP was 159 ± 12. The results suggest that nonlinear modeling scenarios with less data are less likely to yield approximately equal MAP and MMSE accuracies with or

without secondary task measurements.

4.3 Univariate quadratic Bayesian model

In the previous section, we showed an example in which it was reasonable to approximate the accuracy of the MMSE estimator via repeated simulation using the MAP estimator. As we change the model, the number and timing of the observed and predicted data points, and the prior distributions, we would need to repeat these time consuming simulations to ensure that the approximate accuracy was reasonable. In an attempt to address this issue, we may identify classes of nonlinear modeling scenarios where the approximation is justified. We begin by considering a class of Bayesian nonlinear modeling scenarios where the conditional distribution consists of a number of distinct, high and narrow peaks on a background of low likelihood.

We consider a simple case of this class where individual responses are modeled by a quadratic function of a single parameter, and where the error variance is small and the prior variance is large in comparison to the distance between peaks of the conditional distribution. Given a single, observed data point, the conditional distribution consists of two high narrow peaks that occur at values of the parameter where the model intersects the data point. We analytically assess the accuracy of both the MMSE and MAP estimators and evaluate how well the accuracy of the latter approximates that of the former by considering the ratio of the two accuracies.

We begin this section by formulating the quadratic model, and then detail our method of assessing the accuracy. We next show that the application of Bayes rule to assess the accuracy leads to an intractable integral. To remedy this, we approximate the quadratic model by a linear model at the two peaks of the conditional distribution, allowing us to compute the posterior integral using conjugate normality. We next use the posterior to obtain the MMSE and MAP estimators, and compute the posterior expected squared loss for each estimator given the data. Finally, we compute the Bayesian MSE for each estimator and determine the ratio of the accuracies.

Model formulation

Let us specify a scalar model

$$y = \xi(\phi) + \epsilon, \tag{4.3.1}$$

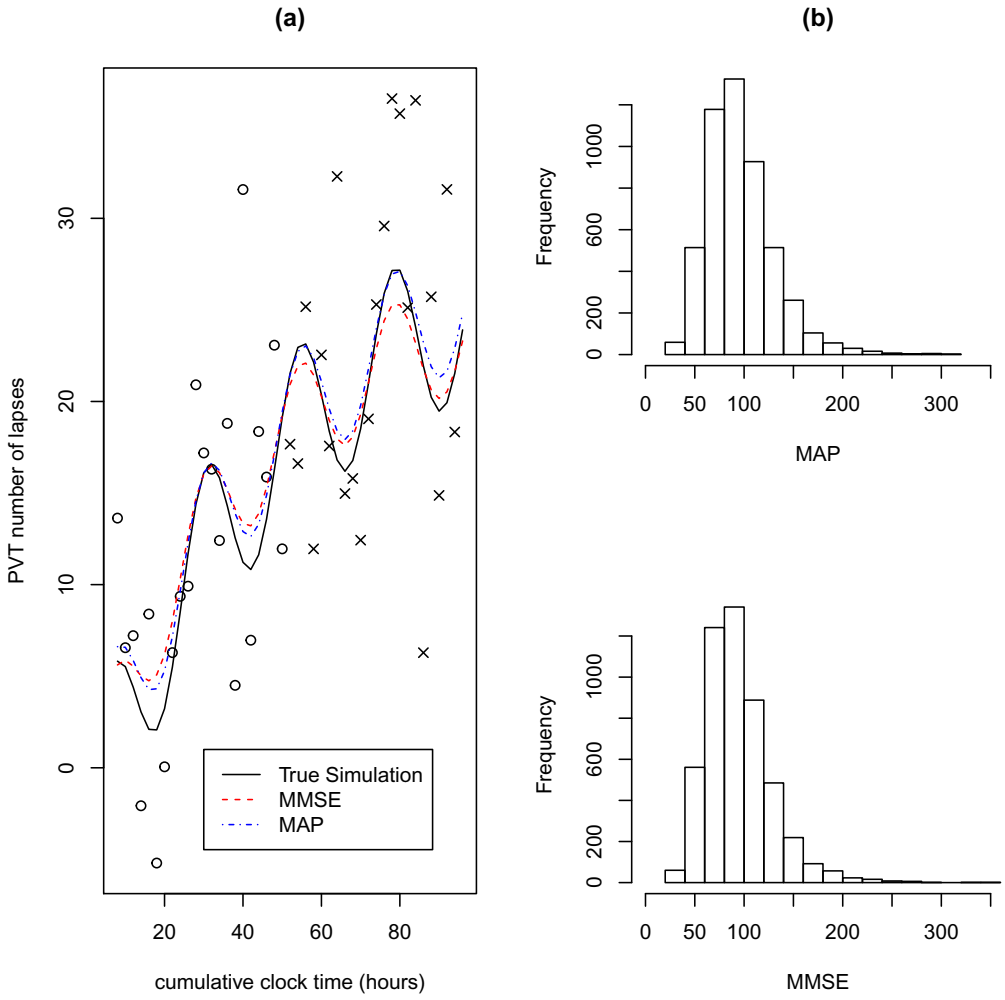

Figure 4.1: (a) Hypothetical PVT responses for a single individual from the parameterized model given in Van Dongen et al. [2007]. MMSE and MAP predictors of performance are determined numerically from the simulated data and plotted (dotted lines) and compared with the response that is obtained from the individual's true parameters (solid line). (b) Frequency plots for the MSE of the MAP and MMSE estimators across a simulation of 5000 individuals.

where

$$\xi(\phi) = \phi + \upsilon\phi^2. \tag{4.3.2}$$

Without loss of generality, we assume that the constant υ is positive. This model has a minimum value of

$$y = \xi_{\min} \equiv -\frac{1}{4\upsilon} \tag{4.3.3}$$

which occurs at

$$\phi = \phi_{\min} \equiv -\frac{1}{2\upsilon}. \tag{4.3.4}$$

We assume that

$$\phi \sim \mathcal{N}\left(\mu, \delta^2\right) \tag{4.3.5}$$

and

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right). \tag{4.3.6}$$

The assumption of a quadratic model will result in a conditional distribution with two peaks in the case that the observed data point $y > \xi_{\min}$ (See figure 4.2.) We also assume that the distance between the two peaks is much greater than the width of each peak and that the width of the prior distribution is much greater than the width of each peak as formalized later in (4.3.16).

Intractable integral in computing the posterior distribution

To obtain the Bayesian MSE ratio, we must (a) obtain $\mathcal{R}\left[\hat{\phi}^A\right]$ and $\mathcal{R}\left[\hat{\phi}^E\right]$, (b) obtain $\tilde{\mathcal{R}}\left[\hat{\phi}^A\right]$ and $\tilde{\mathcal{R}}\left[\hat{\phi}^E\right]$, and finally (c) obtain $\tilde{\mathcal{P}}\left[\hat{\phi}^A, \hat{\phi}^E\right]$. Analytical derivation of $\mathcal{R}\left[\hat{\phi}^A\right]$ and $\mathcal{R}\left[\hat{\phi}^E\right]$ requires computation of an integral containing the posterior distribution $f_{\phi|y}$. As discussed in Section 4.1, we cannot analytically compute the normalization constant of the posterior distribution for an arbitrary nonlinear function (4.3.2). We show here that this remains the case when we specify that nonlinear function to be a quadratic function. The posterior distribution $f_{\phi|y}$ is found using Bayes Theorem,

$$f_{\phi|y} = \frac{f_{y|\phi} f_\phi}{f_y} = \frac{f_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} f_{y|\phi} f_\phi d\phi}, \tag{4.3.7}$$

where f_y represents the marginal distribution of y. It follows from (4.3.6) that the conditional distribution of the data point given the parameter is

$$f_{y|\phi} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y-\left(\phi+\upsilon\phi^2\right)\right)^2}{2\sigma^2}}, \tag{4.3.8}$$

and additionally from (4.3.5) that the prior distribution of ϕ is

$$f_\phi = \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(\phi-\mu)^2}{2\delta^2}}. \tag{4.3.9}$$

The integral in (4.3.7) can therefore be written as

$$f_y = \int_{-\infty}^{\infty} f_{y|\phi} f_\phi d\phi = \frac{1}{2\pi\sigma\delta} \int_{-\infty}^{\infty} e^{-\frac{\left(y-\left(\phi+\upsilon\phi^2\right)\right)^2}{2\sigma^2} - \frac{(\phi-\mu)^2}{2\delta^2}} d\phi, \tag{4.3.10}$$

where the integral contains a quartic polynomial function of ϕ, and in general does not have a closed form solution.

Peak-specific conditional approximation

One solution to the integration problem is to construct a linear approximation to the model function around the two peaks of the conditional distribution, which we will refer to as ϕ_ℓ and ϕ_r, where $\phi_\ell < \phi_r$. The modes of the likelihood are found by setting the residual error to zero,

$$\left(y - \left(\phi + \upsilon\phi^2\right)\right)^2 = 0, \tag{4.3.11}$$

and solving for ϕ. The solutions are

$$\phi_\ell = -\frac{1}{2\upsilon} - \frac{\sqrt{y - \frac{-1}{4\upsilon}}}{\sqrt{\upsilon}} = \phi_{\min} - \frac{\sqrt{y - \xi_{\min}}}{\sqrt{\upsilon}} \tag{4.3.12}$$

and

$$\phi_r = -\frac{1}{2\upsilon} + \frac{\sqrt{y - \frac{-1}{4\upsilon}}}{\sqrt{\upsilon}} = \phi_{\min} + \frac{\sqrt{y - \xi_{\min}}}{\sqrt{\upsilon}}. \tag{4.3.13}$$

Let us define $\tilde{\xi}_x(\phi)$ to represent the Taylor series expansion of $\xi(\phi)$ around one of the modes of the likelihood $\phi_x \in \{\phi_\ell, \phi_r\}$, truncated at first term:

$$\xi_x(\phi) \approx \tilde{\xi}_x(\phi) \equiv \xi(\phi_x) + \xi'(\phi_x)(\phi - \phi_x), \tag{4.3.14}$$

We define $\tilde{f}^x_{y|\phi}$ to represent the corresponding mode-specific first-order approximation to the conditional distribution, also referred to as Laplace's approximation, where the tilde is

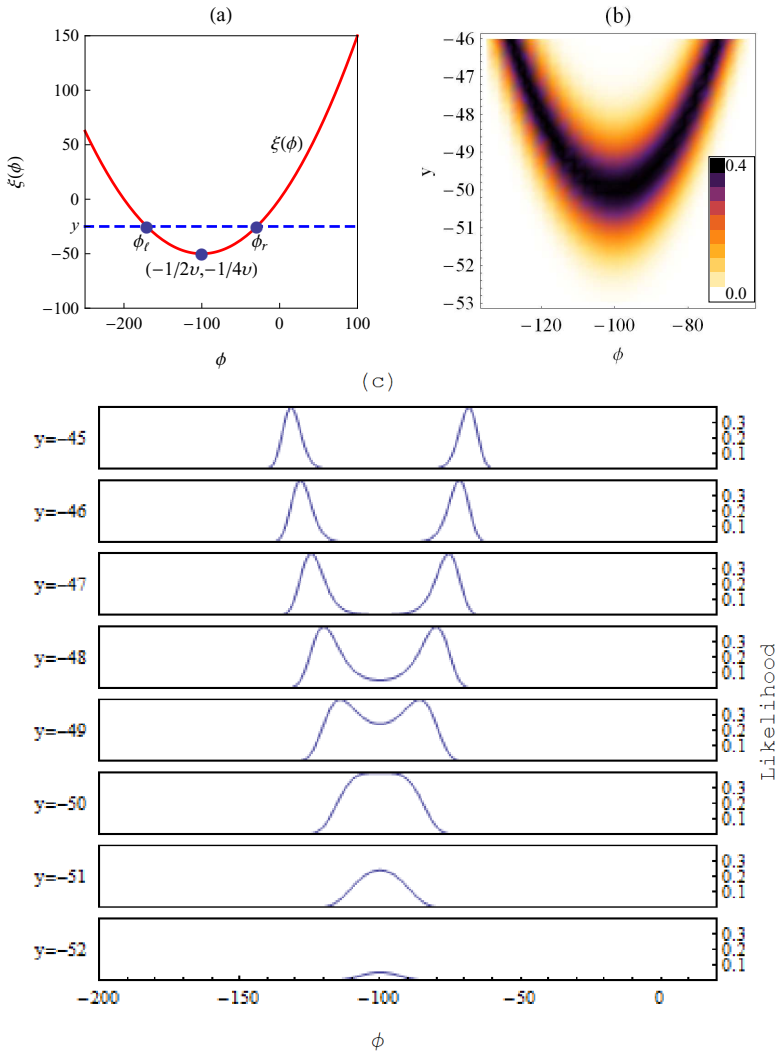

Figure 4.2: Bayesian forecasting example for the model $y = \phi + 0.005\phi^2 + \epsilon$, where $\phi \sim \mathcal{N}\left(-90, 100^2\right)$, $\epsilon \sim \mathcal{N}\left(0, 1^2\right)$, and a single data point is observed as $y = -40$. Parameter values were selected to be consistent with the constraint set of (4.3.16). (a) The solid line represents the quadratic model in ϕ-space, and the dashed line represents the value of a single observed data point y. The vertex is located at $\phi_{\min} = \frac{-1}{2\upsilon} = -100$ and $\xi(\phi) = \frac{-1}{4\upsilon} = -50$. The values $\phi_\ell \approx 145$ and $\phi_r \approx -55$ represent the values of ϕ where the model intersects with the value of the single data point. (b) Plot of the likelihood, where greater likelihood is shown with darker colors, for different values of y and ϕ. This plot focuses on the region near the vertex in (a). (c) Plots of the likelihood (shown on the right side axis) vs. ϕ for different values of y (shown on the left side axis).

used to denote that it is approximating the distribution, the superscript is representative of the neighborhood in which we assume ϕ to lie (left or right mode), and the subscript identifies the probability distribution being approximated. The conditional approximation can be written as

$$\begin{aligned}\tilde{f}^{x}_{y|\phi} &= \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\left(y-\tilde{\xi}_x(\phi)\right)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\left(y-\xi(\phi_x)-\xi'(\phi_x)(\phi-\phi_x)\right)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\left(\xi'(\phi_x)(\phi-\phi_x)\right)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\phi-\phi_x)^2}{2\left(\frac{\sigma}{\xi'(\phi_x)}\right)^2}}.\end{aligned} \tag{4.3.15}$$

Each peak of the conditional distribution has standard deviation $\left|\frac{\sigma}{\xi'(\phi_x)}\right|$, where $\xi'(\phi_x)$ represents the derivative of the model function evaluated at $\phi_x \in \{\phi_\ell, \phi_r\}$. Due to the exponential decrease, the kernel of the conditional distribution is only of significance in a region of size $\alpha\left|\frac{\sigma}{\xi'(\phi_x)}\right|$ around ϕ_x, where α is some constant of moderate size. Outside of this region, we can approximate the kernel of the conditional distribution by zero. The conditional approximation will be valid when the approximations of the conditional distributions do not overlap. This will occur when the distance between the two mode specific approximations is much smaller than the distance between the two modes, or

$$\left(\frac{\sigma}{\xi'(\phi_x)}\right)^2 \ll (\phi_r - \phi_\ell)^2. \tag{4.3.16}$$

Under this condition, we may reasonably approximate the conditional distribution as

$$f_{y|\phi} \approx \tilde{f}^{\ell}_{y|\phi} + \tilde{f}^{r}_{y|\phi}. \tag{4.3.17}$$

Approximating the posterior distribution

Our method of approximating the conditional distribution allows us to now formulate the posterior distribution with analytically computable integrals. Substituting (4.3.17) into (4.3.7), seperating fractions, and multiplying and dividing by $\int_{-\infty}^{\infty} \tilde{f}^{\ell}_{y|\phi} f_\phi d\phi$ and $\int_{-\infty}^{\infty} \tilde{f}^{r}_{y|\phi} f_\phi d\phi$,

respectively, results in

$$\begin{aligned}
f_{\phi|y} &= \frac{f_{y|\phi} f_\phi}{f_y} \\
&= \frac{f_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} f_{y|\phi} f_\phi d\phi} \\
&\approx \frac{\left(\tilde{f}^\ell_{y|\phi} + \tilde{f}^r_{y|\phi}\right) f_\phi}{\int_{-\infty}^{\infty} \left(\tilde{f}^\ell_{y|\phi} + \tilde{f}^r_{y|\phi}\right) f_\phi d\phi} \\
&= \frac{\tilde{f}^\ell_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} \left(\tilde{f}^\ell_{y|\phi} + \tilde{f}^r_{y|\phi}\right) f_\phi d\phi} + \frac{\tilde{f}^r_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} \left(\tilde{f}^\ell_{y|\phi} + \tilde{f}^r_{y|\phi}\right) f_\phi d\phi} \\
&= \frac{\int_{-\infty}^{\infty} \tilde{f}^\ell_{y|\phi} f_\phi d\phi}{\int_{-\infty}^{\infty} \tilde{f}^\ell_{y|\phi} f_\phi d\phi + \int_{-\infty}^{\infty} \tilde{f}^r_{y|\phi} f_\phi d\phi} \frac{\tilde{f}^\ell_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} \tilde{f}^\ell_{y|\phi} f_\phi d\phi} \\
&+ \frac{\int_{-\infty}^{\infty} \tilde{f}^r_{y|\phi} f_\phi d\phi}{\int_{-\infty}^{\infty} \tilde{f}^\ell_{y|\phi} f_\phi d\phi + \int_{-\infty}^{\infty} \tilde{f}^r_{y|\phi} f_\phi d\phi} \frac{\tilde{f}^r_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} \tilde{f}^r_{y|\phi} f_\phi d\phi}.
\end{aligned} \tag{4.3.18}$$

We define

$$\tilde{f}^x_y \equiv \int_{-\infty}^{\infty} \tilde{f}^x_{y|\phi} f_\phi d\phi, \tag{4.3.19}$$

so that

$$\tilde{f}_y = \tilde{f}^\ell_y + \tilde{f}^r_y. \tag{4.3.20}$$

In addition, we define $\tilde{f}^x_{\phi|y}$ to be the approximate posterior probability of ϕ given the value of y and assuming that ϕ is in the neighborhood of ϕ_x, and note that it can be written as

$$\tilde{f}^x_{\phi|y} = \frac{\tilde{f}^x_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} \tilde{f}^x_{y|\phi} f_\phi d\phi}. \tag{4.3.21}$$

We then substitute $\tilde{f}^x_{\phi|y}$ into (4.3.18) to obtain an analytically computable formula to approximate the posterior distribution,

$$f_{\phi|y} \approx \frac{\tilde{f}^\ell_y}{\tilde{f}^\ell_y + \tilde{f}^r_y} \tilde{f}^\ell_{\phi|y} + \frac{\tilde{f}^r_y}{\tilde{f}^\ell_y + \tilde{f}^r_y} \tilde{f}^r_{\phi|y}. \tag{4.3.22}$$

We then simplify the notation by defining the notations

$$p_\ell \equiv \frac{\tilde{f}_y^\ell}{\tilde{f}_y^\ell + \tilde{f}_y^r} \tag{4.3.23}$$

and

$$p_r \equiv \frac{\tilde{f}_y^r}{\tilde{f}_y^\ell + \tilde{f}_y^r}. \tag{4.3.24}$$

We also find that

$$\frac{\tilde{f}_y^\ell}{\tilde{f}_y^\ell + \tilde{f}_y^r} = \frac{e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}}}{e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}} + e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}} \tag{4.3.25}$$

and

$$\frac{\tilde{f}_y^r}{\tilde{f}_y^\ell + \tilde{f}_y^r} = \frac{e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}}{e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}} + e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}}. \tag{4.3.26}$$

To evaluate (4.3.22), we compute the integral from (4.3.19) using Equation (D.0.9) to be

$$\begin{aligned}
\tilde{f}_y^x &= \int_{-\infty}^{\infty} \tilde{f}_{y|\phi}^x f_\phi d\phi \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\phi-\phi_x)^2}{2\left(\frac{\sigma}{\xi'(\phi_x)}\right)^2}} \frac{1}{\sqrt{2\pi}\delta} e^{\frac{-(\phi-\mu)^2}{2\delta^2}} d\phi \\
&= \frac{1}{2\pi\sigma\delta} \int_{-\infty}^{\infty} e^{\frac{-(\phi-\phi_x)^2}{2\left(\frac{\sigma}{\xi'(\phi_x)}\right)^2} - \frac{(\phi-\mu)^2}{2\delta^2}} d\phi \\
&= \frac{1}{2\pi\sigma\delta} \int_{-\infty}^{\infty} e^{\frac{-\left[\delta^2(\phi-\phi_x)^2 + \left(\left(\frac{\sigma}{\xi'(\phi_x)}\right)^2(\phi-\mu)^2\right)\right]}{2\left[\frac{\sigma}{\xi'(\phi_x)}\right]^2\delta^2}} d\phi \\
&= \frac{1}{2\pi\sigma\delta} \int_{-\infty}^{\infty} e^{\frac{-\phi^2\left(\delta^2 + \left[\frac{\sigma}{\xi'(\phi_x)}\right]^2\right) + 2\phi\left[\delta^2\phi_x + \left(\frac{\sigma}{\xi'(\phi_x)}\right)^2\mu\right] - \delta^2\phi_x^2 - \left(\frac{\sigma}{\xi'(\phi_x)}\right)^2\mu^2}{2\left(\frac{\sigma}{\xi'(\phi_x)}\right)^2\delta^2}} d\phi \\
&= \frac{e^{-\frac{\left(\xi_x'\right)^2(\phi_r-\mu)^2}{2\left(\delta^2\left(\xi_x'\right)^2+\sigma^2\right)}}}{\sqrt{2\pi}\delta\sqrt{\frac{1}{\delta^2} + \frac{\left(\xi_x'\right)^2}{\sigma^2}\sigma}}.
\end{aligned} \tag{4.3.27}$$

Further, we evaluate $f_{\phi|y}$ under the conditions of (4.3.16), by determining the value of

$\tilde{f}_y^x$ in the limit as $\sigma \to 0$ to be

$$\tilde{f}_y^x \approx \frac{e^{-\frac{(\phi_x-\mu)^2}{2\delta^2}}}{\sqrt{2\pi}\delta\,|1+2\upsilon\phi_x|}. \tag{4.3.28}$$

We have thus used the approximate conditional distribution to obtain an approximation for the parameter posterior distribution in (4.3.22). We can now use the posterior distribution to obtain the parameter MMSE and MAP estimators.

Parameter MMSE estimator

With the posterior distribution now determined, we continue towards approximating the Bayesian mean squared parameter error ratio $\tilde{\mathcal{P}}\left[\hat{\phi}^A, \hat{\phi}^E\right]$ by first computing the estimator $\hat{\phi}^E$, which is necessary to obtain $\mathcal{R}\left[\hat{\phi}^E\right]$. Since we will need to compute additional integrals containing $f_{\phi|y}$ in this chapter, we first make the following general determination. Let us define

$$E_{\phi|y}^x[g(\phi)] \equiv \int_{-\infty}^{\infty} g(\phi)\tilde{f}_{\phi|y}^x d\phi \tag{4.3.29}$$

and

$$\mathrm{Var}_{\phi|y}^x[\phi] \equiv \int_{-\infty}^{\infty} \left(g(\phi) - E_{\phi|y}^x[g(\phi)]\right)^2 \tilde{f}_{\phi|y}^x d\phi. \tag{4.3.30}$$

We find that the expectation of an arbitrary function of ϕ w.r.t. the distribution $f_{\phi|y}$ can be approximated by a weighted sum of expectations,

$$\begin{aligned} E_{\phi|y}[g(\phi)] &\approx \int_{-\infty}^{\infty} g(\phi)\left(p_\ell \tilde{f}_{\phi|y}^\ell + p_r \tilde{f}_{\phi|y}^r\right) d\phi \\ &= p_\ell \int_{-\infty}^{\infty} g(\phi)\tilde{f}_{\phi|y}^\ell d\phi + p_r \int_{-\infty}^{\infty} g(\phi)\tilde{f}_{\phi|y}^r d\phi \\ &= p_\ell E_{\phi|y}^\ell[g(\phi)] + p_r E_{\phi|y}^r[g(\phi)]. \end{aligned} \tag{4.3.31}$$

To compute an analytic approximation for $\hat{\phi}^E$, we use (4.3.31) with $g(\phi) = \phi$ to obtain

$$\hat{\phi}^E \approx p_\ell E_{\phi|y}^\ell[\phi] + p_r E_{\phi|y}^r[\phi]. \tag{4.3.32}$$

To determine $E_{\phi|y}^\ell[\phi]$ and $E_{\phi|y}^r[\phi]$, we use Appendix D, Theorem (**??**) to find that

$$E_{\phi|y}^x[\phi] = \frac{\left(\frac{y}{\sigma^2} + \frac{\xi'(\phi_x)\mu - \upsilon\phi_x^2}{(a\delta)^2}\right)\left(\frac{1}{\sigma^2} + \frac{1}{(a\delta)^2}\right)^{-1} - \upsilon\phi_x^2}{\xi'(\phi_x)}. \tag{4.3.33}$$

Applying (4.3.33) under the conditions of (4.3.16), we can approximate $E^{\ell}_{\phi|y}[\phi]$ and $E^{r}_{\phi|y}[\phi]$ by taking the limit as $\sigma \to 0$, resulting in

$$\begin{aligned} E^{\ell}_{\phi|y}[\phi] &\approx \frac{y + \upsilon\phi_{\ell}^2}{\xi'(\phi_{\ell})} \\ &= -\frac{1 + 4y\upsilon + \sqrt{1+4y\upsilon}}{2\upsilon\sqrt{1+4y\upsilon}} \\ &= -\frac{1+\xi_r^{'}}{2\upsilon} \\ &= -\frac{1}{2\upsilon} - \frac{\xi_r^{'}}{2\upsilon} \\ &= \phi_{\min} - \frac{\sqrt{1+4y\upsilon}}{2\upsilon} \\ &= \phi_{\min} - \frac{\sqrt{\frac{1}{4\upsilon} + y}}{\sqrt{\upsilon}} \\ &= \phi_{\min} - \frac{\sqrt{y - \xi_{\min}}}{\sqrt{\upsilon}} \\ &= \phi_{\ell} \end{aligned} \tag{4.3.34}$$

and

$$\begin{aligned} E^{r}_{\phi|y}[\phi] &\approx \frac{y + \upsilon\phi_{r}^2}{\xi'(\phi_{r})} \\ &= \frac{1 + 4y\upsilon - \sqrt{1+4y\upsilon}}{2\upsilon\sqrt{1+4y\upsilon}} \\ &= -\frac{1-\xi_r^{'}}{2\upsilon} \\ &= -\frac{1}{2\upsilon} + \frac{\xi_r^{'}}{2\upsilon} \\ &= \phi_{\min} + \frac{\sqrt{1+4y\upsilon}}{2\upsilon} \\ &= \phi_{\min} + \frac{\sqrt{\frac{1}{4\upsilon} + y}}{\sqrt{\upsilon}} \\ &= \phi_{\min} + \frac{\sqrt{y - \xi_{\min}}}{\sqrt{\upsilon}} \\ &= \phi_{r}. \end{aligned} \tag{4.3.35}$$

Therefore, the MMSE can be approximated as

$$\begin{aligned}\hat{\phi}^E &\approx p_\ell \phi_\ell + p_r \phi_r \\ &= \frac{e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}}\phi_\ell + e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}\phi_r}{e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}} + e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}}.\end{aligned} \tag{4.3.36}$$

Parameter MAP estimator

Substituting $f_{y|\phi}$ from (4.3.7) into the definition of the parameter MAP (1.2.5), we find that

$$\hat{\phi}^A = \underset{\phi}{\operatorname{argmax}} \left[\frac{f_{y|\phi} f_\phi}{\int_{-\infty}^{\infty} f_{y|\phi} f_\phi d\phi} \right]. \tag{4.3.37}$$

Using the knowledge that the kernels described by $f_{y|\phi}$ are seperate (i.e., for a particular value of ϕ, either $\tilde{f}^\ell_{y|\phi} \approx 0$ or $\tilde{f}^r_{y|\phi} \approx 0$), we determine that the parameter MAP will be the mean of either the left or right kernel,

$$\hat{\phi}^A \approx E^\ell_{\phi|y} \tag{4.3.38}$$

or

$$\hat{\phi}^A \approx E^r_{\phi|y}. \tag{4.3.39}$$

Furthermore, using the approximation for $E^x_{\phi|y}$ given in (4.3.34), we determine that the MAP will occur approximately at one of the values of ϕ where the graph of the model function intersects the value of the data point,

$$\hat{\phi}^A \approx \phi_\ell \tag{4.3.40}$$

or

$$\hat{\phi}^A \approx \phi_r. \tag{4.3.41}$$

Since $f_{y|\phi}(y|\phi = \phi_\ell) = f_{y|\phi}(y|\phi = \phi_r)$, whether the MAP is the mean of the left or right peak is determined completely by the value of the prior f_ϕ at the two intersection points. We find that

the parameter MAP is ϕ_ℓ if $f_\phi(\phi = \phi_\ell) > f_\phi(\phi = \phi_r)$ and ϕ_r otherwise. Substituting in the prior distribution, this condition becomes

$$(\phi_\ell - \mu)^2 < (\phi_r - \mu)^2, \tag{4.3.42}$$

which further simplifies to

$$\mu < \phi_{\min}. \tag{4.3.43}$$

Therefore, the MAP estimator is

$$\hat{\phi}^A \approx \underset{\phi}{\operatorname{argmax}} \left[\begin{cases} \phi_\ell & \text{if} \mu < \phi_{\min} \\ \phi_r & \text{otherwise} \end{cases} \right]. \tag{4.3.44}$$

Posterior expected squared loss for the parameter MMSE estimator

Now that we have determined analytic approximations for $\hat{\phi}^E$ and $\hat{\phi}^A$, we continue towards obtaining the Bayesian mean squared error ratio $\tilde{\mathcal{P}}\left[\hat{\phi}^A, \hat{\phi}^E\right]$ by obtaining the posterior expected squared loss $\mathcal{R}\left[\hat{\phi}^E\right]$ for the MMSE estimator,

$$\begin{aligned} \mathcal{R}\left[\hat{\phi}^E\right] &= E_{\phi|y}\left[\left(\phi - E_{\phi|y}[\phi]\right)^2\right] \\ &\approx p_\ell E^\ell_{\phi|y}\left[\left(\phi - E_{\phi|y}[\phi]\right)^2\right] + p_r E^r_{\phi|y}\left[\left(\phi - E_{\phi|y}[\phi]\right)^2\right]. \end{aligned} \tag{4.3.45}$$

To compute the unknown terms in this expression, we apply properties of variance and expectation

$$\begin{aligned} E^x_{\phi|y}\left[\left(\phi - E_{\phi|y}[\phi]\right)^2\right] &= \mathrm{Var}^x_{\phi|y}[\phi] + \left(E^x_{\phi|y}\left[\phi - E_{\phi|y}[\phi]\right]\right)^2 \\ &= \mathrm{Var}^x_{\phi|y}[\phi] + \left(E^x_{\phi|y}[\phi] - E_{\phi|y}[\phi]\right)^2 \end{aligned} \tag{4.3.46}$$

Therefore,

$$\begin{aligned} \mathcal{R}\left[\hat{\phi}^E\right] \approx & p_\ell \left(\mathrm{Var}^\ell_{\phi|y}[\phi] + \left(\phi_\ell - E_{\phi|y}[\phi]\right)^2\right) \\ & + p_r \left(\mathrm{Var}^r_{\phi|y}[\phi] + \left(\phi_r - E_{\phi|y}[\phi]\right)^2\right). \end{aligned} \tag{4.3.47}$$

We further simplify by showing that $\mathrm{Var}^\ell_{\phi|y}[\phi] = \mathrm{Var}^r_{\phi|y}[\phi]$. To accomplish this, first note that $\xi'(\phi_\ell)$, the derivative of the model function w.r.t. ϕ at ϕ_ℓ, is the negative of $\xi'(\phi_r)$, the derivative of the model function w.r.t. ϕ at ϕ_r :

$$\xi'(\phi_\ell) = 1 + 2\upsilon\phi_\ell = 1 + 2\upsilon\left(-\frac{1}{2\upsilon} - \frac{\sqrt{y - \frac{-1}{4\upsilon}}}{\sqrt{\upsilon}}\right) = -\sqrt{1 + 4y\upsilon} \tag{4.3.48}$$

and

$$\xi'(\phi_r) = 1 + 2\upsilon\phi_r = 1 + 2\upsilon\left(-\frac{1}{2\upsilon} + \frac{\sqrt{y - \frac{-1}{4\upsilon}}}{\sqrt{\upsilon}}\right) = \sqrt{1 + 4y\upsilon}. \tag{4.3.49}$$

The squared derivatives are thus equal:

$$\left(\xi'\left(\phi_\ell\right)\right)^2 = \left(\xi'\left(\phi_r\right)\right)^2 = 1 + 4y\upsilon. \tag{4.3.50}$$

Therefore,

$$\mathrm{Var}^{\ell}_{\phi|y}[\phi] = \mathrm{Var}^{r}_{\phi|y}[\phi] = \left(\frac{\left(\xi'\left(\phi_\ell\right)\right)^2}{\sigma^2} + \frac{1}{\delta^2}\right)^{-1}. \tag{4.3.51}$$

We can also simplify $\mathcal{R}\left[\hat{\phi}_e\right]$ by showing that the differences,

$$\begin{aligned} \phi_\ell - E_{\phi|y}[\phi] &\approx \phi_\ell - \left(p_\ell\phi_\ell + p_r\phi_r\right) \\ &= p_r\left(\phi_\ell - \phi_r\right) \end{aligned} \tag{4.3.52}$$

and

$$\begin{aligned} \phi_r - E_{\phi|y}[\phi] &\approx \phi_r - \left(p_\ell\phi_\ell + p_r\phi_r\right) \\ &= p_\ell\left(\phi_r - \phi_\ell\right). \end{aligned} \tag{4.3.53}$$

Substituting both simplifications in, we find that

$$\begin{aligned} \mathcal{R}\left[\hat{\phi}_e\right] &\approx \left(\frac{\left(\xi'\left(\phi_\ell\right)\right)^2}{\sigma^2} + \frac{1}{\delta^2}\right)^{-1} + p_\ell\left(p_r\left(\phi_\ell - \phi_r\right)\right)^2 \\ &\quad + p_r\left(p_\ell\left(\phi_r - \phi_\ell\right)\right)^2 \\ &= \left(\frac{\left(\xi'\left(\phi_\ell\right)\right)^2}{\sigma^2} + \frac{1}{\delta^2}\right)^{-1} + p_\ell p_r\left(p_\ell + p_r\right)\left(\phi_\ell - \phi_r\right)^2 = \\ &= \left(\frac{\left(\xi'\left(\phi_\ell\right)\right)^2}{\sigma^2} + \frac{1}{\delta^2}\right)^{-1} + p_\ell p_r\left(\phi_\ell - \phi_r\right)^2. \end{aligned} \tag{4.3.54}$$

Under the conditions of (4.3.16), $\left(\frac{\left(\xi'(\phi_\ell)\right)^2}{\sigma^2} + \frac{1}{\delta^2}\right)^{-1}$will be small compared with the other term in the expression (4.3.54). Consequently, using (4.3.23), we find that

$$\mathcal{R}\left[\hat{\phi}^E\right] = E\left[\left(\phi - E_{\phi|y}[\phi]\right)^2\right] \approx p_\ell p_r\left(\phi_\ell - \phi_r\right)^2 = \frac{e^{\frac{-\left(\phi_\ell-\mu\right)^2}{2\delta^2} + \frac{-\left(\phi_r-\mu\right)^2}{2\delta^2}}\left(1 + 2\upsilon\phi_r\right)^2}{\left(e^{\frac{-\left(\phi_\ell-\mu\right)^2}{2\delta^2}} + e^{\frac{-\left(\phi_r-\mu\right)^2}{2\delta^2}}\right)^2\upsilon^2}. \tag{4.3.55}$$

Posterior expected squared loss for the parameter MAP estimator

Having now determined $\mathcal{R}\left[\hat{\phi}^E\right]$, we continue towards $\tilde{\mathcal{P}}\left[\hat{\phi}^A, \hat{\phi}^E\right]$ by determining the posterior expected squared parameter loss $\mathcal{R}\left[\hat{\phi}^A\right]$ of the MAP estimator:

$$\begin{aligned}\mathcal{R}\left[\hat{\phi}^A\right] &= E_{\phi|y}\left[(\phi-\phi_x)^2\right] \\ &= \mathrm{Var}_{\phi|y}\left[\phi-\phi_x\right] + \left(E_{\phi|y}\left[\phi-\phi_x\right]\right)^2 \\ &= \mathrm{Var}_{\phi|y}[\phi] + \left(E_{\phi|y}[\phi]-\phi_x\right)^2 \\ &= \mathcal{R}\left[\hat{\phi}^E\right] + \left(E_{\phi|y}[\phi]-\phi_x\right)^2.\end{aligned} \tag{4.3.56}$$

In the case when $\hat{\phi}^A \approx \phi_\ell$ $\left(i.e., \mu < -\frac{1}{2\upsilon}\right)$, we find that

$$\mathcal{R}\left[\hat{\phi}^A\right] - \mathcal{R}\left[\hat{\phi}^E\right] = \left(E_{\phi|y}[\phi]-\phi_\ell\right)^2 \approx \left(p_\ell\phi_\ell + p_r\phi_r - \phi_\ell\right)^2 = p_r^2\left(\phi_r-\phi_\ell\right)^2, \tag{4.3.57}$$

and

$$\mathcal{R}\left[\hat{\phi}^A\right] \approx p_r\left(\phi_r-\phi_\ell\right)^2 = \frac{e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}}\left(1+2\upsilon\phi_r\right)^2}{\left(e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}} + e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}\right)\upsilon^2}. \tag{4.3.58}$$

In the case when $\hat{\phi}^A \approx \phi_r$ $\left(i.e., \mu > -\frac{1}{2\upsilon}\right)$, we find

$$\mathcal{R}\left[\hat{\phi}^A\right] - \mathcal{R}\left[\hat{\phi}^E\right] = \left(E_{\phi|y}[\phi]-\phi_r\right)^2 \approx \left(p_\ell\phi_\ell + p_r\phi_r - \phi_r\right)^2 = p_\ell^2\left(\phi_r-\phi_\ell\right)^2, \tag{4.3.59}$$

and

$$\mathcal{R}\left[\hat{\phi}^A\right] \approx p_\ell\left(\phi_r-\phi_\ell\right)^2 = \frac{e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}\left(1+2\upsilon\phi_r\right)^2}{\left(e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}} + e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}\right)\upsilon^2}. \tag{4.3.60}$$

Bayesian mean squared error for the parameter MMSE estimator

Now having found the the posterior expected squared loss for both estimators $\mathcal{R}\left[\hat{\phi}^E\right]$ and $\mathcal{R}\left[\hat{\phi}^A\right]$ we move to determining the Bayesian mean squared parameter errors $\tilde{\mathcal{R}}\left[\hat{\phi}^E\right]$ and $\tilde{\mathcal{R}}\left[\hat{\phi}^A\right]$ and the Bayesian mean squared parameter error ratio $\tilde{\mathcal{P}}\left[\hat{\phi}^A, \hat{\phi}^E\right]$. Obtaining $\tilde{\mathcal{R}}\left[\hat{\phi}^E\right]$ requires computing an integral over y; however, examination of (4.3.55) suggests that it would be easier to evaluate this integral after a transformation of variable to ϕ_r (or ϕ_ℓ). We define $\tilde{f}_\phi^r$ to be the approximate distribution of ϕ_r defined over the region $\phi_r > \phi_{\min}$. Using the method of transforming distributions (see Appendix D), equation (4.3.28) and equation (4.3.20), and noting that

$$\phi_\ell = 2\phi_{\min} - \phi_r, \tag{4.3.61}$$

we find that

$$\begin{aligned}\tilde{f}_{\phi}^{r} &= \tilde{f}_{y}\left|\frac{\partial y}{\partial \phi_r}\right| \\ &= \frac{e^{-\frac{(\phi_\ell-\mu)^2}{2\delta^2}}+e^{-\frac{(\phi_r-\mu)^2}{2\delta^2}}}{\sqrt{2\pi}\delta\left(1+2\upsilon\phi_r\right)}\left(1+2\upsilon\phi_r\right) \\ &= \frac{e^{-\frac{(\phi_\ell-\mu)^2}{2\delta^2}}+e^{-\frac{(\phi_r-\mu)^2}{2\delta^2}}}{\sqrt{2\pi}\delta} \\ &= \frac{e^{-\frac{(\phi_r-\mu)^2}{2\delta^2}}+e^{-\frac{(\mu-(\phi_{\min}-(\phi_r-\phi_{\min})))^2}{2\delta^2}}}{\sqrt{2\pi}\delta}.\end{aligned} \tag{4.3.62}$$

Having obtained the distribution of ϕ_r, we now take the expectation over the distribution of ϕ_r to determine

$$\begin{aligned}\tilde{\mathcal{R}}\left[\hat{\phi}^{E}\right] &= \int_{-\infty}^{\infty} E_{\phi|y}\left[\left(\hat{\phi}^{E}-\phi\right)^2\right]dy \\ &= \int_{\phi_{\min}}^{\infty} E\left[\left(\phi-E_{\phi|y}[\phi]\right)^2\right]\tilde{f}_{\phi}^{r}d\phi_r \\ &= \int_{\phi_{\min}}^{\infty}\frac{\left(e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}+\frac{-(\phi_r-\mu)^2}{2\delta^2}}\right)\left(1+2\upsilon\phi_r\right)^2}{\left(e^{\frac{-(\phi_\ell-\mu)^2}{2\delta^2}}+e^{\frac{-(\phi_r-\mu)^2}{2\delta^2}}\right)^2\upsilon^2}\frac{\left(e^{-\frac{(\phi_\ell-\mu)^2}{2\delta^2}}+e^{-\frac{(\phi_r-\mu)^2}{2\delta^2}}\right)}{\sqrt{2\pi}\delta}d\phi_r \\ &= \int_{\phi_{\min}}^{\infty}\frac{\left(1+2\upsilon\phi_r\right)^2}{\left(e^{\frac{(\phi_\ell-\mu)^2}{2\delta^2}}+e^{\frac{(\phi_r-\mu)^2}{2\delta^2}}\right)\sqrt{2\pi}\delta\upsilon^2}d\phi_r\end{aligned}. \tag{4.3.63}$$

Per (4.3.16), δ^2 is large and

$$e^{\frac{(\phi_\ell-\mu)^2}{2\delta^2}}+e^{\frac{(\phi_r-\mu)^2}{2\delta^2}}\approx 2. \tag{4.3.64}$$

Therefore,

$$\tilde{\mathcal{R}}\left[\hat{\phi}^{E}\right]\approx\int_{\phi_{\min}}^{\infty}\frac{\left(1+2\upsilon\phi_r\right)^2}{2\sqrt{2\pi}\delta\upsilon^2}d\phi_r. \tag{4.3.65}$$

Bayesian mean squared error for the parameter MAP estimator

To obtain $\tilde{\mathcal{P}}\left[\hat{\phi}^A, \hat{\phi}^E\right]$ we need also to compute $\tilde{\mathcal{R}}\left[\hat{\phi}^A\right]$:

$$\begin{aligned}
\tilde{\mathcal{R}}\left[\hat{\phi}^A\right] &= E_{\phi|y}\left[(\phi-\phi_x)^2\right] \\
&= \int_{\phi_{\min}}^{\infty} E\left[(\phi-\phi_x)^2\right] \tilde{f}_{\phi}^{r} d\phi_r \\
&= \int_{\phi_{\min}}^{\infty} \frac{e^{\frac{(\phi_x-\mu)^2}{2\delta^2}}(1+2\upsilon\phi_r)^2}{\left(e^{\frac{(\phi_\ell-\mu)^2}{2\delta^2}}+e^{\frac{(\phi_r-\mu)^2}{2\delta^2}}\right)\upsilon^2} \frac{e^{-\frac{(\phi_\ell-\mu)^2}{2\delta^2}}+e^{-\frac{(\phi_r-\mu)^2}{2\delta^2}}}{\sqrt{2\pi}\delta} d\phi_r \qquad (4.3.66) \\
&= \int_{\phi_{\min}}^{\infty} \frac{e^{-\frac{(2\phi_{\min}-\phi_x-\mu)^2}{2\delta^2}}(1+2\upsilon\phi_r)^2}{\sqrt{2\pi}\delta\upsilon^2} d\phi_r.
\end{aligned}$$

By (4.3.16), δ^2 is large and

$$e^{-\frac{(\phi_x-2\phi_{\min}-\mu)^2}{2\delta^2}} \approx 1. \qquad (4.3.67)$$

Thus,

$$\tilde{\mathcal{R}}\left[\hat{\phi}^A\right] \approx \int_{\phi_{\min}}^{\infty} \frac{(1+2\upsilon\phi_r)^2}{\sqrt{2\pi}\delta\upsilon^2} d\phi_r = 2\tilde{\mathcal{R}}\left[\hat{\phi}^E\right] \qquad (4.3.68)$$

Therefore, for both the case $\hat{\phi}^A \approx \phi_\ell$ and $\hat{\phi}^A \approx \phi_r$, the marginal mean squared parameter error ratio is determined to be

$$\tilde{\mathcal{P}}\left[\hat{\phi}^A, \hat{\phi}^E\right] \approx 2, \qquad (4.3.69)$$

implying that the Bayesian mean squared parameter error for the MAP is twice as large as that for the MMSE.

To analytically differentiate between cases where the accuracy of the MMSE estimator may be reasonably assessed with the MAP estimator, we considered a Bayesian model for a single observed response where the model is quadratic in a single parameter. Construction of the posterior distribution resulted in incomputable integrals for arbitrary values of the error and prior variance. However, in the limiting case with small error variance and large prior variance, we found these integrals to be approximately computable. We utilized this special case to determine that the Bayesian mean squared error of the MAP estimator $\tilde{\mathcal{R}}\left[\hat{\phi}_a\right]$ was approximately twice that of the Bayesian mean squared error of the MMSE estimator $\tilde{\mathcal{R}}\left[\hat{\phi}_e\right]$. Therefore, in this scenario, we have found that the parameter MAP and

MMSE accuracies are not approximately equal.

Note that this result is equivalent to what would be readily obtained in a simpler scenario had we taken the limit as $\sigma \to 0$ initially to approximate the peaks of the conditional distribution with delta functions. Using this approach substantially simplifies the analytics, yielding equivalent final results with more readily computable integrals.

Chapter 5

Discussion

5.1 Summary

Cognitive fatigue due to sleep loss is a major risk in today's 24/7 society. Biomathematical models can be used to help mitigate such risks by predicting the timing at which high levels of fatigue will occur. Due to the large degree of individual variation in performance under sleep loss, the typically used group-average predictions are often inaccurate for a given individual. However, since individual differences are trait-like, between subjects variation can be captured by individualizing model parameters using the technique of Bayesian forecasting. Accuracy of predictions resulting from the individualization procedure depend on the amount of data collected on the individual at hand. This data is often limited by factors such as cost and availability. However, with the availability of correlated secondary performance measures, information may be included via a multivariate Bayesian forecasting framework to further enhance or meet required levels of prediction accuracy.

In Section 1.2, we consider the Bayesian MAP and MMSE estimators and predictors, and formulate measures of the accuracy of these estimators. To detail the accuracy assessment procedure, we derive the Bayesian MSE for the univariate Bayesian linear model of subject means in Section 2.1. We then generalize the derivation to a linear model for an arbitrary number of performance measures and covariates in Section 3.1. Interpreting how the accuracy depends on the sampling strategy (e.g., the number of measurements from the primary and secondary tasks) is difficult when this accuracy is specified in the matrix forms seen in Equations (3.1.22) and (3.2.15). To clarify our understanding in the simplest case that displays random effects correlation between tasks, we determine the accuracy for the

Bayesian bivariate linear model of subject means in Section 3.3. For this model, we further assume that observations from each performance measure cost a fixed price per data point, and use this assumption to determine the number of measurements of each variable needed to minimize the cost while still obtaining no less than the desired level of accuracy.

To aid in extending the findings from the linear case to state of the art nonlinear biomathematical fatigue models, we focus on obtaining the parameter estimate accuracy for the nonlinear case. Computing this accuracy analytically is often infeasible without reliance on model approximations. Model simulations can be used to compute this accuracy; however, such simulations can be time consuming, especially for models that lack analytic solutions and require that a system of differential equations be solved to produce model dynamics. Much of this computational burden in assessing estimator accuracy, however, is produced by using the Bayesian MMSE estimator, and could be reduced by taking advantage of the quicker to compute Bayesian MAP estimator. We show a nonlinear modeling example in which repeated simulation and estimation with the MAP estimator yields a reasonable estimate of the accuracy obtained using the MMSE estimator. Still, for any given case, determination of whether the MMSE accuracy can be approximated with the MAP accuracy requires these time consuming simulations. So as to remove the need for simulation with the MMSE, we begin to analytically identify classes of models where the MMSE accuracy can be approximated by the MAP accuracy. We consider a class of quadratic Bayesian models, and find that for these models the MAP and MMSE are not approximately equal.

5.2 Bayesian quadratic model

In Section 4.3, we discuss a class of Bayesian quadratic models, the motivation of which was to obtain an analytic method for determining whether the MMSE accuracy could reasonably be replaced with the MAP accuracy for a general Bayesian nonlinear model. To obtain an approximation of the nonlinear model, we considered using a Taylor series expansion. The focus on the quadratic model came out of the realization in Section 4.1 that a linear approximation to the nonlinear model can result in the misleading conclusion of equal MMSE and MAP accuracies. The quadratic model represented the simplest Taylor series approximation that resulted in different accuracies for the MAP and MMSE.

Assessment of the accuracy of the MMSE and MAP estimators for the quadratic Bayesian model resulted in the need to compute the integral of an exponential containing a quartic polynomial of the parameter. Finding this integral to have no closed form solution, we

searched instead for an approximate solution. A solution which led to the conclusion that the MAP accuracy was approximately twice that of the MMSE was readily found when we considered the conditions of an uninformative prior and small error variance. However; the value of this solution is limited, as the quadratic model is utilized over a region of parameter space of width $\phi_r - \phi_\ell$, covering both peaks of the conditional distribution (see Figure 4.1). Taylor's theorem specifies that a Taylor series approximation will be guaranteed accurate as we make the region of approximation arbitrarily small; however, the distance between peaks is specified to be large in comparison to the standard deviation of the individual peaks, $\left(\frac{\sigma}{\xi'(\phi_x)}\right)^2 \ll (\phi_r - \phi_\ell)^2$. Therefore, the condition of a large prior and small error variance will only be of interest in the specific case where we encounter a quadratic model with a large prior and small error variance.

The more useful case, where the prior and error variances are of comparatively moderate values, cannot be approximated using the same techniques as those used to solve the case of a large prior and small error variance. Alternatively, exponential integrals containing a quartic polynomial of the parameter can be expressed as sums of Bessel functions. Complications arise concerning the order at which to truncate these Bessel functions. Further work could be done to determine whether approximating the exponential integrals using Bessel functions could yield to quicker determination of whether it is reasonable to approximate the accuracy of the parameter or response MMSE estimator with that of the MAP. Finally, an additional possibility is as follows. When the MAP and MMSE do approximately correspond with each other, it may take fewer simulations to determine that the variability in the difference between the MAP and MMSE is small compared with the overall MMSE variability. Therefore, a few simulations with the MMSE may be done to determine correspondence, and the remainder of the simulations may be done with the MAP to assess the MMSE accuracy. Of course, if there are no limitation on computational resources and time, the MMSE may be repeatedly simulated to determine its own accuracy.

5.3 Random effects selection

In using a biomathematical model of performance to construct a population model, we make a choice concerning which model parameters will be random effects. In the Bayesian forecasting stage, the parameters designated random effects will be estimated for a new individual using subject-specific data. These parameters will likely differ in how much individual variation they explain over subjects on average. The ability to predict is limited by the proportion of the individual variability explained to begin with by the chosen random

effects. Therefore, choosing random effects that explain a substantive amount of individual variation is an important aspect for consideration. In this thesis, we focus our attention on accuracy of Bayesian forecasting predictions that assume an error-free, pre-specified population model; however, this framework could be further extended to include the dependence of Bayesian forecasting accuracy on the misspecification of random effects.

The choice of and correlation structure of random effects are specified through the design of the between-subjects variance/covariance matrix. This design is often chosen with a stepwise selection procedure; however, there is disagreement as to how to structure of covariance matrix to initialize the selection (Bonate, 2011). Suggestions include starting with a fully parameterized covariance matrix, starting with a diagonal covariance matrix, starting with only a single random effect, and motivating the initial choice using confidence plots of parameters for separately fit subjects (Pinheiro & Bates, 2009). As the final choice of random effects often depends on the initial covariance design, it should be noted that different population models may simultaneously be considered optimal by different individuals or modeling groups.

5.4 Nonlinear sampling designs of minimal cost

For the case of the bivariate linear model of subject means in Section 3.3, we showed how to determine the number of measurements on primary and secondary task variables that will minimize the cost of data collection while still meeting a specified level of prediction accuracy on a primary variable.

Additional complexities arise when assessing the accuracy of the analogous bivariate nonlinear model with a single random effect for each task. Firstly, as discussed in Section 4.1, the nonlinearity of the model will often result in an inability to analytically compute the prediction accuracy. The issue may be addressed in some cases by constructing one or more linear approximations to the model at hand, and thereby analytically approximating the prediction accuracy. Bates and Watts (1988) provide guidance on whether a linear approximation is justified. For models where such linear approximations are not justified, running repeated simulations may be the only way to make a reasonable assessment of the accuracy.

A second complication arises for nonlinear models that have time-dependency. In such cases, the accuracy of the predictions will likely depend not only on the amount of data collected, but also on the times at which the measurements and predictions are made.

Consequently, instead of minimizing a function of two variables to determine the optimal sampling strategy, the minimization will need to be performed over both the number of data points sampled from each task, and the sampling times for each of these data points.

In addressing these issues, we may reconsider the scenario as a Bayesian optimal design problem, and then make use of results from the optimal design literature. Bayesian optimal design problems are solved based on design criteria denoted by letters of the alphabet, which represent different cost functions to optimize. Of major consideration are Bayesian D-optimality, Bayesian A-optimality and Bayesian c-optimality. Bayesian D-optimality maximizes the expected Kullback-Leibler distance between the prior and posterior distribution (Chaloner & Verdinelli, 1995), and is suggested when inference on the parameters is the main goal of the study. In this thesis, we have considered minimizing the Bayesian MSE of either a subset of the parameters or a subset of the predictions. For these aims, Bayesian A-optimality and Bayesian c-optimality are most useful. Both types of criteria minimize the expected squared loss when estimating a linear function of the parameters. Bayesian A-optimality minimizes this loss for a function of the parameters $\boldsymbol{c'\theta}$, where $\boldsymbol{c}$ is assumed to arise from some probability distribution. Bayesian c-optimality, is a special case of Bayesian A-optimality, where $\boldsymbol{c}$ is assumed fixed. Chaloner & Verdinelli (1995) consider the linear Bayesian model of Section 3.1 with a uniform-variance, diagonal measurement error matrix. With respect to model (**??**), both Bayesian A-optimality and Bayesian c-optimality would minimize $\boldsymbol{c'}(\boldsymbol{C}_\theta^{-1} + \boldsymbol{H'C}_w^{-1}\boldsymbol{H})^{-1}\boldsymbol{c}$.

The simplest manner of deriving an optimal design comes from fixing the number of samples which can be taken, and hence the number of rows in $\boldsymbol{H}$. Chaloner (1984) defines the Fréchet directional derivative at

$$M_0 = \boldsymbol{C}_\theta^{-1} + \boldsymbol{H}_0^{'}\boldsymbol{C}_w^{-1}\boldsymbol{H}_0 \tag{5.4.1}$$

in the direction

$$M_1 = \boldsymbol{C}_\theta^{-1} + \boldsymbol{H}_1^{'}\boldsymbol{C}_w^{-1}\boldsymbol{H}_1 \tag{5.4.2}$$

to be

$$F_\phi\left(M_0, M_1\right) = \lim_{\epsilon \to 0}\left[\phi\left\{(1-\epsilon)M_0 + \epsilon \mathrm{M}_1\right\} - \phi\left\{M_0\right\}\right]. \tag{5.4.3}$$

An equivalence theorem is then derived that gives conditions on the Fréchet directional derivative that are necessary and sufficient for an A-optimal design. These conditions allow us to determine whether a given covariate matrix $\boldsymbol{H}_0$ yields an A-optimal design. Chaloner (1984) shows that for a convex design space $\mathcal{X}$, the c-optimal criterion is reached by designs $\boldsymbol{H}_0^{'}\boldsymbol{H}_0 = \sum_{i=1}^{m} \boldsymbol{h}_i\boldsymbol{h}_i^{'}$ for which $(\boldsymbol{C}_\theta^{-1} + \boldsymbol{H'C}_w^{-1}\boldsymbol{H})^{-1}\boldsymbol{c}$ is normal to a supporting hyperplane of

the convex hull $\mathcal{X}$ at $\boldsymbol{h}_i$ and $-\boldsymbol{h}_i$ for $i = 1, \ldots, m$. Chaloner (1984) further reveals that if $\mathcal{X}$ is convex, then a c-optimal design concentrated at a single point may be found. However, designs with only a few support points are often undesirable, as with such designs it is difficult to check for model lack of fit.

In this thesis, rather than minimizing the total number of samples, our interest has been on minimizing the total cost of experimentation. Chaloner & Verdinelli (1995) note that through a simple linear transformation, the problem of minimal cost can be transformed into one of minimizing the total number of experimental observations.

For the general nonlinear case, Bayesian c-optimality may be obtained by first constructing a linear approximation to the nonlinear model. Then the approximate expected utility can be obtained using the equation

$$\phi_2(\eta) = -\int \boldsymbol{c}(\theta)'\{n\mathcal{J}(\theta,\eta)\}^{-1}\boldsymbol{c}(\theta)p(\theta)d\theta, \tag{5.4.4}$$

where $n\mathcal{J}(\theta,\eta)$ represents the expected Fisher information matrix for a model with unknown parameters θ, design η, and sample size of n. Clyde (1993) suggests that care be taken to construct a design that will result in asymptotic normality of the posterior distribution, since the computation of the expected utility relies on this approximation. When asymptotic normality is in question, Müller and Parmigiani (1996) suggest the alternative of using Markov Chain Monte Carlo methods to estimate the expected utility. Given an apriori set of candidate designs, we may determine which of these designs is optimal by comparing the expected utility of each design.

5.5 Comparing accuracy with other approaches

We now compare the methods in this thesis to other methods such as the Kalman filter, and those detailed in Chandler et al. (2013), which can also be used to include information from secondary variables to make individualized predictions of a primary variable. Individualized predictors may be compared using a given accuracy criterion. To compare them on the Bayesian MSE, we must assume a Bayesian process by which the data are created. We will hitherto refer to this set of assumptions as the *process model*. Individualized predictors are constructed by applying particular population and individual estimation procedures (i.e., MAP, MMSE, maximum likelihood) to a model assumed for the purpose of estimation. We will refer to this as the *estimation model*. Two methods of constructing individual performance predictions may be compared in Bayesian MSE for a particular process model. The

Bayesian MSE will depend on the process model, the estimation model, and the population and individual estimation methods.

We first consider comparing the MMSE for the general linear Bayesian model of Section 3.1 with estimates obtained using the Kalman filter. As it happens, the two methods produce equivalent estimates. We argue this by first noting that the Kalman filter is the sequential MMSE estimator of a signal embedded in noise (Kay, 1994, pg. 419). We then show that the Kalman filter model (1.3.9) can be reformulated as the general linear Bayesian model (3.1.1). The reformulation is done as follows. First we solve the recursive equations,

$$s[0] = As[-1] + Bu[0] \tag{5.5.1}$$

$$\begin{aligned} s[1] &= As[0] + Bu[1] \\ &= A[As[-1] + Bu[0]] + Bu[1] \\ &= A^2 s[-1] + ABu[0] + Bu[1] \end{aligned} \tag{5.5.2}$$

$$\begin{aligned} s[2] &= As[1] + Bu[2] \\ &= A\left[A^2 s[-1] + ABu[0] + Bu[1]\right] + Bu[2] \\ &= A^3 s[-1] + A^2 Bu[0] + ABu[1] + Bu[2] \end{aligned} \tag{5.5.3}$$

to find that

$$\begin{aligned} s[n] &= A^{n+1} s[-1] + \sum_{j=0}^{n} A^{n-j} Bu[j] \\ &= \begin{pmatrix} A^{n+1} & A^n B & A^{n-1} B & \dots \end{pmatrix} \begin{pmatrix} s[-1] \\ u[0] \\ u[1] \\ \vdots \end{pmatrix}. \end{aligned} \tag{5.5.4}$$

Letting

$$s = \begin{pmatrix} s[0] \\ s[1] \\ \vdots \end{pmatrix}, \tag{5.5.5}$$

we can then put this in matrix form as

$$s = \begin{pmatrix} A & B & 0 & \dots \\ A^2 & AB & B & \dots \\ A^3 & A^2 B & AB & \ddots \\ \dots & \dots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} s[-1] \\ u[0] \\ u[1] \\ \vdots \end{pmatrix}. \tag{5.5.6}$$

We have thus formulated s as a linear function of normal random variables. We know, therefore, that the probability distribution for the vector s will be normally distributed. Finally, we let

$$x = \begin{pmatrix} x[0] \\ x[1] \\ \vdots \end{pmatrix}, H = \begin{pmatrix} \boldsymbol{h}'[0] \\ \boldsymbol{h}'[1] \\ \vdots \end{pmatrix}, w = \begin{pmatrix} w[0] \\ w[1] \\ \vdots \end{pmatrix}. \tag{5.5.7}$$

We can then formulate the model as

$$\boldsymbol{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{w}, \tag{5.5.8}$$

where

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}_\theta, \boldsymbol{C}_{\boldsymbol{\theta}}\right) \tag{5.5.9}$$

and

$$\boldsymbol{w} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{C}_w\right), \tag{5.5.10}$$

which is the form of the general linear Bayesian model. We therefore conclude that properties derived about the MMSE of a general linear Bayesian model, such as the Bayesian MSE, will also hold for the Kalman filter.

We next compare the methods of including covariate information in this thesis with that of the Chandler et al. (2013). Unlike the subject-specific models considered in this thesis, the model of Chandler et al. (2013) does not capture individual differences in vulnerability to sleep loss with subject-specific parameters. Instead, individual differences are predicted by including secondary task variables as covariates in the GLM discussed in Section 1.3. A major drawback to this approach is that it lacks the ability to relate information about individual differences between different points in time. Consequently, individualized forecasts at particular times cannot be made in the absence of secondary data at such times, and predictions at times where secondary measurements are collected cannot be enhanced with secondary measurements at other times. By considering specific process and prediction models, we may determine precisely by how much the accuracy of predictions obtained by including a secondary variable in the manner of Chandler et al. (2013) falls short of including a secondary variable using a Bayesian model with assumed parameter correlations.

We consider a bivariate linear Bayesian model of subject means to represent the process model,

$$\begin{aligned} y_{\mathrm{i1k}} &= \phi_{\mathrm{i1}} + \epsilon_{\mathrm{i1k}} \\ y_{\mathrm{i2k}} &= \phi_{\mathrm{i2k}} + \epsilon_{\mathrm{i2k}} \end{aligned} \tag{5.5.11}$$

$$\begin{pmatrix} \phi_{\text{i1}} \\ \phi_{\text{i2}} \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu}, \begin{pmatrix} \delta_1^2 & \rho\delta_1\delta_2 \\ \rho\delta_1\delta_2 & \delta_2^2 \end{pmatrix}\right) \tag{5.5.12}$$

$$\begin{pmatrix} \epsilon_{\text{i1}} \\ \epsilon_{\text{i2}} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\boldsymbol{I} \end{pmatrix}\right). \tag{5.5.13}$$

We compare two methods of estimation via simulation. We first make use of response forecasts that are constructed using the MMSE where the estimation model is the same as the process model. Secondly, we consider an estimation model that uses the Chandler et al. (2013) method of including a secondary variable by modeling the responses on the primary task as a linear function of the primary task population mean and the secondary task covariate,

$$y_{\text{i1k}} = \phi_1 + \beta y_{\text{i2k}} + \epsilon_{\text{i1k}}. \tag{5.5.14}$$

As in this thesis we assume that population parameters are estimated without error, we now assume for this model that the population parameters β and ϕ_1 are also estimated without error (specifically, at their maximum likelihood values in the population estimation stage using an unlimited amount of data.) We construct maximum likelihood estimates of these population parameters from simulated data for 5000 subjects from model (5.5.11). For each individual, we simulate two measurements on the primary task, and two measurements on the secondary task, assuming the fixed parameters $\delta_1 = 1.0, \delta_2 = 1.0, \sigma_1^2 = 0.25, \sigma_2^2 = 1.0$, and $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. To construct individual forecasts, we again simulate 5000 individuals. From this simulation, we use two observations on the secondary task to construct primary task predictions and one observation on the primary task to assess the accuracy of these predictions.

The results showed the same accuracy between the two methods for $\rho = 0$. As ρ was increased to 1, the accuracy of predictions improved for both methods, but more so for the MMSE estimator. For this simulation, the accuracy of the Chandler method corresponded exactly with what the accuracy of the MMSE would have been had only a single data point been collected on the secondary task. These results suggest that for the bivariate linear Bayesian model of subject means, the Chandler method produces the Bayesian MMSE estimator which assumes only a single data point is collected on the secondary task. If this is indeed the case, then the accuracy of predictions using the Chandler method, when

forecasts are made using only observations from the secondary task will be

$$\mathcal{R}\left[\hat{y}_{\mathrm{i}1}\right] = \left(\frac{1}{\delta_1^2} + \frac{\rho^2}{\delta_1^2\left(1-\rho^2\right) + \sigma_2^2\frac{\delta_1^2}{\delta_2^2}}\right)^{-1} + \sigma_1^2, \tag{5.5.15}$$

whereas for the MMSE the accuracy will be

$$\mathcal{R}\left[\hat{y}_{\mathrm{i}1}\right] = \left(\frac{1}{\delta_1^2} + \frac{m_{\mathrm{i}2}\rho^2}{m_{\mathrm{i}2}\delta_1^2\left(1-\rho^2\right) + \sigma_2^2\frac{\delta_1^2}{\delta_2^2}}\right)^{-1} + \sigma_1^2. \tag{5.5.16}$$

In the case where we assume that $\rho = 1$, $\delta_1 = \delta_2$, and δ_1 is large, the prediction accuracy of the Chandler method will be approximately

$$\mathcal{R}\left[\hat{y}_{\mathrm{i}1}\right] \approx \sigma_2^2 + \sigma_1^2, \tag{5.5.17}$$

whereas for the MMSE the accuracy will be

$$\mathcal{R}\left[\hat{y}_{\mathrm{i}1}\right] \approx \frac{\sigma_2^2}{m_{\mathrm{i}2}} + \sigma_1^2. \tag{5.5.18}$$

Therefore, for the bivariate linear Bayesian model of subject means, the Chandler method will have accuracy that is limited by the error variance on both tasks, whereas the MMSE prediction accuracy will be limited by the primary task only.

We have so far considered a comparison of estimators by first assuming that the data are truely generated by the general linear Bayesian model with uncorrelated errors. The advantages/disadvantages of using one estimation model and estimator over another, however, will depend on the underlying process model being considered. Kay (1993) considers a general linear Bayesian model with an arbitrary error covariance matrix. Such a framework allows us to make performance estimates that in addition to parameter correlations, account for autocorrelation and cross correlation in the errors.

Making use of the notation of Kay (1994) detailed in Section 1.3, we now consider how to make a comparison of the accuracy of Bayesian MMSE estimators using different error correlation structures. Let us represent the error covariance structure of the process model with $\boldsymbol{C}_w$ and the reduced covariance structure with Λ. We consider the mean squared error matrix $\boldsymbol{M}_{\hat{\theta}}$ defined as

$$\boldsymbol{M}_{\hat{\theta}} \equiv E_{x,\theta}\left[\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)'\right], \tag{5.5.19}$$

where $\hat{\boldsymbol{\theta}}$ represents the MMSE estimator made with the assumption of a reduced correlation

structure. In Appendix E, we derive the mean squared error matrix to be

$$\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}} + \boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}' \left(\left(\mathbf{H}\mathbf{C}_{\theta}H' + \boldsymbol{C}_w\right)^{-1} - 2\left(\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\boldsymbol{H}' + \boldsymbol{\Lambda}\right)^{-1} + \left(\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\boldsymbol{H}' + \boldsymbol{\Lambda}\right)^{-1}\left(\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\boldsymbol{H}' + \boldsymbol{C}_{\boldsymbol{w}}\right)\left(\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\boldsymbol{H}' + \boldsymbol{\Lambda}\right)^{-1} \right) \mathbf{H}\mathbf{C}'_{\theta}, \tag{5.5.20}$$

where $\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}}$ represents the minimum mean squared error matrix for the MMSE estimator assuming the same estimation and process model with error covariance $\boldsymbol{C}_w$, and is given in (1.3.4). The Bayesian MSE for each parameter can then be obtained from the diagonal elements of this matrix. The term on the right hand side represents the additional MSE due to assuming a reduced covariance structure for the errors. We note that when $\boldsymbol{\Lambda} = \boldsymbol{C}_w$ the term on the right dissappears. For particular, patterned correlation structures, it may be possible to determine a scalar closed form equation for how the Bayesian MSE for a given parameter estimate depends on the parameters of the error covariance matrix.

Appendix A

Direct sum results

Theorem A.0.1. *The matrix product of two direct sums*

$$\left(\bigoplus_{i=1}^{n} A_i\right)\left(\bigoplus_{i=1}^{n} B_i\right), \tag{A.0.1}$$

where $C[A_i] = R[B_i]$, and where $C[A_i]$ represents the number of columns in A_i and $R[B_i]$ represents the number of rows in B_i, is

$$\bigoplus_{i=1}^{n} A_i B_i. \tag{A.0.2}$$

Proof.

$$\begin{aligned}\left(\bigoplus_{i=1}^{n} A_i\right)\left(\bigoplus_{i=1}^{n} B_i\right) &= \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & A_3 & \\ & & & \ddots \end{bmatrix}\begin{bmatrix} B_1 & & & \\ & B_2 & & \\ & & B_3 & \\ & & & \ddots \end{bmatrix} \\ &= \begin{bmatrix} A_1B_1 & & & \\ & A_2B_2 & & \\ & & A_3B_3 & \\ & & & \ddots \end{bmatrix} \\ &= \bigoplus_{i=1}^{n} A_i B_i.\end{aligned} \tag{A.0.3}$$

□

Theorem A.0.2. *The transpose of $\bigoplus_{i=1}^{n} A_i$ is equal to the direct sum of the transposes of the A_is.*

Proof.

$$\left(\bigoplus_{i=1}^{n} A_i\right)' = \begin{bmatrix} A_1 & & \\ & A_2 & \\ & & \ddots \end{bmatrix}' = \begin{bmatrix} A_1' & & \\ & A_2' & \\ & & \ddots \end{bmatrix} \tag{A.0.4}$$

$$\bigoplus_{i=1}^{n} A_i'. \tag{A.0.5}$$

□

Theorem A.0.3. *The inverse of $\bigoplus_{i=1}^{n} A_i$ is equal to the direct sum of the inverses of each of the A_i's.*

Proof.

$$\begin{aligned} \left(\bigoplus_{i=1}^{n} A_i\right)\left(\bigoplus_{i=1}^{n} A_i^{-1}\right) &= \bigoplus_{i=1}^{n} A_i A_i^{-1} \\ &= \bigoplus_{i=1}^{n} I_{n_i}, \end{aligned} \tag{A.0.6}$$

where $n_i = R[A_i] = I_{\sum n_i}$. □

Appendix B

Multivariate results

Let X, Y be continuous random variables. The conditional probability density function (pdf) for Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{B.0.1}$$

and the marginal distribution of Y is defined as

$$f_Y(y) = \int_x f_{X,Y}(x,y)dx. \tag{B.0.2}$$

These definitions also have conditional versions. Let Z be another continuous random variable, then the conditional probability density function (pdf) for Y given $X = x$ conditioning on $Z = z$ is defined as

$$f_{Y|X,Z}(y|x,z) = \frac{f_{X,Y|Z}(x,y|z)}{f_{X|Z}(x|z)}, \tag{B.0.3}$$

and the marginal distribution of Y conditioning on $Z = z$ is defined as

$$f_{Y|Z}(y|z) = \int_x f_{X,Y|Z}(x,y|z)dx. \tag{B.0.4}$$

Let $f_{Y|X}(y|x)$ be the posterior probability density function of the random variable $Y|X$ on the support $-\infty < x < \infty$. The Maximum a Posteriori estimate is defined as

$$\underset{y}{\text{argmax}} \left[f_{Y|X}(y|x) \right]. \tag{B.0.5}$$

Theorem B.0.4. (Multivariate marginal theorem) *Let* $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, and let* $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, *then* $\boldsymbol{y}_1 \sim N(\mu_{\mathbf{1}}, \boldsymbol{\Sigma_{11}})$.

Proof. Without loss of generality, let $\boldsymbol{y}$ be partitioned as $\boldsymbol{y}' = (\boldsymbol{y}_1', \boldsymbol{y}_2')$, where $\boldsymbol{y}_1$ is the $r \times 1$ subvector of interest. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be partitioned accordingly:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \tag{B.0.6}$$

Define $\boldsymbol{A} = (\boldsymbol{I}_r, \boldsymbol{O})$, where $\boldsymbol{I}_r$ is an $r \times r$ identity matrix and $\boldsymbol{O}$ is an $r \times (p-r)$ matrix of zeros. Then $A\boldsymbol{y} = \boldsymbol{y}_1$, and by Theorem 4.41 (ii), $\boldsymbol{y}_1$ is distributed as $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$. □

Theorem B.0.5. *Let*

$$g\left(\boldsymbol{y}|\boldsymbol{\mu}\right) = c_0 \cdot \exp\left[-\frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} - \boldsymbol{B}y\right)\right]. \tag{B.0.7}$$

Then $\boldsymbol{y}|\boldsymbol{\mu} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$*, where*

$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}, \tag{B.0.8}$$

$$\boldsymbol{\mu} = \frac{1}{2}\Sigma\boldsymbol{B}'. \tag{B.0.9}$$

Proof. Let $\boldsymbol{y}$ be multivariate normal with pdf given by

$$f\left(\boldsymbol{y}|\boldsymbol{\mu}\right) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right]}{(2\pi)^n|\boldsymbol{\Sigma}|^{\frac{1}{2}}}. \tag{B.0.10}$$

Expanding, we find that

$$= c_0 \cdot \exp\left[-\frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - 2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]. \tag{B.0.11}$$

Next we let

$$c_0 = \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right] \tag{B.0.12}$$

be constant with respect to $\boldsymbol{y}$. Then

$$f\left(\boldsymbol{y}|\boldsymbol{\mu}\right) = c_0 \cdot \exp\left[-\frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - 2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right)\right]. \tag{B.0.13}$$

Making substitutions (B.0.8) and (B.0.9), we arrive at (B.0.7) □

Theorem B.0.6. *Let* $\boldsymbol{Y}_{n\times 1} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. Let* $\boldsymbol{A}$ *be a* $p \times n$ *matrix of constants, and let* $\boldsymbol{b}$ *be a* $p \times 1$ *vector of constants. Then the* $p \times 1$ *random vector* $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$.

Appendix C

Bivariate model of subject means results

Theorem C.0.7. *The posterior precision is*

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \lambda\left(m_{i2}\right). \tag{C.0.1}$$

Proof. Using the variance (3.3.7) and the definition of precision from (3.3.10), we know that

$$\eta = \frac{\left(\frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2(1-\rho^2)}\right)\left(\frac{m_{i2}}{\sigma_2^2} + \frac{1}{\delta_2^2(1-\rho^2)}\right) - \frac{\rho^2}{\delta_1^2\delta_2^2(1-\rho^2)^2}}{\frac{m_{i2}}{\sigma_2^2} + \frac{1}{\delta_2^2(1-\rho^2)}}. \tag{C.0.2}$$

Separating the fraction and canceling terms yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2\left(1-\rho^2\right)} - \frac{\rho^2}{\delta_1^2\delta_2^2\left(1-\rho^2\right)^2}\left(\frac{m_{i2}}{\sigma_2^2} + \frac{1}{\delta_2^2\left(1-\rho^2\right)}\right)^{-1}. \tag{C.0.3}$$

Combining fractions yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2\left(1-\rho^2\right)} - \frac{\rho^2}{\delta_1^2\delta_2^2\left(1-\rho^2\right)^2}\left(\frac{m_{i2}\delta_2^2\left(1-\rho^2\right) + \sigma_2^2}{\sigma_2^2\delta_2^2\left(1-\rho^2\right)}\right)^{-1}. \tag{C.0.4}$$

Simplifying yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2\left(1-\rho^2\right)} - \frac{\rho^2\sigma_2^2\delta_2^2\left(1-\rho^2\right)}{\delta_1^2\delta_2^2\left(1-\rho^2\right)^2\left(m_{i2}\delta_2^2\left(1-\rho^2\right) + \sigma_2^2\right)}. \tag{C.0.5}$$

Canceling terms yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2\left(1-\rho^2\right)} - \frac{\rho^2\sigma_2^2}{\delta_1^2\left(1-\rho^2\right)\left(m_{i2}\delta_2^2\left(1-\rho^2\right) + \sigma_2^2\right)}. \tag{C.0.6}$$

Applying a partial fraction expansion yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)} - \frac{\rho^2\sigma_2^2}{\delta_1^2\left(1-\rho^2\right)\left(m_{i2}\delta_2^2\left(1-\rho^2\right)+\sigma_2^2\right)}. \tag{C.0.7}$$

Factoring yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{\rho^2}{1-\rho^2}\left(\frac{1}{\delta_1^2} - \frac{\sigma_2^2}{\delta_1^2\left(m_{i2}\delta_2^2\left(1-\rho^2\right)+\sigma_2^2\right)}\right). \tag{C.0.8}$$

Combining fractions yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{\rho^2}{1-\rho^2}\left(\frac{\delta_1^2\left(m_{i2}\delta_2^2\left(1-\rho^2\right)+\sigma_2^2\right) - \delta_1^2\sigma_2^2}{\delta_1^2\delta_1^2\left(m_{i2}\delta_2^2\left(1-\rho^2\right)+\sigma_2^2\right)}\right). \tag{C.0.9}$$

Expanding yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{\rho^2}{1-\rho^2}\left(\frac{\delta_1^2 m_{i2}\delta_2^2\left(1-\rho^2\right)+\delta_1^2\sigma_2^2 - \delta_1^2\sigma_2^2}{\delta_1^2 m_{i2}\delta_2^2\left(1-\rho^2\right)+\delta_1^2\sigma_2^2}\right). \tag{C.0.10}$$

Canceling terms yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{\rho^2 m_{i2}\delta_2^2}{\delta_1^2 m_{i2}\delta_2^2\left(1-\rho^2\right)+\delta_1^2\sigma_2^2}. \tag{C.0.11}$$

Dividing numerator and denominator by $\delta_1^2\delta_2^2\left(1-\rho^2\right)$ yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{m_{i2}\frac{\rho^2\delta_2^2}{\delta_1^2\delta_2^2\left(1-\rho^2\right)}}{m_{i2}+\frac{\delta_1^2\sigma_2^2}{\delta_1^2\delta_2^2\left(1-\rho^2\right)}}. \tag{C.0.12}$$

Canceling terms yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \frac{m_{i2}\frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)}}{m_{i2}+\frac{\sigma_2^2}{\delta_2^2\left(1-\rho^2\right)}}. \tag{C.0.13}$$

And finally, substituting

$$\lambda\left(m_{i2}\right) = \frac{m_{i2}\lambda_{\max}^m}{m_{i2}+m_h}, \tag{C.0.14}$$

where

$$\lambda_{\max}^m = \frac{\rho^2}{\delta_1^2\left(1-\rho^2\right)}, \tag{C.0.15}$$

and

$$m_h = \frac{\sigma_2^2}{\delta_2^2\left(1-\rho^2\right)}, \tag{C.0.16}$$

yields

$$\eta = \frac{m_{i1}}{\sigma_1^2} + \frac{1}{\delta_1^2} + \lambda\left(m_{i2}\right). \tag{C.0.17}$$

□

Theorem C.0.8. *The Fisher Information on the mean is* $\frac{m_i}{\sigma^2}$.

Proof. The Fisher Information is given by

$$\mathcal{I}(\boldsymbol{b}_i) = -E_{\boldsymbol{y}_i|\boldsymbol{b}_i}\left[\frac{\partial^2}{\partial \boldsymbol{b}_i^2}\log f(\boldsymbol{y}_i|\boldsymbol{b}_i)\right]. \tag{C.0.18}$$

When $\boldsymbol{y}_i|\boldsymbol{b}_i$ is normally distributed with zero mean and variance σ^2,

$$f(\boldsymbol{y}_i|\boldsymbol{b}_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m_i}\exp\left[\frac{\sum_{j=1}^{m_i}(\boldsymbol{y}_{ij}-\boldsymbol{b}_i)^2}{2\sigma^2}\right]. \tag{C.0.19}$$

$$\mathcal{I}(\boldsymbol{b}_i) = -E\left[\frac{\partial^2}{\partial \boldsymbol{b}_i^2}\left[\log\left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m_i}\right] - \frac{\sum_{j=1}^{m_i}(y_{ij}-b_i)^2}{2\sigma^2}\right]\right] \tag{C.0.20}$$

$$= -E\left[\frac{\partial}{\partial \boldsymbol{b}_i}\left(\frac{\sum_{j=1}^{m_i}\boldsymbol{y}_{ij}-\boldsymbol{b}_i}{\sigma^2}\right)\right] \tag{C.0.21}$$

$$= -E\left[\frac{-m_i}{\sigma^2}\right] \tag{C.0.22}$$

$$= \frac{m_i}{\sigma^2}. \tag{C.0.23}$$

□

Appendix D

Nonlinear results

$Failed

Theorem D.0.9.

$$\int_{-\infty}^{\infty} e^{-(ax^2+bx+c)}dx = \sqrt{\frac{\pi}{a}}e^{\left(\frac{b}{2\sqrt{a}}\right)^2-c}$$

Proof. We first prove that

$$\int_{-\infty}^{\infty} e^{-(x^2+fx+g)}dx = \sqrt{\pi}e^{\left(\frac{f}{2}\right)^2-g}$$

To do this, we complete the square, separate the exponents, and factor constants out of the integral.

$$= \int_{-\infty}^{\infty} e^{-\left(\left(x+\frac{f}{2}\right)^2+g-\left(\frac{f}{2}\right)^2\right)}dx$$

$$= e^{\left(\frac{f}{2}\right)^2-g}\int_{-\infty}^{\infty} e^{-\left(x+\frac{f}{2}\right)^2}dx$$

Making the substitution $y = x + \frac{f}{2}$, we find

$$= e^{\left(\frac{f}{2}\right)^2-g}\int_{-\infty}^{\infty} e^{-y^2}dy$$

$$= \sqrt{\pi}e^{\left(\frac{f}{2}\right)^2-g}$$

Let $z = \sqrt{a}x$, and substituting in for x we find

$$\int_{-\infty}^{\infty} e^{-(ax^2+bx+c)}dx = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} e^{-\left(z^2+\frac{b}{\sqrt{a}}z+c\right)}dz$$

Finally, using the results from the first part of this proof we find that

$$\int_{-\infty}^{\infty} e^{-(ax^2+bx+c)}dx = \sqrt{\frac{\pi}{a}}e^{\left(\frac{b}{2\sqrt{a}}\right)^2-c}.$$

□

Appendix E

Comparing estimators

In this section, we derive the MSE matrix for the MMSE derived from assuming a general linear Bayesian model with correlation structure $\mathbf{\Lambda}$, assuming data arises from a general linear Bayesian model with correlation structure $\boldsymbol{C}_w$. The MMSE estimator using the correlation structure $\mathbf{\Lambda}$ is

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu}_\theta + \boldsymbol{C}_\theta \boldsymbol{H}'(\mathbf{H}\mathbf{C}_\theta \boldsymbol{H}' + \mathbf{\Lambda})^{-1} (\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu}_\theta), \tag{E.0.1}$$

and the covariance of the estimation error will be

$$\boldsymbol{M}_{\hat{\theta}} = E_{x,\theta}\left[\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)'\right]. \tag{E.0.2}$$

Using iterated expectations, we find that

$$\boldsymbol{M}_{\hat{\theta}} = E_x\left[E_{\theta|x}\left[\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)'\right]\right]. \tag{E.0.3}$$

Adding and subtracting $E(\boldsymbol{\theta}|\boldsymbol{x})$ and expanding terms results in

$$\begin{aligned}
\boldsymbol{M}_{\hat{\theta}} &= E_x\left[E_{\theta|x}\left[\left(\boldsymbol{\theta} - \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) + \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) - \hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) + \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) - \hat{\boldsymbol{\theta}}\right)'\right]\right] \\
&= E_x\left[\begin{array}{l} E_{\theta|x}[(\boldsymbol{\theta} - \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}))(\boldsymbol{\theta} - \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}))'] + E_{\theta|x}\left[\left(\boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) - \hat{\boldsymbol{\theta}}\right)(\boldsymbol{\theta} - \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}))'\right] \\ + E_{\theta|x}\left[(\boldsymbol{\theta} - \boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}))\left(\boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) - \hat{\boldsymbol{\theta}}\right)'\right] + E_{\theta|x}\left[\left(\boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) - \hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{E}(\boldsymbol{\theta}|\boldsymbol{x}) - \hat{\boldsymbol{\theta}}\right)'\right] \end{array}\right].
\end{aligned} \tag{E.0.4}$$

Applying the expectations, we find that the term $\boldsymbol{E(\theta|x)-\hat{\theta}}$ does not depend on $\boldsymbol{\theta}$ when $\boldsymbol{x}$ is given, and the expected value of $\boldsymbol{\theta - E(\theta|x)}$ is just $\mathbf{0}$. Therefore,

$$\boldsymbol{M}_{\hat{\theta}} = E_x\left[E_{\theta|x}[(\boldsymbol{\theta}-\boldsymbol{E(\theta|x)})(\boldsymbol{\theta}-\boldsymbol{E(\theta|x)})']+\mathbf{0}+\mathbf{0}+E_{\theta|x}\left[\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)'\right]\right]. \tag{E.0.5}$$

Noting that the first term is simply the posterior covariance matrix, we have that

$$\begin{aligned}\boldsymbol{M}_{\hat{\theta}} &= E_x\left[\boldsymbol{C}_{\theta|x}+E_{\theta|x}\left[\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)'\right]\right]\\ &= \boldsymbol{M}_{E[\theta|x]}+E_x\left[E_{\theta|x}\left[\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)'\right]\right].\end{aligned} \tag{E.0.6}$$

Since $\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)'$ does not depend on $\boldsymbol{\theta}$ when $\boldsymbol{x}$ is given, we find that

$$\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon}+E_x\left[\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{E(\theta|x)}-\hat{\boldsymbol{\theta}}\right)'\right]. \tag{E.0.7}$$

Substituting in for $\boldsymbol{E(\theta|x)}$ and $\hat{\boldsymbol{\theta}}$, we find that

$$\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon}+E_x\left[\begin{array}{c}\left(\begin{array}{l}\boldsymbol{\mu_\theta+C_\theta H'(HC_\theta H'+C_w)^{-1}(x-H\mu_\theta)}\\ -\boldsymbol{\mu}_\theta-\boldsymbol{C}_\theta\boldsymbol{H'}(\mathbf{H}\boldsymbol{C}_\theta\boldsymbol{H'}+\boldsymbol{\Lambda})^{-1}(\boldsymbol{x}-\mathbf{H}\boldsymbol{\mu}_\theta)\end{array}\right)\\ \left(\begin{array}{l}\boldsymbol{\mu_\theta+C_\theta H'(HC_\theta H'+C_w)^{-1}(x-H\mu_\theta)}\\ -\boldsymbol{\mu}_\theta-\boldsymbol{C}_\theta\boldsymbol{H'}(\mathbf{H}\boldsymbol{C}_\theta\boldsymbol{H'}+\boldsymbol{\Lambda})^{-1}(\boldsymbol{x}-\mathbf{H}\boldsymbol{\mu}_\theta)\end{array}\right)'\end{array}\right]. \tag{E.0.8}$$

Canceling the $\boldsymbol{\mu}_\theta$ terms yields

$$\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon}+E_x\left[\begin{array}{l}\boldsymbol{(C_\theta H'(HC_\theta H'+C_w)^{-1}(x-H\mu_\theta)}-\boldsymbol{C}_\theta\boldsymbol{H'}(\mathbf{H}\boldsymbol{C}_\theta\boldsymbol{H'}+\boldsymbol{\Lambda})^{-1}(\boldsymbol{x}-\mathbf{H}\boldsymbol{\mu}_\theta))\\ \boldsymbol{(C_\theta H'(HC_\theta H'+C_w)^{-1}(x-H\mu_\theta)}-\boldsymbol{C}_\theta\boldsymbol{H'}(\mathbf{H}\boldsymbol{C}_\theta\boldsymbol{H'}+\boldsymbol{\Lambda})^{-1}(\boldsymbol{x}-\mathbf{H}\boldsymbol{\mu}_\theta))'\end{array}\right]. \tag{E.0.9}$$

Expanding terms and separating expectations results in

$$\begin{aligned}\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon} &+ E_x\left[\boldsymbol{C}_\theta\boldsymbol{H'}(\mathbf{H}\boldsymbol{C}_\theta\boldsymbol{H'}+\boldsymbol{C_w})^{-1}(\boldsymbol{x}-\mathbf{H}\boldsymbol{\mu_\theta})\boldsymbol{(C_\theta H'(HC_\theta H'+C_w)^{-1}(x-H\mu_\theta))'}\right]\\ &- E_x\left[\boldsymbol{C_\theta H'(HC_\theta H'+C_w)^{-1}(x-H\mu_\theta)(C_\theta H'(HC_\theta H'+\Lambda)^{-1}(x-H\mu_\theta))'}\right]\\ &- E_x\left[\boldsymbol{C_\theta H'(HC_\theta H'+\Lambda)^{-1}(x-H\mu_\theta)(C_\theta H'(HC_\theta H'+C_w)^{-1}(x-H\mu_\theta))'}\right]\\ &+ E_x\left[\boldsymbol{C_\theta H'(HC_\theta H'+\Lambda)^{-1}(x-H\mu_\theta)(C_\theta H'(HC_\theta H'+\Lambda)^{-1}(x-H\mu_\theta))'}\right].\end{aligned} \tag{E.0.10}$$

Applying the transpose, we get

$$\begin{aligned}\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon} + E_x\Big[\boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}\mathbf{HC}'_\theta\Big] \\ - E_x\Big[\boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'(\mathbf{HC_\theta H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta\Big] \\ - E_x\Big[\boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{\Lambda})^{-1}(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}\mathbf{HC}'_\theta\Big] \\ + E_x\Big[\boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{\Lambda})^{-1}(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'(\mathbf{HC_\theta H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta\Big]\end{aligned} \tag{E.0.11}$$

Removing terms from the expectation yields

$$\begin{aligned}\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon} + \boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}E_x\Big[(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'\Big](\mathbf{HC_\theta}H' + \boldsymbol{C_w})^{-1}\mathbf{HC}'_\theta \\ - \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)^{-1}E_x\Big[(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'\Big](\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta \\ - \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}E_x\Big[(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'\Big](\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)^{-1}\mathbf{HC}'_\theta \\ + \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}E_x\Big[(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})(\boldsymbol{x} - \mathbf{H}\boldsymbol{\mu_\theta})'\Big](\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta.\end{aligned} \tag{E.0.12}$$

Evaluating the expectation results in

$$\begin{aligned}\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon} + \boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}\mathbf{HC}'_\theta \\ - \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)^{-1}(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta \\ - \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)^{-1}\mathbf{HC}'_\theta \\ + \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta.\end{aligned} \tag{E.0.13}$$

Canceling terms with their inverses yields

$$\begin{aligned}\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon} + \boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}\mathbf{HC}'_\theta \\ - \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta \\ - \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta \\ + \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta.\end{aligned} \tag{E.0.14}$$

Combining like terms, we find that

$$\begin{aligned}\boldsymbol{M}_{\hat{\theta}} = \boldsymbol{C}_{\epsilon} + \boldsymbol{C_\theta H'}(\mathbf{HC_\theta H'} + \boldsymbol{C_w})^{-1}\mathbf{HC}'_\theta \\ - 2\boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta \\ + \boldsymbol{C}_\theta H'(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{C}_w)(\mathbf{HC}_\theta\boldsymbol{H'} + \boldsymbol{\Lambda})^{-1}\mathbf{HC}'_\theta\end{aligned} \tag{E.0.15}$$

and finally that

$$\begin{aligned} M_{\hat{\theta}} = C_{\epsilon} + \boldsymbol{C_{\theta} H'}((\mathbf{H C_{\theta} H' + C_w})^{\mathbf{-1}} - 2\left(\mathbf{H C_{\theta} H' + \Lambda}\right)^{\mathbf{-1}} \\ + \left(\mathbf{H C_{\theta} H' + \Lambda}\right)^{-1}\left(\mathbf{H C_{\theta} H' + C_w}\right)\left(\mathbf{H C_{\theta} H' + \Lambda}\right)^{\mathbf{-1}})\mathbf{H C}'_{\theta}. \end{aligned} \tag{E.0.16}$$

Acknowledgements

I like to take this opportunity to thank several individuals who helped me to complete my Ph.D. program and this thesis. Firstly I would like to thank Dr. Hans Van Dongen for the amount of time which he spent helping to discuss this dissertation and the positive feedback that he gave to keep me motivated throughout the process. Secondly I would like to thank my primary advisor Dr. Leonid Kalachev for his encouragement, support and direction throughout the process.

Borbély, AA. A two process model of sleep regulation. Human *Neurobiology*, 1982;1:195-204.

Bates DM, Watts DG. Nonlinear Regression Analysis and Its Applications. New York: Wiley, 1988.

Bonate PL. Pharmacokinetic-Pharmacodynamic Modeling and Simulation. New York: Springer, 2006.

Chaloner K. Optimal Bayesian Experimental Design for Linear Models. The Annals of Statistics, 1984;12:283-300.

Chaloner K, Verdinelli I. Bayesian Experimental Design: A Review. *Statistical Science* 1995;10(3):273-304.

Clyde, MA. Bayesian optimal designs for approximate normality. Ph.D. dissertation, Univ. Minnesota 1993.

Davidian M, Giltinan D. Nonlinear Models for Repeated Measurement Data. New York: Chapman & Hall, 1995.

Forsman PM, Vila BJ, Short RA, Mott CG, Van Dongen HPA. Efficient driver drowsiness detection at moderate levels of drowsiness. *Accid. Anal. Prev.* 2012;50:341-50.

Hall DB, Clutter M. Multivariate multilevel nonlinear mixed effects models for timber yield predictions. *Biometrics*, 2004;60:16-24.

Kay S. Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory. Upper Saddle River: Prentice Hall, 1993.

Mallis MM, Mejdal S, Nguyen TT, Dinges DF. Summary of the key features of seven biomathematical models of human fatigue and performance. *Aviat Space Environ Med* 2004; 75(3,Suppl.): A4-14

McCauley P, Kalachev LV, Smith AD, Belenky G, Dinges DF, Van Dongen HPA. A new mathematical model for the homeostatic effects of sleep loss on neurobehavioral perfor-

mance. *J Theor Biol.* 2009;256:227-39.

McCauley P, Kalachev LV, Mollicone DJ, Banks S, Dinges DF, Van Dongen HPA. Dynamic circadian modulation in a biomathematical model for the effects of sleep and sleep loss on waking neurobehavioral performance. *Sleep* 2013 Dec;36(12)1987-97.

Müller P, Parmigiani G. Numerical evaluation of information theoretic measures. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* New York: Wiley, 1996.

Pinheiro J, Bates D. Mixed-Effects Models in S and S-PLUS. New York: Springer, 2009.

Van Dongen HPA, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep* 2003;26(2),117-126.

Van Dongen HPA, Dinges DF. Sleep, circadian rhythms, and psychomotor vigilance. *Clinics in Sports Medicine* 2005;24(2):237-49.

Van Dongen HPA, Mott CG, Huang JK, Mollicone DJ, McKenzie FD, and Dinges DF. 2007. Optimization of biomathematical model predictions for cognitive performance impairment in individuals: accounting for unknown traits and uncertain states in homeostatic and circadian processes. *Sleep.* Vol. 30, pp. 1129-43.