Frequentist size of Bayesian inequality tests

David M. Kaplan* Longhao Zhuo†

February 12, 2018

Abstract

Bayesian and frequentist criteria are fundamentally different, but often posterior and sampling distributions are asymptotically equivalent (e.g., Gaussian). For the corresponding limit experiment, we characterize the frequentist size of a certain Bayesian hypothesis test of (possibly nonlinear) inequalities. If the null hypothesis is that the (possibly infinite-dimensional) parameter lies in a certain half-space, then the Bayesian test's size is α; if the null hypothesis is a subset of a half-space, then size is above α (sometimes strictly); and in other cases, size may be above, below, or equal to α. Two examples illustrate our results: testing stochastic dominance and testing curvature of a translog cost function.

JEL classification: C11, C12

Keywords: Bernstein–von Mises theorem, limit experiment, nonstandard inference, stochastic dominance, translog

1 Introduction

Although Bayesian and frequentist properties are fundamentally different, in many cases we can (approximately) achieve both. In other cases, the Bayesian and frequentist summaries of the data differ greatly, and practitioners must carefully consider which to prefer. We provide results on the role of null hypothesis "shape" in determining such differences. We hope to alert practitioners to situations prone to large differences and to foster understanding of why such differences can be large.

*Corresponding author. Email: kaplandm@missouri.edu. Mail: Department of Economics, University of Missouri, 118 Professional Bldg, 909 University Ave, Columbia, MO 65211-6040, United States. Many thanks to Tim Armstrong, Jim Berger, Jeremy Fox, Patrik Guggenberger, Zack Miller, Stephen Montgomery-Smith, Andriy Norets, Iosif Pinelis, and Andres Santos for helpful discussion, comments, examples, and references. Thanks also to anonymous reviewers for detailed comments, references, and the generalization beyond normality.

†Bank of America. Email: longhao.zhuo@gmail.com.

Economic theory prominently features inequality restrictions, often nonlinear.[1] For example, inequality of CDFs characterizes first-order stochastic dominance (SD1), an important concept for welfare analysis. SD1 conclusions from frequentist tests (that have exact asymptotic size) and Bayesian tests may differ greatly, and the direction of the difference partly depends on whether the null hypothesis is dominance or non-dominance. Inequalities also characterize second-order stochastic dominance, which is also used for portfolio comparison in finance. An example of nonlinear inequalities is curvature constraints on production, cost, indirect utility, and other functions. Such constraints usually result from optimization, like utility or profit maximization. We illustrate our results with SD1 and cost function curvature in Section 4.

Further motivation for studying Bayesian–frequentist differences is that deriving frequentist tests for general nonlinear inequalities is notoriously difficult; e.g., see Wolak (1991). In contrast, it is (relatively) simple to compute the posterior probability that the parameter satisfies certain nonlinear inequalites, by computing the proportion of draws from the parameter's posterior in which the constraints are satisfied. Perhaps especially in the absence of a feasible frequentist method, it is helpful to understand if the Bayesian test's size differs greatly from the nominal α.

Statistically, we consider cases where the sampling distribution of an estimator is asymptotically normal, while the asymptotic posterior is also normal, with the same covariance matrix (i.e., a Bernstein–von Mises theorem holds). More generally, any symmetric distribution suffices, and only a certain functional must have the equivalent sampling and posterior distributions, making it easier to treat infinite-dimensional parameters. We examine the corresponding limit experiment, with the general null hypothesis that the parameter belongs to a specified subspace of the parameter space.

In this limit experiment, to quantify Bayesian–frequentist differences, we characterize the frequentist size of a particular Bayesian test. This test rejects the null hypothesis when the posterior probability is below α. In addition to being intuitive and practically salient, there are decision-theoretic reasons to examine this test, as detailed in Section 2.1. Although size gives no insight into admissibility, it captures a practical difference between reporting Bayesian and frequentist inferences. Our results provide insight even if posterior probabilities are reported instead of a hypothesis test; e.g., they say in which cases the posterior probability of H_0 may be below α with greater than α (frequentist) probability even if H_0 is true.

[1]Nonlinear inequalities also come from other sources. For example, $H_0\colon \theta_1\theta_2 \geq 0$ can be used to test stability of the sign of a parameter over time (or geography), or whether a treatment attenuates the effect of another regressor; see Kaplan (2015) for details.

Specifically, we describe how the Bayesian test's size depends on the shape of the null hypothesis, H_0. By "the shape of H_0," we mean the shape of the parameter subspace where H_0 is satisfied (which may be infinite-dimensional). If H_0 is a half-space, then the Bayesian test has size α exactly. If H_0 is strictly smaller than a half-space (in a certain sense), then the Bayesian test's size is strictly above α. If H_0 is not contained within a half-space (i.e., does not have a supporting hyperplane), then the Bayesian test's size may be above, equal to, or below α. An immediate corollary of these results is that the Bayesian test has exact size when testing a single linear inequality constraint, whereas it is size-distorted when testing two or more linear inequality constraints.

Our results beg the question: if inferences on H_0 can disagree while credible and confidence sets coincide, why not simply report the credible or confidence set?[2] If interest is primarily in the parameters themselves, then reporting a credible or confidence set may indeed be better than a posterior or p-value for H_0. However, sometimes interest is in testing implications of economic theory or in specification testing. Other times, inequalities provide economically relevant summaries of a high-dimensional parameter, like whether a certain income distribution stochastically dominates another.

Literature Many papers compare Bayesian and frequentist inference, in a variety of setups. Here, we highlight examples of different types of conclusions: sometimes frequentist inference is more conservative, sometimes Bayesian, sometimes neither.

Some of the literature documents cases where frequentist inference is "too conservative" from a Bayesian perspective. For testing linear inequality constraints of the form $H_0\colon \boldsymbol{\theta} \geq \mathbf{0}$ with $\boldsymbol{\theta} \in \mathbb{R}^d$, Kline (2011, §4) provides examples showing frequentist testing to be more conservative (e.g., his Figure 1), especially as the dimension d grows; his examples are consistent with our general theoretical results. As another example, under set identification, asymptotically, frequentist confidence sets for the true parameter (Imbens and Manski, 2004; Stoye, 2009) are strictly larger than the estimated identified set, whereas Bayesian credible sets are strictly smaller, as shown by Moon and Schorfheide (2012, Cor. 1).[3] Our setup is not directly comparable to theirs since a Bayesian credible set cannot be inverted into a test. For testing the null of a unit root in autoregression, Sims and Uhlig (1991) say frequentist tests "accept the null more easily" (p. 1592), and they determine (sample-dependent) priors

[2]Berger (2003) also notes this possible simultaneous agreement on credible/confidence sets but disagreement on testing. However, he writes, "The disagreement occurs primarily when testing a 'precise' hypothesis" (p. 2), whereas we find disagreements even with inequality hypotheses. Also, Casella and Berger (1987b, p. 344) opine, "Interval estimation is, in our opinion, superior to point null hypothesis testing," although they do not mention composite null hypotheses like in this paper.

[3]There seems to be a typo in the statement of Corollary 1(ii), switching the frequentist and Bayesian sets from their correct places seen in the Supplemental Material proof.

that equate p-values and posterior probabilities. Our setup excludes unit root testing since the parameter may be on the boundary (and there is no Bernstein–von Mises theorem).

Other papers document cases where frequentist inference is "too aggressive" from a Bayesian perspective. Perhaps most famously, in Lindley's (1957) paradox, the frequentist test rejects while the Bayesian test does not. Berger and Sellke (1987) make a similar argument. In both cases, as noted by Casella and Berger (1987b), the results follow primarily from having a large prior probability on a point (or "small interval") null hypothesis, specifically $\mathrm{P}(H_0) = 1/2$. Arguing that $\mathrm{P}(H_0) = 1/2$ is "objective," Berger and Sellke (1987, p. 113) consider even $\mathrm{P}(H_0) = 0.15$ to be "blatant bias toward H_1." Casella and Berger (1987b) disagree, saying $\mathrm{P}(H_0) = 1/2$ is "much larger than is reasonable for most problems" (p. 344).

In yet other cases, Bayesian and frequentist inferences are similar or even identical. Casella and Berger (1987a) compare Bayesian and frequentist one-sided testing of a location parameter, given a single draw of X from an otherwise fully known density. They compare the p-value, $p(x)$, to the infimum of the posterior $\mathrm{P}(H_0 \mid x)$ over various classes of priors. In many cases, the infimum is attained by the improper prior of Lebesgue measure on $(-\infty, \infty)$ and equals $p(x)$ (p. 109). Berger, Brown, and Wolpert (1994) establish an equivalence of Bayesian and *conditional* frequentist testing of a simple null hypothesis against a simple alternative. Goutis, Casella, and Wells (1996) consider jointly testing multiple one-sided hypotheses. In a single-draw Gaussian shift experiment (similar to this paper), further assuming all components of the vector $\mathbf{X}$ are mutually independent, they consider the Bayesian posterior on H_0 when the (improper, uninformative) prior is adjusted to have $\mathrm{P}(H_0) = 1/2$. In this case, the posterior is proportional to one of the frequentist p-values they consider, but it is (weakly) smaller. This complements our setting where we impose neither independence nor $\mathrm{P}(H_0) = 1/2$, and we do not restrict the shape of the null hypothesis subspace.

Paper structure and notation Section 2 presents the setup and assumptions. Section 3 contains our main results and discussion. Section 4 illustrates our results with stochastic dominance and cost function curvature examples. Appendix A contains proofs. Appendices B–D contain details on testing equality of parameters' signs, translog cost functions, and infinite-dimensional Bernstein–von Mises theorems, respectively. Acronyms used include those for cumulative distribution function (CDF), data generating process (DGP), negative semidefinite (NSD), posterior expected loss (PEL), probability density function (PDF), rejection probability (RP), and first-order stochastic dominance (SD1). Notationally, $\subseteq$ is subset and $\subset$ is proper subset; scalars, (column) vectors, and matrices are respectively formatted as X, $\mathbf{X}$, and $\underline{\mathbf{X}}$; $0(\cdot)$ denotes the zero function, i.e., $0(t) = 0$ for all t.

2 Setup and assumptions

In Section 2.1, a specific Bayesian hypothesis test is described along with the decision-theoretic context. In Section 2.2, the assumptions used for the results in Section 3 are presented and discussed. Section 2.3 contains addition details and references about the Bernstein–von Mises theorem.

2.1 The Bayesian test and decision-theoretic context

The Bayesian test rejects the null hypothesis when its posterior probability is below α.

Method 1 (Bayesian test)**.** Reject H_0 if $\mathrm{P}(H_0 \mid \mathbf{X}) \leq \alpha$; otherwise, accept H_0.

In addition to seeming intuitive, Method 1 is a generalized Bayes decision rule that minimizes posterior expected loss (PEL) for the loss function taking value $1-\alpha$ for type I error, α for type II error, and zero otherwise. To see this, let $\mathrm{P}(\cdot \mid \mathbf{X})$ denote the posterior probability given observed data $\mathbf{X}$. The PEL for the decision to reject H_0 is $(1-\alpha)\,\mathrm{P}(H_0 \mid \mathbf{X})$, i.e., the type I error loss times the posterior probability that rejecting H_0 is a type I error. Similarly, the PEL of accepting H_0 is $\alpha[1-\mathrm{P}(H_0 \mid \mathbf{X})]$, the type II error loss times the probability that accepting H_0 is a type II error. PEL is thus minimized by rejecting H_0 if $\mathrm{P}(H_0 \mid \mathbf{X}) \leq \alpha$ and accepting H_0 otherwise.

Our results compare the frequentist size of the Bayesian test in Method 1 to α (instead of some other value) for practical and decision-theoretic reasons. Practically, the Bayesian test can be seen as treating the posterior as a (frequentist) p-value; we want to know if this is valid, similar to Casella and Berger (1987a). Decision-theoretically, the same loss function used to compute PEL could be used to compute a minimax risk decision rule that is closely related to a frequentist hypothesis test. Specifically, if an unbiased frequentist test with size α exists, then it is the minimax risk decision rule. Even without unbiasedness, this is approximately true given conventional values of α (below).

The minimax risk decision rule can be characterized and compared to an unbiased frequentist test. Let $\boldsymbol{\theta} \in \Theta$, $H_0 \colon \boldsymbol{\theta} \in \Theta_0 \subset \Theta$, $H_1 \colon \boldsymbol{\theta} \notin \Theta_0$. "Minimax risk" minimizes

$$\begin{aligned} &\max\Big\{(1-\alpha) \sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}),\ \alpha \sup_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{accept})\Big\} \\ &= \max\Big\{(1-\alpha) \sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}),\ \alpha[1 - \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})]\Big\}, \end{aligned} \tag{1}$$

where $\mathrm{P}_{\boldsymbol{\theta}}(\cdot)$ is probability under $\boldsymbol{\theta}$. Consider a test with exact size $\gamma_0 = \sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})$. If the test is unbiased, then $\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \leq \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})$. If the power function is

continuous in $\boldsymbol{\theta}$, then $\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \geq \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})$. Thus,

$$\gamma_0 = \overbrace{\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \leq \underbrace{\inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})}_{\text{continuity}} \leq \sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})}^{\text{unbiasedness}} = \gamma_0,$$

and (1) becomes $\max\{(1-\alpha)\gamma_0, \alpha(1-\gamma_0)\}$. The minimum value $\alpha(1-\alpha)$ is attained with $\gamma_0 = \alpha$; if $\gamma_0 < \alpha$, then $\alpha(1-\gamma_0) > \alpha(1-\alpha)$, and if $\gamma_0 > \alpha$, then $(1-\alpha)\gamma_0 > \alpha(1-\alpha)$. Without unbiasedness, $\inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \leq \gamma_0$, so $\alpha[1 - \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})] \geq \alpha(1-\gamma_0)$, weakly increasing the minimax risk. Thus, assuming continuity of the power function, the minimax risk decision rule is an unbiased frequentist hypothesis test with exact size α. See also Lehmann and Romano (2005, Problem 1.10) on unbiased tests as minimax risk decision rules.

Without unbiasedness, the minimax-risk-optimal size of a given test is above α, but the magnitude of the difference is very small for conventional α. As a function of γ_0, let $\gamma_1(\gamma_0) \equiv \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})$, so (1) is $\max\{(1-\alpha)\gamma_0, \alpha(1-\gamma_1(\gamma_0))\}$. Continuity restricts $\gamma_1(\gamma_0) \leq \gamma_0$, but we now drop the unbiasedness restriction $\gamma_1(\gamma_0) \geq \gamma_0$. In the extreme, $\gamma_1(\gamma_0) = 0$ for all γ_0, and the maximum risk is α for any test with $\gamma_0 \leq \alpha/(1-\alpha)$, while maximum risk is larger than α if $\gamma_0 > \alpha/(1-\alpha)$. If instead $\gamma_1(\gamma_0)$ is strictly increasing in γ_0 but $\gamma_1(\alpha) < \alpha$, then minimax risk is achieved at some $\gamma_0 \in (\alpha, \alpha/(1-\alpha))$. For example, rounding to two significant digits, if $\alpha = 0.05$, then $\gamma_0 \in (0.050, 0.053)$, or if $\alpha = 0.1$, then $\gamma_0 \in (0.10, 0.11)$. Such small divergence of γ_0 from α is almost imperceptible in practice.

Ideally, a single decision rule minimizes both the maximum risk in (1) and the PEL. However, if the Bayesian test's size is significantly above or below α, this is not possible. In such cases, it may help to use both Bayesian and frequentist inference and to carefully consider the differences in optimality criteria.

2.2 Assumptions

Assumption A1 states conditions on the sampling and posterior distributions of a functional $\phi(\cdot)$ in the (limit) experiment we consider. As usual, the sampling distribution treats the (functional of the) data $\phi(\mathbf{X})$ as random and conditions on the parameter $\boldsymbol{\theta}$, whereas the posterior treats the (functional of the) parameter $\phi(\boldsymbol{\theta})$ as random and conditions on the data $\mathbf{X}$.

Assumption A1. Let $F(\cdot)$ be a continuous CDF with support $\mathbb{R}$ and symmetry $F(-x) = 1 - F(x)$. Let the (lone) observation $\mathbf{X}$ and the parameter $\boldsymbol{\theta}$ both belong to a Banach space of possibly infinite dimension. Let $\phi(\cdot)$ denote a continuous linear functional, with sampling

distribution $\phi(\mathbf{X}) - \phi(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \sim F$ and posterior $\phi(\boldsymbol{\theta}) - \phi(\mathbf{X}) \mid \mathbf{X} \sim F$.

Assumption A1 can be interpreted as a limit experiment where $\boldsymbol{\theta}$ is a local mean parameter. The limit distribution F is often $\mathrm{N}(0, \sigma^2)$, satisfying the continuity, support, and symmetry conditions. For example, consider a simple asymptotic setup leading to scalar X and θ with $\phi(X) = X$ and $\phi(\theta) = \theta$. If $Y_{ni} \overset{iid}{\sim} \mathrm{N}(\mu_n, 1)$, $i = 1, \ldots, n$, and $\sqrt{n}\mu_n \to \theta$, then $\sqrt{n}\bar{Y}_n = n^{-1/2} \sum_{i=1}^n Y_{ni} \xrightarrow{d} \mathrm{N}(\theta, 1)$; more generally, if $Y_{ni} \overset{iid}{\sim} \mathrm{N}(m + \mu_n, \sigma^2)$ and $\sqrt{n}\mu_n \to \theta$, then $\sqrt{n}(\bar{Y}_n - m)/\hat{\sigma} \xrightarrow{d} \mathrm{N}(\theta, 1)$ for any consistent estimator $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. This type of result holds for a wide variety of models, estimators, and sampling assumptions; it is most commonly used for local power analysis but has been used for purposes like ours in papers like Andrews and Soares (2010, eqn. (4.2)). Since θ is the *local* mean parameter, assuming $\theta \in \mathbb{R}$ does not require that $\mathbb{R}$ is the original parameter space (e.g., the space for $m + \mu_n$ in the example), but it does exclude boundary points. Results for posterior asymptotic normality date back to Laplace (1820), as cited in Lehmann and Casella (1998, §6.10, p. 515).

Seeing A1 as a limit experiment, implicitly the prior has no asymptotic effect on the posterior, as in a Bernstein–von Mises theorem, which is any result establishing the same asymptotic sampling distribution (of the estimator, scaled and centered at the true value) as the asymptotic posterior distribution (of the parameter, scaled and centered at the estimator). Bernstein–von Mises theorems are discussed further in Section 2.3.

For our purpose of approximating the finite-sample frequentist size of a Bayesian test, considering a fixed DGP and drifting centering parameter can be just as helpful as considering a fixed centering parameter and drifting DGP. This allows the relevant Bernstein–von Mises theorem to hold only for fixed DGPs (which is usually how such results are stated), instead of the stricter requirement of holding for drifting DGPs. For example, in $\mathbb{R}^d$,

$$\mathbf{X}_n = \sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{0,n}) = \overbrace{\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})}^{\xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}})} + \overbrace{\sqrt{n}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0,n})}^{\to \boldsymbol{\theta}} \xrightarrow{d} \overbrace{\mathbf{X} \sim \mathrm{N}(\boldsymbol{\theta}, \underline{\boldsymbol{\Sigma}})}^{\text{limit experiment}}, \tag{2}$$

where $\boldsymbol{\mu}$ is the true parameter value, $\hat{\boldsymbol{\mu}}$ is a $\sqrt{n}$-normal estimator (as seen), $\underline{\boldsymbol{\Sigma}}$ is the known or consistently estimable asymptotic covariance matrix, and $\mathbf{X}_n$ is a statistic based on $\hat{\boldsymbol{\mu}}$ and centered at the deterministic sequence $\boldsymbol{\mu}_{0,n}$ that satisfies $\sqrt{n}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0,n}) \to \boldsymbol{\theta}$, the local mean parameter. This does not have a literal meaning like "we must change $\boldsymbol{\mu}_0$ if our sample size increases," just as a drifting DGP does not mean literally that "the population distribution changes as we collect more data"; rather, it is simply a way to capture the idea of $\boldsymbol{\mu}_0$ being "close to" the true $\boldsymbol{\mu}$ in the asymptotics. For the posterior, letting $\boldsymbol{\theta}_n = \sqrt{n}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0,n})$ and

again $\mathbf{X}_n = \sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{0,n})$,

$$\boldsymbol{\theta}_n - \mathbf{X}_n = \overbrace{\sqrt{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})}^{\text{Bernstein–von Mises theorem}}. \tag{3}$$

The common case of a Gaussian limit (plus a Bernstein–von Mises theorem) generally satisfies A1. When the Banach space in A1 is $\mathbb{R}^d$, continuous linear functionals are simply linear combinations $\phi(\mathbf{X}) = \mathbf{c}'\mathbf{X}$ for some constant vector $\mathbf{c} \in \mathbb{R}^d$. With multivariate Gaussian $\mathbf{X}$ and $\boldsymbol{\theta}$, linear combinations are (scalar) Gaussian random variables, satisfying the assumption. More generally, including infinite-dimensional spaces, if $X(\cdot)$ is a Gaussian process in some Banach space and $\phi(\cdot)$ belongs to the dual of that space, then $\phi\big(X(\cdot)\big)$ is a scalar Gaussian random variable; e.g., see Definition 2.2.1(ii) in Bogachev (1998, p. 42) and van der Vaart and Wellner (1996, pp. 376–377).

2.3 Bernstein–von Mises theorems

Seeing Assumption A1 as a limit experiment, implicitly the prior has no asymptotic effect on the posterior, as in the Bernstein–von Mises theorem. In the limit experiment, this is equivalent to using an improper uninformative prior. For example, with sampling distribution $X \mid \theta \sim \mathrm{N}(\theta, 1)$ and prior $\theta \sim \mathrm{N}(m, \tau^2)$, the posterior is

$$\theta \mid X \sim \mathrm{N}\left(\frac{\tau^2 X + m}{\tau^2 + 1}, \frac{\tau^2}{\tau^2 + 1}\right),$$

and taking $\tau^2 \to \infty$ yields the posterior $\theta \mid X \sim \mathrm{N}(X, 1)$, satisfying A1. The improper prior is fine with inequality testing because only posterior probabilities are used (see Method 1), unlike with point null hypothesis testing based on Bayes factors that involve a ratio of prior probabilities (e.g., Bayarri, Berger, Forte, and García-Donato, 2012).

There are versions of the Bernstein–von Mises theorem for parametric, semiparametric, and nonparametric models. The first two are discussed below. The third is not necessary (though it can be sufficient) for Assumption A1 since the functional ϕ is finite-dimensional, so discussion is relegated to Appendix D.

Parametric versions of the Bernstein–von Mises theorem are the oldest and can be found in textbooks. They differ in regularity conditions and in how to quantify the distance between two distributions, but they share the requirement that the prior density be both continuous and positive at the true value. For example, see Theorem 10.1 in van der Vaart (1998, §10.2) and Theorems 20.1–3 in DasGupta (2008, §20.2), where Theorem 20.3 allows non-iid sampling.

General semiparametric versions of the Bernstein–von Mises theorem have been established, too. For example, see Shen (2002), Bickel and Kleijn (2012), and Castillo and Rousseau (2015), who allow non-iid sampling. There are also earlier papers for specific models like GMM and quantile regression; see Hahn (1997, Thm. G and footnote 13), Kwan (1999, Thm. 2), Kim (2002, Prop. 1), Lancaster (2003, Ex. 2), Schennach (2005, p. 36), Sims (2010, Sec. III.2), and Norets (2015, Thm. 1), among others.

3 Results and discussion

Theorem 1 contains this paper's main results. Discussion and special cases follow.

Theorem 1. *Let Assumption A1 hold. Consider testing* $H_0\colon \boldsymbol{\theta} \in \Theta_0$ *against* $H_1\colon \boldsymbol{\theta} \notin \Theta_0$ *with the Bayesian test in Method 1, where* Θ_0 *is a subset of the Banach space in A1.*

(i) *If there exists a* $\phi(\cdot)$ *satisfying A1 and value* $c_0 \in \mathbb{R}$ *such that* $\Theta_0 = \{\boldsymbol{\theta} : \phi(\boldsymbol{\theta}) \leq c_0\}$ *(i.e.,* Θ_0 *is a half-space), then the Bayesian test's size is* α*, and its type I error rate is* α *when* $\phi(\boldsymbol{\theta}) = c_0$*.*

(ii) *If there exists a* $\phi(\cdot)$ *satisfying A1 and value* $c_0 \in \mathbb{R}$ *such that* $\Theta_0 \subseteq \{\boldsymbol{\theta} : \phi(\boldsymbol{\theta}) \leq c_0\}$ *with* $c_0 \in \phi(\overline{\Theta}_0)$ *(where* $\overline{\Theta}_0$ *denotes the closure), then the Bayesian test's size is* α *or greater.*

(iii) *Continuing from Theorem 1(ii), assume there exists a* $\phi_2(\cdot)$ *satisfying A1 with distribution* F_2 *over* $\mathbb{R}^d$ *and satisfying the properties below;* $\phi(\cdot) = \phi_2(\cdot)$ *if* $d = 1$*. Assume the* $\phi(\cdot)$ *from Theorem 1(ii) may be written as* $\phi(\cdot) = \phi_3(\phi_2(\cdot))$ *for some* $\phi_3(\cdot)$*. Assume there exists (in* $\mathbb{R}^d$*) a set* $\Phi_2 \equiv \{\phi_2(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_0\} \subset \{\phi_2(\boldsymbol{\theta}) : \phi(\boldsymbol{\theta}) \leq c_0\}$*. If* $\boldsymbol{\theta} \in \mathbb{R}^d$*, then one may set* $\phi_2(\boldsymbol{\theta}) = \boldsymbol{\theta}$*,* $\phi_3(\cdot) = \phi(\cdot)$*, and* $\Phi_2 = \Theta_0$*. Further assume (a) the set* $\Delta \equiv \{\mathbf{p} : \mathbf{p} \in \mathbb{R}^d, \phi_3(\mathbf{p}) \leq c_0, \mathbf{p} \notin \Phi_2\}$ *has positive Lebesgue measure, (b)* F_2 *has a strictly positive PDF over* $\mathbb{R}^d$*, and (c)* $\mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X}))$ *is continuous in* $\phi_2(\mathbf{X})$*. Then, the Bayesian test's rejection probability is strictly above* α *when* $\phi(\boldsymbol{\theta}) = c_0$*, and its size is strictly above* α*.*

(iv) *If, contrary to Theorem 1(ii), there do not exist any* $\phi(\cdot)$ *and* c_0 *such that* $\Theta_0 \subseteq \{\boldsymbol{\theta} : \phi(\boldsymbol{\theta}) \leq c_0\}$ *(i.e.,* Θ_0 *is not a subset of a half-space), then the Bayesian test's size may be greater than, equal to, or less than* α*, and it may depend on the sampling/posterior distribution.*

3.1 Discussion of results

Intuitively, Theorem 1(i) holds by the symmetry of F. It holds for higher-dimensional parameters by finding a single inequality on a scalar-valued functional that is both necessary and sufficient for H_0. Theorem 1(ii) and Theorem 1(iii) hold because when parts of the half-space are carved away to make Θ_0 smaller, the posterior probability of H_0 (at any $\mathbf{X}$) becomes smaller, making the Bayesian test more likely to reject. For infinite-dimensional parameters, this logic is essentially applied to a test of a finite-dimensional necessary condition for H_0.

Theorem 1(ii) has a geometric interpretation: if Θ_0 has a supporting hyperplane, then the Bayesian test's size is at least α. Theorem 1(iii) gives sufficient conditions for the Bayesian test's RP to be strictly above α for any $\boldsymbol{\theta}$ that is a support point of Θ_0.

Theorem 1(iii) is partly a result of the prior $\mathrm{P}(H_0)$ being "small" when Θ_0 is small. That is, the prior over the *parameter* is always the same (regardless of Θ_0), so the implicit prior $\mathrm{P}(H_0)$ shrinks when Θ_0 shrinks. (Technically, the limit experiment's prior is an improper constant prior, so $\mathrm{P}(H_0)$ is not well-defined, but the qualitative idea remains.) Unless Θ_0 is a half-space, this differs from Berger and Sellke (1987) and others who only consider "objective" priors with $\mathrm{P}(H_0) = 0.5$. Whether placing a prior on the null (like $\mathrm{P}(H_0) = 0.5$) or on the parameter is more appropriate depends on the empirical setting; e.g., do we have prior reason to suspect SD1? Often it is easier computationally not to set a specific $\mathrm{P}(H_0)$; for example, one may use the same Bayesian bootstrap posterior advocated by Chamberlain and Imbens (2003) to compute probabilities of many different hypotheses. However, for hypotheses like SD1, this can lead to a very small "$\mathrm{P}(H_0)$" and consequently very large rejection probabilities (and size distortion).

Although the implicit $\mathrm{P}(H_0)$ partially explains Theorem 1(iii), the shape of Θ_0 still plays an important role. For example, in $\mathbb{R}^2$, let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $H_0 \colon \theta_1\theta_2 \geq 0$, so Θ_0 comprises the first and third quadrants (and thus is not contained in any half-space). This Θ_0 is the same "size" as the half-space $\{\boldsymbol{\theta} : \theta_1 \geq 0\}$. However, with bivariate normal sampling and posterior distributions, Theorem 1(i) implies the Bayesian test of the half-space has exact size α, whereas the size of the Bayesian test of $H_0 \colon \theta_1\theta_2 \geq 0$ may be either strictly above or equal to α, depending on the correlation. For example, let $\mathbf{X} = (X_1, X_2)$ have a bivariate normal sampling distribution with $\mathrm{Corr}(X_1, X_2) = -1$. Then the test is equivalent to a scalar test where H_0 is a finite, closed interval, in which case the Bayesian test's size strictly exceeds α by Theorem 1(iii). The same holds for other negative correlations, but size decreases to α as the correlation gets closer to zero; see Appendix B as well as Kaplan (2015, §3.2) for details.

Theorem 1(ii) and Theorem 1(iii) can apply to $\Theta_0 = \{\boldsymbol{\theta} : g(\boldsymbol{\theta}) \leq g_0\}$ when $g(\cdot)$ is directionally differentiable, as in Fang and Santos (2015) and others. For example, all the examples in Section 4 of Fang and Santos (2015) concern $\boldsymbol{\theta}$ belonging to a convex set. They

provide a frequentist resampling scheme that is consistent and corresponding hypothesis tests that control asymptotic size. Thus, in cases like this where Theorem 1(iii) applies, not only is the Bayesian test's (asymptotic) size strictly above α, but it is also strictly above the size of an available frequentist test.

The condition of Θ_0 being a subset of a half-space holds for many economic examples. The examples of stochastic dominance and curvature constraints (on functions describing cost, production, etc.) are explored in Section 4. As noted above, the examples in Section 4 of Fang and Santos (2015) also satisfy this condition, "encompass[ing] tests of moment inequalities, shape restrictions, and the validity of random utility models" (§4, p. 25), the latter referring to Stoye and Kitamura (2013). The general moment inequality null hypothesis $H_0\colon \mathbb{E}[\mathbf{W}] \leq \mathbf{0}$ with $\mathbf{W} \in \mathbb{R}^d$ as in equation (58) of Fang and Santos (2015) includes special cases like testing discrete (or ordinal) first-order stochastic dominance and testing the performance of financial trading rules against a benchmark as in equations (4) and (5) of Sullivan, Timmermann, and White (1999, 2001), among many other applications. Wolak (1989, §6) considers shape/monotonicity restrictions corresponding to $H_0\colon (-\beta_1, \beta_2, \beta_3) \geq \mathbf{0}$ in his model of residential electricity demand. Example 4.2 of Fang and Santos (2015) considers shape restrictions on the infinite-dimensional regression quantile process, e.g., the restriction that the coefficient on the regressor of interest is monotonic in the quantile index. In finance, Patton and Timmermann (2010) test asset return monotonicity (of various types) using null hypotheses that are all strict subsets of half-spaces in $\mathbb{R}^d$, such as (with altered notation) $H_0\colon \boldsymbol{\theta} \leq \mathbf{0}$ in their (5), $H_0\colon \theta_{ij} \leq \theta_{i-1,j}, \theta_{ij} \leq \theta_{i,j-1}$ for all i, j in their (13), and $H_0\colon \theta_{jN} \leq \theta_{jN-1} \leq \cdots \leq \theta_{j0}$ in their (19).

Whether Θ_0 is treated as H_0 or H_1 affects the size of the Bayesian test: if Θ_0 satisfies Theorem 1(iii), then its complement does not. Combining Theorem 1(iii) and Theorem 1(iv), the Bayesian test of $H_0\colon \boldsymbol{\theta} \in \Theta_0$ may have size strictly above α while the Bayesian test of $H_0\colon \boldsymbol{\theta} \in \Theta_0^{\complement}$ has size strictly below or equal to α. This is indeed the case for SD1 (in Section 4.1).

Many nonlinear inequalities could be recast as linear inequalities, but at the expense of additional approximation error. For example, for $g\colon \mathbb{R}^d \mapsto \mathbb{R}$ and underlying (non-local) parameter $\boldsymbol{\mu} \in \mathbb{R}^d$, the nonlinear $H_0\colon g(\boldsymbol{\mu}) \leq 0$ could be written as $H_0\colon \beta \leq 0$ with $\beta \equiv g(\boldsymbol{\mu})$. By the delta method (e.g., Hansen, 2018, Thm. 6.12.3), if $g(\cdot)$ is continuously differentiable in a neighborhood of $\boldsymbol{\mu}$ and $\mathbf{G} \equiv \frac{\partial}{\partial \mathbf{u}} g(\mathbf{u})\big|_{\mathbf{u}=\boldsymbol{\mu}}$, then $\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\mathbf{V}})$ implies $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathrm{N}(0, \mathbf{G}'\underline{\mathbf{V}}\mathbf{G})$. To be concrete, imagine $d = 2$ and $g(\boldsymbol{\mu}) = \mu_1^2 + \mu_2^2 - 1$, so $\{\mathbf{u} : g(\mathbf{u}) \leq 0\}$ is the unit disk in $\mathbb{R}^2$; the delta method gives $\sqrt{n}(g(\hat{\boldsymbol{\mu}}) - g(\boldsymbol{\mu})) \xrightarrow{d} \mathrm{N}(0, 4\boldsymbol{\mu}'\underline{\mathbf{V}}\boldsymbol{\mu})$. Further imagine $\hat{\boldsymbol{\mu}} \sim \mathrm{N}(\boldsymbol{\mu}, \underline{\mathbf{V}}/n)$ in a finite sample of n observations, with the corresponding posterior. Then the Bayesian test's size is strictly above α, as in Theorem 1(iii); as the sampling variance

$\underline{\mathbf{V}}/n$ grows, size grows to 1. In apparent contradiction, $H_0\colon \beta \leq 0$ suggests the Bayesian test has exact asymptotic size, by Theorem 1(i). The "contradiction" is simply that the asymptotic result is less accurate, due to the delta method's linear approximation of $g(\cdot)$. We avoid this layer of delta method approximation by treating nonlinear inequalities directly, providing better finite-sample insights. These insights remain practically helpful for any Θ_0 as long as the sampling and posterior distributions are close to their limits.

3.2 Special cases of results

In the special case when $X, \theta \in \mathbb{R}$, similar results to Theorem 1(i) are found in the literature, like in Casella and Berger (1987a). Less general versions of Theorem 1(ii) and Theorem 1(iii) have also been given when $X, \theta \in \mathbb{R}$.

Theorem 1(iii) covers a special case explored in examples by Kline (2011, §4). Let $\boldsymbol{\theta} \in \mathbb{R}^d$, with $H_0\colon \boldsymbol{\theta} \geq \mathbf{0}$ (elementwise) against $H_1\colon \boldsymbol{\theta} \ngeq \mathbf{0}$ (i.e., at least one element $\theta_j < 0$). Kline (2011, p. 3136) explains the possible divergence of Bayesian and frequentist conclusions as the dimension d grows, when the distribution is multivariate normal with identity matrix covariance. He gives the example of observing $\mathbf{X} = \mathbf{0}$, where for large d the Bayesian $\mathrm{P}(H_0 \mid \mathbf{X} = \mathbf{0}) \approx 0$ while the frequentist p-value is near one. Inverting his example illustrates Theorem 1(iv). If H_0 and H_1 are switched to get $H_0\colon \boldsymbol{\theta} \ngeq \mathbf{0}$ and $H_1\colon \boldsymbol{\theta} \geq \mathbf{0}$, then the divergence is in the opposite direction: $\mathrm{P}(H_0 \mid \mathbf{X} = \mathbf{0}) \approx 1$, and large $\mathrm{P}(H_0 \mid \mathbf{X})$ can occur even when the p-value is near zero. For example, the point $\mathbf{X} = (1.64, 1.64, \ldots, 1.64) \in \mathbb{R}^d$ is the corner of the rejection region for the likelihood ratio test with size $\alpha = 0.05$, but the corresponding $\mathrm{P}(H_0 \mid \mathbf{X}) = 0.40$ when $d = 10$, 0.72 when $d = 25$, and 0.99 when $d = 90$.

Theorem 1 also includes the special case of linear inequalities in $\mathbb{R}^d$. Theorem 1(i) states that for a single linear inequality $H_0\colon \mathbf{c}'\boldsymbol{\theta} \leq c_0$, the Bayesian test has size α. Theorem 1(iii) states that for multiple linear inequalities, the Bayesian test's size is strictly above α, and its RP is strictly above α at every boundary point of Θ_0.

4 Examples

We illustrate Theorem 1 through examples of first-order stochastic dominance and curvature constraints in Sections 4.1 and 4.2, respectively.

4.1 Example: first-order stochastic dominance

For testing first-order stochastic dominance (SD1), let $X_i \overset{iid}{\sim} F_X(\cdot)$, $Y_i \overset{iid}{\sim} F_Y(\cdot)$ and independent of the X_i sample, and $F_0(\cdot)$ is non-random, where all distributions are continuous.

One-sample SD1 is $F_X(\cdot) \leq F_0(\cdot)$; two-sample SD1 is $F_X(\cdot) \leq F_Y(\cdot)$; "non-SD1" means SD1 is not satisfied, so either SD1 or non-SD1 (but not both) must be true.

First, we show how Theorem 1 applies to SD1. Second, we provide analytic results from the limit experiment. Third, we show simulated finite-sample results.

4.1.1 SD1: application of Theorem 1

As we show below, Theorem 1 implies that the Bayesian test's asymptotic size is strictly above α when the null hypothesis is SD1 but that this may not hold when the null is non-SD1. This subsection shows how SD1 and non-SD1 satisfy the conditions of Theorem 1(iii) and Theorem 1(iv), respectively.

Although $F_X(\cdot)$ is infinite-dimensional, only finite-dimensional marginal distributions are required to apply Theorem 1. Consider the simpler Theorem 1(ii) first. Let $X_n(\cdot) \equiv \sqrt{n}\big(\hat{F}_X(\cdot) - F_{0,n}(\cdot)\big)$ and $\theta_n(\cdot) \equiv \sqrt{n}\big(F_X(\cdot) - F_{0,n}(\cdot)\big)$, where (from the frequentist view) $\theta_n(\cdot) \to \theta(\cdot)$, the local mean parameter. SD1 is equivalent to $\theta(\cdot) \leq 0(\cdot)$. Although limits of the full infinite-dimensional sequences are tractable (see Section 4.1.2), only a scalar-valued functional is needed for Theorem 1(ii). Let $\phi(\theta(\cdot)) = \theta(x)$ for some (any) $x \in \mathbb{R}$, so $\theta(\cdot) \leq 0(\cdot) \implies \phi(\theta(\cdot))$, satisfying the condition of Theorem 1(ii) that $\Theta_0 \subseteq \{\theta(\cdot) : \phi(\theta(\cdot)) \leq 0\}$. Let $D_n \equiv \phi(X_n(\cdot)) = \sqrt{n}(\hat{F}_X(x) - F_{0,n}(x))$ and $\gamma_n \equiv \phi(\theta_n(\cdot)) = \sqrt{n}(F_X(x) - F_{0,n}(x))$. Writing $Z_i = \mathbb{1}\{X_i \leq x\}$, then $F_X(x) = \mathbb{E}(Z_i)$ and $\hat{F}_X(x) = n^{-1}\sum_{i=1}^n Z_i$, i.e., we are concerned only with the mean of a random variable. The asymptotic sampling distribution of $D_n - \gamma_n$ is $\mathrm{N}(0, F_X(x)[1 - F_X(x)])$, satisfying the continuity, support, and symmetry conditions in Assumption A1. The remainder of A1 is satisfied if a semiparametric Bernstein–von Mises theorem for the mean holds. This and even stronger results hold with a Dirichlet process prior, as in Lo (1983).

For Theorem 1(iii), we only need strengthen the Bernstein–von Mises theorem from a scalar to a bivariate vector ($d = 2$), which again holds with a Dirichlet process prior (Lo, 1983), for example. Here, let $\phi_2(\theta(\cdot)) = (\theta(x_1), \theta(x_2))$, and $\phi_3(\mathbf{p}) = p_1$ so $\phi_3(\phi_2(\theta(\cdot))) = \theta(x_1) = \phi(\theta(\cdot))$. Also, $\theta(\cdot) \leq 0(\cdot) \implies (\theta(x_1), \theta(x_2)) \leq \mathbf{0}$, satisfying the condition in Theorem 1(iii) on $\Phi_2 = \{\mathbf{p} : p_1 \leq 0, p_2 \leq 0\}$. Continuing to follow the notation from Theorem 1(iii), $\Delta = \{\mathbf{p} : p_1 \leq 0, p_2 > 0\}$, which has positive (indeed infinite) Lebesgue measure as required. The bivariate asymptotic distribution F_2 is bivariate normal, which again satisfies A1 as well as the continuity and strictly positive PDF requirement.

For testing non-SD1, Theorem 1(iv) applies. Non-SD1 is satisfied in the entire half-space $\{\theta(\cdot) : \theta(x) \geq 0\}$, as well as most of the complement half-space (e.g., if $\theta(x) < 0$ but $\theta(x_2) \geq 0$), so it cannot be contained in any half-space.

For the two-sample setting, the infinite-dimensional limits in Section 4.1.2 are more than

sufficient to establish the scalar and bivariate conditions of Theorem 1(iii), and again non-SD1 cannot be contained in any half-space.

4.1.2 SD1: results from limit experiment

We derive the infinite-dimensional limit experiment and then compute certain results. Continuing some notation from Section 4.1.1, consider the one-sample setup with $X_i \overset{iid}{\sim} F_X(\cdot)$. Again let $\theta_n(\cdot) \equiv \sqrt{n}\big(F_X(\cdot) - F_{0,n}(\cdot)\big) \to \theta(\cdot)$, the local parameter. SD1 of F_X over F_0 can be written equivalently as

$$F_X \text{ SD}_1 \ F_0 \iff F_X(\cdot) \leq F_0(\cdot) \iff \theta(\cdot) \leq 0(\cdot). \tag{4}$$

Since (by Donsker's theorem) $\sqrt{n}\big(\hat{F}_X(\cdot) - F_X(\cdot)\big) \rightsquigarrow B\big(F_X(\cdot)\big)$ for standard Brownian bridge $B(\cdot)$, similar to (2),

$$\begin{aligned} X_n(\cdot) &\equiv \sqrt{n}\big(\hat{F}_X(\cdot) - F_{0,n}(\cdot)\big) \\ &= \sqrt{n}\big(\hat{F}_X(\cdot) - F_X(\cdot)\big) + \sqrt{n}\big(F_X(\cdot) - F_{0,n}(\cdot)\big) \\ &\rightsquigarrow B\big(F_X(\cdot)\big) + \theta(\cdot), \end{aligned} \tag{5}$$

so the limit experiment has $X(\cdot) - \theta(\cdot) \mid \theta(\cdot) \sim B\big(F_X(\cdot)\big)$. Note $B\big(F_X(\cdot)\big)$ is a mean-zero Gaussian process with covariance function $\text{Cov}(t_1, t_2) = F_X(t_1)[1 - F_X(t_2)]$ for $t_1 \leq t_2$. Although $F_X(\cdot)$ is unknown, $\hat{F}_X(\cdot) \overset{a.s.}{\to} F_X(\cdot)$ uniformly by the Glivenko–Cantelli theorem, so asymptotically the covariance function is known while the (local) mean function $\theta(\cdot)$ remains unknown. Analogously, for the posterior, similar to (3),

$$\theta_n(\cdot) - X_n(\cdot) = \sqrt{n}\big(F_X(\cdot) - \hat{F}_X(\cdot)\big) \rightsquigarrow B\big(F_X(\cdot)\big), \tag{6}$$

using the Bernstein–von Mises theorem in Lo (1983, 1987) for Dirichlet process prior Bayesian inference or Theorem 4 of Castillo and Nickl (2014).

The two-sample setting is similar since we assume the samples are independent. For notational simplicity, assume both samples have n observations. Let $\Delta(\cdot) \equiv F_X(\cdot) - F_Y(\cdot)$, the true CDF difference function. Let $\Delta_{0,n}(\cdot)$ be the centering functions satisfying $\sqrt{n}(\Delta(\cdot) - \Delta_{0,n}(\cdot)) \to \theta(\cdot)$, the local parameter. SD1 of F_X over F_Y is

$$F_X \text{ SD}_1 \ F_Y \iff F_X(\cdot) \leq F_Y(\cdot) \iff \theta(\cdot) \leq 0(\cdot). \tag{7}$$

For the sampling distribution,

$$X_n(\cdot) \equiv \sqrt{n}(\hat{F}_X(\cdot) - \hat{F}_Y(\cdot) - \Delta_{0,n}(\cdot))$$

$$= \sqrt{n}(\hat{F}_X(\cdot) - F_X(\cdot)) - \sqrt{n}(\hat{F}_Y(\cdot) - F_Y(\cdot)) + \sqrt{n}(\Delta(\cdot) - \Delta_{0,n}(\cdot))$$
$$\rightsquigarrow B_1(F_X(\cdot)) - B_2(F_Y(\cdot)) + \theta(\cdot), \tag{8}$$

where $B_1(\cdot)$ and $B_2(\cdot)$ are independent standard Brownian bridges. For the posterior, using the independence of samples and Bernstein–von Mises theorem,

$$\theta_n(\cdot) - X_n(\cdot) = \sqrt{n}(F_X(\cdot) - \hat{F}_X(\cdot)) - \sqrt{n}(F_Y(\cdot) - \hat{F}_Y(\cdot))$$
$$\rightsquigarrow B_1(F_X(\cdot)) - B_2(F_Y(\cdot)). \tag{9}$$

First, consider the Bayesian posterior probability of SD1 when $X(\cdot) = 0(\cdot)$. The finite-sample analog is when $\hat{F}_X(\cdot) \approx F_0(\cdot)$ or $\hat{F}_X(\cdot) \approx \hat{F}_Y(\cdot)$. The value $0(\cdot)$ is at the very "corner" of Θ_0, and it is a very pointy corner, so a ball centered at $0(\cdot)$ contains very little of Θ_0. In fact, "very little" means "zero probability," as the next result states.

Proposition 2. *Consider the limit experiment posterior for one-sample SD1 testing in* (6) *and for two-sample SD1 testing in* (9)*. Given $X(\cdot) = 0(\cdot)$, the posterior probability of SD1 is zero in both the one-sample and two-sample setting.*

Second, similar intuition and arguments lead to the SD1 Bayesian test's size being one.

Proposition 3. *Consider the limit experiment for one-sample SD1 testing in* (5) *and* (6) *and for two-sample SD1 testing in* (8) *and* (9)*. Consider the Bayesian test from Method 1 that rejects $H_0\colon \theta(\cdot) \leq 0(\cdot)$ iff the posterior probability of H_0 is below α. Then, the Bayesian test's frequentist size equals one, with type I error rate equal to one when $\theta(\cdot) = 0(\cdot)$.*

Third, the following result for non-SD1 rejection probability is immediate.

Corollary 4. *Consider same setup as in Proposition 3. When $\theta(\cdot) = 0(\cdot)$, the Bayesian test's probability of rejecting non-SD1 is zero.*

4.1.3 SD1: finite-sample simulations

The following simulation results reflect the theoretical results from the limit experiment (discussed in Section 4.1.1). Code for replication is provided.

Table 1 shows Bayesian posterior probabilities of SD1 and non-SD1 in datasets near the "corner" of SD1, similar to the setup of Proposition 2. In the one-sample case, this means $\hat{F}_X(\cdot)$ nearly equals the Unif$(0, 1)$ CDF. In the two-sample case, this means $\hat{F}_X(\cdot)$ nearly equals $\hat{F}_Y(\cdot)$. Specifically, $X_i = i/(n+1)$ for $i = 1, \ldots, n$, and in the two-sample case, $Y_i = i/n$ for $i = 1, \ldots, n-1$ (there are $n-1$ observations in the second sample). When

Table 1: Bayesian posterior probabilities of H_0: X SD_1 Unif$(0,1)$ and H_0: X SD_1 Y.

		Comparison distribution	
H_0	n	Unif$(0,1)$	Y
SD1	10	0.103	0.097
SD1	40	0.028	0.025
SD1	100	0.009	0.010
SD1	∞	0.000	0.000
non-SD1	10	0.897	0.903
non-SD1	40	0.972	0.975
non-SD1	100	0.991	0.990
non-SD1	∞	1.000	1.000

$n < \infty$, the Bayesian bootstrap variant of Banks (1988) is used. When $n = \infty$, the results are from Proposition 2.

Table 1 illustrates the Bayesian interpretation of a draw near $X(\cdot) = 0(\cdot)$. This interpretation differs greatly from a frequentist interpretation and illuminates the rejection probabilities seen in Table 2. The subspace of distribution functions $F_X(\cdot)$ satisfying SD1 has a very sharp "corner" at $F_0(\cdot)$, or equivalenty at $\theta(\cdot) = 0(\cdot)$. Consequently, when $\hat{F}_X(\cdot) \approx F_0(\cdot)$, or when $\hat{F}_X(\cdot) \approx \hat{F}_Y(\cdot)$, the Bayesian posterior places nearly zero probability on SD1 and (equivalently) almost all probability on non-SD1. Table 1 shows finite-sample posterior probabilities when $n = 100$ to be very close to the limit as $n \to \infty$. Opposite the Bayesian interpretation, a frequentist p-value for the null of SD1 would be near one when the estimated $\hat{F}_X(\cdot)$ is near $F_0(\cdot)$ or $\hat{F}_Y(\cdot)$. These results are qualitatively similar to those for the one-sample, finite-dimensional example in Kline (2011, §4).

Table 2: Bayesian test rejection probabilities, $\alpha = 0.1$, 1000 replications.

		Comparison distribution	
H_0	n	Unif$(0,1)$	Y
SD1	10	0.740	0.655
SD1	40	0.935	0.917
SD1	100	0.981	0.977
SD1	∞	1.000	1.000
non-SD1	10	0.000	0.005
non-SD1	40	0.000	0.000
non-SD1	100	0.000	0.000
non-SD1	∞	0.000	0.000

Table 2 shows rejection probabilities of the Bayesian test when $\theta(\cdot) = 0(\cdot)$. This is the

"least favorable configuration" for the null of SD1 (but not for non-SD1). The DGP has $X_i \overset{iid}{\sim} \text{Unif}(0,1)$ for $i = 1, \ldots, n$. For one-sample testing, $F_0(\cdot)$ is the true (standard uniform) CDF of X_i. For two-sample testing, $Y_i \overset{iid}{\sim} \text{Unif}(0,1)$ for $i = 1, \ldots, n$, identical to X_i. The hypotheses, methods, and notation are the same as for Table 1. The entries for $n = \infty$ use Proposition 3 and Corollary 4.

Table 2 shows the same patterns as Table 1. When H_0 is SD1, the Bayesian type I error rate is well above α even with $n = 10$, with rejection probability increasing to 100% as n grows; consequently, size is also above α. The opposite occurs when H_0 is non-SD1, which is not a subset of a half-space, with type I error rates of zero.[4]

4.2 Example: curvature constraints

One common nonlinear inequality hypothesis in economics is a "curvature" constraint like concavity. Such constraints come from economic theory, often the second-order condition of an optimization problem like utility maximization or cost minimization. As noted by O'Donnell and Coelli (2005), the Bayesian approach is appealing for imposing or testing curvature constraints due to its (relative) simplicity. However, according to Theorem 1, since curvature is usually satisfied in a parameter subspace much smaller than a half-space, Bayesian inference similar to Method 1 may be much less favorable toward the curvature hypothesis than frequentist inference would be; i.e., the size of the Bayesian test in Method 1 may be well above α. This is true in Table 3 below.

Our example concerns concavity of a cost function with the "translog" functional form (Christensen, Jorgenson, and Lau, 1973). This has been a popular way to parameterize cost, indirect utility, and production functions. The translog is more flexible than many traditional functional forms, allowing violation of certain implications of economic theory (such as curvature) without reducing such constraints to the value of a single parameter. Since Lau (1978), there has been continued interest in methods to impose curvature constraints during estimation, as well as methods to test such constraints. Although "flexible," the translog is still parametric, so violation of curvature constraints may come from misspecification (of the functional form) rather than violation of economic theory.[5]

Specifically, we test concavity of cost in input prices as follows.[6] With output y, input

[4] Although the type I error rate for non-SD1 is near zero with this DGP, the test's size is actually α, which is attained when there is a single "contact point" with $F_X(r) = F_Y(r)$ and the inequalities are strict for all other $t \neq r$, thus reducing the test (asymptotically) to a single, scalar inequality.

[5] With a nonparametric model, one may more plausibly test the theory itself, although there are always other assumptions that may be violated; see Dette, Hoderlein, and Neumeyer (2016) for nonparametrically testing negative semidefiniteness of the Slutsky substitution matrix.

[6] The "translog" example on page 346 of Dufour (1989) is even simpler but appears to ignore the fact that second derivatives are not invariant to log transformations.

prices $\mathbf{w} = (w_1, w_2, w_3)$, and total cost $C(y, \mathbf{w})$, the translog model is

$$\begin{aligned}\ln(C(y,\mathbf{w})) = a_0 + a_y \ln(y) + (1/2)a_{yy}[\ln(y)]^2 + \sum_{k=1}^{3} a_{yk} \ln(y) \ln(w_k) \\ + \sum_{k=1}^{3} b_k \ln(w_k) + (1/2) \sum_{k=1}^{3} \sum_{m=1}^{3} b_{km} \ln(w_k) \ln(w_m).\end{aligned} \tag{10}$$

Standard economic assumptions imply that $C(y, \mathbf{w})$ is concave in $\mathbf{w}$ (as in Kreps, 1990, §7.3), which corresponds to the Hessian matrix (of C with respect to $\mathbf{w}$) being negative semidefinite (NSD), which in turn corresponds to all the Hessian's principal minors of order p (for all $p = 1, 2, 3$) having the same sign as $(-1)^p$ or zero.

For simplicity, we consider local concavity at the point $(1, 1, 1, 1)$:

$$H_0 \colon \underline{\mathbf{H}} \equiv \left. \frac{\partial^2 C(y, \mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}'} \right|_{(y,\mathbf{w})=(1,1,1,1)} \text{ is NSD.} \tag{11}$$

This is necessary but not sufficient for global concavity; rejecting local concavity implies rejection of global concavity. In Appendix C, we show that even this weaker constraint corresponds to a set of parameter values much smaller than a half-space, so Theorem 1(iii) applies.

Our simulation DGP is as follows. To impose homogeneity of degree one in input prices, we use the normalized model (with error term ϵ added)

$$\begin{aligned}\ln(C/w_3) = a_0 + a_y \ln(y) + (1/2)a_{yy}[\ln(y)]^2 + \sum_{k=1}^{2} a_{yk} \ln(y) \ln(w_k/w_3) \\ + \sum_{k=1}^{2} b_k \ln(w_k/w_3) + (1/2) \sum_{k=1}^{2} \sum_{m=1}^{2} b_{km} \ln(w_k/w_3) \ln(w_m/w_3) + \epsilon\end{aligned} \tag{12}$$

for both data generation and inference.[7] The parameter values are $b_1 = b_2 = 1/3$, $b_{11} = b_{22} = 2/9 - \delta$ (more on δ below), and $b_{12} = -1/9$ to make some of the inequality constraints in H_0 close to binding, as well as $a_0 = 1$, $a_y = 1$, $a_{yy} = 0$, $a_{yk} = 0$. The other parameter values follow from imposing symmetry ($b_{km} = b_{mk}$) and homogeneity. When $\delta = 0$, $\underline{\mathbf{H}}$ is a matrix of zeros, on the boundary of being NSD in that each principal minor equals zero (and none are strictly negative). When $\delta > 0$, all principal minors are strictly negative (other than $\det(\underline{\mathbf{H}}) = 0$, which is always true under homogeneity). We set $\delta = 0.001$. In each simulation replication, an iid sample is drawn, where $\ln(y)$ and all $\ln(w_k)$ are $\mathrm{N}(0, \sigma = 0.1)$,

[7] Alternatively, cost share equations may be used. Shephard's lemma implies that the demand for input k is $x_k = \partial C/\partial w_k$. The cost share for input k is then $s_k = x_k w_k/C = (\partial C/\partial w_k)(w_k/C) = \partial \ln(C)/\partial \ln(w_k) \equiv r_k = b_k + a_{yk} \ln(y) + \sum_{j=1}^{3} b_{jk} \ln(w_j)$.

$\epsilon \sim \mathrm{N}(0, \sigma_\epsilon)$, and all variables are mutually independent. There are $n = 100$ observations per sample, 500 simulation replications, and 200 posterior draws per replication. The local monotonicity constraints $b_1, b_2, b_3 \geq 0$ were satisfied in 100.0% of replications overall.

Table 3 reports values from two methods. For the method denoted "Bayesian bootstrap" in the table header, the posterior probability of H_0 is computed by a nonparametric Bayesian method with improper Dirichlet process prior, i.e., the Bayesian bootstrap of Rubin (1981) based on Ferguson (1973) and more recently advocated in economics by Chamberlain and Imbens (2003). For the method denoted "Normal," the parameter vector is sampled from a normal distribution with mean equal to the ordinary least squares (OLS) estimate and covariance matrix equal to the corresponding (homoskedastic) asymptotic covariance matrix estimate; this is the posterior from a homoskedastic normal linear regression model with improper prior (or asymptotically). To accommodate numerical imprecision, we deem an inequality satisfied if it is within 10^{-7}. The simulated type I error rate is the proportion of simulated samples for which the posterior probability of H_0 was below α.

Table 3: Simulated type I error rate Bayesian tests of local NSD.

		Bayesian	
α	σ_ϵ	bootstrap	Normal
0.05	0.00	0.000	0.000
0.05	0.10	0.112	0.084
0.05	0.20	0.354	0.318
0.05	0.30	0.554	0.532
0.05	0.40	0.676	0.694
0.05	0.50	0.764	0.772
0.10	0.00	0.000	0.000
0.10	0.10	0.186	0.160
0.10	0.20	0.530	0.526
0.10	0.30	0.740	0.764
0.10	0.40	0.844	0.872
0.10	0.50	0.890	0.910

Table 3 shows the type I error rate of the Bayesian tests of (11) given our DGP. The values of α and σ_ϵ are varied as shown in the table. The two Bayesian tests are very similar, always within a few percentage points of each other. As a sanity check, when $\sigma_\epsilon = 0$, the RP is zero since the constraints are satisfied by construction. As σ_ϵ increases, the RP increases well above α, even over 50%.[8] Although the Bayesian test's size distortion with the null of

[8]The results with $\delta = 0.01$ and $\sigma_\epsilon \in [0, 1]$ are similar to Table 3; with $\delta = 0$, RP jumps to over 80% even with $\sigma_\epsilon = 0.001$.

local NSD is clearly bad from a frequentist perspective, it reflects the Bayesian method's need for great evidence to conclude in favor of local NSD, which may be reasonable since the translog form does not come from economic theory and since only a small part of the parameter space satisfies local NSD. Either way, it is helpful to understand the behavior of Bayesian inference in this situation.

5 Conclusion

We have explored the frequentist properties of Bayesian inference on general nonlinear inequality constraints, providing formal results on the role of the shape of the null hypothesis parameter subspace. Investigation of approaches like Müller and Norets (2016) applied to nonlinear inequality testing remains for future work. It would also be valuable to extend this paper's analysis to allow (proper) priors with $\mathrm{P}(H_0) = 1/2$ or other values. Moreover, one could explore how to achieve correct frequentist size by adjusting the prior $\mathrm{P}(H_0)$ or by adjusting the relative weight of type I and II errors in the loss function, or how to achieve a posterior probability of H_0 equal to the p-value from a common frequentist method.

References

Andrews, D. W. K., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78 (1), 119–157.
URL https://www.jstor.org/stable/25621398

Banks, D. L., 1988. Histospline smoothing the Bayesian bootstrap. Biometrika 75 (4), 673–684.
URL https://www.jstor.org/stable/2336308

Bayarri, M. J., Berger, J. O., Forte, A., García-Donato, G., 2012. Criteria for Bayesian model choice with application to variable selection. Annals of Statistics 40 (3), 1550–1577.
URL https://projecteuclid.org/euclid.aos/1346850065

Berger, J. O., 2003. Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science 18 (1), 1–32.
URL https://doi.org/10.1214/ss/1056397485

Berger, J. O., Brown, L. D., Wolpert, R. L., 1994. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. Annals of Statistics 22 (4), 1787–1807.
URL https://www.jstor.org/stable/2242484

Berger, J. O., Sellke, T., 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. Journal of the American Statistical Association 82 (397), 112–122.
URL https://doi.org/10.1080/01621459.1987.10478397

Bickel, P. J., Kleijn, B. J. K., 2012. The semiparametric Bernstein–von Mises theorem.

Annals of Statistics 40 (1), 206–237.
URL https://projecteuclid.org/euclid.aos/1333029963

Birnbaum, Z. W., Tingey, F. H., 1951. One-sided confidence contours for probability distribution functions. Annals of Mathematical Statistics 22 (4), 592–596.
URL https://www.jstor.org/stable/2236929

Bogachev, V. I., 1998. Gaussian Measures. Vol. 62 of Mathematical Surveys and Monographs. American Mathematical Society.

Casella, G., Berger, R. L., 1987a. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association 82 (397), 106–111.
URL https://www.jstor.org/stable/2289130

Casella, G., Berger, R. L., 1987b. Testing precise hypotheses: Comment. Statistical Science 2 (3), 344–347.
URL https://www.jstor.org/stable/2245777

Castillo, I., Nickl, R., 2013. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. Annals of Statistics 41 (4), 1999–2028.
URL https://projecteuclid.org/euclid.aos/1382547511

Castillo, I., Nickl, R., 2014. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. Annals of Statistics 42 (5), 1941–1969.
URL https://projecteuclid.org/euclid.aos/1410440630

Castillo, I., Rousseau, J., 2015. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. Annals of Statistics 43 (6), 2353–2383.
URL https://projecteuclid.org/euclid.aos/1444222078

Chamberlain, G., Imbens, G. W., 2003. Nonparametric applications of Bayesian inference. Journal of Business & Economic Statistics 21 (1), 12–18.

Christensen, L. R., Jorgenson, D. W., Lau, L. J., 1973. Transcendental logarithmic production frontiers. Review of Economics and Statistics 55 (1), 28–45.
URL https://www.jstor.org/stable/1927992

DasGupta, A., 2008. Asymptotic Theory of Statistics and Probability. Springer, New York.

Dette, H., Hoderlein, S., Neumeyer, N., 2016. Testing multivariate economic restrictions using quantiles: The example of Slutsky negative semidefiniteness. Journal of Econometrics 191 (1), 129–144.
URL https://doi.org/10.1016/j.jeconom.2015.07.004

Dufour, J.-M., 1989. Nonlinear hypotheses, inequality restrictions, and non-nested hypotheses: Exact simultaneous tests in linear regressions. Econometrica 57 (2), 335–355.
URL https://www.jstor.org/stable/1912558

Fang, Z., Santos, A., 2015. Inference on directionally differentiable functions, working paper, available at https://arxiv.org/abs/1404.3763.

Ferguson, T. S., 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1 (2), 209–230.
URL https://doi.org/10.1214/aos/1176342360

Freedman, D., 1999. On the Bernstein–von Mises theorem with infinite-dimensional parameters. Annals of Statistics 27 (4), 1119–1140.
URL http://projecteuclid.org/euclid.aos/1017938917

Ghosal, S., van der Vaart, A., 2017. Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University

Press.

Goutis, C., Casella, G., Wells, M. T., 1996. Assessing evidence in multiple hypotheses. Journal of the American Statistical Association 91 (435), 1268–1277.
URL https://www.jstor.org/stable/2291745

Hahn, J., 1997. Bayesian bootstrap of the quantile regression estimator: A large sample study. International Economic Review 38 (4), 795–808.
URL https://www.jstor.org/stable/2527216

Hansen, B. E., 2018. Econometrics, unpublished textbook, available at https://www.ssc.wisc.edu/~bhansen/econometrics/.

Hirano, K., Porter, J. R., 2009. Asymptotics for statistical treatment rules. Econometrica 77 (5), 1683–1701.
URL https://www.jstor.org/stable/25621374

Imbens, G. W., Manski, C. F., 2004. Confidence intervals for partially identified parameters. Econometrica 72 (6), 1845–1857.
URL https://doi.org/10.1111/j.1468-0262.2004.00555.x

Kaplan, D. M., 2015. Bayesian and frequentist tests of sign equality and other nonlinear inequalities, working paper, available at https://faculty.missouri.edu/~kaplandm.

Kim, J.-Y., 2002. Limited information likelihood and Bayesian analysis. Journal of Econometrics 107 (1), 175–193.
URL https://doi.org/10.1016/S0304-4076(01)00119-1

Kline, B., 2011. The Bayesian and frequentist approaches to testing a one-sided hypothesis about a multivariate mean. Journal of Statistical Planning and Inference 141 (9), 3131–3141.
URL https://doi.org/10.1016/j.jspi.2011.03.034

Kreps, D. M., 1990. A Course in Microeconomic Theory. Princeton University Press.

Kwan, Y. K., 1999. Asymptotic Bayesian analysis based on a limited information estimator. Journal of Econometrics 88 (1), 99–121.
URL https://doi.org/10.1016/S0304-4076(98)00024-4

Lancaster, T., 2003. A note on bootstraps and robustness, available at SSRN: https://ssrn.com/abstract=896764.
URL https://doi.org/10.2139/ssrn.896764

Laplace, P.-S., 1820. Théorie Analytique des Probabilités, 3rd Edition. V. Courcier, Paris.

Lau, L. J., 1978. Testing and imposing monotonicity, convexity, and quasi-convexity constraints. In: Fuss, M., McFadden, D. (Eds.), Production Economics: A Dual Approach to Theory and Applications. Vol. 1 of Contributions to Economic Analysis. North-Holland, Amsterdam, Ch. A.4, pp. 409–453.

Lehmann, E. L., Casella, G., 1998. Theory of Point Estimation, 2nd Edition. Springer, New York.

Lehmann, E. L., Romano, J. P., 2005. Testing Statistical Hypotheses, 3rd Edition. Springer Texts in Statistics. Springer.
URL http://books.google.com/books?id=Y7vSVW3ebSwC

Lindley, D. V., 1957. A statistical paradox. Biometrika 44 (1–2), 187–192.
URL https://www.jstor.org/stable/2333251

Lo, A. Y., 1983. Weak convergence for Dirichlet processes. Sankhyā: The Indian Journal of Statistics, Series A 45 (1), 105–111.

URL https://www.jstor.org/stable/25050418

Lo, A. Y., 1987. A large sample study of the Bayesian bootstrap. Annals of Statistics 15 (1), 360–375.
URL http://projecteuclid.org/euclid.aos/1176350271

Moon, H. R., Schorfheide, F., 2012. Bayesian and frequentist inference in partially identified models. Econometrica 80 (2), 755–782.
URL https://www.jstor.org/stable/41493833

Müller, U. K., Norets, A., 2016. Credibility of confidence sets in nonstandard econometric problems. Econometrica 84 (6), 2183–2213.
URL https://doi.org/10.3982/ECTA14023

Norets, A., 2015. Bayesian regression with nonparametric heteroskedasticity. Journal of Econometrics 185 (2), 409–419.
URL https://doi.org/10.1016/j.jeconom.2014.12.006

O'Donnell, C. J., Coelli, T. J., 2005. A Bayesian approach to imposing curvature on distance functions. Journal of Econometrics 126 (2), 493–523.
URL https://doi.org/10.1016/j.jeconom.2004.05.011

Patton, A. J., Timmermann, A., 2010. Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. Journal of Financial Economics 98 (3), 605–625.
URL https://doi.org/10.1016/j.jfineco.2010.06.006

Rubin, D. B., 1981. The Bayesian bootstrap. Annals of Statistics 9 (1), 130–134.
URL http://projecteuclid.org/euclid.aos/1176345338

Schennach, S. M., 2005. Bayesian exponentially tilted empirical likelihood. Biometrika 92 (1), 31–46.
URL https://www.jstor.org/stable/20441164

Shen, X., 2002. Asymptotic normality of semiparametric and nonparametric posterior distributions. Journal of the American Statistical Association 97 (457), 222–235.
URL https://doi.org/10.1198/016214502753479365

Sims, C. A., 2010. Understanding non-Bayesians, unpublished book chapter, available at http://sims.princeton.edu/yftp/UndrstndgNnBsns/GewekeBookChpter.pdf.

Sims, C. A., Uhlig, H., 1991. Understanding unit rooters: A helicopter tour. Econometrica 59 (6), 1591–1599.
URL https://www.jstor.org/stable/2938280

Smirnov, N. V., 1939. Sur les écarts de la courbe de distribution empirique. Recueil Mathématique [Mathematicheskii Sbornik] 6(48) (1), 3–26.
URL http://mi.mathnet.ru/eng/msb5810

Stoye, J., 2009. More on confidence intervals for partially identified parameters. Econometrica 77 (4), 1299–1315.
URL https://www.jstor.org/stable/40263861

Stoye, J., Kitamura, Y., 2013. Nonparametric analysis of random utility models: Testing. In: Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung, Session: Microeconometrics, No. D20–V3. pp. 1–43.
URL http://hdl.handle.net/10419/79753

Sullivan, R., Timmermann, A., White, H., 1999. Data-snooping, technical trading rule per-

formance, and the bootstrap. Journal of Finance 54 (5), 1647–1691.
URL https://doi.org/10.1111/0022-1082.00163

Sullivan, R., Timmermann, A., White, H., 2001. Dangers of data mining: The case of calendar effects in stock returns. Journal of Econometrics 105 (1), 249–286.
URL https://doi.org/10.1016/S0304-4076(01)00077-X

van der Vaart, A. W., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.
URL https://books.google.com/books?id=UEuQEM5RjWgC

van der Vaart, A. W., Wellner, J. A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics. Springer, New York.

Wolak, F. A., 1989. Testing inequality constraints in linear econometric models. Journal of Econometrics 41 (2), 205–235.
URL https://doi.org/10.1016/0304-4076(89)90094-8

Wolak, F. A., 1991. The local nature of hypothesis tests involving inequality constraints in nonlinear models. Econometrica 59 (4), 981–995.
URL https://www.jstor.org/stable/2938170

A Proofs

Proof of Theorem 1. For Theorem 1(i): the Bayesian test rejects iff

$$\alpha \geq \mathrm{P}\big(\phi(\boldsymbol{\theta}) \leq c_0 \mid \mathbf{X}\big) = \mathrm{P}\big(\phi(\boldsymbol{\theta}) - \phi(\mathbf{X}) \leq c_0 - \phi(\mathbf{X}) \mid \mathbf{X}\big) \equiv F\big(c_0 - \phi(\mathbf{X})\big).$$

Given any $\boldsymbol{\theta}$ such that $\phi(\boldsymbol{\theta}) \leq c_0$ (so H_0 holds), the RP is

$$\begin{aligned}
\mathrm{P}\big(F(c_0 - \phi(\mathbf{X})) \leq \alpha \mid \boldsymbol{\theta}\big) &= \mathrm{P}\big(\overbrace{1 - F(\phi(\mathbf{X}) - c_0)}^{\text{by symmetry}} \leq \alpha \mid \boldsymbol{\theta}\big) \\
&= \mathrm{P}\big(F(\phi(\mathbf{X}) - c_0) \geq 1 - \alpha \mid \boldsymbol{\theta}\big) \\
&\overbrace{\leq \mathrm{P}\big(F(\phi(\mathbf{X}) - \phi(\boldsymbol{\theta})) \geq 1 - \alpha \mid \boldsymbol{\theta}\big)}^{\text{since } \phi(\boldsymbol{\theta}) \leq c_0 \text{ under } H_0} \\
&= \alpha
\end{aligned}$$

since $F\big(\phi(\mathbf{X}) - \phi(\boldsymbol{\theta})\big) \mid \boldsymbol{\theta} \sim \mathrm{Unif}(0,1)$. If $\phi(\boldsymbol{\theta}) = c_0$, then the $\leq$ becomes $=$.

For Theorem 1(ii): because $\Theta_0 \subseteq \{\boldsymbol{\theta} : \phi(\boldsymbol{\theta}) \leq c_0\}$, then for any $\mathbf{X}$,

$$\mathrm{P}(\boldsymbol{\theta} \in \Theta_0 \mid \mathbf{X}) \leq \mathrm{P}\big(\phi(\boldsymbol{\theta}) \leq c_0 \mid \mathbf{X}\big).$$

Consequently, the rejection region for $H_0 \colon \boldsymbol{\theta} \in \Theta_0$ is at least as big as the rejection region

for H_0: $\phi(\boldsymbol{\theta}) \leq c_0$: for some $r \in \mathbb{R}$,

$$\begin{aligned} &\mathcal{R}_1 \subseteq \mathcal{R}_2,\ \mathcal{R}_1 \equiv \{\mathbf{X} : \mathrm{P}(\phi(\boldsymbol{\theta}) \leq c_0 \mid \mathbf{X}) \leq \alpha\} = \{\mathbf{X} : \phi(\mathbf{X}) \geq r\}, \\ &\mathcal{R}_2 \equiv \{\mathbf{X} : \mathrm{P}(\boldsymbol{\theta} \in \Theta_0 \mid \mathbf{X}) \leq \alpha\}. \end{aligned} \tag{13}$$

Given any $\boldsymbol{\theta} \in \Theta_0$, the probability that $\mathbf{X}$ falls in the new, larger rejection region ($\mathcal{R}_2$) is at least as big as the probability that $\mathbf{X}$ falls in the old, smaller rejection region ($\mathcal{R}_1$) from Theorem 1(i). In particular, when $\phi(\boldsymbol{\theta}) = c_0$, the RP was exactly α in Theorem 1(i). Since the new rejection region is weakly larger, the new RP when $\phi(\boldsymbol{\theta}) = c_0$ must be at least α. If $c_0 \in \phi(\Theta_0)$, then the proof is complete. Otherwise, with $\mathcal{R}_1$ and $\mathcal{R}_2$ from (13), and $\boldsymbol{\theta}^*$ any value such that $\phi(\boldsymbol{\theta}^*) = c_0$ (with the limit formed by a sequence of $\boldsymbol{\theta}$ within Θ_0),

$$\begin{aligned} &\sup_{\boldsymbol{\theta} \in \Theta_0} \mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}) \overbrace{\geq \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta})}^{\text{since } c_0 \in \phi(\overline{\Theta}_0)} \overbrace{\geq \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}(\mathbf{X} \in \mathcal{R}_1 \mid \boldsymbol{\theta})}^{\text{by (13)}} \\ &= \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}\big(\phi(\mathbf{X}) \geq r \mid \boldsymbol{\theta}\big) = \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}(\phi(\mathbf{X}) - \phi(\boldsymbol{\theta}) \geq r - \phi(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) = \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} 1 - F\big(r - \phi(\boldsymbol{\theta})\big) \\ &\overbrace{= 1 - F(r - c_0)}^{\text{by continuity of } F, \phi} \quad \overbrace{= \alpha}^{\text{by Theorem 1(i)}} \quad . \end{aligned} \tag{14}$$

For Theorem 1(iii): given the stated assumption that the posterior distribution of $\phi_2(\boldsymbol{\theta})$ has a strictly positive PDF for any $\phi_2(\mathbf{X})$, and the assumption that Δ has positive Lebesgue measure, then

$$\mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Delta \mid \phi_2(\mathbf{X})) > 0 \text{ for any } \phi_2(\mathbf{X}). \tag{15}$$

Similar to (13), for some $r \in \mathbb{R}$, let

$$\begin{aligned} &\mathcal{R}_a \subseteq \mathcal{R}_b,\ \mathcal{R}_a \equiv \{\phi_2(\mathbf{X}) : \mathrm{P}(\phi_3(\phi_2(\boldsymbol{\theta})) \leq c_0 \mid \phi_2(\mathbf{X})) \leq \alpha\} = \{\phi_2(\mathbf{X}) : \phi_3(\phi_2(\mathbf{X})) \geq r\}, \\ &\mathcal{R}_b \equiv \{\phi_2(\mathbf{X}) : \mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X})) \leq \alpha\}. \end{aligned} \tag{16}$$

Let $\mathbf{p}^*$ be any value such that $\phi_3(\mathbf{p}^*) = r$. Then,

$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta} \in \Theta_0 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*) &\overbrace{\leq \mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}^{\text{since } \boldsymbol{\theta} \in \Theta_0 \implies \phi_2(\boldsymbol{\theta}) \in \Phi_2} \\ &= \overbrace{\mathrm{P}(\phi_3(\phi_2(\boldsymbol{\theta})) \leq c_0 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}^{=\alpha \text{ by Theorem 1(i)}} - \overbrace{\mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Delta \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}^{>0 \text{ by (15)}} \\ &< \alpha. \end{aligned}$$

By the assumption that $\mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X}))$ is continuous in $\phi_2(\mathbf{X})$, there is some ϵ-ball $\mathcal{B}$ around $\mathbf{p}^*$ for which $\mathrm{P}(\boldsymbol{\theta} \in \Theta_0 \mid \phi_2(\mathbf{X}) \in \mathcal{B}) < \alpha$, too. The ball $\mathcal{B}$ has positive Lebesgue measure, as does the part of it lying outside $\mathcal{R}_a$ (i.e., $\mathcal{B} \cap \mathcal{R}_a^{\complement}$) since $\mathbf{p}^*$ is on the boundary

of $\mathcal{R}_a$. Since the sampling distribution of $\phi_2(\mathbf{X})$ given any $\boldsymbol{\theta}$ has a strictly positive PDF (by assumption),

$$\mathrm{P}\big(\phi_2(\mathbf{X}) \in (\mathcal{B} \cap \mathcal{R}_a^{\complement}) \mid \boldsymbol{\theta}\big) > 0 \text{ for any } \boldsymbol{\theta}. \tag{17}$$

Also, using the assumption in A1 that the distribution of $\phi_2(\boldsymbol{\theta})$ only depends on $\mathbf{X}$ through $\phi_2(\mathbf{X})$, as well as the assumption that $\boldsymbol{\theta} \in \Theta_0 \implies \phi_2(\boldsymbol{\theta}) \in \Phi_2$,

$$\mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X})) = \mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \mathbf{X}) \geq \mathrm{P}(\boldsymbol{\theta} \in \Theta_0 \mid \mathbf{X}),$$

so $\mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X})) \leq \alpha \implies \mathrm{P}(\boldsymbol{\theta} \in \Theta_0 \mid \mathbf{X}) \leq \alpha$. Consequently,

$$\phi_2(\mathbf{X}) \in \mathcal{R}_b \implies \mathbf{X} \in \mathcal{R}_2. \tag{18}$$

Letting $\boldsymbol{\theta}^*$ be a value such that $\phi_3(\phi_2(\boldsymbol{\theta}^*)) = \phi(\boldsymbol{\theta}^*) = c_0$, the Bayesian test's rejection probability is bounded from below by

$$\begin{aligned}
\mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}^*) &\overbrace{\geq \mathrm{P}(\phi_2(\mathbf{X}) \in \mathcal{R}_b \mid \boldsymbol{\theta}^*)}^{\text{by (18)}} \\
&= \overbrace{\mathrm{P}(\phi_2(\mathbf{X}) \in \mathcal{R}_a \mid \boldsymbol{\theta}^*)}^{=\alpha \text{ by Theorem 1(i)}} + \overbrace{\mathrm{P}\big(\phi_2(\mathbf{X}) \in (\mathcal{B} \cap \mathcal{R}_a^{\complement}) \mid \boldsymbol{\theta}^*\big)}^{>0 \text{ by (17)}} \\
&> \alpha.
\end{aligned}$$

As in the proof of Theorem 1(ii), if $\boldsymbol{\theta}^* \in \Theta_0$, then the test's size is bounded below by $\mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}^*)$ and the proof is complete. Otherwise, as before, the assumed continuity and $c_0 \in \phi(\overline{\Theta}_0)$ imply

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}) \geq \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}) = \mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}^*).$$

For Theorem 1(iv), some examples suffice. First, consider $H_0\colon \phi(\boldsymbol{\theta}) \neq 0$. Given any $\mathbf{X}$, $\mathrm{P}(\phi(\boldsymbol{\theta}) \neq 0 \mid \mathbf{X}) = 1$ since F is continuous, so the Bayesian test never rejects and its size is zero. Thus, size may be strictly below α. Second, consider $H_0\colon \phi(\boldsymbol{\theta}) \in \mathbb{Z}$ (the integers). This H_0 has zero posterior probability given any $\mathbf{X}$, so the Bayesian test always rejects and has size equal to one. Thus, size may be strictly above α. Third, consider the bivariate example with $H_0\colon \theta_1 \leq 0$ or $\theta_2 \leq 0$. Let the sampling/posterior distribution be bivariate normal with unit variances and known correlation ρ. If $\rho = 1$, then $X_1 - \theta_1 = X_2 - \theta_2$ in every draw of (X_1, X_2), i.e., it becomes a one-dimensional problem. Consequently, if $\theta_1 - \theta_2 \geq 0$, then $X_1 - X_2 \geq 0$ in every draw, and we may simply test $H_0\colon \theta_2 \leq 0$ using $X_2 \sim \mathrm{N}(\theta_2, 1)$, the Bayesian test of which has size exactly α. Similarly, if $\theta_1 - \theta_2 \leq 0$, then $X_1 - X_2 \leq 0$ in every draw, and the Bayesian test of $H_0\colon \theta_1 \leq 0$ has size exactly α. Fourth, Appendix B provides

an example where the Bayesian test's size may be strictly above α or equal to α depending on the sampling/posterior distribution. Another (less practically interesting) example is if $H_0\colon \theta_2 \neq -\theta_1$ or $\theta_1 = \theta_2 = 0$. Using the bivariate normal distribution above, if $\rho = -1$ and $\theta_1 = \theta_2 = 0$, then the problem reduces to a point null hypothesis for which this Bayesian test has 100% size. Conversely, if $\rho = 1$, then the test's size is zero since the posterior probability of the complement of Θ_0 is zero given any (X_1, X_2). □

Proof of Proposition 2. In the one-sample case, using (6),

$$\begin{aligned} \mathrm{P}(F_X \text{ SD}_1 \ F_0 \mid X(\cdot) = 0(\cdot)) &= \mathrm{P}(\theta(\cdot) \leq 0(\cdot) \mid X(\cdot) = 0(\cdot)) = \mathrm{P}(B(F_X(\cdot)) \leq 0(\cdot)) \\ &= \mathrm{P}(B(\cdot) \leq 0(\cdot)) = \mathrm{P}(\sup_{t\in[0,1]} B(t) \leq 0) = 0. \end{aligned} \tag{19}$$

The final equality holds because the distribution of the supremum of a mean-zero Brownian bridge is continuous and has non-negative support; e.g., see Theorem 2 in Smirnov (1939) or equation (1.1) in Birnbaum and Tingey (1951).

For two-sample testing, the result in (19) extends readily if we assume $F_X = F_Y$ since then $B_1(F_X(\cdot)) - B_2(F_Y(\cdot)) = B_1(F(\cdot)) - B_2(F(\cdot)) \overset{d}{=} \sqrt{2}B(F(\cdot))$ for another independent Brownian bridge $B(\cdot)$. More generally,[9] let

$$T(\cdot) \equiv B_1(F_X(\cdot)) - B_2(F_Y(\cdot)),$$

the distribution of $\theta(\cdot)$ conditional on $X(\cdot) = 0(\cdot)$. Using (9),

$$\begin{aligned} \mathrm{P}(F_X \text{ SD}_1 \ F_Y \mid X(\cdot) = 0(\cdot)) &= \mathrm{P}(\theta(\cdot) \leq 0(\cdot) \mid X(\cdot) = 0(\cdot)) = \mathrm{P}(T(\cdot) \leq 0(\cdot)) \\ &= \mathrm{P}(\sup_{r\in\mathbb{R}} T(r) \leq 0). \end{aligned} \tag{20}$$

Let L denote the smaller of the lower bounds of the distributions F_X and F_Y, allowing $L = -\infty$ if both have unbounded support. Let $W(\cdot)$, $W_1(\cdot)$, and $W_2(\cdot)$ denote independent standard Brownian motion processes. We may write

$$\begin{gathered} B_1(t) = W_1(t) - tW_1(1), \quad B_2(t) = W_2(t) - tW_2(1), \\ V(\cdot) \equiv W_1(F_X(\cdot)) - W_2(F_Y(\cdot)) \overset{d}{=} \sqrt{2}W((F_X(\cdot) + F_Y(\cdot))/2), \\ Z(\cdot) \equiv V(\cdot) - T(\cdot) = F_X(\cdot)W_1(1) - F_Y(\cdot)W_2(1). \end{gathered}$$

Looking at $T(r) = V(r) - Z(r)$ as $r \downarrow L$, the $Z(r)$ becomes negligibly small, while the $V(r)$

[9] Thanks to Iosif Pinelis for help extending to $F_X \neq F_Y$: `https://mathoverflow.net/a/292716/120669`

varies sufficiently to attain a strictly positive supremum almost surely. Specifically,

$$\lim_{r \downarrow L} \frac{Z(r)}{\sqrt{F_X(r) + F_Y(r)}} = 0,$$

so continuing from (20) with the $\leq$ changed to $>$,

$$\begin{aligned}
\mathrm{P}(\sup_{r \in \mathbb{R}} T(r) > 0) &\geq \mathrm{P}\left(\limsup_{r \downarrow L} \frac{T(r)}{\sqrt{F_X(r) + F_Y(r)}} = \infty \right) \\
&= \mathrm{P}\left(\limsup_{r \downarrow L} \frac{V(r) - Z(r)}{\sqrt{F_X(r) + F_Y(r)}} = \infty \right) \\
&= \mathrm{P}\left(\limsup_{r \downarrow L} \frac{V(r)}{\sqrt{F_X(r) + F_Y(r)}} = \infty \right) \\
&= \mathrm{P}\left(\limsup_{r \downarrow L} \frac{\sqrt{2}}{\sqrt{F_X(r) + F_Y(r)}} W\left(\frac{F_X(\cdot) + F_Y(\cdot)}{2} \right) = \infty \right) = 1
\end{aligned}$$

by the (local) law of iterated logarithm.[10] □

Proof of Proposition 3. We show that the type I error rate is one when $\theta(\cdot) = 0$, which directly implies the size is one, too. For the two-sample case, $\theta(\cdot) = 0(\cdot)$ implies $F_X(\cdot) = F_Y(\cdot)$, so the limit experiment simplifies since $B_1(F_X(\cdot)) - B_2(F_Y(\cdot)) \stackrel{d}{=} \sqrt{2} B(F(\cdot))$ for $F(\cdot) = F_X(\cdot) = F_Y(\cdot)$ and standard Brownian bridge $B(\cdot)$. Thus, both one-sample and two-sample limiting distributions can be written as $cB(\cdot)$, where $c = 1$ for one-sample and $c = \sqrt{2}$ for two-sample.

The Bayesian test rejects when the posterior is below α, so the probability of *not* rejecting when $\theta(\cdot) = 0$ is

$$\begin{aligned}
&\mathrm{P}(X(\cdot) \in \{x(\cdot) : \mathrm{P}(\theta(\cdot) \leq 0(\cdot) \mid X(\cdot) = x(\cdot)) > \alpha\} \mid \theta(\cdot) = 0(\cdot)) \\
&= \mathrm{P}(X(\cdot) \in \{x(\cdot) : \mathrm{P}(cB(F(\cdot)) \leq -x(\cdot)) > \alpha\} \mid \theta(\cdot) = 0(\cdot)).
\end{aligned}$$

This can be shown to be zero via the unconditional probability

$$\mathrm{P}(cB(F(\cdot)) + cB(F(\cdot)) \leq 0(\cdot)) = \mathrm{P}(\sqrt{2} cB(F(\cdot)) \leq 0(\cdot)) = \mathrm{P}(B(\cdot) \leq 0(\cdot)) = 0, \tag{21}$$

again using (19), where $B_1(\cdot)$, $B_2(\cdot)$, and $B(\cdot)$ are independent standard Brownian bridges.

[10] E.g., Corollary 5.3 in `https://www.stat.berkeley.edu/~peres/bmbook.pdf`

For any set S,

$$\begin{aligned}
&\mathrm{P}(\theta(\cdot) \leq 0(\cdot) \mid X(\cdot) \in S)\,\mathrm{P}(X(\cdot) \in S) \\
&= \mathrm{P}(cB_1(F(\cdot)) + X(\cdot) \leq 0(\cdot) \mid X(\cdot) \in S)\,\mathrm{P}(X(\cdot) \in S \mid \theta(\cdot) = 0) \\
&= \mathrm{P}(cB_1(F(\cdot)) + X(\cdot) \leq 0(\cdot) \text{ and } X(\cdot) \in S \mid \theta(\cdot) = 0) \\
&\leq \mathrm{P}(cB_1(F(\cdot)) + X(\cdot) \leq 0(\cdot) \mid \theta(\cdot) = 0) = \mathrm{P}(cB_1(F(\cdot)) + cB_2(F(\cdot)) \leq 0(\cdot) \mid \theta(\cdot) = 0) \\
&= 0
\end{aligned}$$

by (21), where $B_2(F(\cdot))$ is the sampling distribution of $X(\cdot)$ given $\theta(\cdot) = 0(\cdot)$, and in the posterior $\theta(\cdot) \sim B_1(F(\cdot)) + X(\cdot)$. Consequently,

$$\mathrm{P}(\theta(\cdot) \leq 0(\cdot) \mid X(\cdot) \in S)\,\mathrm{P}(X(\cdot) \in S) \leq 0. \tag{22}$$

Specifically, let S be the complement of the test's rejection region:

$$S = \{x(\cdot) : \mathrm{P}(\theta(\cdot) \leq 0(\cdot) \mid X(\cdot) = x(\cdot)) > \alpha\}.$$

If $\mathrm{P}(X(\cdot) \in S) > 0$, then the left-hand side of (22) is the product of two strictly positive terms (assuming $\alpha > 0$), which is strictly positive. This contradicts (22) since the right-hand side is zero. Consequently, $\mathrm{P}(X(\cdot) \in S) = 0$ and thus $\mathrm{P}(X(\cdot) \notin S) = 1$, i.e., the rejection probability is one. (This does not mean S is empty, just that it is a zero-probability set.) □

Proof of Corollary 4. With $\theta(\cdot) = 0(\cdot)$, by Proposition 3, the probability of rejecting SD1 is 100%; that is, the posterior probability of SD1 is below α with probability one (with respect to the distribution of $X(\cdot)$). Since the posterior probabilities of SD1 and non-SD1 sum to one, this implies the posterior probability of non-SD1 is above $1 - \alpha$ (and thus the test does not reject) with probability one. □

B Example: bivariate normal sign equality test

Here, we provide additional mathematical and simulation details for one of the examples mentioned in Section 3.1. Consider a bivariate normal model where the null hypothesis is that the two parameters have the same sign (letting zero count either way), $H_0\colon \theta_1\theta_2 \geq 0$. Specifically,

$$\mathbf{X} = (X_1, X_2)' \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} \equiv (\theta_1, \theta_2)', \quad \boldsymbol{\Sigma} \equiv \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{23}$$

By symmetry, the test's size is the supremum of the test's rejection probability over $\theta_2 \in (-\infty, 0]$ with $\theta_1 = 0$. The parameter values leading to the biggest rejection probability depend on ρ; e.g., when $\rho = 1$, the supremum comes from $\theta_2 \to -\infty$, whereas when $\rho = -1$, it comes at $\theta_2 = 0$.

To compute the Bayesian test, the posterior probability of H_0 given any (X_1, X_2) must be computed. Let the joint, conditional, and marginal PDFs of the $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution be

$$\begin{aligned}
\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) : f(t_1, t_2) &\equiv \det(2\pi\boldsymbol{\Sigma})^{-1/2} \exp\left\{-\frac{(t_1, t_2)\Sigma^{-1}(t_1, t_2)'}{2}\right\} \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(1-\rho^2)}\right\}, \\
\mathrm{N}(\rho t_2, 1-\rho^2) : f(t_1 \mid t_2) &\equiv \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\{-(t_1 - \rho t_2)^2/(2-2\rho^2)\}, \\
\mathrm{N}(0,1) : f(t_2) &\equiv \frac{1}{\sqrt{2\pi}} \exp\{-t_2^2/2\} = \phi(t_2),
\end{aligned} \tag{24}$$

where $\phi(\cdot)$ is the standard normal PDF. Given (24),

$$\int_{-\infty}^{-X_1} f(t_1 \mid t_2)\, dt_1 = \Phi\left(\frac{-X_1 - \rho t_2}{\sqrt{1-\rho^2}}\right), \tag{25}$$

where $\Phi(\cdot)$ is the standard normal CDF. Using the prior expressions and the symmetry $f(t_1, t_2) = f(-t_1, -t_2) = f(t_2, t_1)$,

$$\begin{aligned}
\mathrm{P}(H_0 \mid X_1, X_2) &= \int_{-\infty}^{-X_2}\int_{-\infty}^{-X_1} f(t_1, t_2)\, dt_1\, dt_2 + \int_{-X_2}^{\infty}\int_{-X_1}^{\infty} f(t_1, t_2)\, dt_1\, dt_2 \\
&= \int_{-\infty}^{-X_2}\int_{-\infty}^{-X_1} f(t_1 \mid t_2) f(t_2)\, dt_1\, dt_2 + \int_{-X_2}^{\infty}\int_{-X_1}^{\infty} f(t_1 \mid t_2) f(t_2)\, dt_1\, dt_2 \\
&= \int_{-\infty}^{-X_2} \Phi\left(\frac{-X_1 - \rho t_2}{\sqrt{1-\rho^2}}\right)\phi(t_2)\, dt_2 + \int_{-X_2}^{\infty}\left[1 - \Phi\left(\frac{-X_1 - \rho t_2}{\sqrt{1-\rho^2}}\right)\right]\phi(t_2)\, dt_2.
\end{aligned}$$

Unfortunately, there is no closed-form expression for this $\mathrm{P}(H_0 \mid X_1, X_2)$. However, it is easily simulated. The function $\Phi(\cdot)$ is available in any modern statistical software (e.g., R). After drawing many $Z_j \overset{iid}{\sim} \mathrm{N}(0,1)$ for $j = 1, \ldots, J$, letting $W_j \equiv \Phi((-X_1 - \rho Z)/\sqrt{1-\rho^2})$ if $Z < -X_2$ and $W_j \equiv 1 - \Phi((-X_1 - \rho Z)/\sqrt{1-\rho^2})$ if $Z \geq -X_2$, then $\mathrm{P}(H_0 \mid X_1, X_2) \approx J^{-1}\sum_{j=1}^{J} W_j$, with the approximation error going to zero as $J \to \infty$.

Table 4 shows type I error rates of the Bayesian test of $H_0\colon \theta_1\theta_2 \geq 0$ for different ρ and different θ_2, with $\theta_1 = 0$. When $\rho = -1$, it reduces to a one-dimensional setting with a finite, convex Θ_0, as in Theorem 1(iii); size equals one, well above α. When $\rho = 1$, it is also essentially one-dimensional, but with non-convex Θ_0, as in Theorem 1(iv); the type I

Table 4: Bayesian test type I error rates, $H_0\colon \theta_1\theta_2 \geq 0$, $\theta_1 = 0$, $\alpha = 0.1$, 10,000 replications.

	ρ						
θ_2	-1	-0.99	-0.9	-0.5	0	0.5	1
0.00	1.000	0.966	0.264	0.058	0.011	0.000	0.000
-0.25	1.000	0.717	0.253	0.059	0.011	0.000	0.000
-0.50	0.253	0.305	0.221	0.062	0.014	0.001	0.000
-4.00	0.100	0.102	0.102	0.101	0.099	0.094	0.093
-10.00	0.100	0.102	0.102	0.101	0.100	0.100	0.100
Max	1.000	0.966	0.264	0.101	0.100	0.100	0.100

error rate is (near) zero at many points on the boundary of Θ_0, but size equals α. When $\rho \in (-1, 1)$, we are also in the setting of Theorem 1(iv); size is strictly above α when ρ is close to -1, but decreases to α somewhere between $\rho = -0.9$ and $\rho = -0.5$.

C Derivation of translog constraints

The Hessian is a nonlinear function of the translog parameters, and it depends on $(y, \mathbf{w})$. Letting[11]

$$r_k \equiv \frac{\partial \ln(C)}{\partial \ln(w_k)} = a_{yk} \ln(y) + b_k + \sum_{j=1}^{3} b_{jk} \ln(w_j), \tag{26}$$

a general element of $\underline{\mathbf{H}}$ is

$$\begin{aligned} H_{mk} &= \frac{\partial^2 C}{\partial w_m \partial w_k} = \frac{\partial}{\partial w_m}\frac{\partial C}{\partial w_k} = \frac{\partial}{\partial w_m}(r_k C/w_k) = \frac{\partial r_k}{\partial w_m}(C/w_k) + \frac{\partial C}{\partial w_m}(r_k/w_k) + \frac{\partial w_k^{-1}}{\partial w_m}(r_k C) \\ &= (b_{mk}/w_m)(C/w_k) + r_m(C/w_m)(r_k/w_k) - \mathbb{1}\{k = m\} w_k^{-2}(r_k C) \\ &= C\frac{b_{mk} + r_m r_k - \mathbb{1}\{k = m\} r_k}{w_m w_k}. \end{aligned}$$

Since each element is proportional to $C > 0$, the value of C does not affect whether or not $\underline{\mathbf{H}}$ is NSD: $\underline{\mathbf{H}}$ is NSD iff $\underline{\mathbf{H}}/C$ is NSD. This may be helpful if the translog parameters are estimated from cost share equations and C is not directly observed.

The local NSD condition in (11) corresponds to a set of parameter values much smaller than a half-space. A necessary (but not sufficient) condition for NSD is that all the principal minors of order $p = 1$ are non-positive, i.e., that $H_{11} \leq 0$, $H_{22} \leq 0$, and $H_{33} \leq 0$. In terms of the parameters, using (26), $H_{11} \leq 0$ iff $b_{11} + r_1^2 - r_1 \leq 0$, i.e., $b_{11} \leq r_1(1 - r_1)$. With $(y, \mathbf{w}) = (1, 1, 1, 1)$, $r_k = b_k$, so $H_{11} \leq 0$ iff $b_{11} \leq b_1(1 - b_1)$. After imposing symmetry

[11] Some notation is from O'Donnell and Coelli (2005).

($b_{mk} = b_{km}$) and homogeneity of degree one in input prices ($b_{m1} + b_{m2} + b_{m3} = 0$, $m = 1, 2, 3$), all b_{mk} can be written in terms of b_{11}, b_{12}, and b_{22}: $b_{21} = b_{12}$, $b_{13} = -b_{11} - b_{12}$, etc. Also from homogeneity, $b_1 + b_2 + b_3 = 1$, and from monotonicity, $b_k = r_k \geq 0$, so $0 \leq b_1 \leq 1$. Thus, $b_1(1 - b_1) \in [0, 0.25]$, so $b_{11} \leq b_1(1 - b_1)$ is larger than the half-space defined by $b_{11} \leq 0$ but smaller than the half-space defined by $b_{11} \leq 0.25$. A similar argument for $H_{22} \leq 0$ at $(1, 1, 1, 1)$ yields $b_{22} \leq b_2(1 - b_2) \leq 0.25$. From the constraints on H_{11} and H_{22} alone, Θ_0 is a subset of the "quarter-space" defined by $b_{11} \leq 0.25$ and $b_{22} \leq 0.25$. In the notation of Theorem 1, we could use $\phi(\boldsymbol{\theta}) = b_{11}$ (or b_{22}) and $c_0 = 0.25$. Adding the constraints for the other principal minors of $\underline{\mathbf{H}}$ makes Θ_0 even smaller.

Since the local concavity H_0 in (11) corresponds to a subset of a *quarter*-space in the parameter space, Theorem 1(iii) suggests that we expect the Bayesian test's size to exceed α. The results in Table 3 show this to be the case here.

D Infinite-dimensional Bernstein–von Mises theorems

As noted, an infinite-dimensional Bernstein–von Mises theorem is not needed to satisfy Assumption A1, which only concerns the distribution of a finite-dimensional functional. However, the results and references below may be helpful or insightful in some cases.

For estimators of functions, it is common to have a (frequentist) Gaussian process limit with sample paths continuous with respect to the covariance semimetric; e.g., see van der Vaart and Wellner (1996). A natural question is whether the (asymptotic, limit experiment) sampling and posterior distributions are ever equivalent in the sense of

$$X(\cdot) - \theta(\cdot) \mid \theta(\cdot) \sim \mathbb{G}, \quad \theta(\cdot) - X(\cdot) \mid X(\cdot) \sim \mathbb{G},$$

where $\mathbb{G}$ is a mean-zero Gaussian process with known covariance function.

Unfortunately, as discussed by Freedman (1999) and others, such a Bernstein–von Mises result does not hold in great generality with infinite-dimensional spaces. As explained by Hirano and Porter (2009, p. 1696), in finite dimensions the prior often behaves locally like Lebesgue measure (if its PDF is continuous and positive at the true parameter value), but in infinite-dimensional Banach spaces there is no analog of Lebesgue measure, let alone one that most priors would satisfy.

Fortunately, some nonparametric Bernstein–von Mises theorems do exist. As in the (semi)parametric case, there are different ways to define "asymptotically equivalent distributions"; see Definition 2 in Castillo and Nickl (2014, p. 1950) for an example. The most general results to date seem to be provided by Castillo and Nickl (2013, 2014); see also Sections 12.4.1 and 12.2 of Ghosal and van der Vaart (2017).

One important special case where a Bernstein–von Mises theorem holds is for inference on a CDF. On the frequentist side, assuming iid sampling,

$$\sqrt{n}\big(\hat{F}(\cdot) - F(\cdot)\big) \rightsquigarrow B\big(F(\cdot)\big), \tag{27}$$

where $B(\cdot)$ is a standard Brownian bridge and $\rightsquigarrow$ denotes weak convergence in $\ell^\infty(\bar{\mathbb{R}})$; e.g., see van der Vaart and Wellner (1996, Ex. 2.1.3). For weak convergence under sequences $F_n(\cdot) \to F(\cdot)$, see Sections 2.8.3 and 3.11 and especially Theorem 3.10.12 in van der Vaart and Wellner (1996). For a nonparametric Bayesian method using the Dirichlet process prior of Ferguson (1973), Lo (1983, Thm. 2.1) shows that a centered (at $\hat{F}(\cdot)$) and $\sqrt{n}$-scaled version of the posterior converges to the same limit as in (27) if the prior dominates $F(\cdot)$. Even with an improper prior, i.e., using the Bayesian bootstrap of Rubin (1981), Lo (1987, Thm. 2.1) establishes the same result. A closely related result is Theorem 12.2 of Ghosal and van der Vaart (2017). An analogous conclusion is found in Theorem 4 of Castillo and Nickl (2014), but as a special case of their more general results. They can provide Bernstein–von Mises theorems for (certain) collections of integral functionals of the PDF, $\int_0^1 g_t(x) f(x)\,dx$, where $f(\cdot)$ is the PDF with support $[0,1]$ and t indexes the collection; for the CDF, $g_t(x) = \mathbb{1}\{x \le t\}$ for $t \in [0,1]$.

For the infinite-dimensional case with convergence rate n^r, similar to (2),

$$\begin{gathered} n^r\big(\hat{\mu}(\cdot) - \mu(\cdot)\big) \rightsquigarrow \mathbb{G}(\cdot), \\ X(\cdot) = n^r\big(\hat{\mu}(\cdot) - \mu_{0,n}(\cdot)\big) = n^r\big(\hat{\mu}(\cdot) - \mu(\cdot)\big) + \overbrace{n^r\big(\mu(\cdot) - \mu_{0,n}(\cdot)\big)}^{\to\theta(\cdot)} \rightsquigarrow \mathbb{G}(\cdot) + \theta(\cdot), \end{gathered} \tag{28}$$

where $\mathbb{G}(\cdot)$ is a mean-zero Gaussian process and $\theta(\cdot)$ is the local mean parameter. For the posterior, with $\theta(\cdot) = n^r\big(\mu(\cdot) - \mu_{0,n}(\cdot)\big)$ and $X(\cdot) = n^r\big(\hat{\mu}(\cdot) - \mu_{0,n}(\cdot)\big)$ analogous to (3),

$$\begin{gathered} n^r\big(\mu(\cdot) - \hat{\mu}(\cdot)\big) \rightsquigarrow \mathbb{G}(\cdot), \\ \theta(\cdot) = n^r\big(\mu(\cdot) - \mu_{0,n}(\cdot)\big) = X(\cdot) + n^r\big(\mu(\cdot) - \hat{\mu}(\cdot)\big) \rightsquigarrow \mathbb{G}(\cdot) + X(\cdot). \end{gathered} \tag{29}$$