Non-Bayesian Inference and Prediction

Di Xiao

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Non-Bayesian Inference and Prediction

Di Xiao

In this thesis, we first propose a coherent inference model that is obtained by distorting the prior density in Bayes' rule and replacing the likelihood with a so-called pseudo-likelihood. This model includes the existing non-Bayesian inference models as special cases and implies new models of base-rate neglect and conservatism. We prove a sufficient and necessary condition under which the coherent inference model is processing consistent, i.e., implies the same posterior density however the samples are grouped and processed retrospectively. We show that processing consistency does not imply Bayes' rule by proving a sufficient and necessary condition under which the coherent inference model can be obtained by applying Bayes' rule to a false stochastic model. We then propose a prediction model that combines a stochastic model with certain parameters and a processing-consistent, coherent inference model. We show that this prediction model is processing consistent, which states that the prediction of samples does not depend on how they are grouped and processed prospectively, if and only if this model is Bayesian. Finally, we apply the new model of conservatism to a car selection problem, a consumption-based asset pricing model, and a regime-switching asset pricing model.

Table of Contents

List of Figures

List of Tables

Key words: Non-Bayesian inference and prediction, processing consistency, distortion, pseudo-likelihood, false-Bayesian models, conservatism, base-rate neglect, consumer choice, asset pricing

JEL Codes: D03, D83, G02

Acknowledgments

Seven years ago, when I first came to Columbia with an undergraduate background in finance and economics, it is impossible for me to imagine composing a purely quantitative thesis on Behavioral Modeling. After 7 years of training in IEOR, I have realized my initial goal of envisioning the finance world in a purely mathematical approach and contributing to the research world my original ideas. This thesis would be impossible without the support and help of many people from IEOR department, my friends and my family.

I would like to first thank my supervisor Dr. Xuedong He.

In addition, I would like to thank my committee members: Dr. Karl Sigman, Dr. Xunyu Zhou, Dr. Agostino Capponi, Dr. Olympia Hadjiliadis.

Thanks Miguel Garrido Garcia, Zhipeng Liu, Xiao Xu, Lin Chen, Jing Guo and Xingye Wu.

Thanks my family.

To My Family

Introduction

In various contexts, individuals need to infer unknown parameters or unknown probabilities of certain events from historical data. A rational model for such inferences is Bayes' rule, which states that the posterior probability density of an unknown parameter is proportional to the likelihood of the parameter given the historical data multiplied the prior probability density of the parameter. However, abundant experimental evidence reveals that individuals often violate or ignore Bayes' rule when attempting to reach an inference. Examples of non-Bayesian behavior include, but are not limited to, representativeness [Tversky and Kahneman, 1974], conservatism [Edwards, 1968], and base-rate neglect [Bar-Hillel, 1980]. Such violation and ignorance are systematic and exhibit certain patterns, making it possible to model them.

Several non-Bayesian inference models have been proposed to describe particular non-Bayesian behavior. For example, Rabin [2002] proposes a model to describe the law of small numbers. Rabin and Vayanos [2010] propose a model for the gambler's and hot-hand fallacies. Recently, Benjamin *et al.* [2016] propose a model of non-belief in the law of large numbers (NBLLN). However, all these models are ad hoc in the sense that they are built for describing particular non-Bayesian behaviors and have some limitations. Indeed, the models proposed by Rabin [2002] and Benjamin *et al.* [2016] are applied only to binary samples. The random samples in the model of Rabin and Vayanos [2010] are time series with random noises following the normal

distribution.

The first goal of this thesis is to propose a general non-Bayesian inference model, which we refer to as the coherent inference model. This model is constructed by applying distortion to the prior density in Bayes' rule and replacing the likelihood with a pseudo-likelihood. Moreover, this model is dynamic: part of a sample sequence can be processed first to obtain an updated density of the parameter, and this updated density serves the prior density when processing the next piece of the sample sequence. The coherent inference model is general in three respects: First, it includes, as special cases, the aforementioned non-Bayesian inference models from the literature. Second, it allows any types of samples, including non-i.i.d. samples. Third, it implies new models of conservatism and base-rate neglect. These new models are tractable and thus widely applicable to many economic and financial problems.

The second goal of this thesis is to understand whether the means of individuals' processing data affects the inference result. In Bayes' rule, after the sample points are observed, whether they are processed retrospectively as a group or one by one does not affect the posterior density of the parameters, a property referred to as *processing consistency*. Benjamin *et al.* [2016], however, find that the model of NBLLN is processing inconsistent. Moreover, some experimental studies have also revealed processing-inconsistent behavior [Shu and Wu, 2003; Kraemer and Weber, 2004]. Two theoretical questions then arise: First, can we find an easy condition to check whether a non-Bayesian inference model is processing consistent? Second, does processing consistency imply Bayes' rule? In this thesis, we attempt to answer these questions. More precisely, we provide a sufficient and necessary condition under which the coherent inference model is processing consistent.

The third goal of this thesis is to study whether processing consistency in inference is equivalent to the use of Bayes' rule. Note that Bayes' rule can be applied to a false

stochastic model, leading to a false-Bayesian model. We provide a sufficient and necessary condition under which the coherent inference model is essentially a false-Bayesian model. Moreover, we provide examples that are processing consistent but not false Bayesian. These examples show clearly that processing consistency does not imply Bayes' rule.

In addition to processing sample points retrospectively after they are observed, in many situations individuals also need to process sample points prospectively before they are observed so as to predict them. Thus, processing inconsistency can arise not only in retrospective data processing (leading to inference) but also in prospective data processing (leading to prediction). The fourth goal of this thesis is to study when individuals are processing consistent in prediction. More precisely, we consider an individual who needs to predict incoming sample points based on historical samples. The individual has a stochastic model for the dynamics of the sample points with an unknown parameter and also has an inference model for the parameter. The individual then predicts incoming sample points by combining his inference model and prediction model with known parameter, leading to a general prediction model. We prove that this prediction model is processing consistent if and only if it uses the Bayes' updating rule.

Finally, we apply non-Bayesian inference models to various economic problems. We first apply the model of conservatism to a consumer choice problem in which an agent needs to decide whether to purchase consumer reports that signal the quality of cars in two brands and then decide which car to purchase. We find that when the agent becomes more conservative, he tends to under-infer more the quality of the cars from the purchased reports and thus are less willing to pay of the reports. We then apply the model of conservatism to asset pricing in a standard consumption-based asset pricing framework. We find that when the representative agent becomes more

conservative, the risk-free return and the price-dividend ratio become less sensitive to the number of good signals in the historical dividend data. Finally, we combine the model of conservatism with a regime-switching asset pricing model in Barberis *et al.* [1998] and find that the more conservative the agent is, the less profound the effect of short-term momentum and long-term reversal.

The remainder of the thesis is organized as follows. In Chapter 1, we achieve the first three goals of the thesis. In Chapter 2, we achieve the fourth goal and study the car selection problem. In Chapter 3, we consider the two asset pricing problems. All proofs are placed in Appendix A. Appendices B and C provide additional results regarding to the coherent inference model and processing consistency.

Chapter 1

Non-Bayesian Inference Model

1.1 Introduction

Bayes' rule is regarded as a rational model for statistical inference regarding unknown parameters of a stochastic model. In this rule, the posterior density of an unknown parameter is proportional to its likelihood given the observed sample multiplied by the prior density of the parameter. However, abundant experimental evidence reveals that individuals often violate or ignore Bayes' rule when attempting to reach an inference. Examples of non-Bayesian behavior include, but are not limited to, representativeness [Tversky and Kahneman, 1974], conservatism [Edwards, 1968], and base-rate neglect [Bar-Hillel, 1980].

Several non-Bayesian inference models have been proposed to describe particular non-Bayesian behavior. For example, Rabin [2002] proposes a model to describe the law of small numbers. Rabin and Vayanos [2010] propose a model for the gambler's and hot-hand fallacies. Recently, Benjamin *et al.* [2016] propose a model of non-belief in the law of large numbers (NBLLN). Note that the models proposed by Rabin [2002] and Benjamin *et al.* [2016] are applied only to binary samples. The random samples

in the model of Rabin and Vayanos [2010] are time series with random noises following the normal distribution.

Inference, as understood here, refers to the process of inferring unknown parameters of a stochastic model after observing sample points. In Bayes' rule, whether the sample points are processed retrospectively as a group or one by one does not affect the posterior density of the parameters, a property referred to as *processing consistency*. Benjamin *et al.* [2016], however, find that the model of NBLLN is processing inconsistent. Moreover, some experimental studies have also revealed processing-inconsistent behavior [Shu and Wu, 2003; Kraemer and Weber, 2004]. Two theoretical questions then arise: First, can we find an easy condition to check whether a non-Bayesian inference model is processing consistent? Second, does processing consistency imply Bayes' rule? The present paper attempts to answer these questions.

First, we propose a general non-Bayesian inference model, which we refer to as the coherent inference model. This model is constructed by applying distortion to the prior density in Bayes' rule and replacing the likelihood with a pseudo-likelihood. Moreover, this model is dynamic: part of a sample sequence can be processed first to obtain an updated density of the parameter, and this updated density serves the prior density when processing the next piece of the sample sequence. The coherent inference model is general in three respects: First, it includes, as special cases, the aforementioned non-Bayesian inference models from the literature. Second, it allows any types of samples, including non-i.i.d. samples. Third, it implies new models of conservatism and base-rate neglect.

We then provide a sufficient and necessary condition under which the coherent inference model is processing consistent. Literally, this condition states that (i) one cannot distort prior densities that are obtained after processing part of a sample sequence and (ii) the information contained in a sample sequence, which is measured by

the log pseudo-likelihood ratio, is additive when the sequence is divided into multiple components that can be processed sequentially. This sufficient and necessary condition is easy to verify and thus helps us to check whether a non-Bayesian inference model is processing consistent. Moreover, this condition highlights two causes of processing inconsistency: First, individuals indirectly distort the information contained in part of a sample sequence through distorting the density that is obtained after processing this part and used to process subsequent parts. Second, individuals measure the information contained in a sample sequence when it is processed as a whole to be different from the aggregate information contained in multiple components of the sample sequence that are processed sequentially.

Using the sufficient and necessary condition, we find that the models proposed by Rabin [2002] and Rabin and Vayanos [2010] are processing consistent. Indeed, these two models are obtained by applying Bayes' rule to particular false underlying stochastic models. Such models are called *false-Bayesian models*, and they constitute special cases of the coherent inference model. Due to Bayes' rule, false-Bayesian models are processing consistent. On the other hand, we confirm the observation by Benjamin *et al.* [2016] that the model of NBLLN is processing inconsistent, and show that the inconsistency arises from the non-additivity of the sample information.

By applying power distortion to the prior density and retaining the likelihood, we obtain a model of base-rate neglect. By using a power transformation of the likelihood as the pseudo-likelihood and not distorting the prior density, we obtain a new model of conservatism. These two models are again special cases of the coherent inference model. Moreover, the model of base-rate neglect is processing inconsistent if the distortion is also applied to prior densities that are obtained after processing part of a sample sequence. Consequently, the inconsistency in this case arises from distorting sample information indirectly through prior densities. On the other hand,

the model of conservatism is processing consistent. Compared to other models that can describe conservatism, such as those proposed by Rabin [2002], Benjamin *et al.* [2016], and Epstein *et al.* [2010], this model, although it is incapable of generating both under- and over-inference simultaneously, has the advantage of being tractable and general enough to allow for all types of samples; see the detailed discussion in Section 1.3.3.4.

The aforementioned examples of the coherent inference model can be combined to generate new examples. For instance, if we combine the models of base-rate neglect and conservatism, we obtain a new inference model, which is the same one implied by the regression analysis performed by Grether [1980] in his experimental test of representativeness.

Finally, we study whether processing consistency implies Bayes' rule. Note that Bayes' rule can be applied to a false stochastic model, leading to a false-Bayesian model. We introduce the following notions: The coherent inference model is false Bayesian in the strong sense if there exists a false-Bayesian model such that these two models imply the same posterior density for any prior density and sample; it is false Bayesian in the weak sense if, for any prior density, there exist a false prior density and a false-Bayesian model such that these two models imply the same posterior density for any sample. We provide sufficient and necessary conditions under which the coherent inference model is false Bayesian both in the strong sense and in the weak sense. Moreover, we provide examples that are processing consistent but not false Bayesian in the weak sense, false Bayesian in the weak sense but not in the strong sense, and false Bayesian in the strong sense, respectively. These examples show clearly that processing consistency does not imply Bayes' rule.

To summarize, the contribution of our work is three-fold: First, we propose the coherent inference model, which is general enough to allow for arbitrary types of

samples, to cover the existing non-Bayesian inference models, and to imply new inference models. Second, we provide a sufficient and necessary condition under which the coherent inference model is processing consistent, and this condition helps us to understand the causes of processing inconsistency and to easily check whether an inference model is processing consistent. Third, by proving a sufficient and necessary condition under which the coherent inference model is false Bayesian, we show that processing consistency does not imply Bayes' rule.

The coherent inference model is descriptive rather than normative; i.e., it is built directly to describe individuals' behavior in inference rather than obtained from a set of normative axioms on individuals' preferences. Descriptive models have been used frequently in the literature to study non-Bayesian inference; see, among others, Rabin and Schrag [1999], Rabin [2002], Rabin and Vayanos [2010], Benjamin *et al.* [2016], Mullainathan [2002], Gennaioli and Shleifer [2010]. The coherent inference model is formulated by applying distortion to the prior density in the Bayes' rule and replacing the likelihood with a pseudo-likelihood. This formulation allows us to study processing consistency analytically, but at a cost of not being able to nest all non-Bayesian inference models in the literature. However, the coherent inference model is still general enough in the sense that it extends some of those existing models and implies new descriptive models for non-Bayesian inference. Thus, the results obtained in this chapter are useful because (i) they provide analytical tools to verify processing consistency and whether a processing consistent model is false Bayesian in a large class of non-Bayesian inference models that can be nested in the coherent inference model, and (ii) they highlight that processing inconsistency might be caused by the indirect distortion of information through distortion of prior beliefs and by the nonadditivity of the measurement of the information contained in different pieces of a sample sequence.

Non-Bayesian inference has also been studied in the literature based on *decision-theoretic* models; see for instance Epstein [2006], Epstein *et al.* [2008], and Ortoleva [2012]. In this approach, the preferences of an individual are assumed to follow a set of axioms from which a preference model is derived, and the individual's inference about random events is embedded in her preference model. Consequently, the individual's non-Bayesian inference behavior is explained by her preferences. The coherent inference model proposed in this chapter is descriptive of some observed non-Bayesian behaviors and we do not attempt to explain why individuals exhibit these behaviors. Rather, we focus on the issue of when processing consistency holds in this model. Another difference of the coherent inference model from the those decision-theoretic models in the literature lies in that the former is mainly in the setting of objective risk where the true prior probability can be manipulated by the experimenters and thus agreed by the subjects. In the latter models, however, the prior probability is subjective and can be different across agents. Thus, the coherent inference model cannot be directly compared to the decision-theoretic models; in particular, the former does not nest the latter.

The remainder of the chapter is organized as follows. In Section 1.2, we propose the coherent inference model and study the issue of when this model is processing consistent. In Section 1.3, we provide several non-Bayesian inference models that can be nested in the coherent inference model. In Section 1.4, we prove a sufficient and necessary condition under which the coherent inference model is false Bayesian. Section 1.5 concludes and all proofs are in Appendix A.

1.2 A Coherent Inference Model

1.2.1 Model

Suppose that an individual observes a sequence of sample points and infers an unknown parameter from this sequence. The sample sequence can be processed in various ways. For instance, the individual can process the sequence as a whole or divide it into two subsequences and process them one by one. In the following, we propose an inference model that is dynamic in the sense that how many sample points have been processed previously and how many sample points will be processed next as a whole are recorded.

We use a standard and general setting. Assume that sample points take values in a topologically complete and separate metric space $\mathbb{X}$.[1] Denote $\mathbb{X}^n$ as the product space of n copies of $\mathbb{X}$, which stands for the space of sample sequences of size n, and denote $\mathbb{X}^\infty$ as the product space of countably infinite copies of $\mathbb{X}$. The sample distribution is parameterized by θ, which lives in a topologically complete and separate metric space Θ and is unknown.

From a Bayesian perspective, parameter θ is random and has a prior distribution, and the estimation of θ is achieved by computing the posterior distribution of θ given an observed sample sequence. We consider only distributions of θ that are absolutely continuous with respect to a given σ-finite measure ν on Θ. As a result, any distribution of θ can be characterized by its density π with respect to ν.

We introduce the following notations that will be frequently used in the paper: for each $\mathbf{x} = (x_1, x_2, \dots) \in \mathbb{X}^\infty$ and for each $m \geq 0$ and $n \geq 1$, denote $\mathbf{x}_{m,n} :=$

[1] A metric space is separable if it has a countable dense subset. A metric space with metric d is topologically complete if there exists another metric d' which defines the same topology as d and renders the space complete.

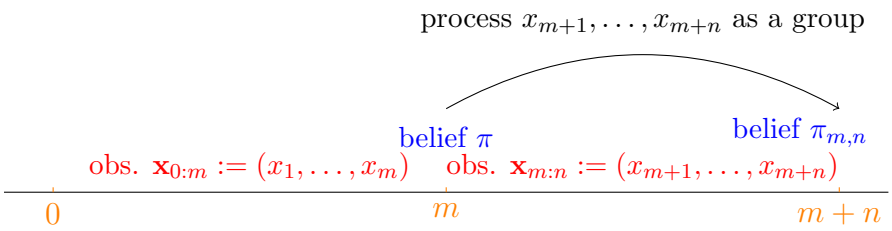


Figure 1.1: Dynamic inference model. Sample points $x_1, x_2, \ldots, x_m$ have been processed and the updated belief then is represented by π. Subsequence n sample points, $x_{m+1}, x_{m+2}, \ldots x_{m+n}$, are processed as a group and the updated belief after processing these n samples is represented by $\pi_{m,n}$.

$(x_{m+1}, \ldots, x_{m+n})$. Using this notation, we have $\mathbb{X}^n = \{\mathbf{x}_{0,n} | \mathbf{x} \in \mathbb{X}^\infty\}$, i.e., we consider $\mathbb{X}^n$ as a projection of $\mathbb{X}^\infty$.

Figure 1.1 illustrates how an individual *dynamically* updates her belief regarding the value of a parameter, e.g., θ, when sample points arrive dynamically or are divided into multi-groups and processed subsequently. Suppose that the individual has processed the first m sample points $\mathbf{x}_{0,m}$ and her belief regarding θ has been updated and is represented by a density π. The individual then processes the subsequent n sample points $\mathbf{x}_{m,n}$ *as a group* and updates her belief regarding θ from π to $\pi_{m,n}$. Note that when processing $\mathbf{x}_{m,n}$, the individual uses π, which contains the information of $\mathbf{x}_{0,m}$, as her prior belief, so the information contained in $\mathbf{x}_{0,m}$ is passed to $\pi_{m,n}$ indirectly through π. In other words, we do not record how $\mathbf{x}_{0,m}$ is processed, e.g., as a group or one-by-one, and the information of $\mathbf{x}_{0,m}$ regarding the value of θ has been coded into π.[2]

If the individual is a Bayesian agent, she will apply Bayes' rule to process sample

[2]This is also the reason why we denote the prior density used in processing $\mathbf{x}_{m,n}$ as π instead of $\pi_{0,m}$: the latter stands for the individual's belief after she processes $\mathbf{x}_{0,m}$ as a group, but the former does not imply how $\mathbf{x}_{0,m}$ is processed.

points; i.e.,

$$\pi_{m,n}(\theta \mid \mathbf{x}) = \frac{\ell_{m,n}(\theta|\mathbf{x})\pi(\theta)}{\int_\Theta \ell_{m,n}(\tilde{\theta}|\mathbf{x})\pi(\tilde{\theta})\nu(d\tilde{\theta})}, \quad \theta \in \Theta, \tag{1.1}$$

where $\ell_{m,n}(\theta|\mathbf{x})$ is the *likelihood* of θ given the first m sample points $\mathbf{x}_{0,m}$ and the subsequent n sample points $\mathbf{x}_{m,n}$. In other words, if she is using Bayes' rule to process $\mathbf{x}_{m,n}$ as a whole, the *posterior* density $\pi_{m,n}$, which stands for the agent's belief regarding θ after processing $\mathbf{x}_{m,n}$, is proportional to the likelihood of θ times the *prior* density π, which stands for the agent's belief regarding θ before processing $\mathbf{x}_{m,n}$. In other words, π stands for the agent's belief regarding θ after processing the first m sample points, i.e., is the posterior density obtained by processing the first m sample points.

In Bayesian theory, $\ell_{m,n}(\theta|\mathbf{x})$ is nothing but the conditional probability density of $\mathbf{x}_{m,n}$ given θ and $\mathbf{x}_{0,m}$. Formally, for each parameter θ, a measure Π_θ is defined on $\mathbb{X}^\infty$, representing the distribution of sample sequences under θ. Thus, we have defined a mapping from Θ to the space of probability measures on $\mathbb{X}^\infty$, which maps θ to Π_θ, and we assume this mapping to be one-to-one and measurable. It is commonly assumed in the Bayesian literature that there exists a σ-finite measure ν_X on $\mathbb{X}$ such that for any $\theta \in \Theta$ and any $n \geq 1$ the projection of Π_θ onto $\mathbb{X}^n$, which represents the distributions of sample sequences of size n, is absolutely continuous with respect to ν_X^n, the product measure of ν_X on $\mathbb{X}^n$. Then, $\ell_{m,n}(\theta|\mathbf{x})$ is the density of the conditional distribution of $\mathbf{x}_{m,n}$ given θ and $\mathbf{x}_{0,m}$ with respect to ν_X^n.

When the individual believes that sample points are i.i.d. given parameter θ, $\ell_{m,n}(\theta|\mathbf{x})$ does not depend on $\mathbf{x}_{0,m}$. Nonetheless, $\pi_{m,n}$ in (1.1) still reflects the information contained in $\mathbf{x}_{0,m}$ because this information has been coded into π and thus passed to $\pi_{m,n}$.

Bayes' rule (1.1) constitutes a rational model in the sense that the posterior density is computed from a probabilistic model: the posterior density stands for the conditional distribution of the unknown parameter given the observed samples. Given the likelihood, however, Bayes' rule can be regarded as a mapping from the space of probability densities and sample sequences to the space of probability densities. Formally, denote $\mathcal{P}(\Theta)$ as the set of probability densities on Θ.[3] Then, Bayes' rule can be represented by mappings $\mathcal{I}^B_{m,n}, m \geq 0, n \geq 1$ from $\mathbb{X}^\infty \times \mathcal{P}(\Theta)$ to $\mathcal{P}(\Theta)$, where, for each $\mathbf{x} \in \mathbb{X}^\infty$ and $\pi \in \mathcal{P}(\Theta)$, $\pi_{m,n}(\cdot|\mathbf{x}) := \mathcal{I}^B_{m,n}(\mathbf{x}, \pi)$ is defined through (1.1).

Now, we propose a new inference model, named *coherent inference model*. Again, suppose the individual has processed $\mathbf{x}_{0,m}$ and her belief has been updated to π. She is then processing subsequent sample points $\mathbf{x}_{m,n}$ as a whole and her belief will become $\pi_{m,n}$ afterwards. In the coherent inference model,

$$\pi_{m,n}(\theta|\mathbf{x}) = \frac{q_{m,n}(\theta|\mathbf{x}) g_m(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x}) g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}, \quad \theta \in \Theta. \tag{1.2}$$

We represent the coherent inference model as the set of mappings $\mathcal{I}^C_{m,n}$, $m \geq 0, n \geq 1$ from $\mathbb{X}^\infty \times \mathcal{P}(\Theta)$ to $\mathcal{P}(\Theta)$: For each $\mathbf{x} \in \mathbb{X}^\infty$ and $\pi \in \mathcal{P}(\Theta)$, $\mathcal{I}^C_{m,n}(\mathbf{x}, \pi) = \pi_{m,n}(\cdot|\mathbf{x})$ where $\pi_{m,n}$ is defined as in (1.2).

We discuss the three key components of the coherent inference model one by one. First, the model is dynamic, consisting of mappings indexed by m and n. Here, m stands for the number of sample points that have been processed previously and n stands for the number of subsequent data points that are being processed as a whole. The dynamic setting is necessary in order to study how different ways of processing samples affect the inference of unknown parameters.

[3] One may replace $\mathcal{P}(\Theta)$ with a subset of $\mathcal{P}(\Theta)$. When all bounded densities with support on finite-measured sets are under consideration, all the results in this chapter remain true.

Second, after processing $\mathbf{x}_{0,m}$, the individual's belief is represented by π and it is used as the prior belief when processing the subsequent sample points $\mathbf{x}_{m,n}$. A *distortion function* g_m is applied to the prior density π when processing $\mathbf{x}_{m,n}$. Note that the distortion function does not depend on sample points, but it can depend on m, i.e., the number of sample points that have been processed.

Third, $q_{m,n}(\theta|\mathbf{x})$ is the *pseudo-likelihood* of θ given $\mathbf{x}_{0,m}$ that has been processed and $\mathbf{x}_{m,n}$ that is being processed. The pseudo-likelihood can be regarded as the carrier of the information of observed samples on parameters, but it can be different from the likelihood in the Bayesian model.

In short, compared to the Bayesian model, the coherent inference model imposes distortion on prior densities and replaces the likelihood with a pseudo-likelihood, so it includes the Bayesian model as a special case. The coherent inference model, however, does not have a probabilistic interpretation; it can only be understood as mappings that take samples and prior densities as input and posterior densities as output. The pseudo-likelihood $\{q_{m,n}\}_{m\geq 0,n\geq 1}$ and the distortion functions $\{g_m\}_{m\geq 0}$ can then be regarded as the *parameters* of these mappings.

The motivation of the coherent inference model is two-fold: First, this model is general enough to accommodate many non-Bayesian inference models in the literature and to describe some non-Bayesian behavior; see the examples in Section 1.3. Second, this model is a consequence of assuming coherence and separability. An inference model is *coherent* if the resulting posterior density is indeed a probability density. In the coherent inference model, with nonrestrictive assumptions on $q_{m,n}$ and g_m that we will present shortly, the posterior density $\pi_{m,n}$ satisfies $\pi_{m,n}(\theta|\mathbf{x}) \geq 0, \forall \theta \in \Theta$ and $\int_\Theta \pi_{m,n}(\theta|\mathbf{x})\nu(d\theta) = 1$, showing that $\pi_{m,n}$ is a probability density. On the other hand,

in the coherent inference model, for any $\theta_1, \theta_2 \in \Theta$, we have

$$\frac{\pi_{m,n}(\theta_1|\mathbf{x})}{\pi_{m,n}(\theta_2|\mathbf{x})} = \frac{q_{m,n}(\theta_1|\mathbf{x})}{q_{m,n}(\theta_2|\mathbf{x})} \times \frac{g_m(\pi(\theta_1))}{g_m(\pi(\theta_2))}, \tag{1.3}$$

provided the denominators are nonzero. Thus, the posterior odds of θ_1 in favor of θ_2 are equal to the pseudo-likelihood ratio multiplied by the distorted prior odds. In other words, the prior density and the observed sample determine the posterior density *separately*.[4]

Before we proceed, we make the following assumption, which will be in force throughout the paper.

Assumption 1 *1. The number of elements of Θ is more than two and the support of ν is Θ. Furthermore, $\mathcal{A}_\mu := \{\theta|\nu(\{\theta\}) > 0\}$ is a closed set and ν has no atom on $\Theta/\mathcal{A}_\mu$.*

2. For each $m \geq 0$, g_m is continuous and strictly increasing and satisfies $g_m(0) = 0$.

3. For each $m \geq 0$, $n \geq 1$, (a) for each $\theta \in \Theta$, $q_{m,n}(\theta|\mathbf{x})$ depends on $\mathbf{x}_{0,m+n}$ only and is measurable in $\mathbf{x}$, and, for each $\mathbf{x} \in \mathbb{X}^\infty$, $q_{m,n}(\theta|\mathbf{x})$ is continuous in θ; and (b) for each $\mathbf{x} \in \mathbb{X}^\infty$, $q_{m,n}(\theta|\mathbf{x}) > 0$ for ν-almost everywhere (a.e.) $\theta \in \Theta$ and $\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x}) g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta}) < +\infty$ for any $\pi \in \mathcal{P}(\Theta)$.

Let us comment on this assumption. Because in most inference problems Θ contains

[4]Coherence and separability may not always apply when individuals make inferences. Indeed, Marks and Clarkson [1972] and Teigen [1974] found that subjects' estimates of posterior probabilities in their experiments were not coherent. Nevertheless, these two properties are convenient in inference modeling, and the coherent inference model satisfying them is general enough to accommodate many non-Bayesian inference models, so we do not explore, in this chapter, cases in which these two properties fail. To have a model without the coherence property, one can generalize the coherent inference model by imposing distortion on posterior densities as well. Indeed, we can prove that this generalized model is coherent if and only if there is no distortion on posterior densities; proofs can be found in Appendix B.

more than two elements, for convenience we exclude the case in which Θ contains only two elements.[5] On the other hand, it is nonrestrictive to assume that the support of ν is the whole space Θ. Otherwise, we can replace Θ with the support of ν because all the prior distributions under consideration are dominated by ν. The set $\mathcal{A}_\mu$ contains all the singletons in Θ with positive measures, i.e., contains all the atoms of ν, and is countable. Thus, it is nonrestrictive to assume that ν has no atom on $\Theta/\mathcal{A}_\nu$. Assuming $\mathcal{A}_\nu$ to be closed is of technical importance, but it is not restrictive: in many inference models, $\mathcal{A}_\mu$ is either the empty set or Θ and thus is closed.

The monotonicity and continuity of g_m are reasonable assumptions because this function represents distortion on the prior density. On the other hand, $g_m(0) = 0$ if and only if $\pi_{m,n}(\theta|\mathbf{x}) = 0$ for any $\theta \in \Theta$ such that $\pi(\theta) = 0$. Thus, the assumption $g_m(0) = 0$ essentially stipulates that if a particular θ is impossible under the prior belief, then it is also impossible under the posterior belief.

Part 3-(a) of Assumption 1 is standard. On the other hand, to make the coherent inference model (1.2) well-defined, we need to assume that $0 < \int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x}) g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta}) < +\infty$ for any prior density π and any sample $\mathbf{x}$. Note that $\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x}) g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta}) > 0$ for any prior density π if and only if $q_{m,n}(\theta|\mathbf{x}) > 0$ for ν-a.e. $\theta \in \Theta$. This assumption is satisfied by many interesting examples; see Section 1.3.3.3. However, it also fails in some cases. For instance, let $\Theta = [0,1]$ and $X = [0,1]$, and for each $\theta \in \Theta$, samples are i.i.d. and follow the uniform distribution on $[0,\theta]$. Then, the likelihood function

[5] In some inference models, such as the model of confirmatory bias proposed by Rabin and Schrag [1999], Θ contains two elements θ_1 and θ_2 only. In this case, Proposition 1 still holds with condition (i) replaced by the one that $g_m(x)/g_m(y) = g'_m(x)/g'_m(y)$ for any $x, y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$. Theorem 1 still holds with condition (i) replaced by the one that $g_m(x)/g_m(y) = x/y$ for any $x, y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$ and any $m \geq 1$. In addition, Theorem 2 also holds with the condition that g_0 is a linear function in part (i) of the theorem is replaced by the one that $g_0(x)/g_0(y) = x/y$ for any $x, y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$. Note that the condition $g_m(x)/g_m(y) = x/y$ stipulates that the distortion g_m does not affect the prior odds of θ_1 in favor of θ_2. Note also that this condition does not imply that g_m is a linear function. Proofs can be found in Appendix C.

is $\frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x)$. Given x, the likelihood is zero for $\theta \in [0, x)$.

It worth noting that the coherent inference model is general enough to nest many existing non-Bayesian inference models and imply new models of conservatism and base-rate neglect (see Section 1.3), but at the same time this model has its restrictions on belief updating. Indeed, as discussed above, in the coherent inference model, the prior density and the observed sample determine the posterior density separably. Due to such separability, the distortion applied to the prior probability is independent of the observed sample points. This is in contrast to the model proposed by Epstein [2006] and the model proposed by Ortoleva [2012] in which the prior density in the Bayes' rule is replaced by another density that can depend on the observed sample points.[6] Another restriction of the coherent inference model is that an impossible parameter value under the prior belief is also impossible under the posterior belief. In the model proposed by Ortoleva [2012], however, an impossible event under the initial prior belief chosen by the agent can become possible under the posterior belief after observing a sample point.

Before we proceed, let us define the effective domain of g_m. One can see that the prior density, which appears as the argument of g_m in the coherent inference model (1.2), may not be able to take all nonnegative real values. For instance, if $\Theta = \{\theta_1, \ldots, \theta_n\}$ and $\nu(\{\theta_i\}) > 0, i = 1, \ldots, n$, the maximum value that a density can take is $\max\{1/\nu(\{\theta_i\}) | i = 1, \ldots, n\}$. Therefore, in this case the definition of

[6]The preference models proposed by Epstein [2006] and by Ortoleva [2012] are based on several axioms of individuals' preferences, so they cannot be directly compared to the coherence inference model, which is a descriptive model of individuals' inference behavior. The interpretation of the preference models therein, however, suggests that these models can imply inference behavior that cannot be nested in the coherent inference model. In the model proposed by Epstein [2006], the agent retroactively applies a posterior probability measure that can be arbitrary, and this measure may not result from the coherent inference model. In the model proposed by Ortoleva [2012], the agent abandons the initial prior she chooses if this prior fails to pass a test. In this case, the agent chooses a new prior, and this choice depends on the observed sample points.

$g_m(z)$ for z beyond this maximum value is irrelevant, so the *effective domain* of g_m is from 0 to this maximum value. In general, one can see that the effective domain of g_m is $[0, M] \cap [0, +\infty)$, where $M := \max\{1/\nu(\{\theta\}) | \theta \in \Theta\}$ with $1/0 := +\infty$.

Proposition 1 *For any fixed $m \geq 0$ and $n \geq 1$, consider two pairs of pseudo-likelihood and distortion, $(q_{m,n}(\theta|\mathbf{x}), g_m)$ and $(q'_{m,n}(\theta|\mathbf{x}), g'_m)$, that satisfy Assumption 1. Then, $(q_{m,n}(\theta|\mathbf{x}), g_m)$ and $(q'_{m,n}(\theta|\mathbf{x}), g'_m)$ lead to the same posterior density $\pi_{m,n}$ in the coherent inference model (1.2) for any prior density π and any sample sequence $\mathbf{x}$ if and only if (i) there exists $C_m > 0$ such that $g_m(z) = C_m g'_m(z)$ for all z in the effective domain of g_m and g'_m, and (ii) for any $\mathbf{x} \in \mathbb{X}^\infty$, there exists $C_{m,n}(\mathbf{x}) > 0$ such that $q_{m,n}(\theta|\mathbf{x}) = C_{m,n}(\mathbf{x}) q'_{m,n}(\theta|\mathbf{x})$ for all $\theta \in \Theta$.*

Proposition 1 shows that if the inference behavior of an individual can be represented by the coherent inference model (1.2), the distortion g_m and the pseudo-likelihood $q_{m,n}$ in this representation are uniquely determined up to a positive scaling constant.

1.2.2 Processing Consistency

Imagine that the daily return rates of a stock are i.i.d. and the mean of the return rates is to be estimated. Suppose the daily returns in the past ten days are available. These daily returns can be processed in different ways to lead to the posterior estimate of the mean. For instance, one investor processes the ten returns as a group, i.e., simultaneously, to obtain his posterior belief while another investor processes them one by one, i.e., sequentially updates his belief after observing each return. In the Bayesian model, these two methods of data-processing result in the same posterior belief. However, this is not necessarily the case in the coherent inference model or in other non-Bayesian models.

The issue of consistency across various methods of data-processing is also observed by Benjamin *et al.* [2016] in the discussion of non-belief in the law of large numbers (NBLLN). As mentioned by Benjamin *et al.* [2016], there is little experimental evidence to show whether or not individuals are consistent when processing data in different ways; the only published findings of which we are aware of are those of Shu and Wu [2003] and Kraemer and Weber [2004], where the experimental results indicate inconsistency.

If an individual is processing inconsistent when making an inference, various issues can arise. In particular, it becomes relevant to model how the individual groups the data she received; see a full discussion in Section 5 and Appendix A of Benjamin *et al.* [2016].[7] In this chapter, we attempt to provide a characterization of processing consistency for the coherent inference model. Because of the various issues and increased modeling complexity, such as the modeling of how individuals group and process data, that arise from processing inconsistency, one might want to look for a processing-consistent non-Bayesian inference model that is able to describe some non-Bayesian behavior to some extent and at the same time retains tractability. In this case, the characterization of processing consistency becomes useful.

We first define processing consistency formally:

Definition 1 The coherent inference model (1.2) is *processing consistent* if for each $m \geq 1$, $n \geq 1$, any $\pi \in \mathcal{P}(\Theta)$, and any $\mathbf{x} \in \mathbb{X}^{\infty}$,

$$\pi_{0,m+n}(\theta|\mathbf{x}) = \pi_{m,n}(\theta|\mathbf{x}), \quad \nu\text{-a.e. } \theta \in \Theta,$$

[7] We consider only the case in which the individual groups and processes the data retrospectively. Benjamin *et al.* [2016] consider also how the individual processes the data prospectively so as to predict them. Inconsistency can also arise when the individual predicts the data. Because we are only concerned about inference, we focus on retrospective data processing.

where $\pi_{0,m+n} := \mathcal{I}^C_{0,m+n}(\mathbf{x}, \pi)$ and $\pi_{m,n} := \mathcal{I}^C_{m,n}(\mathbf{x}, \mathcal{I}^C_{0,m}(\mathbf{x}, \pi))$.

According to Definition 1, if the coherent inference model is processing consistent, then for any sample $\mathbf{x}_{0,m+n}$, processing the whole sample sequence simultaneously or dividing the sequence into two groups $\mathbf{x}_{0,m}$ and $\mathbf{x}_{m,n}$ and processing them consecutively will result in the same posterior distribution. One can see that dividing the sequence into multiple groups and processing them consecutively will lead to the same posterior distribution as well.

Note that here we assume samples points arrive or are arranged in order. An individual can partition a sample sequence in multiple pieces but cannot shuffle the sample points. This assumption is reasonable if the individual believes that the data are the sample points of a time series. When the data are i.i.d. sample points, such as the outcomes of repeated experiments, it is valid to discuss whether the individual is processing consistent if she can even shuffle the sample points. In this chapter, we focus on the case in which the individual cannot shuffle the sample points.

Theorem 1 *The coherent inference model* (1.2) *is processing consistent if and only if*

(i) for each $m \geq 1$, g_m *is a linear function in its effective domain, i.e., there exists constant* $K_m > 0$ *such that* $g_m(z) = K_m z$ *for all* z *in its effective domain; and*

(ii) for each $m \geq 1$ *and* $n \geq 1$ *and any* $\mathbf{x} \in \mathbb{X}^\infty$, *there exists* $C_{m,n}(\mathbf{x}) > 0$ *such that*

$$q_{0,m+n}(\theta|\mathbf{x}) = C_{m,n}(\mathbf{x}) q_{0,m}(\theta|\mathbf{x}) q_{m,n}(\theta|\mathbf{x}), \quad \forall \theta \in \Theta. \tag{1.4}$$

Let us explain conditions (i) and (ii) in Theorem 1 one by one. Condition (i) stipulates that the density that is obtained after processing part of a sample sequence and used as the prior density when processing the subsequent parts of the

sequence cannot be distorted. To understand this condition, fix a sample sequence $\mathbf{x}_{0,m+n}$ and consider two scenarios of processing it: processing $\mathbf{x}_{0,m}$ and $\mathbf{x}_{m,n}$ subsequently and processing $\mathbf{x}_{0,m+n}$ as a whole. The difference in the posterior densities in these two scenarios arises from 1) the difference in the pseudo-likelihood, i.e., $q_{0,m}(\theta|\mathbf{x})q_{m,n}(\theta|\mathbf{x})/q_{0,m+n}(\theta|\mathbf{x})$ and 2) the difference in whether or not distortion g_m is applied, i.e., $g_m\big(\pi_{0,m}(\theta|\mathbf{x})\big)/\pi_{0,m}(\theta|\mathbf{x})$ is a constant or not. To have processing consistency, these two sources of difference must offset each other. For fixed $\mathbf{x}_{0,m+n}$, the difference in the pseudo-likelihood is also fixed, so the second source of difference, i.e., $g_m\big(\pi_{0,m}(\theta|\mathbf{x})\big)/\pi_{0,m}(\theta|\mathbf{x})$, should be the same for any initial prior density π_0 (i.e., the individual's belief before observing $\mathbf{x}_{0,m+n}$) and thus the same for any $\pi_{0,m}$. In consequence, g_m must be a linear function.

To further understand condition (i), we consider $\Theta = \{1/4, 1/2, 3/4\}$ and i.i.d. 0-1 samples points such that the probability that each sample point takes value 1 is θ. Assume the pseudo-likelihood in the coherent inference model to be the true likelihood and $g_m(z) = \sqrt{z}$, $m \geq 0$. Suppose that two sample points, 0 and 1, are observed and the prior density π_0 before observing them is flat, i.e., $\pi_0(1/4) = \pi_0(1/2) = \pi_0(3/4) = 1/3$. If the two sample points are processed as a group, then the posterior density is

$$\pi_{0,2}(1/2) = \frac{1/4}{(3/16) + (1/4) + (3/16)} = \frac{2}{5}, \quad \pi_{0,2}(1/4) = \pi_{0,2}(3/4) = \frac{3}{10}.$$

Intuitively, π_0 is flat so the distortion applied on it has no effect. In consequence, the two sample points indicate that $\theta = 1/2$ is most likely. On the other hand, if the two sample points are processed subsequently, then the belief after processing the first sample point becomes

$$\pi_{0,1}(1/4) = \frac{3/4}{(3/4) + (1/2) + (1/4)} = \frac{1}{2}, \quad \pi_{0,1}(1/2) = \frac{1}{3}, \quad \pi_{0,1}(3/4) = \frac{1}{6}.$$

After processing the second sample point, the belief is further updated to

$$\tilde{\pi}_{0,2}(1/4) = \frac{(1/4)\sqrt{1/2}}{(1/4)\sqrt{1/2} + (1/2)\sqrt{1/3} + (3/4)\sqrt{1/6}} = \frac{\sqrt{3}}{\sqrt{3} + 2\sqrt{2} + 3},$$
$$\tilde{\pi}_{0,2}(1/2) = \frac{2\sqrt{2}}{\sqrt{3} + 2\sqrt{2} + 3}, \quad \tilde{\pi}_{0,2}(3/4) = \frac{3}{\sqrt{3} + 2\sqrt{2} + 3}.$$

Thus, when processing the two sample points separately, the individual believes that $\theta = 3/4$ is most likely. The intuition is as follows: when processing the two sample points separately, the information of the first sample point, 0, is passed to the final posterior density $\tilde{\pi}_{0,2}$ through $\pi_{0,1}$ and thus is discounted due to the distortion g_1. In consequence, the agent underestimates the information of the first sample point 0 and thus overestimates the chance of $\theta = 3/4$. By contrast, when processing the two sample points, 0 and 1, as a whole, their information is processed directly and correctly, so the individual finds that $\theta = 1/2$ is most likely.

We have showed that to have processing consistency, the density that is obtained after processing a part of a sample sequence and used as the prior density when processing the subsequent parts of the sequence cannot be distorted because it contains the information of the sequence. Note, however, that the initial prior density before processing any part of a sample sequence can be distorted because it does not contain any information of the sample sequence. Indeed, distorting initial prior density π with distortion function g_0 is equivalent to using the following prior density without distorting it:

$$\tilde{\pi}(\theta) := \frac{g_0(\pi(\theta))}{\int_\Theta g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}, \quad \theta \in \Theta. \tag{1.5}$$

Thus, one cannot distinguish between the distortion and the use of a false prior density.

Next, we explain part (ii) of the sufficient and necessary condition in Theorem 1, which is a *product rule* for the pseudo-likelihood. According to (1.2), the information contained in $\mathbf{x}_{0,m+n}$ determining the odds of θ_1 in favor of θ_2 is quantified as the log pseudo-likelihood ratio $\log[q_{0,m+n}(\theta_1|\mathbf{x})/q_{0,m+n}(\theta_2|\mathbf{x})]$ if $\mathbf{x}_{0,m+n}$ is processed as a whole. On the other hand, when $\mathbf{x}_{0,m}$ and $\mathbf{x}_{m,n}$ are processed sequentially, the information contained in $\mathbf{x}_{0,m}$ is quantified as $\log[q_{0,m}(\theta_1|\mathbf{x})/q_{0,m}(\theta_2|\mathbf{x})]$ and the information contained $\mathbf{x}_{m,n}$ is quantified as $\log[q_{m,n}(\theta_1|\mathbf{x})/q_{m,n}(\theta_2|\mathbf{x})]$. As already discussed, to achieve processing consistency, we cannot have distortion on the prior density when processing $\mathbf{x}_{m,n}$. Consequently, the information contained in $\mathbf{x}_{0,m}$ is transferred without distortion in the determination of the posterior odds given $\mathbf{x}_{0,m+n}$. As a result, the total information of $\mathbf{x}_{0,m+n}$ is $\log[q_{0,m}(\theta_1|\mathbf{x})/q_{0,m}(\theta_2|\mathbf{x})] + \log[q_{m,n}(\theta_1|\mathbf{x})/q_{m,n}(\theta_2|\mathbf{x})]$. To acheive processing consistency, this information should be the same as $\log[q_{0,m+n}(\theta_1|\mathbf{x})/q_{0,m+n}(\theta_2|\mathbf{x})]$. One can see that this is true for any θ_1 and θ_2 if and only if the product rule (1.4) holds.

To summarize, an individual can manifest processing inconsistency for two reasons. First, she believes that the information contained in a sample sequence regarding the odds of a parameter value in favor of another one, which is subjectively measured by the individual in the coherent inference model as the log pseudo-likelihood ratio of these two parameter values, is not additive; i.e., the information contained in the sample sequence as a whole is not equal to the aggregate information contained in multiple pieces of the sequence that are processed sequentially. Second, when processing a sample sequentially, the individual distorts the odds of a parameter value in favor of another given part of the sample, and this leads indirectly to the distortion of the sample information.

Finally, the Bayesian model is processing consistent because there is no distortion on prior densities and the likelihood satisfies the product rule. Indeed, likelihood

$\ell_{m,n}(\theta|\mathbf{x})$ stands for the conditional probability density of $\mathbf{x}_{m,n}$ given θ and $\mathbf{x}_{0,m}$, so it must satisfy the product rule as a result of the conditional probability formula.

1.3 Examples

1.3.1 False-Bayesian Models

If one replaces the true likelihood $\ell_{m,n}$ in the Bayesian model (1.1) with the likelihood $\tilde{\ell}_{m,n}$ of a "false" model of the underlying stochastic process driving the random samples, the resulting model is a special case of the coherent inference model and is processing consistent. We call such a model a *false-Bayesian model*. It turns out that many models in the literature fall in this category, as illustrated in the following.

Barberis *et al.* [1998] propose a model of predicting future dividend payments from historical dividend payments in order to depict investors' short-term under-reaction and long-term over-reaction to market news and to explain the medium-term momentum and long-term reversal of equity prices. The true underlying model of the dividend payments is a random walk, but the investors believe that the dividend payments follow a regime switching process with two modes, namely, mean reverting and trend following. Assuming the investors to believe in this false model, the authors are able to explain the momentum and reversal effects. Although the model proposed by Barberis *et al.* [1998] does not involve the inference of model parameters, the use of a false model is in the same flavor of a false-Bayesian inference model.

Rabin [2002] proposes a false-Bayesian inference model for the *law of small numbers*. More precisely, an urn contains red balls and blue balls and the percentage of red balls θ is the parameter to infer after a sequence of balls is drawn, with replacement, from the urn. Therefore, the true underlying model of the drawn balls

is a sequence of i.i.d. random variables with Bernoulli distribution. More precisely, let X_n be the color of the n-th ball drawn from the urn with "1" standing for blue and "0" standing for red. The likelihood of the true model is, for any $m \geq 0$, $\ell_{m,1}(\theta|\mathbf{x}) = \theta^{x_{m+1}}(1-\theta)^{1-x_{m+1}}$, $\mathbf{x} \in \{0,1\}^\infty$, $\theta \in [0,1]$. Rabin [2002] assumes that believers in the law of small numbers employ the following likelihood of a false model: for even m,

$$\tilde{\ell}_{m,1}(\theta|\mathbf{x}) = \theta^{x_{m+1}}(1-\theta)^{1-x_{m+1}}, \quad \mathbf{x} \in \{0,1\}^\infty, \theta \in [0,1]$$

and for odd m,

$$\tilde{\ell}_{m,1}(\theta|\mathbf{x}) = \begin{cases} \left(\frac{N\theta-1}{N}\right)^{x_{m+1}} \left(1-\frac{N\theta-1}{N}\right)^{1-x_{m+1}}, & x_m = 1, \\ \left(1-\frac{N(1-\theta)-1}{N}\right)^{x_{m+1}} \left(\frac{N(1-\theta)-1}{N}\right)^{1-x_{m+1}}, & x_m = 0, \end{cases} \quad \mathbf{x} \in \{0,1\}^\infty, \theta \in [0,1].$$

The parameter N measures the degree of the law of small numbers: the smaller N, the more firmly the individual believes in this law; see further the discussion in Rabin [2002].

Rabin and Vayanos [2010] study the gambler's and hot-hand fallacies using a false-Bayesian model. The observable signals are $s_t = \theta_t + \epsilon_t, t \geq 1$ where $\theta_t = \rho\theta_{t-1} + (1-\rho)(\mu + \eta_t), t \geq 1$ and $\{\eta_t\}_{t\geq 1}$ is a sequence of normal shocks with mean zero and variance σ_η^2. In the true model, ϵ_t's are i.i.d. shocks with zero mean and are independent of η_t's. However, individuals subject to the gambler's fallacy have the mistaken belief that the sequence $\{\epsilon_t\}_{t\geq 1}$ exhibits reversals, i.e., $\epsilon_t = \omega_t - \alpha\rho\sum_{k=0}^{\infty}(\delta\rho)^k\epsilon_{t-1-k}, t \geq 1$, where $\{\omega_t\}_{t\geq 1}$ is a sequence of i.i.d. normal shocks with mean zero and variance σ_ω^2. In other words, individuals subject to the gambler's fallacy believe that high realizations in the past make a low re-

alization more likely today. The parameter α measures the strength of the gambler's fallacy and δ measures the relative influence of recent realizations. Note that the observable signals take values in $\mathbb{X} := \mathbb{R}$. The parameter to be estimated is $\theta := (\rho, \mu, \sigma_\eta^2, \sigma_\omega^2) \in \Theta := [0,1] \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$. Using Kalman filtering, the likelihoods of the true model and of the false model can be computed explicitly and are actually normal densities.

1.3.2 Model of Base-Rate Neglect

Base-rate neglect is the tendency to overlook base rates, i.e., prior probabilities; see for instance Bar-Hillel [1980, 1983]. Although base-rate neglect is related to the *representativeness heuristic* [Tversky and Kahneman, 1974], the former is not necessarily caused by the latter. In addition, although in some situations base-rate neglect can be explained by the confusion between posterior probabilities and conditional probabilities (i.e., likelihood), not all experimental results showing base-rate neglect can be accounted for by such a confusion. There are many factors that contribute to base-rate neglect, such as causality, specificity, and vividness; see the detailed discussion in Bar-Hillel [1983].

We consider a special case of the coherent inference model by choosing the pseudo-likelihood as the true likelihood and setting $g_m(z) = z^\alpha, m \geq 0$. Then, the posterior density is computed as follows:

$$\pi_{m,n}(\theta|\mathbf{x}) = \frac{\ell_{m,n}(\theta|\mathbf{x})\pi(\theta)^\alpha}{\int_\Theta \ell_{m,n}(\tilde{\theta}|\mathbf{x})\pi(\tilde{\theta})^\alpha \nu(d\tilde{\theta})}, \quad \theta \in \Theta, \quad m \geq 0, n \geq 1. \tag{1.6}$$

According to Theorem 1, model (1.6) is processing inconsistent, and the inconsistency arises from the distortion of prior densities.[8] We claim that (1.6) models base-rate

[8]In model (1.6), base rates (i.e., prior densities) are neglected to the same extent whether or

neglect when $\alpha < 1$.

Consider two parameter values θ_1 and θ_2. The posterior odds of θ_1 in favor of θ_2 are

$$\frac{\pi_{m,n}(\theta_1|\mathbf{x})}{\pi_{m,n}(\theta_2|\mathbf{x})} = \frac{\ell_{m,n}(\theta_1|\mathbf{x})}{\ell_{m,n}(\theta_2|\mathbf{x})} \cdot \left(\frac{\pi(\theta_1)}{\pi(\theta_2)}\right)^{\alpha}.$$

When $\alpha = 1$, the posterior odds are the same as in the Bayesian case. When $\alpha < 1$,

$$\left|\ln\left[(\pi(\theta_1)/\pi(\theta_2))^{\alpha}\right]\right| \leq \left|\ln\left[\pi(\theta_1)/\pi(\theta_2)\right]\right|,$$

showing that the base rate has less impact on the posterior density than in the Bayesian model. Furthermore, $\lim_{\alpha\to 0} \ln\left[(\pi(\theta_1)/\pi(\theta_2))^{\alpha}\right] = 0$, showing that the base rate is fully neglected in the extreme case $\alpha = 0$.

When $\alpha > 1$, the posterior density in (1.6) weights the prior density more than in the Bayesian case; so, in this case, base-rate overweighting is modeled. In addition, model (1.6) is tractable: As illustrated by (1.5), applying the power distortion on a prior density π is equivalent to replacing π with another density $\tilde{\pi}$. Moreover, for many distribution classes, if π belongs to one of them, $\tilde{\pi}$ belongs to the same class; see Table 1.1.

---

not they contain sample information, i.e., whether or not they are obtained after part of a sample sequence is processed. In the experimental literature revealing base-rate neglect, experiments were designed in a static setting so that the subjects inferred unknown parameters for one time after observing samples. Therefore, there is little empirical evidence regarding whether individuals neglect base rates that are obtained after processing part of a sample sequence. If base rates (i.e., prior densities) are neglected only when they do not contain any sample information, we can set $g_m, m \geq 1$ as the identity function and $g_0(z) = z^{\alpha}$, and the resulting inference model is processing consistent according to Theorem 1.

Table 1.1: True prior density π and distorted prior density $\tilde{\pi} \propto \pi^{\alpha}$ for some $\alpha \geq 0$. Beta(a,b) stands for Beta distribution with density $\pi(z) \propto z^{a-1}(1-z)^{b-1}, z \in (0,1)$. Norm$(\mu,\sigma^2)$ stands for normal likelihood $f(x)$, where $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], x \in \mathbb{R}$. Gamma$(a,b)$ stands for Gamma distribution with density $\pi(z) \propto z^{a-1}e^{-bz}, x \geq 0$.

Prior Density π	Distorted Prior Density $\tilde{\pi}$
Beta(a,b)	Beta$((a-1)\alpha+1,(b-1)\alpha+1)$
Norm$(a,1/b)$	Norm$(a,1/(b\alpha))$
Gamma(a,b)	Gamma$((a-1)\alpha+1,b\alpha)$

1.3.3 Model of Conservatism

1.3.3.1 Model

Edwards [1968] conducted the following experiment: there were two urns, urn A containing 3 blue balls and 7 red ones, and urn B containing 7 blue balls and 3 red ones. A sequence of 12 balls was drawn, with replacement, from one of these two urns and the outcome is 8 reds and 4 blues. The subjects were asked to estimate the probability of those 12 balls being drawn from urn A. The prior probability is naturally 0.5, so the Bayesian posterior probability is 0.97. However, the mean estimate of the subjects was around 0.7. This example shows that the subjects over-emphasized the base rate, i.e., prior probability, relative to the sample. This phenomenon is referred to as *conservatism* and has been confirmed in many other experiments; see for instance Beach [1968], Beach *et al.* [1970], Chinnis Jr and Peterson [1968], Dave and Wolfe [2003], De Swart [1972a,b], Donnell and Du Charme [1975], Griffin and Tversky [1992], Kraemer and Weber [2004], Marks and Clarkson [1972], Nelson *et al.* [2001], Peterson and Miller [1965], Peterson and Swensson [1968], Peterson *et al.* [1965], Phillips and Edwards [1966], Sanders [1968], and Wheeler and Beach [1968].[9]

We set $g_m, m \geq 0$ in the coherent inference model (1.2) as the identity function

[9]There are also a few studies that did not find significant conservatism, such as Camerer [1987].

and choose the pseudo-likelihood to be a power transformation of the true likelihood, i.e., $q_{m,n} = \ell_{m,n}^{\beta}$ for some $\beta \geq 0$. The resulting posterior density is

$$\pi_{m,n}(\theta|\mathbf{x}) = \frac{\ell_{m,n}(\theta|\mathbf{x})^{\beta}\pi(\theta)}{\int_{\Theta} \ell_{m,n}(\tilde{\theta}|\mathbf{x})^{\beta}\pi(\tilde{\theta})\nu(d\tilde{\theta})}, \quad \theta \in \Theta, \quad \mathbf{x} \in \mathbb{X}^{\infty}, \quad m \geq 0, n \geq 1. \tag{1.7}$$

According to Theorem 1, model (1.7) is processing consistent. Moreover, for any θ_1, θ_2,

$$q_{m,n}(\theta_1|\mathbf{x})/q_{m,n}(\theta_2|\mathbf{x}) = \left(\ell_{m,n}(\theta_1|\mathbf{x})/\ell_{m,n}(\theta_2|\mathbf{x})\right)^{\beta},$$

showing that the absolute value of the log pseudo-likelihood ratio is increasing with respect to β and is the same as the absolute value of the log likelihood ratio when $\beta = 1$. Thus, the observed sample in model (1.7) with $\beta < 1$ is less informative than in the Bayesian model, which is the same effect as conservatism. On the other hand, model (1.7) with $\beta > 1$ leads to over-weighting the information of observed samples, and we call this effect *over-inference*.

Model (1.7) is considered implicitly in the aforementioned empirical studies of conservatism. Indeed, in these studies, the authors define the *accuracy ratio* as the ratio of the inferred log likelihood ratio to the true log likelihood ratio, and this accuracy ratio turns out to be the parameter β in model (1.7).

1.3.3.2 Statistical Consistency

The classical Bayesian literature shows that under certain conditions the posterior distribution of the parameter converges to the true parameter value as the sample size goes to infinity, a property known as *statistical consistency*. We show that model (1.7) is statistically consistent.

We restrict ourselves to i.i.d. sample points; i.e., we consider the following true likelihood: $\ell_{m,m+1}(\theta|\mathbf{x}) = f(x_{m+1}, \theta), \mathbf{x} \in \mathbb{X}^\infty, \theta \in \Theta, m \geq 0$ for some density function $f(\cdot, \theta)$. As in Assumption 1, we assume that for each $x \in \mathbb{X}$, $f(x, \theta)$ is continuous in θ and is strictly positive for ν-a.e. $\theta \in \Theta$. For each $\theta \in \Theta$, f is measurable in x.

Denote Π_θ as the probability measure on $\mathbb{X}^\infty$ associated with the likelihood given θ defined above. On the other hand, δ_θ denotes the Dirac measure, i.e., the point mass, at θ. Let $\Omega := \Theta \times \mathbb{X}^\infty$ and define canonical random variables $X_n(\omega) := x_n, \omega = (\theta, x_1, x_2, \dots) \in \Omega, n \geq 1$. Then, X_n's are random sample points. For each $\theta \in \Theta$, define the probability measure on Ω as $P_\theta := \Pi_\theta \times \delta_\theta$ and denote the corresponding expectation operator as $\mathbb{E}_\theta(\cdot)$. Then, under the probability measure P_θ, the sample points are i.i.d. with density function $f(\cdot, \theta)$.

Proposition 2 *Assume that Θ is a compact space. Fix any $\theta_0 \in \Theta$ such that $f(x, \theta_0) > 0$ for ν_X-a.e. $x \in \mathbb{X}$. Assume*

$$\mathbb{E}_{\theta_0}\left[\sup_{\theta \in \Theta} |\ln(f(X_i, \theta)/f(X_i, \theta_0))|\right] < \infty. \tag{1.8}$$

Then, model (1.7) is statistically consistent at θ_0; i.e., $\pi_{0,n}(\cdot|X_1, X_2, \dots)$, computed from (1.7), converges to the point mass δ_{θ_0} almost surely under P_{θ_0} as n goes to infinity.

Condition (1.8) is due to Wald [1949]. With additional assumptions regarding the prior density, this condition can be weakened. Assuming Θ to be compact is technically convenient. When Θ is not compact, with additional assumptions, the statistical consistency is still true; see Ghosh and Ramamoorthi [2003, pp. 27–28, Remarks 1.35–1.37].

Table 1.2: Posterior distribution in the model of conservatism/over-inference (1.7). Suppose the observed sample points are $x_1, \ldots, x_n$. Denote $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$. Bino$(m,p)$ stands for the binomial likelihood $f(x) = \binom{m}{x} p^x (1-p)^{m-x}, x = 0, 1, \ldots, m$. NBino$(r,p)$ stands for the negative binomial likelihood $f(x) = \binom{x+r-1}{x} p^x (1-p)^r, x \in \mathbb{Z}_+$. Poisson$(\lambda)$ stands for Poisson likelihood $f(x) = \lambda^x e^{-\lambda}/x!, x \in \mathbb{Z}_+$. Exp$(\lambda)$ stands for exponential likelihood $f(x) = \lambda e^{-\lambda x}, x \geq 0$. Norm$(\mu, \sigma^2)$ stands for normal likelihood $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], x \in \mathbb{R}$. Beta$(a,b)$ stands for Beta distribution with density $\pi(z) \propto z^{a-1}(1-z)^{b-1}, z \in (0,1)$. Gamma$(a,b)$ stands for Gamma distribution with density $\pi(z) \propto z^{a-1}e^{-bz}, x \geq 0$.

likelihood	parameter	prior	posterior
Bino(m,p)	p	Beta(a,b)	Beta$(a+\beta n\bar{x}, b+\beta n(m-\bar{x}))$
NBino(r,p)	p	Beta(a,b)	Beta$(a+\beta n\bar{x}, b+\beta nr)$
Poisson(λ)	λ	Gamma(a,b)	Gamma$(a+\beta n\bar{x}, b+\beta n)$
Exp(λ)	λ	Gamma(a,b)	Gamma$(a+\beta n, b+\beta n\bar{x})$
Norm$(\mu, 1/\tau)$	μ	Norm$(a, 1/b)$	Norm$\left(\frac{b}{b+\beta n\tau}a + \frac{\beta n\tau}{b+\beta n\tau}\bar{x}, \frac{1}{b+\beta n\tau}\right)$
Norm$(\mu, 1/\tau)$	τ	Gamma(a,b)	Gamma$\left(a+\frac{\beta n}{2}, b+\frac{\beta\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right)$

1.3.3.3 Tractability

Model (1.7) is tractable. Indeed, for almost all of the inference problems in which the Bayesian posterior distribution has a closed form, the posterior distribution in model (1.7) is also in closed form. Table 1.2 summarizes these cases. For instance, for the binomial likelihood, when the prior distribution is a Beta distribution, so is the posterior distribution in (1.7). Moreover, the adjustment of the prior mean to the posterior mean is $\frac{\beta n}{a+b+\beta nm}\left(\bar{x} - \frac{a}{a+b}m\right)$, and the magnitude of the adjustment is increasing with respect to β. Indeed, with a smaller β, an individual following inference model (1.7) becomes more conservative, so her belief in the informative character of the data set reduces and, hence, so does her degree of adjustment from the prior distribution.

1.3.3.4 Comparison to the Literature

Several models of conservatism have been proposed in the literature. Rabin [2002] proposes a false-Bayesian model of the law of small numbers. In this model, the samples are i.i.d. Bernoulli signals (taking either value a or value b), and the probability of a signal taking value a is to be estimated. When the observed sample contains the same number of a's and b's, Rabin [2002, Proposition 2] shows that conservatism is present as a result of the law of small numbers.

In the model of NBLLN by Benjamin *et al.* [2016], *on average*, the agent with NBLLN will under-infer when the sample size is larger than one, thus showing conservatism. However, under certain realizations, the agent may over-infer from the samples.[10]

Unlike the models proposed by Rabin [2002] and Benjamin *et al.* [2016], the inference model (1.7) with $\beta < 1$ leads to under-inference from samples, i.e., to conservatism, under all realizations. Let us emphasize that empirical findings show that individuals do *not* under-infer in any realization and, indeed, in some situations they over-infer. Thus, the models proposed by Rabin [2002] and Benjamin *et al.* [2016] are descriptively more accurate than the conservatism model (1.7) in this chapter. We regard our model as a convenient choice of conservatism modeling because it offers tractability, as seen in Section 1.3.3.3, and because it separates conservatism from other non-Bayesian behavior. Moreover, unlike the models of Rabin [2002] and Benjamin *et al.* [2016], which can be applied only to i.i.d. Bernoulli samples, our model can be applied to all types of samples.

[10] Although the models proposed by Rabin [2002] and Benjamin *et al.* [2016] can imply under-inference under certain realizations, these two models are more than describing conservatism. Indeed, they capture two fundamental non-Bayesian behaviors in inference, i.e., the law of small number and the non-belief in the law of large number. Under-inference can be regarded as a consequence of these two behaviors.

Based on the decision-theoretic models proposed by Epstein [2006] and Epstein *et al.* [2008], Epstein *et al.* [2010] consider non-Bayesian belief updating due to individuals' over-weighting of prior beliefs, and thus to their under-reacting to new samples, which they designate as *under-reaction*. These authors also consider the opposite of under-reaction, namely *over-reaction*. Note that under-reaction is similar to conservatism, while over-reaction is similar to over-inference in the context of our investigation. Epstein *et al.* [2010] propose the following model of under-reaction/over-reaction: the posterior distribution Π_{t+1} at time $t+1$ is calculated as $\Pi_{t+1} = (1-\gamma_t)BU(\Pi_t; x_{t+1}) + \gamma_t \Pi_t$, where Π_t is the belief in the previous period (i.e., at time t), x_{t+1} is the sample observed in the current period, $BU(\Pi_t; x_{t+1})$ is the Bayesian posterior belief obtained from Π_t and x_{t+1}, and γ_t is a number less than or equal to one. In other words, the posterior distribution is a weighted average of the Bayesian posterior belief and the prior distribution. When $\gamma_t = 0$, this model becomes a Bayesian model. When $\gamma_t < 0$, the posterior distribution places excess weight on the Bayesian posterior, so over-reaction is modeled. When $0 < \gamma_t \leq 1$, excess weight is placed on the prior distribution, so under-reaction is modeled.

Our model of conservatism/over-inference differs from and has some advantage over the model proposed by Epstein *et al.* [2010]. First, our model has tractability in the computation of posterior distributions, but theirs does not. For instance, consider i.i.d. 0-1 Bernoulli samples with the probability of the sample taking 1 to be estimated. If we assume the prior distribution to be a Beta distribution, then the Bayesian posterior is also a Beta distribution. However, a linear combination of two Beta distributions does not belong to any known distribution class. Secondly, the output of our model is always a probability density, while γ_t in their model must not be excessively negative so that Π_{t+1} remains a probability measure, and such restriction is too tight in some applications. For instance, consider i.i.d. normal

samples with unknown mean μ and known variance $1/\tau$. The prior distribution Π of μ is normal with mean a and variance $1/b$. After observing one sample x, the Bayesian posterior distribution $BU(\Pi; x)$ is normal with mean $a + (\tau/(b+\tau))(x-a)$ and variance $1/(b+\tau)$. As a result, the density of $(1-\gamma)BU(\Pi; x) + \gamma\Pi$ is

$$\begin{aligned} f(\mu) &= (1-\gamma)\frac{\sqrt{b+\tau}}{\sqrt{2\pi}} e^{-\frac{1}{2}(b+\tau)(\mu-a-(\tau/(b+\tau))(x-a))^2} + \gamma\frac{\sqrt{b}}{\sqrt{2\pi}} e^{-\frac{1}{2}b(\mu-a)^2} \\ &= \frac{\sqrt{b+\tau}}{\sqrt{2\pi}} e^{-\frac{1}{2}(b+\tau)(\mu-a-(\tau/(b+\tau))(x-a))^2} \left[1 + \gamma\left(\sqrt{\frac{b}{b+\tau}} e^{\frac{\tau}{2}\left((\mu-x)^2 - \frac{b}{b+\tau}(x-a)^2\right)} - 1\right)\right]. \end{aligned}$$

Therefore, if $\gamma < 0$, $f(\mu) < 0$ for sufficiently large μ, which shows that $(1-\gamma)BU(\Pi; x) + \gamma\Pi$ is not a probability measure.

Recall that the coherent inference model satisfies separability in the sense of (1.3). We show that if the inference model proposed by Epstein *et al.* [2010] satisfies a similar separability property that the posterior odds of any parameter value θ_1 in favor of another θ_2 depends on the sample sequence only through the likelihood ratio of θ_1 over θ_2, then it must be Bayesian or the one that never updates beliefs. To see this, consider $\Theta = \{\theta_1, \theta_2, \theta_3\}$ and $\nu(\theta_i) = 1$, $i = 1, 2, 3$. Consider prior density π such that $\pi(\theta_i) = 1/3$, $i = 1, 2, 3$. Consider one-step likelihood $\ell_{0,1}$ and $x_1, x_2 \in \mathbb{X}$ such that

$$\begin{aligned} &\ell_{0,1}(\theta_1|x_1) = 0.3, \quad \ell_{0,1}(\theta_2|x_1) = 0.1, \quad \ell_{0,1}(\theta_3|x_1) = 0.6, \\ &\ell_{0,1}(\theta_1|x_2) = 0.6, \quad \ell_{0,1}(\theta_2|x_2) = 0.2, \quad \ell_{0,1}(\theta_3|x_2) = 0.2. \end{aligned}$$

Then, the Bayesian posterior density, denoted as $\pi^B_{0,1}(\theta|x)$, is

$$\begin{aligned} &\pi^B_{0,1}(\theta_1|x_1) = 0.3, \quad \pi^B_{0,1}(\theta_2|x_1) = 0.1, \quad \pi^B_{0,1}(\theta_3|x_1) = 0.6, \\ &\pi^B_{0,1}(\theta_1|x_2) = 0.6, \quad \pi^B_{0,1}(\theta_2|x_2) = 0.2, \quad \pi^B_{0,1}(\theta_3|x_2) = 0.2. \end{aligned}$$

The posterior density $\pi_{0,1}(\theta|x)$ in the model proposed by Epstein *et al.* [2010] is given by $\pi_{0,1}(\theta|x) = (1-\gamma_0)\pi^B_{0,1}(\theta|x) + \gamma_0\pi(\theta)$ for some constant $\gamma_0 \leq 1$. In consequence, we have

$$\pi_{0,1}(\theta_1|x_1) = 0.3(1-\gamma_0) + (1/3)\gamma_0, \quad \pi_{0,1}(\theta_2|x_1) = 0.1(1-\gamma_0) + (1/3)\gamma_0,$$
$$\pi_{0,1}(\theta_1|x_2) = 0.6(1-\gamma_0) + (1/3)\gamma_0, \quad \pi_{0,1}(\theta_2|x_2) = 0.2(1-\gamma_0) + (1/3)\gamma_0.$$

In consequence,

$$\frac{\pi_{0,1}(\theta_1|x_1)}{\pi_{0,1}(\theta_2|x_1)} = \frac{0.3(1-\gamma_0) + (1/3)\gamma_0}{0.1(1-\gamma_0) + (1/3)\gamma_0}, \quad \frac{\pi_{0,1}(\theta_1|x_2)}{\pi_{0,1}(\theta_2|x_2)} = \frac{0.6(1-\gamma_0) + (1/3)\gamma_0}{0.2(1-\gamma_0) + (1/3)\gamma_0}.$$

It is straightforward to see that $\pi_{0,1}(\theta_1|x_1)/\pi_{0,1}(\theta_2|x_1) = \pi_{0,1}(\theta_1|x_2)/\pi_{0,1}(\theta_2|x_2)$ if and only if $\gamma_0 = 1$ or $\gamma_0 = 0$. Because $\ell_{0,1}(\theta_1|x_1)/\ell_{0,1}(\theta_2|x_1) = \ell_{0,1}(\theta_1|x_2)/\ell_{0,1}(\theta_2|x_2)$, we conclude that the separability property implies that $\gamma_0 = 0$ or $\gamma_0 = 1$. The former case corresponds to the Bayesian updating rule and the latter corresponds to the case in which the belief is never updated. Note that the latter case is a special case of the coherent inference model with $q_{m,n} \equiv 1$.

1.3.3.5 Base-Rate Neglect and Conservatism

Model (1.6) leads to $\ln[\pi_{m,n}(\theta_1|\mathbf{x})/\pi_{m,n}(\theta_2|\mathbf{x})] = \ln[\ell_{m,n}(\theta_1|\mathbf{x})/\ell_{m,n}(\theta_2|\mathbf{x})] + \alpha \ln[\pi(\theta_1)/\pi(\theta_2)]$ and model (1.7) leads to $\ln[\pi_{m,n}(\theta_1|\mathbf{x})/\pi_{m,n}(\theta_2|\mathbf{x})] = \beta \ln[\ell_{m,n}(\theta_1|\mathbf{x})/\ell_{m,n}(\theta_2|\mathbf{x})] + \ln[\pi(\theta_1)/\pi(\theta_2)]$. It seems that model (1.6) with $\alpha < 1$ is the same as model (1.7) with $\beta > 1$: In both cases observed samples are weighted more than prior densities in determining posterior densities. However, these two models are different.

For a non-informative sample, i.e., for $\mathbf{x}$ such that $\ln[\ell_{m,n}(\theta_1|\mathbf{x})/\ell_{m,n}(\theta_2|\mathbf{x})] = 0$, model (1.7) with $\beta > 1$ implies the same posterior density as in the Bayesian

model. However, the posterior density in model (1.6) with $\alpha < 1$ is different from the Bayesian posterior density unless the prior density is non-informative, i.e., unless $\ln[\pi(\theta_1)/\pi(\theta_2)] = 0$. Therefore, these models differ. Experiments reported in Lyon and Slovic [1976] and Tversky and Kahneman [1974] show that even with a non-informative sample, base-rate neglect is still present. Therefore, model (1.6) with $\alpha < 1$ is more suitable for modeling base-rate neglect than model (1.7) with $\beta > 1$.

Similarly, model (1.7) with $\beta < 1$ differs from model (1.6) with $\alpha > 1$: The former leads to a non-Bayesian posterior density even with a non-informative prior density, while the latter cannot. In the experiment performed by Edwards [1968], which illustrates conservatism, the prior distribution is set as non-informative, so model (1.7) with $\beta < 1$ describes conservatism more accurately than model (1.6) with $\alpha > 1$.

1.3.4 Hybrid Models

Grether [1980] performed a regression analysis on experimental data in order to test the representativeness heuristic in inference. This regression analysis implies an inference model, which turns out to be the combination of the model of base-rate neglect (1.6) with the model of conservatism (1.7).

One can also combine a false-Bayesian model with the model of conservatism (1.7). The resulting model is processing consistent according to Theorem 1.

1.3.5 Non-Belief in the Law of Large Numbers

Individuals tend to believe that "even in very large random samples, proportions might depart significantly from the overall population rate" [Benjamin *et al.*, 2016, p.2]. This phenomenon is referred to as *non-belief in the law of large numbers*

(NBLLN) and is modeled in Benjamin *et al.* [2016].

In this model, the sample points are i.i.d. Bernoulli random variables, so $\mathbb{X} = \{0, 1\}$. The probability of a sample point taking value 1 is θ, which is to be inferred from observed samples, so $\Theta = [0, 1]$. The true likelihood is

$$\ell_{m,n}(\theta|\mathbf{x}) = \left[f\left(\frac{1}{n} \sum_{i=m+1}^{m+n} x_i, \theta \right) \right]^n, \quad \mathbf{x} \in \mathbb{X}^\infty, \theta \in \Theta, m \geq 0, n \geq 1,$$

where $f(a, \theta) := \theta^a (1 - \theta)^{1-a}$. An individual with NBLLN believes in the following pseudo-likelihood:

$$q_{m,n}(\theta|\mathbf{x}) = \int_\Theta \left[f\left(\frac{1}{n} \sum_{i=m+1}^{m+n} x_i, \tilde{\theta} \right) \right]^n h(\tilde{\theta}|\theta) d\tilde{\theta}, \tag{1.9}$$

where, for each fixed θ, $h(\cdot|\theta)$ is a probability density on Θ such that $\int_\Theta \tilde{\theta} h(\tilde{\theta}|\theta) d\tilde{\theta} = \theta$. In other words, the individual pretends that the parameter value is a random variable with mean θ. The individual then applies this pseudo-likelihood and sets $g_m, m \geq 0$ as the identity function in the coherent inference model (1.2) in order to calculate the posterior density. This is the model of NBLLN proposed by Benjamin *et al.* [2016].

Benjamin *et al.* [2016] observe that the model of NBLLN is processing inconsistent. The authors thus discuss several plausible ways of grouping samples; see discussions in Benjamin *et al.* [2016]. With the help of the sufficient and necessary condition in Theorem 1, we are able to prove the processing inconsistency of the model of NBLLN for any density $h(\cdot|\theta)$.

Proposition 3 *The model of NBLLN is not processing consistent.*

Note that prior densities are not distorted in the model of NBLLN. Thus, the processing inconsistency of this model arises because the pseudo-likelihood does not sat-

isfy the product rule (1.4), i.e., because sample information contained in the pseudo-likelihood is not additive.

1.4 Processing Consistency Does Not Imply Bayes' Rule

We have shown that false-Bayesian models are processing consistent because Bayes' rule is applied in these models. Next, we study whether a processing-consistent inference model is false Bayesian. It is interesting to clarify whether processing consistency fully characterizes Bayes' rule and to draw out their difference if not.

Recall that we represent the coherent inference model (1.2) as a family of mappings $\{\mathcal{I}^C_{m,n}\}_{m\geq 0,n\geq 1}$. Similarly, we can represent a false-Bayesian model with a false underlying stochastic model (with likelihood $\tilde{\ell}_{m,n}$) as $\{\tilde{\mathcal{I}}^B_{m,n}\}_{m\geq 0,n\geq 1}$.

Definition 2 (i) The coherent inference model (1.2) is *false Bayesian in the strong sense* if there exists a false-Bayesian model $\{\tilde{\mathcal{I}}^B_{m,n}\}_{m\geq 0,n\geq 1}$ such that $\mathcal{I}^C_{m,n}(\mathbf{x},\pi) = \tilde{\mathcal{I}}^B_{m,n}(\mathbf{x},\pi)$, $m\geq 0, n\geq 1$ for any $\mathbf{x}\in\mathbb{X}^\infty$ and $\pi\in\mathcal{P}(\Theta)$.
(ii) The coherent inference model (1.2) is *false Bayesian in the weak sense* if, for any $\pi\in\mathcal{P}(\Theta)$, there exist a false-Bayesian model $\{\tilde{\mathcal{I}}^B_{m,n}\}_{m\geq 0,n\geq 1}$ and a false prior density $\tilde{\pi}\in\mathcal{P}(\Theta)$ such that $\mathcal{I}^C_{m,n}(\mathbf{x},\pi) = \tilde{\mathcal{I}}^B_{m,n}(\mathbf{x},\tilde{\pi})$, $m\geq 0, n\geq 1$ for any $\mathbf{x}\in\mathbb{X}^\infty$.

In most contexts, only one prior density is involved, and it is unobservable and subjective. Thus, when an observer tries to tell whether an individual uses the Bayesian model to make an inference, she has no knowledge of the prior density used by the individual. In this case, if the coherent inference model is false Bayesian in the weak sense, it cannot be distinguished from a false-Bayesian model. On the other hand, the examination of whether the coherent inference model is false Bayesian in the

strong sense is of theoretical interest. In addition, this examination is relevant in experimental contexts where prior densities can be controlled.

Theorem 2 *Denote the one-step pseudo-likelihood* $q_{m,1}(\theta|\mathbf{x})$ *in the coherent inference model as* $q_{m,1}(\theta|\mathbf{x}_{0,m}, x_{m+1})$.

(i) The coherent inference model (1.2) *is false Bayesian in the strong sense if and only if it is processing consistent,* g_0 *is a linear function in its effective domain, and, for each* $m \geq 0$*, there exists a measurable function* $\varphi_m(\mathbf{x}_{0,m}, x) > 0, \mathbf{x}_{0,m} \in \mathbb{X}^m, x \in \mathbb{X}$ *such that*

$$\int_{\mathbb{X}} \varphi_m(\mathbf{x}_{0,m}, \tilde{x}) q_{m,1}(\theta|\mathbf{x}_{0,m}, \tilde{x}) \nu_X(d\tilde{x}) = 1, \quad \nu\text{-a.e. } \theta \in \Theta, \quad \forall \mathbf{x}_{0,m} \in \mathbb{X}^m. \tag{1.10}$$

(ii) The coherent inference model (1.2) *is false Bayesian in the weak sense if and only if it is processing consistent and, for each* $m \geq 1$*, there exists a measurable function* $\varphi_m(\mathbf{x}_{0,m}, x) > 0, \mathbf{x}_{0,m} \in \mathbb{X}^m, x \in \mathbb{X}$ *such that* (1.10) *holds.*

Theorem 2 provides sufficient and necessary conditions under which the coherent inference model (1.2) is false Bayesian in both the strong and the weak sense. First, in both cases, processing consistency is necessary. As a result, the posterior density can be calculated by processing sample points one by one. Second, for the coherent inference model to be false Bayesian, one must be able to recast the quasi-likelihood in a way that it can be interpreted probabilistically. Mathematically, it means that the one-step pseudo-likelihood can be normalized into a likelihood, which is equivalent to condition (1.10). Third, if the coherent inference model is false Bayesian only in the weak sense, the initial prior density can be distorted and the one-step pseudo-likelihood used to process the first sample point can be arbitrary because the resulting

posterior density cannot be differentiated from one that is obtained by using a false prior density. If the coherent inference model is false Bayesian in the strong sense, however, it results in the same posterior density as a false Bayesian model for any prior density. In consequence, the initial prior density cannot be distorted and one must be able to interpret the one-step pseudo-likelihood used to process the first sample point probabilistically.

Condition (1.10) is easy to verify, as the following two examples illustrate. Moreover, these two examples show that processing consistency does not imply a false-Bayesian model even in the weak sense.

Example 1 Let $\mathbb{X} = \{0, 1\}$ and let the samples be i.i.d. Bernoulli random variables. The probability of each sample point taking value 1 is $\theta \in \Theta := [0, 1]$, which is to be estimated. The measure ν_X defined on $\mathbb{X}$ is the discrete measure, i.e., $\nu_X(\{0\}) = \nu_X(\{1\}) = 1$, and the measure ν on Θ is the Lebesgue measure. In this case, the true likelihood is

$$\ell_{m,n}(\theta|\mathbf{x}) = \theta^{\sum_{i=m+1}^{m+n} x_i}(1-\theta)^{n-\sum_{i=m+1}^{m+n} x_i}, \quad \mathbf{x} \in \mathbb{X}^\infty, \theta \in [0, 1].$$

The pseudo-likelihood in model (1.7) then becomes

$$q_{m,n}(\theta|\mathbf{x}) = \ell_{m,n}(\theta|\mathbf{x})^\beta = \theta^{\beta\sum_{i=m+1}^{m+n} x_i}(1-\theta)^{\beta(n-\sum_{i=m+1}^{m+n} x_i)}, \quad \mathbf{x} \in \mathbb{X}^\infty, \theta \in [0, 1].$$

This model is processing consistent.

Suppose there exists $\varphi_m(\mathbf{x}_{0,m}, x)$ such that $\int_{\mathbb{X}} \varphi_m(\mathbf{x}_{0,m}, \tilde{x}) q_{m,1}(\theta|\mathbf{x}_{0,m}, \tilde{x}) \nu_X(\tilde{dx}) = 1$ for ν-a.e. $\theta \in \Theta$. Then, we have $\varphi_m(\mathbf{x}_{0,m}, 1)\theta^\beta + \varphi_m(\mathbf{x}_{0,m}, 0)(1-\theta)^\beta = 1$ for ν-a.e. $\theta \in \Theta$. This is impossible, so model (1.7) with Bernoulli likelihood, although processing consistent, is not false Bayesian in the weak sense.

Example 2 Set $\mathbb{X} = \mathbb{R}$ and set the samples to be i.i.d. normal random variables with variance σ^2 and mean $\theta \in \Theta := \mathbb{R}$. The mean θ is to be estimated. Both ν_X and ν are the Lebesgue measure. In this case, the true likelihood is

$$\ell_{m,n}(\theta|\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=m+1}^{m+n}(x_i-\theta)^2\right], \quad \mathbf{x} \in \mathbb{X}^\infty, \theta \in \Theta.$$

The pseudo-likelihood in model (1.7) then becomes

$$q_{m,n}(\theta|\mathbf{x}) = \ell_{m,n}(\theta|\mathbf{x})^\beta = (2\pi\sigma^2)^{-\beta n/2} \exp\left[-\frac{\beta}{2\sigma^2}\sum_{i=m+1}^{m+n}(x_i-\theta)^2\right], \quad \mathbf{x} \in \mathbb{X}^\infty, \theta \in \Theta.$$

This model is false Bayesian in the strong sense. Indeed, g_0 is the identify function in this model, and, by choosing $\varphi_m \equiv (2\pi\sigma^2)^{(\beta-1)/2}\beta^{1/2}$, we have

$$\int_{\mathbb{X}} \varphi_m(\mathbf{x}_{0,m}, \tilde{x}) q_{m,1}(\theta|\mathbf{x}_{0,m}, \tilde{x}) \nu_X(d\tilde{x}) = \int_{\mathbb{X}} (2\pi\sigma^2/\beta)^{-1/2} \exp\left[-\frac{\beta(x-\theta)^2}{2\sigma^2}\right] \nu_X(d\tilde{x}) = 1.$$

Therefore, according to Theorem 2, this model is false Bayesian in the strong sense. Indeed, the likelihood of the corresponding false-Bayesian model is

$$\tilde{\ell}_{m,1}(\theta|\mathbf{x}) = \left(2\pi\frac{\sigma^2}{\beta}\right)^{-1/2} \exp\left[-\frac{\beta(x_{m+1}-\theta)^2}{2\sigma^2}\right], \quad \mathbf{x} \in \mathbb{X}^\infty, \theta \in \Theta.$$

In other words, given the mean θ, the variance of the sample changes from σ^2 in the true model to σ^2/β in the false model.

Finally, if we apply a nonlinear distortion g_0 to the prior density as well, the resulting model is false Bayesian in the weak but not in the strong sense.

1.5 Conclusions

In this chapter we have proposed a dynamic coherent inference model that is sufficiently general to include the existing non-Bayesian inference models as special cases and to imply new models of base-rate neglect and conservatism. The coherent inference model is constructed by distorting the prior density in Bayes' rule and replacing the likelihood with the pseudo-likelihood.

We have proved a sufficient and necessary condition under which the coherent inference model is processing consistent, i.e., the posterior density in the model is invariant to how the samples are grouped and processed. This condition can be used to check whether an inference model is processing consistent and also helps us to understand the causes of processing inconsistency: Individuals can be processing inconsistent if they distort sample information indirectly through prior densities or if they measure sample information in a non-additive way.

We have proved sufficient and necessary conditions for the coherent inference model to be false Bayesian, i.e., to be obtained by applying Bayes' rule to a false stochastic model, both in the strong and in the weak sense. Moreover, we have provided examples of inference models that are processing consistent but not false Bayesian. As a result, we have shown that processing consistency does not imply Bayes' rule.

Chapter 2

Processing Consistency in Prediction

2.1 Introduction

Inference refers to the process of inferring unknown parameters of a stochastic model after observing sample points. Processing consistency in inference refers to the property that the inference of the unknown parameters does not depend on however the observed sample points are processed *retrospectively*. In many situations, individuals need to process sample points *prospectively* before they are observed, a process named *prediction*. For instance, when deciding whether to purchase a group of sample points from some data vendor, an individual needs to predict the sample points he might receive from the vendor. In the Chapter 1, we proposed a general framework for non-Bayesian inference, named coherent inference model, and showed that this inference model is processing consistent if and only if (i) the individual cannot distort the prior density that is obtained by processing part of a sample sequence and (ii) the information contained in a sample sequence is additive when the sequence is di-

vided into multiple pieces. In this chapter, we study when an individual's prediction is processing consistent, i.e., when the individual's prediction of a group of sample points does not depend on how he processes them prospectively.

In a related work, Epstein and Le Breton [1993] show that a preference relation that satisfies certain axioms on decision under risk and the so-called dynamic consistency axiom must be represented by expected utility under a subjective probability measure, and this implies that the individual with this preference relation must use the Bayes' rule to update his belief. Our study is different from theirs in that we do not consider a decision-theoretic framework; indeed, we consider a setting in which the agent can make non-Bayesian inferences. In another related work, Benjamin *et al.* [2016] show that in the model of NBLLN, individuals are processing inconsistent in prediction. Compared to their work, we consider a general framework of prediction with unknown parameters and study the issue of processing consistency in this framework.

Imagine that an investor believes that the monthly returns of a stock follow a distribution with an unknown parameter, such as the mean of the return. The investor wants to predict the stock return in the following month based on historical data. Because he knows the distribution of the stock return as long as the parameter is known, it is natural for him to first estimate the parameter using the historical data and then use the estimate to predict the stock return in the following month. We formalize the investor's thinking as a general prediction model.

More precisely, we consider an individual who needs to predict incoming sample points based on historical samples. The individual has a stochastic model for the dynamics of the sample points with a known parameter. In other words, the individuals has a *prediction model with known parameter* for the sample points; i.e., given the parameter value, the individual knows the distribution of the sample points. The in-

dividual also has an inference model for the parameter. The individual then predicts incoming sample points by combining his inference model and prediction model with known parameter, leading to a general prediction model.

In the above prediction model, at each time, the agent needs to process the historical sample points *retrospectively* so as to make an inference of the unknown parameter and process the incoming sample points *prospectively* so as to predict these sample points. Inconsistency can arise in both the retrospective data processing period and the prospective data processing period; the former leads to inconsistency in inference and the latter leads to inconsistency in prediction. In this chapter, we focus on the issue of processing consistency in prediction. To single out this issue, we assume that the individual uses a processing consistent inference model. More precisely, we assume that the individuals uses the processing-consistent, coherent inference model (1.2) when processing data retrospectively. Note that because of Theorem 1, this inference model is fully characterized by quasi-likelihood and distortion of initial priors, and the quasi-likelihood satisfies the product rule (1.4).

The main result in this chapter shows that the above prediction model is processing consistent, i.e., the prediction of sample points do not depend on how they are processed prospectively, if and only if the quasi-likelihood in the inference model is proportional to likelihood implied by the stochastic model of the data process. In other words, the prediction model is processing consistent if and only if the individual uses the same likelihood in inference and prediction, which means that the prediction model must be described by a probabilistic model and thus the individual uses Bayes' update to predict sample points.

The above result shows that individuals must be processing inconsistent in prediction when they do not use the Bayes' rule. Given that non-Bayesian behavior is commonly observed in experimental and field studies, our result shows that how-

ever individuals process sample points prospectively matters. Then, when applying non-Bayesian prediction models, we need to specify the way of individuals processing data, and in many situations, this depends on how individuals perceive the way that the data is passed to them.

In the second part of this chapter, we consider a consumer choice model to study the impact of conservatism on consumer choices. In this model, we need to describe how individuals process data prospectively. More precisely, we consider a car selection problem in which an agent decides between two car brands. The agent can choose to purchase reviewer reports from Consumer Report to gain information about the quality of the cars in these two brands. After receiving the data, the agent decides which brand to choose. We consider two cases: in the first case, the agent decides whether to purchase a fixed amount of reports at one time. In the second case, the agent can purchase one report each time and decide when to stop purchasing additional reports. In the first case, the agent perceives that he will receive certain number of reports at one time, so when predicting these reports in order to assess their values, he tends to process them as a group. In the second case, the agent perceives that he will receive just one report every time, so he tends to process the incoming reports one by one. In both cases, we find that when the agent becomes more conservative, he tends to under-infer the quality of the cars from the purchased reports and thus (i) are less likely to revise his initial assessment of the car quality when deciding which car to choose and (ii) are less willing to pay of the reports.

2.2 Model

Following the setting in Chapter 1, we assume that sample points take values in a topologically complete and separate metric space $\mathbb{X}$. Denote $\mathbb{X}^n$ as the product space

of n copies of $\mathbb{X}$, and denote $\mathbb{X}^\infty$ as the product space of countably infinite copies of $\mathbb{X}$. Assume that there exists a σ-finite measure ν_X on $\mathbb{X}$ and denote ν_X^n as the product measure of ν_X on $\mathbb{X}^n$. The sample distribution is parametrized by θ, which is in a topologically complete and separate metric space Θ and is unknown. θ has a prior distribution, and estimation of θ is achieved by computing the posterior distribution of θ given an observed sample sequence. We consider only the distribution of θ that is absolutely continuous with respect to a given $\sigma-$finite measure ν on Θ. As a result, any distribution of θ can be characterized by its density π with respect to ν.

We consider the following dynamic prediction model: suppose the agent already observed sample points $\mathbf{x}_{0,m} := (x_1, \ldots, x_m)$ and attempts to predict the following n sample points $\mathbf{x}_{m,n} := (x_{m+1}, \ldots, x_{m+n})$. We assume that the agent believes in a stochastic model of the dynamics of the sample points parameterized by θ; consequently, he knows the distribution of the sample points to arrive if he knows θ. The agent also has an inference model for θ, which leads to a distribution θ based on the sample points he already observed. Then, the agent combines the inference model and the stochastic model to form the prediction of sample points to arrive. To study the issue of the means of prospective data-processing on prediction, we further assume that the agent is processing consistent when making the inference of θ. More precisely, we assume that the inference of θ follows the coherent inference model in Chapter 1 and is processing consistent. Consequently, we can simply denote the posterior density of θ as $\pi_m(\theta|\mathbf{x})$ without recording the means by which the agent processes $\mathbf{x}_{0,m}$ (e.g., as a group or one by one). Moreover, π_m is associated with a quasi-likelihood function q that satisfies the product rule as in (1.4) and a possibly distorted prior density $\tilde{\pi}_0$; see (1.5). On the other hand, given θ, the agent believes that the sample points follow a stochastic process parameterized by θ. Consequently, given θ and $\mathbf{x}_{0,m}$, the agent believes that the density of $\mathbf{x}_{m,n}$ (with respect to ν_X^n) is $\ell_{m,n}(\theta|\mathbf{x})$. Note that given θ,

$\ell_{m,n}$ represents the conditional probability density of $\mathbf{x}_{m,n}$ give $\mathbf{x}_{0,m}$, so when fixing $\mathbf{x}$, we can view $\ell_{m,n}$ as a likelihood function; consequently, ℓ satisfies the product rule.

Now, suppose the agent processes $\mathbf{x}_{m,n}$ prospectively as a group, we assume that his prediction of $\mathbf{x}_{m,n}$ given $\mathbf{x}_{0,m}$ is based on his inference of θ given $\mathbf{x}_{0,m}$ and his prediction of $\mathbf{x}_{m,n}$ given $\mathbf{x}_{0,m}$ and θ. More precisely, the agent predicts that the density of $\mathbf{x}_{m,n}$ is

$$\mu_{m,n}(\mathbf{x}) = \int_{\Theta} \ell_{m,n}(\theta|\mathbf{x})\pi_m(\theta|\mathbf{x})\nu(d\theta). \tag{2.1}$$

It is clear that given $\mathbf{x}_{0,m}$, $\mu_{m,n}$ is a probability density on $\mathbb{X}^n$.

Definition 3 Given a processing-consistent, coherent inference model $\{\pi_m\}$ associated with quasi-likelihood ratio $\{q_{m,n}\}$ and distortion $g_0(\cdot)$ on initial priors π_0, and a likelihood function $\{\ell_{m,n}\}$ that is used for prediction with known parameters, the prediction model (2.1) is *processing consistent* if for any $m \geq 0$, $n_1 \geq 1$, $n_2 \geq 1$, any initial prior $\pi_0 \in \mathcal{P}(\Theta)$, and any $\mathbf{x} \in \mathbb{X}^\infty$, we have

$$\mu_{m,n_1+n_2}(\mathbf{x}) = \mu_{m,n_1}(\mathbf{x})\mu_{m+n_1,n_2}(\mathbf{x}). \tag{2.2}$$

Note that given $\mathbf{x}_{0,m}$, the left-hand side of (2.2) stands for the density of $\mathbf{x}_{m,n_1+n_2}$ that is predicted by the agent when he processes $\mathbf{x}_{m,n_1+n_2}$ as a group prospectively. On the right-hand side, $\mu_{m,n_1}(\mathbf{x})$ stands for the density of $\mathbf{x}_{m,n_1}$ conditioning on $\mathbf{x}_{0,m}$ that is predicted by the agent when he processes $\mathbf{x}_{m,n_1}$ as a group prospectively. Similarly, $\mu_{m+n_1,n_2}(\mathbf{x})$ stands for the density of $\mathbf{x}_{m+n_1,n_2}$ conditioning on $\mathbf{x}_{0,m+n_1}$ that is predicted by the agent when he processes $\mathbf{x}_{m+n_1,n_2}$ as a group prospectively. Thus, the right-hand side of (2.2) represents the agent's prediction of $\mathbf{x}_{m,n_1+n_2}$ when he divides the sequence as two subsequences, $\mathbf{x}_{m,n_1}$ and $\mathbf{x}_{m+n_1,n_2}$, and process them

sequentially. Then, for the prediction model to be processing consistent, (2.1) needs to hold for any $\mathbf{x}$, π_0, m, n_1, and n_2.

Let the following assumption be in force in this section:

Assumption 2 *Suppose*

(i) Assumption 1 holds.

(ii) The inference model is processing consistent; i.e., (i) and (ii) in Theorem 1 hold.

(iii) For each $m \geq 0$, $n \geq 1$, and $\mathbf{x} \in \mathbb{X}^\infty$, $\ell_{m,n}(\theta|\mathbf{x})$ depends on $\mathbf{x}_{0,m+n}$ only, is continuous in θ, and $\ell_{m,n}(\theta|\mathbf{x}) > 0$ for ν-almost everywhere (a.e.) $\theta \in \Theta$. Moreover, for any ν-a.e. $\theta_1 \neq \theta_2$, $\ell_{m,n}(\theta_1|\mathbf{x})/\ell_{m,n}(\theta_2|\mathbf{x})$ is not a constant in $\mathbf{x}$.

(iv) $\{\ell_{m,n}\}$ is a likelihood; i.e., $\{\ell_{m,n}\}$ satisfies the product rule (1.4) and

$$\int_{\mathbb{X}} \ell_{m,m+1}(\theta|\mathbf{x})\nu_X(d\mathbf{x}_{m,m+1}) = 1, \quad \forall \theta \in \Theta, \mathbf{x}_{0,m} \in \mathbb{X}^m, \ m \geq 0.$$

We have the following main theorem:

Theorem 3 *The prediction model (2.1) is processing consistent if and only if for any $m \geq 0$, and $\mathbf{x}$, there exists $C_{0,m}(\mathbf{x}) > 0$ such that*

$$q_{0,m}(\theta|\mathbf{x}) = C_{0,m}(\mathbf{x}) l_{0,m}(\theta|\mathbf{x}). \tag{2.3}$$

Note that when (2.3) holds, the inference $\{\pi_m\}$ is the same as the one obtained in a Bayesian model with likelihood function to be $\{\ell_{m,n}\}$. Consequently, Theorem 3 shows that the prediction model is processing consistent if and only it is a Bayesian model, i.e., if and only if $\mu_{m,n}$ is the conditional probability density of $\mathbf{x}_{m,n}$ conditioning

on $\mathbf{x}_{0,m}$ in the canonical probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \mathbb{X}^\infty$, $\mathcal{F}$ is the Borel product of countably infinite copies the Borel σ-algebra of $\mathbb{X}$, and

$$\mathbb{P} = \int_\Theta \mathbb{P}_\theta \nu(d\theta),$$

where $\mathbb{P}_\theta$ is the probability measure on $(\Omega, \mathcal{F})$ induced by the likelihood function $\{\ell_{m,n}(\theta|\mathbf{x})\}$.

2.3 Example: Normal Samples with Known Variance

Next, we illustrate Theorem 3 by an example. Suppose that the samples are i.i.d. having normal distribution with unknown mean θ and known variance $1/\tau$. The prior distribution of θ is Normal with mean a and variance $1/b$. Suppose the agent uses the model of conservatism/over-inference (1.7) to infer θ from sample points. Given θ, the agent knows that the samples are i.i.d. and normally distributed with mean θ and variance $1/\tau$. Then, the agent applies (2.1) to make predictions. Suppose the agent already observed m sample points $x_1, \ldots, x_m$ and attempts to predict the distribution of the following n sample points, $x_{m+1}, \ldots, x_{m+n}$, as a group. Then, according to Table 1.2, the posterior distribution of θ, denoted as π_m, is normal distribution with mean

$$\frac{b}{b + \beta m \tau} a + \frac{\beta m \tau}{b + \beta m \tau} \bar{x}_{0,m},$$

where $\bar{x}_{0,m} := \sum_{i=1}^{m} x_i/m$, and variance

$$\frac{1}{b+\beta m\tau}.$$

Consequently,

$$\begin{aligned}\mu_{m,n}(\mathbf{x}) &= \int_{\Theta} \ell_{m,n}(\theta|\mathbf{x})\pi_m(\theta|\mathbf{x})d\theta \\ &= \int_{\Theta} \left(\sqrt{\frac{\tau}{2\pi}}\right)^n e^{-\frac{\tau}{2}\sum_{j=1}^{n}(x_j-\theta)^2} \sqrt{\frac{b+\beta m\tau}{2\pi}} e^{-\frac{b+\beta m\tau}{2}(\theta-\frac{b}{b+\beta m\tau}a-\frac{\beta m\tau}{b+\beta m\tau}\bar{x}_{0,m})^2} d\theta.\end{aligned}$$

By the property of normal distribution, we conclude that $\mu_{m,n}(\mathbf{x})$ follows normal distribution with mean vector

$$\left(\frac{b}{b+\beta m\tau}a + \frac{\beta m\tau}{b+\beta m\tau}\bar{x}_{0,m}\right)\hat{\mathbf{1}}_n,$$

where $\hat{\mathbf{1}}_n$ stands for the n-dimensional column vector with all its components to be 1, and covariance matrix

$$\frac{1}{b+\beta m\tau}\hat{\mathbf{1}}_n\hat{\mathbf{1}}_n^{\top} + \frac{1}{\tau}\mathbb{I}_n,$$

where $\mathbb{I}_n$ stands for the n-dimensional identity matrix.

Now, suppose $\mathbf{x}_{0,m}$ has already been observed and the agent predicts $\mathbf{x}_{m,n_1+n_2}$ by processing it as a group, then, he believes that the distribution of $\mathbf{x}_{m,n_1+n_2}$ is

$$\mu_{m,n_1+n_2}(\mathbf{x}) \sim \mathcal{N}\left(\left(\frac{b}{b+\beta m\tau}a + \frac{\beta m\tau}{b+\beta m\tau}\bar{x}_{0,m}\right)\hat{\mathbf{1}}_{n_1+n_2}, \frac{1}{b+\beta m\tau}\hat{\mathbf{1}}_{n_1+n_2}\hat{\mathbf{1}}_{n_1+n_2}^{\top} + \frac{1}{\tau}\mathbb{I}_{n_1+n_2}\right),$$

where $\mathcal{N}(\hat{\mathbf{d}}, \mathbb{C})$ denotes the normal distribution with mean vector $\hat{\mathbf{d}}$ and covariance matrix $\mathbb{C}$. On the other hand, if the agent predicts $\mathbf{x}_{m,n_1}$ first, which yields the

following distribution of $\mathbf{x}_{m,n_1}$ given $\mathbf{x}_{0,m}$

$$\mu_{m,n_1}(\mathbf{x}) \sim \mathcal{N}\left(\left(\frac{b}{b+\beta m\tau}a + \frac{\beta m\tau}{b+\beta m\tau}\bar{x}_{0,m}\right)\hat{\mathbf{1}}_{n_1}, \frac{1}{b+\beta m\tau}\hat{\mathbf{1}}_{n_1}\hat{\mathbf{1}}_{n_1}^{\top} + \frac{1}{\tau}\mathbb{I}_{n_1}\right),$$

and then predicts $\mathbf{x}_{m+n_1,n_2}$, which yields the following distribution of $\mathbf{x}_{m+n_1,n_2}$ given $x_{0,m+n_1}$

$$\mu_{m+n_1,n_2}(\mathbf{x}) \sim \mathcal{N}\Bigg(\left(\frac{b}{b+\beta(m+n_1)\tau}a + \frac{\beta(m+n_1)\tau}{b+\beta(m+n_1)\tau}\bar{x}_{0,m+n_1}\right)\hat{\mathbf{1}}_{n_2}, \frac{1}{b+\beta(m+n_1)\tau}\hat{\mathbf{1}}_{n_2}\hat{\mathbf{1}}_{n_2}^{\top} + \frac{1}{\tau}\mathbb{I}_{n_2}\Bigg).$$

Consequently, when processing $\mathbf{x}_{m,n_1}$ and $\mathbf{x}_{m+n_1,n_2}$ sequentially, the agent predicts $\mathbf{x}_{m,n_1+n_2}$, given $\mathbf{x}_{0,m}$, by the following distribution

$$\mu_{m,n_1}(\mathbf{x})\mu_{m+n_1,n_2}(\mathbf{x}) \sim \mathcal{N}\Bigg(\left(\frac{b}{b+\beta m\tau}a + \frac{\beta m\tau}{b+\beta m\tau}\bar{x}_{0,m}\right)\hat{\mathbf{1}}_{n_1+n_2}, \frac{1}{b+\beta m\tau}\hat{\mathbf{1}}_{n_1+n_2}\hat{\mathbf{1}}_{n_1+n_2}^{\top} + \frac{1}{\tau}\mathbb{I}_{n_1+n_2} - D\Bigg),$$

where

$$D = \begin{pmatrix} 0 & \frac{1-\beta}{b+\beta m\tau}\hat{\mathbf{1}}_{n_1}\hat{\mathbf{1}}_{n_2}^{\top}, \\ \frac{1-\beta}{b+\beta m\tau}\hat{\mathbf{1}}_{n_2}\hat{\mathbf{1}}_{n_1}^{\top}, & \frac{\beta(1-\beta)}{(b+\beta m\tau)(b+\beta(m+n_1)\tau)}\mathbb{I}_{n_2} \end{pmatrix}.$$

We can observe that $\mu_{m,n_1+n_2}(\mathbf{x})$ and $\mu_{m,n_1}(\mathbf{x})\mu_{m+n_1,n_2}(\mathbf{x})$ have the same mean, but differ in the covariance matrix by D. Consequently, they are the same if and only if $\beta = 1$, which corresponds to the Bayesian updating case. We further observe that with conservatism, which is modeled by setting $\beta < 1$, processing $\mathbf{x}_{m,n_1}$ and $\mathbf{x}_{m+n_1,n_2}$ sequentially reduces the covariance between $\mathbf{x}_{m,n_1}$ and $\mathbf{x}_{m+n_1,n_2}$ and the variance of $\mathbf{x}_{m+n_1,n_2}$ compared to processing $\mathbf{x}_{m,n_1}$ and $\mathbf{x}_{m+n_1,n_2}$ as a group. This

can be explained as follows: when processing $\mathbf{x}_{m,n_1}$ and $\mathbf{x}_{m+n_1,n_2}$ sequentially, the agent first infers the unknown mean θ from $\mathbf{x}_{m,n_1}$. Because of conservatism, the agent under-infers, meaning that he underestimates the information regarding θ contained in $\mathbf{x}_{m,n_1}$; consequently, he underestimates the information contained in $\mathbf{x}_{m,n_1}$ regarding the distribution of $\mathbf{x}_{m+n_1,n_2}$, thus underestimating (i) the covariance between $\mathbf{x}_{m,n_1}$ and $\mathbf{x}_{m+n_1,n_2}$ and (ii) the variability of $\mathbf{x}_{m+n_1,n_2}$ due to the variability of $\mathbf{x}_{m,n_1}$.

2.4 Consumption Choice Problem

We consider two variants of the consumer choice model in Benjamin *et al.* [2016] to illustrate the impact of conservatism on consumer choice and how the way of prospectively processing data affects consumer choices.

2.4.1 One-off Purchase of Signals

Consider an agent who wants to buy a car and chooses between two brands: Lada and Volvo. A typical Volvo car is better than a typical Lada car with probability θ. Denote $\omega = 1$ as the case in which the Volvo car the agent is about to choose is better than the Lada car and denote $\omega = 0$ as the case in which the Lada car is better. Denote $\mu = 1$ as the action of the agent to choose Volvo and $\mu = 0$ as the action of the agent to choose Lada. Assume the utility of the agent is $u(\mu, \omega) = \mu\omega + (1-\mu)(1-\omega)$. In other words, if the agent chooses the car which turns out to be the better one, he has one unit of utility; otherwise, he has zero utility.

The agent cannot observe θ, but he can purchase signals X_n's about the qualities of these two cars from Consumer Reports. The agent can decide whether to purchase a bunch of N signals for cost $C > 0$. To solve this decision problem, we need to model the agent's belief of the distribution of the signals.

The agent's belief consists of three parts: (i) the belief of the quality of the cars the agent is about to purchase given that θ is known; (ii) the inference of θ if the N signals are observed; and (iii) the prediction of the N signals before the agent purchases them. The agent faces a two-step decision: in the first step, he needs to decide whether to buy the N signals. In the second step, he decides which car to buy. The second decision depends only on parts (i) and (ii) of the agent's belief because this decision is made after the N signals have been observed. The first decision, however, also depends on part (iii) of the agent's belief because the prediction of the signals is needed in order to make this decision.

Given θ, the agent believes that the probability of the Volvo car being better than the Lada car is θ. For part (ii) of the agent's belief, we assume that the agent's prior belief of θ is Beta(a, b), i.e., is Beta distribution with density $\pi(z) \propto z^{a-1}(1-z)^{b-1}, z \in (0, 1)$, and he uses model (1.7) to infer θ after purchasing and observing these N signals. As a result, the agent believes that the posterior distribution of θ after observing the N signals is Beta$(a+\beta N\bar{X}_N, b+\beta N(1-\bar{X}_N))$ and the posterior mean is $Z_N := (a+\beta N\bar{X}_N)/(a+b+\beta N)$, where $\bar{X}_N := \frac{1}{N}\sum_{i=1}^{N} X_i$.

Denote $\mathcal{F}_N := \{X_1, \ldots, X_N\}$. The expected utility of the agent given the N signals is

$$\begin{aligned}\mathbb{E}\left[u(\mu,\omega)|\mathcal{F}_N\right] &= \mathbb{E}\left[\mathbb{E}\left(\mu\omega + (1-\mu)(1-\omega)|\theta, \mathcal{F}_N\right)|\mathcal{F}_N\right] \\ &= \mathbb{E}\left[\mu\theta + (1-\mu)(1-\theta)|\mathcal{F}_N\right] = \mu Z_N + (1-\mu)(1-Z_N).\end{aligned}$$

The agent chooses μ to be 1 or 0 after observing $\mathcal{F}_N$, so her choice must be 1 when $Z_N \geq 1 - Z_N$ and 0 otherwise. Note that $Z_N \geq 1 - Z_N$ if and only if $\bar{X}_N \geq \frac{1}{2} + \frac{b-a}{2\beta N}$. Therefore, the agent chooses Volvo if the sample average of the N signals is larger than the threshold $\frac{1}{2} + \frac{b-a}{2\beta N}$. When $a = b$, this threshold is independent of β, showing

that the agent's decision does not depend on the degree of conservatism, i.e., the value of β. This is because the agent's prior belief is indifferent between Volvo and Lada in this case, and the agent will choose Volvo as long as he concludes from the observed samples that Volvo is the better option. When $a > b$, the threshold is increasing in β, showing that the more conservative the agent is (i.e., the smaller β is), the more likely the agent chooses Volvo after observing the N signals. Indeed, in this case, the agent favors Vovlo in her prior belief. The more conservative he is, the less adjustment of the belief he makes after observing the signals, and consequently, the less likely he switches from Volvo to Lada. Similarly, when $a < b$, the more conservative the agent is, the more likely he chooses Lada.

Finally, we study whether the agent purchases the N signals. To this end, we need to model part (iii) of the agent's belief. In this problem, the agent knows that he will receive the N signals at one time if he buys them. Therefore, it seems reasonable to assume that the agent processes the N signals simultaneously when predicting them. We assume that the agent believes the signals are i.i.d. Bernoulli samples if θ is known. In addition, her prior belief of θ is Beta(a, b). Therefore, it is natural to model the agent's prediction of the N signals as

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_N = x_n\right) := \int_0^1 \theta^{\sum_{i=1}^N x_i}(1-\theta)^{\sum_{i=1}^N (1-x_i)} \frac{1}{\mathrm{B}(a,b)} \theta^{a-1}(1-\theta)^{b-1} d\theta, \tag{2.4}$$

which is the average of the probability densities of the signals with known θ. Here, $\mathrm{B}(a, b)$ is the Beta function.

We have shown that after observing the signals, the agent chooses $\mu = 1$ if $Z_N \geq 1 - Z_N$ and chooses $\mu = 0$ otherwise. Therefore, her maximum expected utility given $\mathcal{F}_N$ is $\max(Z_N, 1 - Z_N)$. Then, $\mathbb{E}[\max(Z_N, 1 - Z_N)]$ turns out to be the maximum

utility the agent can expect if he buys the N signals. If the agent does not buy the signals, her expected utility is obviously $\max(Z_0, 1 - Z_0)$. Therefore, the maximum price the agent is willing to pay for the signals is $\mathbb{E}[\max(Z_N, 1-Z_N)]-\max(Z_0, 1-Z_0)$.

Proposition 4 $\mathbb{E}[\max(Z_N, 1 - Z_N)]$ *is increasing in* β.

Proposition 4 shows that the more conservative the agent is, the lower price he is willing to pay for the signals. The intuition is clear: the more conservative the agent is, the less he weights the signals, and as a result the less valuable the signals are from her perspective.

Benjamin *et al.* [2016] proposed two facets of how individuals group data: (i) individuals group data *retrospectively* after they have already received data; and (ii) individuals group data *prospectively* before they receive data so as to forecast the data they will receive. In our car example, the agent processes the purchased signals retrospectively in order to decide which car to buy. To decide whether to buy the signals, the agent processes the signals prospectively so as to predict the qualities of the signals. Because we use a processing-consistent inference model, however the agent processes the signals retrospectively, he will make the same decision regarding which car to buy.

For the decision of whether to buy the signals, however, how the agent processes the signals prospectively matters. For instance, suppose the agent processes the signals one by one prospectively when predicting the signals. After having processed n signals, the agent first infers θ based on these n signals using model (1.7). Then, he predicts the $(n + 1)$-th signal based on the updated belief of θ. Therefore, the agent believes that

$$\widetilde{\mathbb{P}}\left(X_{n+1} = 1|X_1, \ldots, X_n\right) = Z_n, \quad \widetilde{\mathbb{P}}\left(X_{n+1} = 0|X_1, \ldots, X_n\right) = 1 - Z_n, \tag{2.5}$$

where $Z_n := (a + \beta \sum_{i=1}^{n} X_i)/(a + b + \beta n)$ is the posterior mean of θ given the first n signals.

Note in this case, (2.5) based on processing signals prospectively one by one is generally different from the prediction (2.4)., i.e.,

$$\mathbb{P}(X_1 = x_1, \cdots, X_N = x_n) \neq \widetilde{\mathbb{P}}(X_1 = x_1, \cdots, X_N = x_n)$$

where

$$\widetilde{\mathbb{P}}(X_1 = x_1, \cdots, X_N = x_n) = \widetilde{\mathbb{P}}(X_1 = x_1)\widetilde{\mathbb{P}}(X_2 = x_2|X_1 = x_1) \cdots \widetilde{\mathbb{P}}(X_N = x_n|X_{N-1} \cdots X_1)$$

through one-by-one processing, as shown in Theorem 3.

Under $\widetilde{\mathbb{P}}$, we also have

$$\widetilde{\mathbb{E}}\left[u(\mu, \omega)|\mathcal{F}_N\right] = \mu Z_N + (1 - \mu)(1 - Z_N),$$

so the agent's decision after purchasing N signals is the same as under $\mathbb{P}$. Same as the previous case, to maximize utility after observing N signals, μ should be 1 if $Z_N \geq 1 - Z_N$ and should be 0 otherwise. Hence

$$\widetilde{\mathbb{E}}(u(\mu, \omega)|\mathcal{F}_N] = \max(Z_N, 1 - Z_N) = \begin{cases} Z_N = \frac{a+\beta S_N}{a+b+\beta N}, & S_N \geq \frac{b-a+\beta N}{2\beta}, \\ 1 - Z_N = \frac{b+\beta(N-S_N)}{a+b+\beta N}, & S_N < \frac{b-a+\beta N}{2\beta}. \end{cases}$$

$$\widetilde{\mathbb{E}}[\max(Z_N, 1 - Z_N)] = \widetilde{\mathbb{E}}[Z_N I(S_N \geq \frac{b - a + \beta N}{2\beta})] + \widetilde{\mathbb{E}}[(1 - Z_N) I(S_N < \frac{b - a + \beta N}{2\beta})]$$

same as before, but the fact that $\widetilde{\mathbb{P}}$ uses one-by-one processing causes the prediction to be different from $\mathbb{P}$, hence $\widetilde{\mathbb{E}}$ is different from $\mathbb{E}$ because the distribution of S_N is

different.

In the following, we provide a numerical example to illustrate this. We set $N = 5$ and $a = b = 1$. We also considered the cases of $a < b$ and of $a > b$ and the results are similar, so we do not report them here. The numerical results are summarized in Figure 2.1 showing the price the agent is willing to pay for the signals with respect to β under two different ways of processing the signals prospectively: processing them as a group simultaneously and processing them one by one. First, in both processing patterns, the price is increasing with respect to β, showing that the more conservative the agent is, the less he is willing to pay for the signals. Secondly, these two processing patterns lead to the same price when $\beta = 1$ and when $\beta = 0$. In the case of $\beta = 1$, the inference model (1.7) becomes the Bayesian model, so how the agent processes the signals prospectively does not matter. In the case of $\beta = 0$, the agent believes that the signals have no information when he infers θ after observing the signals. Thus, in this case, although different patterns of prospective signal processing lead to different prediction of the signals, the agent always thinks these signals to be useless and thus is not willing to pay anything for the signals. Finally, the price is more sensitive to β when the signals are processed one by one prospectively. For instance, when the agent is conservative, i.e., $\beta < 1$, the price is lower if the agent processes the signals one by one than if he processes the signals as a group. This is because in the former case, more conservatism takes effect when the agent predicts the signals, making her believe that the signals are less informative.

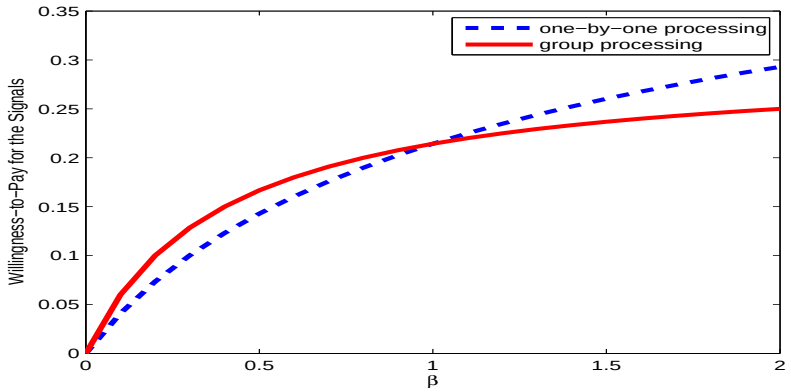


Figure 2.1: Maximum price the agent is willing to pay for the signals. That is, $\mathbb{E}[\max(Z_N, 1-Z_N)] - \mathbb{E}[\max(Z_0, 1-Z_0)]$ under two different updating schemes. The number of signals $N = 5$. The prior distribution of θ is Beta(1,1). The solid line stands for the maximum price the agent is willing to pay for the signals if he processes them as a group prospectively. The dashed line represents the price when the agent processes the signals one by one prospectively.

2.4.2 Sequential Purchase of Signals

In the previous setting, the agent decides whether to purchase a given number of signals at one time, so it is natural for the agent to process the signals he is about to buy as a group. If the agent can purchase the signals one by one, however, it is reasonable for him to process the signals one by one.

Consider a variant of the car selection problem in the previous section. We assume that the agent can purchase signals one by one at unit price $c > 0$ and decide at any time whether to stop purchasing additional signals. In this case, it is natural for the agent to infer the car quality based on the signals he already purchased and then decide whether to purchase the next signal. Thus, the agent predicts signals by processing them one by one; i.e., the agent uses one-step transition probability to update his belief. Consequently, given the n signals the consumer already purchased, the agent predicts the distribution of the next signal X_{n+1} to follow the distribution

as given in (2.5).

The consumer needs to decide the number of signals τ, which is a stopping time, he wants to purchase. At each time n, i.e., after already purchasing n signals, the consumer decides whether to stop purchasing the next signal or not. If he chooses not to purchase, then based on the n signals, the consumer believes that the probability that Volvo is better is Z_n, and the economic value of this information is $\max(Z_n, 1 - Z_n)$. If the consumer decides to purchase, then with new signal X_{n+1}, the believes the probability that Volvo is better is Z_{n+1}, and denote the continuation value as $f(n+1, Z_{n+1})$. On the other hand, the purchase incurs a cost c. Assume that the consumer is risk neutral. Then, the consumer's problem is

$$f(n, Z_n) := \max\left\{\max(Z_n, 1 - Z_n), \tilde{\mathbb{E}}[f(n+1, Z_{n+1})|\mathcal{F}_n] - c\right\}.$$

Recalling that

$$\begin{aligned} Z_{n+1} =& \frac{a + \beta S_{n+1}}{a + b + \beta(n+1)} \\ =& \frac{a + \beta(S_n + X_{n+1})}{a + b + \beta(n+1)} \\ =& \frac{a + \beta S_n}{a + b + \beta n} \frac{a + b + \beta n}{a + b + \beta(n+1)} + \frac{\beta}{a + b + \beta(n+1)} X_{n+1} \\ =& (1 - w_{n+1}) Z_n + w_{n+1} X_{n+1}, \end{aligned}$$

where $w_n = \frac{\beta}{a+b+\beta n}$, and the distribution of X_{n+1} given by (2.5), we have

$$\begin{aligned} \tilde{\mathbb{E}}[f(n+1, Z_{n+1})|\mathcal{F}_n] = & f(n+1, (1 - w_{n+1}) Z_n + w_{n+1}) Z_n \\ & + f(n+1, (1 - w_{n+1}) Z_n)(1 - Z_n). \end{aligned}$$

Therefore, the consumer solves the following dynamic programming equation

$$f(n, Z_n) = \max\Big\{ \max(Z_n, 1 - Z_n), \\ f(n+1, (1-w_{n+1})Z_n + w_{n+1})Z_n + f(n+1, (1-w_{n+1})Z_n)(1 - Z_n) - c\Big\}. \tag{2.6}$$

Note that when (i) $Z_n = 0$ or $Z_n = 1$ or (ii) $n \to +\infty$, an additional signal does not change the assessment of θ, so the optimal decision in this case must be to stop purchasing signals and, consequently, we have

$$f(n, 0) = f(n, 1) = 1, \quad \forall n \geq 0, \tag{2.7}$$

$$\lim_{n \to +\infty} f(n, z) = \max(z, 1 - z), \quad \forall z \in [0, 1]. \tag{2.8}$$

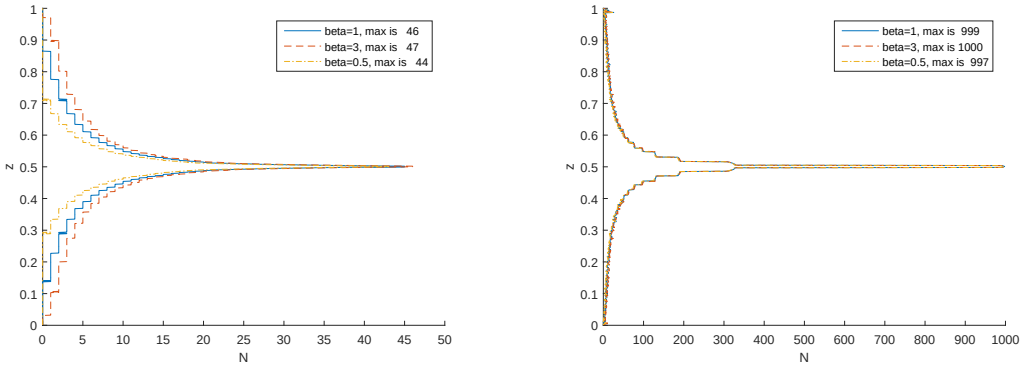
We solve (2.6)–(2.8) numerically and plot the resulting stopping region, which consists of two boundaries, $z_n^{*,u}$ and $z_n^{*,d}$, $n \geq 0$: the optimal stopping time is

$$\tau^* := \inf\{n \geq 0 \mid Z_n \geq z_n^{*,u} \text{ or } Z_n \leq z_n^{*,d}\}.$$

Setting $a = b = 1$, which implies that the prior belief of θ is the uniform distribution, we plot the boundaries of the stopping region in Figure 2.2, with the left panel corresponding to the case in which $c = 0.01$ and the right panel corresponding to the case in which $c = 0.0001$. In each panel, we consider three values of β: 0.5, 1, and 3.

We have the following observations: First, the closer to 0 or 1 the current estimate of θ is, the less likely the consumer purchases the next signal. This is because the consumer is more certain about the quality of the cars, so the value of an additional signal becomes smaller. Second, a higher cost makes the consumer purchase less signals. Third and most importantly, the smaller β is, the less signals the consumer

purchases. This is because when the consumer becomes more conservative, he believes that signals contain less information about car qualities and thus values the signals less.



(a) Boundaries for three different values of β when $c = 0.01$.

(b) Boundaries for three different values of β when $c = 0.0001$.

Figure 2.2: Upper and lower boundaries $z_n^{*,u}$ and $z_n^{*,d}$ of the stopping region of the consumer's signal purchasing problem. The consumer stops to purchase signals when his estimate Z_n of the probability θ that Volvo is better hits one of the two boundaries. a and b are set to be 1 so that the prior belief of θ is the uniform distribution. The unit cost of purchasing a signal is set to be 0.01 in the left panel and 0.0001 in the right panel. In each panel, β takes three values: 0.5, 1, and 3. 'Max' stands for the lowest number n such that given n signals have been purchased, the consumer will not purchase the next signal even if his current estimate of θ is 0.5.

2.5 Conclusion

In many situations, individuals believe in some stochastic models for sample points with certain parameter. When the parameter is unknown, to predict samples points to arrive, the individuals first infer the unknown parameter by processing the sample points that are already observed and then process the sample points to arrive by combining the stochastic model and the estimate of the parameter; the former stage is called retrospective data processing so as to make inferences and the latter is

prospective data processing so as to make predictions. In this chapter, assuming processing consistency in retrospective data processing, we prove that the individuals' prediction is processing consistent if and only if they use a probabilistic model, i.e., uses the Bayes' rule to make predictions. This result highlights the need of modeling the means of data processing in a non-Bayesian prediction model.

We then consider a car selection model in which an agent needs to decide whether to purchase reports that signal the quality of cars in two brands and then decide which brand to choose. We assume that the agent is conservative when making inferences. We consider two cases: in the first case, the agent needs to make a one-off purchase of a certain number of reports; in the second case, the agent can purchase one signal at one time and decides when to stop purchasing additional signals. Thus, in the first case, the agent processes reports he is about to purchase as a group and in the second case, he processes reports one by one. We find that due to conservatism, the consumer undervalues the signals and thus pay less for the signals or purchase fewer signals.

Chapter 3

Asset Pricing Applications

3.1 Introduction

In traditional consumption-based asset pricing models, it is assumed that the distributions of the consumption and dividend streams are known to investors. In practice, however, such distributions are unknown. For instance, investors may not know the mean of the growth rate of the dividend of a stock and have to estimate it from historical data. Thus, statistical inference necessarily plays a role in investors' portfolio selection and thus affects asset prices. In this chapter, we consider two asset pricing models in which some model parameters are unknown and thus must be learnt by the investors from historical data. The investors can be irrational, e.g., subject to the conservatism bias, when learning the parameters, and we study the impact of the irrationality on asset prices.

In the first model, we generalize the classical consumption-based asset pricing model by assuming the mean of the consumption and dividend growth rates to be unknown. The representative agent in the market must learn the mean from the historical data, but he may be conservative when making the inference. By assuming

Bernoulli distribution for the dividend growth rates with unknown mean, we solve the equilibrium asset prices in closed form. We find that when the representative agent becomes more conservative, the risk-free return and the price-dividend ratio become less sensitive to the number of good signals in the historical dividend data, and the risk premium becomes lower.

In the second model, we generalize the asset pricing model in Barberis *et al.* [1998] by assuming some of the model parameters to be unknown. More precisely, as in Barberis *et al.* [1998], we assume that the true model of the earnings of a stock is a random walk, but the representative investor in the market wrongly believes in a regime switching model: one is a trending regime and the other one is a mean-reverting regime. Unlike Barberis *et al.* [1998], however, we assume that the model parameters, i.e., the probability of generating a good earning in these two regimes and the transition probability from one regime to the other, are unknown. We assume that the agent uses the model of conservatism to infer the parameters and then evaluate the stock price by the standard dividend discount model. We compute the stock price in closed form and study whether this model generates short-term momentum and long-term reversal as in Barberis *et al.* [1998] where the model parameters are known. We find that the more conservative the agent is, the less profound the effect of short-term momentum and long-term reversal. Also, by comparing with the results of historical simulation using annual return and net income data of U.S. stocks from 1974 to 2016, we find that investors over-extrapolate the signals in annual income when making an inference.

3.2 Consumption-Based Asset Pricing Model

3.2.1 Model

We consider a representative agent whose preferences for discrete-time consumption stream $\{C_t\}$ is represented by expected utility; i.e., the preference value at time t of the consumption stream $C_s, s \geq t$ is

$$V_t(C) := \mathbb{E}_t \left[\sum_{s=t}^{\infty} \rho^{s-t} u(C_s) \right],$$

where $\mathbb{E}_t$ stands for the expectation conditioning on the information at time t, β is the discount rate, and the utility function

$$u(x) := \begin{cases} \frac{x^{1-\gamma}-1}{1-\gamma}, & \gamma \neq 1, \\ \log(x), & \gamma = 1, \end{cases}$$

for some $\gamma > 0$ that represents the degree of relative risk aversion of the agent.

We assume that the agent can trade a stock with price P_t at time t and a risk-free asset with gross return rate in period t to $t+1$ to be $R_{f,t+1}$. The stock pays a dividend D_t at time t, so the gross return of holding the stock in period t to $t+1$ is

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t}.$$

We assume that there is no endowment other than the stock dividend in the market, so the aggregate consumption must be the same as the stock dividend in equilibrium. We assume that the log dividend growth rates $d_{t+1} := \ln(D_{t+1}/D_t)$, $t \geq 0$ are i.i.d., but the distribution is unknown to the agent. More precisely, we assume that d_{t+1} follows Bernoulli distribution that takes two values y and z with

probabilities p and $1-p$, respectively, and $y > z$. We assume that y are z are known, but p is unknown and thus must be learnt by the agent using historical data.

When p is known, the risk-free return and the stock price can be computed by the Euler equations. More precisely, the risk-free rate in period t to $t+1$ given parameter p is

$$\begin{aligned}\mathcal{R}_{f,t+1}(p) &= \left(\mathbb{E}_t\left[\rho\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} \mid p\right]\right)^{-1} \\ &= \left(\mathbb{E}_t\left[\rho e^{-\gamma d_{t+1}} \mid p\right]\right)^{-1} \\ &= \left(\rho\left(e^{-\gamma y}p + e^{-\gamma z}(1-p)\right)\right)^{-1}.\end{aligned}$$

The stock price given p is

$$\begin{aligned}\mathcal{P}_t(p) &= \mathbb{E}_t\left[\sum_{j=1}^{\infty}\rho^j(\frac{C_{t+j}}{C_t})^{-\gamma}(\frac{D_{t+j}}{D_t})D_t \mid p\right] \\ &= D_t\sum_{j=1}^{\infty}\rho^j\left(\mathbb{E}\left[(\frac{C_{t+1}}{C_t})^{-\gamma}(\frac{D_{t+1}}{D_t}) \mid p\right]\right)^j \\ &= D_t\sum_{j=1}^{\infty}\rho^j\left(\mathbb{E}\left[e^{(1-\gamma)d_{t+1}} \mid p\right]\right)^j \\ &= D_t\frac{\rho[e^{(1-\gamma)y}p + e^{(1-\gamma)z}(1-p))]}{1-\rho[e^{(1-\gamma)y}p + e^{(1-\gamma)z}(1-p))]}.\end{aligned}$$

When p is unknown, the agent estimates p from the historical data. We assume that the agent's prior distribution of p is Beta with parameter (a, b) and the agent uses the model of conservatism (1.7) to infer p. Consequently, the asset prices are

$$R_{f,t+1} = \mathbb{E}_t\left[\mathcal{R}_{f,t+1}(\tilde{p})\right], \tag{3.1}$$

$$\frac{P_t}{D_t} = \mathbb{E}_t\left[\frac{\rho[e^{(1-\gamma)y}\tilde{p} + e^{(1-\gamma)z}(1-\tilde{p}))]}{1-\rho[e^{(1-\gamma)y}\tilde{p} + e^{(1-\gamma)z}(1-\tilde{p}))]}\right], \tag{3.2}$$

where $\tilde{p}$ follows the agent's posterior distribution of p given the information at time t. More precisely, according to Table 1.2, the distribution of $\tilde{p}$ is

$$\text{Beta}\left(a+\beta\sum_{i=1}^{t}x_i, b+\beta(t-\sum_{i=1}^{t}x_i)\right) \tag{3.3}$$

where

$$x_i := \begin{cases} 1, & d_i = y, \\ 0, & d_i = z. \end{cases}$$

Note that $\tilde{p}$ can take any values in $[0,1]$, so for the stock price to be well defined, we need the make the following assumption on the dividend growth rate:

Assumption 3 *$\rho e^{(1-\gamma)y} < 1$ and $\rho e^{(1-\gamma)z} > 1$.*

3.2.2 Numerical Simulation

Next, we study the impact of conservatism on asset pricing. We set $\rho = 0.9$, $\gamma = 2$, $y = 0.1$, and $z = -0.1$. We assume that the agent's prior belief of p is Beta(1,1), i.e., the uniform distribution, and he observed $t = 10$ signals, among which $\sum_{i=1}^{t} x_i$ are good signals.

3.2.2.1 Risk-free Return

We plot the risk-free gross return $R_{f,t+1}$ as a function of $\sum_{i=1}^{t} x_i$, the number of good signals, for three values of β, 0.5, 1, and 3, in Figure 3.1. We can see that conservatism ($\beta < 1$) leads to a higher risk-free rate when most of the signals are bad but to a lower risk-free rate when most of the signals are good; in other words, the risk-free rate is less sensitive with respect to the signals. Intuitively, this is the

case because the agent is conservative and thus believes that the signals contains less information about the dividend growth rate than it really is.

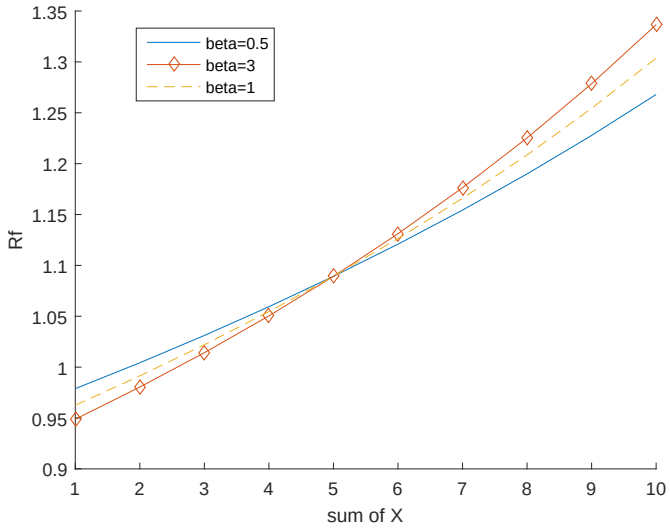


Figure 3.1: Risk-free gross return as a function of $\sum_{i=1}^{t} x_i$. Here, we set $\rho = 0.9$, $\gamma = 2$, $y = 0.1$, $z = -0.1$, and $t = 10$. β takes three values: 0.5, 1, and 3.

3.2.2.2 Price-Dividend Ratio

Next, we plot the price dividend ratio P_t/D_t as a function of β and as a function of $\sum_{i=1}^{t} x_i$ in the left and right panels, respectively, of Figure 3.2. As we can see from the right panel, the price-dividend ratio is less sensitive to the signals when the agent is more conservative, i.e., when β is smaller.

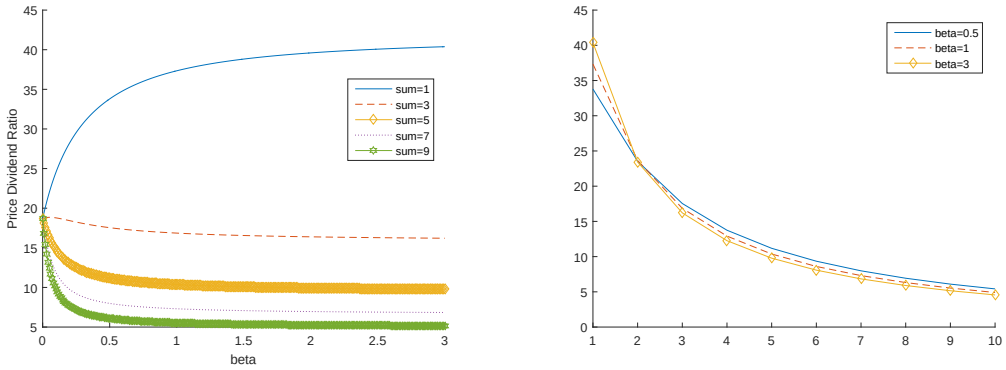


Figure 3.2: Price-dividend ratio as a function of β (left panel) and as a function of $\sum_{i=1}^{t} x_i$ (right panel). Here, we set $\rho = 0.9$, $\gamma = 2$, $y = 0.1$, $z = -0.1$, and $t = 10$. $\sum_{i=1}^{t} x_i$ takes five values: 1, 3, 5, 7, and 9 in the left panel and β takes three values: 0.5, 1, and 3 in the right panel.

3.2.2.3 Risk Premium

Note that the gross return rate of the stock is

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} = \frac{P_{t+1}/D_{t+1} + 1}{P_t/D_t} \cdot \frac{D_{t+1}}{D_t}.$$

To compute the expected return $\mathbb{E}_t[R_{t+1}]$, we first compute

$$P_{t+1}/D_{t+1} = f(d_1, \ldots, d_t, d_{t+1})$$

as a function of d_{t+1}, with $d_1, \ldots, d_t$ given. Then,

$$\begin{aligned}\mathbb{E}_t\left[R_{t+1}\right] &= \mathbb{E}_t\left[\frac{f(d_1, \ldots, d_t, d_{t+1})}{P_t/D_t} \cdot e^{d_{t+1}}\right] \\ &= \mathbb{E}_t\left[\frac{f(d_1, \ldots, d_t, y)}{P_t/D_t} \cdot e^{y}\tilde{p} + \frac{f(d_1, \ldots, d_t, z)}{P_t/D_t} \cdot e^{z}(1-\tilde{p})\right],\end{aligned}$$

where $\tilde{p}$ follows distribution (3.3).

We plot the conditional one-period risk premium, $\mathbb{E}_t[R_{t+1}] - R_{f,t+1}$, as a function of $\sum_{i=1}^{t} x_i$ in Figure 3.3. We can observe that conservatism leads to a lower equity premium.

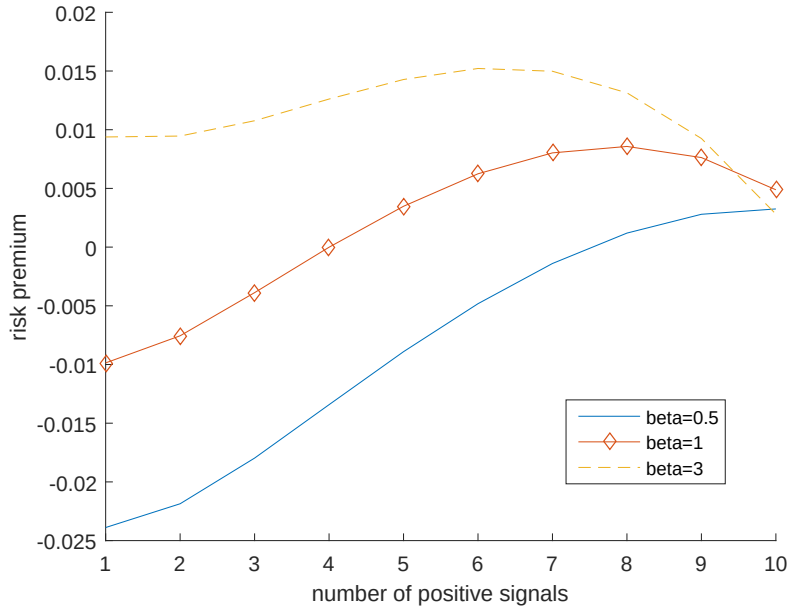


Figure 3.3: Risk Premium as a function of $\sum_{i=1}^{t} x_i$. Here, we set $\rho = 0.9$, $\gamma = 2$, $y = 0.1$, $z = -0.1$, and $t = 10$. β takes three values: 0.5, 1, and 3.

3.2.2.4 Return Volatility

Next, we compute the conditional variance of the one-period stock return R_{t+1}. Recall that

$$R_{t+1} = \frac{f(d_1, \ldots, d_t, d_{t+1}) + 1}{P_t/D_t} \cdot e^{d_{t+1}}$$

and $\tilde{p}$ follows distribution (3.3). We can compute the variance of R_{t+1} conditioning on information at time t. We plot the variance as a function of $\sum_{i=1}^{t} x_i$ in Figure 3.4. We can observe that conservatism leads to a lower conditional variance of the stock return when most signals are bad and to a higher conditional variance when most

signals are good.

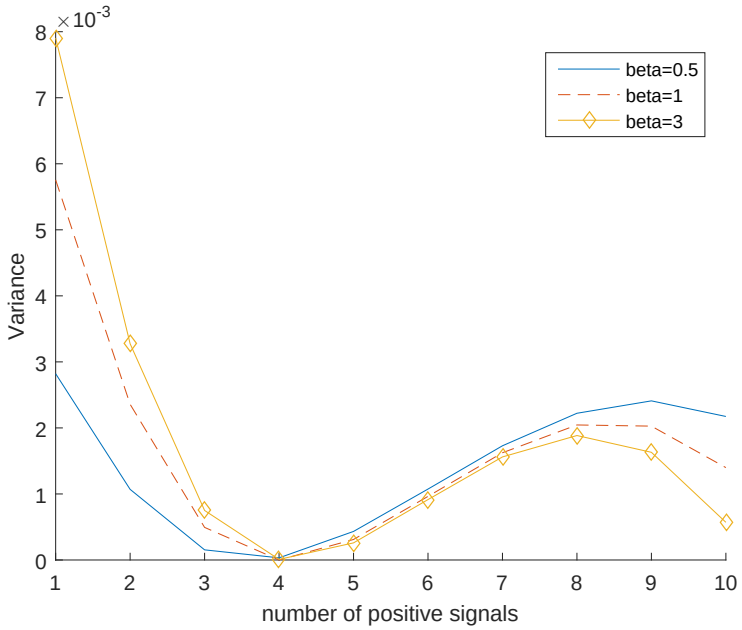


Figure 3.4: Conditional variance of one-period stock return as a function $\sum_{i=1}^{t} x_i$. Here, we set $\rho = 0.9$, $\gamma = 2$, $y = 0.1$, $z = -0.1$, and $t = 10$. β takes three values: 0.5, 1, and 3.

3.3 BSV Model with Learning

3.3.1 Model

Barberis *et al.* [1998] propose an asset pricing model that can generate both underreaction and overreaction of asset prices to market news. More precisely, they consider a stock with earning stream N_t, $t \geq 0$, and all earnings are paid out as dividends. Denoting $Y_t = N_t - N_{t-1}$ as the shock to the earning at time t, the authors assume that Y_t can take only two values: y and $-y$ for some constant $y > 0$. The authors further assume that the true process for the earnings is a random walk. The representative agent, however, believes that the earnings follow a regime-switching

Regime 1	$y_{t+1} = y$	$y_{t+1} = -y$
$y_t = y$	π_L	$1 - \pi_L$
$y_t = -y$	$1 - \pi_L$	π_L

Regime 2	$y_{t+1} = y$	$y_{t+1} = -y$
$y_t = y$	π_H	$1 - \pi_H$
$y_t = -y$	$1 - \pi_H$	π_H

Table 3.1: Transition matrices in two regimes of the earning process believed by the representative agent.

	$s_{t+1} = 1$	$s_{t+1} = 2$
$s_t = 1$	$1 - \lambda_1$	λ_1
$s_t = 2$	λ_2	$1 - \lambda_2$

Table 3.2: Transition matrices in two regimes of the earning process believed by the representative agent.

model with two regimes: 1 and 2. In both regimes, the earnings are Markov processes, with transition matrices listed in Table 3.1. Here, without loss of generality, we assume $0 < \pi_L < 0.5$ and $0.5 < \pi_H < 1$, so Regime 1 becomes a *mean-reverting* regime and Regime 2 is a *trending* regime. We denote $s_t = i$ as the case in which the market is under Regime i at time t, and we assume that the agent believes that the transition matrix for the regime process $\{s_t\}$ is given by Table 3.2.

Barberis *et al.* [1998] assume that the model parameters, λ_i, $i = 1, 2$ and π_j, $j \in \{H, L\}$ are known. Then, the authors compute the distribution of the dividend process, i.e., the distribution of $\{Y_t\}$, given the set of model parameters $\theta := (\lambda_1, \lambda_2, \pi_H, \pi_L)$. More precisely, denoting $\mathcal{F}_t := \{Y_s, s \leq t\}$ as the information available at time t, the authors first compute the probability that the market is under Regime 1 at time t:

$$q_t(\theta) = \mathbb{P}(s_t = 1 | \mathcal{F}_t, \theta).$$

It turns out that $\{q_t\}$ can be computed recursively:

$$q_{t+1}(\theta) = \begin{cases} \frac{\big((1-\lambda_1)q_t(\theta)+\lambda_2\big(1-q_t(\theta)\big)\big)\pi_L}{\big((1-\lambda_1)q_t(\theta)+\lambda_2\big(1-q_t(\theta)\big)\big)\pi_L+\big(\lambda_1 q_t(\theta)+(1-\lambda_2)\big(1-q_t(\theta)\big)\big)\pi_H}, & Y_{t+1}\cdot Y_t = 1, \\ \frac{\big((1-\lambda_1)q_t(\theta)+\lambda_2\big(1-q_t(\theta)\big)\big)(1-\pi_L)}{\big((1-\lambda_1)q_t(\theta)+\lambda_2\big(1-q_t(\theta)\big)\big)(1-\pi_L)+\big(\lambda_1 q_t(\theta)+(1-\lambda_2)\big(1-q_t(\theta)\big)\big)(1-\pi_H)}, & Y_{t+1}\cdot Y_t = -1; \end{cases} \tag{3.4}$$

see Barberis *et al.* [1998, p. 323]. Consequently, the one-step transition probability of $\{Y_t\}$ is

$$\begin{aligned} L_{t,1}(\theta|\mathbf{y}) :&= \mathbb{P}(Y_{t+1}=y_{t+1}|\mathcal{F}_t,\theta) = \mathbb{P}(Y_{t+1}=y_{t+1}|\mathcal{F}_t,\theta,s_t=1)\mathbb{P}(s_t=1|\mathcal{F}_t,\theta) \\ &\quad + \mathbb{P}(Y_{t+1}=y_{t+1}|\mathcal{F}_t,\theta,s_t=2)\mathbb{P}(s_t=2|\mathcal{F}_t,\theta) \qquad (3.5) \\ &= q_t(\theta)\mathbb{P}(Y_{t+1}=y_{t+1}|\mathcal{F}_t,\theta,s_t=1) + (1-q_t(\theta))\mathbb{P}(Y_{t+1}=y_{t+1}|\mathcal{F}_t,\theta,s_t=2) \\ &= \begin{cases} q_t(\theta)\pi_L + (1-q_t(\theta))\pi_H, & y_{t+1}\cdot y_t = 1, \\ q_t(\theta)(1-\pi_L) + (1-q_t(\theta))(1-\pi_H), & y_{t+1}\cdot y_t = -1. \end{cases} \qquad (3.6) \end{aligned}$$

Barberis *et al.* [1998] assume that the representative agent values the stock by

$$P_t(\theta) = \mathbb{E}_t\left[\sum_{s=1}^{\infty}\frac{N_{t+s}}{(1+\delta)^s} \mid \theta\right],$$

where $\delta > 0$ is a discount rate. Barberis *et al.* [1998] find $P_t(\theta)$ in closed form. More precisely,

$$P_t(\theta) = \frac{N_t}{\delta} + y_t(p_1(\theta) - p_2(\theta)q_t(\theta)), \tag{3.7}$$

where

$$p_1(\theta) = \frac{1}{\delta}(\gamma_0^\top(1+\delta)[\mathbb{I}_4(1+\delta) - Q]^{-1}Q\gamma_1),$$
$$p_2(\theta) = -\frac{1}{\delta}(\gamma_0^\top(1+\delta)[\mathbb{I}_4(1+\delta) - Q]^{-1}Q\gamma_2),$$

$\mathbb{I}_4$ is the 4-by-4 identity matrix, and

$$\gamma_0 = (1,-1,1,-1)^\top, \quad \gamma_1 = (0,0,1,0)^\top, \quad \gamma_2 = (1,0,-1,0)^\top,$$
$$Q = \begin{pmatrix} (1-\lambda_1)\pi_L & (1-\lambda_1)(1-\pi_L) & \lambda_1\pi_H & \lambda_1(1-\pi_H) \\ (1-\lambda_1)(1-\pi_L) & (1-\lambda_1)\pi_L & \lambda_1(1-\pi_H) & \lambda_1\pi_H \\ \lambda_2\pi_L & \lambda_2(1-\pi_L) & (1-\lambda_2)\pi_H & (1-\lambda_2)(1-\pi_H) \\ \lambda_2(1-\pi_L) & \lambda_2\pi_L & (1-\lambda_2)(1-\pi_H) & (1-\lambda_2)\pi_H \end{pmatrix}.$$

Barberis *et al.* [1998] find that the stock price underreacts with earning news but overreacts with a series of good or bad earning news.

We consider the case in which θ is unknown. In this case, the representative agent needs to infer θ from historical data. We further assume that the agent uses the model of conservatism (1.7) to infer θ. Note that the likelihood of θ is given (3.6), so given $\mathbf{y}$ and a prior distribution of θ, we can compute the posterior density of θ.

Because the parameter θ is unknown, when evaluating the stock price, the agent is assumed to first evaluate the price with known θ, as in Barberis *et al.* [1998], and then take expectation with respect to θ under the posterior distribution of θ. In other words, the stock price is

$$P_t = \mathbb{E}\left[P_t(\tilde{\theta})\right] = \frac{N_t}{\delta} + y_t\mathbb{E}\left[(p_1(\tilde{\theta}) - p_2(\tilde{\theta})q_t(\tilde{\theta}))\right],$$

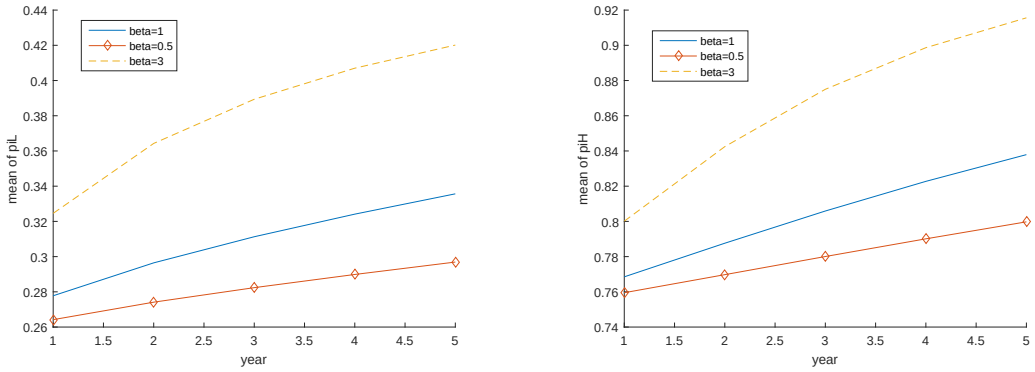
where $\tilde{\theta}$ follows the posterior distribution given $\mathcal{F}_t$.

3.3.2 Numerical Simulation

In the following, we conduct numerical analysis to study the impact of conservatism on asset pricing. We assume that λ_1 and λ_2 are known to be 0.1 and 0.3, respectively, but π_H and π_L are unknown. We assume that the prior distribution of π_L is uniform on $[0, 0.5]$ and the prior distribution of π_H is uniform on $[0.5, 1]$. We consider three return scenarios: (i) consecutive positive returns $y_i = 1$, $i \geq 1$, (ii) consecutive positive returns $y_i = -1$, $i \geq 1$ and (iii) alternating returns $y_{2i-1} = -1$, $y_{2i} = 1$, $i \geq 1$. Then, we calculate the parameters estimated by the agent after observing the first n signals of the return series, $n \geq 1$ and study the resulting return of the stock. We choose three values for β: 0.5, 1, and 3, corresponding to conservatism, Bayesian, and overextrapolation, respectively.

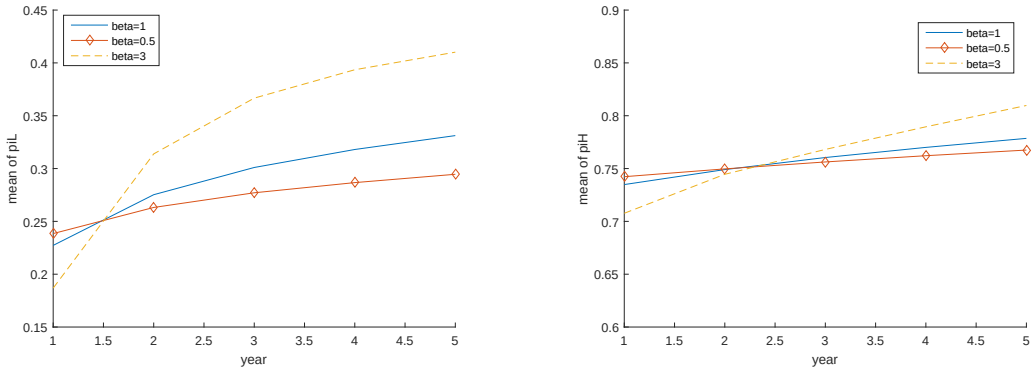
3.3.2.1 Parameter Estimation

We plot the posterior means of π_L and π_H with respect to the number of signals observed in three cases: consecutive positive returns, consecutive negative returns, and alternating returns, as depicted in Figure 3.5, Figure 3.6, and Figure 3.7, respectively.



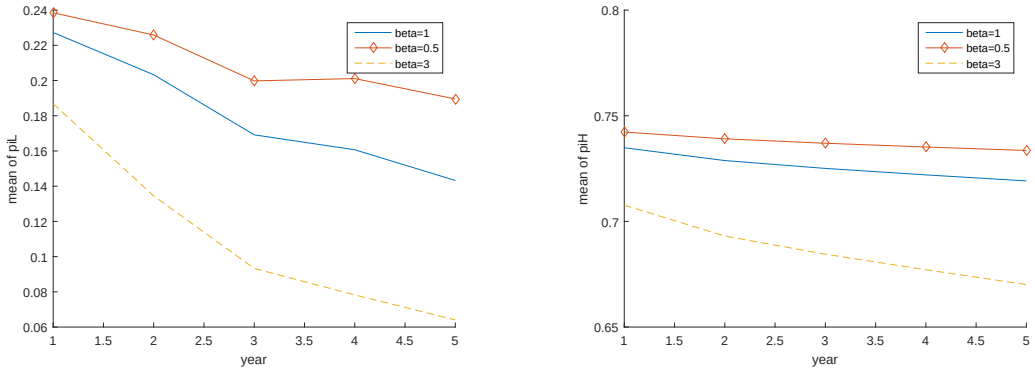
(a) Posterior mean of π_L. (b) Posterior mean of π_H.

Figure 3.5: Posterior means of π_L and π_H with respect to the number of consecutive positive signals. We set λ_1 and λ_2 to be 0.1 and 0.3, respectively, and the prior distributions of π_L and π_H to be uniform distributions on $[0, 0.5]$ and on $[0.5, 1]$, respectively. We choose three values for β: 0.5, 1, and 3, corresponding to conservatism, Bayesian, and overextrapolation, respectively.



(a) Posterior mean of π_L. (b) Posterior mean of π_H.

Figure 3.6: Posterior means of π_L and π_H with respect to the number of consecutive negative signals. We set λ_1 and λ_2 to be 0.1 and 0.3, respectively, and the prior distributions of π_L and π_H to be uniform distributions on $[0, 0.5]$ and on $[0.5, 1]$, respectively. We choose three values for β: 0.5, 1, and 3, corresponding to conservatism, Bayesian, and overextrapolation, respectively.



(a) Posterior mean of π_L.

(b) Posterior mean of π_H.

Figure 3.7: Posterior means of π_L and π_H with respect to the number of alternating signals. We set λ_1 and λ_2 to be 0.1 and 0.3, respectively, and the prior distributions of π_L and π_H to be uniform distributions on $[0, 0.5]$ and on $[0.5, 1]$, respectively. We choose three values for β: 0.5, 1, and 3, corresponding to conservatism, Bayesian, and overextrapolation, respectively.

3.3.2.2 Inference of Regimes

Next, we compute the agent's estimate of the probability that the current market is under regime 1 after observing n signals, i.e., compute

$$q_n = \mathbb{E}_n[q_n(\tilde{\theta})],$$

where $\tilde{\theta}$ follows the posterior distribution of the unknown parameter after observing n signals. Table 3.3, Table 3.4, and Table 3.5 show this probability when observing consecutive positive signals, consecutive negative signals, and alternating signals, respectively.

	$\beta = 0.5$	$\beta = 1$	$\beta = 3$
$n = 1$	0.3229	0.3333	0.3664
$n = 2$	0.2485	0.2655	0.3152
$n = 3$	0.2219	0.2423	0.2970
$n = 4$	0.2142	0.2367	0.2922
$n = 5$	0.2136	0.2379	0.2928

Table 3.3: Agent's estimate of the probability, q_n, that the market is under Regime 1 after observing n consecutive positive signals. We set λ_1 and λ_2 to be 0.1 and 0.3, respectively, and the prior distributions of π_L and π_H to be uniform distributions on $[0, 0.5]$ and on $[0.5, 1]$, respectively. We choose three values for β: 0.5, 1, and 3, corresponding to conservatism, Bayesian, and overextrapolation, respectively.

	$\beta = 0.5$	$\beta = 1$	$\beta = 3$
$n = 1$	0.8206	0.8182	0.8100
$n = 2$	0.5247	0.5391	0.5828
$n = 3$	0.3492	0.3790	0.4620
$n = 4$	0.2706	0.3057	0.3975
$n = 5$	0.2403	0.2762	0.3616

Table 3.4: Agent's estimate of the probability, q_n, that the market is under Regime 1 after observing n consecutive negative signals. We set λ_1 and λ_2 to be 0.1 and 0.3, respectively, and the prior distributions of π_L and π_H to be uniform distributions on $[0, 0.5]$ and on $[0.5, 1]$, respectively. We choose three values for β: 0.5, 1, and 3, corresponding to conservatism, Bayesian, and overextrapolation, respectively.

	$\beta = 0.5$	$\beta = 1$	$\beta = 3$
$n = 1$	0.8206	0.8182	0.8100
$n = 2$	0.9182	0.9168	0.9118
$n = 3$	0.9438	0.9434	0.9404
$n = 4$	0.9505	0.9506	0.9479
$n = 5$	0.9525	0.9530	0.9496

Table 3.5: Agent's estimate of the probability, q_n, that the market is under Regime 1 after observing n alternating negative signals. We set λ_1 and λ_2 to be 0.1 and 0.3, respectively, and the prior distributions of π_L and π_H to be uniform distributions on $[0, 0.5]$ and on $[0.5, 1]$, respectively. We choose three values for β: 0.5, 1, and 3, corresponding to conservatism, Bayesian, and overextrapolation, respectively.

3.3.2.3 Overreaction and Underreaction

Next, we conduct a simulation experiment similar to the one in Barberis *et al.* [1998] to study the impact of conservatism on asset pricing. More precisely, we choose an initial level of earnings N_1 and use the true random walk model to simulate 2000 independent earnings sequences, each one starting at N_1. We think of a period in our model as corresponding to a year and choose the absolute value of the earnings change y to be small relative to the initial earnings level N_1 so as to avoid generating negative earnings.

Now, for each n-year period in the sample, where n ranges from one to seven, we form two portfolios. One portfolio consists of all firms with positive earnings changes in each of n years, and the other of all the firms with negative earnings changes in each of the n years. We then compare the difference in average gains for each one of the stocks in these two portfolios, that is, we compute $(\bar{P}_{t+1}^{n+} - \bar{P}_t^{n+}) - (\bar{P}_{t+1}^{n-} - \bar{P}_t^{n-})$. The results are shown in Table 3.6. We can observe that fixed a value of β, the

difference is positive when n is small and negative when n is large, showing that the stock prices underreact with respect to earning news in short terms but overreact with respect to earning news in long terms. In addition, conservatism (i.e., $\beta < 1$) lowers and overextrapolation (i.e., $\beta > 1$) increases the magnitude of the short-term underreaction and long-term overreaction.

Table 3.6: Difference in Returns of a portfolio of firms with n consecutive positive earnings and a portfolio of firms with n consecutive negative earnings. β takes three values: 0.5, 1, 3 and 10, corresponding to conservatism, Bayesian, and overextrapolation (for $\beta = 3, 10$), respectively.

	$\beta = 0.5$	$\beta = 1$	$\beta = 3$	$\beta = 10$
n				
1	0.4149	0.4474	0.5433	0.6312
2	0.1322	0.1217	0.0727	-0.0518
3	-0.0441	-0.0789	-0.1964	-0.3571
4	-0.1431	-0.2037	-0.3759	-0.5762

Next we study the overreaction and underreaction through numerical simulation. We pick the annual stock return from 1974 to 2016 as well as the yearly net income statistics for each of the firms. We exclude firms whose stock prices is less than 5 dollars and market capitalization less than 10 million. We pick up firms with n consecutive earning increases/decreases in the past n years, where n ranges from 1 to 4, and look at their stock return in the coming year. The following is the result of the historical simulation exercise. We report the difference in average returns in the two samples of stocks, that is, samples with n consecutive positive earnings and samples with consecutive negative earnings.

Table 3.7: Difference in returns of a sample of firms with n consecutive positive earnings and a sample of firms with n consecutive negative earnings. The sample comes from yearly stock returns in the U.S. stock market from 1980 to 2015

	Return Difference	Sample Size
n		
1	0.001673264	23586
2	-0.015758347	8501
3	-0.019152459	2779
4	-0.00515	798

As we can see, the return difference turns negative when $n = 2$, consistent with $\beta = 10$ case in Table 3.6. People are overly extrapolating the signals contained in earnings when making an inference. This can be explained as follows: when yearly earnings statement comes out, people read a large amount of research reports and news coverage, thus overestimating the information contained in annual reports and pushing the price to more extreme levels than the Bayesian case following consecutive signals of the same sign, making future returns reversed.

3.4 Conclusion

In this chapter, we considered two asset pricing model with non-Bayesian inference. In the first model, we generalize the classical consumption-based asset pricing model by assuming the distribution of the dividend growth rate to be unknown. We assume that the representative agent is conservative in inferring the distribution from historical data, and we use (1.7) to model conservatism. We find that when the representative agent becomes more conservative, the risk-free return and the price-dividend

ratio become less sensitive to the number of good signals in the historical dividend data, and the risk premium becomes lower.

In the second model, we generalize the regime-switching asset pricing model in Barberis *et al.* [1998] by assuming some of the model parameters to be unknown. Again, we assume that the representative agent is conservatism in inferring the parameters. We find that this model can generate short-term momentum and long-term reversal, similar to the model in Barberis *et al.* [1998] where the model parameters are known. We also find that the more conservative the agent is, the less profound the effect of short-term momentum and long-term reversal. We finally compare the numerical result with the data on U.S. stock returns and find that people tend to overextrapolate the information contained in annual reports when making an inference.

Bibliography

Maya Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233, 1980.

Maya Bar-Hillel. The base rate fallacy controversy. *Decision making under uncertainty*, pages 39–61, 1983.

Nicholas Barberis, Andrei Shleifer, and Robert Vishny. A model of investor sentiment. *Journal of Financial Economics*, 49(3):307 – 343, 1998.

Lee Roy Beach, James A Wise, and Scott Barclay. Sample proportions and subjective probability revisions. *Organizational Behavior and Human Performance*, 5(2):183–190, 1970.

Lee R Beach. Probability magnitudes and conservative revision of subjective probabilities. *Journal of Experimental Psychology*, 77(1):57–63, 1968.

Daniel J. Benjamin, Matthew Rabin, and Collin Raymond. A model of non-belief in the law of large numbers. *Journal of the European Economics Association*, 14(2):515–544, 2016.

Colin F Camerer. Do biases in probability judgment matter in markets? experimental evidence. *American Economic Review*, pages 981–997, 1987.

James O Chinnis Jr and Cameron R Peterson. Inference about a nonstationary process. *Journal of Experimental Psychology*, 77(4):620–625, 1968.

Chetan Dave and Katherine W Wolfe. On confirmation bias and deviations from bayesian updating. Working Paper, 2003.

JH De Swart. Conservatism as a function of bag composition. *Acta Psychologica*, 36(3):197–206, 1972.

JH De Swart. Effects of diagnosticity and prior odds on conservatism in a bookbag-and-pokerchip situation. *Acta Psychologica*, 36(1):16–31, 1972.

Michael L Donnell and Wesley M Du Charme. The effect of bayesian feedback on learning in an odds estimation task. *Organizational Behavior and Human Performance*, 14(3):305–313, 1975.

Ward Edwards. Conservatism in human information processing. In Benjamin Kleinmuntz, editor, *Formal representation of human judgment*, pages 17–52. Wiley, New York, 1968.

Larry G Epstein and Michel Le Breton. Dynamically consistent beliefs must be bayesian. *Journal of Economic Theory*, 61(1):1–22, 1993.

Larry G Epstein, Jawwad Noor, and Alvaro Sandroni. Non-bayesian updating: a theoretical framework. *Theoretical Economics*, 3:193–229, 2008.

Larry G Epstein, Jawwad Noor, and Alvaro Sandroni. Non-Bayesian learning. *B.E. Journal of Theoretical Economics*, 10(1):1–20, 2010.

Larry G Epstein. An axiomatic model of non-bayesian updating. *Review of Economic Studies*, 73(2):413–436, 2006.

Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, Berlin, 2nd edition, 2004.

Nicola Gennaioli and Andrei Shleifer. What comes to mind. *Quarterly Journal of Economics*, 125(4):1399–1433, 2010.

Jayanta K Ghosh and RV Ramamoorthi. *Bayesian nonparametrics*. Springer, New York, 2003.

David M. Grether. Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics*, 95(3):537–557, 1980.

Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3):411–435, 1992.

Carlo Kraemer and Martin Weber. How do people take into account weight, strength and quality of segregated vs. aggregated data? experimental evidence. *Journal of Risk and Uncertainty*, 29(2):113–142, 2004.

Don Lyon and Paul Slovic. Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40(4):287–298, 1976.

DF Marks and JK Clarkson. An explanation of conservatism in the bookbag-and-pokerchips situation. *Acta Psychologica*, 36(2):145–160, 1972.

Sendhil Mullainathan. Thinking through categories. available at https://www.chicagobooth.edu/research/workshops/theoryoforg/archive/PDF /mullainathan.pdf, 2002.

Mark W Nelson, Robert Bloomfield, Jeffrey W Hales, and Robert Libby. The effect of information strength and weight on behavior in financial markets. *Organizational Behavior and Human Decision Processes*, 86(2):168–196, 2001.

Pietro Ortoleva. Modeling the change of paradigm: Non-bayesian reactions to unexpected news. *American Economic Review*, 102(6):2410–2436, 2012.

Cameron R Peterson and Alan J Miller. Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, 70(1):117–121, 1965.

Cameron R Peterson and Richard G Swensson. Intuitive statistical inferences about diffuse hypotheses. *Organizational Behavior and Human Performance*, 3(1):1–11, 1968.

Cameron R Peterson, Robert J Schneider, and Alan J Miller. Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, 69(5):522–527, 1965.

Lawrence D Phillips and Ward Edwards. Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3):346–354, 1966.

Matthew Rabin and Joel L. Schrag. First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, 114(1):37–82, 1999.

Matthew Rabin and Dimitri Vayanos. The gambler's and hot-hand fallacies: Theory and applications. *Review of Economic Studies*, 77(2):730–778, 2010.

Matthew Rabin. Inference by believers in the law of small numbers. *Quarterly Journal of Economics*, 117(3):775–816, 2002.

AF Sanders. Choice among bets and revision of opinion. *Acta Psychologica*, 28:76–83, 1968.

Suzanne Shu and George Wu. Belief bracketing: Can partitioning information change consumer judgments? Working Paper, Oct. 2003.

Karl Halvor Teigen. Subjective sampling distributions and the additivity of estimates. *Scandinavian Journal of Psychology*, 15(1):50–55, 1974.

Amos Tversky and Daniel Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, Sep. 1974.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20(4):595–601, 1949.

Gloria Wheeler and Lee Roy Beach. Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance*, 3(1):36–46, 1968.

Appendix A

Proofs

A.1 Proofs in Chapter 1

Proof of Proposition 1 The sufficiency is trivial, so we only prove the necessity. We first prove that there exists $C_m > 0$ such that $g_m(z) = C_m g'_m(z)$ for all z in the effective domain of g_m and g'_m.

We consider the case in which $\Theta/\mathcal{A}_\nu \neq \emptyset$ first. In this case, the effective domain of g_m and g'_m is $[0, +\infty)$. Thus, we only need to show that for any $z_1 > z_2 > 0$, $g_m(z_1)/g'_m(z_1) = g_m(z_2)/g'_m(z_2)$.

Fix any $\mathbf{x} \in \mathbb{X}^\infty$. Because $q_{m,n}$ and $q'_{m,n}$ satisfy Assumption 1, there exists $\theta_0 \in \Theta/\mathcal{A}_\nu$ and a neighbourhood of θ_0, denoted as $\mathcal{N}$, such that (i) $\mathcal{N} \subset \Theta/\mathcal{A}_\nu$, (ii) $\nu(\mathcal{N}) > 0$, and (iii) $q_{m,n}(\theta|\mathbf{x}) > 0, q'_{m,n}(\theta|\mathbf{x}) > 0$ for all $\theta \in \mathcal{N}$. Because ν is σ-finite, we can find $\mathcal{N}_0 \subset \mathcal{N}$ such that $0 < \nu(\mathcal{N}_0) < +\infty$.

We then choose two different elements in $\mathcal{N}_0$ and construct their neighbourhoods $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively, such that $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$, $\mathcal{N}_i \subset \mathcal{N}_0$, and $0 < \nu(\mathcal{N}_i) < 1/(2z_1)$, $i = 1, 2$. Such a construction is possible because ν has no atom on $\Theta/\mathcal{A}_\mu$; see for instance Föllmer and Schied [2004] Proposition A.27. Define

$$\pi(\theta) := z_1 \mathbf{1}_{\mathcal{N}_1} + z_2 \mathbf{1}_{\mathcal{N}_2} + \frac{1 - z_1\nu(\mathcal{N}_1) - z_2\nu(\mathcal{N}_2)}{\nu(\mathcal{N}_0/(\mathcal{N}_1 \cup \mathcal{N}_2))} \mathbf{1}_{\mathcal{N}_0/(\mathcal{N}_1 \cup \mathcal{N}_2)},$$
$$\tilde{\pi}(\theta) := z_1 \mathbf{1}_{\mathcal{N}_1 \cup \mathcal{N}_2} + \frac{1 - z_1\nu(\mathcal{N}_1 \cup \mathcal{N}_2)}{\nu(\mathcal{N}_0/(\mathcal{N}_1 \cup \mathcal{N}_2))} \mathbf{1}_{\mathcal{N}_0/(\mathcal{N}_1 \cup \mathcal{N}_2)}.$$

Then, π and $\tilde{\pi}$ are two probability densities on Θ.

Denote the posterior densities resulted from (1.2) with the prior density π and the pairs of pseudo-likelihood and distortion $(q_{m,n}(\theta|\mathbf{x}), g_m)$ and $(q'_{m,n}(\theta|\mathbf{x}), g'_m)$ as $\pi_{m,n}$

and $\pi'_{m,n}$, respectively. Define $\tilde{\pi}_{m,n}$ and $\tilde{\pi}'_{m,n}$ similarly. Then, we have $\pi_{m,n}(\theta|\mathbf{x}) = \pi'_{m,n}(\theta|\mathbf{x})$ and $\tilde{\pi}_{m,n}(\theta|\mathbf{x}) = \tilde{\pi}'_{m,n}(\theta|\mathbf{x})$ for ν-almost surely $\theta \in \Theta$. Because $\nu(\mathcal{N}_1) > 0, \nu(\mathcal{N}_2) > 0$, there exist $\theta_1 \in \mathcal{N}_1$ and $\theta_2 \in \mathcal{N}_2$ such that $\pi_{m,n}(\theta_i|\mathbf{x}) = \pi'_{m,n}(\theta_i|\mathbf{x})$ and $\tilde{\pi}_{m,n}(\theta_i|\mathbf{x}) = \tilde{\pi}'_{m,n}(\theta_i|\mathbf{x})$, $i = 1, 2$. In consequence, we have $\pi_{m,n}(\theta_1|\mathbf{x})/\pi_{m,n}(\theta_2|\mathbf{x}) = \pi'_{m,n}(\theta_1|\mathbf{x})/\pi'_{m,n}(\theta_2|\mathbf{x})$, which yields

$$\frac{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi(\theta_1))}{q_{m,n}(\theta_2|\mathbf{x})g_m(\pi(\theta_2))} = \frac{q'_{m,n}(\theta_1|\mathbf{x})g'_m(\pi(\theta_1))}{q'_{m,n}(\theta_2|\mathbf{x})g'_m(\pi(\theta_2))}$$

Because $\pi(\theta_1) = z_1$ and $\pi(\theta_2) = z_2$, we conclude from the above equation that

$$\frac{q_{m,n}(\theta_1|\mathbf{x})q'_{m,n}(\theta_2|\mathbf{x})}{q_{m,n}(\theta_2|\mathbf{x})q'_{m,n}(\theta_1|\mathbf{x})} = \frac{g_m(z_2)g'_m(z_1)}{g'_m(z_2)g_m(z_1)}.$$

Similarly, we can conclude from $\tilde{\pi}_{m,n}(\theta_i|\mathbf{x}) = \tilde{\pi}'_{m,n}(\theta_i|\mathbf{x})$, $i = 1, 2$ that

$$\frac{q_{m,n}(\theta_1|\mathbf{x})q'_{m,n}(\theta_2|\mathbf{x})}{q_{m,n}(\theta_2|\mathbf{x})q'_{m,n}(\theta_1|\mathbf{x})} = 1.$$

As a result, $g_m(z_1)/g'_m(z_1) = g_m(z_2)/g'_m(z_2)$.

Next, we consider the case in which $\mathcal{A}_\nu = \Theta$. Fix arbitrary $\theta_1, \theta_2, \theta_3 \in \mathcal{A}_\mu$ that are mutually different. Without loss of generality, we assume that $\nu(\{\theta_1\}) \geq \nu(\{\theta_2\}) \geq \nu(\{\theta_3\})$. We first show that $g_m(z)/g'_m(z)$ is constant for $z \in (0, 1/\nu(\{\theta_2, \theta_3\})]$. Fix any $z_1 < z_2$ in this interval. Define

$$\pi(\theta) := z_1\mathbf{1}_{\{\theta_2\}} + z_2\mathbf{1}_{\{\theta_3\}} + \frac{1 - z_1\nu(\{\theta_2\}) - z_2\nu(\{\theta_3\})}{\nu(\{\theta_1\})}\mathbf{1}_{\{\theta_1\}},$$
$$\tilde{\pi}(\theta) := z_1\mathbf{1}_{\{\theta_2,\theta_3\}} + \frac{1 - z_1\nu(\{\theta_2, \theta_3\})}{\nu(\{\theta_1\})}\mathbf{1}_{\{\theta_1\}}.$$

Note that $\pi(\cdot)$ and $\tilde{\pi}(\cdot)$ are well-defined probability densities because $0 < z_1 < z_2 \leq 1/\nu(\{\theta_2, \theta_3\})$. The same argument as in the case $\Theta/\mathcal{A}_\nu \neq \emptyset$ yields $g_m(z_1)/g'_m(z_1) = g_m(z_2)/g'_m(z_2)$, so $g_m(z)/g'_m(z)$ is constant for $z \in (0, 1/\nu(\{\theta_2, \theta_3\})]$.

Next, for any $z_2 < 1/\nu(\{\theta_3\})$, we can always find $z_1 \in (0, 1/\nu(\{\theta_2, \theta_3\})]$ such that both $\pi(\cdot)$ and $\tilde{\pi}(\cdot)$ defined above are probability densities. Repeating the previous argument, we conclude that $g_m(z_2)/g'_m(z_2) = g_m(z_1)/g'_m(z_1)$, so $g_m(z)/g'_m(z)$ is constant in $(0, 1/\nu(\{\theta_3\})]$. Because θ_i's are arbitrarily chosen, $g_m(z)/g'_m(z)$ is constant in the effective domain of g_m and g'_m.

Finally, because ν is σ-finite, we can find a probability density π such that $\pi(\theta) >$

$0, \theta \in \Theta$. Because the posterior densities obtained from this prior and the two pairs $(q_{m,n}(\theta|\mathbf{x}), g_m)$ and $(q'_{m,n}(\theta|\mathbf{x}), g'_m)$ are the same, we conclude, for any $\mathbf{x} \in \mathbb{X}^\infty$,

$$\frac{q_{m,n}(\theta|\mathbf{x})g_m(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta})} = \frac{q'_{m,n}(\theta|\mathbf{x})g'_m(\pi(\theta))}{\int_\Theta q'_{m,n}(\tilde{\theta}|\mathbf{x})g'_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}$$

for ν-almost surely $\theta \in \Theta$, which then holds for any $\theta \in \Theta$ due to the continuity of $q_{m,n}$ and $q'_{m,n}$ in θ. Because we already showed that $g_m = C_m g'_m$ for some $C_m > 0$, we conclude that

$$q_{m,n}(\theta|\mathbf{x}) = \frac{C_m \int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}{\int_\Theta q'_{m,n}(\tilde{\theta}|\mathbf{x})g'_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta})} q'_{m,n}(\theta|\mathbf{x}) =: C_{m,n}(\mathbf{x})q'_{m,n}(\theta|\mathbf{x})$$

for any $\theta \in \Theta$. □

Proof of Theorem 1 We first prove the sufficiency. For any $\mathbf{x} \in \mathbb{X}^\infty$ and prior density π, denote $\pi_{0,m} := \mathcal{I}^C_{0,m}(\mathbf{x}, \pi)$, $\pi_{m,n} := \mathcal{I}^C_{m,n}(\mathbf{x}, \pi_{0,m})$, and $\pi_{0,m+n} := \mathcal{I}^C_{0,m+n}(\mathbf{x}, \pi)$. On the one hand,

$$\begin{aligned}\pi_{m,n}(\theta|\mathbf{x}) &= \frac{q_{m,n}(\theta|\mathbf{x})g_m(\pi_{0,m}(\theta|\mathbf{x}))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})g_m(\pi_{0,m}(\tilde{\theta}|\mathbf{x}))\nu(d\tilde{\theta})} = \frac{q_{m,n}(\theta|\mathbf{x})\pi_{0,m}(\theta|\mathbf{x})}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})\pi_{0,m}(\tilde{\theta}|\mathbf{x})\nu(d\tilde{\theta})} \\ &= \frac{q_{m,n}(\theta|\mathbf{x})q_{0,m}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})q_{0,m}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})},\end{aligned}$$

where the second equality is the case because $g_m(z) = K_m z$ and the third equality follows from the definition of $\pi_{0,m}$. On the other hand,

$$\pi_{0,m+n}(\theta|\mathbf{x}) = \frac{q_{0,m+n}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{0,m+n}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})} = \frac{q_{m,n}(\theta|\mathbf{x})q_{0,m}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})q_{0,m}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})},$$

where the second equality follows from (1.4). Therefore, $\pi_{m,n}(\theta|\mathbf{x}) = \pi_{0,m+n}(\theta|\mathbf{x})$ for all $\theta \in \Theta$.

Next, we prove the necessity. We first prove condition (i). We consider first the case in which $\Theta/\mathcal{A}_\nu \neq \emptyset$. We only need to show that for each fixed $m \geq 1$, $g_m(z_1)/z_1 = g_m(z_2)/z_2$ for any $z_1, z_2 > 0$.

Fix any $z_1 > z_2 > 0$ and any $\mathbf{x} \in \mathbb{X}^\infty$. Choose $\theta_0 \in \Theta/\mathcal{A}_\nu$ such that $q_{0,m+n}(\theta_0|\mathbf{x}) > 0$, $q_{0,m}(\theta_0|\mathbf{x}) > 0$, and $q_{m,n}(\theta_0|\mathbf{x}) > 0$. Choose a neighbourhood $\mathcal{N}$ of θ_0 such that $\mathcal{N} \subset \Theta/\mathcal{A}_\nu$ and $q_{0,m+n}(\theta|\mathbf{x}), q_{0,m}(\theta|\mathbf{x}), q_{m,n}(\theta|\mathbf{x}) \in [\frac{1}{C}, C], \forall \theta \in \mathcal{N}$ for some $C > 0$. Because ν is σ-finite with support Θ, we can find $\mathcal{N}_0 \subset \mathcal{N}$ such that $0 < \nu(\mathcal{N}_0) < +\infty$.

We first show that, for any bounded probability density p that is null outside $\mathcal{N}_0$, we can find a prior density π such that

$$p(\theta) = \frac{q_{0,m}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{0,m}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}, \quad \theta \in \Theta. \tag{A.1}$$

To this end, for each $\delta \geq 0$, define $h_\delta(\theta) := g_0^{-1}\big(\delta p(\theta)/q_{0,m}(\theta|\mathbf{x})\big)\mathbf{1}_{\mathcal{N}_0}(\theta)$, $\theta \in \Theta$, where $g_0^{-1}(y) := +\infty$ for $y \geq \lim_{z\to+\infty} g_0(z)$. Because $1/C \leq q_{0,m}(\theta|\mathbf{x}) \leq C$ for all $\theta \in \mathcal{N}_0$, h_z is well-defined. Furthermore, because $\nu(\mathcal{N}_0) < +\infty$ and p is bounded, the monotone convergence theorem leads to the continuity of $\lambda(\delta) := \int_\Theta h_\delta(\theta)\nu(d\theta)$. Because $g_0(0) = 0$, we have $\lambda(0) = 0$. On the other hand, because $\nu(\mathcal{N}_0) > 0$, we have $\lim_{\delta\to+\infty}\lambda(\delta) = +\infty$. As a result, there exists $\delta_0 > 0$ such that $\lambda(\delta_0) = 1$. Define $\pi(\theta) = h_{\delta_0}(\theta), \theta \in \Theta$, then π is a probability density. Furthermore, because $g_m(0) = 0$ and $p(\theta) \propto q_{0,m}(\theta|\mathbf{x})g_0(\pi(\theta))$, (A.1) holds.

Now, choose two different elements in $\mathcal{N}_0$ and construct their neighbourhoods $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively, such that $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$, $\mathcal{N}_i \subset \mathcal{N}_0$, and $0 < \nu(\mathcal{N}_i) < 1/(2z_1)$, $i = 1, 2$. Such a construction is possible because ν has no atom on $\Theta/\mathcal{A}_\mu$; see for instance Föllmer and Schied [2004]. Define

$$p(\theta) := z_1\mathbf{1}_{\mathcal{N}_1} + z_2\mathbf{1}_{\mathcal{N}_2} + \frac{1 - z_1\nu(\mathcal{N}_1) - z_2\nu(\mathcal{N}_2)}{\nu(\mathcal{N}_0/(\mathcal{N}_1 \cup \mathcal{N}_2))}\mathbf{1}_{\mathcal{N}_0/(\mathcal{N}_1\cup\mathcal{N}_2)}.$$

It is clear that p is a bounded probability density and is null outside $\mathcal{N}_0$. Therefore, we can construct a prior density π such that (A.1) holds. Similarly, for density

$$\tilde{p}(\theta) := z_1\mathbf{1}_{\mathcal{N}_1\cup\mathcal{N}_2} + \frac{1 - z_1\nu(\mathcal{N}_1 \cup \mathcal{N}_2)}{\nu(\mathcal{N}_0/(\mathcal{N}_1 \cup \mathcal{N}_2))}\mathbf{1}_{\mathcal{N}_0/(\mathcal{N}_1\cup\mathcal{N}_2)},$$

we can construct a prior density $\tilde{\pi}$ such that $\tilde{p}$ and $\tilde{\pi}$ are related through (A.1).

Next, denote

$$\begin{aligned}
&\pi_{0,m} := \mathcal{I}^C_{0,m}(\mathbf{x}, \pi), \pi_{m,n} := \mathcal{I}^C_{m,n}(\mathbf{x}, \pi_{0,m}), \pi_{0,m+n} := \mathcal{I}^C_{0,m+n}(\mathbf{x}, \pi),\\
&\tilde{\pi}_{0,m} := \mathcal{I}^C_{0,m}(\mathbf{x}, \tilde{\pi}), \tilde{\pi}_{m,n} := \mathcal{I}^C_{m,n}(\mathbf{x}, \tilde{\pi}_{0,m}), \tilde{\pi}_{0,m+n} := \mathcal{I}^C_{0,m+n}(\mathbf{x}, \tilde{\pi}).
\end{aligned}$$

Then, $\pi_{0,m} = p$ and $\tilde{\pi}_{0,m} = \tilde{p}$. Due to processing consistency, $\pi_{0,m+n}(\theta|\mathbf{x}) = \pi_{m,n}(\theta|\mathbf{x})$ and $\tilde{\pi}_{0,m+n}(\theta|\mathbf{x}) = \tilde{\pi}_{m,n}(\theta|\mathbf{x})$ for ν-almost surely $\theta \in \Theta$. Because $\nu(\mathcal{N}_1) > 0, \nu(\mathcal{N}_2) > 0$, there exist $\theta_1 \in \mathcal{N}_1$ and $\theta_2 \in \mathcal{N}_2$ such that $\pi_{0,m+n}(\theta_i|\mathbf{x}) = \pi_{m,n}(\theta_i|\mathbf{x})$ and $\tilde{\pi}_{0,m+n}(\theta_i|\mathbf{x}) = \tilde{\pi}_{m,n}(\theta_i|\mathbf{x})$, $i = 1, 2$. Furthermore, because $\tilde{\pi}_{0,m}(\theta_i|\mathbf{x}) = \tilde{p}(\theta_i) > 0$ and $\pi_{0,m}(\theta_i|\mathbf{x}) = p(\theta_i) > 0$, $i = 1, 2$, and $q_{0,m+n}(\theta|\mathbf{x}) > 0, q_{0,m}(\theta|\mathbf{x}) > 0, q_{m,n}(\theta|\mathbf{x}) > 0$ for any $\theta \in \mathcal{N}_0$,

we conclude that $\pi_{0,m+n}(\theta_i|\mathbf{x}) = \pi_{m,n}(\theta_i|\mathbf{x}) > 0$ and $\tilde{\pi}_{0,m+n}(\theta_i|\mathbf{x}) = \tilde{\pi}_{m,n}(\theta_i|\mathbf{x}) > 0$, $i = 1, 2$. Therefore, we have

$$\frac{\pi_{0,m+n}(\theta_1|\mathbf{x})}{\pi_{0,m+n}(\theta_2|\mathbf{x})} = \frac{\pi_{m,n}(\theta_1|\mathbf{x})}{\pi_{m,n}(\theta_2|\mathbf{x})}, \quad \frac{\tilde{\pi}_{0,m+n}(\theta_1|\mathbf{x})}{\tilde{\pi}_{0,m+n}(\theta_2|\mathbf{x})} = \frac{\tilde{\pi}_{m,n}(\theta_1|\mathbf{x})}{\tilde{\pi}_{m,n}(\theta_2|\mathbf{x})}$$

which leads to

$$\begin{aligned}
\frac{q_{0,m+n}(\theta_1|\mathbf{x})\big(p(\theta_1)/q_{0,m}(\theta_1|\mathbf{x})\big)}{q_{0,m+n}(\theta_2|\mathbf{x})\big(p(\theta_2)/q_{0,m}(\theta_2|\mathbf{x})\big)} &= \frac{q_{m,n}(\theta_1|\mathbf{x})g_m(p(\theta_1))}{q_{m,n}(\theta_2|\mathbf{x})g_m(p(\theta_2))}, \\
\frac{q_{0,m+n}(\theta_1|\mathbf{x})\big(\tilde{p}(\theta_1)/q_{0,m}(\theta_1|\mathbf{x})\big)}{q_{0,m+n}(\theta_2|\mathbf{x})\big(\tilde{p}(\theta_2)/q_{0,m}(\theta_2|\mathbf{x})\big)} &= \frac{q_{m,n}(\theta_1|\mathbf{x})g_m(\tilde{p}(\theta_1))}{q_{m,n}(\theta_2|\mathbf{x})g_m(\tilde{p}(\theta_2))}.
\end{aligned}$$

Because $p(\theta_1) = z_1$, $p(\theta_2) = z_2$, and $\tilde{p}(\theta_1) = \tilde{p}(\theta_2) = z_1$, we conclude that

$$\begin{aligned}
\frac{q_{0,m+n}(\theta_1|\mathbf{x})/\big(q_{0,m}(\theta_1|\mathbf{x})q_{m,n}(\theta_1|\mathbf{x})\big)}{q_{0,m+n}(\theta_2|\mathbf{x})/\big(q_{0,m}(\theta_2|\mathbf{x})q_{m,n}(\theta_2|\mathbf{x})\big)} &= \frac{g_m(p(\theta_1))/p(\theta_1)}{g_m(p(\theta_2))/p(\theta_2)} = \frac{g_m(z_1)/z_1}{g_m(z_2)/z_2}, \\
\frac{q_{0,m+n}(\theta_1|\mathbf{x})/\big(q_{0,m}(\theta_1|\mathbf{x})q_{m,n}(\theta_1|\mathbf{x})\big)}{q_{0,m+n}(\theta_2|\mathbf{x})/\big(q_{0,m}(\theta_2|\mathbf{x})q_{m,n}(\theta_2|\mathbf{x})\big)} &= \frac{g_m(\tilde{p}(\theta_1))/\tilde{p}(\theta_1)}{g_m(\tilde{p}(\theta_2))/\tilde{p}(\theta_2)} = \frac{g_m(z_1)/z_1}{g_m(z_1)/z_1} = 1.
\end{aligned}$$

As a result, $g_m(z_1)/z_1 = g_m(z_2)/z_2$.

We next consider the case $\mathcal{A}_\nu = \Theta$. Fix arbitrary $\theta_1, \theta_2, \theta_3 \in \mathcal{A}_\mu$ that are mutually different. Without loss of generality, we assume that $\nu(\{\theta_1\}) \geq \nu(\{\theta_2\}) \geq \nu(\{\theta_3\})$. We first show that $g_m(z)/z$ is constant for $z \in (0, 1/\nu(\{\theta_2, \theta_3\})]$. Fix any $z_1 < z_2$ in this interval. As in the case $\Theta/\mathcal{A}_\nu \neq \emptyset$, we can show that, for any bounded probability density $p(\cdot)$ with support within $\{\theta_i | i = 1, 2, 3\}$, we can find a prior density π such that (A.1) holds. Construct prior densities π and $\tilde{\pi}$ to generate densities p and $\tilde{p}$, respectively, where

$$\begin{aligned}
p(\theta) &:= z_1\mathbf{1}_{\{\theta_2\}} + z_2\mathbf{1}_{\{\theta_3\}} + \frac{1 - z_1\nu(\{\theta_2\}) - z_2\nu(\{\theta_3\})}{\nu(\{\theta_1\})}\mathbf{1}_{\{\theta_1\}}, \\
\tilde{p}(\theta) &:= z_1\mathbf{1}_{\{\theta_2,\theta_3\}} + \frac{1 - z_1\nu(\{\theta_2, \theta_3\})}{\nu(\{\theta_1\})}\mathbf{1}_{\{\theta_1\}}.
\end{aligned}$$

Note that $p(\cdot)$ and $\tilde{p}(\cdot)$ are well-defined probability densities because $0 < z_1 < z_2 \leq 1/\nu(\{\theta_2, \theta_3\})$. The same argument as in the case $\Theta/\mathcal{A}_\nu \neq \emptyset$ yields $g_m(z_1)/z_1 = g_m(z_2)/z_2$, so $g_m(z)/z$ is constant for $z \in (0, 1/\nu(\{\theta_2, \theta_3\})]$.

Next, for any $z_2 < 1/\nu(\{\theta_3\})$, we can always find $z_1 \in (0, 1/\nu(\{\theta_2, \theta_3\})]$ such that both $p(\cdot)$ and $\tilde{p}(\cdot)$ defined above are probability densities. Repeating the pre-

vious argument, we conclude that $g_m(z_2)/z_2 = g_m(z_1)/z_1$, so $g_m(z)/z$ is constant in $(0, 1/\nu(\{\theta_3\})]$. Because θ_i's are arbitrarily chosen, $g_m(z)/z$ is constant in the effective domain of g_m, i.e., condition (i) holds.

Finally, we prove condition (ii). Again, for any $\mathbf{x} \in \mathbb{X}^\infty$ and prior density π, denote $\pi_{0,m} := \mathcal{I}^C_{0,m}(\mathbf{x}, \pi)$, $\pi_{m,n} := \mathcal{I}^C_{m,n}(\mathbf{x}, \pi_{0,m})$, and $\pi_{0,m+n} := \mathcal{I}^C_{0,m+n}(\mathbf{x}, \pi)$. Because $g_m(z) = K_m z, z \geq 0$ for some $K_m > 0$, we have

$$\pi_{m,n}(\theta|\mathbf{x}) = \frac{q_{m,n}(\theta|\mathbf{x})\pi_{0,m}(\theta|\mathbf{x})}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})\pi_{0,m}(\tilde{\theta}|\mathbf{x})\nu(d\tilde{\theta})} = \frac{q_{m,n}(\theta|\mathbf{x})q_{0,m}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})q_{0,m}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}.$$

On the other hand, we have

$$\pi_{0,m+n}(\theta|\mathbf{x}) = \frac{q_{0,m+n}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{0,m+n}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}.$$

As a result, for ν-a.e. $\theta \in \Theta$, we have

$$\frac{q_{m,n}(\theta|\mathbf{x})q_{0,m}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})q_{0,m}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})} = \frac{q_{0,m+n}(\theta|\mathbf{x})g_0(\pi(\theta))}{\int_\Theta q_{0,m+n}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}.$$

Because ν is σ-finite, we can construct a prior density π such that $\pi(\theta) > 0$, $\forall \theta \in \Theta$. As a result,

$$q_{0,m+n}(\theta|\mathbf{x}) = C_{m,n}(\mathbf{x})q_{m,n}(\theta|\mathbf{x})q_{0,m}(\theta|\mathbf{x}), \quad \nu\text{-a.e. } \theta \in \Theta$$

where

$$C_{m,n}(\mathbf{x}) = \frac{\int_\Theta q_{0,m+n}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})q_{0,m}(\tilde{\theta}|\mathbf{x})g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}.$$

Because $q_{0,m}, q_{0,m+n}$, and $q_{m,n}$ are continuous in θ, we conclude that condition (ii) holds. □

Proof of Proposition 2 Denote $T(x, \theta) := \ln(f(x, \theta)/f(x, \theta_0))$ and $\mu(\theta) := \mathbb{E}_{\theta_0}[T(X_i, \theta)]$. Then, because of (1.8) and the dominated convergence theorem, we conclude that μ is continuous. Furthermore, $\mu(\theta) \leq 0$ for all $\theta \in \Theta$. In addition, because $\theta \to \Pi_\theta$ is a one-to-one mapping, $\mu(\theta) = 0$ if and only if $\theta = \theta_0$.

Fix any open neighbourhood U of θ_0 and define $K := \Theta/U$. Then, K is closed and thus compact. For sufficiently small δ, define $V_\delta := \{\theta|\mu(\theta) \geq -\delta\}$. Because μ is continuous, V_δ is closed and thus compact. Define $A_1 := \inf_{\theta \in V_\delta} \mu(\theta)$ and $A_2 :=$

$\sup_{\theta\in K}\mu(\theta)$. Because μ is continuous, $\mu(\theta_0)=0$, and $\mu(\theta)<0$ for any $\theta\neq\theta_0$, we can choose $\delta>0$ such that $V_\delta\subset U$ and $A_1>A_2$.

For notational simplicity, denote $Z_n(\theta):=\frac{1}{n}\sum_{i=1}^n T(X_i,\theta)$. According to Ghosh and Ramamoorthi [2003, Theorem 1.3.3], $\lim_{n\to+\infty}\sup_{\theta\in\Theta}\left|Z_n(\theta)-\mu(\theta)\right|=0$ for P_{θ_0}-a.s. ω.[1] Therefore, for any $0<\epsilon<(A_1-A_2)/2$ there exists $n_0(\omega),\omega\in\Omega$ such that for P_{θ_0}-a.s. ω, $|Z_n(\theta)-\mu(\theta)|<\epsilon$ for any $n\geq n_0(\omega)$ and $\theta\in V_\delta\cup K$. As a result,

$$\begin{aligned}\int_U \pi_{0,n}(\tilde{\theta}|X_1,X_2,\dots)\nu(d\tilde{\theta}) &= \frac{\int_U\left(\prod_{i=1}^n f(X_i,\tilde{\theta})^\beta\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}{\int_\Theta\left(\prod_{i=1}^n f(X_i,\tilde{\theta})^\beta\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}=\frac{\int_U\exp\left(\beta nZ_n(\tilde{\theta})\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}{\int_\Theta\exp\left(\beta nZ_n(\tilde{\theta})\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}\\ &=1-\frac{\int_K\exp\left(\beta nZ_n(\tilde{\theta})\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}{\int_\Theta\exp\left(\beta nZ_n(\tilde{\theta})\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}\\ &\geq 1-1\Big/\left(\frac{\int_{V_\delta}\exp\left(\beta nZ_n(\tilde{\theta})\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}{\int_K\exp\left(\beta nZ_n(\tilde{\theta})\right)\pi(\tilde{\theta})\nu(d\tilde{\theta})}+1\right)\\ &\geq 1-1\Big/\left(\frac{\exp\left(\beta n(A_1-\epsilon)\right)\int_{V_\delta}\pi(\tilde{\theta})\nu(d\tilde{\theta})}{\exp\left(\beta n(A_2+\epsilon)\right)}+1\right),\end{aligned}$$

which goes to 1 as n goes to infinity. Thus, model (1.7) is statistically consistent at θ_0. □

Proof of Proposition 3 We use the following notations for the pseudo-likelihood: $q_{0,1}(\theta|x_1)$, $q_{0,2}(\theta|x_1,x_2)$, and $q_{1,1}(\theta|x_1,x_2)$ for $x_1,x_2\in\mathbb{X}$.

If the model of NBLLN is processing consistent, by Theorem 1, there exists $C(x_1,x_2)>0$ such that

$$q_{0,2}(\theta|x_1,x_2)=C(x_1,x_2)q_{0,1}(\theta|x_1)q_{1,1}(\theta|x_1,x_2),\quad x_1,x_2\in\mathbb{X},\theta\in\Theta. \tag{A.2}$$

For each θ, denote Z_θ as a random variable living in some probability space and taking values in Θ with probability density $h(\cdot|\theta)$. Then, we have $\mathbb{E}[Z_\theta]=\theta$. Furthermore,

$$q_{0,1}(\theta|x_1)=\mathbb{E}\left[f(x_1,Z_\theta)\right],q_{1,1}(\theta|x_1,x_2)=\mathbb{E}\left[f(x_2,Z_\theta)\right],q_{0,2}(\theta|x_1,x_2)=\mathbb{E}\left[f(x_1,Z_\theta)f(x_2,Z_\theta)\right].$$

[1] In Ghosh and Ramamoorthi [2003, Theorem 1.3.3], it is assumed that X_i's take real values. However, this assumption is not needed in the proof of the theorem.

Set $x_1 = x_2 = 1$ and denote $C_1 := C(1,1)$. In this case, we have

$$q_{0,1}(\theta|1) = \mathbb{E}[Z_\theta] = \theta, q_{1,1}(\theta|1,1) = \mathbb{E}[Z_\theta] = \theta, q_{0,2}(\theta|1,1) = \mathbb{E}[Z_\theta^2].$$

Then, (A.2) leads to $\mathbb{E}[Z_\theta^2] = C_1(\mathbb{E}[Z_\theta])^2 = C_1\theta^2$. Because of Jensen's inequality, we must have $C_1 > 1$.

On the other hand, set $x_1 = 1, x_2 = 0$ and denote $C_2 := C(1,0)$. In this case, we have

$$q_{0,1}(\theta|1) = \mathbb{E}[Z_\theta] = \theta, q_{1,1}(\theta|1,0) = \mathbb{E}[1 - Z_\theta] = 1 - \theta, q_{0,2}(\theta|1,0) = \mathbb{E}[Z_\theta(1 - Z_\theta)] = \theta - \mathbb{E}[Z_\theta^2].$$

Then, from (A.2), we have $\theta - \mathbb{E}[Z_\theta^2] = C_2(\theta - \theta^2)$. Recalling that $\mathbb{E}[Z_\theta^2] = C_1\theta^2$, we conclude

$$(1 - C_2)\theta - (C_1 - C_2)\theta^2 = 0, \quad \forall \theta \in \Theta.$$

This is true if and only if $C_1 = C_2 = 1$, and is contradicted because $C_1 > 1$. Thus, the pseudo-likelihood in the model of NBLLN does not satisfy the product rule (1.4), and, as a result, the model of NBLLN is not processing consistent. □

Proof of Theorem 2 *Part (i):* We first prove the sufficiency. For each $m \geq 0$, define

$$\tilde{\ell}_{m,1}(\theta|\mathbf{x}_{0,m}, x) := \varphi_m(\mathbf{x}_{0,m}, x) q_{m,1}(\theta|\mathbf{x}_{0,m}, x), \quad \mathbf{x}_{0,m} \in \mathbb{X}^m, x \in \mathbb{X}, \theta \in \Theta \tag{A.3}$$

and $\tilde{\ell}_{m,n} := \prod_{i=m+1}^{m+n} \tilde{\ell}_{i-1,1}$. Because of (1.10), for any $m \geq 0$, $\int_{\mathbb{X}} \tilde{\ell}_{m,1}(\theta|\mathbf{x}_{0,m}, \tilde{x})\nu_X(d\tilde{x}) = 1$ for ν-a.e. $\theta \in \Theta$. Therefore, $\{\tilde{\ell}_{m,n}\}_{m\geq 0, n\geq 1}$ is the likelihood of some stochastic model. Denote the corresponding false-Bayesian model as $\{\tilde{\mathcal{I}}^B_{m,n}\}_{m\geq 0,n\geq 1}$. Because the coherent inference model is assumed to be processing consistent and the false-Bayesian model is also processing consistent, we only need show that for any $\mathbf{x} \in \mathbb{X}^\infty$ and $\pi \in \mathcal{P}(\Theta)$, $\tilde{\mathcal{I}}^B_{n-1,1}(\mathbf{x}, \tilde{\mathcal{I}}^B_{0,n-1}(\mathbf{x}, \pi)) = \mathcal{I}^C_{n-1,1}(\mathbf{x}, \mathcal{I}^C_{0,n-1}(\mathbf{x}, \pi))$ for any $n \geq 1$. Because g_0 is a linear function, we have

$$\tilde{\mathcal{I}}^B_{0,1}(\mathbf{x}, \pi)(\theta) = \frac{\tilde{\ell}_{0,1}(\theta|x_1)\pi(\theta)}{\int_\Theta \tilde{\ell}_{0,1}(\tilde{\theta}|x_1)\pi(\tilde{\theta})\nu(d\tilde{\theta})} = \frac{q_{0,1}(\theta|x_1)\pi(\theta)}{\int_\Theta q_{0,1}(\tilde{\theta}|x_1)\pi(\tilde{\theta})\nu(d\tilde{\theta})} = \mathcal{I}^C_{0,1}(\mathbf{x}, \pi)(\theta)$$

for ν-a.e. $\theta \in \Theta$, where the second inequality is the case due to (A.3). The standard mathematical induction then completes the proof.

Next, we prove the necessity. Suppose there exists a false-Bayesian model $\{\tilde{\mathcal{I}}^B_{m,n}\}_{m\geq 0,n\geq 1}$ with likelihood $\{\tilde{\ell}_{m,n}\}_{m\geq 0,n\geq 1}$ such that $\tilde{\mathcal{I}}^B_{m,n}(\mathbf{x}, \pi) = \mathcal{I}^C_{m,n}(\mathbf{x}, \pi)$ for any $\mathbf{x} \in \mathbb{X}^\infty$,

$\pi \in \mathcal{P}(\Theta)$, and $m \geq 0, n \geq 1$. Then, it is clear that $\{\mathcal{I}^C_{m,n}\}$ is processing consistent.

Setting $m = 0$ and $n = 1$, we have

$$\frac{\tilde{\ell}_{0,1}(\theta|x)\pi(\theta)}{\int_\Theta \tilde{\ell}_{0,1}(\tilde{\theta}|x)\pi(\tilde{\theta})\nu(d\tilde{\theta})} = \frac{q_{0,1}(\theta|x)g_0(\pi(\theta))}{\int_\Theta q_{0,1}(\tilde{\theta}|x)g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}, \quad \nu\text{-a.e. } \theta \in \Theta$$

for any $x \in \mathbb{X}$ and $\pi \in \mathcal{P}(\Theta)$. Using a similar argument to that employed in the proof of Theorem 1, we can show that g_0 is a linear function in its effective domain. As a result, we have

$$\frac{\tilde{\ell}_{0,1}(\theta|x)\pi(\theta)}{\int_\Theta \tilde{\ell}_{0,1}(\tilde{\theta}|x)\pi(\tilde{\theta})\nu(d\tilde{\theta})} = \frac{q_{0,1}(\theta|x)\pi(\theta)}{\int_\Theta q_{0,1}(\tilde{\theta}|x)\pi(\tilde{\theta})\nu(d\tilde{\theta})}, \quad \nu\text{-a.e. } \theta \in \Theta$$

for any $x \in \mathbb{X}$ and $\pi \in \mathcal{P}(\Theta)$. Because ν is σ-finite, we can construct a prior density π such that $\pi(\theta) > 0, \forall \theta \in \Theta$. Plugging this density into the previous equality, setting

$$\varphi_0(x) := \frac{\int_\Theta \tilde{\ell}_{0,1}(\tilde{\theta}|x)\pi(\tilde{\theta})\nu(d\tilde{\theta})}{\int_\Theta q_{0,1}(\tilde{\theta}|x)\pi(\tilde{\theta})\nu(d\tilde{\theta})}, \quad x \in \mathbb{X},$$

and noting that $\int_\mathbb{X} \tilde{\ell}_{0,1}(\theta|x)\nu_X(dx) = 1$ for any $\theta \in \Theta$, we immediately conclude that (1.10) holds for $m = 0$.

Finally, due to processing consistency, we have

$$\mathcal{I}^C_{n,1}(\mathbf{x}, \mathcal{I}^C_{0,n}(\mathbf{x}, \pi)) = \mathcal{I}^C_{0,n+1}(\mathbf{x}, \pi) = \tilde{\mathcal{I}}^B_{0,n+1}(\mathbf{x}, \pi) = \tilde{\mathcal{I}}^B_{n,1}(\mathbf{x}, \tilde{\mathcal{I}}^B_{0,n}(\mathbf{x}, \pi)).$$

Furthermore, if $\mathcal{I}^C_{0,n}(\mathbf{x}, \pi)(\theta)$ and $\tilde{\mathcal{I}}^B_{0,n}(\mathbf{x}, \pi)(\theta)$ are positive for ν-a.e. $\theta \in \Theta$, so are $\mathcal{I}^C_{n,1}(\mathbf{x}, \mathcal{I}^C_{0,n}(\mathbf{x}, \pi))(\theta)$ and $\tilde{\mathcal{I}}^B_{n,1}(\mathbf{x}, \tilde{\mathcal{I}}^B_{0,n}(\mathbf{x}, \pi))(\theta)$. Then, the standard mathematical induction shows that (1.10) holds for any $m \geq 0$.

Part (ii): We first prove the necessity. Suppose for any $\pi \in \mathcal{P}(\Theta)$ there exist a false-Bayesian model $\{\tilde{\mathcal{I}}^B_{m,n}\}_{m\geq 0, n\geq 1}$ with likelihood $\{\tilde{\ell}_{m,n}\}_{m\geq 0, n\geq 1}$ and a false prior density $\tilde{\pi}$ such that $\tilde{\mathcal{I}}^B_{m,n}(\mathbf{x}, \tilde{\pi}) = \mathcal{I}^C_{m,n}(\mathbf{x}, \pi)$ for any $\mathbf{x} \in \mathbb{X}^\infty$ and $m \geq 0, n \geq 1$. Then, it is clear that $\{\mathcal{I}^C_{m,n}\}$ is processing consistent.

Choose any $\pi \in \mathcal{P}(\Theta)$ such that $\pi(\theta) > 0$ for ν-a.e. $\theta \in \Theta$. By assumption, $\tilde{\mathcal{I}}^B_{0,1}(\mathbf{x}, \tilde{\pi}) = \mathcal{I}^C_{0,1}(\mathbf{x}, \pi)$. Furthermore, we have $\mathcal{I}^C_{0,1}(\mathbf{x}, \pi)(\theta) > 0$, ν-a.e. $\theta \in \Theta$. Because for any $m \geq 1$

$$\mathcal{I}^C_{m,1}(\mathbf{x}, \mathcal{I}^C_{0,m}(\mathbf{x}, \pi)) = \mathcal{I}^C_{0,m+1}(\mathbf{x}, \pi) = \tilde{\mathcal{I}}^B_{0,m+1}(\mathbf{x}, \tilde{\pi}) = \tilde{\mathcal{I}}^B_{m,1}(\mathbf{x}, \tilde{\mathcal{I}}^B_{0,m}(\mathbf{x}, \tilde{\pi}))$$

holds for any $\mathbf{x} \in \mathbb{X}^\infty$, following the same proof as for part (i) of this theorem, we

conclude (1.10) holds for any $m \geq 1$.

Next, we prove the sufficiency. Fix any $\pi \in \mathcal{P}(\Theta)$. Because ν_X is σ-finite and $0 < \int_\Theta q_{0,1}(\tilde{\theta}|x) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta}) < +\infty$ for each $x \in \mathbb{X}$, we can find a countable partition $A_n, n \geq 1$ of $\mathbb{X}$ and a sequence of positive numbers $a_n, n \geq 1$ such that $\int_\Theta q_{0,1}(\tilde{\theta}|x) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta}) \leq a_n, x \in A_n$ and $\nu_X(A_n) \leq a_n$, $n \geq 1$. Define $h_0(x) := \sum_{n=1}^\infty \frac{1}{2^n a_n^2} \mathbf{1}_{A_n}(x), x \in \mathbb{X}$. Then,

$$\begin{aligned}
&\int_{\mathbb{X}} \int_\Theta h_0(x) q_{0,1}(\tilde{\theta}|x) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})\nu_X(dx) \\
=&\int_{\mathbb{X}} h_0(x) \int_\Theta q_{0,1}(\tilde{\theta}|x) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})\nu_X(dx) \\
=&\sum_{n=1}^\infty \int_{A_n} h_0(x) \int_\Theta q_{0,1}(\tilde{\theta}|x) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})\nu_X(dx) \\
\leq&\sum_{n=1}^\infty \frac{1}{2^n a_n^2} a_n \nu(A_n) \leq 1 < +\infty.
\end{aligned}$$

Consequently, $g_0(\pi(\theta)) \int_{\mathbb{X}} h_0(x) q_{0,1}(\theta|x)\nu_X(dx) < +\infty$ for ν-a.e. $\theta \in \Theta$. Denote $\Theta_0 := \{\theta \in \Theta | \pi(\theta) = 0\}$. For any $\theta \in \Theta \backslash \Theta_0$, $g_0(\pi(\theta)) > 0$, so $\int_{\mathbb{X}} h_0(x) q_{0,1}(\theta|x)\nu_X(dx) < +\infty$. As a result, for any $\theta \in \Theta \backslash \Theta_0$, we can define

$$\tilde{\pi}(\theta) := \frac{g_0(\pi(\theta)) \int_{\mathbb{X}} h_0(\tilde{x}) q_{0,1}(\theta|\tilde{x})\nu_X(d\tilde{x})}{\int_{\mathbb{X}} \int_\Theta h_0(\tilde{x}) q_{0,1}(\tilde{\theta}|\tilde{x}) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})\nu_X(d\tilde{x})}, \quad \tilde{\ell}_{0,1}(\theta|x) := \frac{h_0(x) q_{0,1}(\theta|x)}{\int_{\mathbb{X}} h_0(\tilde{x}) q_{0,1}(\theta|\tilde{x})\nu_X(d\tilde{x})}.$$

For $\theta \in \Theta_0$, define $\tilde{\pi}(\theta) = 0$ and $\tilde{\ell}_{0,1}(\theta|x) = p(x), x \in \mathbb{X}$ for an arbitrarily chosen p such that $p(x) \geq 0, x \in \mathbb{X}$ and $\int_{\mathbb{X}} p(x)\nu_X(dx) = 1$.

Now, define $\tilde{\ell}_{m,1}$ for $m \geq 1$ as in (A.3) and define $\tilde{\ell}_{m,n} := \prod_{i=m+1}^{m+n} \tilde{\ell}_{i-1,1}$ for $m \geq 0$ and $n \geq 1$. One can verify that $\int_{\mathbb{X}} \tilde{\ell}_{m,1}(\theta|\mathbf{x}_{0,m}, \tilde{x})\nu_X(d\tilde{x}) = 1$ for any $\mathbf{x}_{0,m} \in \mathbb{X}^m$, $\theta \in \Theta$, and $m \geq 0$, so $\{\tilde{\ell}_{m,n}\}$ is the likelihood of a false model. Denote by $\{\tilde{\mathcal{I}}^B_{m,n}\}$ the false-Bayesian model associated with this false likelihood. Then, we have

$$\begin{aligned}
\tilde{\mathcal{I}}^B_{0,1}(\mathbf{x}, \tilde{\pi})(\theta) &= \frac{\tilde{\ell}_{0,1}(\theta|x_1)\tilde{\pi}(\theta)}{\int_\Theta \tilde{\ell}_{0,1}(\tilde{\theta}|x_1)\tilde{\pi}(\tilde{\theta})\nu(d\tilde{\theta})} = \frac{h_0(x_1) q_{0,1}(\theta|x_1) g_0(\pi(\theta))}{\int_\Theta h_0(x_1) q_{0,1}(\tilde{\theta}|x_1) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})} \\
&= \frac{q_{0,1}(\theta|x_1) g_0(\pi(\theta))}{\int_\Theta q_{0,1}(\tilde{\theta}|x_1) g_0(\pi(\tilde{\theta}))\nu(d\tilde{\theta})} = \mathcal{I}^C_{0,1}(\mathbf{x}, \pi)(\theta), \quad \nu\text{-a.e. } \theta \in \Theta.
\end{aligned}$$

Finally, using the same argument as in the proof of part (i) of the theorem, we can

conclude that for any $m \geq 1$

$$\mathcal{I}^C_{0,m+1}(\mathbf{x}, \pi) = \mathcal{I}^C_{m,1}(\mathbf{x}, \mathcal{I}^C_{0,m}(\mathbf{x}, \pi)) = \tilde{\mathcal{I}}^B_{m,1}(\mathbf{x}, \tilde{\mathcal{I}}^B_{0,m}(\mathbf{x}, \tilde{\pi})) = \tilde{\mathcal{I}}^B_{0,m+1}(\mathbf{x}, \tilde{\pi}). \quad \square$$

A.2 Proofs in Chapter 2

Proof of Theorem 3 We first prove the sufficiency; i.e., suppose (2.3) holds and we prove that (2.2) holds. To this end, we only need to prove that (2.2) holds for any $m \geq 0$ and $n_1 = n_2 = 1$. In the following proof, when $m = 0$, π_m stands for the distorted prior as defined in (1.5).

Because both $\{\ell_{m,n}\}$ and $\{q_{m,n}\}$ satisfy the product rule and (2.3) holds, there exists $\bar{C}_m(\mathbf{x}) > 0$ that is independent of θ such that

$$q_{m,1}(\theta|\mathbf{x}) = \bar{C}_m(\mathbf{x})\ell_{m,1}(\theta|\mathbf{x}), \quad \theta \in \Theta.$$

Consequently,

$$\begin{aligned}
\mu_{m,1}(\mathbf{x})\mu_{m+1,1}(\mathbf{x}) &= \int_\Theta \ell_{m,1}(\tilde{\theta}|\mathbf{x})\pi_m(\tilde{\theta}|\mathbf{x})\nu(d\tilde{\theta}) \int_\Theta \ell_{m+1,1}(\theta|\mathbf{x})\pi_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&= \int_\Theta \ell_{m,1}(\tilde{\theta}|\mathbf{x})\pi_m(\tilde{\theta}|\mathbf{x})\nu(d\tilde{\theta}) \int_\Theta \ell_{m+1,1}(\theta|\mathbf{x}) \frac{q_{m,1}(\theta|\mathbf{x})\pi_m(\theta|\mathbf{x})}{\int_\Theta q_{m,1}(\tilde{\theta}|\mathbf{x})\pi_m(\tilde{\theta}|\mathbf{x})\nu(d\tilde{\theta})}\nu(d\theta) \\
&= \int_\Theta \ell_{m,1}(\tilde{\theta}|\mathbf{x})\pi_m(\tilde{\theta}|\mathbf{x})\nu(d\tilde{\theta}) \int_\Theta \ell_{m+1,1}(\theta|\mathbf{x}) \frac{\ell_{m,1}(\theta|\mathbf{x})\pi_m(\theta|\mathbf{x})}{\int_\Theta \ell_{m,1}(\tilde{\theta}|\mathbf{x})\pi_m(\tilde{\theta}|\mathbf{x})\nu(d\tilde{\theta})}\nu(d\theta) \\
&= \int_\Theta \ell_{m+1,1}(\theta|\mathbf{x})\ell_{m,1}(\theta|\mathbf{x})\pi_m(\theta|\mathbf{x})\nu(d\theta) \\
&= \int_\Theta \ell_{m,2}(\theta|\mathbf{x})\pi_m(\theta|\mathbf{x})\nu(d\theta) \\
&= \mu_{m,2}(\mathbf{x}).
\end{aligned}$$

Next, we prove the necessity. We first consider the case in which $\mathcal{A}_\nu = \Theta$. In this case, consider any $\theta_1, \theta_2 \in \Theta$. Fixing $m \geq 0$ and $\mathbf{x}_{0,m} \in \mathbb{X}^m$, for any $y \geq 0$ and $z \geq 0$ such that $\nu(\{\theta_1\})y + \nu(\{\theta_2\})z = 1$, we can find a prior density π_m such that

$\pi_m(\theta_1|\mathbf{x}) = y$ and $\pi_m(\theta_2|\mathbf{x}) = z$; see the proof of Theorem 1. Then,

$$\pi_{m+1}(\theta_1|\mathbf{x}) = \frac{q_{m,1}(\theta_1|\mathbf{x})y}{q_{m,1}(\theta_1|\mathbf{x})\nu(\{\theta_1\})y + q_{m,1}(\theta_2|\mathbf{x})\nu(\{\theta_2\})z},$$
$$\pi_{m+1}(\theta_2|\mathbf{x}) = \frac{q_{m,1}(\theta_2|\mathbf{x})z}{q_{m,1}(\theta_1|\mathbf{x})\nu(\{\theta_1\})y + q_{m,1}(\theta_2|\mathbf{x})\nu(\{\theta_2\})z}.$$

Consequently, because (2.2) holds for $n_1 = n_2 = 1$, we have

$$\begin{aligned}
&\ell_{m,2}(\theta_1|\mathbf{x})\nu(\{\theta_1\})y + \ell_{m,2}(\theta_2|\mathbf{x})\nu(\{\theta_2\})z \\
=&\big(\ell_{m,1}(\theta_1|\mathbf{x})\nu(\{\theta_1\})y + \ell_{m,1}(\theta_2|\mathbf{x})\nu(\{\theta_2\})z\big) \\
&\times \frac{\ell_{m+1,1}(\theta_1|\mathbf{x})q_{m,1}(\theta_1|\mathbf{x})\nu(\{\theta_1\})y + \ell_{m+1,1}(\theta_2|\mathbf{x})q_{m,1}(\theta_2|\mathbf{x})\nu(\{\theta_2\})z}{q_{m,1}(\theta_1|\mathbf{x})\nu(\{\theta_1\})y + q_{m,1}(\theta_2|\mathbf{x})\nu(\{\theta_2\})z},
\end{aligned}$$

which, together with $\nu(\{\theta_1\})y + \nu(\{\theta_2\})z = 1$, implies

$$\begin{aligned}
&[\ell_{m,2}(\theta_1|\mathbf{x})\nu(\theta_1)y + \ell_{m,2}(\theta_2|\mathbf{x})(1-\nu(\theta_1)y)][q_{m,1}(\theta_1|\mathbf{x})\nu(\theta_1)y + q_{m,1}(\theta_2|\mathbf{x})(1-\nu(\theta_1)y)] \\
=&[\ell_{m,1}(\theta_1|\mathbf{x})\nu(\theta_1)y + \ell_{m,1}(\theta_2|\mathbf{x})(1-\nu(\theta_1)y)]\times \\
&[\ell_{m+1,1}(\theta_1|\mathbf{x})q_{m,1}(\theta_1|\mathbf{x})\nu(\theta_1)y + \ell_{m+1,1}(\theta_2|\mathbf{x})q_{m,1}(\theta_2|\mathbf{x})(1-\nu(\theta_1)y)].
\end{aligned}$$

Note that the above equation holds for any y such that $\nu(\{\theta_1\})y \in [0,1]$, so each of the y^2 term, y term, and constant term in the above equation must be zero. This implies

$$\begin{aligned}
&q_{m,1}(\theta_1|\mathbf{x})\ell_{m,2}(\theta_2|\mathbf{x}) + q_{m,1}(\theta_2|\mathbf{x})\ell_{m,2}(\theta_1|\mathbf{x}) \\
=&\ell_{m,1}(\theta_2|\mathbf{x})q_{m,1}(\theta_1|\mathbf{x})\ell_{m+1,1}(\theta_1|\mathbf{x}) + \ell_{m,1}(\theta_1|\mathbf{x})q_{m,1}(\theta_2|\mathbf{x})\ell_{m+1,1}(\theta_2|\mathbf{x}). \qquad \text{(A.4)}
\end{aligned}$$

Because $\ell_{m,2}(\theta|\mathbf{x}) = \ell_{m,1}(\theta|\mathbf{x})\ell_{m+1,1}(\theta|\mathbf{x})$ for any $\theta \in \Theta$, (A.4) yields

$$\begin{aligned}
&\ell_{m,1}(\theta_1|\mathbf{x})\ell_{m+1,1}(\theta_1|\mathbf{x})q_{m,1}(\theta_2|\mathbf{x}) + q_{m,1}(\theta_1|\mathbf{x})\ell_{m,1}(\theta_2|\mathbf{x})\ell_{m+1,1}(\theta_2|\mathbf{x}) \\
=&\ell_{m,1}(\theta_1|\mathbf{x})\ell_{m+1,1}(\theta_2|\mathbf{x})q_{m,1}(\theta_2|\mathbf{x}) + \ell_{m,1}(\theta_2|\mathbf{x})\ell_{m+1,1}(\theta_1|\mathbf{x})q_{m,1}(\theta_1|\mathbf{x}). \qquad \text{(A.5)}
\end{aligned}$$

Now, define $LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x}) = \frac{\ell_{m+1,1}(\theta_1|\mathbf{x})}{\ell_{m+1,1}(\theta_2|\mathbf{x})}$, $QR_{m,1}(\theta_1,\theta_2|\mathbf{x}) = \frac{q_{m,1}(\theta_1|\mathbf{x})}{q_{m,1}(\theta_2|\mathbf{x})}$, $LR_{m,1}(\theta_1,\theta_2|\mathbf{x}) = \frac{\ell_{m,1}(\theta_1|\mathbf{x})}{\ell_{m,1}(\theta_2|\mathbf{x})}$. Dividing both sides of (A.5) by $\ell_{m+1,1}(\theta_2|\mathbf{x})\ell_{m,1}(\theta_2|\mathbf{x})q_{m,1}(\theta_2|\mathbf{x})$, we obtain

$$\begin{aligned}
&LR_{m,1}(\theta_1,\theta_2|\mathbf{x})LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x}) + QR_{m,1}(\theta_1,\theta_2|\mathbf{x}) \\
=\ &LR_{m,1}(\theta_1,\theta_2|\mathbf{x}) + LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x})QR_{m,1}(\theta_1,\theta_2|\mathbf{x}),
\end{aligned}$$

which implies

$$(LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x})-1)(LR_{m,1}(\theta_1,\theta_2|\mathbf{x})-QR_{m,1}(\theta_1,\theta_2|\mathbf{x}))=0. \tag{A.6}$$

We claim that (A.6) implies that $LR_{m,1}(\theta_1,\theta_2|\mathbf{x})=QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$. Suppose for the sake of contradiction that $LR_{m,1}(\theta_1,\theta_2|\mathbf{x})\neq QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$. Then, we must have $LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x})=1$. Note that $LR_{m,1}(\theta_1,\theta_2|\mathbf{x})$ and $QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$ depend on $\mathbf{x}_{m+1}$ only, but $LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x})$ depends on $\mathbf{x}_{m+1,1}$. Because given $\mathbf{x}_{m+1}$, $LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x})$ cannot be constant with respect to $\mathbf{x}_{m+1,1}$, we conclude that $LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x})$ cannot be equal to 1. Consequently, we have $LR_{m,1}(\theta_1,\theta_2|\mathbf{x})=QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$.

We have shown that for any $\theta_1,\theta_2\in\Theta=\mathcal{A}_\nu$, $LR_{m,1}(\theta_1,\theta_2|\mathbf{x})=QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$. It is straightforward to see that this is equivalent to the existence of $C_{m,1}(\mathbf{x})>0$ such that

$$q_{m,1}(\theta|\mathbf{x})=C_{m,1}(\mathbf{x})\ell_{0,1}(\theta|\mathbf{x}),\quad ,\theta\in\Theta.$$

Because both $\{q_{m,n}\}$ and $\{\ell_{m,n}\}$ satisfy the product rule, we conclude (2.3) holds for any m.

Next, we consider the general case. Again, we only need to prove that for $m\geq 0$, any $\mathbf{x}\in\mathbb{X}^\infty$, and any $\theta_1,\theta_2\in\Theta$, $LR_{m,1}(\theta_1,\theta_2|\mathbf{x})=QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$.

Given $m\geq 0$, for the sake of contradiction, suppose that there exists $\mathbf{x}$ and $\theta_1,\theta_2\in\Theta$ such that $LR_{m,1}(\theta_1,\theta_2|\mathbf{x})\neq QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$. We can find $\tilde{\mathbf{x}}$ such that (i) $\tilde{\mathbf{x}}_{0,m+1}=\mathbf{x}_{0,m+1}$, which implies that $\ell_{m,1}(\theta|\mathbf{x})=\ell_{m,1}(\theta|\tilde{\mathbf{x}})$ and $q_{m,1}(\theta|\mathbf{x})=q_{m,1}(\theta|\tilde{\mathbf{x}})$ for any $\theta\in\Theta$, and (ii) $LR_{m+1,1}(\theta_1,\theta_2|\mathbf{x})\neq LR_{m+1,1}(\theta_1,\theta_2|\tilde{\mathbf{x}})$.

If $\theta_i\in\mathcal{A}_\nu$, we define $\mathcal{N}_i:=\{\theta_i\}$. If $\theta_i\notin\mathcal{A}_\nu$, for any $\epsilon>0$, with Assumption 2 in place, we can find a neighbourhood $\mathcal{N}_i$ of θ_i such that (i) $\nu(\mathcal{N}_i)\in(0,\epsilon)$, (ii) $\mathcal{N}_1\cap\mathcal{N}_2=\emptyset$, (iii) $|\ell_{m,1}(\theta|\mathbf{x})-\ell_{m,1}(\theta_i|\mathbf{x})|<\epsilon, |q_{m,1}(\theta|\mathbf{x})-q_{m,1}(\theta_i|\mathbf{x})|<\epsilon, |\ell_{m+1,1}(\theta|\mathbf{x})-\ell_{m+1,1}(\theta_i|\mathbf{x})|<\epsilon, |\ell_{m+1,1}(\theta|\tilde{\mathbf{x}})-\ell_{m+1,1}(\theta_i|\tilde{\mathbf{x}})|<\epsilon$, $\forall\theta\in\mathcal{N}_i$.

Now, for any $y\geq 0$ and $z\geq 0$ such that $\nu(\mathcal{N}_1)y+\nu(\mathcal{N}_2)z=1$, we can find a prior density π_0 such that $\pi_m(\theta|\mathbf{x})=y$, $\theta\in\mathcal{N}_1$ and $\pi_m(\theta|\mathbf{x})=z$, $\theta\in\mathcal{N}_2$; see the proof of Theorem 1. Then,

$$\pi_{m+1}(\theta|\mathbf{x})=\frac{q_{m,1}(\theta|\mathbf{x})y\mathbf{1}_{\mathcal{N}_1}+q_{m,1}(\theta|\mathbf{x})z\mathbf{1}_{\mathcal{N}_2}}{y\int_{\mathcal{N}_1}q_{m,1}(\theta|\mathbf{x})\nu(d\theta)+z\int_{\mathcal{N}_2}q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}.$$

Consequently, because (2.2) holds for $n_1 = n_2 = 1$, we have

$$y \int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) + z \int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) = \frac{y \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) + z \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{y \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) + z \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)} \times \left(y \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) + z \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \right),$$

which, together with $\nu(\mathcal{N}_1)y + \nu(\mathcal{N}_2)z = 1$, implies

$$\begin{aligned}
&(\int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) - \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta))y^2 \\
&+ (\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) + \int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&- \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&- \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta)) \frac{1 - y\nu(\mathcal{N}_1)}{\nu(\mathcal{N}_2)} y \\
&+ \left(\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) - \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \right) \\
&\quad \times \frac{(1 - y\nu(\mathcal{N}_1))^2}{\nu(\mathcal{N}_2)^2} = 0
\end{aligned}$$

The above equation holds for any $y > 0$ such that $\nu(\mathcal{N}_1)y \leq 1$, so each of the y^2, y,

and constant terms should be zero. Thus, we have the following three equations:

$$\begin{aligned}
&\int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) - \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&- \frac{\nu(\mathcal{N}_1)}{\nu(\mathcal{N}_2)} \Bigg[\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) + \int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&\quad - \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&\quad - \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \Bigg] \\
&+ \left(\frac{\nu(\mathcal{N}_1)}{\nu(\mathcal{N}_2)}\right)^2 \Bigg[\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \\
&\quad - \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \Bigg] = 0, \qquad \text{(A.7)} \\
&- 2\frac{\nu(\mathcal{N}_1)}{\nu(\mathcal{N}_2)^2} \Bigg[\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \\
&\quad - \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \Bigg] \\
&+ \frac{1}{\nu(\mathcal{N}_2)} \Bigg[\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) + \int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&\quad - \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\
&\quad - \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \Bigg] = 0, \qquad \text{(A.8)} \\
&\frac{1}{\nu(\mathcal{N}_2)^2} \Bigg[\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \\
&\quad - \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \Bigg] = 0. \qquad \text{(A.9)}
\end{aligned}$$

Note that (A.9) is equivalent to

$$\begin{aligned}
&\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \\
&- \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) = 0. \qquad \text{(A.10)}
\end{aligned}$$

Plugging (A.10) into (A.8), we obtain

$$\begin{aligned}\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) + \int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\ - \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\ - \int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) = 0.\end{aligned} \tag{A.11}$$

Plugging (A.11) and (A.10) into (A.7), we obtain

$$\begin{aligned}\int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta) \\ - \int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta) = 0.\end{aligned} \tag{A.12}$$

We then conclude from (A.12) and (A.10) the following:

$$\frac{\int_{\mathcal{N}_i} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_i} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta)} = \frac{\int_{\mathcal{N}_i} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_i} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta)}, \quad i = 1, 2. \tag{A.13}$$

Dividing both sides of (A.11) by $\int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)$, we obtain

$$\begin{aligned}&\frac{\int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)} + \frac{\int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta)} \\ =&\frac{\int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)} \\ &+ \frac{\int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta) \int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)},\end{aligned}$$

which, together with (A.13), implies

$$\begin{aligned}&\frac{\int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)} + \frac{\int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta)} \\ =& \frac{\int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta)} + \frac{\int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}{\int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x}) q_{m,1}(\theta|\mathbf{x})\nu(d\theta)}.\end{aligned} \tag{A.14}$$

Multiply both sides by $\frac{\nu(\mathcal{N}_2)}{\nu(\mathcal{N}_1)}$, we obtain

$$\begin{aligned}&\frac{\int_{\mathcal{N}_1} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_1)}{\int_{\mathcal{N}_2} q_{m,1}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_2)}+\frac{\int_{\mathcal{N}_1} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_1)}{\int_{\mathcal{N}_2} \ell_{m,2}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_2)}\\ =&\frac{\int_{\mathcal{N}_1} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_1)}{\int_{\mathcal{N}_2} \ell_{m,1}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_2)}+\frac{\int_{\mathcal{N}_1} \ell_{m+1,1}(\theta|\mathbf{x})q_{m,1}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_1)}{\int_{\mathcal{N}_2} \ell_{m+1,1}(\theta|\mathbf{x})q_{m,1}(\theta|\mathbf{x})\nu(d\theta)/\nu(\mathcal{N}_2)}.\end{aligned} \tag{A.15}$$

Note that the above equation holds for $\mathbf{x}$ replaced by $\tilde{\mathbf{x}}$ as well.

Now, by sending $\epsilon \to 0$, we conclude from (A.15) that

$$\begin{aligned}&LR_{m,1}(\theta_1,\theta_2|\tilde{\mathbf{x}})LR_{m+1,1}(\theta_1,\theta_2|\tilde{\mathbf{x}})+QR_{m,1}(\theta_1,\theta_2|\tilde{\mathbf{x}})\\ =\;&LR_{m,1}(\theta_1,\theta_2|\tilde{\mathbf{x}})+LR_{m+1,1}(\theta_1,\theta_2|\tilde{\mathbf{x}})QR_{m,1}(\theta_1,\theta_2|\tilde{\mathbf{x}}).\end{aligned}$$

Now, the same argument as in the case in which $\mathcal{A}_\nu = \Theta$ shows that $LR_{m,1}(\theta_1,\theta_2|\mathbf{x}) = QR_{m,1}(\theta_1,\theta_2|\mathbf{x})$, a contradiction. Thus, the proof completes. □

Proof of Proposition Proposition 4 Denote $p(k), k = 0, 1, \ldots, N$ as the probability mass function of $S_N := \sum_{i=1}^N X_i$. Then,

$$\begin{aligned}h(\beta) :&= \mathbb{E}[\max(Z_N, 1-Z_N)] = \mathbb{E}\left[\max\left(\frac{a+\beta S_N}{a+b+\beta N}, 1-\frac{a+\beta S_N}{a+b+\beta N}\right)\right]\\ &= \sum_{k=\lfloor \frac{N}{2}+\frac{b-a}{2\beta}\rfloor+1}^{N} \frac{a+\beta k}{a+b+\beta N}p(k) + \sum_{k=0}^{\lfloor \frac{N}{2}+\frac{b-a}{2\beta}\rfloor\wedge N}\left(1-\frac{a+\beta k}{a+b+\beta N}\right)p(k),\end{aligned}$$

where $\lfloor x\rfloor$ stands for the largest integer dominated by x and $x\wedge y := \min(x,y)$. For any $\beta > 0$ such that $\frac{N}{2}+\frac{b-a}{2\beta}$ is not an integer, we have

$$h'(\beta) = \frac{1}{(a+b+\beta N)^2}\varphi\left(\lfloor\frac{N}{2}+\frac{b-a}{2\beta}\rfloor\wedge N\right),$$

where

$$\varphi(m) := \left[\sum_{k=m+1}^{N}(k(a+b)-Na)p(k) + \sum_{k=0}^{m}(Na-k(a+b))p(k)\right].$$

Therefore, to show that h is increasing in β, we only need to show that $\varphi(m) \geq 0, m = 0, 1\ldots, N$.

Note that

$$\sum_{k=0}^{N} k(a+b)p(k) = \mathbb{E}[(a+b)S_N] = (a+b)\mathbb{E}[S_N] = (a+b)N\frac{a}{a+b} = aN.$$

Thus, denoting $Y := (a+b)S_N$, we have

$$\begin{aligned}\varphi(m) &= \mathbb{E}\left[(Y - \mathbb{E}[Y])\mathbf{1}_{Y>(a+b)m}\right] + \mathbb{E}\left[(\mathbb{E}[Y] - Y)\mathbf{1}_{Y\le(a+b)m}\right] \\ &= \mathbb{E}\left[Y - \mathbb{E}[Y]\right] + 2\mathbb{E}\left[(\mathbb{E}[Y] - Y)\mathbf{1}_{Y\le(a+b)m}\right] \\ &= 2\mathbb{E}\left[(\mathbb{E}[Y] - Y)\mathbf{1}_{Y\le(a+b)m}\right].\end{aligned}$$

It is obvious that $\phi(y) := \mathbb{E}\left[(\mathbb{E}[Y] - Y)\mathbf{1}_{Y\le y}\right]$ is increasing when $y \le \mathbb{E}[Y]$ and decreasing when $y \ge \mathbb{E}[Y]$. Because $\phi(-\infty) = \phi(+\infty) = 0$, we conclude that $\phi(y) \ge 0$ for any y. Consequently, $\varphi(m) \ge 0$ for any m. $\square$

Appendix B

Coherence

In this section, we consider a generalization of (B.1) by dropping the coherence property. More precisely, we consider the following *generic inference model*

$$f_m(\pi_{m,n}(\theta|\mathbf{x})) = \frac{q_{m,n}(\theta|\mathbf{x})g_m(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x})g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}, \quad \theta \in \Theta. \tag{B.1}$$

It is obvious that the difference between (B.1) and (1.2) is the distortion function $f_m(\cdot)$. We show that the generic inference model (B.1) is coherent if and only if f_m is the identity function for any $m \geq 0$. Consequently, (B.1) becomes the coherent inference model (1.2).

For each $m \geq 0$, we assume that $f_m(\cdot)$ is continuous and strictly increasing. In addition, we assume that $f_m(0) = 0$. We can see that $f_m(0) = 0$ if and only if $\lim_{k\to+\infty} \pi_{m,n}(\theta_k|\mathbf{x}) = 0$ for any $\theta_k \in \Theta, k \geq 1$ and $\mathbf{x} \in \mathbb{X}$ such that $\lim_{k\to+\infty} q_{m,n}(\theta_k|\mathbf{x}) = 0$. Thus, this assumption stipulates that if the observed sample is impossible under certain parameter value, this parameter value is also impossible under the posterior belief. Finally, for technical convenience, we assume $\lim_{x\to+\infty} f_m(x) = +\infty$. Assumption 1 is also in force.

Before we proceed, let us define the effective domain of $f_m(\cdot)$. The right-hand of (B.1) is a probability density on Θ. Depending on the structure of (Θ, ν), a probability density on Θ may not be able to take all possible values in $\mathbb{R}_+$. For instance, if $\Theta = \{\theta_1, \ldots, \theta_n\}$ and $\nu(\{\theta_i\}) > 0, i = 1, \ldots, n$, the maximum value a density can take is $a := \max\{1/\nu(\{\theta_i\})|i = 1, \ldots, n\}$. In this case, the definition of $f_m^{-1}(\cdot)$ is only relevant in $[0, a]$, so the effective domain of $f_m(\cdot)$ is $[0, f_m^{-1}(a)]$. In general, one can see that the effective domain of $f_m(\cdot)$ is $[0, +\infty)$ when $\Theta/\mathcal{A}_\nu \neq \emptyset$ and is $[0, f_m^{-1}(M)] \cap [0, +\infty)$ when $\mathcal{A}_\nu = \Theta$, where $M := \max\{1/\nu(\{\theta\})|\theta \in \Theta\}$.

Theorem 4 *The following are equivalent:*

(a) For each $m \geq 0$, $n \geq 1$, any $\mathbf{x} \in \mathbb{X}^\infty$, and any $\pi(\cdot) \in \mathcal{P}(\Theta)$, $\pi_{m,n}$ defined as in (B.1) is a probability density on Θ.

(b) For each $m \geq 0$, $f_m(\cdot)$ is the identity function in its effective domain.

The following lemma is needed:

Lemma 1 *Suppose $f(\cdot)$ is a continuous function on $[0, M]$ such that $f(0) = 0, f(M) = M$, and for some $\lambda \geq 1$,*

$$f(\lambda x + y) = \lambda f(x) + f(y), \quad \forall x, y \geq 0, \lambda x + y \leq M.$$

Then, $f(x) = x, x \in [0, M]$.

Proof Without loss of generality, we assume $M = 1$. Otherwise, we can set $\tilde{f}(x) := f(Mx)/M, x \in [0, 1]$. When $\lambda = 1$, it is well-known that f is the identity function, so we assume $\lambda > 1$ in the following.

For any $n \geq 1$ and any $x_i \geq 0, i = 0, \ldots n$ such that $\sum_{i=0}^n \lambda^i x_i \leq 1$, we have

$$f\left(\sum_{i=0}^n \lambda^i x_i\right) = \sum_{i=0}^n \lambda^i f(x_i). \tag{B.2}$$

Choosing $x_i = 1/\sum_{i=0}^n \lambda^i, i = 0, \ldots, n$ in (B.2) and recalling $f(1) = 1$, we conclude that $f\left(1/\sum_{i=0}^n \lambda^i\right) = 1/\sum_{i=0}^n \lambda^i$. For any $k \in \{1, \ldots, n\}$, choose $x_k = 1/\sum_{i=0}^n \lambda^i$ and $x_i = 0, i \neq k$ in (B.2). Then, because $f(0) = 0$, we have

$$f\left(\frac{\lambda^k}{\sum_{i=0}^n \lambda^i}\right) = \lambda^k f\left(\frac{1}{\sum_{i=0}^n \lambda^i}\right) = \frac{\lambda^k}{\sum_{i=0}^n \lambda^i}.$$

For each $m \geq 1$, define

$$B_{n,m} := \left\{ \frac{\sum_{j=0}^m \lambda^{k_j}}{\sum_{i=0}^n \lambda^i} \leq 1 | k_j \in \{0, 1, \ldots n\}, j = 1, \ldots, m \text{ and } k_j = 0 \text{ for at most one } j. \right\}$$

We show by induction that $f(x) = x, x \in B_{n,m}$ for any $m \geq 1$. We have shown that $f(x) = x, x \in B_{n,1}$. Suppose $f(x) = x, x \in B_{n,m}$. For any $x = \sum_{j=0}^{m+1} \lambda^{k_j} / \sum_{i=0}^n \lambda^i \in B_{n,m+1}$, we assume $k_{m+1} > 0$ without loss of generality. Let $x_1 := \lambda^{k_{m+1}-1} / \sum_{i=0}^n \lambda^i$ and $x_2 := \sum_{j=0}^m \lambda^{k_j} / \sum_{i=0}^n \lambda^i$, then $x = \lambda x_1 + x_2$. Recalling that $x_1 \in B_{n,m}, x_2 \in B_{n,1}$

and $f(\lambda x_1 + x_2) = \lambda f(x_1) + f(x_2)$, we conclude that

$$f(x) = f(\lambda x_1 + x_2) = \lambda f(x_1) + f(x_2) = \lambda x_1 + x_2 = x.$$

We have shown that $f(x) = x$ for all $x \in B := \cup_{n=0}^{\infty} B_n$ where $B_0 := \{0, 1\}$ and $B_n := \cup_{m=1}^{\infty} B_{n,m}, n \geq 1$. Next, we show that B is dense in $[0, 1]$. For any $x_0 \in [0, 1]$ and any $\epsilon > 0$, choose $n \geq 1$ such that $\frac{\lambda}{\sum_{i=0}^{n} \lambda^i} < \epsilon$. If $x_0 \leq \frac{1}{\sum_{i=0}^{n} \lambda^i}$, then $|x_0 - 0| < \epsilon$. Otherwise, let k_1 be the largest k such that $\frac{\lambda^k}{\sum_{i=0}^{n} \lambda^i} \leq x_0$. Denote $x_1 := \frac{\lambda^{k_1}}{\sum_{i=0}^{n} \lambda^i}$, then $x_1 \in B_n$ and $x_0 - x_1 \geq 0$. If $k_1 = 0$, then $x_0 - x_1 \leq \frac{\lambda}{\sum_{i=0}^{n} \lambda^i}$. Otherwise, let k_2 be the largest k such that $\frac{\lambda^k}{\sum_{i=0}^{n} \lambda^i} \leq x_0 - x_1$ and denote $x_2 := x_1 + \frac{\lambda^{k_2}}{\sum_{i=0}^{n} \lambda^i}$. Then, $x_2 \in B_n$ and $x_0 - x_2 \geq 0$. If $k_2 = 0$, then $x_0 - x_2 \leq \frac{\lambda}{\sum_{i=0}^{n} \lambda^i}$. Otherwise, let k_3 be the largest k such that $\frac{\lambda^k}{\sum_{i=0}^{n} \lambda^i} \leq x_0 - x_2$ and denote $x_3 := x_2 + \frac{\lambda^{k_3}}{\sum_{i=0}^{n} \lambda^i}$. Repeating this construction, after finite number of steps, e.g., ℓ, we must have $k_\ell = 0$. As a result, $0 \leq x_0 - x_\ell \leq \frac{\lambda}{\sum_{i=0}^{n} \lambda^i} < \epsilon$ and $x_\ell \in B_n$. Thus, B is dense in $[0, 1]$.

Because $f(x) = x, x \in B$, B is dense in $[0, 1]$, and $f(\cdot)$ is continuous on $[0, 1]$, we have $f(x) = x, x \in [0, 1]$. $\square$

Proof of Theorem 4 The direction (b)⇒(a) is obvious, so we only need to prove (a)⇒(b). Let us fix $m \geq 0$. We first consider the case in which $\Theta/\mathcal{A}_\nu \neq \emptyset$. In this case, $\nu(\Theta/\mathcal{A}_\mu) \in (0, +\infty]$ because $\Theta/\mathcal{A}_\mu$ is open and the support of ν is Θ. Fix any $\mathbf{x} \in \mathbb{X}^\infty$, then we can find $\theta_0 \in \Theta/\mathcal{A}_\nu$ such that $q_{m,n}(\theta_0|\mathbf{x}) > 0$. Because $q_{m,n}$ is continuous in θ, we can find a neighbourhood of θ_0, denoted as $\mathcal{N}_0$, such that $\mathcal{N}_0 \subset \Theta/\mathcal{A}_\nu$ and $1/C \leq q_{m,n}(\theta|\mathbf{x}) \leq C, \forall \theta \in \mathcal{N}_0$ for some $C > 0$. Because the support of ν is Θ, we have $\nu(\mathcal{N}_0) \in (0, +\infty]$. Because ν is σ-finite, we can find $\mathcal{N}_1 \subset \mathcal{N}_0$ such that $\nu(\mathcal{N}_1) \in (0, +\infty)$.

Next, we show that for any bounded density function $p(\cdot)$ on Θ (i.e., $0 \leq p(\theta) \leq K, \forall \theta \in \Theta$ for some $K > 0$ and $\int_\Theta p(\theta)\nu(d\theta) = 1$) such that $p(\theta) = 0, \theta \in \Theta/\mathcal{N}_1$, we can construct a prior density $\pi(\cdot)$ such that

$$p(\theta) = \frac{q_{m,n}(\theta|\mathbf{x}) g_m(\pi(\theta))}{\int_\Theta q_{m,n}(\tilde{\theta}|\mathbf{x}) g_m(\pi(\tilde{\theta}))\nu(d\tilde{\theta})}, \quad \forall \theta \in \Theta. \tag{B.3}$$

To this end, for each $z \geq 0$, define

$$h_z(\theta) := g_m^{-1}\left(\frac{zp(\theta)}{q_{m,n}(\theta|\mathbf{x})}\right)\mathbf{1}_{\mathcal{N}_1}(\theta), \quad \theta \in \Theta.$$

Because $1/C \leq q_{m,n}(\theta|\mathbf{x}) \leq C, \forall \theta \in \mathcal{N}_1$, $h_z(\cdot)$ is well-defined. Furthermore, because

$\nu(\mathcal{N}_1) < +\infty$ and $p(\cdot)$ is bounded, the monotone convergence theorem shows that $\lambda(z) := \int_\Theta h_z(\theta)\nu(d\theta)$ is continuous in z. Because $g_m(0) = 0$, we have $\lambda(0) = 0$. On the other hand, because $\nu(\mathcal{N}_1) > 0$, we have $\lim_{z\to+\infty} \lambda(z) = +\infty$. As a result, there exist $z_0 > 0$ such that $\lambda(z_0) = 1$. Define $\pi(\theta) = h_{z_0}(\theta), \theta \in \Theta$, then π is a probability density. Furthermore, thanks to $g_m(0) = 0$, one can check that $p(\theta) \propto q_{m,n}(\theta|\mathbf{x})g_m(\pi(\theta))$ and thus (B.3) holds.

It is assumed that $\pi_{m,n}$ defined as in (B.1) is a probability density for any $\pi \in \mathcal{P}(\Theta)$. In addition, any bounded density function $p(\cdot)$ such that $p(\theta) = 0, \theta \in \Theta/\mathcal{N}_1$ can be constructed in the form of (B.3) from some prior density π. Therefore, we conclude that for any such $p(\cdot)$, the function $\pi_{m,n}$ solved from

$$f_m(\pi_{m,n}(\theta|\mathbf{x})) = p(\theta), \quad \theta \in \Theta \tag{B.4}$$

must satisfy $\int_\Theta \pi_{m,n}(\theta|\mathbf{x})\nu(d\theta) = 1$.

Now, for any $y \geq 1/\nu(\mathcal{N}_1)$, because ν has no atom on $\Theta/\mathcal{A}_\nu$, we can find $\mathcal{N}_2 \subset \mathcal{N}_1$ such that $\nu(\mathcal{N}_2) = 1/y$; see for instance Föllmer and Schied [2004, Proposition A.27]. Define $p(\theta) := y\mathbf{1}_{\mathcal{N}_2}(\theta), \theta \in \Theta$. One can see that $p(\cdot)$ is a bounded density on Θ and takes zero outside $\mathcal{N}_2$. Thanks to $f_m(0) = 0$, $\pi_{m,n}(\theta|\mathbf{x})$ defined as in (B.4) can be solved as follows:

$$\pi_{m,n}(\theta|\mathbf{x}) = f_m^{-1}(p(\theta)) = f_m^{-1}(y)\mathbf{1}_{\mathcal{N}_2}(\theta), \quad \theta \in \Theta.$$

As a result,

$$1 = \int_\Theta \pi_{m,n}(\theta|\mathbf{x})\nu(d\theta) = f_m^{-1}(y)\nu(\mathcal{N}_2) = f_m^{-1}(y)/y,$$

which shows that $f_m^{-1}(y) = y$.

For $y \in (0, 1/\nu(\mathcal{N}_1))$, consider $p(\theta) := y\mathbf{1}_{\mathcal{N}_2}(\theta) + \tilde{y}\mathbf{1}_{\mathcal{N}_1/\mathcal{N}_2}(\theta), \theta \in \Theta$, where $\mathcal{N}_2$ is chosen to be a subset of $\mathcal{N}_1$ such that $\nu(\mathcal{N}_2) = \frac{1}{2}\nu(\mathcal{N}_1)$ and $\tilde{y} := \frac{2}{\nu(\mathcal{N}_1)} - y$. One can verify that $p(\cdot)$ is a bounded density function. Furthermore, because $y < 1/\nu(\mathcal{N}_1)$, we have $\tilde{y} > 1/\nu(\mathcal{N}_1)$ and thus $f_m^{-1}(\tilde{y}) = \tilde{y}$. Consider the corresponding $\pi_{m,n}(\theta|\mathbf{x})$ defined through (B.4). One can calculate that

$$\pi_{m,n}(\theta|\mathbf{x}) = f_m^{-1}(p(\theta)) = f_m^{-1}(y)\mathbf{1}_{\mathcal{N}_2}(\theta) + f_m^{-1}(\tilde{y})\mathbf{1}_{\mathcal{N}_1/\mathcal{N}_2}(\theta), \quad \theta \in \Theta.$$

Then,

$$1 = \int_\Theta \pi_{m,n}(\theta|\mathbf{x})\nu(d\theta) = f_m^{-1}(y)\nu(\mathcal{N}_2) + f_m^{-1}(\tilde{y})\nu(\mathcal{N}_1/\mathcal{N}_2) = \frac{\nu(\mathcal{N}_1)}{2}(f_m^{-1}(y) + f_m^{-1}(\tilde{y})).$$

As a result,

$$f_m^{-1}(y) = \frac{2}{\nu(\mathcal{N}_1)} - f_m^{-1}(\tilde{y}) = \frac{2}{\nu(\mathcal{N}_1)} - \tilde{y} = \frac{2}{\nu(\mathcal{N}_1)} - \left(\frac{2}{\nu(\mathcal{N}_1)} - y\right) = y.$$

To summarize, we have shown that when $\Theta/\mathcal{A}_\mu \neq \emptyset$, $f_m^{-1}(y) = y$ for any $y > 0$, i.e., $f_m(\cdot)$ is the identity function.

Next, we consider the case in which $\mathcal{A}_\mu = \Theta$. Fixing any $\mathbf{x} \in \mathbb{X}^\infty$ and any mutually different $\theta_1, \theta_2, \theta_3$, we have $q_{m,n}(\theta_i|\mathbf{x}_{0,m}) > 0, i = 1, 2, 3$. Applying a similar argument to the one used in the case in which $\Theta/\mathcal{A}_\nu \neq \emptyset$, we can show that for any bounded density $p(\cdot)$ with support on $\{\theta_1, \theta_2, \theta_3\}$, there exist a prior density π such that (B.3) holds. As a result, for any such $p(\cdot)$, $\pi_{m,n}(\theta|\mathbf{x})$ solved through (B.4) must satisfy $\int_\Theta \pi_{m,n}(\theta|\mathbf{x})\nu(d\theta) = 1$. Consequently, for any $p_i \geq 0, i = 1, 2, 3$ such that $\sum_{i=1}^3 p_i\nu(\{\theta_i\}) = 1$, by setting $p(\theta_i) = p_i, i = 1, 2, 3$, we have

$$1 = \int_\Theta \pi_{m,n}(\theta|\mathbf{x})\nu(d\theta) = \int_\Theta f_m^{-1}(p(\theta))\nu(d\theta) = \sum_{i=1}^3 f_m^{-1}(p_i)\nu(\{\theta_i\}). \tag{B.5}$$

In particular, $f_m^{-1}\left(\frac{1}{\nu(\{\theta_i\})}\right) = 1, i = 1, 2, 3$. Without loss of generality, we assume $\frac{1}{\nu(\{\theta_1\})} \leq \frac{1}{\nu(\{\theta_2\})} \leq \frac{1}{\nu(\{\theta_3\})}$. In the following, we show that $f_m^{-1}(y) = y, 0 \leq y \leq \frac{1}{\nu(\{\theta_3\})}$.

For any x and y such that $x\nu(\{\theta_2\})+y\nu(\{\theta_3\}) \leq 1$, by setting $p_1 = \frac{1-x\nu(\{\theta_2\})-y\nu(\{\theta_3\})}{\nu(\{\theta_1\})}$, $p_2 = x$, $p_3 = y$, $\tilde{p}_1 = \frac{1-x\nu(\{\theta_2\})-y\nu(\{\theta_3\})}{\nu(\{\theta_1\})}$, $\tilde{p}_2 = 0$, and $\tilde{p}_3 = \frac{\nu(\{\theta_2\})}{\nu(\{\theta_3\})}x + y$, we have

$$\begin{aligned} f_m^{-1}(x)\nu(\{\theta_2\}) + f_m^{-1}(y)\nu(\{\theta_3\}) &= 1 - f_m^{-1}\left(\frac{1 - x\nu(\{\theta_2\}) - y\nu(\{\theta_3\})}{\nu(\{\theta_1\})}\right) \\ &= f_m^{-1}\left(\frac{\nu(\{\theta_2\})}{\nu(\{\theta_3\})}x + y\right)\nu(\{\theta_3\}). \end{aligned}$$

Defining $\lambda := \frac{\nu(\{\theta_2\})}{\nu(\{\theta_3\})}$, we conclude that for any $x, y \geq 0$ such that $\lambda x + y \leq \frac{1}{\nu(\{\theta_3\})}$, $f_m^{-1}(\lambda x + y) = \lambda f_m^{-1}(x) + f_m^{-1}(y)$. Lemma 1 shows that $f_m^{-1}(\cdot)$ is the identity function on $\left[0, \frac{1}{\nu(\{\theta_3\})}\right]$. Because $\theta_i, i = 1, 2, 3$ are arbitrarily chosen in $\mathcal{A}_\nu$, $f_m(\cdot)$ is the identity function in its effective domain. □

Appendix C

The Two-Element Case

In this chapter, we assume that $\Theta = \{\theta_1, \theta_2\}$ and Assumption 1-(ii) and (iii) hold.

C.1 Main Results in Chapter 1

Corollary 1 *For any fixed $m \geq 0$ and $n \geq 1$, consider two pairs of pseudo-likelihood and distortion, $(q_{m,n}(\theta|\mathbf{x}), g_m)$ and $(q'_{m,n}(\theta|\mathbf{x}), g'_m)$, that satisfy Assumption 1. Then, $(q_{m,n}(\theta|\mathbf{x}), g_m)$ and $(q'_{m,n}(\theta|\mathbf{x}), g'_m)$ lead to the same posterior density $\pi_{m,n}$ in the coherent inference model (1.2) for any prior density π and any sample sequence $\mathbf{x}$ if and only if (i) $g_m(x)/g_m(y) = g'_m(x)/g'_m(y)$ for any $x, y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$ and (ii) for any $\mathbf{x} \in \mathbb{X}^\infty$, there exists $C_{m,n}(\mathbf{x}) > 0$ such that $q_{m,n}(\theta|\mathbf{x}) = C_{m,n}(\mathbf{x})q'_{m,n}(\theta|\mathbf{x})$ for all $\theta \in \Theta$.*

Proof Because $\Theta = \{\theta_1, \theta_2\}$, the distribution of θ is simply determined by the odds of θ_1 in favor of θ_2 under this distribution. The posterior odds in the coherent inference model is

$$\mathcal{O}_{m,n}(\mathbf{x}) := \frac{\pi_{m,n}(\theta_1|\mathbf{x})}{\pi_{m,n}(\theta_1|\mathbf{x})} = \frac{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi(\theta_1))}{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi(\theta_2))}.$$

We first prove sufficiency. Denote $\mathcal{O}_{m,n}(\mathbf{x})$ and $\mathcal{O}'_{m,n}(\mathbf{x})$ as the posterior odds in

(1.2) with $(q_{m,n}(\theta|\mathbf{x}), g_m)$ and $(q'_{m,n}(\theta|\mathbf{x}), g'_m)$, respectively. Then,

$$\begin{aligned}\mathcal{O}_{m,n}(\mathbf{x}) &= \frac{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi(\theta_1))}{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi(\theta_2))} \\ &= \frac{C_{m,n}(\mathbf{x})q'_{m,n}(\theta_1|\mathbf{x})g'_m(\pi(\theta_1))}{C_{m,n}(\mathbf{x})q'_{m,n}(\theta_2|\mathbf{x})g'_m(\pi(\theta_2))} \\ &= \mathcal{O}'_{m,n}(\mathbf{x}).\end{aligned}$$

Next, we prove the necessity. For $\mathcal{O}_{m,n}(\mathbf{x})$ and $\mathcal{O}'_{m,n}(\mathbf{x})$ to be equal for any $\mathbf{x}$ and any π, we must have

$$\frac{g_m(\pi(\theta_1))}{g_m(\pi(\theta_2))} \Big/ \frac{g'_m(\pi(\theta_1))}{g'_m(\pi(\theta_2))} = \frac{q'_{m,n}(\theta_1|\mathbf{x})}{q'_{m,n}(\theta_2|\mathbf{x})} \Big/ \frac{q_{m,n}(\theta_1|\mathbf{x})}{q_{m,n}(\theta_1|\mathbf{x})}$$

for any π. Because fixing $\mathbf{x}$, $(\pi(\theta_1), \pi(\theta_2))$ can take any pair of values (x, y) such that $x, y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$, and the right-hand side does not depend on (x, y), we immediately conclude that

$$\frac{g_m(x)}{g_m(y)} \Big/ \frac{g'_m(x)}{g'_m(y)} = C, \quad \forall x > 0, y > 0 \text{ such that } x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$$

for some constant C. By setting $x = y$, we immediately conclude that $C = 1$. This further implies that

$$\frac{q'_{m,n}(\theta_1|\mathbf{x})}{q'_{m,n}(\theta_2|\mathbf{x})} \Big/ \frac{q_{m,n}(\theta_1|\mathbf{x})}{q_{m,n}(\theta_1|\mathbf{x})} = 1, \quad \forall \mathbf{x},$$

which is equivalent to condition (ii) in the Corollary. □

Corollary 2 *The coherent inference model (1.2) is processing consistent if and only if*

(i) for each $m \geq 1$, $g_m(x)/g_m(y) = x/y$ for any $x, y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$; and

(ii) for each $m \geq 1$ and $n \geq 1$ and any $\mathbf{x} \in \mathbb{X}^\infty$, there exists $C_{m,n}(\mathbf{x}) > 0$ such that (1.4) holds.

Proof Because $\Theta = \{\theta_1, \theta_2\}$, the distribution of θ is simply determined by the odds of θ_1 in favor of θ_2 under this distribution. The posterior odds in the coherent inference

model is

$$\mathcal{O}_{m,n}(\mathbf{x}) := \frac{\pi_{m,n}(\theta_1|\mathbf{x})}{\pi_{m,n}(\theta_1|\mathbf{x})} = \frac{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi(\theta_1))}{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi(\theta_2))}.$$

We first prove sufficiency. Denote $\mathcal{O}_{m,n}(\mathbf{x};\pi_0)$ and $\mathcal{O}_{0,m+n}(\mathbf{x};\pi_0)$, respectively as the posterior odds in (1.2) when processing $\mathbf{x}_{0,m}$ and $\mathbf{x}_{m,n}$ sequentially and when processing $\mathbf{x}_{0,m+n}$ as a group. Denote $\pi_{0,m}(\theta|\mathbf{x};\pi_0)$ as the posterior density obtained after processing $\mathbf{x}_{0,m}$ as a group. Then,

$$\begin{aligned}
\mathcal{O}_{m,n}(\mathbf{x};\pi_0) &= \frac{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi_{0,m}(\theta_1|\mathbf{x};\pi_0))}{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi_{0,m}(\theta_2|\mathbf{x};\pi_0))} \\
&= \frac{q_{m,n}(\theta_1|\mathbf{x})\pi_{0,m}(\theta_1|\mathbf{x};\pi_0)}{q_{m,n}(\theta_1|\mathbf{x})\pi_{0,m}(\theta_2|\mathbf{x};\pi_0)} \\
&= \frac{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})g_0(\pi_0(\theta_1))}{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})g_0(\pi_0(\theta_2))} \\
&= \frac{C_{m,n}(\mathbf{x})^{-1}q_{0,m+n}(\theta_1|\mathbf{x})g_0(\pi_0(\theta_1))}{C_{m,n}(\mathbf{x})^{-1}q_{0,m+n}(\theta_2|\mathbf{x})g_0(\pi_0(\theta_2))} \\
&= \mathcal{O}_{0,m+n}(\mathbf{x};\pi_0).
\end{aligned}$$

Next, we prove necessity. For $\mathcal{O}_{m,n}(\mathbf{x};\pi_0)$ and $\mathcal{O}_{0,m+n}(\mathbf{x};\pi_0)$ to be equal for any π_0, we must have

$$\frac{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi_{0,m}(\theta_1|\mathbf{x};\pi_0))}{q_{m,n}(\theta_1|\mathbf{x})g_m(\pi_{0,m}(\theta_2|\mathbf{x};\pi_0))} = \frac{q_{0,m+n}(\theta_1|\mathbf{x})g_0(\pi_0(\theta_1))}{q_{0,m+n}(\theta_2|\mathbf{x})g_0(\pi_0(\theta_2))}$$

for any π_0, which is the case if and only if

$$\frac{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})g_m(\pi(\theta_1))}{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})g_m(\pi(\theta_2))} = \frac{q_{0,m+n}(\theta_1|\mathbf{x})\pi(\theta_1)}{q_{0,m+n}(\theta_2|\mathbf{x})\pi(\theta_2)} \tag{C.1}$$

holds for any density π because (i)

$$\frac{\pi_{0,m}(\theta_1|\mathbf{x};\pi_0)}{\pi_{0,m}(\theta_2|\mathbf{x};\pi_0)} = \frac{q_{0,m}(\theta_1|\mathbf{x})g_0(\pi_0(\theta_1))}{q_{0,m}(\theta_2|\mathbf{x})g_0(\pi_0(\theta_2))}$$

and (ii) for any fixed $\mathbf{x}$ and any density π, we can find π_0 such that $\pi_{0,m}(\theta|\mathbf{x};\pi_0) = \pi(\theta), \forall\theta \in \Theta$. Note that (C.1) holds for any π if and only if

$$\frac{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})}{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})} \Big/ \frac{q_{0,m+n}(\theta_1|\mathbf{x})}{q_{0,m+n}(\theta_2|\mathbf{x})} = \frac{\pi(\theta_1)}{\pi(\theta_2)} \Big/ \frac{g_m(\pi(\theta_1))}{g_m(\pi(\theta_2))}$$

for any density π. Because fixing $\mathbf{x}$, $\big(\pi(\theta_1), \pi(\theta_2)\big)$ can take pair of values (x, y) for any x, y such that $x > 0$, $y > 0$, and $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$, we immediately conclude that

$$\frac{x}{y} \Big/ \frac{g_m(x)}{g_m(y)} = C, \quad \forall x > 0, y > 0 \text{ such that } x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1.$$

for some constant C. By setting $x = y$, we further conclude $C = 1$, thus obtaining the necessary condition (i). We further conclude that

$$\frac{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})}{q_{m,n}(\theta_1|\mathbf{x})q_{0,m}(\theta_1|\mathbf{x})} \Big/ \frac{q_{0,m+n}(\theta_1|\mathbf{x})}{q_{0,m+n}(\theta_2|\mathbf{x})} = 1$$

for any $\mathbf{x}$, and this is the case if and only if the necessary condition (ii) holds. □

Corollary 3 *Denote the one-step pseudo-likelihood $q_{m,1}(\theta|\mathbf{x})$ in the coherent inference model as $q_{m,1}(\theta|\mathbf{x}_{0,m}, x_{m+1})$.*

(i) The coherent inference model (1.2) is false Bayesian in the strong sense if and only if it is processing consistent, $g_0(x)/g_0(y) = x/y$ for any $x, y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$, and, for each $m \geq 0$, there exists a measurable function $\varphi_m(\mathbf{x}_{0,m}, x) > 0, \mathbf{x}_{0,m} \in \mathbb{X}^m, x \in \mathbb{X}$ such that (1.10) holds.

(ii) The coherent inference model (1.2) is false Bayesian in the weak sense if and only if it is processing consistent and, for each $m \geq 1$, there exists a measurable function $\varphi_m(\mathbf{x}_{0,m}, x) > 0, \mathbf{x}_{0,m} \in \mathbb{X}^m, x \in \mathbb{X}$ such that (1.10) holds.

Proof The proof of this Corollary follows from the proof of Corollary 2 and the proof of Theorem 2. □

C.2 Generic Inference Model

Next, we study the issue of when the generic inference model B.1 is coherent in the two-element case.

Corollary 4 *The generic inference model B.1 is coherent, i.e., for each $m \geq 0$, $n \geq 1$, any $\mathbf{x} \in \mathbb{X}^\infty$, and any $\pi(\cdot) \in \mathcal{P}(\Theta)$, $\pi_{m,n}$ defined as in (B.1) is a probability density on Θ, if and only if for each $m \geq 0$ and $x > 0$, $y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$, we have $f_m^{-1}(x)\nu(\{\theta_1\}) + f_m^{-1}(y)\nu(\{\theta_2\}) = 1$.*

Proof Following the proof of Theorem 4, we can show that for any density p, we can find a prior density π such that (B.3) holds. Consequently, model B.1 is coherent if and only if $\int_\Theta f_m^{-1}(p(\theta))\nu(d\theta) = 1$ for any density p. Because $\Theta = \{\theta_1, \theta_2\}$, the proof completes. □

Let us comment that the condition that $f_m^{-1}(x)\nu(\{\theta_1\}) + f_m^{-1}(y)\nu(\{\theta_2\}) = 1$ for any $x > 0$, $y > 0$ such that $x\nu(\{\theta_1\}) + y\nu(\{\theta_2\}) = 1$ does not imply that f_m is the identity function. For example, suppose $\nu(\{\theta_1\}) = \nu(\{\theta_2\}) = 1$, the above condition is satisfied by f on $[0, 1]$ such that $f(x) = f(1 - x), x \in [0, 1/2]$.

C.3 Main Results in Chapter 2

Finally, we comment that the necessary condition (i) in Corollary 2 stipulates that the distortion g_m does not affect the prior odds of θ_1 in favor of θ_2. Thus, even in the case in which $\Theta = \{\theta_1, \theta_2\}$, we can assume without loss of generality that in a processing-consistent, coherent inference model, g_m is a linear function for any $m \geq 1$, i.e., the necessary condition (i) in Theorem 1 holds. Consequently, all the results in Section 2.2 still hold in this case.