Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **00**, 1–26
https://doi.org/10.1093/jrsssb/qkad050

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series B Statistical Methodology

Original Article

A kernel Stein test for comparing latent variable models

Heishiro Kanagawa[1], Wittawat Jitkrittum[2,3], Lester Mackey[4], Kenji Fukumizu[5] and Arthur Gretton[1]

[1]Gatsby Computational Neuroscience Unit, University College London, London, UK
[2]Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]Google Research, New York, NY, USA
[4]Microsoft Research New England, Cambridge, MA, USA
[5]The Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan

Address for correspondence: Heishiro Kanagawa, Gatsby Computational Neuroscience Unit, 46 Cleveland Street, London W1T 4JG, UK. Email: heishiro.kanagawa@gmail.com

Abstract

We propose a kernel-based nonparametric test of relative goodness of fit, where the goal is to compare two models, both of which may have unobserved latent variables, such that the marginal distribution of the observed variables is intractable. The proposed test generalizes the recently proposed kernel Stein discrepancy (KSD) tests (Liu et al., *Proceedings of the 33rd international conference on machine learning* (pp. 276–284); Chwialkowski et al., (2016), In *Proceedings of the 33rd international conference on machine learning* (pp. 2606–2615); Yang et al., (2018), In *Proceedings of the 35th international conference on machine learning* (pp. 5561–5570)) to the case of latent variable models, a much more general class than the fully observed models treated previously. The new test, with a properly calibrated threshold, has a well-controlled type-I error. In the case of certain models with low-dimensional latent structures and high-dimensional observations, our test significantly outperforms the relative maximum mean discrepancy test, which is based on samples from the models and does not exploit the latent structure.

Keywords: hypothesis testing, kernel methods, mixture models, model selection, Stein's method

1 Introduction

A major approach to statistical modeling is the use of variables representing quantities that are unobserved but thought to underlie the observed data: well-known instances include probabilistic PCA (Roweis, 1997; Tipping & Bishop, 1999), factor analysis (see, e.g., Basilevsky, 1994), mixture models (see, e.g., Gilks et al., 1995), topic models for text (Blei et al., 2003), and hidden Markov models (HMMs) (Rabiner, 1989). The hidden structure in these generative models serves multiple purposes: it allows interpretability and understanding of model features [e.g., the topic proportions in a latent Dirichlet allocation (LDA) model of text], and it facilitates modeling by leveraging simple low-dimensional dynamics of phenomena observed in high dimensions (e.g., HMMs with a low-dimensional hidden state). Statistical modelers ultimately use such models to reason about the data; thus, in order to guarantee the validity of the inference, tools for comparing models and evaluating model fit are required.

This article addresses the problem of evaluating and comparing generative probabilistic models, in cases where the models have a latent variable structure, and the marginals over the observed data are intractable. In this scenario, one strategy for evaluating a generative model is to draw samples from it and to compare these samples to the modeled data using a two-sample test: for

Received: June 25, 2021. **Revised:** April 4, 2023. **Accepted:** April 6, 2023

instance, Lloyd and Ghahramani (2015) use a test based on the maximum mean discrepancy (MMD) (Gretton et al., 2012). This approach has two disadvantages, however: it is not computationally efficient due to the sampling step, and it does not take advantage of the information that the model supplies, for instance, the dependence relations among the variables.

Recently, an alternative model evaluation strategy based on Stein's method (Barbour, 1988; L. H. Y. Chen, 1975; Götze, 1991; Stein, 1972, 1986) has been proposed, which directly employs a closed-form expression for the unnormalized model. Stein's method is a technique from probability theory developed to prove central limit theorems with explicit rates of convergence (see, e.g., Ross, 2011). The core of Stein's method is that it characterizes a distribution with a *Stein operator*, which, when applied to a function, causes the expectation of the function to be zero under the distribution. For our purposes, we will use the result that a model-specific Stein operator may be defined, to construct a measure of the model's discrepancy. Notably, Stein operators may be obtained without computing the normalizing constant.

Stein operators have been used to design integral probability metrics (IPMs) (Müller, 1997) to test the goodness of fit of models. IPMs specify a *witness function* which has a large difference in expectation under the sample and model, thereby revealing the difference between the two. When a Stein operator is applied to the IPM function class, the expectation under the model is zero, leaving only the expectation under the sample. A Stein-modified $W^{2,\infty}$ Sobolev ball was used as the witness function class in Gorham and Mackey (2015) and Gorham et al. (2019). Subsequent work in Chwialkowski et al. (2016), Liu et al. (2016), and Gorham and Mackey (2017) used as the witness function class a Stein-transformed reproducing kernel Hilbert ball, as introduced by Oates et al. (2017): the resulting goodness-of-fit statistic is known as the kernel Stein discrepancy (KSD). Conditions for using the KSD in convergence detection were obtained by Gorham and Mackey (2017). While the foregoing work applies in continuous domains, the approach may also be used for models on a finite domain, where Stein operators (Bresler & Nagaraj, 2019; Hodgkinson et al., 2020; Ranganath et al., 2016; Reinert & Ross, 2019; Shi et al., 2022; Yang et al., 2018) and associated goodness-of-fit tests (Yang et al., 2018) have been established. Note that it is also possible to use Stein operators to construct feature dictionaries for comparing models, rather than using an IPM: examples include a test based on Stein features constructed in the sample space so as to maximize test power (Jitkrittum et al., 2018, 2017) and a test based on Stein-transformed random features (Huggins & Mackey, 2018). While the aforementioned tests address simple hypotheses, composite tests that use Stein characterizations have been proposed for specific parametric families including gamma (Betsch & Ebner, 2019b; Henze et al., 2012) and normal distributions (Betsch & Ebner, 2019c; Henze & Visagie, 2019), and general univariate parametric families (Betsch & Ebner, 2019a) (note that these tests are not based on IPMs).

While testing goodness of fit alone may be desirable for models of simple phenomena, it will often be the case that in complex domains, no model will fit the data perfectly. In this setting, it is more constructive to ask which model fits better, either within a class of models or in comparing different model classes. A likelihood-ratio test would be an ideal choice for this task, since it is uniformly most powerful (Lehmann & Romano, 2005). If the models contain latent variables, however, a likelihood-ratio test requires evaluating marginal densities of the models, which are typically intractable. A number of Monte Carlo techniques have been developed to estimate marginal densities or log-density ratios (see, e.g., Friel & Wyse, 2012, for a review). Constructing a test with such techniques is challenging, however; e.g., estimating each marginal density induces a bias in the likelihood ratio, which is difficult to characterize when designing a calibrated threshold (see Section 3.5 for a detailed discussion). Addressing the intractability of the likelihood, Bounliphone et al. (2016) proposed a purely sample-based relative goodness-of-fit test, which compares maximum mean discrepancies between the samples from two rival models with a reference real-world sample. A second relative test was proposed by Jitkrittum et al. (2018), generalizing (Jitkrittum et al., 2017) and learning the Stein features for which each model outperforms the other; this test requires marginal densities up to normalizing constants and does not apply to latent variable models.

In the present work, we introduce a novel relative goodness-of-fit test for latent variable models (LVMs), which compares models by computing approximate kernel Stein discrepancies. Our contribution is to provide a frequentist test of relative goodness of fit, with an approximate U-statistic

of the kernel Stein discrepancy difference as our test statistic. The statistic is expressed by a posterior expectation of the latent given an observation and is amenable to standard Markov Chain Monte Carlo techniques: in particular, it does not suffer from the challenges in characterizing bias observed in likelihood-ratio estimates for LVMs. Note that our approach differs from Bayesian model selection (Jeffreys, 1961; Kass & Raftery, 1995; Schwarz, 1978; Watanabe, 2013), which reports posterior odds (or Bayes factors) and does not concern controlling frequentist risks such as type-I error rates. To the best of our knowledge, our test represents the first general-purpose, frequentist, relative test for LVMs.

We recall the Stein operator and kernel Stein discrepancy in Section 2, and the notion of relative tests in Section 3. Our main theoretical contributions, also in Section 3, are twofold: first, we derive an appropriate test threshold to account for the randomness in the test statistic caused by sampling the latent variables. Second, we provide guarantees that the resulting test has the correct type-I level (i.e., that the rate of false positives is properly controlled) and that the test is consistent under the alternative: the number of false negatives drops to zero as we observe more data. Finally, in Section 4, we demonstrate our relative test of goodness of fit on a variety of LVMs. Our main point of comparison is the relative MMD test (Bounliphone et al., 2016), where we sample from each model. We demonstrate that the relative Stein test outperforms the relative MMD test in the particular case where the low-dimensional structure of the latent variables can be exploited.

2 The kernel Stein discrepancy and LVMs

In this section, we recall the definition of the Stein operator as used in goodness-of-fit testing, as well as the kernel Stein discrepancy, a measure of goodness of fit based on this operator. We will then introduce LVMs, which will bring us to the setting of relative goodness of fit with competing models in Section 3.

Before proceeding, we call attention to our setting: in this article, we treat both continuous- and discrete-valued observations, as formally defined at the outset of Section 2.1. It is our intention to study these two data modalities as they admit the same treatment. The subsequent definitions and analysis of our test are independent of whether a continuous or discrete Stein operator is used, apart from experiments concerning discrete-valued observations. Thus, the detail about discrete models in Section 2.1 may be initially skipped if desired.

2.1 Stein operators and kernel Stein discrepancies

Let $\mathcal{X}$ be the space in which the data take values; for $D \geq 1$, the space $\mathcal{X}$ is either the Euclidean space $\mathbb{R}^D$ or a finite lattice $\{0, \ldots, L-1\}^D$ for some $L > 1$. Depending on $\mathcal{X}$, we shall assume that the densities below are all defined with respect to the Lebesgue measure or the counting measure; i.e., the term *density* includes probability mass functions (pmfs).

Continuous-valued observations. Suppose that we are given data $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} R$ from an unknown distribution R, and we wish to test the goodness of fit of a model P. We first consider the case where the probability distributions P, R are defined on $\mathbb{R}^D$ and have respective probability densities p, r, where all density functions considered in this paper are assumed strictly positive and continuously differentiable. We treat the case of densities defined on bounded domains in the Online Supplementary Material, Section B.1. For differentiable density functions, we define the *score function*,

$$\mathbf{s}_p(x) \in \mathbb{R}^D := \frac{\nabla p(x)}{p(x)} = \nabla \log p(x),$$

where the gradient operator is $\nabla := [\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^D}]^\top$. The score is independent of the normalizing constant for p, making it computable even when p is known only up to normalization. Using this score, we define the *Langevin Stein operator* on a space $\mathcal{F}$ of differentiable functions from $\mathbb{R}^D$ to $\mathbb{R}^D$ (Gorham & Mackey, 2015; Oates et al., 2017),

$$\left[\mathcal{A}_P f\right](x) = \langle \mathbf{s}_p(x), f(x) \rangle + \langle \nabla, f(x) \rangle, \quad f \in \mathcal{F}.$$

A kernel discrepancy may be defined based on the Stein operator (Chwialkowski et al., 2016; Gorham & Mackey, 2017; Liu et al., 2016), which allows us to measure the departure of a distribution R from a model P. We define $\mathcal{F}$ to be a space comprised of D-dimensional vectors of functions $f = (f_1, \ldots f_D)$ where the dth function f_d is in a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Steinwart & Christmann, 2008, Definition 4.18) with a positive definite kernel $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (we use the same kernel for each dimension). The inner product on $\mathcal{F}$ is $\langle f, g \rangle_{\mathcal{F}} := \sum_{d=1}^{D} \langle f_d, g_d \rangle_{\mathcal{F}_k}$, and $\mathcal{F}_k$ denotes an RKHS of real-valued functions with kernel k.

The (Langevin) *kernel Stein discrepancy* (KSD) between P and R is defined as

$$\mathrm{KSD}(P \,\|\, R) = \sup_{\|f\|_{\mathcal{F}} \leq 1} |\mathbb{E}_{x \sim R} \mathcal{A}_P f(x) - \mathbb{E}_{y \sim P} \mathcal{A}_P f(y)|. \tag{1}$$

Under appropriate conditions on the kernel and measure P, the expectation $\mathbb{E}_{y \sim P} \mathcal{A}_P f(y) = 0$ for any $f \in \mathcal{F}$. To ensure this property, we will require that $k \in C^{(1,1)}$, the set of continuous functions on $\mathcal{X} \times \mathcal{X}$ with continuous first derivatives and that $\mathbb{E}_{y \sim P}[\|\mathbf{s}_p(y)\|_2] < \infty$ with $\|\cdot\|_2$ the Euclidean norm. We further assume that the following tail condition holds outside a bounded set: $p(x)\sqrt{k(x, x)} \leq C\|x\|_2^{\delta}$ for some constants $C > 0$ and $\delta > D - 1$ (see the clarification by South et al., 2021, p. 12, on the tail condition for the Stein's identity). With the vanishing expectation $\mathbb{E}_{y \sim P} \mathcal{A}_P f(y) = 0$, the KSD reduces to $\mathrm{KSD}(P \,\|\, R) = \sup_{\|f\|_{\mathcal{F}} \leq 1} |\mathbb{E}_{x \sim R} \mathcal{A}_P f(x)|$. The use of an RKHS as the function class yields a closed-form expression of the discrepancy by the kernel trick (Chwialkowski et al., 2016; Gorham & Mackey, 2017, Proposition 2),

$$\mathrm{KSD}^2(P \| R) = \mathbb{E}_{x, x' \sim R \otimes R}[h_p(x, x')],$$

if $\mathbb{E}_{x \sim R}[h_p(x, x)^{1/2}] < \infty$. Here, the symbol $R \otimes R$ denotes the product measure of two copies of R (so x and x' are independent random variables identically distributed with the law R). The function h_p (called a *Stein kernel*) is expressed in terms of the RKHS kernel k and the score function $\mathbf{s}_p$,

$$h_p(x, x') = \mathbf{s}_p(x)^{\top} \mathbf{s}_p(x') k(x, x') + \mathbf{s}_p(x)^{\top} k_1(x', x) + \mathbf{s}_p(x')^{\top} k_1(x, x') + k_{12}(x, x'),$$

where we have defined

$$k_1(a, b) := \nabla_x k(x, x')|_{x=a, x'=b},$$
$$k_{12}(a, b) := \nabla_x^{\top} \nabla_{x'} k(x, x')|_{x=a, x'=b}.$$

For a given i.i.d. sample $\{x_i\}_{i=1}^n \sim R$, the discrepancy has a simple closed-form finite sample estimate,

$$\mathrm{KSD}^2(P \| R) \approx \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j), \tag{2}$$

which is a U-statistic (Hoeffding, 1948). When the kernel is integrally strictly positive definite (ISPD) (Sriperumbudur et al., 2011, Section 2), and R admits a density r that satisfies $\mathbb{E}_{x \sim R} \|\nabla \log (p(x)/r(x))\|_2 < \infty$, we have that $\mathrm{KSD}(P \,\|\, R) = 0$ iff $P = R$ (Barp et al., 2019, Proposition 1). The earlier results of Chwialkowski et al. (2016) and Liu et al. (2016) require more stringent integrability conditions. Gorham and Mackey (2017, Theorem 7) showed that KSD can distinguish any Borel measure R from P by assuming conditions such as distant dissipativity (satisfied by finite Gaussian mixtures) (Gorham et al., 2019, Section 3). However, such conditions may be difficult to validate for LVMs. Thus, hereafter, we assume the former condition on the data distribution R.

Discrete-valued observations. We next recall the kernel Stein discrepancy in the discrete setting where distributions are defined on $\mathcal{X} = \{0, \ldots, L-1\}^D$ with $L > 1$, as introduced by Yang et al. (2018). In place of derivatives, we specify Δ_k as the cyclic forward difference w.r.t. kth coordinate: $\Delta_k f(x) = f(x^1, \ldots, \tilde{x}^k, \ldots, x^D) - f(x^1, \ldots, x^k, \ldots, x^D)$ where $\tilde{x}^k = x^k + 1 \bmod L$, with the

corresponding vector-valued operator $\Delta = (\Delta_1, \ldots, \Delta_D)$. The inverse operator Δ_k^{-1} is given by the backward difference $\Delta_k^{-1} f(x) = f(x^1, \ldots, x^k, \ldots, x^D) - f(x^1, \ldots, \bar{x}^k, \ldots, x^D)$, where $\bar{x}^k = x^k - 1 \bmod L$, and $\Delta^{-1} = (\Delta_1^{-1}, \ldots, \Delta_D^{-1})$. The score is then $\mathbf{s}_p(x) := p(x)^{-1} \Delta p(x)$, where it is assumed that the pmf is strictly positive (i.e., it is never zero). The difference Stein operator is then defined as $\mathcal{A}_P f(x) = \mathrm{tr}[f(x)\mathbf{s}_p(x)^\top + \Delta^{-1} f(x)]$, where it can be shown that $\mathbb{E}_{x\sim P}[\mathcal{A}_P f(x)] = 0$ (Yang et al., 2018, Theorem 2) (note that we include a trace for consistency with the continuous case—this does not affect the test statistic (Yang et al., 2018, Equation 10). We have defined the Stein operator and the score function slightly differently from Yang et al. (2018); the change is only in their signs, but this results in the same discrepancy. The difference Stein operator is not the only allowable Stein operator on discrete spaces: other alternatives are given by Yang et al. (2018, Theorem 3), Hodgkinson et al. (2020), and Shi et al. (2022). Although we focus on the Stein operator above, in practice, one might want to consider different Stein operators depending on the application. For instance, the score function $\mathbf{s}_p$ can be numerically unstable, as it contains the reciprocal $1/p(x)$; this can occur when the support of the model is severely mismatched to that of the data. In this particular case, one might choose the Barker–Stein operator proposed by Shi et al. (2022), an instance of the Zanella–Stein operator of Hodgkinson et al. (2020, Example 2). See Online Supplementary Material, Appendix B.2 for details. We compare this operator to the difference operator in an experiment where this mismatch occurs (Section 4.3.3).

As in the continuous case, the KSD can be defined as an IPM, given a suitable choice of reproducing kernel Hilbert space for the discrete domain. An example of kernel is the exponentiated Hamming kernel, $k(x, x') = \exp(-d_H(x, x'))$, where $d_H(x, x') = D^{-1} \sum_{d=1}^{D} \mathbb{I}(x^d \neq x'^d)$. The population KSD is again given by the expectation of the Stein kernel, $\mathrm{KSD}^2(P\|R) = \mathbb{E}_{(x,x')\sim R\otimes R}[h_p(x, x')]$, where h_p is defined as

$$h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') + \mathbf{s}_p(x)^\top k_1(x', x) + \mathbf{s}_p(x')^\top k_1(x, x') + k_{12}(x, x'),$$

and the kernel gradient is replaced by the inverse difference operator, e.g., $k_1(x, x') = \Delta_x^{-1} k(x, x')$, where Δ_x^{-1} indicates that the operator Δ^{-1} is applied with respect to the argument x. From Yang et al. (2018, Lemma 8), we have that $\mathrm{KSD}(P \,\|\, R) = 0$ iff $P = R$, under the conditions that the probability mass functions for P and R are positive and that the Gram matrix defined over all the configurations in $\mathcal{X}$ is strictly positive definite (i.e., the kernel is integrally strictly positive definite). One can define a kernel satisfying the required condition, for example, by embedding $\mathcal{X}$ into $\mathbb{R}^{L\times D}$ with one-hot encoding and using a Taylor-type kernel such as the exponentiated quadratic kernel (Christmann & Steinwart, 2010, Theorem 2.2).

2.2 Kernel Stein discrepancies of LVMs

Our objective is to use the KSD to evaluate LVMs, and here we formally specify our target model class. Let $\mathcal{L}_{\mathcal{X}|\mathcal{Z}} = \{p(\cdot \mid z) : z \in \mathcal{Z}\}$ be a family of probability density functions on $\mathcal{X}$ (we call these *likelihood* functions), which are indexed by elements of a set $\mathcal{Z}$. A LVM P is specified by such a family $\mathcal{L}_{\mathcal{X}|\mathcal{Z}}$ and a (prior) probability measure P_Z over $\mathcal{Z}$. The combination of these defines the marginal density function $p(x) = \int p(x \mid z)\, \mathrm{d}P_Z(z)$ and the posterior distribution $P_Z(\mathrm{d}z \mid x) = \{p(x \mid z)/p(x)\} P_Z(\mathrm{d}z)$. The distribution P induced by the former acts as a model of the distribution R underlying the observation, and the latter enables us to draw an inference over the unobserved variable.

Remark In our notation, the variable z can represent multiple latent variables. The likelihood $p(x \mid z)$ often contains parameters, but the dependency on these is suppressed here. If a prior is defined on a parameter, we may treat it as a latent variable; this consideration is relevant to predictive distributions. The likelihood and the prior in a model may be conditioned on some fixed data (e.g., they can be posterior predictive distributions), which we require to be independent of the data used for testing—in such a case, we omit the dependency on the held-out data. For examples, we refer the reader to Section 4.

The definition of the KSD remains the same for LVMs, but an additional difficulty arises in its estimation. Unfortunately, the U-statistic estimator given in (2) requires the score function of the marginal p, which is challenging to obtain due to the intractability of marginalizing out the latent variable. We will address this challenge by rewriting the score function in terms of the posterior distribution of the latent. In the following, we focus on continuous variable models, but the same conclusion holds for discrete counterparts by replacing gradient operation with cyclic differences.

Under a regularity condition, the score function can be expressed as

$$\mathbf{s}_p(x) = \mathbb{E}_{z|x}[\mathbf{s}_p(x \mid z)], \tag{3}$$

where $\mathbf{s}_p(x \mid z)$ is the score function of the conditional $p(x \mid z)$; i.e., $\mathbf{s}_p(x \mid z) = p(x \mid z)^{-1}\nabla_x p(x \mid z)$ for continuous-valued x. The reasoning is as follows:

$$\begin{aligned}\frac{\nabla_x p(x)}{p(x)} &= \frac{1}{p(x)}\int \nabla_x p(x \mid z)\, \mathrm{d}P_Z(z) \\ &= \int \frac{\nabla_x p(x \mid z)}{p(x \mid z)} \cdot \frac{p(x \mid z)\, \mathrm{d}P_Z(z)}{p(x)} = \mathbb{E}_{z|x}[\mathbf{s}_p(x \mid z)],\end{aligned}$$

where we have assumed the exchangeability of differentiation and integration: $\nabla_x p(x) = \int \nabla_x p(x \mid z)\, \mathrm{d}P_Z(z)$. The identity (3) is an analogue of Fisher's identity (Dempster et al., 1977; Fisher, 1925), which pertinently formed the basis for Stein control variate methodology in Friel et al. (2016), parameter inference for doubly intractable models via score matching (Vértes & Sahani, 2016), and Bayesian model selection with a Hyvärinen score (Dawid & Musio, 2015; Shao et al., 2019). Note that the conditional score $\mathbf{s}_p(x \mid z)$ is typically possible to evaluate. For example, consider the following simple form of an exponential family density $p(x \mid z) \propto \exp(T(x)\eta(z))$ defined on $\mathbb{R}^D$ with $T(x) : \mathbb{R}^D \to \mathbb{R}$ and $\eta : \mathcal{Z} \to \mathbb{R}$; for this density, $\mathbf{s}_p(x \mid z) = \eta(z)\nabla_x T(x)$. As can be seen in this example, the formula (3) does not require the likelihood $p(x \mid z)$ to be normalized. This feature eliminates the need for estimating the normalizing constant of $p(x \mid z)$ for each z, which is required to compute goodness-of-fit measures based on the marginal density $p(x)$ (Friel & Wyse, 2012); Online Supplementary Material, Section C.2 in the supplementary presents a use case with a truncated model.

With this identity, the KSD is rewritten as follows.

Lemma 1 Let

$$\begin{aligned}H_p[(x, z), (x', z')] = \mathbf{s}_p(x \mid z)^\top \mathbf{s}_p(x' \mid z')k(x, x') + \mathbf{s}_p(x \mid z)^\top k_1(x', x) \\ + k_1(x, x')^\top \mathbf{s}_p(x' \mid z') + k_{12}(x, x').\end{aligned} \tag{4}$$

Assume $\mathbb{E}_{(x,z),(x',z')\sim\tilde{R}\otimes\tilde{R}}|H_p[(x, z), (x', z')]| < \infty$ with the joint distribution $\tilde{R}(\mathrm{d}(x, z)) = P_Z(\mathrm{d}z \mid x)R(\mathrm{d}x)$. If the formula (3) holds, then,

$$\mathrm{KSD}^2(P\|R) = \mathbb{E}_{(x,z),(x',z')\sim\tilde{R}\otimes\tilde{R}} H_p[(x, z), (x', z')].$$

Proof. Substituting the formula (3) in the definition of KSD gives the required equation by the Tonelli–Fubini theorem. □

Remark The integrability assumption holds trivially if the input space $\mathcal{X}$ is finite, while care needs to be taken otherwise. The condition can be checked by examining the absolute integrability of each term in (4). The integrability assumption on the fourth term is mild and is satisfied by common kernels, e.g., the exponentiated quadratic or the inverse multi-quadratic kernels. The condition on the other

terms needs to be checked on a model-by-model basis. It can be shown that the example models in Section 4 satisfy the assumption (please see Online Supplementary Material, Section B.3 in the supplementary material for details).

The new KSD expression is an expectation of a computable symmetric kernel, and constructing an unbiased estimate is straightforward once we obtain a sample. In practice, when the model is complex, sampling from the posterior distribution generally requires simulation, as the posterior is not available in a closed form. Therefore, we propose to approximate the expectation by Markov Chain Monte Carlo (MCMC) methods and construct an approximate U-statistic estimator as follows. Let $\mathbf{z}_i^{(t)} = (z_{i,1}^{(t)}, \ldots, z_{i,m}^{(t)}) \in \mathcal{Z}^m$ be a latent sample of size m drawn by an MCMC method having $P_Z(\cdot \mid x_i)$ as its invariant measure after t burn-in iterations. Let $\bar{\mathbf{s}}_p(x_i \mid \mathbf{z}_i^{(t)}) = \frac{1}{m}\sum_{j=1}^m \mathbf{s}_p(x_i \mid z_{i,j}^{(t)})$. Given a joint sample $\{(x_i, \mathbf{z}_i^{(t)})\}_{i=1}^n$, we estimate the KSD by

$$U_n^{(t)}(P) := \frac{1}{n(n-1)} \sum_{i \neq j} \bar{H}_p[(x_i, \mathbf{z}_i^{(t)}), (x_j, \mathbf{z}_j^{(t)})], \tag{5}$$

where

$$\begin{aligned}\bar{H}_p[(x_i, \mathbf{z}_i^{(t)}), (x_j, \mathbf{z}_j^{(t)})] = \bar{\mathbf{s}}_p(x_i \mid \mathbf{z}_i^{(t)})^\top \bar{\mathbf{s}}_p(x_j \mid \mathbf{z}_j^{(t)}) k(x_i, x_j) + \bar{\mathbf{s}}_p(x_i \mid \mathbf{z}_i^{(t)})^\top k_1(x_j, x_i) \\ + k_1(x_i, x_j)^\top \bar{\mathbf{s}}_p(x_j \mid \mathbf{z}_j^{(t)}) + k_{12}(x_i, x_j),\end{aligned}$$

and the sum is taken over all distinct sample pairs. If $P_Z^{(t)}(\mathrm{d}\mathbf{z} \mid x)$ denotes the distribution of an MCMC sample $\mathbf{z}^{(t)} = (z_1^{(t)}, \ldots, z_m^{(t)})$, then this estimator is indeed a U-statistic, but its expectation is that of kernel $\bar{H}_p$ with respect to $P_Z^{(t)}(\mathrm{d}\mathbf{z} \mid x)R(\mathrm{d}x)$ instead of $P_Z(\mathrm{d}\mathbf{z} \mid x)R(\mathrm{d}x)$. Thus, the estimator is biased against the target estimand, the model's KSD, for a finite burn-in period t, and can therefore be seen an approximation to the *true* U-statistic $U_n^{(\infty)}$. Designing a statistical test requires understanding the behavior of the statistic (5), and we will provide its analysis in the next section. Although we focus on MCMC for its approximate unbiasedness in our proposed test, different posterior approximations may be considered in other applications; for example, with a more computationally efficient approach (e.g., variational approximation), the new KSD expression in Lemma 1 might allow us to consider parameter estimation for unnormalized statistical models with latent variables (Barp et al., 2019).

3 A relative goodness-of-fit test

We now address the setting of statistical testing for model comparison. We begin this section with our problem settings and notation, and then define a test by showing the asymptotic normality of approximate U-statistics.

3.1 Problem setup

We consider the case where we have two LVMs P and Q, and we wish to determine which is a closer approximation of the distribution R generating our data $\{x_i\}_{i=1}^n$. The respective density functions of the models are given by the integrals $p(x) = \int p(x \mid z)\, \mathrm{d}P_Z(z)$ and $q(x) = \int q(x \mid w)\, \mathrm{d}Q_W(w)$. As with P, the latent variable w is assumed to take values in a set $\mathcal{W}$ with prior Q_W. We assume that $p(x)$ and $q(x)$ cannot be tractably evaluated, even up to their normalizing constants. Our goal is to determine the *relative* goodness of fit of the models by comparing each model's discrepancy from the data distribution. Our problem is formulated as the following hypothesis test:

$$\begin{aligned} &H_0 : \mathrm{KSD}(P \,\|\, R) \leq \mathrm{KSD}\big(Q \,\|\, R\big) \text{ (null hypothesis)}, \\ &H_1 : \mathrm{KSD}(P \,\|\, R) > \mathrm{KSD}\big(Q \,\|\, R\big) \text{ (alternative)}. \end{aligned} \tag{6}$$

In other words, the null hypothesis is that the fit of P to R (in terms of KSD) is as good as Q, or better. Note that the KSD in (6) is defined by a particular reproducing kernel, and thus different

kernels yield distinct hypotheses. For kernel selection, we refer the reader to Section 3.4.

We next provide an overview of the formal assumptions made throughout the paper. Let $(\Omega, \mathcal{S}, \Pi)$ be a probability space, where Ω is a sample space, $\mathcal{S}$ is a σ-algebra, and Π is a probability measure. All random variables (for example, data points x_i and draws $\mathbf{z}_i^{(t)}$ from a Markov chain sampler) are understood as measurable functions from the sample space Ω. The input space $\mathcal{X}$ is equipped with the Borel σ-algebra generated by its standard topology. We assume that $\mathcal{Z}$, $\mathcal{W}$ are Polish spaces with the Borel σ-algebras defined by their respective topologies, on which the priors P_Z, Q_W are defined. Finally, we require that the two models are distinct; i.e., their marginal densities disagree on a set of positive R-measure.

3.2 Estimating kernel Stein discrepancies of LVMs

The hypotheses in (6) can be equally stated in terms of the difference of the (squared) KSDs, $\mathrm{KSD}^2(P\|R) - \mathrm{KSD}^2(Q\|R)$, which motivates us to design a test statistic by estimating each term. Let $U_n^{(t)}(P, Q) := U_n^{(t)}(P) - U_n^{(t)}(Q)$ be the difference of KSD estimates, where $U_n^{(t)}(Q)$ is defined as for $U_n^{(t)}(P)$ in (5). Note that $U_n^{(t)}(P, Q)$ is an *approximate* U-statistic (in the sense of the final paragraph in Section 2.2) defined by the difference kernel

$$\bar{H}_{p,q}[(x, \mathbf{z}, \mathbf{w}), (x', \mathbf{z}', \mathbf{w}')] := \bar{H}_p[(x, \mathbf{z}), (x', \mathbf{z}')] - \bar{H}_q[(x, \mathbf{w}), (x', \mathbf{w}')]$$

evaluated on the joint sample $\{(x_i, \mathbf{z}_i^{(t)}, \mathbf{w}_i^{(t)})\}_{i=1}^n$. The statistic takes as input random variables with evolving laws, and defining a test require us to understand the behavior of such statistics. This section delivers an analysis in a general setting.

We first characterize the asymptotic distribution of an approximate U-statistic. The following theorem shows that such a statistic is asymptotically normal around the expectation of the true U-statistic provided its bias vanishes fast.

Theorem 1 (Asymptotic normality). Let $\{\gamma_t\}_{t=1}^\infty$ be a sequence of Borel probability measures on a Polish space $\mathcal{Y}$ and γ be another Borel probability measure. Let $\{Y_i^{(t)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \gamma_t$, and for a symmetric function $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, define a U-statistic and its mean by

$$U_n^{(t)} = \frac{1}{n(n-1)} \sum_{i \neq j} h(Y_i^{(t)}, Y_j^{(t)}), \quad \theta_t = \mathbb{E}_{(Y,Y') \sim \gamma_t \otimes \gamma_t}[h(Y, Y')].$$

Let $\theta = \mathbb{E}_{(Y,Y') \sim \gamma \otimes \gamma}[h(Y, Y')]$. Let $v_t := \mathbb{E}_{(Y,Y') \sim \gamma_t \otimes \gamma_t}[|\tilde{h}_t(Y, Y')|^3]^{1/3}$ with $\tilde{h}_t = h - \theta_t$, and assume $\limsup_{t \to \infty} v_t < \infty$. Assume that $\sigma_t^2 = 4\mathrm{Var}_{Y' \sim \gamma_t}[\mathbb{E}_{Y \sim \gamma_t}[h(Y, Y')]]$ converges to a constant σ^2. Assume that we have $\theta_t \to \theta$ as $t \to \infty$. Then, in the limit of large n and of t growing as a function of n such that $\sqrt{n}(\theta_t - \theta) \to 0$, the following two statements hold: if $\sigma > 0$, we have

$$\sqrt{n}\left(U_n^{(t)} - \theta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where $\xrightarrow{d}$ denotes convergence in distribution; in the case $\sigma = 0$, $\sqrt{n}(U_n^{(t)} - \theta) \to 0$ in probability.

The proof is in Online Supplementary Material, Section A in the supplement. Note that in the preceding and following results, the limit of n and t is taken simultaneously rather than sequentially, such that the condition $\sqrt{n}(\theta_t - \theta) \to 0$ holds: see discussion below and in Section 3.3. By letting $Y_i^{(t)} = (x_i, \mathbf{z}_i^{(t)}, \mathbf{w}_i^{(t)})$ and $h = \bar{H}_{p,q}$ in the foregoing theorem, we obtain the same conclusion for the difference estimate $U_n^{(t)}(P, Q)$.

The asymptotic normality allows us to define a test procedure. Theorem 1 involves unknown variance σ^2, however; in order to construct a test, we need to be able to estimate it consistently. For our test, we propose to use the following jackknife variance estimator:

$$v_{n,t} := (n-1)\sum_{i=1}^{n}(U_{n,-i}^{(t)} - U_n^{(t)})^2, \tag{7}$$

where $U_n^{(t)}$ is defined as in Theorem 1, and $U_{n,-i}^{(t)}$ the U-statistic computed on the sample with the ith data point removed. We defer the discussion on this choice until we introduce our test procedure in Section 3.3. Here, we present the required consistency, the proof of which can be found in Online Supplementary Material, Appendix A.2 (see Online Supplementary Material, Lemma 1).

Lemma 2 (Jackknife consitency). Define symbols as in Theorem 1 and the jackknife variance estimator as in (7). Assume

$$\limsup_{t\to\infty} \mathbb{E}_{(Y,Y')\sim\gamma_t\otimes\gamma_t}[h(Y, Y')^4] < \infty.$$

Let $\sigma^2 = \lim_{t\to\infty}\sigma_t^2$ where $\sigma_t^2 = 4\zeta_{1,t} = 4\mathrm{Var}_{Y\sim\gamma_t}[\mathbb{E}_{Y'\sim\gamma_t}[h(Y, Y')]]$. Then, we have the double limit $\mathbb{E}(v_{n,t} - \sigma^2)^2 \to 0$ as $n, t \to \infty$. In particular, the limit holds regardless of the growth rate of t as a function of n.

We have shown that the jackknife estimator allows consistent estimation of the asymptotic variance of $U_n^{(t)}(P, Q)$. Using the results obtained in this section, we present our test procedure in the next section.

3.3 Test procedure

We are finally ready to define the test procedure. Recall that our objective is to compare model discrepancies, which can be accomplished by estimating the difference $\mathrm{KSD}^2(P\|R) - \mathrm{KSD}^2(Q\|R)$. The previous section has established the asymptotic normality of the difference estimate $U_n^{(t)}(P, Q)$ and provides a consistent estimator of its asymptotic variance. Therefore, we define our test statistic to be

$$T_{n,t} = \sqrt{n}\frac{U_n^{(t)}(P, Q)}{\sqrt{v_{n,t}}}, \tag{8}$$

with $v_{n,t}$ being the jackknife variance estimator in (7) computed using the joint sample $\{(x_i, \mathbf{z}_i^{(t)}, \mathbf{w}_i^{(t)})\}_{i=1}^n$ and kernel $h = \bar{H}_{p,q}$.

The following property follows from Theorem 1 and Slutsky's lemma (see e.g., van der Vaart, 2000, p. 13) along with the consistency of $v_{n,t}$ from Lemma 2.

Corollary 1 Let $\mu_{P,Q} = \mathrm{KSD}^2(P\|R) - \mathrm{KSD}^2(Q\|R)$. Let $\mathbf{y}_1^{(t)}$ and $\mathbf{y}_2^{(t)}$ be i.i.d. variables; $\mathbf{y}_1^{(t)}$ represents a copy of random variables $(x, \mathbf{z}^{(t)}, \mathbf{w}^{(t)})$; the variables $\mathbf{z}^{(t)}, \mathbf{w}^{(t)}$ are draws from the respective Markov chains of P and Q conditioned on x after t burn-in steps, and x obeys R. Assume $\limsup_{t\to\infty}\mathbb{E}[\bar{H}_{p,q}[\mathbf{y}_1^{(t)}, \mathbf{y}_2^{(t)}]^4] < \infty$. If the assumptions in Theorem 1 hold for the statistic $U_n^{(t)}(P, Q)$ with asymptotic variance $\sigma_{P,Q}^2 > 0$, we have $\sqrt{n}(U_n^{(t)}(P, Q) - \mu_{P,Q})/\sqrt{v_{n,t}} \xrightarrow{d} \mathcal{N}(0, 1)$ as $n, t \to \infty$, where the required growth of t as a function of n is as in Theorem 1.

Remark Corollary 1 holds for any choice of the Markov chain sample size $m \geq 1$. However, in practice, a small value of m leads to large variance of the score estimates $\bar{\mathbf{s}}_p, \bar{\mathbf{s}}_q$, and hence the test statistic $T_{n,t}$, which results in a conservative test. To improve the test's sensitivity, we therefore recommend using as large an m as possible.

Corollary 1 leads to the following simple model comparison test (summarized in Algorithm 1): for a given significance level $\alpha \in (0, 1)$, we compare the test statistic $T_{n,t}$ against the $(1-\alpha)$-quantile $\tau_{1-\alpha}$ of the standard normal and reject the null if $T_{n,t}$ exceeds $\tau_{1-\alpha}$. By this design, under the null hypothesis $H_0 : \mu_{P,Q} \leq 0$, we have $\lim_{n,t\to\infty} \Pi(T_{n,t} > \tau_{1-\alpha} \mid H_0) \leq \alpha$, and the test is therefore asymptotically level α for each fixed R satisfying H_0. On the other hand, under any fixed alternative $H_1 : \mu_{P,Q} > 0$, it follows from $\sqrt{n}\mu_{P,Q} \to \infty$ $(n \to \infty)$ that we have $\lim_{n,t\to\infty} \Pi(T_{n,t} > \tau_{1-\alpha} \mid H_1) = 1$, indicating that the test is consistent in power.

We remark that the above analysis will not apply in particular extreme cases, where both models are identical, or both perfectly match the data distribution. When these occur, then $\sigma_{P,Q} = 0$ and $\mu_{P,Q} = 0$ (note that if $\mu_{p,Q} \neq 0$, the test statistic diverges as the sample size increases). Applying our procedure as above to this setting, the normal approximation might fail to correctly capture the variability of the test statistic, and the type-I error could exceed the significance level. To detect this failure mode, we would need to independently check that the two models are not identical, either by inspection or via two-sample testing. A more systematic treatment could be performed, e.g., by preventing degeneracy using a sample splitting technique as proposed by Schennach and Wilhelm (2017), and we leave this refinement for future work.

We empirically found that our choice of the variance estimator acted as a safeguard against the failure mode mentioned above. The jackknife estimator is nonnegative, while individually estimating the variances and covariance of the two U-statistics might yield a negative estimate. The jackknife is also known to overestimate the variance (Efron & Stein, 1981), and its use may result in a more conservative test. This estimator is not the only allowable choice, as the variance estimation in U-statistics has been extensively studied (for other concrete estimators, see, e.g., Maesono, 1998, and references therein). In our preliminary analysis, we considered two other estimators, but the jackknife estimator controlled type-I errors better than these alternatives in the *near degenerate* case. For details, we refer the reader to experiments in Online Supplementary Material, Sections C.4 and C.5 in the supplement.

The limiting behaviors of the test are only guaranteed when an appropriate double limit is taken with respect to the burn-in size t and the sample size n. Theorem 1 suggests that the bias of the statistic $U_n^{(t)}(P, Q)$ should decay faster than $1/\sqrt{n}$ in the limit of t. Our practical recommendation is to take a burn-in period as long as the computational budget allows; this heuristic is justified if the bias vanishes as $t \to \infty$. For $\mathrm{KSD}^2(P\|R)$ and its estimate, the bias is due to that of the score estimate $\mathbf{s}_p^{(t)}(x) = \mathbb{E}_{\mathbf{z}\mid x}^{(t)}[\bar{\mathbf{s}}_p(x \mid \mathbf{z})]$, where $\mathbb{E}_{\mathbf{z}\mid x}^{(t)}$ denotes the expectation with respect to $P_Z^{(t)}(\mathrm{d}\mathbf{z} \mid x)$. If the score's bias is confirmed to converge to zero, we can check the bias of the KSD estimate by examining the convergence of $\mathbb{E}_{(x,x')\sim R\otimes R}[h_{p,t}(x, x')]$, with $h_{p,t}(x, x')$ a Stein kernel defined by the approximate score $\mathbf{s}_p^{(t)}$. The convergence of $\mathbf{s}_p^{(t)}$ can be established by assuming appropriate conditions on $\mathbf{s}_p(x \mid z)$ and the sampler; for instance, for the exponential family likelihood $p(x \mid z) \propto \exp(T(x)\eta(z))$, if the natural parameter η is a continuous bounded function, the weak convergence of the sampler implies that the desired convergence (the score $\mathbf{s}_p(x \mid z)$ is factorized as $\eta(z)\nabla T(x)$). The quantification of the required growth rate of t relative to n needs more stringent conditions on the employed MCMC sampler, which we discuss in the supplement, Online Supplementary Material, Section B.4. Admittedly, it is often not straightforward to theoretically establish an explicit relation between the growth rates of t and n. We therefore experimentally evaluate the finite-sample performance of our test in Section 4.

The overall computational cost of the proposed test is $O\{n^2 + n(t+m)\}$, assuming that the cost of sampling a latent is constant. The test statistic in (8) requires evaluating the U-statistic kernel $\bar{H}_{p,q}$ on all distinct sample pairs. Note that we need to perform this computation only once if we memoize the evaluated values; in particular, the cost of the variance estimate (2) can be made $O(n^2)$ with memoization. Thus, assuming that we have evaluated and stored the score values $\{\bar{\mathbf{s}}_p(x_i|\mathbf{z}_i^{(t)}), \bar{\mathbf{s}}_q(x_i|\mathbf{w}_i^{(t)})\}_{i=1}^n$, the cost of evaluating the U-statistic kernel is $O(n^2)$, but this operation can be easily parallelized over sample pairs. The additional $O\{n(t+m)\}$ cost comes from evaluating the approximate score functions, as it requires running Markov chains for each data point (see the loop between Lines 3.3–3.3 in Algorithm 1). We can improve the sample-size n dependency in score evaluation by parallelization, since MCMC can be performed independently over sample points x_i.

Algorithm 1: Test procedure

Input: Data $\{x_i\}_{i=1}^n$, models P, Q, and significance level α

Result: Test the null H_0

/*Form a joint sample $\{(x_i, \mathbf{z}_i^{(t)}, \mathbf{w}_i^{(t)})\}_{i=1}^n$ */

for $i \leftarrow 1$ **to** n **do**

  Generate m samples $\mathbf{z}_i^{(t)} = (z_{i,1}^{(t)}, \ldots, z_{i,m}^{(t)})$ after t burn-in steps with an MCMC algorithm to simulate $P_{Z|}(\mathrm{d}z \mid x_i)$;

  Generate m samples $\mathbf{w}_i^{(t)} = (w_{i,1}^{(t)}, \ldots, w_{i,m}^{(t)})$ after t burn-in steps with an MCMC algorithm to simulate $Q_{\mathbb{W}}(\mathrm{d}w \mid x_i)$;

end

$\tau_{1-\alpha} \leftarrow (1-\alpha)$-quantile of $\mathcal{N}(0, 1)$;

/*Compute test statistic $T_{n,t}$ in equation (8) */

Compute KSD difference estimate $U_n^{(t)}(P, Q)$;

Compute variance estimate $v_{n,t}$;

/*Direct computation of $T_{n,t} = \sqrt{n}U_n^{(t)}(P, Q)/\sqrt{v_{n,t}}$ can be numerically unstable */

If $U_n^{(t)}(P, Q) > (\sqrt{v_{n,t}}/\sqrt{n}) \cdot \tau_{1-\alpha}$ **then** Reject the null H_0;

3.4 Kernel choice

A discrepancy measure such as KSD encodes a particular sense of how two distributions differ. In the case of KSD, the magnitude of this discrepancy is affected not only by evaluated models but also the choice of a reproducing kernel. Ideally, we should choose a kernel that makes the KSD reflect the discrepancy of features relevant to the problem at hand. We provide general guidance on kernel selection as follows:

Continuous observations: As mentioned in Section 2, ISPD kernels enable the KSD to distinguish any two distributions satisfying certain regularity conditions. Of ISPD kernels, in the light of practical performances reported in prior work in goodness-of-fit testing (Gorham & Mackey, 2017) and distribution approximation (Chen et al., 2019; Riabiz et al., 2022), we advocate for the use of the preconditioned IMQ kernel (Chen et al., 2019)

$$k(x, x') = (c^2 + \|\Lambda^{-1/2}(x - x')\|_2^2)^{-\beta}, \tag{9}$$

with Λ being a strictly positive definite matrix, and scalars $c > 0$ and $0 < \beta < 1$; as a default choice, we recommend to take $\beta = 1/2$ and $c = 1$. Following the kernel method literature, we recommend to choose the pre-conditioner Λ in a data-dependent way so that the KSD can capture relevant features of the data. We suggest two default options: the median heuristic, where $\Lambda = \lambda^2 I$ with $\lambda = \mathrm{median}\{\|x_i - x_j\|_2 : 1 \le i < j \le n\}$ and I the identity matrix; the sample covariance $\Lambda = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top/(n-1)$ with $\bar{x} = \sum_{i=1}^n x_i/n$, which should be suitably regularized. Each of these choices has its own merits, as we illustrate in a simple example with Gaussian distributions in the supplement (Online Supplementary Material, Section B.7). Moreover, in general, the KSD is not invariant to a change of coordinates representing the data. The above choices partially address this issue, as they ensure that the KSD is invariant to rotation and displacement (see, Online Supplementary Material, Section B.9.2 in the supplement; Matsubara et al., 2022, Section 5.1). For continuous observations, we additionally need to examine the integrability of the Stein kernel to use the KSD expression of Lemma 1. To this end, one might want to make an assumption about the tail decay of the data distribution. The integrability condition can be alternatively enforced by reweighting the reproducing kernel so that the Stein kernel is uniformly bounded; i.e., for a kernel k, define a new kernel k_w by $k_w(x, x') = k(x, x')w(x)w(x')$ where $w : \mathcal{X} \to (0, \infty)$ is some decreasing function dominating the growth of the score function. We discuss how to choose such w in the supplement, Online Supplementary Material, Section B.3. This reweighting might reduce the sensitivity of the KSD and break the aforementioned property of coordinate-choice independence, however.

Discrete observations: We have given a condition for a kernel to be ISPD at the end of Section 2; e.g., the exponentiated quadratic kernel on one-hot encoding, which can be efficiently

implemented in a sparse tensor format. In general, however, it is challenging to compute such an ISPD kernel for discrete objects in high-dimensions. Note that ISPD-ness is only required to distinguish *any* two distributions. In practice, we only require the KSD to capture aspects relevant to model evaluation, and might therefore choose a kernel insensitive to some differences, as long as they represent computationally affordable alternatives suited to the given problem. An instructive example is testing on distributions over graphs. Graphs of V nodes can be represented as adjacency matrices that are elements of $\{0, 1\}^{V\times V}$. The Dirac delta kernel that examines if two graphs are identical is an ISPD kernel but computationally intractable (no polynomial algorithm is known). This notion of graph identification is in practice too restrictive, and therefore one typically uses kernels that convey other relevant graph properties (see, e.g., Borgwardt et al., 2020, for more details). We also demonstrate this trade-off in our experiment with latent Dirichlet allocation models in Section 4.3, where we can ignore the sequential structure of the data.

Use of multiple kernel functions: As we have seen, there are often multiple choices of the kernel function, and they might represent distinct features. Our recommendation is to test the hypotheses corresponding to the kernel choices simultaneously, as it makes evaluation more rigorous. However, one has to correct for multiple comparisons such as controlling the family-wise error rate. It should be noted that a correction typically makes the test more conservative as the number of kernels grows. The user thus needs to control the number of kernels to be used [e.g., using a handful of values of scale parameter λ for the above IMQ kernel with $\Lambda = \mathrm{diag}(\lambda, \ldots, \lambda)$].

Finally, we note that specific kernels can be employed that encode domain-specific expertise in particular problem settings: for instance, kernels have been defined on groups (Fukumizu et al., 2008) and graphs (Borgwardt et al., 2020). KSDs and associated statistical tests can likewise be defined for certain of these cases (e.g., Xu & Matsuda, 2020). That being said, it may sometimes be preferable to favor an MMD with goal-specific features over an omnibus KSD test.

3.5 Challenges with likelihood-ratio tests based on marginal density estimation

As noted in the introduction, a likelihood-ratio test is an alternative choice for comparing LVMs. This section details challenges associated with likelihood-ratio tests. There are two paths to designing such tests, depending on how we estimate the (log-)density ratio. One approach is to estimate the marginal density and take the ratio, and the other is to directly estimate the (log-)density ratio. We refer the reader to Friel and Wyse (2012) for a review of estimation techniques.

Methods such as annealed importance sampling (Neal, 2001) belong to the first category. The problem with this approach is that an estimator ratio is a (typically heavily) biased estimator (let alone the difference of log density estimates); thus, deriving a calibrated test threshold from such an estimator is often challenging. In contrast, in our test, we can characterize the bias due to MCMC relatively straightforwardly, by evaluating the bias of the approximate Stein kernel.

The second category includes techniques such as reversible jump MCMC (Green, 1995) and thermodynamical integration. It is anecdotally known that reversible jump MCMC can be challenging to implement, especially for complex models. Thermodynamic integration uses the formula

$$\log p(x) = \int_0^1 \mathbb{E}_{z\,|\,x,t}[\log p(x\,|\,z)]\mathrm{d}t,$$

where $\mathbb{E}_{z\,|\,x,t}$ is the expectation with respect to the posterior over z defined by the tempered likelihood $p(x\,|\,z)^t$. This technique can in principle be used to construct a likelihood-ratio test, as in our approach. For example, we may uniformly sample t from the unit interval to approximate the outer integral, while conducting MCMC to estimate the inner expectation. We are not aware of a frequentist test based on this construction, however. That said, the computational cost of such an approach would be significant: given data of size n, we would generate T samples for the temperature t, then m samples for the inner MCMC, giving a (naive) computational complexity of $O(nmT)$ (here we compute the sum of n log density evaluations rather than the evidence conditioned on the data batch).

4 Experiments

We evaluate the proposed test (LKSD, hereafter) through simulations. Our goal is to show the utility of the KSD in model comparison. To this end, we compare our test with the relative MMD test

(Bounliphone et al., 2016), a kernel-based frequentist test that supports a great variety of LVMs. Note that this test and ours address different hypotheses, as the MMD and LKSD tests use different discrepancy measures; it is indeed possible that they reach conflicting conclusions (e.g., a model is better in terms of KSD but worse in MMD). To align the judgement of both tests, we construct problems using models with controllable parameters; for a given class, a reference distribution, from which a sample is drawn, is chosen by fixing the model parameter; two candidate models are then formed by perturbing the reference's parameter such that a larger perturbation yields a worse model for both tests. We show that there are cases where the MMD fails to detect model differences whereas the KSD succeeds. For completeness, we provide the detail of our implementation of the MMD test in the supplement (Online Supplementary Material, Section B.5), since it requires modification to yield satisfactory performance in our setting.

Following are details shared by the experiments below. All results below are based on 300 trials, except for the experiment in Section 4.2 (see the section for details). In the light of the discussion in Section 3, unless specified, the MMD test draws $n_{\text{model}} = m + t$ samples for each model so that its cost matches the additional computation afforded to the LKSD test, which is of $O\{n(m+t)\}$ from MCMC.

4.1 Probabilistic principal component analysis

We first consider a simple model in which the score of its marginal is tractable. This allows us to separately assess the impact of employing a score function approximation. Probabilistic principal component analysis (PPCA) models serve this purpose since the marginals are given by Gaussian distributions. Let $\mathcal{X} = \mathbb{R}^D$ and $\mathcal{Z} = \mathbb{R}^{D_z}$ with $1 \leq D_z < D$. A PPCA model PPCA(A, ψ) is defined by

$$p(x \mid z, A, \psi) = \mathcal{N}(Az, \psi^2 I_x), \quad P_Z = \mathcal{N}(\mathbf{0}, I_z),$$

where $A \in \mathbb{R}^{D \times D_z}$, $I_x \in \mathbb{R}^{D \times D}$, $I_z \in \mathbb{R}^{D_z \times D_z}$ are the identity matrices, ψ is a positive scalar, and $\mathbf{0}$ is a vector of zeros. The conditional score function is $\mathbf{s}_p(x \mid z) = -(x - Az)/\psi^2$. In particular, the marginal density is given by $p(x) = \mathcal{N}(\mathbf{0}, AA^\top + \psi^2 I_x)$.

While the posterior in this model is tractable, it is instructive to see how KSD estimation is performed by MCMC. By using an MCMC method, such as the metropolis adjusted Langevin algorithm (MALA) (Besag, 1994; Roberts & Tweedie, 1996) or Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011), we obtain latent samples $\mathbf{z}_i \in \mathcal{Z}^m$ for each x_i, which forms a joint sample $\{(x_i, \mathbf{z}_i)\}_{i=1}^n$; samples $\mathbf{z}_i$ are used to compute a score estimate at each point x_i,

$$\bar{\mathbf{s}}_p(x_i \mid \mathbf{z}_i) = -\left\{ x_i - A\left(\frac{1}{m} \sum_{j=1}^{m} z_{i,j} \right) \right\} / \psi^2,$$

and these approximate score values are used to compute the U-statistic estimate in (5). By choosing suitably decaying kernels (Online Supplementary Material, Section B.3), we can guarantee the integrability condition in Lemma 1. The vanishing bias assumption in Theorem 1 corresponds to the convergence in mean, which can be measured by the Kantorovich–Rubinstein distance (Kantorovich, 2006) (also known as the L^1-Wasserstein distance (see, e.g., Villani, 2009, Chapter 6). Note that the negative logarithm of the unnormalized posterior density is strongly convex, and its gradient is Lipschitz; the strong convexity- and Lipschitz constants are independent of x. Therefore, using HMC for example, by appropriately choosing a duration parameter and a discretization step size, we can show that the bias of the above score estimate diminishes uniformly over x (Bou-Rabee et al., 2020).

4.1.1 Type-I error and test power

We investigate the finite-sample performance of the proposed test in terms of type-I error and power rates. We generate data from a PPCA model $R = \text{PPCA}(A, \psi)$. The dimensions of the observable and the latent are set to $D = 100$, $D_z = 10$, respectively. Each element of the weight matrix A is drawn from a uniform distribution $U[0, 1]$ and fixed. The variance parameter ψ is set to 1.

As PPCA models have tractable marginals, we also compare our test with the KSD test using exact score functions (i.e., no MCMC simulation), which serves as the performance upper-bound. The MCMC sampler we use is HMC; more precisely, we use the NumPyro (Phan et al., 2019) implementation of No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014); we take $t = 200$ burn-in samples and $m = 500$ consecutive draws for computing a score estimate $\bar{\mathbf{s}}_p$.

We use two kernel functions: (a) the exponentiated quadratic (EQ) kernel $k(x, x') = \exp\{-\|x - x'\|_2^2/(2\lambda^2)\}$, and (b) the IMQ kernel (9) with $\beta = 0.5$, $c = 1$, and $\Lambda = \lambda^2 I$. All three tests use the same kernel function, which allows us to investigate the effect of using the Stein-modified kernel. The length scale parameter λ is set to the median of the pairwise (Euclidean) distances of holdout samples from R so that the parameter (and thus the hypothesis) is fixed across trials. We include the EQ kernel in our comparison, as the population MMD is possible to compute, allowing us to verify the hypothesis in advance.

We simulate null and alternative cases by perturbing the weight parameter A; we add a positive value $\delta > 0$ to the (1, 1)-entry of A. Let us denote a perturbed weight by A_δ. Note that the data PPCA model has a Gaussian marginal $\mathcal{N}(\mathbf{0}, AA^\top + \psi^2 I_x)$. Therefore, this perturbation gives a model $\mathcal{N}(\mathbf{0}, A_\delta A_\delta^\top + \psi^2 I_x)$, where the first row and column of $A_\delta A_\delta^\top$ deviate from those of $AA^\top$. The perturbation is additive and increasing in δ, as each element of A is positive. We create a problem by specifying perturbation parameters (δ_P, δ_Q) for (P, Q). For the EQ-kernel MMD, we numerically confirmed that the perturbation gives a worse model for a larger perturbation. While the population KSD is not analytically tractable, this perturbation affects the score function through the covariance matrix, and the same behavior is expected for KSD; see Online Supplementary Material, Section B.6 in the supplement for details.

Problem 1 (null). We create a null scenario by choosing $(\delta_P, \delta_Q) = (1, 1 + 10^{-5})$ (P has a smaller covariance perturbation and is closer to R than Q). For different null settings, we refer the reader to Online Supplementary Material, Section C.4 in the supplement. We run the tests with significance levels $\alpha = 0.01, 0.05$. Table 1 reports the finite-sample size of the three tests for significance level $\alpha = 0.05$. The result for $\alpha = 0.01$ is omitted as none of the tests rejected the hypotheses. The size of the proposed LKSD test is indeed controlled. The extremely small type-I errors of the KSD tests are caused by the sensitivity of KSD to this perturbation; the population KSD value is negative and far from zero, and the test statistics easily fall in the acceptance region. The other two tests also have their error rates lower than the significance level. Note that their test thresholds are determined by treating the population discrepancy differences as zero, resulting in conservative tests.

Problem 2 (alternative). We investigate the power of the proposed test. We set up an alternative scenario by fixing $\delta_P = 2$ for P and $\delta_Q = 1$ for Q. For comparison with different parameter settings, please see Online Supplementary Material, Appendix C.1. The significance level α is fixed at 0.05. All the other parameters are chosen as in Problem 1. Figure 1 shows the plot of the test power against the sample size in each problem. The KSD reaches a near 100 percent rejection rate relatively quickly, indicating that information from the score function is helpful for these problems.

Table 1. Type-I errors the MMD test of Bounliphone et al. (2016), the proposed LKSD test, and the KSD test in PPCA problem 1

Sample size n	Rejection rates					
	EQ-med			IMQ-med		
	MMD	LKSD	KSD	MMD	LKSD	KSD
100	0.000	0.013	0.000	0.000	0.010	0.000
200	0.000	0.000	0.000	0.000	0.000	0.000
300	0.003	0.007	0.000	0.003	0.003	0.000
400	0.003	0.007	0.000	0.003	0.000	0.000
500	0.007	0.013	0.000	0.007	0.007	0.000

Note. Rejection rates are computed on 300 trials with significance level $\alpha = 0.05$. The columns EQ-med and IMQ-med denote EQ and IMQ kernels with the median bandwidth, respectively.

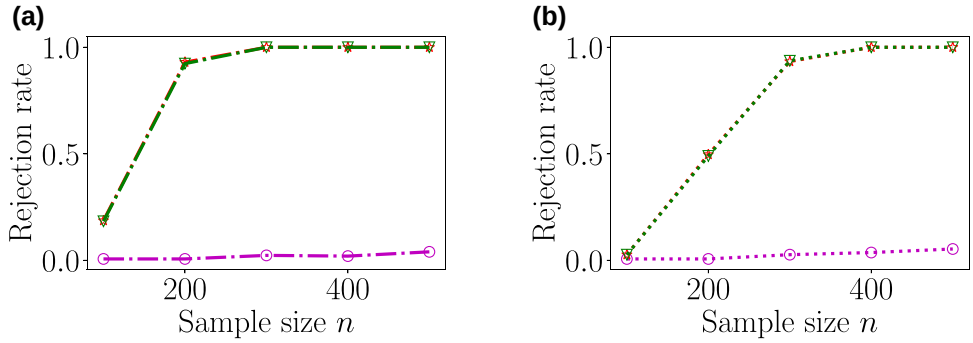

Figure 1. Power curves of the MMD test of Bounliphone et al. (2016), the proposed LKSD test, and the KSD test with the exact score function in PPCA Problem 2. The perturbation parameters are set as (δ_P, $\delta_Q = 2$, 1). each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: ▽ (the LKSD test); ☆ (the KSD test); ○ (the relative MMD test). (a) EQ kernel with median scaling and (b) IMQ kernel with median scaling.

The effect of the score approximation is negligible in this experiment, as the power curve of the LKSD test overlaps with that of KSD. The power of the MMD test is substantially lower than the other tests, indicating that the MMD is insensitive to this perturbation to the covariance.

4.1.2 Effect of kernel parameter choice

Dependency on scaling parameter. Using Problem 2 above, we examine how the test power is affected by the scaling parameter. We use the EQ and IMQ kernels as above, and choose their scaling parameter λ^2 from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. For each $n \in \{100, 300\}$, we run 300 trials and estimate the test power of the LKSD and MMD tests. Figure 2 plots the power curves of the tests. We can see that the high-power region of the EQ kernel is localized while the IMQ kernel's power curves are flat, indicating that the IMQ kernel does not depend on the parameter as much as the EQ. Therefore, for this problem, the IMQ kernel can be seen as more robust against misspecification of the scaling parameter. Nonetheless, with the right choice of the scaling parameter, the EQ kernel yields higher power for both MMD and KSD tests. It can be considered that the distinction arises because of the local nature of the difference between the two distributions; the EQ kernel is more sensitive in choosing features used to compute the KSD (see Online Supplementary Material, Section B.7).

Different parameterization. We also consider a different parameter choice for the preconditioning matrix. Here, we compare the median-scaled IMQ kernel with the same kernel having a covariance preconditioning matrix, as suggested in Section 3.4. Figure 3 shows the power curves of the three tests. Here, the relation between the MMD and KSD tests is overturned, and the KSD test struggles to detect the perturbation to the covariance. This result demonstrates that certain kernel choices can make the testing problem more challenging than others. Using multiple kernels, rather than relying on a single choice, could therefore robustify the evaluation, at the expense of a loss of power due to multiple testing correction.

4.1.3 Quality of Markov chain samplers

The asymptotic property of our test (Corollary 1) hinges on the quality of the Markov chain samplers. This section studies the effect of these Markov chains on the inference. We vary the burn-in size t and the score approximation sample size m, which is expected to affect the type-I error rate and the power of the test. In the experiments below, we set $\alpha = 0.05$. We choose t from $\{50, 100, \ldots, 600\}$ and m from $\{1, 10, 100, 1{,}000\}$.

In our first experiment, as in the previous sections, we use the NUTS with the same initialization strategy for both models. With $n = 300$, we run the test using Problems 1 and 2 above. Figure 4 shows rejection rates of the test for different settings of t and m. In both cases, the burn-in length t does not affect the test's performance, indicating the fast convergence of the sampler. The importance of a larger value of m can be seen when the alternative hypothesis holds, since the test power improves as m increases. The improved performance is likely due to reduced variance.

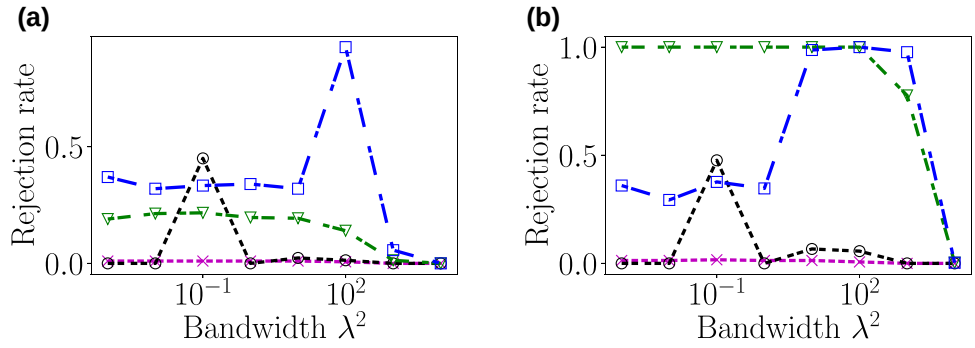

Figure 2. Power curves of the proposed LKSD test and the MMD test in PPCA Problem 2. The perturbation parameters are set as (δ_P, $\delta_Q = 2$, 1). each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: ▽ (LKSD test with IMQ kernel); □ (LKSD test with EQ kernel); ○ (MMD test with IMQ kernel); × (MMD test with EQ kernel). (a) $n = 100$ and (b) $n = 300$.

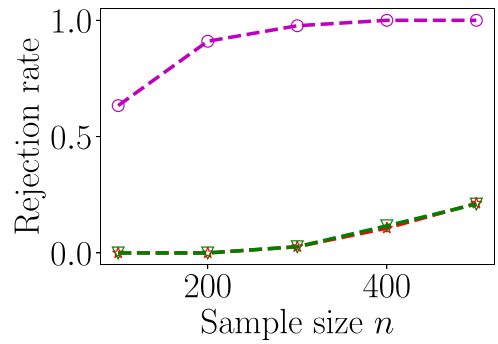

Figure 3. Power curves of the MMD test, the proposed LKSD test, and the KSD test in PPCA Problem 2. All the test use the covariance-preconditioned IMQ kernel. The perturbation parameters are set as (δ_P, $\delta_Q = 2$, 1). Each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: ▽ (the LKSD test); ☆ (the KSD test); ○ (the relative MMD test).

We next consider a slow-converging sampler for which the burn-in length t becomes crucial. We consider the null case (Problem 1) and replace the sampler for the first model P with MALA. We set the step size for the MALA sampler to make its convergence slow; we use the step size of $10^{-4} D_z^{-1/3}$. We initialize the two samplers differently to make sure that the resulting distributions differ when the samplers have not converged: the MALA sampler for P is initialized with samples from a Gaussian $\mathcal{N}\{(1, \ldots, 1), I_z\}$ and the NUTS sampler for Q a uniform distribution $U[-2, 2]^{D_z}$. Figure 5 demonstrates the relation between type-I error rates and choices of t and m. In contrast to the previous experiment, the burn-in has a clear effect on the type-I error: insufficient burn-in leads to uncontrolled error rates. The right panel ($n = 300$) shows that the test has substantially higher type-I error rates than in the left ($n = 100$). Comparison between these cases illustrates that a larger sample size n requires more intensive burn-in, as the test becomes more confident to reject. A large value of m improves the test as in the previous experiment. It can be understood that the contribution of burn-in samples is negligible in the score approximation. Although our analysis in Corollary 1 requires long burn-in, taking large m appears to be more important in practice, especially under a computational budget constraint. This experiment thus confirms the importance of the quality of the sampler.

4.2 Dirichlet process mixtures

Our next experiment applies our test to a Dirichlet process mixtures (DPM) model. Let $\psi(x \mid z)$ be a probability density function on $\mathbb{R}^D$. We consider a mixture density

$$\int \psi(x \mid z) \, \mathrm{d}\rho(z), \tag{10}$$

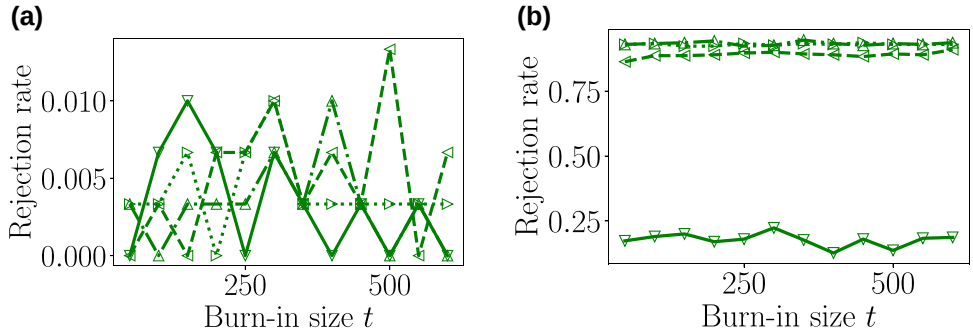

Figure 4. The effect of MCMC quality on the test's performance. Rejection rates against burn-in size t with varying Markov chain sample size m. PPCA Problems 1 and 2 with $\alpha = 0.05$. Both samplers use NUTS. Markers: $\triangledown$ ($m = 1$); $\triangleleft$ ($m = 10$); $\triangle$ ($m = 100$); $\triangleright$ ($m = 1{,}000$). (a) Problem 1 (null H_0 is true) and (b) Problem 2 (alternative H_1 is true).

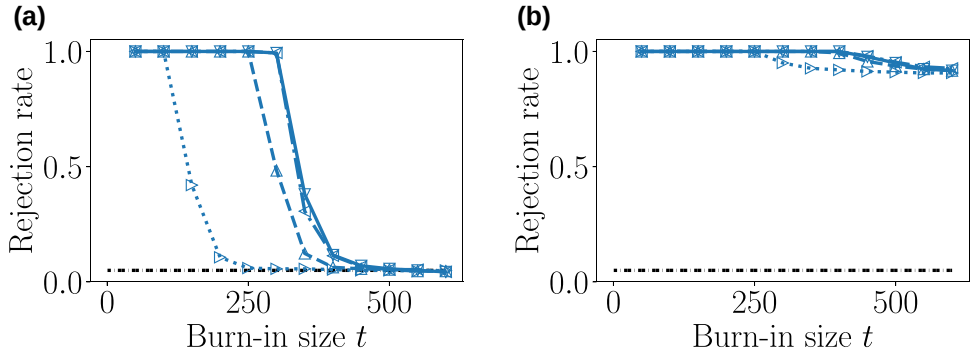

Figure 5. The effect of a poor MCMC sampler on the test. Type-I error rates against the burn-in size t with varying Markov chain sample size m. PPCA Problem 1 (the null H_0 is true). The dark dashed line indicates the significance level $\alpha = 0.05$. The samplers for P and Q are respectively MALA and NUTS. Markers: $\triangledown$ ($m = 1$); $\triangleleft$ ($m = 10$); $\triangle$ ($m = 100$); $\triangleright$ ($m = 1{,}000$). (a) $n = 100$ and (b) $n = 300$.

where ρ is a Borel probability measure on a Polish space $\mathcal{Z}$. A DPM model (Ferguson, 1983) places a Dirichlet process prior DP(a) on the mixing distribution ρ. Thus, a DPM model DPM(a) assumes the following generative process:

$$x_i | z_i, \phi, F \overset{\text{ind.}}{\sim} \psi(x \mid z_i), \quad z_i | F \overset{\text{i.i.d.}}{\sim} F, \quad F \sim \mathrm{DP}(a).$$

Here, a is a finite Borel measure on $\mathcal{Z}$. Note that the marginal density (on a single observation) is given by

$$\mathbb{E}_F\left[\int \psi(x \mid z)\, \mathrm{d}F(z)\right].$$

Although the prior has an infinite-dimensional component, the required conditional score function is simply $\mathbf{s}_\psi(x \mid z, \phi)$; thus we only need to sample from a finite-dimensional posterior $P_Z(\mathrm{d}z \mid x)$. If a model is conditioned on held-out data $\mathcal{D}$, then the predictive density $p(x|\mathcal{D})$ is $\mathbb{E}_{F|\mathcal{D}}[\int \psi(x \mid z)\, \mathrm{d}F(z)]$, which may be used to estimate the density (10). The score function is given by the expectation of $\mathbf{s}_\psi(x \mid z)$ with respect to the posterior

$$\frac{\psi(x \mid z, \phi)}{p(x \mid \mathcal{D})} \bar{F}_{\mathcal{D}}(\mathrm{d}z)$$

with $\bar{F}_{\mathcal{D}}$ being the mean measure of $P_F(\mathrm{d}F|\mathcal{D})$. Sampling from the posterior can be performed with a combination of the Metropolis–Hastings algorithm and Gibbs sampling (see, e.g., Ghosal & van

der Vaart, 2017, Chapter 5). For the score formula and the MCMC procedure, we refer the reader to Online Supplementary Material, Section B.8 in the supplement. By setting ψ to an isotropic normal density, for example, we can guarantee the integrability assumption in Lemma 1 (see Online Supplementary Material, Section B.3).

Our problem below considers comparing the predictive densities defined by two models with different Dirichlet process priors. Note that since candidate models P, Q here are point estimates derived from their respective posterior means of F, we discard some aspects of uncertainty in estimating the target (10); our setting only concerns evaluating the quality of those point estimates in approximating the data generating distribution R.

Experiment details. For the data distribution R, we use the mixture (10) defined by $\psi(x \mid z) = \mathcal{N}(x; z, 2I)$ and $\rho = \mathcal{N}(0, I)$; this choice yields $R = \mathcal{N}(\mathbf{0}, 3I)$. We consider the following simple Gaussian DPM model GDPM(μ):

$$x_i \overset{\text{ind.}}{\sim} \mathcal{N}(z_i, 2I), \quad z_i \overset{\text{i.i.d.}}{\sim} F, F \sim \mathrm{DP}(a), \quad a = \mathcal{N}(\mu, I),$$

where $\mu \in \mathbb{R}^D$. Note that without conditioning on observations, the model's marginal density is simply a Gaussian distribution $\mathcal{N}(\mu, 3I)$, which does not require approximation.

We therefore compare predictive distributions, i.e., we compare two GDPM models conditioned on *training data* $\mathcal{D}_{\mathrm{tr}} = \{\tilde{x}_i\}_{i=1}^{n_{\mathrm{tr}}} \overset{\text{i.i.d.}}{\sim} R$. We consider two GDPM models with *wrong* priors, where their prior means are shifted. Specifically, we take two models chosen as $Q = \mathrm{GDPM}(\bar{\mathbf{1}})$ and $P = \mathrm{GDPM}(\delta\bar{\mathbf{1}})$ with $\bar{\mathbf{1}} = \mathbf{1}/\sqrt{D}$. Unlike the preceding experiments, we condition the two models on the training data and obtain the predictive distributions, denoted by $P_{\mathcal{D}_{\mathrm{tr}}}$ and $Q_{\mathcal{D}_{\mathrm{tr}}}$, respectively; our problem is thus the comparison between $P_{\mathcal{D}_{\mathrm{tr}}}$ and $Q_{\mathcal{D}_{\mathrm{tr}}}$. The distributions now require simulating their posterior, and we use a random-scan Gibbs sampler and the Metropolis algorithm with a burn-in period $t = 1{,}000$ and the size of the latents $m = 500$. For sampling observables from the models, we use a random-scan Gibbs sampler with a burn-in period 2,000. We expect that if the training sample size n_{tr} is small, a larger perturbation would give a worse model as the effect of the prior is still present; we thus set $n_{\mathrm{tr}} = 5$. Due to the small sample size, the expected model relation might not hold, depending on the draw of $\mathcal{D}_{\mathrm{tr}}$. Therefore, we examine the rejection rates of the LKSD and MMD tests, averaged over 50 draws; for each draw of $\mathcal{D}_{\mathrm{tr}}$, we estimate the rejection rates based on 100 trials. Our problem is formed by varying the perturbation scale δ for $P_{\mathcal{D}_{\mathrm{tr}}}$, which is chosen from a regular grid $\{0.5, 0.6, \ldots, 0.9, 1.1, \ldots, 1.5\}$. This construction gives a null case when $\delta < 1$, the alternative otherwise. We set the dimension D to 10 and the significance level α to 0.05. As in Section 4.1, we use the IMQ kernel with median scaling.

Figure 6 reports the rejection rates of the two tests for each of $n \in \{50, 100, 200\}$. Note that the curves in the graph do not represent type-I errors nor power, as they are rejection rates *averaged* over draws $\mathcal{D}_{\mathrm{tr}}$, each of which forms a different problem. It can be seen that on average, both tests have correct sizes ($\delta < 1$). In the alternative regime ($\delta < 1$), the LKSD test underperforms the MMD with a small sample size ($n = 50$); however, its improvement in power is faster and exceeds the MMD at $n = 200$. These results imply that the LKSD estimate has a large variance for a small sample size, whereas its estimand (the population difference) is also larger, and thus the mean of the test statistic diverges faster. Thus, it may be understood that the KSD is more sensitive to model differences in this setting.

4.3 LDA

Our final experiment studies the behavior of the LKSD test on discrete data using LDA models. LDA is a mixed-membership model (Airoldi et al., 2014) for grouped discrete data such as text corpora. We follow Blei et al. (2003) and use the terminology of text data for ease of exposition. Accordingly, the following terms are defined using our notation. A word is an element in a discrete set (a vocabulary) $\{0, \ldots, L-1\}$ of size L. A document x is a sequence of D words, i.e., $x \in \{0, \ldots, L-1\}^D$ is a D-dimensional discrete vector. A prominent feature of LDA is that it groups similar words assuming that they come from a shared latent *topic*, which serves as a mixture component. An LDA model assumes the following generative process on a corpus of documents $\{x_i\}_{i=1}^n$:

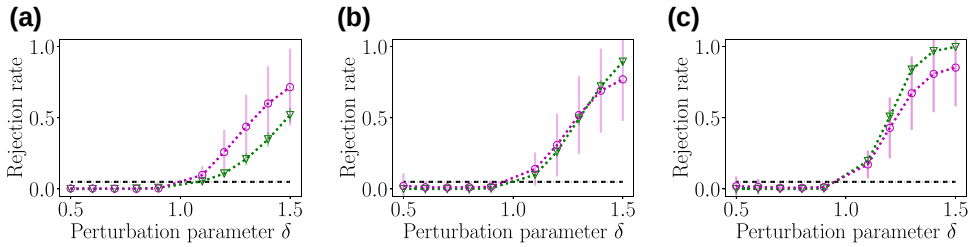

Figure 6. Comparison in Gaussian Dirichlet mixture models. Rejection rates plotted against the perturbation parameter δ. The sample size n is chosen from {50, 100, 200}. The rejection rates are averaged over draws of $\mathcal{D}_{\mathrm{tr}}$. The supposed null and alternative regimes are $\delta < 1$ and $\delta > 1$, respectively. Markers: ▽ (the LKSD test); ○ (the relative MMD test). The dark dashed line indicates the significance level $\alpha = 0.05$. The errorbars indicate the standard deviations of the estimated rejections rates. (a) $n = 50$. (b) $n = 100$ and (c) $n = 200$.

1. For each document $i \in \{1, \ldots, n\}$, generate a distribution over K topics $\theta_i \overset{\text{i.i.d.}}{\sim} \mathrm{Dir}(a)$ (the Dirichlet distribution), where θ_i is a probability vector of size $K \geq 1$.
2. For the jth word x_i^j, $j \in \{1, \ldots, D\}$ in a document i,
   (a) Choose a topic $z_i^j \overset{\text{i.i.d.}}{\sim} \mathrm{Cat}(\theta_i)$.
   (b) Draw a word from $x_i^j \overset{\text{i.i.d.}}{\sim} \mathrm{Cat}(b_k)$, where b_k is the distribution over words for topic k, and the topic assignment $z_i^j = k$.

Here, $a = (a_1, \ldots, a_K)$ is a vector of positive real numbers, and $b = (b_1, \ldots, b_K)^\top \in [0, 1]^{K \times L}$ represents a collection of K distributions over L words. In summary, an LDA model $P = \mathrm{LDA}(a, b)$ assumes the factorization

$$\prod_{i=1}^{n} p(x_i \mid z_i, \theta_i; a, b) p(z_i, \theta_i; a, b) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{D} p(x_i^j \mid z_i^j, b) p_z(z_i^j \mid \theta_i) \right\} p_\theta(\theta_i \mid a),$$

where z_i and θ_i act as latent variables.

Because of the independence structure over words, the conditional score function is simply given as

$$\mathbf{s}_p(x \mid z, \theta, a, b) = \mathbf{s}_p(x \mid z, b) = \left(\frac{p(\tilde{x}^j \mid z^j, b)}{p(x^j \mid z^j, b)} - 1 \right)_{j=1,\ldots,D}, \quad \text{where } \tilde{x}^j = x^j + 1 \bmod L.$$

Score approximation requires the posterior distribution $p(z \mid x; a, b)$ with respect to z. Marginalization of θ renders latent topics dependent on each other, and thus the posterior is intractable. A latent topic is conjugate to the corresponding topic distribution given all other topics. Therefore, an MCMC method such as collapsed Gibbs sampling allows us to sample from $p(z \mid x; a, b)$. As the observable and the latent are supported on finite sets, the use of Lemma 1 is justified; the finite moment assumptions in Corollary 1 are guaranteed; and the consistency of the population mean and variance of the test statistic follows from the convergence of $\mathbb{E}^{(t)}_{\mathbf{z} \mid x}[\bar{\mathbf{s}}_p(x \mid \mathbf{z})]$ and $\mathbb{E}^{(t)}_{\mathbf{w} \mid x}[\bar{\mathbf{s}}_q(x \mid \mathbf{w})]$ for each $x \in \mathcal{X}$.

4.3.1 Synthetic data—prior sparsity perturbation

In the below two problems, we observe a sample $\{x_i\}_{i=1}^n$ from an LDA model $R = \mathrm{LDA}(a, b)$. The number of topics is $K = 3$. The hyper-parameter a is chosen as $a = (a_0, a_0, a_0)$; for model R, we set $a_0 = 0.1$. Each of three rows in $b = (b_1, b_2, b_3)^\top \in [0, 1]^{3 \times L}$ is fixed at a value drawn from the symmetric Dirichlet distribution with all the concentration parameters one, and the vocabulary size is $L = 10{,}000$. Each $x_i \in \{0, \ldots, L-1\}^D$ is a document consisting of $D = 50$ words.

We design problems by perturbing the sparsity parameter a_0. Recall that Dir(a) is a distribution on the $(K-1)$-probability simplex. A small $a_0 < 1$ makes the prior $p_\theta(\theta_i \mid a) = \mathrm{Dir}(a)$ concentrate its mass on the vertices of the simplex; the case $a_0 = 1$ corresponds to the uniform distribution on the simplex; choosing $a_0 > 1$ leads to the prior mass concentrated on the center of the simplex. The data distribution R (with $a_0 = 0.1$) is thus intended to draw sparse topic proportions θ_i, and a document x_i is likely to have words from a particular topic. By increasing a_0, we can design a departure from this behavior. Therefore, as in the PPCA experiments, we additively perturb a_0 with parameters (δ_P, δ_Q) for respective candidate models (P, Q).

As LDA disregards word order, we need a kernel that respects this structure. We use the Bag-of-Words (BoW) IMQ kernel $k(x, x') = (1 + \|B(x) - B(x')\|_2^2)^{-1/2}$; it is simply the IMQ kernel computed in the BoW representation $B(x) \in \{0, 1, 2, \ldots, D\}^L$ whose ℓth entry (counting from 0 to $L-1$) is the count of the occurrences of word $\ell \in \{0, \ldots, L-1\}$ in a document x. By Online Supplementary Material, Lemma 8 in Section B.9.1, this choice ensures that arbitrary reordering of text sequences does not change the KSD value; i.e., the KSD does not assess models by their ability to generate sequences. We also tested differing input-scaling values and found that the bandwidth of the IMQ kernel did not have a significant effect on the test power (Online Supplementary Material, Section C.3.2).

For score estimation in the LKSD test, we use a random scan Gibbs sampler; we generate $m = 1{,}000$ latent samples after $t = 4{,}000$ burn-in iterations.

Problem 1 (null). We create a null situation by having $(\delta_P, \delta_Q) = (0.5, 0.6)$. In this case, Q's prior on θ is less sparse than that of P. Table 2 shows the size of the different tests for significance levels $\alpha = 0.05$; the result for $\alpha = 0.01$ is omitted as both tests did not reject the hypothesis. It can be seen that the rejection rates of both tests are bounded by the nominal level.

Problem 2 (alternative). We consider an alternative case in which the sparsity parameters are chosen as $(\delta_P, \delta_Q) = (1.0, 0.5)$ (we consider other parameter choices in Online Supplementary Material, Appendix C.3). Here, the model Q is expected to have less mixed topic proportions. Table 2 demonstrates the power of the MMD and LKSD tests. The power of the LKSD test improves as the sample size n increases, whereas the MMD has almost no power in this case. In this problem, the topics b are not sparse enough for each topic to have a sufficiently distinctive vocabulary. Thus, the problem is challenging for the MMD, as it is unable to find distinguishing words, in addition to the high-dimensionality. By contrast, the KSD is able to distinguish the models by taking advantage of their underlying structure.

4.3.2 *Synthetic data—topic perturbation*

We provide a negative example to illustrate a failure mode of the LKSD test for discrete data. The data are generated as in the previous section, whereas we construct two models differently. We set up a model by perturbing the topics of the data model R. That is, a model is given by LDA(a, b_δ) with $b_\delta = (1-\delta)b + \delta b_{\mathrm{ptb}}$ with $0 < \delta < 1$. We choose b_{ptb} as we did for b; the value is drawn independently of b. We set the perturbation parameter for Q as $\delta = 0.01$ and vary it for P, where the value is chosen from $\{0.06, 0.11, \ldots, 0.51\}$. Thus, P is morphed from b to b_{ptb} and therefore expected to underperform Q as perturbation δ increases. We run trials with $n = 300$. For score estimation, we take $m = 10{,}000$ and $t = 4{,}000$.

Figure 7 shows the plot of rejection rates against perturbation parameters. We see that the power of the LKSD test degrades as the perturbation increases. As P's topic becomes close to b_{ptb}, some words in the target's topic b become rare and therefore fall in the low probability region of P. This situation leads to increasing variance of the test statistic as δ increases, because the score function contains the reciprocal $1/p(x)$. The LKSD test can therefore fail when the support of the model is severely mismatched to that of the data, since the high variance of the statistic makes it difficult to detect significant departures from the null. Note that this observation does not apply to the continuous counterpart as the score can be written as the gradient of the logarithm of the density, which is typically numerically stable.

4.3.3 *Comparing topic models for arxiv articles*

Our final experiment investigates the test's performance using the arXiv dataset (Cornell University, 2020). The dataset consists of meta information of scholarly articles on the e-print

Table 2. Rejection rates of the MMD test and the LKSD test in LDA experiments

(a) Type-I errors of the KSD and MMD tests in LDA Problem 1; $(\delta_P, \delta_Q) = (0.5, 0.6)$. The significance level $\alpha = 0.05$.

Sample size n	Rejection rates	
	MMD	LKSD
100	0.003	0.013
200	0.010	0.007
300	0.007	0.003
400	0.003	0.007
500	0.007	0.010

(b) Power of the KSD and MMD tests in LDA Problem 2; $(\delta_P, \delta_Q) = (1.0, 0.5)$. The significance level α is chosen from $\{0.01, 0.05\}$.

Sample size n	Rejection rates			
	Level $\alpha = 0.01$		Level $\alpha = 0.05$	
	MMD	LKSD	MMD	LKSD
100	0.000	0.010	0.007	0.070
200	0.003	0.030	0.010	0.183
300	0.000	0.097	0.003	0.283
400	0.000	0.197	0.010	0.463
500	0.000	0.280	0.007	0.570

Note. Each result is based on 300 trials.

service arXiv. We treat the abstract of an article as a document and use paper categories to set up a problem. Specifically, we construct a problem by choosing three paper categories for model P, Q and the data distribution R. Unlike the preceding experiments, for a model category, we fit an LDA model to the dataset of abstracts in the category. As the KSD requires the number of words to be fixed, then for a given data category, we extract abstracts of length no less than $D = 100$ and subsample excess words. This process yields a dataset of articles of equal length D; for each trial, we obtain the data $\{x_i\}_{i=1}^n$ by subsampling from the larger set of articles. Thus, our problem is to compare two LDA models trained on different article sets and assess their fit to the dataset.

In the following experiments, we examine the power of LKSD and MMD tests. We vary the sample size n from 100 to 500. We fix the dataset category to stat.TH (statistics theory) and inspect two combinations of model categories. To train an LDA model LDA(a, b), we use the Gensim implementation (Rehurek & Sojka, 2011) of the variational algorithm of M. Hoffman et al. (2010). For sparsity parameters a, we use the parameter returned by this algorithm; we point-estimate topics b using the mean of the topics under the variational distribution. The number of topics is set to 100. The vocabulary set is comprised of words that appear in the abstracts of three chosen categories. As in the previous experiments, we use the IMQ-BoW kernel for both tests. We fix the significance level α at 0.05.

As we have seen the numerical instability issue in the previous section, we also consider an alternative KSD that is stable but computationally more expensive, as mentioned in Section 2.1 and the supplement (Online Supplementary Material, Section B.2). For this, we take a burn-in size $t = 500$ and a Markov chain size $m = 1{,}000$. We denote this method by LKSD-stable.

Probability theory vs statistical methodology. We choose math.PR (mathematics probability theory) for P and stat.ME (statistics methodology) for Q. In addition to the taxonomic proximity

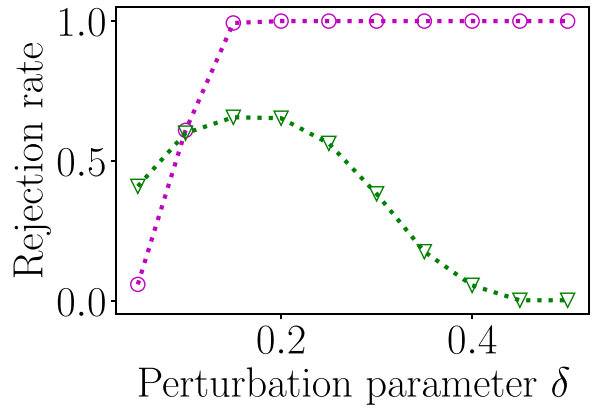

Figure 7. Power estimates plotted against perturbation parameters δ. The significance level $\alpha = 0.05$; the sample size $n = 300$. Markers: ▽ (the LKSD test); ○ (the MMD test).

to stat.TH, the category stat.ME has a larger proportion of articles shared with the target category: 3, 121 of 18, 973 (stat.ME) vs. 2, 884 of 46, 769 (math.PR). Thus, we expect Q to outperform P. This combination results in a vocabulary set of size $L = 126, 190$. For score estimation, we set the burn-in length t to 500 and the Markov chain sample size m to 5,000. Additionally, we run the LKSD test with $m = 15{,}000$ (labelled LKSD-extra) and the MMD test with the model sample size $n_{\text{model}} = 10{,}000$ (labelled MMD-extra). The sample size n_{model} is thresholded at 10,000 as the computational cost exceeds that of the LKSD test (in fact, sampling in this case makes the MMD by an order of magnitude slower due to the large vocabulary size).

Table 3 summarizes the result. The MMD test underperforms all the KSD-based tests; extra sampling did not lead to a significant improvement. We can see that increasing the Markov chain size m boosts the LKSD test, as it reduces the variance of the score estimator. The low power of the MMD test indicates that the model difference is too subtle to discern from the word compositions of generated documents; the LKSD tests offer a different viewpoint based on the model information.

Machine learning vs statistical methodology. Our second experiment uses cs.LG (computer science machine learning) for P, while Q uses the same category as the previous experiment. With this combination, the vocabulary size L is 208, 671. By the same reasoning as above, the second model Q is expected to be better than P. We run the same tests as above and compare their performances.

Table 4 summarizes the result. This experiment serves as a negative case study for the LKSD test: the MMD tests achieved power 1 for all sample-size choices (MMD-extra is omitted here), whereas the power of the LKSD test does not exceed even the significance level α for most sample size settings (LKSD-extra is omitted as increasing the Markov chain size did not improve the power). We attribute this failure to the unmatched support of the model P in the test distribution. This reasoning is supported by the high power of the MMD, as the BoW feature easily detects deviation of document patterns in this case. Thus, as we noted in the synthetic experiment in Section 4.3.2, the LKSD test fails when there is a severe mismatch in data and model support. The stable LKSD test approaches the same level as the MMD at $n = 500$, but still underperforms. While stable, the KSD used for this test can also suffer from the mismatch of the support, since it depends on the same density ratio as in the unstable counterpart.

5 Conclusion

We have developed a test of relative goodness of fit for latent variable models based on the kernel Stein discrepancy. The proposed test applies to a wide range of models, since the requirements of the test are mild: (a) models have MCMC samplers for inferring their latent variables, and (b) likelihoods have evaluable score functions. The proposed test complements existing model

Table 3. Rejection rates of the MMD test and the LKSD test in the math.PR vs. stat.ME experiment

Sample size n	Rejection rates				
	MMD	MMD-extra	LKSD	LKSD-extra	LKSD-stable
100	0.150	0.157	0.333	0.673	0.437
200	0.160	0.167	0.807	0.880	0.845
300	0.197	0.207	0.913	0.980	0.950
400	0.180	0.187	0.950	0.986	0.970
500	0.267	0.263	0.966	0.993	0.983

Note. Each result is based on 300 trials.

Table 4. Rejection rates of the MMD test and the LKSD test in the cs.LG vs. stat.ME experiment

Sample size n	Rejection rates		
	MMD	LKSD	LKSD-stable
100	1	0.000	0.287
200	1	0.007	0.643
300	1	0.013	0.833
400	1	0.013	0.873
500	1	0.113	0.923

Note. Each result is based on 300 trials.

evaluation techniques by providing a different means of model comparison, which takes advantage of the known model structure. Our experimental results confirm this view—the relative MMD test was unable to detect subtle differences between models in several of our benchmark experiments.

Our asymptotic analysis of the test statistic indicates that the test could suffer from bias if the mixing of the deployed MCMC sampler is slow. Removing the assumptions on the bias and the moments in Theorem 1 is certainly desirable; we envision that the recent development of unbiased MCMC (Jacob et al., 2020) could be used to construct an alternative unbiased KSD estimator, and leave this possibility as future work. While we have focused on comparing two models, extensions to ranking multiple models are possible as in Lim et al. (2019). Finally, the technique used in this article can be applied to other Stein discrepancies requiring the score function (Barp et al., 2019; Xu & Matsuda, 2020); one interesting application would be the KSD for directional data (Xu & Matsuda, 2020), where densities with computable normalizing constants are scarce.

Acknowledgments

We would like to thank the three anonymous referees for their constructive feedback that greatly improved the manuscript; Chris Oates for pointing us to the Fisher's identity in the early stage of this project; Jiaxin Shi for suggesting the Barker–Stein operator; Bharath Sriperumbudur, Motonobu Kanagawa, and Tamara Fernandez for helpful comments and discussions. H. Kanagawa and A. Gretton acknowledge support of the Gatsby Charitable Foundation. K. Fukumizu was supported in part by JST CREST Grant Number JPMJCR2015 and JSPS KAKENHI Grant Number 22H05106.

Conflicts of interest: None declared.

Data availability

Code to reproduce all the results is available at https://github.com/noukoudashisoup/lkgof.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series B* online.

References

Airoldi E. M., Blei D., Erosheva E. A., & Fienberg S. E. (2014). Introduction to mixed membership models and methods. In E. M. Airoldi, D. Blei, E. A. Erosheva, & S. E. Fienberg (Eds.), *Handbook of mixed membership models and their applications* chap. 1, (pp. 3–10). Chapman and Hall/CRC.

Aronszajn N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*(3), 337. https://doi.org/10.1090/S0002-9947-1950-0051437-7

Barbour A. D. (1988). Stein's method and Poisson process convergence. *Journal of Applied Probability*, *25*(A), 175–184. https://doi.org/10.2307/3214155

Barp A., Briol F.-X., Duncan A., Girolami M., & Mackey L. (2019). Minimum Stein discrepancy estimators. In *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Basilevsky A. (1994). *Statistical factor analysis and related methods*. John Wiley & Sons, Inc.

Besag J. (1994). Comments on 'Representations of knowledge in complex systems' by U. Grenander and M.I. Miller. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *56*(4), 591–592.

Betsch S., & Ebner B. (2019a). Fixed point characterizations of continuous univariate probability distributions and their applications. *Annals of the Institute of Statistical Mathematics*, *73*(1), 31–59. https://doi.org/10.1007/s10463-019-00735-1

Betsch S., & Ebner B. (2019b). A new characterization of the gamma distribution and associated goodness-of-fit tests. *Metrika*, *82*(7), 779–806. https://doi.org/10.1007/s00184-019-00708-7

Betsch S., & Ebner B. (2019c). Testing normality via a distributional fixed point property in the Stein characterization. *TEST*, *29*(1), 105–138. https://doi.org/10.1007/s11749-019-00630-0

Blei D., Ng A., & Jordan M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borgwardt K., Ghisu E., Llinares-López F., O'Bray L., & Rieck B. (2020). Graph kernels: State-of-the-art and future challenges. *Foundations and Trends® in Machine Learning*, *13*(5–6), 531–712. https://doi.org/10.1561/2200000076

Bounliphone W., Belilovsky E., Blaschko M. B., Antonoglou I., & Gretton A. (2016, May 2–4). A test of relative similarity for model selection in generative models. In *4th international conference on learning representations, ICLR 2016, conference track proceedings*.

Bou-Rabee N., Eberle A., & Zimmer R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, *30*(3), 1209–1250. https://doi.org/10.1214/19-AAP1528

Bresler G., & Nagaraj D. (2019). Stein's method for stationary distributions of Markov chains and application to Ising models. *Annals of Applied Probability*, *29*(5), 3230–3265. https://doi.org/10.1214/19-AAP1479

Chen L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability*, *3*(3), 534–545. https://doi.org/10.1214/aop/1176996359

Chen W. Y., Barp A., Briol F., Gorham J., Girolami M. A., Mackey L. W., & Oates C. J. (2019). Stein point Markov chain Monte Carlo. In *Proceedings of the 36th international conference on machine learning, ICML* (pp. 1011–1021).

Christmann A., & Steinwart I. (2010). Universal kernels on non-standard input spaces. In *Advances in neural information processing systems* (pp. 406–414). Curran Associates, Inc.

Chwialkowski K., Strathmann H., & Gretton A. (2016). A kernel test of goodness of fit. In *Proceedings of the 33rd international conference on machine learning* (pp. 2606–2615).

Cornell University (2020). arXiv Dataset version 68. Kaggle. https://www.kaggle.com/datasets/Cornell-University/arxiv. Accessed on 14 March 2022.

Dawid A. P., & Musio M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, *10*(2), 479–499. https://doi.org/10.1214/15-BA942

Dempster A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Duane S., Kennedy A., Pendleton B. J., & Roweth D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222. https://doi.org/10.1016/0370-2693(87)91197-X

Efron B., & Stein C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, *9*(3), 586–596. https://doi.org/10.1214/aos/1176345462

Ferguson T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics* (pp. 287–302). Elsevier.

Fisher R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725. https://doi.org/10.1017/S0305004100009580

Friel N., Mira A., & Oates C. J. (2016). Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods. *Bayesian Analysis*, *11*(1), 215–245. https://doi.org/10.1214/15-BA948
Friel N., & Wyse J. (2012). Estimating the evidence—A review. *Statistica Neerlandica*, *66*(3), 288–308. https://doi.org/10.1111/j.1467-9574.2011.00515.x
Fukumizu K., Gretton A., Schölkopf B., & Sriperumbudur B. K. (2008). Characteristic kernels on groups and semigroups. In *Advances in neural information processing systems* (Vol. 21). Curran Associates, Inc.
Ghosal S., & van der Vaart A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.
Gilks W. R., Richardson S., & Spiegelhalter D. J. (1995). *Markov chain Monte Carlo in practice*. Chapman & Hall.
Gorham J., Duncan A. B., Vollmer S. J., & Mackey L. (2019). Measuring sample quality with diffusions. *Annals of Applied Probability*, *29*(5), 2884–2928. https://doi.org/10.1214/19-AAP1467
Gorham J., & Mackey L. (2015). Measuring sample quality with Stein's method. In *Advances in neural information processing systems* (pp. 226–234). Curran Associates, Inc.
Gorham J., & Mackey L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th international conference on machine learning* (pp. 1292–1301). PMLR.
Götze F. (1991). On the rate of convergence in the multivariate CLT. *The Annals of Probability*, *19*(2), 724–739. https://doi.org/10.1214/aop/1176990448
Green P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732. https://doi.org/10.1093/biomet/82.4.711
Gretton A., Borgwardt K., Rasch M., Schölkopf B., & Smola A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*, 723–773.
Henze N., Meintanis S. G., & Ebner B. (2012). Goodness-of-fit tests for the gamma distribution based on the empirical Laplace transform. *Communications in Statistics—Theory and Methods*, *41*(9), 1543–1556. https://doi.org/10.1080/03610926.2010.542851
Henze N., & Visagie J. (2019). Testing for normality in any dimension based on a partial differential equation involving the moment generating function. *Annals of the Institute of Statistical Mathematics*, *72*(5), 1109–1136. https://doi.org/10.1007/s10463-019-00720-8
Hodgkinson L., Salomone R., & Roosta F. (2020). The reproducing Stein kernel approach for post-hoc corrected sampling.
Hoeffding W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, *19*(3), 293–325. https://doi.org/10.1214/aoms/1177730196
Hoffman M., Bach F., & Blei D. (2010). Online learning for latent Dirichlet allocation. In *Advances in neural information processing systems* (Vol. 23). Curran Associates, Inc.
Hoffman M. D., & Gelman A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.
Huggins J., & Mackey L. (2018). Random feature Stein discrepancies. In *Advances in neural information processing systems* (Vol. 31, pp. 1899–1909). Curran Associates, Inc.
Jacob P. E., O'Leary J., & Atchadé Y. F. (2020). Unbiased Markov Chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(3), 543–600. https://doi.org/10.1111/rssb.12336
Jeffreys H. (1961). *Theory of probability* (3rd ed.). Clarendon Press.
Jitkrittum W., Kanagawa H., Sangkloy P., Hays J., Schölkopf B., & Gretton A. (2018). Informative features for model comparison. In *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc.
Jitkrittum W., Xu W., Szabo Z., Fukumizu K., & Gretton A. (2017). A linear-time kernel goodness-of-fit test. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.
Kantorovich L. V. (2006). On the translocation of masses. *Journal of Mathematical Sciences*, *133*(4), 1381–1382. https://doi.org/10.1007/s10958-006-0049-2
Kass R. E., & Raftery A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572
Lehmann E. L., & Romano J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer.
Lim J. N., Yamada M., Schölkopf B., & Jitkrittum W. (2019). Kernel Stein tests for multiple model comparison. In *Advances in neural information processing systems* (Vol. 32, pp. 2240–2250). Curran Associates, Inc.
Liu Q., Lee J., & Jordan M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd international conference on machine learning* (pp. 276–284). PMLR.
Lloyd J., & Ghahramani Z. (2015). Statistical model criticism using kernel two sample tests. In *Advances in neural information processing systems* (pp. 829–837). Curran Associates, Inc.
Maesono Y. (1998). Asymptotic mean square errors of variance estimators for U-statistics and their Edgeworth expansions. *Journal of the Japan Statistical Society*, *28*(1), 1–19. https://doi.org/10.14490/jjss1995.28.1

Matsubara T., Knoblauch J., Briol F.-X., & Oates C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 997–1022. https://doi.org/10.1111/rssb.12500

Müller A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, *29*(2), 429–443. https://doi.org/10.2307/1428011

Neal R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*(2), 125–139. https://doi.org/10.1023/A:1008923215028

Neal R. M. (2011). *Handbook of Markov Chain Monte Carlo, Chapter 5* (1st ed.). Chapman and Hall/CRC.

Oates C. J., Girolami M., & Chopin N. (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(3), 695–718. https://doi.org/10.1111/rssb.12185

Phan D., Pradhan N., & Jankowiak M. (2019). 'Composable effects for flexible and accelerated probabilistic programming in NumPyro', arXiv, arXiv:1912.11554, preprint: not peer reviewed.

Rabiner L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–285. https://doi.org/10.1109/5.18626

Ranganath R., Tran D., Altosaar J., & Blei D. (2016). Operator variational inference. In *Advances in neural information processing systems* (Vol. 29, pp. 496–504). Curran Associates, Inc.

Rehurek R., & Sojka P. (2011). Gensim–Python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Reinert G., & Ross N. (2019). Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *Annals of Applied Probability*, *29*(5), 3201–3229. https://doi.org/10.1214/19-AAP1478

Riabiz M., Chen W. Y., Cockayne J., Swietach P., Niederer S. A., Mackey L., & Oates C. J. (2022). Optimal thinning of MCMC output. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(4), 1059–1081. https://doi.org/10.1111/rssb.12503

Roberts G. O., & Tweedie R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, *2*(4), 341. https://doi.org/10.2307/3318418

Ross N. (2011). Fundamentals of Stein's method. *Probability Surveys*, *8*, 210–293. https://doi.org/10.1214/11-PS182

Roweis S. T. (1997). EM algorithms for PCA and SPCA. In *Advances in neural information processing systems* (pp. 626–632). The MIT Press.

Schennach S. M., & Wilhelm D. (2017). A simple parametric model selection test. *Journal of the American Statistical Association*, *112*(520), 1663–1674. https://doi.org/10.1080/01621459.2016.1224716

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shao S., Jacob P. E., Ding J., & Tarokh V. (2019). Bayesian model comparison with the Hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, *114*(528), 1826–1837. https://doi.org/10.1080/01621459.2018.1518237

Shi J., Zhou Y., Hwang J., Titsias M. K., & Mackey L. (2022). Gradient estimation with discrete Stein operators. In *Advances in neural information processing systems* (Vol. 35, pp. 25829–25841). Curran Associates, Inc.

South L. F., Riabiz M., Teymur O., & Oates C. J. (2021). Postprocessing of MCMC. *Annual Review of Statistics and Its Application*, *9*, 529-555. https://doi.org/10.1146/annurev-statistics-040220-091727

Sriperumbudur B. K., Fukumizu K., & Lanckriet G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, *12*, 2389–2410.

Stein C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability (University of California, Berkeley, CA, 1970/1971), Vol. II: Probability theory* (pp. 583–602).

Stein C. (1986). *Approximate computation of expectations*, vol. 7 of Institute of mathematical statistics lecture notes—monograph series, institute of mathematical statistics.

Steinwart I., & Christmann A. (2008). *Support vector machines* (1st ed.). Springer Publishing Company, Incorporated.

Tipping M., & Bishop C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)*, *61*(3), 611–622. https://doi.org/10.1111/1467-9868.00196

van der Vaart A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Vértes E., & Sahani M. (2016). Learning doubly intractable latent variable models via score matching. In *Symposium on advances in approximate Bayesian inference (AABI)*.

Villani C. (2009). *Optimal transport: Old and new*. Springer Berlin Heidelberg.

Watanabe S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897.

Xu W., & Matsuda T. (2020). A Stein goodness-of-fit test for directional distributions. In *Proceedings of the twenty third international conference on artificial intelligence and statistics* (Vol. 108, pp. 320–330).

Yang J., Liu Q., Rao V., & Neville J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *Proceedings of the 35th international conference on machine learning* (pp. 5561–5570).