Frequentist properties of Bayesian inequality tests*

David M. Kaplan[†]   Longhao Zhuo[‡]

April 7, 2020

Abstract

Bayesian and frequentist criteria fundamentally differ, but often posterior and sampling distributions agree asymptotically (e.g., Gaussian with same covariance). For the corresponding single-draw 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; and in other cases, size may be above, below, or equal to α. Rejection probabilities at certain points in the parameter space are also characterized. Two examples illustrate our results: translog cost function curvature and ordinal distribution relationships.

JEL classification: C11, C12

Keywords: generalized Bayes rule, limit experiment, minimax, nonstandard inference, posterior

1   Introduction

1.1   Motivation

Although Bayesian and frequentist properties fundamentally differ, often we can (approximately) achieve both. In other cases, Bayesian and frequentist summaries of the data differ greatly. We characterize the role of null hypothesis "shape" in determining such differences.

This paper came out of an empirical question, inspired by Deaton and Paxson (1998a,b): how can health inequality be assessed when only an ordinal health measure is observed?

Specifically, the ordinal, five-category "self-reported health status" is found in many health and economics datasets, but interest is usually in the underlying (latent) health. Various relationships between two latent distributions can be "identified" by various relationships between the corresponding ordinal distributions. Although ordinal first-order stochastic dominance is relatively simple (i.e., all CDF differences are non-positive), other relationships correspond to more complex subsets of the parameter space (of ordinal CDFs). Despite such complexity, it is simple to compute Bayesian posterior probabilities for all types of relationships, even simultaneously. Indeed, often such Bayesian probabilities are straightforward to compute, whereas deriving frequentist tests for general nonlinear inequalities is notoriously difficult (e.g., Wolak, 1991). This raises the question: do these posterior probabilities have any frequentist interpretation, like a p-value? Existing results comparing Bayesian and frequentist inference do not cover these more complex parameter sets; see Section 2. For details of this ordinal health example, see Section 5.1 and Kaplan and Zhuo (2019).

The ordinal health example is a special case of testing whether a certain property holds for all possible partially identified parameter values (or, at least one) within the identified set, as discussed more generally by Kline and Tamer (2016, p. 330). The ordinal distributions can be seen as point-identified parameters that map to identified sets of latent distributions. Certain ordinal distribution pairs imply that all latent distribution pairs possess some relationship (or, that at least one does). For example, under certain conditions, latent first-order stochastic dominance implies ordinal first-order stochastic dominance. Thus, ordinal non-dominance implies latent non-dominance. The hypothesis of ordinal non-dominance can be interpreted as, "There is no latent distribution pair in the identified set that satisfies first-order stochastic dominance," and ordinal dominance can be interpreted as, "There is at least one latent distribution pair in the identified set that satisfies first-order stochastic dominance."

The hypotheses considered in this paper include all the types of hypotheses about the identified set described in Kline and Tamer (2016, p. 330). In addition to the hypotheses in the prior paragraph, they mention the hypothesis that a specified value is in the identified set, which translates to the point-identified parameter being in some region. They also mention specification testing, where the null hypothesis is that the identified set is nonempty. This translates to yet another region of the point-identified parameter space, and its complement (where the identified set is empty) is the null hypothesis of *mis*specification. If the mapping from point-identified parameter to identified set is complex, these regions generally have complex shapes not covered by the existing literature comparing Bayesian and frequentist inference; see Section 2.

Inequality restrictions (often nonlinear) also feature prominently in economic theory.[1] For example, inequalities involving cumulative distribution functions characterize first-order and second-order stochastic dominance, important concepts in welfare analysis and financial portfolio comparisons. Another 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. Some examples are studied in detail in Section 5 and Appendix G. Additional economic examples like bifurcation in dynamic macroeconomic models are reviewed in Section 4.2.

If (as will be shown) Bayesian and frequentist tests of the null hypothesis can disagree while credible and confidence sets coincide, why not simply report the credible or confidence set?[2] This makes sense if interest is only in the parameter values themselves. 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 or its properties.

1.2   Contributions

Theoretically, we consider a single-draw experiment with equivalent posterior and sampling distributions. This equivalence often holds asymptotically, as in the various Bernstein–von Mises theorems in the literature. For example, after centering and scaling, a finite-dimensional parameter estimator often has an asymptotic Gaussian distribution with the same covariance matrix as the asymptotically Gaussian posterior for the same parameter. We also allow certain non-Gaussian distributions and infinite-dimensional parameters.

To quantify Bayesian–frequentist differences, we characterize the frequentist size and rejection probabilities of a particular Bayesian test. This test rejects the null hypothesis H_0 when the posterior probability of H_0 is below α, i.e., it treats the posterior like a p-value. Besides being intuitive and practically salient, there are decision-theoretic reasons to examine this test; see Section 3.1. Although size gives no insight into admissibility, it captures a practical difference between reporting Bayesian and frequentist inferences.

Specifically, we describe how the Bayesian test's size and certain rejection probabilities depend on the shape of the parameter subspace Θ_0 where the null hypothesis is satisfied. If

---

[1]Nonlinear inequalities come from other sources, too. For example, as in Kaplan (2015), $H_0\colon \theta_1\theta_2 \geq 0$ can test stability of the sign of a parameter over time (or geography), the slope coefficient's sign in IV regression, or whether a treatment attenuates the effect of another regressor.

[2]Berger (2003) also notes this possible disagreement/agreement, but 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.

Θ_0 is a half-space, then the Bayesian test has size α, with rejection probability α at every point on the boundary of Θ_0. If Θ_0 is strictly smaller than a half-space (in a certain sense), then the Bayesian test's size is strictly above α, with rejection probability strictly above α at certain boundary points of Θ_0 (like support points). If Θ_0 is not contained within any half-space, then the Bayesian test's size may be above, equal to, or below α, with rejection probability strictly below α at certain boundary points of Θ_0 (like support points of its complement).

Paper structure and notation   Section 2 reviews the literature. Section 3 presents the setup and assumptions. Section 4 contains our main results and discussion. Section 5 illustrates our results with economic examples. Appendix A contains proofs. Appendices B–G contain additional details and examples. 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 single crossing (SC). Notationally, $\subseteq$ is subset and $\subset$ is proper subset; $\supseteq$ is superset and $\supset$ is proper superset; $\mathcal{A}^{\complement}$ is the complement of set $\mathcal{A}$, and $\overline{\mathcal{A}}$ the closure; 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   Literature

Many papers compare Bayesian and frequentist inference. Here, we highlight examples of different types of conclusions (not all directly comparable to inequality testing): sometimes frequentist inference is more conservative, sometimes Bayesian, sometimes neither. Then, we discuss three more closely related papers in more detail.

2.1   Bayesian–frequentist comparisons

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 $\boldsymbol{\theta} \in \mathbb{R}^d$, $d > 1$, Kline (2011) shows frequentist testing to be more conservative, especially with large d; see Section 2.2 for details. As another example, under set identification, asymptotically, frequentist confidence sets for the true parameter (e.g., Imbens and Manski, 2004; Stoye, 2009) are strictly larger than the estimated identified set, whereas Bayesian credible sets are strictly smaller when the prior is over the structural parameter of interest,

as shown by Moon and Schorfheide (2012, Cor. 1).[3] For testing the null of a unit root in autoregression, Sims and Uhlig (1991) say frequentist tests "accept the null more easily" (p. 1592).

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 can reject when 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 $P(H_0) = 1/2$. Arguing that $P(H_0) = 1/2$ is "objective," Berger and Sellke (1987, p. 113) consider even $P(H_0) = 0.15$ to be "blatant bias toward H_1." Casella and Berger (1987b) disagree, saying $P(H_0) = 1/2$ is "much larger than is reasonable for most problems" (p. 344).

In yet other cases, Bayesian and frequentist inferences can be similar or even identical, depending on the prior. For unit root testing, Sims and Uhlig (1991) determine sample-dependent priors that equate p-values and posterior probabilities. 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 to the infimum of the posterior $P(H_0 \mid X)$ over various classes of priors. In many cases, the infimum is attained by the improper prior of Lebesgue measure on $(-\infty, \infty)$ and equals the p-value (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. They propose "an impartial Bayes rule" (Section 3.2) that sets the prior $P(H_0) = 1/2$ to avoid "undesirable bias against H_0" (p. 1270). (This also makes the prior odds $P(H_0)/P(H_1) = 1$, which simplifies the Bayes factor to the posterior odds, although they only examine the posterior probability of H_0.) If the components of $\mathbf{X}$ are mutually independent, then taking the (improper) limit as a symmetric prior's variance goes to infinity, the posterior is proportional to one of the frequentist p-values they consider, but it is (weakly) smaller (p. 1271). This complements our setting where we do not impose independence, $P(H_0) = 1/2$, finite dimensionality, or restricted null hypothesis shape.

There are also many papers showing asymptotic equivalence of Bayesian credible intervals and frequentist confidence sets, due to asymptotic equivalence of posterior and sampling distributions. Our setup has such an equivalence. Besides results for point-identified parameters, asymptotically exact frequentist coverage has been established for certain Bayesian

---

[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.

credible sets of identified sets, when the prior is (only) over some point-identified parameter and not the partially identified parameter; e.g., see Kline and Tamer (2016). However, these are less directly related to our results on inequality testing.

2.2   More closely related papers

Kline and Tamer (2016) mention many of the issues related to our paper, although they do not focus on frequentist properties of Bayesian tests. Their Sections 3 and 4 describe and interpret different statements about the identified set, but the focus is on computation and consistency of Bayesian posterior probabilities. Asymptotic approximations based on normal variables are given (Theorems 1(iii) and 3(iii,vi)) but not compared to frequentist p-values. In Section 5, in the context of a certain credible set, they raise the question of equality of posterior probability and p-value (pp. 348–349). They state this holds for interval-identified parameters but otherwise just refer to Kline (2011).

Kline (2011) studies a special case of our setting. Similar to our setting (although we allow certain non-normal distributions), his posterior is multivariate normal (Definition 1, p. 3133), with covariance matrix equal to that of the multivariate normal sampling distribution. Kline (2011) motivates this equivalence from estimating a vector of means, where either the likelihood is normal and the prior is flat, or there is a Dirichlet process prior under which the asymptotic posterior is equivalent to the asymptotic sampling distribution (Sections 2.1–2.2). Most restrictive compared to our setting, the null hypothesis is that all components of the vector are non-negative.

Most relevant to our paper are Kline's (2011) Lemmas 1 and 4 (pp. 3134, 3140). They show that if the null hypothesis is satisfied and at least two components of the mean vector equal zero (on the boundary of the null), then the asymptotic sampling distribution of the posterior probability of the null is first-order stochastically dominated by the standard uniform distribution. That is, if $\hat{\pi}$ is the posterior probability of H_0, then $P(\hat{\pi} \leq \alpha) > \alpha$ for all $\alpha \in (0, 1)$. Consequently, the test that rejects when the posterior is below α has type I error rate (and thus size) above α in such a situation. Although not discussed in the original paper, this also implies $P(\hat{\pi} \geq 1 - \alpha) < \alpha$, so the type I error rate is below α if the null and alternative hypotheses are switched. Kline (2011, §4) also provides numerical examples and simulations showing that the magnitude of the discrepancy increases with the dimension.

Kline's (2011) results are consistent with ours, but we study a more general null hypothesis. By restricting the null hypothesis, Kline (2011) can more precisely quantify the Bayesian test's rejection probabilities (Lemma 1) and can propose a frequentist test using the posterior probability of H_0 as the test statistic, but with critical value smaller than α

(which can be interpreted as an adjusted loss function, in light of our Section 3.1). Our null hypothesis is simply that the parameter lies in some subset of the parameter space, with no restrictions on the shape of the subset. Consequently, our results apply to more settings. For example, the applications in Section 5 and most others mentioned in Sections 1 and 4 are not covered by the results of Kline (2011). In particular, we allow hypotheses about identified sets without restricting the mapping from point-identified parameter to identified set.

Efron and Tibshirani (1998) consider the confidence level for a parameter $\boldsymbol{\mu}$ being in a certain region of the parameter space. They call this the "problem of regions." In particular, they focus on the confidence level of the region in which the point estimate $\hat{\boldsymbol{\mu}}$ lies. They note (p. 1698) that their "first-order bootstrap" confidence value is essentially a Bayesian posterior probability given a flat prior. Although their main contribution is a bootstrap adjustment to better match the "correct" frequentist confidence level, most related to our paper is their discussion of how the Bayesian (i.e., first-order bootstrap) and frequentist measures differ depending on the curvature of the boundary between the regions. Qualitatively, they note the Bayesian posterior probability is higher than the frequentist confidence level when the boundary "curves away from $\hat{\boldsymbol{\mu}}$," whereas it is lower when the boundary "curves toward $\hat{\boldsymbol{\mu}}$" (p. 1690). In terms of hypothesis testing, this implies the Bayesian test is more likely (than a frequentist test) to reject the null when the boundary of the null region "curves away from $\hat{\boldsymbol{\mu}}$," i.e., has a convex boundary near $\hat{\boldsymbol{\mu}}$. Similarly, it implies the Bayesian test is less likely to reject if the null region's boundary is non-convex near $\hat{\boldsymbol{\mu}}$. Efron and Tibshirani (1998) also discuss the better agreement of Bayesian and frequentist confidence when i) the boundary can be written as a level set of a smooth function of the underlying parameter and ii) the Welch–Peers prior (Welch and Peers, 1963) can be used (e.g., §4 or Remark B on p. 1714).

However, the theoretical results in Efron and Tibshirani (1998) stipulate an asymptotic approximation local to a smooth boundary of the null region. More precisely, up to smaller-order terms, the boundary is treated as quadratic, as in their (2.10). This greatly simplifies matters; e.g., the distance from $\hat{\boldsymbol{\mu}}$ to the nearest boundary point has a normal distribution with mean and variance related to the quadratic coefficient matrix, as in their (2.12). Asymptotically, the smooth boundary looks nearly flat, with only $O(n^{-1/2})$ curvature (p. 1692 before (2.14)). Consequently, even with no adjustment, Bayesian and frequentist confidence levels differ by merely $O_p(n^{-1/2})$; see (2.24) and following text (p. 1694). Further, the approximate connection with Welch–Peers priors in Section 4 assumes the boundary is a level set of some scalar-valued function of $\boldsymbol{\mu}$. In contrast, our results do not restrict the boundary of the null hypothesis region in any way. This is important since many null hypotheses entail non-smooth boundaries, such as joint tests of multiple inequalities (as in Kline, 2011). Of

course, the benefit of their stronger assumptions is that Efron and Tibshirani (1998) can more precisely approximate the quantitative difference between frequentist confidence and Bayesian posterior probability, as in their (2.21) or (2.24).

3 Setup and assumptions

In Section 3.1, a specific Bayesian hypothesis test is described along with the decision-theoretic context. In Section 3.2, the assumptions used for the results in Section 4 are presented and discussed. Section 3.3 discusses an alternative approach for this setting.

3.1 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 $\mathrm{P}(H_0 \mid \text{data}) \leq \alpha$; otherwise, accept H_0.

Besides seeming intuitive, Method 1 is a generalized Bayes decision rule. Specifically, it 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. Letting $H = 0$ if H_0 is true and $H = 1$ if H_0 is false, the loss function is

$$
\begin{aligned}
L(H = 0, \text{reject}) &= 1 - \alpha, \quad L(H = 1, \text{accept}) = \alpha, \\
L(H = 1, \text{reject}) &= L(H = 0, \text{accept}) = 0.
\end{aligned} \tag{1}
$$

Thus, given a particular decision but unknown H,

$$
\begin{aligned}
L(H, \text{reject}) &= (1 - \alpha)\, \mathbb{1}\{H = 0\} + (0)\, \mathbb{1}\{H = 1\} = (1 - \alpha)\, \mathbb{1}\{H_0 \text{ true}\}, \\
L(H, \text{accept}) &= (0)\, \mathbb{1}\{H = 0\} + (\alpha)\, \mathbb{1}\{H = 1\} = \alpha\, \mathbb{1}\{H_0 \text{ false}\}.
\end{aligned} \tag{2}
$$

To compute PEL, let $\mathrm{P}(\cdot \mid \mathbf{X})$ denote the posterior probability given observed data $\mathbf{X}$, and $\mathrm{E}(\cdot \mid \mathbf{X})$ the posterior expectation. The PEL for the decision to reject H_0 is

$$
\mathrm{E}[L(H, \text{reject}) \mid \mathbf{X}] = \mathrm{E}[\overbrace{(1 - \alpha)\, \mathbb{1}\{H_0 \text{ true}\}}^{\text{from (2)}} \mid \mathbf{X}] = (1 - \alpha)\, \mathrm{P}(H_0 \mid \mathbf{X}), \tag{3}
$$

i.e., the type I error loss $(1 - \alpha)$ times the posterior probability that H_0 is true and thus "reject" is a type I error. Similarly, the PEL of accepting H_0 is

$$
\mathrm{E}[L(H, \text{accept}) \mid \mathbf{X}] = \mathrm{E}[\overbrace{\alpha\, \mathbb{1}\{H_0 \text{ false}\}}^{\text{from (2)}} \mid \mathbf{X}] = \alpha[1 - \mathrm{P}(H_0 \mid \mathbf{X})], \tag{4}
$$

i.e., the type II error loss (α) times the probability that H_0 is false and thus "accept" 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; i.e., (3) is smaller than (4) if and only if $\mathrm{P}(H_0 \mid \mathbf{X}) \leq \alpha$, with equality when $\mathrm{P}(H_0 \mid \mathbf{X}) = \alpha$.

The Bayesian test's type I error rates and size are compared to α (instead of some other value) for both 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, given the same loss function from (1) under which the Bayesian test in Method 1 is a generalized Bayes rule, any unbiased frequentist test with size α is a minimax risk decision rule; see Appendix C.

Ideally, a single decision rule minimizes both maximum risk and PEL. However, if the Bayesian test's size is above or below α, this is not possible.

3.2   Assumptions

The two main assumptions are stated and discussed in Sections 3.2.1 and 3.2.2.

3.2.1   Posterior and sampling distributions

Assumption A1 states conditions on the sampling and posterior distributions of a functional $\phi(\cdot)$ in the single-draw 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 single 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 $\phi(\boldsymbol{\theta}) - \phi(\mathbf{X}) \mid \mathbf{X} \sim F$, for every $\boldsymbol{\theta}$ and every $\mathbf{X}$ in the space.

The hope is that the conditions in A1 are reasonable approximations of the conditions encountered in practice. The use of a single draw of $\mathbf{X}$ has precedent in papers like Casella and Berger (1987a) and Goutis et al. (1996). Usually, $\mathbf{X}$ in A1 represents some estimator $\hat{\boldsymbol{\theta}}$. For example, in a particular empirical setting, if the estimator (centered at the true parameter) and posterior (centered at the estimate) are both approximately Gaussian with approximately the same covariance, then our results should provide some insight. The conditions in A1 are now further discussed.

Assumption A1 includes the common case of a Gaussian distribution. In $\mathbb{R}^d$, continuous linear functionals are simply linear combinations $\phi(\mathbf{X}) = \mathbf{c}'\mathbf{X}$ for constant vector $\mathbf{c} \in \mathbb{R}^d$. If $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\theta}, \underline{\mathbf{V}})$ and $\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{X}, \underline{\mathbf{V}})$, then the sampling distribution of $\mathbf{c}'\mathbf{X} - \mathbf{c}'\boldsymbol{\theta}$ and the posterior distribution of $\mathbf{c}'\boldsymbol{\theta} - \mathbf{c}'\mathbf{X}$ are both $\mathrm{N}(0, \mathbf{c}'\underline{\mathbf{V}}\mathbf{c})$, satisfying the equivalence and symmetry conditions in A1. In 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). For example, given Gaussian process $X(\cdot)$ and $r, r_1, r_2 \in \mathbb{R}$, both $\phi(X(\cdot)) = X(r)$ and $\phi(X(\cdot)) = X(r_2) - X(r_1)$ are Gaussian (relevant for first-order stochastic dominance and monotonicity, respectively).

The equivalence of posterior and sampling distributions in A1 requires that the prior not influence the posterior. Implicitly, A1 entails an improper uninformative (flat) prior. For example, take scalar $X, \theta \in \mathbb{R}$, so $\phi(X) = X$. With sampling distribution $X \mid \theta \sim \mathrm{N}(\theta, 1)$ and prior $\theta \sim \mathrm{N}(m, \tau^2)$, the posterior is

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

The equivalence condition is satisfied (only) in the limit as $\tau^2 \to \infty$, yielding the posterior $\theta \mid X \sim \mathrm{N}(X, 1)$. Alternatively, the improper "flat" prior (equal to 1 for all θ) yields the same posterior. This happens because the posterior is proportional to the likelihood times the prior, so the flat prior disappears and leaves the posterior proportional to the likelihood, which in this example is $(2\pi)^{-1/2} \exp\{-(X - \theta)^2/2\}$. Interpreted as a posterior, this gives the PDF of θ given X as that of a $\mathrm{N}(X, 1)$ distribution. This improper prior is fine for Method 1 because only posterior probabilities are used, unlike with Bayes factors that involve a ratio of prior probabilities (e.g., Bayarri, Berger, Forte, and García-Donato, 2012).

The uninformative prior implicit in A1 is restrictive but often relevant. It is restrictive because in practice the prior may influence the posterior greatly. Even in large samples, a sufficiently strong prior can always influence the posterior. However, to gain any general understanding, there must be some restriction on the prior, otherwise it can arbitrarily affect the results (e.g., a dogmatic prior on a point inside Θ_0, or outside). Further, in practice, often the prior does not influence the posterior much. This may be due to a large sample and/or an uninformative prior on $\boldsymbol{\theta}$. For example, in the application to partially identified parameters in a binary entry game, Kline and Tamer (2016, p. 355) specify "an uninformative conjugate Dirichlet prior," similar to that in our application in Section 5.1. Finally, our results can be seen as trying to isolate the effect of the Bayesian or frequentist framework, apart from any particular prior.

Assumption A1 can be interpreted asymptotically. The asymptotic equivalence of posterior and sampling distributions is the conclusion of a Bernstein–von Mises theorem. Many Bernstein–von Mises theorems have been established in different parametric, semiparametric, and even nonparametric settings; see Appendix B. Most directly, such results could approximately justify the posterior and sampling distribution equivalence in A1, where $\boldsymbol{\theta}$ is the parameter of interest estimated by $\hat{\boldsymbol{\theta}} = \mathbf{X}$. Alternatively, if A1 describes a limit experiment, then $\boldsymbol{\theta}$ is a *local* mean parameter. Then, the "size" of the test is really the "local size," i.e., the supremum type I error in a local neighborhood of some parameter value; but statements about rejection probability retain the same interpretation. Also, the shape of the null hypothesis region may change substantially when "zooming in" on a particular local neighborhood; see below for further discussion.

The following is a simple example of an asymptotic limit experiment with a local parameter. Let $X, \theta \in \mathbb{R}$ with $\phi(v) = v$. If $Y_{ni} \overset{iid}{\sim} \mathrm{N}(\mu_n, 1)$, $i = 1, \ldots, n$, and $\sqrt{n}\mu_n \to \theta$, then $\sqrt{n}\bar{Y}_n = n^{-1/2} \sum_{i=1}^n Y_{ni} \overset{d}{\to} \mathrm{N}(\theta, 1)$. More generally, if $Y_{ni} \overset{iid}{\sim} \mathrm{N}(m + \mu_n, \sigma^2)$ and $\sqrt{n}\mu_n \to \theta$, then

$$X_n \equiv \sqrt{n}(\bar{Y}_n - m)/\hat{\sigma} \overset{d}{\to} X \sim \mathrm{N}(\theta, 1) \tag{5}$$

given consistent estimator $\hat{\sigma}^2 \overset{p}{\to} \sigma^2$. 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. 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)).

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 (not drifting) DGPs. For example, in $\mathbb{R}^d$,

$$\mathbf{X}_n = \sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{0,n}) = \overset{\overset{d}{\to}\mathrm{N}(\mathbf{0},\boldsymbol{\Sigma})}{\overbrace{\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})}} + \overset{=\boldsymbol{\theta}_n \to \boldsymbol{\theta}}{\overbrace{\sqrt{n}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0,n})}} \overset{d}{\to} \overset{\text{limit experiment}}{\overbrace{\mathbf{X} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})}}, \tag{6}$$

where $\boldsymbol{\mu}$ is the true parameter value, $\hat{\boldsymbol{\mu}}$ is a $\sqrt{n}$-normal estimator (as seen), $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix (known or consistently estimable), 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

11

of $\boldsymbol{\mu}_0$ being "close to" the true $\boldsymbol{\mu}$ in the asymptotics. Analogously, for the posterior, letting $\boldsymbol{\theta}_n = \sqrt{n}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0,n})$ and again $\mathbf{X}_n = \sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{0,n})$, assuming a Bernstein–von Mises theorem,

$$\boldsymbol{\theta}_n - \mathbf{X}_n = \sqrt{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}}). \tag{7}$$

3.2.2 Shape of null hypothesis region

Assumption A2. The null hypothesis is $H_0 \colon \boldsymbol{\theta} \in \Theta_0$, where Θ_0 is a subspace of the Banach space in A1; the alternative hypothesis is that $\boldsymbol{\theta} \notin \Theta_0$.

Assumption A2 is very general. The shape of Θ_0 is not restricted at all. It does assume each $\boldsymbol{\theta}$ satisfies either the null or alternative hypothesis (i.e., there is no third region), but this is often true.

In practice, it matters whether $\boldsymbol{\theta}$ is interpreted as the original parameter or as a local parameter as in (5) or (7). In our results, the shape of Θ_0 matters; should this be the "global shape" of the null hypothesis region, or the "local shape" in a neighborhood of the true parameter? The answer is not obvious. In practice, one idea is to construct something like a 99.9% credible set or confidence set to get an idea of the "local" shape, where "local" is based on the data, similar in spirit to the idea of Romano, Shaikh, and Wolf (2014).

Consider some examples in which the relevant shape of Θ_0 depends on the sample size as well as the true parameter value. Let $\mathbf{X}, \boldsymbol{\theta} \in \mathbb{R}^2$. First, let $H_0 \colon \theta_1 \leq 0, \theta_2 \leq 0$. Thus, Θ_0 is the third quadrant, with a convex corner at the origin and flat sides extending down each axis to $-\infty$. If we look at a neighborhood of the origin (or equivalently scale by $\sqrt{n}$), the shape is identical; no matter how much we zoom in, the null hypothesis is the third quadrant since $(\theta_1, \theta_2) \leq (0,0) \iff (\sqrt{n}\theta_1, \sqrt{n}\theta_2) \leq (0,0)$. However, consider a neighborhood of $(\theta_1, \theta_2) = (-m, 0)$. Perhaps with small sample size n, the standard deviation is much larger than m, so the original Θ_0 shape with the corner is appropriate. But with large enough n, as the standard deviation shrinks to zero, a $n^{-1/2}$-neighborhood around $(-m, 0)$ no longer includes the corner at the origin. Instead, the null hypothesis boundary appears flat. That is, it becomes obvious that $\theta_1 < 0$, so the null hypothesis reduces to $H_0 \colon \theta_2 \leq 0$. This is the same idea behind moment inequality testing procedures like in Romano et al. (2014). They hope to learn which inequalities might be binding, which is equivalent to the local shape of Θ_0.

Consider a variant of the above example. Remove a notch from Θ_0 near the origin: if $\theta_1 > -r$ and $\theta_2 > -r$, then the point (θ_1, θ_2) is excluded from Θ_0 even if $(\theta_1, \theta_2) \leq (0,0)$. If r is very small compared to the standard deviation, then this notch's effect is negligible; Θ_0 still looks like the third quadrant. However, if the sample size is large enough that r is much bigger than the standard deviation, a local neighborhood of $(-r, -r)$ looks very different:

the null hypothesis is satisfied in quadrants 2, 3, and 4, not just quadrant 3. Instead of a convex corner, the null region has a non-convex corner, which leads to very different behavior of the Bayesian test.

This difference between global and local shape (and its practical importance) is illustrated further in the example of Section 5.1.

3.3   Least favorable prior approach

Comparing Bayes and minimax decision rules evokes the least favorable prior. Most simply, if there exists a (least favorable) prior such that the Bayes risk of the corresponding Bayes rule equals the supremum risk (over the parameter space) of the same Bayes rule, then that rule is minimax (Lehmann and Casella, 1998, Thm. 1.4 in Ch. 5, p. 310). More generally, there is a least favorable *sequence* of priors, with a corresponding sequence of Bayes rules and their Bayes risks (against the prior); any decision rule whose supremum risk equals the limit of any such Bayes risk sequence is minimax (Lehmann and Casella, 1998, Thm. 1.12 in Ch. 5, p. 316).

Deriving a least favorable prior (sequence) would be valuable, although there are two reasons we do not pursue it in depth here. First, for hypothesis testing, the (limiting) least favorable prior is never the flat prior implicit in Assumption A1. Even if such a prior is found, it is still helpful to understand the Bayesian test's behavior under A1. Second, for very general problems like in Assumption A2, minimax results are notoriously difficult to obtain. For example, before discussing least favorable priors, Lehmann and Casella (1998, §5.1, p. 309) remark, "We shall not be able to determine minimax estimators for large classes of problems but, rather, will treat problems individually." (Hypothesis testing is essentially "estimation" of the population object $\mathbb{1}\{\boldsymbol{\theta} \notin \Theta_0\}$.)

For reference, a simple example is worked out in Appendix D. The null hypothesis is $H_0: \theta \leq 0$, with scalar $X \sim \mathrm{N}(\theta, 1)$. Despite the simple setup, it is more complicated than a simple estimation problem, where any sequence of priors approaching the flat (improper) prior is least favorable. Unlike with estimation, the testing decision is most difficult near $\theta = 0$, so the least favorable prior sequence must (in the limit) concentrate mass near that point. Otherwise, if for example there is significant prior mass on $\theta > 10$, it is obvious that the test should reject in such cases, which lowers the posterior expected loss below the minimax risk. The prior also cannot put all mass on $\theta = 0$, otherwise the Bayesian test never rejects and has zero Bayes risk. One least favorable prior sequence places mass on only two points: $\theta = 0$, and a strictly positive value. In the limit, the strictly positive value converges to zero, and the prior probability of $\theta = 0$ converges to α in a particular way (depending on

the value of the positive support point). We show the limit of the corresponding Bayes risk sequence equals the minimax risk $\alpha(1 - \alpha)$ from Appendix C. Details and discussion are in Appendix D.

Although the example in Appendix D works out intuitively, it is unclear how well this approach generalizes. Nonetheless, it is worth investigation in future work.

4 Results and discussion

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

Theorem 1. *Let Assumptions A1 and A2 hold. Consider the Bayesian test of Method 1.*

(i) *For any $\phi(\cdot)$ satisfying A1, if there exists $c_0 \in \mathbb{R}$ such that $\Theta_0 = \{\boldsymbol{\theta} : \phi(\boldsymbol{\theta}) \le c_0\}$ (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) *For any $\phi(\cdot)$ satisfying A1, if there exists $c_0 \in \mathbb{R}$ such that $\Theta_0 \subseteq \{\boldsymbol{\theta} : \phi(\boldsymbol{\theta}) \le c_0\}$ 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), further assume that there exists a $\phi_2(\cdot)$ satisfying $\phi_2(\mathbf{X}) - \phi_2(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \sim F_2$ and $\phi_2(\boldsymbol{\theta}) - \phi_2(\mathbf{X}) \mid \mathbf{X} \sim F_2$ for every $\boldsymbol{\theta}$ and $\mathbf{X}$ and satisfying the properties below. Assume the $\phi(\cdot)$ from Theorem 1(ii) may be written as $\phi(\cdot) = \phi_3(\phi_2(\cdot))$ for some projection $\phi_3(\cdot) \colon \mathbb{R}^d \mapsto \mathbb{R}$. Assume there exists (in $\mathbb{R}^d$) a set $\Phi_2 \equiv \{\phi_2(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_0\} \subset \{\phi_2(\boldsymbol{\theta}) : \phi(\boldsymbol{\theta}) \le c_0\}$. Further assume (a) the set $\Delta \equiv \{\mathbf{p} : \mathbf{p} \in \mathbb{R}^d, \phi_3(\mathbf{p}) \le c_0, \mathbf{p} \notin \Phi_2\}$ 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. If $\boldsymbol{\theta} \in \mathbb{R}^d$, then $\phi_2(\boldsymbol{\theta}) = \boldsymbol{\theta}$, $\phi_3(\cdot) = \phi(\cdot)$, and $\Phi_2 = \Theta_0$, and the rejection probability is strictly above α for any $\boldsymbol{\theta}$ that is a support point of the closure of Θ_0.*

(iv) *If, contrary to Theorem 1(ii), there do not exist any $\phi(\cdot)$ and c_0 such that $\Theta_0 \subseteq \{\boldsymbol{\theta} : \phi(\boldsymbol{\theta}) \le c_0\}$ (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 (in addition to the shape of Θ_0).*

(v) *Continue the notation and conditions from Theorem 1(iii), but redefine $\Phi_2 \equiv \{\phi_2(\boldsymbol{\theta}) : \boldsymbol{\theta} \notin \Theta_0\}^{\complement} \supset \{\phi_2(\boldsymbol{\theta}) : \phi(\boldsymbol{\theta}) \le c_0\}$, again with $\Phi_2 = \Theta_0$ if $\boldsymbol{\theta} \in \mathbb{R}^d$; and redefine*

$\Delta \equiv \{\mathbf{p} : \mathbf{p} \in \Phi_2, \phi_3(\mathbf{p}) > c_0\}$ *(the "extra" part of Φ_2 beyond the half-space in $\mathbb{R}^d$).* *Then, the Bayesian test's rejection probability is strictly below α when $\phi(\boldsymbol{\theta}) = c_0$.* *With $\boldsymbol{\theta} \in \mathbb{R}^d$, this implies the rejection probability is strictly below α for any $\boldsymbol{\theta}$ that is a support point of the closure of the complement of Θ_0, i.e., a support point of $\overline{\Theta_0^{\complement}}$.*

4.1 Intuition for results

Some intuition for Theorem 1 is now provided. The focus is on the finite-dimensional case, specifically $\mathbb{R}^2$; the infinite-dimensional argument essentially finds a finite-dimensional projection in which the same logic goes through. For simplicity, here the inverse CDF $F^{-1}(\cdot)$ is assumed to exist.

The one-dimensional version of Theorem 1(i) is well known but worth mentioning since it underlies all other results. The key is the symmetry of $F(\cdot)$ in Assumption A1. Without symmetry, as seen below, the Bayesian test's size would not generally equal α. Consider $H_0 \colon \theta \leq 0$. The posterior probability of H_0 is strictly decreasing in X, and it equals α when $X = -F^{-1}(\alpha)$, as shown below, so the Bayesian test rejects when $X > -F^{-1}(\alpha)$. To verify this critical value, since the posterior is $\theta - X \mid X \sim F$, and $\theta \leq 0$ can be written $\theta - X \leq -X$,

$$
\begin{aligned}
\mathrm{P}(H_0 \mid X = -F^{-1}(\alpha)) &= \mathrm{P}(\theta \leq 0 \mid X = -F^{-1}(\alpha)) \\
&= \mathrm{P}(\theta - X \leq -X \mid X = -F^{-1}(\alpha)) \\
&= F(-(-F^{-1}(\alpha))) \\
&= F(F^{-1}(\alpha)) \\
&= \alpha.
\end{aligned}
$$

The power function is increasing in θ, so the type I error rate is largest at $\theta = 0$. Since the sampling distribution of $X - \theta$ (which is just X when $\theta = 0$) also has CDF $F(\cdot)$, size is
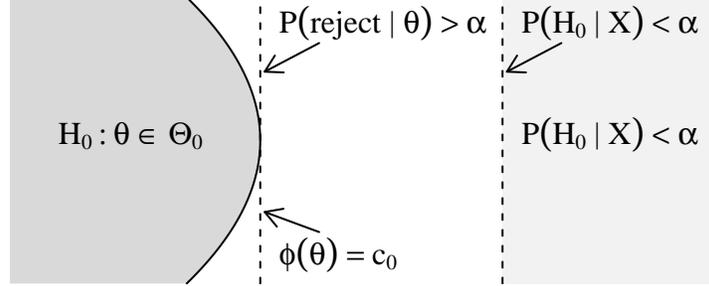
$$
\begin{aligned}
\sup_{\theta \leq 0} \mathrm{P}(X > -F^{-1}(\alpha) \mid \theta) &= \mathrm{P}(X > -F^{-1}(\alpha) \mid \theta = 0) \\
&= 1 - \mathrm{P}(X - \theta \leq -F^{-1}(\alpha) - \theta \mid \theta = 0) \\
&= 1 - F(-F^{-1}(\alpha)).
\end{aligned}
$$

Without symmetry, this does not equal α. With the symmetry condition $F(-x) = 1 - F(x)$ in A1, however, the Bayesian test's size simplifies to
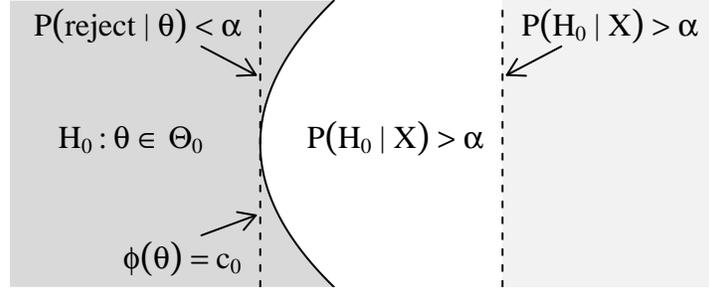
$$
1 - F(-F^{-1}(\alpha)) = 1 - [1 - F(F^{-1}(\alpha))] = \alpha.
$$
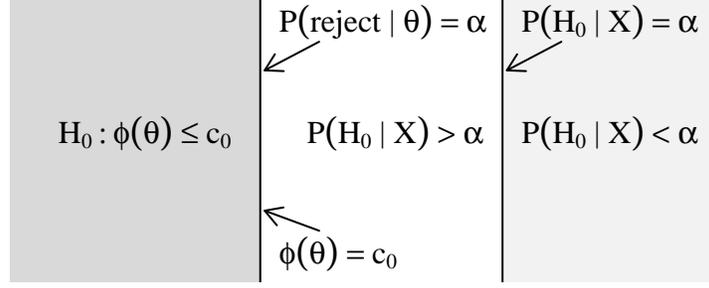
(a) Illustration of Theorem 1(i)

(b) Illustration of Theorem 1(iii)

(c) Illustration of Theorem 1(v)

Figure 1: Illustrations of Theorem 1 for $\mathbf{X}, \boldsymbol{\theta} \in \mathbb{R}^2$.

Figure 1 illustrates the intuition behind Theorem 1. Here, $\mathbf{X}, \boldsymbol{\theta} \in \mathbb{R}^2$, so $\phi_2(\boldsymbol{\theta}) = \boldsymbol{\theta}$, $\phi_3(\cdot) = \phi(\cdot)$, and $\Phi_2 = \Theta_0$.

Figure 1(a) illustrates Theorem 1(i). Here, $\mathbf{X} = (X_1, X_2)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where X_1 and θ_1 are on the horizontal axis while X_2 and θ_2 are on the vertical axis. Further, $\phi(\boldsymbol{\theta}) = \theta_1$, and imagine $c_0 = 0$ for simplicity. The left vertical line shows $\phi(\boldsymbol{\theta}) = c_0$, i.e., $\theta_1 = 0$. Thus, the null is $H_0: \theta_1 \leq 0$, reducing to the one-dimensional case. Assumption A1 says $\phi(\mathbf{X}) - \phi(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \sim F$ and $\phi(\boldsymbol{\theta}) - \phi(\mathbf{X}) \mid \mathbf{X} \sim F$, which here is $X_1 - \theta_1 \mid \boldsymbol{\theta} \sim F$ and $\theta_1 - X_1 \mid \mathbf{X} \sim F$, with symmetric $F(-x) = 1 - F(x)$. When $X_1 = -F^{-1}(\alpha)$, for any X_2, $\mathrm{P}(H_0 \mid \mathbf{X}) = \alpha$ exactly; this is indicated by the right vertical line in the figure. For smaller values of X_1 (to the left of the vertical line), $\mathrm{P}(H_0 \mid \mathbf{X}) > \alpha$, whereas for larger X_1 (to the right), $\mathrm{P}(H_0 \mid \mathbf{X}) < \alpha$, as indicated in the figure. For any $\boldsymbol{\theta}$ with $\phi(\boldsymbol{\theta}) = c_0$, i.e., with $\theta_1 = 0$

(along the left vertical line), the rejection probability $P(X_1 > -F^{-1}(\alpha) \mid \boldsymbol{\theta}) = \alpha$ exactly, as indicated in the figure, so the test's size is also α. This is essentially the one-dimensional result for X_1 and θ_1, applied to all X_2 and θ_2.

Figure 1(b) illustrates Theorem 1(iii). The setup is the same as for Figure 1(a) but with a different null hypothesis $H_0 \colon \boldsymbol{\theta} \in \Theta_0$ (the shaded region on the left). The set Δ from Theorem 1(iii) is the white area between the shaded Θ_0 and the vertical (dashed) line $\phi(\boldsymbol{\theta}) = c_0$, which again can be thought of as $\theta_1 = 0$. Consider any point on the right vertical line, where previously $P(H_0 \mid \mathbf{X}) = \alpha$ in Figure 1(a). Since Δ has positive posterior probability given any $\mathbf{X}$ on that line, H_0 must have strictly smaller posterior probability than in Figure 1(a), i.e., strictly smaller than α. Thus, the boundary of the rejection region must shift to the left by some amount, enlarging the rejection region. Consequently, for any point on the left vertical line where previously the rejection probability was exactly α, the rejection probability is now strictly above α. Most such points are now outside Θ_0 (so rejection probability is power, not type I error rate), but there is one such point on the boundary of Θ_0.

Figure 1(b) could be drawn for any support point of Θ_0, by redefining $\phi(\cdot)$. That is, consider any point on the boundary of Θ_0 that is also on a supporting hyperplane (here, tangent line) $\phi(\boldsymbol{\theta}) = c_0$, meaning Θ_0 is entirely on one side of the line. The picture will look qualitatively similar at any such point. The key is that after Δ is carved away from the supporting half-space $\phi(\boldsymbol{\theta}) \leq c_0$, then all points with $\phi(\boldsymbol{\theta}) = c_0$ have rejection probability above α, including the point on the boundary of Θ_0.

Figure 1(c) illustrates Theorem 1(v). The setup is again the same except the null hypothesis. Now Θ_0 is larger than the original half-space, so Δ is the shaded area within Θ_0 to the right of the left vertical line ($\phi(\boldsymbol{\theta}) = c_0$). Reversing the logic from before, this *increases* the posterior probability of H_0 for all points $\mathbf{X}$ on the right vertical line. This in turn shifts the boundary of the rejection region to the right, making it smaller. Thus, the rejection probability for any $\boldsymbol{\theta}$ on the left vertical line is smaller than in Figure 1(a), i.e., strictly below α, including the point on the boundary of Θ_0. Again, by redefining $\phi(\cdot)$, essentially the same picture could be drawn for other points on the boundary of Θ_0, specifically any support points.

4.2 Discussion of results

Theorem 1(iii) is partly a result of the prior $P(H_0)$ being "small" when Θ_0 is small. That is, since the prior over the *parameter* is always the same (regardless of Θ_0), the implicit prior $P(H_0)$ shrinks when Θ_0 shrinks. (Technically, the prior in Assumption A1 is improper, so

$P(H_0)$ 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 $P(H_0) = 1/2$. Whether placing a prior on the null (like $P(H_0) = 1/2$) or on the parameter is more appropriate depends on the empirical setting. Often it is easier computationally not to set a specific $P(H_0)$; e.g., one may use the same posterior to compute probabilities of many different hypotheses. However, this can lead to a small $P(H_0)$ and thus large type I error rate.

Although the implicitly small $P(H_0)$ partially explains Theorem 1(iii), the shape of Θ_0 also plays an important role. For example, in $\mathbb{R}^2$, let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $H_0: \theta_1\theta_2 \geq 0$, 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 size α, whereas the size of the Bayesian test of $H_0: \theta_1\theta_2 \geq 0$ 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 $\text{Corr}(X_1, X_2) = -1$. 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 strong negative correlations, although size decreases to α as the correlation increases; see Appendix E as well as Kaplan (2015, §3.2) for details. Also, simulation results in Section 5.1 show size can exceed α even with prior $P(H_0) = 1/2$.

As a special case, 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 (2018) and others. For example, all the examples in Section 4 of Fang and Santos (2018) 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 size strictly above α, but it is also strictly above the size of an available frequentist test.

Many economic examples satisfy the condition of Θ_0 being a subset of a half-space as in Theorem 1(ii) and Theorem 1(iii). Switching the null and alternative hypotheses then generates examples where Θ_0 contains a half-space as in Theorem 1(iv) and Theorem 1(v). Often both are relevant; e.g., both the null of first-order stochastic dominance (e.g., Barrett and Donald, 2003) and the null of non-dominance (Davidson and Duclos, 2013) have been studied. For example, the null hypothesis of satisfaction of curvature constraints on economic functions (cost, production, etc.) has Θ_0 smaller than a half-space, and thus the null of non-satisfaction has Θ_0 larger than a half-space. The ordinal distribution relationships in Section 5.1 and Kaplan and Zhuo (2019) also have Θ_0 smaller than a half-space. The

examples in Section 4 of Fang and Santos (2018) 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 Kitamura and Stoye (2018). The general moment inequality null hypothesis $H_0\colon \mathrm{E}[\mathbf{W}] \leq \mathbf{0}$ with $\mathbf{W} \in \mathbb{R}^d$ as in equation (58) of Fang and Santos (2018) covers many other applications, but it has been studied already by Kline (2011) (see Section 2.2); and although technically not covered by his results, first-order stochastic dominance with continuous distributions is essentially an infinite-dimensional extension of this setup. Example 4.2 of Fang and Santos (2018) 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 dynamic macroeconomic models, certain parameter values lead to a stable equilibrium while others do not. Studies of such bifurcation find that 95% confidence sets usually include values with an unstable equilibrium; more directly, equilibrium (in)stability could be tested. Even in relatively simple models, the inequalities characterizing stability involve nonlinear functions of multiple parameters. See for example Barnett and Duzhak (2010), Barnett and Eryilmaz (2016), and references therein. Finally, as described in the introduction, many general hypotheses about identified sets correspond to the point-identified parameter lying in some region Θ_0.

Whether a set Θ is treated as the null or alternative hypothesis affects the properties of the Bayesian test. Combining Theorem 1(iii) and Theorem 1(iv), the Bayesian test of $H_0\colon \boldsymbol{\theta} \in \Theta$ may have size strictly above α while the Bayesian test of $H_0\colon \boldsymbol{\theta} \notin \Theta$ has size strictly below or equal to α. Moreover, if $\boldsymbol{\theta}$ is a support point of Θ, then the rejection probability of $H_0\colon \boldsymbol{\theta} \in \Theta$ is strictly above α by Theorem 1(iii); but the rejection probability of $H_0\colon \boldsymbol{\theta} \notin \Theta$ is strictly below α by Theorem 1(v).

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 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 (first-order) delta method (e.g., Hansen, 2020, Thm. 6.8), 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 $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathrm{N}(0, \mathbf{G}'\underline{\mathbf{V}}\mathbf{G})$, where $\hat{\beta} \equiv g(\hat{\boldsymbol{\mu}})$. However, this approximation may be poor in finite samples.

To be concrete, consider the following delta method example, continuing the same notation. Let $d = 2$ and $\beta \equiv g(\boldsymbol{\mu}) = \mu_1^2 + \mu_2^2 - 1$, 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). Further, as the variance $\underline{\mathbf{V}}/n$ grows, $\mathrm{P}(H_0 \mid \hat{\boldsymbol{\mu}}) \to 0$ given any $\hat{\boldsymbol{\mu}}$, so size grows to 1. In apparent contradiction, $H_0\colon \beta \leq 0$ suggests the Bayesian test has approximately correct size, by Theorem 1(i). The "contradiction" is simply that the

asymptotic approximation 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. Another drawback of the delta method is the existence of singularity points; e.g., with $\boldsymbol{\mu} = \mathbf{0}$ in this example, the delta method's distribution of $g(\hat{\boldsymbol{\mu}})$ is degenerate even if the distribution of $\hat{\boldsymbol{\mu}}$ is not.

4.3 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 $X, \theta \in \mathbb{R}$.

A special case of Theorem 1(iii) with $\boldsymbol{\theta} \in \mathbb{R}^d$ and $H_0 \colon \boldsymbol{\theta} \geq \mathbf{0}$ was explored by Kline (2011), as detailed in Section 2.2.

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.

5 Examples

We illustrate Theorem 1 through two specific examples; another is in Appendix G.

5.1 Example: ordinal distributions

5.1.1 Background

The original motivation for this paper was inference on relationships between ordinal distributions. Kaplan and Zhuo (2019) connect such ordinal relationships to relationships between corresponding latent distributions. Here, we focus on one such result.

Consider the ordinal variable self-reported health status (SRHS), commonly used in health economics. Let random variables X and Y denote SRHS from two different populations, like different socioeconomic or demographic groups. Let $X = 1$ denote "poor" health (the worst category), $X = 5$ "excellent" health (the best), and $X = 2, 3, 4$ the ordered intermediate categories (fair, good, very good), and similarly for Y. These numbers have no cardinal meaning; they are only for notational convenience. Imagine they are determined by an individual's underlying latent health X^* or Y^*, where $X = 1$ is reported if

$X^* \leq \gamma_1$, $X = 2$ is reported if $\gamma_1 < X^* \leq \gamma_2$, etc. The ordinal CDFs are $F_X(\cdot)$ and $F_Y(\cdot)$; the latent CDFs are $F_X^*(\cdot)$ and $F_Y^*(\cdot)$.

If Y values are determined from Y^* using the same γ_j thresholds as X, possibly shifted by a common constant $(\gamma_j + c)$, then it is possible to learn about the relative dispersion of latent X^* and Y^* from the ordinal X and Y. Specifically, if there is a "single crossing" of $F_Y(\cdot)$ by $F_X(\cdot)$ from below, then certain interquantile ranges of X^* are smaller than those of Y^*. (Other interquantile ranges lack evidence for either direction.) "Single crossing" (SC) means that $F_X(\cdot)$ is below $F_Y(\cdot)$ up to a certain category, and then is above thereafter. That is, for some category m, $F_X(j) < F_Y(j)$ for $j < m$ and $F_X(j) > F_Y(j)$ for $j > m$.

This can also be interpreted in terms of partial identification. The two ordinal CDFs are point-identified, but the two latent CDFs are only partially identified. That is, there is a mapping from a pair of ordinal CDFs to a (very large) set of pairs of latent CDFs that are consistent with the ordinal pair. The ordinal SC implies that *all* pairs of latent CDFs in the identified set have certain interquantile ranges of X^* smaller than those of Y^*.

5.1.2 Connections to Theorem 1

The null hypothesis of SC (of Y by X) can be written in terms of the parameters as follows. There are eight unknown ordinal parameters: $F_X(j)$ and $F_Y(j)$ for $j = 1, 2, 3, 4$, since $F_X(5) = F_Y(5) = 1$. Define the four CDF difference parameters

$$\theta_j \equiv F_X(j) - F_Y(j). \tag{8}$$

SC requires $\theta_1 < 0$ and $\theta_4 > 0$, but there are multiple possibilities for θ_2 and θ_3. Specifically, SC rules out $\theta_2 > 0 > \theta_3$, but the other three quadrants of (θ_2, θ_3) satisfy SC. We ignore the difference between $<$ and $\leq$ since they are statistically indistinguishable.

Writing the SC null hypothesis as $H_0 \colon \boldsymbol{\theta} \in \Theta_0$, the shape of Θ_0 has both convex and non-convex features. First, consider the first and fourth components. Since $\theta_1 < 0 < \theta_4$ is a necessary (not sufficient) condition,

$$\Theta_0 \subset \{\boldsymbol{\theta} : \theta_1 < 0 < \theta_4\} \subset \{\boldsymbol{\theta} : \theta_1 < 0\}, \tag{9}$$

where $\phi(\boldsymbol{\theta}) = \theta_1$ is a continuous linear functional of $\boldsymbol{\theta}$. That is, Θ_0 is a subset of a quarter-space, which is a subset of a half-space that satisfies the conditions in Theorem 1(iii). Thus, the Bayesian test's size should be strictly above α, with rejection probability above α for DGPs with $\boldsymbol{\theta}$ on the boundary of Θ_0 near the corner $(\theta_1, \theta_4) = (0, 0)$.

Second, looking at only (θ_2, θ_3) reveals local non-convexity. Imagine $\theta_1 < 0 < \theta_4$ by a large margin, and the sample size is large enough that the posterior of $\theta_1 < 0 < \theta_4$ is near

21

one. Then, the test practically only relies on (θ_2, θ_3). But Θ_0 only excludes one quadrant of the (θ_2, θ_3) subspace, the one with $\theta_2 > 0 > \theta_3$. When the conditions on (θ_1, θ_4) are satisfied, this leaves three quadrants of the (θ_2, θ_3) subspace, which is more than a half-space. In this case, the Bayesian test's size probably still equals α, but it may reject much less under certain DGPs, like with (θ_2, θ_3) near the origin. That is, Theorem 1(v) suggests asymptotic type I error rates strictly below α for DGPs local to $(\theta_2, \theta_3) = (0, 0)$ with $\theta_1 < 0 < \theta_4$, but since Θ_0 is not larger than a half-space globally, this may not be true in small samples.

Instead of SC, the null hypothesis could be the opposite, i.e., non-SC. Then, when $(\theta_1, \theta_4) = (0, 0)$, instead of Theorem 1(iii) implying type I error rate above α, Theorem 1(v) implies type I error rate below α. When $(\theta_2, \theta_3) = (0, 0)$ and $\theta_1 < 0 < \theta_4$, instead of Theorem 1(v) applying asymptotically, Theorem 1(iii) suggests asymptotic type I error rate above α.

5.1.3 Simulation setup

The following simulations show the importance of the shape of Θ_0 and sometimes sample size. Independent, iid samples of X and Y are taken, with sample size n for both. The results show the simulated rejection probability (RP) of the Bayesian test of X single crossing Y. DGP 1 is at the convex "corner" with $(\theta_1, \theta_4) = (0, 0)$, where the Bayesian test's RP should be above α. Specifically, $F_X(1) = F_Y(1) = 0.2$, $F_X(2) = 0.39 < 0.4 = F_Y(2)$, $F_X(3) = 0.5 < 0.6 = F_Y(3)$, and $F_X(4) = F_Y(4) = 0.8$. DGP 2 is at the local non-convexity with $(\theta_2, \theta_3) = (0, 0)$, where the RP may be below α with large enough n. Specifically, $F_X(1) = 0.18 < 0.22 = F_Y(1)$, $F_X(2) = F_Y(2) = 0.4$, $F_X(3) = F_Y(3) = 0.6$, and $F_X(4) = 0.82 > 0.78 = F_Y(4)$. Since both DGPs are on the boundary of SC and non-SC, the null hypothesis can be considered either SC or non-SC, and the rejection probability can be interpreted as a type I error rate. Each simulation had 1000 replications, with 1000 posterior draws each.

The Bayesian test is computed as follows. Let $\mathbf{p}_X$ be the vector with entry j equal to $P(X = j)$, $j = 1, \dots, 5$, and similarly for $\mathbf{p}_Y$. Let $\mathbf{n}_X$ be the vector of data counts with entry j equal to the number of observations with $X_i = j$ (out of $i = 1, \dots, n$), and similarly for $\mathbf{n}_Y$. With a uniform prior, the posterior distributions are Dirichlet; specifically, with $\mathbf{1} \equiv (1, 1, 1, 1, 1)$,

$$\mathbf{p}_X \mid \mathbf{n}_X \sim \text{Dir}(\mathbf{n}_X + \mathbf{1}), \quad \mathbf{p}_Y \mid \mathbf{n}_Y \sim \text{Dir}(\mathbf{n}_Y + \mathbf{1}). \tag{10}$$

The Bayesian test takes draws from these posteriors and computes the proportion of draws in which SC holds, rejecting the SC null hypothesis if the proportion is below α, and rejecting the null of non-SC if the proportion is above $1 - \alpha$.

Additionally, to help isolate the role of "shape" from "size" of Θ_0, a test with adjusted prior is computed. It is identical except that the prior is adjusted to have $1/2$ probability on H_0. Here, the original probability of H_0 can be simulated by drawing from the uniform prior, and then the prior is normalized using this original probability. The prior remains uniform within Θ_0, and uniform within $\Theta_0^{\complement}$, but at two different levels instead of the same level. See Goutis et al. (1996) or Kaplan and Zhuo (2019) for details.

Table 1: Simulated type I error rate of Bayesian tests of ordinal single crossing, $\alpha = 0.1$.

DGP	n	H_0: SC		H_0: non-SC	
		RP	RP (adj)	RP	RP (adj)
1	20	0.422	0.170	0.000	0.014
1	100	0.484	0.186	0.000	0.010
1	500	0.531	0.249	0.000	0.008
2	20	0.246	0.056	0.001	0.050
2	100	0.128	0.022	0.009	0.148
2	500	0.011	0.002	0.130	0.501

5.1.4 Simulation results

Table 1 shows the results. "RP" is the rejection probability of the first Bayesian test, and "RP (adj)" is the rejection probability of the Bayesian test with the adjusted prior. Some patterns emerge.

For DGP 1, RP is indeed well above α when the null hypothesis is SC, consistent with Theorem 1(iii). Here, the RP actually increases as n increases; it is not simply a finite-sample difference. Even with the adjusted prior, RP is well above α and again increasing with n. In contrast, when the null is non-SC, the RP is near zero, consistent with Theorem 1(v).

For DGP 2, the sample size plays an important role. When the sample size is small, the posterior and sampling distributions are very dispersed, so the global shape of Θ_0 matters most. Globally, for the null of SC, Θ_0 is a subset of a quarter-space; Theorem 1(iii) suggests DGPs near the boundary may have RP above α. Indeed, RP is again well above α for small n. With the null of non-SC, instead Theorem 1(v) suggests RP below α, which is seen in the table. As n increases, the local shape of Θ_0 matters more, so Theorem 1(v) is more appropriate for the null of SC and Theorem 1(iii) is more appropriate for the null of non-SC. In this case, with the largest sample size, RP is far below α for the SC null, opposite the result with small n. Similarly, the RP becomes above α for the non-SC null, also opposite the small n results. Like with DGP 1, the RP with adjusted prior is different, but it is not always closer to α.

5.2   Example: curvature constraints

5.2.1   Setting and connections to Theorem 1

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 2 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.[4]

Specifically, we test concavity of cost in input prices as follows.[5] With output y, input prices $\mathbf{w} = (w_1, w_2, w_3)$, and total cost $C(y, \mathbf{w})$, the translog model is

$$
\begin{aligned}
\ln(C(y, \mathbf{w})) = a_0 + a_y \ln(y) + (1/2)a_{yy}[\ln(y)]^2 + \sum_{k=1}^{3} a_{yk} \ln(y) \ln(w_k) \\
+ \sum_{k=1}^{3} b_k \ln(w_k) + (1/2) \sum_{k=1}^{3} \sum_{m=1}^{3} b_{km} \ln(w_k) \ln(w_m).
\end{aligned}
\tag{11}
$$

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

---

[4]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.

[5]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.

p (for all $p = 1, 2, 3$) having the same sign as $(-1)^p$ or zero.[6]

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

$$H_0 \colon \underline{\mathbf{H}} \equiv \left. \frac{\partial^2 C(y, \mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}'} \right|_{(y, \mathbf{w}) = (1,1,1,1)} \quad \text{is NSD.} \tag{12}$$

This is necessary but not sufficient for global concavity; rejecting local concavity implies rejection of global concavity. In Appendix F, 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.

5.2.2 Simulation setup and results

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)

$$\ln(C/w_3) = a_0 + a_y \ln(y) + (1/2)a_{yy}[\ln(y)]^2 + \sum_{k=1}^{2} a_{yk} \ln(y) \ln(w_k/w_3)$$
$$+ \sum_{k=1}^{2} b_k \ln(w_k/w_3) + (1/2) \sum_{k=1}^{2} \sum_{m=1}^{2} b_{km} \ln(w_k/w_3) \ln(w_m/w_3) + \epsilon \tag{13}$$

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. When $\delta > 0$, all principal minors satisfy NSD and are non-zero (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 $N(0, \sigma = 0.1)$, $\epsilon \sim N(0, \sigma_\epsilon)$, 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 draws among all replications.

Table 2 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)

---

[6]In some cases, only leading principal minors are required; see Mandy (2018).

[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 $s_k = x_k w_k/C = (\partial C/\partial w_k)(w_k/C) = \partial \ln(C)/\partial \ln(w_k) = b_k + a_{yk} \ln(y) + \sum_{j=1}^{3} b_{jk} \ln(w_j)$.

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

α	σ_ϵ	Bayesian 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 2 shows the type I error rate of the Bayesian tests of (12) 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] This is consistent with Theorem 1(iii).

6   Conclusion

We have explored the frequentist properties of Bayesian inference on general nonlinear inequality constraints, including hypotheses about identified set properties, providing formal results on the role of the shape of the null hypothesis parameter subspace. Future work could include investigation of approaches like Müller and Norets (2016) applied to nonlinear

---

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

inequality testing. Moreover, one could explore how to match frequentist size by adjusting the prior or loss function.

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

Barnett, W. A., Duzhak, E. A., 2010. Empirical assessment of bifurcation regions within New Keynesian models. Economic Theory 45 (1–2), 99–128.
URL https://www.jstor.org/stable/40864839

Barnett, W. A., Eryilmaz, U., 2016. An analytical and numerical search for bifurcations in open economy New Keynesian models. Macroeconomic Dynamics 20 (2), 482–503.
URL https://doi.org/10.1017/S1365100514000224

Barrett, G. F., Donald, S. G., 2003. Consistent tests for stochastic dominance. Econometrica 71 (1), 71–104.
URL https://www.jstor.org/stable/3082041

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.
URL https://doi.org/10.1090/surv/062

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.

URL https://www.jstor.org/stable/1392346

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.

URL https://doi.org/10.1007/978-0-387-75971-5

Davidson, R., Duclos, J.-Y., 2013. Testing for restricted stochastic dominance. Econometric Reviews 32 (1), 84–125.

URL https://doi.org/10.1080/07474938.2012.690332

Deaton, A., Paxson, C., 1998a. Aging and inequality in income and health. American Economic Review (Papers and Proceedings) 88 (2), 248–253.

URL https://www.jstor.org/stable/116928

Deaton, A., Paxson, C., 1998b. Health, income, and inequality over the life cycle. In: Frontiers in the Economics of Aging. University of Chicago Press, pp. 431–462.

URL https://www.nber.org/chapters/c7309

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

Efron, B., Tibshirani, R., 1998. The problem of regions. Annals of Statistics 26 (5), 1687–1718.

URL https://projecteuclid.org/euclid.aos/1024691353

Fang, Z., Santos, A., 2018. Inference on directionally differentiable functions. Review of Economic Studies 86 (1), 377–412.

URL https://doi.org/10.1093/restud/rdy049

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.
   URL https://doi.org/10.1017/9781139029834

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., 2020. Econometrics, textbook draft.
   URL 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.

Kaplan, D. M., Zhuo, L., 2019. Comparing latent inequality with ordinal data, 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

Kitamura, Y., Stoye, J., 2018. Nonparametric analysis of random utility models. Econometrica 86 (6), 1883–1909.
   URL https://doi.org/10.3982/ECTA14478

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

Kline, B., Tamer, E., 2016. Bayesian inference in a class of partially identified models. Quantitative Economics 7 (2), 329–366.
   URL https://doi.org/10.3982/QE399

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.
URL https://eml.berkeley.edu/~mcfadden/prodecon1.html

Lehmann, E. L., Casella, G., 1998. Theory of Point Estimation, 2nd Edition. Springer, New York.
URL https://books.google.com/books?id=4f24CgAAQBAJ

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

Mandy, D. M., 2018. Leading principal minors and semidefiniteness. Economic Inquiry 56 (2), 1396–1398.
URL https://doi.org/10.1111/ecin.12536

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

Romano, J. P., Shaikh, A. M., Wolf, M., 2014. A practical two-step method for testing moment inequalities. Econometrica 82 (5), 1979–2002.
URL https://www.jstor.org/stable/24029299

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

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.
URL https://doi.org/10.1007/978-1-4757-2545-2

Welch, B. L., Peers, H. W., 1963. On formulae for confidence points based on integrals of weighted likelihoods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 25 (2), 318–329.
URL https://www.jstor.org/stable/2984298

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

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

31

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

$$
\begin{aligned}
&\mathcal{R}_1 \subseteq \mathcal{R}_2, \ \mathcal{R}_1 \equiv \{\mathbf{X} : \mathrm{P}(\phi(\boldsymbol{\theta}) \leq c_0 \mid \mathbf{X}) \leq \alpha\} = \{\mathbf{X} : \phi(\mathbf{X}) \geq r\}, \\
&\mathcal{R}_2 \equiv \{\mathbf{X} : \mathrm{P}(\boldsymbol{\theta} \in \Theta_0 \mid \mathbf{X}) \leq \alpha\}.
\end{aligned}
\tag{A.1}
$$

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 (A.1), 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),

$$
\sup_{\boldsymbol{\theta} \in \Theta_0} \mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}) \overbrace{\geq \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta})}^{\text{since } c_0 \in \phi(\overline{\Theta_0})} \overbrace{\geq \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}(\mathbf{X} \in \mathcal{R}_1 \mid \boldsymbol{\theta})}^{\text{by (A.1)}}
$$

$$
= \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}\big(\phi(\mathbf{X}) \geq r \mid \boldsymbol{\theta}\big) = \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \mathrm{P}(\phi(\mathbf{X}) - \phi(\boldsymbol{\theta}) \geq r - \phi(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) = \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} 1 - F\big(r - \phi(\boldsymbol{\theta})\big)
$$

$$
\overbrace{= 1 - F\big(r - c_0\big)}^{\text{by continuity of } F, \phi} \overbrace{= \alpha}^{\text{by Theorem 1(i)}} .
\tag{A.2}
$$

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

$$\mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Delta \mid \phi_2(\mathbf{X})) > 0 \text{ for any } \phi_2(\mathbf{X}). \tag{A.3}$$

Similar to (A.1), for some $r \in \mathbb{R}$, let

$$
\begin{aligned}
&\mathcal{R}_a \subseteq \mathcal{R}_b, \ \mathcal{R}_a \equiv \{\phi_2(\mathbf{X}) : \mathrm{P}(\phi_3(\phi_2(\boldsymbol{\theta})) \leq c_0 \mid \phi_2(\mathbf{X})) \leq \alpha\} = \{\phi_2(\mathbf{X}) : \phi_3(\phi_2(\mathbf{X})) \geq r\}, \\
&\mathcal{R}_b \equiv \{\phi_2(\mathbf{X}) : \mathrm{P}(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X})) \leq \alpha\}.
\end{aligned}
\tag{A.4}
$$

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

$$P(\boldsymbol{\theta} \in \Theta_0 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*) \overbrace{\leq P(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}^{\text{since } \boldsymbol{\theta} \in \Theta_0 \implies \phi_2(\boldsymbol{\theta}) \in \Phi_2}$$

$$= \overbrace{P(\phi_3(\phi_2(\boldsymbol{\theta})) \leq c_0 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}^{=\alpha \text{ by Theorem 1(i)}} - \overbrace{P(\phi_2(\boldsymbol{\theta}) \in \Delta \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}^{>0 \text{ by (A.3)}}$$

$$< \alpha.$$

By the assumption that $P(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X}))$ is continuous in $\phi_2(\mathbf{X})$, there is some ϵ-ball $\mathcal{B}$ around $\mathbf{p}^*$ for which $P(\boldsymbol{\theta} \in \Theta_0 \mid \phi_2(\mathbf{X}) \in \mathcal{B}) < \alpha$, 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),

$$P\big(\phi_2(\mathbf{X}) \in (\mathcal{B} \cap \mathcal{R}_a^\complement) \mid \boldsymbol{\theta}\big) > 0 \text{ for any } \boldsymbol{\theta}. \tag{A.5}$$

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 $\boldsymbol{\theta} \in \Theta_0 \implies \phi_2(\boldsymbol{\theta}) \in \Phi_2$,

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

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

$$\phi_2(\mathbf{X}) \in \mathcal{R}_b \implies \mathbf{X} \in \mathcal{R}_2. \tag{A.6}$$

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

$$P(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}^*) \overbrace{\geq P(\phi_2(\mathbf{X}) \in \mathcal{R}_b \mid \boldsymbol{\theta}^*)}^{\text{by (A.6)}}$$

$$\geq \overbrace{P(\phi_2(\mathbf{X}) \in \mathcal{R}_a \mid \boldsymbol{\theta}^*)}^{=\alpha \text{ by Theorem 1(i)}} + \overbrace{P\big(\phi_2(\mathbf{X}) \in (\mathcal{B} \cap \mathcal{R}_a^\complement) \mid \boldsymbol{\theta}^*\big)}^{>0 \text{ by (A.5)}}$$

$$> \alpha.$$

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

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

For Theorem 1(iv), some examples suffice. First, consider $H_0\colon \phi(\boldsymbol{\theta}) \neq 0$. Given any $\mathbf{X}$,

$P(\phi(\boldsymbol{\theta}) \neq 0 \mid \mathbf{X}) = 1$ 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 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 α. Fourth, Appendix E 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).

For Theorem 1(v): the argument is the same as for Theorem 1(iii), but with Θ_0 larger than a half-space instead of smaller. The first divergence from the proof of Theorem 1(iii) is the change from $\leq$ and $<$ to $\geq$ and $>$ in

$$P(\boldsymbol{\theta} \in \Theta_0 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*) \overset{\text{since } \boldsymbol{\theta} \in \Theta_0 \Longleftarrow \phi_2(\boldsymbol{\theta}) \in \Phi_2}{\geq} P(\phi_2(\boldsymbol{\theta}) \in \Phi_2 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)$$

$$= \overset{=\alpha \text{ by Theorem 1(i)}}{\overbrace{P(\phi_3(\phi_2(\boldsymbol{\theta})) \leq c_0 \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}} + \overset{>0 \text{ by (A.3)}}{\overbrace{P(\phi_2(\boldsymbol{\theta}) \in \Delta \mid \phi_2(\mathbf{X}) = \mathbf{p}^*)}}$$

$$> \alpha.$$

Analogous to before, there is an ϵ-ball $\mathcal{B}$ around $\mathbf{p}^*$ for which $P(\boldsymbol{\theta} \in \Theta_0 \mid \phi_2(\mathbf{X}) \in \mathcal{B}) > \alpha$. The same arguments as before lead to a statement parallel to (A.5) (simply replacing $\mathcal{R}_a^{\complement}$ with $\mathcal{R}_a$),

$$P\big(\phi_2(\mathbf{X}) \in (\mathcal{B} \cap \mathcal{R}_a) \mid \boldsymbol{\theta}\big) > 0 \text{ for any } \boldsymbol{\theta}. \tag{A.7}$$

Continuing in parallel, now $\boldsymbol{\theta} \in \Theta_0 \Longleftarrow \phi_2(\boldsymbol{\theta}) \in \Phi_2$ (instead of $\Longrightarrow$), so

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

so (A.6) is replaced by

$$\phi_2(\mathbf{X}) \in \mathcal{R}_b \Longleftarrow \mathbf{X} \in \mathcal{R}_2. \tag{A.8}$$

The final argument becomes

$$
\mathrm{P}(\mathbf{X} \in \mathcal{R}_2 \mid \boldsymbol{\theta}^*) \overset{\text{by (A.8)}}{\leq \mathrm{P}(\phi_2(\mathbf{X}) \in \mathcal{R}_b \mid \boldsymbol{\theta}^*)}
$$

$$
\leq \overset{=\alpha \text{ by Theorem 1(i)}}{\mathrm{P}(\phi_2(\mathbf{X}) \in \mathcal{R}_a \mid \boldsymbol{\theta}^*)} - \overset{>0 \text{ by (A.7)}}{\mathrm{P}\big(\phi_2(\mathbf{X}) \in (\mathcal{B} \cap \mathcal{R}_a) \mid \boldsymbol{\theta}^*\big)}
$$

$$
< \alpha. \qquad\qquad \square
$$

B   Bernstein–von Mises theorems

A Bernstein–von Mises theorem establishes the asymptotic equivalence of the sampling and posterior distributions. This can help verify Assumption A1, although it is not strictly necessary. Results for posterior asymptotic normality date back to Laplace (1820), as cited in Lehmann and Casella (1998, §6.10, p. 515).

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 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.

In the case of infinite-dimensional $\mathbf{X}$ and $\boldsymbol{\theta}$, as noted, an infinite-dimensional Bernstein–von Mises theorem is not necessary to satisfy Assumption A1, which only concerns the distribution of a scalar-valued 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

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

35

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,

$$\sqrt{n}\big(\hat{F}(\cdot) - F(\cdot)\big) \rightsquigarrow B\big(F(\cdot)\big), \tag{B.1}$$

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 (B.1) 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 $t \in [0, 1]$.

C Details on minimax risk

The following shows that an unbiased frequentist test with size α minimizes "max" (supremum) risk. Let $\boldsymbol{\theta} \in \Theta$, $H_0\colon \boldsymbol{\theta} \in \Theta_0 \subset \Theta$, $H_1\colon \boldsymbol{\theta} \notin \Theta_0$. From the frequentist perspective, given (1), the loss function given $\boldsymbol{\theta}$ is $(1 - \alpha)\,\mathbb{1}\{\text{reject}\}$ if $\boldsymbol{\theta} \in \Theta_0$ or $\alpha\,\mathbb{1}\{\text{accept}\}$ if $\boldsymbol{\theta} \notin \Theta_0$. Thus, risk (expected loss) given $\boldsymbol{\theta}$ is $(1 - \alpha)\,\mathrm{P}_{\boldsymbol{\theta}}(\text{reject})$ if $\boldsymbol{\theta} \in \Theta_0$ or $\alpha\,\mathrm{P}_{\boldsymbol{\theta}}(\text{accept})$ if $\boldsymbol{\theta} \notin \Theta_0$,

so "minimax risk" minimizes

$$\max\left\{(1-\alpha)\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}),\ \alpha \sup_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{accept})\right\}$$
$$= \max\left\{(1-\alpha)\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}),\ \alpha[1-\inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})]\right\}, \tag{C.1}$$

where $\mathrm{P}_{\boldsymbol{\theta}}(\cdot)$ is probability under $\boldsymbol{\theta}$. "Unbiased" means the test satisfies

$$\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \leq \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}). \tag{C.2}$$

If the power function is continuous in $\boldsymbol{\theta}$, then (writing $\partial\Theta_0$ for the boundary of Θ_0, which is also the boundary of its complement)

$$\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \geq \sup_{\boldsymbol{\theta}\in\partial\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \geq \inf_{\boldsymbol{\theta}\in\partial\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) \geq \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}). \tag{C.3}$$

Thus, combining (C.2) and (C.3),

$$\sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) = \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}).$$

Defining the test's size as

$$\gamma_0 \equiv \sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}), \tag{C.4}$$

(C.1) becomes

$$\max\{(1-\alpha)\gamma_0, \alpha(1-\gamma_0)\}. \tag{C.5}$$

The minimum possible value $\alpha(1-\alpha)$ is attained with size $\gamma_0 = \alpha$: if instead $\gamma_0 < \alpha$, then $\alpha(1-\gamma_0) > \alpha(1-\alpha)$, or if $\gamma_0 > \alpha$, then $(1-\alpha)\gamma_0 > \alpha(1-\alpha)$. Without unbiasedness, $\inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject}) < \gamma_0$, so $\alpha[1-\inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})] > \alpha(1-\gamma_0)$, increasing the minimax risk. Thus, given (C.3), a minimax risk decision rule is any unbiased frequentist hypothesis test with 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 test is above α, but the magnitude of the difference is very small for conventional α. As a function of γ_0, let $\gamma_1(\gamma_0) \equiv \inf_{\boldsymbol{\theta}\notin\Theta_0} \mathrm{P}_{\boldsymbol{\theta}}(\text{reject})$. Then (C.1) equals

$$\max\{R_0, R_1\}, \quad R_0 \equiv (1-\alpha)\gamma_0, \quad R_1 \equiv \alpha(1-\gamma_1(\gamma_0)). \tag{C.6}$$

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 $\gamma_0 > \alpha/(1-\alpha)$. More commonly, $\gamma_1(\gamma_0)$ is strictly

increasing in γ_0 but $\gamma_1(\alpha) < \alpha$. This implies R_0 is increasing in γ_0 while R_1 is decreasing in γ_0, so the γ_0 that minimizes (C.6) sets $R_0 = R_1$. If $\gamma_0 = \alpha$, then $R_0 < R_1$; if $\gamma_0 = \alpha/(1-\alpha)$, then $R_0 = \alpha > R_1$. Hence, $R_0 = R_1$ (and thus minimax risk) is achieved in between, 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. Without unbiasedness, such a test is no longer a minimax rule (since the trivial randomized test "reject with probability α" achieves lower $\alpha(1-\alpha)$ maximum risk), but the difference in the resulting maximum risk in (C.6) is quantitatively small.

D Least favorable prior example

For intuition, consider a simple setting. Let $X, \theta \in \mathbb{R}$, with $H_0 \colon \theta \leq 0$ (so $\Theta_0 = \{t : t \leq 0\}$), and $X \sim \mathrm{N}(\theta, 1)$, so the likelihood is $\ell(X \mid \theta) = (2\pi)^{-1/2} \exp\{-(X - \theta)^2/2\}$. The prior sequence is $\pi_m(\theta)$ for $m = 1, 2, \ldots$, with posterior denoted $\pi_m(\cdot \mid X)$. The corresponding Bayes rules are δ_m, and the corresponding Bayes risks are r_m. Given the limit $r_m \to r$, any decision rule whose supremum (over $\theta \in \mathbb{R}$) risk equals r is minimax. The posteriors are $\pi_m(\theta \mid X)$. Below, the flat prior is shown not least favorable, and a least favorable sequence is derived. The Bayes rules are from Method 1, i.e., the loss function still takes value $1 - \alpha$ for type I error and α for type II error as in (1).

There is no single, fixed least favorable prior. Consider the Bayes risk of the usual frequentist test that rejects when $X > \Phi^{-1}(1 - \alpha)$, the $(1 - \alpha)$-quantile of the $\mathrm{N}(0, 1)$ distribution. Risk equals $\alpha(1 - \alpha)$ if $\theta = 0$, but risk is strictly below $\alpha(1 - \alpha)$ for any $\theta \neq 0$. Thus, unless $\theta = 0$ has prior probability one, the frequentist test's Bayes risk is strictly below $\alpha(1 - \alpha)$, and by definition the Bayes rule's Bayes risk is weakly lower than that. In the case where $\theta = 0$ indeed has prior probability one, the posterior also puts all probability on $\theta = 0$, so the Bayes rule is "never reject," and the Bayes risk is zero (since "never reject" is always correct if $\theta = 0$), also strictly below $\alpha(1 - \alpha)$.

Next, consider the sequence of normal priors with improper limit, $\theta \sim \mathrm{N}(0, m^2)$ as $m \to \infty$. With large m, the posterior is approximately $\mathrm{N}(X, 1)$, so the Bayes rule rejects when $X > \Phi^{-1}(1 - \alpha)$. Again, when $\theta = 0$, risk is $\alpha(1 - \alpha)$, but when $\theta \neq 0$, the risk is strictly below $\alpha(1 - \alpha)$. For example, when $\theta = -4$, $\mathrm{P}(X > \Phi^{-1}(1 - \alpha)) < 0.00003$ for any $\alpha < 0.5$. Thus, at $\theta = -4$, risk is below $0.00003(1 - \alpha)$. The Bayes risk against π_m averages the risk across all different values of θ, so it will be below the maximum risk of $\alpha(1 - \alpha)$. Further, as $m \to \infty$, values of θ very far from zero become increasingly likely, so the Bayes risk becomes smaller and smaller, $r_m \to 0$. For example, even with $m = 10$, $\mathrm{P}(\theta \leq -4) = 0.34$.

Coincidentally, the limiting (generalized) Bayes rule happens to be minimax, but this is not the least favorable prior sequence. Arguably, it is the "most favorable" prior sequence because it places less and less weight on θ near the boundary point of Θ_0 (i.e., near $\theta = 0$), which is where risk is highest.

The least favorable prior sequence can be reverse-engineered from the unbiased frequentist test. The prior must have $0 < \pi_m(\Theta_0) < 1$: both the null and alternative must have strictly positive probability, otherwise the test will always reject or always accept and have zero Bayes risk. Given that the Bayes rule against π_m rejects when $X > c_m$ for some $c_m \in \mathbb{R}$, there cannot be positive probability on both $\theta = 0$ and $\theta < 0$ (in the limit), or else the risk given $\theta < 0$ is below risk given $\theta = 0$. The simplest remaining option is a two-point prior:

$$\pi_m(0) = p_m, \quad \pi_m(a_m) = 1 - p_m \text{ for some } a_m > 0. \tag{D.1}$$

In the limit, presumably the risk at $\theta = 0$ is $\alpha(1 - \alpha)$ because the rejection probability is α. If so, it must be that $a_m \downarrow 0$; otherwise, the rejection probability at a_m would remain strictly larger than α, in which case risk at $\theta = a_m$ would remain strictly smaller than $\alpha(1 - \alpha)$.

A sequence p_m in (D.1) can be solved for given the sequence a_m. For example, let $a_m = 1/m \to 0^+$. We can solve for p_m (below) such that the Bayes rule matches the frequentist test; it turns out $p_m \to \alpha$ as $m \to \infty$. In that case, the risk at $\theta = 0$ is $\alpha(1 - \alpha)$ for all m, while the risk at $\theta = a_m$ approaches $\alpha(1 - \alpha)$ from below as $m \to \infty$. Since the Bayes risk averages these two risks (weighted by p_m and $1 - p_m$), and both approach $\alpha(1 - \alpha)$, the limit of the sequence of Bayes risks is $\alpha(1 - \alpha)$. Thus, given that the usual frequentist test's supremum (over θ) risk is $\alpha(1 - \alpha)$ (the risk at $\theta = 0$), the conclusion is that the frequentist test is a minimax risk decision rule.

The value of p_m mentioned above can be found as follows. The goal is for the Bayesian test to reject when $X > \Phi^{-1}(1 - \alpha)$, same as the frequentist test. Since the posterior probability of Θ_0 decreases monotonically (and strictly) with X, it suffices to solve for X such that the posterior equals α exactly, to find the boundary of the rejection region. Since the prior has only two points with positive probability, so does the posterior. Using the usual Bayes law of posterior proportional to likelihood times prior,

$$\pi_m(\theta = 0 \mid X) \propto b_{0m} \equiv p_m \phi(X), \quad \pi_m(\theta = a_m \mid X) \propto b_{1m} \equiv (1 - p_m)\phi(X - a_m), \tag{D.2}$$

where $\phi(\cdot)$ is the standard normal PDF. Normalizing so these sum to one yields

$$\pi_m(\theta = 0 \mid X) = b_{0m}/(b_{0m} + b_{1m}), \quad \pi_m(\theta = a_m \mid X) = b_{1m}/(b_{0m} + b_{1m}). \tag{D.3}$$

We want $\pi_m(\theta = 0 \mid X) = \alpha$ when $X = \Phi^{-1}(1-\alpha)$, i.e.,

$$\alpha = \pi_m(0 \mid X = \Phi^{-1}(1-\alpha)) = \frac{p_m\phi\big(\Phi^{-1}(1-\alpha)\big)}{p_m\phi\big(\Phi^{-1}(1-\alpha)\big) + (1-p_m)\phi\big(\Phi^{-1}(1-\alpha) - a_m\big)}. \quad \text{(D.4)}$$

This is easily solved numerically for p_m given α and a_m. Further, as $m \to \infty$ and $a_m \to 0$, the right hand side converges to $p_m/(p_m + (1 - p_m)) = p_m$, so $p_m \downarrow \alpha$. The risk at $\theta = 0$ is $\alpha(1-\alpha)$, where $1 - \alpha$ is the loss for type I error and α is the rejection probability, i.e., probability $X > \Phi^{-1}(1-\alpha)$ given $X \sim \mathrm{N}(0,1)$. The risk at $\theta = a_m$ is $\alpha\Phi(\Phi^{-1}(1-\alpha) - a_m)$, where α is the type II error loss and the rest is the probability of accepting H_0 (i.e., of $X \le \Phi^{-1}(1-\alpha)$) when $\theta = a_m$, given $X \sim \mathrm{N}(a_m, 1)$. Altogether, the Bayes risk is

$$p_m(1-\alpha)\alpha + (1-p_m)\alpha\Phi(\Phi^{-1}(1-\alpha) - a_m) \to (\alpha)(1-\alpha)\alpha + (1-\alpha)\alpha(1-\alpha) = \alpha(1-\alpha). \quad \text{(D.5)}$$

Consequently, the test that rejects when $X > \Phi^{-1}(1-\alpha)$ is minimax since its supremum risk matches the Bayes risk limit in (D.5):

$$\max\{\sup_{\theta \le 0}(1-\alpha)\,\mathrm{P}(\text{reject} \mid \theta), \quad \sup_{\theta > 0}\alpha\,\mathrm{P}(\text{accept} \mid \theta)\}$$

$$= \max\{\sup_{\theta \le 0}(1-\alpha)\,\mathrm{P}(X > \Phi^{-1}(1-\alpha) \mid \theta), \quad \sup_{\theta > 0}\alpha\,\mathrm{P}(X \le \Phi^{-1}(1-\alpha) \mid \theta)\}$$

$$= \max\{(1-\alpha)\,\mathrm{P}(X > \Phi^{-1}(1-\alpha) \mid \theta = 0), \quad \alpha\lim_{\theta \downarrow 0}\mathrm{P}(X \le \Phi^{-1}(1-\alpha) \mid \theta)\}$$

$$= \max\{(1-\alpha)\alpha, \quad \alpha(1-\alpha)\}$$

$$= \alpha(1-\alpha).$$

E   Example: bivariate normal sign equality test

Here, we provide additional mathematical and simulation details for one of the examples mentioned in Section 4.2. Consider a bivariate normal model where the null hypothesis is that the two parameters have the same sign (letting zero count either way), $H_0\colon \theta_1\theta_2 \ge 0$. Specifically,

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

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 $N(\mathbf{0}, \mathbf{\underline{\Sigma}})$ distribution be

$$
N(\mathbf{0}, \mathbf{\underline{\Sigma}}) : f(t_1, t_2) \equiv \det(2\pi\mathbf{\underline{\Sigma}})^{-1/2} \exp\left\{-\frac{(t_1, t_2)\Sigma^{-1}(t_1, t_2)'}{2}\right\}
$$

$$
= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(1-\rho^2)}\right\},
$$

$$
N(\rho t_2, 1 - \rho^2) : f(t_1 \mid t_2) \equiv \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\{-(t_1 - \rho t_2)^2/(2 - 2\rho^2)\},
$$

$$
N(0, 1) : f(t_2) \equiv \frac{1}{\sqrt{2\pi}} \exp\{-t_2^2/2\} = \phi(t_2),
$$

(E.2)

where $\phi(\cdot)$ is the standard normal PDF. Given (E.2),

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

(E.3)

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

$$
P(H_0 \mid X_1, X_2) = \int_{-\infty}^{-X_2} \int_{-\infty}^{-X_1} f(t_1, t_2) \, dt_1 \, dt_2 + \int_{-X_2}^{\infty} \int_{-X_1}^{\infty} f(t_1, t_2) \, dt_1 \, dt_2
$$

$$
= \int_{-\infty}^{-X_2} \int_{-\infty}^{-X_1} f(t_1 \mid t_2) f(t_2) \, dt_1 \, dt_2 + \int_{-X_2}^{\infty} \int_{-X_1}^{\infty} f(t_1 \mid t_2) f(t_2) \, dt_1 \, dt_2
$$

$$
= \int_{-\infty}^{-X_2} \Phi\left(\frac{-X_1 - \rho t_2}{\sqrt{1-\rho^2}}\right) \phi(t_2) \, dt_2 + \int_{-X_2}^{\infty} \left[1 - \Phi\left(\frac{-X_1 - \rho t_2}{\sqrt{1-\rho^2}}\right)\right] \phi(t_2) \, dt_2.
$$

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

Table 3 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 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 $\rho = -0.5$.

Table 3: Bayesian test type I error rates, $H_0 \colon \theta_1\theta_2 \geq 0$, $\theta_1 = 0$, $\alpha = 0.1$, 10,000 replications.

θ_2	-1	-0.99	-0.9	-0.5	0	0.5	1
				ρ			
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

F   Derivation of translog constraints

The Hessian is a nonlinear function of the translog parameters, and it depends on $(y, \mathbf{w})$. Letting[9]

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

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

$$H_{mk} = \frac{\partial^2 C}{\partial w_m \partial w_k} = \frac{\partial}{\partial w_m}\frac{\partial C}{\partial w_k} = \frac{\partial}{\partial w_m}(r_k C/w_k) = \frac{\partial r_k}{\partial w_m}(C/w_k) + \frac{\partial C}{\partial w_m}(r_k/w_k) + \frac{\partial w_k^{-1}}{\partial w_m}(r_k C)$$

$$= (b_{mk}/w_m)(C/w_k) + r_m(C/w_m)(r_k/w_k) - \mathbb{1}\{k = m\}w_k^{-2}(r_k C)$$

$$= C\frac{b_{mk} + r_m r_k - \mathbb{1}\{k = m\}r_k}{w_m w_k}.$$

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 (12) 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 (F.1), $H_{11} \leq 0$ 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 $(b_{mk} = b_{km})$ 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 $b_{11} \leq 0$

---

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

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 (12) 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 2 show this to be the case here.

G 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.

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.

G.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 $X_n(\cdot) \equiv \sqrt{n}(\hat{F}_X(\cdot) - F_{0,n}(\cdot))$ and $\theta_n(\cdot) \equiv \sqrt{n}(F_X(\cdot) - F_{0,n}(\cdot))$, 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 Appendix G.2), only a scalar-valued functional is needed for Theorem 1(ii). Let $\phi(\theta(\cdot)) = \theta(r)$ for some (any) $r \in \mathbb{R}$, so $\theta(\cdot) \leq 0(\cdot) \implies \phi(\theta(\cdot)) \leq 0$, satisfying the condition of Theorem 1(ii) that $\Theta_0 \subseteq \{\theta(\cdot) : \phi(\theta(\cdot)) \leq 0\}$. Let $D_n \equiv \phi(X_n(\cdot)) = \sqrt{n}(\hat{F}_X(r) - F_{0,n}(r))$ and $\gamma_n \equiv \phi(\theta_n(\cdot)) = \sqrt{n}(F_X(r) - F_{0,n}(r))$. Writing $Z_i = \mathbb{1}\{X_i \leq r\}$, then $F_X(r) = \mathrm{E}(Z_i)$ and $\hat{F}_X(r) = n^{-1}\sum_{i=1}^{n} Z_i$, i.e., we are concerned only with the mean of a random variable. The asymptotic sampling distribution of $D_n - \gamma_n$ is $\mathrm{N}\big(0, F_X(r)[1 - F_X(r)]\big)$, 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 $\phi_2(\theta(\cdot)) = (\theta(r), \theta(t))$ for $r, t \in \mathbb{R}$, and $\phi_3(\mathbf{p}) = p_1$ so $\phi_3(\phi_2(\theta(\cdot))) = \theta(r) = \phi(\theta(\cdot))$. Also, $\theta(\cdot) \leq 0(\cdot) \implies (\theta(r), \theta(t)) \leq \mathbf{0}$, 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 $\{\theta(\cdot) : \theta(r) \geq 0\}$, as well as most of the complement half-space (e.g., if $\theta(r) < 0$ but $\theta(t) \geq 0$), so it cannot be contained in any half-space.

For the two-sample setting, the infinite-dimensional limits in Appendix G.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.

G.2   SD1: results from limit experiment

We derive the infinite-dimensional limit experiment and then compute certain results. Continuing some notation from Appendix G.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

$$F_X \text{ SD}_1 F_0 \iff F_X(\cdot) \leq F_0(\cdot) \iff \theta(\cdot) \leq 0(\cdot). \tag{G.1}$$

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 (6),

$$\begin{aligned} X_n(\cdot) &\equiv \sqrt{n}\big(\hat{F}_X(\cdot) - F_{0,n}(\cdot)\big) \\ &= \sqrt{n}\big(\hat{F}_X(\cdot) - F_X(\cdot)\big) + \sqrt{n}\big(F_X(\cdot) - F_{0,n}(\cdot)\big) \\ &\rightsquigarrow B\big(F_X(\cdot)\big) + \theta(\cdot), \end{aligned} \tag{G.2}$$

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 (7),

$$\theta_n(\cdot) - X_n(\cdot) = \sqrt{n}\big(F_X(\cdot) - \hat{F}_X(\cdot)\big) \rightsquigarrow B\big(F_X(\cdot)\big), \tag{G.3}$$

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

$$F_X \; \mathrm{SD}_1 \; F_Y \iff F_X(\cdot) \le F_Y(\cdot) \iff \theta(\cdot) \le 0(\cdot). \tag{G.4}$$

For the sampling distribution,

$$\begin{aligned} X_n(\cdot) &\equiv \sqrt{n}(\hat{F}_X(\cdot) - \hat{F}_Y(\cdot) - \Delta_{0,n}(\cdot)) \\ &= \sqrt{n}(\hat{F}_X(\cdot) - F_X(\cdot)) - \sqrt{n}(\hat{F}_Y(\cdot) - F_Y(\cdot)) + \sqrt{n}(\Delta(\cdot) - \Delta_{0,n}(\cdot)) \\ &\rightsquigarrow B_1(F_X(\cdot)) - B_2(F_Y(\cdot)) + \theta(\cdot), \end{aligned} \tag{G.5}$$

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,

$$\begin{aligned} \theta_n(\cdot) - X_n(\cdot) &= \sqrt{n}(F_X(\cdot) - \hat{F}_X(\cdot)) - \sqrt{n}(F_Y(\cdot) - \hat{F}_Y(\cdot)) \\ &\rightsquigarrow B_1(F_X(\cdot)) - B_2(F_Y(\cdot)). \end{aligned} \tag{G.6}$$

Three results are now stated, followed by proofs.

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* (G.3) *and for two-sample SD1 testing in* (G.6). *Given $X(\cdot) = 0(\cdot)$, the posterior probability of SD1 is zero in both one-sample and two-sample settings.*

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

Proposition 3. *Consider limit experiments for one-sample SD1 testing in* (G.2) *and* (G.3) *and for two-sample SD1 testing in* (G.5) *and* (G.6). *Consider the Bayesian test from*

Method 1 that rejects $H_0 \colon \theta(\cdot) \le 0(\cdot)$ iff the posterior probability of H_0 is below α. Then, the Bayesian test's 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.*

Proof of Proposition 2. In the one-sample case, using (G.3),

$$\mathrm{P}(F_X \text{ SD}_1 \ F_0 \mid X(\cdot) = 0(\cdot)) = \mathrm{P}(\theta(\cdot) \le 0(\cdot) \mid X(\cdot) = 0(\cdot)) = \mathrm{P}(B(F_X(\cdot)) \le 0(\cdot))$$

$$= \mathrm{P}(B(\cdot) \le 0(\cdot)) = \mathrm{P}(\sup_{t \in [0,1]} B(t) \le 0) = 0. \qquad \text{(G.7)}$$

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 (G.7) 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,[10] 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 (G.6),

$$\mathrm{P}(F_X \text{ SD}_1 \ F_Y \mid X(\cdot) = 0(\cdot)) = \mathrm{P}(\theta(\cdot) \le 0(\cdot) \mid X(\cdot) = 0(\cdot)) = \mathrm{P}(T(\cdot) \le 0(\cdot))$$

$$= \mathrm{P}(\sup_{r \in \mathbb{R}} T(r) \le 0). \qquad \text{(G.8)}$$

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

$$B_1(t) = W_1(t) - tW_1(1), \quad B_2(t) = W_2(t) - tW_2(1),$$

$$V(\cdot) \equiv W_1(F_X(\cdot)) - W_2(F_Y(\cdot)) \overset{d}{=} \sqrt{2}W((F_X(\cdot) + F_Y(\cdot))/2),$$

$$Z(\cdot) \equiv V(\cdot) - T(\cdot) = F_X(\cdot)W_1(1) - F_Y(\cdot)W_2(1).$$

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

---

[10]Thanks to Iosif Pinelis for help extending to $F_X \ne F_Y$: 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 (G.8) 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.[11] $\qquad\square$

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

$$\mathrm{P}(cB_1(F(\cdot)) + cB_2(F(\cdot)) \leq 0(\cdot)) = \mathrm{P}(\sqrt{2}cB(F(\cdot)) \leq 0(\cdot)) = \mathrm{P}(B(\cdot) \leq 0(\cdot)) = 0, \quad \text{(G.9)}$$

again using (G.7), where $B_1(\cdot)$, $B_2(\cdot)$, and $B(\cdot)$ are independent standard Brownian bridges.

---

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

For any set S, with $X(\cdot) \sim cB_2(F(\cdot))$ since $\theta(\cdot) = 0(\cdot)$,

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

by (G.9), where $cB_2(F(\cdot))$ is the sampling distribution of $X(\cdot)$ given $\theta(\cdot) = 0(\cdot)$, and in the posterior $\theta(\cdot) \sim cB_1(F(\cdot)) + X(\cdot)$. Consequently,

$$
P(\theta(\cdot) \leq 0(\cdot) \mid X(\cdot) \in S) \, P(X(\cdot) \in S) \leq 0. \tag{G.10}
$$

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

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

If $P(X(\cdot) \in S) > 0$, then the left-hand side of (G.10) is the product of two strictly positive terms (assuming $\alpha > 0$), which is strictly positive. This contradicts (G.10) since the right-hand side is zero. Consequently, $P(X(\cdot) \in S) = 0$ and thus $P(X(\cdot) \notin S) = 1$, i.e., the rejection probability is one. (This does not mean S is empty, just that it is a zero-probability set.) $\qquad\square$

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. $\qquad\square$

G.3 SD1: finite-sample simulations

The following simulation results reflect the theoretical results from the limit experiment (discussed in Appendix G.1). Code for replication is provided.

Table 4 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, we set $X_i = i/(n+1)$ 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 $n < \infty$, the Bayesian bootstrap variant of Banks (1988) is used. When $n = \infty$, the results are from Proposition 2.

Table 4: 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

Table 4 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 5. The same intuition before Proposition 2 applies here. 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 4 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 5: 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 5 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 4. The entries for $n = \infty$ use Proposition 3 and Corollary 4.

Table 5 shows the same patterns as Table 4. 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.[12]

---

[12]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.