# Bayesian Model Uncertainty and Foundations

by

Víctor Peña

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
James O. Berger, Supervisor

_____
Merlise A. Clyde

_____
Fan Li

_____
Gonzalo García-Donato

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2018

# Abstract

## Bayesian Model Uncertainty and Foundations

by

Víctor Peña

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
James O. Berger, Supervisor

_____
Merlise A. Clyde

_____
Fan Li

_____
Gonzalo García-Donato

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2018

# Abstract

This dissertation contains research on Bayesian model uncertainty and foundations of statistical inference.

In Chapter 2, we study the properties of constrained empirical Bayes (EB) priors on regression coefficients. Unrestricted EB procedures can have undesirable properties when their "estimates" correspond to hyperparameters that would be seen as overly informative in an actual Bayesian analysis. For that reason, we propose constraining EB procedures so that they are at least as vague as proper Bayesian lower bounds (which can be either informative or "noninformative"). The main emphasis of the chapter is studying the properties of a constrained EB prior that has Zellner's $g-$prior with $g = n$ as its lower bound. We show that it avoids some of the pitfalls of unconstrained EB priors and the lower bound, and see that it behaves similarly to the Bayesian Information Criterion (BIC).

In Chapter 3, we take a close look at "information inconsistency." Information inconsistency is said to occur when there is overwhelming evidence in favor of a hypothesis in finite sample sizes, but the Bayes factor in its favor is finite. In Chapter 3, we investigate when it occurs (and when it does not) in normal linear models. Our conclusion is that conjugate priors are usually information-inconsistent, but thick-tailed priors and empirical Bayes procedures avoid the issue. The chapter also includes a discussion of the different formalizations of information inconsistency that have appeared in the literature, which are not equivalent.

In Chapter 4, we turn to "limit consistency," which is an asymptotic property of two-sample tests. Suppose the sample size of one of the groups goes to infinity while the sample size of the other one stays fixed. According to our definition, limit consistency occurs if, under this asymptotic regime, the decision rule of the two-sample test converges to the decision rule of the one-sample test we would have performed had we known the parameters of the group with "infinite" data. In Chapter 4, we study limit consistency in the context of comparing whether two normal means are equal. We conclude that parametrizations where the 2 groups have common parameters are generally limit-consistent when the prior on the common parameters is flat.

Finally, the goal of Chapter 5 is discussing 2 articles that cast doubt on the correctness and applicability of Birnbaum's theorem, which implies that statisticians that wish to respect the sufficiency and conditionality principle must accept the likelihood principle. This result, which was proved in 1962, is still highly controversial because some statisticians believe that sufficiency and conditionality are appealing, but the likelihood principle is not (for example, the likelihood principle precludes the use of $p$-values, which are highly popular in common statistical practice). In Chapter 5, we provide counterarguments to the criticisms and put them in historical context.

# Contents

# List of Tables

# List of Figures

# 1

# Introduction

In this dissertation, we present research on Bayesian model uncertainty and foundations of statistical inference. The goal of this chapter is introduce and motivate our work.

Chapters 2, 3, and 4 are concerned with prior choice under model uncertainty. Consider the hypothesis test $H_0 : y \mid \alpha, \sigma^2 \sim N_n(1_n\alpha, \sigma^2 I_n)$ against $H_1 : y \mid \alpha, \beta, \sigma^2 \sim N_n(1_n\alpha + X\beta, \sigma^2 I_n)$, where $y \in \mathbb{R}^n$ is the outcome, $1_n = (1, 1, \ldots, 1)' \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ is the design matrix and $\beta \in \mathbb{R}^p$ is a vector of regression coefficients. The posterior probabilities of $H_0$ and $H_1$ depend on the data only through the so-called Bayes factor, which is a ratio of integrated likelihoods. In our example, the Bayes factor of $H_1$ to $H_0$ is

$$B_{10} = \frac{\int N_n(y \mid 1_n\alpha + X\beta, \sigma^2 I)\pi_1(\alpha, \beta, \sigma^2 \mid H_1)\,\mathrm{d}(\alpha, \beta, \sigma^2)}{\int N_n(y \mid 1_n\alpha, \sigma^2 I)\pi_0(\alpha, \sigma^2 \mid H_0)\,\mathrm{d}(\alpha, \sigma^2)},$$

where $\pi_0(\alpha, \sigma^2 \mid H_0)$ and $\pi_1(\alpha, \beta, \sigma^2 \mid H_1)$ are prior densities under $H_0$ and $H_1$, respectively. Our work in Chapters 2, 3, and 4 can be summarized in this context as follows: Chapter 2 studies the properties of (empirical Bayes) procedures which estimate the prior scale of $\beta \mid H_1$ subject to linear matrix constraints; Chapter 3

studies the behavior of $B_{10}$ in a limiting case where $\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2 \to \infty$; and Chapter 4 compares Bayes decisions in two-sample problems when the sample size of one of the groups goes to infinity.

Chapter 5 is about Birnbaum's theorem, which connects 3 statistical principles: sufficiency, conditionality, and the likelihood principle. A classical result, proved in Birnbaum (1962), shows that a statistician that is willing to accept the principles of sufficiency and conditionality must accept the likelihood principle. The correctness of the proof, its implications, and its meaning are debated to this day. As we explain in Section 1.4, these principles have strong implications in hypothesis testing, so Chapter 5 is arguably not completely disconnected to the other chapters in the dissertation. The main goal of Chapter 5 is discussing the objections raised in 2 articles: Evans (2013) and Mayo (2014).

Chapter 6 is the last chapter of the dissertation, and it has conclusions and brief descriptions of current and future research. The next subsections provide more detailed descriptions of our work.

## 1.1 Constrained empirical Bayes priors on regression coefficients

In the example we presented at the beginning of the chapter, the choice of the prior scale of $\beta \mid H_1$ can strongly affect Bayes decisions. If it is arbitrarily vague, the posterior probability of $H_0$ can be arbitrarily large (Jeffreys-Lindley paradox); on the other hand, if we "let the data speak for themselves" and estimate the prior scale from the data, the procedure can be overly favorable to $H_1$.

To give a concrete example, suppose that $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$ and that the prior on $\alpha, \sigma^2$ is the improper prior $\pi(\alpha, \sigma^2) \propto 1/\sigma^2$ under both $H_0$ and $H_1$. Suppose further that $\beta \mid \sigma^2, H_1 \sim N_n(b_0, \sigma^2 g A)$, where $A$ is a given positive definite matrix. If $g \to \infty$, the posterior probability of $H_0$ goes to 1. However, if we set $g$ by maximizing the marginal likelihood under $H_1$, the posterior probability of the alternative hypothesis

is always greater than 1/2.

In Chapter 2, we study empirical Bayes priors for $\beta \mid H_1$ which have restrictions that ensure that the procedures are "at least as vague" as a proper default Bayesian prior. The main emphasis of the chapter is describing the properties of the empirical Bayes prior $\beta \mid \sigma^2, H_1 \sim N_p(0_p, \sigma^2 \widehat{W})$, where $\widehat{W}$ maximizes the marginal likelihood of $H_1$ subject to the linear matrix constraint $\widehat{W} \succ n(X'X)^{-1}$. The choice of lower bound can be justified as follows: the Fisher information of $\beta$ is $(X'X)/\sigma^2$, so the lower bound contains roughly as much information about $\beta$ as a typical observation in the sample. Then, the linear matrix constraint ensures that the empirical Bayes procedure yields a prior that is "at least as vague" as the lower bound. For example, the constraint implies that the highest posterior density regions of the empirical Bayes prior have higher volume than those of the lower bound. The procedure behaves similarly to the Bayesian Information Criterion (Schwarz, 1978). The chapter also studies an analogue for generalized linear models and examples discussed in Berger et al. (2003) and Shibata (1983).

## 1.2 On information consistency

The term "information consistency" was coined in Liang et al. (2008) to describe a situation where there is overwhelming evidence in favor of a hypothesis in finite sample sizes. This hypothetical scenario has been studied in the Bayesian hypothesis testing literature since at least Jeffreys (1939), but it has been formalized in different ways. In Chapter 3, we favor the definition proposed in Som et al. (2016), where $X, \alpha$, and $\epsilon$ are fixed and $\|\beta\| \to \infty$. Under this limit, the true signal becomes arbitrarily large, the error stays constant, the likelihood ratio in favor of $H_1$ goes to infinity, and $R^2 \to 1$. Information consistency is said to occur if, under this asymptotic regime, $B_{10}$ does not go to infinity. In Chapter 3, we prove that conjugate normal priors are often

information inconsistent, but thicker-tailed priors and empirical Bayes approaches (such as the one studied in Chapter 2) are information-consistent.

## 1.3 On limit consistency

Limit consistency is an asymptotic property of two-sample tests that was introduced in Alexander Ly's dissertation (Ly, 2017). Suppose that we have two groups, named $A$ and $B$. We assume that the sample size for group $A$ ($n_A$) is fixed, but the sample size for group $B$ ($n_B$) goes to infinity. According to the definition in (Ly, 2017), a Bayes factor is limit-consistent if the limiting Bayes factor as $n_B \to \infty$ is finite. The motivating example in (Ly, 2017) is testing whether the rates of 2 homogeneous Poisson processes are equal. In Chapter 4, we consider the problem of testing whether 2 normal means are equal, and propose a stronger definition of limit consistency. We describe the limiting behavior of Bayes decisions under different parametrizations and prior specifications, and recommend parametrizations that have common parameters to the null and alternative hypotheses along with flat (improper) priors on the common parameters.

## 1.4 On Birnbaum's theorem

Birnbaum's theorem shows that a statistician that is willing to accept the sufficiency principle and the weak conditionality principle must abide by the likelihood principle. Before we explain our contributions in Chapter 5, we define the principles at stake and give examples that involve testing statistical hypotheses.

The sufficiency principle states if the data come from a model with a sufficient statistic $T$, our inferences upon observing two samples $X$ and $Y$ with $T(X) = T(Y)$ should be the same. For example, if we have independent and identically distributed observations from a normal model, the sufficiency principle implies that we should make identical inferences from samples $X = (X_1, X_2, \ldots, X_n)$ and $Y =$

$(Y_1, Y_2, \ldots, Y_n)$ if $\sum_{i=1}^n X_i = \sum_{i=1}^n Y_i$ and $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n Y_i^2$. Therefore, statisticians who want to respect the sufficiency principle should not use rank-based tests (or the sample median) if they truly believe that the data are generated from a normal model.

The weak conditionality principle roughly states that our inferences should not depend on irrelevant experiments that were not performed. Formally, it is defined in terms of a mixture experiment. Suppose that we perform a mixture experiment to learn about a parameter $\theta$: we choose experiment $A$ with probability $\pi_A$ or experiment $B$ with probability $1-\pi_A$ ($\pi_A$ does not depend on $\theta$). The weak conditionality principle states that our final inferences should only depend on the experiment that was performed, and not on the fact that another experiment could have been performed. A classical example in Cox (1958) (which is also discussed in Chapter 5) shows that the most powerful test for a mixture experiment need not be equal to performing the optimal conditional tests after randomization.

Finally, the likelihood principle states that inferences for a parameter $\theta$ from models with proportional likelihoods (with respect to $\theta$) should be the same. A good pedagogical example is Example 12 in Berger (1985). Suppose that two experimenters are interested in the probability $\theta$ that a coin shows heads. Their hypotheses are $H_0 : \theta = 1/2$ against $H_1 : \theta > 1/2$, and they observe 9 heads and 3 tails. After analyzing the data, they reach different conclusions: one of them claims that she cannot reject $H_0$ at significance level 0.05, whereas the other one claims that the result is significant at the same significance level. The disagreement occured because the first experimenter found a $p-$value assuming that the data were Binomial$(13, \theta)$, whereas the second experimenter found a $p-$value assuming that the data are Negative-Binomial$(3, \theta)$. The data are exactly the same, the likelihoods are proportional (with respect to $\theta$), but the conclusions differ.

Standard Bayesian analysis with subjective priors satisfies all 3 principles. Ob-

jective Bayesian analysis can fail to satisfy them: for example, the reference prior for a binomial and a negative binomial example are different.

In Chapter 5, we discuss the articles Evans (2013) and Mayo (2014). We also include a brief discussion highlighting some difficulties in applying the ancillarity principle (which is a generalization of the weak conditionality principle) in frequentist statistics.

# 2

# Constrained empirical Bayes priors on regression coefficients

## 2.1 Introduction

In this chapter, we study the properties of constrained empirical Bayes procedures under model uncertainty. In particular, we restrict our attention to parametric type II maximum likelihood (type II ML), which in general proceeds as follows: (1) start with a sampling distribution for the data $y$ given a parameter $\theta$, which we assume has a density $f(y \mid \theta)$ (with respect to an appropriate dominating measure), and a prior $\pi_\eta(\theta)$ that depends on a hyperparameter $\eta \in \mathcal{C}$ and (2) set $\eta$ by maximizing the marginal likelihood $m(y)$ of the model, that is

$$\eta = \arg\max_{\eta \in \mathcal{C}} \int f(x \mid \theta) \pi_\eta(\theta) \, \mathrm{d}\theta = \arg\max_{\eta \in \mathcal{C}} m(y).$$

Type II ML was first proposed in Good (1965), and it is a particular instance of empirical Bayes which, in general, "estimates" the hyperparameter $\eta$ from the data, although not necessarily by maximizing the marginal likelihood (a popular alternative being the method of moments).

Unconstrained type II ML procedures can be problematic. In regression models, if the prior inclusion probabilities of the predictors are estimated via type II ML and the null or the full model have the highest marginal likelihood, the type II ML estimates are all exactly 0 or 1, respectively (Scott and Berger, 2010). In the same context, if the scale parameter of Zellner's $g$-prior (Zellner, 1986) is estimated via type II ML (George and Foster, 2000), the procedure is not model selection consistent under the null model (Liang et al., 2008).

A general problem with type II ML is that it can produce estimates that would never be chosen as hyperparameters in a *bona fide* Bayesian analysis because they would be regarded as overly informative. We propose avoiding this issue by maximizing marginal likelihoods subject to constraints that ensure that the estimates are at least as "vague" as the hyperparameters of proper default priors.

In the context of estimation, DasGupta and Studden (1989), Leamer (1978), and Polasek (1985) study priors that resemble our type II ML prior for normal linear models which bound the prior covariance matrix above and below (for known $\sigma^2$). To the best of our knowledge, the properties of constrained empirical Bayes procedures under model uncertainty have not been investigated elsewhere.

The examples we cover are regression coefficients in normal linear models (Section 2.2), generalized linear models (Section 2.3), high-dimensional ANOVA (Section 2.4), and the nonparametric regression example in Shibata (1983) (Section 2.5). The chapter ends with conclusions and possible directions for future work. Proofs are relegated to Appendix A.

## 2.2 Normal linear models

Consider the linear model

$$Y = X_0\beta_0 + X\beta + \epsilon,\ \epsilon \sim N_n(0_n, \sigma^2 I_n),$$

where $Y \in \mathbb{R}^n$, $X_0 \in \mathbb{R}^{n \times p_0}$, and $X \in \mathbb{R}^{n \times p}$. We assume that the predictors are linearly independent and that the blocks $X_0$ and $X$ are orthogonal, so that $X_0'X = 0_{p_0 \times p}$. The predictors in $X_0$ are understood to be a set of "common" predictors that are active in all the models in the model space (although, for now, we are focusing on a particular model with a fixed set of predictors). Our treatment covers the case where $X_0$ is an intercept and the predictors are centered by setting $X_0 = 1_n$, and the case where there are no common predictors by setting the set of common predictors to $\emptyset$. The prior on $\beta_0, \sigma^2$ is $\pi(\beta_0, \sigma^2) \propto 1/\sigma^2$, which is supported by group invariance arguments in Berger et al. (1998) and Bayarri et al. (2012).

In this section, we study the properties of a constrained type II ML procedure which estimates the prior covariance of $\beta$. Specifically, we start with $\beta \mid \sigma^2 \sim N_p(0_p, \sigma^2 W)$ and set $W$ by maximizing the marginal likelihood of the model with subject to the linear matrix inequality $W \succeq B$. Formally, $W \succeq B$ if and only if $W - B$ is positive semidefinite (this ordering is sometimes referred to as the Löwner order). In practice, the constraint implies conditions on the estimated prior covariance, which we denote $\widehat{W}$, that can be interpreted as "$\widehat{W}$ is at least as disperse as $B$." For example, the constraint ensures that $\mathrm{tr}(\widehat{W}) \geq \mathrm{tr}(B)$ and $\det(\widehat{W}) \geq \det(B)$. [Traces and determinants are used in design of experiments ($A$- and $D$-optimal designs) and multivariate analysis (in MANOVA, for instance) for measuring "total variablity," "dispersion" or "size" of matrices.] The restriction also implies a convex stochastic ordering for the type II ML prior. Let $\pi_{\mathsf{LB}}$ be the lower bound $N_p(\beta \mid 0_p, B)$ and $\pi_{\mathsf{ML}}$ be the type II ML prior. Then, for any convex function $f$, $\mathbb{E}_{\pi_{\mathsf{LB}}}[f(\beta)] \leq \mathbb{E}_{\pi_{\mathsf{ML}}}[f(\beta)]$ (Müller, 2001). In particular, the volume of highest probability density (HPD) regions under $\pi_{\mathsf{ML}}$ is greater or equal than the volume of the HPD regions under $\pi_{\mathsf{LB}}$. We focus on a default choice of $B$, but before we proceed we would like to remark that if prior information is available, one can define analogous procedures with informative lower bounds (see Section 2.5).

Our default choice of lower bound is $B = n(X'X)^{-1}$, under which we retrieve Zellner's $g$-prior (Zellner, 1986) with $g = n$. The choice $g = n$ corresponds to the prior covariance of the unit information prior (Kass and Wasserman, 1995; Raftery et al., 1997; Hoff, 2009), which is intuitively justified in the literature as follows. The Fisher information of $\beta$ is $(X'X)/\sigma^2$, so one can argue that $(X'X)/(n\sigma^2)$ contains as much information as a "typical" observation in the sample.

Under the model we described, for a fixed positive definite $W$ and $n \geq p + p_0$, the marginal likelihood is

$$\begin{aligned} m_W(Y) &= \int_{\mathbb{R}} \int_{\mathbb{R}^p} \int_{\mathbb{R}_+} N_n(Y \mid X_0\beta_0 + X\beta, \sigma^2 I_n)\, N_p(\beta \mid 0, \sigma^2 W)\, 1/\sigma^2 \,\mathrm{d}\beta_0 \,\mathrm{d}\beta \,\mathrm{d}\sigma^2 \\ &= \frac{\Gamma\left(\frac{n-p_0}{2}\right) \pi^{-(n-p_0)/2}}{\left(|X'X|\,|X_0'X_0|\,|(X'X)^{-1} + W|\right)^{1/2}} [\mathsf{SSE} + \widehat{\beta}'[W + (X'X)^{-1}]^{-1}\widehat{\beta}]^{-\frac{(n-p_0)}{2}}, \end{aligned} \tag{2.1}$$

where $\widehat{\beta} = (X'X)^{-1}X'Y$, $\mathsf{P}_X = X(X'X)^{-1}X'$, $\mathsf{P}_{X_0} = X_0(X_0'X_0)^{-1}X_0'$, and $\mathsf{SSE} = Y'(I_n - \mathsf{P}_{X_0} - \mathsf{P}_X)Y$. We study the properties of the constrained type II ML prior

$$\begin{aligned} \beta \mid \sigma^2 &\sim N_p(0_p, \sigma^2\, \widehat{W}) \\ \widehat{W} &= \arg\max_{W \succeq n(X'X)^{-1}} m(Y). \end{aligned}$$

Proposition 1 below gives an explicit closed-form expression for $\widehat{W}$. It is a linear combination of the unconstrained maximum over all positive semidefinite matrices, which is proportional to $\widehat{\beta}\widehat{\beta}'$, and the lower bound $n(X'X)^{-1}$.

**Proposition 1.** *Let $Y \sim N_n(X_0\beta_0 + X\beta, \sigma^2)$, where $X_0'X = 0_{p_0 \times p}$, $p(\beta_0, \sigma^2) \propto 1/\sigma^2$ and $\beta \mid \sigma^2 \sim N_p(0_p, \sigma^2 W)$. Let $m_W(y)$ be the marginal likelihood of the model, as defined in Equation 2.1. For $n > p + p_0$, the solution to the optimization problem*

$$\begin{aligned} &\text{maximize } m_W(y) \\ &\text{subject to } W \succeq n(X'X)^{-1} \end{aligned}$$

*can be written as*

$$\widehat{W} = a\,\widehat{\beta}\widehat{\beta}' + n(X'X)^{-1}$$

$$a = \max(0, (n - p_0 - 1)/\mathsf{SSE} - (n+1)/\mathsf{SSR})$$

$$\mathsf{SSR} = \widehat{\beta}'(X'X)\widehat{\beta}.$$

When the signal (represented by $\mathsf{SSR}$) is weak relative to the noise (represented by $\mathsf{SSE}$), $\widehat{W}$ is equal to the lower bound. As the signal to noise ratio increases (i.e. as $\mathsf{SSR}$ grows or $\mathsf{SSE}$ decreases),$\widehat{W}$ and the lower bound become more different. Indeed, the constant $a$ increases as the signal to noise ratio increases, giving more weight to the rank-1 matrix $\widehat{\beta}\widehat{\beta}'$ in the linear combination.

### *2.2.1 Model uncertainty and selection*

Let $X_i$ be a design matrix that includes a subset with $p_i$ out of the $p$ predictors in $X$, with $i \in \{1, 2, \ldots, 2^p\}$ ($p_i$ can be 0, which corresponds to the null model, which we denote $\mathcal{M}_0$), and let $\mathcal{M}_i$ be the model $Y = X_0\beta_0 + X_i\beta_i + \epsilon_i$, $\epsilon_i \sim N_n(0, \sigma^2 I_n)$. Throughout, we set prior covariances of locally – that is, each model $\mathcal{M}_i$ is assigned its own $\widehat{W}_i$. The local approach to empirical Bayes model selection is justified through information-theoretical arguments in Hansen and Yu (2003), and competing approaches are reviewed in Consonni et al. (2008). The posterior probability of a model $\mathcal{M}_i$ can be written as $\mathbb{P}(\mathcal{M}_i \mid Y) = \mathbb{P}(\mathcal{M}_i)B_{ib}/[\mathbb{P}(\mathcal{M}_0) + \sum_{j=1}^{2^p}\mathbb{P}(\mathcal{M}_j)B_{jb}]$, which depends on the data only through the Bayes factor of $\mathcal{M}_j$ to $\mathcal{M}_b$, which is equal to the ratio of the marginal likelihood of $\mathcal{M}_j$ to that of model $\mathcal{M}_b$. If common predictors are defined as "predictors that are active in both models," our prior specification would depend on the choice of base model (Liang et al., 2008). We assume that the set of common predictors $X_0$ is the same across all models or, analogously, we restrict our attention to null-based Bayes factors, namely

$$B_{i0} = \frac{\int N_n(Y \mid X_0\beta_0 + X_i\beta, \sigma^2 I_n)\pi_{\mathsf{ML}}(\beta_0, \beta_i, \sigma^2)\pi_0(\beta_0, \sigma^2)\, d(\beta_0, \beta_i, \sigma^2)}{\int N_n(Y \mid X_0\beta_0, \sigma^2 I_n)\pi_0(\beta_0, \sigma^2)\, d(\beta_0, \sigma^2)} = \frac{m_i(Y)}{m_0(Y)}.$$

The constrained type II ML null-based Bayes factor of $\mathcal{M}_i$ is

$$B_{i0} = \begin{cases} (n+1)^{\frac{n-p_0-p_i}{2}} \left[n(1-R_i^2)+1\right]^{-(n-p_0)/2} & \text{if } R_i^2 \leq \frac{n+1}{2n-p_0} \\ \varphi(n)^{-1/2} \left(R_i^2\right)^{-1/2} (1-R_i^2)^{-(n-p_0-1)/2} & \text{if } R_i^2 > \frac{n+1}{2n-p_0}, \end{cases} \tag{2.2}$$

where $\varphi(n) = \left(\frac{(n+1)^{p_i-1}(n-p_0)^{n-p_0}}{(n-p_0-1)^{n-p_0-1}}\right)$ and $R_i^2 = 1 - \mathsf{SSE}_i/\mathsf{SSE}_0$. The first case corresponds to the null-based Bayes factor with the lower bound $W = \sigma^2 n(X'X)^{-1}$ (which is Zellner's $g$-prior with $g = n$.

Before we study the properties of the procedure in more detail, we present an example with $p = 2$ predictors to introduce some geometric intuition. It also serves as motivation to compare $\pi_{\mathsf{LB}}$ and the type II ML prior $\pi_{\mathsf{ML}}$ to the Bayesian Information Criterion (BIC; Schwarz (1978)), which is defined as

$$\log N(Y \mid X_0 + \widehat{\beta}_0 + X\widehat{\beta}, \widehat{\sigma}^2 I_n) - (p/2)\log n,$$

where $\widehat{\beta}_0, \widehat{\beta}$, and $\widehat{\sigma}^2$ are the maximum likelihood estimators of $\beta_0, \beta$ and $\sigma^2$, respectively. Throughout, we treat $\exp(\mathsf{BIC})$ as an approximate marginal likelihood, which leads to the following null-based Bayes factors:

$$B_{i0,\mathsf{BIC}} = n^{-p_i/2}(1-R_i^2)^{-n/2}.$$

**Example 1.** *(Correlated predictors) Consider a model with 2 standardized (centered and scaled) predictors and an intercept* $Y = 1_n + X\beta + \epsilon$ *where* $\beta = (\beta_1, \beta_2)'$ *and* $\epsilon \sim N_n(0_n, \sigma^2 I_n)$ *and* $\sigma^2 = 1$. *Since the predictors are standardized, their (uncorrected) sample correlation is the off-diagonal entry of* $(X'X)/n$, *which we denote* $r$. *The prior covariance between* $X_1$ *and* $X_2$ *implied by the prior* $\beta \sim N_p(0_p, n(X'X)^{-1})$ *is* $-r/(1-r^2)$ *(Ghosh and Ghattas, 2015). Therefore, if* $X_1$ *and* $X_2$ *are positively correlated, the prior covariance between* $\beta_1$ *and* $\beta_2$ *is negative (and conversely for negative correlations). We set* $n = 10$, $\beta = (5, 5)'$ *and consider the cases* $r = 0.9$ *and* $r = -0.9$. *In order to isolate the effect of changing the sign of* $r$ *as much as*

*possible, we use the same random $\epsilon$ in both cases and the same $N_1(0,1)$ random numbers for generating the design matrices before transforming them (deterministically, via PCA scores times the Choleski of the target sample covariance) to correlated predictors with the desired $r$. Figure 2.1 shows contours of $N_p(0_p, n(X'X)^{-1})$ (solid blue) and $N_p(0_p, \widehat{W})$ (solid green), the type II ML prior for a simulated dataset. It also shows the contours of $N_p(\widehat{\beta}, n(X'X)^{-1})$ (dashed red) because: (1) BIC is a $O_p(n^{-1/2})$ approximation to the marginal likelihood under this data-dependent prior and the approximation is remarkably close in finite samples (Raftery, 1995), and (2) the likelihood (as a function of $\beta$) is proportional to $N_p(\widehat{\beta}, (X'X)^{-1})$, so it has the same shape.*

*When $r = -0.9$, the marginal likelihood of the true model is high with all the priors. If $r = 0.9$, the highest density regions of the likelihood are assigned relatively low probability density under $N_p(0_p, n(X'X)^{-1})$. Table 2.1 confirms this geometric intuition – for sample sizes ranging from 5 to 15 and after 1000 simulations, the average posterior probability that the lower bound (LB) assigns to the true model is lower than with BIC (we obtain similar results with $N_p(\widehat{\beta}, n(X'X)^{-1})$) or the type II ML prior (ML). This intuition can be formalized. If $\sigma^2$ is known,*

$$\log\left(\frac{B_{i0,\mathsf{BIC}}}{B_{i0,\mathsf{LB}}}\right) = \frac{p_i}{2}\log\left(\frac{n+1}{n}\right) + \frac{1}{2\sigma^2}\left(1 - \frac{n}{n+1}\right)\mathsf{SSR}_i > 0, \tag{2.3}$$

*which is positive and depends on the data only through $\mathsf{SSR}_i$. If the full model is true, $\mathbb{E}[\mathsf{SSR}_i] = 2\sigma^2 + n[\beta_1^2 + 2\beta_1\beta_2 r + \beta_2^2] > 0$. Since in our example $\beta_1$ and $\beta_2$ have the same sign, $\mathbb{E}[\log(B_{i0,\mathsf{BIC}}/B_{i0,\mathsf{LB}})]$ increases as $r$ increases. If $\beta_1$ and $\beta_2$ had different signs, we expect differences to be smaller as $r$ increases. The Zellner-Siow (ZS) prior, which is* $\text{Cauchy}_p(0_p, \sigma^2 n(X'X)^{-1})$ *is less sensitive to the sign of $r$ than the lower bound, despite the fact that they are both centered at $0_p$ and have the same prior scale.*

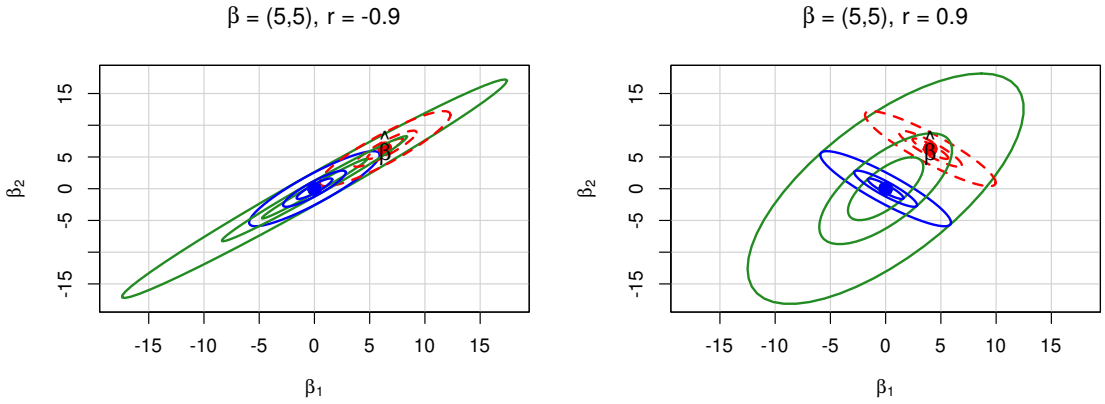


FIGURE 2.1: HPD regions (20%, 50%, 95%) of the lower bound $N_p(0_p, n(X'X)^{-1})$ (dotted blue), BIC prior $N_p(\widehat{\beta}, n(X'X)^{-1})$ (dashed red), and the type II ML prior $N_p(0_p, \widehat{W})$ (solid green). The MLE is indicated with a $\widehat{\beta}$ symbol.

Table 2.1: Average posterior probability assigned to the true model (full model), $B = 1000$ simulations.

| | $r = -0.9$ | | | | $r = 0.9$ | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | BIC | ML | LB | ZS | BIC | ML | LB | ZS |
| 5 | 0.957 | 0.802 | 0.668 | 0.505 | 0.980 | 0.915 | 0.361 | 0.310 |
| 10 | 0.996 | 0.981 | 0.974 | 0.949 | 0.998 | 0.995 | 0.606 | 0.966 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.979 | 1.000 |
| 20 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Intuitively, the type II ML prior is a compromise between BIC and the lower bound: it maximizes the marginal likelihood so that the asymmetry in Figure 2.1 does not occur, but at the same time it is centered at $0_p$ and "more vague" than the lower bound. This is made precise in Proposition 2 below.

**Proposition 2.** *Let $\mathcal{M}_i$ be the model with conditional mean $X_0\beta_0 + X_i\beta_i$, where $X_i$ are the model-specific predictors and $X_0$ are the common predictors, with $X_0'X_i = 0_{p_0 \times p}$ (if $i = 0$, we recover the null model $\mathcal{M}_0$). If $\mathcal{M}_j \supset \mathcal{M}_i$, then*

$$B_{ji,\mathsf{BIC}} \geq B_{ji,\mathsf{ML}} \geq B_{ji,\mathsf{LB}},$$

*where $B_{ji,\mathsf{BIC}} = n^{-(p_j - p_i)/2}[(1 - R_i^2)/(1 - R_j^2)]^{-n/2}$. Let $\mathcal{M}_f$ be the full model (which*

*includes all p predictors) and $\mathcal{M}_0$ be the null model. If the prior on the model space is the same in all cases, the inequality above implies*

$$\mathbb{P}_{\mathsf{BIC}}(\mathcal{M}_f \mid Y) \geq \mathbb{P}_{\mathsf{ML}}(\mathcal{M}_f \mid Y) \geq \mathbb{P}_{\mathsf{LB}}(\mathcal{M}_f \mid Y)$$

$$\mathbb{P}_{\mathsf{BIC}}(\mathcal{M}_0 \mid Y) \leq \mathbb{P}_{\mathsf{ML}}(\mathcal{M}_0 \mid Y) \leq \mathbb{P}_{\mathsf{LB}}(\mathcal{M}_0 \mid Y).$$

BIC assigns more posterior probability to the full model than the type II ML prior and the lower bound, and the lower bound assigns more posterior probability to the null model than the type II ML prior and BIC. However, there is another interesting asymmetry. When the true model is the null model, the differences between the lower bound and BIC tend to be small, whereas if the true model is not the full model, the differences can be rather large. If we assume that $\sigma^2$ is known, the expression for $\log(B_{i0,\mathsf{BIC}}/B_{i0,\mathsf{LB}})$ is given in Equation 2.3. If the true model is $\mathcal{M}_i$ and $\beta_*$ is the true value of $\beta$, we can write $\mathbb{E}[\mathsf{SSR}_i] = p_i\sigma^2 + \beta_*' X_i' X_i \beta_*$, which is positive and clearly depends on the sample covariance between the predictors. For fixed $X_i$ and $X_0\beta_{0*}$, $\mathbb{E}[\log(B_{i0,\mathsf{BIC}}/B_{i0,\mathsf{LB}})]$ is minimized when $\beta_* = 0_{p_*}$; that is, the Bayes factors with BIC and the lower bound are most similar when the null model is true. The fact that the difference between the lower bound and BIC is increasing with $\mathsf{SSR}_i$ is consistent with our interpretation of $\widehat{W}$ at the end of Section 2.2. Also note that $\mathbb{E}[\mathsf{SSR}_i]$ is increasing in $p_i$, which implies that the expected (log) differences between the lower bound and BIC grow as the number of predictors grows (this is seen empirically in simulation studies in Section 2.2.3). For unknown $\sigma^2$, $\log(B_{i0,\mathsf{BIC}}/B_{i0,\mathsf{LB}})$ is increasing in $R_i^2$, which is also consistent with our intuition.

Now, we check whether the type II ML prior satisfies the desiderata in Bayarri et al. (2012) for objective priors in model selection.

1. **Basic criterion:** The basic criterion is satisfied if the prior is proper. The restriction ensures that the type II ML prior is proper and non-singular.

2. **Model selection consistency:** Let the true model be $\mathcal{M}_* : Y = X_0\beta_0 +$

$X_*\beta_* + \epsilon_i$, $\epsilon_i \sim N_n(0, \sigma^2 I_n)$, where $X_*$ is a subset of the $p$ predictors in $X$. Model selection consistency is satisfied if $\mathbb{P}(\mathcal{M}_* \mid Y) \to 1$ in probability as $n \to \infty$. The type II ML prior is model-selection consistent under the following regularity condition, which is commonly made in the literature (Fernandez et al., 2001; Liang et al., 2008; Guo and Speckman, 2009; Maruyama and George, 2011; Bayarri et al., 2012). For any model that does not nest the true model, assume that

$$\lim_{n\to\infty} \frac{\beta_*' X_*'(I - \mathsf{P}_{X_j}) X_* \beta_*}{n} = b_j \in (0, \infty).$$

The assumption can be interpreted as that the models are distinguishable in the limit (Bayarri et al., 2012).

3. **Information consistency:** Assume that, for a fixed $n$, $R_i^2 \to 1$. This is a situation where there is overwhelming evidence in favor of $\mathcal{M}_i$ (Liang et al., 2008). Information consistency holds if $B_{i0} \to \infty$, which is satisfied by the type II ML prior.

4. **Intrinsic consistency:** A prior satisfies intrinsic consistency if, as $n$ grows, it converges to a proper prior which does not depend on model-specific parameters or $n$. The type II ML prior does not satisfy this criterion. To see this, assume that $(X_i' X_i)/n \to \Xi_i$ for a positive definite matrix $\Xi_i$, which holds if there is a fixed design or the covariates are drawn from a distribution with finite second moments (Bayarri et al., 2012). Then, the prior covariance $\widehat{W}_*$ has the following limiting behavior (in probability)

$$\widehat{W}_* \to \begin{cases} \Xi_*^{-1} & \text{if } \beta_*'\Xi_*\beta_* \leq \sigma_*^2 \\ \left(\frac{1}{\sigma_*^2} - \frac{1}{\beta_*'\Xi_*\beta_*}\right)\beta_*\beta_*' + \Xi_*^{-1} & \text{if } \beta_*'\Xi_*\beta_* > \sigma_*^2 \end{cases},$$

which depends on $\beta_*$ and $\sigma_*^2$.

5. **Null and dimensional predictive matching:** In both cases, the notion of minimal training sample size is central to the definition. For any model $\mathcal{M}_i$, the minimal training sample size is the smallest sample size $n_i^*$ such that the marginal likelihood of the model is finite. Null predictive matching is achieved if, for any model $\mathcal{M}_i$, we have $B_{i0} = 1$ for all possible samples of size $n_i^*$. Dimensional predictive matching is achieved if, for any pair models of the same dimension $\mathcal{M}_i$ and $\mathcal{M}_j$, we have $B_{ij} = 1$ whenever $n_i^* = n_j^*$. The type II ML prior is not null or dimensional predictive matching. For $p > 1$, the minimal training sample size for the type II ML prior is $n = p + p_0 + 1$. [If $p = 1$, the marginal likelihood does not depend on the choice of $W$.] When $n = p+p_0$, the marginal likelihood is finite for any given $W$, but one can choose $W \succeq n(X'X)^{-1}$ so that the marginal goes to $\infty$ (this is shown in Appendix A). Null predictive matching is not satisfied: in fact, $B_{i0}$ goes to $\infty$ as $R_i^2 \to 1$ when $n = p + p_0 + 1$. Similarly, it is easy to see that dimensional predictive matching is not satisfied either; different models will have different $R_i^2$, yielding Bayes factors that are different than 1.

6. **Invariance:** The type II ML prior is invariant with respect to linear transformations of the design matrix (in particular, this is true for changes of measurement units). More explicitly, let $H$ be an invertible $p \times p$ matrix and $\tilde{X} = XH$. Let $\beta$ and $\tilde{\beta}$ be the regression coefficients of the linear model if the design matrices are $X$ and $\tilde{X}$, respectively. If the type II ML prior is put on $\beta$ and $\tilde{\beta}$, then $\beta$ and $H\tilde{\beta}$ are equal in distribution (a proof can be found in Appendix A).

George and Foster (2000) propose a version of Zellner's $g$-prior where $\widehat{g}$ is set by maximizing the marginal likelihood subject to $g \geq 0$. This type II ML prior has undesirable features that are a byproduct of maximizing the marginal likelihood subject to a lower bound on $g$ that is not bounded away from 0: it is inconsistent

Table 2.2: Comparison of model selection desiderata for different approaches.

| | ML | BIC | LB | ZS | $\widehat{g}$ |
|---|---|---|---|---|---|
| Proper | yes | - | yes | yes | yes |
| Model selection consistency | yes | yes | yes | yes | no |
| Information consistency | yes | yes | no | yes | yes |
| Intrinsic consistency | no | - | yes | yes | no |
| Predictive matching | no | no | yes | yes | no |
| Invariance | yes | - | yes | yes | yes |

under the null model (Liang et al., 2008), and it is straightforward to show that null-based Bayes factors are always greater or equal to 1. Our constrained type II ML prior avoids both of these undesirable features.

We conclude the subsection with a comparison with BIC, the lower bound (LB), the Zellner-Siow prior (ZS), and prior proposed in George and Foster (2000), which we denote $\widehat{g}$ (see Table 2.2). None of the approaches dominates the rest.

### *2.2.2 Estimation and prediction*

For simplicity, we omit model subscripts and assume that the model is $Y \sim N_n(X_0\beta_0 + X\beta, \sigma^2 I_n)$, $X_0'X = 0_{p_0 \times p}$. If we put the right-Haar prior $\pi(\alpha, \sigma^2) \propto 1/\sigma^2$ on the common parameters and the type II ML prior on $\beta \mid \sigma^2$, the posterior mean of $\beta$ is

$$\tilde{\beta} = \mathbb{E}(\beta \mid Y) = \begin{cases} \frac{n}{n+1}\widehat{\beta} & \text{if } R^2 \leq \frac{n+1}{2n-p_0} \\ \left(1 - \frac{1-R^2}{(n-p_0-1)R^2}\right)\widehat{\beta} & \text{if } R^2 > \frac{n+1}{2n-p_0}. \end{cases}$$

The expression can be derived by applying the Sherman-Morrison formula twice to $\mathbb{E}(\beta \mid Y) = [\widehat{W}^{-1} + (X'X)]^{-1}X'Y$. The properties of an analogous estimator in the Normal means problem (for known $\sigma^2$) are studied in DasGupta and Studden (1989), where it is shown that it is minimax with respect to the square loss. Proposition 3 shows that $\mathbb{E}(\beta \mid Y)$ is minimax with respect to a (scaled) predictive loss because it belongs to the class of minimax estimators characterized in Strawderman (1973).

**Proposition 3.** *Let $p \geq 3$ and $n > p + p_0$. The estimator $\tilde{\beta} = \mathbb{E}(\beta \mid Y)$ is minimax with respect to the (scaled) squared predictive loss*

$$L(\beta, \delta) = (\beta - \delta)'(X'X)(\beta - \delta)/\sigma^2$$

The mean squared error of the posterior mean of the lower bound is

$$\mathbb{E}[\|n\,\widehat{\beta}/(n+1) - \beta\|^2] = \sigma^2 \left(\frac{n}{n+1}\right)^2 \mathrm{tr}\left((X'X)^{-1}\right) + \frac{\|\beta\|^2}{(n+1)^2},$$

which is increasing in $\|\beta\|$. On the other hand, the mean squared error of $\hat{\beta}$ is

$$\mathbb{E}[\|\widehat{\beta} - \beta\|^2] = \sigma^2 \mathrm{tr}\left((X'X)^{-1}\right),$$

which is constant in $\|\beta\|$. The estimator $\tilde{\beta}$ is equal to $n\widehat{\beta}/(n+1)$ when $R^2$ is low, and very close to $\widehat{\beta}$ when $R^2$ is high. Therefore, $\tilde{\beta}$ avoids "selecting" $n\widehat{\beta}/(n+1)$ in cases where it has high mean squared error (that is, whenever $\|\beta\|$ and $R^2$ are high).

### *2.2.3 Simulation study*

We simulate data from $Y = 1_n\,\alpha + X\beta + \epsilon$, $\epsilon \sim N_n(0_n, \sigma^2 I_n)$, where $n = 20$, $\alpha = 2$, $\sigma^2 = 1$, and $\beta$ is 8-dimensional with $k$ nonzero elements, for $k \in \{0, 1, 2, \ldots, 8\}$. We consider 2 different types of correlation between the predictors: an orthogonal case $X'X = I_p$ and an AR(1) structure

$$\frac{1}{n-1}(X'X) = \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^p \\ \rho & 1 & \rho & \ldots & \rho^{p-1} \\ \rho^2 & \rho & 1 & \ldots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^p & \rho^{p-1} & \rho^{p-2} & \ldots & 1 \end{pmatrix}$$

for $\rho = 0.9$. For all $k$, we generate $\beta_k \sim N_k(0_k, gI_k)$. The location of the $k$ zeros in the $\beta$ vector is drawn at random (according to the uniform distribution). We use $g \in \{5, 25\}$ as in Cui and George (2008) and Liang et al. (2008), representing

weak and strong signal-to-noise ratios, and evaluate performance with respect to the predictive squared loss function

$$L(\beta, \delta) = (\beta - \delta)'(X'X)(\beta - \delta),$$

where $\delta$ is an estimator of $\beta$ (this is the loss function that was used in the simulation studies in Cui and George (2008) and Liang et al. (2008)). The estimators that are considered are the posterior means (and $\widehat{\beta}$ in the case of BIC) of the highest probability model (HPM), median probability model (MPM), and marginal coefficients after model averaging (BMA) with BIC, the lower bound (LB), the type II ML prior (ML), and Zellner-Siow (ZS). We simulate 1000 datasets for all scenarios, and the results are reported in Figures 2.2 and 2.3.

In the orthogonal case, BIC and the type II ML prior have very similar behavior. Their losses are slightly higher than those of the LB or ZS when the number of true predictors is between 1 and 6; when the number of true predictors is 7 or 8, BIC and the ML have slightly smaller losses than the LB.

The results with the AR(1) correlation structure show bigger discrepancies. As the number of true predictors increases, the loss of the LB is substantially higher than the loss with any other prior, especially when $g = 25$. In the cases where the LB is clearly outperformed, its posterior distribution over the model space is closer to the uniform distribution than the other posteriors, as evidenced in the first panel in Figure 2.4, which shows the average entropy of the posterior distributions over the model space. ZS and the LB select HPMs and MPMs with fewer predictors than BIC and the type II ML prior (see second and third panel in Figure 2.4, which show the percentage of times the MPM equals the true model and the average size of the MPM, respectively). The HPM and MPM with BIC and the type II ML prior tend to be the same model, and they coincide with the models selected with the LB in the cases where the signal is low, as expected. On the other hand, when the signal is high,

the LB assigns more probability to wrong models than the other approaches, and sometimes the HPM and MPM end up being an egregiously wrong model, resulting in a substantially higher average loss. Note that ZS, which also has $n(X'X)^{-1}$ as its prior scale but has thicker tails, does not seem to be nearly as affected by this issue as the LB, especially when the signal is high enough (e.g. $g = 25$).

## 2.3 Generalized linear models

We study the properties of a type II ML prior for generalized linear models (GLMs; McCullagh and Nelder (1989)) that builds upon the work of Li and Clyde (2015). The prior has a very similar form and similar properties as the type II ML prior introduced in Section 2.2. Let the outcomes $y_i$ be independent with density

$$p(y_i \mid \theta) = \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a(\phi_0)} + c(y_i, \phi_0)\right\}$$

for $i \in \{1, 2, \ldots, n\}$. The predictors enter the model through $\theta_i = h(\eta_i)$, where $\eta_i = 1_n\alpha + X\beta$. The function $h$ is the so-called link function, and the identity $h(x) = x$ is referred to as the canonical link. For simplicity, we assume that $a(\phi_0) = \phi_0/w$ with known $\phi_0$ and $w$ (which includes binary and Poisson regression), but the results here can be extended as in Li and Clyde (2015) in a straightforward fashion. Throughout, we assume that the design matrices $X$ are full-rank and the maximum likelihood estimators exist and are unique.

A plethora of priors for GLMs has been proposed; see Section 2.1 in Li and Clyde (2015) for a lucid review. Some of the priors on $\beta$ that have been studied are either normal or mixtures or normals (Sabanés-Bové et al., 2011; Copas, 1983; Held et al., 2015; Li and Clyde, 2015). In normal linear models, the Fisher information of $\beta$ does not depend on $\beta$ but, in general, the Fisher information of $\theta$ in GLMs can depend on $\theta$ itself. As a consequence, there are priors in the literature that resemble

Zellner's $g$-prior but have different prior covariances: while some use the expected information matrix (Kass and Wasserman, 1995; Hansen and Yu, 2003; Marin and Robert, 2007; Sabanés-Bové et al., 2011) others favor the observed information Wang and George (2007); and whereas Copas (1983); Sabanés-Bové et al. (2011) propose evaluating the information parameters at $0_p$, Li and Clyde (2015) propose evaluating information matrices at the maximum likelihood estimates. In this section we build directly upon Li and Clyde (2015), but similar type II ML priors can be defined using other prior covariances as lower bounds. Li and Clyde (2015) propose putting a flat prior $p(\alpha) \propto 1$ on the intercept and

$$\beta \mid g \sim N_p(0_p, g\mathcal{J}_n(\widehat{\beta})^{-1}),$$

where

$$\begin{aligned}
\widehat{\mathsf{P}}_1 &= 1_n(1_n'\mathcal{J}_n(\widehat{\eta})1_n)^{-1}1_n'\mathcal{J}_n(\widehat{\eta}) \\
\mathcal{J}_n(\widehat{\eta}) &= \mathrm{diag}(d_i), \;\; d_i = -Y_i\theta''(\widehat{\eta}) + (b \circ \theta)''(\widehat{\eta}) \\
\mathcal{J}_n(\widehat{\beta}) &= X'(I_n - \widehat{\mathsf{P}}_1)'\mathcal{J}_n(\widehat{\eta})(I_n - \widehat{\mathsf{P}}_1)X.
\end{aligned}$$

Li and Clyde (2015) find approximate Bayes factors through a Laplace approximation, which results in an approximate marginal likelihood that is proportional to $N_p(\beta \mid \widehat{\beta}, A)$, where $A = \mathcal{J}_n(\widehat{\beta})^{-1}$. If $\tilde{m}(Y)$ is the approximate marginal likelihood, a type II ML prior which is analogous to the one in Section 2.2 is

$$\beta \sim N_p(0_p, W), \;\; \arg\max_{W \succeq cA} \tilde{m}(y)$$

for $c > 0$. Mirroring our work in Section 2.2, we can set $c = n$. However, recall that the motivation behind this choice was that, in the normal linear model, the Fisher information of $\beta$ is $(X'X)/\sigma^2$, so the information in a typical observation is roughly $(X'X)/(n\sigma^2)$. It is less clear that this relation should hold approximately in GLMs. [This is in direct connection with the issue of determining the effective

sample size for a parameter (Berger et al., 2014).] For simplicity, we take $c = n$ in the sequel, but it is important to stress that more research is needed in this direction. Proposition 4 below shows the explicit expression of the optimum (a proof can be found in Appendix B.2).

**Proposition 4.** *The covariance matrix that maximizes the approximate marginal likelihood* $\tilde{m}(y)$ *is*

$$\widehat{W} = nA + a\,\widehat{\beta}\widehat{\beta}',$$

*where*

$$a = \max\left(0, 1 - \frac{n+1}{Q}\right), \;\; Q = \widehat{\beta}' A^{-1} \widehat{\beta}.$$

Plugging in the solution in Propositions 4 and adding model subscripts as in Subsection 2.2.1, the null-based Bayes factors are

$$B_{i0} = \frac{p(Y \mid \widehat{\alpha}_i, \widehat{\beta}_i)}{p(Y \mid \widehat{\alpha}_0)} \left[\frac{\mathcal{J}(\widehat{\alpha}_0)}{\mathcal{J}(\widehat{\alpha}_i)}\right]^{1/2} (n+1)^{-p/2}\, \Omega_{\rangle\prime},$$

where

$$\Omega_{i0} = \mathbb{1}(a_i = 0) \exp\left(-\frac{1}{2}\frac{Q_i}{n+1}\right) + \mathbb{1}(a_i \neq 0)\,[Q_i/(n+1)]^{-1/2} \exp\left(-\frac{1}{2}\right).$$

As we did in Section 2.2, we study whether the type II ML prior satisfies the desiderata in Bayarri et al. (2012).

1. **Basic criterion:** The type II ML prior is proper, so the basic criterion is satisfied.

2. **Model selection consistency:** Under the regularity conditions in Li and Clyde (2015), the type II ML prior is consistent.

3. **Information consistency:** As Li and Clyde (2015) point out, the likelihoods of some GLMs are bounded (as a function of the data), and it is unclear whether this criterion applies in general.

4. **Intrinsic consistency and predictive matching:** The type II ML prior does not satisfy intrinsic consistency or predictive matching. This can be seen following the same reasoning used in Section 2.2.

5. **Invariance:** The type II ML prior is invariant to linear transformations of the design matrix (in particular, this is true for changes of measurement units).

The relationship between this type II ML prior, BIC, and the lower bound is very similar to the one described in normal linear models.

## 2.4 High-dimensional ANOVA

We revisit the high-dimensional one-way ANOVA example Berger et al. (2003). In this case, the choice of restriction can have a strong effect in the properties of the procedure. Let $p$ be the number of groups and $r$ the replicates within each group. We assume that $r$ is fixed and $p$ grows to infinity. We have observations

$$y_{ij} = \mu_i + \epsilon_{ij},\ \epsilon_{ij} \overset{\text{iid}}{\sim} N_1(0, 1)$$

where $i \in \{1, 2, \ldots, p\}$ (groups) and $j \in \{1, 2, \ldots, r\}$ (replicates).

We only consider the null model and the full model

$$\mathcal{M}_1 : \mu = 0_p,\ \mathcal{M}_2 : \mu \in \mathbb{R}^p.$$

Under the true model, we assume (as in Berger et al. (2003)) that

$$\liminf_{p\to\infty} \|\mu\|^2/p = \tau^2 > 0.$$

Let $\ell$ be the log-likelihood function of the full model and $\widehat{\mu}$ the maximum likelihood estimate of $\mu$. If BIC is defined as

$$\ell(\widehat{\mu}) - (p/2)\log n,$$

it is inconsistent under $\mathcal{M}_2$ (Stone, 1979). Berger et al. (2003) show that if the prior on $\mu$ under $\mathcal{M}_2$ is $\mu \mid g \sim N_p(0, gI_p)$ along with a mixing density $g$ with support $(0, +\infty)$, consistency holds; however, if $g$ has constrained support $(0, T)$ for $T < \infty$, there is a region of inconsistency under $\mathcal{M}_2$.

In this example, the prior scale has to be chosen carefully. A naive parallel of the type II ML prior in Section 2.2 would set the lower bound to be $n(X'X)^{-1} = \frac{n}{r}I_p = pI_p$. However, it is easy to see (using Markov's inequality) that any normal prior whose scale increases as $p$ increases is inconsistent under $\mathcal{M}_2$. Since the expected sample size of $\mu$ in this problem is $r$ instead of $n$ (Berger et al., 2014), we use $r(X'X)^{-1} = I_p$ as our lower bound. In the same vein, BIC can be redefined appropriately by taking $\log r$ as the penalty instead of $\log n$. The asymptotic behavior of the modified versions of the lower bound and BIC can be summarized as follows:

- Normal prior with $I_p$ as the lower bound: Under $\mathcal{M}_1$, consistency for all $r$ and $\tau^2$. Under $\mathcal{M}_2$, inconsistency if $\tau^2 \leq (1+r)\log(1+r)/r^2 - 1/r$ and consistency otherwise. For example, if $\tau^2 = 0.25$, consistency holds under $\mathcal{M}_2$ for $r \geq 5$, and consistency holds for all $r$ if $\tau^2 > 0.386$.

- BIC with $\log r$ as penalty: Under $\mathcal{M}_1$, inconsistency if $r \in \{1, 2\}$ and consistency otherwise. Under $\mathcal{M}_2$, inconsistency if $\tau^2 \leq (\log r - 1)/r$ and consistency otherwise. The condition is most stringent at $r = e^2$, so consistency holds for all $r$ if $\tau^2 > 1/e^2 \approx 0.15$.

Under $\mathcal{M}_2$, the region of inconsistency of BIC is contained in the region of inconsistency of the normal prior; however, BIC can be inconsistent under $\mathcal{M}_1$.

The type II ML prior

$$\mu \sim N_p(0_p, \widehat{W}),\ \widehat{W} = \arg\max_{W \succeq I} m(Y)$$

is consistent under $\mathcal{M}_1$ and has the same region of inconsistency as BIC under $\mathcal{M}_2$. Therefore, it is a desirable compromise between the normal prior and BIC but, unfortunately, still has a region of inconsistency that mixtures of Normal priors avoid.

## 2.5 Shibata example

We revisit the example in Shibata (1983), which was also studied in Barbieri and Berger (2004), and observe that informative lower bounds in type II ML priors can lead to better inferences. The goal is estimating the function $f(x) = -\log(1-x)$, $-1 \leq x \leq 1$ from noisy observations $y = f(x) + \epsilon$, where $\epsilon \sim N_n(0_n, \sigma^2 I_n)$, where $\sigma^2$ is known. To that end, we fit (nested) models

$$\mathcal{M}_j : Y \mid \alpha, \beta_j, \sigma^2 \sim N_n(1_n\alpha + X_j\beta_j, \sigma^2 I_n)$$

for $j \in \{1, 2, 3, \ldots, k\}$, where the design matrices $X_j$ have dimension $n \times j$ and the columns are given by the Chebyshev polynomials of the first kind evaluated at the knots $x_i = \cos(\pi(n - i + 1/2)/n)$, for $i \in \{1, 2, \ldots, n\}$. The true coefficients in an infinite orthogonal expansion are $\alpha = \log 2$ and $\beta_j = 2/j$. The design matrices are orthogonal with $X_j'X_j = (n/2)I_j$ and $1_n'X_j = 0_j'$. We put a uniform prior on the model space and compare the squared predictive loss $L(f, \widehat{f}) = \int_{-1}^{1}(f(x) - \widehat{f}(x))^2 \, \mathrm{d}x$ using different local type II ML priors. More specifically, we consider priors $\beta \sim$

$N(0, A)$, where $A$ is equal to

$$D_{\text{ord}} = \arg\max_{A=\text{diag}(d_1,d_2,\ldots,d_p),d_1\geq d_2\geq\ldots\geq d_p\geq 0} m(Y|A)$$

$$D_{i^{-a}} = \arg\max_{A=\text{diag}(ci^{-a}),\ c,a\geq 0}\ m(Y|A)$$

$$\widehat{W} = \arg\max_{A\succeq\sigma^2 n(X'X)^{-1}} m(Y|A) = n\sigma(X'X)^{-1} + \max\left(0, 1 - \frac{\sigma^2(n+1)}{\widehat{\beta}'X'X\widehat{\beta}}\right)\widehat{\beta}\widehat{\beta}'.$$

We also consider AIC ($\ell(\widehat{\beta}_j) - j$) and BIC ($\ell(\widehat{\beta}_j) - (j/2)\log n$), treating exp(AIC) and exp(BIC) as approximate marginal likelihoods. The results are summarized in Table 2.5. The lower bounds that use prior information outperform the others. The results with the prior that imposes $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ are slightly worse than those with the prior with exponentially decreasing scales $ci^{-a}$, which is the optimal scale – a piece of prior information that might not be available in a realistic application. We compare the predictive losses of Bayesian model averaging (BMA), the median probability model (MPM; Barbieri and Berger (2004)), and the highest probability model (HPM). BMA outperforms the rest, followed by the MPM and the HPM. In Barbieri and Berger (2004), it is claimed that the type II ML procedure therein is "considerably better" than AIC and BIC. In that analysis, AIC and BIC were not considered for model averaging (so the results concerning AIC and BIC in Barbieri and Berger (2004) correspond only to the HPM section of Table 2.5). After our analysis, it is apparent that the performance of AIC and BIC is comparable to that of (uninformative) type II ML priors if they are averaged as approximate marginal likelihoods, and that adding prior information is clearly advantageous.

## 2.6 Conclusions and future work

We have studied the properties of some constrained type II ML priors under model uncertainty. In normal linear models, we studied a normal prior whose prior covariance is set by maximizing the marginal likelihood of the model subject to a

Table 2.3: Predictive loss, 1000 simulations. Average model sizes in square brackets.

| **HPM** | $D$ ord | $ci^{-a}$ | $\widehat{W}$ | AIC | BIC |
|---|---|---|---|---|---|
| $n = 30, k = 29, \sigma^2 = 1$ | 0.933 [7] | **0.894** [10] | 1.968 [29] | 1.069 [7] | 1.133 [4] |
| $n = 100, k = 79, \sigma^2 = 1$ | 0.589 [8] | **0.472** [23] | 0.687 [7] | 0.594 [13] | 0.687 [7] |
| **MPM** | | | | | |
| $n = 30, k = 29, \sigma^2 = 1$ | 0.943 [6] | **0.833** [16] | 1.094 [4] | 1.010 [7] | 1.088 [4] |
| $n = 100, k = 79, \sigma^2 = 1$ | 0.598 [8] | **0.438** [44] | 0.673 [7] | 0.575 [14] | 0.672 [7] |
| **BMA** | | | | | |
| $n = 30, k = 29, \sigma^2 = 1$ | 0.846 | **0.830** | 0.990 | 0.908 | 0.984 |
| $n = 100, k = 79, \sigma^2 = 1$ | 0.568 | **0.437** | 0.623 | 0.521 | 0.621 |

linear matrix constraint. The lower bound, which is Zellner's $g$-prior with $g = n$, is information-inconsistent and can have undesirable behavior if the predictors are strongly correlated (especially if most predictors are active). The restricted type II ML prior avoids both issues, and so do thick-tailed priors such as the Zellner-Siow prior. The type II ML procedure is remarkably close to BIC, which might be seen as a new argument for Bayesians to use the latter (since it is similar to a procedure with a robust Bayesian interpretation).

In generalized linear models, we built upon the work of Li and Clyde (2015) and defined an analogous type II ML prior, although the appropriateness of the restriction is less clear because the effective sample size (Berger et al., 2014) in, say, binary regression need not be the same as in the normal linear model. We also studied an example in high-dimensional one-way ANOVA where the number of groups goes to infinity; in this case, choosing the restriction appropriately is important. Finally, we revisited the nonparametric regression example in Shibata (1983) and concluded that informative restrictions can help us make better inferences, as expected. This chapter is an initial exploration of constrained type II maximum likelihood procedures, so we cannot make any overarching claims about their properties or desirability. An obvious limitation of our treatment is that we only covered normal priors on regression coefficients. However, we think our results are still interesting because,

if nothing else, normal priors are still wildly used, and these constrained empirical Bayes procedures avoid some of their undesirable issues.

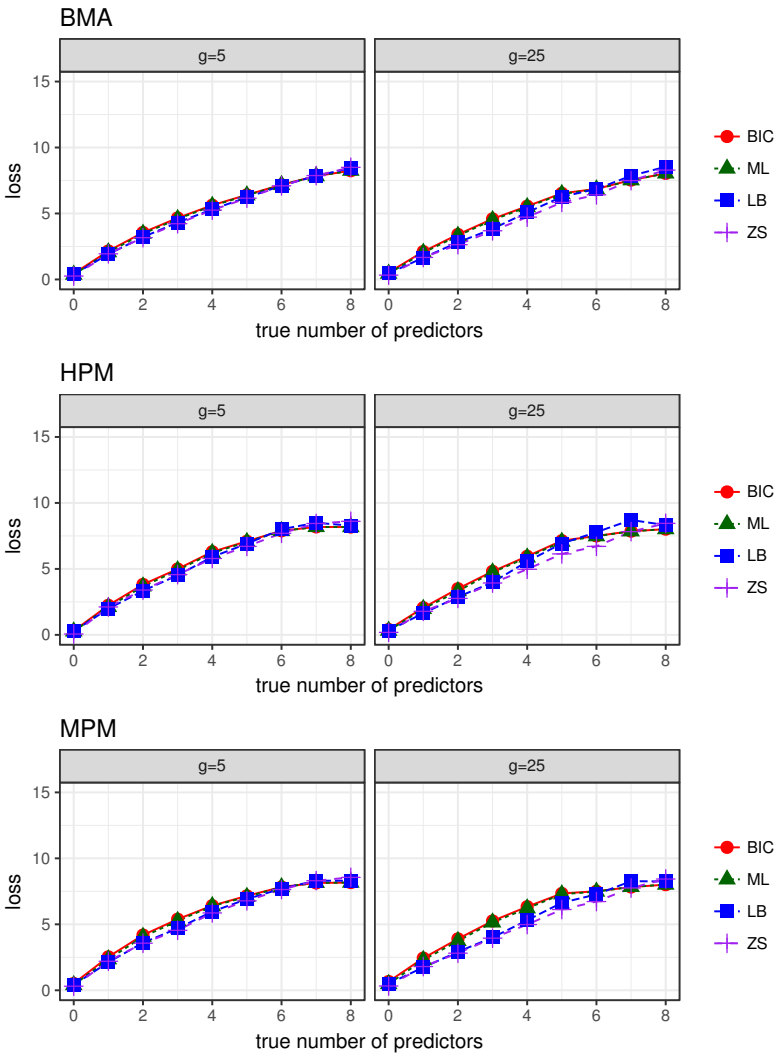


FIGURE 2.2: Simulation study: comparison between highest probability model (HPM), median probability model (MPM), and coefficients after model averaging (BMA) with BIC, the lower bound of the type II ML prior (LB), the type II ML prior (ML), and ZS. Orthogonal design.

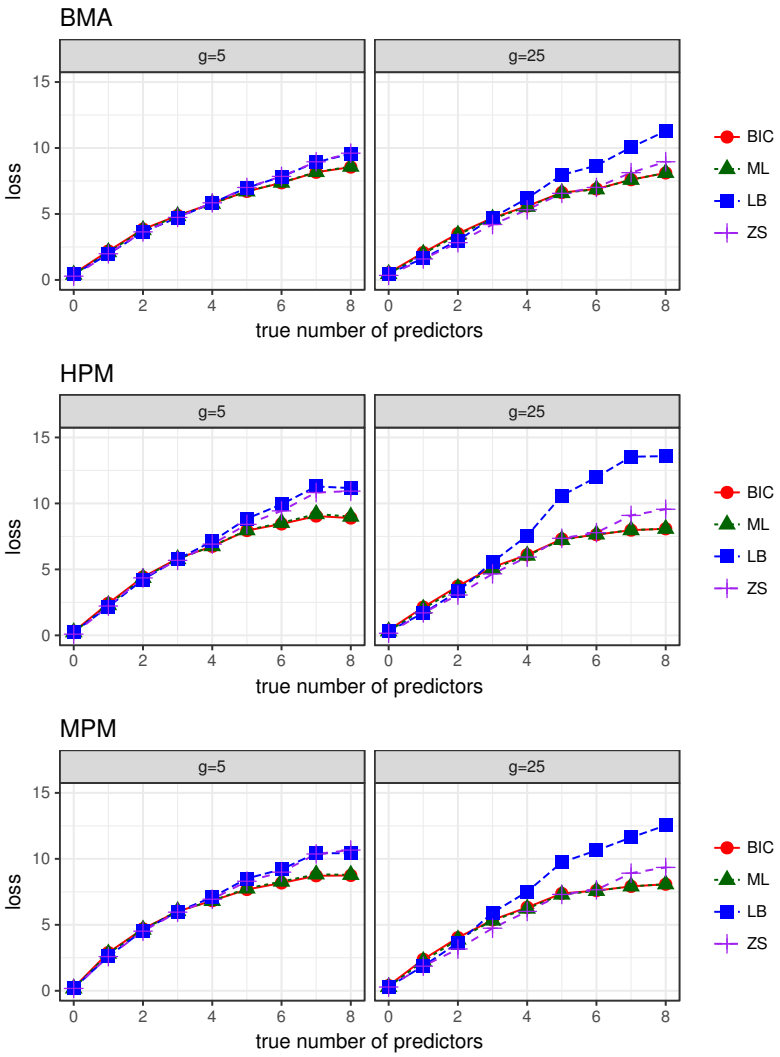


FIGURE 2.3: Simulation study: comparison between highest probability model (HPM), median probability model (MPM), and coefficients after model averaging (BMA) with BIC, the lower bound of the type II ML prior (LB), the type II ML prior (ML), and ZS. AR(1) correlation structure in design matrix.

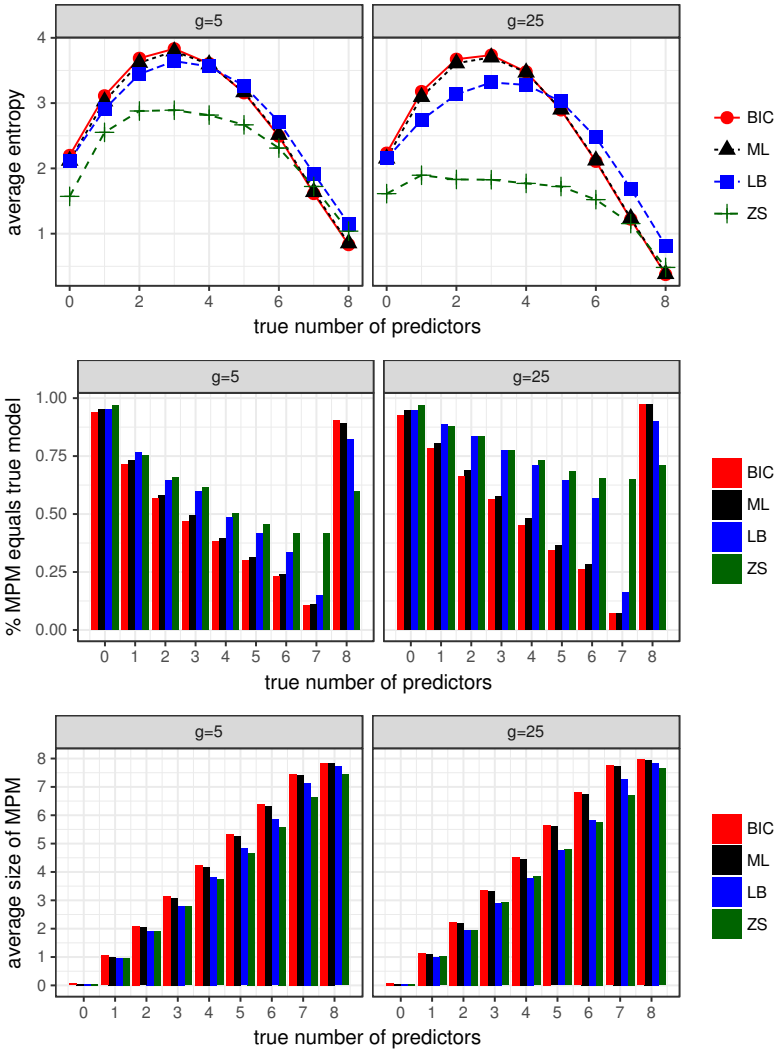


FIGURE 2.4: First panel: Average entropy of the posterior distribution over the model space; second panel: % of times the median probability model equals the true model; third panel: average size of median probability model. AR(1) correlation structure in design matrix.

# 3

# On information consistency

This chapter contains results that are part of Mulder et al. (2017), of which I am third author. The original article has been edited, and the sections to which I have not contributed are not included. The discussion on the different definitions of information consistency that coexist in the literature does not appear in the original article.

## 3.1 Introduction

Prior choice in Bayesian hypothesis testing is a delicate issue. This is well-documented since at least the work of Jeffreys in the monograph Jeffreys (1939), which identifies desirable properties that Bayesian hypothesis tests should satisfy and recommends specific priors that satisfies them.

The term "information inconsistency" was coined in Liang et al. (2008), but the concept behind it already appears in Jeffreys (1939). Information inconsistency occurs when, for a fixed sample size, there is overwhelming evidence in favor of a hypothesis but the Bayes factor in its favor does not go to infinity. Jeffreys' motivating

example is testing whether a normal mean is 0 (null hypothesis) or not (alternative hypothesis) when the population variance is unknown and has an improper prior $p(\sigma^2) \propto \sigma^{-2}$. Quoting Jeffreys' "exact agreement of even two observations would be interpreted as an indication that $\sigma = 0$ and therefore $\lambda = \overline{x} \neq 0$" [$\sigma$ and $\lambda$ are the population standard deviation and mean, respectively, and $\overline{x}$ is the sample mean]. Then, Jeffreys finds the condition that the prior on $\lambda/\sigma$ must satisfy so that the Bayes factor in favor of the alternative goes to 0, and proposes a Cauchy prior as "the simplest function" that satisfies it.

Information consistency is formalized in Liang et al. (2008) and Bayarri et al. (2012) in a slightly different manner than in Jeffreys (1939): it is said to occur when the (maximum) likelihood ratio in favor of a hypothesis goes to infinity and the Bayes factor in its favor is finite. In this chapter, we use a third definition that was introduced in Som et al. (2016). In our recurring example, it is equivalent to considering sequences $y_k = 1_n\mu_k + \epsilon$ with a fixed error term $\epsilon$ and $\mu_k \to \infty$.

Our motivation behind using population parameters instead of maximum likelihood estimates is that a definition involving the former is more likely to be compelling to both objective and subjective Bayesians. While objective Bayesians are generally interested in finding cases where frequentist and Bayesian tests disagree, it is not obvious that a subjective Bayesian should care about agreeing with a frequentist likelihood ratio (or find any statements regarding maximum likelihood estimates relevant). However, if one of the models is true and the "distance" between the truth and the other models grows to infinity, it seems natural that we should be able to pick the true model unequivocally. Also, we do not consider sequences where $\epsilon$ goes to 0 because, in such case, there is no uncertainty and a hypothesis test is not necessary.

The chapter is organized as follows. First, we define the normal linear model with dependent errors and introduce notation (Section 3.2). In Sections 3.3 and 3.4 we cover information consistency in point-null and one-sided hypothesis tests,

respectively. In Section 3.5, we compare the 3 definitions that have been introduced here. Finally, conclusions and recommendations are given in Section 3.6. All proofs are relegated to Appendix B.

## 3.2 Normal linear model with dependent errors

Consider the normal linear model with dependent errors,

$$y = X\beta + \epsilon, \text{ with } \epsilon \sim N_n(0_n, \sigma^2\Sigma), \tag{3.1}$$

where $y$ is the outcome, $X$ is an $n \times K$ design matrix ($n > K$), and $\epsilon$ is the error term, which is normally distributed with 0 mean and and covariance matrix $\sigma^2\Sigma$, where $\Sigma$ is known.

In Section 3.3, we consider point-null hypothesis tests estimable for linear combinations on $\beta$. That is, $H_0 : R\beta = 0_{r_1}$ against $H_1 : R\beta \neq 0_{r_1}$, where $R$ is an $r_1 \times K$ matrix with known constants ($r_1 \leq K$). In Section 3.4 we cover one-sided hypotheses of the type $H_0 : R\beta \leq 0$ against $H_1 : R\beta \not\leq 0$, where "$\not\leq$" implies that at least one componentwise inequality goes in the other direction.

We reparametrize the model so that $\theta = R\beta$ is null-orthogonal to a $r_2 = K - p_1$ common parameter $\gamma = D\beta$ (the joint Fisher information matrix for $\beta$ and $\gamma$ is block-diagonal). We rewrite

$$\begin{bmatrix} \theta \\ \gamma \end{bmatrix} = \begin{bmatrix} R \\ D \end{bmatrix} \beta = T\beta,$$

where $D$ contains $r_2 = K - r_1$ independent rows of $P_R^{\perp} X'\Sigma^{-1}X$, where $P_R^{\perp} = I_K - R'\left(RR'\right)^{-1} R$ is the perpendicular projection operator onto the rowspace of $R$. We rewrite the model as

$$y = X_\theta\theta + X_\gamma\gamma + \epsilon,$$

where $X_\theta$ contains the first $r_1$ columns of $XT^{-1}$ that are regressed on $\theta$ and $X_\gamma$ contains the remaining $r_2 = K - r_1$ columns of $XT^{-1}$ that are regressed on $\gamma$.

The point-null hypothesis test becomes $H_0 : \theta = 0$ against $H_1 : \theta \in \mathbb{R}^{r_1}$, while the one-sided test can be written as $H_0 : \theta \leq 0$ versus $H_1 : \theta \not\leq 0$. Under this parametrization, the maximum likelihood estimators $\widehat{\theta} = (X_\theta' \Sigma^{-1} X_\theta)^{-1} X_\theta' \Sigma^{-1} y$ and $\widehat{\gamma} = (X_\gamma' \Sigma^{-1} X_\gamma)^{-1} X_\gamma' \Sigma^{-1} y$ are independent.

## 3.3 Point-null hypothesis testing

In this section, we cover point-null hypothesis tests for conjugate priors, mixtures of conjugate priors, and empirical Bayes procedures, which estimate the prior scale from the data.

As we mentioned in the introduction, we favor a definition of information inconsistency which is stated in terms of population parameters, in contrast with the definitions in Jeffreys (1939) and Bayarri et al. (2012), which are stated in terms of maximum likelihood estimators. Our preferred definition was introduced in Som et al. (2016) for a "conditional" version of information inconsistency, where the true norm of one block of predictors goes to infinity and the norm of another stays constant.

**Definition 1.** *The Bayes factor of $H_1 : \theta \neq 0_{r_1}$ against $H_0 : \theta = 0_{r_1}$ is information inconsistent if the limit as $\|\theta\| \to \infty$ for a fixed design matrix $X$, nuisance parameter $\gamma$, and error term $\epsilon$ is finite.*

The lemma below shows that, under this asymptotic regime, the residual sum of squares under the alternative stays constant, while $\|\widehat{\theta}\|^2$ and the explained sum of squares of the alternative model go to $\infty$.

**Lemma 1.** *Let $X, \gamma$, and $\epsilon$ be fixed and take the limit as $\|\theta\| \to \infty$. Then,*

$$s_y^2 = (y - X_\gamma\widehat{\gamma} - X_\theta\widehat{\theta})'\Sigma^{-1}(y - X_\gamma\widehat{\gamma} - X_\theta\widehat{\theta}) \text{ fixed.}$$

$$\|\widehat{\theta}\|^2 = \|(X_\theta\Sigma^{-1}X_\theta')^{-1}X_\theta'\Sigma^{-1}y\|^2 \to \infty$$

$$\widehat{\theta}'X_\theta'\Sigma^{-1}X_\theta\widehat{\theta} \to \infty$$

*3.3.1 Conjugate priors*

Consider the following conjugate priors under the alternative ($\pi_1$) and the null ($\pi_0$), respectively:

$$\begin{aligned}
\pi_1(\theta, \gamma_1, \sigma_1^2) &= \pi_1(\theta \mid \sigma_1^2) \times \pi_1(\gamma_1) \times \pi_1(\sigma_1^2) \\
&\propto N_{r_1}(\theta \mid 0_{r_1}, \sigma_1^2\Omega) \times 1 \times \text{InvGam}(\sigma^2 \mid \nu_1/2, \nu_1 s_1^2/2) \qquad (3.2) \\
\pi_0(\gamma_0, \sigma_0^2) &= \pi_0(\gamma_0) \times \pi_0(\sigma_0^2) \\
&\propto 1 \times \text{InvGam}(\sigma^2 \mid \nu_0/2, \nu_0 s_0^2/2)\,. \qquad (3.3)
\end{aligned}$$

This prior specification includes the right-Haar prior $\pi(\sigma^2) \propto \sigma^{-2}$ as special case by letting $\nu_1 \to 0$, and Zellner's $g$-priors by setting $\Omega = g\,(X_\theta'\Sigma^{-1}X_\theta)^{-1}$. The Bayes factor of $H_1$ against $H_0$ is

$$B_{10} = C_1 \times \frac{\left(s_1^2\nu_1 + s_y^2 + \widehat{\theta}'\left((X_\theta'\Sigma^{-1}X_\theta)^{-1} + \Omega\right)^{-1}\widehat{\theta}\right)^{-(n+\nu_1-r_2)/2}}{\left(s_0^2\nu_0 + s_y^2 + \widehat{\theta}'X_\theta'\Sigma^{-1}X_\theta\widehat{\theta}\right)^{-(n+\nu_0-r_2)/2}}\,, \qquad (3.4)$$

where the constant is

$$C_1 = \frac{(\nu_1/2)^{\nu_1/2}s_1^{\nu_1}\Gamma\left(\frac{\nu_0}{2}\right)\Gamma\left(\frac{n+\nu_1-r_2}{2}\right)}{(\nu_0/2)^{\nu_0/2}s_0^{\nu_0}\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{n+\nu_0-r_2}{2}\right)}2^{(\nu_1-\nu_0)/2}|\Omega + (X_\theta'\Sigma^{-1}X_\theta)^{-1}|^{-\frac{1}{2}}|X_\theta'\Sigma^{-1}X_\theta|^{-\frac{1}{2}}\,.$$

The lemma below shows that the Bayes factor is information inconsistent unless $\nu_0 < \nu_1$ for any choice of $\Omega$, and it is a generalization of the well-known results for $g-$priors that appear in Berger and Pericchi (2001) and Liang et al. (2008).

**Lemma 2.** *The limit of the Bayes factor in* (3.4) *as* $\|\theta\| \to \infty$ *with* $X$ *and* $\sigma^2$ *fixed is* 0 *if* $\nu_0 < \nu_1$*;* $\infty$ *if* $\nu_0 > \nu_1$*; and, if* $\nu_0 = \nu_1$*,*

$$B_{10} \leq C_1 \left( \limsup_{|\widehat{\theta}| \to \infty} \frac{\widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta}}{\widehat{\theta}' \left( (X_\theta' \Sigma^{-1} X_\theta)^{-1} + \Omega \right)^{-1} \widehat{\theta}} \right)^{\frac{(n+\nu-r_2)}{2}} = C_1 \ (1 + \lambda_{\max})^{(n+\nu-r_2)/2} < \infty \,,$$

*where* $\lambda_{\max}$ *is the largest eigenvalue of* $X_\theta' \Sigma^{-1} X_\theta \Omega$*.*

Interestingly, when $\nu_0 < \nu_1$, the Bayes factor concludes that the null hypothesis is unequivocally true. This is in stark disagreement with the frequentist likelihood ratio of $H_1$ against $H_0$, which goes to infinity as $\|\theta\| \to \infty$.

To gain some intuition on how problematic the limiting Bayes factors can be in practice and gauge the effect of the correlation between observations, we study the behavior of the Bayes factor for a one-sample normal means test with correlated errors in the example below.

**Example 2.** *(One sample test with correlated errors) We want to test* $H_0 : \theta = 0$ *against* $H_1 : \theta \neq 0$ *with correlated observations. Specifically, let* $r_1 = 1$*,* $r_2 = 0$*,* $X_\theta = 1_n$*, and* $\Sigma$ *be the correlation matrix with identical correlations* $\rho$ *in the off-diagonal elements. Our choice of hyperparameters is* $\nu_0 = \nu_1 = 0$ *and* $\Omega = 1$*. The Bayes factor in* (3.4) *can then be written as a function of the t-statistic* $t = \frac{\widehat{\theta}\sqrt{1_n' \Sigma^{-1} 1_n}}{s_y/\sqrt{n-1}}$*:*

$$B_{10} = \left( 1 + \frac{n}{1 + (n-1)\rho} \right)^{-1/2} \left( 1 - \frac{nt^2}{[t^2 + n - 1][n + 1 + (n-1)\rho]} \right)^{-n/2} .$$

*The limit of the Bayes factor as* $\theta$ *goes to infinity, which implies* $|t| \to \infty$*, is*

$$\begin{aligned} \lim_{\theta \to \infty} B_{10} &= \left( 1 + \frac{n}{1 + (n-1)\rho} \right)^{-1/2} \left( 1 - \frac{n}{n + 1 + (n-1)\rho} \right)^{-n/2} \\ &= \begin{cases} (1+n)^{(n-1)/2}, & \text{if } \ \rho = 0; \\ \left(1 + \frac{2n}{n+1}\right)^{-1/2} \left(\frac{3n+1}{n+1}\right)^{n/2} \approx 3^{(n-1)/2}, & \text{if } \ \rho = 0.5; \\ 2^{(n-1)/2}, & \text{if } \ \rho = 1. \end{cases} \end{aligned}$$

*Table 3.1 shows limiting Bayes factors as $|t| \to \infty$ for different values of $\rho$ and sample sizes ranging from $n = 2$ to $n = 20$. When $\rho = 0$, the limit $(n+1)^{(n-1)/2}$ is large for $n \geq 6$, so that information inconsistency is not problematic in practice. On the other hand, the limit can be relatively small when $\rho \approx 1$.*

Table 3.1: Limiting values of the Bayes factor for a univariate normal means test as $|t| \to \infty$ for different sample sizes $n$ and correlations $\rho$.

| | $n = 2$ | $n = 5$ | $n = 7$ | $n = 10$ | $n = 20$ |
|---|---|---|---|---|---|
| $\rho = 0$ | 1.73 | 36 | 512 | $4.85 \times 10^4$ | $1.79 \times 10^{11}$ |
| $\rho = 0.5$ | 1.53 | 7.10 | 20.8 | 106 | $2.01 \times 10^4$ |
| $\rho \approx 1$ | 1.41 | 4 | 8 | 22.6 | 724 |

### *3.3.2 Mixtures of conjugate priors*

It has long been argued (starting with Jeffreys (1939)) that priors that that have thicker tails than conjugate priors should be used for testing. A popular choice is the class of scale mixtures of conjugate priors, which results in information consistent Bayes factors if the prior on $g$ is thick enough. This is shown in the following lemmas, which generalize the result in Liang et al. (2008) for $\nu_0 = \nu_1 = 0$, $\Sigma = I$, and $\Omega = g(X_\theta' \Sigma^{-1} X_\theta)^{-1}$.

**Lemma 3.** *Let $\theta \mid g, \gamma_1, \sigma_1^2 \sim N_{r_1}(0_{r_1}, g\sigma_1^2 A)$, where $g$ has a prior with density $\pi(g)$. The prior on $\sigma^2$ and $\gamma$ is the same as in Equations 3.2 and 3.3. If $\nu_0 > \nu_1$, any $\pi(g)$ with positive support yields an information-consistent $B_{10}$. If $\nu_0 = \nu_1$, the condition*

$$\int_0^\infty (g+1)^{(n-r_1-r_2+\nu_1)/2} \pi(g) \, \mathrm{d}g = \infty$$

*is necessary and sufficient for information consistency, and it is necessary whenever $\nu_0 < \nu_1$.*

The maximum number of finite moments that the prior on $g$ can have to achieve information consistency increases with the sample size $n$ and decreases with the

number of predictors $K = r_1 + r_2$. Lemma 3 gives necessary and sufficient conditions for information consistency for $\nu_0 \geq \nu_1$, but only gives us a necessary condition for information consistency for $\nu_0 < \nu_1$. The lemma below characterizes the behavior of polynomial-tailed priors on $g$ in this latter case, and provides partial results for priors with thinner- and thicker-than-polynomial priors on $g$.

**Lemma 4.** *Suppose $\nu_0 < \nu_1$ and let $\theta \mid g, \gamma_1, \sigma_1^2 \sim N_{r_1}(0_{r_1}, g\sigma_1^2 A)$, where $g$ has a prior with density $\pi(g)$. The prior on $\sigma^2$ and $\gamma$ is the same as in Equations 3.2 and 3.3. Then, the following are true:*

1. *If there exist $0 < M < \infty$ and $0 < K < \infty$ such that for all $g \geq M$, $\pi(g) \geq Kg^{-\alpha}$ for $\alpha > 1$, $B_{10}$ is information consistent whenever $\alpha < (n - r_1 - r_2 + \nu_0)/2 + 1$.*

2. *If there exist $0 < M' < \infty$ and $0 < K' < \infty$ such that for all $g \geq M'$, $\pi(g) \leq K'g^{-\alpha}$ for $\alpha > 1$, $B_{10}$ is information inconsistent whenever $\alpha \geq (n - r_1 - r_2 + \nu_0)/2 + 1$.*

Both the Zellner-Siow prior (Zellner and Siow, 1980) and hyper-$g$ priors (Liang et al., 2008) satisfy the conditions of Lemma 4 because they have polynomial tails.

### *3.3.3 Empirical Bayes approaches*

Another approach to Bayesian hypothesis testing is empirical Bayes, where some hyperparameters are estimated from the data (some examples are George and Foster (2000), Hansen and Yu (2001), and Chapter 2 of this dissertation). The following lemmas show that some empirical Bayes approaches are successful in avoiding information inconsistency, even when they are based on normal priors that are inconsistent when their scale parameters are fixed.

The following lemma generalizes the result in Liang et al. (2008) for $\nu_0 = \nu_1 = 0$, $\Sigma = I$, and $\Omega = g(X_\theta' \Sigma^{-1} X_\theta)^{-1}$.

**Lemma 5.** *Let $\theta \mid g, \gamma_1, \sigma_1^2 \sim N_{r_1}(0_{r_1}, g\sigma_1^2 A)$. The prior on $\sigma^2$ and $\gamma$ is the same as in Equations 3.2 and 3.3. If $g > 0$ is set by maximizing $B_{10}$, information consistency holds.*

Lemma 5 establishes information consistency for all $\nu_0$ and $\nu_1$. This is in contrast with the results in previous sections, where the behavior of $B_{10}$ depends (sometimes rather strongly) on $\nu_0$ and $\nu_1$.

We conclude this subsection with a lemma that shows that the restricted empirical Bayes procedure described in Section 2.2 of Chapter 2 is also information information consistent for any $\nu_0$ and $\nu_1$.

**Lemma 6.** *Let $\theta \mid g, \gamma_1, \sigma_1^2 \sim N_{r_1}(0_{r_1}, \sigma_1^2 A)$, where $A$ is set by maximizing $B_{10}$ subject to $A \succeq n(X_\theta' \Sigma^{-1} X_\theta)^{-1}$. The prior on $\sigma^2$ and $\gamma$ is the same as in Equations 3.2 and 3.3. Then, the Bayes factor is information consistent.*

## 3.4 One-sided hypothesis testing

In this section, we study information inconsistency for one-sided hypothesis tests of the type $H_0 : \theta \leq 0_{r_1}$ against $H_1 : \theta \not\leq 0_{r_1}$. As we did in the previous section, we consider conjugate priors, mixtures of conjugate priors, and empirical Bayes approaches. We use the following definition of information consistency, which is very similar to the one in Section 3.3.

**Definition 2.** *The Bayes factor of $H_0 : \theta \leq 0_{r_1}$ against $H_1 : \theta \not\leq 0_{r_1}$ is information inconsistent if the limit as $\|\theta\| \to \infty$ with at least one coordinate of $\theta$ going to $\infty$ is finite, or if all the coordinates of $\theta$ go to $-\infty$ and the Bayes factor is greater than 0.*

### *3.4.1 Conjugate prior*

We work with encompassing priors, as in Berger and Mortera (1999) and Klugkist and Hoijtink (2007). The starting point is a prior centered at the boundary between

the subspaces of interest, which in this case is $0_{r_1}$ :

$$\pi(\theta, \gamma, \sigma^2) \propto N_{r_1}(\theta \mid 0_{r_1}, \sigma^2\Omega) \times \text{InvGam}(\sigma^2 \mid \nu/2, \nu s^2/2), \tag{3.5}$$

with a flat improper prior for $\gamma$. The priors for $H_0$ and $H_1$ are found by conditioning the encompasisng prior on the appropriate subspaces. That is, for $t \in \{0, 1\}$, $\Theta_0 = \{\theta \leq 0\}$, and $\Theta_1 = \{\theta \not\leq 0\}$ :

$$\pi_t(\theta \mid \sigma^2) \;\; = \;\; \pi(\theta \mid \sigma^2) I_{\Theta_t}(\theta) / P_\pi(\theta \in \Theta_t \mid \sigma^2), \tag{3.6}$$

The priors on $\sigma^2$ and $\gamma$ are the same under $H_0$ and $H_1$ and equal to their priors under the encompassing prior: $\pi_t(\sigma^2) = \pi(\sigma^2)$ and $\pi_t(\gamma) = \pi(\gamma)$.

Under this prior specification, the Bayes factor for the one-sided hypothesis test can be written as

$$B_{10} = \left(P_\pi(\theta \leq 0 \mid \sigma^2 = 1)^{-1} - 1\right)^{-1} \left(P_\pi(\theta \leq 0 \mid y)^{-1} - 1\right). \tag{3.7}$$

The derivation is similar to that in Mulder (2014). This expression shows that the Bayes factor for a one-sided hypothesis test is information consistent if and only if $P_\pi(\theta \leq 0 \mid y) \to 0$, and we use this fact for proving the results in this subsection.

The lemma below shows that the Bayes factor is information inconsistent under the encompassing conjugate prior.

**Lemma 7.** *$P_\pi(\theta \leq 0 \mid y)$ is bounded away from 0 and 1 for all $y$. Hence $B_{10}$ is information inconsistent.*

*If $\widehat{\theta} = cv$ and $c \to \infty$, then*

$$P_\pi(\theta \leq 0 \mid y) \to P_\pi(\xi \leq 0 \mid y),$$

*where $\xi$ has a multivariate t distribution with mean*

$$v^* = \frac{(X_\theta' \Sigma^{-1} X_\theta + \Omega^{-1})^{-1} X_\theta' \Sigma^{-1} X_\theta v}{(n + \nu - r_2)^{-1/2} (v'((X_\theta' \Sigma^{-1} X_\theta)^{-1} + \Omega)^{-1} v)^{1/2}},$$

*scale matrix $(X_\theta' \Sigma^{-1} X_\theta + \Omega^{-1})^{-1}$, and $n + \nu - r_2$ degrees of freedom.*

**Example 3.** *We return to the problem of testing a normal mean with correlated errors. The hypotheses are* $H_0 : \theta \leq 0$ *and* $H_1 : \theta > 0$*, with* $\nu = 0$*,* $r_1 = 1$*,* $r_2 = 0$*,* $X_\theta = 1$*,* $\Omega = 1$*, and* $\Sigma = \rho J_n + (1-\rho) I_n$*, so that* $P_\pi(\theta \leq 0 \mid \sigma^2) = \frac{1}{2}$*. By applying Lemma 7, we find that the Bayes factor is*

$$\begin{aligned} B_{10} &= T_n\left(-\sqrt{\frac{n^2}{1+(n-1)\rho + t^{-2}(n-1)(1+n+(n-1)\rho)}}\right)^{-1} - 1 \\ &\to T_n\left(-n(1+(n-1)\rho)^{-\frac{1}{2}}\right)^{-1} - 1. \end{aligned} \tag{3.8}$$

*Again,* $t$ *is the* $t-$*statistic, which goes to infinity as* $\theta \to \infty$*, and* $T_\nu(\cdot)$ *is the cumulative distribution function of a Student-t distribution with* $\nu$ *degrees of freedom. Note that as* $t \to -\infty$*,* $B_{10}$ *converges to the reciprocal of* (3.8)*.*

*Table 3.2 provides the limiting values of the Bayes factors for different values of* $n$ *and* $\rho$*. When comparing Table 3.2 with Table 3.1, it is clear that the practical importance of information inconsistency for one-sided hypothesis testing is less problematic than in null hypothesis tests.*

Table 3.2: Limiting values of the Bayes factor for a one-sided univariate normal mean test for different sample sizes $n$ and correlations $\rho$.

| | $n=2$ | $n=5$ | $n=7$ | $n=10$ | $n=20$ |
|---|---|---|---|---|---|
| $\rho = 0$ | 9.90 | 486 | $9.45 \times 10^3$ | $1.26 \times 10^6$ | $1.85 \times 10^{14}$ |
| $\rho = 0.5$ | 7.19 | 57.2 | 199 | $1.21 \times 10^3$ | $4.02 \times 10^5$ |
| $\rho \approx 1$ | 5.83 | 25.5 | 59.3 | 197 | $8.57 \times 10^4$ |

### *3.4.2 Mixtures of conjugate priors*

We provide a necessary and sufficient condition for information consistency for scale mixtures of conjugate priors in a one-sided hypothesis test. The lemma requires extra conditions for general $\Omega$, but not for $\Omega \propto (X_\theta' \Sigma^{-1} X_\theta)^{-1}$, which includes Zellner's $g$-priors and univariate $\theta$.

**Lemma 8.** *Consider the setup in Equations 3.5 and 3.6. Let* $\theta \mid g, \sigma^2 \sim N_{r_1}(0_{r_1}, g\sigma^2 A)$, *where* $g$ *has a prior with density* $\pi(g)$, *and let* $w = \mathbb{E}(\theta \mid g, y)$. *Assume that if there exists* $i$ *such that* $\widehat{\theta}_i \to +\infty$, *there exists* $j$ *such that* $w_j > 0$. *Alternatively, assume that if* $\widehat{\theta}_i \to -\infty$ *for all* $i$, *then* $w_i < 0$ *for all* $i$. *Then, the condition*

$$\int_0^\infty (g+1)^{(n-r_1-r_2+\nu)/2} \pi(g) \mathrm{d}g = \infty$$

*is necessary and sufficient for information consistency.*

### *3.4.3 Empirical Bayes approaches*

In this subsection, we show that an empirical Bayes approach is information consistent. This is in agreement with the positive results for point-null hypothesis testing.

**Lemma 9.** *Consider the setup in Equations 3.5 and 3.6. The Bayes factor based on the g-prior, with* $g_{\max} = \arg\max_g\{B_{01}\}$ *if* $\widehat{\theta} \leq 0$ *and* $g_{\max} = \arg\max_g\{B_{10}\}$ *if* $\widehat{\theta} \not\leq 0$, *is information consistent for one-sided hypothesis testing.*

The choice of $g$ that maximizes the Bayes factor goes to infinity as $\|\theta\| \to \infty$. Letting $g \to \infty$ in this context was already proposed in Mulder (2014) without reference to empirical Bayes.

## 3.5 Definitions of information inconsistency

As we mentioned in the introduction, there are 3 different definitions of information inconsistency in the literature. In this section, we state them and compare them in the context of Section 3.3.

- **Jeffreys':** The Bayes factor $B_{10}$ is information inconsistent if $s_y^2 = 0$ and $\widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} \neq 0$ but $B_{10}$ is finite. This condition is a generalization of the definition in Jeffreys (1939) described in Section 3.1.

- **Likelihood ratio:** The Bayes factor $B_{10}$ is information inconsistent if the (maximum) likelihood ratio of $H_1$ against $H_0$ goes to $\infty$, but $B_{10}$ is finite. In this context, the condition is equivalent to $\widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} \to \infty$. This definition is used in Liang et al. (2008) and Bayarri et al. (2012).

- **Parameter going to $\infty$:** The Bayes factor $B_{10}$ is information consistent if $\|\theta\| \to \infty$ for fixed design matrix $X$ and error term $\epsilon$, but $B_{10}$ is finite. This was introduced in Som et al. (2016) and it is the definition we favor in this chapter.

The condition stated in terms of the likelihood ratio is weaker than the others, in the sense that the likelihood ratio is infinite both when $s_y^2 = 0$ with $\widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} \neq 0$ and under the limit as $\|\theta\| \to \infty$ for fixed $X$ and $\epsilon$ (see Lemma 1).

Consider the Bayes factor in Equation 3.4 and let $s_y^2 = 0$, $\widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} \neq 0$. Then,

$$B_{10} \propto \frac{\left( s_1^2 \nu_1 + \widehat{\theta}' \left( \left( X_\theta' \Sigma^{-1} X_\theta \right)^{-1} + \Omega \right)^{-1} \widehat{\theta} \right)^{-(n+\nu_1-r_2)/2}}{\left( s_0^2 \nu_0 + \widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} \right)^{-(n+\nu_0-r_2)/2}},$$

which is finite for any $s_0^2, \nu_0, s_1^2, \nu_1$. Therefore, information inconsistency always occurs in this context, according to Jeffreys' definition. In general, the limit when the likelihood ratio goes to infinity does not exist. If $\widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} \to \infty$ and $s_y^2$ is bounded away from 0, the limit is equal to the limit stated in Lemma 2. However, if $0 < \widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} < \infty$ and $s_y^2 \to 0$, $B_{10}$ is finite. This shows that the limit does not exist, because in both situations the likelihood ratio goes to infinity, but the limits need not be equal. The limit exists if $\nu_0 = \nu_1 = 0$ (right-Haar prior) and $\Omega \propto (X_\theta \Sigma^{-1} X_\theta)^{-1}$ (this choice includes Zellner's $g$-priors and their mixtures), in which case the Bayes factor depends on the data only through the likelihood ratio.

As argued in Li and Clyde (2015), extending any of these definitions to models that have bounded likelihoods (such as logistic regression) is problematic. The likelihood ratio in that scenario is bounded, so the second definition cannot be used. The first and third definition seem hard to justify. For example, observing 3 successes in 3 independent Bernoulli($\theta$) trials is hardly "overwhelming evidence" in favor of $\theta \neq 1/2$ against $\theta = 0$. However, a naive application Jeffreys' definition would suggest that $B_{10} \to \infty$. The same argument applies to the third definition; if $\theta \to 1$, we would see 3 successes in 3 trials, and a naive application of the definition would recommend $B_{10} \to \infty$.

## 3.6 Conclusions

In normal linear models, information inconsistency under conjugate priors is common in point- and one-sided hypothesis testing. The practical consequences depend on the strength of the correlation between the observations and how small the sample size is, relative to the number of observations and predictors.

Thick-tailed priors and empirical Bayes approaches are typically information-consistent. This is in a agreement with previous results in the literature, such as the ones in Liang et al. (2008) for mixtures of $g-$priors.

We have discussed the implications and differences between existing definitions of information consistency, and favor one that is stated in terms of population parameters and not maximum likelihood estimates or likelihood ratios. This definition avoids technical problems such as nonexistence of limits but, more importantly, we believe that it is more likely to be convincing to non-objective Bayesians. If one truly believes their subjective prior and it turns out to be information inconsistent, there is no logical flaw in using it. However, it seems unlikely that features such as their tail behavior (which is key to information consistency) can be elicited reliably, so subjective Bayesians might also find this criterion useful in choosing their priors.

# 4

# On limit consistency

## 4.1 Introduction

In Chapter 3, we studied information inconsistency, which is a criterion that can help choosing priors for regression coefficients in normal linear models. In this chapter, we turn to another criterion, namely "limit consistency", which was introduced in Chapter 6 of Ly (2017). Ly motivates limit consistency in a hypothesis test where the interest is in testing whether the rates of two homogeneous Poisson processes are equal:

> "Assume that the measurement of the first process is terminated early, whereas the measurement of the second process continuous indefinitely. In the limit, knowledge about the second process will reach perfection, but knowledge about the interrupted process will remain incomplete. Consequently, there exists a bound on the level of evidence that can be obtained in a test that compares the two processes. As measurement for the second process continues, the Bayes factors ought to approach a finite limit."

This requirement can be generalized for comparing any two distributions after introducing some notation. Let $P_{\theta_A}$ and $P_{\theta_B}$ be the distributions of groups A and B, where $\theta_A \in \Theta_A$ and $\theta_B \in \Theta_B$ are unknown and $\Theta_A \cap \Theta_B \neq \emptyset$. Throughout, we assume that the statistical models are well-specified and that the true values of the parameters, $\theta_A^*$ and $\theta_B^*$, can be estimated consistently. Suppose we are interested in testing $H_0 : P_{\theta_A} = P_{\theta_B}$ against $H_1 : P_{\theta_A} \neq P_{\theta_B}$ after collecting $n_A$ observations $y_A = \{y_{1A}, y_{2A}, \ldots, y_{n_A,A}\}$ from group A and $n_B$ observations $y_B = \{y_{1B}, y_{2B}, \ldots, y_{n_B,B}\}$ from group B. The joint density of the data (with respect to an appropriate dominating measure) under $H_0$ and $H_1$ are denoted $p_0(y_A, y_B \mid \theta)$ and $p_1(y_A, y_B \mid \theta_A, \theta_B)$. Lastly, let $\pi_0(\mathrm{d}\theta)$ and $\pi_1(\mathrm{d}\theta_{A,B})$ be the prior probability measures under $H_0$ and $H_1$, which have support $\Theta_A \cap \Theta_B$ and $(\Theta_A \cap \Theta_B)^c$, respectively. The Bayes factor of $H_1$ to $H_0$ is defined as

$$B_{10} = \frac{\int_{\theta_A \cap \theta_B} p(y_A, y_B \mid \theta_{1:2}) \pi_1(\mathrm{d}\theta_{1:2})}{\int_{(\theta_A \cap \theta_B)^c} p(y_A, y_B \mid \theta) \pi_0(\mathrm{d}\theta)} = \frac{m_1}{m_0},$$

where $m_1$ and $m_0$ are the marginal distributions under $H_1$ and $H_0$. Using this notation, the criterion can be formalized as follows.

**Definition 1.** *Consider a scenario where $y_A$ is fixed and the sample size of group B ($n_B$) goes to infinity. Then, $B_{10}$ is limit consistent for testing $H_0 : P_{\theta_A} = P_{\theta_B}$ against $H_1 : P_{\theta_A} \neq P_{\theta_B}$ if it converges to a finite limit $B_{10}^*$ as $n_B \to \infty$.*

While we agree with Ly that the limiting Bayes factor should be finite, we believe that more should be true: as $n_B \to \infty$, the Bayes factor of the two-sample test should converge to the Bayes factor of a one-sample test for testing $H_0 : P_{\theta_A} = P_{\theta_B^*}$ against $H_1 : P_{\theta_A} \neq P_{\theta_2^*}$. Moreover, the limiting Bayes factor should arise under an appealing prior specification for the one-sample problem.

Another problem with Definition 1 is that it only applies to Bayes factors, but the general idea could be applied more generally to decision rules (that are not

necessarily Bayesian). This motivates our defining limit consistency in terms of the decision rules of the one-sample and two-sample tests.

**Definition 2.** *Consider the following hypothesis tests:*

1. *One-sample test:* $H_{01} : P_{\theta_A} = P_{\theta_B^*}$ *against* $H_{11} : P_{\theta_A} \neq P_{\theta_B^*}$ *with* $P_{\theta_B^*}$ *known.*

2. *Two-sample test:* $H_{02} : P_{\theta_A} = P_{\theta_B}$ *against* $H_{12} : P_{\theta_A} \neq P_{\theta_B}$ *with* $\theta_A$ *and* $\theta_B$ *unknown.*

*Let* $\delta_1$ *be the decision rule for the one-sample test and* $\delta_2$ *for the decision rule for the two-sample test. If* $y_A$ *is fixed and* $n_B \to \infty$, $\delta_1$ *and* $\delta_2$ *are limit consistent if* $\delta_2 \to \delta_*$ *such that* $\delta_* = H_{h2}$ *if and only if* $\delta_1 = H_{h1}$ *for* $h \in \{0, 1\}$

The chapter is organized as follows: in Section 4.2, we introduce the notation and decision-theoretical framework that we use throughout the chapter. Then, Section 4.3 studies limit consistency of Bayesian two-sample normal means tests under different parametrizations and prior specifications. Finally, Section 4.4 closes the chapter with conclusions and pointers to future work.

## 4.2 Notation and preliminaries

In this chapter, we restrict our attention to testing 2 normal means with known variance. That is, the data are $y_{ij} \sim N_1(\mu_j, \sigma^2)$ independently with $\sigma^2$ known, where $j \in \{A, B\}$ indexes group membership and $i \in \{1, 2, ..., n_j\}$ indexes observations within groups. The total sample size is denoted $n = n_A + n_B$, the sample means of groups A and B are denoted $\overline{y}_A$ and $\overline{y}_B$, and we use the notation $\kappa = n_A n_B / n$. The overall mean is denoted $\overline{y} = (n_A \overline{y}_A + n_B \overline{y}_B)/(n_A + n_B)$.

We compare the Bayes decisions of the two-sample test $H_{02} : \mu_A = \mu_B$ against $H_{12} : \mu_A \neq \mu_B$ with the Bayes decisions of the analogous one-sample test where

$\mu_B = \mu_B^*$ is known. More precisely, we study the behavior of Bayes decisions under the losses $L_j$ ($j = 1$ is the one-sample test and $j = 2$ the two-sample test):

$$L_j(\mu_A, \mu_B, H_0) = \mathbb{1}(\mu_A \neq \mu_B)\gamma_1 f(|\mu_A - \mu_B|); \quad L_j(\mu_A, \mu_B, H_1) = \mathbb{1}(\mu_A = \mu_B)\gamma_0,$$

where $\gamma_0, \gamma_1, f(|\mu_A - \mu_B|) > 0$. It should be understood that, in $L_1$, $\mu_B$ is replaced by the known value $\mu_B^*$. We assume that the prior probabilities of the hypotheses are the same for the two-sample and one-sample problem, which we denote $\mathbb{P}_1(H_{01}) = \mathbb{P}_2(H_{02}) = \pi_0$. This framework includes decisions under $\{0, 1\}$ loss by letting $\gamma_0 = \gamma_1$ and $f(|\mu_A - \mu_B|) = 1$ but other loss functions are possible, such as adaptations of the losses studied in Robert and Casella (1994), or losses based on divergences between distributions as in Robert (1996). Given this framework, the Bayes decisions are

$$\delta_j^\pi = H_{0j}\, \mathbb{1}\left(\mathbb{P}_j(H_0 \mid y) \geq \frac{\gamma_1 \overline{f}_j}{\gamma_0 + \gamma_1 \overline{f}_j}\right) + H_{1j}\, \mathbb{1}\left(\mathbb{P}_j(H_0 \mid y) < \frac{\gamma_1 \overline{f}_j}{\gamma_0 + \gamma_1 \overline{f}_j}\right) H_{1j},$$

where $\overline{f}_1 = \mathbb{E}_1[f_1(|\mu_A - \mu_B|) \mid H_{11}, y]$ is the posterior expectation of $f(|\mu_A - \mu_B|)$ for the one-sample problem and $\overline{f}_2$ the posterior expectation of $f(|\mu_A - \mu_B|)$ for the two-sample problem. The posterior probabilities are $\mathbb{P}_j(H_{0j} \mid y) = [1 + (1-\pi_0)B_{10,j}/\pi_0]^{-1}$, which depend on the data only through the Bayes factors $B_{10,j}$:

$$B_{10,1} = \frac{\int N_1(\overline{y}_A \mid \mu_A, \sigma^2/n_A)\pi_1(\mu_A \mid H_{11})\, \mathrm{d}\mu_A}{N_1(\overline{y}_A \mid \mu_B^*, \sigma^2/n_A)} = \frac{m_{11}}{m_{01}}$$

$$B_{10,2} = \frac{\int N_1(\overline{y}_A \mid \mu_A, \sigma^2/n_A) N_1(\overline{y}_B \mid \mu_B, \sigma^2/n_B)\pi_2(\mu_A, \mu_B \mid H_{12})\, \mathrm{d}(\mu_A, \mu_B)}{\int N_1(\overline{y}_A \mid \mu, \sigma^2/n_A) N_1(\overline{y}_B \mid \mu, \sigma^2/n_B)\pi_2(\mu \mid H_{02})\, \mathrm{d}\mu} = \frac{m_{12}}{m_{02}}.$$

In Section 4.3 we reparametrize the problem in different ways, and the expressions above will be recast in terms of the appropriate parametrizations as needed.

## 4.3 Two-sample tests

Before we turn to the discussion about limit consistency of Bayesian tests, we note that the two-sample and one-sample $z-$tests are limit-consistent according to Definition 2. Indeed, let $Z_n = \sqrt{\kappa}(\overline{y}_A - \overline{y}_B)/\sigma$, $Z_* = \sqrt{n_A}(\overline{y}_A - \mu_B^*)/\sigma$, $z_\alpha/2$ be the

$(1-\alpha/2)\%$ quantile of a $N_1(0,1)$. Then, the decision rules $\delta_2 = H_{02}\,\mathbb{1}\{Z_n^2 \leq z_{\alpha/2}^2\} + H_{12}\,\mathbb{1}\{Z_n^2 > z_{\alpha/2}^2\}$ and $\delta_1 = H_{01}\,\mathbb{1}\{Z_*^2 \leq z_{\alpha/2}^2\} + H_{11}\,\mathbb{1}\{Z_*^2 > z_{\alpha/2}^2\}$ are limit-consistent as $n_B \to \infty$ because $Z_n^2 \to_d Z_*^2$ and the indicators converge.

### *4.3.1 Independent priors*

Suppose that our priors for the two-sample test are

$$\pi_2(\mu \mid H_{02}) = N_1(\mu \mid 0, \sigma^2/\omega),$$
$$\pi_2(\mu_A, \mu_B \mid H_{12}) = N_1(\mu_A \mid 0, \sigma^2/\omega_A)\, N_1(\mu_B \mid 0, \sigma^2/\omega_B).$$

The Bayes factor of $H_{12} : \mu_A \neq \mu_B$ to $H_{02} : \mu_A = \mu_B$ is

$$B_{10,2} = \left[\frac{\omega_A\omega_B(\omega+n)}{(\omega_A+n_A)(\omega_B+n_B)\omega}\right]^{1/2} \exp\left\{\frac{1}{2\sigma^2}\left[\frac{n_A^2\overline{y}_A^2}{n_A+\omega_A} + \frac{n_B^2\overline{y}_B^2}{n_B+\omega_B} - \frac{n^2\overline{y}^2}{n+\omega}\right]\right\}.$$

As $n_B \to \infty$,

$$B_{10,2} \to_d B_l = \left(\frac{\omega}{\omega_B} + \frac{\omega n_A}{\omega_A\omega_B}\right)^{-1/2} \exp\left\{\frac{1}{2}\left[Z_*^2 + \frac{(\omega-\omega_B)(\mu_B^*)^2}{\sigma^2} - \frac{n_A\omega_A\overline{y}_A^2}{\sigma^2(n_A+\omega_A)}\right]\right\}.$$

In the one-sample test $H_{11} : \mu_A \neq \mu_B^*$ against $H_{02} : \mu_A = \mu_B^*$ , the Bayes factor can be written as

$$B_{10,1} = \frac{\int N_1(\overline{y}_A \mid \mu_A, \sigma^2/n_A)\pi_1(\mu_A \mid H_{11})\mathrm{d}\mu_A}{N_1(\overline{y}_A \mid \mu_B^*, \sigma^2/n_A)}.$$

The integral of the numerator with respect to $\overline{y}_A$ must be equal to 1 because

$$\int N_1(\overline{y}_A \mid \mu_A, \sigma^2/n_A)\pi_1(\mu_A \mid H_{11})\mathrm{d}\mu_A = \int p_1(\overline{y}_A, \mu_A \mid H_{11})\mathrm{d}\mu_A = p_1(\overline{y}_A \mid H_{11}).$$

If $B_l$ were to be $B_{10,1}$, the numerator should be equal to

$$N_1\left(\overline{y}_A \mid \mu_B^*, \frac{\sigma^2}{n_A}\right) B_l = \left(\frac{\omega}{\omega_B}\right)^{-1/2} \exp\left\{\frac{(\omega-\omega_B)(\mu_B^*)^2}{2\sigma^2}\right\} N_1\left(\overline{y}_A \mid 0, \frac{\sigma^2(n_A+\omega_A)}{n_A\omega_A}\right).$$

The integral of the expression above is not equal to 1 unless the product of the terms that are not the normal density are equal to 1. For any $\omega > 0$, there are only 2

choices of $\omega_B > 0$ that satisfy the property. One depends on $\mu_B^*$, so it cannot be picked *a priori*, and the other one is $\omega_B = \omega$. In the latter case, the limit is equal to the Bayes factor under $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid 0, \sigma^2/\omega)$. This prior seems hard to justify as a default choice unless $\mu_B^* = 0$, since it assumes a preference of values of $\mu_A$ near 0 whenever $\mu_A \neq \mu_B^*$. For example, one of the consequences of centering the prior at 0 is that the value of $\overline{y}_A$ that minimizes $B_{10,1}$ is $\overline{y}_A = (1 + \omega/n_A)\mu_B$ instead of $\overline{y}_A = \mu_B^*$.

The joint posterior distribution of $\mu_A$ and $\mu_B$ under $H_{12}$ is

$$\pi_2(\mu_A, \mu_B \mid y, H_1) = N_1\left(\mu_A \mid \frac{n_A\overline{y}_A}{n_A + \omega_A}, \frac{\sigma^2}{n_A + \omega_A}\right) N_1\left(\mu_B \mid \frac{n_B\overline{y}_B}{n_B + \omega_B}, \frac{\sigma^2}{n_B + \omega_B}\right).$$

This posterior converges weakly (almost surely) to $\pi_*(\mu_A, \mu_B \mid y, H_1) = N_1(\mu_A \mid n_A\overline{y}_A/(n_A + \omega_A), \sigma^2/(n_A + \omega_A))\mathbb{1}(\mu_B = \mu_B^*)$, which is the posterior of the one-sample test if the prior on $\mu_A$ under $H_{11}$ is $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid 0, \sigma^2/\omega_A)$. Under this prior specification, $\overline{f}_2 \to_d \overline{f}_1$ for continuous and bounded $f$ (which includes decisions under $\{0, 1\}$ loss) and it is possible to show that convergence holds for some unbounded functions such as $(\mu_A - \mu_B)^2$ on a case-to-case basis (see Section 4.3.4 for more details).

Therefore, limit consistency does not hold unless the prior for the unknown mean in the one-sample test is $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid 0, \sigma^2/\omega_A)$, which is centered at 0 and not the hypothesized value under the null $\mu_B^*$. We centered the priors at 0 for simplicity, but the same phenomenon occurs if they are at centered at some arbitrary $\mu_0$ : limit consistency would not occur unless $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid \mu_0, \sigma^2/\omega_A)$.

### *4.3.2 Effect-size, baseline, and sum-to-zero priors*

In this section, we consider alternative parametrizations that are limit-consistent under an appealing prior specification for the one-sample problem.

First, we parametrize the two-sample problem in terms of a grand mean $\mu$ and an effect size $\delta$, so that the mean in the first group is $\mu_A = \mu - \delta/2$ and the mean of the second group is $\mu_B = \mu + \delta/2$. Then, the hypotheses become $H_{02} : \delta = 0$ against $H_{12} : \delta \neq 0$. Our prior on $\mu$ under both $H_{02}$ and $H_{12}$ is $\pi_2(\mu) \propto 1$, whereas our prior on $\delta$ is $\pi_2(\delta \mid H_{12}) = N_1(\delta \mid 0, \sigma^2/\omega)$. Then, the Bayes factor of $H_{12}$ to $H_{02}$ can be written as

$$B_{10,2} = (1 + \kappa/\omega)^{-1/2} \exp\left\{\frac{\kappa Z_n^2}{2(\kappa + \omega)}\right\},$$

As $n_B \to \infty$, $B_{10,2}$ converges in distribution to the Bayes factor of the one-sample test $H_{01} : \mu_A = \mu_B^*$ against $H_{11} : \mu_A \neq \mu_B^*$ under the prior $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid \mu_B^*, \sigma^2/\omega)$, that is

$$B_{10,2} \to_d (1 + n_A/\omega)^{-1/2} \exp\left\{\frac{n_A\, Z_*^2}{2(n_A + \omega)}\right\}.$$

The prior implied by the limit is centered at the hypothesized value under the null, and it is a natural analogue of the prior set for the two-sample test: for instance, if $\omega = \kappa/g$, we recover Zellner's $g-$prior (Zellner, 1986). It is straightforward to check the posterior of $\delta = (\mu_B - \mu_A)$ is distributed as $N_1(\kappa(\overline{y}_B - \overline{y}_A)/(\kappa + \omega), \sigma^2/(\kappa + \omega)$, which converges weakly to $N_1(\delta \mid n_A(\mu_B^* - \overline{y}_A)/(n_A + \omega), \sigma^2/(n_A + \omega))$. The limit is equal in distribution to $(\mu_B^* - \mu_A) \mid y_A, H_{11}$ under $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid \mu_B^*, \sigma^2/\omega)$. Thus, limit consistency for this choice of priors follows for continuous and bounded $f$.

Now, consider an alternative parametrization where one of the groups is a baseline and the other one has a group-specific effect: that is, the mean of group A is $\mu_A$ and the mean of group B is $\mu_A + \alpha$. Then the hypothesis test can be rewritten as $H_{02} : \alpha = 0$ against $H_{12} : \alpha \neq 0$. Our prior on $\mu_A$ under the null and alternative is $\pi_2(\mu_A) \propto 1$, whereas $\pi_2(\alpha \mid H_{12}) = N_1(\alpha \mid 0, \sigma^2/\omega)$. Then, the Bayes factor of $H_{12}$ to $H_{02}$ is identical to $B_{10,2}$ above [NB: This is true in general for any transformation that

can be written as a translation of the design matrix]. Crucially, the result is invariant to the choice of baseline. The posterior distribution of $\alpha$ behaves analogously to the posterior of $\delta$, so limit consistency follows for continuous and bounded $f$ if $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid \mu_B^*, \sigma^2/\omega)$.

Finally, we parametrize the problem in terms of a grand mean $\mu$ and group specific effects $\tau_A$ and $\tau_B$ that are restricted to $\tau_A + \tau_B = 0$. The hypothesis test can be written as $H_{02} : \tau_A = \tau_B = 0$ against $H_{12} : \tau_A \neq \tau_B$. Our prior on $\mu$ is flat under both hypotheses, just as with the other parametrizations. We construct the restricted prior on $\tau = (\tau_A, \tau_B)'$ as follows. We start with an unrestricted prior $\tau \sim N_2(0_2, (\sigma^2/\psi)I_2)$. Then, we note that $\tau$ and $\tau_A + \tau_B$ are jointly singular normal, and condition on the event $\tau_A + \tau_B = 0$. The resulting prior is the singular normal

$$\tau \mid \{\tau_A + \tau_B = 0\} \sim N_2(0_2, V),\ V = (\sigma^2/\psi)(I_2 - 1_2 1_2'/2),$$

where $1_2 = (1, 1)'$ and $I_2$ is the $2 \times 2$ identity matrix. These parameters can be computed with the usual formulae for deriving conditional distributions from multivariate normal distributions, substituting matrix inverses by generalized inverses as needed (Marsaglia, 1964). The density of the singular normal with respect to Lebesgue measure on $\{\tau_A + \tau_B = 0\}$ can be written as (see, for example, Chapter 20 of Seber (2008))

$$\pi(\tau \mid \{\tau_A + \tau_B = 0\}) = (2\pi)^{-1/2} |V|_+^{-1/2} \exp\{-\tau' V^- \tau/2\}\, \mathbb{1}(\{\tau_A + \tau_B = 0\}),$$

where $|V|_+$ is equal to the product of the non-zero eigenvalues of $V$ and $V^-$ is a generalized inverse of $V$ (in this case, $|V|_+ = \sigma^2/\psi$). The Bayes factor of $H_{12}$ to $H_{02}$ is

$$B_{10,2} = (1 + 2\kappa/\psi)^{-1/2} \exp\left\{ \frac{\kappa Z_n^2}{2(\kappa + \psi/2)} \right\},$$

which is equivalent to the Bayes factor under the effect-size and baseline parametrizations if $\omega = \psi/2$. Moreover, the posterior of $\tau_B - \tau_A$ is $N_1(\kappa(\overline{y}_B - \overline{y}_A)/(\kappa + \psi/2), \sigma^2/(\kappa +$

$\psi/2)$, which is the same as the posterior distribution of $\delta$ in the effect-size parametrization if $\psi = 2\omega$. Therefore, limit consistency holds if the prior of the one-sample problem is $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid \mu_B^*, \sigma^2/\omega)$ and $f$ is continuous and bounded.

### *4.3.3 Normal prior on common parameter*

In the previous subsection, the priors on the common parameters were improper (flat) in all cases. In this subsection, we study what happens if the prior on the common parameter is a proper normal prior instead.

We start with the effect-size parametrization. That is, we parametrize in terms of grand mean $\mu$ and an effect size $\delta$, where the mean of group A is $\mu - \delta/2$ and the mean of group B is $\mu + \delta/2$. Let the prior on the grand mean be $\pi_2(\mu) = N_1(\mu \mid 0, \sigma^2/\lambda)$ and the prior on the effect size be $\pi(\delta) = N_1(\delta \mid 0, \sigma^2/\omega)$. Then, the Bayes factor of $H_{12}$ to $H_{02}$ is

$$B_{10,2}^{\delta} = (1 + \kappa_\delta/\omega)^{-1/2} \exp\left\{ \frac{\kappa_\delta Z_\delta^2}{2(\kappa_\delta + \omega)} \right\}$$

$$\kappa_\delta = \frac{\lambda n + 4 n_A n_B}{4(n + \lambda)}$$

$$Z_\delta = \frac{(\lambda + 2n_A) n_B \overline{y}_B - (\lambda + 2n_B) n_A \overline{y}_A}{2\sigma\sqrt{k_\lambda}(n + \lambda)}.$$

If $\lambda = 0$, we recover the Bayes factor in Section 4.3.2. As $n_B \to \infty$,

$$B_{10,2}^{\delta} \to_d B_l = (1 + \kappa_l/\omega)^{-1/2} \exp\left\{ \frac{\kappa_l}{2(\kappa_l + \omega)} Z_l^2 \right\}$$

$$\kappa_l = \lambda/4 + n_A$$

$$Z_l = \frac{n_A(\mu_B^* - \overline{y}_A)}{\sigma\sqrt{\kappa_l}} + \frac{\lambda \mu_B^*}{2\sigma\sqrt{\kappa_l}}.$$

If $\lambda \neq 0$, in general there is no prior distribution $\pi_1(\mu_A \mid H_{11})$ that yields $B_l$ as the Bayes factor. Following the same argument as in Section 4.3.1, the integral of

$N_1(\overline{y}_A \mid \mu_B^*, \sigma^2/n_A)B_l$ with respect to $\overline{y}_A$ should be equal to 1, but

$$N_1\left(\overline{y}_A \mid \mu_B^*, \frac{\sigma^2}{n_A}\right) B_l = \left[1 + \frac{\lambda}{4\omega}\right]^{-1/2} \exp\left\{\frac{\lambda^2(\mu_B^*)^2}{2(\lambda+4\omega)\sigma^2}\right\} N\left(\overline{y}_A \mid m, \frac{(\kappa_l+\omega)\sigma^2}{n_A(\kappa_l+\omega-n_A)}\right)$$

$$m = \frac{1-(\lambda+2n_A)/(2(\kappa_l+\omega))}{1-n_A/(\kappa_l+\omega)}\mu_B^*.$$

There are only two values of $\eta = 1+\lambda/(4\omega)$ that make the extra term go to 1. One of them is $\eta = 1$, which includes the uninteresting case $\omega \to \infty$ (the prior for $\delta$ converges to a point mass at 0) and the aforementioned case where $\lambda = 0$. The other solution is the unique root of $4\omega(\mu_B^*)^2(\eta-1)^2/\sigma^2 = \eta\log\eta$ that is strictly greater than 1. This choice depends on $\mu_B^*$, so it cannot be picked *a priori* for the two-sample problem (and it seems unlikely that it would be chosen for the one-sample problem). This shows that parametrizing in terms of $\mu$ and $\delta$ is not a guarantee to have limit consistency. The posterior of $\mu_A = \mu - \delta/2$ converges weakly to $N_1(n_A\overline{y}_A + (\omega-\lambda/4)\mu_B^*)/(n_A+\lambda/4+\omega), \sigma^2/(n_A+\lambda/4+\omega))$. There exist priors for the one-sample problem that yield that posterior if $\lambda = 0$ or $\lambda = 4\omega$. The former case is the one discussed in the previous subsection; in the latter case, the prior would be $\pi_1(\mu_A \mid 0, \sigma^2/(2\omega))$, which is centered at 0 (not the hypothesized value under the null).

The baseline parametrization is arguably even more problematic. Suppose we parametrize the problem so that the mean of group A is $\mu_A$ and the mean of group B is $\mu_A + \alpha$, so we want to test $H_{02} : \alpha = 0$ against $H_{12} : \alpha \neq 0$. In this case, the prior on $\mu_A$ is $\pi_2(\mu_A) = N_1(\mu_A \mid 0, \sigma^2/\lambda)$ and $\pi_2(\alpha \mid H_{12}) = N_1(\alpha \mid 0, \sigma^2/\omega)$. Then, the Bayes factor of $H_{12}$ to $H_{02}$ is

$$B_{10,2}^{\alpha} = (1+\kappa_\alpha/\omega)^{-1/2} \exp\left\{\frac{\kappa_\alpha Z_\alpha^2}{2(\kappa_\alpha+\omega)}\right\}$$

$$\kappa_\alpha = \frac{\lambda n_B + n_A n_B}{n+\lambda}$$

$$Z_\alpha = \frac{n_B n_A(\overline{y}_B - \overline{y}_A) + \lambda n_B \overline{y}_B}{2\sigma\sqrt{k_\alpha}(n+\lambda)}.$$

The Bayes factor depends the choice of baseline unless $\lambda = 0$, in which case we recover the Bayes factor in the previous subsection. Additionally, the limit of $B^{\alpha}_{10,2}$ as $n_B \to \infty$ is not equal to the Bayes factor of the one-sample test under any realistic prior unless $\lambda \neq 0$ (in the same sense that we argued for the effect-size parametrization). The posterior for $\mu$ as $n_B$ goes to infinity converges weakly to $N_1((n_A\overline{y}_A + \omega\mu_B)/(n_A + \omega + \lambda), \sigma^2/(n_A + \omega + \lambda))$, which is not a posterior for the one-sample problem if $\lambda \neq 0$.

Lastly, with the sum-to-zero parametrization, one can show that the Bayes factor and the posterior for $\tau_A - \tau_B$ is the same as the Bayes factor and posterior for $\delta$ with the effect size parametrization with $\omega = \psi/2$, so the same arguments apply in this case.

To sum up, we have seen that putting a proper prior on the common parameter is problematic: the Bayes factors are not limit-consistent and, in the case of the baseline parametrization, the inferences depend on the choice of baseline, which is undesirable.

### *4.3.4 Non-local priors*

Consider the hypothesis test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. A prior for $\theta \mid H_1$ is said to be non-local if it vanishes as $\theta$ approaches $\Theta_0$ (Johnson and Rossell, 2010). Bayes factors under non-local priors have been seen to attain faster rates of convergence to the "truth" under the null hypothesis (Johnson and Rossell, 2010, 2012), and can be expressed as a product of a local prior times a penalty term near $\Theta_0$ (Rossell and Telesca, 2017). To see this, if $\pi_{\mathrm{NL}}$ is the non-local prior and $\pi_{\mathrm{L}}$ is a local prior,

$$\pi_{\mathrm{NL}}(\theta) = \frac{\pi_{\mathrm{NL}}(\theta)}{\pi_{\mathrm{L}}(\theta)}\pi_{\mathrm{L}}(\theta) = q(\theta)\pi_L(\theta).$$

Clearly, $q(\theta)$ vanishes at $\Theta_0$ because $\pi_{\mathrm{NL}}$ does by definition. If the conditional distribution of the data given $\theta$ has density $p(y \mid \theta)$ and the marginal distribution under

the local prior is $m_L(y)$, the marginal of the non-local prior can be expressed as

$$m_{\text{NL}}(y) = \int q(\theta)\pi_L(\theta)p(y \mid \theta)\,\mathrm{d}\theta = \mathbb{E}_{\text{L}}[q(\theta) \mid y]\, m_L(\theta),$$

which, again, can be interpreted as "penalty $\times$ local prior." As noted by Consonni et al. (2013), Bayes decisions under non-local priors are equivalent to Bayes decisions under local priors with respect to a loss function that involves $q(\theta)$. If we adapt the notation described in Section 4.2, the Bayes decision under $\pi_{\text{NL}}$ with respect to the loss

$$L_{\text{NL}}(\theta, H_0) = \mathbb{1}(\theta \in \Theta_1)\gamma_1; \quad L_{\text{L}}(\theta, H_1) = \mathbb{1}(\theta \in \Theta_0)\gamma_0,$$

is equivalent to the Bayes decision under $\pi_{\text{L}}$ with respect to the loss

$$L_{\text{L}}(\theta, H_0) = \mathbb{1}(\theta \in \Theta_1)\gamma_1 q(\theta); \quad L_{\text{L}}(\theta, H_1) = \mathbb{1}(\theta \in \Theta_0)\gamma_0.$$

Thus, users of local priors can be compelled to use non-local priors as long as they believe that $q(\theta)$ is a reasonable "penalty" for wrongly rejecting the alternative.

This general discussion applies to our problem by changing the notation appropriately; indeed, the results in previous sections for local priors carry over to non-local priors with continuous and bounded penalties $q(|\mu_A - \mu_B|)$ under the loss $L_{\text{NL}}$ (which includes $\{0, 1\}$ loss). Two of the most popular choices of non-local priors have unbounded penalties, but it is possible to show that the correspondence holds for them as well. For concreteness, we illustrate this with the effect-size parametrization, but the same line of reasoning can be used for the baseline and sum-to-zero parametrizations.

The 3 most popular classes of non-local priors are the so-called moment, exponential, and inverse-moment priors. In our context, they correspond to the penalties:

$$q_M(\delta) = \omega\delta^2/\sigma^2$$

$$q_E(\delta) = \exp\left\{\sqrt{2} - \frac{\sigma^2}{\omega\delta^2}\right\}$$

$$q_I(\delta) = \frac{\sqrt{2}\sigma^2}{\omega\delta^2}\exp\left\{-\frac{\sigma^2}{\omega\delta^2} + \frac{\omega\delta^2}{2\sigma^2}\right\}.$$

Since $q_E$ is continuous and bounded, our results apply directly. Although $q_M$ is unbounded, it is easy to check that its expectation with respect to $\pi_2(\delta \mid H_{12}) = N_1(\delta \mid 0, \sigma^2/\omega)$ converges to the corresponding posterior expectation under $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid \mu_B^*, \sigma^2/\omega)$ as $n_B \to \infty$ :

$$\begin{aligned}
\mathbb{E}_{\pi_2}[q_M(\delta) \mid y_A, y_B] &= \frac{\omega}{\sigma^2}\left[\frac{\kappa^2(\overline{y}_B - \overline{y}_A)^2}{(\kappa+\omega)^2} + \frac{\sigma^2}{\kappa+\omega}\right] \\
&\to_d \frac{\omega}{\sigma^2}\left[\frac{n_A^2(\mu_B^* - \overline{y}_A)^2}{(n_A+\omega)^2} + \frac{\sigma^2}{n_A+\omega}\right] \\
&= \mathbb{E}_{\pi_1}[q_M(|\mu_B^* - \mu_A|) \mid y_A].
\end{aligned}$$

Showing that the posterior expectation of $q_I$ under $\pi_2$ converges to that under $\pi_1$ is a bit more elaborate, but it can be shown to be true (see Appendix C; it amounts to showing that the sequence of posterior measures are uniformly integrable almost surely). Therefore, limit consistency for the loss $L_{\mathrm{NL}}$ follows for the moment and exponential non-local priors implied by $q(\theta)\pi_2(\delta \mid H_{12})$ and $q(\theta)\pi_1(\mu_A \mid H_{11})$.

### *4.3.5 Mixtures of g−priors and empirical Bayes*

So far, our discussion has been limited to normal priors. However, conjugate priors can have undesirable behavior that thicker-tailed priors and empirical Bayes approaches avoid (see, for example, Chapters 2 and 3 of this dissertation).

In Section 4.3.2, we showed that what we called effect-size, baseline, and sum-to-zero parametrizations yield equivalent Bayes factors if the prior on the common parameter is flat and the prior on the model-specific parameter (the additional parameter under the alternative) is conjugate. Using the same notation as in Section 4.3.2, let $\delta \sim N_1(0, \sigma^2 g/\kappa)$, $\alpha \sim N_1(0, \sigma^2 g/\kappa)$, or $\tau \mid \{\tau_A + \tau_B = 0\} \sim N_2(0_2, g\sigma^2(I_2 - 1_2 1_2'/2)/(2\kappa))$ and suppose that $g$ has a prior distribution with density $\pi(g)$. Then,

$$B_{10,2} = \int (g+1)^{-1/2} \exp\left\{\frac{g Z_n^2}{2(g+1)}\right\} \pi(g)\, \mathrm{d}g.$$

Given $n_A$ and $\overline{y}_A$, the integral is a continuous function of $Z_n^2$ (see Appendix C), so

we can apply the continuous mapping theorem to conclude that

$$B_{10,2} \to \int (g+1)^{-1/2} \exp\left\{\frac{gZ_*^2}{2(g+1)}\right\} \pi(g)\,\mathrm{d}g.$$

Therefore, conditional on $y_A$, mixtures of conjugate priors are limit-consistent under $\{0,1\}$-loss in the same conditions as in Section 4.3.2.

Empirical Bayes approaches are also well-behaved. If the scale parameter $\omega$ is set by maximizing $B_{10,2}$ in Section 4.3.2, the maximizer is $\widehat{\omega}_n = \kappa/(Z_n^2 - 1)$. If the prior for the one-sample problem is $\pi_1(\mu_A \mid H_{11}) = N_1(\mu_A \mid \mu_B^*, \sigma^2/\widehat{\omega}^*)$, where $\widehat{\omega}^*$ is set by maximizing the corresponding $B_{10,1}$, we find $\widehat{\omega}^* = n_A/(Z_*^2 - 1)$. Since $\widehat{\omega}_n \to_d \widehat{\omega}^*$, limit consistency follows for continuous and bounded $f$.

Similar arguments can be made for the priors in Section 4.3.3 substituting $\kappa$ by $\kappa_\delta$, $\kappa_\alpha$, and $\kappa_\tau$. The conclusion is the same: limit consistency does not hold in general and the inferences with baseline parametrization depend on the choice of baseline.

## 4.4 Conclusions and future work

In this chapter, we have defined limit consistency and, in our view, have learned some lessons that can be summarized as follows:

1. It is important to parametrize the problem so that there is a common parameter between the groups. If the priors for groups A and B are independent, limit consistency does not hold unless the prior for the one-sample problem is centered at 0 instead of the hypothesized value under the null.

2. It is important to put a flat prior on the common parameter. Otherwise, Bayes decisions depend on the choice of parametrization and are not limit consistent. If one favors the baseline parametrization, choosing a normal prior on the common parameter breaks the invariance of decisions with respect to the choice of baseline.

3. Conjugate priors, non-local priors, mixtures of $g-$priors, and empirical Bayes approaches are limit-consistent as long as one follows lessons 1. and 2.

We have assumed that the variance is known, but we do not regard this assumption as critical. For example, all the results in this chapter go through if $\sigma^2$ is replaced by a consistent estimator (for example, the sample variance).

The paramerizations discussed in this article are connected to the usual parametrizations in analysis of variance (ANOVA). Independent priors can be seen as a form of what usually goes under the name "cell-means," whereas the baseline and sum-to-zero parametrizations go under that very same name in that literature. An obvious next step is considering comparisons of more than two groups. The baseline and sum-to-zero parametrizations can be extended in a straightforward manner to more than 2 groups (the effect-size parametrization does not generalize easily). When comparing more than 2 groups, interesting questions arise. For example, does the choice of prior covariance matrix for the model-specific parameters matter?

Another item that should be studied in more detail is the connection between the sum-to-zero and the baseline parametrizations. The correspondence between baseline and effect-size was immediate because it is well known that, under a flat prior on the common parameter, "Bayesian answers" are invariant up to translations of the design matrix. To the best of our knowledge, there is no general result in the literature connecting sum-to-zero and baseline parametrizations, although we do suspect that they can be shown to be equivalent.

Lastly, we believe that limit consistency can be used to improve prior choice in problems that are not testing normal means, such as comparing proportions, variances, etc. Moreover, it could potentially be applied in Bayesian nonparametric tests for comparing two groups like those in Holmes et al. (2015) and Soriano and Ma (2017).

# 5

# On Birnbaum's theorem

## 5.1 Introduction

Birnbaum's theorem (Birnbaum, 1962) states that two statistical principles that are intuitively reasonable, the weak conditionality principle and the sufficiency principle, imply the likelihood principle, which is violated by statistical procedures such as $p$-values or reference priors. Ever since the result was published, there has been a lively discussion on its validity and implications. The monograph Berger and Wolpert (1988) contains a defense of the likelihood principle and responses to criticisms up to the date it was published, but the flow of articles has not stopped in the fields of statistics and philosophy of science (for example, Helland (1995), Bjørnstad (1996), Robins and Wasserman (2000), Sweeting (2001), Wechsler et al. (2008), Grossman (2011), Gandenberger (2014)). The articles Evans (2013) and Mayo (2014) have cast doubt on the validity and implications of Birnbaum's theorem, and the goal of this chapter is to review and discuss their content.

First, we introduce our basic notation and definitions for statistical experiment, inference bases, and informative inference:

- **Statistical experiment:** A triplet $E = \{\mathcal{X}_E, \Theta_E, p_{\theta,E}\}$, where $\mathcal{X}_E$ is the sample space of the experiment, $\Theta_E$ is the parameter space, and $p_{\theta,E}$ is the sampling distribution of $E$ for $\theta \in \Theta$. As it is usual in the literature, we avoid measure-theoretical details by considering experiments with a discrete support (see Section 3.4 in Berger and Wolpert (1988) for generalizations).

- **Inference tuple:** A tuple $(E, x)$ where $E$ is a statistical experiment and $x \in \mathcal{X}_E$ is an outcome from $E$.

- **Informative inference:** $\mathbf{Ev}(E, x)$ is the informative inference (or conclusion) made by an agent given $(E, x)$. If $\mathcal{I}$ is the space of inference bases, one can think of Ev as a function from $\mathcal{I}$ to a set $\mathcal{D}$ of possible inferences.

- **Inferentially equivalent:** Two inference bases $(E, x)$ and $(E', x')$ are inferentially equivalent if $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$ (the same inferences are made given $(E, x)$ or $(E', x')$).

Given the definitions above, we define the statistical principles at stake: the weak conditionality principle (WCP), ancillarity principle (AP), sufficiency principle (SP), and the likelihood principle (LP):

- **Weak Conditionality Principle (WCP):** Consider the statistical experiments $E_1 = (\mathcal{X}_{E_1}, \Theta, p_{\theta,E_1})$, $E_2 = (\mathcal{X}_{E_2}, \Theta, p_{\theta,E_2})$ and a 50-50 mixture between $E_1$ and $E_2$, which we denote $E_{\text{mix}}$. Conceptually, one can imagine that a fair coin is tossed: if it lands heads, $E_1$ is performed; if it lands tails. $E_2$ is performed. Formally, the outcome of the mixture experiment will be a pair $(j, x)$, where $j$ indicates the experiment that was performed ($j = 1$ if $E_1$ was performed, and $j = 2$ if $E_2$ was performed instead), and $x \in \mathcal{X}_{E_1} \cup \mathcal{X}_{E_2}$ is the outcome of the experiment that was performed. The WCP states that the informative inference given $(E_{\text{mix}}, (j, x))$ from the mixture experiment should be

equal to the informative inference given the inference base of the component experiment $(E_j, x)$; that is, $\mathbf{Ev}(E_{\text{mix}}, (j, x)) = \mathbf{Ev}(E_j, x)$.

- **Ancillarity Principle (AP):** Let $U$ be an ancillary statistic for $\theta$ (the distribution of $U$ does not depend on $\theta$) for which the value $u$ is observed. Then, $\mathbf{Ev}(E, (u, x)) = \mathbf{Ev}(E_{|U=u}, x)$, where the sampling distribution associated with $E_{|U=u}$ is $p_{\theta,E|U=u}(\cdot)$ (the conditional probability mass function of $x$ given $U = u$). In words, the ancillarity principle states that conditioning on an ancillary statistic should not change our informative inference. This principle is also known as the (strong) conditionality principle. Clearly, the selection of the component in WCP is an example of an ancillary statistic, so AP implies WCP.

- **Sufficiency Principle (SP):** If $(E, x)$ and $(E, x')$ are such that $T(x) = T(x')$ for a sufficient statistic $T$ for $\theta$, then $\mathbf{Ev}(E, x) = \mathbf{Ev}(E, x')$.

- **Likelihood Principle (LP):** If $(E, x)$ and $(E', x')$ are such that $p_{\theta,E}(x) = c\, p_{\theta,E'}(x')$ for $c > 0$ that does not depend on $\theta$, then $\mathbf{Ev}(E, x) = \mathbf{Ev}(E, x')$.

In our framework, **Ev** can be any function from the space of inference bases to inferences, and the mathematical role of statistical principles is restricting the set of functions that one is allowed to use. As explained in more detail in Section 5.2, Evans' objections arise because a map **Ev** is not introduced. Conversely, in Section 5.3 we show that the definition of the sufficiency principle in Mayo (2014) is different from SP (as defined in the paragraph above) and blocks Birnbaum's proof.

## 5.2 Evans' objections

Evans defines the statistical principles as the following set relations on $\mathcal{I} \times \mathcal{I}$:[1]

[1] We use a slightly different notation than in Evans (2013). We define $C$ as a formalization of WCP and $A$ as a formalization of AP. However, Evans (2013) does not consider WCP at all and

- $C$: $(E, x) \sim_C (E', x')$ if and only if $E = E_{\text{mix}}$, $x = (j, x_j)$, $E' = E_j$, and $x' = x_j$ as in the definition of WCP in Section 5.1 (or with roles of $(E, x)$ and $(E', x')$ reversed).

- $A$: $(E, x) \sim_A (E', x')$ if and only if $x = (u, x')$ and $E' = E_{|U=u}$, where $U = u$ and $E_{|U=u}$ are as defined for AP in Section 5.1 (or with roles of $(E, x)$ and $(E', x')$ reversed).

- $S$: $(E, x) \sim_S (E', x')$ if and only if there exists a sufficient statistic $T$ for $\theta$ such that $T(x) = T(x')$.

- $L$: $(E, x) \sim_L (E', x')$ if and only if $p_{\theta,E}(x) = c\, p_{\theta,E'}(x')$ for a constant $c > 0$ which does not depend on $\theta$.

This approach is different from the one taken in Section 5.1 and the one in Birnbaum (1962) because **Ev** is not defined or used at all in the definitions. Nonetheless, the set relations are very similar to the principles defined in Section 5.1: they are of the form $(E, x) \sim_P (E', x')$ if and only if $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$ by an application of a principle statistical $P$. According to Evans, "A basic step missing in Birnbaum (1962) was to formulate the principles as relations on the set $\mathcal{I}$ of all model and data combinations." But the definition of a function **Ev** automatically induces an equivalence relation on $\mathcal{I} \times \mathcal{I}$ (the kernel of **Ev**): $(E, x) \sim (E, x')$ if and only if $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$. If we accept WCP and SP as defined in Section 5.1, the equivalence relation on $\mathcal{I} \times \mathcal{I}$ induced by accepting WCP and SP implies that bases with proportional likelihoods are equivalent because they map to the same value.

Evans shows that statistical principles defined as set relations need not be equivalence relations (for instance, $A$ and $C$ as defined above are not). Even if two statistical principles are equivalence relations, their union may not be because it could fail to

defines a set relation (which is denoted $C$ in Evans' article) which is equivalent to our $A$. We apologize for the possible confusion that this change might cause.

be transitive: if $P_1$ and $P_2$ are set relations formalizing statistical principles, it is possible that the inference bases $(E_1, x_1)$ and $(E_2, x_2)$ are inferentially equivalent with respect to $P_1$ and $(E_2, x_2)$ and $(E_3, x_3)$ are inferentially equivalent according to $P_2$ but $(E_1, x_1)$ and $(E_3, x_3)$ are not inferentially equivalent according to either $P_1$ or $P_2$ alone.

Birnbaum's argument is a neat illustration of this phenomenon: the inference bases with proportional likelihoods are shown to be equivalent by a chain of applications of WCP and SP, but they are not equivalent according to either WCP or SP individually. Therefore, $L \neq S \cup C$, and the correct result is that $L$ is equal to the smallest equivalence relation generated by $S \cup C$, and Evans argues that extending statistical principles that are originally defined as set relations to equivalence relations requires further justification.

Thie extension can be justified as follows. If we define the sufficiency principle and the weak conditionality principle as the set relations $S$ and $C$ and introduce a function **Ev** with the minimal requirement that $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$ if and only if $(E, x) \sim_C (E', x')$ or $(E, x) \sim_S (E', x')$, the equivalence relation on $\mathcal{I} \times \mathcal{I}$ generated by **Ev** is precisely the smallest equivalence relation generated by $S \cup C$, which in this case is $L$. In general, if we define statistical principles $P_1, P_2, \ldots, P_k$ as set relations on $\mathcal{I} \times \mathcal{I}$ and introduce **Ev** with the property $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$ if and only if $(E, x) \sim_{P_i} (E', x')$ for some $i \in \{1, 2, \ldots k\}$, the equivalence relation on $\mathcal{I} \times \mathcal{I}$ induced by **Ev** is equal to the smallest equivalence relation generated by $P_1, P_2, \ldots, P_k$. Defining statistical principles as set relations on $\mathcal{I} \times \mathcal{I}$ and introducing **Ev** as we just did is equivalent to stating the definitions in terms of **Ev** in the first place as in Section 5.1.

Since the notation **Ev** is very explicit in Birnbaum (1962), we believe that the definition of the principles in terms of set relations was implied by the fact that **Ev** is a function. But even within a framework where **Ev** is not defined, the smallest

equivalence relation generated by a collection of principles has a straightforward interpretation: its elements are (exclusively) the result of a chain of applications of the principles we wish to respect. Rejecting the extension implies rejecting the equivalence of inference bases that can be shown to be equivalent by a number of applications of our principles. We believe, then, that the extension is also justified if **Ev** is not introduced.

Now we turn to an example in Evans (2013) that shows that $A$ is not transitive and illustrates some of the issues that were discussed in the paragraphs above.

**Example 4.** *(Evans (2013), pg. 2651) Let* $\mathcal{X}_E = \{1,2\} \times \{1,2\}$, $\Theta_E = \{1,2\}$, *with* $p_{E,\theta}$ *given in Table 5.1. Both* $U(x_1,x_2) = x_1$ *and* $V(x_1,x_2) = x_2$ *are ancillary, and the conditional models upon observing* $U = 1$ *and* $V = 1$ *are given in Tables 5.2 and 5.3. This example shows that* $A$ *is not transitive:* $(E,(x_1,x_2)) \sim_A (E_{|U},x_2)$ *and* $(E,(x_1,x_2)) \sim_A (E_{|V},x_1)$, *but* $(E_{|U},x_2) \not\sim_A (E_{|V},x_1)$ *because there is no ancillary statistic linking the two conditional models. However, using the definitions in Section 5.1 (or equivalently, using* $A$ *and introducing Ev with the property* $\mathbf{Ev}(E,x) = \mathbf{Ev}(E',x')$ *if and only if* $(E,x) \sim_A (E',x')$*), we have* $\mathbf{Ev}(E,(x_1,x_2)) = \mathbf{Ev}(E_{|U},x_2)$ *and* $\mathbf{Ev}(E,(x_1,x_2)) = \mathbf{Ev}(E_{|V},x_1)$, *so* $\mathbf{Ev}(E_{|U},x_2) = \mathbf{Ev}(E_{|V},x_1)$.

Table 5.1: Unconditional model (rows: sampling distributions for $\theta \in \{1,2\}$)

| $(x_1,x_2)$ | $(1,1)$ | $(1,2)$ | $(2,1)$ | $(2,2)$ |
|---|---|---|---|---|
| $f_{E,\theta=1}(x_1,x_2)$ | $1/6$ | $1/6$ | $2/6$ | $2/6$ |
| $f_{E,\theta=2}(x_1,x_2)$ | $1/12$ | $3/12$ | $5/12$ | $3/12$ |

Table 5.2: Conditional model when $U = 1$ (rows: sampling distributions for $\theta \in \{1,2\}$)

| $(x_1,x_2)$ | $(1,1)$ | $(1,2)$ | $(2,1)$ | $(2,2)$ |
|---|---|---|---|---|
| $f_{E,\theta=1}(x_1,x_2 \mid U=1)$ | $1/2$ | $1/2$ | $0$ | $0$ |
| $f_{E,\theta=2}(x_1,x_2 \mid U=1)$ | $1/4$ | $3/4$ | $0$ | $0$ |

Table 5.3: Conditional model when $V = 1$ (rows: sampling distributions for $\theta \in \{1, 2\}$)

| $(x_1, x_2)$ | $(1,1)$ | $(1,2)$ | $(2,1)$ | $(2,2)$ |
|---|---|---|---|---|
| $f_{E,\theta=1}(x_1, x_2 \mid V = 1)$ | $1/3$ | $0$ | $2/3$ | $0$ |
| $f_{E,\theta=2}(x_1, x_2 \mid V = 1)$ | $1/6$ | $0$ | $5/6$ | $0$ |

Quoting Evans (2013): "Saying that such models [the conditional models in Tables 5.2, 5.3] contain an equivalent amount of statistical information is clearly a substantial generalization of $[A]$. To measure the accuracy of this estimate we can compute the conditional probabilities based on the two inference bases, namely,

$$\mathbb{P}_{\theta=1}(\widehat{\theta} = 1 \mid U = 1) = 1/2, \;\; \mathbb{P}_{\theta=2}(\widehat{\theta} = 1 \mid V = 1) = 3/4$$

and so the accuracy of $\widehat{\theta}$ is quite different depending on whether we [condition on $U$ or $V$]. It seems unlikely that we would interpret these inference bases as containing an equivalent amount of information in a frequentist formulation of statistics."

Concluding that the inference bases are equivalent with respect to $A$ is a consequence of introducing **Ev** with the property $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$ if and only if $(E, x) \sim_A (E', x')$. Also, the likelihood ratio of $\theta = 1$ to $\theta = 2$ equals 2 if we condition on either $U$ or $V$, which is unsurprising because, as Evans proves, AP equals LP. We agree with Evans in that this example shows that accepting AP can be problematic for frequentist statisticians: there are two ancillary statistics we can condition on, there is no apparent reason one should prefer one over the other and, unfortunately, standard errors and $p$-values depend on the choice of ancillary. We return to this point in Section 5.4.

After showing that AP is equivalent to LP, Evans concludes that SP is redundant in Birnbaum's argument. Then, Example 4 leads him to cast doubt on the impact of Birnbaum's result because he believes that many statisticians would not accept AP (or equivalently, $A$ and the equivalences generated by the principle). But SP is certainly not redundant if only WCP is assumed (recall that WCP only requires

equivalence of 50-50 mixtures), and WCP and SP also imply LP. In Example 4, the conditional experiments are not equivalent according to WCP, and the smallest equivalence relation containing $C$ would only add cases where mixture experiments with different components (or different probabilities of performing them) were considered, but the same component experiment was performed and the same result was obtained. Finally, we agree with Evans that accepting statistical principles may induce unexpected equivalences between inference bases, which is precisely what makes Birnbaum's result surprising and relevant.

## 5.3 Mayo's objections

In our view, the objections to Birnbaum's proof in Mayo (2014) stem from using a definition for the sufficiency principle that is different from that in Section 5.1. We believe that introducing new notation that makes an explicit distinction between the output of methods and the inference made by an agent that is using them is helpful for understanding the arguments:

- $\mathbf{M}(E, x)$: Result of a applying a method $M$ to the inference base $(E, x)$.
- $\mathbf{Ev}(E, x)$: Inference made by an agent given $(E, x)$ (as in Section 5.1).

Given $(E, x)$, the agent makes informative inferences $\mathbf{Ev}(E.x)$ by means of $\mathbf{M}(E', x')$ for some $(E', x')$ which may not be equal to $(E, x)$. The interpretation of $\mathbf{M}(E, x) = \mathbf{M}(E', x')$ is that the "output" of applying a method $\mathbf{M}$ to $(E, x)$ and $(E', x')$ is the same (one can imagine that $\mathbf{M}$ is a function in some programming language that takes $E$ and $x$ as inputs), whereas $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$ means that an agent makes the same informative inferences given $(E, x)$ or $(E', x')$. This distinction is somewhat obscured in Mayo (2014). The definitions are

- $\text{Infr}_E[x]$: The parametric statistical inference from a given or known $(E, x)$.

- $(E', x') \Rightarrow \text{Infr}_E[x]$: An informative parametric inference about $\theta$ from a given $(E, x)$ is to be computed by means of $\text{Infr}_E[x]$.

The definition of $\text{Infr}_E[x]$ and the name "Infr" suggest that $\text{Infr}_E[x] = \mathbf{Ev}(E, x)$. However, the second definition implies that $\text{Infr}_E[x]$ need not be equal to the final inference $\mathbf{Ev}(E, x)$. This is explicit in Mayo's definition of the weak conditionality principle (WCP):

- **WCP:** Given $(E_{\text{mix}}, (j, x_j))$, condition on the $E_j$ producing the result: $(E_{\text{mix}}, (j, x_j)) \Rightarrow \text{Infr}_{E_j}[x_j]$. Do not use the unconditional formulation: $(E_{\text{mix}}, (j, x_j)) \not\Rightarrow \text{Infr}_{E_{\text{mix}}}[(j, x_j)]$.

Using our notation, this definition is equivalent to the WCP in Section 5.1. However, Mayo defines the sufficiency principle as follows

- **SP2**: If there exists a sufficient statistic $T$ for $\theta$ and $T(x) = T(x')$, then $\text{Infr}_E[x] = \text{Infr}_E[x']$,

which is different from SP, and can be recast as

- **SP2:** If there exists a sufficient statistic $T$ for $\theta$ and $T(x) = T(x')$, then $\mathbf{M}(E, x) = \mathbf{M}(E, x')$.

The key point is that WCP is a property of **Ev** and SP2 is a property of **M**. If this distinction is made, LP does not follow. The distinction between **Ev** and **M** is not made in Birnbaum (1962) or Section 5.1. The following example, which is a slight modification of the example presented in Section 4. in Mayo (2010), puts the notation in context and makes clear why WCP and SP2 do not imply LP.

**Example 5.** *Consider binomial and negative binomial experiments*

$$E_1 = \{\{0, 1, 2, \ldots, n\}, \Theta, \text{Binomial}(n, \theta)\}, \quad E_2 = \{\{0, 1, 2, \ldots.\}, \Theta, \text{NegBinomial}(k, \theta)\}.$$

*Suppose that a fair coin is flipped and $E_1$ is performed if the coin lands heads and $E_2$ is performed if it lands tails. Let $E_{mix}$ denote the "mixture" experiment. The*

*outcome of $E_{mix}$ is $(j, x)$, with $j \in \{1, 2\}$ ($j = 1$ if $E_1$ is performed and $j = 2$ if $E_2$ is performed) and $x = (k, n - k)$, where $k$ and $n - k$ are the number of successes and failures observed after performing $E_j$. The statistical method $M(E, x)$ is the one-sided p-value for testing $\theta = \theta_0$ against $\theta > \theta_0$:*

$$\mathbf{M}(E_1, x) = \mathbb{P}(\text{Binomial}(n, \theta_0) \geq x)$$

$$\mathbf{M}(E_2, x) = \mathbb{P}(\text{NegBinomial}(r, \theta_0) \geq x)$$

$$\mathbf{M}(E_{\text{mix}}, x) = 0.5\,\mathbb{P}(\text{Binomial}(n, \theta_0) \geq x) + 0.5\,\mathbb{P}(\text{NegBinomial}(k, \theta_0) \geq x).$$

*We assume that the agent makes inference using the rule $\mathbf{Ev}(E_{\text{mix}}, (j, x)) = \mathbf{M}(E_j, x)$. The statistic $T(j, x) = (1, x)$ is sufficient for $\theta$ with respect to $E_{mix}$, and it satisfies both $T(1, x) = T(2, x)$ and $\mathbf{M}(E_{mix}, (1, x)) = \mathbf{M}(E_{mix}, (2, x))$, so SP2 is respected. WCP is automatically satisfied because the inference rule is $\mathbf{Ev}(E_{mix}, (j, x)) = \mathbf{M}(E_j, x) = \mathbf{Ev}(E_j, x_j)$ (the inference rule is chosen so that WCP is respected). It follows that WCP and SP2 do not imply LP.*

According to the definitions in Mayo (2014), WCP and SP2 do not imply LP, as seen in the example above. Where does Birnbaum's proof go wrong? With WCP as stated, the mixture experiments are inferentially equivalent to the performed components: $\mathbf{Ev}(E_{\text{mix}}, (1, x_1)) = \mathbf{Ev}(E_1, x_1)$ and $\mathbf{Ev}(E_{\text{mix}}, (2, x_2)) = \mathbf{Ev}(E_2, x_2)$. However, SP2 does not imply $\mathbf{Ev}(E_{\text{mix}}, (1, x_1)) = \mathbf{Ev}(E_{\text{mix}}, (2, x_2))$: instead, it requires $\mathbf{M}(E_{\text{mix}}, (1, x_1)) = \mathbf{M}(E_{\text{mix}}, (2, x_2))$. However, $\mathbf{Ev}(E_{\text{mix}}, (1, x_1))$ need not be equal to $\mathbf{Ev}(E_{\text{mix}}, (2, x_2))$. These definitions allow Mayo to claim that, in Example 5, reporting the conditional $p$-value according to the sampling distribution of the component experiment that was performed does not violate the sufficiency principle. In contrast, reporting the conditional $p$-value is a violation of SP as defined in Section 5.1 (and the proof of WCP and SP implies LP goes through as usual). Critically, note that SP states that if there exists a sufficient statistic, the inferences bases are inferentially equivalent, but there is no requirement that said sufficient

statistic be used for our final inferences. If that were the case, it would imply that SP instructs to use of the unconditional $p$-value. The reason that Birnbaum's proof does not go through in this framework hinges on the distinction of **Ev** and **M**: if we define a new WCP2 as a property of **M** (so that both SP2 and WCP2 were properties of **M**), reporting the conditional $p$-value in Example 5 would violate WCP2, as $\mathbf{M}(E_j, x_j) \neq \mathbf{M}(E_{\text{mix}}, (j, x_j))$ (and WCP2 and SP2 would, of course, imply a version of LP written in terms of **M**).

## 5.4 Can ancillaries be used in frequentist statistics?

We briefly discuss the applicability of the ancillarity principle in frequentist inference, motivated by comments in Cox and Mayo (2010), Evans (2013), and Mayo (2014). Since AP is equivalent to LP, frequentist statisticians that want to make conditional frequentist statements have to propose restricted versions of AP. Additionally, it is of utmost importance to find well-defined criteria for choosing among ancillaries because, as we have seen in Example 4, there are instances where there are multiple ancillaries one can condition on that give rise to different conditional $p$-values or standard errors. Some authors have proposed restricting the set of ancillaries to condition on (Durbin (1970), Kalbfleisch (1975)), but this approach is problematic because there are examples where several ancillaries satisfy the restrictions (see Basu (1964) for examples and Dawid (2011) for lucid review on the ancillarity principle and the issues discussed in this paragraph).

To the best of our knowledge, there is no (restricted) formulation of AP that instructs which ancillary one should use for any given problem, and as a result there is no adequate definition for a restricted ancillarity principle that is not equivalent to the likelihood principle [Cox (1971) provides a heuristic that works when applied to an example in Basu (1964), but it does not give a definite answer in other problems and it is not regarded as a general solution to this problem]. Another issue is that there

are examples where a conditional analysis is clearly desirable, but useful ancillary statistics are not available. We present two examples below.

**Example 6.** *(Example 8 in Berger and Wolpert (1988)) Let* $\Theta = [0, 1)$*,* $P(X = \theta) = 1 - \theta$*, and* $P(X = 0) = \theta$*. Consider the confidence set* $C = \{X\}$*. Unconditionally,* $P(\theta \in C) = 1 - \theta$*. However, if* $X > 0$*, we know that* $C = \{X\}$ *contains* $\theta$ *with probability 1, but* $X$ *is not ancillary, so the ancillarity principle would not allow conditioning on its value.*

**Example 7.** *Let* $X_1, X_2$ *be independent and identically distributed random variables with* $P(X_i = \theta - 1) = P(X_i = \theta + 1) = 1/2$ *for* $i \in \{1, 2\}$*. Let* $D = |X_1 - X_2|/2$*, which is ancillary with* $P(D = 1) = P(D = 0) = 1/2$*. Suppose we want to evaluate the quality of the estimator* $T = X_{(1)} + 1$ *(*$X_{(1)}$ *is the minimum of* $X_1$ *and* $X_2$*). Conditioning on* $D$*, we know that* $P(T = \theta \mid D = 1) = 1$ *and* $P(T = \theta \mid D = 0) = 1/2$*, and Cox and Mayo (2010) would propose reporting inferences conditional on* $D$ *because it is more informative than an unconditional analysis. But now consider the following modification:* $P(X_i = \theta + 1) = 1/2 + \theta\epsilon$ *and* $P(X_i = \theta - 1) = 1/2 - \theta\epsilon$ *for a known* $\epsilon \in [0, 1]$ *and* $\theta \in [-1/(2\epsilon), 1/(2\epsilon)]$*. The original example is a particular case with* $\epsilon = 0$*. If* $\epsilon \neq 0$*,* $D$ *is not ancillary anymore, despite the fact that if* $\epsilon$ *is small (say* $\epsilon = 10^{-100}$*) we are essentially in the same situation as if* $\epsilon = 0$*. In addition, if* $\epsilon \neq 0$*, there are (even more) cases where we can retrieve* $\theta$ *with probability 1 given the data. Indeed, if* $X_1 \neq X_2$ *we still have that* $\theta = X_{(1)} + 1$*, but now there are cases where we know the value of* $\theta$ *exactly even if* $X_{(1)} = X_{(2)}$*. Let* $A_{\theta-1} = [-1/(2\epsilon) - 1, 1/(2\epsilon) - 1]$ *and* $A_{\theta+1} = [-1/(2\epsilon) + 1, 1/(2\epsilon) + 1]$*. If* $X_{(1)} \in A_{\theta-1} \setminus A_{\theta+1}$*, then* $\theta = X_{(1)} + 1$*; analogously,* $\theta = X_{(1)} - 1$ *whenever* $X_{(1)} \in A_{\theta+1} \setminus A_{\theta-1}$*. Note that if* $\epsilon > 1/2$*,* $A_{\theta-1} \cap A_{\theta+1} = \emptyset$ *and we can always retrieve the value of* $\theta$*. If we want to assess the*

*performance of $T$ conditionally, we know that*

$$P(T = \theta \mid X_{(1)} \neq X_{(2)}) = 1$$

$$P(T = \theta \mid X_{(1)} = X_{(2)}, X_{(1)} \in (A_{\theta-1} \setminus A_{\theta+1})) = 1$$

$$P(T = \theta \mid X_{(1)} = X_{(2)}, X_{(1)} \in (A_{\theta+1} \setminus A_{\theta-1})) = 0$$

$$P(T = \theta \mid X_{(1)} = X_{(2)}, X_{(1)} \in A_{\theta-1} \cap A_{\theta+1}) = \frac{(1/2 - \theta\epsilon)^2}{(1/2 - \theta\epsilon)^2 + (1/2 + \theta\epsilon)^2},$$

*but unconditionally $P(T = \theta) = 1 - (1/2 + \theta\epsilon)^2$, which depends on $\theta$ and ranges from 1 to 0 for $\theta \in [-1/(2\epsilon), 1/(2\epsilon)]$. Therefore, the confidence level of the set $C = \{T\}$ is* $\inf P_\theta(\theta \in C) = 0$*, which is clearly undesirable and misleading (especially in cases where $\epsilon > 1/2$, where a conditional analysis reveals if $T = \theta$ with probability 0 or 1 depending on the data). As an aside, a modified estimator that takes on the value $X_{(1)} - 1$ whenever $X_{(1)} = X_{(2)}$ and $X_{(1)} > 0$ has better performance, but we used $T$ for illustrative purposes.*

Finally, we note that applying the ancillarity principle can be suboptimal according to strictly frequentist criteria: in practice, there are cases where an unconditional test is preferable to a conditional test, as in the following example inspired by an Example in Cox (1958).

**Example 8.** *Suppose a production line is periodically tested to see if it is operating correctly. If correct, it produces a part of diameter 1. Periodically it goes out of line and then produces parts with diameter 1.1. In the testing, the parts are measured with one of two measuring instruments, an old one which produces a normal observation with mean the true diameter of the part and standard deviation 0.1, and a new measuring instrument which produces a normal observation with mean the true diameter and standard deviation 0.05. The old and new measuring instruments are each available with probability 1/2 (as there is another production line for which they are also used). If the production line is deemed to be out of line, it must be shut*

*down and reset, at considerable expense. The company does a cost-benefit analysis and determines that it will be optimal to control overall Type I error in the testing at the 0.05 level. This is a scenario in which frequentist analysis is absolutely appropriate, in that there is true long-term repetition of the test. Also, the cost-benefit analysis is presumably carried out in a Bayes-frequentist sense, since historical levels of in-line and out-of-line must be taken into account. If the company followed WCP, they would do the 0.05 level test conditional on which measuring instrument is being used at each test. But this will lose the company money, as the power of this test for detecting an out-of-line process (which is 0.646) is 9% less than that of the most powerful test (which is 0.694). This most powerful test corresponds to using Type I error probabilities of 0.099 and 0.001 for the old and new measuring instruments, respectively.*

The example above is interesting in that it suggests that, for frequentists, the only way to implement the conditionality principle is to use a method that is compatible with Bayesian reasoning (as the unconditional test would be equivalent to the Bayes rule with respect to the loss function implied by the cost-benefit analysis). This is not surprising, given the complete class theorems that show that optimal frequentist decision procedures are necessarily Bayesian.

## 5.5 Conclusions

The articles Evans (2013) and Mayo (2014) contain thought-provoking discussions about the conditions under which the result in Birnbaum (1962) is valid, but neither of them show that WCP and SP do not imply LP according to the definitions in Section 5.1, which, in our view, are equivalent to the definitions in Birnbaum (1962).

Evans (2013) avoids introducing **Ev**, which is central in Birnbaum's argument, and defines statistical principles as set relations on the (product) space of infer-

ences. If **Ev** is introduced with the property $\mathbf{Ev}(E, x) = \mathbf{Ev}(E', x')$ if and only if $(E, x) \sim_C (E', x')$ or $(E, x) \sim_S (E', x')$, Birnbaum's result follows. If we stick to Evans' framework, the union of the set relation defined by the sufficiency principle ($S$) and the conditionality principle ($C$) does not equal the set relation defined by the likelihood principle ($L$). This result might seem surprising at first glance but, if it were true, two inference bases with proportional likelihoods would be equivalent according to either the sufficiency principle or the conditionality principle individually, which is clearly false. What is true is that the smallest equivalence relation generated by $S \cup C$ equals $L$. As explained in Section 4, the equivalence relation generated by $S \cup C$ only contains inference bases that are equivalent to a chain of applications of the principles.

Mayo (2014) defines statistical principles making a distinction between the output of methods (**M**) and the inferences that are made by an agent using them (**Ev**): the weak conditionality principle is defined as a property of **Ev**, whereas the sufficiency principle is defined as a property of **M**. This distinction is not inconsequential. For example, in a mixture experiment where a Negative Binomial or Binomial experiment is selected with equal probability, reporting the conditional $p$-value does not result in a violation of the sufficiency principle (see Example 5). In the framework of Mayo (2014), the weak conditionality principle and the sufficiency principle do not imply the likelihood principle, but the definition of the sufficiency principle differs from that in Birnbaum (1962).

# 6

# Conclusions

We conclude the dissertation with a summary of our findings and pointers to ongoing and future work.

## 6.1 Constrained empirical Bayes priors on regression coefficients

In Chapter 2, we showed that constrained empirical Bayes procedures can avoid problematic behavior that unrestricted empirical Bayes procedures have, such as inconsistency under the null model and generally favoring bigger models over more parsimonious ones. They can also avoid some of the issues associated with the proper Bayesian lower bounds: in Section 2.2, the empirical Bayes procedure is information consistent while the lower bound is not, and in Example 1 (and the similation study in Section 2.2.3) the empirical Bayes procedure is less affected by collinearity than the lower bound. In Section 2.5, we used informative lower bounds to our advantage, by constraining the scales of the parameters in an orthogonal expansion to be decreasing.

It would be interesting to quantify the robustness to multicollinearity of the empirical Bayes procedure described in Section 2.2 more generally, and contrast it to that of the lower bound and Zellner-Siow's prior (Zellner and Siow, 1980). Another

avenue for future work is considering restricted empirical Bayes procedures in other contexts, such as inclusion probabilities. In Scott and Berger (2010) it is shown that unrestricted empirical Bayes procedures are problematic in this context, but can we define contrained empirical Bayes procedures that are well-behaved?

## 6.2 On information consistency

The concept of information consistency has been present in the Bayesian hypothesis testing literature since at least Jeffreys (1939), but under different definitions. In Section 3.5, we have shown that the definitions can disagree, and that a definition stated in terms of the likelihood ratio can be problematic (in the sense that the limit may not exist). We believe that the most natural definition is the one given in Som et al. (2016), which is stated in terms of population parameters.

The main message of the chapter is that information inconsistency is pervasive in normal linear models if conjugate priors are used, but it can be avoided using empirical Bayes approaches and thick-tailed priors.

Further work is needed on how to extend the definition of information consistency beyond normal linear models. The work presented in Chapter 3 can be extended to larger classes of priors (such non-local priors) or, in general, information consistency could be studied under a decision-theoretical framework similar to the one devised in Chapter 4.

## 6.3 On limit consistency

In Chapter 4, we elaborated on the work in Ly (2017) on limit consistency, which we believe to be a valuable criterion to determine the adequacy of prior specifications in two-sample problems.

In the two normal means problem, it is important to choose a parametrization such that the null and alternative hypothesis share a common parameter. If the

prior on the common parameter is flat, what we called "effect-size", baseline, and sum-to-zero parametrizations are equivalent and limit consistency holds. The correspondence between effect-size and baseline parametrizations is a well-known consequence of our choice of priors; however, the correspondence between the sum-to-zero parametrization and the others is, to the best of our knowledge, less well-known. If the prior on the common parameter is a (proper) normal prior, the parametrizations are not equivalent and limit consistency does not hold in general. Lastly, we showed that empirical Bayes priors, mixtures of $g-$priors, and non-local priors satisfy limit consistency insofar they are built upon our recommendations.

We are currently working on extending the results to more than two groups and to other two-sample tests, such as comparing two variances, proportions, etc. As we mentioned in Section 4.4, we would also like to investigate whether the non-parametric two-sample tests in Holmes et al. (2015) and Soriano and Ma (2017) are limit-consistent.

## 6.4 On Birnbaum's theorem

The main goal of the chapter was dicussing the articles Evans (2013) and Mayo (2014).

In Evans (2013), it is claimed that the proof in Birnbaum (1962) includes a tacit condition that is undesirable. In Section 5.2, we argued that the condition is rather explicit and not unreasonable: extending a union of statistical principles (as set relations) to the smallest equivalence relation containing them is equivalent to accepting that the principles can be applied successively, even if the result of the chain of equivalences leads to considering that two inference bases are equivalent even if they are not equivalent according to the application of one of the principles alone.

On the other hand, Mayo (2014), claims that the proof in Birnbaum (1962) is

incorrect. We argued that the disagreement stems from introducing a distinction between informative inference and the "output" of statistical methods.

In Section 5.4, we argued that applying the ancillarity principle in frequentist statistics can be difficult. First, it is not obvious how to choose among them when there is more than one maximal ancillary, but inferences depend on the choice of ancillary. Second, one can construct examples where the dependence on the parameter can be made arbitrarily small (see Example 7), so we wonder if it is sensible to have such a hard rule on what one is allowed to condition on.

At this point, we believe that it would be useful to write an article classifying the types of existent criticisms to the likelihood principle. Another possible, but admittedly difficult, area of investigation would be studying the consequences of a weaker version of the likelihood principle that includes objective Bayesian analyses.

# Appendix A

## Proofs for Chapter 2

**Proposition 1.** *For $n > p + p_0$, the solution to the optimization problem*

$$\begin{aligned} &\text{maximize } m_W(y) \\ &\text{subject to } W \succeq n(X'X)^{-1} \end{aligned}$$

*can be written as*

$$\begin{aligned} \widehat{W} &= a\,\widehat{\beta}\widehat{\beta}' + n(X'X)^{-1} \\ a &= \max(0, (n - p_0 - 1)/\mathsf{SSE} - (n+1)/\mathsf{SSR}) \\ \mathsf{SSR} &= \widehat{\beta}'(X'X)\widehat{\beta}. \end{aligned}$$

*Proof.* The structure of the proof is similar to that of Proposition 3.1 in DasGupta and Studden (1989). Let $\Sigma = (X'X)^{-1} + W$. Then, the restriction $W \succeq n(X'X)^{-1}$ is equivalent to $\Sigma \succeq (n+1)(X'X)^{-1}$ and the marginal likelihood can be rewritten as

$$m(Y \mid \Sigma) \propto |\Sigma|^{-1/2}(\mathsf{SSE} + \widehat{\beta}'\Sigma^{-1}\widehat{\beta})^{-(n-p_0)/2}.$$

Now, let $A = (X'X)^{1/2}\Sigma(X'X)^{1/2}$, $b = (X'X)^{1/2}\widehat{\beta}$. Then, the restriction is $A \succeq (n+1)I_n$ and

$$m(Y|A) \propto |A|^{-1/2}(\mathsf{SSE} + b'A^{-1}b)^{-(n-p_0)/2}.$$

Let $P$ be an orthogonal matrix such that

$$P = \begin{pmatrix} P_{1\cdot} \\ P_{2\cdot} \\ \vdots \\ P_{p-1\,\cdot} \\ b/\|b\| \end{pmatrix}$$

where $\|P_{i\cdot}\|^2 = 1$ and $P_{i\cdot}' P_{j\cdot} = 0$ for $i \neq j$ ($P_{i\cdot}$ can be constructed via Gram-Schmidt). This implies that $Pb = (0, 0, \ldots, ||b||)'$ and

$$b'A^{-1}b = (Pb)'PA^{-1}P'(Pb)$$

If we define

$$R = PA^{-1}P' = \begin{pmatrix} R_{11} & u' \\ u & r_{pp} \end{pmatrix}.$$

Then, $b'A^{-1}b = r_{pp}\|b\| = r_{pp}\mathsf{SSR}$, so we can rewrite

$$m(Y \mid R) \propto |R|^{1/2}(\mathsf{SSE} + \mathsf{SSR}\, r_{pp})^{-(n-p_0)/2},$$

with the restriction $R \preceq I_p/(n+1)$. The determinant is $|R| = |R_{11}|(r_{pp} - u'R_{11}^{-1}u)$. For fixed $r_{pp}$, DasGupta and Studden (1989) show that the maximizers are $u = 0_{1\times(p-1)}$ and $|R_{11}| = I_{p-1}/(n+1)$. Now, we can find the optimal $r_{pp}$ by maximizing $\log(r_{pp}) - (n-p_0)\log(\mathsf{SSE} + \mathsf{SSR}\, r_{pp})$ subject to $r_{pp} \leq 1/(n+1)$. The solution is $r_0 = \min(1/(n+1), \mathsf{SSE}/[(n-p_0-1)\mathsf{SSR}])$. Therefore, $\widehat{\Sigma} = (X'X)^{-1/2}(P'RP)^{-1}(X'X)^{-1/2}$. First, note that

$$R = I_p/(n+1) + \begin{pmatrix} 0 & 0 \\ 0 & r_0 - 1/(n+1) \end{pmatrix}$$

and

$$P' \begin{pmatrix} 0 & 0 \\ 0 & r_0 - 1/(n+1) \end{pmatrix} P = [r_0 - 1/(n+1)]bb'/\mathsf{SSR},$$

which implies

$$P'RP = I_p/(n+1) + [r_0 - 1/(n+1)]bb'/\mathsf{SSR}.$$

Then, we can compute $(P'RP)^{-1}$ with the Sherman-Morrison formula and easily find $\widehat{W} = n(X'X)^{-1} + a\widehat{\beta}\widehat{\beta}'$, as reqired.

□

**Proposition 2.** *Let $\mathcal{M}_\gamma$ be the model with conditional mean $X_0\beta_0 + X_\gamma\beta_\gamma$, where $X_\gamma$ are the model-specific predictors and $X_0$ are the common predictors, with $X_0'X_\gamma = 0_{p_0\times p}$ (if $\gamma = \emptyset$, we recover the null model $\mathcal{M}_0$). If $\mathcal{M}_j \supset \mathcal{M}_i$, then*

$$\mathsf{BF}_{ji,\mathsf{BIC}} \geq \mathsf{BF}_{ji,\mathsf{ML}} \geq \mathsf{BF}_{ji,\mathsf{LB}},$$

*where $\mathsf{BF}_{ji,\mathsf{BIC}} = n^{-(p_j-p_i)/2}[(1-R_i^2)/(1-R_j^2)]^{-n/2}$. Let $\mathcal{M}_f$ be the full model (which includes all $p$ predictors) and $\mathcal{M}_0$ be the null model. If the prior on the model space is the same in all cases, the inequality above implies*

$$\mathbb{P}_{\mathsf{BIC}}(\mathcal{M}_f \mid Y) \geq \mathbb{P}_{\mathsf{ML}}(\mathcal{M}_f \mid Y) \geq \mathbb{P}_{\mathsf{LB}}(\mathcal{M}_f \mid Y)$$
$$\mathbb{P}_{\mathsf{BIC}}(\mathcal{M}_0 \mid Y) \leq \mathbb{P}_{\mathsf{ML}}(\mathcal{M}_0 \mid Y) \leq \mathbb{P}_{\mathsf{LB}}(\mathcal{M}_0 \mid Y).$$

Case $\mathcal{M}_i = \mathcal{M}_0$

**LB vs type II ML.** If $R_j^2 \leq (n+1)/(2n-p_0) = \ell$, the type II ML and the LB are equal, so the inequality is satisfied. If $R_j^2 \geq \ell$, then

$$\mathsf{BF}_{j0,\mathsf{ML}} = \frac{\max_{W\succeq n(X'X)^{-1}} m_j(Y)}{m_0(Y)} \geq \frac{m_{j,W=n(X'X)^{-1}}(Y)}{m_0(Y)} = \mathsf{BF}_{j0,\mathsf{LB}},$$

as required.

**Type II ML vs BIC.** If $R_j^2 \leq (n+1)/(2n-p_0)$, we must show that

$$n^{-p_j}(1-R_j^2)^{-n} \geq (n+1)^{n-p_0-p_j}[1+n(1-R_j^2)]^{-(n-p_0)}.$$

First, note that $n^{-p_j} \geq (n+1)^{-p_j}$. Then, it suffices to show that

$$\left[\frac{1}{1-R_j^2}\right]^n \geq \left[\frac{n+1}{1+n(1-R_j^2)}\right]^{n-p_0},$$

which is satisfied because the base on the LHS is greater than the base on the RHS. Now, if $R_j^2 \geq (n+1)/(2n-p_0)$, we must show that

$$n^{-p_j}(1-R_j^2)^{-n} \geq \varphi_j(n)^{-1}\left[\frac{1-R_j^2}{R_j^2}\right](1-R_j^2)^{-(n-p_0)}.$$

Clearly, $(1-R_j^2)^{-n} \geq (1-R_j^2)^{-(n-p_0)}$, so it suffices to show that

$$n^{-p_j} \geq \varphi_j(n)^{-1}\left[\frac{1-R_j^2}{R_j^2}\right].$$

Since $R_j^2 \geq (n+1)/(2n-p_0)$, we have

$$\left[\frac{1-R_j^2}{R_j^2}\right] \leq \frac{n-p_0-1}{n+1}.$$

The result follows because $n^{-p_j} \geq (n+1)^{-p_j}\left[(n-p_0-1)/(n-p_0)\right]^{n-p_0}$.

Case $\mathcal{M}_i \neq \mathcal{M}_0$

**LB vs type II ML.** Let $\ell = (n+1)/(2n-p_0)$. We consider 3 cases: (1) $R_j^2 \leq \ell$; (2) $R_i^2 \leq \ell$ and $R_j^2 \geq \ell$; (3) $R_i^2 \geq \ell$. In case (1), the type II ML and LB are the same, so the inequalities are trivially satisfied. In case (2), $\mathsf{BF}_{j0,\mathsf{ML}} > \mathsf{BF}_{j0,\mathsf{LB}}$, so

$$\mathsf{BF}_{ji,\mathsf{ML}} = \mathsf{BF}_{j0,\mathsf{ML}}/\mathsf{BF}_{i0,\mathsf{LB}} \geq \mathsf{BF}_{j0,\mathsf{LB}}/\mathsf{BF}_{i0,\mathsf{LB}} = \mathsf{BF}_{ji,\mathsf{LB}}.$$

Finally, in case (3) we have to show that

$$\left[\frac{\varphi_i(n)}{\varphi_j(n)}\right]^{1/2}\left[\frac{R_i^2}{R_j^2}\right]^{1/2}\left[\frac{1-R_i^2}{1-R_j^2}\right]^{(n-p_0-1)/2} \geq (n+1)^{(p_i-p_j)/2}\left[\frac{1+n(1-R_i^2)}{1+n(1-R_j^2)}\right]^{(n-p_0)/2}$$

We have $\varphi_i(n)/\varphi_j(n) = (n+1)^{(p_i-p_j)/2}$, so it suffices to show that

$$\frac{R_i^2}{R_j^2}\left[\frac{1-R_i^2}{1-R_j^2}\right]^{n-p_0-1} \geq \left[\frac{1+n(1-R_i^2)}{1+n(1-R_j^2)}\right]^{n-p_0},$$

which is true. Indeed, let $x = R_i^2$ and $y = R_j^2$. We need to show that

$$\frac{x}{y}\left[\frac{1-x}{1-y}\right]^{n-p_0-1} \geq \left[\frac{1+n(1-x)}{1+n(1-y)}\right]^{n-p_0}, \ell \leq x \leq y \leq 1$$

For any value of $x$, the inequality is satisfied for any $y \to 1^-$, so we only need to show the inequality for $\ell \leq x \leq y < 1$. Taking logarithms,

$$\log\left(\frac{x}{1-x}\right)+(n-p_0)\log\left(\frac{1-x}{1+n(1-x)}\right) \geq \log\left(\frac{y}{1-y}\right)+(n-p_0)\log\left(\frac{1-y}{1+n(1-y)}\right),$$

and the function

$$f(x) = \log\left(\frac{x}{1-x}\right) + (n-p_0)\log\left(\frac{1-x}{1+n(1-x)}\right)$$

is decreasing for $\ell \leq x < 1$, so the result follows.

**Type II ML vs BIC.** Again, let $\ell = (n+1)/(2n-p_0)$ and consider the cases (1) $R_j^2 \leq \ell$; (2) $R_i^2 \leq \ell$ and $R_j^2 \geq \ell$; (3) $R_i^2 \geq \ell$. In the first case, $\mathsf{BF}_{ji,\mathsf{ML}} = \mathsf{BF}_{ji,\mathsf{LB}}$, so we need to show

$$n^{(p_i-p_j)/2}\left(\frac{1-R_j^2}{1-R_i^2}\right)^{-n/2} \geq (n+1)^{(p_i-p_j)/2}\left(\frac{1+n(1-R_j^2)}{1+n(1-R_i^2)}\right)^{-(n-p_0)/2}.$$

Since $n^{(p_i-p_j)/2} > (n+1)^{(p_i-p_j)/2}$, it suffices to show that

$$\left(\frac{1-R_i^2}{1-R_j^2}\right)^{n} \geq \left(\frac{1+n(1-R_i^2)}{1+n(1-R_j^2)}\right)^{n-p_0}.$$

The bases on the LHS and RHS are both greater than 1 because $1-R_i^2 \geq 1-R_j^2$. Then, since the exponent on the LHS is greater than the exponent on the RHS, it suffices to show that

$$\frac{1-R_i^2}{1-R_j^2} \geq \frac{1+n(1-R_i^2)}{1+n(1-R_j^2)},$$

which is clearly satisfied. In case (2), we have to show that

$$n^{p_i-p_j}\left[\frac{1-R_i^2}{1-R_j^2}\right]^n \geq \left[\frac{(n+1)^{p_i+p_0-n}}{\varphi_j(n)}\right]\left[\frac{1-R_j^2}{R_j^2}\right]\left[\frac{1+n(1-R_i^2)}{1-R_j^2}\right]^{n-p_0}.$$

First, note that

$$n^{p_i-p_j} \geq (n+1)^{p_i-p_j+1}\frac{(n-p_0-1)^{n-p_0-1}}{(n-p_0)^{n-p_0}}\left[\frac{1-R_j^2}{R_j^2}\right]$$

because $(1-R_j^2)/R_j^2 \leq (n-p_0-1)/(n+1)$, so the result follows because clearly

$$\frac{1-R_i^2}{1-R_j^2} \geq \frac{1+n(1-R_i^2)}{(n+1)[1-R_j^2]}.$$

Finally, in case (3) we have to show that

$$n^{p_i-p_j}\left[\frac{1-R_i^2}{1-R_j^2}\right]^n \geq \left[\frac{\varphi_i(n)}{\varphi_j(n)}\right]\left[\frac{R_i^2}{R_j^2}\right]\left[\frac{1-R_i^2}{1-R_j^2}\right]^{n-p_0-1},$$

which is satisfied because $n^{p_i-p_j} \geq (n+1)^{p_i-p_j}$ and $(1-R_i^2)/(1-R_j^2) \geq R_i^2/R_j^2$.

Posterior probabilities

If $\mathcal{M}_f$ is the full model,

$$\mathbb{P}(\mathcal{M}_f \mid Y) = \frac{\mathbb{P}(\mathcal{M}_f)}{\mathbb{P}(\mathcal{M}_f) + \sum_\gamma \mathbb{P}(\mathcal{M}_\gamma)\mathsf{BF}_{\gamma f}}.$$

From this expression and the fact that for any $\gamma$

$$\mathsf{BF}_{f\gamma,\mathsf{BIC}} \geq \mathsf{BF}_{f\gamma,\mathsf{ML}} \geq \mathsf{BF}_{f\gamma,\mathsf{LB}},$$

it follows that

$$\mathbb{P}_{\mathsf{BIC}}(\mathcal{M}_f \mid Y) \geq \mathbb{P}_{\mathsf{ML}}(\mathcal{M}_f \mid Y) \geq \mathbb{P}_{\mathsf{LB}}(\mathcal{M}_f \mid Y).$$

An analogous argument can be used to show that

$$\mathbb{P}_{\mathsf{BIC}}(\mathcal{M}_0 \mid Y) \leq \mathbb{P}_{\mathsf{ML}}(\mathcal{M}_0 \mid Y) \leq \mathbb{P}_{\mathsf{LB}}(\mathcal{M}_0 \mid Y),$$

noting that the null model is nested in all the other models.

**Proposition 3.** *Let $p \geq 3$ and $n > p + p_0$. The estimator $\tilde{\beta} = \mathbb{E}(\beta \mid Y)$ is minimax with respect to the (scaled) squared predictive loss*

$$L(\beta, \delta) = (\beta - \delta)'(X'X)(\beta - \delta)/\sigma^2$$

*Proof.* Let $Z = (X'X)^{1/2}\widehat{\beta}$, $\theta = (X'X)^{1/2}\beta$, and $\tilde{\theta} = (X'X)^{1/2}\tilde{\beta}$. Then, $L(\beta, \tilde{\beta})$ becomes $L(\theta, \tilde{\theta}) = \|\theta - \tilde{\theta}\|^2/\sigma^2$, with $\tilde{\delta} = (1 - r(F, s)/F)Z$ for

$$r(F, s) = \begin{cases} \frac{F}{n+1} & \text{if } F \leq \frac{n}{n-p_0} \\ \frac{1}{n-p_0-1} & \text{if } F > \frac{n}{n-p_0} \end{cases},$$

where $F = \|Z\|^2/\mathsf{SSE}$, $\mathsf{SSE} \sim \sigma^2\chi^2_{n-p-p_0}$, and $\mathsf{SSE}$ is independent of $Z$. For any fixed $F$, $r(\cdot, \cdot)$ does not depend on $s$. Now, we show that for any fixed $s$, $r(F, s)$ is monotone nondecreasing in $F$. Clearly, $r$ is monotonely increasing in $F$ for $F \leq n/(n - p_0)$, and $n/(n - p_0)$ is the maximum value it can take on in the first case. Since we have the inequality

$$\frac{1}{n - p_0 - 1} > \frac{n}{(n - p_0)(n + 1)}$$

for $n > p_0/2$, which is implied by the assumption $n > p + p_0$, it follows that $r$ is monotonely nondecreasing in $F$. It remains to show that $0 \leq r(\cdot, \cdot) \leq 2(p - 2)(n - p - p_0 + 2)$. It is clear that $r(\cdot, \cdot) \geq 0$, and, finally, $r(\cdot, \cdot) \leq 2(p - 2)(n - p - p_0 + 2)$ holds for $n \geq p_0 - 1$ for $p \geq 3$, which is implied by the assumption that $n > p + p_0$. By the characterization in Strawderman (1973), it follows that $\tilde{\beta}$ is minimax with respect to $L(\beta, \delta)$. □

**Proposition 4.** *The covariance matrix that maximizes the approximate marginal likelihood $\tilde{m}(y)$ is*

$$\widehat{W} = nA + a\,\widehat{\beta}\widehat{\beta}',$$

*where*

$$a = \max\left(0, 1 - \frac{n+1}{Q}\right), \quad Q = \widehat{\beta}'A^{-1}\widehat{\beta}.$$

*Proof.* It can be proved following the same steps as in Proposition 1, but with the approximate marginal likelihood instead. Let $\Sigma = A + W$. Then, the restriction $W \succeq cA$ is equivalent to $\Sigma \succeq (c+1)A$. Now, let $B = A^{-1/2}\Sigma A^{-1/2}$ and $b = A^{-1/2}\widehat{\beta}$. We can rewrite the restriction as $B \succeq (c+1)I_p$ and

$$\tilde{m}(Y) \propto |B|^{-1/2} \exp\left(-\frac{1}{2} b' B^{-1} b\right).$$

Let $P$ be an orthogonal matrix such that $Pb = (0, 0, \ldots, \|b\|)'$, so

$$b' B^{-1} b = (Pb)' P B^{-1} P' (Pb).$$

If we define

$$R = P B^{-1} P' = \begin{pmatrix} R_{11} & u' \\ u & r_{pp} \end{pmatrix},$$

we can rewrite

$$b' B^{-1} b = \|b\|^2 r_{pp} = Q r_{pp}.$$

The linear matrix restriction becomes $R \preceq I_p/(c+1)$ and

$$\tilde{m}(Y) \propto |R|^{1/2} \exp\left(-\frac{1}{2} Q r_{pp}\right)$$

The determinant is $|R| = |R_{11}|(r_{pp} - u' R_{11}^{-1} u)$. As argued in DasGupta and Studden (1989), for fixed $r_{pp}$, the maximizers are $u = 0$ and $|R_{11}| = I_{p-1}/(c+1)$. Then, we can find $r_{pp}$ by maximizing

$$\log r_{pp} - Q r_{pp}$$

subject to $r_{pp} \leq 1/(c+1)$, which yields $r_{pp} = \max(1/(c+1), 1/Q)$. Therefore,

$$\Sigma = A^{1/2} (P' R P)^{-1} A^{1/2}$$

We can compute $(P'RP)^{-1}$ using the Sherman-Morrison formula:

$$(P'RP)^{-1} = (c+1)I_p + \alpha A^{-1/2} \widehat{\beta}\widehat{\beta}' A^{-1/2},$$

so

$$\Sigma = (c+1)A + \alpha\widehat{\beta}\widehat{\beta}'.$$

□

## A.1 Other proofs

### *A.1.1 Normal linear models*

**Proposition 5.** ***Model selection consistency.*** *Let the true model be* $\mathcal{M}_i : Y = X_0\beta_0 + X_i\beta_i + \epsilon_i$, $\epsilon_i \sim N_n(0, \sigma^2 I_n)$ *for some* $X_i$ *which includes a subset of predictors in* $X$. *For any* $\mathcal{M}_j \not\supset \mathcal{M}_i$, *assume*

$$\lim_{n\to\infty} \frac{\beta_i' X_i'(I - \mathsf{P}_{X_j})X_i\beta_i}{n} = b_j \in (0, \infty). \tag{A.1}$$

*Then, the type II ML prior is model selection consistent: that is,* $\mathbb{P}(\mathcal{M}_i \,|\, Y) \to 1$ *in probability as* $n \to \infty$.

*Proof.* The proof is essentially immediate given Proposition 2. The lower bound is consistent (Fernandez et al., 2001) and so is BIC (see e.g. Claeskens et al. (2008)). If the true model is the null model $\mathsf{BF}_{\mathsf{ML},i0} \leq \mathsf{BF}_{\mathsf{BIC},i0} \to 0$ in probability for any model $\mathcal{M}_i$, so the procedure is consistent. If the true model $\mathcal{M}_j$ is not the null model, we consider 2 cases:

- $\mathcal{M}_i \supset \mathcal{M}_j$ or $\mathcal{M}_i \subset \mathcal{M}_j$ : In either case, $\mathsf{BF}_{ij} \to 0$ follows directly by Proposition 2.

- $\mathcal{M}_i \not\supset \mathcal{M}_j$ and $\mathcal{M}_i \not\subset \mathcal{M}_j$ : If $\mathcal{M}_i \cap \mathcal{M}_j \neq \emptyset$, both $R_i^2 \to r_i$ and $R_j^2 \to r_j$ in probability with $0 < r_i < r_j < 1$ (Guo and Speckman, 2009). If $r_j \leq 1/2$, then $\mathsf{BF}_{\mathsf{ML},ij} = \mathsf{BF}_{\mathsf{LB},ij}$ and consistency follows. If $r_i > 1/2$, both models are in the unrestricted case, and consistency follows because $[(1 - R_j^2)/(1 - R_i^2)]^{(n-p_0-1)/2}$ converges to 0 exponentially fast. In the intermediate case $r_i < 1/2$ and $r_j >$

1/2, we have $\mathsf{BF}_{\mathsf{ML},ij} \leq \mathsf{BF}_{\mathsf{LB},ij} \to 0$ in probability. If $\mathcal{M}_i \cap \mathcal{M}_j = \emptyset$, then $R_i^2 \to 0$ in probability (Liang et al., 2008), so the relevant case is $\mathsf{BF}_{\mathsf{ML},i0} = \mathsf{BF}_{\mathsf{LB},i0}$ in which case $\mathsf{BF}_{\mathsf{ML},ij} \leq \mathsf{BF}_{\mathsf{LB},ij} \to 0$.

□

**Proposition 6.** ***Minimal training sample size.*** *When $n = p + p_0$, the marginal likelihood is finite for any given $W$, but one can choose $W \succeq n(X'X)^{-1}$ so that the marginal goes to $\infty$*

*Proof.* Let $n = p + p_0$. As a function of $W$, the marginal likelihood is proportional to

$$f_1(W) = |(X'X)^{-1} + W|^{-1/2}[\widehat{\beta}'[W + (X'X)^{-1}]^{-1}\widehat{\beta}]^{-\frac{p}{2}}$$

Define $\Sigma = (X'X)^{-1} + W$, and consider its eigendecomposition $\Sigma = ODO'$, where $O$ is orthogonal and $D = \text{diag}(d_1, d_2, \ldots, d_p)$. Therefore, maximizing $f_1$ is equivalent to minimizing

$$\prod_{i=1}^{p} d_i^{1/p} \sum_{i=1}^{p} \frac{\tilde{\beta}_i^2}{d_i},$$

where $\tilde{\beta} = O'\widehat{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p)'$. We can construct a matrix $\Sigma$ that satisfies the linear matrix inequality constraint $\Sigma \succeq (n+1)(X'X)^{-1}$ and makes $f_1$ infinite by taking $d_1 \to \infty$, $d_j \geq n + 1$ for $2 \leq j \leq p$, and choosing $O$ such that

$$O'\widehat{\beta} = (\|\widehat{\beta}\|, 0, \ldots, 0)'.$$

Then, we can construct $O$ as follows. Let $w = (\|\widehat{\beta}\|, 0, \ldots, 0)'$; then, define $u = (\widehat{\beta} - w)/\|\widehat{\beta} - w\|^2$ and finally, $O' = I_p - 2uu'$. □

**Proposition 7.** ***Invariance.*** *Let $X$ be an $n \times p$ matrix and $H$ be a an invertible $p \times p$ matrix. Define $\tilde{X} = XH$. For a fixed outcome $Y$, we use the notation $\widehat{W}_Z$ to denote the estimated prior covariance if the design matrix is $Z$, which is obtained by*

*solving the optimization problem in Proposition 1. Then,* $H\tilde{\beta} \sim N_p(0_p, \sigma^2 H W_{\tilde{X}} H')$ *and* $\beta \sim N_p(0_p, \sigma^2 W_X)$ *are equal in distribution.*

*Proof.* We need to show that $W_X = H W_{\tilde{X}} H'$. First, note that $R^2$ is invariant with respect to invertible linear transformations. Indeed, it only depends on $X$ through $\mathsf{P}_X$, and for $\tilde{X} = XH$

$$\mathsf{P}_{\tilde{X}} = XH(H'X'XH)^{-1}H'X = X(X'X)^{-1}X' = \mathsf{P}_X.$$

If $R^2 \leq (n+1)/(2n - p_0)$, $W_X$ and $W_{\tilde{X}}$ are equal to their respective lower bounds, and

$$H' W_{\tilde{X}} H = nH(H'X'XH)^{-1}H' = n(X'X)^{-1} = W_X,$$

as required. If $R^2 > (n+1)/(2n-p_0)$, $a_X = a_{\tilde{X}}$ because $\mathsf{P}_X = \mathsf{P}_{\tilde{X}}$ (see above). Lastly,

$$H\widehat{\beta}_{\tilde{X}}\widehat{\beta}'_{\tilde{X}}H' = [H(H'X'XH)^{-1}H'XY][H(H'X'XH)^{-1}H'XY]' = \widehat{\beta}_X\widehat{\beta}'_X,$$

so the result follows. □

### *A.1.2 Generalized linear models*

**Proposition 8.** ***Invariance.*** *The constrained empirical Bayes prior introduced in Section 2.3 is invariant to linear transformations of the design matrix. Let* $H$ *be an invertible* $p \times p$ *matrix and* $\tilde{X} = XH$*. Let* $\tilde{\beta}$ *be the prior if the design matrix is* $\tilde{X}$*. Then* $H\tilde{\beta}$ *and* $\beta$ *are equal in distribution.*

*Proof.* Let $W_X$, $W_{\tilde{X}}$ be the prior covariances if the design matrices are $X$ and $\tilde{X}$, respectively. We want to show that $H W_{\tilde{X}} H' = W_X$. That is,

$$cHA_{\tilde{X}}H' + a(H\widehat{\beta}_{\tilde{X}})(H\widehat{\beta}_{\tilde{X}})' = cHA_XH' + a\widehat{\beta}_X\widehat{\beta}'_X$$

From Li and Clyde (2015), we know that $HA_{\tilde{X}}H' = A_X$. Maximum likelihood estimators are invariant with respect to linear transformations, so $H\widehat{\beta}_{\tilde{X}} = \widehat{\beta}_X$. Finally,

$a$ depends on the design matrix only through $Q$, which satisfies $Q_{\tilde{X}} = Q_X$. Indeed,

$$\begin{aligned} Q_{\tilde{X}} &= \widehat{\beta}'_{\tilde{X}} A_{\tilde{X}}^{-1} \widehat{\beta}_{\tilde{X}} \\ &= \widehat{\beta}'_X (H A_{\tilde{X}} H')^{-1} \widehat{\beta}_X \\ &= Q_X. \end{aligned}$$

The result follows. □

**Proposition 9.** ***Model selection consistency.*** *If, as $n \to \infty$, $c_n \to \infty$ and $c_n / \exp(an) \to 0$ for all $a > 0$, we have model selection consistency under the same conditions as in Li and Clyde (2015).*

*Proof.* We prove this result by cases, borrowing results from Li and Clyde (2015).

Case $\mathcal{M}_T \neq \mathcal{M}_\emptyset$: Assume $\mathcal{M} \not\supset \mathcal{M}_T$, then:

- By Lemma 4 in Li and Clyde (2015), the likelihood ratio is $O_{\mathbb{P}}(\exp(a_{\mathcal{M}} n))$ for $a_{\mathcal{M}} > 0$
- By Lemma 1 in Li and Clyde (2015), $[\mathcal{J}_{\mathcal{M}}(\widehat{\alpha}) / \mathcal{J}_{\mathcal{M}_T}(\widehat{\alpha})]^{1/2}$ is $O_{\mathbb{P}}(n^{-(-1-\tau_{\mathcal{M}})/2})$ for $0 \leq \tau_M \leq 1$.
- By Lemma 5 in Li and Clyde (2015), $Q_{\mathcal{M}_T} \in O_{\mathbb{P}}(n)$ and $Q_{\mathcal{M}} \in O_p(n^{\xi_{\mathcal{M}}})$, for $0 \leq \xi_{\mathcal{M}} \leq 1$.

Let $a_{\mathcal{M},\infty} = \text{plim}_{n\to\infty} \max(0, 1-(c_n+1)/Q_{\mathcal{M}})$, then $a_{\mathcal{M},\infty} \neq 0$ whenever $\text{plim}_{n\to\infty} Q_{\mathcal{M}}/(c_n+1) = K > 1$ or $\text{plim}_{n\to\infty} Q_{\mathcal{M}}/(c_n+1) = \infty$; alternatively, $\text{plim}_{n\to\infty} a_{\mathcal{M},\infty} = 0$ whenever $\text{plim}_{n\to\infty} Q_{\mathcal{M}}/(c_n+1) = K < 1$ or $\text{plim}_{n\to\infty} Q_{\mathcal{M}}/(c_n+1) = 0$ (the limits are guaranteed to exist by Lemma 5 in Li and Clyde (2015) and $c_n \to \infty$). With that in consideration, we can study the behavior of $\Omega_{\mathcal{M}_t,\mathcal{M}}$. The worst case is when $a_{\mathcal{M}_T,\infty} \neq 0$ and $a_{\mathcal{M},\infty} = 0$, where

$$\Omega_{\mathcal{M}_T,\mathcal{M}} \propto [Q_{\mathcal{M}_T}]^{-1/2} \exp\left\{ \frac{Q_{\mathcal{M}}}{2(c+1)} \right\} \in O_p(n^{-1/2})$$

Therefore, model selection consistency follows because the likelihood ratio is $O_p(\exp(a_{\mathcal{M}}n))$, $\exp(a_{\mathcal{M}}n)/(c_n+1)^{p/2} \to 0$ by our assumption, and the remaining terms go to zero (at most) at polynomial rate. Now, assume $\mathcal{M} \supset \mathcal{M}_T$, then:

- By Lemma 4 in Li and Clyde (2015), the likelihood ratio is $O_{\mathbb{P}}(1)$.
- By Lemma 1 in Li and Clyde (2015), $[\mathcal{J}_{\mathcal{M}}(\widehat{\alpha})/\mathcal{J}_{\mathcal{M}_T}(\widehat{\alpha})]^{1/2}$ is $O_{\mathbb{P}}(1)$.
- By Lemma 5 in Li and Clyde (2015), both $Q_{\mathcal{M}_T}/n$ and $Q_{\mathcal{M}}/n$ converge in probability to the same constant, and it is straightforward to show that $\Omega_{\mathcal{M},\mathcal{M}_T} \in O_{\mathbb{P}}(1)$

We have $p_{\mathcal{M}} > p_{\mathcal{M}_T}$, so $(c+1)^{-(p_{\mathcal{M}_T}-p_{\mathcal{M}})/2}$ combined with the results above implies consistency.

Case $\mathcal{M}_T = \mathcal{M}_\emptyset$ In this case:

- By Lemma 4 in Li and Clyde (2015), the likelihood ratio is $O_{\mathbb{P}}(1)$.
- By Lemma 1 in Li and Clyde (2015), $[\mathcal{J}_{\mathcal{M}}(\widehat{\alpha})/\mathcal{J}_{\mathcal{M}_T}(\widehat{\alpha})]^{1/2}$ is $O_{\mathbb{P}}(1)$.
- By Lemma 5 in Li and Clyde (2015), $Q_{\mathcal{M}} \in O_{\mathbb{P}}(1)$ and $\Omega_{\mathcal{M}_T,\mathcal{M}} \in O_{\mathbb{P}}(1)$.

Therefore, the term $(c+1)^{p_{\mathcal{M}}/2}$ combined with the results above implies consistency.

□

# Appendix B

## Proofs for Chapter 3

### B.1 Proof of Lemma 1

We show the result for $\widehat{\theta}$ first. Since $X_\gamma' \Sigma^{-1} X_\theta = 0_{\dim(\gamma)\times n}$, $\widehat{\theta} = \theta + A\epsilon$, where $A = (X_\theta' \Sigma X_\theta)^{-1} X_\theta' \Sigma^{-1}$. Then,

$$\begin{aligned}
\lim_{\|\theta\|\to\infty} \frac{\|\widehat{\theta}\|^2}{\|\theta\|^2} &= \lim_{\|\theta\|\to\infty} \frac{\|\theta\|^2 + \epsilon' A' A\epsilon + 2\epsilon' A'\theta}{\|\theta\|^2} \\
&\leq \lim_{\|\theta\|\to\infty} 1 + \frac{\epsilon' A' A\epsilon}{\|\theta\|^2} + \frac{\|2\epsilon' A'\|}{\|\theta\|} \\
&= 1.
\end{aligned}$$

Thus,

$$\lim_{\|\theta\|\to\infty} \|\widehat{\theta}\|^2 = \lim_{\|\theta\|\to\infty} \frac{\|\widehat{\theta}\|^2}{\|\theta\|^2} \|\theta\|^2 = \infty,$$

as required. Finally, $s_y^2$ is fixed because:

$$X_\gamma \widehat{\gamma} = X_\gamma \gamma + X_\gamma (X_\gamma' \Sigma^{-1} X_\gamma)^{-1} X_\gamma' \Sigma^{-1} \epsilon$$

$$X_\theta \widehat{\theta} = X_\theta \theta + X_\theta (X_\theta' \Sigma^{-1} X_\theta)^{-1} X_\theta' \Sigma^{-1} \epsilon.$$

So

$$Y - X_\gamma \widehat{\gamma} - X_\theta \widehat{\theta} = (I_n - X_\gamma (X_\gamma' \Sigma^{-1} X_\gamma)^{-1} X_\gamma' \Sigma^{-1} - X_\theta (X_\theta' \Sigma^{-1} X_\theta)^{-1} X_\theta' \Sigma^{-1}) \epsilon,$$

which does not depend on $\theta$.

## B.2 Proof of Lemma 3

Denote:

$$\begin{aligned}
\widehat{\theta} &= \left(X_\theta' \Sigma^{-1} X_\theta\right)^{-1} X_\theta' \Sigma^{-1} y \\
s_y^2 &= (y - X_\theta \widehat{\theta} - X_\gamma \widehat{\gamma})' \Sigma^{-1} (y - X_\theta \widehat{\theta} - X_\gamma \widehat{\gamma}) \\
\mathsf{SSE}_0 &= s_0^2 \nu_0 + s_y^2 \\
\mathsf{SSE}_1 &= s_1^2 \nu_1 + s_y^2 \\
\mathsf{SSR} &= \widehat{\theta}' X_\theta' \Sigma^{-1} X_\theta \widehat{\theta} \\
\mathcal{I}_\theta &= X_\theta' \Sigma^{-1} X_\theta \\
p_0 &= r_2 - \nu_0 \\
p_1 &= r_2 - \nu_1
\end{aligned}$$

Throughout, we use the following notation for functions $a, b$:

- $a(g, \widehat{\theta}) \lesssim b(g, \widehat{\theta})$ if and only if there exists $0 < M < \infty$ which doesn't depend on $g$ or $\widehat{\theta}$ such that $a(g, \widehat{\theta}) \leq M b(g, \widehat{\theta})$.
- $a(g, \widehat{\theta}) \gtrsim b(g, \widehat{\theta})$ if and only if there exists $0 < M < \infty$ which doesn't depend on $g$ or $\widehat{\theta}$ such that $a(g, \widehat{\theta}) \geq M b(g, \widehat{\theta})$.
- $a(g, \widehat{\theta}) \asymp b(g, \widehat{\theta})$ if and only if $a(g, \widehat{\theta}) \lesssim b(g, \widehat{\theta})$ and $a(g, \widehat{\theta}) \gtrsim b(g, \widehat{\theta})$.

Before we prove Lemma 3, we prove an auxiliary result

**Lemma 1.** *Let*

$$h(g) = |g\Omega + \mathcal{I}_\theta^{-1}|^{-1/2}[\mathsf{SSE}_1 + \widehat{\theta}'(g\Omega + \mathcal{I}_\theta^{-1})^{-1}\widehat{\theta}]^{-(n-p_1)/2},$$

*then, there exist* $0 < d_l < d_u < \infty$ *such that*

$$\frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}} \lesssim h(g) \lesssim \frac{(g+d_u)^{(n-p_1-r_1)/2}}{[(g+d_u)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}$$

*Proof:* Consider the matrix factorization

$$\mathcal{I}_\theta^{-1} + g\Omega = \Omega^{1/2}[\Omega^{-1/2}\mathcal{I}_\theta^{-1}\Omega^{-1/2} + gI_{r_1}]\Omega^{1/2},$$

and take the eigendecomposition $\Omega^{-1/2}\mathcal{I}_\theta^{-1}\Omega^{-1/2} = ODO'$, where $O$ is orthogonal and $D$ diagonal with elements $0 < d_l < d_i < d_u < \infty$. Then, we can rewrite

$$\mathcal{I}_\theta^{-1} + g\Omega = \Omega^{1/2}O[D + gI_{r_1}]O'\Omega^{1/2}.$$

We can bound

$$\widehat{\theta}'\Omega^{-1}\widehat{\theta}/(d_u+g) \leq \widehat{\theta}'(g\Omega + \mathcal{I}_\theta^{-1})^{-1}\widehat{\theta} \leq \widehat{\theta}'\Omega^{-1}\widehat{\theta}/(d_l+g)$$

and

$$|g\Omega + \mathcal{I}_\theta^{-1}|^{-1/2} \propto |D + gI_{r_1}|^{-1/2} \in [(g+d_u)^{-r_1/2}, (g+d_l)^{-r_1/2}],$$

so

$$h(g) \lesssim \frac{(d_u+g)^{(n-p_1)/2}(d_l+g)^{-r_1/2}}{[(d_u+g)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}} \lesssim \frac{(d_u+g)^{(n-p_1-r_1)/2}}{[(d_u+g)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}.$$

Similarly, we can find the lower bound

$$h(g) \gtrsim \frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}.$$

Now, we prove Lemma 3 arguing by cases.

Case $\nu_0 > \nu_1$ Applying the lower bound in Lemma 1,

$$B_{10} \gtrsim \frac{[\mathsf{SSE}_0 + \mathsf{SSR}]^{(n-p_0)/2}}{(\widehat{\theta}'\Omega^{-1}\widehat{\theta})^{(n-p_1)/2}} \int_0^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p_1)/2}}\pi(\mathrm{d}g).$$

Since $p_0 < p_1$, the term outside the integral goes to infinity as $\|\widehat{\theta}\|^2 \to \infty$, and by Fatou's lemma,

$$\liminf_{\|\widehat{\theta}\|^2\to\infty} \int_0^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p_1)/2}}\pi(\mathrm{d}g) \geq \int_0^\infty (g+d_l)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g),$$

which is clearly bounded away from 0 for any prior on $g$ with positive support, so any such prior yields an information-consistent $B_{10}$ whenever $\nu_0 > \nu_1$.

Case $\nu_0 = \nu_1$ Applying the lower bound in Lemma 1 and Fatou's lemma as we did for the case $\nu_0 > \nu_1$:

$$\lim_{\|\widehat{\theta}\|^2\to\infty} B_{10} \gtrsim \int_0^\infty (g+d_l)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g) \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{[\mathsf{SSE}_0 + \mathsf{SSR}]^{(n-p_0)/2}}{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}.$$

The limit is $O(1)$, so a sufficient condition for information consistency is

$$\int_0^\infty (g+d_l)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g) \asymp \int_0^\infty (g+1)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g) = \infty,$$

as required.

Case $\nu_0 < \nu_1$ In this case, we apply the upper bound in Lemma 1:

$$B_{10} \lesssim \frac{[\mathsf{SSE}_0 + \mathsf{SSR}]^{(n-p_0)/2}}{(\widehat{\theta}'\Omega^{-1}\widehat{\theta})^{(n-p_1)/2}} \int_0^\infty \frac{(g+d_u)^{(n-p_1-r_1)/2}}{[(g+d_u)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p_1)/2}}\pi(\mathrm{d}g).$$

The term outside the integral goes to 0, so a necessary condition for information consistency is that the integral be infinite. We can bound the integral:

$$\int_0^\infty \frac{(g+d_u)^{(n-p_1-r_1)/2}}{[(g+d_u)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p_1)/2}}\pi(\mathrm{d}g) \leq \int_0^\infty (g+d_u)^{(n-p_1-r_1)/2}\,\pi(\mathrm{d}g),$$

so a necessary condition for information consistency is

$$\int_0^\infty (g+d_u)^{(n-p_1-r_1)/2}\,\pi(\mathrm{d}g) \asymp \int_0^\infty (g+1)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g) = \infty,$$

as required.

## B.3 Proof of Lemma 4

Throughout, we use the notation in Appendix B.2.

Case 1. Suppose there exists $M < \infty$ such that for all $g \geq M$, $\pi(g) \gtrsim g^{-\alpha}$ for $\alpha > 1$ and $p_0 > p_1$. Then, we apply the lower bound in Lemma 1:

$$B_{10} \gtrsim [\mathsf{SSE}_0 + \mathsf{SSR}]^{(n-p_0)/2} \int_M^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}\mathrm{d}g.$$

Now, note that for any $K, d > 0$ with $1 - d < K$,

$$0 \leq \lim_{\|\widehat{\theta}\|^2\to\infty} \int_{\min(0,1-d)}^K \frac{(g+d)^{(n-p_1-r_1)/2-\alpha}}{[(g+d)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}\mathrm{d}g \lesssim \lim_{\|\widehat{\theta}\|^2\to\infty} [\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p_1)/2} = 0,$$

so

$$\lim_{\|\widehat{\theta}\|^2\to\infty} \int_M^\infty \frac{(g+d_l)^{\frac{n-p_1-r_1}{2}-\alpha}\,\mathrm{d}g}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}} = \lim_{\|\widehat{\theta}\|^2\to\infty} \int_{1-d_l}^\infty \frac{(g+d_l)^{\frac{n-p_1-r_1}{2}-\alpha}\,\mathrm{d}g}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}.$$

Plugging in:

$$\begin{aligned}\lim_{\|\widehat{\theta}\|^2\to\infty} B_{10} &\gtrsim \lim_{\|\widehat{\theta}\|^2\to\infty} [\mathsf{SSE}_0 + \mathsf{SSR}]^{(n-p_0)/2} \int_{1-d_l}^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}\mathrm{d}g \\ &\propto \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{(\mathsf{SSE}_0 + \mathsf{SSR})^{(n-p_0)/2}}{\mathsf{SSE}_1^{(n-p_1)/2}} {}_2F_1\left(\frac{n-p_1}{2}, \frac{r_1}{2}+\alpha-1; \frac{r_1}{2}+\alpha; \frac{-\widehat{\theta}'\Omega^{-1}\widehat{\theta}}{\mathsf{SSE}_1}\right).\end{aligned}$$

Using the identity

$$ {}_2F_1(a,b;c;z) = (1-z)^{-b}\,{}_2F_1\left(b, c-a; c; \tfrac{z}{z-1}\right), $$

we have

$$ \lim_{\|\widehat{\theta}\|^2\to\infty} B_{10} \gtrsim \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{(\mathsf{SSE}_0+\mathsf{SSR})^{(n-p_0)/2}\,{}_2F_1\left(\frac{r_1}{2}+\alpha-1, \frac{r_1-(n-p_1)}{2}+\alpha; \frac{r_1}{2}+\alpha; R^2\right)}{\mathsf{SSE}_1^{(n-p_1)/2}\left[1+\frac{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}{\mathsf{SSE}_1}\right]^{(r_1/2)+\alpha-1}}, $$

where $R^2 = \widehat{\theta}'\Omega^{-1}\widehat{\theta}/(\widehat{\theta}'\Omega^{-1}\widehat{\theta}+\mathsf{SSE}_1) \to 1$ as $\|\widehat{\theta}\|^2\to\infty$. If $\alpha < (n-p_1-r_1)/2+1$ (which is satisfied because $\alpha < (n-p_0-r_1)/2$ and $p_0 > p_1$ by assumption), the limit of the hypergeometric function as $R^2\to 1$ is a constant (by Gauss' theorem). From here, it is immediate to conclude that $B_{10}$ is information consistent whenever the lower bound is infinite, which occurs for $\alpha < (n-p_0-r_1)/2+1$, as required.

Case 2. Suppose there exists $M' < \infty$ such that for all $g \geq M'$, $\pi(g) \lesssim g^{-\alpha}$ for $\alpha > 1$ and $p_0 > p_1$. Then, by Lemma 1:

$$ B_{10} \lesssim [\mathsf{SSE}_0+\mathsf{SSR}]^{(n-p_0)/2} \int_M^\infty \frac{(g+d_u)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_u)\mathsf{SSE}_1+\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}\mathrm{d}g. $$

As argued in Case 1, the limit of the integral is equal to

$$ \lim_{\|\widehat{\theta}\|^2\to\infty} \int_{1-d_u}^\infty \frac{(g+d_u)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_u)\mathsf{SSE}_1+\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}}\mathrm{d}g, $$

and carrying out the same computations as in Case 1:

$$ \lim_{\|\widehat{\theta}\|^2\to\infty} B_{10} \lesssim \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{(\mathsf{SSE}_0+\mathsf{SSR})^{(n-p_0)/2}\,{}_2F_1\left(\frac{r_1}{2}+\alpha-1, \frac{r_1-(n-p_1)}{2}+\alpha; \frac{r_1}{2}+\alpha; R^2\right)}{\mathsf{SSE}_1^{(n-p_1)/2}\left[1+\frac{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}{\mathsf{SSE}_1}\right]^{(r_1/2)+\alpha-1}}. $$

If $(n-p_0-r_1)/2+1 \leq \alpha < (n-p_1-r_1)/2+1$, the limit of the hypergeometric function is $O(1)$ and $B_{10}$ is information inconsistent. If $\alpha \geq (n-p_1-r_1)/2+1$, the necessary condition of Lemma 3 implies that $B_{10}$ is information inconsistent. Therefore, $B_{10}$ is information inconsistent whenever $\alpha \geq (n-p_0-r_1)/2+1$, as required.

## B.4 Proof of Lemma 5

Using the notation in Appendix B.2 and applying Lemma 1:

$$B_{10} \gtrsim \frac{(\mathsf{SSE}_0 + \mathsf{SSR})^{(n-p_0)/2}}{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{n-p_1/2}} \frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\mathsf{SSE}_1/\widehat{\theta}'\Omega^{-1}\widehat{\theta}+1]^{(n-p_1)/2}}$$

For $g > 0$, the right-hand side is maximized at

$$\widehat{g} = \max(0, (n-p_1-r_1)\widehat{\theta}'\Omega^{-1}\widehat{\theta}/(r_1\mathsf{SSE}) - d_l)$$

. Then,

$$\begin{aligned}
\lim_{\|\widehat{\theta}\|^2\to\infty} \max_{g\geq 0} B_{10} &\gtrsim \frac{(\mathsf{SSE}_0 + \mathsf{SSR})^{(n-p_0)/2}}{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p_1)/2}} \frac{(\widehat{g}+d_l)^{(n-p_1-r_1)/2}}{[(\widehat{g}+d_l)\mathsf{SSE}_1/\widehat{\theta}'\Omega^{-1}\widehat{\theta}+1]^{(n-p_1)/2}} \\
&\propto \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{(\mathsf{SSE}_0 + \mathsf{SSR})^{(n-p_0)/2}}{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{r_1/2}\mathsf{SSE}^{(n-p_1-r_1)/2}} \\
&= \infty,
\end{aligned}$$

so the adaptive prior is information consistent.

## B.5 Proof of Lemma 8

Throughout, we use the notation in Appendix B.2.

Sufficient condition:

We start with the case where there exists $\widehat{\theta}_i \to +\infty$; we treat the case where all $\widehat{\theta}_i \to -\infty$ later.

We can write:

$$\begin{aligned}
\lim_{\|\widehat{\theta}\|^2\to\infty} P(\theta \leq 0 \mid y) &= \lim_{\|\widehat{\theta}\|^2\to\infty} \int_0^\infty P(\theta \leq 0 \mid g, y) p(g \mid y) \mathrm{d}g \\
&= \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{1}{p(y)} \int_0^\infty P(\theta \leq 0 \mid g, y) p(y \mid g) \pi(\mathrm{d}g) \\
&\propto \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{1}{p(y)} \int_0^\infty P(\theta \leq 0 \mid g, y) h(g)\, \pi(\mathrm{d}g),
\end{aligned}$$

with $h$ as defined in Lemma 1 (but noting that, in this case, the notation is $\nu_1 = \nu$). Letting $p = \nu - r_2$ and using the upper bound in Lemma 1,

$$\lim_{\|\widehat{\theta}\|^2 \to \infty} P(\theta \leq 0 \mid y) \lesssim \lim_{\|\widehat{\theta}\|^2 \to \infty} \frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}}{p(y)} \int_0^\infty \frac{P(\theta \leq 0 \mid g, y)\,(g + d_u)^{(n-p-r_1)/2}}{[(g + d_u)\mathsf{SSE}_1/\widehat{\theta}'\Omega^{-1}\widehat{\theta} + 1]^{(n-p)/2}}\,\pi(\mathrm{d}g)$$

From Lemma 7, we know that

$$P(\theta \leq 0 \mid g, y) = P(\xi \leq 0 \mid g, y),$$

where $\xi$ has a multivariate Student-$t$ distribution, with location and scale

$$w = (\mathcal{I}_\theta + \Omega^{-1}/g)^{-1}\mathcal{I}_\theta\widehat{\theta}$$

$$m = \frac{(n + \nu - r_2)^{1/2}\, w}{[\mathsf{SSE}_1 + \widehat{\theta}'(\mathcal{I}_\theta^{-1} + g\Omega)^{-1}\widehat{\theta}]^{1/2}}$$

$$S = (\mathcal{I}_\theta + \Omega^{-1}/g)^{-1}.$$

We factor

$$S = \Omega^{1/2}(\Omega^{1/2}\mathcal{I}_\theta\Omega^{1/2} + I_{r_1}/g)^{-1}\Omega^{1/2} = \Omega^{1/2}O'(D^{-1} + I_{r_1}/g)^{-1}O\Omega^{1/2},$$

where $O$ is orthogonal and $D$ is diagonal (with positive entries) as defined in Lemma 1. Therefore, for a fixed coordinate $j$,

$$S_{jj} \in \left[\frac{g}{g/d_l + 1}\Omega_{jj}, \frac{g}{g/d_u + 1}\Omega_{jj}\right],$$

so $0 < S_{jj} < \infty$ for $g > 0$. Using the same factorizations, we obtain $\|w\|^2 \propto \widehat{\theta}'\Omega\Omega\widehat{\theta}$ for $g > 0$. Plugging this in and factorizing the denominator in $m$ in a similar manner, we obtain

$$m = \frac{(n + \nu - r_2)^{1/2}\,\|w\|}{[\mathsf{SSE}_1 + \widehat{\theta}'(\mathcal{I}_\theta^{-1} + g\Omega)^{-1}\widehat{\theta}]^{1/2}} \frac{w}{\|w\|}$$

$$\propto \frac{(\widehat{\theta}'\Omega\Omega\widehat{\theta})^{1/2}}{[\mathsf{SSE}_1 + \widehat{\theta}'(\mathcal{I}_\theta^{-1} + g\Omega)^{-1}\widehat{\theta}]^{1/2}} \frac{w}{\|w\|}.$$

If we choose a coordinate $j$ such that $w_j > 0$ (which exists by assumption), using the lower bound in Lemma 1,

$$m_j \gtrsim \frac{(g+d_l)^{1/2}(\widehat{\theta}'\Omega\Omega\widehat{\theta})^{1/2}}{\left[(g+d_l)\mathsf{SSE}_1 + \widehat{\theta}\Omega^{-1}\widehat{\theta}\right]^{1/2}} \gtrsim \frac{(g+d_l)^{1/2}}{\left[(g+d_l)\mathsf{SSE}_1/\widehat{\theta}'\Omega^{-1}\widehat{\theta}+1\right]^{1/2}}$$

Now,

$$\lim_{\|\widehat{\theta}\|^2\to\infty} P(\theta \le 0 \mid y) \lesssim \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}}{p(y)} \int_0^\infty \frac{P(T_{n-p} \ge \frac{m_j}{\sqrt{S_{jj}}})\,(g+d_u)^{\frac{n-p-r_1}{2}}\,\pi(\mathrm{d}g)}{[(g+d_u)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p)/2}}.$$

where $T_{n-p}$ is a central Student-$t$ with $n-p$ degrees of freedom. Let $\varepsilon > 0$, then

$$\int_0^\varepsilon \frac{P(T_{n-p} \ge m_j/\sqrt{S_{jj}})\,(g+d_u)^{(n-p-r_1)/2}}{[(g+d_u)\mathsf{SSE}_1/\widehat{\theta}'\Omega^{-1}\widehat{\theta}+1]^{(n-p)/2}}\,\pi(\mathrm{d}g) \le (\varepsilon+d_u)^{(n-p-r_1)/2},$$

so

$$\lim_{\|\widehat{\theta}\|^2\to\infty} P(\theta \le 0 \mid y) \lesssim \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}}{p(y)} \int_\varepsilon^\infty \frac{P(T_{n-p} \ge \frac{m_j}{\sqrt{S_{jj}}})\,(g+d_u)^{\frac{n-p-r_1}{2}}\,\pi(\mathrm{d}g)}{[(g+d_u)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p)/2}}.$$

Therefore, we can plug in our bounds for $m_j$ and $S_{jj}$, which are bounded away from 0 whenever $g > 0$. Using the tail bound

$$P(T_{n-p} \ge x) \lesssim \frac{1}{x(1+x^2/\nu)^{(n-p-1)/2}} \lesssim x^{-(n-p)}$$

and our previous work, we obtain

$$\lim_{\|\widehat{\theta}\|^2\to\infty} P(\theta \le 0 \mid y) \lesssim \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}}{p(y)} \int_\varepsilon^\infty (g+d_u)^{-r_1/2}\,\pi(\mathrm{d}g)$$

$$\propto \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}}{p(y)}.$$

Clearly

$$\lim_{\|\widehat{\theta}\|^2\to\infty} \frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}}{p(y)} = 0 \Leftrightarrow \lim_{\|\widehat{\theta}\|^2\to\infty} [\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2}\,p(y) = \infty$$

and

$$\begin{aligned}
\lim_{\|\widehat{\theta}\|^2\to\infty}[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2}\,p(y) &= \lim_{\|\widehat{\theta}\|^2\to\infty}[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2}\int_0^\infty p(y\mid g)\,\pi(\mathrm{d}g)\\
&\propto \lim_{\|\widehat{\theta}\|^2\to\infty}[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2}\int_0^\infty h(g)\,\pi(\mathrm{d}g)\\
&\gtrsim \lim_{\|\widehat{\theta}\|^2\to\infty}\int_0^\infty \frac{(g+d_l)^{(n-p-r_1)/2}}{[(g+d_l)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p)/2}}\pi(\mathrm{d}g)\\
&\gtrsim \int_0^\infty \liminf_{\|\widehat{\theta}\|^2\to\infty}\frac{(g+d_l)^{(n-p-r_1)/2}}{[(g+d_l)\frac{\mathsf{SSE}_1}{\widehat{\theta}'\Omega^{-1}\widehat{\theta}}+1]^{(n-p)/2}}\pi(\mathrm{d}g)\\
&= \int_0^\infty (g+d_l)^{(n-p-r_1)/2}\pi(\mathrm{d}g)\\
&\asymp \int_0^\infty (g+1)^{(n-p-r_1)/2}\pi(\mathrm{d}g).
\end{aligned}$$

Therefore, if the integral above is infinite, $\lim_{\|\widehat{\theta}\|^2\to\infty}P(\theta\le 0\mid y)=0$, as required.

Now we turn to the case where $\widehat{\theta}_i\to-\infty$ for all $i$, in which case we assume that $w_i<0$ for all $i$. Then, a Fréchet bound ensures that

$$P(\theta\le 0\mid y)=P(\theta_1\le 0,\theta_2\le 0,\ldots,\theta_{r_1}\le 0\mid y)\ge\sum_{i=1}^{r_1}P(\theta_i\le 0\mid y)-(r_1-1).$$

Therefore,

$$\lim_{\|\widehat{\theta}\|^2\to\infty}P(\theta_i\ge 0\mid y)=0,\ 1\le i\le r_1\Rightarrow\lim_{\|\widehat{\theta}\|^2\to\infty}P(\theta\le 0\mid y)=1.$$

Then, we can work with the conditional probabilities exactly as we did for the previous case:

$$\begin{aligned}
\lim_{\|\widehat{\theta}\|^2\to\infty}P(\theta_i\ge 0\mid y) &= \lim_{\|\widehat{\theta}\|^2\to\infty}\int_0^\infty P(\theta_i\ge 0\mid g,y)p(g\mid y)\,\mathrm{d}g\\
&\lesssim \lim_{\|\widehat{\theta}\|^2\to\infty}\frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}\int_\varepsilon^\infty \frac{P(T_{n-p}\ge -m_j/\sqrt{S_{jj}})\,(g+d_u)^{(n-p-r_1)/2}}{[(g+d_u)\mathsf{SSE}_1/\widehat{\theta}'\Omega^{-1}\widehat{\theta}+1]^{(n-p)/2}}\,\pi(\mathrm{d}g)}{p(y)}.
\end{aligned}$$

Since $-m_j$ is positive, the subsequent steps in the proof for the previous case allow us to conclude that $\lim_{\|\widehat{\theta}\|^2\to\infty} P(\theta_i \geq 0 \mid y) = 0$, as required.

Necessary condition:

In the sequel we assume that there is at least one $i$ such that $\widehat{\theta}_i \to +\infty$. The case where all coordinates go to $-\infty$ can be dealt with the same way we did for the sufficient condition. We can write:

$$\lim_{\|\widehat{\theta}\|^2\to\infty} P(\theta \leq 0 \mid y) = \lim_{\|\widehat{\theta}\|^2\to\infty} \frac{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2} \int_0^\infty P(\theta \leq 0 \mid g, y) p(y \mid g) \pi(\mathrm{d}g)}{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2} p(y)}.$$

First, we show that the limit of the numerator is bounded away from 0. Applying Fatou's lemma and one of the bounds in Lemma 1,

$$\begin{aligned}\lim_{\|\widehat{\theta}\|^2\to\infty} \frac{\int_0^\infty P(\theta \leq 0 \mid g, y) p(y \mid g) \pi(\mathrm{d}g)}{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{-(n-p)/2}} &\geq \int_0^\infty \liminf_{\|\widehat{\theta}\|^2\to\infty} \frac{P(\theta \leq 0 \mid g, y)\, h(g)}{[\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2}} \pi(\mathrm{d}g) \\ &\gtrsim \int_0^\infty (g + d_l)^{(n-p-r_1)/2} \liminf_{\|\widehat{\theta}\|^2\to\infty} P(\theta \leq 0 \mid g, y) \pi(\mathrm{d}g),\end{aligned}$$

and for any $g$,

$$\liminf_{\|\widehat{\theta}\|^2\to\infty} P(\theta \leq 0 \mid g, y) = \liminf_{\|\widehat{\theta}\|^2\to\infty} P(\xi \leq 0 \mid g, y)$$

where $\xi$ is a multivariate Student-$t$ as in Lemma 7. Lemma 7 shows that $P(\xi \leq 0 \mid g, y)$ is bounded away from 0, which implies that the numerator is bounded away from 0, as claimed. A necessary condition for $\lim_{\|\widehat{\theta}\|^2\to\infty} P(\theta \leq 0 \mid y) = 0$ is that $\lim_{\|\widehat{\theta}\|^2\to\infty} [\widehat{\theta}'\Omega^{-1}\widehat{\theta}]^{(n-p)/2} p(y) = \infty$ which, as we saw in the proof of the sufficient condition, is equivalent to

$$\int_0^\infty (g+1)^{(n-p-r_1)/2} \pi(\mathrm{d}g) = \infty,$$

as required.

# Appendix C

## Proofs for Chapter 4

### C.1 Limit consistency of inverse-moment prior.

The sequence of posteriors $\pi_2(\delta \mid y_A, y_B)$ converges weakly to $\pi_1(\mu_A \mid \mu_B^*, y_A)$ (almost surely), so limit consistency for the inverse-moment prior follows if the posterior expectations converge; that is, if

$$\int \frac{\sqrt{2}\sigma^2 e^{-\frac{\sigma^2}{\omega\delta^2}+\frac{\omega\delta^2}{2\sigma^2}}}{\omega\delta^2}\pi_2(\delta \mid y_A, y_B)\,\mathrm{d}\delta \to \int \frac{\sqrt{2}\sigma^2 e^{-\frac{\sigma^2}{\omega(\mu_B^*-\mu_A)^2}+\frac{\omega(\mu_B^*-\mu_A)^2}{2\sigma^2}}}{\omega(\mu_B^*-\mu_A)^2}\pi_1(\mu_A \mid \mu_B^*, y_A)\,\mathrm{d}\mu_A.$$

A sufficient condition (see e.g. Chapter 3 of Billingsley (2013)) is that there exists $\varepsilon > 0$ such that

$$\int \left| \frac{\sqrt{2}\sigma^2 e^{-\frac{\sigma^2}{\omega\delta^2}+\frac{\omega\delta^2}{2\sigma^2}}}{\omega\delta^2} \right|^{1+\varepsilon} \pi_2(\delta \mid y_A, y_B)\,\mathrm{d}\delta < \infty.$$

with $\mathbb{P}_{y_B}-$probability 1. This is satisfied if $0 < \varepsilon < n_A/((n_A+1)\omega)$. Indeed,

$$\int \left| \frac{\sqrt{2}\sigma^2 e^{-\frac{\sigma^2}{\omega\delta^2}+\frac{\omega\delta^2}{2\sigma^2}}}{\omega\delta^2} \right|^{1+\varepsilon} \pi_2(\delta \mid y_A, y_B)\,\mathrm{d}\delta \leq \left[\frac{\sqrt{2}}{e}\right]^{1+\varepsilon} \int e^{\frac{(1+\varepsilon)\omega\delta^2}{2\sigma^2}} \pi_2(\delta \mid y_A, y_B)\,\mathrm{d}\delta.$$

The integral can be written as the moment generating function of a non-central $\chi^2$ distribution evaluated at $\omega(1+\varepsilon)/[2(\kappa+\omega)]$, which is finite if $\omega(1+\varepsilon)/[2(\kappa+\omega)] < 1/2$. This condition is equivalent to picking $\varepsilon < \kappa/\omega$, and since $\kappa > n_A/(n_A+1)$, our choice of $\varepsilon$ ensures that the integral is finite, as required.

## C.2 Limit consistency for mixtures of $g-$priors

We want to show that the function

$$\int (g+1)^{-1/2} \exp\left\{\frac{gZ_n^2}{2(g+1)}\right\} \pi(g)\,\mathrm{d}g.$$

is continuous in $Z_n$, so that we can apply the continuous mapping theorem and conclude that the results we have for $g-$priors with fixed $g$ apply to mixtures. To that end, it suffices to show that (1) $f(g, Z_n) = (g+1)^{-1/2} \exp\{gZ_n^2/[2(g+1)]\}$ is continuous for $g \in \mathbb{R}_+$ for all $Z_n^2 \in \mathbb{R}_+$ and (2) there exists $h(g) > f(g, Z_n)$ such that $\int h(g)\pi(g)\mathrm{d}g < \infty$ (see e.g. Theorem 2.27 of Folland (2013)). The first condition is clearly satisfied. The second condition is satisfied for fixed $y_A$ (so that the randomness in $Z_n$ only comes from $y_B$) since $f(g, Z_n) > \exp\{n_A\overline{y}_A^2/[2\sigma^2]\}$, which is constant because $n_A$ and $\overline{y}_A^2$ are treated as fixed.

# Bibliography


Barbieri, M. M. and Berger, J. O. (2004), "Optimal predictive model selection," *Annals of Statistics*, pp. 870–897.

Basu, D. (1964), "Recovery of Ancillary Information," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 3–16.

Bayarri, M., Berger, J., Forte, A., García-Donato, G., et al. (2012), "Criteria for Bayesian model choice with application to variable selection," *The Annals of Statistics*, 40, 1550–1577.

Berger, J., Bayarri, M., and Pericchi, L. (2014), "The effective sample size," *Econometric Reviews*, 33, 197–217.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer Science & Business Media.

Berger, J. O. and Mortera, J. (1999), "Default Bayes factors for nonnested hypothesis testing," *Journal of American Statistical Association*, 94, 542–554.

Berger, J. O. and Pericchi, L. (2001), "Objective Bayesian methods for model selection: Introduction and comparison (with discussion)," in *Model Selection*, ed. P. Lahiri, vol. 38 of *Monograph Series*, pp. 135–207, Beachwood Ohio, institute of mathematical statistics lecture notes edn.

Berger, J. O. and Wolpert, R. L. (1988), "The likelihood principle," *Lecture notes-Monograph series*.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998), "Bayes factors and marginal distributions in invariant situations," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 307–321.

Berger, J. O., Ghosh, J. K., and Mukhopadhyay, N. (2003), "Approximations and consistency of Bayes factors as model dimension grows," *Journal of Statistical Planning and Inference*, 112, 241–258.

Billingsley, P. (2013), *Convergence of probability measures*, John Wiley & Sons.

Birnbaum, A. (1962), "On the foundations of statistical inference," *Journal of the American Statistical Association*, 57, 269–306.

Bjørnstad, J. F. (1996), "On the generalization of the Likelihood function and the likelihood principle," *Journal of the American Statistical Association*, 91, 791–806.

Claeskens, G., Hjort, N. L., et al. (2008), "Model selection and model averaging," *Cambridge Books*.

Consonni, G., Veronese, P., et al. (2008), "Compatibility of prior specifications across linear models," *Statistical Science*, 23, 332–353.

Consonni, G., Forster, J. J., and La Rocca, L. (2013), "The whetstone and the alum block: Balanced objective Bayesian comparison of nested models for discrete data," *Statistical Science*, pp. 398–423.

Copas, J. B. (1983), "Regression, prediction and shrinkage," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 311–354.

Cox, D. (1971), "The choice between alternative ancillary statistics," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 251–255.

Cox, D. and Mayo, D. G. (2010), "Objectivity and Conditionality in Frequentist Inference," *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*, 276.

Cox, D. R. (1958), "Some problems connected with statistical inference," *The Annals of Mathematical Statistics*, pp. 357–372.

Cui, W. and George, E. I. (2008), "Empirical Bayes vs. fully Bayes variable selection," *Journal of Statistical Planning and Inference*, 138, 888–900.

DasGupta, A. and Studden, W. J. (1989), "Frequentist behavior of robust Bayes estimates of normal means," *Statistics and Decisions*, 7, 333–361.

Dawid, P. (2011), "Basu on ancillarity," in *Selected Works of Debabrata Basu*, pp. 5–8, Springer.

Durbin, J. (1970), "On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood," *Journal of the American Statistical Association*, 65, 395–398.

Evans, M. (2013), "What does the proof of Birnbaum's theorem prove?" *Electronic Journal of Statistics*, 7, 2645–2655.

Fernandez, C., Ley, E., and Steel, M. F. (2001), "Benchmark priors for Bayesian model averaging," *Journal of Econometrics*, 100, 381–427.

Folland, G. B. (2013), *Real analysis: modern techniques and their applications*, John Wiley & Sons.

Gandenberger, G. (2014), "A new proof of the likelihood principle," *The British Journal for the Philosophy of Science*, p. axt039.

George, E. and Foster, D. P. (2000), "Calibration and empirical Bayes variable selection," *Biometrika*, 87, 731–747.

Ghosh, J. and Ghattas, A. E. (2015), "Bayesian variable selection under collinearity," *The American Statistician*, 69, 165–173.

Good, I. J. (1965), *The estimation of probabilities: An essay on modern Bayesian methods*, vol. 30, MIT press.

Grossman, J. (2011), "The Likelihood Principle," *Philosophy of Statistics*, 1, 553.

Guo, R. and Speckman, P. L. (2009), "Bayes factor consistency in linear models," in *The 2009 International Workshop on Objective Bayes Methodology*.

Hansen, M. H. and Yu, B. (2001), "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, 96, 746–774.

Hansen, M. H. and Yu, B. (2003), "Minimum description length model selection criteria for generalized linear models," *Lecture Notes-Monograph Series*, pp. 145–163.

Held, L., Sabanés-Bové, D., Gravestock, I., et al. (2015), "Approximate Bayesian model selection with the deviance statistic," *Statistical Science*, 30, 242–257.

Helland, I. S. (1995), "Simple counterexamples against the conditionality principle," *The American Statistician*, 49, 351–356.

Hoff, P. D. (2009), *A first course in Bayesian statistical methods*, Springer Science & Business Media.

Holmes, C. C., Caron, F., Griffin, J. E., Stephens, D. A., et al. (2015), "Two-sample Bayesian nonparametric hypothesis testing," *Bayesian Analysis*, 10, 297–320.

Jeffreys, H. (1939), *Theory of Probability-1st ed*, New York: Oxford University Press.

Johnson, V. E. and Rossell, D. (2010), "On the use of non-local prior densities in Bayesian hypothesis tests," *Journal of the Royal Statistical Society Series B*, 72, 143–170.

Johnson, V. E. and Rossell, D. (2012), "Bayesian model selection in high-dimensional settings," *Journal of the American Statistical Association*, 107, 649–660.

Kalbfleisch, J. D. (1975), "Sufficiency and Conditionality," *Biometrika*, 62, 251–259.

Kass, R. E. and Wasserman, L. (1995), "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," *Journal of the American Statistical Association*, 90, 928–934.

Klugkist, I. and Hoijtink, H. (2007), "The Bayes Factor for Inequality and About Equality Constrained Models," *Computational Statistics and Data Analysis*, 51, 6367–6379.

Leamer, E. E. (1978), *Specification searches: Ad hoc inference with nonexperimental data*, John Wiley & Sons Inc.

Li, Y. and Clyde, M. A. (2015), "Mixtures of g-priors in Generalized Linear Models," *arXiv preprint arXiv:1503.06913*.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

Ly, A. (2017), "Bayes Factors for Research Workers," Ph.D. thesis, Universiteit van Amsterdam.

Marin, J.-M. and Robert, C. (2007), *Bayesian core: a practical approach to computational Bayesian statistics*, Springer Science & Business Media.

Marsaglia, G. (1964), "Conditional means and covariances of normal variables with singular covariance matrix," *Journal of the American Statistical Association*, 59, 1203–1204.

Maruyama, Y. and George, E. I. (2011), "Fully Bayes factors with a generalized g-prior," *The Annals of Statistics*, 39, 2740–2765.

Mayo, D. G. (2010), "An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle," *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, p. 305.

Mayo, D. G. (2014), "On the Birnbaum Argument for the Strong Likelihood Principle," *Statistical Science*, 29, 227–239.

McCullagh, P. and Nelder, J. A. (1989), "Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability," .

Mulder, J. (2014), "Bayes factors for testing inequality constrained hypotheses: Issues with prior specification," *British Journal of Mathematical and Statistical Psychology*, 67, 153–171.

Mulder, J., Berger, J. O., Peña, V., and Bayarri, M. (2017), "On the Ubiquity of Information Inconsistency for Conjugate Priors," *arXiv preprint arXiv:1710.09700*.

Müller, A. (2001), "Stochastic ordering of multivariate normal distributions," *Annals of the Institute of Statistical Mathematics*, 53, 567–575.

Polasek, W. (1985), "Sensitivity analysis for general and hierarchical linear regression models," *Bayesian Inference and Decision Techniques with Applications*.

Raftery, A. E. (1995), "Bayesian model selection in social research," *Sociological methodology*, 25, 111–164.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92, 179–191.

Robert, C. P. (1996), "Intrinsic losses," *Theory and decision*, 40, 191–214.

Robert, C. P. and Casella, G. (1994), "Distance weighted losses for testing and confidence set evaluation," *Test*, 3, 163–182.

Robins, J. and Wasserman, L. (2000), "Conditioning, likelihood, and coherence: a review of some foundational concepts," *Journal of the American Statistical Association*, 95, 1340–1346.

Rossell, D. and Telesca, D. (2017), "Nonlocal priors for high-dimensional estimation," *Journal of the American Statistical Association*, 112, 254–265.

Sabanés-Bové, D., Held, L., et al. (2011), "Hyper-$g$ priors for generalized linear models," *Bayesian Analysis*, 6, 387–410.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Scott, J. G. and Berger, J. O. (2010), "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem," *The Annals of Statistics*, 38, 2587–2619.

Seber, G. A. (2008), *A matrix handbook for statisticians*, vol. 15, John Wiley & Sons.

Shibata, R. (1983), "Asymptotic mean efficiency of a selection of regression variables," *Annals of the Institute of Statistical Mathematics*, 35, 415–423.

Som, A., Hans, C. M., and MacEachern, S. N. (2016), "A conditional Lindley paradox in Bayesian linear models," *Biometrika*, 103, 993–999.

Soriano, J. and Ma, L. (2017), "Probabilistic multi-resolution scanning for two-sample differences," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 547–572.

Stone, M. (1979), "Comments on model selection criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 276–278.

Strawderman, W. E. (1973), "Proper Bayes minimax estimators of the multivariate normal mean vector for the case of common unknown variances," *The Annals of Statistics*, pp. 1189–1194.

Sweeting, T. J. (2001), "Coverage probability bias, objective Bayes and the likelihood principle," *Biometrika*, 88, 657–675.

Wang, X. and George, E. I. (2007), "Adaptive Bayesian criteria in variable selection for generalized linear models," *Statistica Sinica*, pp. 667–690.

Wechsler, S., Pereira, C. A. d. B., et al. (2008), "Birnbaum's theorem redux," in *Bayesian Inference and Maximum Entropy methods in Science and Engineering: Proceedings of the 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, vol. 1073, pp. 96–100, AIP Publishing.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6, 233–243.

Zellner, A. and Siow, A. (1980), "Posterior odds ratios for selected regression hypotheses," *Trabajos de estadística y de investigación operativa*, 31, 585–603.

# Biography

Víctor Peña was born on November 2, 1989 in Barcelona, Spain. He received a bachelor's degree (2011) and a master's degree (2013) from Universitat Politècnica de Catalunya. In August of 2013, he started his doctorate in Statistical Science at Duke University under the supervision of James O. Berger. He defended his dissertation on July 16th of 2018, and graduated in the summer of 2018.