Invariant Procedures for Model Checking, Checking for Prior-Data Conflict and Bayesian Inference

by

Gun Ho Jang

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Statistics
University of Toronto

Copyright © 2010 by Gun Ho Jang
Abstract

Invariant Procedures for Model Checking, Checking for Prior-Data Conflict and Bayesian Inference

Gun Ho Jang
Doctor of Philosophy
Graduate Department of Statistics
University of Toronto
2010

We consider a statistical theory as being \textit{invariant} when the results of two statisticians’ independent data analyses, based upon the same statistical theory and using effectively the same statistical ingredients, are the same. We discuss three aspects of invariant statistical theories.

Both model checking and checking for prior-data conflict are assessments of single null hypothesis without any specific alternative hypothesis. Hence, we conduct these assessments using a measure of surprise based on a discrepancy statistic. For the discrete case, it is natural to use the probability of obtaining a data point that is less probable than the observed data. For the continuous case, the natural analog of this is not invariant under equivalent choices of discrepancies. A new method is developed to obtain an invariant assessment. This approach also allows several discrepancies to be combined into one discrepancy via a single $P$-value.

Second, Bayesians developed many noninformative priors that are supposed to contain no information concerning the true parameter value. Any of these are data dependent or improper which can lead to a variety of difficulties. Gelman (2006) introduced the notion of the weak informativity as a compromise between informative and noninformative priors without a precise definition. We give a precise definition of weak informativity using a measure of prior-data conflict that assesses whether or not a prior places its mass around
the parameter values having relatively high likelihood. In particular, we say a prior $\Pi_2$ is weakly informative relative to another prior $\Pi_1$ whenever $\Pi_2$ leads to fewer prior-data conflicts a priori than $\Pi_1$. This leads to a precise quantitative measure of how much less informative a weakly informative prior is.

In Bayesian data analysis, highest posterior density inference is a commonly used method. This approach is not invariant to the choice of dominating measure or reparametrizations. We explore properties of relative surprise inferences suggested by Evans (1997). Relative surprise inferences which compare the belief changes from \textit{a priori} to \textit{a posteriori} are invariant under reparametrizations. We mainly focus on the connection of relative surprise inferences to classical Bayesian decision theory as well as important optimalities.
Dedication

To My wife Jae Young
Acknowledgements

When I came to Canada to study statistics, I did not have clear understanding of statistics. My supervisor, Professor Michael Evans, has provided me chances to make sound understandings of statistics by arguing why statistical theories or methodologies arose and what are the virtues and defects of them. Without his knowledge, inspiration, passion and patience, I could not achieve my accomplishment. I am very happy to have this space to express my sincerest heartfelt gratitude to my supervisor, Michael Evans.

Many thanks are due to Professor Jeffrey Rosenthal for encouragement for my researches. I wish to thank many other professors. Particularly, I want to thank Professor Andrey Feuerverger for discussing with me lots of academic and nonacademic issues, Professor Nancy Reid, Professor David Brenner and Professor Bálint Virág for discussing many probability problems and Professor Jaeyong Lee of the Seoul National University for leading and guiding me to the world of Bayesian nonparametrics. Also I express my thank to Professor R.V. Ramamoorthi who gave me advise concerning study abroad.

Also thank to members of the administrative staff, Andrea Carter, Ram Mohabir and Dermot Wheland who helped me sort out many administrative matters.

I also would like to thank my dear colleages in the Department of Statistics for sharing their knowledge. Among them, I thank to research room members, Yan Bai, Shelley Yun Cao, Lan Gong, Panpan Wu, and Lizhen Xu, for allowing me to use computing resource in the room.

Finally, I am deeply grateful to all my family members, parents, and parents-in-law for their endless love, concern and support. Especially to my wife, Jae Young Yu, I am thankful for her absolute patience, support and understanding and her love that cannot be expressed within these short sentences. Without you, I could never make this accomplishment.

Toronto, Ontario

Gun Ho Jang

April 13, 2010
2.2 Model-based $P$-values with Discrete $P$ ........................................ 23
2.3 Model-based $P$-values with General $P$ ........................................ 27
2.4 Invariant $P$-values for General Statistics ................................. 33
2.5 Applications ................................................................. 38
  2.5.1 Model Checking ......................................................... 39
  2.5.2 Checking for Prior-Data Conflict .................................... 46
2.6 Volume Distortions and Computations .................................... 49
  2.6.1 Derivations of $J_T$ ....................................................... 49
  2.6.2 Disappearance of Volume Distortion .................................. 50
  2.6.3 Computations .............................................................. 52
2.7 Conclusions ................................................................. 53
2.8 Proofs ....................................................................... 56

3 Weakly Informative Priors ................................................. 67
  3.1 Introduction ................................................................. 67
  3.2 Comparing Priors .......................................................... 71
  3.3 Deriving Weakly Informative Priors ................................. 74
    3.3.1 Comparing Normal Priors ........................................... 74
    3.3.2 Comparing a $t$ Prior with a Normal Prior ...................... 77
    3.3.3 Comparing Inverse Gamma Priors ............................... 81
  3.4 Applications .............................................................. 83
    3.4.1 Weakly Informative Beta Priors for the Binomial .......... 83
    3.4.2 Weakly Informative Priors for a Location-Scale Model .... 84
    3.4.3 Weakly Informative Priors for Logistic Regression ........ 85
  3.5 Refinements Based Upon Ancillarity .................................. 89
  3.6 Conclusions ............................................................... 91
  3.7 Proofs ................................................................. 92
# 4 Invariant Bayesian Inference

4.1 Relative Surprise Inference .............................................. 110
4.2 Properties of Relative Surprise ....................................... 113
4.3 Decision Theoretic Interpretation ...................................... 124
4.4 Asymptotic Properties of the LRS Estimator ....................... 133
   4.4.1 Consistency of the LRS Estimator ................................. 133
   4.4.2 Bernstein-von Mises Theorem ..................................... 137
4.5 Discussion ........................................................................ 139
4.6 Proofs ............................................................................ 140

# 5 Inferences for Infinite Dimensional Models

5.1 Consistency of the Sieve LRS Estimator ............................... 145
5.2 Models with Increasing Number of Incidental Parameters ........ 151
5.3 Future Work ..................................................................... 158

# A Appendix

# Bibliography
List of Tables

2.1 Conditional probability function for Fisher’s exact test in Example 2.2 . 27

3.1 Evaluations of (3.9) for some λ values. . . . . . . . . . . . . . . . . . . . 78

3.2 Bioassay data from Racine et al. (1986). . . . . . . . . . . . . . . . . . . 86
List of Figures

2.1 Densities and invariant \( P \)-values for test of skewness for various sample sizes \( n \) when sampling from normal. ........................................... 42
2.2 Densities and invariant \( P \)-values for test of kurtosis for various sample sizes \( n \) when sampling from normal. ........................................... 42
2.3 Densities and invariant \( P \)-values for Jarque-Bera test for various sample sizes \( n \) when sampling from normal. ........................................... 43

3.1 Plot of (3.4) versus \( \gamma \) for \( t_{1}(0, \sigma^{2}_{2}, 3) \) priors relative to a \( N(0, 1) \) prior ... 80
3.2 Weakly informative \( \text{Gamma}_{\text{rate}}(\alpha, \beta) \) priors relative to \( \text{Gamma}_{\text{rate}}(2, 2) \) . 82
3.3 Weakly informative \( \text{Beta}(\alpha, \beta) \) priors relative to \( \text{Beta}(6, 6) \) in Section 3.4.1 83
3.4 Weakly informative \( \Pi_2 \) priors relative to \( \Pi_1 \) for the bioessay data .... 87
3.5 Reduction levels of \( N(0, \sigma^{2}_{0}) \times N(0, \sigma^{2}_{1}) \) priors relative to \( N(0, 10^{2}) \times N(0, 2.5^{2}) \) prior ................................................................. 88
3.6 Weakly informative \( \text{Beta}(\alpha, \beta) \) priors relative to \( \text{Beta}(20, 20) \) prior in Example 3.1 ................................................................. 90

4.1 Densities, relative surprises, and Bayes factors of Example 4.4. ........... 124
Chapter 1

Introduction

Reproducibility is one of the most important principles of scientific research. A reproducible result can be used for other research and can be a foundation for a universal theorem. If a research is not reproducible, it may be considered incidental and could be discarded. Needless to say, Statistics should possess reproducibility.

We could consider a statistical theory as being reproducible when the results of two statisticians’ independent data analyses, based upon the same statistical theory and using effectively the same statistical ingredients, are the same. Then, the statistical theory is called invariant. However, as we will see, many statistical theories are not invariant. In this chapter, we discuss why non-invariance emerges and how invariant statistical theories can be achieved.

1.1 Reproducibility and Invariance

Reproducibility is one of the essential principles of scientific research. Findings without reproducibility can be temporary and cannot be universal. Namely, such results cannot be applicable for a general situation.

A similar principle is required in Statistics. Roughly speaking, statistical data analyses have three components—data, models and inference methods. Once data are given to
the statistician, they are non-conditional ingredients, while models and inference methods can be chosen by statisticians. If different models or inference methods are chosen, then the corresponding results can be different. However, two data analyses of the same data using the same inference method, under the same model, should be the same. We call this the *principle of invariance*.

Two data analyses processed by two statisticians under the same statistical ingredients (data, models, inference methods) must be the same as the reproducibility of scientific research dictates. Suppose two data analyses under the same statistical ingredients show different results, that is, data analyses are different by statisticians. Then the inference method used in this data analyses cannot be reliable because any reliable inference methods should bring out the same results under the same statistical ingredients. In a worst case, one analysis shows evidence against a null hypothesis and the other shows evidence for the null hypothesis. Then, at least one of those analyses is wrong. These unclear results would not happen if the inference method is invariant. This is a reason why invariant inferences are required in statistics.

Many statistical theories satisfy the principle of invariance. However, it is not difficult to find non-invariant statistical procedures. Consider the following examples.

**Example 1.1** (Maximum Likelihood Estimation). The maximum likelihood estimator is one of the most used inference methods. Since *the method of maximum likelihood* was introduced by Fisher (1922), many properties of the maximum likelihood estimator including the existence, consistency, and efficiency have been studied. One of the properties of the maximum likelihood estimator is that it is invariant under any one-to-one smooth data transformation as well as any one-to-one smooth reparametrization.

Consider a model $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, namely, a collection of densities. The maximum likelihood estimator is the parameter value $\widehat{\theta}$ maximizing the likelihood function $L(\theta | x) = f_\theta(x)$. If a one-to-one smooth data transformation $T : \mathcal{X} \rightarrow \mathcal{T}$ is made on the model, then the transformed model is $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$ where $g_\theta(t) =$
\( f_\theta(T^{-1}(t)) J_T(T^{-1}(t)) \) and \( J_T \) is the reciprocal of the Jacobian determinant of \( T \). The maximum likelihood estimator of the transformed model is the maximizer of \( L(\theta | t) = g_\theta(t) = f_\theta(T^{-1}(t)) J_T(T^{-1}(t)) = f_\theta(x) J_T(x) \) when \( t = T(x) \) is observed. Hence, the maximum likelihood estimator is invariant under any one-to-one smooth data transformation.

In a similar vein, the maximum likelihood estimator is invariant under any one-to-one reparametrization, \( \Psi : \Theta \rightarrow \Psi \). Let \( \mathcal{H} = \{ h_\psi : \psi \in \Psi \} \) be the reparametrized model where \( h_\psi(x) = f_{\Psi^{-1}(\psi)}(x) \). It is easy to see that \( L(\psi | x) = h_\psi(x) = f_\theta(x) = L(\theta | x) \) for \( \psi = \Psi(\theta) \). Hence we have \( \hat{\psi}_{\text{MLE}} = \Psi(\hat{\theta}_{\text{MLE}}) \). The likelihood order is also preserved under reparametrization, that is, \( L(\theta_1 | x) < L(\theta_2 | x) \) implies \( L(\psi_1 | x) = L(\theta_1 | x) < L(\psi_2 | x) = L(\theta_2 | x) \) where \( \psi_i = \Psi(\theta_i) \) for \( i = 1, 2 \). Hence, likelihood regions, defined by \( \{ \theta : L(\theta | x) > c \} \) for \( c > 0 \), are also invariant, i.e., \( \{ \psi : L(\psi | x) > c \} = \Psi(\{ \theta : L(\theta | x) > c \}) \) for all \( c > 0 \).

Likelihood functions can be extended by multiplying by a nonnegative function of the data, that is, \( \tilde{L}(\theta | x) = w(x) f_\theta(x) \) is also a likelihood function for the model \( \mathcal{F} \). Then the maximum likelihood estimator based on \( \tilde{L} \) is \( \hat{\theta}_{\text{MLE}} = \text{arg max}_\theta \tilde{L}(\theta | x) = \text{arg max}_\theta w(x) f_\theta(x) = \text{arg max}_\theta f_\theta(x) \) because \( w(x) \) does not depend on \( \theta \). Hence, the maximum likelihood estimator is independent of the choice of a likelihood function.

Therefore, the maximum likelihood estimator is invariant under one-to-one smooth data transformation, one-to-one smooth reparametrization, and the choice of a likelihood function. Similar invariance properties of likelihood functions and maximum likelihood estimators are discussed in Sections 2.8 and 2.9 of Pawitan (2001).

**Example 1.2 (Shortest Confidence Interval).** The term ‘confidence interval’ was coined by Neyman (1934). The asymptotically most efficient confidence interval is studied by Wilks (1938). Wilks found the shortest average confidence interval under asymptotic normality. However, we consider the shortest confidence interval having exact coverage probability rather than having asymptotic coverage probability.

Consider a model \( \mathcal{F} = \{ N(0, \sigma^2) : \sigma^2 \in (0, \infty) \} \). The question of interest is to
obtain a confidence interval for a measure of the variability. In this statistical context, two statisticians, statistician A and B, may approach the problem in different ways. Statistician A measures the variability via the variance. So statistician A finds the shortest confidence interval for $\sigma^2$ based on a consistent estimator $T = n^{-1}(X_1^2 + \cdots + X_n^2)$. Statistician B uses the scale of the distribution as the measure of the variability. Then statistician B finds the shortest confidence interval for $\sigma$ based on a consistent estimator $U = T^{1/2}$. It is easy to see that $\sigma^2$ and $\sigma$ are equivalent parametrizations. Also $T$ and $U$ are equivalent data transformations. Both $T$ and $U$ are minimal sufficient for the model $\mathcal{F}$. From the fact $nT/\sigma^2 \sim \chi^2(n)$, densities of $T$ and $U$ are

$$f_{\theta,T}(t) = \left( \frac{n}{2\sigma^2} \right)^{n/2} \frac{1}{\Gamma(n/2)} t^{n/2-1} \exp(-nt/2\sigma^2) \quad \text{and} \quad f_{\theta,U}(u) = 2uf_{\theta,T}(u^2).$$

Since $T/\sigma^2 \sim \text{Gamma}_{\text{scale}}(n/2, 2/n)$, where Gamma$\_\text{scale}(\alpha, \beta)$ is a scale gamma distribution having density $(\Gamma(\alpha)\beta^\alpha)^{-1}x^{\alpha-1}\exp(-x/\beta)$ on $(0, \infty)$, the shortest confidence intervals are $I_A = [t/c_{2,A}, t/c_{1,A}]$ for $\sigma^2$ and $I_B = [u/c_{2,B}, u/c_{1,B}]$ for $\sigma$ for some $c_{i,j}$ for $i = 1, 2, j = A, B$. In order to compare two confidence intervals, $I_B$ is squared or $I_A$ is square-rooted. If the confidence interval of statistician B is squared, that is, $I_B^2 = [u^2/c_{2,B}^2, u^2/c_{1,B}^2]$, then surely the length of $I_B^2$ is bigger than that of $I_A$. Similarly, if the confidence interval of statistician A is square-rooted, that is, $I_A^{1/2} = [t^{1/2}/c_{2,A}^{1/2}, t^{1/2}/c_{1,A}^{1/2}]$, then surely the length of $I_A^{1/2}$ is bigger than that of $I_B$. For example, for $n = 3$, the shortest 90%-confidence intervals are $I_A = [0.1701, 5.1539]t$, $I_B^2 = [0.2280, 5.3062]t$, $I_A^{1/2} = [0.4124, 2.22702]u$, and $I_B = [0.4775, 2.3035]u$. Hence, $|I_A| = 4.9838t < 5.0781t = |I_B^2|$ and $|I_A^{1/2}| = 1.8578u > 1.8260u = |I_B|$

So we are left with the question: what is the right shortest confidence interval for assessing variability?

Example 1.2 shows that the shortest confidence interval is not invariant under one-to-one smooth reparametrizations. In other words, the shortest confidence interval depends on the parametrization.
Even though two confidence intervals are different for finite \( n \), the difference of two confidence intervals converges to 0 as the sample size \( n \) increases. Asymptotic distributions of \( T \) and \( U \) are \( \sqrt{n}(T - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4) \) and \( \sqrt{n}(U - \sigma) \xrightarrow{d} N(0, \sigma^2/2) \). Hence, we have \( I_A \approx [(1 + z(1+\gamma)/2/\sqrt{n/2})^{-1}, (1 - z(1+\gamma)/2/\sqrt{n/2})^{-1}]t \) and \( I_B^{1/2} \approx [(1 + z(1+\gamma)/2/\sqrt{2n})^{-1}, (1 + z(1+\gamma)/2/\sqrt{2n})^{-1}]u \) where \( z_\gamma \) is the \( \gamma \)-quantile of a standard normal distribution. Since \((1 \pm z(1+\gamma)/2/\sqrt{2n})^2 = 1 \pm z(1+\gamma)/2/\sqrt{n/2} + z^2(1+\gamma)/2/(2n)\), we have that \( I_B = [(1 + z(1+\gamma)/2/\sqrt{n/2})^{-1}, (1 - z(1+\gamma)/2/\sqrt{n/2})^{-1}]t + O_p(n^{-1}) \approx I_A \). Thus, the two confidence intervals are asymptotically equivalent.

There are many examples of invariant procedures used in Statistics: the conditional distribution of whole data given a minimal sufficient statistic is invariant, likelihood inferences are invariant under the choice of the likelihood function, conditional inferences are independent to the choice of a maximal ancillary among all equivalent ancillaries, the Neyman-Pearson fundamental lemma is invariant under the choice of the dominating measure.

1.2 Causes of Non-Invariance

Statistical inference requires three components, namely, data, models, and inference methods, as mentioned in Section 1.1. Also, Example 1.2 shows that inferences may not be invariant even though those three components are essentially the same. This means, that the theory used is not invariant. Essentially this is due to the statistician’s subjective choice of the parametrization used. So we investigate what can be chosen by a statistician in her/his data analysis and how an inference can be invariant under such statistician’s subjective choices. For doing this, we need to understand how these three components are related. Then, we investigate how the noninvariance of statistical data analyses arise.

A model \( \mathcal{F} \) is a collection of probability measures that are supposed to explain the
data $X$. That is, the data $X$ supposedly arise from a probability measure $P$ that is contained in $\mathcal{F}$. Typically, the model $\mathcal{F}$ is indexed by parameters as a convenience, that is, $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$ where $\theta$ is called the model parameter and $\Theta$ is called the parameter space. Since arbitrary probability measures are difficult to deal with, we generally assume that the model is dominated, so a density function $f_\theta(x) = \frac{dP_\theta}{d\mu}(x)$ with respect to a $\sigma$-finite measure $\mu$ is generally considered. A likelihood function is then defined as $L(\theta|x) = f_\theta(x)$, where the data $x$ is fixed and $\theta$ is variable. So we note that statisticians choose a parametrization of the model in their analysis. If there are things that statisticians can choose, then these things might cause noninvariance of their inferences. So we focus on the statistical ingredients that can be chosen. Of course, the model $\mathcal{F}$ is itself chosen by the statistician. But a key difference is that we can check whether or not this choice makes sense by seeing whether or not the model makes sense in light of the data observed. In fact, we do not think of the data as affirming the model only not providing evidence against the model. Such a check would seem to be necessary for any subjective choice made by a statistician.

**Parametrizations.** A key advantage of parametrizations is that an inference question can be written as a statement about parameters once the model is parametrized. For example, an inference about the question of whether or not the central location of a distribution is zero, can be stated as $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$ where $\mu$ is a location parameter. This convenience leads statisticians to parametrize models. And there is almost no way to resist this convenience. However, this subjective choices of a parametrization may result in noninvariance of data analysis, as in Example 1.2. Further these does appear to be a way to check whether a particular parametrization makes sense in light of the data.

**Data transformation.** Parametric inferences usually depend on the relationship between sampling models and parameters. One of the best descriptions of this relationship
is the principle of sufficiency. We follow the definition in Evans, Fraser and Monette (1986) given as follows:

**The Principle of Sufficiency.** The evidential contents of inferential bases \( I = (\mathcal{F}_x, x) \) and \( I_T = (\mathcal{F}_T, t) \) are the same when \( t = T(x) \) is a sufficient statistic for the model \( \mathcal{F}_x = \{ P_{\theta, x} : \theta \in \Theta \} \) for \( x \) and \( \mathcal{F}_T = \{ P_{\theta, T} : \theta \in \Theta \} \) is the model for \( t \) derived from \( \mathcal{F}_x \).

As this principle states, every sufficient statistic contains all the information in the data for inferences. Since a minimal sufficient statistic has the maximal reduction of data among all sufficient statistics, the usage of minimal sufficient statistics in inferences is encouraged. Then, statisticians can choose a minimal sufficient statistic. This might cause a non-invariant inference. In Example 1.2, both \( T = n^{-1}(X_1^2 + \cdots + X_n^2) \) and \( U = T^{1/2} \) are minimal sufficient but the shortest confidence intervals based on them are different. This noninvariance turns out to be important when we consider methods for checking for prior-data conflict, see Chapter 2, i.e., we check to see whether or not the prior makes sense in light of the data observed.

**Dominating Measures.** A dominating measure for the model is necessary in order to get a likelihood function. Since a dominating measure can be chosen by statisticians in their problems, likelihood inferences might depend on the choice of a dominating measure. In fact, likelihood based inferences are invariant under the choice of a dominating measure. For example, maximum likelihood inferences are invariant under the choice of the dominating measure, see Section 2.8 of Pawitan (2001) or Example 1.1 discussed in Section 1.1.

Another interesting example is penalized likelihood inferences. Akaike (1974) introduced the Akaike information criteria (AIC) for model selection problems. Akaike (1974) suggested to choose a model having the smallest AIC given by

\[
-2 \log(\text{maximum likelihood}) + 2 (\# \text{ of parameters}).
\]
The fact that minimizing the AIC is equivalent to maximizing the likelihood penalized by the number of parameters encouraged the development of general penalized likelihoods, namely, using the logarithm of the likelihood minus nonnegative functions of the parameters. So,

\[
\text{penalized likelihood} = \log(\text{likelihood}) - \text{penalty}
\]

where the penalty is a nonnegative function of parameters. Hence, the choice of a dominating measure results in a constant function difference between penalized likelihood functions. Suppose \( \nu \) and \( \mu \) are absolutely continuous with respect to each other and are dominating measures of the considered model. Two penalized likelihoods are

\[
\log \frac{dP_\theta}{d\nu}(x) - \beta(\theta) \quad \text{and} \quad \log \frac{dP_\theta}{d\mu}(x) - \beta(\theta)
\]

(1.2)

where \( \beta(\theta) \) is the common penalty. Thus, the difference of two penalized likelihood functions is \( \log(d\mu/d\nu)(x) \) which does not depend on \( \theta \).

It is easy to see from (1.2) that the maximum penalized likelihood estimator is the posterior mode using a prior density proportional to \( \exp(-\beta(\theta)) \) with respect to a \( \sigma \)-finite measure \( \Lambda \). When the penalty term \( \beta(\theta) \) is fixed, changes of \( \Lambda \) results in changes of the prior because the prior probability of a set \( B \) is proportional to \( \int_B \exp(-\beta(\theta)) \Lambda(d\theta) \) which surely depends on \( \Lambda \). Hence, penalized likelihood inference in a Bayesian sense is not invariant because the change of a dominating measure changes the prior. Therefore, Bayesian versions of penalized likelihood inferences are non-invariant.

The main causes of noninvariance are thus choices of reparametrizations, data transformations and dominating measures. Also each inference method might have its own cause of noninvariance, e.g., posterior means as point estimators are not invariant to reparametrization.

In Chapter 2, the invariance under data transformation and the choice of the dominating measure of the model are studied when we are concerned with model checking.
Our results are also applicable in Chapter 3 when we are concerned with the effects of the prior. The invariance under reparametrization and the choice of the dominating measure for the prior are studied in Chapter 4 when we are concerned with inferences about $\theta$. In all chapters, we focus on tests of null hypotheses using measures of surprise. Brief summaries concerning measures of surprise and tests of null hypotheses are provided in Sections 1.3.

1.3 Measures of Surprise and Hypothesis Assessment

Statistical hypothesis testing generally has two characteristics, the level of the test and the power of the test. The concept of a most powerful test was introduced by Neyman and Pearson (1933). When null and alternative hypotheses are simple, a level $\alpha$ test having the biggest power is constructed using the densities of the two hypotheses. Roughly speaking, the most powerful test rejects the null hypothesis if the observed data is in the region where the alternative hypothesis is more probable relative to the null hypothesis. In other words, when the observed data are in the rejection region of a test, the data is surprising if it really comes from the null hypothesis. This idea underlies most hypothesis testing procedures.

In many problems no alternatives are considered to the null hypothesis. For example, model checking procedures test the model without any alternative models. Further many model checking procedures involve comparing the observed data against a fixed distribution. For example, if $T(x)$ is a minimal sufficient statistic, then comparing the data $x$ to the conditional distribution given $T(x)$ is a check of the model. So we consider the problem of assessing the null hypothesis $H : X \sim P$ with no alternative hypothesis. Such an assessment is concerned with measuring surprise.

Weaver (1948, 1963) noted the following concerning the concept of surprise.

... one concludes that probability and degree of rarity are essentially identical
Chapter 1. Introduction

concepts: that a *rare* event is *interesting* or not depending on whether you consider it interesting or not and that an event is *surprising* only providing its probability is very small as compared with the probabilities of the other accessible alternatives. This requires that a *surprising* event be a *rare* event, but it does not at all require that a *rare* event be a *surprising* event.

Then, he defined the *surprise index* for a discrete model by

\[
SI(x) = \frac{\mathbb{E}(p(Y))}{p(x)} = \frac{\sum y p(y)^2}{p(x)}
\]

where \( p \) is the probability mass function of \( P \) and \( Y \sim P \). Surprise indices have been further developed by I. J. Good in a sequence of papers, Good (1954, 1956, 1957, 1971, 1982a,b, 1983a,b,c, 1985, 1989). Good generalized (1.3) for \( \lambda \in [0, 1] \), via

\[
SI_{\lambda}(x) = \begin{cases} 
\left[ \mathbb{E}(p(Y)^{\lambda}) \right]^{1/\lambda}/p(x), & \text{for } 0 < \lambda \leq 1, \\
- \log p(x) + \mathbb{E}[\log p(Y)], & \text{for } \lambda = 0.
\end{cases}
\]

Related references concerning the surprise index can be found in Bartlett (1952) (an entropy approach leading to \( SI_0 \)), Good (1955) (a computational method), Kvalseth (1987) (an application for reaction time) and Redheffer (1951) (\( SI_1 \) for Binary and Poisson models). Among these, Kvalseth (1987) argued that the surprise index for a single event in a discrete model must be

\[
SI_K(x_i) = \sum_{j \neq i} p_j^2/[p_i(1 - p_i)]
\]

rather than Weaver’s definition \( SI_1 = \sum_{j=1}^n p_j^2/p_i \). Also Kvalseth pointed out that \( SI_K \) has more meaning than \( SI_1 \) by showing that \( SI_K = 1 \) when \( X \) has a uniform distribution on \( x_1, \ldots, x_n \) and \( SI_K \) tends to 0 or \( \infty \) as \( p_i \) approaches to 1 or 0, respectively.

Weaver’s surprise index (1.3) does not have an absolute interpretation, that is, interpretations of two surprise indices having the same value are different depending on models. Good and Kvalseth suggested some ways to measure surprise in order to have
better interpretations, but their interpretations still need calibration. So, for a surprise measure, we use the $P$-value as this is easily calibrated as a probability and has the virtue of common usage. Cox and Hinkley (1974) defined a $P$-value at an observed value $x_0$, and using discrepancy statistic $T$, as

$$P\{\{x : T(x) \geq T(x_0)\}\} \quad (1.6)$$

where $T$ must satisfy the property that larger values of $T$ should “indicate” greater departure from the null hypothesis.

There are, however, several problems with (1.6). For example, it is clear that $T(x_0)$ can be a surprising value from $P_T$ in more ways than just lying out in the right tail. For example, $T(x_0)$ could be in the left tail or for that matter lie in a shallow antimode. Further, (1.6) is restricted to univariate $T$ and it is very natural for multivariate $T$ to arise, e.g., we are using several univariate discrepancy statistics for model checking and we need to take into account their dependencies when computing a joint measure of surprise.

In the discrete case the $P$-value

$$P\{\{x : p_T(T(x)) \leq p_T(T(x_0))\}\} \quad (1.7)$$

seems very natural as the appropriate approach to assess whether or not $T(x)$ is surprising and it deals with all the problems raised above. The continuous analog of this is

$$P\{\{x : f_T(T(x)) \leq f_T(T(x_0))\}\} \quad (1.8)$$

but there are problems with (1.8) involving invariance under change of variables. This issue will be addressed in Chapter 2.

A Bayesian model can also be checked using a measure of surprise. For example, prior predictive distributions are suggested by Box (1980) and Evans and Moshonov (2006, 2007), and posterior predictive distributions are suggested by Rubin (1984) and Gelman, Meng and Stern (1996).
There are some related references: Barndorff-Nielsen (1976) developed something called *plausibility inference* which is very similar to the surprise inference, Bayarri and Berger (1997) summarized surprise inferences concerning Bayesian models, Bayarri and Morales (2003) dealt with outlier detection problems using surprise inferences as defined in Bayarri and Berger (1997), and Baldi developed a measure of surprise using entropies in Baldi (2005) and Itti and Baldi (2009). Baldi’s surprise measures are used in capturing eye movements in video streams.

1.4 Summary of the Results in this Thesis

1.4.1 Invariant $P$-value for Model Checking and Checking for Prior-Data Conflict

Both model checking and checking for prior-data conflict are assessments of a null hypothesis without any specific alternative hypothesis. Hence, we conduct these assessments using a measure of surprise as discussed in Section 1.3. There are, however, several issues that must be solved. Among them we present three important issues.

First, the interpretation of Good’s surprise index (1.4) is not clear. So we choose to measure surprise via $P$-values. Moreover, when we want to assess the hypothesis $H_0 : x \sim P$ via statistic $T$ then we use the $P$-value given by (1.7) when the discrete case holds. As discussed, this assesses whether values in any tail or antimodes are surprising and is applicable for multidimensional case. Of course, (1.7) is invariant under one-to-one data transformations.

Second, the $P$-value for continuous distributions would appear to be given by (1.8), as the natural analog of (1.7). However, if $W : \mathcal{T} \to \mathcal{W}$ is one-to-one and smooth, then the $P$-value (1.8) based on $W$ is given by

$$P_W(\{w : f_W(w) \leq f_W(w_0)\}) = P_T(\{t : f_T(t)J_W(t) \leq f_T(t_0)J_W(t_0)\})$$

(1.9)
and it is clear that (1.9) will not equal (1.8) unless \( J_W \) is constant. So (1.8) is not invariant and this leads to serious doubt, as to its validity.

In Chapter 2, we address this problem and provide a definition of a \( P \)-value that conforms to (1.7) in the discrete case and is invariant in the continuous case. This definition is motivated as follows. Whenever we use a continuous model we are in essence approximating a true underlying discrete model. This is because the original response \( x \) arises as the result of same measurement process and any measurement has some finite accuracy. This entails that, whenever we use a continuous model we acknowledge the underlying discrete reality and do not allow the effects of continuity to influence inferences. In particular, we must adjust our \( P \)-values so that the effects of volume distortions caused by the transformation \( T \) do not affect the \( P \)-value. Taking this into account leads directly to an invariant \( P \)-value.

Third, when there are several discrepancies \( T_1, \ldots, T_k \), it is perhaps not obvious how we should combine these to compute a single \( P \)-value. It is clear, however, that this should involve the joint distribution \( (T_1, \ldots, T_k) \) as this takes into account the dependencies among these statistics. An invariant analog of (1.8) is presented as the natural approach to this problem.

The material in Chapter 2 has been published as Evans and Jang (2008) and a subset of this has appeared as Evans and Jang (2010).

### 1.4.2 Weak Informativity and the Information in One Prior Relative to Another

One of the criticisms of Bayesian data analysis is subjectivity. A subjective prior could put a large amount of information into the data analysis. One resolution of subjectivity is to find noninformative priors which are supposed to contain no information concerning the true value of the parameter. There are several recommendations for noninformative priors, for example, reference priors are introduced in Bernardo (1979), Berger and
Bernardo (1992) and Berger, Bernardo and Sun (2009). But reference priors are not the only recommended noninformative prior, for example, Fraser et al. (2009) introduced the default prior which leads to higher order inferences. In general there is some ambiguity as to the correct definition of “noninformative.”

A sampling model may not explain the observed data appropriately if it does not pass model checking. Inferences based on such a model are then suspect. Similarly, in a Bayesian data analysis, if a prior placed on a sampling model conflicts with the observed data, then inferences are again suspect. So before making inferences, we check the sampling model first and then check for prior-data conflict. Note that we check prior-data conflict under the assumption that the sampling model passed model checking. The underlying idea of checking for prior-data conflict is measuring the surprise of the observed data $T(x_0)$, where $T$ is a minimal sufficient statistic for the model, via a $P$-value based on the prior predictive measure $M_T$ for $T$. For further discussion and motivation for this see Evans and Moshonov (2006, 2007).

Noninformative priors are not easy to deal with and can lead to anomalous results. Furthermore, it is not clear that they are actually noninformative, and so the current common use of the terminology “reference prior.” So Gelman (2006) and Gelman et al. (2008) introduced the intuitively satisfying notion of weakly informative priors as a compromise between informative and noninformative priors. These papers did not, however, give a precise definition of weakly informative prior. In Chapter 3, we give a precise definition of weak informativity. This involves the notion of a prior-data conflict and make use of the $P$-value developed in Evans and Moshonov (2006, 2007) to assess this. In particular, we say a prior $\Pi_2$ is weakly informative relative to a prior $\Pi_1$, if $\Pi_2$ leads to fewer prior-data conflicts than $\Pi_1$, a priori $\Pi_1$. Here $\Pi_1$ plays the role of a base or “true” prior that has been elicited. The motivation for choosing $\Pi_2$ is then that the analyst wishes to be conservative with respect to the subjective information they place into the analysis but still reflect some of the information contained in $\Pi_1$. Note that a
truly noninformative prior would never produce any prior-data conflict.

We note that the $P$-values developed in Evans and Moshonov (2006, 2007) need to be modified, using the approach of Chapter 2, so that they are invariant in the continuous case.

Intuitively, a prior having bigger variance is less informative. Using weak informativity, we prove that this intuition is correct for location normal models and normal linear regression models with normal priors but not when $t$ priors are used. We also show, however, that this is not true for logistic regression models in a very dramatic way. In general, prior variances are not a surrogate for assessing weak informativity.

When $T$ is a complete sufficient statistic, the marginal prior predictive distribution $M_T$ is the relevant distribution to use for assessing prior-data conflict. But if $T$ is not complete and there is an ancillary $U$, then we must remove the ancillary variation before making the assessment, i.e., use $M_T(\cdot | U)$. If there are many maximal ancillaries that are not equivalent, then the conditional inferences are different depending on the choice of the maximal ancillary. This phenomenon also affects prior-data conflict as well as weak informativity. We address this issue in Chapter 3.

The results in Chapter 3 have been published in Evans and Jang (2009).

1.4.3 An Invariant Bayesian Inference - Relative Surprise Inference

One very popular method for Bayesian data analysis is highest posterior density (HPD) inference. But this method is not invariant to the choice of dominating measures or reparametrizations. Evans (1997) suggested relative surprise inferences which compare belief changes from a priori to a posteriori. Relative surprise inferences are invariant under reparametrizations. Some representative properties of relative surprise inferences are summarized in Chapter 4.

The ratio of belief a posteriori to a priori is called a relative belief ratio and is
the basis for relative surprise inferences. Relative belief ratios are very closely related to Bayes factors which also measure change in belief but on the odds scale as opposed to the probability scale. For example, $RB(H) = \Pi(H \mid x)/\Pi(H)$ and $BF(H) = [\Pi(H \mid x)/\Pi(H^c \mid x)]/[\Pi(H)/\Pi(H^c)] = RB(H)/RB(H^c)$. Also $BF(H) \approx RB(H)$ when $\Pi(H) \approx 0$.

HPD regions depend on the choice of the dominating measure. If an optimal property of HPD regions is not invariant to reparametrization, we cannot say this property is truly optimal. Relative surprise regions are optimal with respect to the prior probability measure, that is, relative surprise regions have the smallest prior probability content among all credible regions having the same or larger posterior probability content. Also relative surprise regions have the biggest relative belief ratio and Bayes factor among all credible regions having the same posterior probability.

Bayesian point estimators are usually obtained via classical Bayesian decision theory. Such estimators are called Bayes rules. We show in Chapter 4 that the point estimator arising from relative surprise inferences, namely, the least relative surprise estimator (LRSE), is the limit of Bayes rules. In essence this shows that relative surprise inferences also have a decision theoretic interpretation. The loss function in this case has the appealing property of depending on the prior.

We prove the consistency of least relative surprise estimators for broad classes of distributions. In order to construct asymptotic credible regions, Bayesian asymptotic normality is also discussed.

### 1.4.4 Inferences for Infinite Dimensional Models

Statistical models are getting bigger and more complicated as data in applications are getting more complex. Models for such data are often nonparametric/semiparametric models or models having an increasing number of parameters as the size of data increases. Data analysis for such models is somewhat harder than data analysis for parametric models.
models. Only a few results are known for these models when using relative surprise inferences.

In Chapter 5, we show consistency of the least relative surprise estimator and its relationship to posterior consistency, that is, convergence of relative surprise regions imply posterior consistency. Also conservative prior selection is studied for some models having increasing number of incidental parameters.

There are still many issues to be solved in this area. For example Bayesian asymptotic normality for these models are still needed to be studied. Some aspects of future research problems are also discussed in this chapter.
Chapter 2

Invariant $P$-value for Model Checking and Checking for Prior-Data Conflict

$P$-values have been the focus of considerable criticism based on various considerations. Still the $P$-value represents one of the most commonly used statistical tools. When we are assessing the suitability of a single hypothesized distribution, it is not clear that there is a better choice as a measure of surprise. This chapter is concerned with the definition of appropriate model-based $P$-values for model checking and checking for prior-data conflict.

2.1 Introduction

The use of $P$-values is common in statistical practice. Despite this it is reasonable to say that the logical foundations for the $P$-value are somewhat weak. This has lead to a variety of criticisms of $P$-values and even to doubts as to their correctness. See, for example, the discussions in Berger (1985), Bernardo and Smith (2000), Morrison and Henkel (1970), and Royall (1997). The purpose of this chapter is to examine the foundations of the
$P$-value concept and attempt to provide a version of the $P$-value that addresses at least some of the issues raised concerning their validity.

While many of the criticisms concern applications of $P$-values where there are a number of distributions to choose among as candidates for the true distribution, we focus on a special case. The following situation arises in many statistical contexts and could be considered almost the archetypal statistical problem. Suppose we observe a value $x_0 \in \mathcal{X}$, where $\mathcal{X}$ is a sample space which is assumed to be a subset of a finite dimensional Euclidean space throughout this chapter, and this value was presumed to have been generated via a prescribed probability measure $P$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. The question of interest is then: given the evidence presented by $x_0$, is $P$ a reasonable choice? In certain situations we could answer this with a categorical no, that is, suppose that $P$ concentrates on $C$ and $x_0 \notin C$. While this can arise, it is typical in applications that $x_0$ is a possible value from $P$ but, if $x_0$ is in a region where $P$ assigns relatively little probability, then we feel we have evidence against $P$. Note that $x_0$ not occurring in such a region is not evidence in favor of $P$, as there are many probability distributions with this property and we are not selecting among them. In general, we are only looking here for evidence that suggests that a specific $P$ is inappropriate.

As an example of this, consider model checking where $P$ corresponds to the conditional distribution of the data given a minimal sufficient statistic, or where $P$ corresponds to the distribution of an ancillary statistic. Then evidence against $P$ is evidence against assumptions we have made as part of a statistical analysis. Model checking is an important and necessary part of statistical analyses. In Bayesian analyses $P$ could correspond to the prior predictive distribution of a minimal sufficient statistic given an ancillary and, as discussed in Evans and Moshonov (2006, 2007), we want to assess whether or not the observed value of the minimal sufficient statistic is a reasonable value from this distribution. This is a check as to whether or not the prior is in conflict with the data.

In general then, we are looking for a measure of how surprising the observed value $x_0$
is as a possible value from $P$. A common approach to this is via a \textit{discrepancy statistic} as explained in Cox and Hinkley (1974). A discrepancy statistic is a real-valued statistic $T : \mathcal{X} \rightarrow \mathbb{R}$, which measures how divergent the value $x$ is. The discrepancy statistic $T$ must satisfy the property that the bigger the value $T(x)$ is, the more surprising $x$ is. Then the $P$-value of the observed data $x_0$ is computed using

$$P(\{x : T(x) \geq T(x_0)\}). \quad (2.1)$$

This $P$-value measures the probability of observing more surprising data. So if this $P$-value is small, then we interpret this as evidence that $x_0$ is a surprising value and so we have evidence against $P$. In general, no guidance is provided as to how the statistic $T$ is to be chosen with respect to $P$.

In the choice of $T$, we usually consider characteristics of $P$ rather than $P$ itself since $P$ is not easy to deal with. Suppose we check whether or not $x_0$ comes from $P$ via a characteristic of $P$ as measured by statistic $T$. For example, we could consider the sample mean, the sample variance, or sample quantile points as a characteristic of $P$. If $T$ satisfies the requirement the bigger value $T(x)$ is the more surprising $x$ is, then $T$ can be used as a discrepancy statistic. In general $T$ does not satisfy this. So (2.1) does not have an appropriate interpretation in general. In particular, the right tail of the distribution of $T$ should be the only region that has relatively low probability. Otherwise, we could have an extreme value of $T(x_0)$ in the left tail, or a value that occurs near a shallow anti-mode, that leads to a reasonable value of (2.1) and yet $T(x_0)$ could still be considered as surprising and so evidence against $P$.

As an example of this consider the situation where $\#(\mathcal{X}) = k < \infty$, $x_0$ corresponds to a sample of $n$ and we use the $\chi^2$ statistic $T$ as the discrepancy statistic. Then, for $n$ large, (2.1) corresponds to the right-tail of a $\chi^2(k-1)$ distribution. If $k$ is also large, however, a value of $T$ in the left-tail does not provide evidence against $P$ via (2.1), even though we know that it is very unlikely for $P$ to have produced such a value. We will see in Example 2.1, however, that this apparent ambiguity can be explained when we are
more careful about defining $P$-values.

The choice of the discrepancy statistic $T$ also poses some problems. It seems clear that the choice of such a statistic should be made prior to seeing the data. Further, if we think of choosing $T$ to check for some particular characteristics of $P$, then there seems to be no reason why we should be restricted to a specific choice but may well have $T = (T_1, \ldots, T_m)'$ where the $T_i$ are different characteristics. Clearly, it is more appropriate then to compare $T(x_0)$ to $P_T$ rather than compute (2.1) for each $T_i$, that is, the dependencies among the $T_i$ should be taken into account when making the assessment, as $T(x_0)$, the collection of several characteristics, may be a surprising value even when each characteristic, $T_i(x_0)$, isn’t. We note, however, that (2.1) does not tell us how to construct a discrepancy statistic when we have chosen multiple characteristics of $P$ to compare with.

In Section 2.2, we present the basic motivation and definition of a $P$-value as a measure of surprise for $x_0$ as a possible value from $P$, when $P$ is discrete. For a characteristic $T$ with discrete marginal probability measure $P_T$, this leads to the $P$-value given by

$$P(p_T(T(x)) \leq p_T(T(x_0)))$$

where $p_T$ is the probability function associated with $P_T$. It is easily seen that this $P$-value corrects the difficulties with (2.1) that we have just described. We then argue that (2.2) is the basis for the development of $P$-values for the problems considered in this chapter and, in Sections 2.3 and 2.4, develop a general approach for the computation of $P$-values based on an observed measured response. The essential difficulty involves developing a version of (2.2) when $P_T$ is continuous, as naïve analogs will not be invariant under smooth, one-to-one transformations of $T$ (It can be generalized to diffeomorphisms). The central idea that leads to a resolution of this problem, and so to the development of an invariant $P$-value under such transformations, is that volume distortions must not affect the computation of $P$-values. We will refer to such a $P$-value as a model based $P$-value. We will also argue that there are contexts where we do not want the statistician to be
free to choose the discrepancy measures and their associated $P$-values, but want them to be determined by the model. This arises in checking for prior-data conflict and we will discuss this issue in Section 2.5. In Section 2.6.3 we discuss some computational issues.

Many other criticisms of $P$-values are often cited. In particular, a common complaint, is that in reality the data is never distributed exactly as $P$, but $P$ may be adequate for the application at hand in the sense that it provides a good approximation. If we observe enough data, however, any reasonable $P$-value will detect the discrepancy and lead to evidence against $P$. Of course, this cannot be viewed as a criticism of the $P$-value in question, as it is doing the right thing. Rather it suggests that in such problems we really do have to say what size of discrepancies are meaningful and then assess whether or not the discrepancy detected is to be taken seriously. So the $P$-value is not the end of the story in model assessment and cannot cover-up modelling inadequacies, namely, situations where we really can’t say what discrepancies are meaningful. We do not view the necessity of taking into account practical significance, as opposed to statistical significance, as a criticism of the $P$-value.

In Schervish (1996) the use of certain frequentist $P$-values was examined as a measure of support for a hypothesis. The analysis there demonstrated convincingly that there is little logical support for this. As mentioned, we are using $P$-values as measures of surprise, not support, and restricting our discussion to the situation where we have a single $P$. In Berger and Delampady (1987) and Berger and Sellke (1987) comparisons are made between frequentist $P$-values and Bayesian measures of evidence in the context of hypothesis testing, i.e., assessing the evidence in favor of a point null hypothesis $H_0 : \theta = \theta_0$. Arguments are presented there in favor of the Bayesian measures. For our discussion here, however, we are restricting to a single $P$ and agree that we might proceed very differently in situations where alternatives to $P$ are prescribed, that is, the computation of the $P$-values we discuss here may well not be appropriate in such situations, as we have more information available. In Section 2.7 we provide some discussion of the relevance
of the developments here to Bayesian contexts.

### 2.2 Model-based $P$-values with Discrete $P$

Suppose that $P$ is discrete with probability function given by $p(x) = P(\{x\})$. Then an obvious model-based $P$-value for checking whether or not $x_0$ is a surprising value from $P$ is given by

$$P(\{x : p(x) \leq p(x_0)\}).$$

(2.3)

We see that (2.3) is the probability of observing a value whose probability of occurrence is no greater than the probability of occurrence of the observed $x_0$. If (2.3) is small then it seems clear that $x_0$ is surprising and we have evidence against $P$. In effect (2.3) serves as the basic definition of a $P$-value for this surprise inference.

Note that the appropriate inequality in (2.3) is less than or equal to, as we want no evidence against $P$, for any $x_0 \in \mathcal{X}$, when $P$ is uniform on $\mathcal{X}$. Further, we see that (2.3) handles values in any tail region, values that lie between modes and also multidimensional $x$. It seems reasonable to refer to (2.3) as a pure model-based $P$-value. Values that are surprising are identified by the model and not by the statistician’s choices.

We may, however, have a characteristic of interest measured in $T$, for example, the sample mean, the sample median, lower 5th percentile, etc. Then $P_T$ is discrete with probability function $p_T$ and the model-based $P$-value induced by $T$ is given by (2.2). For example, suppose that $C \subset \mathcal{X}$ is such that $P(C)$ is very small. Then, with $T = I_C$, (2.2) equals $P(C)$ and we have evidence against $P$ when $x_0 \in C$.

We see that (2.3) determines whether or not $x_0$ is a surprising value based solely on the smallness of its probability of occurrence when compared to the probabilities of occurrence of other values. If $p(x_0)$ is very small compared to the other possibilities, then (2.3) seems like an appropriate measure of surprise. Consider a situation, however, where $\mathcal{X}$ has a large finite cardinality $k$ and two probability mass functions $p(x_0) = (1 + \epsilon)/k$,
Chapter 2. Invariant P-value for Model Checking

\( q(x_0) = (1 - \epsilon)/k, \ p(x) = 1/k - \epsilon/(k(k - 1)), \ q(x) = 1/k + \epsilon/(k(k - 1)) \) otherwise and \( \epsilon \) is very small. In the first case, \( p \), the \( P \)-value is \( 1 - (1 + \epsilon)/k \) and we have no evidence against \( P \), while in the second case, \( q \), it is \( (1 - \epsilon)/k \) and we have evidence against \( P \). So even though the probability distributions are very similar, the \( P \)-values are quite different. This points to the need generally to consider discrepancy statistics as checks on \( P \) rather than relying solely on (2.3).

We now consider an important example.

**Example 2.1** (Multinomial(1; \( \theta_1, \ldots, \theta_k \))). Suppose we observe a sample \( x_{10}, \ldots, x_{n0} \) that is supposed to have come from the Multinomial(1, \( \theta_1, \ldots, \theta_k \)) distribution where \( \theta_1, \ldots, \theta_k \) are known values that are all positive. Then, denoting the cell counts by \( T(x_{10}, \ldots, x_{n0}) = (t_{10}, \ldots, t_{k0}) \), the \( P \)-value (2.3) is given by

\[
P_T(\{ t : \theta_1^{t_1} \cdots \theta_k^{t_k} \leq \theta_1^{t_{10}} \cdots \theta_k^{t_{k0}} \})
\]

where \( P_T \) is the Multinomial(n, \( \theta_1, \ldots, \theta_k \)) distribution. We can write (2.4) as

\[
P_T(\{ t : \sqrt{n} \sum_{i=1}^{k} (\log \theta_i)(t_i/n - \theta_i) \leq \sqrt{n} \sum_{i=1}^{k} (\log \theta_i)(t_{i0}/n - \theta_i) \}).
\]

Putting \( \sigma^2 = \sum_{i=1}^{k} (\log \theta_i)^2 \theta_i - (\sum_{i=1}^{k} \theta_i \log \theta_i) \), then

\[
\sqrt{n} \sum_{i=1}^{k} (\log \theta_i)(t_i/n - \theta_i) \xrightarrow{d} N(0, \sigma^2)
\]

so (2.4) is equivalent in probability to \( \Phi(\sqrt{n} \sum_{i=1}^{k} (\log \theta_i)(t_{i0}/n - \theta_i)/\sigma) \) and the joint asymptotic normality of the \( t_{i0}/n \) implies that (2.4) is asymptotically uniform, when the probabilities \( \theta_1, \ldots, \theta_k \) are correct.

Now observe that \( \sum_{i=1}^{k} (\log \theta_i)(t_{i0}/n - \theta_i) \xrightarrow{a.s.} \sum_{i=1}^{k} (\log \theta_i)(p_i - \theta_i) \) for some \( p_i \), and so we will find evidence against the probabilities \( \theta_1, \ldots, \theta_k \), for large enough \( n \), whenever \( \sum_{i=1}^{k} (\log \theta_i)(p_i - \theta_i) < 0 \). Note that this holds whenever the expected value of \( -\log \theta_i \) under the true distribution is greater than the entropy of the assumed distribution. If we take \( E(-\log \theta_i) \) as a measure of diffuseness of a Multinomial(1, \( p_1, \ldots, p_k \)) distribution,
then this says we will inevitably find evidence against \( P \) whenever the true distribution is more diffuse, but not otherwise. For example, when \( k = 2 \) and \( \theta \neq 1/2 \), this is equivalent to
\[
p \log(\theta/(1 - \theta)) < \theta \log(\theta/(1 - \theta))
\]
which occurs when \( \theta > 1/2 \) and \( p < \theta \) or when \( \theta < 1/2 \) and \( p > \theta \). When each \( \theta_i = 1/k \), then we will never find evidence against the uniform distribution and this makes sense as the uniform distribution is the most diffuse distribution.

The check based on (2.4) will only detect certain discrepancies and this is true of most discrepancy statistics. Of course, we can also consider other discrepancy statistics and perhaps it is natural to consider \( T \) itself. So in this case we need to evaluate
\[
P_T\{t : p_T(t) \leq p_T(t_0)\}
\tag{2.5}
\]
where \( p_T(t) = \binom{n}{t_1 \ldots t_k} \theta_1^{t_1} \cdots \theta_k^{t_k} \). The following result, proved in the Appendix, gives an asymptotic approximation to (2.5) and, while the result seems close to a multinomial version of De Moivre’s theorem, an additional complexity arises due to the dependence of \( p_T(t_0) \) on \( n \) as well.

**Theorem 2.1.** Suppose that \( \theta_i \neq 0 \) for \( i = 1, \ldots, k \) and we have a sample of \( n \) from a Multinomial(1, \( p_1, \ldots, p_k \)) distribution. Then, as \( n \to \infty \),

(i) when \( \theta_i = p_i \) for all \( i \),
\[
-\log p_T(t) - \frac{1}{2} \sum_{i=1}^{k} \log \theta_i + \frac{k - 1}{2} \log 2\pi n \to \sum_{i=1}^{k} \frac{(t_i - n\theta_i)^2}{2n\theta_i}
\]
and so has a limiting \( \chi^2(k - 1) \) distribution,

(ii) when \( \theta_i = p_i \) for all \( i \), the \( P \)-value (2.5) converges in probability to \( P(X \geq \sum_{i=1}^{k} (t_{i0} - n\theta_i)^2/n\theta_i) \) where \( X \sim \chi^2(k - 1) \),

(iii) when \( \theta_i \neq p_i \) for some \( i \), the \( P \)-value (2.5) converges in probability to 0.

We note that (iii) says that the model-based \( P \)-value based on \( T \) will always detect when the assumed distribution is wrong, provided \( n \) is large enough. Also, we see that
the Pearson’s \( \chi^2 \)-test statistic arises directly as an approximation when computing the model-based \( P \)-value (2.5) and this adds support for the use of this statistic. In the typical development of the \( \chi^2 \)-test, the statistic is developed via intuition and then its asymptotic distribution is derived using an integration approximation while, in Pearson (1900), the statistic is developed from the quadratic form in the exponential of a multivariate normal density approximating the multinomial distribution. Finally, we see that computing the right-tail probability for the chi-squared test is the correct approximation to (2.5). So, for large \( n \), if the chi-squared statistic is small, then (2.5) is large. Bayesian uses of the chi-squared test statistic for model checking are discussed in Johnson (2004).

We consider another application of (2.2) where we are checking a model comprised of a number of distributions.

**Example 2.2** (Fisher’s exact test). Consider a Multinomial(1, \( \alpha_1 \beta_1, \alpha_1 \beta_2, \ldots, \alpha_k \beta_l \)) distribution where \( \alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1, \beta_j \geq 0, \sum_{j=1}^{l} \beta_j = 1 \) and these values are unknown. Then, for a sample of size \( n \), the cell counts satisfy

\[
T = (T_{11}, \ldots, T_{kl}) \sim \text{Multinomial}(n, \alpha_1 \beta_1, \alpha_2 \beta_1, \ldots, \alpha_k \beta_l).
\]

Further, it is easy to show that the marginal counts \( U = (T_{1+}, \ldots, T_{k+}) \) and \( V = (T_{+1}, \ldots, T_{+l}) \) comprise a minimal sufficient statistic. Accordingly, we have that the conditional distribution of \( T \) given \( U \) and \( V \), is completely independent of the \( \alpha_i \) and \( \beta_j \) and so is a fixed known distribution. If the observed value of \( T \) is surprising when compared to this conditional distribution, then we have evidence against the model, or in this case, the independence of the classifying variables. Then (2.2) is given by

\[
P_T(\{t : p_T(t \mid U = u^*, V = v^*) \leq p_T(t^* \mid U = u^*, V = v^*)\} \mid U = u^*, V = v^*)
\]

where \( t^*, u^*, \) and \( v^* \) are the observed values of \( T, U, \) and \( V \) respectively, and

\[
p_T(t \mid U = u^*, V = v^*) = \left\{ \prod_{i=1}^{k} \prod_{j=1}^{l} t_{ij} \right\}^{-1}.
\]
We then need to determine those $t$ that satisfy $\prod_{i=1}^{k} \prod_{j=1}^{l} t_{ij}! \geq \prod_{i=1}^{k} \prod_{j=1}^{l} t_{ij}^{*}!$ and the marginal constraints $t_{i+} = t_{i+}^{*}, t_{+j} = t_{+j}^{*}$ for all $i$ and $j$, to determine the $P$-value.

For example, when $k = l = 2$, then $(T_{11}, T_{21}, T_{12}, T_{22})$ is completely determined by $T_{11}$ when $(T_{1+}, T_{2+})$ and $(T_{+1}, T_{+2})$ are given and, as is well-known, the conditional distribution of $T_{11}$ is Hypergeometric($n, T_{1+}, T_{1+}$). When $n = 20, T_{1+} = 6$, and $T_{+1} = 12$, then the conditional probability function of $T_{11}$ is given by:

<table>
<thead>
<tr>
<th>$t$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{T}(\cdot</td>
<td>U,V)$</td>
<td>0.001</td>
<td>0.017</td>
<td>0.119</td>
<td>0.318</td>
<td>0.358</td>
<td>0.163</td>
</tr>
</tbody>
</table>

Table 2.1: Conditional probability function for Fisher’s exact test in Example 2.2

When $T_{11} = 2$, then (2.2) equals $(0.001 + 0.017 + 0.119) + 0.024 = 0.161$. Note that we have added values asymmetrically from the left and right tails of this distribution to obtain the $P$-value.

### 2.3 Model-based $P$-values with General $P$

Additional considerations arise when $P$ is a continuous measure. We restrict our attention to the absolutely continuous case so that $P$ has density $f$ with respect to a support measure $\mu$. The natural analog of (2.3) is then

$$P(\{x : f(x) \leq f(x_0)\})$$

(2.6)

and this is the probability of observing a value whose density is no greater than the density of the observed $x_0$.

The $P$-value (2.6), however, has a disturbing feature. Suppose we change the support measure from $\mu$ to $\nu$ where $\nu(A) = \int_A g(x) \mu(dx)$ for some integrable, nonnegative $g$. Then the density of $P$ with respect to $\nu$ is $f/g$ and (2.6) becomes $P(\{x : f(x)/g(x) \leq f(x_0)/g(x_0)\})$ which will generally be quite different than (2.6).
Another manifestation of the nonuniqueness of (2.6) arises when we consider one-to-one transformations of $\mathbf{x}$. Suppose that $\mathcal{X}$ is an open subset of $\mathbb{R}^k$, $\mu_k$ is volume measure, and $W : \mathcal{X} \to \mathcal{X}$ is one-to-one and sufficiently smooth. Then the density of $w = W(x)$ with respect to $\mu_k$ is given by $f_W(w) = f(W^{-1}(w))J_W(W^{-1}(w))$ where $J_W(x)$ is the reciprocal of the Jacobian determinant of $W$ at $x$. We see that (2.6) applied to $w$ becomes

$$P_W(\{w : f_W(w) \leq f_W(w_0)\}) = P(\{x : f(x)J_W(x) \leq f(x_0)J_W(x_0)\})$$

where $w_0 = W(x_0)$ and again this will typically be different than (2.6) unless $J_W$ is constant, e.g., when $\mathcal{X} = \mathbb{R}^k$ and $W$ is an affine transformation.

So it is clear that we cannot just write down a density and compute (2.6) as a model-based $P$-value. Still, the fact that we can do this in a satisfactory way in the discrete case, leads us to believe that there must be an appropriate resolution of this problem in more general contexts.

In measure-theoretic terms a density $f$, with respect to a support measure $\mu$, is seen simply as a device to compute probabilities. In statistical contexts, however, a density plays a somewhat greater role. For example, if $f(x_1) > f(x_2)$, then we want to say that the probability of $x_1$ occurring is greater than the probability of $x_2$ occurring. For this to hold we can’t allow $f$ to be defined in an arbitrary fashion. In effect we need to have that $P(A)/\mu(A) \to f(x)$, as the set $A$ converges to $\{x\}$, as then $P(A) \approx f(x)\mu(A)$ when $A$ is close to $\{x\}$. Further, to compare the probabilities of two points $x_1$ and $x_2$ we need $A_i \to \{x_i\}$ with $\mu_k(A_1) = \mu_k(A_2)$ and then, for example, we can say that the probability of $x_1$ occurring is greater than the probability of $x_2$ occurring when $f(x_1) > f(x_2)$. The mathematics of making this precise is discussed, for example, in Rudin (1974), under the topic of differentiating one measure with respect to another. To use this here we suppose that $\mathcal{X}$ is an open subset of $\mathbb{R}^k$ and $P$ is absolutely continuous with respect to volume measure $\mu_k$.

It then seems natural to choose $\mu = \mu_k$ as it weights sample points equally and so $f(x)$
expresses the essence of how the probability measure is behaving at $x$. This is analogous to using counting measure as the support measure in the discrete case as then $f(x)$ has a direct interpretation as the probability of $x$. In fact any measure that weights points equally will be a positive multiple of volume measure and (2.6) is invariant under these choices for $\mu$. More generally $\mathcal{X}$ could be a manifold with locally Euclidean structure and with $\mu$ being geometric measure—the analog of volume measure on such a space—see Section 14 of Tjur (1974) for more details.

Given that we have settled on a specific support measure, the issue is then how to deal with the noninvariance of (2.6) under smooth, one-to-one transformations $W : \mathcal{X} \rightarrow \mathcal{X}$. It might seem that the only way to obtain invariance in general is to add an ingredient to the problem. We argue, however, that such an ingredient is actually implicitly part of any statistical problem where we are using a continuous distribution to model a measured response $x$. When we take this into account we can derive a version of (2.6) that is invariant under smooth transformations and that serves as a sensible definition of a model-based $P$-value. For note that, when we use a continuous probability distribution to model a variable that is being measured as part of some observational process, we are in fact thinking of the distribution as an approximation to an underlying discrete reality. For example, we measure variables to some fixed accuracy and so there is an underlying discreteness to the sample space.

To develop an invariant $P$-value we first show that (2.6) arises as an approximation to a $P$-value based on an appropriate discrete response. Suppose then that the underlying discreteness translates into a value $x$ lying in a set $B_n(x)$ such that \{\(B_n(x) : x \in \mathcal{X}\)\} forms a partition of $\mathcal{X}$ with $\mu_k(B_n(x))$ finite and constant in $x$, and such that $B_n(x)$ shrinks ‘nicely’ (see Chapter 7 of Rudin (1987)) to $x$ as $n \rightarrow \infty$. We then have that $P(B_n(x))/\mu_k(B_n(x)) \rightarrow f(x)$ as $n \rightarrow \infty$ as long as $f$ is continuous at $x$. So for $n$ large, $\frac{P(B_n(x))}{\mu_k(B_n(x))} \approx f(x)\mu_k(B_n(x))$ and $f(x)$ serves as surrogate for the probability of $x$, at least when we are comparing the probabilities of different values of $x$ occurring. Note that the
constancy of $\mu_k(B_n(x))$ in $x$ is necessary for this interpretation of $f(x)$. As a particular example of this, suppose that $\mathcal{X} = \mathbb{R}^1$ and we partition $\mathbb{R}^1$ using $\{(i-1)/n, i/n] : i \in \mathbb{Z}\}$ and $B_n(x)$ is the set $\{(i-1)/n, i/n\}$ that contains $x$.

Rather than observing $x$, the essential discreteness of the problem means that we will observe some $x_n(x) \in B_n(x)$ and the probability of observing $x_n(x)$ is $P(B_n(x))$. Note that implicitly $x_0$ is one of the values assumed by $x_n$. Then for the discrete response variable $x_n$, the appropriate $P$-value (2.3) is given by

$$
\sum_{\{x_n(x): P(B_n(x)) \leq P(B_n(x_0))\}} P(B_n(x)).
$$

We then want to show that (2.6) serves as an approximation to (2.7).

While such a result seems intuitively plausible, a general proof is not straightforward. We require some regularity conditions, as we cannot expect such an approximation to hold if we allow $f$ and the partition $\{B_n(x) : x \in \mathcal{X}\}$ to be too general (see Example 2.3). For this we use the theory of contented sets and functions as discussed in Loomis and Sternberg (1968) where it is used to develop the Riemann integral. Essentially, a bounded set $A$ is contented if its $\mu_k$-measure can be approximated arbitrarily closely by the $\mu_k$-measure of a finite union of disjoint rectangles contained in $A$ and also by the $\mu_k$-measure of a finite union of disjoint rectangles containing $A$. A bounded function $f$ with compact support is contented if it can be approximated arbitrarily closely by step functions. Further, we say that a function $f$ is locally constant at $x$ if we can find an open set containing $x$ on which $f$ is constant. For $x_0 \in \mathcal{X}$ let $LC(x_0) = \{x : f(x) = f(x_0), f \text{ is locally constant at } x\}$. In Section 2.8 we prove the following result.

**Theorem 2.2.** Suppose that

(i) $\mathcal{X}$ is a contented subset of $\mathbb{R}^k$ with positive content,

(ii) $B_n(x)$ is a rectangle containing $x$ with $\mu_k(B_n(x))$ finite and constant in $x$, and such that $B_n(x)$ shrinks nicely to $x$ as $n \to \infty$, 

(iii) \( \{ B_n(x) : x \in \mathbb{R}^k \} \) forms a partition of \( \mathbb{R}^k \) with \( \{ B_{n+1}(x) : x \in \mathbb{R}^k \} \) a subpartition of \( \{ B_n(x) : x \in \mathbb{R}^k \} \) and \( \sup_{x \in \mathbb{R}^k} \text{diam}(B_n(x)) \to 0 \) as \( n \to \infty \),

(iv) \( f \) is a continuous density function on \( \mathcal{X} \) with \( f^{-1}(A) \) contented for any interval \( A \) and such that \( LC(x_0) \) is contented with \( \mu_k(\{ f^{-1}(f(x_0)) \cap LC(x_0) \text{c} \}) = 0 \).

Then (2.7) converges to (2.6) as \( n \to \infty \).

Theorem 2.2 establishes that the appropriate discrete \( P \)-value, in the sense that we will always be measuring \( x \) to some finite accuracy, is indeed approximated by the continuous version given by (2.6), provided \( n \) is large enough. It is implicit in Theorem 2.2 that the accuracy of the discretization is effectively the same across the sample space.

We could allow for this accuracy to vary across \( \mathcal{X} \) and this would determine a different approximation to (2.7). While this is reasonable, we do not pursue this further here but note that the situation we have considered is very common.

Although the result will hold under weaker conditions, e.g., we could allow \( f \) to be piecewise smooth and allow for more general sets than rectangles in \( \{ B_n(x) : x \in \mathbb{R}^k \} \), the conditions specified seem to apply in typical applications. As previously mentioned, some restrictions are required on the density \( f \) as (2.6) does not approximate (2.7) in general. Consider the following example.

**Example 2.3** (Failure of the approximation). The fact that we cannot allow \( f \) to defined arbitrarily at points is easily illustrated. For example, suppose that \( P \) is the uniform measure on \((0, 1)\) and we take the density \( f \) to be everywhere 1 except at \( x_0 \) where \( f(x_0) = 0 \). Such an \( f \) is a valid density function but (2.7) will converge to 1 while (2.6) is 0.

We might suppose that simply adding the condition that \( f \) be everywhere continuous will suffice, but the following situation illustrates that this is not the case. Let \( q_1, q_2, \ldots \) be a listing of all rational numbers in \( \mathcal{X} = (0, 1) \). Fix a \( \delta > 0 \) sufficiently small and let \( A_0 = \mathcal{X} \cap \bigcup_{i=1}^{\infty} (q_i - \delta 2^{-i}, q_i + \delta 2^{-i}) \). For \( x \in A_0 \), there is an interval \( (a,b) \subset \mathcal{X} \) such that \( x \in (a,b) \subset A_0 \). For \( x \in A_0 \) define \( a(x) = \inf\{ a \in [0,x) \mid (a,x) \subset A_0 \} \) and
\[ b(x) = \sup \{ b \in [x, 1] \mid (x, b) \subset A_0 \}. \] Then, \((a(x), b(x)) \subset A_0\) for all \(x \in A_0\). Since \(A_0\) is a countable union of intervals, there are countably many \(x_i\)’s in \(A_0\) such that \(A_0 = \bigcup_{i=1}^{\infty} (a(x_i), b(x_i))\) and note that these intervals are disjoint.

The intervals \((a(x_i), b(x_i))\)’s can be ordered to obtain the class of intervals \(\{(a_i, b_i) : i = 1, 2, \ldots\}\) where the ordering is such that \((a_i, b_i)\) satisfies \(i < j\) whenever \(b_i - a_i > b_j - a_j\), or \(a_i < a_j\) when \(b_i - a_i = b_j - a_j\). Now let \(A = \mathcal{X} \cap A_0^c\) and define probability density \(f(x)\) on \(\mathcal{X}\) as

\[
f(x) = \begin{cases} 
1 & \text{if } x \in A, \\
1 + \frac{(b_i-a_i)^2}{\pi} \sin \frac{2\pi(x-a_i)}{b_i-a_i} & \text{if } x \in (a_i, b_i) \text{ for some } i.
\end{cases}
\]

We see immediately that \(f\) is a continuous density on \(\mathcal{X}\) and so, for every \(x \in \mathcal{X}\),

\[
\lim_{n \to \infty} P(B_n(x)) = f(x) \text{ where } B_n(x) = \mathcal{X} \cap ((\lfloor 2^n x \rfloor - 1)2^{-n}, [2^n x]2^{-n}].
\]

It can be shown, however, that whenever \(x_0 = b_i\) for some \(i\), then (2.6) equals \(1 - \alpha - \beta\), while (2.7) either doesn’t converge or converges to \(\alpha - \beta\) where \(\alpha = \sum_{i=1}^{\infty} (b_i - a_i)/2\), \(\beta = \sum_{i=1}^{\infty} (b_i - a_i)^3/\pi^2\) and \(\alpha - \beta \neq 1 - \alpha - \beta\), see Lemma 2.11 in Section 2.8 for details.

Clearly, the problem here is the excessive oscillatory behavior of \(f\) such that (iv) does not hold. This condition can be substantially weakened if \(P(\{x : f(x) = c\}) = 0\) for every \(c\). In essence the distribution of \(f(x)\) can have a discrete component, but our conditions imply that this can really only arise by \(f\) being constant on sets where it is locally constant.

We may, however, base the \(P\)-value on a statistic \(T\), such as a discrepancy statistic, and use the observed value \(T(x_0)\). The question then is: given the initial discretization on \(\mathcal{X}\) as determined by the measurement process, how should we take this into account? For, even if \(T\) is 1-1, it will give rise to volume distortions and we do not want these volume distortions to affect our \(P\)-value. This is the heart of the invariance issue and we discuss this in the next section.
2.4 Invariant $P$-values for General Statistics

Suppose $T : \mathcal{X} \to \mathcal{T}$ is a general statistic, and we want to compare $t_0 = T(x_0)$ to $P_T$ to assess whether or not we have evidence against $P$. When $P_T$ is discrete, it would seem that the relevant $P$-value is as discussed in Section 2.2 (see Example 2.6). In the continuous case, however, additional complexities arise. This is because $T$ may distort volume and we need to ensure that the $P$-value we use does not depend on these distortions.

Suppose first that $\mathcal{X}$ and $\mathcal{T}$ are open subsets of $\mathbb{R}^k$ and that $T$ is 1-1 and smooth. Then a partition element $B_n(x) \subset \mathcal{X}$, with measure $\mu_k(B_n(x))$, is transformed into $T(B_n(x))$ with measure $\mu_k(T(B_n(x))) = \mu_k(B_n(x))J_T^{-1}(x')$ for some $x' \in B_n(x)$, while the density of the transformed response with respect to $\mu_k$ (Euclidean measure on $\mathbb{R}^k$) is $f_T(t) = f(T^{-1}(t))J_T(T^{-1}(t))$. Accordingly, we cannot use the $P$-value $P_T(\{t : f_T(t) \leq f_T(t_0)\})$ to assess whether or not $t_0$, or equivalently $x_0 = T^{-1}(t_0)$ is surprising, since the density $f_T(t)$ depends on volume distortions and the sets $T(B_n(x))$ are no longer necessarily of equal volume. There is clearly an easy fix for this, however, as we simply correct for this volume distortion and compute the $P$-value

\[ P_T(\{t : f_T(t)/J_T(T^{-1}(t)) \leq f_T(t_0)/J_T(T^{-1}(t))\}) = P(\{x : f(x) \leq f(x_0)\}). \quad (2.8) \]

With this refinement the $P$-value introduced in Section 2.3 becomes invariant under 1-1, smooth transformations of the response, that is, we retain as part of the problem prescription how the continuous probability model is approximating an essentially discrete response.

In general, however, $T$ will not be 1-1. Suppose then, that $\mathcal{X}$ is an open subset of $\mathbb{R}^k$ and $\mathcal{T}$ is an open subset of $\mathbb{R}^l$ where $l \leq k$. Let $f_T$ denote the density of $P_T$ with respect to $\mu_l$ and suppose this is continuous. Suppose that $T$ is sufficiently smooth so that for each $t \in \mathcal{T}$ the set $T^{-1}(t)$ is a Riemann manifold with geometric measure on $T^{-1}(t)$ denoted by $\nu_{T^{-1}(t)}$. For example, when $T$ is 1-1, then $T^{-1}(t)$ is a 0-dimensional Riemann manifold and $\nu_{T^{-1}(t)}$ is counting measure. Results in Tjur (1974) show that, in
general,

\[ f_T(t) = \int_{T^{-1}(t)} f(x) \left| \det(dT(x) \circ dT(x)') \right|^{-1/2} \nu_{T^{-1}(t)}(dx) \]  

(2.9)

where \(dT\) is the differential of \(T\). Formula (2.9) shows directly how \(f_T\) is affected by volume distortions. For, at \(x \in T^{-1}(t)\) the contribution to the density value \(f_T(t)\) is distorted by the factor \(J_T(x) = \left| \det(dT(x) \circ dT(x)') \right|^{-1/2}\). Accordingly, just as we do in the 1-1 case, we adjust the integrand in (2.9) by dividing by the factor \(J_T(x)\) to obtain

\[ f^*_T(t) = \int_{T^{-1}(t)} f(x) \nu_{T^{-1}(t)}(dx) \]  

(2.10)

as the appropriate density to use. Note that the ratio \(f_T(t)/f^*_T(t)\) measures the effect of the volume distortion induced by \(T\) on the density \(f_T\). We then compute the \(P\)-value

\[ P_T(\{t : f_T(t) \leq f^*_T(t_0)\}) = P(\{x : f^*_T(T(x)) \leq f^*_T(T(x_0))\}) \]  

(2.11)

to assess whether or not \(t_0 = T(x_0)\) is a surprising value from \(P_T\). We see that (2.11) depends only on the density assignment \(f\) on the original response space, which is determined by how we are approximating an essentially discrete response, and the preimage sets of \(T\).

We have the following simple but significant result for (2.11).

**Theorem 2.3.** When \(\mathcal{X}\) is an open subset of \(\mathbb{R}^k\), \(T : \mathcal{X} \rightarrow \mathcal{T}\) is onto with \(\mathcal{T} \subset \mathbb{R}^l\) open, and \(T\) is sufficiently smooth, then the \(P\)-value given by (2.11) is invariant under 1-1 smooth transformations on \(\mathcal{T}\).

**Proof of Theorem 2.3.** Suppose \(W\) is a 1-1, smooth transformation defined on \(\mathcal{T}\) and \(w = W(t)\). Then, \((W \circ T)^{-1}(w) = T^{-1}(t)\) and \(f^*_{W \circ T}(w) = \int_{T^{-1}(t)} f(x) \nu_{T^{-1}(t)}(dx) = f^*_T(t)\). \(\square\)

We now consider some applications and note that these support (2.11) as the appropriate definition of an invariant \(P\)-value.
Example 2.4 (A smooth one-to-one transformation). Suppose that $T : \mathcal{X} \to \mathcal{Y}$ is one-to-one. Then $T^{-1}(t)$ is a singleton set. Any discrete set of points is a 0-dimensional Riemann manifold and geometric measure is counting measure. Therefore, $f^*_T(t) = \int_{T^{-1}(t)} f(x) \nu_{T^{-1}(t)}(dx) = f(T^{-1}(t))$ and (2.11) equals (2.8).

Example 2.5 (A smooth $k$-to-one transformation). Suppose that $T^{-1}(t) = \{x_1(t), \ldots, x_k(t)\}$ for each $t$. Then we have that $f_T(t) = \sum_{i=1}^k f(x_i(t))J_T(x_i(t))$ and note that the volume distortion $J_T(x_i(t))$ could vary with $i$. In this case, we have that $\nu_{T^{-1}(t)}$ is counting measure and the corrected density is $f^*_T(t) = \sum_{i=1}^k f^*(x_i(t))$. As in Example 2.4, it seems clear here how we need to correct for volume distortions and, as such, it provides strong support for (2.11) as the relevant $P$-value.

The following example shows that (2.11) gives the correct answer in the discrete case as well.

Example 2.6 ($P_T$ is discrete). First suppose that $P$ is discrete so we can consider $\mathcal{X}$ as a 0-dimensional Riemann manifold with geometric measure equal to counting measure and similarly for $\mathcal{Y}$. In this case $\nu_{T^{-1}(t)}$ is counting measure on $T^{-1}(t)$, $J_T(x) \equiv 1$ and so $f^*_T(t) = \int_{T^{-1}(t)} f(x) \nu_{T^{-1}(t)}(dx) = \sum_{x \in T^{-1}(t)} P(X = x) = p_T(t)$ is the probability function of $T$. Note $dT$ is just the identity so there is no volume distortion.

When $P$ is continuous then, for those $t$ with $p_T(t) > 0$, we have that $\nu_{T^{-1}(t)}$ is $\mu_k$ restricted to $T^{-1}(t)$. Accordingly, $p_T(t) = \int_{T^{-1}(t)} f(x) \nu_{T^{-1}(t)}(dx) = f^*_T(t)$. So, in general, we obtain the $P$-value discussed in Section 2.2.

Example 2.7 ($J_T(x)$ is constant). Note that $J_T(x)$ is constant for all $x$ whenever $T(x)$ is an affine transformation. So we could have $T(x) = a + Bx$ for some $a \in \mathbb{R}^l$ and $B \in \mathbb{R}^{l \times k}$ when $\mathcal{X} \subset \mathbb{R}^k$. Also notice that when $x \in \mathbb{R}^n$ and $T(x)$ is the order statistic then $J_T(x)$ is constant for all $x$. It is then clear from (2.9) that, in this case, we can compute (2.11) as $P_T(\{t : f_T(t) \leq f_T(t_0)\})$. 

For example, when $T(x) = \bar{x}$ we simply use the density of $\bar{x}$ to compute the $P$-value. When $P$ is the $N(0, 1)$ distribution, then (2.11) is $2(1 - \Phi(\bar{x}))$.

As another example, suppose that $T$ is projection on the $i$-th coordinate, so $J_T(x) \equiv 1$. Then $T^{-1}(t)$ is the set of points in $\mathcal{X}$ with $i$-th coordinate equal to $t$, $\nu_{T^{-1}(t)}$ is Euclidean volume on this set, and $f^*_{T}(t)$ is the marginal density of the $i$-th coordinate. Of course, this generalizes to arbitrary coordinate projections.

**Example 2.8** ($J_T(x)$ is constant for $x \in T^{-1}(t)$). In some ways Example 2.8 is the simplest situation as the volume distortion induced by $T$ is constant on $\mathcal{X}$. We now allow for the possibility that the volume distortion is constant in $T^{-1}(t)$ but may vary with $t$.

Put $J^*_T(t) = J_T(x)$ for $x \in T^{-1}(t)$. From (2.9) we have that $f_T(t) = f^*_T(t)/J_T^*(t)$ and so (2.11) can be computed as

$$P_T(\{t : f_T(t)/J_T^*(t) \leq f_T(t_0)/J_T^*(t_0)\})$$

This permits us to avoid the integration involved in calculating $f^*_T(t)$ when we know the distribution of $T$ and can compute $J_T(x)$ easily.

As an example, suppose that $T(x) = x'x$. Then $T^{-1}(t)$ is a $(k-1)$-dimensional sphere in $\mathbb{R}^k$. Now $dT(x) = 2(x_1, \ldots, x_k)$, so $dT(x) \circ dT(x)' = 4x'x = 4t$ and $J_T(x) = (4t)^{-1/2}$ is constant for $x \in T^{-1}(t)$ for every $t$. Note that the adjustment factor applied to $f_T(t)$ is to multiply by $2t^{1/2}$ and this is precisely the distortion caused by the “quadratic” part of the transformation. The appropriate $P$-value is $P_T(\{t : f_T(t)t^{1/2} \leq f_T(t_0)t_0^{1/2}\})$. We see that in this case we must modify the usual density that we work with.

As a particular case, suppose that $x \sim N_k(0, I)$. Then $T(x) \sim \chi^2(k)$ with density $f_T(t) = \Gamma(k/2)^{-1}2^{-k/2}t^{(k/2)-1}e^{-t/2}$. Therefore, the invariant $P$-value is given by $P_T(\{t : t^{(k-1)/2}e^{-t/2} \leq t_0^{(k-1)/2}e^{-t_0/2}\})$ and only when $k = 1$ is this equivalent to $P_T(t \geq t_0)$.

Notice that when we directly observe $T \sim \chi^2(k)$, in the sense that it is a measured variable, and we discretize using equal length intervals, then the relevant $P$-value is $P_T(\{t : t^{(k/2)-1}e^{-t/2} \leq t_0^{(k/2)-1}e^{-t_0/2}\})$. As just shown, when we take into account that $T$ arises as a transformation of a measured variable, the $P$-value changes. Further, both
of these $P$-values are two-sided when $k > 1$. In contrast, the approximate $P$-value that arises in Example 2.1, for a multinomial with $k + 1$ categories, is a right-tail only $P$-value for the $\chi^2(k)$ distribution, and this follows directly from our theory.

In Section 2.5 we discuss some further examples and, in particular, some examples where $J_T(x)$ varies with $x \in T^{-1}(t)$. There are also computational issues that need to be addressed in such contexts and we discuss these in Section 2.6.3.

In Section 2.5 we also discuss another use of a transformation $W$ to assess surprise. This involves comparing the observed $x_0$ with the conditional distribution of $x$ given that $W(x) = W(x_0) = w_0$. In this case the conditional density of $x$, with respect to geometric measure on $W^{-1}(w_0)$, is given by $f(x)J_W(x)/f_W(w_0)$ and it is clear that the volume distortion at $x$, induced by the conditioning, is given by $J_W(x)$. Accordingly the relevant $P$-value, based on the full data, is given by $P\{\{x : f(x)/f_W(w_0) \leq f(x_0)/f_W(w_0)\} \mid W(x) = w_0\} = P\{\{x : f(x) \leq f(x_0)\} \mid W(x) = w_0\}$. If we have a transformation $T$ of $x$, then the relevant $P$-value is given in the following result.

**Lemma 2.4.** Suppose that $X$, $W$ and $T$ are Riemann manifolds, with geometric measures $\mu_X$, $\mu_W$ and $\mu_T$ respectively, and $W : X \to W$, $T : X \to T$ are onto and smooth. Let $\nu_{T,W,t,w}$ denote geometric measure on $T^{-1}(t) \cap W^{-1}(w)$. Then the relevant conditional $P$-value based on $T$, given $W(x) = w_0$, is

$$P_T(\{t : f^*_{T,W}(t \mid w_0) \leq f^*_{T,W}(t_0 \mid w_0)\} \mid W(x) = w_0) \quad (2.12)$$

where $t_0 = T(x_0)$, and $f^*_{T,W}(t \mid w_0) = \int_{T^{-1}(t) \cap W^{-1}(w_0)} f(x) \nu_{T,W,t,w}(dx)$.

*Proof of Lemma 2.4.* The conditional density of $T$ given $W = w$ is given by $f_{T,W}(t \mid w) = \int_{T^{-1}(t) \cap W^{-1}(w)} (f(x)/f_W(w))J_{(T,W)}(x) \nu_{T,W,t,w}(dx)$. Therefore, volume distortion induced by the transformations is $J_{(T,W)}(x)$ and the result follows.

We will need the following result concerning the composition of mappings.
Lemma 2.5. Suppose that $\mathcal{X}, \mathcal{U}$ and $\mathcal{T}$ are Riemann manifolds, with geometric measures $\mu_\mathcal{X}$, $\mu_\mathcal{U}$ and $\mu_\mathcal{T}$ respectively, and $U : \mathcal{X} \to \mathcal{U}, T : \mathcal{U} \to \mathcal{T}$ are onto smooth mappings, then

$$f_{T\circ U}(t) = \int_{T^{-1}(t)} J_T(u) \int_{U^{-1}(u)} f(x) J_{T\circ U}^{-1}(u) J_U(x) \nu_{U^{-1}(u)}(dx) \nu_{T^{-1}(u)}(du),$$

where $\nu_{U^{-1}(u)}$ and $\nu_{T^{-1}(u)}$ are the geometric measures on $U^{-1}(u)$ and $T^{-1}(t)$.

Proof of Lemma 2.5. Suppose that $g : \mathcal{X} \to \mathbb{R}$ is nonnegative, $\int_A g(x) \mu_\mathcal{X}(dx)$ is finite for compact $A$ and let $B \subset \mathcal{T}$ be open. By the measure decomposition theorem (see Tjur, 1974, Theorem 15.1) applied to $g(x) \mu_\mathcal{X}(dx)$ and $T \circ U$, we have that

$$\int_{\mathcal{X}} I_B(T(U(x))) g(x) \mu_\mathcal{X}(dx) = \int_B \int_{(T\circ U)^{-1}(t)} g(x) J_{T\circ U}(x) \nu_{(T\circ U)^{-1}(t)}(dx) \mu_T(dt).$$

We apply the measure decomposition theorem in Tjur (1974) first to $g(x) \mu_\mathcal{X}(dx)$ and $U$ and then to $\int_{U^{-1}(u)} I_B(T(U(x))) g(x) J_U(x) \nu_{U^{-1}(u)}(dx) \mu_\mathcal{U}(du)$ and $T$ to obtain

$$\int_{\mathcal{X}} I_B(T(U(x))) g(x) \mu_\mathcal{X}(dx) = \int_U \int_{U^{-1}(u)} I_B(T(U(x))) g(x) J_U(x) \nu_{U^{-1}(u)}(dx) \mu_\mathcal{U}(du)$$

$$= \int_B \int_{T^{-1}(t)} \int_{U^{-1}(u)} g(x) J_U(x) \nu_{U^{-1}(u)}(dx) J_T(u) \nu_{T^{-1}(u)}(du) \mu_T(dt).$$

From this we conclude that

$$\int_{(T\circ U)^{-1}(t)} g(x) J_{T\circ U}(x) \nu_{(T\circ U)^{-1}(t)}(dx) \mu_T(dt) = \int_{T^{-1}(t)} \int_{U^{-1}(u)} g(x) J_U(x) \nu_{U^{-1}(u)}(dx) J_T(u) \nu_{T^{-1}(u)}(du) \mu_T(dt).$$

Putting $g(x) = f(x) J_{T\circ U}^{-1}(x)$ establishes the result.

2.5 Applications

Suppose that we have a statistical model $\{P_\theta : \theta \in \Theta\}$ where $P_\theta$ is a probability measure on $\mathcal{X}$ with density $f_\theta$ with respect to support measure $\mu_k$, and $\Pi$ is a prior probability measure on $\Theta$. Let $M(A) = \int_\Theta P_\theta(A) \Pi(d\theta)$ denote the prior predictive on $\mathcal{X}$ with density $m(x) = \int_\Theta f_\theta(x) \Pi(d\theta)$ with respect to $\mu_k$. We will investigate here the relevant $P$-values for assessing the model and checking for prior-data conflict in light of an observed $x_0$.

As we will see, these $P$-values are invariant and depend only on the densities $f_\theta$. In
particular, the $P$-values do not depend on any choice of density for the prior. This makes sense because we do not directly measure the variable $\theta$, only the response variable $x$.

### 2.5.1 Model Checking

If $W: \mathcal{X} \rightarrow \mathcal{W}$ is a minimal sufficient statistic, then the conditional distribution of the data given $W$ is independent of $\theta$ and is denoted $P(\cdot | W(x) = w_0)$. Suppose we wish to check if the model makes sense in light of the observed $x_0$. By the converse of the factorization theorem we have that $f_\theta(x) = g_\theta(W(x))h(x)$. Lemma 2.4 and (2.12) give an invariant $P$-value that assesses $f_\theta$ for each $\theta$. We have the following result.

**Theorem 2.6.** The $P$-value (2.12) is given by

$$P_T \left( h_{T,W}(t | w_0) \leq h_{T,W}(t_0 | w_0) \mid W(x) = w_0 \right),$$

where $h_{T,W}(t | w_0) = \int_{T^{-1}(t) \cap W^{-1}(w_0)} h(x) \nu_{T^{-1}(t)}(dx)$, i.e., it is independent of $\theta$, and (2.13) is independent of the choice of $h$.

**Proof of Theorem 2.6.** In the continuous case we assume that each density is continuous at any observed $x_0$ and restrict attention to those $x_0$ for which $f_\theta(x_0) > 0$. When $f_\theta(x_0) > 0$, then $g_\theta(W(x_0)) > 0$ and $g_\theta(W(x)) = g_\theta(W(x_0))$ for the event $W(x) = w_0 = W(x_0)$. We have that (2.12) equals

$$P_T \left( \int_{T^{-1}(t) \cap W^{-1}(w_0)} f_\theta(x) \nu_{T^{-1}(t)}(dx) \leq \int_{T^{-1}(t_0) \cap W^{-1}(w_0)} f_\theta(x) \nu_{T^{-1}(t_0)}(dx) \mid W(x) = w_0 \right)$$

$$= P \left( h_{T,W}(t | w_0) \leq h_{T,W}(t_0 | w_0) \mid W(x) = w_0 \right).$$

Further, if $g_\theta(W(x))h(x) = g_\theta'(W(x))h'(x)$, then

$$P_T \left( h_{T,W}(t | w_0) \leq h_{T,W}(t_0 | w_0) \mid W(x) = w_0 \right)$$

$$= P_T \left( \int_{T^{-1}(t) \cap W^{-1}(w_0)} g_\theta(W(x))h(x) \nu_{T^{-1}(t)}(dx) \leq \int_{T^{-1}(t_0) \cap W^{-1}(w_0)} g_\theta(W(x))h(x) \nu_{T^{-1}(t_0)}(dx) \mid W(x) = w_0 \right).$$
Suppose that

\[ P_T \left( \int_{T^{-1}(t) \cap W^{-1}(w_0)} g_\theta(W(x)) h'(x) \, \nu_{T^{-1}(t)}(dx) \right| W(x) = w_0 \right) \leq \int_{T^{-1}(t) \cap W^{-1}(w_0)} g_\theta(W(x)) h'(x) \, \nu_{T^{-1}(t)}(dx) \]

and we are done.

\[ P_T (h'_{T,W}(t \mid w_0) \leq h'_{T,W}(t_0 \mid w_0) \mid W(x) = w_0) \]

We now consider an application of this result.

**Example 2.9** (Model checking for the location-scale normal model). Suppose that \( x = (x_1, \ldots, x_n) \) is a sample of \( n \) from the \( \mathcal{N}(\mu, \sigma^2) \) distribution with \( \mu \in \mathbb{R} \) and \( \sigma^2 > 0 \) unknown. Then \( W(x) = (\bar{x}, ||x - \bar{x}1_n||) \) is minimal sufficient. Putting \( d(x) = (x - \bar{x}1_n)/||x - \bar{x}1_n|| \), we can write \( x = \bar{x} + ||x - \bar{x}1_n||d \) and note that \( \bar{x}, ||x - \bar{x}1_n|| \) and \( d \) are statistically independent with \( d \) uniformly distributed on \( S^{n-1} \cap L^\perp \{1_n\} \). In this case \( h \) is constant (so we can take it to be 1) and \( W^{-1}(\bar{x}_0, ||x_0 - \bar{x}_01_n||) \) is the \((n-2)\)-dimensional sphere \( \bar{x}_01_n + ||x_0 - \bar{x}_01_n||(S^{n-1} \cap L^\perp \{1_n\}) \).

It is natural here to consider functions of \( d \) as discrepancy statistics for checking the model. For example, the family \( T_p \circ d = \sum_{i=1}^n d_i^p \) is of some interest as this gives effectively the skewness and kurtosis statistics when \( p = 3 \) and 4, respectively. In this case, \( h_{T,W}(t \mid w_0) \) is the volume of the \((n-3)\)-dimensional submanifold of \( \bar{x}_01_n + ||x_0 - \bar{x}_01_n||(S^{n-1} \cap L^\perp \{1_n\}) \) given by \((T_p \circ d)^{-1}(t) \cap W^{-1}(\bar{x}_0, ||x_0 - \bar{x}_01_n||)\). Alternatively, from the proof of Theorem 2.6, we can compute the invariant \( P \)-value by assuming \((\mu, \sigma) = (0, 1)\), letting \( f \) denote the density of a sample of \( n \) from the \( \mathcal{N}(0, 1) \) distribution and computing

\[ P_{(0,1)}(f_{T_p \circ d}^*(T_p(d(x)))) \leq f_{T_p \circ d}^*(T_p(d(x)))) (\bar{x}_0, ||x_0 - \bar{x}_01_n||)) = P_d(f_{T_p \circ d}^*(T_p(d))) \leq f_{T_p \circ d}^*(T_p(d))) \]

where \( f_{T_p \circ d}^*(t) = \int_{(T_p \circ d)^{-1}(t)} f(x) \nu_{T^{-1}(t)}(dx) \) and \( d \) is uniformly distributed on \( S^{n-1} \cap L^\perp \{1_n\} \). The volume distortion induced by \( T_p(d) \) can be computed explicitly as \( J_{T_p \circ d}(x) = ||x - \bar{x}1_n||/|p|T_{2p-2}(d(x)) - T_{2p-1}(d(x))/n - T_p^2(d(x)))^{-1/2} \), see Example 2.12 for details. We see that this is not a function of \( T_p \circ d \) and also \( J_{T_p \circ d}(x) = J_{T_p \circ d}(-x) \). Since \( f(x) = f(-x) \) and \((T_p \circ d)^{-1}(-t) = (-1)p(T_p \circ d)^{-1}(t) \) when \( p \) is a nonnegative integer, we have that \( f_{T_p \circ d}^*(t) \) and the density \( f_{T_p \circ d}^*(t) = \int_{(T_p \circ d)^{-1}(t)} f(x)J_{T_p \circ d}(x) \nu_{T^{-1}(t)}(dx) \)
are symmetric about 0 when \( p \) is odd. If both \( f^*_{T_p \circ d} \) and \( f_{T_p \circ d} \) are unimodal, then this implies that the \( P \)-values based on the densities, tail probabilities of \( |T_p \circ d| \) and the invariant \( P \)-values are the same when \( p \) is odd. It would appear that in general, the differences among these \( P \)-values disappear with increasing \( n \), so the need to correct for volume distortion vanishes with large sample sizes in this case, see Section 2.6.2.

In Figure 2.1 we have plotted the densities and invariant \( P \)-values for tests of skewness for several sample sizes \( n \). The \( P \)-values are two-sided. The invariant \( P \)-values are the same as those based on the density of \( T_3(d) \) and tail probabilities of \( |T_3(d)| \), for all cases except \( n = 3 \). We note that the asymptotic approximation to the exact \( P \)-value can be poor but is quite good for \( n = 100 \). When \( n = 3 \) the invariant \( P \)-value is identically equal to 1, i.e., we never find evidence against the model. In this case, we can show that \( J_{T_3}^{-2}(x) = 1/6 - T_3^2(x) \) and the density of \( T_3 \) at \( t \) is proportional to \( (1 - 6t^2)^{-1/2} \) for \(-1/\sqrt{6} < t < 1/\sqrt{6}\), i.e., all the density is due to the volume distortion caused by \( T_3 \). Notice too that the density is U-shaped with infinite singularities at the end-points. Accordingly, if we were to use the density for the \( P \)-value we would reject the model for values of \( T_3 \) near 0 and this doesn’t make sense. It wouldn’t seem to make sense to reject for large values of \( |T_3| \) either, at least based on the shape of the distribution. The invariant \( P \)-value is telling us that there is no test for skewness based on \( T_3 \) when \( n = 3 \). Intuitively this seems reasonable as we need two degrees of freedom for location and scale and to check for skewness, we need at least two more to see if there is asymmetry about the center.

In Figure 2.2 we have plotted the densities and invariant \( P \)-values for tests of kurtosis for several sample sizes \( n \). The densities are quite irregular for small sample sizes and skewed. The invariant \( P \)-values, those based on the density and tail probabilities are all different in this case. The \( P \)-values based on the densities and asymptotics are quite similar for \( n = 100 \) while this is not the case for the invariant \( P \)-values. This indicates that the volume distortion is still having an appreciable effect when \( n = 100 \).
Figure 2.1: Densities and invariant \( P \)-values for test of skewness for various sample sizes \( n \) when sampling from normal.

Figure 2.2: Densities and invariant \( P \)-values for test of kurtosis for various sample sizes \( n \) when sampling from normal.
In Figure 2.3 we have plotted the densities and invariant $P$-values based on the Jarque-Bera test statistic $n(nT^2_3/6 + (nT_4 - 3)^2/24)$ for several sample sizes $n$. This is clearly an attempt to assess both skewness and kurtosis simultaneously. The densities are quite irregular for small sample sizes and skewed. The $P$-values based on the densities and asymptotics are quite different for $n = 100$ while this is not the case for the invariant $P$-values. Again this indicates that the volume distortion is still appreciable when $n = 100$.

Figure 2.3: Densities and invariant $P$-values for Jarque-Bera test for various sample sizes $n$ when sampling from normal.

When $U$ is an ancillary statistic, then a function of $U$ can be used to assess the model. If we consider the transformation $T \circ U$, then we must evaluate $\int_{(T \circ U)^{-1}(t)} f_\theta(x) \nu_{T^{-1}(t)}(dx)$ and, in general, there is no reason to suppose that this is independent of $\theta$. When the distribution of $T \circ U$ is discrete, however, then $\int_{(T \circ U)^{-1}(t)} f_\theta(x) \nu_{T^{-1}(t)}(dx)$ is the probability function of $T \circ U$ and as such is independent of $\theta$. Theorem 2.7 will show that the $P$-value based on $\int_{(T \circ U)^{-1}(t)} f_\theta(x) \nu_{T^{-1}(t)}(dx)$ is independent of $\theta$ for a very broad class of ancillaries.

Consider the following example which will serve as an archetype for a common situa-
Section where ancillaries arise.

**Example 2.10** (Location-scale models). Suppose we have \( x \in \mathbb{R}^n \) and the model is \( x = \mu 1_n + \sigma z \) where \( z \) is distributed with density \( f \) with respect to volume measure on \( \mathbb{R}^n \), and \( \mu \in \mathbb{R}^1, \sigma > 0 \) are unknown. Then \( x \) has density \( f_{\mu,\sigma}(x) = \sigma^{-n} f((x - \mu 1_n)/\sigma) \).

We take the parameter space to be \( \Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}^1, \sigma > 0 \} \) and note that we have a group product defined on \( \Theta \) via \((\mu_1, \sigma_1)(\mu_2, \sigma_2) = (\mu_1 + \sigma_1 \mu_2, \sigma_1 \sigma_2)\). This group acts on \( \mathbb{R}^n \) via \((\mu, \sigma)x = \mu 1_n + \sigma x \). A maximal invariant is then given by \( U(x) = (x - \bar{x} 1_n)/||x - \bar{x} 1_n|| \) and this is ancillary. Note that \( U^{-1}(u) = \{x : x = a 1_n + cu, \text{ for some } (a, c) \in \Theta \} = \Theta u \), i.e., \( U^{-1}(u) \) is an orbit of the group action. Clearly, this orbit is half of a 2-dimensional plane in \( \mathbb{R}^n \) and so geometric measure \( \nu_{U^{-1}(u)} \) is just area.

If we wish to base our checking on \( U \) itself, then we must evaluate

\[
 f_{\mu,\sigma,U}(d_0) = \int_{U^{-1}(u)} f_{\mu,\sigma}(x) \nu_{U^{-1}(u)}(dx) = \int_0^\infty \int_{-\infty}^\infty f_{\mu,\sigma}(a 1_n + cu) \sqrt{\sigma} da \ dc \\
= \int_0^\infty \int_{-\infty}^\infty \sigma^{-n} f \left( \frac{a - \mu}{\sigma} 1_n + \frac{c}{\sigma} u \right) \sqrt{\sigma} da \ dc \\
= \sigma^{-(n-2)} \int_0^\infty \int_{-\infty}^\infty f(a 1_n + cu) \sqrt{\sigma} da \ dc.
\]

Accordingly the \( P \)-value for model checking is given by

\[
 P_U(f_{\mu,\sigma,U}(u) \leq f_{\mu,\sigma,U}(u_0)) = P_U \left( \int_0^\infty \int_{-\infty}^\infty f(a 1_n + cu) \ da \ dc \leq \int_0^\infty \int_{-\infty}^\infty f(a 1_n + cu_0) \ da \ dc \right)
\]

and this is independent of the model parameter and we have a valid \( P \)-value for checking the model. If instead we use a function \( T(U) \) then, an application of Lemma 2.5 shows that (2.11) is independent of \((\mu, \sigma)\) by the same argument, as the Jacobian factors do not depend on the parameter. Note that when \( f \) is the \( N(0,1) \) density, then basing model checking on the ancillary \( d \) or the conditional distribution of the data given a minimal sufficient statistic produces the same results.

More generally suppose we have a group model \( \{f_g : g \in G\} \) where \( G \) is a group, with a smooth product, acting freely and smoothly on \( \mathcal{X} \) and \( f_g(x) = f(g^{-1}x)J_g(g^{-1}x) \) for
some fixed density \( f \). Now suppose that \([\cdot] : \mathcal{X} \to G\) is smooth and satisfies \([gx] = g[x]\) so \( U(x) = [x]^{-1}x\) is a maximal invariant and is thus ancillary. So \( u = U(x) \in \mathcal{X}, x = [x]U(x) \) and \( U^{-1}(u) \) is the orbit \( \{gu : g \in G\} \). Now if \( \nu^*_G \) denotes geometric measure on \( G \) we have that \( \nu_{U^{-1}(u)} = K(u)\nu^*_G \) for some positive function \( K \). Let \( z = g^{-1}x \) so that \([z] = g^{-1}[x]\) and let \( J^*_g([z]) \) denote the Jacobian of the transformation \([z] \to [x]\). Then we have that \( f^*_g(U(u)) = \int_{U^{-1}(u)} f_g(x) \nu_{U^{-1}(u)}(dx) = \int_{\{gu : g \in G\}} f_g([x]u) \nu_{U^{-1}(u)}(dx) = K(u) \int_G f(g^{-1}[x]u)J_g(g^{-1}[x]u) \nu^*_G(d[x]) = K(u) \int_G f([z]u)J_g(u)J^*_g([z]) \nu^*_G(d[z]) \). Now, if we can write \( J_g(u)J^*_g(z) = L(u)m(g) \) for some positive functions \( L \) and \( m \), then we have that the invariant \( P\)-value \( P_U(f^*_g(U(u)) \leq f^*_g(U(u_0)) \) is indeed independent of \( g \). Further, by Lemma 2.5 this will also hold for \( T \circ U \) as well. For example, in Example 2.7 \( J_{(\mu,\sigma)}(u) = \sigma^{-n} \) and, with \([x] = [\bar{x},||x - \bar{x}1_n||]\), then \( J^*_g([z]) = \sigma^2 \) and this condition is satisfied. More generally, this condition is satisfied in a wide range of group models, such as those discussed in Fraser (1979). Accordingly the following result is broadly applicable.

**Theorem 2.7.** Suppose that \( \{f_g : g \in G\}\) is a family of densities with respect to geometric measure \( \mu,\sigma \) on \( \mathcal{X} \), where \( G \) is a group with a smooth product, a smooth action defined on \( \mathcal{X} \), and \( f_g(x) = f(g^{-1}x)J_g(g^{-1}x) \). Further suppose that there exists smooth \([\cdot] : \mathcal{X} \to G\) satisfying \([gx] = g[x]\) and let \( J^*_g([z]) \) denote the Jacobian of the transformation \([z] \to [x]\) where \( x = gz \). If there exist positive functions \( L \) and \( m \) so that \( J_g(u)J^*_g(z) = L(u)m(g) \), then we have that the \( P\)-value (2.11) based on the ancillary \( T \circ U \), with \( U(x) = [x]^{-1}x \) and \( T \) smooth, is independent of the model parameter and is thus a valid check on the model.

Although we have not been able to find an example, there may exist cases where the invariant \( P\)-value based on an ancillary \( T \circ U \) is not independent of the parameter. In such a case it is perhaps difficult to accept \( T \circ U \) as a true ancillary, because its ancillarity is dependent on the way the transformation is distorting volumes in some essential way.
2.5.2 Checking for Prior-Data Conflict

In Evans and Moshonov (2006, 2007), a methodology was developed for investigating the existence of a conflict between the prior probability assignments made for the model parameter $\theta$, and the values of $\theta$ deemed relevant by the likelihood. If $T$ is a minimal sufficient statistic and $U(T)$ is a maximal ancillary, then the assessment is made based upon comparing the observed value $t_0 = T(x_0)$ with $M_T(\cdot \mid u_0)$, the conditional prior predictive distribution of $T$ given $U(t_0) = u_0$. The comparison was based upon the $P$-value $M_T(m_{T\mid t}(t_0) \leq m_{T\mid t}(m_T(t)|u_0))$ where $m_{T\mid t}$ is the prior predictive density of $T$ given $U(T) = u_0$, based on either counting measure or volume measure depending on whether $M_T(\cdot \mid u_0)$ was discrete or continuous. This choice of $P$-value was made primarily because there was no theory that dictated how such an assessment was to be made, but concern was expressed about the lack of invariance in the continuous case. We can now use the approach developed here to derive an appropriate invariant $P$-value.

The prior predictive density of $x$ is given by $m(x) = \int_\Theta f_\theta(x) \Pi(d\theta)$ and note that this is just an average of the density values $f_\theta(x)$ with respect to the prior. There is no volume distortion involved in this, for if $f_\theta(x)\mu_k(B_n(x))$ is the probability of observing the discretized response $x_n(x)$ when $\theta$ is true, then $m(x)\mu_k(B_n(x))$ is the probability of observing $x_n(x)$ when $\theta \sim \Pi$ and $x \sim P_\theta$. Furthermore,

$$m^*_T(t) = \int_{T^{-1}(t)} m(x) \nu_{T^{-1}(t)}(dx) = \int_{T^{-1}(t)} \int_\Theta f_\theta(x) \Pi(d\theta) \nu_{T^{-1}(t)}(dx)$$

$$= \int_\Theta \int_{T^{-1}(t)} f_\theta(x) \nu_{T^{-1}(t)}(dx) \Pi(d\theta) = \int_\Theta f^*_T(x) \Pi(d\theta)$$

and so $m^*_T$ is obtained by averaging the densities appropriate to checking each $P_\theta$ measure individually based on any statistic $T$.

If $T$ is a complete minimal sufficient statistic, then we can ignore ancillaries and the relevant $P$-value is $M_T(m^*_T(t) \leq m^*_T(t_0))$. The following result shows that we need a slight modification for the general situation.
Chapter 2. Invariant $P$-value for Model Checking

**Theorem 2.8.** Suppose that $\{f_\theta : \theta \in \Theta\}$ is a family of densities with respect to geometric measure $\mu_X$ on $X$, $\Pi$ is a prior probability measure on $\Theta$, $T : X \rightarrow T$ is a smooth mapping onto manifold $T$ with geometric measure $\mu_T$ and $U : T \rightarrow U$ is a smooth mapping onto manifold $U$ with geometric measure $\mu_U$. Further suppose that $T$ is minimal sufficient and $U \circ T$ is ancillary. Then the invariant $P$-value based on the conditional prior predictive distribution of $T$ given $U$ is

$$M_T(m_T^*(t) \leq m_T^*(t_0) | U = u_0)$$

where

$$m_T^*(t) = \int_\Theta f^*_\theta(x) \Pi(d\theta).$$

*Proof of Theorem 2.8.* For $t \in U^{-1}(u_0)$ we have that $m_T(t | u_0) = \int_\Theta f_{\theta T}(t | u_0) \Pi(d\theta)$.

Note that $T^{-1}(t) \cap T^{-1}(U^{-1}(u_0)) = T^{-1}(t)$ when $t \in U^{-1}(u_0)$ and is the empty set otherwise. If $t \in U^{-1}(u_0)$, then $f_{\theta T}(t | u_0) = f_{\theta T}(t) J_U(t) / f_U(u_0) = (J_U(t) / f_U(u_0)) \int_{T^{-1}(t)} f_{\theta}(x) J_T(x) \nu_{T^{-1}(t)}(dx)$. Therefore, removing the volume distortions due to $T$ and $U$, we have that

$$m_T^*(t | u_0) = \int_\Theta \int_{T^{-1}(t)} (f_{\theta}(x) / f_U(u_0)) \nu_{T^{-1}(t)}(dx) \Pi(d\theta) = m_T^*(t) / f_U(u_0)$$

and the result follows. 

So (2.14) is obtained by averaging, with respect to the prior, the relevant functions for checking each $P_\theta$ measure and then, comparing the observed value of this function with its distribution under the prior predictive given the ancillary $U(T)$. As argued in Evans and Moshonov (2006, 2007), conditioning on the ancillary is appropriate when assessing prior-data conflict, as this removes variation from the assessment that has nothing to do with the prior.

For the examples included in Evans and Moshonov (2006, 2007) the only $P$-values that will change, when we use these invariant $P$-values, are those recorded for the location-scale models. In all the other examples the volume distortions are constant, either because of discreteness or because the model was a location model. The change in the $P$-values
for location-scale models is illustrated by the following example, and we see that this is very small.

**Example 2.11** (Prior-data conflict for the location-scale normal model). For a sample \( x \) of size \( n \) from the \( N(\mu, \sigma^2) \) model, \( T(x) = (\bar{x}, (n-1)^{-1}||x - \bar{x}1_n||^2) = (\bar{x}, s^2) \) is minimal sufficient. Then for prior \( \pi \), the prior predictive density is

\[
m_T(\bar{x}, s^2) = \int_0^\infty \int_{-\infty}^{\infty} \int_{T^{-1}(\bar{x}, s^2)} f(x | \mu, \sigma^2) J_T(x) \nu_{T^{-1}(\bar{x}, s^2)}(dx) \pi(\mu, \sigma^2) d\mu d\sigma^2
\]

where \( f(\cdot | \mu, \sigma^2) \) is the joint density of the sample, \( \nu_{T^{-1}(\bar{x}, s^2)} \) is surface area measure on the \((n-2)\)-dimensional sphere \( \bar{x}1_n + ||x - \bar{x}1_n||(S^{n-1} \cap L_1(1_n)) \), and \( J_T(x) = n^{1/2} (n-1)^{1/2} / (2s) \). So \( m_T^*(\bar{x}, s^2) = 2s m_T(\bar{x}, s^2) / n^{1/2} (n-1)^{1/2} \) and the \( P \)-values based on \( m_T \) and \( m_T^* \) differ by very little. In fact this difference disappears as \( n \) grows. The arbitrariness of the \( P \)-value based on the density is demonstrated by the fact that, if we had instead chosen the minimal sufficient statistic to be \( T(x) = (n\bar{x}, ||x - \bar{x}1_n||) \), then the \( P \)-value based on \( m_T \) equals the invariant \( P \)-value.

It is not clear that the general use of discrepancy statistics is appropriate when checking for prior-data conflict. For there is no sense in which we think of the prior as being wrong, as it represents some individual’s (or individuals’) beliefs about what the true distribution is. Rather we simply look to see if there is a conflict between what the data says about \( \theta \), as expressed by the likelihood function, and the prior. In Evans and Moshonov (2006, 2007) this was assessed by determining whether or not the observed likelihood function is a surprising value from its prior predictive distribution given an ancillary or, equivalently, whether or not the observed value of a minimal sufficient statistic is a surprising value from its prior predictive distribution given an ancillary. With the use of the invariant \( P \)-values developed here, this assessment becomes independent of the choice of a particular form chosen for the minimal sufficient statistic.

The assessment for prior-data conflict becomes more involved when a prior is specified hierarchically. For example, suppose the prior is specified component-wise as \( \pi_2(\theta_2 | \theta_1) \pi_1(\theta_1) \), where the model parameter equals \((\theta_1, \theta_2)\), or where \( \theta_2 \) is the model parameter and
\( \theta_1 \) corresponds to hyperparameters. In such a case, choices are made for the \( \pi_i \) and, as discussed in Evans and Moshonov (2006, 2007), we wonder if perhaps only part of this specification leads to a prior-data conflict. Of course, the \( P \)-values developed there for such cases should also be modified to use invariant \( P \)-values.

We have restricted our discussion to determining appropriate \( P \)-values for checks on the sampling model, based on the conditional distribution given a minimal sufficient statistic or based on an ancillary statistic, and separate checks for prior-data conflict, based on the conditional prior predictive of a minimal sufficient statistic given an ancillary. Other authors, such as Bayarri and Berger (2000), Bayarri and Castellanos (2007), Box (1980), Gelman, Meng and Stern (1996), and Meng (1994) have recommended \( P \)-values for Bayesian model checking that combine elements of the prior and the model. We feel that our developments are also relevant to the checks recommended by these authors.

2.6 Volume Distortions and Computations

Some topics concerning \( f^* \) are discussed in this section.

2.6.1 Derivations of \( J_T \)

The value \( J_T(x) \) denotes the volume distortion rate of \( T \) at \( x \). The evaluation of \( J_T(x) \) is not difficult usually. For example, suppose the range of \( T = (T_1, \ldots, T_k) \) contains an open set of \( \mathbb{R}^k \). Then, for \( x = (x_1, \ldots, x_n) \), \( dT(x) = (\partial T_i(x)/\partial x_j)_{i,j} \) and

\[
J_T(x) = |\det(dT(x) \circ dT(x)')|^{-1/2} = \left| \det \left( \begin{pmatrix} \partial T_i(x) \\ \partial x_j \end{pmatrix} \right) \right|^{-1/2}.
\]

**Example 2.12** \((J_{T_{\text{prod}}} \text{ and } J_{JB} \text{ in Example 2.9})\). Let \( r^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = ||x - \bar{x}1_n||^2 \), \( d_i = (x_i - \bar{x})/r \), \( U_p(x) = (T_p \circ d)(x) = \sum_{i=1}^{n} d_i^k \). We need to specify differentials of \( U_p \).

\[
\frac{\partial U_p}{\partial x_i} = \frac{p}{r} (d_i^{p-1} - \frac{1}{n} U_{p-1} - d_i U_p).
\]
Hence we have
\[
J_{T,\rho}(\mathbf{x}) = J_{U_p}(\mathbf{x}) = \left( \det \left( \begin{pmatrix} \frac{\partial U_p}{\partial x_1} \\ \vdots \\ \frac{\partial U_p}{\partial x_i} \end{pmatrix} \right) \right)^{-1/2} = \left( \frac{p^2}{r^2} \sum_{i=1}^{n} \left( d_i^{p-1} - \frac{U_{p-1}}{n} - d_i U_p \right)^2 \right)^{-1/2}
\]
\[
= \left( r/p \right) \left( U_{2p-2} - \frac{U_{p-1}}{n} - U_p^2 \right)^{-1/2}.
\]

Also for \( JB = \frac{n^2}{6} U_3^2 + \frac{n}{24} (n U_4 - 3)^2 \) we have
\[
\frac{\partial JB}{\partial x_i} = \frac{n^2}{3} \frac{\partial U_3}{\partial x_i} + \frac{n^2}{12} (n U_4 - 3) \frac{\partial U_4}{\partial x_i}
\]
\[
= \frac{n^2}{r} \left[ (n U_4/3 - 1) d_i^3 + U_3 d_i^2 - (T_3^2 + n U_4^2/3 - U_4) d_i - U_3 U_4/3 \right]
\]
and similar computation gives
\[
J_{JB}(\mathbf{x})^{-2} = \frac{n^4}{r^2} \left[ \left( \frac{n}{3} T_4 - 1 \right)^2 T_6 + 2 \left( \frac{n}{3} T_4 - 1 \right) T_3 T_5 - \left( T_3^2 + \frac{n}{3} T_4^2 - T_4 \right)^2 + T_3^2 T_4 - \frac{n}{9} T_3^2 T_4^2 \right].
\]

Computation of \( J_T(x) \) is not easy if the range space of \( T = (T_1, \ldots, T_k) \) is not \( k \)-dimensional. For example, in Example 2.9, \( d \) itself can be considered as \( T \). Then, the range space of \( d = x - \bar{x} 1_n \) is contained in \( \mathbb{R}^n \) but it is \( (n - 2) \)-dimensional space. In this case, the computation of \( J_T(x) \) requires some aspects of the theory of Riemann manifolds. Related topics can be found in Sections 13 and 15 of Tjur (1974).

### 2.6.2 Disappearance of Volume Distortion

The integration notation of \( f^* \) in (2.10) is not easy to access. So we decompose \( f^* \) into two parts in the following lemma.

**Lemma 2.9.** Suppose that \( \mathcal{X} \) and \( \mathcal{T} \) are manifolds, with geometric measures \( \mu_{\mathcal{X}} \) and \( \mu_{\mathcal{T}} \) respectively, and \( T : \mathcal{X} \to \mathcal{T} \) is an onto, smooth mapping. If \( \mathbf{E}(J_T^{-1}(X)) < \infty \), then
\[
f_T^*(t) = f_T(t) \mathbf{E}(J_T^{-1}(X) \mid T = t).
\]

**Proof of Lemma 2.9.** For \( B \subset \mathcal{T} \), by the measure decomposition theorem, we have that
\[
\mathbf{E}(J_T^{-1}(X)I_B(T(X))) = \int_{\mathcal{X}} J_T^{-1}(x)I_B(T(x)) f(x) \mu_{\mathcal{X}}(dx)
\]
\[
= \int_B \int_{T^{-1}(t)} f(x) \nu_{T^{-1}(t)}(dx) \mu_T(dt) = \int_B f_T^*(t) \mu_T(dt)
\]
and of course \( \mathbf{E}(J_T^{-1}(X)I_B(T(X))) = \int_B \mathbf{E}(J_T^{-1}(X) \mid T = t) f_T(t) \mu_T(dt) \).

Hence, \( f_T^* \) can be decomposed into the density \( f_T \) of \( T \) times the average correction for volume distortion given by \( \mathbf{E}(J_T(X)^{-1} \mid T = t) \).

Volume distortion affects a \( P \)-value when the sample size is small. The effect of volume distortion disappears as sample size grows in many cases.

**Theorem 2.10.** Let \( X_1, X_2, \ldots, X_n \sim i.i.d. P_0 \). Consider a statistic \( T = n^{-1} \sum_{i=1}^{n} g(X_i) \) where \( g \) is a measurable function having the derivative \( \dot{g} \). If (i) \( \mu_1 = \mathbf{E}_{P_0} \|g(X)\| < \infty \) and (ii) \( \mu_2 = \mathbf{E}_{P_0} \|\dot{g}(X)\|^2 < \infty \), then, for a fixed \( t_* \), the ratio of the average volume distortion rates \( \mathbf{E}(J_T(X)^{-1} \mid T = t)/\mathbf{E}(J_T(X)^{-1} \mid T = t_*) \) converges to a constant as the sample size increases. So the effect of volume distortion factor disappears as the sample size increases.

**Proof of Theorem 2.10.** The proof for the general case \( X_i \in \mathbb{R}^p \) and \( g: \mathbb{R}^p \to \mathbb{R}^q \) is very similar to that of \( p = 1 \) and \( q = 1 \). So we prove this theorem for \( p = q = 1 \). The volume distortion factor is computed using the differentiability of \( g \) by

\[
\frac{\partial T}{\partial x_i}(x) = \frac{1}{n} \dot{g}(x_i) \quad \text{and} \quad |\det(dT(x) \circ dT(x'))| = \frac{1}{n^2} \sum_{i=1}^{n} (\dot{g}(x_i))^2.
\]

Hence, \( J_T(x) = n \left( \sum_{i=1}^{n} (\dot{g}(x_i))^2 \right)^{-1/2} \). It is easy to check that, for fixed \( n \),

\[
\mathbf{E}[J_T(X)^{-1}] = \frac{1}{n} \mathbf{E} \left[ \left( \sum_{i=1}^{n} (\dot{g}(x_i))^2 \right)^{1/2} \right] \leq \frac{1}{n} \mathbf{E} \left[ \sum_{i=1}^{n} (\dot{g}(x_i))^2 \right]^{1/2} = \frac{1}{n} (n \mu_2)^{1/2} < \infty
\]

where Jensen’s inequality is used in the first inequality. Thus, \( f_T^* \) exists and \( f_T^*(t) = f_T(t) \mathbf{E}(J_T(X)^{-1} \mid T = t) \) by Lemma 2.9. Let \( U_n = T - \mu_1 = n^{-1} \sum_{i=1}^{n} (g(X_i) - \mu_1) \) and \( V_n = n^{-1} \sum_{i=1}^{n} (\dot{g}(X_i)^2 - \mu_2) \). Then, by the strong law of large numbers \( U_n = o_p(1) \) and \( V_n = o_p(1) \). Also, \( T = \mu_1 + U_n \) and \( J_T(x) = n^{1/2}(\mu_2 + V_n)^{-1/2} \).

To show the disappearance of the volume distortion, we show that \( n^{1/2} \mathbf{E}(J_T(X)^{-1} \mid T) \) converges to a constant. Note that \( n^{1/2} J_T(x)^{-1} = (\mu_2 + V_n)^{1/2} \leq (1 + \mu_2 + |V_n|)/2 \) by Lemma A.7. Since \( \mathbf{E}(|V_n|) \to 0 \) by Lemma A.8, we have \( \mathbf{E}(|V_n| \mid T) \to 0 \) as well.
as $E((1 + \mu_2 + |V_n|)/2 | T) \to (1 + \mu_2)/2$. Then, the generalized Lebesgue dominated convergence theorem implies $n^{1/2} E(J_T(X)^{-1} | T) \to \sqrt{\mu_2}$. This proves the theorem.

The conditions in Theorem 2.10 can be weakened, as in the following example.

**Example 2.13** (Example 2.12 continued). Under the model assumption $X_i \sim N(0, 1)$, both $N_3 = nU_3/\sqrt{6}$ and $N_4 = \sqrt{n/24(nU_4 - 3)}$ converge to $N(0, 1)$ in distribution. Also $N_3$ and $N_4$ are asymptotically independent. Then, for $T = U_3$,

$$n E(J_T(X)^{-1} | T) = n E[(3/r)(U_4 - 1/n - U_3^2)^{1/2} | T)$$

$$= n E(3/r) E((3/n + \sqrt{24n^{-3/2}N_4 - 1/n - 6n^{-2}N_3^2})^{1/2} | N_3)$$

$$= \frac{3\Gamma((n - 2)/2)}{\Gamma((n - 1)/2)\sqrt{2/n}} E((2 + \sqrt{24}n^{-1/2}N_4 - 6n^{-3/2}N_3) | N_3) \to \sqrt{12}.$$  

Hence, the volume distortion rate of $T = U_3$ disappears.

### 2.6.3 Computations

Implementation of invariant $P$-values will sometimes require the numerical evaluation of $f^*_T$. From Lemma 2.9, $f^*_T(t) = f_T(t) E(J_T(X)^{-1} | T = t)$. Hence, numerical evaluation can be separated into two parts, i.e., the density estimation of $f_T(t)$ and nonparametric regression of $E(J_T(X)^{-1} | T = t)$.

We present an estimation method for $E(J_T(X)^{-1} | T)$ using importance sampling. For the computation we assume that $T$ is real-valued and both $f_T(t)$ and $E(J_T(X)^{-1} | T = t)$ are continuous in $t$. Note that $E(J_T(X)^{-1}) = E[E(J_T(X)^{-1} | T)] = \int E(J_T(X)^{-1} | T = t) f_T(t) \mu(dt)$. On $(t - h, t + h)$ for small $h > 0$, we have $E(J_T(X)^{-1}I_{(t-h,t+h)}(T(X))) = \int E(J_T(X)^{-1} | T = u)I_{(t-h,t+h)}(u) f_T(u) \mu(du) \approx 2hf_T(t) E(J_T(X)^{-1} | T = t)$. Thus we get an estimator of $E(J_T(X)^{-1} | T = t)$ given by

$$\hat{E}(J_T(X)^{-1} | T = t) = \frac{\hat{E}(J_T(X)^{-1}I_{(t-h,t+h)}(T(X)))}{2hf_T(t)}.$$  

(2.15)
Both denominator and numerator of (2.15) can be approximated by simple importance sampling. Suppose $X_1, \ldots, X_n$ is a random sample from the density $f$. Then,

$$
E(J_T(X)^{-1}I_{(t-h,t+h)}(T(X))) \approx \frac{1}{n} \sum_{i=1}^{n} J_T(X_i)^{-1}I_{(t-h,t+h)}(T(X_i)),
$$

$$
2hf_T(t) \approx E(I_{(t-h,t+h)}(T(X))) \approx \frac{1}{n} \sum_{i=1}^{n} I_{(t-h,t+h)}(T(X_i)).
$$

For simplicity let $K_h(t) = I_{(-1,1)}(t)/(2h)$. Then, we have

$$
\hat{f}_T(t) = \frac{1}{n} \sum_{i=1}^{n} K_h(T(x_i) - t) \quad \text{and} \quad \hat{E}(J_T^{-1}(X) \mid T = t) = \frac{\sum_{i=1}^{n} J_T^{-1}(x_i)K_h(T(x_i) - t)}{\sum_{i=1}^{n} K_h(T(x_i) - t)}.
$$

The first part is a kernel density estimator of $f_T$ and the second part is the Nadaraya-Watson estimator of $E(J_T(X)^{-1} \mid T = t)$. By changing kernel $K_h$, better estimators can be obtained. The graphs shown in Example 2.9 were obtained from local polynomial regression that performs better than the Nadaraya-Watson regression estimator. Further details on this estimator and kernel regression can be found in Wand and Jones (1995).

### 2.7 Conclusions

The use of $P$-values is somewhat of a controversial topic in statistics. In many ways, it seems to us, however, that an appropriately defined $P$-value represents the best way of assessing whether or not an observed value $x_0$ from a distribution $P$ is surprising. Perhaps the first concern about $P$-values arises here, as it is not clear exactly how such a $P$-value should be computed. Some may argue that we must make use of a real-valued discrepancy statistic $T(x)$ and compute $P_T(T \geq T(x_0))$. While there is intuitive support for this, it only seems justified when the region of relatively low probability for $T$ is just the right-tail. Further, it doesn’t really help at all when $T$ is multivariate.

We have argued that there is a logical basis for the development of appropriate $P$-values for discrete models. Further, we can carry this development over to suitably regular continuous models provided that we acknowledge that our continuous models
are approximations to a discrete reality and, that we make sure that volume distortions induced by transformations do not affect our $P$-values. This leads to results that are the same or very similar, in many examples, to the way we currently compute $P$-values based simply on intuition. This is satisfying, as a radical change in such a fundamental tool would make us wary. Perhaps the most important consequence is that we feel that we have resolved the issue of the noninvariance of the $P$-value in the continuous case, and this makes us more confident that these are appropriate measures of surprise for the problems discussed.

There are many other contexts where $P$-values are computed that do not correspond to the situation where we have an observation $x_0$ ostensibly from a single known distribution $P$, as discussed in this paper. For example, suppose we have a statistical model $\{P_\theta : \theta \in \Theta\}$ and a proper subset $H_0 \subset \Theta$ and want to assess whether or not $x_0$ could have come from a $P_\theta$ with $\theta \in H_0$. We could consider looking for a minimal sufficient statistic or ancillaries for the model $\{P_\theta : \theta \in H_0\}$ and then proceed as we have described here. But such an approach ignores the fact that, given that we have accepted the model $\{P_\theta : \theta \in \Theta\}$ as correct, then $H_0$ being false implies that $H_c^0$ is true. It seems that any method of assessing $H_0$ should take this into account and methods based on the conditional distribution given a minimal sufficient statistic or ancillaries for $\{P_\theta : \theta \in H_0\}$ would not. In such problems, likelihood ratio methods are the methods of choice for many, and we note that the associated $P$-values are invariant under smooth, one-to-one transformations of the data. These methods are not available, however, for the problems discussed in this paper.

An alternative to likelihood inferences, when we want to assess $\{P_\theta : \theta \in H_0\}$, is to use a Bayesian approach. The situation then becomes somewhat complicated. For example, when we have a prior $\Pi$ on $\Theta$, $H_0 = \psi^{-1}\psi_0$ for $\psi_0 \in \Psi$ and $\psi : \Theta \to \Psi$, then a possibility is to compute the Bayesian $P$-value $\Pi(\pi_\psi(\psi(\theta) | x) \leq \pi_\psi(\psi_0 | x) | x)$ where $\pi_\psi(\cdot | x)$ is the posterior density of $\psi$. Such $P$-values arise naturally from constructing highest posterior
density regions for $\psi$. In the continuous case, however, highest posterior density regions and the associated $P$-values, will not be invariant under 1-1 transformations of $\psi$. Clearly, a possible fix for this is to follow the recommendation in this paper and adjust for volume distortions to obtain invariant $P$-values and regions. We note, however, a subtle distinction between this situation and that discussed in this paper. Our argument was based on the fact that the response $x$ is essentially discrete and this discreteness, and its fineness, is determined by the measurement process employed in the application. For the Bayesian application we do not directly observe $\theta$ and no discretization seems necessarily implicit. So an alternative argument is needed to justify such an adjustment of the Bayesian $P$-value to obtain invariance.

Perhaps the most common Bayesian approach to assessing $H_0$, is to simply compute $\Pi(H_0 \mid x_0)$ and conclude we have evidence against $H_0$ when this is small. There is a problem with this approach, however, that arises most obviously when $\Pi$ is continuous and $H_0$ is a lower dimensional subset of $\Theta$. In this case $\Pi(H_0) = 0$, which implies that $\Pi(H_0 \mid x_0) = 0$, and so we would always find evidence against $H_0$ no matter what data is observed. More generally we can expect a similar problem whenever $\Pi$ places very little probability on $H_0$, as we may need a lot of data to get a reasonable value for $\Pi(H_0 \mid x_0)$ when $H_0$ is true. A common solution to this problem is to insist at least that $\Pi(H_0) > 0$. This usually involves using a discrete mixture of priors on $H_0$ and $H_0^c$. When this is done, however, a sharp discrepancy has been noted, see Berger and Delampady (1987) and Berger and Sellke (1987), between the evidence presented by classical, frequentist $P$-values and $\Pi(H_0 \mid x_0)$. This has lead many to have considerable doubts concerning the value of $P$-values.

In Evans (1997) an alternative approach was developed to hypothesis assessment and region construction when there is a prior, based upon the relative surprise principle. These inferences are similar to highest posterior density inferences but the $P$-value equals $\Pi(\pi_\psi(\psi(\theta) \mid x_0) / \pi_\psi(\psi(\theta))) \leq \pi_\psi(\psi_0 \mid x_0) / \pi_\psi(\psi_0) \mid x)$ where $\pi_\psi$ is the prior density of $\psi$, i.e.,
inferences are based on how the data change beliefs from \textit{a priori} to \textit{a posteriori}. It is immediate that these inferences are invariant under smooth, 1-1 transformations on $\psi$. In Evans, Guttman and Swartz (2006) and Evans and Shakhatreh (2008) a number of optimal properties of these inferences are developed that show a strong similarity with the likelihood-based theory of hypothesis testing, e.g., a result much like the Neyman-Pearson theorem plays a key role. Furthermore, as with the Bayesian $P$-value, there is no need to put a discrete mass on $H_0$ and sequences of diffuse continuous priors commonly lead to limiting $P$-values that correspond to classical $P$-values. In essence, the relative surprise approach lends support to these classical $P$-values. Also, if we do use a prior that puts a discrete mass on $H_0$, then the relative surprise approach leads to a $P$-value equal to $\Pi(H_0 \mid x_0)$, when the Bayes factor in favour of $H_0$ is less than 1, and is equal to 1 otherwise, i.e., we get a minor modification of the usual Bayesian answer. One could argue then, that the debate about the relevance of the $P$-value is rather a debate about which prior we should use in such a problem and, in particular, whether or not a discrete mass on $H_0$ is essential. It is our view that it is not and the relative surprise approach is our recommendation for Bayesian hypothesis assessment problems in general. We acknowledge, however, that there are a considerable variety of opinions on this issue.

2.8 Proofs

Proof of Theorem 2.1

(i) We need to consider the distribution of

$$\log p_T(t) = \log \left( \frac{n}{t_1 \ldots t_k} \right) \theta_{t_1}^{t_1} \ldots \theta_{t_k}^{t_k} = \sum_{i=1}^{k} (t_i \log \theta_i - \log t_i!) + \log n!$$

when $(t_1, \ldots, t_k) \sim \text{Multinomial}(n, \theta_1, \ldots, \theta_k)$. For $M > 0$, let

$$C_M = \{ (t_1, \ldots, t_k) : \max_{i=1,\ldots,k} |t_i - n\theta_i|/n^{1/2} \leq M \text{ and no } t_i = 0 \}.$$
Since \((z_1, \ldots, z_k) = (t_1 - n\theta_1, \ldots, t_k - n\theta_k)/n^{1/2} \xrightarrow{d} N_k(0, \Sigma)\) as \(n \to \infty\), where \(\sigma_{ij} = \theta_i(1 - \theta_i)\) and \(\sigma_{ij} = -\theta_i\theta_j\) when \(i \neq j\), and since no \(\theta_i = 0\), there exist \(M_e\) and \(n_{M_e}\), such that for all \(n \geq n_{M_e}\) then \(P_T(C_{M_e}) > 1 - \epsilon\).

Suppose that \((t_1, \ldots, t_k) \in C_{M_e}\). By Stirling’s formula \(\log t_i! = (1/2) \ln 2\pi t_i + t_i \log t_i - t_i + \lambda(t_i)\) where \(|\lambda(t_i)| < 1/12t_i\) and, using \(\sum_{i=1}^{k} t_i = n\),

\[
\sum_{i=1}^{k} (t_i \log \theta_i - \log t_i!)
\]

\[
= -\sum_{i=1}^{k} t_i \log \frac{t_i}{n\theta_i} - \frac{1}{2} \sum_{i=1}^{k} \log \frac{t_i}{n} + \sum_{i=1}^{k} \lambda(t_i) - n \log n + n - \frac{k}{2} \ln 2\pi n. \tag{2.16}
\]

Now \(t_i \log (t_i/n\theta_i) = (n\theta_i + \sqrt{n}z_i) \log (1 + z_i/\sqrt{n}\theta_i)\) and since \(|z_i| \leq M_e\), we can choose \(n\) larger than \(n_{M_e}\) so that \(|z_i|/\sqrt{n}\theta_i \leq 1/2\) for all \(i\). When \(|z| \leq 1/2\),

\[
|\log(1 + z) - z + z^2/2| = \left| \sum_{i=1}^{\infty} (-1)^{i+1} \frac{z^i}{i} - z + z^2/2 \right| \leq \sum_{i=3}^{\infty} \frac{|z|^i}{i} \leq \frac{2}{3} |z|^3,
\]

so \(\log(1 + z) = z - z^2/2 + O(|z|^3)\). Applying this to the first term in (2.16), gives

\[
\sum_{i=1}^{k} t_i \log \frac{t_i}{n\theta_i} = \sum_{i=1}^{k} (n\theta_i + \sqrt{n}z_i) \log (1 + z_i/\sqrt{n}\theta_i)
\]

\[
= \sum_{i=1}^{k} \left( n\theta_i + \sqrt{n}z_i \right) \left( \frac{z_i}{\sqrt{n}\theta_i} - \frac{z_i^2}{2n\theta_i^2} + O\left( \frac{|z_i|^3}{n^{3/2}} \right) \right)
\]

\[
= \sum_{i=1}^{k} \frac{z_i^2}{2\theta_i} + \sum_{i=1}^{k} O\left( \frac{|z_i|^3}{n^{1/2}} \right) + \sum_{i=1}^{k} O\left( \frac{|z_i|^4}{n} \right).
\]

When \(|z| \leq 1/2\), then \(|\log(1 + z)| \leq \sum_{i=1}^{\infty} |z|^i / i = |z| \sum_{i=0}^{\infty} |z|^i / (i + 1) \leq 2 |z|\) and, applying this to the second term in (2.16), gives

\[
\frac{1}{2} \sum_{i=1}^{k} \log \frac{t_i}{n} = \frac{1}{2} \sum_{i=1}^{k} \log \theta_i + \frac{1}{2} \sum_{i=1}^{k} \log \left( 1 + \frac{z_i}{\sqrt{n}\theta_i} \right) = \frac{1}{2} \sum_{i=1}^{k} \log \theta_i + \sum_{i=1}^{k} O\left( \frac{|z_i|}{\sqrt{n}} \right).
\]

The third term in (2.16) satisfies

\[
\left| \sum_{i=1}^{k} \lambda(t_i) \right| \leq \sum_{i=1}^{k} \frac{1}{12t_i} = \frac{1}{12} \sum_{i=1}^{k} \frac{1}{n\theta_i + \sqrt{n}z_i} = O\left( \frac{1}{n} \right).
\]
Combining all this and $|\log n! - n \log n + n - (1/2) \ln 2\pi n| \leq 1/12n$ gives

$$- \log p_T(t) = \sum_{i=1}^{k} \left( \frac{(t_i - n\theta_i)^2}{2n\theta_i} \right) + \frac{1}{2} \sum_{i=1}^{k} \log \theta_i + \frac{k - 1}{2} \ln 2\pi n + r_n(t)$$

where

$$r_n(t) = \sum_{i=1}^{k} O \left( \frac{|z_i|^3}{n^{1/2}} \right) + \sum_{i=1}^{k} O \left( \frac{|z_i|^4}{n} \right) + O \left( \frac{1}{n} \right) = O \left( \frac{1}{n^{1/2}} \right).$$

when $(t_1, \ldots, t_k) \in C_M$, since $|z_i| \leq M$. Then, for $\eta > 0$, we have that

$$P_T(|r_n(t)| > \eta) \leq P_T(|r_n(t)| > \eta | C_M) P_T(C_M) + P_T(C^c_M) \leq \epsilon$$

since $P_T(|r_n(t)| > \eta | C_M) = 0$ for all $n$ large enough and so $r_n(t) \xrightarrow{P} 0$. Since $\sum_{i=1}^{k} (t_i - n\theta_i)^2 / n\theta_i \xrightarrow{d} X$ where $X \sim \chi^2(k-1)$ we have proved (i).

(ii) We can write

$$P_T(\log p_T(t) \leq \log p_T(t_0)) = P_T \left( \sum_{i=1}^{k} \left( \frac{(t_i - n\theta_i)^2}{2n\theta_i} + r_n(t) \right) \geq \sum_{i=1}^{k} \frac{(t_0 - n\theta_i)^2}{n\theta_i} + r_n(t_0) \right).$$

Then, letting $G_{k-1}$ denote the $\chi^2(k-1)$ distribution function,

$$P_{T_0} \left( P_T(\log p_T(t) \leq \log p_T(t_0)) - \left\{ 1 - G_{k-1} \left( \sum_{i=1}^{k} \frac{(t_0 - n\theta_i)^2}{n\theta_i} \right) \right\} > \eta \right)$$

$$\leq P_{T_0} \left( P_T \left( \left\{ \sum_{i=1}^{k} \left( \frac{(t_i - n\theta_i)^2}{2n\theta_i} + r_n(t) \right) \geq \sum_{i=1}^{k} \frac{(t_0 - n\theta_i)^2}{n\theta_i} + r_n(t_0) \right\} \right) > \eta / 2 \right) +$$

$$P_{T_0} \left( \left\{ 1 - G_{k-1} \left( \sum_{i=1}^{k} \frac{(t_0 - n\theta_i)^2}{n\theta_i} + r_n(t_0) \right) \right\} > \eta / 2 \right) \quad (2.17)$$

Now let $\epsilon > 0$ satisfy $\epsilon < \eta / 2$. When $X$ has a continuous distribution and $X_n \xrightarrow{d} X$ then, for any $\epsilon > 0$, we have that $\sup_x |P(X \leq x) - P(X_n \leq x)| \leq \epsilon$ for all $n$ large enough.

So, since $r_n(t) \xrightarrow{P} 0$, the first term on the right in (2.17) equals 0 for all $n$ large enough.

Since $r_n(t_0) \xrightarrow{P} 0$, then

$$G_{k-1} \left( \sum_{i=1}^{k} \frac{(t_0 - n\theta_i)^2}{n\theta_i} + r_n(t_0) \right) \xrightarrow{P} G_{k-1} \left( \sum_{i=1}^{k} \frac{(t_0 - n\theta_i)^2}{n\theta_i} \right).$$
Combining all this we have proved (ii).

(iii) First we note that if \( p_i = 0 \), then \( P_T(t_i = 0) = 1 \) and so \( p_i \log(\theta_i/p_i) = t_i \log(\theta_i/p_i) = 0 \). Now \((z_1, \ldots, z_k) = (t_1 - np_1, \ldots, t_k - np_k)/n^{1/2} \overset{d}{\to} N_k(0, \Sigma) \) as \( n \to \infty \), where \( \sigma_{ii} = p_i(1 - p_i) \) and \( \sigma_{ij} = -p_ip_j \) when \( i \neq j \). We have that

\[
\log P_T(t_0) = \sum_{i=1}^{k} t_{i0}(\log \theta_i - \log p_i) + \sum_{i=1}^{k} (t_{i0} \log p_i - \log t_{i0}) + \log n!. \tag{2.18}
\]

We can apply the same analysis to \( \sum_{i=1}^{k} (t_{i0} \log p_i - \log t_{i0}) \) as we did in (i) but now we must take into account that whenever \( p_i = 0 \), then corresponding terms are dropped. Doing this we obtain (always interpreting \( 0 \cdot \infty = 0 \))

\[
s_n(t_0) = -\log P_T(t_0) - \frac{1}{2} \sum_{i=1}^{k} \log \theta_i - \left( \log n! - n \log n + n - \frac{k}{2} \ln 2\pi n \right)
\]

\[
= -n \sum_{i=1}^{k} \frac{t_{i0}}{n} \log \frac{\theta_i}{p_i} + \sum_{i=1}^{k} \frac{(t_{i0} - np_i)^2}{2np_i} - \frac{1}{2} \sum_{i=1}^{k} \log \theta_i + \frac{1}{2} \sum_{i=1,p_i \neq 0}^{k} \log p_i + \frac{l}{2} \ln 2\pi n + r^*_n(t_0)
\]

where \( l \) is the number of \( p_i = 0 \) and \( r^*_n(t_0) \xrightarrow{P} 0 \) when sampling from the Multinomial\((1, p_1, \ldots, p_k)\) distribution. Arguing as in (ii), we have that the \( P \)-value given by (2.5) converges in probability to \( 1 - G_{k-1}(s_n(t_0)) \). Under sampling from the Multinomial\((1, p_1, \ldots, p_k)\) distribution \( \sum_{i=1}^{k} (t_{i0}/n) \log(\theta_i/p_i) \xrightarrow{a.s.} \sum_{i=1}^{k} p_i \log(\theta_i/p_i) \). This value equals minus the Kullback-Leibler distance between the \( p_i \) and \( \theta_i \) distributions and so is negative when \( p_i \neq \theta_i \) for some \( i \). It is then immediate that \( s_n(t_0) \xrightarrow{P} \infty \) and this completes the proof of (iii).

**Proof of Theorem 2.2**

Let \( B \) be a bounded set formed from a union of elements of \( \{B_1(x) : x \in \mathbb{R}^k\} \), such that \( P(B^c) < \epsilon \) and \( x_0 \in B \). As \( P(B_n(x) \mid B) = P(B_n(x))/P(B) \) when \( B_n(x) \subset B \) and
Chapter 2. Invariant $P$-value for Model Checking

$$P(B_n(x) \mid B) = 0$$ otherwise, we have that

$$\left| \sum_{\{B_n(x) : P(B_n(x)) \leq P(B_n(x_0))\}} P(B_n(x)) - P(f(x) \leq f(x_0)) \right|$$

$$\leq 2\epsilon + \left| \sum_{\{B_n(x) : P(B_n(x)) \leq P(B_n(x_0)) \mid B\}} P(B_n(x) \mid B) - P(f(x) \leq f(x_0) \mid B) \right| P(B).$$

So, if we prove that

$$\sum_{\{B_n(x) : P(B_n(x)) \leq P(B_n(x_0)) \mid B\}} P(B_n(x) \mid B) \rightarrow P(f(x) \leq f(x_0) \mid B)$$

as $n \rightarrow \infty$, then the result will be established. Accordingly, we hereafter assume that $\mathcal{X}$ is contained in a bounded set $B$ with $\{B_n(x) : x \in \mathbb{R}^k\}$ a finite partition of $B$.

Now suppose that $f$ is unbounded on $\mathcal{X}$ and let $\epsilon > 0$. Let $M > 0$ and $\mathcal{X}_M = \{x : f(x) < M\}$. Since $P(\mathcal{X}_M^c) \rightarrow 0$ as $M \rightarrow \infty$, we can find $M$ such that $P(\mathcal{X}_M^c) < \epsilon$. Since

$$\cup\{B_n(x) : B_n(x) \cap \mathcal{X}_M^c \neq \emptyset\} B_n(x)$$

is monotonically decreasing to $\mathcal{X}_M^c$, there exists $n_0$ such that for all $n \geq n_0$, then

$$\left| \sum_{\{B_n(x) : B_n(x) \cap \mathcal{X}_M^c \neq \emptyset\}} P(B_n(x)) - P(\mathcal{X}_M^c) \right| < \epsilon.$$

Therefore, taking $B' = B \setminus \cup\{x : B_{n_0}(x) \cap \mathcal{X}_M^c \neq \emptyset\} B_{n_0}(x)$, and reasoning as in the preceding paragraph with $B'$ replacing $B$, we see that we need only prove the result when $f$ is bounded. We assume $f$ is bounded hereafter.

Since $f$ is continuous on $\mathcal{X}$, we have that $P(B_n(x))/\mu(B_n(x)) \rightarrow f(x)$ for all $x \in \mathcal{X}$. For each $x \in \mathcal{X}$ there exists $x'_n(x) \in B_n(x)$ such that $P(B_n(x)) = f(x'_n(x))\mu(B_n(x))$ and so, since $\mu(B_n(x))$ is finite and constant, (2.7) equals

$$\sum_{\{x_n(x) : f(x'_n(x)) \leq f(x'_n(x_0))\}} f(x'_n(x))\mu(B_n(x)).$$

(2.19)

Since $\mathcal{X}$ is contained in the union of finitely many of the $B_n(x)$, the sum in (2.19) is a
finite sum. Now

\[
\sum_{\{x_0(x): f(x'_n(x)) \leq f(x'_n(x_0))\}} f(x'_n(x)) \mu(B_n(x)) = \sum_{\{x_0(x): f(x'_0) < f(x_0)\}} f(x'_n(x)) \mu(B_n(x)) + \\
\sum_{\{x_0(x): f(x_0) \leq f(x'_n(x)) \leq f(x'_n(x_0))\}} f(x'_n(x)) \mu(B_n(x)) - \\
\sum_{\{x_0(x): f(x'_n(x_0)) \leq f(x'_n(x)) < f(x_0)\}} f(x'_n(x)) \mu(B_n(x)).
\]

We have that

\[
\sum_{\{x_0(x): f(x'_n(x)) < f(x_0)\}} f(x'_n(x)) \mu(B_n(x)) \to P(f(x) < f(x_0))
\]
as \(n \to \infty\) as the left side is an approximating Riemann sum to the right side. Further, \(f(x'_n(x_0)) \to f(x_0)\) as \(n \to \infty\) and so, for \(\epsilon > 0\) we can find \(n_\epsilon\) such that for all \(n \geq n_\epsilon\), then \(|f(x_0) - f(x'_n(x_0))| < \epsilon\). Accordingly,

\[
\sum_{\{x_0(x): f(x'_n(x)) \leq f(x'_n(x)) < f(x_0)\}} f(x'_n(x)) \mu(B_n(x)) \leq \sum_{\{x_0(x): f(x_0) - \epsilon < f(x'_n(x)) < f(x_0)\}} f(x'_n(x)) \mu(B_n(x))
\]

\[
\to P(f(x_0) - \epsilon < f(x) < f(x_0))
\]
as \(n \to \infty\) and this upper bound converges to 0 as \(\epsilon \to 0\).

Now we have that \(1 = I_{LC(x_0)} + I_{f^{-1}f(x_0) \cap LC(x_0)^c} + I_{(f^{-1}f(x_0))^c}\),

\[
\sum_{\{x_0(x): f(x_0) \leq f(x'_n(x)) \leq f(x'_n(x_0))\}} I_{f^{-1}f(x_0) \cap LC(x_0)^c}(x'_n(x)) f(x'_n(x)) \mu(B_n(x)) = f(x_0) \sum_{\{x_0(x): f(x_0) \leq f(x'_n(x)) \leq f(x'_n(x_0))\}} I_{f^{-1}f(x_0) \cap LC(x_0)^c}(x'_n(x)) \mu(B_n(x)) \leq f(x_0) \sum_{x_0(x)} I_{f^{-1}f(x_0) \cap LC(x_0)^c}(x'_n(x)) \mu(B_n(x))
\]

\[
\to f(x_0) \mu(f^{-1}f(x_0) \cap LC(x_0)^c)
\]
where \( \mu(f^{-1}f(x_0) \cap LC(x_0)^c) = 0 \) and

\[
\sum_{\{x_n(x):f(x_0) \leq f(x_n(x)) \leq f(x_n'(x_0))\}} I_{(f^{-1}f(x_0))^c}(x_n'(x))f(x_n'(x))\mu(B_n(x))
\]

\[
\leq \sum_{\{x_n(x):f(x_0) \leq f(x_n'(x)) \leq f(x_0)+\epsilon\}} I_{(f^{-1}f(x_0))^c}(x_n'(x))f(x_n'(x))\mu(B_n(x))
\]

\[
\to P(\{x: f(x_0) \leq f(x) \leq f(x_0)+\epsilon\} \cap (f^{-1}f(x_0))^c)
\]

and this converges to \( P(f^{-1}f(x_0) \cap (f^{-1}f(x_0))^c) = 0 \) as \( \epsilon \to 0 \).

Finally, we consider

\[
\sum_{\{x_n(x):f(x_0) \leq f(x_n'(x)) \leq f(x_n'(x_0))\}} I_{LC(x_0)}(x_n'(x))f(x_n'(x))\mu(B_n(x)).
\]

Now \( LC(x_0) \) is covered by finitely many of the \( B_n(x) \). Let \( LC^\epsilon(x_0) \) be the set of points in \( LC(x_0) \) that lie a distance greater than \( \epsilon \) from \( \partial LC(x_0) \). Since \( LC(x_0) \) is an open set, \( LC^\epsilon(x_0) \uparrow LC(x_0) \) as \( \epsilon \to 0 \) and so \( \mu(LC^\epsilon(x_0)) \uparrow \mu(LC(x_0)) \). We can choose \( n_\epsilon \) so that when \( n \geq n_\epsilon \), then \( x_n(x) \in LC^{2\epsilon}(x_0) \) then \( x_n'(x) \in LC^\epsilon(x_0) \) and so

\[
\sum_{\{x_n(x):f(x_0) \leq f(x_n'(x)) \leq f(x_n'(x_0))\}} I_{LC^{2\epsilon}(x_0)}(x_n'(x))f(x_n'(x))\mu(B_n(x))
\]

\[
= f(x_0) \sum_{x_n(x)} I_{LC^{2\epsilon}(x_0)}(x_n'(x))\mu(B_n(x))
\]

\[
\leq \sum_{\{x_n(x):f(x_0) \leq f(x_n'(x)) \leq f(x_n'(x_0))\}} I_{LC(x_0)}(x_n'(x))f(x_n'(x))\mu(B_n(x))
\]

\[
\leq \sum_{\{x_n(x):f(x_0) \leq f(x_n'(x)) \leq f(x_0)+\epsilon\}} I_{LC(x_0)}(x_n'(x))(f(x_0) + \epsilon)\mu(B_\delta(x)). \tag{2.20}
\]

Now the left-hand side of (2.20) converges to \( f(x_0)\mu(LC^{2\epsilon}(x_0)) \) which converges to \( f(x_0)\mu(LC(x_0)) = P(f(x) = f(x_0)) \) as \( \epsilon \to 0 \). The right-hand side of (2.20) converges to

\[
P(LC(x_0) \cap \{x : f(x_0) \leq f(x) \leq f(x_0)+\epsilon\}) + \epsilon\mu(LC(x_0) \cap \{x : f(x_0) \leq f(x) \leq f(x_0)+\epsilon\})
\]

which converges to \( P(f(x) = f(x_0)) \) as \( \epsilon \to 0 \) and this establishes the result.
Chapter 2. Invariant $P$-value for Model Checking

63

Lemmas Used in Example 2.3

Lemma 2.11. (i) $\mu_k(A) \geq 1 - 2\delta$.

(ii) For any $x_0 \in A$, $P(f(x) \leq f(x_0)) = 1 - \beta - \alpha$ where $\alpha = \sum_{i=1}^{\infty} (b_i - a_i)/2$ and $\beta = \sum_{i=1}^{\infty} (b_i - a_i)^3/\pi^2$.

(iii) For $x_0 = b_j$ for some $j$,

$$\liminf_{n \to \infty} \sum_{B_n(x): P(B_n(x)) \leq P(B_n(x_0))} P(B_n(x)) \to \alpha - \beta.$$  

Proof of Lemma 2.11. Part (i) is obtained from

$$\mu_k(A) = \mu_k((0, 1) \backslash A_0) = \mu_k((0, 1) \cap \bigcap_{i=1}^{\infty} (q_i - \delta 2^{-i}, q_i + \delta 2^{-i})^c) \geq 1 - \sum_{i=1}^{\infty} \mu_k((q_i - \delta 2^{-i}, q_i + \delta 2^{-i})) = 1 - \sum_{i=1}^{\infty} 2\delta 2^{-i} = 1 - 2\delta.$$  

Part (ii). For any $x_0 \in A$, $f(x_0) = 1$ and

$$P(f(x) \leq f(x_0)) = P(f(x) \leq 1) = 1 - P(f(x) > 1) = 1 - P(\bigcup_{i=1}^{\infty} (a_i, (a_i + b_i)/2))$$

$$= 1 - \sum_{i=1}^{\infty} \int_{a_i}^{a_i + b_i} f(x) \, dx = 1 - \sum_{i=1}^{\infty} \left[ \frac{b_i - a_i}{2} + \int_{a_i}^{b_i + a_i} \frac{(b_i - a_i)^2}{\pi} \sin \frac{2\pi(x - a_i)}{b_i - a_i} \, dx \right]$$

$$= 1 - \alpha + \sum_{i=1}^{\infty} \frac{(b_i - a_i)^3}{2\pi^2} \cos \frac{2\pi(a_i + b_i)/2}{b_i - a_i} \bigg|_{x=a_i} = 1 - \alpha - \sum_{i=1}^{\infty} \frac{(b_i - a_i)^3}{\pi^2}$$

$$= 1 - \alpha - \beta.$$  

Part (iii) is proved in the following sequence of lemmas.

Let $A = \{a_i, b_i : i = 1, 2, \ldots \}$ be the collection of all end points of $A_0$ and $l_n(x)$ and $u_n(x)$ be the lower and upper bounds of $B_n(x)$.

Lemma 2.12. (i) Each $x \in A$ is irrational. (ii) The set $N(x) = \{n \in \mathbb{N} : x - l_n(x) > u_n(x) - x \}$ has infinitely many elements for $x \in \mathcal{X}$.

Proof of Lemma 2.12. (i) Any $x \in A$ is also an element of $A$ because of the definition of $A$ given by $A = (0, 1) \backslash \bigcup_{i=1}^{\infty} (a_i, b_i)$. Also $A = (0, 1) \backslash \bigcup_{i=1}^{\infty} (q_i - \delta 2^{-i}, q_i + \delta 2^{-i})$ does not
contain rational numbers in \( X \). If \( x \in X \) is a rational number, then there is \( j \) such that \( x = q_j \). But the interval \((q_j - \delta 2^{-j}, q_j + \delta 2^{-j})\) is excluded from \( A \). So \( A \) does not contain rational numbers. Thus \( A \) does not contain any rational numbers.

**(ii)** The binary expansion of \( x \) is given by \( x = [0.x_1x_2x_3 \cdots ]_2 = \sum_{i=1}^{\infty} x_i 2^{-i} \). Then, \( 2^n x = [x_1 \cdots x_n.x_{n+1}x_{n+2} \cdots ]_2 \). If \( x_{n+k} = 1 \) for some \( k > 0 \), then \( u_n(x) = 2^{-n} + [0.x_1 \cdots x_n]_2 \). Otherwise, i.e., \( x_{n+k} = 0 \) for all \( k \in \mathbb{N} \) or equivalently \( x = m2^{-n} \) for some \( m \in \mathbb{N} \), \( u_n(x) = [0.x_1 \cdots x_n]_2 = x \).

For the second case, i.e., \( u_n(x) = x \). Then, \( u_k(x) = x \) for all \( k \geq n \). Note \( l_k(x) = u_k(x) - 2^{-k} \). Thus, \( x - l_k(x) = u_k(x) - l_k(x) = 2^{-k} > 0 = u_k(x) - x \) for all \( k \geq n \). Hence, \( N(x) \) contains infinitely many elements.

For the first case, i.e., there are infinitely many 1’s in \( x_1, x_2, \ldots \), assume that \( N(x) \) is finite. There is a big number \( M \in \mathbb{N} \) such that \( x - l_n(x) \leq u_n(x) - x \) for all \( n \geq M \). Also \( x - l_n(x) \leq u_n(x) - x \) is equivalent to \( x_{n+1} = 0 \) because the equality event belongs to the second case. Hence, we get \( x_n = 0 \) for all \( n \geq 1 + M \). In other words \( x = [0.x_1 \cdots x_M]_2 = m2^{-M} \) for some \( m \in \mathbb{N} \). It makes a contradiction. Hence \( N(x) \) must contain infinitely many elements.

For any \( x \), there exist two intervals \((a_i, b_i)\) and \((a_j, b_j)\) containing \( l_n(x) \) and \( u_n(x) \) respectively. Then, define \( a_{t,n}(x) = a_i \), \( b_{t,n}(x) = b_i \), \( a_{u,n}(x) = a_j \), and \( b_{u,n}(x) = b_j \). Since \( a_{t,n}(x) < b_{t,n}(x) \), \( a_{u,n}(x) < b_{u,n}(x) \) are irrational and \( l_n(x), u_n(x) \) are rational, we have \( a_{t,n}(x) < l_n(x) < b_{t,n}(x) < a_{u,n}(x) < u_n(x) < b_{u,n}(x) \). To prove part (iii), we need lower and upper bounds of \( P(B_n(x)) \) for \( x \in A \).

Suppose \( x \in A \). Let \( M_n(x) = \{ i \in \mathbb{N} : (a_i, b_i) \cap B_n(x) \neq \emptyset \} \).

\[
P(B_n(x)) = \int_{l_n(x)}^{u_n(x)} f(v)dv = \int_{l_n(x)}^{u_n(x)} 1 + \sum_{i \in M_n(x)} I_{(a_i, b_i)}(v) \frac{(b_i - a_i)^2}{\pi} \sin 2\pi \frac{(v - a_i)}{b_i - a_i}dv
\]

\[
= 2^{-n} + \int_{l_n(x)}^{b_{t,n}(x)} \frac{(b_{t,n}(x) - a_{t,n}(x))^2}{\pi} \sin 2\pi \frac{(v - a_{t,n}(x))}{b_{t,n}(x) - a_{t,n}(x)}dv
\]

\[
+ \int_{a_{u,n}(x)}^{u_n(x)} \frac{(b_{u,n}(x) - a_{u,n}(x))^2}{\pi} \sin 2\pi \frac{(v - a_{u,n}(x))}{b_{u,n}(x) - a_{u,n}(x)}dv
\]
\[ P(B_n(x)) \leq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x))(b_{l,n}(x) - l_n(x))^2 \]

The last line is derived using the equality \( 1 - \cos x = 2 \sin^2(x/2) \).

From the trigonometric inequalities in Lemma A.6, we have

\[ P(B_n(x)) \geq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x))(b_{l,n}(x) - l_n(x))^2 \]

and also from \( 0 \leq \frac{2\pi(b_{l,n}(x) - l_n(x))}{b_{l,n}(x) - a_{l,n}(x)} \leq 2\pi < \sqrt{56} \) and Lemma A.6,

\[ P(B_n(x)) \leq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x))(b_{l,n}(x) - l_n(x))^2 \left(1 - \frac{\pi^2(b_{l,n}(x) - l_n(x))^2}{12(b_{l,n}(x) - a_{l,n}(x))^2}\right) + (b_{u,n}(x) - a_{u,n}(x))(u_n(x) - a_{u,n}(x))^2. \]

**Proof of Lemma 2.11 (iii).** Let \( \mathcal{C}_{1,n} \) be the union of \( B_n(x) \) such that \( B_n(x) \cap \mathcal{A} \neq \emptyset \) and \( P(B_n(x)) \leq P(B_n(x_0)) \). Also let \( \mathcal{C}_{2,n} \) be the union of \( B_n(x) \) such that \( B_n(x) \cap \mathcal{A} = \emptyset \) and \( P(B) \leq P(B_n(x_0)) \). Then, we prove that (a) \( P(\mathcal{C}_{2,n}) \to \alpha - \beta \) as \( n \to \infty \) and (b) \( P(\mathcal{C}_{1,n}) \to 0 \) as \( n \to \infty \) subject to \( n \in N(x_0) \).

(a) By the continuity of \( f \), \( P(B_n(x_0))/\mu_k(B_n(x_0)) \to 1 \) and \( P(B_n(x))/\mu_k(B_n(x)) \to f(x) \) for all \( x \in \bigcup_{i=1}^{\infty} (a_i, b_i) \). It is not difficult to show that \( \liminf_{n \to \infty} \mathcal{C}_{2,n} \supseteq \mathcal{D}_{1,\epsilon} = \{ x \in \mathcal{X} : f(x) < 1 - \epsilon \} \) and \( \limsup_{n \to \infty} \mathcal{C}_{2,n} \subset \mathcal{D}_{2,\epsilon} = \{ x \in \mathcal{X} : f(x) < 1 + \epsilon, a_i < x < b_i \text{ for some } i \} \) for any fixed \( \epsilon > 0 \). Hence, \( \{ x \in \mathcal{X} : f(x) < 1 \} \subset \liminf_{n \to \infty} \mathcal{C}_{2,n} \subset \limsup_{n \to \infty} \mathcal{C}_{2,n} \subset \{ x \in \mathcal{X} : f(x) \leq 1, a_i < x < b_i \text{ for some } i \} \). Note \( P(\{ x \in \mathcal{X} : f(x) = 1, a_i < x < b_i \text{ for some } i \}) = 0 \) because the event set has countably many points. Thus,

\[ P(\mathcal{C}_{2,n}) \to P(\{ x \in \mathcal{X} : f(x) < 1 \}) = \sum_{i=1}^{\infty} \int_{a_i}^{b_i} 1 + \frac{(b_i - a_i)^2}{\pi} \sin \frac{2\pi(x - a_i)}{b_i - a_i} dx = \alpha - \beta. \]

(b) Recall that \( x_0 = b_j \) for some \( j \). Fix \( \epsilon \in (0, 1/4) \). For large \( n \) such that \( b_j - a_j > 2^{-n} \), the sequences \( a_{l,n}(x_0) \) and \( b_{l,n}(x_0) \) are uniquely determined by \( a_{l,n}(x_0) = a_j \) and \( b_{l,n}(x_0) = \)}
Chapter 2. Invariant $P$-value for Model Checking

$b_j$ because $a_j < l_n(b_j) < b_j < u_n(b_j)$. As $n$ increases, $b_{u,n}(x_0) - a_{u,n}(x_0)$ converges to 0 because the length of $B_n(x_0)$ shrinks to 0. Hence, there is a number $N_0 > 0$ such that $b_j - a_j > 2^{-n}$, $2^{-2n} < \epsilon 48(b_j - a_j)/\pi^2$, and $b_{u,n}(b_j) - a_{u,n}(b_j) < \epsilon(b_j - a_j)$ for all $n \geq N_0$. Then, for all $n \geq N_0$ in $N(x_0)$ defined in Lemma 2.12, the upper bound of $P(B_n(x_0))$ in (2.22) becomes

$$P(B_n(x_0)) \leq 2^{-n} - (b_j - a_j)(b_j - l_n(b_j))^2(1 - \frac{\pi^2(b_j - l_n(b_j))^2}{12(b_j - a_j)^2})$$

$$+ (b_{u,n}(b_j) - a_{u,n}(b_j))(u_n(b_j) - a_{u,n}(b_j))^2$$

$$\leq 2^{-n} - (b_j - a_j)(2^{-n}/2)^2(1 - \frac{\pi^2(2^{-n}/2)^2}{12(b_j - a_j)^2}) + \epsilon(b_j - a_j)(2^{-n}/2)^2$$

Since $2^{-2n} < \epsilon 48(b_j - a_j)/\pi^2$ implies $\pi^2 2^{-2n}/[48(b_j - a_j)] < \epsilon$, we have

$$< 2^{-n} - (b_j - a_j)2^{-2n}(1 - 2\epsilon)/4.$$ 

For any $x \in A$, the inequality (2.21) becomes

$$P(B_n(x)) \geq 2^{-n} - (l_{t,n}(x) - a_{t,n}(x))(l_{t,n}(x) - l_{n}(x))^2 \geq 2^{-n} - (l_{t,n}(x) - a_{t,n}(x))2^{-2n}.$$ 

Hence, for all $x \in A$ satisfying $(l_{t,n}(x) - a_{t,n}(x)) < (b_j - a_j)/8$,

$$P(B_n(x)) \geq 2^{-n} - (l_{t,n}(x) - a_{t,n}(x))2^{-2n} > 2^{-n} - (b_j - a_j)2^{-2n}/8$$

$$\geq 2^{-n} - (b_j - a_j)2^{-2n}(1 - 2\epsilon)/4 > P(B_n(x_0)).$$

Thus, for $x \in A$, $P(B_n(x)) \leq P(B_n(x_0))$ implies $b_{t,n}(x) - b_{t,n}(x) \geq (b_j - a_j)/8$.

Since $b_i - a_i \to 0$ as $i \to \infty$, there is $J \in \mathbb{N}$ such that $b_i - a_i < (b_j - a_j)/8$ for all $i > J$. It is easy to check that

$$C_{1,n} = \bigcup \{B_n(x) : B_n(x) \cap A \neq \emptyset, P(B_n(x)) \leq P(B_n(b_j))\}$$

$$= \bigcup \{B_n(x) : x \in A, P(B_n(x)) \leq P(B_n(b_j))\}$$

$$\subset \{B_n(b_i) | i \leq J\}.$$ 

Then, $P(C_{1,n}) \leq P(\cup_{i \leq J} B_n(b_i)) \leq J/2^n \to 0$ as $n \to \infty$ subject to $n \in N_0(x_0)$. Therefore the lemma follows by combining (a) and (b).
Chapter 3

Weak Informativity and the Information in One Prior Relative to Another

A question of some interest is how to characterize the amount of information that a prior puts into a statistical analysis. Rather than a general characterization of this information, we provide an approach to characterizing the amount of information a prior puts into an analysis, when compared to another base prior. The base prior is considered to be the prior that best reflects the current available information. Our purpose then, is to characterize priors that can be used as conservative inputs to an analysis, relative to the base prior, in the sense that they put less information into the analysis. The characterization that we provide is in terms of a priori measures of prior-data conflict.

3.1 Introduction

Suppose we have two proper priors $\Pi_1$ and $\Pi_2$ on a parameter space $\Theta$ for a statistical model $\{P_\theta : \theta \in \Theta\}$. A natural question to ask is: how do we compare the amount
of information each of these priors puts into the problem? While there may seem to be natural intuitive ways to express this, such as prior variances, it seems difficult to characterize this precisely in general. For example, the consideration of several examples in Sections 3.3 and 3.4 makes it clear that using the variance of the prior is not appropriate for this task.

The motivation for this work comes from Gelman (2006) and Gelman et al. (2008), where the intuitively satisfying notion of weakly informative priors is introduced as a compromise between informative and noninformative priors. The basic idea is that we have a base prior $\Pi_1$, perhaps elicited, that we believe reflects our current information about $\theta$, but we choose to be conservative in our inferences and select a prior $\Pi_2$ that puts less information into the analysis. While it is common to take $\Pi_2$ to be a noninformative prior, this can often produce difficulties when $\Pi_2$ is improper, and even when $\Pi_2$ is proper, it seems inappropriate as it completely discards the information we have about $\theta$ as expressed in $\Pi_1$.

To implement the idea of weak informativity we need a precise definition and so our purpose here is to give a definition of what it means for one prior to be weakly informative relative to another. We do this in Section 3.2 and it involves the notion of prior-data conflict. Intuitively, a prior-data conflict occurs when the prior places the bulk of its mass where the likelihood is relatively low. A completely noninformative prior should produce no prior-data conflicts \textit{a priori}. Our definition of weak informativity can be expressed as saying that $\Pi_2$ is weakly informative relative to $\Pi_1$ whenever $\Pi_2$ produces fewer prior-data conflicts \textit{a priori} than $\Pi_1$. This leads to a quantifiable expression of weak informativity that can be used to choose priors. In Section 3.3 we consider this definition in the context of several standard families of priors and it is seen to produce results that are intuitively reasonable. In Section 3.4 we consider applications of this concept in some data analysis problems. While our intuition about weak informativity is often borne out, we also find that in certain situations we have to be careful before calling a prior weakly
informative.

First, however, we establish some notation and then review how we check for prior-data conflict. We suppose that \( P_\theta(A) = \int_A f_\theta(x) \mu(dx) \), i.e., each \( P_\theta \) is absolutely continuous with respect to a support measure \( \mu \) on the sample space \( \mathcal{X} \), with the density denoted by \( f_\theta \). With this formulation a prior \( \Pi \) leads to a prior predictive probability measure on \( \mathcal{X} \) given by

\[
M(A) = \int_\Theta \left[ \int_A P_\theta(A) \Pi(d\theta) \right] = \int_A m(x) \mu(dx),
\]

where \( m(x) = \int_\Theta f_\theta(x) \Pi(d\theta) \).

If \( T \) is a minimal sufficient statistic for \( \{P_\theta : \theta \in \Theta\} \), then it is well known that the posterior is the same whether we observe \( x \) or \( T(x) \). So we will denote the posterior by \( \Pi(\cdot | T) \) hereafter. Since \( T \) is minimal sufficient we know that the conditional distribution of \( x \) given \( T \) is independent of \( \theta \). We denote this conditional measure by \( P(\cdot | T) \).

The joint distribution \( P_\theta \times \Pi \) can then be factored as

\[
P_\theta \times \Pi = M \times \Pi(\cdot | x) = P(\cdot | T) \times M_T \times \Pi(\cdot | T)
\]

where \( M_T \) is the marginal prior predictive distribution of \( T \).

While much of Bayesian analysis focuses on the third factor in (3.1), there are also roles in a statistical analysis for \( P(\cdot | T) \) and \( M_T \). As discussed in Evans and Moshonov (2006, 2007), \( P(\cdot | T) \) is available for checking the sampling model, e.g., if \( x \) is a surprising value from this distribution, then we have evidence that the model \( \{P_\theta : \theta \in \Theta\} \) is incorrect. Further it is argued that, if we conclude that we have no evidence against the model, then the factor \( M_T \) is available for checking whether or not there is any prior-data conflict. So if \( T(x) \) is a surprising value from \( M_T \), then we have evidence that the prior \( \Pi \) is placing most of its mass on \( \theta \) values where the likelihood is relatively low. This is supported by the fact that \( T \) is equivalent to the likelihood map. Finally, if we have no evidence against the model, and no evidence of prior-data conflict, then \( \Pi(\cdot | T) \) is available for probability statements about \( \theta \). Actually the issues involved in model checking and checking for prior-data conflict are more involved than this (see, for example, the cited references and Section 3.5), but (3.1) gives the basic idea that the full information, as expressed by the joint distribution of \((\theta, x)\), splits into components, each
Chapter 3. Weakly Informative Priors

of which is available for a specific purpose in a statistical analysis.

Accordingly we restrict ourselves here, for any discussions about the respective merits of priors, to working with \( M_T \). One issue that needs to be addressed is how one is to compare the observed data \( x_0 \) to \( P(\cdot | T) \) or compare \( t_0 = T(x_0) \) to \( M_T \). In essence we need a measure of surprise. Perhaps the best measure of surprise is the \( P \)-value. Effectively, we are in the situation where we have a value from a single fixed distribution and we need to specify the appropriate \( P \)-value to use. In Evans and Moshonov (2006, 2007) the \( P \)-value for checking for prior-data conflict was based on the prior predictive density \( m_T \), namely,

\[
M_T(m_T(t) \leq m_T(t_0)).
\]  

A difficulty with (3.2) is that it is not generally invariant to the choice of the minimal sufficient statistic \( T \). A general invariant \( P \)-value is developed in Evans and Jang (2008) for such situations. When applied to (3.2), this leads to using the invariant \( P \)-value

\[
M_T(m_T^*(t) \leq m_T^*(t_0))
\]  

instead, where \( m_T^*(t) = \int_\Theta f_{\theta T}^*(t) \Pi(d\theta) \), \( f_{\theta T}(t) = \int_{T^{-1}\{t\}} f_\theta(x) \nu_{T^{-1}\{t\}}(dx) \) and \( \nu_{T^{-1}\{t\}} \) is volume measure on \( T^{-1}\{t\} \). Note that it can be shown that the marginal density of \( T \), with respect to volume measure on the range space for \( T \), is given by \( f_{\theta T}(t) = \int_{T^{-1}\{t\}} f_\theta(x) J_T(x) \nu_{T^{-1}\{t\}}(dx) \) where \( J_T(x) = (\det (dT(x) \circ dT'(x)))^{-1/2} \) and \( dT \) is the differential of \( T \). So \( f_{\theta T}^*(t) \) is an adjustment of \( f_{\theta T}(t) \) where we do not allow the volume distortions induced by \( T \) to affect the density. It is also shown in Evans and Jang (2008) that \( f_{\theta T}^*(t) = f_{\theta,T}(t) E_\theta(J_T(X)^{-1} | T = t) \) so \( f_{\theta T}^*(t) \) can be thought of as a multiplication of the usual marginal density by the expected volume correction given \( T = t \). We will refer to the invariant \( P \)-value (3.3) throughout the remainder of our discussion but note that, for many of the examples in this paper, \( J_T(x) \) is constant, e.g., whenever \( T \) is linear in \( x \), and then (3.2) and (3.3) are equal.
3.2 Comparing Priors

There are a variety of measures of information used in statistics. Several measures have been based on the concept of entropy, e.g., see Lindley (1956) and Bernardo (1979). While these measures have their virtues, we note that their coding theory interpretations can seem somewhat abstract in statistical contexts and they can suffer from nonexistence in certain problems. Also, Kass and Wasserman (1995) contains some discussion concerned with expressing the absolute information content of a prior in terms of additional sample values. Rather than adopting these approaches, we consider here comparing priors based on their tendencies to produce prior-data conflicts. The basic intuitive idea is that a prior which produces fewer prior-data conflicts than another, is putting less information into an analysis. This formulation of the relative amount of information put into an analysis, has a direct interpretation in terms of statistical consequences.

Suppose that an analyst has in mind a prior $\Pi_1$ that they believe represents the information at hand concerning $\theta$. The analyst, however, prefers to use a prior $\Pi_2$ that is somewhat conservative, with respect to the amount of information put into the analysis, when compared to $\Pi_1$. The motivation for this lies with a desire not to put too much information into the analysis via the prior, and avoid the use of completely noninformative priors. Often priors that are characterized as being noninformative are improper and their use can be challenged for a variety of reasons. In such a situation it seems reasonable to consider $\Pi_1$ as a base prior and then compare all other priors to it. This idea comes from Gelman (2006) and leads to the notion of weakly informative priors, but without a precise characterization of this concept.

For a given prior $\Pi_1$ and observed value $t_0 = T(x)$ then, from (3.3), we have that $M_{1T}(m_{1T}(t) \leq m_{1T}(t_0))$ is the relevant quantity for assessing whether or not there is prior-data conflict with $\Pi_1$. Before we observe data, however, we have no way of knowing if we will have a prior-data conflict. Accordingly, since the analyst has determined that $\Pi_1$ best reflects the available information, it is reasonable to consider the prior distribution of
Chapter 3. Weakly Informative Priors

\[ P_1(t_0) = M_{1T}(m^*_{1T}(t) \leq m^*_{1T}(t_0)) \] when \( t_0 \sim M_{1T} \). Of course, this is effectively uniformly distributed (exactly so when \( m^*_{1T}(t) \) has a continuous distribution when \( t \sim M_{1T} \)) and this expresses the fact that all the information about assessing whether or not a prior-data conflict exists, is contained in the \( P \)-value, with no need to compare the \( P \)-value to its distribution.

Consider now, however, the distribution of \( P_2(t_0) = M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0)) \) which is used to check whether or not there is prior-data conflict with respect to \( \Pi_2 \). Given we have identified that \textit{a priori} the appropriate distribution of \( t_0 \) is \( M_{1T} \), at least for inferences about an unobserved value, then \( P_2(t_0) \) is not uniformly distributed. In fact, from the distribution of \( P_2(t_0) \) we can obtain an intuitively reasonable idea of what it means for a prior \( \Pi_2 \) to be weakly informative relative to \( \Pi_1 \). For suppose that the prior distribution of \( P_2(t_0) \) clusters around 1. This implies that, if we were to use \( \Pi_2 \) as the prior when \( \Pi_1 \) is appropriate, then there is a small prior probability that a prior-data conflict would arise. Similarly, if the prior distribution of \( P_2(t_0) \) clusters around 0, then there is a large prior probability that a prior-data conflict would arise. If one prior distribution results in a larger prior probability of there being a prior-data conflict than another, then it seems reasonable to say that the first prior is more informative than the second. In fact, a completely noninformative prior should never produce prior-data conflicts.

So we compare the distribution of \( P_2(t_0) \) when \( t_0 \sim M_{1T} \), to the distribution of \( P_1(t_0) \) when \( t_0 \sim M_{1T} \), and do this in a way that is relevant to the prior probability of obtaining a prior-data conflict. One approach to this comparison is to select a \( \gamma \)-quantile \( x_\gamma \in [0,1] \) of the distribution of \( P_1(t_0) \), and then compute the probability

\[
M_{1T}(P_2(t_0) \leq x_\gamma). \tag{3.4}
\]

The value \( \gamma \) is presumably some cut-off, dependent on the application, where we will consider that evidence of a prior-data conflict exists whenever \( P_1(t_0) \leq \gamma \). Of course, if \( m^*_{1T}(t_0) \) has a continuous distribution when \( t_0 \sim M_{1T} \), then \( x_\gamma = \gamma \). Our criterion for the weak informativity of \( \Pi_2 \) relative to \( \Pi_1 \) will then be that (3.4) is less than or equal to \( x_\gamma \).
This indicates that the prior probability of obtaining a prior-data conflict under $\Pi_2$ is no greater than when $\Pi_1$ is used, at least when we have identified $\Pi_1$ as our correct prior.

**Definition 3.1.** If (3.4) is less than or equal to $x_\gamma$, then $\Pi_2$ is *weakly informative relative to* $\Pi_1$ *at level* $\gamma$. If $\Pi_2$ is weakly informative relative to $\Pi_1$ at level $\gamma$ for every $\gamma \leq \gamma_0$, then $\Pi_2$ is *uniformly weakly informative relative to* $\Pi_1$ *at level* $\gamma_0$. If $\Pi_2$ is weakly informative relative to $\Pi_1$ at level $\gamma$ for every $\gamma$, then $\Pi_2$ is *uniformly weakly informative relative to* $\Pi_1$.

While it would be appealing to be able to choose a prior $\Pi_2$ that is uniformly weakly informative relative to $\Pi_1$, this may not always be preferable. The criterion of uniformly weakly informative at level $\gamma_0$ seems much more generally applicable.

While (3.4) seems difficult to work with, the following result is proved in Section 3.7 and gives a simpler expression.

**Lemma 3.2.** Suppose $P_i(t)$ has a continuous distribution under $M_{iT}$ for $i = 1, 2$. Then there exists $r_\gamma$ such that $M_{1T}(P_2(t) \leq \gamma) = M_{1T}(m_{2T}^*(t) \leq r_\gamma)$, and $\Pi_2$ is weakly informative at level $\gamma$ relative to $\Pi_1$ whenever $M_{1T}(m_{2T}^*(t) \leq r_\gamma) \leq \gamma$. Furthermore, $\Pi_2$ is uniformly weakly informative relative to $\Pi_1$ if and only if $M_{1T}(m_{2T}^*(t) \leq m_{2T}^*(t_0)) \leq M_{2T}(m_{2T}^*(t) \leq m_{2T}^*(t_0))$ for every $t_0$.

Lemma 3.2 typically applies when we are dealing with continuous distributions on $X$. It can also be shown that $P_i(t)$ has a continuous distribution under $M_{iT}$ if and only if $m_{iT}^*(t)$ has a continuous distribution under $M_{iT}$.

Once we have selected $\gamma$, the degree of weak informativity of a prior $\Pi_2$ relative to $\Pi_1$ can be assessed by comparing $M_{1T}(P_2(t_0) \leq x_\gamma)$ to $x_\gamma$ via the ratio

$$1 - M_{1T}(P_2(t_0) \leq x_\gamma)/x_\gamma.$$ (3.5)

If $\Pi_2$ is weakly informative relative to $\Pi_1$ at level $\gamma$, then (3.5) tells us the proportion of fewer prior-data conflicts we can expect *a priori* when using $\Pi_2$ rather than $\Pi_1$. Thus
(3.5) provides a measure of how much less informative \( \Pi_2 \) is than \( \Pi_1 \), e.g., we might ask for a prior \( \Pi_2 \) such that (3.5) equals 50%.

As we will see in the examples it makes sense to talk of one prior being \textit{asymptotically weakly informative at level} \( \gamma \) relative to another prior in the sense that (3.4) is bounded above by \( \gamma \) in the limit as the amount of data increases. In several cases this simplifies matters considerably as an asymptotically weakly informative prior is easy to find and may still be weakly informative for finite amounts of data. For assessing asymptotic weak informativity of a prior relative to another, a useful theorem is stated below and proved in Section 3.7.

**Theorem 3.3.** Consider a full-rank exponential family having density \( \exp(\eta(\theta)'u - A(\eta(\theta)) + h(u)) \) with respect to the Lebesgue measure, \( \mathcal{L} \), on an open of \( \mathbb{R}^k \). Let \( X_1, X_2, \ldots \) be a random sample from this distribution. Let \( U = (U(X_1) + \cdots + U(X_n))/n \) and \( T = \eta^{-1}(U) \). Then, both \( U \) and \( T \) are minimal sufficient. Assume that (i) \( V_{T,n}(t)/V_{T,n}(t^*) \) converges to \( w(t) \) for all \( t \) and a fixed \( t^* \) where \( V_{T,n}(t) = \mathbb{E}(J^{-1}_T(X)|T=t) \), (ii) \( \pi_i(\theta) \) is a continuous density of \( \Pi_i \) with respect to \( \mathcal{L} \), (iii) \( \pi_i(T)w(T) \) has a continuous distribution when \( T \sim \Pi_i \), and (iv) \( \pi_2(T)w(T) \) has a continuous distribution when \( T \sim \Pi_1 \). Then,

\[
M_{1T,n}(P_{2,n}(t) \leq x_{\gamma,n}) \to \Pi_1(P_2(t) \leq \gamma) \quad \text{as} \quad n \to \infty
\]

where \( P_{i,n}(t_0) = M_{iT,n}(m^*_{iT,n}(t) \leq m^*_{iT,n}(t_0)), P_i(t_0) = \Pi_i(\pi_i(t)w(t) \leq \pi_i(t_0)w(t_0)), \) and \( x_{\gamma,n} = \inf\{P_{1,n}(t) : P_{1,n}(t) \geq \gamma\} \).

### 3.3 Deriving Weakly Informative Priors

We consider several examples of priors that arise in applications. These examples support our definition of weak informativity and also lead to some insights into choosing priors.

#### 3.3.1 Comparing Normal Priors

Suppose we have a sample \( x = (x_1, \ldots, x_n) \) from a \( N(\mu, 1) \) distribution where \( \mu \) is unknown. Then \( t = T(x) = \bar{x} \sim N(\mu, 1/n) \) is minimal sufficient and since \( T \) is linear
there is constant volume distortion and so this can be ignored. Suppose that the prior \( \Pi_1 \) on \( \mu \) is a \( N(\mu_0, \sigma_1^2) \) distribution with \( \mu_0 \) and \( \sigma_1^2 \) known. We then have that \( M_{1T} \) is the \( N(\mu_0, 1/n + \sigma_1^2) \) distribution. Now suppose that \( \Pi_2 \) is a \( N(\mu_0, \sigma_2^2) \) distribution with \( \sigma_2^2 \) known. Then \( M_{2T} \) is the \( N(\mu_0, 1/n + \sigma_2^2) \) distribution and

\[
P_2(t_0) = M_{2T}(m_{2T}^*(t) \leq m_{2T}^*(t_0)) = M_{2T}(m_{2T}(t) \leq m_{2T}(t_0))
= M_{2T}((t - \mu_0)^2 \geq (t_0 - \mu_0)^2) = 1 - G_1((t_0 - \mu_0)^2/(1/n + \sigma_2^2)),
\]

where \( G_k \) denotes the \( \chi^2(k) \) distribution function. Now under \( M_{1T} \) we have that \( (t_0 - \mu_0)^2/(1/n + \sigma_2^2) \sim \chi^2(1) \). Therefore,

\[
M_{1T}(P_2(t_0) \leq \gamma) = M_{1T}(1 - G_1((t_0 - \mu_0)^2/(1/n + \sigma_2^2)) \leq \gamma)
= M_{1T} \left( \frac{(t_0 - \mu_0)^2}{1/n + \sigma_2^2} \geq \frac{1/n + \sigma_2^2}{1/n + \sigma_1^2} G_1^{-1}(1 - \gamma) \right)
= 1 - G_1 \left( \frac{1/n + \sigma_2^2}{1/n + \sigma_1^2} G_1^{-1}(1 - \gamma) \right).
\]

We see immediately that (3.6) will be less than \( \gamma \) if and only if \( \sigma_2 > \sigma_1 \). In other words \( \Pi_2 \) will be uniformly weakly informative relative to \( \Pi_1 \) if and only if \( \Pi_2 \) is more diffuse than \( \Pi_1 \). Note that \( M_{1T}(P_2(t_0) \leq \gamma) \) converges to 0 as \( \sigma_2^2 \to \infty \) to reflect noninformativity.

Also, as \( n \to \infty \), then (3.6) increases to \( 1 - G_1((\sigma_2^2/\sigma_1^2)G_1^{-1}(1 - \gamma)) \) and so we could ignore \( n \) and choose \( \sigma_2^2 \) conservatively based on this limit, to obtain an asymptotically uniformly weakly informative prior, as we know this value of \( \sigma_2^2 \) will also be weakly informative for finite \( n \).

If we specify that we want (3.5) to equal \( p \in [0, 1] \), then (3.6) implies that \( \sigma_2^2 = (1/n + \sigma_1^2)(G_1^{-1}(1 - \gamma + p\gamma)/G_1^{-1}(1 - \gamma)) - 1/n \). Such a choice will give a proportion \( p \) fewer prior-data conflicts at level \( \gamma \) than the base prior. This decreases to \( \sigma_1^2 G_1^{-1}(1 - \gamma + p\gamma)/G_1^{-1}(1 - \gamma) \) as \( n \to \infty \) and so the more data we have the less extra variance we need for \( \Pi_2 \) for weak informativity.

We can generalize this to \( t \sim N_k(\mu, n^{-1}I) \) with \( \Pi_1 \) given by \( \mu \sim N_k(\mu_0, \Sigma_1) \). Note we have that \( M_{1T} \) is the \( N_k(\mu_0, n^{-1}I + \Sigma_1) \) distribution. It is then easy to see that
For a sample of Theorem 3.4.

Let $\tilde{X} = (t_0 - \mu_0)'(n^{-1}I + \Sigma_2)^{-1}(t_0 - \mu_0)$ and

$$M_{1T}(P_2(t_0) \leq \gamma) = M_{1T}((t_0 - \mu_0)'(n^{-1}I + \Sigma_2)^{-1}(t_0 - \mu_0) \geq G_k^{-1}(1 - \gamma)).$$

Note that (3.7) increases to the probability that $(t_0 - \mu_0)'\Sigma_2^{-1}(t_0 - \mu_0) \geq G_k^{-1}(1 - \gamma)$, when $t_0 \sim N_k(\mu_0, \Sigma_1)$, as $n \to \infty$. This probability can be easily computed via simulation.

The following result is proved in Section 3.7.

**Theorem 3.4.** For a sample of $n$ from the statistical model $\{N_k(\mu, I) : \mu \in R^k\}$, a $N_k(\mu_0, \Sigma_2)$ prior is uniformly weakly informative relative to a $N_k(\mu_0, \Sigma_1)$ prior if and only if $\Sigma_2 - \Sigma_1$ is positive semidefinite.

The necessary part of Theorem 3.4 is much more difficult than the $k = 1$ case and shows that we cannot have a $N_k(\mu_0, \Sigma_2)$ prior uniformly weakly informative relative to a $N_k(\mu_0, \Sigma_1)$ prior unless $\Sigma_2 \geq \Sigma_1$. For the choice of $\Sigma_2$ we have that, if $\Sigma_1$ and $\Sigma_2$ are arbitrary $k \times k$ positive definite matrices, then $r\Sigma_2 \geq \Sigma_1$ whenever $r \geq \lambda_k(\Sigma_2)/\lambda_1(\Sigma_2)$ where $\lambda_i(\Sigma)$ denotes the $i$-th ordered eigenvalue of $\Sigma$. Also, if $\Sigma_i = QD_iQ'$ is the spectral decomposition of $\Sigma_i$, then $\Sigma_2 \geq \Sigma_1$ whenever $\lambda_i(\Sigma_2) \geq \lambda_i(\Sigma_1)$ for $i = 1, \ldots, k$.

Theorem 3.4 can be generalized to a linear regression model given by

$$Y_i = \beta'x_i + e_i$$

where $e_i \sim i.i.d. \ N(0, \sigma^2)$ for known variance $\sigma^2$. Then, for $Y = (Y_1, \ldots, Y_n)'$ and $X = (x_1, \ldots, x_n)'$, $T = (X'X)^{-1}X'Y$ is a minimal sufficient statistic for this model and is a consistent estimator of $\beta$.

Suppose two priors $\Pi_i \sim N(\mu, \Psi_i)$ are placed on the model. Then, the prior predictive distribution of $T$ is $N(\mu, \sigma^2(X'X)^{-1} + \Psi_i)$ and $(m_{iT}(t) \leq m_{iT}(t_0)) = ((t-\mu)'(\sigma^2(X'X)^{-1} + \Psi_i)^{-1}(t-\mu) \geq (t_0 - \mu)'(\sigma^2(X'X)^{-1} + \Psi_i)^{-1}(t_0 - \mu))$. Considering $(T - \mu)'(\sigma^2(X'X)^{-1} + \Psi_i)^{-1}(T - \mu) \sim \chi^2(k)$ under $M_{iT}$ where $k$ is the rank of $X'X$ and $T$ has a constant volume.
distortion, the $P_i$ functions are given by

$$P_i(t_0) = M_{iT}(m_{iT}(t) \leq m_{iT}(t_0)) = M_{iT}(m_{iT}(t) \leq m_{iT}(t_0))$$

$$= M_{iT}((t - \mu)'(\sigma^2(X'X)^{-1} + \Psi_i)^{-1}(t - \mu) \geq (t_0 - \mu)'(\sigma^2(X'X)^{-1} + \Psi_i)^{-1}(t_0 - \mu))$$

which is the tail probability of $\chi^2(k)$ distribution. Thus, prior-data conflict $P$-value is given by

$$M_{1T}(P_2(t) \leq \gamma) = M_{1T}((t - \mu)'(\sigma^2(X'X)^{-1} + \Sigma_2)^{-1}(t - \mu) \geq \chi^2_{1-\gamma}(k)).$$

If we also assume that $\Sigma_2 - \Sigma$ is positive semi-definite, the so are $(\sigma^2(X'X)^{-1} + \Sigma_2) - (\sigma^2(X'X)^{-1} + \Sigma)$ and $(\sigma^2(X'X)^{-1} + \Sigma)^{-1} - (\sigma^2(X'X)^{-1} + \Sigma_2)^{-1}$. Hence we get

$$M_{1T}(P_2(t) \leq \gamma) \leq M_{1T}((t - \mu)'(\sigma^2(X'X)^{-1} + \Sigma)^{-1}(t - \mu) \geq \chi^2_{1-\gamma}(k)) = \gamma.$$ 

Thus, $\Pi_2$ is uniformly weakly informative relative to $\Pi_1$ if $\Sigma_2 - \Sigma_1$ is positive semi-definite.

Also from the same proof to Theorem 3.4, the weak informativity of $\Pi_2$ relative to $\Pi_1$ implies $\Sigma_2 - \Sigma_1$ is positive semi-definite. So we have the following theorem.

**Theorem 3.5.** Place two priors $\Pi_1 \sim N(\mu, \Sigma_1)$ and $\Pi_2 \sim N(\mu, \Sigma_2)$ on the regression model (3.8) with known $\sigma^2$. Then, $\Pi_2$ is uniformly weakly informative relative to $\Pi_1$ if and only if $\Sigma_2 - \Sigma_1$ is positive semi-definite.

As we can see in Theorem 3.5, diffuse priors on linear regression models are more likely to be uniformly weakly informative relative to a prior having small variance. But diffuseness doesn’t guarantee weak informativity, see Section 3.4.3.

### 3.3.2 Comparing a $t$ Prior with a Normal Prior

It is not uncommon to find $t$ priors being substituted for normal priors on location parameters. Suppose $x = (x_1, \ldots, x_n)$ is a sample from a $N(\mu, 1)$ distribution where $\mu$ is unknown. We take $\Pi_1$ to be a $N(\mu_0, \sigma^2_1)$ distribution and $\Pi_2$ to be a $t_1(\mu_0, \sigma^2_2, \lambda)$
distribution, i.e., $t_1(\mu_0, \sigma_2^2, \lambda)$ denotes the distribution of $\mu_0 + \sigma_2 z$ with $z$ distributed as a 1-dimensional $t$ distribution with $\lambda$ degrees of freedom. We then want to determine $\sigma_2^2$ and $\lambda$ so that the $t_1(\mu_0, \sigma_2^2, \lambda)$ prior is weakly informative relative to the normal prior.

We consider first the limiting case as $n \to \infty$. The limiting prior predictive distribution of the minimal sufficient statistic $T(x) = \bar{x}$ is $N(\mu_0, \sigma_2^2)$ while $P_2(t_0)$ converges in distribution to $1 - H_{1,\lambda}((t_0 - \mu_0)^2/\sigma_2^2)$ where $H_{1,\lambda}$ is the distribution function of an $F_{1,\lambda}$ distribution. This implies that (3.4) converges to $1 - G_1^{-1}(1 - \gamma)/H_{1,\lambda}^{-1}(1 - \gamma)$ and this is less than or equal to $\gamma$ if and only if $\sigma_2^2/\sigma_1^2 \geq G_1^{-1}(1 - \gamma)/H_{1,\lambda}^{-1}(1 - \gamma)$. So to have that $\Pi_2$ is asymptotically weakly informative relative to $\Pi_1$ at level $\gamma$ we must choose $\sigma_2$ large enough. Clearly we have that $\Pi_2$ is asymptotically uniformly weakly informative relative to $\Pi_1$ if and only if

$$\sigma_2^2/\sigma_1^2 \geq \sup_{\gamma \in [0,1]} G_1^{-1}(1 - \gamma)/H_{1,\lambda}^{-1}(1 - \gamma).$$

(3.9)

It is evaluated at some $\lambda$ values in Table 3.1.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>0.5</th>
<th>1</th>
<th>3</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sup_{\gamma \in [0,1]} G_1^{-1}(1 - \gamma)/H_{1,\lambda}^{-1}(1 - \gamma)$</td>
<td>0.4569</td>
<td>0.6366</td>
<td>0.8488</td>
<td>0.9950</td>
</tr>
</tbody>
</table>

Table 3.1: Evaluations of (3.9) for some $\lambda$ values.

We see that for a Cauchy prior we need to have $\sigma_2^2 \geq \sigma_1^2(0.6366)$ for this prior to be uniformly weakly informative with respect to a $N(\mu_0, \sigma_1^2)$ prior. When we have a $t_1(\mu_0, \sigma_2^2, 3)$ prior, then this has variance $3\sigma_2^2$ and, if we choose $\sigma_2$ so that this prior also has variance $\sigma_1^2$, then $\sigma_2^2/\sigma_1^2 = 1/3$ and this is less than 0.8488 and so is not uniformly weakly informative. So a $t_1(\mu_0, \sigma_2^2, 3)$ prior has to have variance at least equal to $(2.5464)\sigma_1^2$ if we want it to be uniformly weakly informative relative to a $N(\mu_0, \sigma_1^2)$ prior. This is somewhat surprising and undoubtedly is caused by the peakedness of the $t$ distribution. Note that $\sup_{\gamma \in [0,1]} G_1^{-1}(1 - \gamma)/H_{1,\lambda}^{-1}(1 - \gamma) \to 1$ as $\lambda \to \infty$ so this increase in variance, for the $t$ prior over the normal prior, decreases as we increase the degrees of freedom.
The situation for finite \( n \) is covered by the following results proved in Section 3.7.

**Theorem 3.6.** For a sample of \( n \) from the statistical model \( \{N(\mu, 1) : \mu \in R^1\} \), a \( t_1(\mu_0, \sigma_2^2, \lambda) \) prior is uniformly weakly informative relative to a \( N_1(\mu_0, \sigma_1^2) \) prior whenever \( \sigma_2^2 \geq \sigma_{0n}^2 \), where \( \sigma_{0n}^2 \) is the unique solution of \( (1/n + \sigma_2^2)^{-1/2} = \int_0^\infty (1/n + \sigma_{0n}^2/u)^{-1/2}k_\lambda(u) \) \( du \) with \( k_\lambda \) the Gamma rate \((\lambda/2, \lambda/2)\) density. Further, \( \sigma_{0n}^2/\sigma_1^2 \) increases to

\[
\sup_{\gamma \in [0,1]} \frac{G_1^{-1}(1 - \gamma)}{H_{1,\lambda}^{-1}(1 - \gamma)} = \frac{2 [\Gamma((\lambda + 1)/2)]^2}{\lambda [\Gamma(\lambda/2)]^2}
\]

(3.10)
as \( n \to \infty \) and so a \( t_1(\mu_0, \sigma_2^2, \lambda) \) prior is asymptotically uniformly weakly informative if and only if \( \sigma_2^2/\sigma_1^2 \) is greater than or equal to (3.10).

Theorem 3.6 establishes that we can conservatively use (3.10) to select a uniformly weakly informative \( t \) prior.

In Figure 3.1 we have plotted the value of (3.4), that arises with various \( t_1(0, \sigma_2^2, 3) \) priors where \( \sigma_2^2 \) is chosen in a variety of ways together with the 45-degree line. A uniformly weakly informative prior will have (3.4) always below the 45-degree line, while uniformly weakly informative prior at level \( \gamma_0 \) will have (3.4) below the 45-degree line to the left of \( \gamma_0 \) and possibly above to the right of \( \gamma_0 \). For example, when \( \sigma_2^2 = 1/3 \) then the \( t_1(0, \sigma_2^2, 3) \) prior and the \( N(0, 1) \) prior have the same variance. We see that this prior is only uniformly weakly informative at level \( \gamma_0 = 0.0357 \) and is not uniformly weakly informative.

Note that (3.5) converges to \( 1 - G_1((\sigma_2^2/\sigma_1^2)H_{1,\lambda}^{-1}(1 - \gamma))/\gamma \) as \( n \to \infty \), and setting this equal to \( p \), implies that \( \sigma_2^2 = \sigma_1^2G_1^{-1}(1 - \gamma + \gamma p)/H_{1,\lambda}^{-1}(1 - \gamma) \) which converges, as \( \lambda \to \infty \), to the result we obtained in Section 3.3.1. So when \( \lambda = 3, \gamma = 0.05 \) and \( p = 0.5 \) we must have \( \sigma_2^2/\sigma_1^2 = 5.0239/10.1280 = 0.49604 \).

Our analysis indicates that one has to be careful about the scaling of the \( t \) prior if we want to say that the \( t \) prior is less informative than a normal prior, at least when we want uniform weak informativity. This is undoubtedly due to the peakedness of the
Chapter 3. Weakly Informative Priors

Figure 3.1: Plot of (3.4) versus $\gamma$ for $t_1(0, \sigma^2, 3)$ priors relative to a $N(0, 1)$ prior when $n = 20$ where $\sigma^2$ is chosen to match variances (thick solid line), match the MAD (dashed line), just achieve uniform weak informativity (dotted line), just achieve asymptotic uniform weak informativity (dash-dot line), and equal to 1 (long-dashed line).

t prior as we are putting more prior probability near the center than with a similarly scaled normal prior.

Consider now comparing a multivariate $t$ prior to a multivariate normal prior. Let $t_k(\mu_0, \Sigma_2, \lambda)$ denote the $k$-dimensional $t$ distribution given by $\mu_0 + z/\sqrt{u/\lambda}$, where $z \sim N_k(0, \Sigma_2)$ and independently $u \sim \chi^2(\lambda)$. This is somewhat more complicated than the normal case but we have proved the following result in Section 3.7 which provides sufficient conditions for the asymptotic uniform weak informativity.

**Theorem 3.7.** When sampling from the statistical model $\{N_k(\mu, I) : \mu \in R^k\}$, a $t_k(\mu_0, \Sigma_2, \lambda)$ prior is asymptotically uniformly weakly informative relative to a $N_k(\mu_0, \Sigma_1)$ prior whenever $\Sigma_2 - \tau^2_\lambda \Sigma_1$ is positive semidefinite, where $\tau^k_\lambda = (2/\lambda)^k/2\Gamma((k+\lambda)/2)/\Gamma(\lambda/2)$.

For the choice of $\Sigma_2$ we have that, if $\Sigma_1$ and $\Sigma_2$ are arbitrary $k \times k$ positive definite matrices, then $r\Sigma_2 \geq \tau^2_\lambda \Sigma_1$ whenever $r \geq \tau^2_\lambda \lambda_k(\Sigma_2)/\lambda_1(\Sigma_2)$ where $\lambda_i(\Sigma)$ denotes the $i$-th ordered eigenvalue of $\Sigma$. Also, if $\Sigma_i = QD_iQ'$ is the spectral decomposition of $\Sigma_i$, then $\Sigma_2 \geq \Sigma_1$ whenever $\lambda_i(\Sigma_2) \geq \tau^2_\lambda \lambda_i(\Sigma_1)$ for $i = 1, \ldots, k.$
3.3.3 Comparing Inverse Gamma Priors

Suppose now that we have a sample \( x = (x_1, \ldots, x_n) \) from a \( N(0, \sigma^2) \) distribution where \( \sigma^2 \) is unknown. Then \( t = T(x) = (x_1^2 + \cdots + x_n^2)/n \) is minimal sufficient and \( T \sim \Gamma(n/2, n/2\sigma^2) \). Now suppose that we take \( \Pi_i \) to be an inverse gamma prior on \( \sigma^2 \), namely, \( \sigma^{-2} \sim \Gamma_i(\alpha_i, \beta_i) \). From this we get that \( \alpha_i T/\beta_i \sim F(n, 2\alpha_i) \) and, since \( J_T(x) = (4x^2/n)^{-1/2} = (4t/n)^{-1/2} \), \( m_{iT,n}^*(t) = m_{iT,n}(t)(4t/n)^{1/2} \propto t^{(n-1)/2}(1 + nt/2\beta_i)^{-n/2-\alpha_i} \), which implies

\[
P_{1,n}(t_0) = M_{iT,n}(t^{(n-1)/2}(1 + nt/2\beta_i)^{-n/2-\alpha_i}) \leq t_0^{(n-1)/2}(1 + nt_0/2\beta_i)^{-n/2-\alpha_i}.
\]

We want to investigate the weak informativity of a \( \Gamma_i(\alpha_2, \beta_2) \) prior relative to a \( \Gamma_i(\alpha_1, \beta_1) \) prior. For finite \( n \) this is a very difficult problem so we simplify this by considering only the asymptotic case. When the prior is \( \Pi_i \), then, as \( n \to \infty \), we have that \( m_{iT,n}(t) \to m_{iT}(t) = (\beta_i^{\alpha_i}/\Gamma(\alpha_i))t^{-\alpha_i-1}e^{-\beta_i/t} \), i.e., \( 1/t \sim \Gamma_i(\alpha_i, \beta_i) \) in the limit. Therefore, \( P_{2,n}(t_0) \to P_{2}(t_0^*) = \Pi_2(t^{-\alpha_2-1/2}e^{-\beta_2/t} \leq t_0^{-\alpha_2-1/2}e^{-\beta_2/t_0}) \) and we want to determine conditions on \( (\alpha_2, \beta_2) \) so that \( \Pi_1(P_2(t) \leq \gamma) \leq \gamma \).

While results can be obtained for this problem it is still rather difficult. It is greatly simplified, however, if we impose a natural restriction on \( (\alpha_2, \beta_2) \). In particular, we want the location of the bulk of the mass for \( \Pi_2 \) to be located roughly in the same place as the bulk of the mass for \( \Pi_1 \). Accordingly, we could require the priors to have the same means or modes but, as it turns out, the constraint that requires the modes of the \( m_{iT}^* \) functions to be the same greatly simplifies the analysis. Actually \( m_{iT,n}^*(t) \) converges to 0 but the \( n \)'s cancel in the inequalities defining \( P_{1,n}(t_0) \) and so we can define \( m_{iT,n}^*(t) = t^{-\alpha_i-1/2}e^{-\beta_i/t} \) which has its mode at \( t = \beta_i/(\alpha_i + 1/2) \). Therefore, we must have \( \beta_2/(\alpha_2 + 1/2) = \beta_1/(\alpha_1 + 1/2) \) so that \( (\alpha_2, \beta_2) \) lies on the line through the points \( (0, \beta_1/2(\alpha_1 + 1/2)) \) and \( (\alpha_1, \beta_1) \). We prove the following result in Section 3.7.

**Theorem 3.8.** Suppose we use a \( \Gamma_i(\alpha_1, \beta_1) \) prior on \( 1/\sigma^2 \) when sampling from the statistical model \( \{N(0, \sigma^2) : \sigma^2 > 0\} \). Then a \( \Gamma_i(\alpha_2, \beta_2) \) prior on \( 1/\sigma^2 \),
Chapter 3. Weakly Informative Priors

Figure 3.2: Weakly informative Gamma\(_{\text{rate}}(\alpha, \beta)\) priors (dark black region) relative to Gamma\(_{\text{rate}}(2, 2)\) at level 0.05 and uniformly weakly informative Gamma\(_{\text{rate}}(\alpha, \beta)\) priors (light green region) relative to Gamma\(_{\text{rate}}(2, 2)\) when (a) \(n = 20\), (b) \(n = 100\) and (c) \(n = 1000\).

with \(\beta_2/(\alpha_2 + 1/2) = \beta_1/(\alpha_1 + 1/2)\), is asymptotically weakly informative relative to the Gamma\(_{\text{rate}}(\alpha_1, \beta_1)\) prior whenever \(\alpha_2 \leq \alpha_1\) and \(\beta_2 = \beta_1(\alpha_2 + 1/2)/(\alpha_1 + 1/2)\) or, equivalently, whenever \(\beta_1/2(\alpha_1 + 1/2) \leq \beta_2 \leq \beta_1\) and \(\alpha_2 = (\alpha_1 + 1/2)\beta_2/\beta_1 - 1/2\).

Of particular interest here is that we cannot reduce the rate parameter \(\beta_2\) arbitrarily close to 0 and be guaranteed asymptotic weak informativity. With a long proof in Section 3.7, Theorem 3.8 can be extended to next theorem.

**Theorem 3.9.** Consider the same model and priors in Theorem 3.8. If hyperparameters \(\alpha_1, \beta_1, \alpha_2, \text{ and } \beta_2\) satisfy (i) \((\Gamma(\alpha_2)/\Gamma(\alpha_1))(\beta_1/\beta_2)^{\alpha_1}(\alpha_2 + 1/2)^{\alpha_1 - \alpha_2}\exp((1 - \beta_1/\beta_2)(\alpha_2 + 1/2)) \geq 1\), (ii) \(\alpha_2 \leq \alpha_1\), (iii) \(\beta \leq \beta_0(1 + 1/(2\alpha))\) and (iv) \(0 \leq \eta \leq 1/2\) for \(\eta = (1 - \beta_1/\beta_2)(\alpha_2 + 1/2) + (\alpha_1 - \alpha_2) \geq 0\), then \(\Pi_2\) is uniformly weakly informative relative to \(\Pi_1\) asymptotically.

The first condition is concerned with the weak informativity at level around 1. It compares the prior predictive densities at the mode of \(m^*_T\). Also (ii) and (iii) assure the weak informativity at level \(\gamma\) around 0. Theorem 3.8 is a special case of Theorem 3.9 with \(\eta = 0\). Figure 3.2 shows weakly informative Gamma\(_{\text{rate}}(\alpha, \beta)\) priors (dark black region)
relative to $\Pi_1 \sim \text{Gamma}_{\text{rate}}(2,2)$ at level $\gamma = 0.05$ and uniformly weakly informative priors (light green region) when $n = 20, 100$ and $1000$. As the sample size increases, the regions of weakly informative priors shrink.

## 3.4 Applications

We consider now some applications of determining weakly informative priors.

### 3.4.1 Weakly Informative Beta Priors for the Binomial

Suppose that $T \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$. This implies that $m_T(t) = \binom{n}{t} \Gamma(\alpha + \beta)\Gamma(t + \alpha)\Gamma(n - t + \beta) / \Gamma(\alpha)\Gamma(\beta)\Gamma(n + \alpha + \beta)$ and from this we can compute (3.4) for various choices of $(\alpha, \beta)$.

![Figure 3.3: Plot of $(\alpha, \beta)$ corresponding to weakly informative priors at level $\gamma = 0.05$ (light and dark shading) and all $(\alpha, \beta)$ corresponding to uniformly weakly informative priors (light shading) for $n = 20$, $n = 100$ (middle), and $n = \infty$ (on the right).](image)

Figure 3.3: Plot of $(\alpha, \beta)$ corresponding to weakly informative priors at level $\gamma = 0.05$ (light and dark shading) and all $(\alpha, \beta)$ corresponding to uniformly weakly informative priors (light shading) for $n = 20$, $n = 100$ (middle), and $n = \infty$ (on the right).

As a specific example, suppose that $n = 20$, the base prior is given by $(\alpha, \beta) = (6, 6)$, and we take $\gamma = 0.05$ so that $x_{0.05} = 0.0588$. As alternatives to this base prior, we consider Beta$(\alpha, \beta)$ priors. In Figure 3.3 we have plotted all the $(\alpha, \beta)$ corresponding to Beta$(\alpha, \beta)$ distributions that are weakly informative with respect to the Beta$(6, 6)$ distribution at level 0.05, together with the subset of all $(\alpha, \beta)$ corresponding to Beta$(\alpha, \beta)$ distributions.
that are uniformly weakly informative relative to the Beta(6,6) distribution. The graph on the left corresponds to \( n = 20 \), the middle graph corresponds to \( n = 100 \), and the graph on the right corresponds to \( n = \infty \). The plot for \( n = 20 \) shows some anomalous effects due to the discreteness of the prior predictive distributions and these effects disappear as \( n \) increases. In such an application we may choose to restrict to symmetric priors as this fixes the primary location of the prior mass. For example, when \( n = 20 \), a Beta(\( \alpha, \alpha \)) prior for \( \alpha \) satisfying \( 1 \leq \alpha \leq 12.3639 \) is uniformly weakly informative with respect to the Beta(6,6) prior and we see that values of \( \alpha > 6 \) are eliminated as \( n \) increases.

3.4.2 Weakly Informative Priors for a Location-Scale Model

Suppose that \( x = (x_1, \ldots, x_n) \) is a sample from a \( N(\mu, \sigma^2) \) distribution where \( \mu \in \mathbb{R}^1 \) and \( \sigma^2 > 0 \) are unknown. Suppose we have a elicited conjugate prior on \( (\mu, \sigma^2) \), namely, \( \mu | \sigma^2 \sim N(\mu_1, \tau_1 \sigma^2) \) and \( 1/\sigma^2 \sim \text{Gamma}_{\text{rate}}(\alpha_1, \beta_1) \), and we would prefer to use a prior that is asymptotically uniformly weakly informative relative to this choice. As discussed in Evans and Moshonov (2006, 2007) it seems that the most sensible way to check for prior-data conflict here is to first check the prior on \( 1/\sigma^2 \), based on the prior predictive distribution of \( s^2 \), and, if no prior-data conflict is found at this stage, then check the prior on \( \mu \) based on conditional prior predictive for \( \bar{x} \) given \( s^2 \), as \( s^2 \) is ancillary for \( \mu \).

We have that \( s^2 | \sigma^2 \sim \text{Gamma}_{\text{rate}}((n - 1)/2, (n - 1)/2\sigma^2) \) and so, as in Section 3.3.3, when \( 1/\sigma^2 \sim \text{Gamma}_{\text{rate}}(\alpha_i, \beta_i) \) the limiting prior predictive distribution of \( 1/s^2 \) is \( \text{Gamma}_{\text{rate}}(\alpha_i, \beta_i) \) as \( n \to \infty \). Furthermore, when \( T(x) = s^2 \), then \( J_T(x) = (4s^2/(n - 1))^{-1/2} \). Therefore, the limiting value of (3.4) in this case is the same as that discussed in Section 3.3.3 and Theorem 3.8 applies to obtain a \( \text{Gamma}_{\text{rate}}(\alpha_2, \beta_2) \) prior uniformly weakly informative relative to the \( \text{Gamma}_{\text{rate}}(\alpha_1, \beta_1) \) prior.

If we consider \( s^2 \) as an arbitrary fixed value from its prior predictive distribution, then the conditional prior predictive distribution of \( \bar{x} | s^2 \) converges to the \( N(\mu_1, \tau_1 s^2) \) distribution. Then by the results in Section 3.3.1 the \( N(\mu_1, \tau_2 \sigma^2) \) prior is asymptotically weakly
informative at level $\gamma$ relative to the $N(\mu_1, \tau_1 \sigma^2)$ prior whenever $1 - G_1((\tau_2 s^2 / \tau_1 s^2)G_1^{-1}(1 - \gamma)) = 1 - G_1((\tau_2 / \tau_1)G_1^{-1}(1 - \gamma)) \leq \gamma$ and this occurs if and only if $\tau_2 \geq \tau_1$. Note that the dependence on the unknown value of $s^2$ disappears. Furthermore, the $N(\mu_1, \tau_2 \sigma^2)$ prior is asymptotically uniformly weakly informative with respect to the $N(\mu_1, \tau_1 \sigma^2)$ if and only if $\tau_2 \geq \tau_1$. We can determine an appropriate value for $\tau_2$ by specifying $p \in [0, 1]$ and then setting $\tau_2 = \tau_1 G_1^{-1}(1 - \gamma + p\gamma)/G_1^{-1}(1 - \gamma)$.

While this analysis is for a normal location-scale model, it is easy to see that the analysis for a general normal linear model will proceed along similar lines.

### 3.4.3 Weakly Informative Priors for Logistic Regression

Suppose we have a single binary valued response variable $Y$ and $k$ quantitative predictors $X_1, \ldots, X_k$, we observe $(Y, X_1, \ldots, X_k)$ at $q$ settings of the predictor variables and have $n_i$ observations at the $i$-th setting of the predictors. The logistic regression model then says that $Y_{ij} \sim \text{Bernoulli}(p_i)$ where $\log(p_i/(1 - p_i)) = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_k(x_{ik} - \bar{x}_k)$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, q$ and the $\beta_i$ are unknown real values. For this model $T = (T_1, \ldots, T_q)$, with $T_i = Y_{i1} + \cdots + Y_{in_i}$, is a minimal sufficient statistic. For the base prior we suppose that $\Pi_1$ is the product of independent priors on the $\beta_i$’s and we consider the problem of finding a prior $\Pi_2$ that is weakly informative relative to $\Pi_1$. For example, we could take $\Pi_1$ to be a product of $N(0, \sigma_{1i}^2)$ priors and $\Pi_2$ to be a product of $N(0, \sigma_{2i}^2)$ priors and choose the $\sigma_{2i}^2$ so that weak informativity is obtained. Note that since $T$ is discrete we can use (3.2) in our computations.

As we will see, it is not the case that choosing the $\sigma_{2i}^2$ very large relative to the $\sigma_{1i}^2$ will necessarily make $\Pi_2$ weakly informative relative to $\Pi_1$. In fact there is only a finite range of $\sigma_{2i}^2$ values where weak informativity will be obtained.

While this can be demonstrated analytically, the argument is somewhat technical and it is perhaps easier to see this in an example. The following bioassay data are from Racine et al. (1986) and were also analyzed in Gelman et al. (2008). These data arise from an
experiment where 20 animals were exposed to four doses of a toxin and the number of deaths recorded.

<table>
<thead>
<tr>
<th>Dose (g/ml)</th>
<th>Number of animals $n_i$</th>
<th>Number of deaths $t_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.422</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>0.744</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>0.948</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>2.069</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3.2: Bioassay data from Racine et al. (1986).

Following Gelman et al. (2008) we took $X_1$ to be the variable formed by calculating the logarithm of dose and then standardizing to make the mean of $X_1$ equal to 0 and its standard deviation equal to 1/2. Gelman et al. (2008) placed independent Cauchy priors scales on the regression coefficients, namely, $\beta_0 \sim t_1(0, 10^2, 1)$ independent of $\beta_1 \sim t_1(0, 2.5^2, 1)$.

We consider four possible scenarios for the investigation of weak informativity at level $\gamma = 0.05$ and uniform weak informativity. In Figure 3.4 (a) we compare $\Pi_2 = N(0, \sigma_0^2) \times N(0, \sigma_1^2)$ priors with the prior $\Pi_1 = N(0, 10^2) \times N(0, 2.5^2)$. The entire region gives the $(\sigma_0, \sigma_1)$ values corresponding to priors that are weakly informative at level $\gamma = 0.05$ while the lighter subregion gives the $(\sigma_0, \sigma_1)$ values corresponding to priors that are uniformly weakly informative. Note that some of the irregularity in the plots is caused by the fact that the prior predictive distributions of $T$ are discrete. The three remaining plots are similar where, in Figure 3.4(b) $\Pi_1 = t_1(0, 10^2, 1) \times t_1(0, 2.5^2, 1)$ and $\Pi_2 = t_1(0, \sigma_0^2, 1) \times t_1(0, \sigma_1^2, 1)$, in Figure 3.4(c) $\Pi_1 = N(0, 10^2) \times N(0, 2.5^2)$ and $\Pi_2 = t_1(0, \sigma_0^2, 1) \times t_1(0, \sigma_1^2, 1)$, and in Figure 3.4(d) $\Pi_1 = t_1(0, 10^2, 1) \times t_1(0, 2.5^2, 1)$ and $\Pi_2 = N(0, \sigma_0^2) \times N(0, \sigma_1^2)$. Note that these plots only depend on the data through the values of $X_1$.

We see clearly from these plots that increasing the scaling on any of the $\beta_i$ does not
necessarily lead to weak informativity and in fact inevitably destroys it. Furthermore, a smaller scaling on a parameter can lead to uniform weak informativity. These plots underscore how our intuition does not work very well with the logistic regression model as it is not clear how priors on the $\beta_i$ ultimately translate to priors on the $p_i$. In fact it can be proven that, if we put independent priors on the $\beta_i$, fix all the scalings but one, say $\Pi_{20}$ on $\beta_0$, and let that one grow arbitrarily large, then the prior predictive distribution of $T$ converges to a distribution degenerate on the event $\{\sum_{i=1}^{q} T_i = 0\} \cup \{\sum_{i=1}^{q} T_i = \sum_{i=1}^{q} n_i\}$, and this is definitely not desirable. This partially explains the results obtained. Specifics of this phenomena are in Proposition 3.10 proven in Section 3.7.

**Proposition 3.10.** Consider the logistic model described in Section 3.4.3. Assume that no $x_{ij} - \bar{x}_{oj}$ is 0 for $i = 1, \ldots, q$ and $j = 1, \ldots, k$. If independent priors $\Pi_{2j}$ on $\beta_j$ for $i = 1, 2$, $j = 1, \ldots, k$ have fixed scalings except one, say $\Pi_{2j}$, and let that one grow
arbitrarily large, then the prior predictive distribution of \( T \) converges to a distribution degenerate on the event \( \{ \sum_{i=1}^{q} T_i I(x_{ij} - \bar{x}_j > 0) = 0 \text{ or } = \sum_{i=1}^{q} n_i \} \). Hence, \( \Pi_2 \) is not uniformly weakly informative relative to \( \Pi_1 \).

Of some interest is how much reduction we actually get, via (3.5), when we employ a weakly informative prior. In Figure 3.5 we have plotted contours of the choices of \((\sigma_0, \sigma_1)\) that give 0%, 25%, 50% and 75% reduction in prior-data conflicts for the case where \( \Pi_2 = N(0, \sigma_0^2) \times N(0, \sigma_1^2) \) and \( \Pi_1 = N(0, 10^2) \times N(0, 2.5^2) \) when \( \gamma = 0.05 \) (this corresponds to \( x_\gamma = 0.0503 \)). Note that a substantial reduction can be obtained.

![Figure 3.5](image)

**Figure 3.5:** Reduction levels of \( N(0, \sigma_0^2) \times N(0, \sigma_1^2) \) priors relative to \( N(0, 10^2) \times N(0, 2.5^2) \) prior using (3.5) when \( \gamma = 0.05 \). (a) The plotted reduction levels are 0% (solid line), 25% (dashed line), 50% (dotted line) and 75% (long dashed line); (b) \( \sigma_0 = 2.5 \) is fixed; and (c) \( \sigma_1 = 2.5 \) is fixed.

We could consider fixing one of the scalings and seeing how much reduction we obtain when varying the other. For example, when we fix \( \sigma_0 = 2.5 \) we find that the maximum reduction is obtained when \( \sigma_1 \) is close to 2.2628 while if we fix \( \sigma_1 = 2.5 \) then the maximum reduction is obtained when \( \sigma_0 \) is close to 0.875.

It makes sense in any application to check to see if any prior-data conflict exists with respect to the base prior. If there is no prior-data conflict this increases our confidence that the weakly informative prior is indeed putting less information into the analysis. This is assessed generally using (3.3) although (3.2) suffices in this example. When \( \Pi_1 = \)
Chapter 3. Weakly Informative Priors

The approach in Section 3.2 works whenever $T$ is a complete minimal sufficient statistic. This is a consequence of Basu’s Theorem as, in such a case, any ancillary is statistically independent of $T$ and so conditioning on such an ancillary is irrelevant. When $U(T)$ is a meaningful ancillary, however, then the variation due to $U(T)$ is independent of $\theta$ and so should be removed from the $P$-value (3.3) when checking for prior-data conflict. Removing this variation is equivalent to conditioning on $U(T)$ and so we replace (3.3) by

$$M_T(m^*_T(t) \leq m^*_T(t_0) \mid U(T)), \tag{3.11}$$

i.e., we use the conditional prior predictive given the ancillary $U(T)$. To remove the maximal amount of ancillary variation we must have that $U(T)$ is a maximal ancillary. Therefore (3.4) becomes

$$M_{1T}(P_2(t_0 \mid U(T)) \leq x_\gamma \mid U(T)), \tag{3.12}$$

i.e., we have replaced $M_{1T}$ by $M_{1T}(\cdot \mid U(T))$ and $P_2(t_0)$ by $P_2(t_0 \mid U(T)) = M_{2T}(m^*_T(t) \leq m^*_T(t_0) \mid U(T))$.

One problem with ancillaries is that multiple maximal ancillaries may exist. When ancillaries are used for frequentist inferences about $\theta$ via conditioning, this poses a problem because it is not clear which multiple ancillary to use and confidence regions depend on the maximal ancillary chosen. For checking for prior-data conflict via (3.11), however, this does not pose a problem. This is because we simply get different checks depending on which maximal ancillary we condition on. For example, if conditioning on maximal ancillary $U_1(T)$ does not lead to prior-data conflict, but conditioning on maximal ancillary
$U_2(T)$ does, then we have evidence against no prior-data conflict existing.

Similarly, when we go to use (3.12), we can also simply look at the effect of each maximal ancillary on the analysis and make our assessment about $\Pi_2$ based on this. For example, we can use the maximum value of (3.12) over all maximal ancillaries to assess whether or not $\Pi_2$ is weakly informative relative to $\Pi_1$. When this maximum is small, we conclude that we have a small prior probability of finding evidence against the null hypothesis of no prior-data conflict when using $\Pi_2$. We illustrate this via an example.

**Example 3.1.** Suppose that we have a sample of $n$ from the Multinomial\((1, (1-\theta)/6, (1+\theta)/6, (2-\theta)/6, (2+\theta)/6)\) distribution where $\theta \in [-1, 1]$ is unknown. Then the counts $(f_1, f_2, f_3, f_4)$ constitute a minimal sufficient statistic and $U_1 = (f_1 + f_2, f_3 + f_4)$ is ancillary as is $U_2 = (f_1 + f_4, f_2 + f_3)$. Then $T = (f_1, f_2, f_3, f_4) | U_1$ is given by $f_1 | U_1 \sim$ Binomial$(f_1 + f_2, (1-\theta)/2)$ independent of $f_3 | U_1 \sim$ Binomial$(f_3 + f_4, (2-\theta)/4)$ giving

$$m_T(f_1, f_2, f_3, f_4 | U_1) = \left(\frac{f_1 + f_2}{f_1}\right)\left(\frac{f_3 + f_4}{f_3}\right) \times \int_{-1}^{1} \left(\frac{1 - \theta}{2}\right)^{f_1} \left(\frac{1 + \theta}{2}\right)^{f_2} \left(\frac{2 - \theta}{4}\right)^{f_3} \left(\frac{2 + \theta}{4}\right)^{f_4} \pi(\theta) d\theta.$$  

We then have two 1-dimensional distributions $f_1 | U_1$ and $f_3 | U_1$ to use for checking for prior-data conflict. A similar result holds for the conditional distribution given $U_2$.

![Figure 3.6: Plot of all ($\alpha, \beta$) corresponding to weakly informative priors at level $\gamma = 0.05$ (light and dark shading) and uniformly weakly informative priors (light shading) in Example 3.1.](image)
For example, suppose $\pi$ is a Beta(20, 20) distribution on $[-1, 1]$, so the prior concentrates about 0, and for a sample of $n = 18$ we have that $U_1 = f_1 + f_2 = 10$ and $U_2 = f_1 + f_4 = 8$. In Figure 3.6 we have plotted all the values of $(\alpha, \beta)$ that correspond to a Beta$(\alpha, \beta)$ prior that is weakly informative relative to the Beta(20, 20) prior at level $\gamma = 0.05$ as well as those that are uniformly weakly informative. So for each such $(\alpha, \beta)$ we have that (3.12) is less than or equal to 0.05 for both $U = U_1$ and $U = U_2$.

### 3.6 Conclusions

We have developed an approach to measuring the amount of information a prior puts into a statistical analysis relative to another base prior. This base prior can be considered as the prior that best reflects current information and our goal is to determine a prior that is weakly informative with respect to it. Our measure is in terms of the prior predictive probability, using the base prior, of obtaining a prior-data conflict. This was applied in several examples where the approach is seen to give intuitively reasonable results.

As noted in several examples, however, we need to be careful when we conceive of a prior being weakly informative relative to another. Ultimately this concept needs to be made precise and we feel our definition is a reasonable proposal. The definition has intuitive support, in terms of avoiding prior-data conflicts, and provides a quantifiable criterion that can be used to select priors. This entails choosing a $\gamma$ and using (3.5).

In any application we should still check for prior-data conflict for the base prior using (3.3). If prior-data conflict is found, a substitute prior that is weakly informative relative to the base prior can then be selected and a check made for prior-data conflict with respect to the new prior. In this way the new prior still incorporates some of the information from the base prior. If this new prior passes the checks for prior-data conflict, then the same theory can be applied to select a weakly informative prior relative to it, so that our inferences can be described as conservative.
We note that in several of the examples we have discussed, e.g., comparing normal priors, it is the case that a prior-data conflict with respect to the base prior can be avoided by a suitably chosen prior that is weakly informative relative to it. It doesn’t seem possible, however, to prove that we can always find a weakly informative prior in a family that will avoid a prior-data conflict found for the base prior.

### 3.7 Proofs

#### Proof of Lemma 3.2

We have that \( x_\gamma = \gamma \) since \( P_1(t) \) has a continuous distribution under \( M_{1T} \). Suppose \( m_{1T}^*(t) \) has a point mass at \( r_0 \) when \( t \sim M_{1T} \). The assumption \( M_{1T}(m_{1T}^*(t) = r_0) > 0 \) implies \( (m_{1T}^*)^{-1}\{r_0\} \neq \emptyset \). Then, pick \( t_{r_0} \in (m_{1T}^*)^{-1}\{r_0\} \) so that \( m_{1T}^*(t_{r_0}) = r_0 \) and let \( \eta_i = P_i(t_{r_0}) \). Then, \( P_i(t_{r_0}) \) has point mass at \( \eta_i \) because \( M_{1T}(P_i(t) = \eta_i) \geq M_{1T}(m_{1T}^*(t) = m_{1T}(m_{1T}^*(t) = r_0) > 0 \). This is a contradiction and so \( m_{1T}^*(t) \) has a continuous distribution when \( t \sim M_{1T} \).

Let \( r_\gamma = \sup\{r \in \mathcal{R} : M_{2T}(m_{2T}^*(t) \leq r) \leq \gamma \} \) where \( \mathcal{R} = \{m_{2T}^*(t) : t \in \mathcal{T} \} \) and \( \mathcal{T} \) is the range space of \( T \). Then, \( M_{2T}(m_{2T}^*(t) \leq r_\gamma) = \gamma \) and \( M_{2T}(m_{2T}^*(t) \leq r_\gamma + \epsilon) > \gamma \) for all \( \epsilon > 0 \). Thus, we have that \( \{t : P_2(t) \leq \gamma\} = \{t : m_{2T}^*(t) \leq r_\gamma\} \), \( M_{1T}(P_2(t) \leq \gamma) = M_{1T}(m_{2T}^*(t) \leq r_\gamma) \), and \( \Pi_2 \) is weakly informative at level \( \gamma \) relative to \( \Pi_1 \) if and only if \( M_{1T}(m_{2T}^*(t) \leq r_\gamma) \leq \gamma \). The fact that \( \{r_\gamma : \gamma \in [0,1]\} \subset \mathcal{R} \) implies the last statement.

#### Proof of Theorem 3.3

The sufficiency of \( T \) implies that \( \mathbb{E}_\theta(J_T(X)^{-1}|T = t) \) does not depend on \( \theta \).

**Lemma 3.11.** \( \int |m_{i,T,n}(t) - \pi_i(t)| \mathcal{L}(dt) \to 0 \) as \( n \to \infty \).

**Proof of Lemma 3.11.** Let \( a_n = (n^{-1} \log n)^{1/2} \). Then, we have
It is easy to show that the first part converges to 0 as \( n \to \infty \), that is,

\[
I_1 = \int f_{\theta,T,n}(t) \pi_i(\theta) I(||t - \theta|| \geq a_n) \mathcal{L}(d\theta) \mathcal{L}(dt)
\]

\[
= \int P_{\theta,T,n}(\sqrt{n}||T - \theta|| \geq (\log n)^{1/2}) \Pi_i(d\theta) \to 0.
\]

The last convergence obtained from \( P_{\theta,T,n}(\sqrt{n}||T - \theta|| \geq (\log n)^{1/2}) \to 0 \) and the bounded convergence theorem. By Lemma A.9 and change of variables, we get

\[
f_{\eta,u,n}(u) = |2\pi \hat{A}(u)/n|^{-1/2} \exp(-2^{-1}n(u - \eta)^T \hat{A}(u)^{-1}(u - \eta) + O(n||u - \eta||^3))
\]

for \( ||u - \eta|| = O(a_n) \). Applying change of variables formula, we obtain

\[
f_{\theta,T,n}(t) = |2\pi \Sigma(t)/n|^{-1/2} \exp(-2^{-1}n(t - \theta)^T \Sigma(t)^{-1}(t - \theta) + O(n||t - \theta||^3))
\]

for \( ||t - \theta|| = O(a_n) \) and \( \Sigma(t) = (\frac{\partial \hat{A}}{\partial u}(u)(\frac{\partial \hat{A}}{\partial u})^T)_{u=\eta(t)} \). Also from the continuity of \( \pi_i \), we get \( |\pi_i(\theta) - \pi_i(t)| = O(||t - \theta||) \). Thus, \( I_3(t) = \int f_{\theta,T,n}(t) \pi_i(\theta) I(||t - \theta|| < a_n) \mathcal{L}(d\theta) \) converges to

\[
I_3(t) = \pi_i(t) \int |2\pi \Sigma(t)/n|^{-1/2} \exp(-2^{-1}n(t - \theta)^T \Sigma(t)^{-1}(t - \theta) + O(n||t - \theta||^3))
\]

\[
\times I(||t - \theta|| < a_n) \mathcal{L}(d\theta)
\]

\[
= \pi_i(t) \exp(O(na_n^3)) \int |2\pi \Sigma(t)|^{-1/2} \exp(-2^{-1}u^T \Sigma(t)^{-1}u) I(||u|| < n^{1/2}a_n) \mathcal{L}(du)
\]

\[
\to \pi_i(t) \quad \text{as} \quad n \to \infty.
\]

Then, the generalized Lebesgue dominated convergence theorem implies

\[
I_2 = \int |I_3(t) - \pi_i(t)| \mathcal{L}(dt) \to 0.
\]

Hence, the lemma holds. \( \square \)
Suppose that $i = j$ or $i = 2, j = 1$ throughout this proof. We define two events $A_{i,n,\epsilon} = \{ t : |m_{iT,n}(t)/V_{i,n}(t) - \pi_i(t)w(t)| > \epsilon \}$ and $B_{i,r} = \{ t : \pi_i(t)w(t) \leq r \}$. Then, we compute upper and lower bounds of $I_{j,i,n,r} = M_{jT,n}(m_{iT,n}(t)/V_{i,n}(t) \leq r)$ and show that the difference of two bounds converges to zero.

By separating $I_{j,i,n,r}$ into two parts, we have an upper bound as follows

$$I_{j,i,n,r} = M_{jT,n}(m_{iT,n}(t)/V_{i,n}(t) \leq r, A_{i,n,\epsilon}) + M_{jT,n}(m_{iT,n}(t)/V_{i,n}(t) \leq r, A_{i,n,\epsilon}^c)$$

$$\leq M_{jT,n}(A_{i,n,\epsilon}) + M_{jT,n}(m_{iT,n}(t)/V_{i,n}(t) \leq r, A_{i,n,\epsilon}^c) \leq M_{jT,n}(A_{i,n,\epsilon}) + J_{j,i,n,r,\epsilon}.$$  

We prove $\Pi_j(A_{i,n,\epsilon}) \rightarrow 0$ using Lemma A.3. Take any subsequence $n_k$. From Lemma 3.11, there exists a further subsequence $n_{k_l}$ such that $m_{iT,n_{k_l}}(t) \rightarrow \pi_i(t)$ a.e.-$\mathcal{L}$ as $l \rightarrow \infty$. Thus, $m_{iT,n_{k_l}}(t)/V_{i,n_{k_l}}(t) \rightarrow \pi_i(t)w(t)$ a.e.-$\mathcal{L}$ as $l \rightarrow \infty$. The convergence also holds in a.e.-$\Pi_j$ sense because of $\Pi_j \ll \mathcal{L}$. Then, $I_{A_{i,n_{k_l}},\epsilon}(t) \rightarrow 0$ a.s.-$\Pi_j$. So the bounded convergence theorem implies

$$\Pi_j(A_{i,n_{k_l}},\epsilon) = \int I_{A_{i,n_{k_l}},\epsilon}(t) \Pi_j(dt) \rightarrow 0 \quad \text{as} \quad l \rightarrow \infty.$$  

Hence, we have $\Pi_j(A_{i,n,\epsilon}) \rightarrow 0$ as $n \rightarrow \infty$ by Lemma A.3. Then, Lemma 3.11 and the convergence $\Pi_j(A_{i,n,\epsilon}) \rightarrow 0$ imply

$$M_{jT,n}(A_{i,n,\epsilon}) = \int I_{A_{i,n,\epsilon}}(t)m_{jT,n}(t) \mathcal{L}(dt) \leq \int I_{A_{i,n,\epsilon}}(t)(\pi_j(t) + |m_{jT,n}(t) - \pi_j(t)|) \mathcal{L}(dt)$$

$$= \Pi_j(A_{i,n,\epsilon}) + \int I_{A_{i,n,\epsilon}}(t)|m_{jT,n}(t) - \pi_j(t)| \mathcal{L}(dt) \rightarrow 0. \quad (3.13)$$

An upper bound of the second term $J_{j,i,n,r,\epsilon}$ is given by

$$J_{j,i,n,r,\epsilon} = M_{jT,n}(m_{iT,n}(t)/V_{i,n}(t) \leq r, A_{i,n,\epsilon}^c) \leq M_{jT,n}(\pi(t)w(t) \leq r + \epsilon, A_{i,n,\epsilon}^c)$$

$$\leq M_{jT,n}(B_{i,r+\epsilon}) = \int I_{B_{i,r+\epsilon}}(t)m_{jT,n}(t) \mathcal{L}(dt)$$

$$\leq \int I_{B_{i,r+\epsilon}}(t)(\pi_j(t) + |m_{jT,n}(t) - \pi_j(t)|) \mathcal{L}(dt)$$

$$\leq \Pi_j(B_{i,r+\epsilon}) + \int |m_{jT,n}(t) - \pi_j(t)| \mathcal{L}(dt).$$
In a similar vein, we have

\[ I_{j,i,n,r} \geq M_{j,T,n}(m_{i,T,n}^*(t)/V_{T,n}(t)) \leq r, A_{i,n,\epsilon} \]

\[ \geq M_{j,T,n}(B_{i,r-\epsilon}, A_{i,n,\epsilon}) \geq M_{j,T,n}(B_{i,r-\epsilon}) - M_{j,T,n}(A_{i,n,\epsilon}) \]

\[ \geq \int I_{B_{i,r-\epsilon}}(\pi_j(t) - |m_{j,T,n}(t) - \pi_j(t)|) \mathcal{L}(dt) - M_{j,T,n}(A_{i,n,\epsilon}). \]

Thus Lemma 3.11 and (3.13) imply

\[ |I_{j,i,n,r} - \Pi_j(B_{i,r})| \leq |\Pi_j(B_{i,r+\epsilon}) - \Pi_j(B_{i,r-\epsilon})| + 2\Pi_j(A_{i,n,\epsilon}) + 2\int |m_{j,T,n}(t) - \pi_j(t)| \mathcal{L}(dt) \]

\[ \rightarrow \Pi_j(|\pi_i(t)w(t) - r| \leq \epsilon) \text{ as } n \rightarrow \infty. \]

Then assumptions (iii) and (iv) imply that \( \Pi_j(|\pi_i(t)w(t) - r| \leq \epsilon) \rightarrow 0 \) as \( \epsilon \rightarrow 0 \). This proves that, for any \( r > 0 \),

\[ M_{j,n,r}(m_{i,T,n}^*(t)/V_{T,n}(t)) \leq r \rightarrow \Pi_j(\pi_i(t)w(t) \leq r) = \Pi_j(B_{i,r}) \text{ as } n \rightarrow \infty. \quad (3.14) \]

Assumption (iii) implies \( P_i(T) \) has a continuous distribution when \( T \sim \Pi_i \). Then, \( P_i(T) \sim \text{Uniform}(0,1) \) when \( T \sim \Pi_i \). Let \( r_{i,2} = \inf\{r : \Pi_i(B_{i,r}) > \gamma\} \) and \( r_{i,1} = \sup\{r : \Pi_i(B_{i,r}) < \gamma\} \) so that \( \Pi_i(B_{i,r_{i,1}+}) < \Pi_i(B_{i,r_{i,1}}) = \gamma < \Pi_i(B_{i,r_{i,2}}) < \Pi_i(B_{i,r_{i,2}+}) \) for all \( \epsilon > 0 \). It is easy to check that \( x_{\gamma,n} \rightarrow \gamma \) because \( I_{1,1,n,r_{i,1}+,2+} \rightarrow \Pi_1(B_{1,r_{i,1}+,2+}) \) as \( n \rightarrow \infty \) and \( \Pi_1(B_{1,r_{i,1}+,2+}) \rightarrow \gamma \) as \( \epsilon \rightarrow 0 \). There is a big number \( N_0 > 0 \) such that

\[ |x_{\gamma,n} - \gamma| < \eta/2, \quad |I_{2,2,n,r_{2,2}+,2+ - \Pi_2(B_{2,r_{2,2}+})}| < \eta/2, \quad |I_{2,2,n,r_{2,2}+,2+ - \Pi_2(B_{2,r_{2,2}+})}| < \eta/2, \quad |I_{2,2,n,r_{2,2}+,2+ - \Pi_1(B_{2,r_{2,2}+})}| < \eta/2 \]

for all \( n \geq N_0 \) where \( \eta = \min(\Pi_1(B_{2,r_{2,2}+}) - \gamma, \gamma - \Pi_1(B_{2,r_{2,1}-})) \). Then, \( I_{2,2,n,r_{2,2}+,2+} > \gamma + \epsilon/2 \) so

\[ \{t : P_{2,n}(t) \leq \gamma + \eta/2\} \subset \{t : m_{2,T,n}^*(t)/V_{T,n}(t) \leq r_{2,2} + \epsilon\}, \]

\[ M_{1,T,n}(P_{2,n}(t) \leq x_{\gamma,n}) \leq M_{1,T,n}(P_{2,n}(t) \leq \gamma + \eta/2) \leq I_{1,2,n,r_{2,2}+,2+} \leq \Pi_1(B_{2,r_{2,2}+}) + \eta/2 \]

and

\[ M_{1,T,n}(P_{2,n}(t) \leq x_{\gamma,n}) \geq M_{1,T,n}(P_{2,n}(t) \leq \gamma - \eta/2) \geq I_{1,2,n,r_{2,1}+,2+} \geq \Pi_1(B_{2,r_{2,1}-}) - \eta/2 \]
Hence, we have
\[
\limsup_{n \to \infty} |M_{1,T,n}(P_{2,n}(t) \leq x_{\gamma,n}) - \Pi_1(P_2(t) \leq \gamma)| \leq \Pi_1(B_{2,r_2,2+\epsilon} - B_{2,r_2,1-\epsilon}) + \eta.
\]
Since \( \eta \to 0 \) and \( \Pi_1(B_{2,r_2,2+\epsilon} - B_{2,r_2,1-\epsilon}) \to 0 \) as \( \epsilon \to 0 \), the theorem follows.

**Proof of Theorem 3.4**

Suppose first that \( \Sigma_1 \leq \Sigma_2 \). We have that \( n^{-1}I + \Sigma_1 \leq n^{-1}I + \Sigma_2 \) and so \( (n^{-1}I + \Sigma_1)^{-1} \geq (n^{-1}I + \Sigma_2)^{-1} \). This implies that (3.7) is less than \( \gamma \) and so the \( N_k(\mu_0, \Sigma_2) \) prior is uniformly weakly informative relative to the \( N_k(\mu_0, \Sigma_1) \) prior.

For the converse put \( V_i = \{ y : y'(n^{-1}I + \Sigma_i)^{-1}y \leq 1 \} \). If \( V_1 \subset V_2 \), then for \( y \in R^k \setminus \{ 0 \} \) there exists \( c > 0 \) such that \( c^2 y'(n^{-1}I + \Sigma_1)^{-1}y = 1 \) which implies \( cy \in V_2 \) and so \( c^2 y'(n^{-1}I + \Sigma_2)^{-1}y \leq 1 \). This implies that \( y'(n^{-1}I + \Sigma_1)^{-1}y \leq y'(n^{-1}I + \Sigma_2)^{-1}y \) and so \( \Sigma_1 \leq \Sigma_2 \) and the result follows. If \( V_2 \subset V_1 \) then the same reasoning says that \( \Sigma_2 \leq \Sigma_1 \) and (3.7) would be greater than \( \gamma \) if \( \Sigma_2 < \Sigma_1 \).

So we need only consider the case where \( V_1 \cap V_2^c, V_1^c \cap V_2 \) both have positive volume, i.e., we are supposing that neither \( \Sigma_2 - \Sigma_1 \) nor \( \Sigma_1 - \Sigma_2 \) is positive semidefinite and then will obtain a contradiction. Let \( \delta = \inf \{ y'(n^{-1}I + \Sigma_1)^{-1}y : y \in V_1 \cap \partial V_2 \} \) and note that \( \delta < 1 \), since \( V_1^\circ \cap \partial V_2 \neq \emptyset \), i.e., there are points in the interior of \( V_1 \) on the boundary of \( V_2 \). Now put \( V_0 = \{ y \in V_1 \cap V_2^c : y'(n^{-1}I + \Sigma_1)^{-1}y \leq (1 + \delta)/2 \} \) and note that \( V_0 \) has positive volume.

Let \( Y \sim N_k(0, n^{-1}I + \Sigma_1) \) and \( \tau_\gamma^2 = G^{-1}_k(1 - \gamma) \). Then \( M_{1T}(P_1(t) \leq \gamma) = P(Y'(n^{-1}I + \Sigma_1)^{-1}Y \geq \tau_\gamma^2) = P(Y \notin \tau_\gamma V_1) = 1 - P_Y(\tau_\gamma(V_1 \cap V_2) \cup \tau_\gamma(V_1 \cap V_2^c)) \) while \( M_{1T}(P_2(t) \leq \gamma) = P(Y'(n^{-1}I + \Sigma_2)^{-1}Y \geq \tau_\gamma^2) = P(Y \notin \tau_\gamma V_2) = 1 - P_Y(\tau_\gamma(V_1 \cap V_2) \cup \tau_\gamma(V_1^c \cap V_2)) \).

Since \( \gamma = M_{1T}(P_1(t) \leq \gamma) \), we need only show that \( P_Y(\tau_\gamma(V_1 \cap V_2^c)) > P_Y(\tau_\gamma(V_1^c \cap V_2)) \) for all \( \gamma \) sufficiently small, to establish the result.

Let \( f(x) = k_1 e^{-x/2} \) be such that \( f(y'(n^{-1}I + \Sigma_1)^{-1}y) \) is the density of \( Y \). Then
\[
P_Y(\tau_\gamma(V_1^c \cap V_2)) = \int_{\tau_\gamma(V_1^c \cap V_2)} f(y'(n^{-1}I + \Sigma_1)^{-1}y) \, dy \leq f(\tau_\gamma^2 y_0'(n^{-1}I + \Sigma_1)^{-1}y_0) \, \text{Vol}(V_1^c \cap V_2)
\]
If $y_* = \arg \min \{ y'(n^{-1}I + \Sigma_1)^{-1}y : y \in V_1^c \cap V_2 \}$. Note it is clear that $y_* \in \partial V_1$ and so $y'_\gamma(n^{-1}I + \Sigma_1)^{-1}y_* = 1$ and $f(\tau_\gamma^2 y'_\gamma(n^{-1}I + \Sigma_1)^{-1}y_*) = k_1 e^{-\tau_\gamma^2/2}$. Also, $P_Y(\tau_\gamma (V_1 \cap V_2^c)) \geq P_Y(\tau_\gamma V_0) = \int_{\tau_\gamma V_0} f(y'(n^{-1}I + \Sigma_1)^{-1}y) dy \geq f(\tau_\gamma^2 (1 + \delta)/2) \text{Vol}(V_0) \tau_\gamma^k$ where $f(\tau_\gamma^2 (1 + \delta)/2) = k_1 e^{-\tau_\gamma^2(1+\delta)/4}$. Therefore, as $\gamma \to 0$,

$$\frac{P_Y(\tau_\gamma (V_1 \cap V_2^c))}{P_Y(\tau_\gamma V_0)} \geq e^{\tau_\gamma^2(1-\delta)/4} \frac{\text{Vol}(V_1^c \cap V_2^c)}{\text{Vol}(V_0)} \to \infty$$

since $\tau_\gamma = (G^{-1}_k(1-\gamma))^{1/2} \to \infty$ as $\gamma \to 0$ and $0 < \delta < 1$.

**Proof of Theorem 3.6**

First note that we can use (3.2) instead of (3.3) in this case as $J_T(x)$ is constant in this case. We assume without loss of generality that $\mu_0 = 0$.

We first establish several useful technical results. If $\Pi_i$ is a probability distribution that is unimodal and symmetric about 0, and $\phi_\nu$ denotes a $N(0, \nu)$ density, we have that $m_i T(t) = \int_R \phi_\nu(t - \mu) \Pi_i(d\mu)$ is unimodal and symmetric about 0. We have the following result.

**Lemma 3.12.** If $T$ is a minimal sufficient statistic, $J_T(x)$ is constant in $x$, $\Pi_1$ and $\Pi_2$ are unimodal and symmetric about 0, the $P_i(t)$ have continuous distributions when $t \sim M_i T, m_1 T(0) > m_2 T(0)$, and $m_1 T(t) = m_2 T(t)$ has a unique solution for $t > 0$, then $\Pi_2$ is uniformly weakly informative relative to $\Pi_1$.

**Proof of Lemma 3.12.** By the unimodality and symmetry of $m_i T(t)$, we have $P_i(t) = M_i T(m_i T(u) \leq m_i T(t)) = M_i T(|u| \geq |t|)$. We show $M_1 T(|t| \geq t_0) \leq M_2 T(|t| \geq t_0)$ for all $t_0 > 0$ because it is equivalent to $\Pi_2$ is uniformly weakly informative relative to $\Pi_1$ by Lemma 3.2. Let $t_s$ be the solution of $m_1 T(t) = m_2 T(t)$ on $(0, \infty)$. From the unique solution assumption, $m_1 T(t) > m_2 T(t)$ for $t \in (0, t_s)$ and $m_1 T(t) < m_2 T(t)$ for $t > t_s$. For $0 \leq t_0 < t_s$, $M_1 T(|t| \geq t_0) = 2 \int_{t_0}^{t_s} m_1 T(t) dt = 1 - 2 \int_{t_0}^{t_0} m_1 T(t) dt \leq 1 - 2 \int_{t_0}^{t_s} m_2 T(t) dt = 2 \int_{t_0}^{t_s} m_2 T(t) dt = M_2 T(|t| \geq t_0)$ and for $t_0 \geq t_s$, $M_1 T(|t| \geq t_0) = 2 \int_{t_0}^{t_0} m_1 T(t) dt \leq 2 \int_{t_0}^{t_0} m_2 T(t) dt = M_2 T(|t| \geq t_0)$. Thus, we are done. \qed
We can apply Lemma 3.12 to comparing normal and \( t \) priors when sampling from a normal.

**Lemma 3.13.** Suppose we have a sample of \( n \) from a location normal model, \( \Pi_1 \) is a \( N(0, \sigma_1^2) \) prior and \( \Pi_2 \) is a \( t_1(0, \sigma_2^2, \lambda) \) prior. If \( m_{1T}(0) > m_{2T}(0) \), then \( \Pi_2 \) is uniformly weakly informative relative to \( \Pi_1 \).

**Proof of Lemma 3.13.** We have that \( m_{1T} = \phi_{1/n+\sigma_1^2} \) and, using the representation of the \( t(\lambda) \) distribution as a gamma mixture of normals, we can write \( m_{2T}(t) = \int_0^\infty \phi_{1/n+\sigma_2^2/u}(t) k_\lambda(u) du \) where \( k_\lambda \) is the density of \( \text{Gamma}_{\text{rate}}(\lambda/2, \lambda/2) \) distribution. By the symmetry of \( \phi_v, m_{2T} \) is symmetric. Also \( \phi_v(t_1) > \phi_v(t_2) \) for \( 0 \leq t_1 < t_2 \) and so \( m_{2T}(t_1) = \int \phi_{1/n+\sigma_2^2/u}(t_1)k_\lambda(u) du \geq \int \phi_{1/n+\sigma_2^2/u}(t_2)k_\lambda(u) du = m_{2T}(t_2) \). Thus, \( m_{2T} \) is decreasing on \((0, \infty)\), i.e., \( m_{2T} \) is unimodal. To show that \( m_{2T}(t) \) is log-convex with respect to \( t^2 \) we prove that \( (d^2/d(t^2)^2) \log m_{2T}(t) \geq 0 \). Note that \( (d/d(t^2))(\phi_v(t)) = (d/d(t^2))[(2\pi v)^{-1/2} \exp(-t^2/2v)] = -\phi_v(t)/2v, \)

\[
\frac{dm_{2T}(t)}{dt^2} = -\int_0^\infty \frac{\phi_{1/n+\sigma_2^2/u}(t)}{2(1/n + \sigma_2^2/u)}k_\lambda(u) du, \\
\frac{d^2 \log m_{2T}(t)}{d(t^2)^2} = \frac{1}{m_{2T}(t)} \int_0^\infty \frac{\phi_{1/n+\sigma_2^2/u}(t)}{2(1/n + \sigma_2^2/u)^2}k_\lambda(u) du \\
- \frac{1}{m_{2T}(t)^2} \left( \int_0^\infty \frac{\phi_{1/n+\sigma_2^2/u}(t)}{2(1/n + \sigma_2^2/u)}k_\lambda(u) du \right)^2
\]

and so \( d^2 \log m_{2T}(t)/d(t^2)^2 = \text{Var}_V((2(1/n + \sigma_2^2/V)/V)^{-1}) \geq 0 \), where \( V \) is the random variable having density \( \phi_{1/n+\sigma_2^2/v}(u)k_\lambda(v)/m_{2T}(t) \). Thus, \( m_{2T}(t) \) is log-convex in \( t^2 \).

The functions \( m_{1T}(t) \) and \( m_{2T}(t) \) meet in at most two points on \((0, \infty)\) because \( \log m_{1T}(t) \) is linear in \( t^2 \) and \( \log m_{2T}(t) \) is convex in \( t^2 \). Also \( m_{1T}(t) \) and \( m_{2T}(t) \) share at least one point on \((0, \infty)\) because \( m_{1T}(0) > m_{2T}(0) \), and the following shows that \( m_{1T}(t) < m_{2T}(t) \) for all large \( t \). Note first that if \( u \geq \sigma_2^2/2\sigma_1^2 \), then \( (1/n + \sigma_1^2)/(1/n +
\[ \sigma^2_2/u \geq 1/2 \text{ and } t^2/(u/n + \sigma^2_2) \geq (2\sigma^2_1/\sigma^2_2)t^2/(1/n + 2\sigma^2_1). \] Then,

\[
m_{2T}(t) \leq \frac{m_{2T}(t)}{m_{1T}(t)} \geq \frac{(\lambda/2)^{\lambda/2}}{\Gamma(\lambda/2)} \frac{(2\pi(1/n + \sigma^2_2/u))^{1/2}}{(2\pi(1/n + \sigma^2_1))^{1/2}} u^{\lambda/2-1} \times \frac{\exp\{-((u/2)(\lambda + t^2/(u/n + \sigma^2_2))\}}{\exp\{-(1/2)t^2/(1/n + \sigma^2_1)\}} du
\]

\[
\geq \frac{(\lambda/2)^{\lambda/2}}{\Gamma(\lambda/2)} \frac{1}{2^{1/2}} \frac{(\sigma^2_2/2\sigma^2_1)^{\lambda/2-1}}{\int_{\sigma^2_2/2\sigma^2_1}^{\infty} \exp\{-(u/2)(\lambda + (2\sigma^2_1/\sigma^2_2)t^2/(1/n + 2\sigma^2_1))\} du
\]

\[
= \frac{(\lambda/2)^{\lambda/2}}{\Gamma(\lambda/2)} \frac{1}{2^{1/2}} \frac{(\sigma^2_2/2\sigma^2_1)^{\lambda/2-1}}{\exp\{-\frac{\lambda \sigma^2_2}{2 \sigma^2_1}\}} \times \exp\{((1/2)t^2((1/n + \sigma^2_1) - (1/n + 2\sigma^2_1)) \rightarrow \infty \}
\]

as \( t^2 \rightarrow \infty. \)

The above conditions together imply that \( m_{1T}(t) \) and \( m_{2T}(t) \) meet in exactly one point on \((0, \infty). \) Therefore, \( \Pi_2 \) is uniformly weakly informative relative to \( \Pi_1 \) by Lemma 3.12.

Since \( \int_0^\infty (1/n + \sigma^2/u)^{-1/2}k_\lambda(u) du \) is strictly decreasing in \( \sigma^2, \) we see that \( m_{1T}(0) = (2\pi(1/n + \sigma^2_1))^{-1/2} \geq m_{2T}(0) = (2\pi)^{-1/2} \int_0^\infty (1/n + \sigma^2_2/u)^{-1/2}k_\lambda(u) du \) is equivalent to \( \sigma_2 \geq \sigma_{0n} \) where \( \sigma_{0n} \) satisfies \( (1/n + \sigma^2_{1})^{-1/2} = \int_0^\infty (1/n + \sigma^2_{0n}/u)^{-1/2}k_\lambda(u) du. \) This proves the first part of Theorem 3.6.

We also need the following results for the remaining parts of Theorem 2.

**Lemma 3.14.** \( \sigma^2_{0n}/\sigma^2_1 \) increases to \((2/\lambda)\Gamma^2((\lambda + 1)/2)/\Gamma^2(\lambda/2)\) as \( n\sigma^2_1 \rightarrow \infty. \)

(i) \( \sigma^2_{0n}/\sigma^2_1 \) increases as \( n\sigma^2_1 \rightarrow \infty, \) (ii) \( \sigma^2_{0n}/\sigma^2_1 \rightarrow (2/\lambda)\Gamma^2((\lambda + 1)/2)/\Gamma^2(\lambda/2) \) as \( n\sigma^2_1 \rightarrow \infty. \)

**Proof of Lemma 3.14.** We have \( n^{-1/2}(1/n + \sigma^2_1)^{-1/2} = n^{-1/2} \int_0^\infty (1/n + \sigma^2_{0n}/u)^{-1/2}k_\lambda(u) du \) and putting \( \alpha = n\sigma^2_1, \beta = n\sigma^2_{0n} \) we can write this as

\[(1 + \alpha)^{-1/2} = \int_0^\infty (1 + \beta/u)^{-1/2}k_\lambda(u) du. \quad (3.15)\]

Differentiating both sides of (3.15) with respect to \( \alpha \) we have \( (1 + \alpha)^{-3/2} = \int_0^\infty (1 + \beta/u)^{-3/2}u^{-1}k_\lambda(u) du \ (d\beta/d\alpha) \). If we let \( U \sim \text{Gamma}_{rate}(\lambda/2, \lambda/2) \), then this integral can
be written as the expectation

\[
E((1 + \beta/U)^{-3/2}U^{-1}) = E((1 + \beta/U)^{-3/2}(\beta/U + 1 - 1)/\beta)
\]

\[
= \beta^{-1}E((1 + \beta/U)^{-1/2}) - \beta^{-1}E((1 + \beta/U)^{-3/2})
\]

\[
\leq \beta^{-1}E((1 + \beta/U)^{-1/2}) - \beta^{-1}\{E((1 + \beta/U)^{-1/2})\}^3
\]

\[
= \beta^{-1}(1 + \alpha)^{-1/2} - \beta^{-1}(1 + \alpha)^{-3/2} = (1 + \alpha)^{-3/2}(\alpha/\beta)
\]

where the inequality follows via Jensen’s inequality. Hence, \(d\beta/d\alpha = (1 + \alpha)^{-3/2}/E((1 + \beta/U)^{-3/2}U^{-1}) \geq \beta/\alpha\) and so \(\beta/\alpha\) is an increasing function of \(\alpha\) because \(d(\beta/\alpha)/d\alpha = \alpha^{-1}(d\beta/d\alpha) - \beta/\alpha^2 \geq 0\). This proves \(\sigma_{\beta\alpha}^2/\sigma_1^2 = n\sigma_{\beta\alpha}^2/n\sigma_1^2 = \beta/\alpha\) increases as \(\alpha = n\sigma_1^2 \to \infty\). It is easy to check that \(\beta = 0\) when \(\alpha = 0\) and \(\beta > 0\) for \(\alpha > 0\). Let \(\alpha_0, \beta_0\) be a pair satisfying \(\alpha_0 > 0\) and (3.15). Then, \(\beta/\alpha \geq \beta_0/\alpha_0 > 0\) for \(\alpha > \alpha_0\) and \(\beta \to \infty\) as \(\alpha \to \infty\). Therefore,

\[
\lim_{\alpha \to \infty} \left(\frac{\beta}{\alpha}\right)^{1/2} = \lim_{\alpha \to \infty} \frac{\sqrt{\beta}}{\sqrt{1 + \alpha}} = \lim_{\alpha \to \infty} E\left(\frac{\sqrt{\beta}}{\sqrt{1 + \beta/U}}\right) = \lim_{\beta \to \infty} E\left(\frac{\sqrt{\beta}}{\sqrt{1 + \beta/U}}\right)
\]

\[
= E(\sqrt{U}) = \int_0^\infty \sqrt{u} k_u(u) \, du = (2/\lambda)^{1/2}\Gamma((\lambda + 1)/2)/\Gamma(\lambda/2)
\]

and this proves the lemma. \(\Box\)

**Lemma 3.15.** Suppose we have a sample of \(n\) from a location normal model, \(\Pi_1\) is a \(N(0, \sigma_1^2)\) prior and \(\Pi_2\) is a \(t_1(0, \sigma_2^2, \lambda)\) prior. Then \(\Pi_2\) is asymptotically uniformly weakly informative relative to \(\Pi_1\) if and only if \(\sigma_2^2/\sigma_1^2 \geq (2/\lambda)[\Gamma((\lambda + 1)/2)/\Gamma(\lambda/2)]^2\).

**Proof of Lemma 3.15.** Suppose that \(\sigma_2^2/\sigma_1^2 \geq (2/\lambda)[\Gamma((\lambda + 1)/2)/\Gamma(\lambda/2)]^2\). Then, by Lemma 3.14, \(\sigma_2^2/\sigma_1^2 \geq \sigma_{\beta\alpha}^2/\sigma_1^2\) for all \(n\) and so \(\Pi_2\) is uniformly weakly informative with respect to \(\Pi_1\) for all \(n\). So (3.4) is bounded above by \(\gamma\) for all \(n\) and so the limiting value of (3.4) is also bounded above by \(\gamma\). This establishes that \(\Pi_2\) is asymptotically uniformly weakly informative relative to \(\Pi_1\).

Suppose now that \(\sigma_2^2/\sigma_1^2 < (2/\lambda)[\Gamma((\lambda + 1)/2)/\Gamma(\lambda/2)]^2\). Note that \(m_{1T}(t) = \lim_{n \to \infty} m_{1T,n}(t) = (2\pi\sigma_1^2)^{-1/2}\exp(-t^2/(2\sigma_1^2))\) and similarly \(m_{2T}(t) = \lim_{n \to \infty} m_{2T,n}(t) = \Gamma((\lambda + \lambda)^{1/2})\).
Chapter 3. Weakly Informative Priors

101

\( (\pi_\lambda \sigma^2_\pi) (1 + t^2 / (\sigma^2_\lambda)) ^{- (\lambda+1)/2} \). Therefore, \( m_{1T}(0) = 1 / \sqrt{2\pi \sigma^2_\lambda} < \Gamma((\lambda + 1)/2) / \Gamma(\lambda/2) = m_{2T}(0) \). Let \( B = \{ t : m_{2T}(t) > m_{1T}(0) \} \) and \( \gamma = M_{2T}(B^c) \). Then, \( m_{1T}(t) \leq m_{1T}(0) \leq m_{2T}(t) \) on \( B \) and \( M_{1T}(P_2(t) \leq \gamma) = M_{1T}(B^c) = 1 - M_{1T}(B) = 1 - \int_B m_{1T}(t) dt \geq 1 - \int_B m_{2T}(t) dt = M_{2T}(B^c) = \gamma \). Hence, \( \Pi_2 \) is not weakly informative relative to \( \Pi_1 \) at level \( \gamma \). Therefore, \( \sigma^2_\pi / \sigma^2_\lambda \geq (2/\lambda) [\Gamma((\lambda+1)/2) / \Gamma(\lambda/2)]^2 \). □

It is now immediate that \( \sup_{\gamma \in [0,1]} G^{-1}_1(1-\gamma) / H^{-1}_1(1-\gamma) = (2/\lambda) [\Gamma((\lambda+1)/2) / \Gamma(\lambda/2)]^2 \) and the proof of Theorem 3.6 is complete.

Proof of Theorem 3.7

Since the minimal sufficient statistic \( T(x) = \bar{x} \) is linear there is no volume distortion and we can use (3.2) instead of (3.3). The limiting prior predictive distribution of \( T(x) = \bar{x} \) under \( \Pi_1 \) is \( N(\mu_0, \Sigma_1) \) and under \( \Pi_2 \) it is \( t_k(\mu_0, \Sigma_2, \lambda) \). It is easy to check that \( U_1 = (T - \mu_0)' \Sigma^{-1}_1 (T - \mu_0) \sim \chi^2(k) \) when \( T \sim \Pi_1 \) and \( U_2 = (T - \mu_0)' \Sigma^{-1}_2 (T - \mu_0) \sim k F_{k,\lambda} \) when \( T \sim \Pi_2 \). This implies that, \( P_{2,n}(t_0) \) converges to \( P_2(t_0) = \Pi_2(\pi_2(t) \leq \pi_2(t_0)) = 1 - H_{k,\lambda}((t_0 - \mu_0)' \Sigma^{-1}_2 (t_0 - \mu_0)/k) \) where \( H_{k,\lambda} \) is the distribution function of an \( F_{k,\lambda} \) distribution. Further, we have that (3.4) converges to \( \Pi_1(P_2(t) \leq \gamma) \).

Let \( V_i = \{ u \in R^k : u' \Sigma^{-1}_i u < 1 \} \) for \( i = 1, 2 \). By the continuity of \( \Pi_2(\pi_2(t) \leq r) \) as a function of \( r \), and the continuity of \( \pi_2(t) \), there exists \( t_0 \) such that \( P_2(t) \leq \gamma \) if and only if \( \pi_2(t) \leq \pi_2(t_0) \). Hence, \( \Pi_2 \) is asymptotically uniformly weakly informative relative to \( \Pi_1 \) if and only if \( \Pi_1(\pi_2(t) \leq \pi_2(t_0)) \leq \Pi_2(\pi_2(t) \leq \pi_2(t_0)) \) for all \( t_0 \in R^k \) by Lemma 3.2.

Since \( \pi_2(t) \) is decreasing in \( u_2 = U_2(t) \), the set \( \{ \pi_2(t) \leq \pi_2(t_0) \} = \{ u_2(t) \geq u_2(t_0) \} = \mu_0 + u_2(t_0)V^c \). So we must prove that \( \Pi_1(\mu_0 + r^{1/2} V^c) \leq \Pi_2(\mu_0 + r^{1/2} V^c) \) for all \( r \geq 0 \).

The positive semidefiniteness of \( \Sigma_2 - \tau^2_\lambda \Sigma_1 \) implies that \( \Sigma^{-1}_1 / \tau^2_\lambda - \Sigma^{-1}_2 \) is positive semidefinite. Then, for \( u \in V^c \), that is, \( u' \Sigma^{-1}_2 u \geq 1 \), we have \( u' \Sigma^{-1}_1 u = \tau^2_\lambda \cdot u' (\Sigma^{-1}_1 / \tau^2_\lambda) u \geq \tau^2_\lambda u' \Sigma^{-1}_1 u \geq \tau^2_\lambda. \) Thus, \( V^c \subset \tau_\lambda V^c \).

Now we prove a stronger inequality \( \Pi_1(\mu_0 + r^{1/2} \tau_\lambda V^c) \leq \Pi_2(\mu_0 + r^{1/2} V^c) \) for all \( r \geq 0 \).
Chapter 3. Weakly Informative Priors

Note that

\[ \Pi_1(\mu_0 + r^{1/2}\tau \lambda V_1^c) = \Pi_1(u_1(t) \geq r^{2}\tau \lambda) = \int_{r^{2}\tau \lambda}^{\infty} \frac{2^{-k/2}}{\Gamma(k/2)} u^{k/2-1} e^{-u/2} \, du, \]

\[ \Pi_2(\mu_0 + r^{1/2}V_2^c) = \Pi_2(u_2(t) \geq r) = \int_{r/\lambda}^{\infty} \frac{\Gamma((k + \lambda)/2)}{\Gamma(k/2)\Gamma(\lambda/2)} \left( \frac{k}{\lambda} \right)^{k/2} \left( 1 + \frac{r}{\lambda} \right)^{-(k+\lambda)/2} \, dt. \]

and set \( f(r) = \Pi_2(\mu_0 + r^{1/2}V_2^c) - \Pi_1(\mu_0 + r^{1/2}\tau \lambda V_1^c). \) Then, \( f(0) = 0 \) and

\[
\begin{align*}
\frac{df(r)}{dr} &= \frac{2^{-k/2}}{\Gamma(k/2)} (r^{2}\tau \lambda)^{k/2-1} e^{-r^{2}\tau \lambda/2} - \frac{\Gamma((k + \lambda)/2)}{\Gamma(k/2)\Gamma(\lambda/2)} \left( \frac{k}{\lambda} \right)^{k/2} \left( 1 + \frac{r}{\lambda} \right)^{-(k+\lambda)/2} \\
&= \frac{(\tau \lambda^2)^{k/2}}{\Gamma(k/2)} r^{k-2} e^{-r^2\tau \lambda/2} - \frac{\Gamma((k + \lambda)/2)}{\Gamma(k/2)\Gamma(\lambda/2)} \left( \frac{k}{\lambda} \right)^{k/2} \left( 1 + \frac{r}{\lambda} \right)^{-(k+\lambda)/2} \left( 1 + \frac{r}{\lambda} \right)^{-(k+\lambda)/2} = p_1 - p_2.
\end{align*}
\]

Note that \( p_1 - p_2 \geq 0 \) is equivalent to \( p_1/p_2 \geq 1 \). Further recalling the definition of \( \tau \lambda^k \)

from the statement of the theorem,

\[
\frac{p_1}{p_2} = \frac{\tau \lambda^k \Gamma(\lambda/2)}{\Gamma((k + \lambda)/2)} \left( \frac{\lambda}{2} \right)^{k/2} \left( 1 + \frac{r}{\lambda} \right)^{-(k+\lambda)/2} \exp \left( -\frac{r^2\tau \lambda}{2} \right) = \left( 1 + \frac{r}{\lambda} \right)^{-(k+\lambda)/2} \exp \left( -\frac{r^2\tau \lambda}{2} \right) \geq 1.
\]

The logarithm of \( p_1/p_2 \) given by \( \log(p_1/p_2) = -r^2\tau \lambda/(2+(k+\lambda)/2) \log(1+r/\lambda) \) is concave as a function of \( r > 0 \). Hence, \( \log(p_1/p_2) = 0 \) has exactly two solutions \( r = 0 \) and \( r = r_u \).

Because of its concavity, the function \( \log(p_1/p_2) \) is positive on \((0, r_u)\) and negative on \((r_u, \infty)\). This implies that \( f(r) \) is increasing on \((0, r_u)\) and decreasing on \((r_u, \infty)\). Since \( f(0) = 0 \) and \( \lim_{r \to \infty} f(r) = 0 \), the function \( f \) is nonnegative, that is, \( f(r) \geq 0 \) for all \( r \geq 0 \). Thus, \( \Pi_1(\mu_0 + r^{1/2}V_2^c) \leq \Pi_1(\mu_0 + r^{1/2}\tau \lambda V_1^c) \leq \Pi_2(\mu_0 + r^{1/2}V_2^c) \) for all \( r \geq 0 \).

**Proof of Theorem 3.8**

Let \( x_c^{-1} = \beta_i/(\alpha_i + 1/2) = \beta_2(\alpha_2 + 1/2) \). For \( i = 1, 2 \), let \( t_i(t_0) = 1/(x_c r_i(t_0)) \) be the two solutions of \( m^*_{2T}(t_i) = m^*_{2T}(t_0) \) (one of the \( t_i \) equals \( t_0 \)) so \( 0 < r_1 \leq 1 \leq r_2 \). Note that \( r_2(t_0) = 1 \) if and only if \( t_0 = x_c^{-1} \) and then \( r_1(t_0) = 1 \) as well. Then, \( \log(r_1/r_2) = r_1 - r_2 \) and \( dr_1/dr_2 = (r_2 - 1)r_1/[(r_1 - 1)r_2] \). Now \( \{ t : m^*_{2T}(t) \leq m^*_{2T}(t_0) \} = \{ t : 1/t \leq x_c r_1(t_0) \text{ or } 1/t \geq x_c r_2(t_0) \} \). By Lemma 3.2 we have that uniform weak informativity is equivalent to \( M_{1T}(m^*_{2T}(t) \leq m^*_{2T}(t_0)) \leq M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0)) \) for all \( t_0 \) and so we must prove that \( M_{1T}(t \notin (t_2(t_0), t_1(t_0))) = M_{1T}(1/t \leq x_c r_1(t_0) \text{ or } 1/t \geq x_c r_2(t_0)) = \)
1 - M_{1T}(x_c r_1(t_0) \leq 1/t \leq x_c r_2(t_0)) \leq 1 - M_{2T}(x_c r_1(t_0) \leq 1/t \leq x_c r_2(t_0))\) for all \(t_0\). Since \(r_1\) is implicitly a function of \(r_2\), it is equivalent to prove that \(M_{1T}(x_c r_1 \leq 1/t \leq x_c r_2) - M_{2T}(x_c r_1 \leq 1/t \leq x_c r_2) \geq 0\) for all \(r_2 \geq 1\). Using \((r_1/r_2)^\alpha = \exp(\alpha (r_1 - r_2))\), we have that the derivatives of the two terms are given by

\[
p_1 = \frac{d}{dr_2} \int_{x_c r_1}^{x_c r_2} c_1 u^{\alpha_1 - 1} e^{-\beta_1 u} \, du = c_1 (x_c r_2)^{\alpha_1 - 1} e^{-\beta_1 x_c r_2} x_c - c_1 (x_c r_1)^{\alpha_1 - 1} e^{-\beta_1 x_c r_1} x_c \frac{dr_1}{dr_2}
\]

\[
p_2 = c_2 x_c^{\alpha_2} r_2^{\alpha_2 - 1} e^{-\beta_2 x_c r_2} \left(1 - \frac{r_2 - 1}{r_1 - 1} \exp((r_2 - r_1)(\beta_2 x_c - \alpha_2))\right)
\]

where \(c_i = \beta_i^{\alpha_i}/\Gamma(\alpha_i)\). Then, recalling the definition of \(x_c\), we have that the ratio \(p_1/p_2 = (c_1/c_2)x_c^{\alpha_1 - \alpha_2} r_2^{\alpha_1 - \alpha_2} e^{(\beta_2 - \beta_1)x_c r_2} = (c_1/c_2)x_c^{\alpha_1 - \alpha_2} (r_2 e^{-r_2})^{\alpha_1 - \alpha_2}\) strictly decreases as \(r_2\) increases from 1 to \(\infty\) when \(\alpha_1 > \alpha_2\) because \(\alpha_1 - \alpha_2 = (\beta_1 - \beta_2)x_c \geq 0\), and is identically 1 when \(\alpha_1 = \alpha_2\). Suppose then that \(\alpha_1 > \alpha_2\) so there is at most one \(r_2\) value where \(p_1 = p_2\) and the derivative is 0. If \((p_1/p_2)|_{r_2=1} < 1\), then \(p_1 - p_2 < 0\) for all \(r_2 \geq 1\) and \(M_{1T}(x_c r_1 \leq 1/t \leq x_c r_2) - M_{2T}(x_c r_1 \leq 1/t \leq x_c r_2)\) strictly decreases from 0. This cannot hold because \(M_{1T}(x_c r_1 \leq 1/t \leq x_c r_2) - M_{2T}(x_c r_1 \leq 1/t \leq x_c r_2) \to 0\) as \(r_2 \to \infty\). Hence, \((p_1/p_2)|_{r_2=1} \geq 1\) and \(M_{1T}(x_c r_1 \leq 1/t \leq x_c r_2) - M_{2T}(x_c r_1 \leq 1/t \leq x_c r_2)\) increases from 0 near \(r_2 = 1\) and decreases to 0 as \(r_2 \to \infty\). Therefore, \(M_{1T}(x_c r_1 \leq 1/t \leq x_c r_2) - M_{2T}(x_c r_1 \leq 1/t \leq x_c r_2)\) goes up from 0 and down to 0 as \(r_2\) increases from 1 to \(\infty\), we have \(M_{1T}(x_c r_1 \leq 1/t \leq x_c r_2) - M_{2T}(x_c r_1 \leq 1/t \leq x_c r_2) \geq 0\) for all \(r_2 \geq 1\).

**Proof of Theorem 3.9**

We use the same notations in the proof of Theorem 3.8.

Let \(v = (r_2 - 1)/(1 - r_1)\) for computational simplicity. The ratio of derivatives is

\[
\frac{p_1}{p_2} = \frac{c_1}{c_2} x_c^{\alpha_1 - \alpha_2} e^{(\beta_2 - \beta_1)x_c} \times (r_2 e^{1-r_2})^{\alpha_1 - \alpha_2} \times e^{\eta(r_2-1)} \frac{1 + v \exp((r_2 - r_1)(1/2 - \eta))}{1 + v \exp((r_2 - r_1)/2)}.
\]

From the assumption, \((c_1/c_2)x_c^{\alpha_1 - \alpha_2} e^{(\beta_2 - \beta_1)x_c} \geq 1\). Also \(r_2 e^{1-r_2}\) is a monotone decreasing function on \(r_2 \in [1, \infty)\). If the third factor is monotone decreasing, then \(p_1/p_2\) is monotone
In the last inequality, it is non-positive if and only if decreasing and so it hits 1 exactly once. That means, $M_{1T}(1/(x_{c}r_2) \leq t \leq 1/(x_{c}r_1)) - M_{2T}(1/(x_{c}r_2) \leq t \leq 1/(x_{c}r_1))$ is non-negative.

The derivative of the logarithm of $e^{\eta r_2(1 + v \exp((r_2 - r_1)(1/2 - \eta)))(1 + v \exp((r_2 - r_1)/2))}$ is given by

$$\eta + \frac{e^{r_2(1/2 - \eta)}(v' + v(1/2 - \eta)(1 - r_1'))}{1 + ve^{r_2(1/2 - \eta)}} - \frac{e^{r_2(1/2 - \eta)}(v' + v(1/2)(1 - r_1'))}{1 + ve^{r_2(1/2 - \eta)}}.$$}

It is non-positive if and only if

$$\frac{1 - (rv_2/r_1)^{1/2 - \eta}r_1/r_2}{1 + v(rv_2/r_1)^{1/2 - \eta}} \leq \left(\frac{v' + v(1 + vr_1/r_2)}{2}ight) - \frac{(rv_2/r_1)^{1/2} - (rv_2/r_1)^{1/2 - \eta}}{(1 + v(rv_2/r_1)^{1/2 - \eta})(1 + v(rv_2/r_1)^{1/2})},$$

if and only if

$$\left(\frac{v' + v(1 + vr_1/r_2)}{2}\right)\left(1 - \left(\frac{r_1}{r_2}\right)^{\eta}\right) - \eta\left(\left(\frac{r_1}{r_2}\right)^{1/2} + v\right)\left(1 - v^2\left(\frac{r_1}{r_2}\right)^{1/2 + \eta}\right)$$

is non-negative for all $r_2 \geq 1$.

**Case I:** $r_2 \in [1, r_*]$. Where $r_* \approx 7.519997$ is the unique solution of $v^2((1/2 - \eta)) = 1$ on $r_2 \in (1, \infty)$. Then, $v^2((1/2 - \eta)) \geq 1$ for $r_2 \in [1, r_*]$ by Lemma 3.16. Thus, $1 - v^2((1/2 - \eta)) \leq 1 - (v_2/r_1)^{\eta}$ and

$$(3.16) \geq \left(\frac{v'}{2}(1 + vr_1/r_2) - \eta\left(\left(\frac{r_1}{r_2}\right)^{1/2} + v\right)\right)\left(1 - \left(\frac{r_1}{r_2}\right)^{\eta}\right)$$

$$= \left(\frac{v' - \eta\left(\frac{r_1}{r_2}\right)^{1/2}}{2} + v\left(\frac{1}{2} - \eta\right) + \frac{vr_1}{2r_2}\right)\left(1 - \left(\frac{r_1}{r_2}\right)^{\eta}\right)$$

$$\geq \left(\frac{2}{3} - \frac{1}{2}\right)\left(1 - \left(\frac{r_1}{r_2}\right)^{\eta}\right) \geq 0.$$}

In the last inequality, $v' \geq 2/3$ and $\eta((1/2 - \eta) \leq \eta \leq 1/2$ is used.

**Case II:** $r_2 > r_*$ and $1/3 \leq \eta \leq 1/2$. Under the condition $\eta \geq 1/3$, $1 - (v_2/r_1)^{\eta} \geq 1 - (v_2/r_1)^{1/3}$. Also $r_2 > r_* \approx 7.52$ implies $r_2/r_1 \leq 0.0005444 \leq 10^{-3}$. Thus, we have

$$(3.16) \geq \left(\frac{v'}{2}(1 + vr_1/r_2) - \eta\left(\left(\frac{r_1}{r_2}\right)^{1/2} + v\right)\right)\left(1 - \left(\frac{r_1}{r_2}\right)^{\eta}\right)$$

$$- \eta\left(\left(\frac{r_1}{r_2}\right)^{1/2} + v\right)\left(\frac{r_1}{r_2}\right)^{\eta}\left(1 - v^2\left(\frac{r_1}{r_2}\right)^{1/2}\right)$$
As a function of $\eta$, $\eta(r_1/r_2)^n$ is maximized at $\eta = 1/(r_2 - r_1)$. On $r_2 \geq r_*, 1/(r_2 - r_1) \leq 0.1331 < 1/3$. Hence, the maximum of $\eta(r_1/r_2)^n$ is achieved at $\eta = 1/3$. Besides $(1 - v^2(r_1/r_2)^{1/2})/(1 - (r_1/r_2)^{1/3}) \leq 1.01$ from Lemma 3.16.

$$\geq \left(\frac{2}{3} - \frac{1}{2} - \frac{1}{2} \cdot 10^{-3/2}\right) \left(1 - \left(\frac{r_1}{r_2}\right)^{1/3}\right) - \frac{1}{3}\left(\left(\frac{r_1}{r_2}\right)^{1/2} + v\right) \left(\frac{r_1}{r_2}\right)^{1/3} \cdot 1.01 \left(1 - \left(\frac{r_1}{r_2}\right)^{1/3}\right)
$$

$$\geq \left(\frac{2}{3} - \frac{1}{2} - \frac{1}{2} \cdot 10^{-1.5} - \frac{1}{3}\left(\left(\frac{r_1}{r_2}\right)^{5/6} + v\left(\frac{r_1}{r_2}\right)^{1/3}\right)\right) \left(1 - \left(\frac{r_1}{r_2}\right)^{1/3}\right)
$$

$$\geq \left(\frac{2}{3} - \frac{1}{2} - \frac{1}{2} \cdot 10^{-1.5} - \frac{1}{3}(10^{-2.5} + 1)\right) \left(1 - \left(\frac{r_1}{r_2}\right)^{1/3}\right) > 0.$$

**Case III:** $r_2 > r_*$ and $0 \leq \eta \leq 1/3$. Note that $\eta(r_1/r_2)^n$ is maximized at $\eta = 1/(r_2 - r_1)$ and the maximum value is $(r_2 - r_1)^{-1}(r_1/r_2)^{1/(r_2-r_1)} = e^{-1}/(r_2 - r_1) \leq 0.0489$ on $r_2 \geq r_*$. By the mean value theorem, $1 - (r_1/r_2)^n = (-\eta)(r_1/r_2)^n(r_1 - r_2)$ where $0 < \eta_* < \eta$. Also on $r_2 \geq r_*$, we get $1 - (r_1/r_2)^n \geq \eta(r_2 - r_1)(r_1/r_2)^n \geq 7\eta(r_1/r_2)^n$. Thus, we have

$$(3.16) \geq \left(v' + \frac{v}{2}(1 + v \frac{r_1}{r_2}) - \eta\left(\left(\frac{r_1}{r_2}\right)^{1/2} + v\right)\right) \left(1 - \left(\frac{r_1}{r_2}\right)^n\right) - \eta\left(\left(\frac{r_1}{r_2}\right)^{1/2} + v\right) \left(\frac{r_1}{r_2}\right)^n
$$

$$\geq \left(\frac{2}{3} + v/2 - (1/3)(1 + v)\right) \left(1 + v\right) - (1 + v)\eta\left(\frac{r_1}{r_2}\right)^n
$$

$$\geq \frac{1 + v}{6} \eta\left(\frac{r_1}{r_2}\right)^n > 0.$$

From cases I, II, and III, we have proved the theorem.

**Lemma 3.16.** Assume two numbers $0 < r_1 \leq 1 \leq r_2$ satisfies $r_1e^{-r_1} = r_2e^{-r_2}$. Let $v = (r_2 - 1)/(1 - r_1)$.

(a) $r_1' = \frac{dr_1}{dr_2} = -v r_1/r_2 \geq -1$ and $r_1'(1) = \frac{dr_1}{dr_2}|_{r_2=1} = -1$.

(b) $r_1'' = \frac{d^2r_1}{dr_2^2} = (r_1/r_2^2) (r_1 + r_2 - 1)(v^2 - 1)/(r_2 - 1) \geq 0$.

(c) $v' = ((v + 1)/r_2)(1 - r_1(v - 1)/(r_2 - 1)) \geq 0$ is increasing, $v'(1) = 2/3$ and $v' \to 1$ as $r_2 \to \infty$.

(d) $v(r_1/r_2)^\delta$ is monotone decreasing if $\delta \geq 1/3$.

(e) $(1 - v^2(r_1/r_2)^{1/2})/(1 - (r_1/r_2)^{1/3}) \leq 1 + 3 \times 10^{-8} \leq 1.01$. The maximum is achieved at $r_2 = 48.1371$. 
Proof of Lemma 3.16. Two numbers \( r_1 \) and \( r_2 \) also always satisfy \( \log(r_1) - r_1 = \log(r_2) - r_2 \). By differentiating both sides with respect to \( r_2 \), we have

\[
\left( \frac{1}{r_1} - 1 \right) \frac{dr_1}{dr_2} = \frac{1}{r_2} - 1.
\]

Hence, the first part of (a) is obtained as follows.

\[
\frac{dr_1}{dr_2} = \left( \frac{1}{r_2} - 1 \right) / \left( \frac{1}{r_1} - 1 \right) = -\frac{r_2 - 1}{r_1 - 1} = -\frac{r_1}{r_2} \leq 0.
\]

As \( r_2 \) decreases to 1, \( r_1 \) increases to 1. Hence,

\[
\left. \frac{dr_1}{dr_2} \right|_{r_2=1} = \lim_{r_2 \to 1} \frac{dr_1}{dr_2} = \lim_{r_2 \to 1} \frac{r_2 - 1}{r_1 - 1} \frac{r_1}{r_2} = \lim_{r_2 \to 1} \frac{dr_2}{dr_1} = 1/ \lim_{r_2 \to 1} \frac{dr_1}{dr_2} = 1/ \left. \frac{dr_1}{dr_2} \right|_{r_2=1}.
\]

So we have \( \frac{dr_1}{dr_2} \mid_{r_2=1} = -1 \). Also \( -1 = r_1'(1) = -v(1)r_1(1)/1 = -v(1) \), that is, \( v(1) = 1 \).

To get \( r_1' \geq -1 \), we need (c) and (b). Using \( r_1' = -vr_1/r_2 \), we have

\[
v' = \frac{dv}{dr_2} = \frac{1}{1 - r_1} \left( 1 - v^2 \frac{r_1}{r_2} \right) = \frac{1}{r_2} (v + 1) - \frac{r_1}{r_2} \frac{v^2 - 1}{v - 1} = \frac{v + 1}{r_2} \left( 1 - r_1 \frac{v - 1}{r_2 - 1} \right).
\]

If \( v < 1 \), then \( v' > (v + 1)/r_2 = (r_2 - r_1)/(r_2 - r_1 r_2) > 1 \). Suppose \( v < 1 \) at \( r_2 = s \). Then, \( s \) must be bigger than 1. And there exists a \( s_2 \in (1, s) \) such that \( v(s_2) = 1 \). That means \( v \) is in increasing status at \( r_2 = s_2 \). Thus, \( v \) bounds up once it hits 1. Hence, \( v \) cannot be less than 1. So we have \( v \geq 1 \).

The second derivative of \( r_1 \) is

\[
r_1'' = \frac{d^2r_1}{dr_2^2} = -\frac{r_1}{r_2} \frac{dv}{dr_2} - \frac{v}{r_2} \frac{dv}{dr_2} + \frac{vr_1}{r_2} = \frac{r_1(r_1 + r_2 - 1)}{r_2} \frac{v^2 - 1}{r_2 - 1} \geq 0.
\]

Besides, \( r_1' \geq r_1'(1) = -1 \). So we have (a) and (b).

Note that \( (r_1 r_2)' = r_1(1-v) \leq 0 \) implies \( r_1 r_2 \leq 1 \), \( (1 - r_1)(r_2 - v) = 1 - r_1 r_2 \geq 0 \), and \( r_2 \geq v \). Then, \( r_1 v (v - 1)/(r_2 - 1) \leq r_1 r_2 \cdot (v - 1)/(r_2 - 1) \leq 1 \cdot 1 = 1 \). Thus, \( v' \geq 0 \).

Also \( v'(1) = \lim_{r_2 \to 1} ((v + 1)/r_2) (1 - r_1 v (v - 1)/(r_2 - 1)) = 2(1 - v'(1)) \) implies \( v'(1) = 2/3 \).

Considering \( (v + 1)/r_2 = ((r_2 - 1)/r_2)/(1 - r_1) - 1/r_2 \to 1 \) and \( r_1 (v - 1)/(r_2 - 1) = r_1/(1 - r_1) - r_1/(r_2 - 1) \to 0 \) as \( r_2 \to \infty \), we have \( v' \to 1 \) as \( r_2 \to \infty \). To prove \( v' \) is non-decreasing, we use the second derivative given by

\[
v'' = \frac{v}{1 - r_1} \frac{r_1}{r_2} \left( \frac{v (v + 1)}{r_2} - 3 v' \right).
\]
Note that \( v''(1) = \lim_{r_2 \to 1} v''(r_2) = \lim_{r_2 \to 1} (v(v+1)/r_2 - 3v')/(r_2 - 1) = -3v''(1) = 0. \)

Since \( v'' > 0 \) for \( r_2 > 4 \), we have \( v'' > 0 \) on \( r_2 \in (1, \infty) \) if \( v'' = 0 \) has no solution on \( r_2 \in (1, \infty) \). We solve \( v'' = 0 \) by solving \( v(v+1)/r_2 = 3v' = 3((v+1)/r_2)(1-r_1v(v-1)/(r_2-1)) \).

By rearranging terms, this equation becomes

\[
\frac{v - 1}{r_2 - 1} = \frac{2}{r_2 - 1 + 3r_1v} = \frac{v - 1 + (1/2) \cdot 2}{r_2 - 1 + (1/2)(r_2 - 1 + 3r_1v)} = \frac{2}{3} \frac{v}{r_2 - 1 + r_1v} = \frac{2}{3}.
\]

In the last equality, \( v = r_2 - 1 + r_1v \) is used. Let \( r_2 = s > 1 \) be the smallest solution of \( v'' = 0 \). Then, \((v(s)-1)/(s-1) = 2/3\). By the mean value theorem, \((v(s)-1)/(s-1) = v'(s_2)\) for some \( 1 < s_2 < s \). Since \( v' \) is continuous and \( v'(1) = v'(s_2) = 2/3 \), there exists \( 1 < s_3 < s_2 \) such that \( v''(s_3) = 0 \). It contradicts to the assumption \( s \) is the smallest solution. Then, \( v'' = 0 \) has no solution on \((1, \infty)\) or infinitely many solutions. Using the smoothness of \( v'' \), we conclude that there is no solution on \((1, \infty)\). Hence, \( v'' \geq 0 \) and \( v' \geq 2/3 \).

Fix \( \delta \geq 1/3 \). Then,

\[
(\log(v(r_1/r_2)^\delta))' = \frac{v'}{v} - \delta \frac{v}{v} \leq \frac{v'}{v} - \frac{1}{3} \left( 1 + v \frac{r_1}{r_2} \right) = \frac{1}{3v} \left( 3v' - \frac{v(r_2 + r_1v)}{r_2} \right) = \frac{1}{3v} \left( 3v' - \frac{v(v+1)}{r_2} \right) \leq 0.
\]

Hence, \( v(r_1/r_2)^\delta \) is monotone decreasing.

(e) is obtained from a numerical computation.

**Proof of Proposition 3.10**

Without loss of generality we can assume that the scale of \( \Pi_{20} \) is increasing while the other parts \( \Pi_{2j} \) is fixed for \( j = 1, \ldots, q \). Suppose the prior \( \Pi_{20} \) on \( \beta_0 \) is a scaled prior of \( \Pi_* \), that is, \( \Pi_{20}(\cdot) = \Pi_*(\cdot/\tau) \) for some scale \( \tau \). Also assume that \( \Pi_*(\beta_0 > 0), \Pi_*(\beta_0 < 0) > 0 \) and \( \Pi_*(\beta_0 = 0) = 0 \). For very small number \( 0 < \epsilon \ll 1 \), there exists \( M > 0 \) such that \( 1/(1+e^{-M}) > 1 - \epsilon/(4n) \), \( 1/(1+e^M) < \epsilon/(4n) \) and \( \Pi_{20}(\|\beta\|_{\infty} \max_{j=1}^n |x_{ij} - \bar{x}_i| > M/k) < \epsilon/(4k) \) for \( i = 1, \ldots, q \) where \( n = n_1 + \cdots + n_q \). To compute prior predictive distribution,
Chapter 3. Weakly Informative Priors

define some sets: $B_1 = (|\beta_0| < 2M)$, $B_2 = (\beta_0 \geq 2M)$, $B_3 = (\beta_0 \leq -2M)$ and $B_4 = \bigcap_{i=1}^{k}(\max_{j=1,\ldots,n_i} |\beta_i(x_{ij} - \bar{x}_i)| \leq M/k)$. Then, $\Pi_2(B_4^c) \leq \epsilon/(4k)$; $p_i \geq 1/(1 + 1 - p_i \leq 1/(1 + e^M)$ on $B_2 \cap B_4$; $p_i \leq 1/(1 + e^M)$ and $1 - p_i \geq 1/(1 + e^{-M})$ on $B_3 \cap B_4$.

Now take $\tau > \tau_0$ where $\tau_0$ satisfies $\Pi_*(|\beta_0| < 2M/\tau_0) < \epsilon/2$.

\[
M_{2T}(t) = \int \cdots \int \prod_{i=1}^{q} \left( \frac{n_i}{t_i} \right) p_i^{t_i}(1-p_i)^{n_i-t_i} \prod_{i=0}^{k} \Pi_2(d\beta_i) \\
\leq \prod_{i=1}^{q} \left( \frac{n_i}{t_i} \right) t_i \left(1 - \frac{t_i}{n_i}\right)^{n_i-t_i} \Pi_2(B_1 \cup B_4^c) + \left[ \prod_{i=1}^{q} \left( \frac{n_i}{t_i} \right) \right] \left[ \left( \frac{1}{1 + e^M} \right)^{\sum_{i=1}^{q} (n_i-t_i)} \Pi_2(B_2 \cap B_4) + \left( \frac{1}{1 + e^M} \right)^{\sum_{i=1}^{q} t_i} \Pi_2(B_3 \cap B_4) \right]
\]

Using $\left( \frac{n_i}{t_i} \right)[1/(1 + e^M)]^{t_i} \leq \left( \frac{n_i}{t_i} \right)[\epsilon/(4n)]^{t_i} \leq \left( \frac{n_i}{t_i} \right)n^{-t_i}(\epsilon/4)^{t_i} \leq (\epsilon/4)^{t_i}$ and $\left( \frac{n_i}{t_i} \right)[1/(1 + e^M)]^{n_i-t_i} = (\frac{n_i}{t_i})[1/(1 + e^M)]^{n_i-t_i} \leq (\epsilon/4)^{n_i-t_i}$, we have

\[
\leq \Pi_2(B_1 \cup B_4^c) + (\epsilon/4)^{\sum_{i=1}^{q} (n_i-t_i)} \Pi_2(B_2 \cap B_4) + (\epsilon/4)^{\sum_{i=1}^{q} t_i} \Pi_2(B_3 \cap B_4) \\
\leq \epsilon/2 + (\epsilon/4)^{\sum_{i=1}^{q} (n_i-t_i)} \Pi_2(B_2) + (\epsilon/4)^{\sum_{i=1}^{q} t_i} \Pi_2(B_3).
\]

From this result, we have $M_{2T}(t) < \epsilon$ for $0 < \sum_{i=1}^{q} t_i < n$. Similar calculation gives us, for $\sum_{i=1}^{q} t_i = 0$,

\[
M_T(t) = \int \cdots \int \prod_{i=1}^{q} (1-p_i)^{n_i} \prod_{i=0}^{k} \Pi_2(d\beta_i) \geq \left( \frac{1}{1 + e^{-M}} \right)^{n} \Pi_2(B_3 \cap B_4) \\
\geq \left( 1 - \frac{\epsilon}{4n} \right)^{n} \Pi_2(B_3) \left( 1 - \frac{\epsilon}{4k} \right)^{k} \geq (1 - \epsilon/2) \Pi_2(B_3)
\]

where $(1 + a)^h \geq 1 + ha$ for positive integer $h$ and $a \geq -1$ is used, and for $\sum_{i=1}^{q} t_i = n$,

\[
M_{2T}(t) = \int \cdots \int \prod_{i=1}^{q} p_i^{n_i} \prod_{i=0}^{k} \Pi_2(d\beta_i) \geq (1 - \epsilon/2) \Pi_2(B_2)
\]

Take $\epsilon$ such that $\epsilon < \min, M_{1T}(t)$ and $\epsilon < \min(\Pi_*(\beta_0 > 0), \Pi_*(\beta_0 < 0))/2$, and also take $\tau$ big enough so that $\Pi_2(B_2) = \Pi_*(\beta_0 \geq 2M/\tau) \geq (1 - \epsilon/2) \Pi_*(\beta_0 > 0)$ and $\Pi_2(B_3) = \Pi_*(\beta_0 < -2M/\tau) \geq (1 - \epsilon/2) \Pi_*(\beta_0 < 0)$. Define $t_0 = (0, \ldots, 0) and
\( t_n = (n_1, \ldots, n_q) \). Then,

\[
M_{2T}(m_{2T}(t) < \min(m_{2T}(t_0), m_{2T}(t_n))) = \sum_t M_{2T}(t) I(0 < t_1 + \cdots + t_q < n)
\]

\[
= 1 - M_{2T}(t_0) - M_{2T}(t_n) \leq 1 - (1 - \epsilon/2)^2 \Pi_s(\beta_0 > 0) - (1 - \epsilon/2)^2 \Pi_s(\beta_0 < 0) < \epsilon.
\]

Then, consider \( \gamma = \min_t M_{1T}(t) > \epsilon \). By definition, \( x_\gamma = \gamma \) and

\[
M_{1T}(P_2(t) \leq x_\gamma) \geq M_{1T}(\{t_0, t_n\}^c) > \gamma.
\]

Hence, \( \Pi_2 \) is not weakly informative at level \( \gamma \) which implies \( \Pi_2 \) is not uniformly weakly informative at every level \( \gamma \in (0, 1) \).
Chapter 4

An Invariant Bayesian Inference - Relative Surprise Inference

A posterior probability is invariant under reparametrization, that is, for a one-to-one smooth transformation $W : \Theta \rightarrow \mathcal{H}$, the posterior probability of the transformed hypothesis is $\Pi_{\mathcal{H}}(W(H) \mid x_n) = \Pi_{\Theta}(H \mid x_n)$ for all $H \in \mathcal{B}(\Theta)$. Thus, posterior probabilities are invariant under reparametrization. As a result, Bayesian inference seems to be invariant. Unfortunately, counterexamples can be found very easily as we show in Section 4.1. Also relative surprise inferences are introduced in Section 4.1. Some properties of relative surprise inferences are summarized in Section 4.2. A decision theoretic approach for relative surprise inferences is dealt with in Section 4.3. In Section 4.4, consistency and the Bernstein-von Mises theorem for the least relative surprise estimator are dealt with. Conclusions and discussion are provided in Section 4.5

4.1 Relative Surprise Inference

A posterior probability is invariant under one-to-one smooth reparametrizations as already mentioned. A test of a hypothesis with zero prior probability may not be invariant. A singleton hypothesis $H : \theta = \theta_0$ cannot be tested using posterior probability because
\[ \Pi(H) = \Pi(H \mid x_n) = 0. \] In this case, highest posterior density (HPD) regions are suggested. For details see Welch and Peers (1963), Stein (1985), and Severini (1991). For this we reject the null hypothesis if the observed highest posterior density region at \( \theta_0 \), denoted by \( HPD(\theta_0) = \{ \theta \in \Theta \mid \pi(\theta \mid x_n) > \pi(\theta_0 \mid x_n) \} \), has large posterior probability, where \( \pi(\cdot \mid x_n) \) is the posterior density with respect to a dominating measure \( \nu \). So, the choice of dominating measure can also change the inference.

Consider a diffeomorphism \( W : \Theta \to \mathcal{W} \) (a one-to-one function such that itself and its inverse are continuously differentiable) and that is not the identity. Then, the density of \( W \) with respect to \( \nu \) is

\[ \pi_W(w) = \pi(W^{-1}(w))J_W(W^{-1}(w)) \]

where \( J_W \) is the reciprocal of the Jacobian of the transformation \( W \).

We have evidence against the hypothesis \( H_0 : W = w_0 (= W(\theta_0)) \) if

\[ \Pi_W(\pi_W(w \mid x_n) > \pi_W(w_0 \mid x_n) \mid x_n) \]

is large, or equivalently if

\[ \Pi(\pi(\theta \mid x_n)J_W(\theta) > \pi(\theta_0 \mid x_n)J_W(\theta_0) \mid x_n) \]

is large. Obviously, the HPD test is not invariant under one-to-one smooth parameter transformation.

An invariant Bayesian inference was defined in Evans (1997) by investigating how beliefs change from \textit{a priori} to \textit{a posteriori}. Suppose \( \Theta \) is the parameter set of the model \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \), \( \psi = \Psi(P_\theta) \) is the parameter of interest and a prior \( \Pi \) is placed on the model. Then, the \textit{relative surprise} at \( \psi_0 \) is defined by

\[ \Pi(\psi \in \Psi(\Theta) : \frac{\pi(\cdot \mid x_n)}{\pi(\psi_0 \mid x_n)} > \frac{\pi(\psi \mid x_n)}{\pi(\psi_0 \mid x_n)}) \mid x_n \]  

(4.1)

where \( \pi_\psi \) and \( \pi_\psi(\cdot \mid x_n) \) are the densities of the marginal prior, \( \Pi_\psi \), and the marginal posterior, \( \Pi_\psi(\cdot \mid x_n) \), with a dominating measure, \( \nu \), that is,

\[ \pi_\psi(\psi) = \frac{d\Pi_\psi}{d\nu}(\psi) \quad \text{and} \quad \pi_\psi(\psi \mid x_n) = \frac{d\Pi_\psi(\cdot \mid x_n)}{d\nu}(\psi). \]
Suppose \( \nu_i \)'s are two dominating measures of \( \Pi_\Psi \) and \( \Pi_\Psi(\cdot | x_n) \). Then, \( \nu_i \)'s are also dominated by \( \nu = \nu_1 + \nu_2 \). Let

\[
f_i(\psi) = \frac{d\nu_i}{d\nu}(\psi), \quad \pi_{i,\Psi,\text{prior}} = \frac{d\Pi_\Psi}{d\nu_i}(\psi), \quad \text{and} \quad \pi_{i,\Psi,\text{posterior}} = \frac{d\Pi_\Psi(\cdot | x_n)}{d\nu_i}(\psi).
\]

Then, it is easy to check that

\[
\frac{\pi_{i,\Psi,\text{posterior}}(\psi)}{\pi_{i,\Psi,\text{prior}}(\psi)} = \frac{\pi_{i,\Psi,\text{posterior}}(\psi)f_i(\psi)}{\pi_{i,\Psi,\text{prior}}(\psi)f_i(\psi)} = \frac{d\Pi_\Psi(\cdot | x_n)}{d\nu_i}(\psi)\frac{d\nu_i}{d\nu}(\psi)\left[\frac{d\Pi_\Psi}{d\nu_i}(\psi)\right] = \frac{d\Pi_\Psi(\cdot | x_n)}{d\nu}(\psi).
\]

Hence, (4.1) does not depend on the choice of a dominating measure.

The invariance under reparametrization is shown in the next theorem.

**Theorem 4.1.** Assume that \( \Psi(\Theta) \) is an open set of \( \mathbb{R}^k \). The relative surprise given in (4.1) is invariant under one-to-one smooth parameter transformation, that is, (4.1) is the same as

\[
\Pi_\Xi\left(\left\{ \xi \in \Xi(\Theta) : \frac{\pi_\Xi(\xi | x_n)}{\pi_\Xi(\xi)} > \frac{\pi_\Xi(\xi_0 | x_n)}{\pi_\Xi(\xi_0)} \right\} \mid x_n \right)
\]

where \( \Xi \) and \( \Psi \) are equivalent to each other, in other words, there exists a diffeomorphism \( h : \Psi(\Theta) \rightarrow \Xi(\Theta) \).

A sketch proof of Theorem 4.1. Suppose \( \Psi(\Theta) = \Xi(\Theta) \) and let \( \nu = \nu_\Psi + \nu_\Xi \). Two densities of \( \Pi_\Psi \) and \( \Pi_\Xi \) with respect to \( \nu \) are \( \pi_\Psi(\psi)\frac{d\psi}{d\nu}(\psi) \) and \( \pi_\Xi(\xi)\frac{d\xi}{d\nu}(\xi) \). The change of variable formula implies that

\[
\pi_\Xi(\xi)\frac{d\psi}{d\nu}(\xi) = \pi_\Psi(h^{-1}(\xi))\frac{d\psi}{d\nu}(h^{-1}(\xi))J_h(h^{-1}(\xi))
\]

In a similar vein, \( \pi_\Xi(\xi | x_n)\frac{d\psi}{d\nu}(\xi) = \pi_\Psi(h^{-1}(\xi | x_n))\frac{d\psi}{d\nu}(h^{-1}(\xi | x_n))J_h(h^{-1}(\xi | x_n)) \). Hence, we get

\[
\frac{\pi_\Xi(\xi | x_n)}{\pi_\Xi(\xi)} = \frac{\pi_\Psi(h^{-1}(\xi | x_n))\frac{d\psi}{d\nu}(h^{-1}(\xi | x_n))J_h(h^{-1}(\xi | x_n))}{\pi_\Psi(h^{-1}(\xi))\frac{d\psi}{d\nu}(h^{-1}(\xi))J_h(h^{-1}(\xi))} = \frac{\pi_\Psi(h^{-1}(\xi | x_n))}{\pi_\Psi(h^{-1}(\xi))} = \frac{\pi_\Psi(\psi | x_n)}{\pi_\Psi(\psi)}.
\]
Therefore, the relative surprise given by (4.1) is invariant under one-to-one, smooth reparametrizations.

The condition $\Psi(\Theta) = \Xi(\Theta)$ can be weakened to the case that $\Psi(\Theta)$ is a finite dimensional Riemann manifold. A formal proof of this generalization is given in Section 4.6.

\section{Properties of Relative Surprise}

The definition of the relative surprise indicates that, if the relative surprise at $\psi_0$ is close to 1, then the parameter value $\psi_0$ is very surprising because the posterior probability of $\psi$ values having a greater change of belief than $\psi_0$ is very large. In essence the data are pointing to other values of $\psi$ as more plausible candidates for the true value. Naturally we choose a parameter value that has the least surprise in the point estimation problem. This gives the following definition.

\textbf{Definition 4.2.} A value $\psi_0$ having the smallest relative surprise (4.1) is called the \textit{least relative surprise estimator (LRSE)} of $\psi$.

While least relative surprise estimators are point estimators, inferences are concerned with intervals or regions. So we need the following definition.

\textbf{Definition 4.3.} The relative change in belief for a set $H$ or a singleton $\psi$ is called the \textit{relative belief ratio (RB)} of $H$ or $\psi$, that is,

\begin{equation}
RB(H) = \frac{\Pi_{\Psi}(H | x_n)}{\Pi_{\Psi}(H)} \quad \text{and} \quad RB(\psi) = \frac{\pi_{\Psi}(\psi | x_n)}{\pi_{\Psi}(\psi)}.
\end{equation}

It is worth noting that the \textit{Bayes factor in favour of $H$}, which is denoted by $BF(H)$, and will be called ‘the Bayes factor of $H$’ without ‘in favour’ hereafter, is

\begin{equation}
BF(H) = \frac{\Pi_{\Psi}(H | x_n)}{1 - \Pi_{\Psi}(H | x_n)} / \frac{\Pi_{\Psi}(H)}{1 - \Pi_{\Psi}(H)} = \frac{\Pi_{\Psi}(H | x_n)}{\Pi_{\Psi}(H)} \frac{\Pi_{\Psi}(H^c | x_n)}{\Pi_{\Psi}(H^c)} = \frac{RB(H)}{RB(H^c)}.
\end{equation}

So relative belief ratios are closely related to Bayes factors.
Relative surprise inferences have many interesting properties. In this section, we present some of those properties which are mostly summarized from or are developed based on Evans (1997), Evans, Guttman and Swartz (2006), Evans and Shakhatreh (2007, 2008), and Evans and Zou (2002).

Property 1. LRSE is invariant under reparametrization.
This holds immediately from Theorem 4.1.

Property 2. The relative belief ratio, $RB(\psi)$, is the infinitesimal form of the Bayes factor.
When a Borel set $H$ shrinks ‘nicely’ (see Folland, 1999, page 98) to $\psi$ that is not an atom of $\Pi_\psi$, the Bayes factor converges to the posterior density with respect to the prior, that is,

$$BF(H) = \frac{\Pi_\psi(H | x_n)}{1 - \Pi_\psi(H | x_n)} / \frac{\Pi_\psi(H)}{1 - \Pi_\psi(H)} \to \frac{\pi_\psi(\psi | x_n)}{\pi_\psi(\psi)} = RB(\psi)$$

whenever both $\pi_\psi(\cdot | x_n)$ and $\pi_\psi$ are continuous in $\psi$. Even for a broad class of priors having discontinuous densities, this property still holds based on Theorem 10.49 of Wheeden and Zygmund (1977). Hence, the Bayes factor and the relative belief ratio are closely related.

Property 3. The Bayes factor and the relative belief ratio have the same sign when they are subtracted from 1. Also $|BF(H) - 1| \geq |RB(H) - 1|$ for all $H$.
A simple computation gives us

$$BF(H) = \frac{\Pi_\psi(H | x_n)}{1 - \Pi_\psi(H | x_n)} / \frac{\Pi_\psi(H)}{1 - \Pi_\psi(H)} = RB(H) \frac{1 - \Pi_\psi(H | x_n)/RB(H)}{1 - \Pi_\psi(H | x_n)}$$

$$= \frac{RB(H) - \Pi_\psi(H | x_n)}{1 - \Pi_\psi(H | x_n)} = 1 + \frac{RB(H) - 1}{\Pi_\psi(H^c | x_n)} .$$

This shows that $BF(H) \geq 1$ if and only if $RB(H) \geq 1$ and that $|BF(H) - 1| = |RB(H) - 1|/\Pi_\psi(H^c | x_n) \geq |RB(H) - 1|$. This means that the posterior odds ratio is greater than the prior odds ratio if the posterior belief is greater than the prior belief. In other words, a relative comparison of Bayes factors is equivalent to a relative comparison of relative beliefs.
**Property 4.** The Bayes factor increases as the relative belief increases when the prior or the posterior probability is fixed.

If the prior probability is fixed, then

\[
BF(H) = RB(H) \frac{1 - \Pi_{\psi}(H)}{1 - \Pi_{\psi}(H|x_n)} = \Pi_{\psi}(H^c) \frac{RB(H)}{1 - RB(H)\Pi_{\psi}(H)}.
\]

Since \(x/(1-ax)\) is increasing for \(0 \leq a \leq 1\) and \(0 \leq x \leq 1/a\), the Bayes factor increases as the relative belief increases when the prior probability is fixed. The Bayes factor obviously increases as the relative belief increases when the posterior probability is fixed, see (4.3).

Hence, the comparison of Bayes factors is equivalent to the comparison of relative beliefs when the prior or the posterior probability is fixed even on different scales.

**Property 5.** The LRSE maximizes the posterior density with respect to the prior when the model is dominated by a \(\sigma\)-finite measure.

The relative surprise at \(\psi_0\) becomes zero when the relative belief ratio at \(\psi\), given by

\[
RB(\psi) = \frac{\pi_{\psi}(\psi | x_n)}{\pi_{\psi}(\psi)}
\]

is essentially maximized at \(\psi_0\).

**Property 6.** The LRSE is the same as the maximum likelihood estimator if the full parameter \(\theta\) is of interest in a dominated model.

The posterior density is \(\pi(\theta | x_n) = f_\theta(x_n)\pi(\theta)/m(x_n)\). From this relationship, we have \(RB(\theta) = \frac{\pi(\theta | x_n)}{\pi(\theta)} = f_\theta(x_n)/m(x_n)\) and this is maximized at the maximum likelihood estimator. If a marginal parameter is of interest, there are similar likelihood based point estimators, see Evans, Guttman and Swartz (2006) for details.

In properties 5 and 6, we assumed that the model is dominated by a \(\sigma\)-finite measure. For any dominating measure \(\Lambda\) of \(\Pi\), that is, \(\Pi \ll \Lambda\), Lemma 4.4 given below implies \(\Pi(\cdot | x) \ll \Pi \ll \Lambda\). Hence, \(\Lambda\) can be used as a common dominating measure when we compute relative surprise.

**Lemma 4.4.** If the model \(\mathcal{F}\) is dominated by a \(\sigma\)-finite measure \(\mu\), then the posterior is absolutely continuous with respect to the prior a.s.-\(M\) where \(M\) is the prior predictive distribution.
The proof of Lemma 4.4 is in Section 4.6. Lemma 4.4 is not true unless the model $\mathcal{F}$ is dominated by a $\sigma$-finite measure, see Example 4.1.

**Example 4.1.** Consider a model $\mathcal{F}$ that is the collection of all probability measures on $\mathbb{R}^1$. Then, place a Dirichlet process $\Pi \sim \text{DP}(\alpha)$ on the model $\mathcal{F}$ where $\alpha$ is an atomless finite measure supported on $\mathbb{R}^1$. Obviously $\mathcal{F}$ contains all degenerate measures $\delta_x$ for $x \in \mathbb{R}^1$. Besides $\Pi$ is concentrated on the discrete probability measures, that is, $\Pi(E) = 1$ where $E$ is the collection of all discrete probability measures on $\mathbb{R}^1$, see Theorem 3.2.3 of Ghosh and Ramamoorthi (2003). Since there are more than countably many points in $\mathbb{R}^1$, the model $\mathcal{F}$ cannot be dominated by a $\sigma$-finite measure.

Let $B = \{P \in \mathcal{F} : P(\{x_1, \ldots, x_n\}) > 0\}$. Theorem 3.2.1 of Ghosh and Ramamoorthi (2003) implies $\Pi(B) \sim \text{Beta}(\alpha(B), \alpha(B^c))$. So we get $\Pi(B) = 0$ because of $\alpha(B) = 0$. While $\Pi(B | x_n) = 1$ from Theorems 3.2.1 and 3.3.2 of Ghosh and Ramamoorthi (2003). Thus, $\Pi(\cdot | x) \not\ll \Pi$.

A construction of a test for a null hypothesis is equivalent to the construction of a collection of nested rejection regions. Rejection regions can be replaced by the complements of credible regions in Bayesian inferences. In order to develop inferences based on relative belief ratios or Bayes factors, we define two credible regions in Definitions 4.5 and 4.6.

**Definition 4.5.** A relative surprise region is a credible region of form $\{\psi : RB(\psi) > c\}$ for some $c \geq 0$. The $\gamma$-relative surprise region denoted by $C_\gamma(x_n)$ is the smallest region among all relative surprise regions having posterior probability at least $\gamma$, that is, the intersection of all relative surprise regions having posterior probability at least $\gamma$.

Since relative surprise regions are nested, every $\gamma$-relative surprise region has the smallest prior probability (or the smallest volume with respect to the volume measure on $\Psi$) among all relative surprise regions having posterior probability at least $\gamma$. Then, the $\gamma$-relative surprise region is given by $C_\gamma(x_n) = C(x_n, c_\gamma(x_n))$ where $c_\gamma(x_n) = \inf\{c \geq$
0 : \( \Pi(\Psi) \geq \gamma \) and \( C(x_n, c) = \{ \psi : RB(\psi) > c \} \). Every \( \gamma \)-relative surprise region has posterior probability at least \( \gamma \) from the definition. And this might be strictly bigger than \( \gamma \) in some cases — it happens if \( \pi(\psi|x_n)/\pi(x_n) \) has a positive posterior probability, i.e., \( \Pi(\{ \psi : \pi(\psi|x_n)/\pi(x_n) = c \} | x_n) > 0 \) for some \( c \). Similarly, Bayes factor regions can be defined.

**Definition 4.6.** The \( \gamma \)-Bayes factor region is a posterior region having the highest Bayes factor among all regions having posterior probability exactly \( \gamma \) if there exists a region having posterior probability exactly \( \gamma \).

The restriction having posterior probability exactly \( \gamma \) is required to define the \( \gamma \)-Bayes factor region appropriately. If a discrete prior is considered in data analysis, Bayes factor regions may not exist for some \( \gamma \)'s. Besides Bayes factor regions may not be nested. In that case, an appropriate \( P \)-value cannot be achieved based on Bayes factors. Hence, comparisons of Bayes factors may not be appropriate for non-nested regions. In general, Bayes factor regions are nested under mild conditions, see Proposition 4.8.

**Property 7.** A relative surprise region \( C_\gamma(x_n) \) has the smallest prior measure among all \( \psi(x_n) \) satisfying \( \Pi(\psi) \geq \Pi(\psi(x_n)) | x_n) \). Also, \( \Pi(C_\gamma(x_n)) \geq \Pi(C) whenever C \) satisfies \( \Pi(C) | x_n) = \Pi(C_\gamma(x_n) | x_n) \).

This means relative surprise regions are optimal with respect to prior belief. Details can be found in Evans, Guttman and Swartz (2006). For \( \eta = \Pi(\psi(x_n) | x_n) \), the \( \eta \)-relative surprise region is given by \( C_\eta(x_n) = C_\gamma(x_n) \) and it also maximizes Bayes factor among all credible regions having posterior probability exactly \( \eta \) by combining Properties 4 and 7. This proves the next proposition.

**Proposition 4.7.** When \( \gamma \)-relative surprise region \( C_\gamma(x_n) \) has posterior probability exactly \( \gamma \), the \( \gamma \)-Bayes factor region is the same as the \( \gamma \)-relative surprise region \( C_\gamma(x_n) \).

Proposition 4.7 implies that there exists a sequence of nested Bayes factor regions whenever the \( \gamma \)-relative surprise region has posterior probability exactly \( \gamma \).
We say that \( \Pi_\Psi \) has the *atomless relative belief ratio property* when the posterior probability of the set of \( \psi \)'s having the same relative belief ratio is always 0, that is,

\[
\Pi_\Psi(\{\psi : RB(\psi) = c\} | x_n) = \Pi_\Psi(\{\psi : \pi_\Psi(\psi | x_n)/\pi_\Psi(\psi) = c\} | x_n) = 0 \text{ for all } c \geq 0.
\]

It is easy to see that every \( \gamma \)-relative surprise region has posterior probability exactly \( \gamma \) for all \( \gamma \in (0, 1) \) if and only if \( \Pi_\Psi \) has the atomless relative belief ratio property. Hence, we have the following proposition.

**Proposition 4.8.** If \( \Pi_\Psi \) has the atomless relative belief ratio property, then every \( \gamma \)-relative surprise region has posterior probability exactly \( \gamma \) and there exist nested \( \gamma \)-Bayes factor regions.

**Property 8.** The relative surprise region gains more belief from a priori to a posteriori, that is, \( RB(C_\gamma(x_n)) > 1 \) for all \( \gamma \).

Also Properties 3 and 8 imply that \( BF(C_\gamma(x_n)) \geq 1 \) for all \( \gamma \). Details can be found in Evans and Shakhatreh (2007).

The posterior odds ratios of relative surprise regions are also bigger than the prior odds ratios using Property 3. The optimality in Properties 7 and 8 can be extended in several ways that are described in the following theorem.

**Theorem 4.9.** Suppose that \( \Pi_\Psi \) satisfies the atomless relative belief ratio property. Let \( C_\gamma(x_n) \) be the \( \gamma \)-relative surprise region. Then,

(i) the relative belief ratio of \( C_\gamma(x_n) \) decreases as \( \gamma \) increases,

(ii) the \( \gamma \)-relative surprise region maximizes relative belief ratio among all posterior regions having posterior probability at least \( \gamma \),

(iii) the \( \gamma \)-relative surprise region maximizes Bayes factor among all posterior regions having posterior probability exactly \( \gamma \).

**Proof of Theorem 4.9.** (i). The proof presented below is similar to the proof of Lemma 6 of Evans and Shakhatreh (2007). Note that \( C_\gamma(x_n) = \{\psi : \pi_\Psi(\psi | x)/\pi_\Psi(\psi) > c_\gamma(x_n)\} \).
Under the atomless relative belief ratio property, $C_\gamma(x_n) \subseteq C_\eta(x_n)$ and $c_\gamma(x_n) > c_\eta(x_n)$ whenever $0 < \gamma < \eta < 1$. By the definition of $C_\gamma(x_n)$, $\pi_\psi(\psi|x_n)/\pi_\psi(\psi) > c_\gamma(x_n)$ on $C_\gamma(x_n)$. So we obtain

$$
\Pi_\psi(C_\gamma(x_n)|x_n) = \int_{C_\gamma(x_n)} \frac{\pi_\psi(\psi|x_n)}{\pi_\psi(\psi)} \Pi_\psi(d\psi) > \int_{C_\gamma(x_n)} c_\gamma(x_n) \Pi_\psi(d\psi) = c_\gamma(x_n) \Pi_\psi(C_\gamma(x_n)).
$$

And this implies

$$RB(C_\gamma(x_n)) = \frac{\Pi_\psi(C_\gamma(x_n)|x_n)}{\Pi(C_\gamma(x_n))} > c_\gamma(x_n).$$

We have $\pi_\psi(\psi|x_n)/\pi_\psi(\psi) \leq c_\gamma(x_n)$ on $C_\eta(x_n) - C_\gamma(x_n)$. Thus,

$$\Pi_\psi(C_\eta(x_n) - C_\gamma(x_n)|x_n) = \int_{C_\eta(x_n) - C_\gamma(x_n)} \pi_\psi(\psi|x_n)d\nu(\psi) \leq \int_{C_\eta(x_n) - C_\gamma(x_n)} c_\gamma(x_n) \pi_\psi(\psi)d\nu(\psi) = c_\gamma(x_n) \Pi_\psi(C_\eta(x_n) - C_\gamma(x_n)).$$

Hence, we have

$$RB(C_\gamma(x_n)) = \frac{\Pi_\psi(C_\gamma(x_n)|x_n)}{\Pi(C_\gamma(x_n))} > c_\gamma(x_n) \geq \frac{\Pi_\psi(C_\eta(x_n) - C_\gamma(x_n)|x_n)}{\Pi_\psi(C_\eta(x_n) - C_\gamma(x_n))}.$$

An application of Lemma A.2 for this inequality implies

$$RB(C_\gamma(x_n)) = \frac{\Pi_\psi(C_\gamma(x_n)|x_n)}{\Pi(C_\gamma(x_n))} > \frac{\Pi_\psi(C_\eta(x_n)|x_n)}{\Pi(C_\eta(x_n))} = RB(C_\eta(x_n)).$$

(ii). Suppose $B$ is a Borel set such that $\eta = \Pi_\psi(B|x_n) \geq \gamma$. Then, we have $\Pi_\psi(C_\eta(x_n)) \leq \Pi_\psi(B)$ by Corollary 4 of Evans, Guttman and Swartz (2006). This implies

$$RB(B) = \frac{\Pi_\psi(B|x_n)}{\Pi_\psi(B)} = \frac{\eta}{\Pi_\psi(B)} \leq \frac{\eta}{\Pi_\psi(C_\eta(x_n))} = \frac{\Pi_\psi(C_\eta(x_n)|x_n)}{\Pi_\psi(C_\eta(x_n))} = RB(C_\eta(x_n)) \leq RB(C_\gamma(x_n)).$$

Therefore, $C_\gamma(x_n)$ obtains the maximum relative belief among all posterior regions having posterior probability at least $\gamma$. 

Let $B$ be a Borel set having posterior probability exactly $\gamma$, that is, $\Pi_\Psi(B \mid x_n) = \gamma$. The Bayes factor of $B$ is
\[
BF(B) = \frac{\Pi_\Psi(B \mid x_n)}{1 - \Pi_\Psi(B \mid x_n)} \frac{1 - \Pi_\Psi(B)}{\Pi_\Psi(B)} = \frac{\gamma}{1 - \gamma \left( \frac{1}{\Pi_\Psi(B)} - 1 \right)}.
\]
Consequently the maximum Bayes factor is achieved when $\Pi_\Psi(B)$ is minimized. Again $\Pi_\Psi(B)$ is minimized when $B = C_\gamma(x_n)$ by Corollary 4 of Evans, Guttman and Swartz (2006). Therefore, we have the theorem.

The atomless relative belief ratio property ensures that each $\gamma$-relative surprise region has posterior probability exactly $\gamma$ for all $\gamma \in (0, 1)$. When this condition is dropped, the same results hold for the $\gamma$'s having $\gamma = \Pi_\Psi(C_\gamma(x_n) \mid x_n)$. A similar phenomenon is discussed in Corollary 4 of Evans, Guttman and Swartz (2006).

As the credible level $\gamma$ is getting higher, the corresponding credible region is also getting bigger. Then, the relative belief of the higher credible region must decrease because the credible region must contain points having smaller and smaller relative belief ratios. That means, when $\gamma$ is getting higher, the $\gamma$-relative surprise region contains less believable points. Hence, the relative belief of the $\gamma$-relative surprise region must not be greater than those of low level relative surprise regions. So the first result is very sensible.

In contrast, the Bayes factor of the $\gamma$-Bayes factor region increases as $\gamma$ increases in many cases. The $\gamma$-Bayes factor region is also the same as the $\gamma$-relative surprise region, $C_\gamma(x_n)$, by Theorem 4.9 (iii). Note that, for $\psi \in C_\gamma(x_n)^c$, we have $\pi_\Psi(\psi \mid x_n)/\pi_\Psi(\psi) \leq c_\gamma(x_n)$. Thus,
\[
\Pi_\Psi(C_\gamma(x_n)^c \mid x_n) = \int I_{C_\gamma(x_n)^c}(\psi) \pi_\Psi(\psi \mid x_n) \nu_\Psi(d\psi) \\
\leq \int I_{C_\gamma(x_n)^c}(\psi) c_\gamma(x_n) \pi_\Psi(\psi) \nu_\Psi(d\psi) = c_\gamma(x_n) \Pi_\Psi(C_\gamma(x_n)^c).
\]
The Bayes factor of the $\gamma$-relative surprise region (or the $\gamma$-Bayes factor region) satisfies
\[
BF(C_\gamma(x_n)) = \frac{RB(C_\gamma(x_n))}{RB(C_\gamma(x_n)^c)} = RB(C_\gamma(x_n)) \frac{\Pi_\Psi(C_\gamma(x_n)^c)}{\Pi_\Psi(C_\gamma(x_n)^c \mid x_n)} \geq RB(C_\gamma(x_n)) \frac{1}{c_\gamma(x_n)}.
\]
As $\gamma \to 1$, clearly $\Pi_\psi(C_\gamma(x_n) \mid x_n) \to 1$. Besides in many problems, $\Pi_\psi(C_\gamma(x_n)) \to 1$ and $c_\gamma(x_n) \to 0$ as $\gamma \to 1$. Hence, $RB(C_\gamma(x_n)) \to 1$ and $BF(C_\gamma(x_n)) \to \infty$ as $\gamma \to 1$.

Note that $\Pi_\psi(C_\gamma(x_n)) \to 1$ as $\gamma \to 1$ is not guaranteed. A radical example can be found when the prior and the posterior are singular (It happens with a nonparametric prior, see Example 4.1). In this case, $\Pi_\psi(C_\gamma(x_n)) = 0$ for all $\gamma \in (0,1)$. Even among dominated models, we can find a counterexample, see Example 4.2.

Example 4.2. Consider $X_i | \theta \sim i.i.d. \text{Uniform}(0, \theta)$ where $\theta \in (0, N)$ for a big number $N > 0$. Then, place a prior $\Pi \sim \text{Uniform}(0, N)$. It is easy to see that the posterior density is $(n - 1)\theta^{-n}/(x_{(n)}^{1-n} - N^{1-n})I(x_{(n)} < \theta < N)$ where $x_{(n)} = \max(x_1, \ldots, x_n)$. Hence, it satisfies the atomless relative belief ratio property. So the $\gamma$-Bayes factor region is the same as the $\gamma$-relative surprise region which is given by $(x_{(n)}, x_{N,n,\gamma})$ where $x_{N,n,\gamma} = x_{(n)}(1 - \gamma(1 - (x_{(n)}/N)^{n-1}))^{-1/(n-1)}$. Since $x_{N,n,\gamma} \to N$ as $\gamma \to 1$, we have $\Pi(C_\gamma(x_n)) \to \Pi((x_{(n)}, N)) = (N - x_{(n)})/N = 1 - x_{(n)}/N < 1$ as $\gamma \to 1$. In this example, $c_\gamma(x_n) = (n - 1)N x_{N,n,\gamma}^{-n} / (x_{(n)}^{1-n} - N^{1-n}) \to (n - 1)/(N/x_{(n)})^{n-1} - 1 > 0$.

In many problems, the Bayes factor of $C_\gamma(x_n)$ is strictly increasing in $\gamma$ while the relative belief $RB(C_\gamma(x_n))$ is decreasing. It is natural to ask whether or not the Bayes factor of $\gamma$-relative surprise region (or $\gamma$-Bayes factor region) increases as $\gamma$ increases.

Lemma 4.10. Suppose $\Pi_\psi(C_\gamma(x_n) \mid x_n) = \gamma$ for all $\gamma \in (0,1)$ and $c_\gamma(x_n)$ is continuous in $\gamma$. Then, the Bayes factor of $C_\gamma(x_n)$ is increasing in $\gamma$ if and only if
\[
c_\gamma(x_n) \geq \frac{\Pi_\psi(C_\gamma(x_n) \mid x_n)(1 - \Pi_\psi(C_\gamma(x_n) \mid x_n))}{\Pi_\psi(C_\gamma(x_n))(1 - \Pi_\psi(C_\gamma(x_n)))} = RB(C_\gamma(x_n))RB(C_\gamma(x_n)^c). \quad (4.4)
\]

The proof of Lemma 4.10 is in Section 4.6. Interestingly, upper and lower bounds of (4.4) are given below
\[
RB(C_\gamma(x_n)^c)c_\gamma(x_n) \leq \frac{\Pi_\psi(C_\gamma(x_n) \mid x_n)(1 - \Pi_\psi(C_\gamma(x_n) \mid x_n))}{\Pi_\psi(C_\gamma(x_n))(1 - \Pi_\psi(C_\gamma(x_n)))} \leq RB(C_\gamma(x_n))c_\gamma(x_n).
\]
But, $RB(C_\gamma(x_n)^c) \leq \min(1, c_\gamma(x_n)) \leq \max(1, c_\gamma(x_n)) \leq RB(C_\gamma(x_n))$ does not prove (4.4). Neither a counterexample to (4.4) nor a proof is known. So we leave this question as an open problem.
Open Problem 4.11. Is the inequality (4.4) true in general?

Property 8 says that a relative surprise region always gains more belief from \textit{a priori} to \textit{a posteriori}. It also increases the Bayes factor. However, this may not be true for HPD regions, see Examples 4.3 and 4.4. Of course, HPD regions achieve bigger Bayes factor for large $n$, see Remark 4.1 for details.

Example 4.3. Suppose $f_j \sim N(j, 1)$ for $j = 1, 2$. Consider a prior $\Pi$ having mass $\pi_1 = e/(e + 1)$ at $f_1$ and $\pi_2 = 1 - \pi_1 = 1/(e + 1)$ at $f_2$ where $e$ is the base of the natural logarithm. Suppose we observed a datum $x$. The posterior probability is

$$\Pi(j|x) = \frac{f_j(x)\pi_j}{f_1(x)\pi_1 + f_2(x)\pi_2}.$$  

When $x = 2$ is observed, the posterior probability becomes

$$\Pi(1|x) = \frac{e^{1/2}}{e^{1/2} + 1} \approx 0.6225 \quad \text{and} \quad \Pi(2|x) = \frac{1}{e^{1/2} + 1} \approx 0.3775.$$

For $\gamma = 0.6$, the $HPD_\gamma(x)$ region with respect to counting measure is $HPD_\gamma(x) = \{1\}$. Then the relative belief ratio of $HPD_\gamma(x)$ is

$$RB(HPD_\gamma(x)) = \frac{\Pi(HPD_\gamma(x)|x)}{\Pi(HPD_\gamma(x))} = \frac{e^{1/2}}{e^{1/2} + 1} \left/ \frac{e}{e + 1} \right. = \frac{e + 1}{e + e^{1/2}} (\approx 0.8514) < 1.$$  

Also the Bayes factor of $HPD_\gamma(x)$ is $RB(HPD_\gamma(x)) = e^{-1/2} (\approx 0.6065) < 1$. Hence, the posterior probability of the $HPD_\gamma(x)$ region is less than the prior probability of the region.

Remark 4.1. Chen (1985) showed that appropriately centered and scaled posterior distribution converges to a normal distribution, that is, $\sqrt{n}(\theta - \hat{\theta})|x_n \xrightarrow{d} N(0, I(\theta_0)^{-1})$ where $\hat{\theta}$ is the posterior mode and $I(\theta)$ is the Fisher information matrix. We call this property \textit{Bayesian asymptotic normality}. Then, asymptotic $\gamma$-HPD regions can be constructed as

$$HPD_{n,\gamma}(x_n) = \{\theta : n(\theta - \hat{\theta}(x_n))'I(\theta_0)(\theta - \hat{\theta}(x_n)) \geq \chi^2_\gamma(k)\}$$  

where $\chi^2_\gamma(k)$ is the $\gamma$-quantile of the $\chi^2(k)$ distribution and $k$ is the rank of $I(\theta_0)$. So $HPD_{n,\gamma}(x_n)$ shrinks to $\theta_0$ with
order $O_P(n^{-1/2})$. Hence, we have $\Pi(HPD_{n,\gamma}(x_n)|x_n) \to \gamma$ and $\Pi(HPD_{n,\gamma}(x_n)) \to 0$ as $n \to \infty$. These also imply, for $n \to \infty$,

$$RB(HPD_{n,\gamma}(x_n)) = \frac{\Pi(HPD_{n,\gamma}(x_n)|x_n)}{\Pi(HPD_{n,\gamma}(x_n))} \to \infty$$

as well as

$$BF(HPD_{n,\gamma}(x_n)) = \frac{\Pi_\psi(HPD_{n,\gamma}(x_n)|x_n)}{1 - \Pi_\psi(HPD_{n,\gamma}(x_n)|x_n)} / \frac{\Pi_\psi(HPD_{n,\gamma}(x_n))}{1 - \Pi_\psi(HPD_{n,\gamma}(x_n))} \to \infty.$$ 

Therefore, HPD regions also have a property similar to Property 8 in the asymptotic sense.

**Example 4.4.** Suppose $f_{\theta} \sim N(\theta, 1)$ for $\theta \in \mathbb{R}$. Assume a prior $\theta \sim p_0 N(0, \sigma^2) + (1 - p_0) N(\theta_0, \sigma^2)$. Suppose we observed a datum $x$. The posterior distribution is

$$\theta|x \sim pN\left(\frac{x}{1 + 1/\sigma^2}, \frac{1}{1 + 1/\sigma^2}\right) + (1 - p)N\left(\frac{x + \theta_0/\sigma^2}{1 + 1/\sigma^2}, \frac{1}{1 + 1/\sigma^2}\right)$$

where $p = p_0 \exp\left(-\frac{x^2}{2(1+\sigma^2)}\right)/(p_0 \exp\left(-\frac{x^2}{2(1+\sigma^2)}\right) + (1 - p_0) \exp\left(-\frac{(x-\theta_0)^2}{2(1+\sigma^2)}\right)$. Fix the hyper-parameters $p_0 = 0.15, \theta_0 = 2, \sigma^2 = 0.16$. Assume that $x = 0.2$ is observed. The prior and posterior densities are given in Figure 4.1.

The posterior mode is approximately 1.7517 (the vertical line in Figure 4.1(a)). The relative belief ratio of the $\gamma$-HPD region given by $HPD_{\gamma}(x) = \{\theta : \pi_\psi(\psi|x_n) > r_\gamma\}$ for some $r_\gamma \ge 0$ is given in Figure 4.1(b).

Both the relative belief and the Bayes factor are less than 1 when $\gamma < 0.4651$. That means that the prior probability of $HPD_{\gamma}(x)$ is greater than the posterior probability of $HPD_{\gamma}(x)$ when $\gamma < 0.4651$. □

This phenomenon occurs because the posterior density is smaller than the prior density at the posterior mode. So the HPD inference may not be plausible if the posterior density at the posterior mode is less than the prior density. As discussed in Remark 4.1, the posterior density gets higher density at the posterior mode as the sample size increases.
Decision Theoretic Interpretation of Inferences

Based upon Surprise

Decision theoretic methods comprise one of the most popular inference methods in Bayesian data analysis. This method also starts from a sample, $x_n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, which comes from an unknown distribution $P_\theta$ where $\theta \in \Theta$. One of main purposes of this methodology is to make a decision among all possible actions, denoted by $\mathcal{A} = \{\beta(\theta) : \theta \in \Theta\}$. The action $\beta(\theta)$ is called a correct action when the sample $x_n$ comes from $P_\theta$. A prior $\Pi$ is placed on $\Theta$. A decision $\delta$ is a random variable having values in $\mathcal{A}$, that is $\delta : \mathcal{X}^n \times \Omega \rightarrow \mathcal{A}$ where $\Omega$ is a probability space that is independent from $P^*_n \times \Pi$.

To find the correct action or an action very close to the correct action, a loss function is defined. In general a loss function is a mapping $L : \Theta \times \mathcal{A} \rightarrow [-K, \infty)$ where $0 \leq K < \infty$. But we impose two restrictions on loss functions in this chapter: (1) it is non-negative, that is, $L(\theta, a) \geq 0$ for all $\theta \in \Theta$ and $a \in \mathcal{A}$, and (2) $L(\theta, \beta(\theta)) = 0$ where $\beta(\theta) \in \mathcal{A}$ is the correct action if the data come from $P_\theta$. The risk of a decision is the expected loss of the
decision function, that is, $R(\theta, \delta) = \mathbf{E}_\theta \mathbf{E}_\Omega L(\theta, \delta(X_n, \omega))$. If $\delta(x_n, \omega)$ has a unique value for all $\omega$, it is called a non-randomized decision. A non-randomized decision becomes a simpler mapping $\delta : \mathcal{X}^n \rightarrow \mathcal{A}$ and the risk function is $R(\theta, \delta) = \mathbf{E}_\theta L(\theta, \delta(x_n))$. The prior risk is the integrated risk with respect to the prior, that is, $r(\delta) = \mathbf{E}_\Pi R(\theta, \delta)$. A decision $\delta$ is called a Bayes rule if it has the smallest finite prior risk. Immediately, a Bayes rule minimizes the posterior risk given by $\mathbf{E}_{\Pi(\cdot|X_n)}(R(\theta, \delta)|x_n)$. But the converse is not true because $\mathbf{E}_M \mathbf{E}_{\Pi(\cdot|X_n)}(R(\theta, \delta)|x_n)$ can be infinity where $M$ is the prior predictive measure. Also a decision $\delta$ is called a generalized Bayes rule if it minimizes the posterior risk.

Bayes rules depend on the choice of the loss function. The posterior mean is the Bayes rule with respect to squared error loss function given by $L(\theta, a) = ||\theta - a||^2$. In the discrete case, the maximum a posteriori (MAP) estimator, or posterior mode, is the Bayes rule when 0-1 loss function $L(\theta, a) = I(\theta \neq a)$ is used. The MAP estimator, the point at which the posterior probability is maximized in a discrete model, is used often in statistical machine learning theory. It can be generalized to the posterior mode in continuous models.

It is well-known that the posterior mode can be interpreted as a limit of Bayes rules using a sequence of special loss functions, details are in Proposition 5.2 of Bernardo and Smith (2000). Also Le Cam (1953) showed that the maximum likelihood estimator is a Bayes rule in an asymptotic sense. Note that the difference between the posterior mode and the least relative surprise estimator is the choice of the dominating measure. Also the maximum likelihood estimator is the least relative surprise estimator if the full parameter is of interest.

In this subsection, we derive relative surprise inferences via a decision-theoretic analysis. We show that the least relative surprise estimator is a limit of Bayes rules with respect to a certain sequence of loss functions. We also show that surprise regions are also limits of lowest posterior loss regions (see Bernardo, 2005, for the definition) for the same sequence of loss functions.
The sampling model is $F = \{ P_\theta | \theta \in \Theta \}$ and a prior $\Pi$ is placed on the parameter space $\Theta$. Suppose that the parameter of interest is $\psi = \Psi(\theta)$ and so $A = \Psi(\Theta)$. The correct action is $\psi = \Psi(\theta)$ when $\theta$ is the true parameter value. For the generalized case, let $\Lambda$ be a finite measure on the action space, $\Psi \subset \mathbb{R}^k$. We need to specify a loss function for a decision theoretic analysis. For two positive numbers $r, \eta > 0$, the loss function is defined by

$$L^{\Lambda}_{r,\eta}(\theta, a) = I(a \notin B_r(\Psi(\theta))) / \max(\eta, \Lambda(B_r(\Psi(\theta))))$$

(4.5)

where $B_r(\psi)$ is the open ball centered at $\psi$ with radius $r > 0$. For notational simplicity, we denote $\psi = \Psi(\theta)$ in this section. The bound on the denominator in (4.5) is required to make the risk function bounded in general. Then, the risk function for deterministic decision function $\delta$ is

$$R^{\Lambda}_{r,\eta}(\theta, \delta) = E_\theta L^{\Lambda}_{r,\eta}(\theta, \delta) = P_\theta(\delta(X) \notin B_r(\psi)) / \max(\eta, \Lambda(B_r(\psi))).$$

So, the prior risk becomes

$$r^{\Lambda}_{r,\eta}(\delta) = E_\Pi R^{\Lambda}_{r,\eta}(\theta, \delta) = \int \frac{1 - P_\theta(\delta(X) \in B_r(\psi))}{\max(\eta, \Lambda(B_r(\psi)))} \Pi(d\theta)$$

$$= \int \frac{1}{\max(\eta, \Lambda(B_r(\psi)))} \Pi_\psi(d\psi) - \int \int \frac{I(\delta(\mathbf{x}_n) \in B_r(\psi))}{\max(\eta, \Lambda(B_r(\psi)))} \Pi_\psi(d\psi | \mathbf{x}_n) M(d\mathbf{x}_n)$$

where $M$ is the prior predictive measure.

To get a Bayes rule, we need some conditions on the measure $\Lambda$.

**SD1** Both $\Pi_\psi$ and $\Pi_\psi(\cdot | \mathbf{x})$ are absolutely continuous with respect to $\Lambda$, that is, $\Pi_\psi \ll \Lambda$ and $\Pi_\psi(\cdot | \mathbf{x}_n) \ll \Lambda$.

**SD2** For all $\psi_0 \in \text{supp}(\Lambda)$,

$$\lim_{r \to 0} \sup_{\psi \in B_r(\psi_0)} \frac{\Lambda(B_r(\psi))}{\Lambda(B_r(\psi_0))} = 1$$

and

$$\lim_{r \to 0} \inf_{\psi \in B_r(\psi_0)} \frac{\Lambda(B_r(\psi))}{\Lambda(B_r(\psi_0))} = 1$$

where $\text{supp}(\Lambda)$ is the topological support of $\Lambda$, that is, the smallest closed set on which the measure is concentrated.
(SD3) There exists a sequence \( \eta(r, \Lambda) \) such that \( \{ \psi : \Lambda(B_r(\psi)) \geq \eta(r, \Lambda) \} \) covers the support of \( \Lambda \) as \( r \to 0 \), that is, \( \Lambda([\liminf_{r \to 0}\{ \psi : \Lambda(B_r(\psi)) \geq \eta(r, \Lambda) \}]^c) = \Lambda([\cup_{s>0} \cap_{0<r<s} \{ \psi : \Lambda(B_r(\psi)) \geq \eta(r, \Lambda) \}]^c) = 0. \)

The first condition is easily satisfied if the considered model \( F \) is dominated by a \( \sigma \)-finite measure and \( \Pi_\Psi \ll \Lambda \). In Proposition 4.13, we show that (SD2) is fulfilled for a very large class of finite measures. The condition (SD3) is required to make \( \Lambda(B_r(\psi)) \geq \eta(r, \Lambda) \) as \( r \to 0 \) for all \( \psi \). If a version of the Radon-Nikodym derivative of \( \Lambda \) with respect to the Lebesgue measure \( L \) is continuous, then there exists \( \eta(r, \Lambda) \) satisfying (SD3), see Lemma 4.18.

In Evans and Zou (2002), the observed surprise based on \( \Lambda \) at \( \psi_0 \) is defined by

\[
OS^\Lambda(\psi_0 | x_n) = \Pi_\Psi\left(\frac{d\Pi_\Psi(\cdot | x_n)}{d\Lambda}(\psi) > \frac{d\Pi_\Psi(\cdot | x_n)}{d\Lambda}(\psi_0) | x_n\right). \tag{4.6}
\]

Also the \( \gamma \)-observed surprise region is defined by \( OS_\gamma(x_n) = \{ \psi : OS^\Lambda(\psi | x_n) \leq \gamma \} \).

Then, the \( \gamma \)-OS region equals the \( \gamma \)-HPD region when \( \Lambda \) is the dominating measure used for HPD inferences, and equals the \( \gamma \)-relative surprise region when \( \Lambda = \Pi_\Psi \). So, the observed surprise inference extends both HPD and RS inferences. However, OS inference may not be invariant. We discuss this issue later. The least observed surprise estimator (LOSE), \( \psi^\Lambda_{LOSE} \), is the marginal parameter value \( \psi_0 \) minimizing the observed surprise \( OS^\Lambda(\psi_0) \). It is not hard to see that the LOSE \( \psi^\Lambda_{LOSE} \) is the ‘essential’ (see Royden, 1988, page 119 for a definition) maximizer of \( g^\Lambda(\psi | x_n) = (d\Pi_\Psi(\cdot | x_n)/d\Lambda)(\psi) \) with respect to \( \Pi_\Psi(\cdot | x_n) \).

We now establish the following result.

**Theorem 4.12.** Assume that \( \Psi \) is an open set of a finite dimensional Euclidean space and that \( g^\Lambda \) has a unique essential supremum. Then, the following holds.

(i) A Bayes rule with respect to the loss function \( L^\Lambda_{r,\eta} \) for fixed \( r \) and \( \eta \) is

\[
\delta^\Lambda_{r,\eta}(x_n) = \text{arg sup}_{\psi \in \Psi} \int_{B_r(\psi)} \frac{1}{\max(\eta, \Lambda(B_r(\psi)))} \Pi_\Psi(d\psi | x_n).
\]
(ii) Under the assumptions (SD1)–(SD3), the Bayes rule \( \delta_r = \delta_{r,\eta(r,\Lambda)}^\Lambda \) converges to \( \delta_{LOSE}^\Lambda \) as \( r \to 0 \).

(iii) Under the assumptions (SD1)–(SD3) and \( \Pi_\Psi(\{\psi : g^\Lambda(\psi \mid x_n) = c\} \mid x_n) = 0 \) for all \( c \geq 0 \), the \( \gamma \)-lowest posterior loss region converges to

\[
OS_{\gamma}^\Lambda(x_n) = \{\psi : g^\Lambda(\psi \mid x_n) > c_{\gamma}^\Lambda(x_n)\}
\]

in \( \Pi_\Psi(\cdot \mid x_n) \)-probability as \( r \to 0 \) where \( c_{\gamma}^\Lambda(x_n) \) is the maximum constant \( c \) satisfying \( \Pi_\Psi(\{\psi : g^\Lambda(\psi \mid x_n) \geq c\} \mid x_n) = \gamma \).

Proof of Theorem 4.12. In this proof, we put \( \psi = \Psi(\theta) \) for notational simplicity.

(i) Since \( L_{r,\eta}^\Lambda(\theta, a) = I(a(x_n) \not\in B_r(\psi))/\max(\eta, \Lambda(B_r(\psi))) \leq 1/\eta \), the prior risk is bounded by \( 1/\eta \), that is,

\[
r_{r,\eta}^\Lambda(a) = E_{\Pi} L_{r,\eta}^\Lambda(\theta, a) \leq E_{\Pi}(1/\eta) \leq 1/\eta < \infty.
\]

Hence, for fixed \( r, \eta \), an estimator minimizing the posterior risk is a Bayes rule. The posterior risk is

\[
r_{r,\eta}^\Lambda(a \mid x_n) = E_{\Pi|X}(L_{r,\eta}^\Lambda(\theta, a) \mid x_n) = \int \frac{1 - I(a \in B_r(\psi))}{\max(\eta, \Lambda(B_r(\psi)))} \Pi_\Psi(d\psi \mid x_n)
\]

\[
= \int \frac{1}{\max(\eta, \Lambda(B_r(\psi)))} \Pi_\Psi(d\psi \mid x_n) - \int_{B_r(a)} \frac{1}{\max(\eta, \Lambda(B_r(\psi)))} \Pi_\Psi(d\psi \mid x_n).
\]

The first term in (4.7) is constant when \( \eta \) and \( r \) are fixed and \( a \) is varying. Thus, the minimizer \( a \) of \( r_{r,\eta}^\Lambda(a \mid x_n) \) is the maximizer of

\[
h(a, r) = \int_{B_r(a)} \frac{1}{\max(\eta, \Lambda(B_r(\psi)))} \Pi_\Psi(d\psi \mid x_n).
\]

(ii) From the assumption (SD1), \( g^\Lambda(\psi \mid x_n) \) is finite a.e.-\( \Lambda \). Fix \( \epsilon \in (0, 1) \). From (SD2) and (SD3), there is \( R_a > 0 \) such that for all \( r < R \), we obtain \( 1 - \epsilon < \Lambda(B_r(\psi))/\Lambda(B_r(a)) < \)
1 + \epsilon for all \psi \in B_r(a) and \Lambda(B_r(a)) \geq \eta(r, \Lambda). Hence,
\[
h(a, r) \leq \int_{B_r(a)} \frac{1}{\max(\eta(r, \Lambda), (1 - \epsilon)\Lambda(B_r(a)))} \Pi_\psi(d\psi \mid x_n) \\
\leq \frac{1}{1 - \epsilon} \int_{B_r(a)} \frac{1}{\max(\eta(r, \Lambda), \Lambda(B_r(a)))} \Pi_\psi(d\psi \mid x_n) = \frac{1}{1 - \epsilon} \frac{\Pi_\psi(B_r(a) \mid x_n)}{\Lambda(B_r(a))} = \frac{1}{1 - \epsilon} g^\Lambda(a)
\]
as \epsilon \to 0, where the convergence is obtained by the measure differentiation theorem, see Folland (1999). Similarly,
\[
h(a, r) \geq \frac{1}{1 + \epsilon} \frac{\Pi_\psi(B_r(a) \mid x_n)}{\max(\eta(r, \Lambda), \Lambda(B_r(a)))} \to \frac{1}{1 + \epsilon} g^\Lambda(a \mid x_n)
\]
as \epsilon \to 0. Hence, \(h(a, r) \to g^\Lambda(a \mid x_n)\).

We now prove \(\delta_r \to \delta^A_{\text{LOOSE}}\) as \(r \to 0\). Let \(\epsilon > 0\). We show there exists \(R_1 > 0\) such that \(||\delta_r - \delta^A_{\text{LOOSE}}|| < \epsilon\) for all \(r \leq R_1\). Since \(\delta^A_{\text{LOOSE}}\) is the unique essential maximizer, \(D_\zeta = \{\psi : g^\Lambda(\psi \mid x_n) > (1 - \zeta)g^A(\delta^A_{\text{LOOSE}} \mid x_n)\}\) converges to \(\delta^A_{\text{LOOSE}}\) as \(\zeta\) converges to 0. Also the convergence of \(D_\zeta\) implies that the diameter of \(D_\zeta\) converges to 0 as \(\zeta \to 0\). So there exists \(\zeta_0 > 0\) such that \(\text{diam}(D_\zeta) < \epsilon/3\) for all \(0 < \zeta \leq \zeta_0\). Since \(||\psi - \delta^A_{\text{LOOSE}}|| \leq \text{diam}(D_\zeta)\) for all \(\psi \in D_\zeta\), we have \(g^\Lambda(\psi \mid x_n) \leq (1 - \zeta)g^A(\delta^A_{\text{LOOSE}} \mid x)\) for all \(\psi \in [B_{\text{diam}(D_\zeta)}(\delta^A_{\text{LOOSE}})]^c \subset D_\zeta^c\). This implies, for \(\psi \in [B_{r + \text{diam}(D_\zeta)}(\delta^A_{\text{LOOSE}})]^c\), \(B_r(\psi) \subset [B_{\text{diam}(D_\zeta)}(\delta^A_{\text{LOOSE}})]^c\) and
\[
\frac{\Pi_\psi(B_r(\psi) \mid x_n)}{\max(\eta(r, \Lambda), \Lambda(B_r(\psi)))} \leq \frac{\Pi_\psi(B_r(\psi) \mid x_n)}{\Lambda(B_r(\psi))} = \frac{1}{\Lambda(B_r(\psi))} \int_{B_r(\psi)} g^A(\psi^* \mid x_n) \Lambda(d\psi^*) \\
\leq \frac{1}{\Lambda(B_r(\psi))} \int_{B_r(\psi)} (1 - \zeta)g^A(\delta^A_{\text{LOOSE}} \mid x_n) \Lambda(d\psi^*) \\
= (1 - \zeta)g^A(\delta^A_{\text{LOOSE}} \mid x_n).
\]
So we obtain \(h(\psi, r) \leq (1 - \zeta_0)g^A(\delta^A_{\text{LOOSE}} \mid x_n)\) for all \(||\psi - \delta^A_{\text{LOOSE}}|| \geq r + \text{diam}(D_\zeta)\). Since \(h(\delta^A_{\text{LOOSE}} \mid r) \to g^A(\delta^A_{\text{LOOSE}} \mid x_n)\), there exists \(R_2 > 0\) such that \(h(\delta^A_{\text{LOOSE}} \mid r) > (1 - \zeta_0)g^A(\delta^A_{\text{LOOSE}} \mid x_n)\) for all \(r \leq R_2\). Then, \(h(\delta, r) \geq h(\delta^A_{\text{LOOSE}} \mid r) > (1 - \zeta_0)g^A(\delta^A_{\text{LOOSE}} \mid x_n)\). Hence, the essential maximizer of \(h(\psi, r)\) is in \(B_{r + \text{diam}(D_\zeta)}(\delta^A_{\text{LOOSE}})\) and we have \(||\delta_r - \delta^A_{\text{LOOSE}}|| < r + \text{diam}(D_\zeta)\) for all \(r \leq R_2\). Therefore, \(||\delta_r - \delta^A_{\text{LOOSE}}|| < \epsilon\) for all \(r \leq R_1 = \min(R_2, \epsilon/2)\). So we obtain \(\delta_r \to \delta^A_{\text{LOOSE}}\) as \(r \to 0\).
(iii) Considering that the first term in (4.7) is constant for fixed $r$ and $\eta = \eta(r, \Lambda)$, the $\gamma$-lowest posterior loss region is

$$O^\Lambda_\gamma(x_n, r) = \{ \psi : r^\Lambda_{r,\eta(r,\Lambda)}(\psi | x_n) \leq c^\Lambda_\gamma(x_n | r) \} = \{ \psi : h(\psi, r) \geq \alpha^\Lambda_\gamma(x_n | r) \}$$

where $\alpha^\Lambda_\gamma(x_n | r) = \int (\max(\eta(r, \Lambda), \Lambda(B_r(\psi))))^{-1} \Pi_\Psi(d\psi | x_n) - \Lambda(B_r(\psi)) | c^\Lambda_\gamma(x_n | r)$ is the minimum constant $c$ satisfying $\Pi_\Psi(\{ \psi : r^\Lambda_{r,\eta(r,\Lambda)}(\psi | x_n) \leq c \} | x_n) \geq \gamma$. We prove that the convergence by showing that $\Pi_\Psi(O S^\Lambda_\gamma(x_n) \Delta O^\Lambda_\gamma(x_n, r) | x_n) \to 0$ as $r \to 0$.

Fix $\epsilon > 0$. Take a big number $R_1 > 0$ such that $\Pi_\Psi(B_{R_1}(\delta^\Lambda_{\text{LOSE}}) | x_n) < \epsilon/8$. Define

$$c_{2\epsilon} = \inf \{ c > 0 : \Pi_\Psi(\{ \psi : g^\Lambda(\psi | x_n) \geq c^\Lambda_\gamma(x_n) + c \} | x_n) \leq \gamma - \epsilon/4 \} \quad (4.8)$$

where $c^\Lambda_\gamma(x_n)$ defines $O S^\Lambda(x_n)$. Also define

$$c_{1\epsilon} = \inf \{ c > 0 : \Pi_\Psi(\{ \psi : g^\Lambda(\psi | x_n) \geq c^\Lambda_\gamma(x_n) - c \} | x_n) \geq \gamma + \epsilon/4 \} \quad (4.9)$$

Then, we get

$$c_{2\epsilon} \to 0 \quad \text{and} \quad c_{1\epsilon} \to c_{\text{range}} \quad \text{as} \quad \epsilon \to 0 \quad (4.10)$$

where $c_{\text{range}} = \inf \{ c > 0 : \Pi_\Psi(\{ \psi : g^\Lambda(\psi | x_n) \in [c^\Lambda_\gamma(x_n) - c, c^\Lambda_\gamma(x_n)] \} | x_n) > 0 \}$, that is, $c_{\text{range}}$ is the maximum value satisfying $\gamma = \Pi_\Psi(O S^\Lambda_\gamma(x_n) | x_n) = \Pi_\Psi(\{ \psi : g^\Lambda(\psi | x_n) > c^\Lambda_\gamma(x_n) - c \} | x_n)$. So we have $c_{\text{range}} = 0$ whenever $g^\Lambda$ is continuous. Also the continuity assumption, $\Pi_\Psi(\{ \psi : g^\Lambda(\psi | x_n) = c \}) = 0$ for all $c \geq 0$, implies

$$\Pi_\Psi(\{ \psi : g^\Lambda(\psi | x_n) \in [c^\Lambda_\gamma(x_n) - c_{\text{range}}, c^\Lambda_\gamma(x_n)] \} | x_n) = 0. \quad (4.11)$$

Also from part (ii), $h(\psi, r)$ converges to $g^\Lambda(\psi | x_n)$ as $r \to 0$. By Egorov’s theorem, there exists a set $E \subset B_{R_1}(\delta^\Lambda_{\text{LOSE}})$ such that $h(\psi, r)$ converges uniformly on $E$ and $\Lambda(B_{R_1}(\delta^\Lambda_{\text{LOSE}}) - E) < \epsilon/[8g^\Lambda(\delta^\Lambda_{\text{LOSE}} | x_n)\Lambda(B_{R_1}(\delta^\Lambda_{\text{LOSE}}))]$. Then,

$$\Pi_\Psi(E^c | x_n) = \Pi_\Psi(B_{R_1}(\delta^\Lambda_{\text{LOSE}})^c | x_n) + \Pi_\Psi(B_{R_1}(\delta^\Lambda_{\text{LOSE}}) - E | x_n) \leq \epsilon/8 + g^\Lambda(\delta^\Lambda_{\text{LOSE}} | x_n)\Lambda(B_{R_1}(\delta^\Lambda_{\text{LOSE}}) - E) < \epsilon/4. \quad (4.12)$$
Since \( h(\psi, r) \) converges to \( g^A(\psi \mid x_n) \) uniformly on \( E \), there exists \( R_2 > 0 \) such that

\[
|h(\psi, r) - g^A(\psi \mid x_n)| < c_{3\epsilon} = \min(c_{1\epsilon}, c_{2\epsilon})
\]

(4.13)

for all \( \psi \in E \) and \( r \leq R_2 \). Let \( E_{1\epsilon} = \{ \psi : h(\psi, r) > c_{1\epsilon}(x_n) - c_{1\epsilon} - c_{3\epsilon} \} \). A lower bound on \( \Pi_\psi(E_{1\epsilon} \mid x_n) \) is given by

\[
\Pi_\psi(E_{1\epsilon} \mid x_n) \geq \Pi_\psi(E \cap E_{1\epsilon} \mid x_n) \geq \Pi_\psi(E \cap \{ \psi : g^A(\psi \mid x_n) > c_{1\epsilon}(x_n) - c_{1\epsilon} \} \mid x_n)
\]

\[
\geq \Pi_\psi(\{ \psi : g^A(\psi \mid x_n) > c_{1\epsilon}(x_n) - c_{1\epsilon} \} \mid x_n) - \Pi_\psi(E^c \mid x_n)
\]

Here, \( P(A \cap B) \geq P(A) - P(B^c) \) is used. From (4.9) and (4.12), we have

\[
> \gamma + \epsilon/4 - \epsilon/4 = \gamma.
\]

By definition, \( E_{1\epsilon} \) is a lowest posterior loss region having posterior probability bigger than \( \gamma \). Hence, we obtain

\[
O^A_\gamma(x_n, r) \subset E_{1\epsilon}
\]

(4.14)

and \( \alpha^A_\gamma(x_n \mid r) > c^A_\gamma(x_n) - c_{1\epsilon} - c_{3\epsilon} \). An upper bound on \( \Pi_\psi(E_{1\epsilon} \mid x_n) \) is given by

\[
\Pi_\psi(E_{1\epsilon} \mid x_n) \leq \Pi_\psi((E \cap E_{1\epsilon}) \cup E^c \mid x_n)
\]

\[
\leq \Pi_\psi(\{ \psi : g^A(\psi \mid x_n) > c^A_\gamma(x_n) - c_{1\epsilon} - 2c_{3\epsilon} \} \cup E^c \mid x_n)
\]

\[
\leq \Pi_\psi(E^c \mid x_n) + \Pi_\psi(\{ \psi : g^A(\psi \mid x_n) > c^A_\gamma(x_n) \} \mid x_n)
\]

\[
+ \Pi_\psi(\{ \psi : c^A_\gamma(x_n) - c_{1\epsilon} - 2c_{3\epsilon} < g^A(\psi \mid x_n) \leq c^A_\gamma(x_n) \} \mid x_n)
\]

\[
\leq \epsilon/4 + \gamma + \Pi_\psi(\{ \psi : g^A(\psi \mid x_n) \in (c^A_\gamma(x_n) - c_{1\epsilon} - 2c_{3\epsilon}, c^A_\gamma(x_n)) \} \mid x_n).
\]

(4.15)

Let \( E_{2\epsilon} = \{ \psi \in E : h(\psi, r) > c^A_\gamma(x_n) + 2c_{2\epsilon} \} \). For any \( \psi \in E_{2\epsilon} \), we automatically have \( \psi \in E \) and \( |h(\psi, r) - g^A(\psi)| < c_{2\epsilon} \) for all \( r \leq R_2 \) by (4.13). Then,

\[
E_{2\epsilon} \subset \{ \psi : g^A(\psi \mid x_n) > c^A_\gamma(x_n) + c_{2\epsilon} \} \subset \{ \psi : g^A(\psi \mid x_n) > c^A_\gamma(x_n) \} = OS^A_\gamma(x_n).
\]

(4.16)
We obtain

\[ \Pi_\Psi(\{ \psi : h(\psi, r) > c_\gamma^A(x_n) + 2c_2 \} | x_n) \]

\[ \leq \Pi_\Psi(E \cap \{ \psi : h(\psi, r) > c_\gamma^A(x_n) + 2c_2 \} | x_n) + \Pi_\Psi(E^c | x_n) \]

\[ < \Pi_\Psi(\{ \psi : g^A(\psi | x_n) > c_\gamma^A(x_n) + c_2 \} | x_n) + \epsilon/4 \leq \gamma. \]

In the last inequality, (4.8) is used. Thus we get

\[ E_{2\epsilon} \subset O_\gamma^A(x_n, r) \] (4.17)

and \( \alpha_\gamma^A(x_n | r) < c_\gamma^A(x_n) + 2c_2 \). A lower bound on \( \Pi_\Psi(E_{2\epsilon} | x_n) \) is given by

\[ \Pi_\Psi(E_{2\epsilon} | x_n) = \Pi_\Psi(E \cap \{ \psi : h(\psi, r) > c_\gamma^A(x_n) + 2c_2 \} | x_n) \]

\[ \geq \Pi_\Psi(E \cap \{ \psi : g^A(\psi | x_n) > c_\gamma^A(x_n) + 3c_2 \} | x_n) \text{ by (4.13)} \]

\[ \geq \Pi_\Psi(\{ \psi : g^A(\psi | x_n) > c_\gamma^A(x_n) + 3c_2 \} | x_n) - \Pi_\Psi(E^c | x_n) \]

\[ = \Pi_\Psi(\{ \psi : g^A(\psi | x_n) > c_\gamma^A(x_n) \} | x_n) \]

\[ - \Pi_\Psi(\{ \psi : g^A(\psi | x_n) \in (c_\gamma^A(x_n), c_\gamma^A(x_n) + 3c_2] \} | x_n) - \Pi_\Psi(E^c | x_n) \]

\[ \geq \gamma - \Pi_\Psi(\{ \psi : g^A(\psi | x_n) \in (c_\gamma^A(x_n), c_\gamma^A(x_n) + 3c_2] \} | x_n) - \epsilon/4. \] (4.18)

From (4.14), (4.16) and (4.17), we have \( E_{2\epsilon} \subset O_\gamma^A(x_n) \) and \( E_{2\epsilon} \subset O_\gamma^A(x_n, r) \subset E_{1\epsilon} \).

So, we have \( O_\gamma^A(x_n) \Delta O_\gamma^A(x_n, r) \subset (O_\gamma^A(x_n) - E_{2\epsilon}) \cup (O_\gamma^A(x_n, r) - E_{2\epsilon}). \) From all these relationships, we obtain

\[ \Pi_\Psi(O_\gamma^A(x_n) \Delta O_\gamma^A(x_n, r) | x_n) \leq \Pi_\Psi(O_\gamma^A(x_n) - E_{2\epsilon} | x_n) + \Pi_\Psi(E_{1\epsilon} - E_{2\epsilon} | x_n) \]

\[ = \Pi_\Psi(O_\gamma^A(x_n) | x_n) + \Pi_\Psi(E_{1\epsilon} | x_n) - 2\Pi_\Psi(E_{2\epsilon} | x_n) \]

Using (4.15) and (4.18), we have

\[ \leq 3\epsilon/4 + 2\Pi_\Psi(\{ \psi : g^A(\psi | x_n) \in [c_\gamma^A(x_n) - c_{1\epsilon} - 2c_{3\epsilon}, c_\gamma^A(x_n) + 3c_{2\epsilon}] \} | x_n). \]

From the continuity assumption, we get, as \( \epsilon \to 0, \)

\[ \Pi_\Psi(\{ \psi : g^A(\psi | x_n) \in [c_\gamma^A(x_n) - c_{1\epsilon} - 2c_{3\epsilon}, c_\gamma^A(x_n) + 3c_{2\epsilon}] \} | x_n) \]

\[ \to \Pi_\Psi(\{ \psi : g^A(\psi | x_n) \in [c_{\text{range}}, c_\gamma^A(x_n)] \} | x_n) = 0 \]
In the last line, we used (4.10) and (4.11). This proves the theorem.

The condition (SD2) holds for a large class of probability measures. This is stated in the following proposition. The proof is provided in Section 4.6.

**Proposition 4.13.** (i) Every discrete probability measure satisfies condition (SD2).
(ii) Every finite measure \( \Lambda \) having a non-zero continuous density function with respect to Lebesgue measure also satisfies condition (SD2).

Note that the limit Bayes rule \( \delta^\Lambda_{\text{LOSE}} \) depends on the choice of \( \Lambda \). So for a specific problem, we need to choose the measure \( \Lambda \). To obtain invariant procedures it is natural to take \( \Lambda = \Pi_\Psi \) and then we have relative surprise inferences. Note that in this case the loss function is dependent on the prior. Values of \( \psi \) that are in the tails of the prior assigned higher loss as \( \eta \to 0 \) and this seems very natural.

### 4.4 Asymptotic Properties of the LRS Estimator

We investigate large sample properties of the least relative surprise estimator.

#### 4.4.1 Consistency of the LRS Estimator

We proved that the relative surprise inference has some optimal properties in Section 4.2. Also, it is natural to assume that relative surprise inferences will give us the right answer as the sample size increases. However, there are some known counterexamples showing that the posterior may not be consistent even though the sample size increases. Detailed discussions can be found in Freedman (1963, 1965), Diaconis and Freedman (1986a,b), Wasserman (1998) and references therein.

Among them Diaconis and Freedman (1986a) showed that a Bayes rule may not be consistent. Even though a naïve looking prior is placed on a location model, the posterior mean does not converge to the true parameter. As we showed in Section 4.3,
Chapter 4. Invariant Bayesian Inference

134

the least relative surprise estimator is also a limit of Bayes estimators. Then, it might be inconsistent for some cases. So we investigate under what condition the least relative surprise estimator is consistent in this section.

There are two consistencies in Bayesian inference, that is, posterior consistency and consistency of Bayes rules, see Section 3 of Diaconis and Freedman (1986a) for a related discussion. To see the difference, consider a model \( \{P_\theta : \theta \in \Theta\} \) on a space \( \mathcal{X} \). Suppose random variables \( X_1, X_2, \ldots \) are independently sampled from \( P_\theta \) with joint probability measure \( P_\theta^\infty \). Furthermore, a prior \( \Pi \) is placed on \( \Theta \). We say that the posterior is consistent at \( \theta \) if \( \Pi(U | X_1, \ldots, X_n) \to 1 \), a.s.-\( P_\theta^\infty \) for any open neighbourhood \( U \) of \( \theta \), and a Bayes estimator \( \hat{\theta} \) is consistent if \( \hat{\theta} \to \theta \), a.s.-\( P_\theta^\infty \). In many cases, the posterior consistency implies the consistency of a Bayes estimator. The posterior probability contracts around the true parameter value by the posterior consistency. Hence, roughly speaking, a Bayes rule will be near the true parameter value. However, this does not mean that any Bayes estimator is consistent whenever the posterior is consistent. Thus, we have to investigate the consistency of a Bayes rule before making an inference.

The least relative surprise estimator is also a Bayes estimator as shown in Section 4.3. So its consistency must be verified before making an inference based on it. We consider a model \( \mathcal{F} = \{P_\theta : \theta \in \Theta\} \) which is dominated by a \( \sigma \)-finite measure, \( \mu \), where \( \Theta \) is the parameter space. We also assume that a prior \( \Pi \) is placed on \( \Theta \) (or on \( \mathcal{F} \) for some cases) throughout this section.

Let \( \pi(\theta | x_n) \) be the Radon-Nikodym derivative of the posterior, \( \Pi(\cdot | x_n) \), with respect to the prior \( \Pi \). Then, the LRSE, \( \hat{\theta}_{\text{LRSE}} \), is the essential maximizer of the posterior density, \( \pi(\theta | x_n) \). Posterior consistency implies \( \Pi(U^c | x_n) \to 0 \) for every open set \( U \subset \Theta \) containing the true parameter value \( \theta_0 \). This implies

\[
\Pi(U^c | x_n) = \int \pi(\theta | x_n) I(\theta \not\in U) \, \Pi(d\theta) \to 0.
\]

In other words, \( \pi(\theta | x_n) I(\theta \not\in U) \) converges to 0 in the \( L_1(\Pi) \) sense. Thus, it seems like any essential maximizer is in \( U \) eventually. However, it may not be true because \( L_1(\Pi) \)
convergence does not imply a.s.-\Pi convergence. So, some restrictions are required to achieve the consistency of the LRSE.

When the full parameter \( \theta \) is of interest, the LRS estimator is also the MLE as shown in Property 6 of Section 4.2. Hence, the consistency of the LRSE as a point estimator is inherited directly from the consistency of the MLE. We summarize a set of sufficient conditions from Chapter 6 of Lehmann and Casella (1998).

\textbf{(AS0)} The parametrization of \( \mathcal{F} = \{ P_\theta : \theta \in \Theta \} \) is identifiable where \( \Theta \) is an open set in \( \mathbb{R}^k \). And there exists a \( \sigma \)-finite measure \( \mu \) such that \( P_\theta \ll \mu \) for all \( \theta \in \Theta \).

\textbf{(AS1)} The distribution \( P_\theta \) have common support, that is, \( \{ x : \frac{dP_\theta}{d\mu}(x) > 0 \} \) is independent of \( \theta \).

\textbf{(AS2)} The \( X_i \)'s are i.i.d. with \( P_\theta \).

\textbf{(AS3)} There exists an open neighborhood \( H \) of the true parameter value \( \theta_0 \) such that, for almost all \( x \), the density \( f(x \mid \theta) \) admits all third derivatives \( (\partial^3/\partial \theta_j \partial \theta_k \partial \theta_l)f(x \mid \theta) \) for all \( \theta \in H \).

\textbf{(AS4)} The first and second logarithmic derivatives of \( f \) satisfy the equations

\[ E_{\theta} \left( \frac{\partial}{\partial \theta} \log f(X \mid \theta) \right) = 0 \]

and

\[ I(\theta) = \text{var}_{\theta} \left( \frac{\partial}{\partial \theta} \log f(X \mid \theta) \right) = -E_{\theta} \left( \frac{\partial^2}{\partial \theta \partial \theta^T} \log f(X \mid \theta) \right). \]

\textbf{(AS5)} The Fisher information matrix \( I(\theta) \) is positive definite for all \( \theta \in H \).

\textbf{(AS6)} For all \( j, k, l \), there exist functions \( K_{jkl} \) such that \( E_{\theta_0} K_{jkl}(X) < \infty \)

\[ \left| \frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} \log f(x \mid \theta) \right| \leq K_{jkl}(x) \quad \text{for all} \; \theta \in H. \]
**Theorem 4.14.** Assume (AS0)–(AS6). Also assume that $\Pi$ has density $\pi$ and $\pi > 0$ on a neighborhood of $\theta_0$. Then, there exists the least relative surprise estimator in $P_{\theta_0}^\infty$-probability and it is consistent.

**Proof of Theorem 4.14.** The theorem holds immediately from Theorem 6.5.1 of Lehmann and Casella (1998).

We may consider a weaker condition for the consistency. But, for asymptotic normal distribution, we need all assumptions. Also when a marginal parameter $\psi = \Psi(\theta)$ is of interest, we need more assumptions to obtain the consistency for the LRSE.

**Theorem 4.15.** Assume (a) all (AS0)–(AS6), (b) $\Pi$ has density $\pi$ and $\pi > 0$ on a neighborhood of $\theta_0$ and (c) the map $\theta \mapsto \Psi(\theta)$ is continuously differentiable. Then, there exists the least relative surprise estimator for $\psi_0 = \Psi(\theta_0)$ in $P_{\theta_0}^\infty$-probability and it is consistent.

**Proof of Theorem 4.15.** Let $\ell(\theta) = \sum_{i=1}^n \log f(x_i | \theta)$. We prove that, for all $r > 0$, there exists $\psi$ such that $||\psi - \psi_0|| < r$ and $RB(\psi) > \sup_{\psi : ||\psi - \psi_0|| > r} RB(\psi)$ in $P_{\theta_0}^\infty$-probability.

From the proof of Theorem 6.5.1 of Lehmann and Casella (1998), there exists $a > 0$ and $\eta > 0$ such that $n^{-1}(\ell(\theta) - \ell(\theta_0)) < -\eta$ for all $\theta \notin B_a(\theta_0)$ in $P_{\theta_0}^\infty$-probability. Particularly, we can take $a > 0$ small enough to $B_a(\theta_0) \subset \{ \theta : ||\Psi(\theta) - \Psi(\theta_0)|| < r \}$ and $B_a \subset H$. Then for any $\psi$ satisfying $||\psi - \psi_0|| \geq r$,

$$RB(\psi) = \frac{\pi_\Psi(\psi | x_n)}{\pi_\Psi(\psi)} = \frac{\int_{\Psi^{-1}(\psi)} f(x_n | \theta) \pi(\theta) J_\Psi(\theta) / m(x_n) \nu_{\Psi^{-1}(\psi)}(d\theta)}{\int_{\Psi^{-1}(\psi)} \pi(\theta) J_\Psi(\theta) \nu_{\Psi^{-1}(\psi)}(d\theta)}$$

$$= e^{\ell(\theta_0)} \frac{\int_{\Psi^{-1}(\psi)} \exp(\ell(\theta) - \ell(\theta_0)) \pi(\theta) J_\Psi(\theta) / m(x_n) \nu_{\Psi^{-1}(\psi)}(d\theta)}{m(x_n) \int_{\Psi^{-1}(\psi)} \pi(\theta) J_\Psi(\theta) \nu_{\Psi^{-1}(\psi)}(d\theta)}$$

$$\leq e^{\ell(\theta_0) - n\eta / m(x_n)}$$ in $P_{\theta_0}^\infty$-probability.

Hence, $\sup_{\psi : ||\psi - \psi_0|| \geq r} RB(\psi) \leq e^{\ell(\theta_0) - n\eta / m(x_n)}$ in $P_{\theta_0}^\infty$-probability.

Then we show there exists $b > 0$ such that $n^{-1}(\ell(\theta) - \ell(\theta_0)) > -\eta$ for all $\theta \in B_b(\theta_0)$
in $P_{\theta_0}^{\infty}$-probability. By the Taylor expansion of we have, for $\theta \in B_a(\theta_0)$,

$$n^{-1}\ell(\theta) = n^{-1}\ell(\theta_0) + n^{-1}\ell(\theta_0)(\theta - \theta_0) + (2n)^{-1}(\theta - \theta_0)^T\tilde{\ell}(\theta_0)(\theta - \theta_0)$$

for some $\theta_* \in \{\theta_0 + u(\theta - \theta_0) : u \in [0, 1]\}$, where $\ell$ and $\tilde{\ell}$ are the first and second derivative of $\ell$ with respect to $\theta$. By the law of large numbers $n^{-1}\ell(\theta_0)$ converges to 0 a.s.-$P_{\theta_0}^{\infty}$. So we have $|n^{-1}\ell(\theta_0)| < \eta/(4a)$ in $P_{\theta_0}^{\infty}$-probability. Hence, we need to show that the last term is bounded by $3\eta/4$. Since $I(\theta)$ is continuous and bounded on $H$, the supremum eigen value of $I(\theta)$ on $\theta \in B_a(\theta_0)$ is finite, say $I_a$. Let $b = \min(a, (\eta/I_a)^{1/2})$.

Then, for any $\theta \in B_b(\theta_0)$, we have

$$|(2n)^{-1}(\theta - \theta_0)^T\tilde{\ell}(\theta_0)(\theta - \theta_0)| \leq I_a b^2/2 \leq \eta/2 \text{ a.s.-} P_{\theta_0}^{\infty}$$. Hence, $|(2n)^{-1}(\theta - \theta_0)^T\tilde{\ell}(\theta_0)(\theta - \theta_0)| \leq 7\eta/12$ in $P_{\theta_0}^{\infty}$-probability. Thus, we obtain $n^{-1}(\ell(\theta) - \ell(\theta_0)) > -5\eta/6$ for $\theta \in B_b(\theta_0)$ in $P_{\theta_0}^{\infty}$-probability. This implies that,

$$RB(\psi_0) = e^{\ell(\theta_0)}\int_{\Psi} \exp(\ell(\theta) - \ell(\theta_0))\pi(\theta)J_\psi(\theta)\nu_{\psi^{-1}\psi}(d\theta)$$

$$= e^{\ell(\theta_0)}\int_{\Psi} \pi(\theta)J_\psi(\theta)\nu_{\psi^{-1}\psi}(d\theta) \geq e^{\ell(\theta_0) - 5\eta/6} \int_{\Psi} \pi(\theta)J_\psi(\theta)\nu_{\psi^{-1}\psi}(d\theta)$$

From the assumption (b), the ratio of integrals in the above inequality is positive. Then, in $P_{\theta_0}^{\infty}$-probability,

$$\sup_{\psi: ||\psi - \psi_0|| > r} RB(\psi) \geq e^{\ell(\theta_0) - 5\eta/6} \int_{\Psi} \pi(\theta)J_\psi(\theta)\nu_{\psi^{-1}\psi}(d\theta) \geq e^{\ell(\theta_0) - n\eta} \int_{\Psi} \pi(\theta)J_\psi(\theta)\nu_{\psi^{-1}\psi}(d\theta)$$

$$= e^{n\eta/6} \int_{\Psi} \pi(\theta)J_\psi(\theta)\nu_{\psi^{-1}\psi}(d\theta) \to \infty \text{ as } n \to \infty.$$

Hence, there exists the LRS estimator $\hat{\psi}_n$ with $P_{\theta_0}^{\infty}$-probability tending to 1. Besides, $\lim_{n \to \infty} P_{\theta_0}(||\hat{\psi}_n - \psi_0|| < r) = 1$ for all $r > 0$. Therefore, the LRSE is consistent.

### 4.4.2 Bernstein-von Mises Theorem

The posterior density with respect to prior may not be easily computed if the dimension of parameter space is big. Naturally, we want to approximate the relative surprise region in these cases as well as the case that the sample size $n$ is sufficiently large. This topic is
connected to the Bayesian asymptotic normality, so called the Bernstein-von Mises theorem. Roughly speaking, the posterior distribution centered at the maximum likelihood estimator has asymptotically the same distribution of the maximum likelihood centered at the true parameter.

The phenomena were known to be reported by Bernstein (1917) and von Mises (1931) even though it has older history. This is the reason it is called the Bernstein-von Mises theorem. The asymptotic normality for Multinomial\( (n; p_1, \ldots, p_k) \) distributions with an arbitrary prior is shown in section 8.7 of von Mises (1931). After that many researches have shown the Bernstein-von Mises theorems. For example, Le Cam (1953) extended this theorem to any model satisfying a regularity condition. It is embroadened to Markov chains by Borwanker, Kallianpur and Prakasa Rao (1971). Also Chen (1985) and Ibragimov and Hasminskii (1981) showed the Bayesian asymptotic normality in weak sense for a broad class.

The Bayesian asymptotic normality requires the posterior consistency. The posterior concentrates high probability at a wrong value if posterior consistency is not fulfilled.

Most proofs of the Bayesian asymptotic normality show the asymptotic normality of the maximum likelihood estimator, then, show the \( L_1 \) convergence of the posterior density to the limit normal density. In the history of the Bayesian asymptotic normality, the posterior distribution was centred at the maximum likelihood estimator. The posterior distribution centred at the posterior mode also has the asymptotic normality. For example, Ghosh, Delampady and Samanta (2006) showed heuristically, and Chen (1985) showed the weak convergence. Besides Kim and Lee (2004), and James (2008) showed that the posterior normality holds when the posterior is centred at the posterior mean in a non-parametric model. We summarize Bayesian asymptotic normality following Ghosh and Ramamoorthi (2003).

**Theorem 4.16.** Assume (a) (AS0)–(AS6), (b) \( \Pi \) has density \( \pi \) with respect to the Lebesgue measure and (c) \( \pi \) is continuous and positive at \( \theta_0 \). Let \( \hat{\theta} = \hat{\theta}(x_n) \) be the least
relative surprise estimator. Then, in \( P_{\theta_0}^{\infty} \)-probability,
\[
\int \left| \frac{f_{\theta,n}(x_n)\pi(\theta)}{m_n(x_n)} - n^{k/2}\phi_1(\theta_0)^{-1}(n^{1/2}(\theta - \hat{\theta})) \right| \Sigma(d\theta) \to 0
\]
where \( I(\theta) \) is the Fisher information matrix and \( \phi_\Sigma \) is the density of \( N(0, \Sigma) \) distribution.

Proof of Theorem 4.16. Note that \( \hat{\theta} \) is also the maximum likelihood estimator. For a proof, see Theorem 1.4.2 of Ghosh and Ramamoorthi (2003).

Theorem 4.16 says that the Bernstein-von Mises theorem holds when a consistent posterior is centered at the LRSE. Simply speaking \( \sqrt{n}(\theta - \hat{\theta})|x_n \to N(0, I(\hat{\theta})^{-1}) \). It is similar to the asymptotic normality of the MLE given by \( \sqrt{n}(\hat{\theta} - \theta_0) \to N(0, I(\theta_0)^{-1}) \).

Using Theorem 4.16, an approximate \( \gamma \)-relative surprise region is obtained by
\[
C_\gamma(x_n) \approx \{ \theta : n(\theta - \hat{\theta})^T I(\hat{\theta})(\theta - \hat{\theta}) \leq \chi^2_\gamma(k) \}.
\]
where \( \chi^2_\gamma(k) \) is the \( \gamma \)-quantile of the \( \chi^2(k) \) distribution. This asymptotic \( \gamma \)-relative surprise regions are also asymptotic \( \gamma \)-confidence region in frequentist sense.

### 4.5 Discussion

Since Bayes (1763) founded Bayesian data analysis, lots of procedures had been developed under the Bayesian philosophy. Among them, it is not easy to find an invariant inference under reparametrizations in a continuous parameter space. We have focused here on Bayes factors and relative surprise regions. Especially, relative surprise regions are invariant under reparametrization.

Relative surprise inferences connect several inference methodologies. Relative surprise regions can be interpreted as the lowest posterior loss region using a sequence of loss functions. So relative surprise inferences are related to decision theoretic Bayesian inference. For point estimation, the least relative surprise estimator is exactly the maximum likelihood estimator when the full parameter is of interest. Also large sample properties of the least relative surprise estimator are applicable for asymptotic credible regions.
Relative surprise inferences do not solve every problem. Without posterior consistency, posterior credible regions may not be reliable, that is, credible region may not contain the true parameter or it is too wide (it does not shrink even though the sample size increases). So the use of a prior having the posterior consistency is still required to achieve a reliable inference. Especially, a prior must be chosen very carefully if the number of nuisance parameter increases.

There are still many issues to be solved. For example, the choice of prior has not been solved yet and problems with a high dimension nuisance parameter must be dealt with. Even though many unsolved issues are still lingering, the relative surprise inference has many desirable properties and unites many methodologies.

4.6 Proofs

Proof of Theorem 4.1

Assume \( \Psi(\Theta) \) is a finite dimensional Riemann manifold. Note that \( \Psi(\Theta) \) itself is a locally compact Hausdorff space and so is \( \Xi(\Theta) \) because \( h : \Psi(\Theta) \to \Xi(\Theta) \) is continuous. Then, \( \Pi_\Psi \) and \( \Pi_\Xi \) are finite Radon measures. By Theorem 20.3 of Hewitt and Stromberg (1965), there exists a function \( w : \Xi(\Theta) \to \mathbb{R} \) such that, for any Lebesgue measurable function \( f \) on \( \Psi(\Theta) \),

\[
\int f(\psi) \pi_\Psi(\psi) \nu_\Psi(d\psi) = \int f(\psi) \Pi_\Psi(d\psi) = \int f(h^{-1}(\xi)) w(\xi) \Pi_\Xi(d\xi) \\
= \int f(h^{-1}(\xi)) w(\xi) \pi_\Xi(\xi) \nu_\Xi(d\xi).
\]

By taking \( f(\psi) = I_B(\psi) \) for any Borel set \( B \in \mathcal{B}(\Psi(\Theta)) \) and \( A = h(B) \in \mathcal{B}(\Xi(\Theta)) \), we have

\[
\int_B \pi_\Psi(\psi) \nu_\Psi(d\psi) = \int_A w(\xi) \pi_\Xi(\xi) \nu_\Xi(d\xi).
\]
In a similar vein, by letting \( f(\psi) = I_B(\psi) \pi_\psi(\psi | \mathbf{x}_n)/\pi_\psi(\psi) \), we have
\[
\int_A w(\xi) \pi_\xi(\xi | \mathbf{x}_n) \nu_\xi(d\xi) = \int_B \pi_\psi(\psi | \mathbf{x}_n) \nu_\psi(d\psi).
\]
As \( B \) shrinking to a point \( \psi \), we have
\[
\frac{\pi_\psi(\psi | \mathbf{x}_n)}{\pi_\psi(\psi)} = \frac{w(\xi) \pi_\xi(\xi | \mathbf{x}_n)}{w(\xi) \pi_\xi(\xi)} = \frac{\pi_\xi(\xi | \mathbf{x}_n)}{\pi_\xi(\xi)}.
\]
Hence, the theorem follows.

**Proof of Lemma 4.4**

If \( m(\mathbf{x}) = \int f_\theta(\mathbf{x}) \Pi(d\theta) = \infty \), then any posterior distribution cannot be defined. So the case \( m(\mathbf{x}) = \infty \) is discarded. If \( m(\mathbf{x}) = 0 \), then such \( \mathbf{x} \) is in a \( M \)-null set because \( M(\{ \mathbf{x} : m(\mathbf{x}) = 0 \}) = \int_{\{\mathbf{x} : m(\mathbf{x}) = 0\}} m(\mathbf{x}) \mu(d\mathbf{x}) = 0 \). Thus we just consider \( 0 < m(\mathbf{x}) < \infty \).

Let \( N \) be a \( \Pi \)-null set. Then,
\[
\Pi(N | \mathbf{x}) = \frac{1}{m(\mathbf{x})} \int_N f_\theta(\mathbf{x}) \Pi(d\theta) = 0
\]
because \( f_\theta(\mathbf{x}) \) is a non-negative integrable function with respect to \( \Pi \) and \( N \) is \( \Pi \)-null set. It proves \( \Pi(\cdot | \mathbf{x}) \ll \Pi \).

Now consider a marginal parameter \( \psi = \Psi(\theta) \) is of interest. Let \( N_\Psi \) be a \( \Pi_\Psi \)-null set, that is, \( \Pi_\Psi(N_\Psi) = 0 \). Since \( 0 = \Pi_\Psi(N_\Psi) = \Pi(\Psi^{-1}(N_\Psi)) \), we have \( \Pi_\Psi(N_\Psi | \mathbf{x}) = \Pi(\Psi^{-1}(N_\Psi | \mathbf{x}) = 0 \) from the above result. Hence, \( \Pi_\Psi(\cdot | \mathbf{x}) \ll \Pi_\Psi \).

**Proof of Lemma 4.10**

First of all, we prove \( \Pi_\Psi(C_\gamma(\mathbf{x}_n)) \) is differentiable with respect to \( \gamma \). Suppose \( 1 > \eta > \gamma > 0 \). Then \( C_\gamma(\mathbf{x}_n) \subset C_\eta(\mathbf{x}_n) \) and \( c_\eta(\mathbf{x}_n) < \pi_\psi(\psi | \mathbf{x}_n)/\pi_\psi(\psi) \leq c_\gamma(\mathbf{x}_n) \) for \( \psi \in C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n) \). Hence, we have \( \eta - \gamma = \Pi_\Psi(C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n) | \mathbf{x}_n) \) and
\[
\eta - \gamma = \Pi_\Psi(C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n) | \mathbf{x}_n) = \int_{C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n)} \frac{\pi_\psi(\psi | \mathbf{x}_n)}{\pi_\psi(\psi)} \Pi_\psi(d\psi)
\]
\[
< \int_{C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n)} c_\gamma(\mathbf{x}_n) \Pi_\psi(d\psi) = c_\gamma(\mathbf{x}_n) \Pi_\psi(C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n)),
\]
\[
\eta - \gamma \geq \int_{C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n)} c_\eta(\mathbf{x}_n) \Pi_\psi(d\psi) = c_\eta(\mathbf{x}_n) \Pi_\psi(C_\eta(\mathbf{x}_n) - C_\gamma(\mathbf{x}_n)).
\]
By summarizing two inequalities, we obtain
\[
\frac{1}{c_\gamma(x_n)} < \frac{\Pi_\psi(C_\eta(x_n)) - \Pi_\psi(C_\gamma(x_n))}{\eta - \gamma} \leq \frac{1}{c_\eta(x_n)}.
\]
By the continuity of $c_\gamma(x_n)$, $\Pi_\psi(C_\gamma(x_n))$ is differentiable with respect to $\gamma$ and the derivative of $\Pi_\psi(C_\gamma(x_n))$ is
\[
\frac{\partial \Pi_\psi(C_\gamma(x_n))}{\partial \gamma} = \lim_{h \to 0} \frac{\Pi_\psi(C_{\gamma+h}(x_n)) - \Pi_\psi(C_\gamma(x_n))}{h} = \frac{1}{c_\gamma(x_n)}.
\]
Since $\Pi_\psi(C_\gamma(x_n)|x_n) = \gamma$, the Bayes factor of $C_\gamma(x_n)$ is
\[
BF(C_\gamma(x_n)) = \frac{\Pi_\psi(C_\gamma(x_n)|x_n)}{1 - \Pi_\psi(C_\gamma(x_n)|x_n)} = 1 - \left(\frac{1 - \Pi_\psi(C_\gamma(x_n))}{\Pi_\psi(C_\gamma(x_n))}\right).
\]
Thus the derivative of $BF(C_\gamma(x_n))$ with respect to $\gamma$ is
\[
\frac{\partial BF(C_\gamma(x_n))}{\partial \gamma} = \frac{1 - \Pi_\psi(C_\gamma(x_n))}{\Pi_\psi(C_\gamma(x_n))} \frac{\partial}{\partial \gamma} \gamma + \frac{\gamma}{1 - \gamma} \frac{\partial}{\partial \gamma} \left(\frac{1 - \Pi_\psi(C_\gamma(x_n))}{\Pi_\psi(C_\gamma(x_n))}\right)
\]
\[
= \frac{\gamma(1 - \gamma)}{(1 - \gamma)^2 \Pi_\psi(C_\gamma(x_n))} - \frac{1}{1 - \gamma \Pi_\psi(C_\gamma(x_n))^2 c_\gamma(x_n)}
\]
Hence $BF(C_\gamma(x_n))$ is increasing if and only if $\partial BF(C_\gamma(x_n))/\partial \gamma \geq 0$ if and only if
\[
c_\gamma(x_n) \geq \frac{\gamma(1 - \gamma)}{\Pi_\psi(C_\gamma(x_n)) (1 - \Pi_\psi(C_\gamma(x_n)))} = \frac{\Pi_\psi(C_\gamma(x_n)|x_n)(1 - \Pi_\psi(C_\gamma(x_n)|x_n))}{\Pi_\psi(C_\gamma(x_n)) (1 - \Pi_\psi(C_\gamma(x_n)))}
\]
\[
= RB(C_\gamma(x_n))RB(C_\gamma(x_n)^c).
\]
Therefore, the lemma holds.

**Proof of Proposition 4.13**

(i) It is enough to show this for the points $\psi_0$ such that $\Lambda(\{\psi_0\}) > 0$. Fix $\epsilon \in (0, 1/2)$. There exists an $R > 0$ such that $\Lambda(B_{2R}(\psi_0)) < (1 + \epsilon)\Lambda(\{\psi_0\})$. For all $r < R$ and $\psi \in B_r(\psi_0)$, we have $\psi_0 \in B_r(\psi)$ and $\Lambda(B_r(\psi)) \geq \Lambda(\{\psi_0\})$. So for $\psi \in B_r(\psi_0), B_r(\psi) \subset B_{2R}(\psi_0)$, then
\[
\frac{\Lambda(B_r(\psi))}{\Lambda(B_r(\psi_0))} \leq \frac{\Lambda(B_{2R}(\psi_0))}{\Lambda(B_r(\psi_0))} < \frac{(1 + \epsilon)\Lambda(\{\psi_0\})}{\Lambda(\{\psi_0\})} = 1 + \epsilon.
\]
Also, we have
\[
\frac{\Lambda(B_r(\psi))}{\Lambda(B_r(\psi_0))} \geq \frac{\Lambda(\{\psi_0\})}{\Lambda(B_{2R}(\psi_0))} > \frac{1}{1 + \epsilon} > 1 - \epsilon.
\]

Hence, (SD2) holds.

(ii) Let $\mathcal{L}$ be the Lebesgue measure on $\Psi$ and $\lambda$ be the density of $\Lambda$ with respect to $\mathcal{L}$. Fix $\epsilon \in (0, 1)$. The assumption says that $\lambda(\psi_0) > 0$ for every $\psi_0 \in \Psi$. Since $\lambda$ is continuous at each point, there is an $R > 0$ such that $|\lambda(\psi) - \lambda(\psi_0)| < \lambda(\psi_0)\epsilon/3$ for all $\psi \in B_R(\psi_0)$. Then, for all $0 < r < R/2$ and $\psi' \in B_{R/2}(\psi_0)$,
\[
\Lambda(B_r(\psi')) = \int_{B_r(\psi')} \lambda(\psi') \mathcal{L}(d\psi') \leq (1 + \epsilon/3)\lambda(\psi_0)\mathcal{L}(B_r(\psi')).
\]
In a similar vein, $\Lambda(B_r(\psi')) \geq (1 - \epsilon/3)\lambda(\psi_0)\mathcal{L}(B_r(\psi'))$. Since $\mathcal{L}(B_r(\psi')) = \mathcal{L}(B_r(\psi_0))$ and
\[
(1 - \epsilon/3)\lambda(\psi_0)\mathcal{L}(B_r(\psi_0)) \leq \Lambda(B_r(\psi_0)) \leq (1 + \epsilon/3)\lambda(\psi_0)\mathcal{L}(B_r(\psi_0)),
\]
we have for $\psi \in B_r(\psi_0)$
\[
1 - \epsilon \leq \frac{1 - \epsilon/3}{1 + \epsilon/3} \leq \frac{\Lambda(B_r(\psi))}{\Lambda(B_r(\psi_0))} \leq \frac{1 + \epsilon/3}{1 - \epsilon/3} \leq 1 + \epsilon.
\]
Therefore, (SD2) holds.

**Other Lemmas**

**Lemma 4.17.** Let $T$ be a consistent estimator of $\theta$ and $\Pi$ be a prior on the parameter space. Then, $M_{T,n} \xrightarrow{d} \Pi$.

*Proof of Lemma 4.17.* Let $U$ be a $\Pi$-continuity set. By the consistency, $P_{\theta,T,n}(U) \rightarrow 1$ for $\theta \in U^\circ$ and $P_{\theta,T,n}(U) \rightarrow 0$ for $\theta \notin U$. Thus,
\[
M_{T,n}(U) = \int_U \int_{\Theta} f_{\theta,T,n}(t)d\Pi(\theta)d\mu(t) = \int_{\Theta} \int_U f_{\theta,T,n}(t)d\mu(t)d\Pi(\theta) = \int_{\Theta} P_{\theta,T,n}(U)\Pi(d\theta) \\
\rightarrow \int_{\Theta} I(\theta \in U)\Pi(d\theta) = \Pi(U).
\]

By Theorem 3, Appendix III, of Billingsley (1968), $M_{T,n} \xrightarrow{d} \Pi$ as $n \rightarrow \infty$. \qed
Lemma 4.18. Suppose a finite Borel measure $\Lambda$ is decomposed into two parts by the Lebesgue decomposition theorem, that is, $\Lambda = \Lambda_s + \Lambda_a$ where $\Lambda_s$ is singular to the Lebesgue measure $\mathcal{L}$ and $\Lambda_a$ is absolutely continuous with respect to $\mathcal{L}$. If the Radon-Nikodym derivative $g(\psi) = \frac{d\Lambda}{d\mathcal{L}}(\psi)$ is continuous, there exists $\eta(r, \Lambda)$ such that
\[
\lim_{s \to 0} \int_{0 < r < s} \{ \psi : \Lambda(B_r(\psi)) \geq \eta(r, \Lambda) \} = \text{supp}(\Lambda).
\]

Proof of Lemma 4.18. Let $g(\psi) = \frac{d\Lambda}{d\mathcal{L}}(\psi)$, $A(0, \epsilon) = \{ \psi : g(\psi) > \epsilon \}$, and $A(r, \epsilon) = \{ \psi : B_r(\psi) \subset A(0, \epsilon) \}$. So $A(0, \epsilon)$ is an open set having density bigger than $\epsilon$ and $A(r, \epsilon)$ is the inside subset of $A(0, \epsilon)$ with distance at least $r$ from the boundary of $A(0, \epsilon)$. Then, $A(0, \epsilon)$ increases to $A(0, 0)$ as $\epsilon$ decreases to 0. Also for fixed $\epsilon \geq 0$, $A(r, \epsilon)$ increases to $A(0, \epsilon)$ as $r$ decreases to 0. Suppose $r \geq s \geq 0$ and $\epsilon \geq \eta \geq 0$. For $\psi \in A(r, \epsilon)$, the density satisfies $g(\psi) > \epsilon \geq \eta$. So $\psi \in A(0, \eta)$. And $B_s(\psi) \subset B_r(\psi) \subset A(0, \epsilon) \subset A(0, \eta)$. Thus, $A(r, \epsilon) \subset A(s, \eta)$ whenever $r \geq s \geq 0$ and $\epsilon \geq \eta \geq 0$. Then, for any decreasing sequence $r_n$ and $\epsilon_n$, we have $A(r_n, \epsilon_n) \subset A(r_{n+1}, \epsilon_{n+1})$. For any $\psi \in A(0, 0)$, we have $r > 0$ such that $B_{2r}(\psi) \subset A(0, 0)$ because $A(0, 0)$ is open. Since $g$ is continuous and $\overline{B_r(\psi)}$ is a compact set, the density has a minimum on this compact set, say the minimum density $\epsilon > 0$. Then, for $\psi^* \in B_r(\psi)$, $g(\psi^*) \geq \epsilon > \epsilon/2 > 0$. Thus, $B_r(\psi) \subset A(r, \epsilon/2)$. Since both $r_n$ and $\epsilon_n$ decrease, there exists $N > 0$ such that $r_n < r$ and $\epsilon_n < \epsilon/2$ for all $n \geq N$. Thus $\psi \in A(r_n, \epsilon_n)$ for all $n \geq N$. So we obtain $A(0, 0) \subset \bigcup_{n=1}^\infty A(r_n, \epsilon_n)$.

Define $\eta(r, \Lambda) = \epsilon_n \mathcal{L}(B_r(\psi))$ for $r_n \leq r < r_{n-1}$. Then, for $0 < r \leq r_n$ and $\psi \in A(r_n, \epsilon_n)$, we have $B_r(\psi) \subset A(0, \epsilon_n)$ and
\[
\Lambda(B_r(\psi)) = \int_{B_r(\psi)} g(\psi^*) \mathcal{L}(d\psi^*) \geq \int_{B_r(\psi)} \epsilon_n \mathcal{L}(d\psi^*) = \epsilon_n \mathcal{L}(B_r(\psi)) \geq \eta(r, \Lambda).
\]
Thus $\Psi = \{ \psi : \Lambda(B_r(\psi)) \geq \eta(r, \Lambda) \} \supset A(r_n, \epsilon_n)$. Then, for any $s > 0$, $\cap_{0 < r < s} \Psi \supset A(r_n, \epsilon_n)$ for all $n$ such that $r_n \geq s$. So we obtain
\[
\bigcup_{s > 0} \cap_{0 < r < s} \Psi \supset \bigcup_{s > 0} \bigcup_{n : r_n \geq s} A(r_n, \epsilon_n) = \bigcup_{n=1}^\infty A(r_n, \epsilon_n) \supset A(0, 0).
\]
Therefore, $\Lambda([\bigcup_{s > 0} \cap_{0 < r < s} \Psi]) \leq \Lambda(A(0, 0)) = \int_{A(0, 0)} g(\psi) \mathcal{L}(d\psi) = 0$ because $g(\psi) = 0$ on $A(0, 0)^c$. \qed
Chapter 5

Inferences for Infinite Dimensional Models

This chapter is concerned with ongoing research concerning Bayesian inferences on infinite dimensional models.

There are many problems where infinitely many parameters arise. For example, in a measurement error model, the true value cannot be measured aside from a surrogate or auxiliary variable. If the true value of each study object is treated as a fixed parameter, the model will contain an increasing number of parameters. If the measurement error of each observation is treated as independent and identically distributed random variable, then this model requires nonparametric or semiparametric function estimation.

In this chapter, we discuss some partial results and future research directions concerning these models.

5.1 Consistency of the Sieve LRS Estimator

We discuss large sample properties of the relative surprise inference for parametric models, in Section 4.4. In this section, we present the consistency of the sieve least relative surprise estimator and a relationship between consistency of the LRSE and posterior
consistency.

Nonparametric models are too big to be parametrized with finite dimensional parameters. So we need a metric on the model that is different from usual Euclidean distance. We consider the Hellinger distance between two probability measures \( P \) and \( Q \) as given by

\[
d_H(P, Q) = \int_X \left( \left( \frac{dP}{d\mu}(x) \right)^{1/2} - \left( \frac{dQ}{d\mu}(x) \right)^{1/2} \right)^2 \mu(dx)
\] (5.1)

where \( \mu \) is a common dominating measure of \( P \) and \( Q \) such as \( \mu = P + Q \). To state a condition for the LRSE to have consistency, we also need the bracketing Hellinger metric entropy of the model \( F \). In this argument, we interpret \( F \) as a collection of densities rather than probability measures. In the theory of empirical processes, entropies are defined as a measurement of the size of a class. From the abundant literature on this subject, we use as references van der Vaart and Wellner (1996). Suppose there is a metric \( d \) on \( F \). The bracket \([l, u]\) for given two functions \( l \) and \( u \) is the set of all functions such that \( l \leq f \leq u \). The bracketing number \( N_{[\cdot]}(\epsilon, F, d) \) is the minimum number of brackets \([l, u]\) with \( d(l, u) < \epsilon \) needed to cover \( F \). The \( \epsilon \)-bracketing entropy is the logarithm of the bracketing number. We consider the Hellinger metric defined by (5.1) on \( F \) and let

\[
H(\epsilon, F) = \log N_{[\cdot]}(\epsilon, F, d_H).
\]

The entropy number \( H(\epsilon, F) \) can be infinite if the model \( F \) is too big. Then, an appropriate estimation on \( F \) may not be feasible due to the size of the model. For instance, there is no maximum likelihood estimator in the density estimation problem. Suppose \( x_1, \ldots, x_n \in \mathbb{R} \) are observed. For \( 0 < \epsilon < \min_{i \neq j} |x_i - x_j|/2 \), the likelihood of \( f_\epsilon(x) = (2n\epsilon)^{-1} \sum_{i=1}^n I(|x - x_i| < \epsilon) \) is \( L(f_\epsilon | x_n) = (2n\epsilon)^{-n} \). Hence, the likelihood \( L(f_\epsilon | x) \) increases to infinity as \( \epsilon \) decreases to 0. This kind of difficulty arises when the model is big. For such a problem, a sieve method is proposed, that is, for a sequence of nested models given by \( F_1 \subset F_2 \subset \cdots \), an estimation method consists of finding a sequence of estimators \( \hat{\theta}_n \) restricted on \( F_n \) where \( n \) is the sample size. The nested models are called a sieve and the sequence of estimators is called the sieve estimator.
Chapter 5. Inferences for Infinite Dimensional Models

**Theorem 5.1.** Suppose a prior $\Pi$ is placed on a model $\mathcal{F}$ that is absolutely continuous with respect to a $\sigma$-finite measure $\mu$. Assume that

(i) $\Pi$ has the Kullback-Leibler property, that is, for each $r > 0$, $\Pi(\{f \in \mathcal{F} : KL(f_0, f) < r\}) > 0$ for all $f_0 \in \mathcal{F}$,

(ii) the posterior is consistent at $f_0$ with respect to a topology $T$ on $\mathcal{F}$,

(iii) for any $\epsilon > 0$, there exist $c > 0$ and an increasing sieve $\mathcal{F}_n \subset \mathcal{F}$ such that $\bigcup_{n=1}^{\infty} \mathcal{F}_n = \mathcal{F}$, $\sum_{n=1}^{\infty} \Pi(\mathcal{F}_n^c) < \infty$ and, for a constant $c > 0$,

$$\int_{c^2}^{\epsilon} H^{1/2}(u, \mathcal{F}_n) du \leq cn^{1/2}\epsilon^2. \quad (5.2)$$

Then, the sieve LRSE is consistent at $f_0$ with respect to the topology $T$.

**Proof of Theorem 5.1.** In this proof, we treat $\mathcal{F}$ itself as a parameter space equipped with the Hellinger metric. The posterior density with respect to the prior is given by

$$\pi(f | x_n) = \frac{1}{m(x_n)} \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{f(x_i)}{f_0(x_i)} \left/ \int \prod_{i=1}^{n} \frac{g(x_i)}{f_0(x_i)} \Pi(dg) \right. \quad (5.3)$$

The denominator term is common for all $f \in \mathcal{F}$ when $n$ is fixed. Thus, the maximizer of the numerator term is the least relative surprise estimator.

Theorem 1 of Wong and Shen (1995) implies that there exists $c_1, c_2 > 0$ such that

$$P_0 \left( \sup_{f \in \mathcal{F}_n, d_H(f, f_0) > \epsilon^2} \prod_{i=1}^{n} \frac{f(X_i)}{f_0(X_i)} \geq \exp(-c_1 n \epsilon^2) \right) \leq 4 \exp(-c_2 n \epsilon^2).$$

By the Borel-Cantelli lemma, we have that $\sup_{f \in \mathcal{F}_n, d_H(f, f_0) > \epsilon^2} (\prod_{i=1}^{n} (f(X_i)/f_0(X_i))) \leq \exp(-c_1 n \epsilon^2) \quad \text{a.s.-} P_0^\infty$.

A lower bound of the denominator in (5.3) can be obtained by Lemma 4.4.1 of Ghosh and Ramamoorthi (2003). That is, under the assumption (i), we have

$$\int \prod_{i=1}^{n} \frac{f(X_i)}{f_0(X_i)} \Pi(df) \geq \exp(-c_1 n \epsilon^2/2). \quad \text{a.s.-} P_0^\infty$$

Thus, we have that

$$\lim_{n \to \infty} \sup_{f \in \mathcal{F}_n, d_H(f, f_0) > \epsilon^2} \pi(f | x_n) \leq \lim_{n \to \infty} \frac{\exp(-c_1 n \epsilon^2)}{\exp(-c_1 n \epsilon^2/2)} = 0. \quad \text{a.s.-} P_0^\infty.$$
Let \( U_\epsilon = \{ f \in \mathcal{F} : d_H(f_0, f) \leq \epsilon^2 \} \). Then, the posterior consistency assumption implies that \( \Pi(U_\epsilon | x_n) \to 1 \). Hence, there exists \( f_n \in U_\epsilon \) such that \( \pi(f_n | x_n) \geq \Pi(U_\epsilon | x_n) / \Pi(U_\epsilon) > 0 \) for sufficiently large \( n \). Since \( \Pi(F_n^c \text{ i.o.}) = 0 \) from the assumption (iii), the maximizer in \( F_n \) is the same as the global maximizer a.s.-\( \Pi \). In other words, the least relative surprise estimator is eventually in \( U_\epsilon \). Then \( \lim_{n \to \infty} d_H(\hat{f}_{\text{LRSE}}, f_0) \leq \epsilon \).

As \( \epsilon \to 0 \), we have the consistency of the sieve LRSE.

If we replace \( \sum_{n=1}^{\infty} \Pi(F_n) < \infty \) by a stronger condition \( \Pi(F_n) < e^{-nr} \) for a constant \( r > 0 \), then we obtain posterior consistency with respect to the Hellinger distance, for details see Wasserman (1998). Hence, the consistency of the LRSE seems weaker than the posterior consistency with respect to the Hellinger distance. But we don’t have any definite answer. We find a relationship between the consistency of the LRS estimator and posterior consistency in Theorem 5.2.

**Theorem 5.2.** Suppose a prior \( \Pi \) is placed on a model \( \mathcal{F} \) that is absolutely continuous with respect to a \( \sigma \)-finite measure \( \mu \). Assume that

(i) a metric \( d \) is defined on the model \( \mathcal{F} \), let \( \mathcal{T} \) be the topology generated by \( d \),

(ii) for fixed \( x \in \mathcal{X} \), the density function \( f(x) \) is continuous with respect to the topology \( \mathcal{T} \) a.e.-\( \mu \),

(iii) the LRS estimator, \( \hat{f}_{\text{LRSE}} \), is consistent and the relative surprise region \( C_\gamma(x_n) \) satisfies \( \limsup_{n \to \infty} C_\gamma(x_n) = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} C_\gamma(x_n) = \{ f_0 \} \) a.s.-\( P_0^\infty \) for all \( \gamma \in (0, 1) \).

Then, the posterior is consistent with respect to the topology \( \mathcal{T} \).

**Proof of Theorem 5.2.** The \( \gamma \)-relative surprise region is defined by \( C_\gamma(x_n) = \{ f \in \mathcal{F} : \pi(f | x_n) > c_\gamma(x_n) \} \) where \( \pi(f | x_n) = \frac{\Pi(c | x_n)(f)}{\Pi(f)} = \frac{\prod_{i=1}^{n} f(x_i)}{m(x_n)} \) and \( m(x_n) = \int \prod_{i=1}^{n} f(x_i) \Pi(df) \). Hence, \( C_\gamma(x_n) = \{ f \in \mathcal{F} : \Pi_{i=1}^{n} f(x_i) > \alpha_\gamma(x_n) \} \) where \( \alpha_\gamma(x_n) = c_\gamma(x_n)m(x_n) \). This implies that every \( C_\gamma(x_n) \) is an open set, so is \( \cup_{m=n}^{\infty} C_\gamma(x_m) \).

Consider an open neighbourhood \( U \) of \( f_0 \). Fix \( \gamma \in (0, 1) \). Then, there exists \( N > 0 \) such that \( \cup_{m=n}^{\infty} C_\gamma(x_m) \subset U \) for all \( n \geq N \). Suppose there is no such \( N \). Then, there
exist \( f_k \in \mathcal{F} \) and \( n_k \) such that \( f_k \in C_\gamma(x_{n_k}) \setminus U \). Take \( \epsilon > 0 \) such that \( B_{f_0}(\epsilon) = \{ f : d(f_0, f) < \epsilon \} \subset U \). The condition (iii) implies \( d(f_k, f_0) \to 0 \). Hence, there is \( N_0 \) such that \( d(f_k, f_0) < \epsilon \) for all \( k \geq N_0 \). For \( k \geq N_0 \), \( f_k \in B_{f_0}(\epsilon) \subset U \). It makes a contradiction. Thus, there exists \( N > 0 \) such that \( \bigcup_{m=n}^\infty C_\gamma(x_m) \subset U \) for all \( n \geq N \).

It is easy to check that, for \( n > N \),

\[
\Pi(U \mid x_n) \geq \Pi(\bigcup_{m=n}^\infty C_\gamma(x_m) \mid x_n) \geq \Pi(C_\gamma(x_n) \mid x_n) \geq \gamma.
\]

As \( \gamma \to 1 \), we have the result. \( \square \)

Theorem 5.2 shows that the consistency of the LRSE together with the convergence of the relative surprise regions implies posterior consistency. In the general theory of posterior consistency, the Kullback-Leibler property is usually assumed, see Barron (1998), Barron, Schervish and Wasserman (1999), Ghosal, Ghosh and Ramamoorthi (1999), or Walker (2004). However, Theorem 5.2 does not need the Kullback-Leibler property. If the convergence of relative surprise regions fails to hold, then the posterior may not be consistent even though the LRSE is consistent, see Example 5.2.

Many statistical problems contain nuisance parameters. In that case, posterior consistency is easily achieved from the posterior consistency of the full parameter. Assume that \( \psi = \Psi(\theta) \) is the parameter of interest and \( \Psi : \Theta \to \Psi(\Theta) \) is continuous with respect to the topology on \( \Theta \). Then, for any open neighbourhood \( U_\psi \) of \( \psi \), we obtain \( \Pi_\psi(U_\psi \mid x_n) = \Pi(\Psi^{-1}(U_\psi) \mid x_n) \to 1 \) a.s.\( -P_\theta^\infty \) because \( \Psi^{-1}(U_\psi) \) is also an open neighbourhood of \( \theta \). Similarly, the LRSE of \( \psi \) is consistent under a mild condition. Considering that the true distribution is \( P_\theta \), rather than a distribution that depends only on the marginal parameter \( \psi \), the continuity condition of \( \Psi \) can be weakened to the condition that \( \Psi^{-1}(\psi) \) contains an open set containing \( \theta \).

**Theorem 5.3.** The parameter of interest is \( \psi = \Psi(f) \) for \( f \in \mathcal{F} \). Assume (a) all conditions in Theorem 5.1, (b) \( \Psi(\mathcal{F}) \) is equipped with a metric \( d_\psi \) such that, for any \( \epsilon > 0 \),
and fixed $f_0$, $\inf\{d_H(f_0, f) : d_{\Psi}(\Psi(f_0), \Psi(f)) > \epsilon\} > 0$ and (c) $\mathcal{F}_n$’s are $\Psi$-measurable sets. Then, the sieve LRSE $\hat{\psi}$ for $\psi$ is consistent.

**Proof of Theorem 5.3.** Let $\pi_{\Psi}(\psi \mid x_n) = \frac{d\Pi_{\Psi(\cdot \mid x_n)}}{d\Pi_{\psi}}(\psi)$. For any $\Psi$-measurable set $B$, we have

$$\int_B \pi_{\Psi}(\psi \mid x_n) \Pi_{\Psi}(d\psi) = \Pi_{\Psi}(B \mid x_n) = \Pi(\Psi^{-1}(B) \mid x_n) = \int_{\Psi^{-1}(B)} \pi(f \mid x_n) \Pi(df)$$

$$\leq \int_{\Psi^{-1}(B)} \sup_{g \in \Psi^{-1}(\Psi(f))} \pi(g \mid x_n) \Pi(df) = \int_B \sup_{g \in \Psi^{-1}(\psi)} \pi(g \mid x_n) \Pi_{\Psi}(d\psi).$$

Thus, we obtain

$$\pi_{\Psi}(\psi \mid x_n) \leq \sup_{f \in \Psi^{-1}(\psi)} \pi(f \mid x_n). \tag{5.4}$$

For any $\epsilon > 0$, there exists $\delta > 0$ such that $d_H(f_0, f) > \delta$ for all $f$ satisfying $d_{\Psi}(\Psi(f_0), \Psi(f)) > \epsilon$ by the assumption (b). Then, the same argument in the proof of Theorem 5.1 shows that

$$\lim_{n \to \infty} \sup_{f \in \mathcal{F}_n : d_H(f_0, f) > \delta} \pi(f \mid x_n) = 0. \quad \text{a.s.}-P_0^\infty.$$ 

Then, using (5.4), we have

$$\lim_{n \to \infty} \sup_{\psi \in \Psi(\mathcal{F}_n) : d_{\Psi}(\psi, \Psi(f_0)) > \epsilon} \pi_{\Psi}(\psi \mid x_n) \leq \lim_{n \to \infty} \sup_{f \in \mathcal{F}_n : d_H(f, f_0) > \delta} \pi(f \mid x_n) = 0. \quad \text{a.s.}-P_0^\infty.$$ 

For any neighborhood $U_{\psi_0}$ of $\psi_0$ in $\{\psi : d_{\Psi}(\psi, \psi_0) < \epsilon\}$, we have $\Pi_{\Psi}(U_{\psi_0} \cap \mathcal{F}_n \mid x_n) \to 1$ a.s.$-P_0^\infty$ and $\Pi_{\Psi}(U_{\psi_0} \cap \mathcal{F}_n) \to \Pi_{\Psi}(U_{\psi_0})$ as $n \to \infty$. Hence, there exists $\psi \in U_{\psi_0} \cap \mathcal{F}_n$ such that $\pi_{\Psi}(\psi \mid x_n) > 1$ a.s.$-P_0^\infty$. Thus, the sieve LRS estimator $\hat{\psi}$ is in the ball $\{\psi : d_{\Psi}(\psi, \psi_0) < \epsilon\}$ for sufficiently large $n$. Therefore $\hat{\psi}$ is consistent. \qed

Asymptotic normality of the least relative surprise estimator for a infinite dimensional (or a functional) parameter hasn’t been studied yet. As a related topic, we studied some examples containing an increasing number of incidental parameters in the following section.
5.2 Models with Increasing Number of Incidental Parameters

Statistical models used in applications are getting bigger and more complicated as the data are becoming more complex. Usually, big models contain many parameters while only a couple of parameters are of interest. The parameters that are not of interest are called nuisance parameters. Nuisance parameters depending only on a study object are called incidental parameters.

For a better understanding, consider an imaginary clinical trial on a medicine. Each experimental unit has a different propensity to absorb the medicine into one’s body. So, we must take into account each individual’s potential. Roughly speaking, the observed effect can be expressed as the sum of true effect, personal potential, and error. The value depending only on each individual is called an incidental parameter. If we ignore such incidental parameters, then the estimated true effect could be biased.

In Bayesian inference, such nuisance parameters are integrated out. Usually, the number of incidental parameters increases as the size of the data grows. So, if an informative prior is placed on incidental parameters, then Bayesian inferences might be misleading in the sense that Bayes rules are not consistent, see Example 5.1. In this section, we suggest a method to construct a less informative prior.

Let $\theta$ be the parameter for a model. Assume there is a reparametrization $\theta \mapsto (\psi, \lambda) = (\Psi(\theta), \Lambda(\theta))$ where $\psi$ is the parameters of interest and $\lambda$ is a nuisance parameter. The dimension of $\lambda$ might increase as the data size grows.

Evans and Moshonov (2006) consider a decomposition of the joint model $P_\theta \times \Pi$ and the usage of each component. When we have a minimal sufficient statistic $T$, we can write

$$P_\theta \times \Pi = P(\cdot \mid T) \times M_T \times \Pi(\cdot \mid T).$$

The first component $P(\cdot \mid T)$ is independent of $\theta$ and the prior $\Pi$. So it can be used for
model checking. The last component $\Pi(\cdot | T)$ is available for inference about $\theta$. $M_T$ can be used for assessing whether or not there is a conflict between the prior and the data. Also, when there is a maximal ancillary $U$, that is a function of the minimal sufficient statistic $T$, we have a further decomposition

$$P_\theta \times \Pi = P(\cdot | T) \times P_U \times M_T(\cdot | U) \times \Pi(\cdot | T).$$

Then, $P_U$ can be also used for assessing whether or not the sampling model is reasonable.

We may extend this data decomposition for the model having nuisance parameters, that is, we may consider a data decomposition $x_n \mapsto (W, U, V)$ such that $(U, V)$ is minimal sufficient for $(\psi, \lambda)$, and $U$ is maximal ancillary for $\lambda$ whenever $\psi$ is given. Such a data decomposition may not exist in general. However, we suppose that this decomposition is available throughout this section. Then, we obtain a density decomposition

$$f(w, u, v | \psi, \lambda) = f(w) f(u | \psi) f(v | u, \psi, \lambda).$$

Hence, we get the prior predictive distribution

$$m(w, u, v) = \int \int f(w) f(u | \psi) f(v | u, \psi, \lambda) \pi(\lambda | \psi) \mathcal{L}(d\lambda) \pi_\psi(\psi) \mathcal{L}(d\psi)$$

$$= f(w) \int f(u | \psi) m(v | u, \psi) \pi_\psi(\psi) \mathcal{L}(d\psi)$$

where $m(v | u, \psi) = \int f(v | u, \psi, \lambda) \pi(\lambda | \psi) \mathcal{L}(d\lambda)$. The relative belief ratio at $\psi$ is

$$RB(\psi) = \frac{\pi_\psi(\psi | w, u, v)}{\pi_\psi(\psi)} = \int \frac{f(w) f(u | \psi) f(v | u, \psi, \lambda)}{m(w, u, v)} \pi_\psi(\psi) \pi(\lambda | \psi) \mathcal{L}(d\lambda) \mathcal{L}(d\psi) / \pi_\psi(\psi)$$

(5.5)

where $m(u, v) = \int f(u | \psi) m(v | u, \psi) \pi_\psi(\psi) \mathcal{L}(d\psi)$.

If the conditional prior $\pi(\lambda | \psi)$ is strongly informative, then this conditional prior affects on relative surprise inference through $m(v | u, \psi)$ among the terms in (5.5). If $m(v | u, \psi)$ is constant when $\psi$ varies, then we get no information concerning $\psi$ from $m(v | u, \psi)$. So we call a conditional prior $\Pi_{\Lambda | \psi}$ having density $\pi(\lambda | \psi)$ noninformative if
\( m(v \mid u, \psi) \) is constant as a function of \( \psi \). Such a noninformative conditional prior may not exist. However, we get non-subjective or less subjective results by using noninformative conditional prior when it exists.

We consider Examples 5.1 and 5.2 where this criterion is applied.

**Example 5.1 (Neyman-Scott Problem).** Neyman and Scott (1948) considered a model given by \( X_{ij} \sim N(\mu_i, \sigma^2) \) where \( j = 1, \ldots, n_i \) and \( i = 1, \ldots, n \). In the paper \( n_i = 2 \) is also assumed. In this example we will consider arbitrary \( n_i \)'s, then restrict \( n_i = m \) for all \( i \). So, the Neyman-Scott problem will be a special case. The parameter of interest in this problem is the common variance \( \sigma^2 \). The \( \mu_i \)'s are nuisance parameters.

A conjugate prior is placed on the whole parameter space, i.e., \( \mu_i \mid \sigma^2, x_n \sim N\left(\frac{n_i \bar{x}_i + \mu_0 / \tau_0^2}{n_i + 1 / \tau_0^2}, \sigma^2 / (n_i + 1 / \tau_0^2)\right) \),

\[ \frac{1}{\sigma^2} \mid x_n \sim \text{Gamma}_{\text{rate}}(\alpha + N/2, \beta + \frac{1}{2} \sum_{i,j} (x_{ij} - \bar{x}_i)^2 + \frac{1}{2} \sum_i \frac{n_i (\bar{x}_i - \mu_0)^2}{1 + n_i \tau_0^2}) \]

where \( N = n_1 + \cdots + n_n \). Hence, the marginal posterior is the gamma distribution described above. The ratio of the posterior density to the prior density of \( 1/\sigma^2 \) is proportional to

\[ \left( \frac{1}{\sigma^2} \right)^{N/2} \exp \left( -\frac{1}{2\sigma^2} \left[ \sum_{i,j} (x_{ij} - \bar{x}_i)^2 + \sum_i \frac{n_i (\bar{x}_i - \mu_0)^2}{1 + n_i \tau_0^2} \right] \right) \]

Therefore the LRSE is given by

\[ \hat{\sigma}^2_{LRSE} = \frac{1}{N} \left[ \sum_{i,j} (x_{ij} - \bar{x}_i)^2 + \sum_i \frac{n_i (\bar{x}_i - \mu_0)^2}{1 + n_i \tau_0^2} \right] \]

Under the model assumption, the mean and variance of the LRSE are

\[ \mathbf{E}(\hat{\sigma}^2_{LRSE}) = \frac{1}{N} \left[ (N - n)\sigma^2 + \sum_i \frac{(\sigma^2 + n_i (\mu_i - \mu_0)^2)}{1 + n_i \tau_0^2} \right], \quad (5.6) \]

\[ \text{var}(\hat{\sigma}^2_{LRSE}) = \frac{2}{N^2} \left[ (N - n)\sigma^4 + \sum_i \frac{\sigma^4 + 2\sigma^2 n_i (\mu_i - \mu_0)^2}{(1 + n_i \tau_0^2)^2} \right]. \quad (5.7) \]
Suppose that the mean given in (5.6) converges as \( n \to \infty \). From the fact \((1+n_i \tau_0^2)^{-1} \leq 1\), we have

\[
(5.7) \leq \frac{2}{N^2} [(N - n) \sigma^4 + \sum_i \frac{\sigma^4 + 2\sigma^2 n_i (\mu_i - \mu_0)^2}{1 + n_i \tau_0^2}] \\
\leq \frac{4\sigma^2}{N} \sum_i \frac{\sigma^4 + n_i (\mu_i - \mu_0)^2}{(1 + n_i \tau_0^2)^2} = \frac{4\sigma^2}{N} E(\hat{\sigma}^2_{LRSE}) \\
\to 0 \quad \text{as} \quad n \to \infty.
\]

In the last convergence, \( N \to \infty \) as \( n \to \infty \) is used. That means, the LRS estimator is consistent by Chebyshev’s inequality if its expectation converges to the true parameter. Set \( n_i = m \) in order to simplify the problem, then, the expectation (5.6) becomes

\[
\frac{1}{m} [(m - 1 + \frac{1}{1 + m \tau_0^2}) \sigma^2 + \frac{m}{1 + m \tau_0^2} \frac{1}{n} \sum_i (\mu_i - \mu_0)^2].
\]

Hence it converges to \( \sigma^2 \) if and only if

\[
\frac{1}{n \tau_0^2} \sum_i (\mu_i - \mu_0)^2 \to \sigma^2 \quad \text{as} \quad n \to \infty. \tag{5.8}
\]

If the data is generated from a sampling model that comes from the prior distribution, then the condition (5.8) is automatically satisfied. So the LRSE is consistent. However, the sampling distribution does not come from the prior distribution in general. That means, the LRSE is inconsistent whenever we think there is some fixed sequence \((\mu_1, \mu_2, \ldots)\) such that (5.8) does not hold.

The inconsistency of the LRSE is caused by the information of the conditional prior. Note that a full likelihood is given by

\[
\prod_{i=1}^n \prod_{j=1}^{n_i} [2\pi \sigma^2]^{-1/2} \exp\left(-\frac{(x_{ij} - \mu_i)^2}{2\sigma^2}\right)
\]

\[
= \frac{[2\pi \sigma^2]^{-(N-n)/2}}{\prod_{i=1}^n n_i^{1/2}} \cdot \prod_{i=1}^n \frac{n_i^{1/2}}{[2\pi \sigma^2]^{1/2}} \exp\left(-\frac{n_i(\bar{x}_i - \mu_i)^2}{2\sigma^2}\right).
\]

If a conditional prior \( \mu_i \sim N(\mu_0, \sigma^2_0^2) \) is used, the conditional prior predictive distribution of \( \bar{x}_i \) is \( \bar{x}_i | \sigma^2 \sim N(\mu_0, \sigma^2(1/n_i + \tau_0^2)) \). It surely conveys information about \( \sigma^2 \).
which comes from the conditional prior and the conditional model. Intuitively we assume that the information about \( \sigma^2 \) in the conditional prior predictive distribution of \( \bar{x}_i \mid \sigma^2 \) disappears. However, for given \( \sigma^2 \),

\[
\prod_{i=1}^{n} \left( 2\pi \sigma^2(1/n + \tau_0^2) \right)^{-1/2} \exp(-\frac{(\bar{x}_i - \mu_0)^2}{2\sigma^2(1/n + \tau_0^2)}) \propto \sigma^{-n} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\bar{x}_i - \mu_0)^2) \\
\rightarrow \sigma^{-n} \quad \text{as} \quad \tau_0^2 \rightarrow \infty.
\]

Hence, the information about \( \sigma^2 \) in the conditional prior predictive distribution is not ignorable. And this information affects any inferences about \( \sigma^2 \).

The inconsistency may arise from prior-data conflict. The prior predictive density of a minimal sufficient statistic \( T = (\bar{x}_1, \ldots, \bar{x}_n, s^2) \) where \( s^2 = (N-n)^{-1} \sum_{i,j}(x_{ij} - \bar{x}_i)^2 \) is

\[
m_T(t) = c(s^2)^{(N-n)/2-1} \left( \beta + \frac{(N-n)s^2}{2} + \frac{m}{2(1+m\tau_0^2)} \sum_{i=1}^{n} (\bar{x}_i - \mu_0)^2 \right)^{-(\alpha+N/2)}
\]

where \( c = \beta^\alpha \Gamma(\alpha + N/2)/[(2\pi)^{N/2}(1 + m\tau_0^2)^{n/2}\Gamma(\alpha)\Gamma((N-n)/2)] \). Then, marginally, 2\( \alpha s^2/\beta \sim F(N-n, 2\alpha) \) and, for given \( s^2, v = (\bar{x}_1, \ldots, \bar{x}_n) \) has a multivariate t distribution with degree of freedom \( N-n+2\alpha \), location \( \mu_0 1_n \) and scale matrix \( \Sigma = \{2(1+m\tau_0^2)(2\beta + (N-n)s^2)/(m(N-n+2\alpha))\}I_n \). Since \( x_n \rightarrow v \) has constant volume distortion, the prior-data conflict \( P \)-value with respect to \( M_{V \mid s^2} \) is

\[
M_{V \mid s^2}(\{v = (x_1, \ldots, x_n) : \sum_{i=1}^{n} (x_i - \mu_0)^2 \geq \sum_{i=1}^{n} (x_{i0} - \mu_0)^2\}).
\]

For computational simplicity, we use

\[
\frac{m}{2(1+m\tau_0^2)} \frac{N-n+2\alpha}{2\beta + (N-n)s^2} \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_i - \mu_0)^2 \sim F(n, N-n+2\alpha). \quad (5.9)
\]

Also assume \( ((\mu_1 - \mu_0)^2 + \cdots + (\mu_n - \mu_0)^2)/n) \rightarrow \sigma_{\mu^*}^2 \). Let \( \xi_i = \bar{x}_i - \mu_i \). Then, \( \xi_i \sim i.i.d.N(0, \sigma^2/m) \) and

\[
\frac{1}{n} \sum_{i=1}^{n} (\bar{x}_i - \mu_0)^2 = \frac{1}{n} \sum_{i=1}^{n} (\xi_i + \mu_i - \mu_0)^2 = \frac{1}{n} \sum_{i=1}^{n} \{\xi_i^2 + 2\xi_i(\mu_i - \mu_0) + (\mu_i - \mu_0)^2\}.
\]

Since \( E(\xi_i^2) = \sigma^2/m, \) \( \sum_{i=1}^{n} 2\xi_i(\mu_i - \mu_0)/n \sim N(0, 4\sum_{i=1}^{n}(\mu_i - \mu_0)^2/n^2) \) and \( \sum_{i=1}^{n}(\mu_i - \mu_0)^2/n \rightarrow \sigma_{\mu^*}^2 \), we have \( n^{-1} \sum_{i=1}^{n} (\bar{x}_i - \mu_0)^2 \xrightarrow{a.s.} \sigma^2/m + \sigma_{\mu^*}^2 \). It is easy to check that
Chapter 5. Inferences for Infinite Dimensional Models

$F(n, N - n + 2\alpha) \overset{d}{\to} 1$ as $n \to \infty$. Then, under the sampling model assumption, the left hand side of (5.9) converges to

$$\frac{m}{2(1 + m\tau_0^2)} \frac{1}{s^2} \left( \frac{\sigma^2}{m} + \sigma_{\mu^*}^2 \right)$$  \hspace{1cm} (5.10)

for fixed $s^2$. Hence, there is prior-data conflict asymptotically unless (5.10) converges to 1. Thus, there is prior-data conflict asymptotically if $|s^2 - (\sigma^2 + m\sigma_{\mu^*}^2)/(2(1 + m\tau_0^2))| > \epsilon$ for any fixed $\epsilon > 0$. By taking $\epsilon$ very small, there is strong asymptotic prior-data conflict. Hence, the inconsistency of the LRSE may be caused by this asymptotic prior-data conflict.

To get rid of this information we place $\mu_i \sim \text{Uniform}(-B, B)$. Then, the conditional prior predictive distribution becomes $m_{\bar{X}_i|\sigma^2}(\bar{x}_i|\sigma^2) = (\Psi((B - \bar{x}_i)/\sqrt{\sigma^2/n_i}) - \Psi((-B - \bar{x}_i)/\sqrt{\sigma^2/n_i}))/B$. Hence, the information of $\sigma^2$ disappears as $B$ increases. Then, as $B \to \infty$, we have $B \cdot m_{\bar{X}_i|\sigma^2}(\bar{x}_i|\sigma^2) \to 1$ and the relative belief ratio converges to

$$\frac{|2\pi\sigma^2|^{-(N-n)/2}}{\prod_{i=1}^{n} n_i^{1/2}} \exp\left(-\frac{\sum_{i,j} (x_{ij} - \bar{x}_i)^2}{2\sigma^2}\right).$$

So the LRS estimator $\hat{\sigma}_{LRSE,B}^2$, which maximizes $\sigma^{-(N-n)} \exp(-(2\sigma^2)^{-1} \sum_{i,j} (x_{ij} - \bar{x}_i)^2) \times \prod_{i=1}^{n} m_{\bar{X}_i}(\bar{x}_i)$, converges to

$$\hat{\sigma}_{LRSE,\infty}^2 = \frac{1}{N-n} \sum_{i,j} (x_{ij} - \bar{x}_i)^2$$

which is an unbiased consistent estimator of $\sigma^2$. Considering $(N-n)\hat{\sigma}_{LRSE,\infty}^2/\sigma^2 \sim \chi^2(N-n)$, we have that an asymptotic normality $\sqrt{N-n}(\hat{\sigma}_{LRSE,B}^2 - \sigma^2) \overset{d}{\to} N(0, 2\sigma^4)$. Also the Bernstein-von Mises theorem holds when a prior having a continuously differentiable density is used.

As we can see in Example 5.1, the choice of the conditional prior is very important when the number of nuisance parameters increases. If the conditional distribution $f(v | u, \psi, \lambda)$ is also a density in $\lambda$, then a uniform prior on a huge compact set works fine.
Example 5.2. Evans and Shakhatreh (2007) presents an interesting example. Namely, the least relative surprise estimator is consistent while many other Bayesian estimators aren’t. The sampling model is given by

\[ X_i | \theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1) \]

and the parameter of interest is \( \psi = (\theta_1^2 + \cdots + \theta_n^2)/n \). The nuisance parameters are \( \eta = (\theta_1, \ldots, \theta_n)/\sqrt{n\psi} \). The density of the sampling model is

\[
\prod_{i=1}^{n} |2\pi|^{-1/2} \exp\left(-\frac{(x_i - \theta_i)^2}{2}\right) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2} ||x_n||^2 + n\psi + \sqrt{n\psi}\eta'x_n\right). \tag{5.11}
\]

In this formula, we need to separate terms into two groups according to whether or not it is related to the nuisance parameter. In (5.11), only \( \sqrt{n\psi}\eta'x_n \) is related to the nuisance parameter \( \eta \). Hence, it is not contained in the marginal likelihood function. The other term \( ||x_n||^2 \) might contain information about \( \psi \). From this, we make a data transformation \( x_n \mapsto (||x_n||^2, y_n) \) where \( y_n = x_n/||x_n|| \). Then, we have the density decomposition

\[
(||x_n||^2)^{n/2-1} \exp\left(-\frac{1}{2} (||x_n||^2 + n\psi)\right) \sum_{i=0}^{\infty} \frac{(n\psi||x_n||^2/4)^i}{i!\Gamma(n/2+i)} (d||x_n||^2) \times \left( \sum_{i=0}^{\infty} \frac{(n\psi||x_n||^2/4)^i}{i!\Gamma(n/2+i)} \right)^{-1} \exp(||x_n||\sqrt{n\psi}\eta'y_n)(dy_n).
\]

where the range of \( y_n \) is the surface of the \((n-1)\)-dimensional sphere \( S^{n-1} \). So the first line can be used for inferences on \( \psi \) and the second line must be integrated out using a conditional prior. Since the second line is also a density of \( \eta \), we place a uniform prior on the surface of \( S^{n-1} \). Then, the least relative surprise estimator is the maximizer of

\[
\exp(-n\psi/2) \sum_{i=0}^{\infty} \frac{(n\psi||x_n||^2/4)^i}{i!\Gamma(n/2+i)}
\]

as presented in Evans and Shakhatreh (2007). Since the LRSE is \((||x_n||^2/n - 1)_+ = \max(0,||x_n||^2/n - 1)\) except an absolute error less than \( 2/n \), the LRSE is asymptotically normal. If a prior having positive and continuous density is placed on \( \psi \), then the
Bernstein-von Mises theorem holds when the posterior is centered at the LRSE and scaled by $\sqrt{n}$.

The prior considered in Evans and Shakhatreh (2007) is $\psi \sim \text{Gamma}_{\text{scale}}(n/2, 2\sigma^2/n)$. Hence, this sequence of priors converges to $\sigma^2$ as $n \to \infty$. This explains why they get that the posterior is consistent if and only if $\psi = \sigma^2$.

The reference prior for full parameter is an improper uniform prior on each $\theta_i$. Then, the conditional prior $\pi(\lambda | \psi)$ is uniform on the surface of $S^{n-1}$. Thus, the LRSE is consistent, however, the prior density on $\psi$ is proportional to $\psi^{n/2-1}$, so the corresponding posterior is $n\psi | x_n \sim \chi^2(n, ||x_n||^2)$. The posterior distribution converges to $2 + \psi_0$ as $n \to \infty$ where $\psi_0$ is the true parameter. This shows that the reference prior may not satisfy posterior consistency.

5.3 Future Work

Infinite dimensional models are becoming more and more commonly used. In this chapter, we extended relative surprise inferences from models with finite dimensional parameter space to infinite dimensional models. For example, consistency can be obtained by restricting the model to a sequence of sieve models. And we showed the similarity of posterior consistency and the consistency of the least relative surprise estimator. Further we discussed a noninformative conditional prior for infinitely many incidental parameters. Even though noninformative conditional priors may not exist in general, it provides a conservative inference when it exists.

There are still many characteristics to be studied. For example, the Bernstein-von Mises theorem for infinite dimensional models and characteristics of relative surprise inferences for semiparametric models, are two problems that are currently being considered.

In Theorem 5.1, the Kullback-Leibler property is used in the consistency of the LRSE. The possibility of weakening the Kullback-Leibler property will be studied. Also asym-
totic properties for marginal parameters are less known and must be studied further. Similar asymptotic properties for semiparametric models introduced in Section 5.2 are also of interest. Finally, applications are being considered for the models containing latent variables, e.g., structural equation models and measurement error models.
Appendix A

Some useful results, used in the thesis, are summarized in this appendix. The author is sure that most results in this chapter are not new findings.

**Lemma A.1.** The inequality \((1 + a)^h \geq 1 + ha\) for positive integer \(h\) and \(a \geq -1\).

**Proof Lemma A.1.** The lemma can be proven using mathematical induction. It is true when \(h = 1\). Suppose it is true for \(h = k\). Then,

\[(1 + a)^{k+1} = (1 + a)^k(1 + a) \geq (1 + ka)(1 + a) = 1 + (k + 1)a + ka^2 \geq 1 + (k + 1)a.\]

Hence it is true for all positive integer \(h\). □

**Lemma A.2.** If positive numbers \(a, b, c\) and \(d > 0\) satisfy \(a/b > c/d\), then

\[
\frac{a}{b} > \frac{a + c}{b + d} > \frac{c}{d}.
\]

**Proof of Lemma A.2.** From \(a/b > c/d\), we have \(ad - bc > 0\). The inequalities are obtained from \(a(b + d) - b(a + c) = ad - bc > 0\) and \((a + c)d - (b + d)c = ad - bc > 0\). □

**Lemma A.3.** A sequence \(x_n\) converges to \(x\) if and only if, for any subsequence \(x_{n_k}\), there exists a further subsequence \(x_{n_{k_l}}\) converging to \(x\).

**Proof of Lemma A.3.** It is a special case of Theorem 20.5 of Billingsley (1995). □

**Lemma A.4.** For \(|z| < 1\) and positive integer \(k\), we have \(|\log(1 - z) + (z + z^2/2 + \cdots + z^k/k)| \leq |z|^{k+1}/[(k + 1)(1 - |z|)]\). If \(|z| \leq 1/2\), then \(|\log(1 - z) + (z + z^2/2 + \cdots + z^k/k)| < (2/k)|z|^{k+1} \).
Proof of Lemma A.4. The Taylor expansion of \( \log(1-z) \) is given by \( \log(1-z) = -\sum_{i=1}^\infty \frac{z^i}{i} \).

Hence we have
\[
\left| \log(1-z) + \sum_{i=1}^k \frac{z^i}{i} \right| \leq \frac{1}{k+1} \sum_{i=k+1}^\infty |z|^i = \frac{1}{k+1} \frac{|z|^{k+1}}{1-|z|}.
\]

If \( |z| \leq 1/2 \), then \( 1/(1-|z|) \leq 2 \). Thus we have the lemma. \( \square \)

Lemma A.5 (Stirling’s Formula). \( |\log \Gamma(z) - \frac{1}{2} \log(2\pi) - (z - \frac{1}{2}) \log z + z - (12z)^{-1}| < |z|^3/360. \)

Proof of Lemma A.5. See Formula 6.1.42 of Abramowitz and Stegun (1972). \( \square \)

Lemma A.6. (i) \( |\sin x| \leq |x| \) for all \( x \in \mathbb{R} \),

(ii) \( \cos x \leq 1 - \frac{x^2}{2} + \frac{x^4}{24} = 1 - (\frac{x^2}{2})(1 - \frac{x^2}{12}) \) for all \( |x| \leq \sqrt{56} \).

Proof of Lemma A.6. (i) The result is well-known.

(ii) The trigonometric function expansion of \( \cos x \) is given by
\[
\cos x = \sum_{i=0}^\infty (-1)^i \frac{x^{2i}}{(2i)!} = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \sum_{j=2}^\infty \frac{x^{4j-2}}{(4j-2)!} \left( 1 - \frac{x^2}{(4j-1)(4j)} \right).
\]

For \( |x| \leq \sqrt{56} \) and \( j \geq 2 \), we have \( 1 - x^2/[(4j-1)4j] \geq 0 \) as well as
\[
\cos x \leq \sum_{i=0}^\infty (-1)^i \frac{x^{2i}}{(2i)!} = 1 - \frac{x^2}{2} + \frac{x^4}{24}.
\]

Hence, the lemma follows. \( \square \)

Lemma A.7 (Arithmetic mean and Geometric mean inequality). For real numbers \( x_1, \ldots, x_n \),
\[
|x_1 \cdots x_n|^{1/n} \leq (|x_1| + \cdots + |x_n|)/n
\]
where the equality holds when \( |x_1| = \cdots = |x_n| \). For \( n = 2 \), it becomes \( ab \leq (a^2 + b^2)/2 \).

Proof of Lemma A.7. This result is well-known. \( \square \)

Lemma A.8. Suppose \( X_i \)'s are an i.i.d. sample and \( \mathbf{E}(|X_i|) < \infty \). Then, for \( S_n = X_1 + \cdots + X_n \), \( S_n/n \xrightarrow{a.s.} \mathbf{E}(X_1) \) and \( \mathbf{E}(|S_n/n - \mathbf{E}(X_1)|) \to 0 \) as \( n \to \infty \).
Proof of Lemma A.8. The almost sure convergence is the strong law of large numbers. For the $L_1$ convergence, we can assume $E(X_1) = 0$ without loss of generality. Then, $|S_n/n| \leq T_n = (|X_1| + \cdots + |X_n|)/n$, $E(T_n) = E(|X_1|)$ and $T_n \xrightarrow{a.s.} E(|X_1|)$. Hence, the generalized Lebesgue dominated convergence theorem implies $E(|S_n/n|) \to 0$. 

Lemma A.9. Suppose an exponential family having density $f_\eta(x) = \exp(\eta x - A(\eta))h(x)$. For an i.i.d sample $X_1, \ldots, X_n$ from $f_\eta$, define $T = (V(X_1) + \cdots + V(X_n))/n$. Then, $f_{\eta,T,n}(t) = |2\pi\tilde{A}(\eta)/n|^{-1/2}\exp(-\frac{t}{2}(\eta - \tilde{A}^{-1}(t))'\tilde{A}(\eta)(\eta - \tilde{A}^{-1}(t)) + o(n||\tilde{A}^{-1}(t) - \eta||^2))$.

Proof of Lemma A.9. Suppose that $||\tilde{A}^{-1}(t) - \eta|| < n^{-1}\log n$. Note that $E_\eta[T] = E_\eta(V(X_1)) = \tilde{A}(\eta)$. The characteristic function of $P_{\eta,T,n}$ is

$$
ch.f_{\eta,T,n}(w) = \int \exp(iw't) \exp(n\eta't - nA(\eta))h(t)\mathcal{L}(dt) = \exp(nA(\eta + iw/n) - nA(\eta)).
$$

By the density inversion formula, we have

$$
f_{\eta,T,n}(t) = (2\pi)^{-k}\int e^{-iw't} \exp(nA(\eta + iw/n) - nA(\eta))\mathcal{L}(dw)
$$

where $k = \dim(T)$. This integral is approximated using the saddlepoint approximation. By solving $\frac{\partial}{\partial w}(-iw't + nA(\eta + iw/n) - nA(\eta)) = -it + n\tilde{A}(\eta + iw/n)i/n = 0$, we have the solution $\tilde{w} = -i(n\tilde{A}^{-1}(t) - \eta)$. The second derivative is $-\tilde{A}(\eta + iw/n)/n$. Hence, we obtain $i\tilde{w}/n = \tilde{A}^{-1}(t) - \eta$ and

$$
\approx (2\pi)^{-k}|2\pi n\tilde{A}(\tilde{A}^{-1}(t))|^{-1/2}\exp(-i\tilde{w}'t + nA(\tilde{A}^{-1}(t)) - nA(\eta))
= |2\pi\tilde{A}(\tilde{A}^{-1}(t))/n|^{-1/2}\exp(-n(\tilde{A}^{-1}(t) - \eta)'t + nA(\tilde{A}^{-1}(t)) - nA(\eta))
$$

The Taylor expansion of $A(\eta)$ around $\tilde{A}^{-1}(t)$ gives us

$$
= |2\pi\tilde{A}(\tilde{A}^{-1}(t))/n|^{-1/2}\exp(-n(\tilde{A}^{-1}(t) - \eta)'t - n(\eta - \tilde{A}^{-1}(t))\tilde{A}(\tilde{A}^{-1}(t)))
\times \exp(-\frac{n}{2}(\eta - \tilde{A}^{-1}(t))'\tilde{A}(\tilde{A}^{-1}(t))(\eta - \tilde{A}^{-1}(t)) + O(||\tilde{A}^{-1}(t) - \eta||^3))
= |2\pi\tilde{A}(\eta)/n|^{-1/2}\exp(-\frac{n}{2}(\eta - \tilde{A}^{-1}(t))'\tilde{A}(\eta)(\eta - \tilde{A}^{-1}(t)) + o(n||\tilde{A}^{-1}(t) - \eta||^2))
$$
In the last equality, \( \log |\tilde{A}(\eta)^{-1} \tilde{A}(\hat{A}^{-1}(t))| = O(||\eta - \hat{A}^{-1}(t)||) \) and \( ||\tilde{A}(\hat{A}^{-1}(t)) - \tilde{A}(\eta)|| = O(||\hat{A}^{-1}(t) - \eta||) \) are used. \( \square \)
Bibliography


Fraser, D., Reid, N., Marras, E., , and Yi, G. (2009). Default priors for Bayesian and frequentist inference. *working paper*.


Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine, Series 5*, 50(302):157–176.


