Journal of Machine Learning Research 24 (2023) 1-72 Submitted 9/21; Revised 9/22; Published 1/23

Bayesian Data Selection

Eli N. Weinstein[*] EW2760@COLUMBIA.EDU
Data Science Institute
Columbia University
New York, NY 10027, USA

Jeffrey W. Miller JWMILLER@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

Editor: Mingyuan Zhou

Abstract

Insights into complex, high-dimensional data can be obtained by discovering features of the data that match or do not match a model of interest. To formalize this task, we introduce the "data selection" problem: finding a lower-dimensional statistic—such as a subset of variables—that is well fit by a given parametric model of interest. A fully Bayesian approach to data selection would be to parametrically model the value of the statistic, nonparametrically model the remaining "background" components of the data, and perform standard Bayesian model selection for the choice of statistic. However, fitting a nonparametric model to high-dimensional data tends to be highly inefficient, statistically and computationally. We propose a novel score for performing data selection, the "Stein volume criterion (SVC)", that does not require fitting a nonparametric model. The SVC takes the form of a generalized marginal likelihood with a kernelized Stein discrepancy in place of the Kullback–Leibler divergence. We prove that the SVC is consistent for data selection, and establish consistency and asymptotic normality of the corresponding generalized posterior on parameters. We apply the SVC to the analysis of single-cell RNA sequencing data sets using probabilistic principal components analysis and a spin glass model of gene regulation.

Keywords: Bayesian nonparametrics, Bayesian theory, consistency, misspecification, Stein discrepancy

∗. Work conducted while at Harvard University.

1. Introduction

Scientists often seek to understand complex phenomena by developing working models for various special cases and subsets. Thus, when faced with a large complex data set, a natural question to ask is where and when a given working model applies. We formalize this question statistically by saying that given a high-dimensional data set, we want to identify a lower-dimensional statistic—such as a subset of variables—that follows a parametric model of interest (the working model). We refer to this problem as "data selection", in counterpoint to model selection, since it requires selecting the aspect of the data to which a given model applies.

For example, early studies of single-cell RNA expression showed that the expression of individual genes was often bistable, which suggests that the system of cellular gene expression might be described with the theory of interacting bistable systems, or spin glasses, with each gene a separate spin and each cell a separate observation. While it seems implausible that such a model would hold in full generality, it is quite possible that there are subsets of genes for which the spin glass model is a reasonable approximation to reality. Finding such subsets of genes is a data selection problem. In general, a good data selection method would enable one to (a) discover interesting phenomena in complex data sets, (b) identify precisely where naive application of the working model to the full data set goes wrong, and (c) evaluate the robustness of inferences made with the working model.

Perhaps the most natural Bayesian approach to data selection is to employ a semi-parametric joint model, using the parametric model of interest for the low-dimensional statistic (the "foreground") and using a flexible nonparametric model to explain all other aspects of the data (the "background"). Then, to infer where the foreground model applies, one would perform standard Bayesian model selection across different choices of the foreground statistic. However, this is computationally challenging due to the need to integrate over the nonparametric model for each choice of foreground statistic, making this approach quite difficult in practice. A natural frequentist approach to data selection would be to perform a goodness-of-fit test for each choice of foreground statistic. However, this still requires specifying an alternative hypothesis, even if the alternative is nonparametric, and ensuring comparability between alternatives used for different choices of foreground statistics is nontrivial. Moreover, developing goodness-of-fit tests for composite hypotheses or hierarchical models is often difficult in practice.

In this article, we propose a new score—for both data selection and model selection—that is similar to the marginal likelihood of a semi-parametric model but does not require one to specify a background model, let alone integrate over it. The basic idea is to employ a generalized marginal likelihood where we replace the foreground model likelihood by an exponentiated divergence with nice properties, and replace the background model's marginal likelihood with a simple volume correction factor. For the choice of divergence, we use a kernelized Stein discrepancy (KSD) since it enables us to provide statistical guarantees and is easy to estimate compared to other divergences—for instance, the Kullback–Leibler divergence involves a problematic entropy term that cannot simply be dropped. The background model volume correction arises roughly as follows: if the background model is well-specified, then asymptotically, its divergence from the empirical distribution converges to zero and all that remains of the background model's contribution is the volume of its effective parameter

space. Consequently, it is not necessary to specify the background model, only its effective dimension. To facilitate computation further, we develop a Laplace approximation for the foreground model's contribution to our proposed score.

This article makes a number of novel contributions. We introduce the data selection problem in broad generality, and provide a thorough asymptotic analysis. We propose a novel model/data selection score, which we refer to as the *Stein volume criterion*, that takes the form of a generalized marginal likelihood using a KSD. We provide new theoretical results for this generalized marginal likelihood and its associated posterior, complementing and building upon recent work on the frequentist properties of minimum KSD estimators (Barp et al., 2019). Finally, we provide first-of-a-kind empirical data selection analyses with two models that are frequently used in single-cell RNA sequencing analysis.

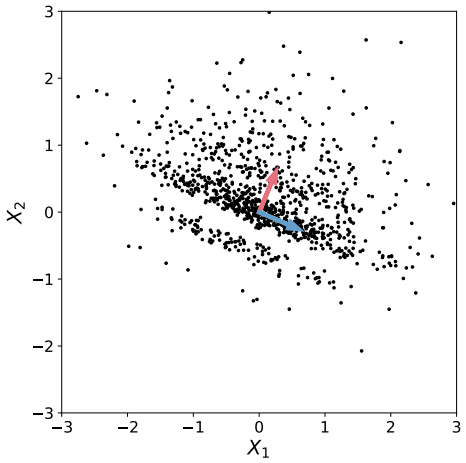
The article is organized as follows. In Section 2, we introduce the data selection problem and our proposed method. In Section 3 we study the asymptotic properties of Bayesian data selection methods and compare to model selection. Section 4 provides a review of related work and Section 5 illustrates the method on a toy example. In Section 6, we prove (a) consistency results for both data selection and model selection, (b) a Laplace approximation for the proposed score, and (c) a Bernstein–von Mises theorem for the corresponding generalized posterior. In Section 7, we apply our method to probabilistic principal components analysis (pPCA), assess its performance in simulations, and demonstrate it on single-cell RNA sequencing (scRNAseq) data. In Section 8, we apply our method to a spin glass model of gene expression, also demonstrated on an scRNAseq data set. Section 9 concludes with a brief discussion.

2. Method

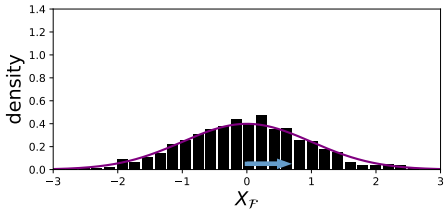
Suppose the data $X^{(1)}, \ldots, X^{(N)} \in \mathcal{X}$ are independent and identically distributed (i.i.d.), where $\mathcal{X} \subseteq \mathbb{R}^d$. Suppose the true data-generating distribution P_0 has density $p_0(x)$ with respect to Lebesgue measure, and let $\{q(x|\theta) : \theta \in \Theta\}$ be a parametric model of interest, where $\Theta \subseteq \mathbb{R}^m$. We are interested in evaluating this model when applied to a projection of the data onto a subspace, $\mathcal{X}_\mathcal{F} \subseteq \mathcal{X}$ (the "foreground" space). Specifically, let $X_\mathcal{F} := V^\top X$ be a linear projection of a datapoint $X \in \mathcal{X}$ onto $\mathcal{X}_\mathcal{F}$, where V is a matrix with orthonormal columns which defines the foreground space. Let $q(x_\mathcal{F}|\theta)$ denote the distribution of $X_\mathcal{F}$ when $X \sim q(x|\theta)$, and likewise, let $p_0(x_\mathcal{F})$ be the distribution of $X_\mathcal{F}$ when $X \sim p_0(x)$. Even when the complete model $q(x|\theta)$ is misspecified with respect to $p_0(x)$, it may be that $q(x_\mathcal{F}|\theta)$ is well-specified with respect to $p_0(x_\mathcal{F})$; see Figure 1 for a toy example. In such cases, the parametric model is only partially misspecified—specifically, it is misspecified on the "background" space $\mathcal{X}_\mathcal{B}$, defined as the orthogonal complement of $\mathcal{X}_\mathcal{F}$ (that is, the set of all vectors that are orthogonal to every vector in $\mathcal{X}_\mathcal{F}$).

Our goal is to find subspaces $\mathcal{X}_\mathcal{F}$ of the data space $\mathcal{X}$ for which the model $q(x_\mathcal{F}|\theta)$ is correctly specified. We are not seeking a subset of datapoints, but rather a projection of all the datapoints.

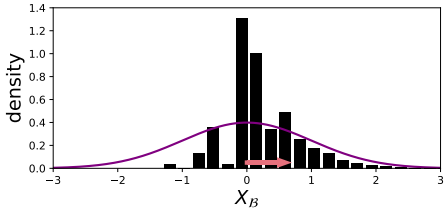
A natural Bayesian solution would be to replace the background component of the assumed model, $q(x_\mathcal{B}|x_\mathcal{F}, \theta)$, with a more flexible component $\tilde{q}(x_\mathcal{B}|x_\mathcal{F}, \phi_\mathcal{B})$ that is guaranteed to be well-specified with respect to $p_0(x_\mathcal{B}|x_\mathcal{F})$, such as a nonparametric model. The resulting



(a) An example for which a bivariate normal model is partially misspecified. Basis vectors for $\mathcal{X}_\mathcal{F}$ (foreground) and $\mathcal{X}_\mathcal{B}$ (background) are blue and red, respectively.



(b) A univariate normal model is well-specified for the data projection onto $\mathcal{X}_\mathcal{F}$.



(c) A univariate normal model is misspecified for the data projection onto $\mathcal{X}_\mathcal{B}$.

Figure 1: A simple example illustrating the data selection problem.

joint model, which we refer to as the "augmented model", is then

$$\begin{aligned} \theta \sim \pi(\theta), \qquad & X_\mathcal{F}^{(i)} \mid \theta \sim q(x_\mathcal{F} \mid \theta), \\ \phi_\mathcal{B} \sim \pi_\mathcal{B}(\phi_\mathcal{B}), \qquad & X_\mathcal{B}^{(i)} \mid X_\mathcal{F}^{(i)}, \phi_\mathcal{B} \sim \tilde{q}(x_\mathcal{B} \mid X_\mathcal{F}^{(i)}, \phi_\mathcal{B}) \end{aligned} \tag{1}$$

independently for $i \in \{1, \ldots, N\}$. In other words, the pairs $(X_\mathcal{F}^{(1)}, X_\mathcal{B}^{(1)}), \ldots, (X_\mathcal{F}^{(N)}, X_\mathcal{B}^{(N)})$ are i.i.d. given θ and $\phi_\mathcal{B}$, with the foreground projections $X_\mathcal{F}^{(i)}$ drawn from the parametric model of interest, and the background projections $X_\mathcal{B}^{(i)}$ drawn from the flexible background model. The standard Bayesian approach to infer $\mathcal{X}_\mathcal{F}$ would be to put a prior on the choice of foreground space $\mathcal{X}_\mathcal{F}$, and compute the posterior over the choice of $\mathcal{X}_\mathcal{F}$. Computing this posterior boils down to computing the Bayes factor $\tilde{q}(X^{(1:N)}|\mathcal{F})/\tilde{q}(X^{(1:N)}|\mathcal{F}')$ for any given pair of foregrounds $\mathcal{F}$ and $\mathcal{F}'$, where $\tilde{q}(X^{(1:N)}|\mathcal{F})$ denotes the marginal likelihood of $\mathcal{F}$ under the augmented model, that is, $\tilde{q}(X^{(1:N)}|\mathcal{F}) = \int\int q(X_\mathcal{F}^{(1:N)}|\theta)\, \tilde{q}(X_\mathcal{B}^{(1:N)}|X_\mathcal{F}^{(1:N)}, \phi_\mathcal{B})\pi(\theta)\pi_\mathcal{B}(\phi_\mathcal{B})d\theta d\phi_\mathcal{B}$.

However, in general, it is difficult to find a background model that (a) is guaranteed to be well-specified with respect to $p_0(x_\mathcal{B}|x_\mathcal{F})$ and (b) can be integrated over in a computationally tractable way to obtain the posterior on the choice of $\mathcal{F}$. Our proposed method, which we introduce next, sidesteps these difficulties while still exhibiting similar guarantees.

2.1 Proposed score for data selection and model selection

In this section, we propose a model/data selection score that is simpler to compute than the marginal likelihood of the augmented model and has similar theoretical guarantees. This score takes the form of a generalized marginal likelihood with a normalized kernelized Stein discrepancy (NKSD) estimate taking the place of the log likelihood. Specifically, our

proposed model/data selection score, termed the "Stein volume criterion" (SVC), is

$$\mathcal{K} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} \int \exp\Big(- \frac{N}{T} \widehat{\text{NKSD}}(p_0(x_{\mathcal{F}}) \| q(x_{\mathcal{F}}|\theta)) \Big) \pi(\theta) d\theta \tag{2}$$

where the "temperature" $T > 0$ is a hyperparameter and $m_{\mathcal{B}}$ is the effective dimension of the background model parameter space. $\widehat{\text{NKSD}}(\cdot\|\cdot)$ is an empirical estimate of the NKSD (Equations 4 and 5), and measures the mismatch between the data and the model over the foreground subspace.

There are three key properties of $\widehat{\text{NKSD}}$ that distinguish it from other estimators of other divergences. First, it estimates the divergence directly, not just up to a data-dependent constant; this is essential for data selection consistency (Section 3.1). For instance, putting the log likelihood in place of $\frac{N}{T}\widehat{\text{NKSD}}$ in Equation 2 fails to provide data selection consistency since it implicitly involves comparing the foreground entropy under P_0. Second, $\widehat{\text{NKSD}}$ converges at a $O(1/N)$ convergence rate when the model is correct; this is essential for nested data selection consistency (Section 3.2). In contrast, even if the foreground entropy under P_0 is known exactly, using a Monte Carlo estimate of the Kullback–Leibler divergence in place of $\frac{N}{T}\widehat{\text{NKSD}}$ fails since the convergence rate is only $O(1/\sqrt{N})$. Third, the NKSD exhibits subsystem independence (Section 6.1), which ensures that the SVC is comparable between foreground spaces of different dimension. We are unaware of any other divergence estimator with all three of these key properties.

The integral in Equation 2 can be approximated using techniques discussed in Section 2.3. The hyperparameter T can be calibrated by comparing the coverage of the standard Bayesian posterior to the coverage of the NKSD generalized posterior (Section A.1). The $(2\pi/N)^{m_{\mathcal{B}}/2}$ factor penalizes higher-complexity background models. In general, we allow $m_{\mathcal{B}}$ to grow with N, particularly when the background model is nonparametric. Crucially, the likelihood of the background model does not appear in our proposed score, sidestepping the need to fit or even specify the background model—indeed, the only place that the background model enters into the SVC is through $m_{\mathcal{B}}$.

Thus, rather than specify a background model and then derive $m_{\mathcal{B}}$, one can simply specify an appropriate value of $m_{\mathcal{B}}$. Reasonable choices of $m_{\mathcal{B}}$ can be derived by considering the asymptotic behavior of a Pitman-Yor process mixture model, a common nonparametric model that is a natural choice for a background model. A Pitman-Yor process mixture model with discount parameter $\alpha \in (0, 1)$, concentration parameter $\nu > -\alpha$, and D-dimensional component parameters will asymptotically have expected effective dimension

$$m_{\mathcal{B}} \sim D \frac{\Gamma(\nu + 1)}{\alpha \Gamma(\nu + \alpha)} N^{\alpha} \tag{3}$$

under the prior, where $a_N \sim b_N$ means that $a_N/b_N \to 1$ as $N \to \infty$ and $\Gamma(\cdot)$ is the gamma function (Pitman, 2002, §3.3). As a default, we recommend setting $m_{\mathcal{B}} = c_{\mathcal{B}}\, r_{\mathcal{B}} \sqrt{N}$, where $r_{\mathcal{B}}$ is the dimension of $\mathcal{X}_{\mathcal{B}}$ and $c_{\mathcal{B}}$ is a constant chosen to match Equation 3 with $\alpha = 1/2$. The $\sqrt{N}$ scaling is particularly nice in terms of asymptotic guarantees; see Section 3.2.

The SVC uses a novel, normalized version of the KSD between densities $p(x)$ and $q(x)$:

$$\text{NKSD}(p(x)\|q(x)) := \frac{\mathbb{E}_{X,Y\sim p}\big[(s_q(X) - s_p(X))^{\top}(s_q(Y) - s_p(Y))k(X,Y)\big]}{\mathbb{E}_{X,Y\sim p}[k(X,Y)]} \tag{4}$$

where $k(x, y) \in \mathbb{R}$ is an integrally strictly positive definite kernel, $s_q(x) := \nabla_x \log q(x)$, and $s_p(x) := \nabla_x \log p(x)$; see Section 6.1 for details. The numerator corresponds to the standard KSD (Liu et al., 2016). The denominator, which is strictly positive and independent of $q(x)$, is a normalization factor that we have introduced to make the divergence comparable across spaces of different dimension. See Section A.2 for kernel recommendations. Extending the technique of Liu et al. (2016), we propose to estimate the normalized KSD using U-statistics:

$$\widehat{\mathrm{NKSD}}(p(x)\|q(x)) = \frac{\sum_{i\neq j} u(X^{(i)}, X^{(j)})}{\sum_{i\neq j} k(X^{(i)}, X^{(j)})} \tag{5}$$

where $X^{(i)} \sim p(x)$ i.i.d., the sums are over all $i, j \in \{1, \ldots, N\}$ such that $i \neq j$, and

$$u(x, y) := s_q(x)^\top s_q(y) k(x, y) + s_q(x)^\top \nabla_y k(x, y) + s_q(y)^\top \nabla_x k(x, y) + \mathrm{trace}(\nabla_x \nabla_y^\top k(x, y)).$$

Importantly, Equation 5 does not require knowledge of $s_p(x)$, which is unknown in practice.

2.2 Comparison with the standard marginal likelihood

It is instructive to compare our proposed model/data selection score, the Stein volume criterion, to the standard marginal likelihood $\tilde{q}(X^{(1:N)}|\mathcal{F})$. In particular, we show that the SVC approximates a generalized version of the marginal likelihood. To see this, first define $H := -\int p_0(x) \log p_0(x) dx$, the entropy of the complete data distribution, and note that if H were somehow known, then the Kullback-Leibler (KL) divergence between the augmented model and the data distribution could be approximated as

$$\widehat{\mathrm{KL}}(p_0(x)\|q(x_\mathcal{F}|\theta)\, \tilde{q}(x_\mathcal{B}|x_\mathcal{F}, \phi_\mathcal{B})) := -\frac{1}{N}\sum_{i=1}^{N} \log q(X_\mathcal{F}^{(i)}|\theta)\, \tilde{q}(X_\mathcal{B}^{(i)}|X_\mathcal{F}^{(i)}, \phi_\mathcal{B}) - H.$$

Since multiplying the marginal likelihoods by a fixed constant does not affect the Bayes factors, the following expression could be used instead of the marginal likelihood $\tilde{q}(X^{(1:N)}|\mathcal{F})$ to decide among foreground subspaces:

$$\frac{\tilde{q}(X^{(1:N)}|\mathcal{F})}{\exp(-NH)} = \int\int \exp\Big(- N\, \widehat{\mathrm{KL}}(p_0(x)\|q(x_\mathcal{F}|\theta)\, \tilde{q}(x_\mathcal{B}|x_\mathcal{F}, \phi_\mathcal{B}))\Big)\pi(\theta)\pi_\mathcal{B}(\phi_\mathcal{B}) d\theta d\phi_\mathcal{B}. \tag{6}$$

Now, consider a generalized marginal likelihood where the NKSD replaces the KL:

$$\tilde{\mathcal{K}} := \int\int \exp\Big(- N \frac{1}{T} \widehat{\mathrm{NKSD}}\big(p_0(x)\|q(x_\mathcal{F}|\theta)\, \tilde{q}(x_\mathcal{B}|x_\mathcal{F}, \phi_\mathcal{B})\big)\Big)\pi(\theta)\pi_\mathcal{B}(\phi_\mathcal{B}) d\theta d\phi_\mathcal{B}. \tag{7}$$

We refer to $\tilde{\mathcal{K}}$ as the "NKSD marginal likelihood" of the augmented model. Intuitively, we expect it to behave similarly to the standard marginal likelihood, except that it quantifies the divergence between the model and data distributions using the NKSD instead of the KL.

However, a key advantage of the NKSD marginal likelihood is that it admits a simple approximation via the SVC when the background model is well-specified, unlike the standard marginal likelihood. For instance, if the foreground and background are independent, that

is, $p_0(x) = p_0(x_\mathcal{F})p_0(x_\mathcal{B})$ and $\tilde{q}(x_\mathcal{B}|x_\mathcal{F}, \phi_\mathcal{B}) = \tilde{q}(x_\mathcal{B}|\phi_\mathcal{B})$, then the theory in Section 6 can be extended to the full augmented model to show that

$$\frac{\log \tilde{\mathcal{K}}}{\log \mathcal{K}} \xrightarrow[N\to\infty]{P_0} 1, \tag{8}$$

where $\mathcal{K}$ is the SVC (Equation 2). Thus, the SVC approximates the NKSD marginal likelihood of the augmented model, suggesting that the SVC may be a convenient alternative to the standard marginal likelihood. Formally, Section 3 shows that the SVC exhibits consistency properties similar to the standard marginal likelihood, even when $p_0(x) \neq p_0(x_\mathcal{F})p_0(x_\mathcal{B})$.

2.3 Computation

Next, we discuss methods for computing the SVC including exact solutions, Laplace/BIC approximation, variational approximation, and comparing many possible choices of $\mathcal{F}$. An attractive feature of the SVC is that, unlike the fully Bayesian augmented model, the computation time required does not grow with the background dimension $m_\mathcal{B}$.

2.3.1 Exact solution for exponential families

When the foreground model is an exponential family, the SVC can be computed analytically. Specifically, in Section A.3, we show if $q(x_\mathcal{F}|\theta) = \lambda(x_\mathcal{F}) \exp(\theta^\top t(x_\mathcal{F}) - \kappa(\theta))$, then

$$\widehat{\mathrm{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta)) = \theta^\top A\, \theta + B^\top \theta + C \tag{9}$$

where A, B, and C depend on the data $X^{(1:N)}$ but not on θ. Therefore, we can compute the SVC in closed form by choosing a multivariate Gaussian for the prior $\pi(\theta)$ in Equation 2; see Section A.3.

2.3.2 Laplace and BIC approximations

The Laplace approximation is a widely-used technique for computing marginal likelihoods. In Theorem 9, we establish regularity conditions under which a Laplace approximation to the SVC is justified by being asymptotically correct. The resulting approximation is

$$\mathcal{K} \approx \frac{\exp\left(-\frac{N}{T}\widehat{\mathrm{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta_N))\right)\pi(\theta_N)}{|\det \frac{1}{T}\nabla^2_\theta \widehat{\mathrm{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta_N))|^{1/2}} \left(\frac{2\pi}{N}\right)^{(m_\mathcal{F}+m_\mathcal{B})/2} \tag{10}$$

where $\theta_N := \mathrm{argmin}_\theta \widehat{\mathrm{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta))$ is the point at which the estimated NKSD is minimized, the "minimum Stein discrepancy estimator" as defined by Barp et al. (2019). Here, θ_N is simply used to help compute the approximation and does not depend on $\pi(\theta)$, which can be any prior that is continuous and positive at the limiting value of θ_N.

We can also make a rougher approximation, analogous to the Bayesian information criterion (BIC), which does not require one to compute second derivatives of $\widehat{\mathrm{NKSD}}$:

$$\mathcal{K} \approx \exp\Big(-\frac{N}{T}\widehat{\mathrm{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta_N))\Big)\left(\frac{2\pi}{N}\right)^{(m_\mathcal{F}+m_\mathcal{B})/2}. \tag{11}$$

This approximation is easy to compute, given a minimum Stein discrepancy estimator θ_N. Like the SVC, it satisfies all of our consistency desiderata (Section B). However, we expect it to perform worse than the SVC when there is not yet enough data for the NKSD posterior to be highly concentrated, that is, when a range of θ values can plausibly explain the data.

2.3.3 Comparing many foregrounds using approximate optima

Often, we would like to evaluate many possible subspaces $\mathcal{X}_\mathcal{F}$ when performing data selection. Even when using the Laplace or BIC approximation to the SVC, this can get computationally prohibitive since we need to re-optimize to find θ_N for every $\mathcal{F}$ under consideration. Here, we propose a way to reduce this cost by making a fast linear approximation. Define $\ell_j(\theta) := \widehat{\text{NKSD}}(p_0(x_{\mathcal{F}_j})\|q(x_{\mathcal{F}_j}|\theta))$ for $j \in \{1, 2\}$. For $w \in [0, 1]$, we can linearly interpolate

$$\theta_N(w) := \operatorname*{argmin}_\theta \ell_1(\theta) + w(\ell_2(\theta) - \ell_1(\theta)). \tag{12}$$

Now, $\theta_N(0)$ and $\theta_N(1)$ are the minimum Stein discrepancy estimators for $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Given $\theta_N(0)$, we can approximate $\theta_N(1)$ by applying the implicit function theorem and a first-order Taylor expansion (Section A.4):

$$\theta_N(1) \approx \theta_N(0) - \nabla^2_\theta \ell_1(\theta_N(0))^{-1} \nabla_\theta \ell_2(\theta_N(0)). \tag{13}$$

Note that the derivatives of ℓ_j are often easy to compute with automatic differentiation (Baydin et al., 2018). Note also that when we are comparing one foreground subspace, such as $\mathcal{X}_{\mathcal{F}_1} = \mathcal{X}$, to many other foreground subspaces $\mathcal{X}_{\mathcal{F}_2}$, the inverse Hessian $\nabla^2_\theta \ell_1(\theta_N(0))^{-1}$ only needs to be computed once. Thus, Equation 13 provides a fast approximate method for computing Laplace or BIC approximations to the SVC for a large number of candidate foregrounds $\mathcal{F}$. We apply this technique in Section 7, where we find that it performs well in simulation studies and in practice.

2.3.4 Variational approximation

Variational inference is a method for approximating both the posterior distribution and the marginal likelihood of a probabilistic model. Since the SVC takes the form of a generalized marginal likelihood, we can derive a variational approximation to the SVC. Let $r_\zeta(\theta)$ be an approximating distribution parameterized by ζ. By Jensen's inequality, we have

$$\begin{aligned}
&\log \int \exp\Big(-\frac{N}{T} \widehat{\text{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta))\Big)\pi(\theta)d\theta \\
&\quad = \log \int \frac{\exp\left(-\frac{N}{T}\widehat{\text{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta))\right)\pi(\theta)}{r_\zeta(\theta)} r_\zeta(\theta)d\theta \\
&\quad \geq \mathbb{E}_{r_\zeta}\left[\log\left(\frac{\exp\left(-\frac{N}{T}\widehat{\text{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta))\right)\pi(\theta)}{r_\zeta(\theta)}\right)\right] \\
&\quad = -\frac{N}{T}\mathbb{E}_{r_\zeta}\left[\widehat{\text{NKSD}}(p_0(x_\mathcal{F})\|q(x_\mathcal{F}|\theta))\right] + \mathbb{E}_{r_\zeta}[\log \pi(\theta)] - \mathbb{E}_{r_\zeta}[\log r_\zeta(\theta)].
\end{aligned} \tag{14}$$

Maximizing this lower bound with respect to the variational parameters ζ, and adding the background correction $(m_\mathcal{B}/2)\log(2\pi/N)$, provides an approximation to the log SVC. Note

that this variational approximation falls within the framework of generalized variational inference proposed by Knoblauch et al. (2022).

This variational approximation to the SVC is particularly useful when we are aiming to find the best subspace $\mathcal{X}_\mathcal{F}$ among a very large number of candidates, since we can jointly optimize the variational parameters ζ and the choice of foreground subspace $\mathcal{X}_\mathcal{F}$. Here, we do not necessarily need to evaluate the SVC for all foreground subspaces $\mathcal{X}_\mathcal{F}$ under consideration, and can instead rely on optimization methods to search for the best $\mathcal{X}_\mathcal{F}$ from among a large set of possibilities (see Section 8 for an example). Practically, we recommend using the local linear approximation in Section 2.3.3 when the goal is to compare SVC values among many not-too-different foreground subspaces $\mathcal{X}_\mathcal{F}$, and using the variational approximation when the goal is to find one best $\mathcal{X}_\mathcal{F}$ from among a large and diverse set.

3. Data selection and model selection consistency

This section presents our consistency results when comparing two different foreground subspaces (data selection) or two different foreground models (model selection). The theory supporting these results is in Sections 6 and B. We consider four distinct properties that a procedure would ideally exhibit: data selection consistency, nested data selection consistency, model selection consistency, and nested model selection consistency; see Section 6.4 for precise definitions. We consider six possible model/data selection scores, and we establish which scores satisfy which properties; see Table 1. The SVC and the full marginal likelihood are the only two of the six scores that satisfy all four consistency properties.

The intuition behind Bayesian model selection is often explained in terms of Occam's razor: a theory should be as simple as possible but no simpler. Data selection and nested data selection encapsulate a complementary intuition: a theory should explain as much of the data as possible but no more. In other words, when choosing between foreground spaces, a consistent data selection score will asymptotically prefer the highest-dimensional space on which the model is correctly specified.

As in standard model selection, a practical concern in data selection is robustness. For instance, if the foreground model is even slightly misspecified on $\mathcal{X}_{\mathcal{F}_2}$, then the empty foreground $\mathcal{X}_{\mathcal{F}_1} = \varnothing$ will be asymptotically preferred over $\mathcal{X}_{\mathcal{F}_2}$. Since the SVC takes the form of a generalized marginal likelihood, techniques for improving robustness with the standard marginal likelihood—such as coarsened posteriors, power posteriors, and BayesBag—could potentially be extended to address this issue (Miller and Dunson, 2019; Huggins and Miller, 2021). We leave exploration of such approaches to future work.

3.1 Data selection consistency

First, consider comparisons between different choices of foreground, $\mathcal{F}_1$ and $\mathcal{F}_2$. When the model is correctly specified over $\mathcal{F}_1$ but not $\mathcal{F}_2$, we refer to asymptotic concentration on $\mathcal{F}_1$ as "data selection consistency" (and vice versa if $\mathcal{F}_2$ is correct but not $\mathcal{F}_1$). For the standard marginal likelihood of the augmented model, we have (see Section B.2)

$$\frac{1}{N} \log \frac{\tilde{q}(X^{(1:N)}|\mathcal{F}_1)}{\tilde{q}(X^{(1:N)}|\mathcal{F}_2)} \xrightarrow[N\to\infty]{P_0} \mathrm{KL}(p_0(x_{\mathcal{F}_2}) \| q(x_{\mathcal{F}_2}|\theta^{\mathrm{KL}}_{2,*})) - \mathrm{KL}(p_0(x_{\mathcal{F}_1}) \| q(x_{\mathcal{F}_1}|\theta^{\mathrm{KL}}_{1,*})) \tag{15}$$

Score	Consistency property			
	d.s.	nested d.s.	m.s.	nested m.s.
$\tilde{q}(X^{(1:N)}\|\mathcal{F})$ full marginal likelihood	✓	✓	✓	✓
$\mathcal{K}^{(\mathrm{a})}$ foreground marg lik, background volume	✗	✗	✓	✓
$\mathcal{K}^{(\mathrm{b})}$ foreground marg NKSD	✓	✗	✓	✓
$\mathcal{K}^{(\mathrm{c})}$ foreground marg KL, background volume	✓	✗	✓	✓
$\mathcal{K}^{(\mathrm{d})}$ foreground NKSD, background volume	✓	✓	✓	✗
$\mathcal{K}$ foreground marg NKSD, background volume	✓	✓	✓	✓

Table 1: Consistency properties satisfied by various model/data selection scores. Only the Stein volume criterion $\mathcal{K}$ and the full marginal likelihood $\tilde{q}(X^{(1:N)}|\mathcal{F})$ satisfy all four desiderata. (d.s. = data selection, m.s. = model selection, marg = marginal, lik = likelihood.)

where $\theta^{\mathrm{KL}}_{j,*} := \operatorname{argmin} \mathrm{KL}(p_0(x_{\mathcal{F}_j})\|q(x_{\mathcal{F}_j}|\theta))$ for $j \in \{1,2\}$, that is, $\theta^{\mathrm{KL}}_{j,*}$ is the parameter value that minimizes the KL divergence between the projected data distribution $p_0(x_{\mathcal{F}_j})$ and the projected model $q(x_{\mathcal{F}_j}|\theta)$. Thus, $\tilde{q}(X^{(1:N)}|\mathcal{F}_j)$ asymptotically concentrates on the $\mathcal{F}_j$ on which the projected model can most closely match the data distribution in terms of KL.

In Theorem 17, we show that under mild regularity conditions, the Stein volume criterion behaves precisely the same way but with the NKSD in place of the KL:

$$\frac{1}{N}\log\frac{\mathcal{K}_1}{\mathcal{K}_2} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta^{\mathrm{NKSD}}_{2,*})) - \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta^{\mathrm{NKSD}}_{1,*})) \quad (16)$$

where $\theta^{\mathrm{NKSD}}_{j,*} := \operatorname{argmin} \mathrm{NKSD}(p_0(x_{\mathcal{F}_j})\|q(x_{\mathcal{F}_j}|\theta))$ for $j \in \{1,2\}$. Therefore, $\tilde{q}(X^{(1:N)}|\mathcal{F})$ and $\mathcal{K}$ both yield data selection consistency. It is important here that the SVC uses a true divergence, rather than a divergence up to a data-dependent constant. If we instead used

$$\mathcal{K}^{(\mathrm{a})} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} q(X^{(1:N)}_{\mathcal{F}}), \quad (17)$$

which employs the foreground marginal likelihood $q(X^{(1:N)}_{\mathcal{F}}) = \int q(X^{(1:N)}_{\mathcal{F}}|\theta)\pi(\theta)d\theta$ and a background volume correction, we would get qualitatively different behavior (Section B.2):

$$\frac{1}{N}\log\frac{\mathcal{K}^{(\mathrm{a})}_1}{\mathcal{K}^{(\mathrm{a})}_2} \xrightarrow[N\to\infty]{P_0} \mathrm{KL}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta^{\mathrm{KL}}_{2,*})) - \mathrm{KL}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta^{\mathrm{KL}}_{1,*})) + H_{\mathcal{F}_2} - H_{\mathcal{F}_1} \quad (18)$$

where $H_{\mathcal{F}_j} := -\int p_0(x_{\mathcal{F}_j})\log p_0(x_{\mathcal{F}_j})dx_{\mathcal{F}_j}$ is the entropy of $p_0(x_{\mathcal{F}_j})$ for $j \in \{1,2\}$. In short, the naive score $\mathcal{K}^{(\mathrm{a})}$ is a bad choice: it decides between data subspaces based not just on how well the parametric foreground model performs, but also on the entropy of the data distribution in each space. As a result, $\mathcal{K}^{(\mathrm{a})}$ does not exhibit data selection consistency.

3.2 Nested data selection consistency

When $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$, we refer to the problem of deciding between subspaces $\mathcal{F}_1$ and $\mathcal{F}_2$ as nested data selection, in counterpoint to nested model selection, where one model is a

subset of another (Vuong, 1989). If the model $q(x|\theta)$ is well-specified over $\mathcal{X}_{\mathcal{F}_1}$, then it is guaranteed to be well-specified over any lower-dimensional sub-subspace $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$; in this case, we refer to asymptotic concentration on $\mathcal{F}_1$ as "nested data selection consistency". In this situation, $\mathrm{KL}(p_0(x_{\mathcal{F}_j}) \| q(x_{\mathcal{F}_j}|\theta^{\mathrm{KL}}_{j,*}))$ and $\mathrm{NKSD}(p_0(x_{\mathcal{F}_j}), q(x_{\mathcal{F}_j}|\theta^{\mathrm{NKSD}}_{j,*}))$ are both zero for $j \in \{1, 2\}$, making it necessary to look at higher-order terms in Equations 15 and 16. In Section B.3, we show that if $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$, $q(x|\theta)$ is well-specified over $\mathcal{X}_{\mathcal{F}_1}$, the background models are well-specified, and their dimensions $m_{\mathcal{B}_1}$ and $m_{\mathcal{B}_2}$ are constant with respect to N, then

$$\frac{1}{\log N} \log \frac{\tilde{q}(X^{(1:N)}|\mathcal{F}_1)}{\tilde{q}(X^{(1:N)}|\mathcal{F}_2)} \xrightarrow[N\to\infty]{P_0} \frac{1}{2}(m_{\mathcal{F}_2} + m_{\mathcal{B}_2} - m_{\mathcal{F}_1} - m_{\mathcal{B}_1}) \tag{19}$$

where $m_{\mathcal{F}_j}$ is the effective dimension of the parameter space of $q(x_{\mathcal{F}_j}|\theta)$. In Theorem 17, we show that under mild regularity conditions, the SVC behaves the same way:

$$\frac{1}{\log N} \log \frac{\mathcal{K}_1}{\mathcal{K}_2} \xrightarrow[N\to\infty]{P_0} \frac{1}{2}(m_{\mathcal{F}_2} + m_{\mathcal{B}_2} - m_{\mathcal{F}_1} - m_{\mathcal{B}_1}). \tag{20}$$

Thus, so long as $m_{\mathcal{F}_2} + m_{\mathcal{B}_2} > m_{\mathcal{F}_1} + m_{\mathcal{B}_1}$ whenever $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$, the marginal likelihood and the SVC asymptotically concentrate on the larger foreground $\mathcal{F}_1$; hence, they both exhibit nested data selection consistency. This is a natural assumption since the background model is generally more flexible—on a per dimension basis—than the foreground model.

The volume correction $(2\pi/N)^{m_{\mathcal{B}}/2}$ in the definition of the SVC is important for nested data selection consistency (Equation 20). An alternative score without that correction,

$$\mathcal{K}^{(\mathrm{b})} := \int \exp\Big(- \frac{N}{T} \widehat{\mathrm{NKSD}}(p_0(x_{\mathcal{F}}) \| q(x_{\mathcal{F}}|\theta)) \Big) \pi(\theta) d\theta, \tag{21}$$

exhibits data selection consistency (Equation 16 holds for $\mathcal{K}^{(\mathrm{b})}$), but not nested data selection consistency; see Sections B.2 and B.3. More subtly, the asymptotics of the SVC in the case of nested data selection also depend on the variance of U-statistics. To illustrate, consider a score that is similar to the SVC but uses $\widehat{\mathrm{KL}}$ instead of $\widehat{\mathrm{NKSD}}$:

$$\mathcal{K}^{(\mathrm{c})} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} \int \exp\Big(- N \widehat{\mathrm{KL}}(p_0(x_{\mathcal{F}}) \| q(x_{\mathcal{F}}|\theta)) \Big) \pi(\theta) d\theta \tag{22}$$

where $\widehat{\mathrm{KL}}(p_0(x_{\mathcal{F}}) \| q(x_{\mathcal{F}}|\theta)) := -\frac{1}{N}\sum_{i=1}^{N} \log q(X^{(i)}_{\mathcal{F}}|\theta) - H_{\mathcal{F}}$ and $H_{\mathcal{F}}$ is required to be known. The score $\mathcal{K}^{(\mathrm{c})}$ exhibits data selection consistency, but not nested data selection consistency. The reason is that the error in estimating the KL is of order $1/\sqrt{N}$ by the central limit theorem, and this source of error dominates the $\log N$ term contributed by the volume correction; see Section B.3. Meanwhile, the error in estimating the NKSD is of order $1/N$ when the model is well-specified, due to the rapid convergence rate of the U-statistic estimator. Thus, in the SVC, this source of error is dominated by the volume correction; see Theorem 12.

The nested data selection results we have described so far assume $m_{\mathcal{B}}$ does not depend on N, or at least $m_{\mathcal{B}_2} - m_{\mathcal{B}_1}$ does not depend on N (Theorem 17). However, in Section 2.1, we suggest setting $m_{\mathcal{B}} = c_{\mathcal{B}}\, r_{\mathcal{B}} \sqrt{N}$ where $c_{\mathcal{B}}$ is a constant and $r_{\mathcal{B}}$ is the dimension of $\mathcal{X}_{\mathcal{B}}$. With this choice, the asymptotics of the SVC for nested data selection become (Theorem 17)

$$\frac{1}{\sqrt{N}\log N} \log \frac{\mathcal{K}_1}{\mathcal{K}_2} \xrightarrow[N\to\infty]{P_0} \frac{1}{2} c_{\mathcal{B}}\, (r_{\mathcal{B}_2} - r_{\mathcal{B}_1}). \tag{23}$$

Since $r_{\mathcal{B}_1} < r_{\mathcal{B}_2}$ when $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$, the SVC concentrates on the larger foreground $\mathcal{F}_1$, yielding nested data selection consistency. Going beyond the well-specified case, Theorem 17 shows that Equation 23 holds when $\text{NKSD}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1} \mid \theta_{1,*}^{\text{NKSD}})) = \text{NKSD}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2} \mid \theta_{2,*}^{\text{NKSD}})) \neq 0$, that is, when the models are misspecified by the same amount as measured by the NKSD. Equation 23 holds regardless of whether $m_{\mathcal{F}_1}$ is equal to $m_{\mathcal{F}_2}$.

3.3 Model selection and nested model selection consistency

Consider comparing different foreground models $q_1(x_{\mathcal{F}}|\theta_1)$ and $q_2(x_{\mathcal{F}}|\theta_2)$ over the same subspace $\mathcal{X}_{\mathcal{F}}$, while using the same background model. We say that a score exhibits "model selection consistency" if it concentrates on the correct model, when one of the models is correctly specified and the other is not. When the two models are nested and both are correct, a score exhibits "nested model selection consistency" if it concentrates on the simpler model.

Like the standard marginal likelihood, the SVC exhibits both types of model selection consistency. The standard marginal likelihood satisfies (Section B.4)

$$\frac{1}{N}\log\frac{\tilde{q}_1(X^{(1:N)}|\mathcal{F})}{\tilde{q}_2(X^{(1:N)}|\mathcal{F})} \xrightarrow[N\to\infty]{P_0} \text{KL}(p_0(x_{\mathcal{F}})\|q_2(x_{\mathcal{F}}|\theta_{2,*}^{\text{KL}})) - \text{KL}(p_0(x_{\mathcal{F}})\|q_1(x_{\mathcal{F}}|\theta_{1,*}^{\text{KL}})) \tag{24}$$

where $\theta_{j,*}^{\text{KL}} := \operatorname{argmin} \text{KL}(p_0(x_{\mathcal{F}})\|q_j(x_{\mathcal{F}}|\theta_j))$ for $j \in \{1,2\}$. Analogously, by Theorem 17,

$$\frac{1}{N}\log\frac{\mathcal{K}_1}{\mathcal{K}_2} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\text{NKSD}(p_0(x_{\mathcal{F}})\|q_2(x_{\mathcal{F}}|\theta_{2,*}^{\text{NKSD}})) - \frac{1}{T}\text{NKSD}(p_0(x_{\mathcal{F}})\|q_1(x_{\mathcal{F}}|\theta_{1,*}^{\text{NKSD}})) \tag{25}$$

where $\theta_{j,*}^{\text{NKSD}} := \operatorname{argmin} \text{NKSD}(p_0(x_{\mathcal{F}})\|q_j(x_{\mathcal{F}}|\theta_j))$ for $j \in \{1,2\}$. Thus, for both scores, concentration occurs on the model that comes closer to the data distribution in terms of the corresponding divergence (KL or NKSD).

For nested model selection, suppose both foreground models are well-specified and $m_{\mathcal{B}_1} = m_{\mathcal{B}_2}$. Letting $m_{\mathcal{F},j}$ be the parameter dimension of $q_j(x_{\mathcal{F}}|\theta_j)$, we have (Section B.5)

$$\frac{1}{\log N}\log\frac{\tilde{q}_1(X^{(1:N)}|\mathcal{F})}{\tilde{q}_2(X^{(1:N)}|\mathcal{F})} \xrightarrow[N\to\infty]{P_0} \frac{1}{2}(m_{\mathcal{F},2} - m_{\mathcal{F},1}). \tag{26}$$

In Theorem 17, we show that the SVC behaves identically:

$$\frac{1}{\log N}\log\frac{\mathcal{K}_1}{\mathcal{K}_2} \xrightarrow[N\to\infty]{P_0} \frac{1}{2}(m_{\mathcal{F},2} - m_{\mathcal{F},1}). \tag{27}$$

Here, a key role is played by the volume of the foreground parameter space, which quantifies the foreground model complexity. The SVC accounts for this by integrating over foreground parameter space. Meanwhile, a naive alternative that ignores the foreground volume,

$$\mathcal{K}^{(\text{d})} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} \exp\Big(-\frac{N}{T}\min_{\theta}\widehat{\text{NKSD}}(p_0(x_{\mathcal{F}})\|q(x_{\mathcal{F}}|\theta))\Big), \tag{28}$$

exhibits model selection consistency (Equation 25 holds for $\mathcal{K}^{(\text{d})}$) but not nested model selection consistency (Section B.5). The Laplace and BIC approximations to the SVC (Equations 10 and 11) explicitly correct for the foreground parameter volume without integrating.

4. Related work

Projection pursuit methods are closely related to data selection in that they attempt to identify "interesting" subspaces of the data. However, projection pursuit uses certain prespecified objective functions to optimize over projections, whereas our method allows one to specify a model of interest (Huber, 1985).

Another related line of research is on Bayesian goodness-of-fit (GOF) tests, which compute the posterior probability that the data comes from a given parametric model versus a flexible alternative such as a nonparametric model. Our setup differs in that it aims to compare among different semiparametric models. Nonetheless, in an effort to address the GOF problem, a number of authors have developed nonparametric models with tractable marginals (Verdinelli and Wasserman, 1998; Berger and Guglielmi, 2001), and using these models as the background component in an augmented model could in theory solve data selection problems. In practice, however, such models can only be applied to one-dimensional or few-dimensional data spaces. In Section 7, we show that naively extending the method of Berger and Guglielmi (2001) to the multi-dimensional setting has fundamental limitations.

There is a sizeable frequentist literature on GOF testing using discrepancies (Gretton et al., 2012; Barron, 1989; Györfi and Van Der Meulen, 1991). Our proposed method builds directly on the KSD-based GOF test proposed by Liu et al. (2016) and Chwialkowski et al. (2016). However, using these methods to draw comparisons between different foreground subspaces is non-trivial, since the set of alternative models considered by the GOF test, though nonparametric, will be different over data spaces with different dimensionality. Moreover, the Bayesian aspect of the SVC makes it more straightforward to integrate prior information and employ hierarchical models.

In composite likelihood methods, instead of the standard likelihood, one uses the product of the conditional likelihoods of selected statistics (Lindsay, 1988; Varin et al., 2011). Composite likelihoods have seen widespread use, often for robustness or computational purposes. However, in composite likelihood methods, the choice of statistics is fixed before performing inference. In contrast, in data selection the choice of statistics is a central quantity to be inferred.

Relatedly, our work connects with the literature on robust Bayesian methods. Doksum and Lo (1990) propose conditioning on the value of an insufficient statistic, rather than the complete data set, when performing inference; also see Lewis et al. (2021). However, making an appropriate choice of statistic requires one to know which aspects of the model are correct; in contrast, our procedure infers the choice of statistic. The NKSD posterior also falls within the general class of Gibbs posteriors, which have been studied in the context of robustness, randomized estimators, and generalized belief updating (Zhang, 2006a,b; Jiang and Tanner, 2008; Bissiri et al., 2016; Jewson et al., 2018; Miller and Dunson, 2019).

Our theoretical results also contribute to the emerging literature on Stein discrepancies (Anastasiou et al., 2021). Barp et al. (2019) recently proposed minimum kernelized Stein discrepancy estimators and established their consistency and asymptotic normality. In Section 6, we establish a Bayesian counterpart to these results, showing that the NKSD posterior is asymptotically normal (in the sense of Bernstein–von Mises) and admits a Laplace approximation. To prove this result, we rely on the recent work of Miller (2021) on the asymptotics of generalized posteriors. Since Barp et al. (2019) show that the kernelized

Stein discrepancy is related to the Hyvärinen divergence in that both are Stein discrepancies, our work bears an interesting relationship to that of Shao et al. (2018), who use a Bayesian version of the Hyvärinen divergence to perform model selection with improper priors. They derive a consistency result analogous to Equation 16, however, their model selection score takes the form of a prequential score, not a Gibbs marginal likelihood as in the SVC, and cannot be used for data selection.

In independent recent work, Matsubara et al. (2022) propose a Gibbs posterior based on the KSD and derive a Bernstein-von Mises theorem similar to Theorem 9 using the results of Miller (2021). Their method is not motivated by the Bayesian data selection problem but rather by (1) inference for energy-based models with intractable normalizing constants and (2) robustness to ϵ-contamination. Their Bernstein-von Mises theorem differs from ours in that it applies to a V-statistic estimator of the KSD rather than a U-statistic estimator of the NKSD.

Our linear approximation to the minimum Stein discrepancy estimator (Section 2.3.3) is inspired by previous work on empirical influence functions and the Swiss Army infinitesimal jackknife (Giordano et al., 2019; Koh and Liang, 2017). These previous methods similarly compute the linear response of an extremum estimator with respect to perturbations of the data set, but focus on the effects of dropping datapoints rather than data subspaces.

5. Toy example

The purpose of this toy example is to illustrate the behavior of the Stein volume criterion, and compare it to some of the defective alternatives listed in Table 1, in a simple setting where all computations can be done analytically (Section A.3). In all of the following experiments, we simulated data from a bivariate normal distribution: $X^{(1)}, \ldots, X^{(N)}$ i.i.d. $\sim \mathcal{N}((0,0)^\top, \Sigma_0)$.

To set up the Stein volume criterion, we set $T = 5$ and we choose a radial basis function kernel, $k(x, y) = \exp(-\frac{1}{2}\|x - y\|_2^2)$, which factors across dimensions. We considered both data set size-independent values of $m_{\mathcal{B}}$ (in particular, $m_{\mathcal{B}} = 5\, r_{\mathcal{B}}$) and data set size-dependent values of $m_{\mathcal{B}}$ (in particular, Equation 3 with $\alpha = 0.5$, $\nu = 1$, and $D = 0.2$, where fractional values of D correspond to shared parameters across components in the Pitman-Yor mixture model), obtaining very similar results in each case (shown in Figures 2 and 10, respectively). These choices of $m_{\mathcal{B}}$ ensure that, except for at very small N, the background model has more parameters per data dimension than each of the foreground models considered below, which have just one. In particular, $m_{\mathcal{B}} > 1\, r_{\mathcal{B}}$ for all N (in the size-independent case) and for $N \geq 5$ (in the size-dependent case).

5.1 Data selection consistency

First, we set Σ_0 to be a diagonal matrix with entries $(1, 1/2)$, that is, $\Sigma_0 = \mathrm{diag}(1, 1/2)$, and for $x \in \mathbb{R}^2$, we consider the model

$$\begin{aligned} q(x|\theta) &= \mathcal{N}(x \mid \theta, I) \\ \pi(\theta) &= \mathcal{N}(\theta \mid (0,0)^\top, 10I) \end{aligned} \tag{29}$$

where I denotes the identity matrix. This parametric model is misspecified, owing to the incorrect choice of covariance matrix. We consider two choices of foreground subspace: the

first dimension (defined by the projection matrix $V_{\mathcal{F}_1} = (1,0)^\top$) or the second dimension (projection matrix $V_{\mathcal{F}_2} = (0,1)^\top$). The model is only well-specified for $\mathcal{F}_1$ (not $\mathcal{F}_2$), so a successful data selection procedure would asymptotically select $\mathcal{F}_1$.

In Figure 2a, we see that the SVC correctly concentrates on $\mathcal{F}_1$ as the number of datapoints N increases, with the log SVC ratio growing linearly in N, as predicted by Equation 16. Meanwhile, the naive alternative score $\mathcal{K}^{(\mathrm{a})}$ (Equation 17) fails since it depends on the foreground entropies, while $\mathcal{K}^{(\mathrm{b})}$ (Equation 21) succeeds since the volume correction is negligible in this case; see Section 3.1 and Table 1.

5.2 Nested data selection consistency

Next, we examine the nested data selection case. We use the same model (Equation 29), but we set $\Sigma_0 = I$ so that the model is well-specified even without being projected. We compare the complete data space ($\mathcal{X}_{\mathcal{F}_1} = \mathcal{X}$, projection matrix $V_{\mathcal{F}_1} = I$) to the first dimension alone (projection matrix $V_{\mathcal{F}_1} = (1,0)^\top$). Nested data selection consistency demands that the higher-dimensional data space $\mathcal{X}_{\mathcal{F}_1}$ be preferred asymptotically, since the model is well-specified for both $\mathcal{X}_{\mathcal{F}_1}$ and $\mathcal{X}_{\mathcal{F}_2}$. Figure 2b shows that this is indeed the case for the Stein volume criterion, with the log SVC ratio growing at a $\log N$ rate when $m_{\mathcal{B}}$ is independent of N, as predicted by Equation 20. When $m_{\mathcal{B}}$ depends on N via the Pitman-Yor expression, the log SVC ratio grows at a $N^\alpha \log N$ rate (Figure 10b). Meanwhile, Figure 2b shows that $\mathcal{K}^{(\mathrm{a})}$ and $\mathcal{K}^{(\mathrm{b})}$ both fail to exhibit nested data selection consistency, in accordance with our theory (Section 3.2 and Table 1).

5.3 Model selection consistency (nested and non-nested)

Finally, we examine model selection and nested model selection consistency. We again set $\Sigma_0 = I$. We first compare the (well-specified) model $q(x|\theta) = \mathcal{N}(x \mid \theta, I)$ to the (misspecified) model $q(x|\theta) = \mathcal{N}(x \mid \theta, 2I)$, using the prior $\pi(\theta) = \mathcal{N}(\theta \mid (0,0)^\top, 10I)$ for both models. As shown in Figure 2c, the SVC correctly concentrates on the first model, with the log SVC ratio growing linearly in N, as predicted by Equation 25. The same asymptotic behavior is exhibited by $\mathcal{K}^{(\mathrm{a})}$, which is equivalent to the standard Bayesian marginal likelihood in this setting (Section 3.3). Finally, to check nested model selection consistency, we compare two well-specified nested models: $q(x) = \mathcal{N}(x \mid (0,0)^\top, I)$ and $q(x|\theta) = \mathcal{N}(x \mid \theta, I)$. Figure 2d shows that the SVC correctly selects the simpler model (that is, the model with smaller parameter dimension) and the log SVC ratio grows as $\log N$ (Equation 27). This, too, matches the behavior of the standard Bayesian marginal likelihood, seen in the plot of $\mathcal{K}^{(\mathrm{a})}$.

6. Theory

In this section we describe our formal theoretical results. We start by studying the NKSD and then the NKSD posterior, before finally establishing data and model selection consistency for the SVC.

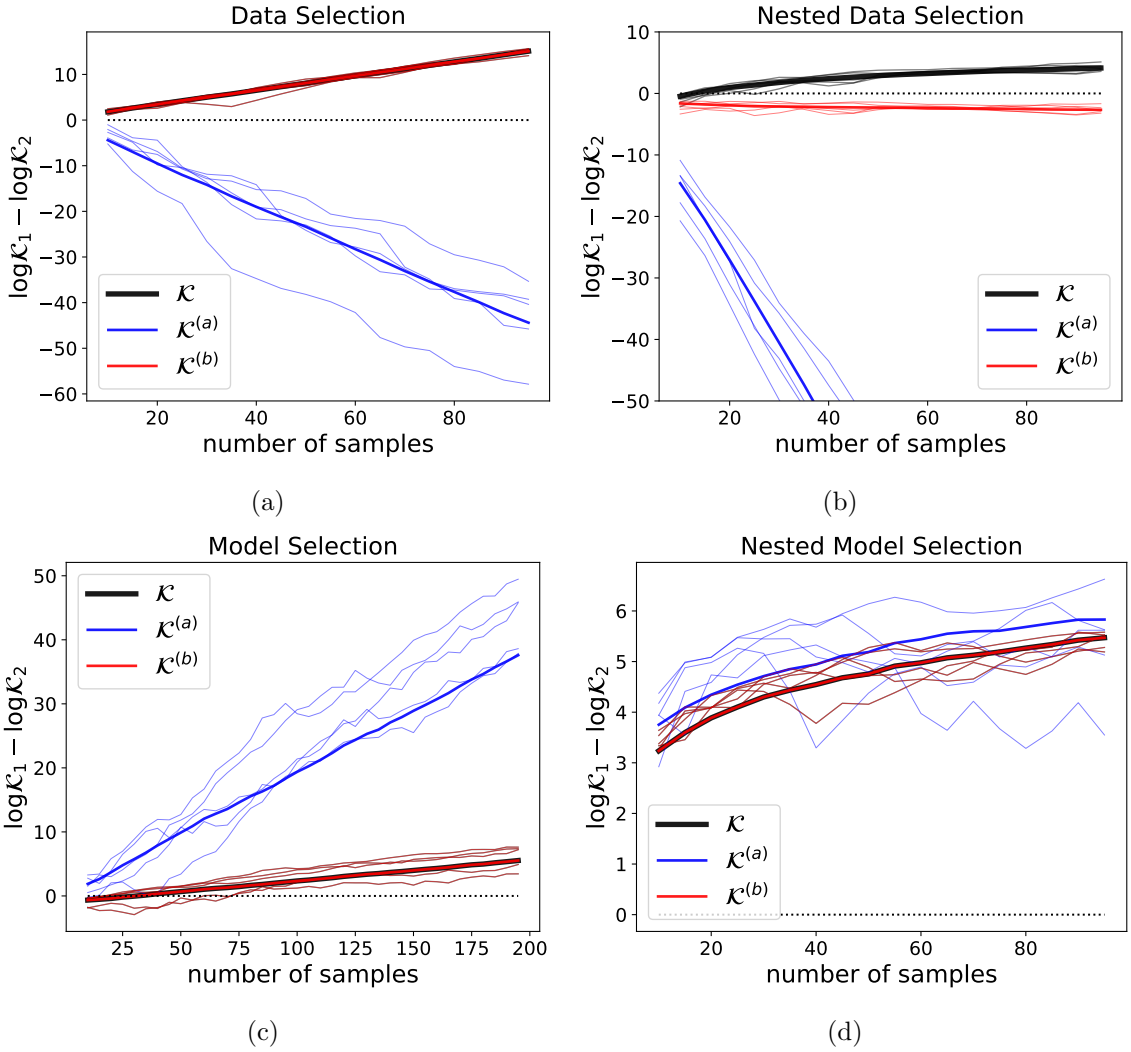


Figure 2: Behavior of the Stein volume criterion $\mathcal{K}$, the foreground marginal likelihood with a background volume correction $\mathcal{K}^{(\mathrm{a})}$, and the foreground marginal NKSD $\mathcal{K}^{(\mathrm{b})}$ on toy examples. Here, we set $m_{\mathcal{B}} = 5\, r_{\mathcal{B}}$. The plots show the results for 5 randomly generated data sets (thin lines) and the average over 100 random data sets (bold lines).

6.1 Properties of the NKSD

Suppose $X^{(1)}, \ldots, X^{(N)}$ are i.i.d. samples from a probability measure P on $\mathcal{X} \subseteq \mathbb{R}^d$ having density $p(x)$ with respect to the Lebesgue measure. Let $L^1(P)$ denote the set of measurable functions f such that $\int \|f(x)\| p(x) dx < \infty$ where $\|\cdot\|$ is the Euclidean norm. We impose the following regularity conditions to use the NKSD to compare P with another probability measure Q having density $q(x)$ with respect to the Lebesgue measure; these are similar to conditions used for the standard KSD in previous work (Liu et al., 2016; Barp et al., 2019).

Condition 1 (Restrictions on p and q) *Assume $s_p(x) := \nabla_x \log p(x)$ and $s_q(x) := \nabla_x \log q(x)$ exist and are continuous for all $x \in \mathcal{X}$, and assume $\mathcal{X}$ is connected and open. Further, assume $s_p, s_q \in L^1(P)$.*

We refer to s_p as the Stein score function of p. Note that existence of $s_p(x)$ implies $p(x) > 0$. Now, consider a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The kernel k is said to be *integrally strictly positive definite* if for any $g : \mathcal{X} \to \mathbb{R}$ such that $0 < \int_{\mathcal{X}} |g(x)|dx < \infty$, we have $\int_{\mathcal{X}} \int_{\mathcal{X}} g(x)k(x,y)g(y)dxdy > 0$. The kernel k is said to *belong to the Stein class of P* if $\int_{\mathcal{X}} \nabla_x(k(x,y)p(x))dx = 0$ for all $y \in \mathcal{X}$.

Condition 2 (Restrictions on k) *Assume the kernel k is symmetric, bounded, integrally strictly positive definite, and belongs to the Stein class of P.*

The following result shows that the NKSD can be written in a way that does not involve s_p; this is particularly useful for estimating the NKSD when P is unknown.

Proposition 3 *If Conditions 1 and 2 hold, then the NKSD is finite and*

$$\text{NKSD}(p(x)\|q(x)) := \frac{\mathbb{E}_{X,Y\sim p}[u(X,Y)]}{\mathbb{E}_{X,Y\sim p}[k(X,Y)]} \tag{30}$$

where

$$u(x,y) = s_q(x)^\top s_q(y)k(x,y) + s_q(x)^\top \nabla_y k(x,y) + s_q(y)^\top \nabla_x k(x,y) + \text{trace}(\nabla_x \nabla_y^\top k(x,y)). \tag{31}$$

The proof is in Section C.1. Next, we show the NKSD satisfies the properties of a divergence.

Proposition 4 *If Conditions 1 and 2 hold, then*

$$\text{NKSD}(p(x)\|q(x)) \geq 0, \tag{32}$$

with equality if and only if $p(x) = q(x)$ almost everywhere.

The proof is in Section C.1. Unlike the standard KSD, but like the KL divergence, the NKSD exhibits subsystem independence (Caticha, 2004, 2011; Rezende, 2018): if two distributions P and Q have the same independence structure, then the total NKSD separates into a sum of individual NKSD terms. This is formalized in Proposition 6.

Condition 5 (Shared independence structure) *Let $x = (x_1^\top, x_2^\top)^\top$ be a decomposition of a vector $x \in \mathbb{R}^d$ into two subvectors, x_1 and x_2. Assume $p(x)$ and $q(x)$ factor as $p(x) = p(x_1)p(x_2)$ and $q(x) = q(x_1)q(x_2)$, and that the kernel k factors as $k(x,y) = k_1(x_1,y_1)k_2(x_2,y_2)$ where k_1 and k_2 both satisfy Condition 2.*

Proposition 6 (Subsystem independence) *If Conditions 1, 2, and 5 hold, then*

$$\text{NKSD}(p(x)\|q(x)) = \text{NKSD}(p(x_1)\|q(x_1)) + \text{NKSD}(p(x_2)\|q(x_2)) \tag{33}$$

where the first term on the right-hand side uses kernel k_1 and the second term uses k_2.

See Section C.1 for the proof. Subsystem independence is powerful since it separates the problem of evaluating the foreground model from that of evaluating the background model. A modified version applies to the estimator $\widehat{\text{NKSD}}(p\|q)$ (Equation 5); see Proposition 20.

6.2 Bernstein–von Mises theorem for the NKSD posterior

In this section, we establish asymptotic properties of the SVC and, more broadly, of its corresponding generalized posterior, which we refer to as the NKSD posterior, defined as

$$\pi_N(\theta) \propto \exp\Big(- \frac{N}{T} \widehat{\mathrm{NKSD}}(p_0(x_{\mathcal{F}}) \| q(x_{\mathcal{F}}|\theta)) \Big) \pi(\theta). \tag{34}$$

In particular, in Theorem 9, we show that the NKSD posterior concentrates and is asymptotically normal, and we establish that the Laplace approximation to the SVC (Equation 10) is asymptotically correct. These results form a Bayesian counterpart to those of Barp et al. (2019), who establish the consistency and asymptotic normality of minimum KSD estimators. Thus, in both the frequentist and Bayesian contexts, we can replace the average log likelihood with the negative KSD and obtain similar key properties. Our results in this section do not depend on whether or not we are working with a foreground subspace, so we suppress the $x_{\mathcal{F}}$ notation.

Let $\Theta \subseteq \mathbb{R}^m$, and let $\{Q_\theta : \theta \in \Theta\}$ be a family of probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$ having densities $q_\theta(x)$ with respect to Lebesgue measure. For notational convenience, we sometimes write $q(x|\theta)$ instead of $q_\theta(x)$. Suppose the data $X^{(1)}, \ldots, X^{(N)}$ are i.i.d. samples from some probability measure P_0 on $\mathcal{X}$ having density $p_0(x)$ with respect to Lebesgue measure. To ensure the NKSD satisfies the properties of a divergence for all q_θ, and that convergence of $\widehat{\mathrm{NKSD}}$ is uniform on compact subsets of Θ (Proposition 21), we require the following.

Condition 7 *Assume Conditions 1 and 2 hold for p_0, k, and q_θ for all $\theta \in \Theta$. Further, assume that the kernel k has continuous and bounded partial derivatives up to and including second order, and $k(x, y) > 0$ for all $x, y \in \mathcal{X}$.*

Now we can set up the generalized posterior. First define

$$f_N(\theta) := \frac{1}{T} \widehat{\mathrm{NKSD}}(p_0(x) \| q(x|\theta)) = \frac{1}{T} \frac{\sum_{i \neq j} u_\theta(X^{(i)}, X^{(j)})}{\sum_{i \neq j} k(X^{(i)}, X^{(j)})}, \tag{35}$$

where $u_\theta(x, y)$ is the $u(x, y)$ function from Equation 5 with q_θ in place of q. For the case of $N = 1$, we define $f_1(\theta) = 0$ by convention. Note that $-N f_N(\theta)$ plays the role of the log likelihood. Also define

$$\begin{aligned} f(\theta) &:= \frac{1}{T} \mathrm{NKSD}(p_0(x) \| q(x|\theta)), \\ z_N &:= \int_\Theta \exp(-N f_N(\theta)) \pi(\theta) d\theta, \\ \pi_N(\theta) &:= \frac{1}{z_N} \exp(-N f_N(\theta)) \pi(\theta), \end{aligned} \tag{36}$$

where $\pi(\theta)$ is a prior density on Θ. Note that $\pi_N(\theta)d\theta$ is the NKSD posterior and z_N is the corresponding generalized marginal likelihood employed in the SVC. Denote the gradient and Hessian of f by $f'(\theta) = \nabla_\theta f(\theta)$ and $f''(\theta) = \nabla^2_\theta f(\theta)$, respectively. To ensure that the NKSD posterior is well defined and has an isolated maximum, we assume the following condition.

Condition 8 *Suppose $\Theta \subseteq \mathbb{R}^m$ is a convex set and (a) Θ is compact or (b) Θ is open and f_N is convex on Θ with probability 1 for all N. Assume $z_N < \infty$ a.s. for all N. Assume f has a unique minimizer $\theta_* \in \Theta$, $f''(\theta_*)$ is invertible, π is continuous at θ_*, and $\pi(\theta_*) > 0$.*

By Proposition 4, f has a unique minimizer whenever $\{Q_\theta : \theta \in \Theta\}$ is well-specified and identifiable, that is, when $Q_\theta = P_0$ for some θ and $\theta \mapsto Q_\theta$ is injective.

In Theorem 9 below, we establish the following results: (1) the minimum $\widehat{\text{NKSD}}$ converges to the minimum NKSD; (2) π_N concentrates around the minimizer of the NKSD; (3) the Laplace approximation to z_N is asymptotically correct; and (4) π_N is asymptotically normal in the sense of Bernstein–von Mises. The primary regularity conditions we need for this theorem are restraints on the derivatives of s_{q_θ} with respect to θ (Condition 10). Our proof of Theorem 9 relies on the theory of generalized posteriors developed by Miller (2021). We use $\|\cdot\|$ for the Euclidean–Frobenius norms: for vectors $A \in \mathbb{R}^D$, $\|A\| = (\sum_i A_i^2)^{1/2}$; for matrices $A \in \mathbb{R}^{D\times D}$, $\|A\| = (\sum_{i,j} A_{i,j}^2)^{1/2}$; for tensors $A \in \mathbb{R}^{D\times D\times D}$, $\|A\| = (\sum_{i,j,k} A_{i,j,k}^2)^{1/2}$; and so on.

Theorem 9 *If Conditions 7, 8, and 10 hold, then there is a sequence $\theta_N \to \theta_*$ a.s. such that:*

1. *$f_N(\theta_N) \to f(\theta_*)$, $f_N'(\theta_N) = 0$ for all N sufficiently large, and $f_N''(\theta_N) \to f''(\theta_*)$ a.s.,*
2. *letting $B_\epsilon(\theta_*) := \{\theta \in \mathbb{R}^m : \|\theta - \theta_*\| < \epsilon\}$, we have*

$$\int_{B_\epsilon(\theta_*)} \pi_N(\theta) d\theta \xrightarrow[N\to\infty]{\text{a.s.}} 1 \text{ for all } \epsilon > 0, \tag{37}$$

3. $$z_N \sim \frac{\exp(-N f_N(\theta_N))\pi(\theta_*)}{|\det f''(\theta_*)|^{1/2}} \left(\frac{2\pi}{N}\right)^{m/2} \tag{38}$$

   almost surely, where $a_N \sim b_N$ means that $a_N/b_N \to 1$ as $N \to \infty$, and

4. *letting h_N denote the density of $\sqrt{N}(\theta - \theta_N)$ when θ is sampled from π_N, we have that h_N converges to $\mathcal{N}(0, f''(\theta_*)^{-1})$ in total variation, that is,*

$$\int_{\mathbb{R}^m} \left| h_N(\tilde{\theta}) - \mathcal{N}(\tilde{\theta} \mid 0, f''(\theta_*)^{-1}) \right| d\tilde{\theta} \xrightarrow[N\to\infty]{\text{a.s.}} 0. \tag{39}$$

The proof is in Section C.2. We write $\nabla_\theta^2 s_{q_\theta}$ to denote the tensor in $\mathbb{R}^{d\times m\times m}$ in which entry (i,j,k) is $\partial^2 s_{q_\theta}(x)_i / \partial\theta_j \partial\theta_k$. Likewise, $\nabla_\theta^3 s_{q_\theta}$ denotes the tensor in $\mathbb{R}^{d\times m\times m\times m}$ in which entry (i,j,k,ℓ) is $\partial^3 s_{q_\theta}(x)_i / \partial\theta_j \partial\theta_k \partial\theta_\ell$. We write $\mathbb{N}$ to denote the set of natural numbers.

Condition 10 (Stein score regularity) *Assume $s_{q_\theta}(x)$ has continuous third-order partial derivatives with respect to the entries of θ on Θ. Suppose that for any compact, convex subset $C \subseteq \Theta$, there exist continuous functions $g_{0,C}, g_{1,C} \in L^1(P_0)$ such that for all $\theta \in C$, $x \in \mathcal{X}$,*

$$\begin{aligned} \|s_{q_\theta}(x)\| &\le g_{0,C}(x), \\ \|\nabla_\theta s_{q_\theta}(x)\| &\le g_{1,C}(x). \end{aligned} \tag{40}$$

Further, assume there is an open, convex, bounded set $E \subseteq \Theta$ such that $\theta_* \in E$, $\bar{E} \subseteq \Theta$, and the sets

$$\Big\{\frac{1}{N}\sum_{i=1}^{N} \|\nabla_\theta^2 s_{q_\theta}(X^{(i)})\| : N \in \mathbb{N}, \theta \in E\Big\}, \tag{41}$$

$$\Big\{\frac{1}{N}\sum_{i=1}^{N} \|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\| : N \in \mathbb{N}, \theta \in E\Big\} \tag{42}$$

are bounded with probability 1.

Next, Theorem 11 shows that in the special case where $q_\theta(x)$ is an exponential family, many of the conditions of Theorem 9 are automatically satisfied.

Theorem 11 *Suppose $\{Q_\theta : \theta \in \Theta\}$ is an exponential family with densities of the form $q_\theta(x) = \lambda(x)\exp(\theta^\top t(x) - \kappa(\theta))$ for $x \in \mathcal{X} \subseteq \mathbb{R}^d$. Assume $\Theta = \{\theta \in \mathbb{R}^m : |\kappa(\theta)| < \infty\}$, and assume Θ is convex, open, and nonempty. Assume $\log\lambda(x)$ and $t(x)$ are continuously differentiable on $\mathcal{X}$, $\|\nabla_x \log\lambda(x)\|$ and $\|\nabla_x t(x)\|$ are in $L^1(P_0)$, and the rows of the Jacobian matrix $\nabla_x t(x) \in \mathbb{R}^{m\times d}$ are linearly independent with positive probability under P_0. Suppose Condition 7 holds, f has a unique minimizer $\theta_* \in \Theta$, the prior π is continuous at θ_*, and $\pi(\theta_*) > 0$. Then the assumptions of Theorem 9 are satisfied for all N sufficiently large.*

The proof is in Section C.2.

6.3 Asymptotics of the Stein volume criterion

The Laplace approximation to the SVC uses the estimate $\widehat{\text{NKSD}}$ and its minimizer θ_N, rather than the true NKSD and its minimizer θ_*. To establish the consistency properties of the SVC, we need to understand the relationship between the two. To do so, we adapt a standard approach to performing such an analysis of the marginal likelihood, for instance, as in Theorem 1 of Dawid (2011).

Theorem 12 *Assume the conditions of Theorem 9 hold, and assume $s_{q_{\theta_*}}$ and $\nabla_\theta\big|_{\theta=\theta_*} s_{q_\theta}$ are in $L^2(P_0)$. Then as $N \to \infty$,*

$$f_N(\theta_N) - f_N(\theta_*) = O_{P_0}(N^{-1}). \tag{43}$$

Further, if $\text{NKSD}(p_0(x)\|q(x|\theta_*)) > 0$ then

$$f_N(\theta_*) - f(\theta_*) = O_{P_0}(N^{-1/2}), \tag{44}$$

whereas if $\text{NKSD}(p_0(x)\|q(x|\theta_*)) = 0$ then

$$f_N(\theta_*) - f(\theta_*) = O_{P_0}(N^{-1}). \tag{45}$$

The proof is in Section C.3. Remarkably, Equation 45 shows that $f_N(\theta_*)$ converges to $f(\theta_*)$ more rapidly when the model is well-specified, specifically, at a $1/N$ rate instead of $1/\sqrt{N}$. This is unusual and is crucial for our results in Section 6.4. The standard log likelihood does not exhibit this rapid convergence; see Section B.1. This property of the NKSD derives from similar properties exhibited by the standard KSD (Liu et al., 2016, Theorem 4.1). Combined with Theorem 9 (part 3), Theorem 12 implies that when the model is misspecified, the leading order term of $\log z_N$ is $-Nf(\theta_*)$, whereas when the model is well-specified, the leading order term is $-\frac{1}{2}\, m \log N$.

6.4 Data and model selection consistency of the SVC

In this section, we establish the asymptotic consistency of the Stein volume criterion (SVC) when used for data selection, nested data selection, model selection, and nested model selection; see Theorem 17. This provides rigorous justification for the claims in Section 3. These results are all in the context of pairwise comparisons between two models or two model projections, M_1 and M_2. Before proving the results, we formally define the consistency properties discussed in Section 3. Each property is defined in terms of a pairwise score $\rho(M_1, M_2)$, such as $\rho(M_1, M_2) = \log(\mathcal{K}_1/\mathcal{K}_2)$. For simplicity, we assume $\rho(M_1, M_2) = -\rho(M_2, M_1)$; this is satisfied for all of the cases we consider. Let $\dim(\cdot)$ denote the dimension of a real space.

Definition 13 (Data selection consistency) *For $j \in \{1, 2\}$, consider foreground model projections $M_j := \{q(x_{\mathcal{F}_j}|\theta) : \theta \in \Theta\}$. We say that ρ satisfies "data selection consistency" if $\rho(M_1, M_2) \to \infty$ as $N \to \infty$ when M_1 is well-specified with respect to $p_0(x_{\mathcal{F}_1})$ and M_2 is misspecified with respect to $p_0(x_{\mathcal{F}_2})$.*

Definition 14 (Nested data selection consistency) *For $j \in \{1, 2\}$, consider foreground model projections $M_j := \{q(x_{\mathcal{F}_j}|\theta) : \theta \in \Theta\}$. We say that ρ satisfies "nested data selection consistency" if $\rho(M_1, M_2) \to \infty$ as $N \to \infty$ when M_1 is well-specified with respect to $p_0(x_{\mathcal{F}_1})$, $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$, and $\dim(\mathcal{X}_{\mathcal{F}_2}) < \dim(\mathcal{X}_{\mathcal{F}_1})$.*

Definition 15 (Model selection consistency) *For $j \in \{1, 2\}$, consider foreground models $M_j := \{q_j(x_{\mathcal{F}}|\theta_j) : \theta_j \in \Theta_j\}$. We say that ρ satisfies "model selection consistency" if $\rho(M_1, M_2) \to \infty$ as $N \to \infty$ when M_1 is well-specified with respect to $p_0(x_{\mathcal{F}})$ and M_2 is misspecified.*

Definition 16 (Nested model selection consistency) *For $j \in \{1, 2\}$, consider foreground models $M_j := \{q_j(x_{\mathcal{F}}|\theta_j) : \theta_j \in \Theta_j\}$. We say that ρ satisfies "nested model selection consistency" if $\rho(M_1, M_2) \to \infty$ as $N \to \infty$ when M_1 is well-specified with respect to $p_0(x_{\mathcal{F}})$, $M_1 \subset M_2$, and $\dim(\Theta_1) < \dim(\Theta_2)$.*

In each case, ρ may diverge almost surely ("strong consistency") or in probability ("weak consistency"). Note that in Definitions 13–14, the difference between M_1 and M_2 is the choice of foreground data space $\mathcal{F}$, whereas in Definitions 15–16, M_1 and M_2 are over the same foreground space but employ different model spaces.

In Theorem 17, we show that the SVC has the asymptotic properties outlined in Section 3. In combination with the subsystem independence properties of the NKSD (Propositions 6 and 20), Theorem 17 also leads to the conclusion that the SVC approximates the NKSD marginal likelihood of the augmented model (Equation 8). Our proof is similar in spirit to previous results for model selection with the standard marginal likelihood, notably those of Hong and Preston (2005) and Huggins and Miller (2021), but relies on the special properties of the NKSD marginal likelihood in Theorem 12.

Theorem 17 *For $j \in \{1, 2\}$, assume the conditions of Theorem 12 hold for model M_j defined on $\mathcal{X}_{\mathcal{F}_j}$, with density $q_j(x_{\mathcal{F}_j}|\theta_j)$ for $\theta_j \in \Theta_j \subseteq \mathbb{R}^{m_{\mathcal{F}_j,j}}$. Let $\mathcal{K}_{j,N}$ be the Stein volume criterion for M_j, with background model penalty $m_{\mathcal{B}_j} = m_{\mathcal{B}_j}(N)$, and let $\theta_{j,*} := \operatorname{argmin}_{\theta_j} \text{NKSD}(p_0(x_{\mathcal{F}_j})\|q_j(x_{\mathcal{F}_j}|\theta_j))$. Then:*

1. *If* $m_{\mathcal{B}_j} = o(N/\log N)$ *for* $j \in \{1,2\}$*, then*

$$\frac{1}{N}\log\frac{\mathcal{K}_{1,N}}{\mathcal{K}_{2,N}} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\text{NKSD}(p_0(x_{\mathcal{F}_2})\|q_2(x_{\mathcal{F}_2}|\theta_{2,*})) - \frac{1}{T}\text{NKSD}(p_0(x_{\mathcal{F}_1})\|q_1(x_{\mathcal{F}_1}|\theta_{1,*})).$$

2. *If* $\text{NKSD}(p_0(x_{\mathcal{F}_1})\|q_1(x_{\mathcal{F}_1}|\theta_{1,*})) = \text{NKSD}(p_0(x_{\mathcal{F}_2})\|q_2(x_{\mathcal{F}_2}|\theta_{2,*})) = 0$ *and* $m_{\mathcal{B}_2} - m_{\mathcal{B}_1}$ *does not depend on* N*, then*

$$\frac{1}{\log N}\log\frac{\mathcal{K}_{1,N}}{\mathcal{K}_{2,N}} \xrightarrow[N\to\infty]{P_0} \frac{1}{2}(m_{\mathcal{F}_2,2} + m_{\mathcal{B}_2} - m_{\mathcal{F}_1,1} - m_{\mathcal{B}_1}).$$

3. *If* $\text{NKSD}(p_0(x_{\mathcal{F}_1})\|q_1(x_{\mathcal{F}_1}|\theta_{1,*})) = \text{NKSD}(p_0(x_{\mathcal{F}_2})\|q_2(x_{\mathcal{F}_2}|\theta_{2,*}))$*,* $m_{\mathcal{B}_1} = c_{\mathcal{B}_1}\sqrt{N}$*, and* $m_{\mathcal{B}_2} = c_{\mathcal{B}_2}\sqrt{N}$*, where* $c_{\mathcal{B}_1}$ *and* $c_{\mathcal{B}_2}$ *are positive and constant in* N*, then*

$$\frac{1}{\sqrt{N}\log N}\log\frac{\mathcal{K}_{1,N}}{\mathcal{K}_{2,N}} \xrightarrow[N\to\infty]{P_0} \frac{1}{2}(c_{\mathcal{B}_2} - c_{\mathcal{B}_1}).$$

The proof is in Section C.4. In particular, assuming the conditions of Theorem 12, we obtain the following consistency results in terms of convergence in probability. Let $D_j := \text{NKSD}(p_0(x_{\mathcal{F}_j})\|q_j(x_{\mathcal{F}_j}|\theta_{j,*}))$ for $j \in \{1,2\}$.

- If $m_{\mathcal{B}_j} = o(N/\log N)$ then the SVC exhibits data selection consistency and model selection consistency. This holds by Theorem 17 (part 1) since $D_2 > D_1 = 0$.
- If $m_{\mathcal{B}_1} = m_{\mathcal{B}_2}$ then the SVC exhibits nested model selection consistency. This holds by Theorem 17 (part 2) since $D_1 = D_2 = 0$, $m_{\mathcal{B}_2} - m_{\mathcal{B}_1} = 0$, and $m_{\mathcal{F}_2,2} > m_{\mathcal{F}_1,1}$.
- Consider a nested data selection problem with $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$. If (A) $m_{\mathcal{B}_2} - m_{\mathcal{B}_1}$ does not depend on N and $m_{\mathcal{F}_2,2} + m_{\mathcal{B}_2} > m_{\mathcal{F}_1,1} + m_{\mathcal{B}_1}$ or (B) $m_{\mathcal{B}_j} = c_{\mathcal{B}_j}\sqrt{N}$ and $c_{\mathcal{B}_2} > c_{\mathcal{B}_1} > 0$, then the SVC exhibits nested data selection consistency. Cases A and B hold by Theorem 17 (parts 2 and 3, respectively) since $D_1 = D_2 = 0$.

7. Application: probabilistic PCA

Probabilistic principal components analysis (pPCA) is a commonly used tool for modeling and visualization. The basic idea is to model the data as linear combinations of k latent factors plus Gaussian noise. The inferred weights on the factors are frequently used to provide low-dimensional summaries of the data, while the factors themselves describe major axes of variation in the data. In practice, pPCA is often applied in settings where it is likely to be misspecified – for instance, the weights are often clearly non-Gaussian. In this section, we show how data selection can be used to uncover sources of misspecification and to analyze how this misspecification affects downstream inferences.

The generative model used in pPCA is

$$\begin{aligned} Z^{(i)} &\sim \mathcal{N}(0, I_k), \\ X^{(i)}|Z^{(i)} &\sim \mathcal{N}(HZ^{(i)}, vI_d), \end{aligned} \tag{46}$$

independently for $i = 1, \ldots, N$, where I_k is the k-dimensional identity matrix, $Z^{(i)} \in \mathbb{R}^k$ is the weight vector for datapoint i, $H \in \mathbb{R}^{d \times k}$ is the unknown matrix of latent factors, and $v > 0$ is the variance of the noise. To form a Laplace approximation for the Stein volume criterion, we follow the approach developed by Minka (2001) for the standard marginal likelihood. Specifically, we parameterize H as

$$H = U(L - vI_k)^{1/2} \tag{47}$$

where U is a $d \times k$ matrix with orthonormal columns (that is, it lies on the Stiefel manifold) and L is a $k \times k$ diagonal matrix. We use the priors suggested by Minka (2001),

$$\begin{aligned} U &\sim \text{Uniform}(\mathcal{U}), \\ L_{ii} &\sim \text{InverseGamma}(\alpha/2,\, \alpha/2), \\ v &\sim \text{InverseGamma}\big((\alpha/2+1)(d-k)-1,\; (\alpha/2)(d-k)\big), \end{aligned} \tag{48}$$

where $\mathcal{U}$ is the set of $d \times k$ matrices with orthonormal columns and L_{ii} is the ith diagonal entry of L. We set $\alpha = 0.1$ in the following experiments, and we use pymanopt (Townsend et al., 2016) to optimize U over the Stiefel manifold (Section D).

7.1 Simulations

In simulations, we evaluate the ability of the SVC to detect partial misspecification. We set $d = 6$, draw the first four dimensions from a pPCA model with $k = 2$ and

$$H = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}, \tag{49}$$

and generate dimensions 5 and 6 in such a way that pPCA is misspecified. We consider two misspecified scenarios: scenario A (Figure 3a) is that

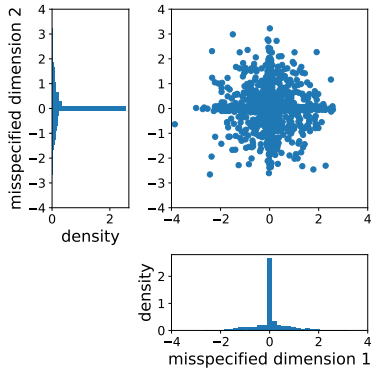
$$\begin{aligned} W^{(i)} &\sim \text{Bernoulli}(0.5), \\ X^{(i)}_{5:6} \mid W^{(i)} &\sim \mathcal{N}\left(0, \Sigma_{W^{(i)}}\right), \end{aligned} \tag{50}$$

where $\Sigma_{W^{(i)}} = (0.05)^{W^{(i)}} I_2$. Scenario B (Figure 3d) is the same but with

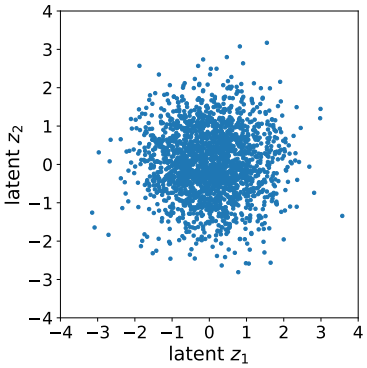
$$\Sigma_{W^{(i)}} = \begin{pmatrix} 1 & (-1)^{W^{(i)}} 0.99 \\ (-1)^{W^{(i)}} 0.99 & 1 \end{pmatrix}. \tag{51}$$

Scenario B is more challenging because the marginals of the misspecified dimensions are still Gaussian, and thus, misspecification only comes from the dependence between X_5 and X_6. As illustrated in Figures 3b and 3e, both kinds of misspecification are very hard to see in the lower-dimensional latent representation of the data.

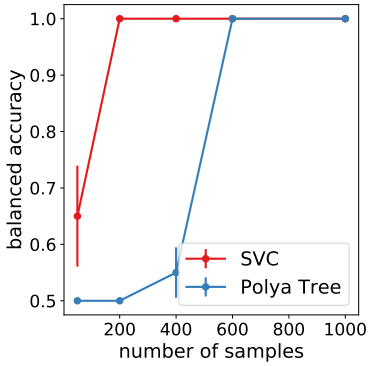
Our method can be used to both (i) detect misspecified subsets of dimensions, and (ii) conversely, find a maximal subset of dimensions for which the pPCA model provides a reasonable fit to the data. We set $T = 0.05$ in the SVC, based on the calibration procedure



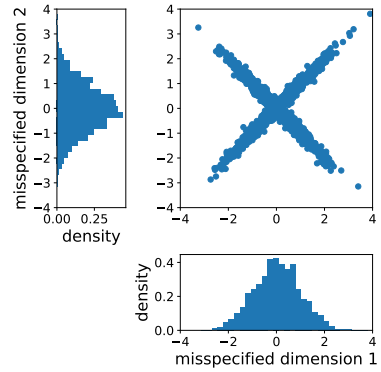
(a) Scenario A, misspecified dimensions.



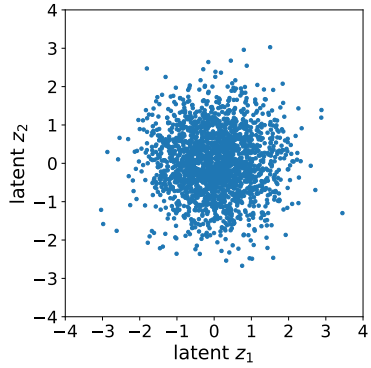
(b) Scenario A, pPCA latent space.



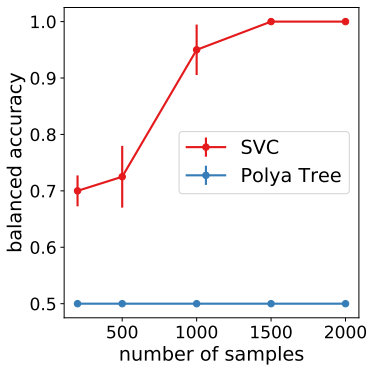
(c) Scenario A, accuracy in detecting misspecified dimensions.



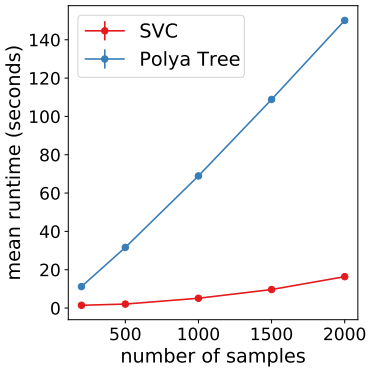
(d) Scenario B, misspecified dimensions.



(e) Scenario B, pPCA latent space.



(f) Scenario B, accuracy in detecting misspecified dimensions.



(g) Mean runtime over 5 repeats.

Figure 3: Data selection in the probabilistic PCA model.

in Section A.1 (Section D.3). We use the Pitman-Yor mixture model expression for the background model dimension (Equation 3), with $\alpha = 0.5$, $\nu = 1$, and $D = 0.2$. This value of D ensures that the number of background model parameters per data dimension is greater than the number of foreground model parameters per data dimension except for at very small N, since there are two foreground parameters for each additional data dimension in the pPCA model, and $m_{\mathcal{B}} > 2\,r_{\mathcal{B}}$ for $N \geq 20$. We performed leave-one-out data selection, comparing the foreground space $\mathcal{X}_{\mathcal{F}_0} = \mathcal{X}$ to foreground spaces $\mathcal{X}_{\mathcal{F}_j}$ for $j \in \{1, \ldots, d\}$, which exclude the jth dimension of the data. We computed the log SVC ratio $\log(\mathcal{K}_j/\mathcal{K}_0) = \log \mathcal{K}_j - \log \mathcal{K}_0$ using the BIC approximation to the SVC (Section 2.3.2) and the approximate optima technique (Section 2.3.3). We quantify the performance of the method in detecting misspecified dimensions in terms of the balanced accuracy, defined as $(TN/N + TP/P)/2$ where TN is the number of true negatives (dimension by dimension), N is the number of negatives, TP is the number of true positives, and P is the number of positives. Experiments were repeated independently five times. Figures 3c and 3f show that as the sample size increases, the SVC correctly infers that dimensions 1 through 4 should be included and dimensions 5 and 6 should be excluded.

7.2 Comparison with a nonparametric background model

To benchmark our method, we compare with an alternative approach that uses an explicit augmented model. The Pólya tree is a nonparametric model with a closed-form marginal likelihood that is tractable for one-dimensional data (Lavine, 1992). We define a flexible background model by sampling each dimension j of the background space independently as

$$X_j \sim \text{PolyaTree}(F, \tilde{F}, \eta), \tag{52}$$

with the Pólya tree constructed as by Berger and Guglielmi (2001) (Section D.4). We set $F = \mathcal{N}(0, 10)$, $\tilde{F} = \mathcal{N}(0, 10)$, and $\eta = 1000$ so that the model is weighted only very weakly towards the base distribution.

We performed data selection using the marginal likelihood of the Pólya tree augmented model, computing the marginal of the pPCA foreground model using the approximation of Minka (2001). The accuracy results for data selection are in Figures 3c and 3f. On scenario A (Equation 50), the Pólya tree augmented model requires significantly more data to detect which dimensions are misspecified. On scenario B (Equation 51) the Pólya tree augmented model fails entirely, preferring the full data space $\mathcal{X}_{\mathcal{F}_0} = \mathcal{X}$ which includes all dimensions (Figure 3f). The reason is that the background model is misspecified due to the assumption of independent dimensions, and thus, the asymptotic data selection results (Equations 15 and 19) do not hold. This could be resolved by using a richer background model that allows for dependence between dimensions, however, computing the marginal likelihood under such a model would be computationally challenging. Even with the independence assumption, the Pólya tree approach is already substantially slower than the SVC (Figure 3g).

7.3 Application to pPCA for single-cell RNA sequencing

Single-cell RNA sequencing (scRNAseq) has emerged as a powerful technology for high-throughput characterization of individual cells. It provides a snapshot of the transcriptional state of each cell by measuring the number of RNA transcripts from each gene. PCA is widely used to study scRNAseq data sets, both as a method for visualizing different cell types in the data set and as a pre-processing technique, where the latent embedding is used for downstream tasks like clustering and lineage reconstruction (Qiu et al., 2017; van Dijk et al., 2018). We applied data selection to answer two practical questions in the application of probabilistic PCA to scRNAseq data: (1) Where is the pPCA model misspecified? (2) How does partial misspecification of the pPCA model affect downstream inferences?

7.3.1 Model criticism

Our first goal was to verify that the SVC provides reasonable inferences of partial model misspecification in practice. We examined two different scRNAseq data sets, focusing for illustration on a data set from human peripheral blood mononuclear cells taken from a healthy donor, and pre-processed the data following standard procedures in the field (Section D.5). We subsampled each data set to 200 genes (selected randomly from among the 2000 most highly expressed) and 2000 cells (selected randomly) for computational tractability, then mean-subtracted and standardized the variance of each gene, again following standard practice in the field. The number of latent components k was set to 3, based on the procedure of Minka (2000). We performed leave-one-out data selection, comparing the foreground space $\mathcal{X}_{\mathcal{F}_0} := \mathcal{X}$ to foreground spaces $\mathcal{X}_{\mathcal{F}_j}$ that exclude the jth gene. We computed the log SVC ratio $\log \mathcal{K}_j - \log \mathcal{K}_0$ using the BIC approximation to the SVC (Section 2.3.2) and the approximate optima technique (Section 2.3.3). We used the same setting of T and of $m_{\mathcal{B}}$ as was used in simulation, resulting in a background model complexity of $m_{\mathcal{B}} = 20\, r_{\mathcal{B}}$ for data sets of this size. Based on the SVC criterion, 162 out of 200 genes should be excluded from the foreground pPCA model, suggesting widespread partial misspecification. Figure 4 compares the histogram of individual genes to their estimated density under the pPCA model inferred for $\mathcal{X}_{\mathcal{F}_0} = \mathcal{X}$. Those genes most favored to be excluded (namely, UBE2V2 and IRF8) show extreme violations of normality, in stark contrast to those genes most favored to be included (MT-CO1 and RPL6).

Next, we compared the results of our data selection approach to a more conventional strategy for model criticism. Criticism of partially misspecified models can be challenging in practice because misspecification of the model over some dimensions of the data can lead to substantial model-data mismatch in dimensions for which the model is indeed well-specified (Jacob et al., 2017). The standard approach to model criticism—first fit a model, then identify aspects of the data that the model poorly explains—can therefore be misleading if our aim is to determine how the model might be improved (e.g., in the context of "Box's loop", Blei, 2014). In particular, standard approaches such as posterior predictive checks will be expected to overstate problems with components of the model that are well-specified and understate problems with components of the model that are misspecified. Bayesian data selection circumvents this issue by evaluating augmented models, which replace potentially misspecified components of the model by well-specified compo-

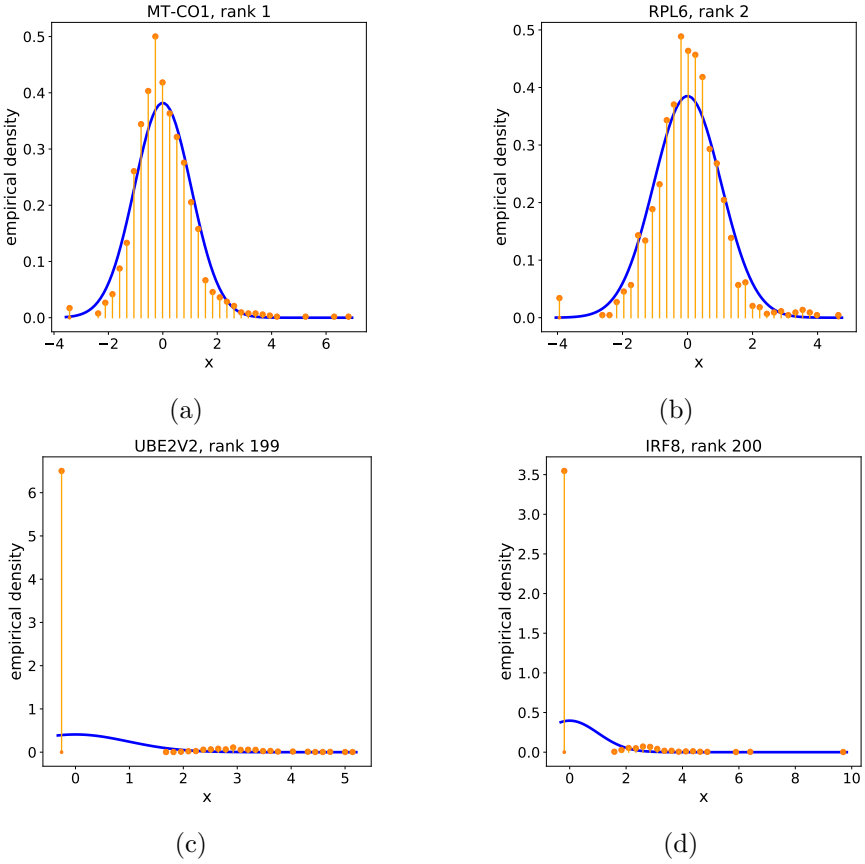


Figure 4: (a,b) Histograms of gene expression (after pre-processing), i.e., $X_j^{(1)}, \ldots, X_j^{(N)}$, for genes j selected to be included in the foreground space based on the log SVC ratio $\log \mathcal{K}_j - \log \mathcal{K}_0$. The estimated density under the pPCA model is shown in blue. (c,d) Histograms of example genes selected to be excluded. Higher ranks (in each title) correspond to larger log SVC ratios.

nents. To illustrate the difference between these approaches in practice, we compared the SVC to a closely analogous measurement of error for the full foreground model (inferred from $\mathcal{X}_{\mathcal{F}_0} = \mathcal{X}$),

$$\log \mathcal{E}_j - \log \mathcal{E}_0 := -\frac{N}{T}\widehat{\text{NKSD}}(p_0(x_{\mathcal{F}_j}) \| q(x_{\mathcal{F}_j}|\theta_{0,N})) + \frac{N}{T}\widehat{\text{NKSD}}(p_0(x) \| q(x|\theta_{0,N})) \tag{53}$$

where $\theta_{0,N} := \operatorname{argmin} \widehat{\text{NKSD}}(p_0(x) \| q(x|\theta))$ is the minimum NKSD estimator for the foreground model when including all dimensions. This model criticism score evaluates the amount of model-data mismatch contributed by the subspace $\mathcal{X}_{\mathcal{B}_j}$ when modeling all data dimensions with the foreground model. For comparison, the BIC approximation to the log

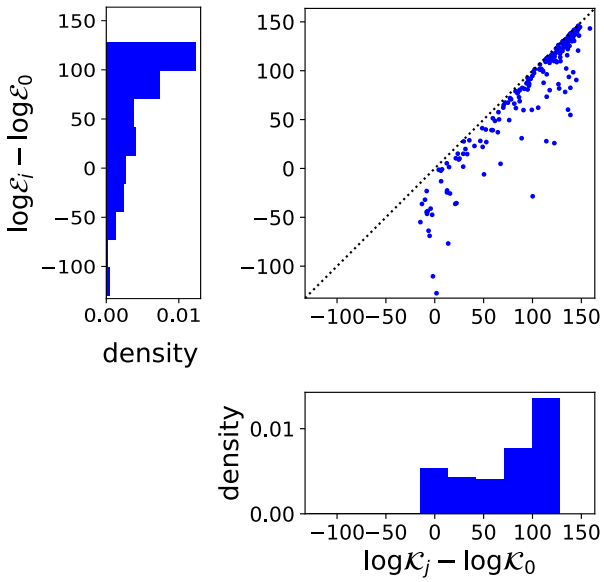


Figure 5: Scatterplot comparison and projected marginals of the leave-one-out log SVC ratio, $\log \mathcal{K}_j - \log \mathcal{K}_0$ (with $m_{\mathcal{B}_j} = m_{\mathcal{F}_0} - m_{\mathcal{F}_j}$), and the conventional full model criticism score, $\log \mathcal{E}_j - \log \mathcal{E}_0$, for each gene.

SVC ratio is

$$\begin{aligned}\log \mathcal{K}_j - \log \mathcal{K}_0 \approx -\frac{N}{T}\widehat{\text{NKSD}}(p_0(x_{\mathcal{F}_j})\|q(x_{\mathcal{F}_j}|\theta_{j,N}) + \frac{N}{T}\widehat{\text{NKSD}}(p_0(x)\|q(x|\theta_{0,N})) \\ + \frac{m_{\mathcal{B}_j} + m_{\mathcal{F}_j} - m_{\mathcal{F}_0}}{2}\log\left(\frac{2\pi}{N}\right)\end{aligned} \tag{54}$$

where $\theta_{j,N} := \operatorname{argmin} \widehat{\text{NKSD}}(p_0(x_{\mathcal{F}_j})\|q(x_{\mathcal{F}_j}|\theta))$ is the minimum NKSD estimator for the projected foreground model applied to the restricted data set, which we approximate as $\theta_{0,N}$ plus the implicit function correction derived in Section 2.3.3. Figure 5 illustrates the differences between the conventional criticism approach ($\log \mathcal{E}_j - \log \mathcal{E}_0$) and the log SVC ratio on an scRNAseq data set. To enable direct comparison of the two methods, we focus on the lower order terms of Equation 54, that is, we set $m_{\mathcal{B}_j} = m_{\mathcal{F}_0} - m_{\mathcal{F}_j}$. We see that the amount of error contributed by $\mathcal{X}_{\mathcal{B}_j}$, as judged by the SVC, is often substantially higher than the amount indicated by the conventional criticism approach, implying that the conventional criticism approach understates the problems caused by individual genes and, conversely, overstates the problems with the rest of the model.

Using the SVC instead of a standard criticism approach can also help clarify trends in where the proposed model fails. A prominent concern in scRNAseq data analysis is the common occurrence of cells that show exactly zero expression of a certain gene (Pierson and Yau, 2015; Hicks et al., 2018). We found a Spearman correlation of $\rho = 0.89$ between the conventional criticism $\log \mathcal{E}_j - \log \mathcal{E}_0$ for a gene j and the fraction of cells with zero expression of that gene j, suggesting that this is an important source of model-data mismatch in this scRNAseq data set, but not necessarily the only source (Figure 6a). However,

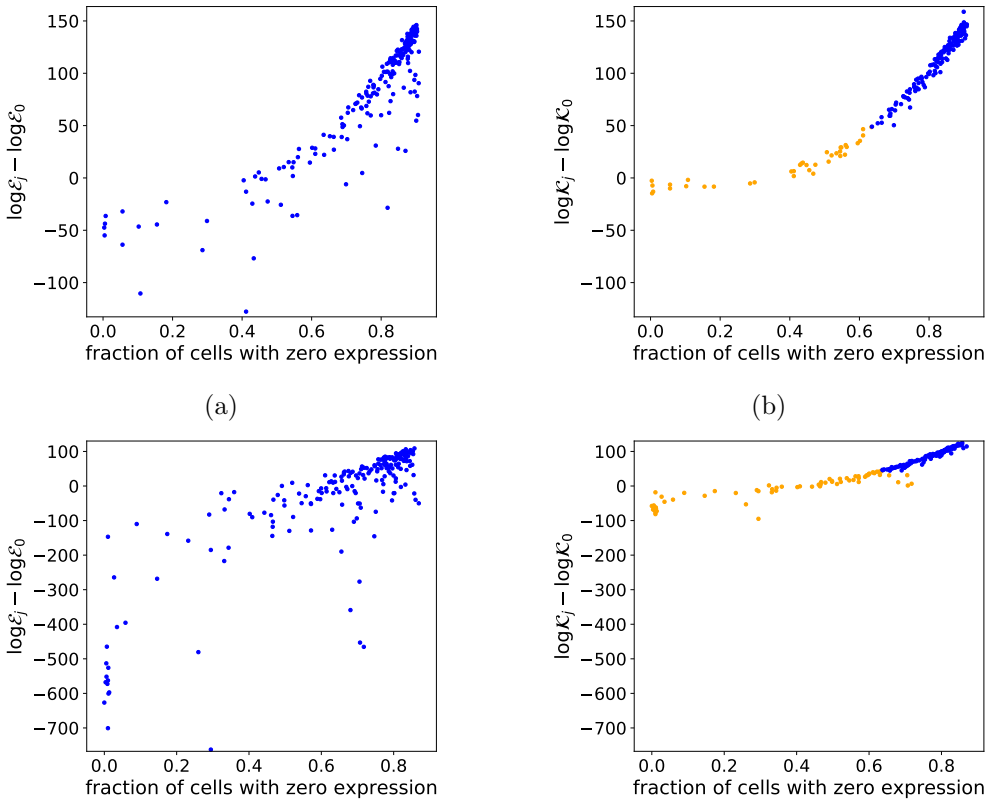


Figure 6: (a) Comparison of the conventional criticism score, for each gene j, and the fraction of cells that show zero expression of that gene j in the raw data. Spearman $\rho = 0.89$, $p < 0.01$. (b) Same as (a) but with the log SVC ratio. Spearman $\rho = 0.98$, $p < 0.01$. In orange are genes that would be included when using a background model with $c_{\mathcal{B}} = 20$ and in blue are genes that would be excluded. (c) Same as (a) for a data set taken from a MALT lymphoma (Section D.5). Spearman $\rho = 0.81$, $p < 0.01$. (d) Same as (b) for the MALT lymphoma data set. Spearman $\rho = 0.99$, $p < 0.01$.

the log SVC ratio yields a Spearman correlation of $\rho = 0.98$, suggesting instead that the amount of model-data mismatch can be entirely explained by the fraction of cells with zero expression (Figure 6b). These observations are repeatable across different scRNAseq data sets (Figure 6c, 6d).

7.3.2 Evaluating robustness

Data selection can also be used to evaluate the robustness of the foreground model to partial model misspecification. This is particularly relevant for pPCA on scRNAseq data, since the inferred latent embeddings of each cell are often used for downstream tasks such

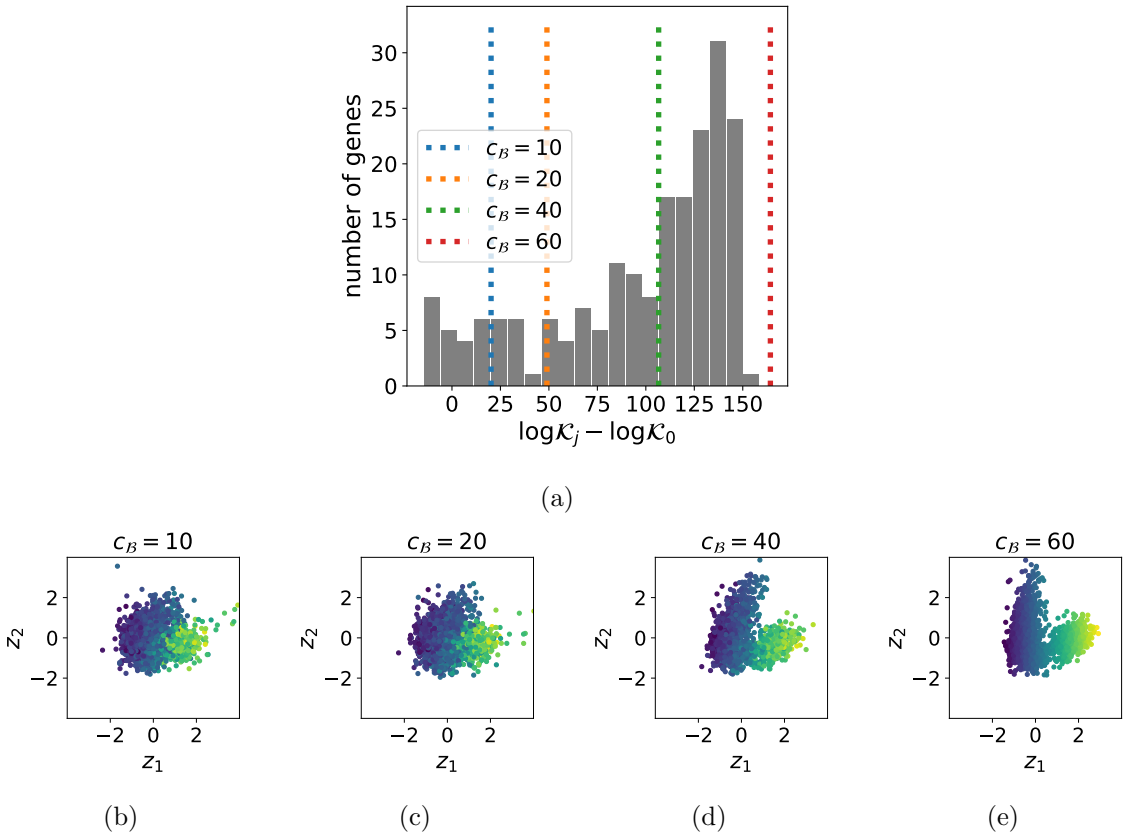
as clustering, lineage reconstruction, and so on. Misspecification may produce spurious conclusions, or alternatively, misspecification may be due to structure in the data that is scientifically interesting. To understand how partial misspecification of the pPCA model affects the latent representation of cells (and thus, downstream inferences), we performed data selection with a sequence of background model complexities $c_{\mathcal{B}}$, where $m_{\mathcal{B}} = c_{\mathcal{B}}\, r_{\mathcal{B}}$ (Figure 7a). We inferred the pPCA parameters based only on genes that the SVC selects to include in the foreground subspace. Figures 7e-7b visualize how the latent representation changes as $c_{\mathcal{B}}$ grows and fewer genes are selected. We can observe the representation morphing into a standard normal distribution, as we would expect in the case where the pPCA model is well-specified. However, the relative spatial organization of cells in the latent space remains fairly stable, suggesting that this aspect of the latent embedding is robust to partial misspecification. We can conclude that, at least in this example, misspecification strongly contributes to the non-Gaussian shape of the latent representation of the data set, but not to the distinction between subpopulations.

8. Application: Glass model of gene regulation

A central goal in the study of gene expression is to discover how individual genes regulate one another other's expression. Early studies of single cell gene expression noted the prevalence of genes that were bistable in their expression level (Shalek et al., 2013; Singer et al., 2014). This suggests a simple physical analogy: if individual gene expression is a two-state system, we might study gene regulation with the theory of interacting two-state systems, namely spin glasses. We can consider for instance a standard model of this type in which each cell i is described by a vector of spins $z_i = (z_{i1}, \ldots, z_{id})^\top$ drawn from an Ising model, specifying whether each gene $j \in \{1, \ldots, d\}$ is "on" or "off". In reality, gene expression lies on a continuum, so we use a continuous relaxation of the Ising model and parameterize each spin using a logistic function, setting $z_{ij1}(x_{ij}, \mu, \tau) = 1/(1 + \exp(-\tau(x_{ij} - \mu)))$ and $z_{ij2}(x_{ij}, \mu, \tau) = 1 - z_{ij1}(x_{ij}, \mu, \tau)$. Here, x_{ij} is the observed expression level of gene j in cell i, the unknown parameter μ controls the threshold for whether the expression of a gene is "on" (such that $z_{ij} \approx (1, 0)^\top$) or "off" (such that $z_{ij} \approx (0, 1)^\top$), and the unknown parameter $\tau > 0$ controls the sharpness of the threshold. The complete model is then given by

$$X^{(i)} \sim p(x_i|H, J, \mu, \tau) := \frac{1}{\mathcal{Z}_{H,J,\mu,\tau}} \exp\Big(\sum_j H_j^\top z_{ij}(x_{ij}, \tau, \mu) + \sum_{j'>j} z_{ij}^\top(x_{ij}, \tau, \mu) J_{jj'} z_{ij'}(x_{ij'}, \tau, \mu) \Big)$$

where $\mathcal{Z}_{H,J,\mu,\tau}$ is the unknown normalizing constant of the model, and the vectors $H_j \in \mathbb{R}^2$ and matrices $J_{jj'} \in \mathbb{R}^{2\times 2}$ are unknown parameters. This model is motivated by experimental observations and is closely related to RNAseq analysis methods that have been successfully applied in the past (Friedman et al., 2000; Friedman, 2004; Ding and Peng, 2005; Chen et al., 2015; Banerjee et al., 2008; Duvenaud et al., 2008; Liu et al., 2009; Huynh-Thu et al., 2010; Moignard et al., 2015; Matsumoto et al., 2017). However, from a biological perspective we can expect that serious problems may occur when applying the model naively to an scRNAseq data set. Genes need not exhibit bistable expression: it is straightforward in theory to write down models of gene regulation that do not have just one



(a) (b) (c) (d) (e)

Figure 7: (a) Histogram of log SVC ratios $\log \mathcal{K}_j - \log \mathcal{K}_0$ for all 200 genes in the data set (with $m_{\mathcal{B}_j} = m_{\mathcal{F}_0} - m_{\mathcal{F}_j}$). Dotted lines show the value of the volume correction term in the SVC for different choices of background model complexity $c_\mathcal{B}$; for each choice, genes with $\log \mathcal{K}_j - \log \mathcal{K}_0$ values above the dotted line would be excluded from the foreground subspace based on the SVC. (b) Posterior mean of the first two latent variables (z_1 and z_2), with the pPCA model applied to the genes selected with a background model complexity of $c_\mathcal{B} = 10$ (keeping 23 genes in the foreground). (c-e) Same as (b), but with $c_\mathcal{B} = 20$ (keeping 38 genes), $c_\mathcal{B} = 40$ (keeping 87 genes) and $c_\mathcal{B} = 60$ (keeping all 200 genes). In (a)-(d), the points are colored using the z_1 value when $c_\mathcal{B} = 60$.

or two steady states—gene expression may fall on a continuum, or oscillate, or have three stable states—and many alternative patterns have been well-documented empirically (Alon, 2019). Interactions between genes may also be more complex than the model assumes, involving for instance three-way dependencies between genes. All of these biological concerns can potentially produce severe violations of the proposed two-state glass model's assumptions. Data selection provides a method for discovering where the proposed model applies.

Applying standard Bayesian inference to the glass model is intractable, since the normalizing constant is unknown (it is an energy-based model). However, the normalizing

constant does not affect the SVC, so we can still perform data selection. We used the variational approximation to the SVC in Section 2.3.4. We placed a Gaussian prior on H and a Laplace prior on each entry of J to encourage sparsity in the pairwise gene interactions; we also used Gaussian priors for μ and τ after applying an appropriate transform to remove constraints (Section E.1). Following the logic of stochastic variational inference, we optimized the SVC variational approximation using minibatches of the data and a reparameterization gradient estimator (Hoffman et al., 2013; Kingma and Welling, 2014; Kucukelbir et al., 2017). We also simultaneously stochastically optimized the set of genes included in the foreground subspace, using Leave-One-Out REINFORCE (Kool et al., 2019; Dimitriev and Zhou, 2021) to estimate log-odds gradients. We implemented the model and inference strategy within the probabilistic programming language Pyro by defining a new distribution with log probability given by the negative NKSD (Bingham et al., 2019). Pyro provides automated, GPU-accelerated stochastic variational inference, requiring less than an hour for inference on data sets with thousands of cells. See Section E.1 for more details on these inference procedures.

We examined three scRNAseq data sets, taken from (i) peripheral blood monocytes (PBMCs) from a healthy donor (2,428 cells), (ii) a MALT lymphoma (7,570 cells), and (iii) mouse neurons (10,658 cells) (Section E.2). We preprocessed the data following standard protocols and focused on 200 high expression, high variability genes in each data set, based on the metric of Gigante et al. (2020). We set $T = 0.05$ as in Section 7, and used the Pitman-Yor expression for $m_{\mathcal{B}}$ (Equation 3) with $\alpha = 0.5$, $\nu = 1$ and $D = 100$. This value of D ensures that the number of background model parameters per data dimension is larger than the number of foreground model parameters per data dimension except for at very small N; in particular, there are 798 foreground model parameter dimensions associated with each data dimension (from the 199 interactions $J_{jj'}$ that each gene has with each other gene, plus the contribution of H_j), and $m_{\mathcal{B}} > 798\, r_{\mathcal{B}}$ for $N \geq 13$. Our data selection procedure selects 65 genes (32.5%) in the PBMC data set, 0 genes in the neuron data set, and 187 genes (93.5%) in the MALT data set; note that for a lower value of $m_{\mathcal{B}}$, in particular using $D = 10$, no genes are selected in the MALT data set. These results suggest substantial partial misspecification in the PBMC and neuron data sets, and more moderate partial misspecification in the MALT data set.

We investigated the biological information captured by the foreground model on the MALT data set. In particular, we looked at the approximate NKSD posterior for the selected 187 genes, and compared it to the approximate NKSD posterior for the model when applied to all 200 genes. (Note that, since the glass model lacks a tractable normalizing constant, we cannot compare standard Bayesian posteriors.) Figure 8 shows, for a subset of selected genes, the posterior mean of the interaction energy $\Delta E_{jj'} := J_{jj'21} + J_{jj'12} - J_{jj'22} - J_{jj'11}$, that is, the total difference in energy between two genes being in the same state versus in opposite states. We focused on strong interactions with $|\Delta E_{jj'}| > 1$, corresponding to just 5% of all possible gene-gene interactions (Figure 12).

One foreground gene with especially large loading onto the top principal component of the ΔE matrix is CD37 (Figure 8). In B-cell lymphomas, of which MALT lymphoma is an example, CD37 loss is known to be associated with decreased patient survival (Xu-Monette et al., 2016). Further, previous studies have observed that CD37 loss leads to high NF-κB pathway activation (Xu-Monette et al., 2016). Consistent with this observation,

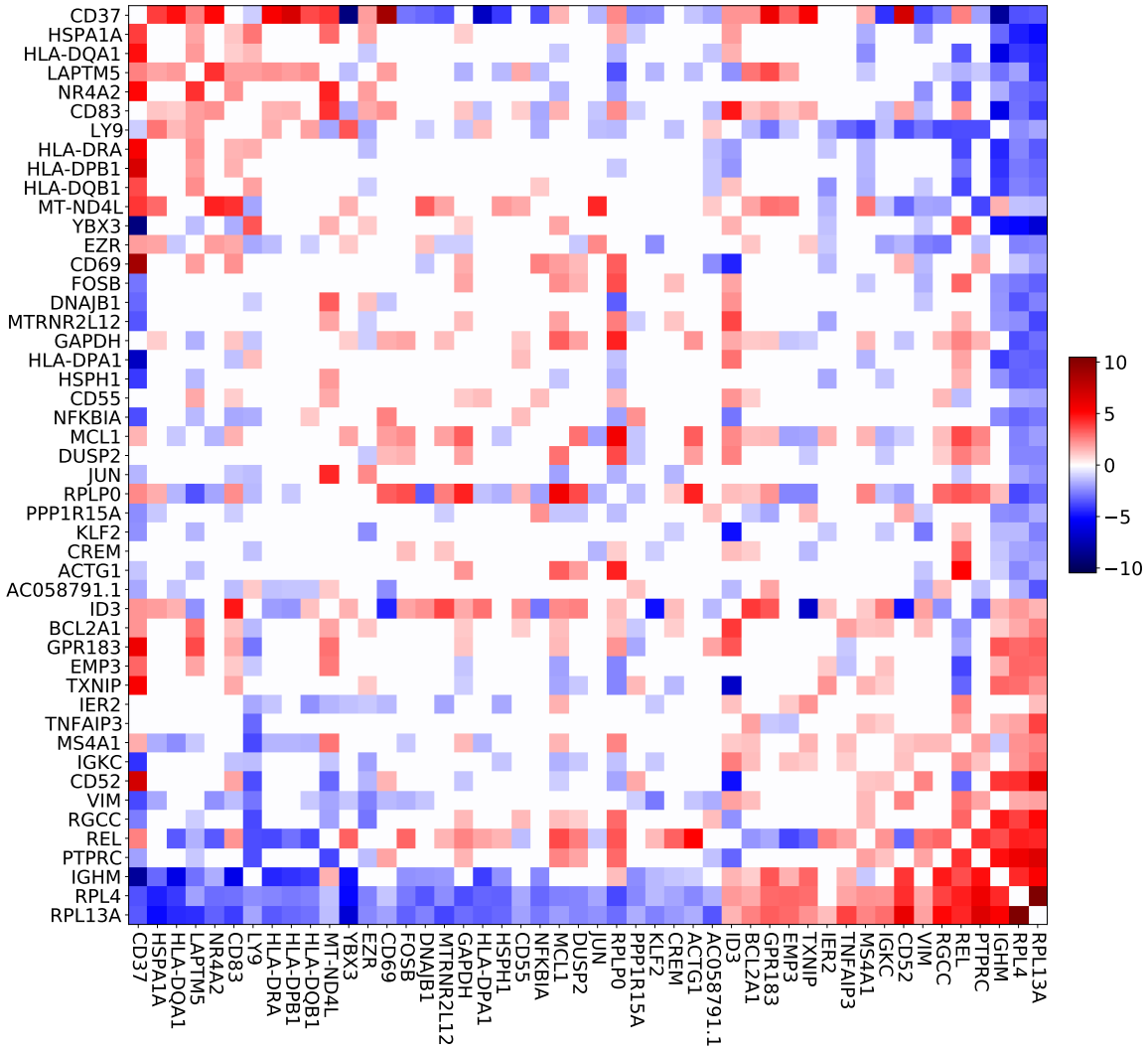


Figure 8: Posterior mean interaction energies $\Delta E_{jj'} := J_{jj'21} + J_{jj'12} - J_{jj'22} - J_{jj'11}$ for a subset of the selected genes. For visualization purposes, weak interactions ($|\Delta E_{jj'}| \leq 1$) are set to zero, and genes with less than 10 total strong connections are not shown. Genes are sorted based on their (signed) projection onto the top principal component of the ΔE matrix.

the estimated interaction energies in our model suggest that decreasing CD37 will lead to higher expression of REL, an NF-κB transcription factor ($\Delta E_{\text{CD37,REL}} = 2.5$), decreased expression of NKFBIA, an NF-κB inhibitor ($\Delta E_{\text{CD37,NKFBIA}} = -3.6$), and higher expression of BCL2A1, a downstream target of the NF-κB pathway ($\Delta E_{\text{CD37,BCL2A1}} = 2.1$). Separately, a knockout study of Cd37 in B-cell lymphoma in mice does not show IgM expression (de Winde et al., 2016), consistent with our model ($\Delta E_{\text{CD37,IGHM}} = -8.2$). The same study does show MHC-II expression, and our model predicts the same result, for

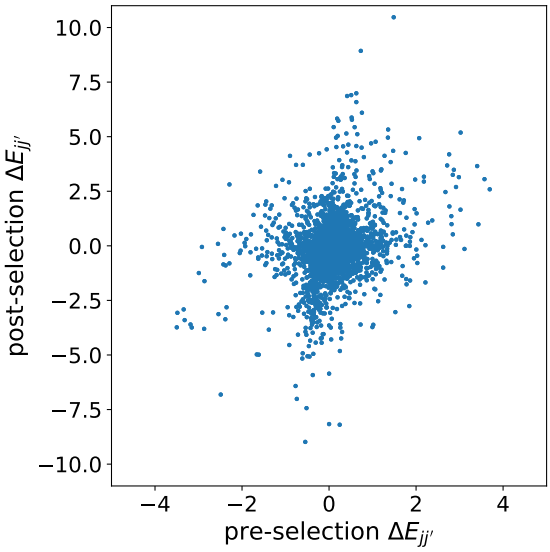


Figure 9: Comparison of posterior mean interaction energies $\Delta E_{jj'}$ for a model applied to all 200 genes (pre-data selection) to those learned from a model applied to the selected foreground subspace (post-data selection). Each point corresponds to a pairwise interaction between two of the selected 187 genes.

HLA-DQ in particular ($\Delta E_{\text{CD37,HLA-DQA1}} = 5.0$, $\Delta E_{\text{CD37,HLA-DQB1}} = 3.7$). These results suggest that the data selection procedure can successfully find systems of interacting genes that can plausibly be modeled as a spin glass, and which, in this case, are relevant for cancer.

To investigate whether data selection provided a benefit in this analysis, we compare with the results obtained by applying the foreground model to the full data set of all 200 genes. All but one of the interactions listed above have $|\Delta E| < 1$ in the full foreground model, and three have opposite signs ($\Delta E_{\text{CD37,NFKBIA}} = +0.7$, $\Delta E_{\text{CD37,IGHM}} = +0.0$, $\Delta E_{\text{CD37,HLA-DQB1}} = -0.6$); see Figure 13. Across all 187 selected genes, we find only a moderate correlation between the interaction energies estimated when using the full foreground model compared with the data selection-based model (Spearman's rho = 0.30, $p < 0.01$; Figure 9). These results show that using data selection can lead to substantially different, and arguably more biologically plausible, downstream conclusions as compared to naive application of the foreground model to the full data set.

As a simple alternative, one might wonder whether genes that are poorly fit by the model could be identified simply by looking their posterior uncertainty under the full foreground model. This simple approach does not work well, however, since it is possible for parameters to have low uncertainty even when the model poorly describes the data. Indeed, we found that examining uncertainty in the glass model does not lead to the same conclusions as performing data selection: the genes excluded by our data selection procedure are not the ones with the highest uncertainty in their interactions (as measured by the mean posterior standard deviation of $\Delta E_{jj'}$ under the NKSD posterior), though they do have above average

uncertainty (Figure 14a). Instead, the genes excluded by our data selection procedure are the ones with the highest fraction of cells with zero expression, violating the assumptions of the foreground model (Figure 14b). These results show how data selection provides a sound, computationally tractable approach to criticizing and evaluating complex Bayesian models.

9. Discussion

Statistical modeling is often described as an iterative process, where we design models, infer hidden parameters, critique model performance, and then use what we have learned from the critique to design new models and repeat the process (Gelman et al., 2013). This process has been called "Box's loop" (Blei, 2014). From one perspective, data selection offers a new criticism approach. It goes beyond posterior predictive checks and related methods by changing the model itself, replacing potentially misspecified components with a flexible background model. This has important practical consequences: since misspecification can distort estimates of model parameters in unpredictable ways, predictive checks are likely to indicate mismatch between the model and the data across the entire space $\mathcal{X}$ even when the proposed parametric model is only partially misspecified. Our method, by contrast, reveals precisely those subspaces of $\mathcal{X}$ where model-data mismatch occurs.

From another perspective, data selection is outside the design-infer-critique loop. An underlying assumption of Box's loop is that scientists want to model the entire data set. As data sets get larger, and measurements get more extensive, this desire has led to more and more complex (and often difficult to interpret) models. In experimental science, however, scientists have often followed the opposite trajectory: faced with a complicated natural phenomenon, they attempt to isolate a simpler example of the phenomenon for close study. Data selection offers one approach to formalizing this intuitive idea in the context of statistical analysis: we can propose a simple parametric model and then isolate a piece of the whole data set—a subspace $\mathcal{X}_\mathcal{F}$—to which this model applies. When working with large, complicated data sets, this provides a method of searching for simpler phenomena that are hypothesized to exist.

There are several directions for future work and improvement upon our proposed data selection approach. First, we have focused in our applied examples on discovering subsets of data dimensions. However, our theoretical results show that one can perform data selection on linear subspaces in general; for instance, in the context of scRNAseq, we might find that a model can describe a certain set of linear gene expression programs. Even more generally, one might be interested in discovering nonlinear features of the data that the model can explain—such as a set of nonlinear gene expression programs—and this would require extending our approach, perhaps by (1) applying a nonlinear volume-preserving map to the data, and then (2) performing standard linear data selection.

Second, we have focused on choosing one best $\mathcal{X}_\mathcal{F}$ from among a finite set of possibilities. A future direction is to provide rigorous asymptotic guarantees when there are infinitely many possible choices of $\mathcal{X}_\mathcal{F}$, such as the set of all linear subspaces of $\mathcal{X}$. Another future direction is to provide uncertainty quantification of $\mathcal{X}_\mathcal{F}$, rather than just point estimation. Here, it is important to consider the uncertainty due to having finite data as well as non-identifiability, since there may exist multiple optimal values of $\mathcal{X}_\mathcal{F}$; for instance, this

can occur if the model is well specified over marginals of the data but not over the joint distribution of the data.

Third, in many applications, researchers will be interested in inferring the parameters θ of the foreground model when applied to the selected subspace $\mathcal{X}_{\mathcal{F}}$. On finite data, it is conceivable that foreground subspaces $\mathcal{X}_{\mathcal{F}}$ that are more likely to be selected are also more likely to have certain values of θ, which could create a "post- data selection bias" in conclusions about θ, analogous to the bias that occurs in post-selection inference (Yekutieli, 2012). The data selection problem does not fit neatly in the framework of post-selection inference, however, so further investigation will be required to understand if, when, and to what extent such bias occurs.

Finally, in comparison to the augmented model marginal likelihood, the SVC makes different judgments as to what types of model-data mismatch are important. The NKSD and the KL divergence are quite different and do not, in general, coincide or tightly bound one another, so a model-data mismatch that looks big to one divergence may not look big to the other, and vice versa (Matsubara et al., 2022). The preference of the NKSD for certain types of errors is not essential to achieving consistent data selection and nested data selection, but is very relevant to the practical use and interpretation of the SVC. One could use another divergence instead of the NKSD in the definition of the SVC, and this would typically be expected to yield consistent model selection and nested model selection (Appendix B.1 and Miller, 2021), however, consistent data selection and nested data selection are more challenging, and depend on a combination of special properties that our NKSD estimator possesses (Section 3). Developing data selection approaches with different model-data mismatch preferences, therefore, remains an open challenge. In summary, Bayesian data selection is a rich area for future work.

Acknowledgments

The authors wish to thank Jonathan Huggins, Pierre Jacob, Andre Nguyen, Elizabeth Wood, and the anonymous reviewers for helpful discussions and suggestions. We would like to thank Debora S. Marks in particular for suggesting the use of a Potts model in RNAseq analysis. E.N.W. was supported by the Fannie and John Hertz Fellowship. J.W.M. is supported by the National Institutes of Health grant 5R01CA240299-02.

Appendix A. Methods details

A.1 Calibrating T

The SVC contains a hyperparameter $T > 0$. To choose an appropriate value of T, we aim, roughly, to match the coverage of the generalized posterior

$$\pi_N^{\mathrm{SVC}}(\theta)d\theta = \frac{1}{z_N}\exp\Big(-\frac{N}{T}\widehat{\mathrm{NKSD}}(p_0(x))\|q(x|\theta))\Big)\pi(\theta)d\theta$$

to the coverage of the standard Bayesian posterior

$$\pi_N^{\mathrm{KL}}(\theta)d\theta = \frac{1}{q(X^{(1:N)})}\exp\Big(\sum_{i=1}^{N}\log q(X^{(i)}|\theta)\Big)\pi(\theta)d\theta$$

when the model is well-specified.

Let θ_* be the true parameter value, such that $p_0(x) = q(x|\theta_*)$ almost everywhere. Let $G^{\mathrm{KL}}(\theta) := \nabla_\theta^2\mathbb{E}_{X\sim p_0}[-\log q(X|\theta)]$ and let $\theta_N^{\mathrm{KL}} := \operatorname{argmax}\sum_{i=1}^{N}\log q(X^{(i)}|\theta)$ be the maximum likelihood estimator. Let h_N^{KL} be the density of $\sqrt{N}(\theta-\theta_N^{\mathrm{KL}})$ when $\theta\sim\pi_N^{\mathrm{KL}}$. Under regularity conditions (Miller, 2021), according to the Bernstein–von Mises theorem, h_N^{KL} converges to a normal distribution in total variation,

$$\int_{\mathbb{R}^m}\Big|h_N^{\mathrm{KL}}(x) - \mathcal{N}\big(x\mid 0, G^{\mathrm{KL}}(\theta_*)^{-1}\big)\Big|dx \xrightarrow[N\to\infty]{\text{a.s.}} 0.$$

According to Theorem 9, the generalized posterior associated with the SVC has analogous behavior. Let $G^{\mathrm{SVC}}(\theta) := \nabla_\theta^2\frac{1}{T}\mathrm{NKSD}(p_0(x)\|q(x|\theta))$ and let $\theta_N^{\mathrm{SVC}} := \operatorname{argmin}\widehat{\mathrm{NKSD}}(p_0(x)\|q(x\mid\theta))$. Let h_N^{SVC} be the density of $\sqrt{N}(\theta-\theta_N^{\mathrm{SVC}})$ when $\theta\sim\pi_N^{\mathrm{SVC}}$. Then by Theorem 9, h_N^{SVC} converges to a normal distribution in total variation,

$$\int_{\mathbb{R}^m}\Big|h_N^{\mathrm{SVC}}(x) - \mathcal{N}\big(x\mid 0, G^{\mathrm{SVC}}(\theta_*)^{-1}\big)\Big|dx \xrightarrow[N\to\infty]{\text{a.s.}} 0.$$

For the uncertainty in each posterior to be roughly the same order of magnitude, we want

$$\det G^{\mathrm{KL}}(\theta_*) \approx \det G^{\mathrm{SVC}}(\theta_*),$$

or equivalently,

$$T \approx \left(\frac{\det\left[\nabla_\theta^2\big|_{\theta=\theta_*}\mathrm{NKSD}(p_0(x)\|q(x|\theta))\right]}{\det\left[\nabla_\theta^2\big|_{\theta=\theta_*}\mathbb{E}_{X\sim p_0}[-\log q(X|\theta)]\right]}\right)^{1/m}.$$

To choose a single T value, we simulate true parameters from the prior, generate data from each simulated true parameter, and take the median of the estimated T values. That is, we use the median $\hat{T}$ across samples drawn as

$$\begin{aligned}\theta_* &\sim \pi(\theta)\\ X^{(i)} &\overset{\text{iid}}{\sim} q(x|\theta_*)\\ \hat{T} &= \left(\frac{|\det\left[\nabla_\theta^2\big|_{\theta=\theta_*}\widehat{\mathrm{NKSD}}(p_0(x)\|q(x|\theta))\right]|}{|\det\left[\nabla_\theta^2\big|_{\theta=\theta_*}\frac{1}{N}\sum_{i=1}^{N}-\log q(X^{(i)}|\theta)\right]|}\right)^{1/m}.\end{aligned} \tag{55}$$

In practice, we find that the order of magnitude of $\hat{T}$ is stable across samples θ_* from $\pi(\theta)$. See Section D.3 for an example.

A.2 Kernel recommendations

To obtain subsystem independence (Proposition 6), we suggest using a kernel that factors across subspaces, such that $k(X, Y) = k_{\mathcal{F}}(X_{\mathcal{F}}, Y_{\mathcal{F}})k_{\mathcal{B}}(X_{\mathcal{B}}, Y_{\mathcal{B}})$ where $k_{\mathcal{F}}$ and $k_{\mathcal{B}}$ are integrally strictly positive definite kernels. In the applications in Sections 7 and 8, we use the following kernel.

Definition 18 *The* factored inverse multiquadric (IMQ) kernel *is defined as*

$$k(x, y) = \prod_{i=1}^{d} \left(c^2 + (x_i - y_i)^2\right)^{\beta/d}$$

for $x, y \in \mathbb{R}^d$, where $\beta \in [-1/2, 0)$ and $c > 0$.

Note that this kernel factors across any subset of dimensions, that is, if $S \subseteq \{1, \ldots, d\}$ and $S^c = \{1, \ldots, d\} \setminus S$, then we can write $k(x, y) = k_S(x_S, y_S)k_{S^c}(x_{S^c}, y_{S^c})$. Thus, if the foreground subspace $\mathcal{X}_{\mathcal{F}}$ is the span of a subset of the standard basis, such that $x_{\mathcal{F}} = V^\top x = x_S$ for some $S \subseteq \{1, \ldots, d\}$, then it follows that k factors as $k(x, y) = k_{\mathcal{F}}(x_{\mathcal{F}}, y_{\mathcal{F}})k_{\mathcal{B}}(x_{\mathcal{B}}, y_{\mathcal{B}})$. Along with this observation, the next result shows that the factored IMQ satisfies the conditions of Theorem 9 that pertain to k alone.

Proposition 19 *The factored IMQ kernel is symmetric, positive, bounded, integrally strictly positive definite, and has continuous and bounded partial derivatives up to order 2.*

Proof It is clear that $k(x, y) = k(y, x)$ and $k(x, y) > 0$. Next, we show that k has continuous and bounded partial derivatives up to order 2. Note that we can write $k(x, y) = \prod_{i=1}^{d} \psi(x_i - y_i)$ where $\psi(r) = (c^2 + r^2)^{\beta/d}$ for $r \in \mathbb{R}$. Differentiating, we have

$$\begin{aligned}
\psi'(r) &= \frac{\beta}{d} \frac{2r}{c^2 + r^2} \psi(r) \\
\psi''(r) &= \left(\frac{\beta^2}{d^2} - \frac{\beta}{d}\right) \left(\frac{2r}{c^2 + r^2}\right)^2 \psi(r) + \frac{\beta}{d} \frac{2}{c^2 + r^2} \psi(r).
\end{aligned}$$

Since $r^2 \geq 0$ and $\beta < 0$, $|\psi(r)| \leq c^{2\beta/d}$ for all $r \in \mathbb{R}$. Further, it is straightforward to verify that $|\psi'(r)|$ and $|\psi''(r)|$ are bounded on $\mathbb{R}$ by using the fact that $|r|/(c^2 + r^2) \leq 1/(2c)$. By the chain rule, it follows that for all i, j, the functions $k(x, y)$, $|\partial k/\partial x_i|$, and $|\partial^2 k/\partial x_i \partial y_j|$ are bounded. Thus, we conclude that k, $\|\nabla k\|$, and $\|\nabla^2 k\|$ are bounded.

Finally, we show that k is integrally strictly positive definite. First, for any d, for $x, y \in \mathbb{R}^d$, the function $(x, y) \mapsto (c^2 + \|x - y\|_2^2)^{\beta/d}$ is an integrally strictly positive definite kernel (see, for example, Section 3.1 of Sriperumbudur et al., 2010); we refer to this as the standard IMQ kernel. Since the factored IMQ is a product of one-dimensional standard IMQ kernels, it defines a kernel on $\mathbb{R}^d$ (Lemma 4.6 of Steinwart and Christmann, 2008) and is positive definite (Theorem 4.16 of Steinwart and Christmann, 2008). By Bochner's theorem (Theorem 3 of Sriperumbudur et al., 2010), a continuous positive definite kernel can be expressed in terms of the Fourier transform of a finite nonnegative Borel measure.

In particular, applying Bochner's theorem to $\psi(r)$, we have

$$k(x,y) = \Psi(x-y) := \prod_{i=1}^{d} \psi(x_i - y_i) = \prod_{i=1}^{d} \int_{\mathbb{R}} \exp\big(-\sqrt{-1}(x_i - y_i)\omega_i\big) d\Lambda^0(\omega_i)$$
$$= \int_{\mathbb{R}^d} \exp\big(-\sqrt{-1}(x-y)^\top \omega\big) d\Lambda(\omega)$$

by Fubini's theorem, where Λ^0 is the finite nonnegative Borel measure on $\mathbb{R}$ associated with $\psi(r)$ and $\Lambda = \Lambda^0 \times \cdots \times \Lambda^0$ is the resulting product measure on $\mathbb{R}^d$. Applying Bochner's theorem in the other direction, we see that Ψ is a positive definite function. Moreover, since the standard IMQ kernel is characteristic (Theorem 7 of Sriperumbudur et al., 2010), it follows that the support of Λ^0 is $\mathbb{R}$ (Theorem 9 of Sriperumbudur et al., 2010), and thus the support of Λ is $\mathbb{R}^d$. This implies that the factored IMQ kernel k is characteristic (Theorem 9 of Sriperumbudur et al., 2010) and, since k is also translation invariant, k must be integrally strictly positive definite (Section 3.4 of Sriperumbudur et al., 2011). ■

Our choice of the factored IMQ kernel is motivated by the analysis of Gorham and Mackey (2017), which suggests that the standard IMQ is a good default choice for the kernelized Stein discrepancy, particularly when working with distributions that are (roughly speaking) very spread out. In particular, it is straightforward to show that the factored IMQ kernel, like the standard IMQ kernel, meets the conditions of Theorem 3.2 of Huggins and Mackey (2018). However, we do not pursue further the question of whether the NKSD with the factored IMQ detects convergence and non-convergence since our statistical setting is different from that of Gorham and Mackey (2017), and we are assuming the data consists of i.i.d. samples from some underlying distribution rather than correlated samples from an MCMC chain which may or may not converge.

A.3 Exact solution for exponential families

Here, we show that when $q(x|\theta)$ is an exponential family, the estimated NKSD has the form

$$\widehat{\text{NKSD}}(p_0(x) \| q(x|\theta)) = \theta^\top A\, \theta + B^\top \theta + C \tag{56}$$

where A, B, and C depend on the data but not on θ. Since $q_\theta(x) = q(x|\theta) = \lambda(x)\exp(\theta^\top t(x) - \kappa(\theta))$, we have $s_{q_\theta}(x) = \nabla_x \log \lambda(x) + (\nabla_x t(x))^\top \theta$ where $(\nabla_x t(x))_{ij} = \partial t_i / \partial x_j$. Thus, we can write

$$\begin{aligned}
&u_\theta(x,y) \qquad\qquad (57)\\
&:= s_{q_\theta}(x)^\top s_{q_\theta}(y) k(x,y) + s_{q_\theta}(x)^\top \nabla_y k(x,y) + s_{q_\theta}(y)^\top \nabla_x k(x,y) + \text{trace}(\nabla_x \nabla_y^\top k(x,y))\\
&= \theta^\top [(\nabla_x t(x))(\nabla_y t(y))^\top k(x,y)]\theta\\
&\quad + [(\nabla_x \log \lambda(x))^\top (\nabla_y t(y))^\top k(x,y) + (\nabla_y \log \lambda(y))^\top (\nabla_x t(x))^\top k(x,y)\\
&\qquad + (\nabla_x k(x,y))^\top (\nabla_y t(y))^\top + (\nabla_y k(x,y))^\top (\nabla_x t(x))^\top]\theta\\
&\quad + [(\nabla_x \log \lambda(x))^\top (\nabla_y \log \lambda(y)) k(x,y) + (\nabla_y \log \lambda(y))^\top (\nabla_x k(x,y))\\
&\qquad + (\nabla_x \log \lambda(x))^\top (\nabla_y k(x,y)) + \text{trace}(\nabla_x \nabla_y^\top k(x,y))]. \qquad (58)
\end{aligned}$$

Then the estimated NKSD takes the form in Equation 56 if we choose

$$
\begin{aligned}
A &:= \frac{1}{\sum_{i\neq j} k(X^{(i)}, X^{(j)})} \sum_{i\neq j} \nabla_x t(X^{(i)}) \nabla_x t(X^{(j)})^\top k(X^{(i)}, X^{(j)}) \\
B^\top &:= \frac{1}{\sum_{i\neq j} k(X^{(i)}, X^{(j)})} \sum_{i\neq j} \big[(\nabla_x \log \lambda(X^{(i)}))^\top \nabla_x t(X^{(j)})^\top k(X^{(i)}, X^{(j)}) \\
&\qquad + (\nabla_x \log \lambda(X^{(j)}))^\top \nabla_x t(X^{(i)})^\top k(X^{(i)}, X^{(j)}) \\
&\qquad + (\nabla_x k(X^{(i)}, X^{(j)}))^\top \nabla_x t(X^{(j)})^\top + (\nabla_y k(X^{(i)}, X^{(j)}))^\top \nabla_x t(X^{(i)})^\top \big] \\
C &:= \frac{1}{\sum_{i\neq j} k(X^{(i)}, X^{(j)})} \sum_{i\neq j} \big[(\nabla_x \log \lambda(X^{(i)}))^\top (\nabla_x \log \lambda(X^{(j)})) k(X^{(i)}, X^{(j)}) \\
&\qquad + (\nabla_x \log \lambda(X^{(j)}))^\top \nabla_x k(X^{(i)}, X^{(j)}) \\
&\qquad + (\nabla_x \log \lambda(X^{(i)}))^\top \nabla_y k(X^{(i)}, X^{(j)}) + \operatorname{trace}(\nabla_x \nabla_y^\top k(X^{(i)}, X^{(j)})) \big].
\end{aligned}
$$

If the prior on θ is $\mathcal{N}(\mu, \Sigma_0)$, then the SVC is

$$
\begin{aligned}
\mathcal{K} &= \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} (2\pi)^{-m_{\mathcal{F}}/2} (\det \Sigma_0)^{-1/2} \\
&\quad \times \int \exp\Big(-\frac{N}{T}[\theta^\top A\, \theta + B^\top \theta + C]\Big) \exp\Big(-\frac{1}{2}(\theta-\mu)^\top \Sigma_0^{-1}(\theta-\mu)\Big) d\theta \\
&= \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} (2\pi)^{-m_{\mathcal{F}}/2} (\det \Sigma_0)^{-1/2} \\
&\quad \times \int \exp\Big(-\frac{1}{2}\theta^\top \Big(\frac{2N}{T}A + \Sigma_0^{-1}\Big)\theta + \Big(-\frac{N}{T}B^\top + \mu^\top \Sigma_0^{-1}\Big)\theta - \frac{N}{T}C - \frac{1}{2}\mu^\top \Sigma_0^{-1}\mu \Big) d\theta \\
&= \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} (\det \Sigma_0)^{-1/2} \Big(\det\Big(\frac{2N}{T}A + \Sigma_0^{-1}\Big)\Big)^{-1/2} \\
&\times \exp\Big(\frac{1}{2}\Big(-\frac{N}{T}B^\top + \mu^\top \Sigma_0^{-1}\Big)^\top \Big(\frac{2N}{T}A + \Sigma_0^{-1}\Big)^{-1} \Big(-\frac{N}{T}B^\top + \mu^\top \Sigma_0^{-1}\Big) - \frac{N}{T}C - \frac{1}{2}\mu^\top \Sigma_0^{-1}\mu \Big).
\end{aligned}
$$

Meanwhile, if $q(x|\theta) = \mathcal{N}(\theta, \Sigma)$ where Σ is a fixed covariance matrix, then we have $\nabla_x \log \lambda(x) = -\Sigma^{-1}x$ and $\nabla_x t(x) = \Sigma^{-1}$.

A.4 Comparing many foregrounds using approximate optima

Here, we justify the technique described in Section 2.3.3. As in Section 2.3.3, define $\ell_j(\theta) = \widehat{\mathrm{NKSD}}(p_0(x_{\mathcal{F}_j}) \| q(x_{\mathcal{F}_j}|\theta))$ for $j \in \{1, 2\}$, and let $\theta_N(w) = \operatorname{argmin}_\theta \mathcal{L}(w, \theta)$ where

$$
\mathcal{L}(w, \theta) := \ell_1(\theta) + w(\ell_2(\theta) - \ell_1(\theta))
$$

for $w \in [0, 1]$. We assume that the conditions of Theorem 9 are met, over both $\mathcal{X}_{\mathcal{F}_1}$ and $\mathcal{X}_{\mathcal{F}_2}$. Since $(\partial\mathcal{L}/\partial\theta_i)(w, \theta_N(w)) = 0$, we have

$$
0 = \frac{\partial}{\partial w}\Big(\frac{\partial \mathcal{L}}{\partial \theta_i}(w, \theta_N(w))\Big) = \frac{\partial^2 \mathcal{L}}{\partial w \partial \theta_i}(w, \theta_N(w)) + \sum_j \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j}(w, \theta_N(w)) \Big(\frac{\partial}{\partial w}\theta_{N,j}(w)\Big),
$$

or equivalently, in matrix/vector notation,

$$0 = \nabla_w(\nabla_\theta \mathcal{L}(w, \theta_N(w))) = \nabla_\theta \nabla_w \mathcal{L}(w, \theta_N) + \nabla_\theta^2 \mathcal{L}(w, \theta_N) \nabla_w(\theta_N(w)).$$

Rearranging, we have

$$\nabla_w \theta_N(w) = -\big(\nabla_\theta^2 \mathcal{L}(w, \theta_N)\big)^{-1} \nabla_\theta \nabla_w \mathcal{L}(w, \theta_N).$$

At $w = 0$ we find, plugging back in the definition of $\mathcal{L}$,

$$\begin{aligned}\nabla_w \theta_N(0) &= -\nabla_\theta^2 \ell_1(\theta_N(0))^{-1}(\nabla_\theta \ell_2(\theta_N(0)) - \nabla_\theta \ell_1(\theta_N(0))) \\ &= -\nabla_\theta^2 \ell_1(\theta_N(0))^{-1} \nabla_\theta \ell_2(\theta_N(0)).\end{aligned}$$

Applying a first-order Taylor series expansion gives us $\theta_N(1) \approx \theta_N(0) + \nabla_w \theta_N(0)$, which yields Equation 13.

Appendix B. Asymptotics of the alternative selection criteria

Theorem 17 shows that the SVC exhibits all four types of consistency: data selection, nested data selection, model selection, and nested model selection. In this section, we establish the consistency properties of the alternative criteria considered in Section 3.

B.1 Setup

We first review the asymptotics of the standard marginal likelihood, discussed in depth by Dawid (2011) and Hong and Preston (2005), for example. Define

$$\begin{aligned} f_N^{\mathrm{KL}}(\theta) &:= -\frac{1}{N}\sum_{i=1}^N \log q(X^{(i)}|\theta), & \theta_N^{\mathrm{KL}} &:= \operatorname*{argmin}_\theta f_N^{\mathrm{KL}}(\theta), \\ f^{\mathrm{KL}}(\theta) &:= -\mathbb{E}_{X\sim p_0}[\log q(X|\theta)], & \theta_*^{\mathrm{KL}} &:= \operatorname*{argmin}_\theta f^{\mathrm{KL}}(\theta). \end{aligned}$$

Let m be the dimension of the parameter space. Under suitable regularity conditions (Miller, 2021), the Laplace approximation to the marginal likelihood is

$$q(X^{(1:N)}) = \int q(X^{(1:N)}|\theta)\pi(\theta)d\theta \sim \frac{\exp\big(-N f_N^{\mathrm{KL}}(\theta_N^{\mathrm{KL}})\big)\pi(\theta_*^{\mathrm{KL}})}{\big|\det \nabla_\theta^2 f^{\mathrm{KL}}(\theta_*^{\mathrm{KL}})\big|^{1/2}} \left(\frac{2\pi}{N}\right)^{m/2} \tag{59}$$

almost surely, where $a_N \sim b_N$ indicates that $a_N/b_N \to 1$ as $N \to \infty$. We can rewrite this as

$$\begin{aligned} \log q(X^{(1:N)}) &+ N(f_N^{\mathrm{KL}}(\theta_N^{\mathrm{KL}}) - f_N^{\mathrm{KL}}(\theta_*^{\mathrm{KL}})) \\ &+ N(f_N^{\mathrm{KL}}(\theta_*^{\mathrm{KL}}) - f^{\mathrm{KL}}(\theta_*^{\mathrm{KL}})) + N f^{\mathrm{KL}}(\theta_*^{\mathrm{KL}}) \\ &+ \frac{m}{2}\log N - \log\left(\frac{\pi(\theta_*^{\mathrm{KL}})(2\pi)^{m/2}}{\big|\det \nabla_\theta^2 f^{\mathrm{KL}}(\theta_*^{\mathrm{KL}})\big|^{1/2}}\right) \xrightarrow[N\to\infty]{\text{a.s.}} 0. \end{aligned} \tag{60}$$

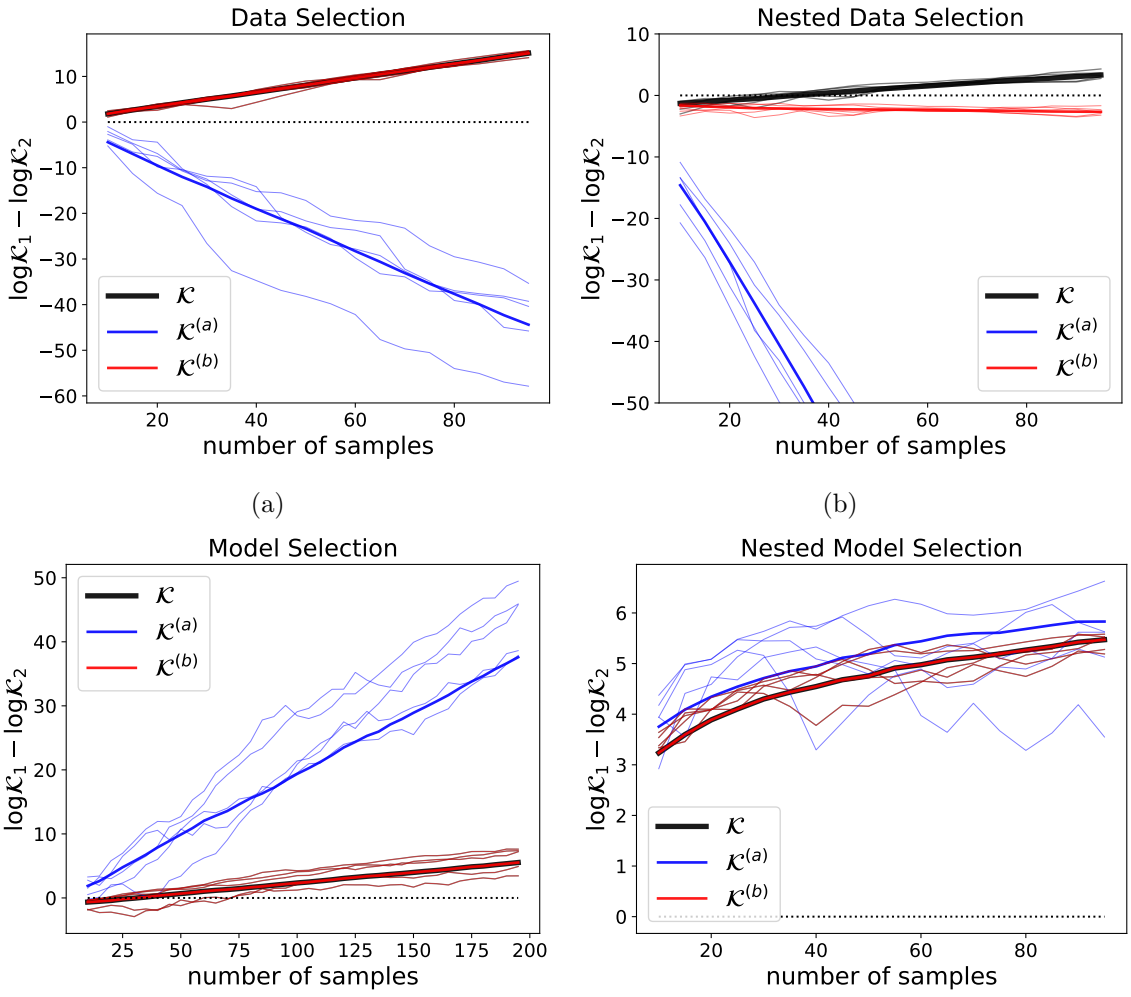


Figure 10: Behavior of the Stein volume criterion $\mathcal{K}$, the foreground marginal likelihood with a background volume correction $\mathcal{K}^{(\mathrm{a})}$, and the foreground marginal NKSD $\mathcal{K}^{(\mathrm{b})}$ on toy examples. The plots show the results for 5 randomly generated data sets (thin lines) and the average over 100 random data sets (bold lines). Here, unlike Figure 2, the Pitman-Yor expression for $m_{\mathcal{B}}$ is used (Equation 3), with $\alpha = 0.5$, $\nu = 1$, and $D = 0.2$.

As shown by Dawid (2011) and Hong and Preston (2005), under regularity conditions,

$$\begin{aligned} N(f_N^{\mathrm{KL}}(\theta_N^{\mathrm{KL}}) - f_N^{\mathrm{KL}}(\theta_*^{\mathrm{KL}})) &= O_{P_0}(1) \\ N(f_N^{\mathrm{KL}}(\theta_*^{\mathrm{KL}}) - f^{\mathrm{KL}}(\theta_*^{\mathrm{KL}})) &= O_{P_0}(\sqrt{N}) \\ N f^{\mathrm{KL}}(\theta_*^{\mathrm{KL}}) &= O_{P_0}(N) \\ \log \left(\frac{\pi(\theta_*^{\mathrm{KL}})(2\pi)^{m/2}}{\left| \det \nabla_\theta^2 f^{\mathrm{KL}}(\theta_*^{\mathrm{KL}}) \right|^{1/2}} \right) &= O_{P_0}(1). \end{aligned} \tag{61}$$

The NKSD marginal likelihood has a similar decomposition. Following Section 6, define

$$f_N^{\mathrm{NKSD}}(\theta) := \frac{1}{T}\widehat{\mathrm{NKSD}}(p_0(x)\|q(x|\theta)), \qquad \theta_N^{\mathrm{NKSD}} := \operatorname*{argmin}_\theta f_N^{\mathrm{NKSD}}(\theta),$$
$$f^{\mathrm{NKSD}}(\theta) := \frac{1}{T}\mathrm{NKSD}(p_0(x)\|q(x|\theta)), \qquad \theta_*^{\mathrm{NKSD}} := \operatorname*{argmin}_\theta f^{\mathrm{NKSD}}(\theta).$$

As shown in Theorem 9,

$$z_N := \int \exp(-N f_N^{\mathrm{NKSD}}(\theta))\pi(\theta)d\theta \sim \frac{\exp(-N f_N^{\mathrm{NKSD}}(\theta_N^{\mathrm{NKSD}}))\pi(\theta_*^{\mathrm{NKSD}})}{\left|\det \nabla_\theta^2 f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})\right|^{1/2}} \left(\frac{2\pi}{N}\right)^{m/2}$$

almost surely as $N \to \infty$. As above, we can rewrite this as

$$\begin{aligned}\log z_N &+ N(f_N^{\mathrm{NKSD}}(\theta_N^{\mathrm{NKSD}}) - f_N^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})) \\ &+ N(f_N^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}}) - f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})) + N f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}}) \\ &+ \frac{m}{2}\log N - \log\left(\frac{\pi(\theta_*^{\mathrm{NKSD}})(2\pi)^{m/2}}{\left|\det \nabla_\theta^2 f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})\right|^{1/2}}\right) \xrightarrow[N\to\infty]{\text{a.s.}} 0.\end{aligned} \tag{62}$$

By Theorem 12, we have

$$\begin{aligned} N(f_N^{\mathrm{NKSD}}(\theta_N^{\mathrm{NKSD}}) - f_N^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})) &= O_{P_0}(1), \\ N(f_N^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}}) - f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})) &= O_{P_0}(\sqrt{N}), \\ N f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}}) &= O_{P_0}(N), \\ \log\left(\frac{\pi(\theta_*^{\mathrm{NKSD}})(2\pi)^{m/2}}{\left|\det \nabla_\theta^2 f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})\right|^{1/2}}\right) &= O_{P_0}(1), \end{aligned} \tag{63}$$

and further, when the model is well-specified, such that $\mathrm{NKSD}(p_0(x)\|q(x|\theta_*^{\mathrm{NKSD}})) = 0$,

$$N(f_N^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}}) - f^{\mathrm{NKSD}}(\theta_*^{\mathrm{NKSD}})) = O_{P_0}(1). \tag{64}$$

For ease of reference, here are the various scores that we consider for model/data selection.

Marginal likelihood of the augmented model (foreground+background):

$$\tilde{q}(X^{(1:N)}|\mathcal{F}) = \int\int q(X_{\mathcal{F}}^{(1:N)}|\theta)\, \tilde{q}(X_{\mathcal{B}}^{(1:N)}|X_{\mathcal{F}}^{(1:N)}, \phi_{\mathcal{B}})\pi(\theta)\pi_{\mathcal{B}}(\phi_{\mathcal{B}})d\theta d\phi_{\mathcal{B}}.$$

Foreground marginal NKSD*, background volume correction (a.k.a. the SVC):*

$$\mathcal{K} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} \int \exp\Big(-\frac{N}{T}\widehat{\mathrm{NKSD}}(p_0(x_{\mathcal{F}})\|q(x_{\mathcal{F}}|\theta))\Big)\pi(\theta)d\theta.$$

Foreground marginal likelihood, background volume correction:

$$\mathcal{K}^{(\mathrm{a})} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2} q(X_{\mathcal{F}}^{(1:N)}).$$

Foreground marginal NKSD*:*

$$\mathcal{K}^{(\mathrm{b})} := \int \exp\left(-\frac{N}{T}\widehat{\mathrm{NKSD}}(p_0(x_{\mathcal{F}})\|q(x_{\mathcal{F}}|\theta))\right)\pi(\theta)d\theta.$$

Foreground marginal KL*, background volume correction:*

$$\mathcal{K}^{(\mathrm{c})} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2}\int \exp\left(-N\widehat{\mathrm{KL}}(p_0(x_{\mathcal{F}})\|q(x_{\mathcal{F}}|\theta))\right)\pi(\theta)d\theta.$$

Foreground NKSD*, background volume correction:*

$$\mathcal{K}^{(\mathrm{d})} := \left(\frac{2\pi}{N}\right)^{m_{\mathcal{B}}/2}\exp\left(-\frac{N}{T}\min_{\theta}\widehat{\mathrm{NKSD}}(p_0(x_{\mathcal{F}})\|q(x_{\mathcal{F}}|\theta))\right).$$

Foreground NKSD*, foreground and background volume correction (a.k.a. BIC for SVC)*

$$\mathcal{K}^{\mathrm{BIC}} := \left(\frac{2\pi}{N}\right)^{(m_{\mathcal{F}}+m_{\mathcal{B}})/2}\exp\Big(-\frac{N}{T}\min_{\theta}\widehat{\mathrm{NKSD}}(p_0(x_{\mathcal{F}})\|q(x_{\mathcal{F}}|\theta))\Big).$$

B.2 Data selection

Assume $m_{\mathcal{B}_j} = o(N/\log N)$ for $j \in \{1,2\}$. By Equations 60 and 61,

$$\begin{aligned}\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{a})}}{\mathcal{K}_2^{(\mathrm{a})}} \xrightarrow[N\to\infty]{P_0} &\mathbb{E}_{X\sim p_0}[-\log q(X_{\mathcal{F}_2}|\theta_{2,*}^{\mathrm{KL}})] - \mathbb{E}_{X\sim p_0}[-\log q(X_{\mathcal{F}_1}|\theta_{1,*}^{\mathrm{KL}})] \\ &= \mathrm{KL}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta_{2,*}^{\mathrm{KL}})) + H_{\mathcal{F}_2} - \mathrm{KL}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta_{1,*}^{\mathrm{KL}})) - H_{\mathcal{F}_1},\end{aligned} \tag{65}$$

so $\mathcal{K}^{(\mathrm{a})}$ does not satisfy data selection consistency. The SVC satisfies data selection consistency by Theorem 17 (part 1). We show that the other scores also satisfy data selection consistency. Since $\mathcal{K}^{(\mathrm{b})} = (2\pi/N)^{-m_{\mathcal{B}}/2}\mathcal{K}$ where $\mathcal{K}$ is the SVC, by Theorem 17 (part 1),

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{b})}}{\mathcal{K}_2^{(\mathrm{b})}} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta_{2,*}^{\mathrm{NKSD}})) - \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta_{1,*}^{\mathrm{NKSD}})). \tag{66}$$

By Equation 65 and the fact that $\mathcal{K}^{(\mathrm{c})} = \exp(NH_{\mathcal{F}})\mathcal{K}^{(\mathrm{a})}$, we have

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{c})}}{\mathcal{K}_2^{(\mathrm{c})}} \xrightarrow[N\to\infty]{P_0} \mathrm{KL}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta_{2,*}^{\mathrm{KL}})) - \mathrm{KL}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta_{1,*}^{\mathrm{KL}})). \tag{67}$$

Since $\mathcal{K}^{(\mathrm{d})} = (2\pi/N)^{m_{\mathcal{B}}/2}\exp(-Nf_N^{\mathrm{NKSD}}(\theta_N^{\mathrm{NKSD}}))$, then by Equation 63,

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{d})}}{\mathcal{K}_2^{(\mathrm{d})}} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta_{2,*}^{\mathrm{NKSD}})) - \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta_{1,*}^{\mathrm{NKSD}})). \tag{68}$$

Similarly, since $\mathcal{K}^{\mathrm{BIC}} = (2\pi/N)^{m_{\mathcal{F}}/2}\mathcal{K}^{(\mathrm{d})}$,

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{\mathrm{BIC}}}{\mathcal{K}_2^{\mathrm{BIC}}} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta_{2,*}^{\mathrm{NKSD}})) - \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta_{1,*}^{\mathrm{NKSD}})). \tag{69}$$

These methods therefore satisfy data selection consistency. For the marginal likelihood of the augmented model, suppose $m_{\mathcal{B}_1}$ and $m_{\mathcal{B}_2}$ do not depend on N. Then by Equation 60,

$$\frac{1}{N}\log\frac{\tilde{q}(X^{(1:N)}|\mathcal{F}_1)}{\tilde{q}(X^{(1:N)}|\mathcal{F}_2)} \xrightarrow[N\to\infty]{P_0} \mathbb{E}_{X_{\mathcal{F}_2}\sim p_0}[-\log q(X_{\mathcal{F}_2}|\theta^{\mathrm{KL}}_{2,*})] + \mathbb{E}_{X\sim p_0}[-\log \tilde{q}(X_{\mathcal{B}_2}|X_{\mathcal{F}_2},\phi^{\mathrm{KL}}_{2,*})] \\ - \mathbb{E}_{X_{\mathcal{F}_1}\sim p_0}[-\log q(X_{\mathcal{F}_1}|\theta^{\mathrm{KL}}_{1,*})] - \mathbb{E}_{X\sim p_0}[-\log \tilde{q}(X_{\mathcal{B}_1}|X_{\mathcal{F}_1},\phi^{\mathrm{KL}}_{1,*})]\big] \tag{70}$$

We can rewrite this in terms of the KL divergence. First note the decomposition,

$$H = -\int p_0(x)\log p_0(x)dx = -\int p_0(x_{\mathcal{F}_j})\log p_0(x_{\mathcal{F}_j})dx_{\mathcal{F}_j} - \int p_0(x)\log p_0(x_{\mathcal{B}_j}|x_{\mathcal{F}_j})dx$$

for $j\in\{1,2\}$. Adding and subtracting the entropy H in Equation 70, and using the fact that the background model is well-specified,

$$\begin{aligned}\frac{1}{N}\log\frac{\tilde{q}(X^{(1:N)}|\mathcal{F}_1)}{\tilde{q}(X^{(1:N)}|\mathcal{F}_2)} &\xrightarrow[N\to\infty]{P_0} \mathrm{KL}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta^{\mathrm{KL}}_{2,*})) + \mathrm{KL}(p_0(x_{\mathcal{B}_2}|x_{\mathcal{F}_2})\|\tilde{q}(x_{\mathcal{B}_2}|x_{\mathcal{F}_2},\phi^{\mathrm{KL}}_{2,*})) \\ &\quad - \mathrm{KL}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta^{\mathrm{KL}}_{1,*})) - \mathrm{KL}(p_0(x_{\mathcal{B}_1}|x_{\mathcal{F}_1})\|\tilde{q}(x_{\mathcal{B}_1}|x_{\mathcal{F}_1},\phi^{\mathrm{KL}}_{1,*})) \\ &= \mathrm{KL}(p_0(x_{\mathcal{F}_2})\|q(x_{\mathcal{F}_2}|\theta^{\mathrm{KL}}_{2,*})) - \mathrm{KL}(p_0(x_{\mathcal{F}_1})\|q(x_{\mathcal{F}_1}|\theta^{\mathrm{KL}}_{1,*})).\end{aligned} \tag{71}$$

B.3 Nested data selection

In nested data selection, we are concerned with situations in which $\mathcal{X}_{\mathcal{F}_2}\subset\mathcal{X}_{\mathcal{F}_1}$ and the model is well-specified over both $\mathcal{X}_{\mathcal{F}_1}$ and $\mathcal{X}_{\mathcal{F}_2}$. Assume further that $m_{\mathcal{B}_2}-m_{\mathcal{B}_1}$ does not depend on N. First, consider $\mathcal{K}^{(\mathrm{d})}$ and $\mathcal{K}^{\mathrm{BIC}}$. Since $\mathcal{K}^{(\mathrm{d})} = (2\pi/N)^{m_{\mathcal{B}}/2}\exp(-Nf^{\mathrm{NKSD}}_N(\theta^{\mathrm{NKSD}}_N))$ and by Theorem 12, $f^{\mathrm{NKSD}}_N(\theta^{\mathrm{NKSD}}_N) = O_{P_0}(1/N)$, we have

$$\frac{1}{\log N}\log\frac{\mathcal{K}^{(\mathrm{d})}_1}{\mathcal{K}^{(\mathrm{d})}_2} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{B}_2}-m_{\mathcal{B}_1}}{2}. \tag{72}$$

Likewise, since $\mathcal{K}^{\mathrm{BIC}} = (2\pi/N)^{m_{\mathcal{F}}/2}\mathcal{K}^{(\mathrm{d})}$, it follows that

$$\frac{1}{\log N}\log\frac{\mathcal{K}^{\mathrm{BIC}}_1}{\mathcal{K}^{\mathrm{BIC}}_2} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F}_2}+m_{\mathcal{B}_2}-m_{\mathcal{F}_1}-m_{\mathcal{B}_1}}{2}. \tag{73}$$

As in Section 6.4, it is natural to assume $m_{\mathcal{B}_2} > m_{\mathcal{B}_1}$ and $m_{\mathcal{F}_2}+m_{\mathcal{B}_2} > m_{\mathcal{F}_1}+m_{\mathcal{B}_1}$, in which case these criteria satisfy nested data selection consistency.

None of $\mathcal{K}^{(\mathrm{a})}$, $\mathcal{K}^{(\mathrm{b})}$, and $\mathcal{K}^{(\mathrm{c})}$ are guaranteed to satisfy nested data selection consistency, because the contribution of background model complexity is negligible or nonexistent. To see this, note that assuming $m_{\mathcal{B}_j} = o(N/\log N)$, by Equation 65 we have

$$\frac{1}{N}\log\frac{\mathcal{K}^{(\mathrm{a})}_1}{\mathcal{K}^{(\mathrm{a})}_2} \xrightarrow[N\to\infty]{P_0} H_{\mathcal{F}_2} - H_{\mathcal{F}_1}. \tag{74}$$

Meanwhile, since $\mathcal{K}^{(\mathrm{b})} = (2\pi/N)^{-m_{\mathcal{B}}/2}\mathcal{K}$ then by Theorem 17 (part 2),

$$\frac{1}{\log N}\log\frac{\mathcal{K}^{(\mathrm{b})}_1}{\mathcal{K}^{(\mathrm{b})}_2} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F}_2}-m_{\mathcal{F}_1}}{2}. \tag{75}$$

Since $\mathcal{X}_{\mathcal{F}_2} \subset \mathcal{X}_{\mathcal{F}_1}$, we have $m_{\mathcal{F}_2} \leq m_{\mathcal{F}_1}$ except perhaps in highly contrived scenarios. If $m_{\mathcal{F}_2} < m_{\mathcal{F}_1}$ then Equation 75 shows that $\log(\mathcal{K}_1^{(\mathrm{b})}/\mathcal{K}_2^{(\mathrm{b})}) \xrightarrow{P_0} -\infty$. On the other hand, if $m_{\mathcal{F}_2} = m_{\mathcal{F}_1}$, then by Equations 62 and 63, $\log(\mathcal{K}_1^{(\mathrm{b})}/\mathcal{K}_2^{(\mathrm{b})}) = O_{P_0}(1)$, so it is not possible to have $\log(\mathcal{K}_1^{(\mathrm{b})}/\mathcal{K}_2^{(\mathrm{b})}) \xrightarrow{P_0} \infty$. Therefore, $\mathcal{K}^{(\mathrm{b})}$ does not satisfy nested data selection consistency.

Since $\mathcal{K}^{(\mathrm{c})} = e^{NH_{\mathcal{F}}}\mathcal{K}^{(\mathrm{a})} = e^{NH_{\mathcal{F}}}(2\pi/N)^{m_{\mathcal{B}}/2}q(X_{\mathcal{F}}^{(1:N)})$, then by Equations 60 and 61,

$$\frac{1}{\sqrt{N}}\log\frac{\mathcal{K}_1^{(\mathrm{c})}}{\mathcal{K}_2^{(\mathrm{c})}} = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\log\frac{p_0(X_{\mathcal{F}_1}^{(i)})}{p_0(X_{\mathcal{F}_2}^{(i)})} - \mathbb{E}\Big(\log\frac{p_0(X_{\mathcal{F}_1})}{p_0(X_{\mathcal{F}_2})}\Big)\right) + O_{P_0}(N^{-1/2}\log N). \quad (76)$$

If $\sigma^2 := \mathbb{V}_{P_0}(\log p_0(X_{\mathcal{F}_1})/p_0(X_{\mathcal{F}_2}))$ is positive and finite, then by the central limit theorem and Slutsky's theorem, $N^{-1/2}\log(\mathcal{K}_1^{(\mathrm{c})}/\mathcal{K}_2^{(\mathrm{c})}) \xrightarrow{D} \mathcal{N}(0,\sigma^2)$. Thus, $\mathcal{K}^{(\mathrm{c})}$ randomly selects $\mathcal{F}_1$ or $\mathcal{F}_2$ with equal probability, and therefore, it does not satisfy nested data selection consistency.

For the marginal likelihood of the augmented model, suppose $m_{\mathcal{B}_1}$ and $m_{\mathcal{B}_2}$ do not depend on N. The marginal likelihood achieves nested data selection consistency because the augmented models are both well-specified and describe the complete data space $\mathcal{X}$; this guarantees that the $O_{P_0}(\sqrt{N})$ terms in the marginal likelihood decomposition cancel. Specifically, $p_0(x) = q(x \mid \theta_{j,*}^{\mathrm{KL}}, \phi_{j,*}^{\mathrm{KL}}, \mathcal{F}_j)$ for $j \in \{1,2\}$, and thus, by Equations 60 and 61 applied to the augmented model,

$$\frac{1}{\log N}\log\frac{\tilde{q}(X^{(1:N)}|\mathcal{F}_1)}{\tilde{q}(X^{(1:N)}|\mathcal{F}_2)} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F}_2} + m_{\mathcal{B}_2} - m_{\mathcal{F}_1} - m_{\mathcal{B}_1}}{2}. \quad (77)$$

Nested data selection consistency follows assuming $m_{\mathcal{F}_2} + m_{\mathcal{B}_2} > m_{\mathcal{F}_1} + m_{\mathcal{B}_1}$ as before. This can be contrasted with Equation 76, where although both foreground models are well-specified, they describe different data ($X_{\mathcal{F}_1}^{(1:N)}$ versus $X_{\mathcal{F}_2}^{(1:N)}$), so the $O_{P_0}(\sqrt{N})$ terms remain.

B.4 Model selection

All of the criteria we consider satisfy model selection consistency. To see this, we apply the same asymptotic analysis as used for data selection in Section B.2, under the same conditions on $m_{\mathcal{B}}$, obtaining

$$\frac{1}{N}\log\frac{\tilde{q}_1(X^{(1:N)}|\mathcal{F})}{\tilde{q}_2(X^{(1:N)}|\mathcal{F})} \xrightarrow[N\to\infty]{P_0} \mathrm{KL}(p_0(x_{\mathcal{F}})\|q_2(x_{\mathcal{F}}|\theta_{2,*}^{\mathrm{KL}})) - \mathrm{KL}(p_0(x_{\mathcal{F}})\|q_1(x_{\mathcal{F}}|\theta_{1,*}^{\mathrm{KL}})), \quad (78)$$

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{a})}}{\mathcal{K}_2^{(\mathrm{a})}} \xrightarrow[N\to\infty]{P_0} \mathrm{KL}(p_0(x_{\mathcal{F}})\|q_2(x_{\mathcal{F}}|\theta_{2,*}^{\mathrm{KL}})) - \mathrm{KL}(p_0(x_{\mathcal{F}})\|q_1(x_{\mathcal{F}}|\theta_{1,*}^{\mathrm{KL}})), \quad (79)$$

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{b})}}{\mathcal{K}_2^{(\mathrm{b})}} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}})\|q_2(x_{\mathcal{F}}|\theta_{2,*}^{\mathrm{NKSD}})) - \frac{1}{T}\mathrm{NKSD}(p_0(x_{\mathcal{F}})\|q_1(x_{\mathcal{F}}|\theta_{1,*}^{\mathrm{NKSD}})), \quad (80)$$

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{c})}}{\mathcal{K}_2^{(\mathrm{c})}} \xrightarrow[N\to\infty]{P_0} \mathrm{KL}(p_0(x_{\mathcal{F}})\|q_2(x_{\mathcal{F}}|\theta_{2,*}^{\mathrm{KL}})) - \mathrm{KL}(p_0(x_{\mathcal{F}})\|q_1(x_{\mathcal{F}}|\theta_{1,*}^{\mathrm{KL}})), \quad (81)$$

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{(\mathrm{d})}}{\mathcal{K}_2^{(\mathrm{d})}} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\mathrm{NKSD}(p_0(x_\mathcal{F})\|q_2(x_\mathcal{F}|\theta_{2,*}^{\mathrm{NKSD}})) - \frac{1}{T}\mathrm{NKSD}(p_0(x_\mathcal{F})\|q_1(x_\mathcal{F}|\theta_{1,*}^{\mathrm{NKSD}})), \quad (82)$$

$$\frac{1}{N}\log\frac{\mathcal{K}_1^{\mathrm{BIC}}}{\mathcal{K}_2^{\mathrm{BIC}}} \xrightarrow[N\to\infty]{P_0} \frac{1}{T}\mathrm{NKSD}(p_0(x_\mathcal{F})\|q_2(x_\mathcal{F}|\theta_{2,*}^{\mathrm{NKSD}})) - \frac{1}{T}\mathrm{NKSD}(p_0(x_\mathcal{F})\|q_1(x_\mathcal{F}|\theta_{1,*}^{\mathrm{NKSD}})). \quad (83)$$

Note that in contrast to the data selection case, $\mathcal{K}^{(\mathrm{a})}$ satisfies model selection consistency since the entropy terms $H_{\mathcal{F}_j}$ cancel due to the fact that $\mathcal{F}$ is fixed. We can think of this as a consequence of the KL divergence's subsystem independence; if we are just interested in modeling a fixed foreground space, there is no problem considering the foreground marginal likelihood alone (Caticha, 2004, 2011; Rezende, 2018).

B.5 Nested model selection

In nested model selection, since both models are well-specified, we have $q_j(x_\mathcal{F}|\theta_{j,*}^{\mathrm{KL}}) = p_0(x_\mathcal{F}) = q_j(x_\mathcal{F}|\theta_{j,*}^{\mathrm{NKSD}})$ for $j \in \{1,2\}$. Thus, the estimated divergences cancel:

$$\begin{aligned}
\widehat{\mathrm{NKSD}}(p_0(x_\mathcal{F})\|q_1(x_\mathcal{F}|\theta_{1,*}^{\mathrm{NKSD}})) &= \widehat{\mathrm{NKSD}}(p_0(x_\mathcal{F})\|q_2(x_\mathcal{F}|\theta_{2,*}^{\mathrm{NKSD}})),\\
\sum_{i=1}^N \log q_1(X_\mathcal{F}^{(i)}|\theta_{1,*}^{\mathrm{KL}}) &= \sum_{i=1}^N \log q_2(X_\mathcal{F}^{(i)}|\theta_{2,*}^{\mathrm{KL}}),\\
\widehat{\mathrm{KL}}(p_0(x_\mathcal{F})\|q_1(x_\mathcal{F}|\theta_{1,*}^{\mathrm{KL}})) &= \widehat{\mathrm{KL}}(p_0(x_\mathcal{F})\|q_2(x_\mathcal{F}|\theta_{2,*}^{\mathrm{KL}})).
\end{aligned}$$

Using this along with Equations 60–64, under the same conditions on $m_\mathcal{B}$ as in Section B.2,

$$\frac{1}{\log N}\log\frac{\tilde{q}_1(X^{(1:N)}|\mathcal{F})}{\tilde{q}_2(X^{(1:N)}|\mathcal{F})} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F},2}-m_{\mathcal{F},1}}{2}, \quad (84)$$

$$\frac{1}{\log N}\log\frac{\mathcal{K}_1^{(\mathrm{a})}}{\mathcal{K}_2^{(\mathrm{a})}} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F},2}-m_{\mathcal{F},1}}{2}, \quad (85)$$

$$\frac{1}{\log N}\log\frac{\mathcal{K}_1^{(\mathrm{b})}}{\mathcal{K}_2^{(\mathrm{b})}} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F},2}-m_{\mathcal{F},1}}{2}, \quad (86)$$

$$\frac{1}{\log N}\log\frac{\mathcal{K}_1^{(\mathrm{c})}}{\mathcal{K}_2^{(\mathrm{c})}} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F},2}-m_{\mathcal{F},1}}{2}, \quad (87)$$

$$\log\frac{\mathcal{K}_1^{(\mathrm{d})}}{\mathcal{K}_2^{(\mathrm{d})}} = O_{P_0}(1), \quad (88)$$

$$\frac{1}{\log N}\log\frac{\mathcal{K}_1^{\mathrm{BIC}}}{\mathcal{K}_2^{\mathrm{BIC}}} \xrightarrow[N\to\infty]{P_0} \frac{m_{\mathcal{F},2}-m_{\mathcal{F},1}}{2}, \quad (89)$$

where we are using the assumption that the background model is the same in the two augmented models $\tilde{q}_1$ and $\tilde{q}_2$ and so $m_{\mathcal{B},1} = m_{\mathcal{B},2}$. Only $\mathcal{K}^{(\mathrm{d})}$ fails to satisfy nested model selection consistency.

Appendix C. Proofs

C.1 Proofs of NKSD properties

Proof of Proposition 3 By assumption, the kernel is bounded, say $|k(x,y)| \leq B$, and $s_p, s_q \in L^1(P)$. Thus, by the Cauchy–Schwarz inequality,

$$\left| \int_{\mathcal{X}} \int_{\mathcal{X}} (s_q(x) - s_p(x))^\top (s_q(y) - s_p(y)) k(x,y) p(x) p(y) dx dy \right| \leq B \Big(\textstyle\int_{\mathcal{X}} \|s_q(x) - s_p(x)\| p(x) dx \Big)^2 < \infty.$$

Since the kernel is integrally strictly positive definite and $|k(x,y)| \leq B$,

$$0 < \int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y) p(x) p(y) dx dy \leq B < \infty. \tag{90}$$

Thus, the NKSD is finite. Equation 30 follows from Theorem 3.6 of Liu et al. (2016). ■

Proof of Proposition 4 The denominator of the NKSD is positive since k is integrally strictly positive definite. Defining $\delta(x) = s_q(x) - s_p(x)$, the numerator of the NKSD is

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \delta(x)^\top \delta(y) k(x,y) p(x) p(y) dx dy = \sum_{i=1}^{d} \int_{\mathcal{X}} \int_{\mathcal{X}} \delta_i(x) \delta_i(y) k(x,y) p(x) p(y) dx dy. \tag{91}$$

If $\delta_i(x)p(x) = 0$ almost everywhere with respect to Lebesgue measure on $\mathcal{X}$, then the ith term on the right-hand side is zero. Meanwhile, if $\delta_i(x)p(x)$ is not a.e. zero, then $\int_{\mathcal{X}} |\delta_i(x)| p(x) dx > 0$, and hence, the ith term is positive since k is integrally strictly positive definite and $\delta_i \in L^1(P)$ by assumption. Hence, the NKSD is nonnegative, and equals zero if and only if $\delta(x)p(x) = 0$ almost everywhere.

Suppose $\delta(x)p(x) = 0$ almost everywhere. Since $p(x) > 0$ on $\mathcal{X}$ by assumption, this implies $s_p(x) = s_q(x)$ a.e., and in fact, $s_p(x) = s_q(x)$ for all $x \in \mathcal{X}$ by continuity. Since $\mathcal{X}$ is open and connected, then by the gradient theorem (that is, the fundamental theorem of calculus for line integrals), $p(x) \propto q(x)$, and hence, $p(x) = q(x)$ on $\mathcal{X}$. Conversely, if $p(x) = q(x)$ almost everywhere, then $\delta(x)p(x) = 0$ almost everywhere. ■

Proof of Proposition 6 Define

$$\delta_1(x_1) := \nabla_{x_1} \log q(x) - \nabla_{x_1} \log p(x) = \nabla_{x_1} \log q(x_1) - \nabla_{x_1} \log p(x_1)$$
$$\delta_2(x_2) := \nabla_{x_2} \log q(x) - \nabla_{x_2} \log p(x) = \nabla_{x_2} \log q(x_2) - \nabla_{x_2} \log p(x_2).$$

Let $X, Y \sim p(x)$ independently. Note that $\mathbb{E}[k_1(X_1, Y_1)] > 0$ and $\mathbb{E}[k_2(X_2, Y_2)] > 0$ since k_1 and k_2 are integrally strictly positive definite by assumption. Therefore,

$$\begin{aligned}
\textsc{nksd}(p(x)\|q(x)) &= \frac{\mathbb{E}[(\nabla_x \log q(X) - \nabla_x \log p(X))^\top (\nabla_x \log q(Y) - \nabla_x \log p(Y)) k(X,Y)]}{\mathbb{E}[k(X,Y)]} \\
&= \frac{\mathbb{E}[\delta_1(X_1)^\top \delta_1(Y_1) k_1(X_1,Y_1)]\mathbb{E}[k_2(X_2,Y_2)]}{\mathbb{E}[k_1(X_1,Y_1)]\mathbb{E}[k_2(X_2,Y_2)]} + \frac{\mathbb{E}[\delta_2(X_2)^\top \delta_2(Y_2) k_2(X_2,Y_2)]\mathbb{E}[k_1(X_1,Y_1)]}{\mathbb{E}[k_1(X_1,Y_1)]\mathbb{E}[k_2(X_2,Y_2)]} \\
&= \frac{\mathbb{E}[\delta_1(X_1)^\top \delta_1(Y_1) k_1(X_1,Y_1)]}{\mathbb{E}[k_1(X_1,Y_1)]} + \frac{\mathbb{E}[\delta_2(X_2)^\top \delta_2(Y_2) k_2(X_2,Y_2)]}{\mathbb{E}[k_2(X_2,Y_2)]} \\
&= \textsc{nksd}(p(x_1)\|q(x_1)) + \textsc{nksd}(p(x_2)\|q(x_2)).
\end{aligned}$$

■

The following modified version applies to the estimator $\widehat{\textsc{nksd}}(p\|q)$ (Equation 5).

Proposition 20

$$\widehat{\textsc{nksd}}(p(x)\|q(x)) = \overline{\textsc{nksd}}(p(x_1)\|q(x_1)) + \overline{\textsc{nksd}}(p(x_2)\|q(x_2)) \tag{92}$$

where

$$\begin{aligned}
\overline{\textsc{nksd}}(p(x_1)\|q(x_1)) :=& \frac{\sum_{i\neq j} u_1(X_1^{(i)}, X_1^{(j)}) k_2(X_2^{(i)}, X_2^{(j)})}{\sum_{i\neq j} k_1(X_1^{(i)}, X_1^{(j)}) k_2(X_2^{(i)}, X_2^{(j)})} \\
u_1(x_1,y_1) :=& s_q(x_1)^\top s_q(y_1) k_1(x_1,y_1) + s_q(x_1)^\top \nabla_{y_1} k_1(x_1,y_1) + s_q(y_1)^\top \nabla_{x_1} k_1(x_1,y_1) \\
&+ \operatorname{trace}(\nabla_{x_1} \nabla_{y_1}^\top k_1(x_1,y_1)) \\
s_q(x_1) :=& \nabla_{x_1} \log q(x_1),
\end{aligned}$$

and vice versa for $\overline{\textsc{nksd}}(p(x_2)\|q(x_2))$ *with the roles of 1 and 2 swapped.*

Proof Recall the definition of $\widehat{\textsc{nksd}}(p(x)\|q(x))$ in Equation 5. Note that $\nabla_{x_1} k(x,y) = k_2(x_2,y_2) \nabla_{x_1} k_1(x_1,y_1)$ and $\nabla_{x_1} \log q(x) = \nabla_{x_1} \log q(x_1)$. Examining $u(x,y)$ term-by-term,

$$\begin{aligned}
\nabla_x \log q(x)^\top \nabla_y \log q(y) k(x,y) =& \left[\nabla_{x_1} \log q(x_1)^\top \nabla_{y_1} \log q(y_1) k_1(x_1,y_1)\right] k_2(x_2,y_2) \\
&+ \left[\nabla_{x_2} \log q(x_2)^\top \nabla_{y_2} \log q(y_2) k_2(x_2,y_2)\right] k_1(x_1,y_1), \\
\nabla_x \log q(x)^\top \nabla_y k(x,y) =& [\nabla_{x_1} \log q(x_1)^\top \nabla_{y_1} k_1(x_1,y_1)] k_2(x_2,y_2) \\
&+ [\nabla_{x_2} \log q(x_2)^\top \nabla_{y_2} k_2(x_2,y_2)] k_1(x_1,y_1), \\
\nabla_x k(x,y)^\top \nabla_y \log q(y) =& [\nabla_{x_1} k_1(x_1,y_1)^\top \nabla_{y_1} \log q(y_1)] k_2(x_2,y_2), \\
&+ [\nabla_{x_2} k_2(x_2,y_2)^\top \nabla_{y_2} \log q(y_2)] k_1(x_1,y_1) \\
\operatorname{trace}(\nabla_x \nabla_y^\top k(x,y)) =& \operatorname{trace}(\nabla_{x_1} \nabla_{y_1}^\top k_1(x_1,y_1)) k_2(x_2,y_2), \\
&+ \operatorname{trace}(\nabla_{x_2} \nabla_{y_2}^\top k_2(x_2,y_2)) k_1(x_1,y_1).
\end{aligned}$$

Thus, defining u_1 and u_2 as in Proposition 20, we have

$$\begin{aligned}
u(x,y) &= u_1(x_1,y_1) k_2(x_2,y_2) + u_2(x_2,y_2) k_1(x_1,y_1), \\
k(x,y) &= k_1(x_1,y_1) k_2(x_2,y_2).
\end{aligned}$$

The result follows. ■

To interpret Proposition 20, note that

$$\frac{\mathbb{E}_{X,Y\sim p}[u_1(X_1,Y_1)k_2(X_2,Y_2)]}{\mathbb{E}_{X,Y\sim p}[k_1(X_1,Y_1)k_2(X_2,Y_2)]} = \frac{\mathbb{E}_{X_1,Y_1\sim p(x_1)}[u_1(X_1,Y_1)]}{\mathbb{E}_{X_1,Y_1\sim p(x_1)}[k_1(X_1,Y_1)]} = \text{NKSD}(p(x_1)\|q(x_1)),$$

so $\overline{\text{NKSD}}(p(x_1)\|q(x_1))$ is an estimator of $\text{NKSD}(p(x_1)\|q(x_1))$, and likewise for $\overline{\text{NKSD}}(p(x_2)\|q(x_2))$.

C.2 Proof of Theorems 9 and 11

Our proofs in this section build on the proof of Theorem 3 of Barp et al. (2019).

Proposition 21 *Under the assumptions of Theorem 9, for any compact convex $C \subseteq \Theta$,*

$$\sup_{\theta\in C} |f_N(\theta) - f(\theta)| \xrightarrow{\text{a.s.}} 0. \tag{93}$$

Proof First, we establish almost sure convergence for the denominator of $f_N(\theta)$. Since k is assumed to be bounded and to have bounded derivatives up to order two, we can choose $B < \infty$ such that $B \geq |k| + \|\nabla_x k\| + \|\nabla_x \nabla_y^\top k\|$. In particular, the expected value of the kernel is finite:

$$\int_{\mathcal{X}}\int_{\mathcal{X}} |k(x,y)| P_0(dx)P_0(dy) \leq B < \infty. \tag{94}$$

By the strong law of large numbers for U-statistics (Theorem 5.4A of Serfling, 2009),

$$\frac{1}{N(N-1)}\sum_{i\neq j} k(X^{(i)},X^{(j)}) \xrightarrow[N\to\infty]{\text{a.s.}} \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)P_0(dx)P_0(dy). \tag{95}$$

Note that the limit is positive since $k(x,y) > 0$ for all $x, y \in \mathcal{X}$. For the numerator, we establish bounds on u_θ and $\nabla_\theta u_\theta$. Let $C \subseteq \Theta$ be compact and convex. By Equation 5, for all $\theta \in C$ and all $x, y \in \mathcal{X}$,

$$\begin{aligned}
|u_\theta(x,y)| &\leq |s_{q_\theta}(x)^\top s_{q_\theta}(y)k(x,y)| + |s_{q_\theta}(x)^\top \nabla_y k(x,y)| \\
&\quad + |s_{q_\theta}(y)^\top \nabla_x k(x,y)| + |\operatorname{trace}(\nabla_x \nabla_y^\top k(x,y))| \\
&\leq \|s_{q_\theta}(x)\|\|s_{q_\theta}(y)\|B + \|s_{q_\theta}(x)\|B + \|s_{q_\theta}(y)\|B + Bd \\
&\leq g_{0,C}(x)g_{0,C}(y)B + g_{0,C}(x)B + g_{0,C}(y)B + Bd \\
&=: h_{0,C}(x,y).
\end{aligned} \tag{96}$$

Similarly, for all $\theta \in C$ and all $x, y \in \mathcal{X}$,

$$\begin{aligned}
\|\nabla_\theta u_\theta(x,y)\| &\leq \|\nabla_\theta(s_{q_\theta}(x)^\top s_{q_\theta}(y))k(x,y)\| + \|\nabla_\theta(s_{q_\theta}(x)^\top \nabla_y k(x,y))\| \\
&\quad + \|\nabla_\theta(s_{q_\theta}(y)^\top \nabla_x k(x,y))\| + \|\nabla_\theta \operatorname{trace}(\nabla_x \nabla_y^\top k(x,y))\| \\
&\leq g_{0,C}(x)g_{1,C}(y)B + g_{0,C}(y)g_{1,C}(x)B + g_{1,C}(x)B + g_{1,C}(y)B \\
&=: h_{1,C}(x,y).
\end{aligned} \tag{97}$$

Note that $h_{0,C}$ and $h_{1,C}$ are continuous and belong to $L^1(P_0 \times P_0)$.

Let $S_1 \subseteq S_2 \subseteq \cdots \subseteq \mathcal{X}$ be a sequence of compact sets such that $\cup_{M=1}^{\infty} S_M = \mathcal{X}$. Note that this implies $\cup_{M=1}^{\infty} S_M \times S_M = \mathcal{X} \times \mathcal{X}$. Suppose for the moment that, for each M, the following collections of functions are equicontinuous on C: (A) $(\theta \mapsto u_\theta(x,y) : x, y \in S_M)$ and (B) $\big(\theta \mapsto \int u_\theta(x,y) P_0(dy) : x \in S_M\big)$. Assuming this, Theorem 1 of Yeo and Johnson (2001) shows that

$$\sup_{\theta \in C} \left| \frac{1}{N(N-1)} \sum_{i \neq j} u_\theta(X^{(i)}, X^{(j)}) - \int_{\mathcal{X}} \int_{\mathcal{X}} u_\theta(x,y) P_0(dx) P_0(dy) \right| \xrightarrow[N\to\infty]{\text{a.s.}} 0, \tag{98}$$

and that $\theta \mapsto \int_{\mathcal{X}} \int_{\mathcal{X}} u_\theta(x,y) P_0(dx) P_0(dy)$ is continuous. (Note that although Yeo and Johnson (2001) assume $\mathcal{X} = \mathbb{R}$, their proof goes through without further modification for any nonempty $\mathcal{X} \subseteq \mathbb{R}^d$.) Combining Equations 95 and 98, we have

$$\frac{\sup_{\theta \in C} \left| \frac{1}{N(N-1)} \sum_{i \neq j} u_\theta(X^{(i)}, X^{(j)}) - \int \int u_\theta(x,y) P_0(dx) P_0(dy) \right|}{\frac{1}{N(N-1)} \sum_{i \neq j} k(X^{(i)}, X^{(j)})} \xrightarrow[N\to\infty]{\text{a.s.}} 0.$$

Thus, it follows that $\sup_{\theta \in C} |f_N(\theta) - f(\theta)| \to 0$ a.s. by Equations 95 and 96. To complete the proof, we must show that (A) and (B) are equicontinuous on C.

(A) Since $\theta \mapsto u_\theta(x,y)$ is differentiable on C, then by the mean value theorem, we have that for all $\theta_1, \theta_2 \in C$ and all $x, y \in S_M$,

$$\begin{aligned} |u_{\theta_1}(x,y) - u_{\theta_2}(x,y)| &\leq \|\nabla_\theta|_{\theta=\tilde{\theta}} u_\theta(x,y)\| \|\theta_1 - \theta_2\| \\ &\leq h_{1,C}(x,y) \|\theta_1 - \theta_2\| \\ &\leq \Big(\sup_{x,y \in S_M} h_{1,C}(x,y) \Big) \|\theta_1 - \theta_2\| < \infty \end{aligned}$$

where $\tilde{\theta} = \gamma\theta_1 + (1-\gamma)\theta_2$ for some $\gamma \in [0,1]$. Here, the second inequality holds since $\tilde{\theta} \in C$ by the convexity of C, and the supremum is finite because a continuous function on a compact set attains its maximum. Therefore, $(\theta \mapsto u_\theta(x,y) : x, y \in S_M)$ is equicontinuous on C.

(B) To see that $\big(\theta \mapsto \int u_\theta(x,y) P_0(dy) : x \in S_M\big)$ is equicontinuous on C, first note that

$$\int |u_\theta(x,y)| P_0(dy) \leq \int h_{0,C}(x,y) P_0(dy) < \infty.$$

Further, due to Equations 96 and 97, we can apply the Leibniz integral rule (Folland, 1999, Theorem 2.27) and find that $\nabla_\theta \int u_\theta(x,y) P_0(dy)$ exists and is equal to $\int \nabla_\theta u_\theta(x,y) P_0(dy)$. Now we apply the mean value theorem and the same reasoning as before to find that for all $\theta_1, \theta_2 \in C$ and all $x \in S_M$,

$$\begin{aligned} \left| \int u_{\theta_1}(x,y) P_0(dy) - \int u_{\theta_2}(x,y) P_0(dy) \right| &\leq \left\| \nabla_\theta|_{\theta=\tilde{\theta}} \int u_\theta(x,y) P_0(dy) \right\| \|\theta_1 - \theta_2\| \\ &\leq \|\theta_1 - \theta_2\| \int \left\| \nabla_\theta|_{\theta=\tilde{\theta}} u_\theta(x,y) \right\| P_0(dy) \\ &\leq \|\theta_1 - \theta_2\| \sup_{x \in S_M} \int h_{1,C}(x,y) P_0(dy) < \infty \end{aligned}$$

where $\tilde{\theta} = \gamma\theta_1 + (1-\gamma)\theta_2$ for some $\gamma \in [0,1]$. The supremum is finite since $x \mapsto \int h_{1,C}(x,y)P_0(dy)$ is continuous, which can easily be seen by plugging in the definition of $h_{1,C}$. Therefore, $\big(\theta \mapsto \int u_\theta(x,y)P_0(dy) : x \in S_M\big)$ is equicontinuous on C. ■

Proposition 22 *Under the assumptions of Theorem 9, $(f_N''' : N \in \mathbb{N})$ is uniformly bounded on E.*

Proof First, for any $x, y \in \mathcal{X}$, if we define $g(\theta) = s_{q_\theta}(x)$ and $h(\theta) = s_{q_\theta}(y)$ then $u_\theta = (g^\top h)k + g^\top(\nabla_y k) + h^\top(\nabla_x k) + \mathrm{trace}(\nabla_x \nabla_y^\top k)$. By differentiating, applying Minkowski's inequality to the resulting sum of tensors, and applying the Cauchy–Schwarz inequality to each term, we have

$$\begin{aligned}\|\nabla_\theta^3 u_\theta(x,y)\| \leq{}& \|\nabla^3 g\|\|h\|k + 3\|\nabla^2 g\|\|\nabla h\|k + 3\|\nabla g\|\|\nabla^2 h\|k + \|g\|\|\nabla^3 h\|k \\ &+ \|\nabla^3 g\|\|\nabla_y k\| + \|\nabla^3 h\|\|\nabla_x k\|.\end{aligned}$$

Using the symmetry of the kernel to combine like terms, this yields that

$$\begin{aligned}&\Big\|\sum_{i\neq j}\nabla_\theta^3 u_\theta(X^{(i)}, X^{(j)})\Big\| \\ &\leq \sum_{i\neq j}\Big(2\|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\|\|s_{q_\theta}(X^{(j)})\|B + 6\|\nabla_\theta^2 s_{q_\theta}(X^{(i)})\|\|\nabla_\theta s_{q_\theta}(X^{(j)})\|B + 2\|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\|B\Big)\end{aligned}$$

where $B < \infty$ such that $B \geq |k| + \|\nabla_x k\| + \|\nabla_x \nabla_y^\top k\|$. Since $f_N(\theta) = 0$ when $N = 1$ by definition, we can assume without loss of generality that $N \geq 2$, so $\frac{1}{N-1} = \frac{1}{N}(1 + \frac{1}{N-1}) \leq 2/N$. Since each term is non-negative, we can add in the $i = j$ terms,

$$\begin{aligned}&\Big\|\frac{1}{N(N-1)}\sum_{i\neq j}\nabla_\theta^3 u_\theta(X^{(i)}, X^{(j)})\Big\| \\ &\quad\leq \frac{2B}{N^2}\sum_{i,j}\Big(2\|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\|\|s_{q_\theta}(X^{(j)})\| + 6\|\nabla_\theta^2 s_{q_\theta}(X^{(i)})\|\|\nabla_\theta s_{q_\theta}(X^{(j)})\| + 2\|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\|\Big) \\ &\quad= 4B\Big(\frac{1}{N}\sum_i \|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\|\Big)\Big(\frac{1}{N}\sum_j \|s_{q_\theta}(X^{(j)})\|\Big) \qquad (99) \\ &\qquad + 12B\Big(\frac{1}{N}\sum_i \|\nabla_\theta^2 s_{q_\theta}(X^{(i)})\|\Big)\Big(\frac{1}{N}\sum_j \|\nabla_\theta s_{q_\theta}(X^{(j)})\|\Big) \\ &\qquad + 4B\Big(\frac{1}{N}\sum_i \|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\|\Big).\end{aligned}$$

By assumption, $\big\{\frac{1}{N}\sum_i \|\nabla_\theta^2 s_{q_\theta}(X^{(i)})\| : N \in \mathbb{N}, \theta \in E\big\}$ is bounded with probability 1, and similarly for $\big\{\frac{1}{N}\sum_i \|\nabla_\theta^3 s_{q_\theta}(X^{(i)})\| : N \in \mathbb{N}, \theta \in E\big\}$. We show the same for $\frac{1}{N}\sum_i \|s_{q_\theta}(X^{(i)})\|$ and $\frac{1}{N}\sum_i \|\nabla_\theta s_{q_\theta}(X^{(i)})\|$. By Equation 40, we have

$$\int \sup_{\theta\in\bar{E}} \|s_{q_\theta}(x)\| P_0(dx) \leq \int g_{0,\bar{E}}(x) P_0(dx) < \infty.$$

Hence, by Theorem 1.3.3 of Ghosh and Ramamoorthi (2003), $\frac{1}{N}\sum_i \|s_{q_\theta}(X^{(i)})\|$ converges uniformly on $\bar{E}$, almost surely. In particular, $\frac{1}{N}\sum_i \|s_{q_\theta}(X^{(i)})\|$ is uniformly bounded on E, almost surely. The same argument holds for $\frac{1}{N}\sum_i \|\nabla_\theta s_{q_\theta}(X^{(i)})\|$ using $g_{1,\bar{E}}(x)$. Therefore, by Equation 99, it follows that $\|\frac{1}{N(N-1)}\sum_{i\neq j}\nabla^3_\theta u_\theta(X^{(i)}, X^{(j)})\|$ is uniformly bounded on E. Since k is positive by assumption, $\frac{1}{N(N-1)}\sum_{i\neq j} k(X^{(i)}, X^{(j)}) > 0$ for all $N \geq 2$ and by Equations 94 and 95, $\frac{1}{N(N-1)}\sum_{i\neq j} k(X^{(i)}, X^{(j)})$ converges a.s. to a finite quantity greater than 0. We conclude that almost surely,

$$\|f'''_N(\theta)\| = \frac{1}{T}\frac{\|\frac{1}{N(N-1)}\sum_{i\neq j}\nabla^3_\theta u_\theta(X^{(i)}, X^{(j)})\|}{\frac{1}{N(N-1)}\sum_{i\neq j} k(X^{(i)}, X^{(j)})}$$

is uniformly bounded on E, for $N \in \{2, 3, \ldots\}$. Recall that for $N = 1$, $f_N(\theta) = 0$ by definition. Therefore, almost surely, $(f'''_N : N \in \mathbb{N})$ is uniformly bounded on E. ■

Proof of Theorem 9 We show that the conditions of Theorem 3.2 of Miller (2021) are met, from which the conclusions of this theorem follow immediately.

By Condition 10 and Equation 35, f_N has continuous third-order partial derivatives on Θ. Let E be the set from Condition 10. With probability 1, $f_N \to f$ uniformly on E (by Proposition 21 with $C = \bar{E}$) and (f'''_N) is uniformly bounded on E (by Proposition 22). Note that f is finite on Θ by Proposition 3. Thus, by Theorem 3.4 of Miller (2021), f' and f'' exist on E and $f''_N \to f''$ uniformly on E with probability 1. Since θ_* is a minimizer of f and $\theta_* \in E$, we know that $f'(\theta_*) = 0$ and $f''(\theta_*)$ is positive semidefinite; thus, $f''(\theta_*)$ is positive definite since it is invertible by assumption.

Case (a): Now, consider the case where Θ is compact. Then almost surely, $f_N \to f$ uniformly on Θ by Proposition 21 with $C = \Theta$. Since θ_* is a unique minimizer of f, we have $f(\theta) > f(\theta_*)$ for all $\theta \in \Theta \setminus \{\theta_*\}$. Let $H \subseteq E$ be an open set such that $\theta_* \in H$ and $\bar{H} \subseteq E$. We show that $\liminf_N \inf_{\theta\in\Theta\setminus\bar{H}} f_N(\theta) > f(\theta_*)$. Since $\Theta \setminus H$ is compact,

$$\inf_{\theta\in\Theta\setminus\bar{H}} f(\theta) - f(\theta_*) =: \epsilon > 0.$$

By uniform convergence, with probability 1, there exists N such that for all $N' > N$, $\sup_{\theta\in\Theta}|f_{N'}(\theta) - f(\theta)| \leq \epsilon/2$, and thus,

$$\inf_{\theta\in\Theta\setminus\bar{H}} f_{N'}(\theta) \geq \inf_{\theta\in\Theta\setminus\bar{H}} f(\theta) - \epsilon/2 = f(\theta_*) + \epsilon/2.$$

Hence, $\liminf_N \inf_{\theta\in\Theta\setminus\bar{H}} f_N(\theta) > f(\theta_*)$ almost surely. Applying Theorem 3.2 of Miller (2021), the conclusion of the theorem follows. Note that $f''_N(\theta_N) \to f''(\theta_*)$ a.s. since $\theta_N \to \theta_*$ and $f''_N \to f''$ uniformly on E.

Case (b): Alternatively, consider the case where Θ is open and f_N is convex on Θ. For all $\theta \in \Theta$, with probability 1, $f_N(\theta) \to f(\theta)$ (by Proposition 21 with $C = \{\theta\}$). However, we need to show that with probability 1, for all $\theta \in \Theta$, $f_N(\theta) \to f(\theta)$. We follow the argument in the proof of Theorem 6.3 of Miller (2021). Let W be a countable dense subset of Θ. Since W is countable, with probability 1, for all $\theta \in W$, $f_N(\theta) \to f(\theta)$. Since f_N is convex, then with probability 1, for all $\theta \in \Theta$, the limit $\tilde{f}(\theta) := \lim_N f_N(\theta)$ exists and is finite,

and $\tilde{f}$ is convex (Theorem 10.8 of Rockafellar, 1970). Since f_N is convex and $f(\theta)$ is finite, $f(\theta)$ is also convex. Since f and $\tilde{f}$ are convex, they are also continuous (Theorem 10.1 of Rockafellar, 1970). Continuous functions that agree on a dense subset of points must be equal. Thus, with probability 1, for all $\theta \in \Theta$, $f_N(\theta) \to f(\theta)$. Applying Theorem 3.2 of Miller (2021), the conclusion of the theorem follows. ■

Proof of Theorem 11 Our proof builds on Appendix D.3 of Barp et al. (2019), which establishes a central limit theorem for the KSD when the model is an exponential family. The outline of the proof is as follows. First, we establish bounds on s_{q_θ} and its derivatives, using the assumed bounds on $\nabla_x t(x)$ and $\nabla_x \log \lambda(x)$. Second, we establish that $f''(\theta)$ is positive definite and independent of θ, and that $f_N''(\theta)$ converges to it almost surely; from this, we conclude that $f''(\theta_*)$ is invertible and $f_N(\theta)$ is convex. These results rely on the convergence properties of U-statistics and on Sylvester's criterion.

The assumption that $\log \lambda(x)$ is continuously differentiable on $\mathcal{X}$ implies that $\lambda(x) > 0$ for $x \in \mathcal{X}$. Since $q_\theta(x) = \lambda(x) \exp(\theta^\top t(x) - \kappa(\theta))$, we have

$$\begin{aligned} s_{q_\theta}(x) &= \nabla_x \log \lambda(x) + (\nabla_x t(x))^\top \theta \\ \nabla_\theta s_{q_\theta}(x) &= (\nabla_x t(x))^\top \in \mathbb{R}^{d \times m} \\ \nabla_\theta^2 s_{q_\theta}(x) &= 0 \in \mathbb{R}^{d \times m \times m} \end{aligned}$$

where $(\nabla_x t(x))_{ij} = \partial t_i / \partial x_j$. Thus, $s_{q_\theta}(x)$ has continuous third-order partial derivatives with respect to θ, and Equations 41 and 42 are trivially satisfied. Equation 40 holds for all compact $C \subseteq \Theta$ since $\|\nabla_x \log \lambda(x)\|$ and $\|\nabla_x t(x)\|$ are continuous functions in $L^1(P_0)$ and

$$\begin{aligned} \|s_{q_\theta}(x)\| &= \|\nabla_x \log \lambda(x) + (\nabla_x t(x))^\top \theta\| \leq \|\nabla_x \log \lambda(x)\| + \|\nabla_x t(x)\| \|\theta\|, \\ \|\nabla_\theta s_{q_\theta}(x)\| &= \|\nabla_x t(x)\|. \end{aligned}$$

Hence, Condition 10 holds. By Equation 36 and Proposition 3,

$$f(\theta) = \frac{1}{T} \mathrm{NKSD}(p_0(x) \| q(x|\theta)) = \frac{1}{TK} \int_{\mathcal{X}} \int_{\mathcal{X}} u_\theta(x, y) P_0(dx) P_0(dy) \tag{100}$$

where $K := \int \int k(x, y) P_0(dx) P_0(dy)$. By Equation 57,

$$u_\theta(x, y) = \theta^\top B_2(x, y) \theta + B_1(x, y)^\top \theta + B_0(x, y) \tag{101}$$

where

$$\begin{aligned} B_2(x, y) &= (\nabla_x t(x))(\nabla_y t(y))^\top k(x, y), \\ B_1(x, y) &= (\nabla_y t(y))(\nabla_x \log \lambda(x)) k(x, y) + (\nabla_x t(x))(\nabla_y \log \lambda(y)) k(x, y) \\ &\quad + (\nabla_y t(y))(\nabla_x k(x, y)) + (\nabla_x t(x))(\nabla_y k(x, y)), \\ B_0(x, y) &= (\nabla_x \log \lambda(x))^\top (\nabla_y \log \lambda(y)) k(x, y) + (\nabla_y \log \lambda(y))^\top (\nabla_x k(x, y)) \\ &\quad + (\nabla_x \log \lambda(x))^\top (\nabla_y k(x, y)) + \mathrm{trace}(\nabla_x \nabla_y^\top k(x, y)). \end{aligned}$$

By Condition 7, $|k(x, y)|$, $\|\nabla_x k(x, y)\|$, and $\|\nabla_x \nabla_y^\top k(x, y)\|$ are bounded by a constant, say, $B < \infty$. Thus, it is straightforward to check that B_2, B_1, and B_0 belong to $L^1(P_0 \times P_0)$ since

$\|\nabla_x t(x)\|$ and $\|\nabla_x \log \lambda(x)\|$ are in $L^1(P_0)$. Further, $0 < K < \infty$ since $0 < k(x,y) \le B < \infty$ by assumption. Thus,

$$f(\theta) = \frac{1}{TK} \int \int \big(\theta^\top B_2(x,y)\theta + B_1(x,y)^\top \theta + B_0(x,y)\big) P_0(dx) P_0(dy) \in \mathbb{R}.$$

Since k is symmetric, $B_2(x,y)^\top = B_2(y,x)$. Hence, $\nabla_\theta(\theta^\top B_2(x,y)\theta) = (B_2(x,y) + B_2(y,x))\theta$, so by Fubini's theorem,

$$f'(\theta) = \frac{1}{TK} \int \int \big(2B_2(x,y)\theta + B_1(x,y)\big) P_0(dx) P_0(dy) \in \mathbb{R}^m,$$
$$f''(\theta) = \frac{2}{TK} \int \int B_2(x,y) P_0(dx) P_0(dy) \in \mathbb{R}^{m \times m}.$$

Here, differentiating under the integral sign is justified simply by linearity of the expectation. Note that $f''(\theta)$ is a symmetric matrix since $B_2(x,y)^\top = B_2(y,x)$. Next, to show $f''(\theta)$ is positive definite, let $v \in \mathbb{R}^m \setminus \{0\}$. By assumption, the rows of $\nabla_x t(x)$ are linearly independent with positive probability under P_0. Thus, there is a set $E \subseteq \mathcal{X}$ such that $P_0(E) > 0$ and $(\nabla_x t(x))^\top v \neq 0$ for all $x \in E$. Define $g(x) = (\nabla_x t(x))^\top v \, p_0(x) \in \mathbb{R}^d$. Then $\int_{\mathcal{X}} |g_i(x)| dx > 0$ for at least one i, and $\int_{\mathcal{X}} |g_i(x)| dx \le \|v\| \int_{\mathcal{X}} \|\nabla_x t(x)\| p_0(x) dx < \infty$ for all i. Thus,

$$v^\top f''(\theta) v = \frac{2}{TK} \int \int g(x)^\top g(y) k(x,y) dx dy = \frac{2}{TK} \sum_{i=1}^{d} \int \int g_i(x) g_i(y) k(x,y) dx dy > 0$$

since k is integrally strictly positive definite. Therefore, $f''(\theta)$ is positive definite. In particular, $f''(\theta_*)$ is invertible.

Finally, we show that with probability 1, for all N sufficiently large, $f_N(\theta)$ is convex. By Equations 35 and 101,

$$f_N(\theta) = \frac{1}{T} \frac{\sum_{i \neq j} \left[\theta^\top B_2(X^{(i)}, X^{(j)})\theta + B_1(X^{(i)}, X^{(j)})^\top \theta + B_0(X^{(i)}, X^{(j)})\right]}{\sum_{i \neq j} k(X^{(i)}, X^{(j)})}.$$

Thus,

$$f_N''(\theta) = \frac{2}{T} \frac{\sum_{i \neq j} B_2(X^{(i)}, X^{(j)})}{\sum_{i \neq j} k(X^{(i)}, X^{(j)})}.$$

By the strong law of large numbers for U-statistics (Theorem 5.4A of Serfling, 2009), we have $f_N''(\theta) \to f''(\theta)$ almost surely, since $\int_{\mathcal{X}} \int_{\mathcal{X}} \|B_2(x,y)\| P_0(dx) P_0(dy) < \infty$ and $0 < K < \infty$. For a symmetric matrix A, let $\lambda_*(A)$ denote the smallest eigenvalue. Since $\lambda_*(A)$ is a continuous function of the entries of A, we have $\lambda_*(f_N''(\theta)) \to \lambda_*(f''(\theta))$ a.s. as $N \to \infty$. Thus, with probability 1, for all N sufficiently large, $f_N''(\theta)$ is positive definite, and hence, f_N is convex. Further, for such N, since f_N is a quadratic function with positive definite Hessian, we have $M_N := \inf_{\theta \in \Theta} f_N(\theta) > -\infty$ and $z_N = \int_\Theta \exp(-N f_N(\theta)) \pi(\theta) d\theta \le \exp(-N M_N) < \infty$. ■

C.3 Proof of Theorem 12

To establish Theorem 12, we use the properties of U-statistics described in Chapter 5.5 of Serfling (2009). When the data distribution matches the model distribution, $\widehat{\text{NKSD}}$ converges more quickly than when it does not match; this same property was used by Liu et al. (2016) to develop a goodness-of-fit test based on the KSD.

Proof We first study the asymptotics of $f'_N(\theta_*)$. Denoting $\nabla_\theta\big|_{\theta=\theta_*} u_\theta$ by $\nabla_\theta u_{\theta_*}$ for brevity,

$$f'_N(\theta_*) = \frac{1}{T}\frac{\frac{1}{N(N-1)}\sum_{i\neq j}\nabla_\theta u_{\theta_*}(X^{(i)}, X^{(j)})}{\frac{1}{N(N-1)}\sum_{i\neq j} k(X^{(i)}, X^{(j)})}.$$

The denominator converges a.s. to a finite positive constant, as in the proof of Proposition 21. It is straightforward to verify that $\mathbb{E}_{X,Y\sim P_0}[\|\nabla_\theta u_{\theta_*}(X,Y)\|^2] < \infty$ since $s_{q_{\theta_*}}$ and $\nabla_\theta\big|_{\theta=\theta_*} s_{q_\theta}$ are in $L^2(P_0)$ by assumption. By Theorems 5.5.1A and 5.5.2 of Serfling (2009),

$$\frac{1}{N(N-1)}\sum_{i\neq j}\nabla_\theta u_{\theta_*}(X^{(i)}, X^{(j)}) - \mathbb{E}_{X,Y\sim P_0}[\nabla_\theta u_{\theta_*}(X,Y)] = O_{P_0}(N^{-1/2}).$$

Further, by the Leibniz integral rule (Folland, 1999, Theorem 2.27),

$$\mathbb{E}_{X,Y\sim P_0}[\nabla_\theta u_{\theta_*}(X,Y)] = \nabla_\theta\big|_{\theta=\theta_*}\mathbb{E}_{X,Y\sim P_0}[u_\theta(X,Y)] = T\,\mathbb{E}_{X,Y\sim P_0}[k(X,Y)]f'(\theta_*) = 0,$$

using the fact that $f'(\theta_*) = 0$ since θ_* is a minimizer of f. Thus,

$$f'_N(\theta_*) = O_{P_0}(N^{-1/2}). \tag{102}$$

Next, we examine the convergence of θ_N to θ_*. For all N sufficiently large, $f'_N(\theta_N) = 0$ by Theorem 9 (part 1), and thus, by Taylor's theorem,

$$0 = f'_N(\theta_N) = f'_N(\theta_*) + f''_N(\theta_N^+)(\theta_N - \theta_*),$$

where θ_N^+ is on the line between θ_N and θ_*. As in the proof of Theorem 9, $f''_N \to f''$ uniformly on the set E defined in Condition 10. Thus, since f''_N is continuous on E and $\theta_N^+ \to \theta_*$,

$$f''_N(\theta_N^+) \xrightarrow[N\to\infty]{\text{a.s.}} f''(\theta_*). \tag{103}$$

In particular, $f''_N(\theta_N^+)$ is invertible for all N sufficiently large, since $f''(\theta_*)$ is invertible by assumption. Hence,

$$\theta_N - \theta_* = -f''_N(\theta_N^+)^{-1} f'_N(\theta_*), \tag{104}$$

and therefore, by Equation 102,

$$\|\theta_N - \theta_*\| \leq \|f''_N(\theta_N^+)^{-1}\|\|f'_N(\theta_*)\| = O_{P_0}(N^{-1/2}). \tag{105}$$

This result matches Theorem 4 in Barp et al. (2019). By Taylor's theorem,

$$f_N(\theta_*) - f_N(\theta_N) = f'_N(\theta_N)^\top(\theta_* - \theta_N) + \frac{1}{2}(\theta_* - \theta_N)^\top f''_N(\theta_N^{++})(\theta_* - \theta_N)$$

$$= \frac{1}{2}(\theta_* - \theta_N)^\top f_N''(\theta_N^{++})(\theta_* - \theta_N)$$

for all N sufficiently large, where θ_N^{++} is on the line between θ_N and θ_*. Therefore, using the same reasoning as for Equations 103 and 105,

$$|f_N(\theta_*) - f_N(\theta_N)| \leq \frac{1}{2}\|f_N''(\theta_N^{++})\|\|\theta_* - \theta_N\|^2 = O_{P_0}(N^{-1}). \tag{106}$$

This proves the first part of the theorem (Equation 43). Next, consider $f_N(\theta_*) - f(\theta_*)$. Recall that

$$f_N(\theta_*) = \frac{1}{T}\frac{\frac{1}{N(N-1)}\sum_{i\neq j} u_{\theta_*}(X^{(i)}, X^{(j)})}{\frac{1}{N(N-1)}\sum_{i\neq j} k(X^{(i)}, X^{(j)})}.$$

It is straightforward to verify that $\mathbb{E}_{X,Y\sim P_0}[|u_{\theta_*}(X,Y)|^2] < \infty$ since $s_{q_{\theta_*}}$ is in $L^2(P_0)$. By Theorems 5.5.1A and 5.5.2 of Serfling (2009),

$$\frac{1}{N(N-1)}\sum_{i\neq j} u_{\theta_*}(X^{(i)}, X^{(j)}) - \mathbb{E}_{X,Y\sim P_0}[u_{\theta_*}(X,Y)] = O_{P_0}(N^{-1/2}).$$

Similarly, since k is bounded,

$$\frac{1}{N(N-1)}\sum_{i\neq j} k(X^{(i)}, X^{(j)}) - \mathbb{E}_{X,Y\sim P_0}[k(X,Y)] = O_{P_0}(N^{-1/2}).$$

It is straightforward to check that the second part of the theorem (Equation 44) follows.

For the third part, our argument follows that of the proof of Theorem 4.1 of Liu et al. (2016). Suppose $\text{NKSD}(p_0(x)\|q(x|\theta_*)) = 0$, and note that $P_0(x) = Q_{\theta_*}(x)$ by Proposition 4. Given a differentiable function $g : \mathbb{R}^d \to \mathbb{R}^d$, define $\nabla_x^\top g(x) := \sum_{i=1}^d \partial g_i(x)/\partial x_i$. Then

$$\begin{aligned}
\mathbb{E}_{X\sim P_0}[u_{\theta_*}(X,y)] &= s_{p_0}(y)^\top \int_{\mathcal{X}} \Big((\nabla_x p_0(x))k(x,y) + p_0(x)(\nabla_x k(x,y))\Big)dx \\
&\quad + \int_{\mathcal{X}} \Big((\nabla_x p_0(x))^\top \nabla_y k(x,y) + p_0(x)(\nabla_x^\top \nabla_y k(x,y))\Big)dx \\
&= s_{p_0}(y)^\top \int_{\mathcal{X}} \nabla_x\big(p_0(x)k(x,y)\big)dx + \int_{\mathcal{X}} \nabla_x^\top \nabla_y (p_0(x)k(x,y))dx.
\end{aligned} \tag{107}$$

The first term on the right-hand side of Equation 107 is zero since, by assumption, k is in the Stein class of P_0 (Condition 2). The second term is also zero since, by the Leibniz integral rule (Folland, 1999, Theorem 2.27), $\int \nabla_y^\top \nabla_x (p_0(x)k(x,y))dx = \nabla_y^\top \int \nabla_x (p_0(x)k(x,y))dx$, which again equals zero because k is in the Stein class of P_0. Therefore, $\mathbb{E}_{X\sim P_0}[u_{\theta_*}(X,y)] = 0$ for all $y \in \mathcal{X}$, and in particular, the variance of this expression is also zero: $\mathbb{V}_{Y\sim P_0}[\mathbb{E}_{X\sim P_0}[u_{\theta_*}(X,Y)]] = 0$. By Theorem 5.5.2 of Serfling (2009), it follows that

$$\frac{1}{N(N-1)}\sum_{i\neq j} u_{\theta_*}(X^{(i)}, X^{(j)}) = O_{P_0}(N^{-1}) \tag{108}$$

since $\mathbb{E}_{X,Y\sim P_0}[u_{\theta_*}(X,Y)] = 0$. Although Serfling (2009) requires $\mathbb{V}_{X,Y\sim P_0}[u_{\theta_*}(X,Y)] > 0$, Equation 108 holds trivially if $\mathbb{V}_{X,Y\sim P_0}[u_{\theta_*}(X,Y)] = 0$. As before, since the denominator of $f_N(\theta_*)$ converges a.s. to a finite positive constant, we have that $f_N(\theta_*) = O_{P_0}(N^{-1})$. Equation 45 follows since $f(\theta_*) = 0$ when $\text{NKSD}(p_0(x)\|q(x|\theta_*)) = 0$. ■

C.4 Proof of Theorem 17

Proof Applying Theorem 9 (part 3) to each foreground model $j \in \{1, 2\}$, we have

$$\log z_{j,N} + N f_{j,N}(\theta_{j,N}) - \log \pi(\theta_{j,*}) + \log |\det f_j''(\theta_{j,*})|^{1/2} - \frac{1}{2} m_{\mathcal{F}_j,j} \log(2\pi/N) \xrightarrow[N\to\infty]{\text{a.s.}} 0.$$

Since $\mathcal{K}_{j,N} = (2\pi/N)^{m_{\mathcal{B}_j}/2} z_{j,N}$, this implies

$$\log \mathcal{K}_{j,N} + N f_{j,N}(\theta_{j,N}) - \frac{1}{2}(m_{\mathcal{F}_j,j} + m_{\mathcal{B}_j}) \log(2\pi/N) + C_j \xrightarrow[N\to\infty]{\text{a.s.}} 0$$

where C_j is a constant that does not depend on N. Hence,

$$\begin{aligned} \log \frac{\mathcal{K}_{1,N}}{\mathcal{K}_{2,N}} &+ N(f_{1,N}(\theta_{1,N}) - f_{2,N}(\theta_{2,N})) \\ &- \frac{1}{2}(m_{\mathcal{F}_1,1} + m_{\mathcal{B}_1} - m_{\mathcal{F}_2,2} - m_{\mathcal{B}_2}) \log(2\pi/N) + C_1 - C_2 \xrightarrow[N\to\infty]{\text{a.s.}} 0. \end{aligned} \tag{109}$$

By Theorem 12, $f_{j,N}(\theta_{j,N}) \xrightarrow{P_0} f_j(\theta_{j,*})$, and therefore,

$$\frac{1}{N} \log \frac{\mathcal{K}_{1,N}}{\mathcal{K}_{2,N}} + f_1(\theta_{1,*}) - f_2(\theta_{2,*}) \xrightarrow[N\to\infty]{P_0} 0.$$

Plugging in the definition of f_j (Equation 36), this proves part 1 of the theorem.

For part 2, suppose $f_1(\theta_{1,*}) = f_2(\theta_{2,*}) = 0$ and $m_{\mathcal{B}_2} - m_{\mathcal{B}_1}$ does not depend on N. Then by Theorem 12, $f_{j,N}(\theta_{j,N}) = O_{P_0}(N^{-1})$. Using this in Equation 109, we have

$$\frac{1}{\log N} \log \frac{\mathcal{K}_{1,N}}{\mathcal{K}_{2,N}} + \frac{1}{2}(m_{\mathcal{F}_1,1} + m_{\mathcal{B}_1} - m_{F_2,2} - m_{\mathcal{B}_2}) \xrightarrow[N\to\infty]{P_0} 0. \tag{110}$$

For part 3, suppose $f_1(\theta_{1,*}) = f_2(\theta_{2,*})$ and $m_{\mathcal{B}_j} = c_{\mathcal{B}_j}\sqrt{N}$. Then by Theorem 12, $f_{j,N}(\theta_{j,N}) = f_j(\theta_{j,*}) + O_{P_0}(N^{-1/2})$. Using this in Equation 109, we have

$$\frac{1}{\sqrt{N}\log N} \log \frac{\mathcal{K}_{1,N}}{\mathcal{K}_{2,N}} + \frac{1}{2}(c_{\mathcal{B}_1} - c_{\mathcal{B}_2}) \xrightarrow[N\to\infty]{P_0} 0. \tag{111}$$

■

Appendix D. Additional probabilistic PCA details

D.1 Optimizing the NKSD

Computing the Laplace or BIC approximation to the SVC requires finding the minimizer of $\widehat{\text{NKSD}}(p_0(x)\|q(x|\theta))$ with respect to θ. In this section, we describe how components of the NKSD can be pre-computed to speed up this optimization process. The generative model used for pPCA can be rewritten using the properties of multivariate normal distributions as

$$X \sim \mathcal{N}(0, HH^\top + vI_d). \tag{112}$$

The Stein score function for the pPCA model is then

$$s_{q_\theta}(x) = \nabla_x \log q(x|H, v) = -(HH^\top + vI_d)^{-1}x.$$

Define the matrices

$$K_{ij} := \mathbb{I}(i \neq j)\, k(X^{(i)}, X^{(j)}),$$
$$\dot{K}_{jb} := \sum_{i=1}^{N} \mathbb{I}(i \neq j)\, \frac{\partial k}{\partial x_b}(X^{(i)}, X^{(j)}),$$

where $\mathbb{I}(E)$ is the indicator function, which equals 1 when E is true and is 0 otherwise. Define the scalars

$$\bar{K} := \sum_{i,j=1}^{N} K_{ij},$$
$$\ddot{K} := \sum_{i,j=1}^{N} \sum_{b=1}^{d} \mathbb{I}(i \neq j)\, \frac{\partial^2 k}{\partial x_b \partial y_b}(X^{(i)}, X^{(j)}).$$

Letting $X \in \mathbb{R}^{N\times d}$ be the data matrix, the NKSD can be written as

$$\widehat{\mathrm{NKSD}}(p_0(x)\|q(x|H, v)) = \frac{1}{\bar{K}}\big[\,\mathrm{trace}(X^\top K X (HH^\top + vI_d)^{-1}(HH^\top + vI_d)^{-1}) - 2\,\mathrm{trace}(X^\top \dot{K}(HH^\top + vI_d)^{-1}) + \ddot{K}\big],$$

where we have used the fact that the kernel is symmetric. The terms $X^\top K X$ and $X^\top \dot{K}$ are the only ones that include sums over the entire data set; these can be pre-computed, before optimizing the parameters H and v.

To compute the matrix inversion $(HH^\top + vI_d)^{-1}$ we follow the strategy of Minka (2001),

$$\begin{aligned}(HH^\top + vI_d)^{-1} - v^{-1}I_d &= (HH^\top + vI_d)^{-1}(I_d - v^{-1}(HH^\top + vI_d)) \\ &= -(HH^\top + vI_d)^{-1}HH^\top v^{-1} \\ &= -(U(L - vI_k)U^\top + vI_d)^{-1}U(L - vI_k)U^\top v^{-1}.\end{aligned}$$

Thus, applying the Woodbury matrix identity and using $I_d U = U = UI_kI_k = UI_kU^\top U$,

$$\begin{aligned}(HH^\top + vI_d)^{-1} - v^{-1}I_d &= -\big[v^{-1}I_d - v^{-2}U\big((L - vI_k)^{-1} + v^{-1}\big)^{-1}U^\top\big]U(L - vI_k)U^\top v^{-1} \\ &= -U[v^{-1}I_k - v^{-2}((L - v)^{-1} + v^{-1})^{-1}](L - vI_k)U^\top v^{-1} \\ &= -UL^{-1}(L - vI_k)U^\top v^{-1} \\ &= U(L^{-1} - v^{-1}I_k)U^\top.\end{aligned}$$

Therefore,

$$(HH^\top + vI_d)^{-1} = U(L^{-1} - v^{-1}I_k)U^\top + v^{-1}I_d.$$

Computing L^{-1} is trivial since the matrix is diagonal. Returning to the NKSD we have

$$\begin{aligned}
&\widehat{\mathrm{NKSD}}(p_0(x)\|q(x|U,L,v)) \\
&\quad= \frac{1}{\bar{K}}\Big[\operatorname{trace}\left(X^\top K X[U(L^{-1}-v^{-1}I_k)^2U^\top + 2v^{-1}U(L^{-1}-v^{-1}I_k)U^\top + v^{-2}I_d]\right) \\
&\qquad\quad - 2\operatorname{trace}\left(X^\top \dot{K}[U(L^{-1}-v^{-1}I_k)U^\top + v^{-1}I_d]\right) + \ddot{K}\Big] \\
&\quad= \frac{1}{\bar{K}}\Big[\operatorname{trace}\left(U^\top X^\top K X U(L^{-1}-v^{-1}I_k)^2\right) \\
&\qquad\quad + \operatorname{trace}\left(U^\top[2v^{-1}X^\top K X - 2X^\top\dot{K}]U(L^{-1}-v^{-1}I_k)\right) \\
&\qquad\quad + v^{-1}\operatorname{trace}\left(v^{-1}X^\top K X - 2X^\top\dot{K}\right) + \ddot{K}\Big].
\end{aligned}$$

We optimized U, L and v using the trust region method implemented in pymanopt (Townsend et al., 2016).

D.2 Data selection with the SVC

We used the approximate optimum technique in Section 2.3.3 to estimate the SVC for different foreground subspaces. Following Section A.2, we used the factored IMQ kernel with $\beta = -0.5$ and $c = 1$.

We focused on foreground subspaces that correspond to subsets of the data dimensions. More specifically, recall that $X_{\mathcal{F}} = V^\top X$; then, we impose the restriction that each column of V is a standard basis vector $e^{(b)} \in \mathbb{R}^d$, where $e^{(b)}_b = 1$ and $e^{(b)}_{b'} = 0$ for $b' \neq b$. A subspace $\mathcal{X}_{\mathcal{F}}$ is then characterized by the set of included dimensions $S_{\mathcal{F}} \subseteq \{1,\ldots,d\}$. The marginal distribution of the model $q(x_{\mathcal{F}}|H,v)$ is now straightforward to compute based on Equation 112 and the properties of multivariate normals:

$$X_{\mathcal{F}} \sim \mathcal{N}(0, H_{S_{\mathcal{F}}}H_{S_{\mathcal{F}}}^\top + vI_{|S_{\mathcal{F}}|})$$

where $H_{S_{\mathcal{F}}}$ is the submatrix consisting of rows of H indexed by $S_{\mathcal{F}}$, and $|S_{\mathcal{F}}|$ is the size of the set $S_{\mathcal{F}}$.

In the projected model, some of the parameters are nuisance variables with no contribution to the likelihood. Since the dimension of a $d \times k$ matrix on the Stiefel manifold is $dk - k(k+1)/2$, the total dimension of the foreground model (including contributions from parameters U, L and v) is $m_{\mathcal{F}} = |S_{\mathcal{F}}|k - k(k+1)/2 + k + 1$, assuming $|S_{\mathcal{F}}| \geq k$.

Code is available at `https://github.com/EWeinstein/data-selection`.

D.3 Calibration

The T hyperparameter was calibrated as in Section A.1. In detail, we sampled 10 independent true parameter values from the prior, with $\alpha = 1$ and $d = 6$. (We used a slightly less disperse prior than during inference, where we set $\alpha = 0.1$, to avoid numerical instabilities in the $\hat{T}$ estimate.) Then, for each of the true parameter values, we simulated $N = 2000$ datapoints. For each simulated true parameter value, we tracked the trend in the $\hat{T}$ estimator (Equation 55) with increasing N (Figure 11). The median estimated T value at $N = 2000$ was 0.052 across the 10 runs.

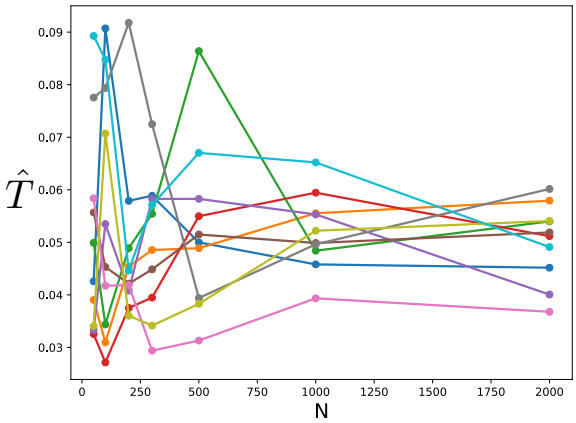


Figure 11: Estimated T for increasing number of data samples, for 10 independent parameter samples from the prior. The median value at $N = 2000$ is $\hat{T} = 0.052$.

D.4 Pólya tree model

In this section, we describe the Pólya tree model (Ferguson, 1974; Mauldin et al., 1992; Lavine, 1992) following the construction of Berger and Guglielmi (2001). Let $\underline{\epsilon}_n := (\epsilon_1, \ldots, \epsilon_n)$ denote a vector of length n, where each $\epsilon_j \in \{0, 1\}$. Each $\underline{\epsilon}_n$ vector indexes an interval in $\mathbb{R}$, given by

$$B_{\underline{\epsilon}_n} := \Big[\tilde{F}^{-1}\big(\textstyle\sum_{j=1}^{n} \epsilon_j/2^j\big),\ \tilde{F}^{-1}\big(\textstyle\sum_{j=1}^{n} \epsilon_j/2^j + 1/2^n\big)\Big],$$

where $\tilde{F}^{-1}$ is the inverse c.d.f. of some probability distribution. For all $n \in \{0, 1, 2, \ldots\}$ and all $\underline{\epsilon}_n \in \{0, 1\}^n$, let

$$Y_{\underline{\epsilon}_n} \sim \text{Beta}(\xi_{\underline{\epsilon}_n 0}, \xi_{\underline{\epsilon}_n 1}),$$

where the ξ's are hyperparameters. We say that a random variable $X \in \mathbb{R}$ is distributed according to a Pólya tree model if

$$P(X \in B_{\underline{\epsilon}_n}) = \prod_{j=1}^{n} (Y_{\underline{\epsilon}_{j-1}})^{\mathbb{I}(\epsilon_j = 0)} (1 - Y_{\underline{\epsilon}_{j-1}})^{\mathbb{I}(\epsilon_j = 1)},$$

where $\mathbb{I}(E)$ is the indicator function, which equals 1 when E is true and is 0 otherwise. We follow Berger and Guglielmi (2001) and use

$$\begin{aligned}
\mu(B_{\underline{\epsilon}_n}) &:= F\big(\tilde{F}^{-1}\big(\textstyle\sum_{j=1}^{n} \epsilon_j/2^j + 1/2^n\big)\big) - F\big(\tilde{F}^{-1}\big(\textstyle\sum_{j=1}^{n} \epsilon_j/2^j\big)\big), \\
\rho(\underline{\epsilon}_n) &:= \frac{1}{\eta}\left(\frac{f(\tilde{F}^{-1}(\sum_{j=1}^{n} \epsilon_j/2^j + 1/2^{n+1}))}{\mu(B_{\underline{\epsilon}_n})}\right)^2, \\
\xi_{\underline{\epsilon}_n 0} &:= \rho(\underline{\epsilon}_n)\sqrt{\frac{\mu(B_{\underline{\epsilon}_n 0})}{\mu(B_{\underline{\epsilon}_n 1})}}, \\
\xi_{\underline{\epsilon}_n 1} &:= \rho(\underline{\epsilon}_n)\sqrt{\frac{\mu(B_{\underline{\epsilon}_n 1})}{\mu(B_{\underline{\epsilon}_n 0})}},
\end{aligned}$$

where F and f are the c.d.f. and p.d.f. respectively of some probability distribution, and $\eta > 0$ is a scale hyperparameter. We denote this complete model as $X \sim \text{PolyaTree}(F, \tilde{F}, \eta)$.

D.5 Data sets and preprocessing

We downloaded two publicly available data sets. The first data set was from human peripheral blood mononuclear cells (PBMCs), available at: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k`. This is a standard data set used in the tutorials for Seurat (Stuart et al., 2019) and Scanpy (Wolf et al., 2018), for example. The second was taken from a dissociated extranodal marginal zone B-cell tumor, specifically a mucosa-associated lymphoid tissue (MALT) tumor: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/malt_10k_protein_v3`.

We pre-processed the data using Scprep (Gigante et al., 2020), following its example: we normalized the total expression of each cell to match the median total expression in the data set, to account for variability in library size, and then square-root transformed the resulting normalized counts.

Appendix E. Additional glass model details

E.1 Glass model inference

We place a standard normal prior on each entry of H_j and a Laplace prior on each entry of $J_{jj'}$ with scale 0.1 to encourage sparsity. To enforce that $\mu \geq 0$ (since scRNAseq counts are nonnegative) and $\tau > 0$, we place priors on a transformed version of these parameters, as follows:

$$\begin{aligned}
\tilde{\mu} &\sim \mathcal{N}(0, 1) \\
\mu &= \log(1 + \exp(\tilde{\mu})) \\
\tilde{\tau} &\sim \mathcal{N}(0, 1) \\
\tau &= \log(1 + \exp(\tilde{\tau})) + 1.
\end{aligned}$$

For posterior inference, we employ a mean-field variational approximation: independent normal distributions for the entries of H_j, normal distributions for $\tilde{\mu}$ and $\tilde{\tau}$, and Laplace distributions for each entry of $J_{jj'}$. We use the factored IMQ kernel for the NKSD, with $\beta = -0.5$ and $c = 1$.

To optimize the variational approximation (Equation 14), we construct stochastic estimates of its gradient. At each optimization step, the expectation $\mathbb{E}_{r_\zeta}\left[\widehat{\text{NKSD}}(p_0(x_\mathcal{F}) \| q(x_\mathcal{F}|\theta))\right]$ is estimated using a minibatch of 200 randomly selected datapoints and a single sample from the variational approximation r_ζ. The rest of the variational inference algorithm follows standard practice in stochastic variational inference, as implemented in Pyro: automatic differentiation to compute gradients, reparameterization estimators for Monte Carlo expectations over the variational distribution, and the Adam optimizer (Kingma and Ba, 2015; Bingham et al., 2019).

We also used stochastic optimization to perform data selection, as follows. Let $I = (I_1, \ldots, I_d)^\top$ be an indicator variable that specifies for each gene j whether it is included in the foreground subspace ($I_j = 1$) or not ($I_j = 0$). We place a distribution on I such that

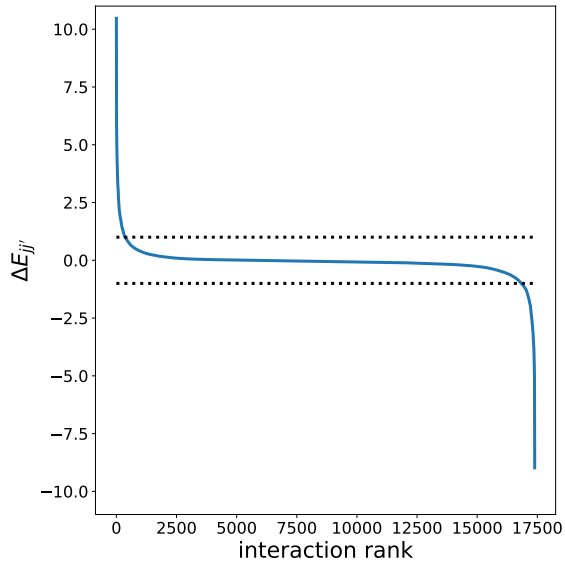


Figure 12: Posterior mean interaction energies $\Delta E_{jj'}$ for all selected genes, sorted. Dotted lines show the thresholds for strong interactions (set by visual inspection).

$I_j \sim \text{Bernoulli}(1/(1+\exp(-\phi_j)))$ for $j = 1, \ldots, d$ independently. Then, to perform data selection over all possible subsets of genes, we optimize

$$\text{argmax}_\phi \, \mathbb{E}(\mathcal{K}(I) \mid \phi) \tag{113}$$

where the expectation is taken with respect to I, where $\mathcal{K}(I)$ is the (estimated) SVC when genes with $I_j = 1$ are included in the foreground space, and $\phi = (\phi_1, \ldots, \phi_d)^\top \in \mathbb{R}^d$ is a vector of log-odds. This stochastic approach to discrete optimization has been used extensively in reinforcement learning and related fields. We use the Leave-One-Out REINFORCE (LOORF) estimator as described in Section 2.1 of Dimitriev and Zhou (2021) to estimate gradients of ϕ, using 8 samples per step.

We interleave updates to the variational approximation and to ϕ, using the Adam optimizer with step size 0.01 for each. We ran the procedure with 4 random initial seeds, taking the result with the largest final estimated SVC. We halt optimization using the stopping rule proposed in Grathwohl et al. (2020), stopping when the estimated mean minus the estimated variance of the SVC begins to decrease, based on the average over 2000 steps.

Code is available at `https://github.com/EWeinstein/data-selection`.

E.2 Data sets and preprocessing

In addition to the two data sets in D.5, we also explored a data set of E18 mouse neurons: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/neuron_10k_v3`.

We preprocessed each data set using Scprep (Gigante et al., 2020) in the same way as in Section D.5. After preprocessing, we used the top 200 most highly expressed genes from

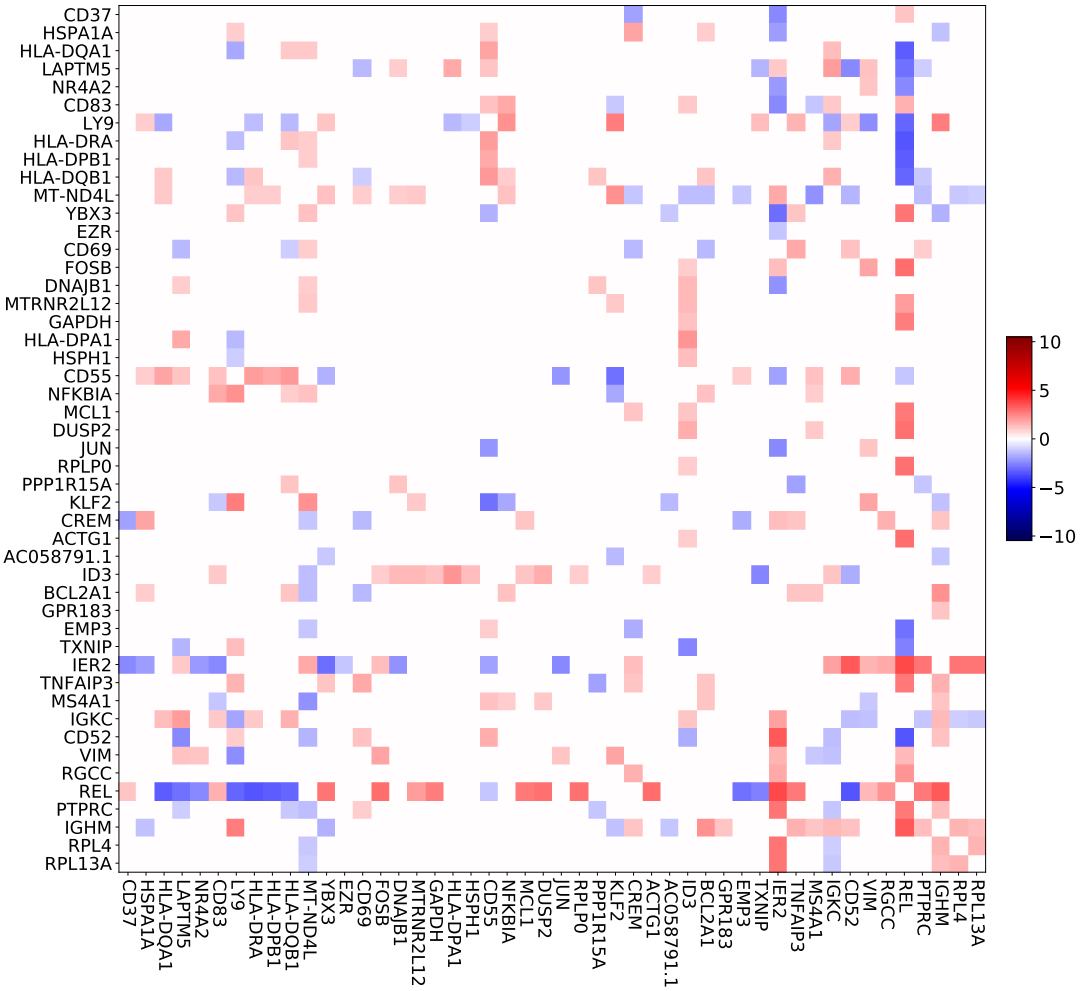
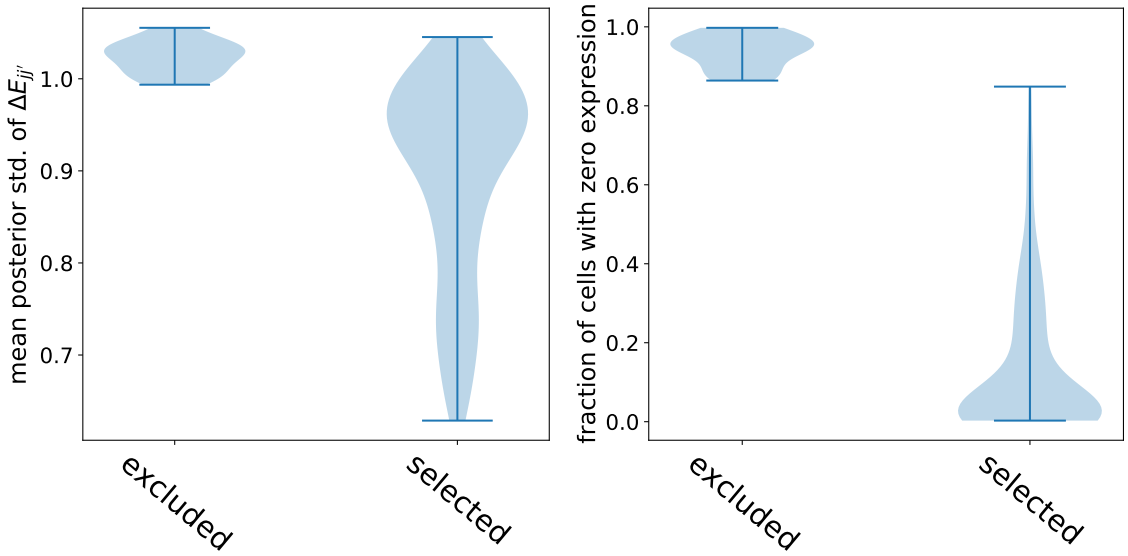


Figure 13: Posterior mean interaction energies $\Delta E_{jj'}$ for the glass model applied to all 200 genes in the MALT data set (rather than the selected 187). Genes shown are the same as in Figure 8, for visual comparison.

among the top 500 most variable genes, according to the Scprep variability score. We log transform the counts, that is we define $x_{ij} = \log(1 + c_{ij})$ where c_{ij} is the expression count for gene j in cell i.



(a) (b)

Figure 14: Comparison of the 187 selected genes and 13 excluded genes using data selection. (a) Violin plot of $\bar{\sigma}_j$ over all excluded and selected genes j, respectively, when applying the model to all 200 genes, where $\bar{\sigma}_j$ is the mean posterior standard deviation of the interaction energies $\Delta E_{jj'}$ for gene j, that is, $\bar{\sigma}_j := \frac{1}{d-1} \sum_{j' \neq j} \mathrm{std}(\Delta E_{jj'} \mid \mathrm{data})$. (b) Violin plot of f_j over all excluded and selected genes j, respectively, where f_j is the fraction of cells with count equal to zero for gene j. The data selection procedure excluded all genes with more than 85% zeros and selected all genes with fewer than 85% zeros.

References

Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, July 2019.

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, Lester Mackey, Chris J Oates, Gesine Reinert, and Yvik Swan. Stein's method meets statistics: A review of some recent developments. *arXiv preprint arXiv:2105.03481*, May 2021.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516, 2008.

Alessandro Barp, Francois-Xavier Briol, Andrew B Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. *arXiv preprint arXiv:1906.08283*, June 2019.

Andrew R Barron. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, 17 (1):107–124, 1989.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153), 2018.

James O Berger and Alessandra Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96(453):174–184, 2001.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.

Pier G Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *J. R. Stat. Soc. Series B Stat. Methodol.*, 78(5):1103–1130, November 2016.

David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1(1):203–232, 2014.

Ariel Caticha. Relative entropy and inductive inference. *AIP Conference Proceedings*, 707 (1):75–96, 2004.

Ariel Caticha. Entropic inference. *AIP Conference Proceedings*, 1305(1):20–29, 2011.

Haifen Chen, Jing Guo, Shital K Mishra, Paul Robson, Mahesan Niranjan, and Jie Zheng. Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development. *Bioinformatics*, 31(7):1060–1066, April 2015.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, pages 2606–2615, 2016.

A Philip Dawid. Posterior model probabilities. In Prasanta S Bandyopadhyay and Malcolm R Forster, editors, *Philosophy of Statistics*, volume 7, pages 607–630. North-Holland, Amsterdam, January 2011.

Charlotte M de Winde, Sharon Veenbergen, Ken H Young, Zijun Y Xu-Monette, Xiao-Xiao Wang, Yi Xia, Kausar J Jabbar, Michiel van den Brand, Alie van der Schaaf, Suraya Elfrink, Inge S van Houdt, Marion J Gijbels, Fons A J van de Loo, Miranda B Bennink, Konnie M Hebeda, Patricia J T A Groenen, J Han van Krieken, Carl G Figdor, and Annemiek B van Spriel. Tetraspanin CD37 protects against the development of B cell lymphoma. *The Journal of Clinical Investigation*, 126(2):653–666, February 2016.

Alek Dimitriev and Mingyuan Zhou. ARMS: Antithetic-REINFORCE-Multi-Sample gradient for binary variables. In *International Conference on Machine Learning (ICML)*, 2021.

Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205, April 2005.

Kjell A Doksum and Albert Y Lo. Consistent and robust Bayes procedures for location based on partial information. *The Annals of Statistics*, 18(1):443–453, 1990.

David Duvenaud, Daniel Eaton, Kevin Murphy, and Mark Schmidt. Causal learning without DAGs. In *NeurIPS workshop on causality*, 2008.

Thomas S Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629, July 1974.

Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 1999.

Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303 (5659):799–805, February 2004.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *J. Comput. Biol.*, 7(3-4):601–620, 2000.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

J K Ghosh and R V Ramamoorthi. *Bayesian Nonparametrics*. Series in Statistics. Springer, 2003.

Scott Gigante, Daniel Burkhardt, Daniel Dager, Jay Stanley, and Alexander Tong. scprep. `https://github.com/KrishnaswamyLab/scprep`, 2020.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A Swiss Army infinitesimal jackknife. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1139–1147. PMLR, 2019.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, pages 1292–1301, Sydney, NSW, Australia, 2017.

Will Grathwohl, Kuan-Chieh Wang, Jorn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning (ICML)*, 2020.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

László Györfi and Edward C Van Der Meulen. A consistent goodness of fit test based on the total variation distance. In George Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 631–645. Springer Netherlands, Dordrecht, 1991.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, October 2018.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Han Hong and Bruce Preston. Nonnested model selection criteria. 2005.

Peter J Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

Jonathan H Huggins and Lester Mackey. Random feature stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jonathan H Huggins and Jeffrey W Miller. Reproducible model selection using bagged posteriors. *arXiv preprint arXiv:2007.14845*, 2021.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5(9), September 2010.

Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert. Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*, 2017.

Jack Jewson, Jim Q Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

Wenxin Jiang and Martin A Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207–2231, 2008.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, April 2014.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free. In *ICLR Workshop: Deep Reinforcement Learning Meets Structured Prediction*, 2019.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:1–45, January 2017.

Michael Lavine. Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235, 1992.

John R Lewis, Steven N MacEachern, and Yoonkyung Lee. Bayesian restricted likelihood methods: Conditioning on insufficient statistics in Bayesian regression. *Bayesian Analysis*, 1(1):1–38, 2021.

Bruce G Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239, 1988.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

Qiang Liu, Jason D Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, volume 33, pages 276–284, 2016.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris J. Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):997–1022, 2022.

Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru S H Ko, Shigeru B H Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, 33(15):2314–2321, August 2017.

R Daniel Mauldin, William D Sudderth, and S C Williams. Polya trees and random distributions. *The Annals of Statistics*, 20(3):1203–1221, 1992.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Thomas Minka. Old and new matrix algebra useful for statistics, 2000.

Thomas P Minka. Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 598–604, 2001.

Victoria Moignard, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti, Shin-Ichi Nishikawa, Nir Piterman, Valerie Kouskoff, Fabian J Theis, Jasmin Fisher, and Berthold Göttgens. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 33(3): 269–276, March 2015.

Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16:241, November 2015.

Jim Pitman. Combinatorial stochastic processes. Technical Report 621, Dept of Statistics, UC Berkeley, 2002.

Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979, 2017.

Danilo Jimenez Rezende. Short notes on divergence measures. July 2018.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, September 2009.

Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, June 2013.

Stephane Shao, Pierre E Jacob, Jie Ding, and Vahid Tarokh. Bayesian model comparison with the Hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, pages 1–24, September 2018.

Zakary S Singer, John Yong, Julia Tischler, Jamie A Hackett, Alphan Altinok, M Azim Surani, Long Cai, and Michael B Elowitz. Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Molecular Cell*, 55(2):319–331, July 2014.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R G Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert R G Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, September 2008.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A Python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(1):4755–4759, 2016.

David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.e27, July 2018.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, January 2011.

Isabella Verdinelli and Larry Wasserman. Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *The Annals of Statistics*, 26(4):1215–1241, August 1998.

Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 57(2):307–333, 1989.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, February 2018.

Zijun Y Xu-Monette, Ling Li, John C Byrd, Kausar J Jabbar, Ganiraju C Manyam, Charlotte Maria de Winde, Michiel van den Brand, Alexandar Tzankov, Carlo Visco, Jing Wang, Karen Dybkaer, April Chiu, Attilio Orazi, Youli Zu, Govind Bhagat, Kristy L Richards, Eric D Hsi, William W L Choi, Jooryung Huh, Maurilio Ponzoni, Andrés J M Ferreri, Michael B Møller, Ben M Parsons, Jane N Winter, Michael Wang, Frederick B Hagemeister, Miguel A Piris, J Han van Krieken, L Jeffrey Medeiros, Yong Li, Annemiek B van Spriel, and Ken H Young. Assessment of CD37 B-cell antigen and cell of origin significantly improves risk prediction in diffuse large B-cell lymphoma. *Blood*, 128 (26):3083–3100, December 2016.

Daniel Yekutieli. Adjusted Bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):515–541, 2012.

In-Kwon Yeo and Richard A Johnson. A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. *Statistics & Probability Letters*, 51 (1):63–69, 2001.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006a.

Tong Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006b.