Robust model selection using likelihood as data

Jongwoo Choi[*1], Neil A. Spencer[1], and Jeffrey W. Miller[2]

[1]*Department of Statistics, University of Connecticut*

[2]*Department of Biostatistics, Harvard T.H. Chan School of Public Health*

Abstract

Model selection is a central task in statistics, but standard methods are not robust in misspecified settings where the true data-generating process (DGP) is not in the set of candidate models. The key limitation is that existing methods— including information criteria and Bayesian posteriors—do not quantify uncertainty about how well each candidate model approximates the true DGP. In this paper, we introduce a novel approach to model selection based on modeling the likelihood values themselves. Specifically, given K candidate models and n observations, we view the $n \times K$ matrix of negative log-likelihood values as a random data matrix and observe that the expectation of each row is equal to the vector of Kullback– Leibler divergences between the K models and the true DGP, up to an additive constant. We use a multivariate normal model to estimate and quantify uncertainty in this expectation, providing calibrated inferences for robust model selection under misspecification. The procedure is easy to compute, interpretable, and comes with theoretical guarantees, including consistency.

Keywords— Bayesian inference; Likelihood as Data; Model misspecification; Robustness

---
[*]Corresponding author, E-mail: jongwoo.choi@uconn.edu

1 Introduction

A good statistical model should be flexible enough to capture the essential structure of the data, yet simple enough to interpret, communicate, and use in practice. Model selection helps balance these goals by choosing among competing models of varying complexity and explanatory power (Claeskens and Hjort, 2008). In most applications, the true data-generating process (DGP) is unknown and all of the candidate models are misspecified (Bernardo and Smith, 2000). In such cases, it is desirable to find a model that is as close as possible to the true DGP while remaining sufficiently simple to interpret and use.

Selection criteria such as the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) are based on the log-likelihood, which can be viewed as estimating the relative closeness of each model to the true DGP. However, they do not quantify uncertainty in these estimates or in the choice of model. Meanwhile, Bayesian posteriors quantify uncertainty in the choice of model (Wasserman, 2000), but posterior probabilities do not measure how far each model is from the truth; a model may receive a high probability simply because the alternatives are worse.

When none of the candidate models are correct, these information criteria and Bayesian posteriors asymptotically concentrate on the model with the smallest Kullback–Leibler (KL) divergence from the true DGP, even if there is a much simpler model that is nearly as good (Berk, 1966). Further, these methods tend to be highly unstable when more than one model attains the minimal KL divergence, even approximately (Huggins and Miller, 2023). Due to these issues, the standard leading methods do not provide well-calibrated inferences about the utility of each model in a way that is robust to misspecification.

To illustrate these limitations, consider the problem of choosing the number of components k in a finite mixture of Gaussians. Figure 1 (left) shows a histogram of radial velocity measurements of 4215 galaxies in the Shapley supercluster, a large concentration of gravitationally interacting galaxies (Drinkwater et al., 2004). We apply AIC, BIC, and a Bayesian model with a prior on the number of components to increasing subsets of this data set. In Figure 1 (middle), we see that the posterior distribution concentrates on larger and larger values of k as the sample size n increases, and does not provide any
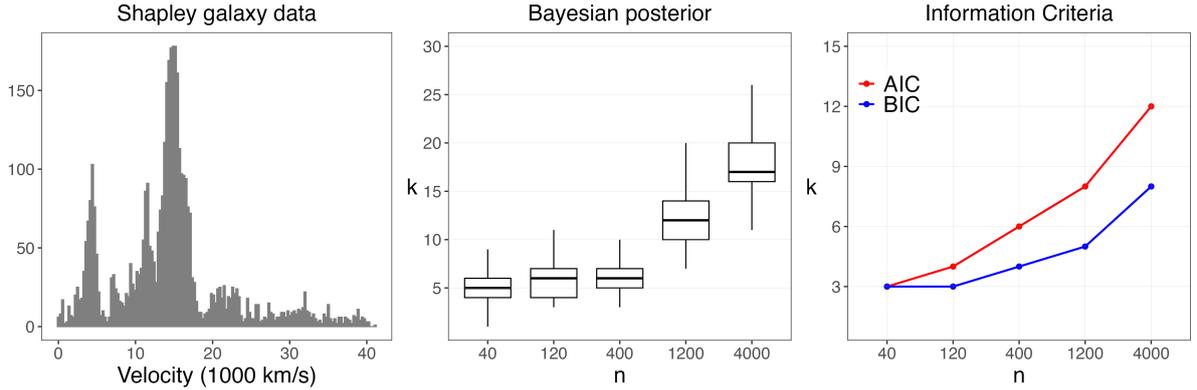
Figure 1: Limitations of existing model selection methods. *(Left)* Histogram of the Shapley galaxy velocities ($\times 10^3$ km/s). We trim the top 0.5% of data points to remove the extremely long right tail extending beyond 45,000 km/s. *(Middle)* Boxplots summarizing draws from the Bayesian posterior on k for a Gaussian mixture model with k components, for sample sizes $n \in \{40, 120, 400, 1200, 4000\}$. As n grows, the posterior shifts toward larger k, a typical phenomenon under misspecification. *(Right)* Values of k selected by AIC (red) and BIC (blue) for each n. These results illustrate lack of robustness and inability to quantify uncertainty in how well a k-component Gaussian mixture can approximate the true DGP.

insight into how close each model is to the true DGP. Similarly, Figure 1 (right) shows that AIC and BIC select increasing values of k as n grows, and furthermore, they do not provide any uncertainty quantification in the choice of k. See Section S3.1 for details.

In this paper, we introduce a new framework for model selection that overcomes these issues by jointly quantifying uncertainty in the expected log-likelihoods of the candidate models. More precisely, let Z_{ik} denote the negative log-likelihood for model k on observation i, assuming for the moment that there are no unknown parameters. Then $\mu_k = E(Z_{ik})$ equals the KL divergence between the true DGP and model k, plus the (unknown) entropy of the true DGP. We propose viewing Z_{ik} as data—hence the term "Likelihood as Data"—and using a Bayesian multivariate normal model to compute the posterior distribution of $\mu = (\mu_1, \ldots, \mu_K)$. This makes it straightforward to perform Bayesian inference for a range of questions, including not only which model is closest to the true DGP, but whether a given model is closer than another model, and which is the simplest model within a specified tolerance of being closest to the true DGP.

A key feature of our approach is that it is fully robust to misspecification of the candidate models since it is based inferring the μ_k's, which are well defined and interpretable without requiring any assumptions of model correctness. We prove that the method is

3

consistent and remains stable under ties – that is, when multiple models are the same distance from the true DGP. We propose a procedure for finding the simplest class of models within a specified tolerance of being closest to the true DGP, and choosing the best-fitting model (or models) within this class. In a range of simulations and real-data applications, we find that the proposed approach yields reliable and informative inferences for model selection in the presence of misspecification.

The paper is organized as follows. Section 2 introduces our proposed methodology, and Section 3 provides algorithms for computing the method. In Section 4, we provide theoretical results. Section 5 contains empirical demonstrations on Gaussian mixture models, sparse multivariate normal models, microbial ecology models, and population structure admixture models in genetics. Section 6 concludes with a brief discussion.

2 Methodology

In this section, we introduce the *Likelihood as Data* framework (Section 2.1), describe our robust model selection criterion (Section 2.2), discuss an interpretation of the selection threshold (Section 2.3), extend the method to parametrized models (Section 2.4), and provide an adjustment for overfitting bias in parametrized models (Section 2.5).

Suppose $X_1, \ldots, X_n \in \mathcal{X}$ are independent and identically distributed (i.i.d.) observations from some unknown distribution P_0, where $\mathcal{X}$ denotes the sample space. Assume P_0 admits a density f_0 with respect to a sigma-finite measure λ, for example, Lebesgue measure for continuous observations or counting measure for discrete observations.

Consider a finite collection of candidate models $P_1, \ldots, P_K$ with associated densities $f_1, \ldots, f_K$ (also with respect to λ), respectively. To clearly explain the main idea of our proposed method, we assume for now that there are no unknown parameters to be estimated in any of the candidate models; however, in Section 2.4, we extend the method to handle unknown parameters. We do not assume P_0 coincides with one of the candidate models, that is, all K models may be misspecified.

2.1 Likelihood as Data (LaD) framework

Let $Z_{ik} := -\log(f_k(X_i))$ be the negative log-likelihood value for each observation $i \in \{1, \ldots, n\}$ and each model $k \in \{1, \ldots, K\}$, and collect these values into a matrix $Z = [Z_{ik}] \in \mathbb{R}^{n \times K}$. The key idea is to treat each vector $Z_i := (Z_{i1}, \ldots, Z_{iK})^{\mathsf{T}} \in \mathbb{R}^K$ as a data point, and analyze the resulting data set, $Z_1, \ldots, Z_n$. In this way, we regard the likelihood values as data in their own right, which is the foundation of the LaD approach.

As with any data set, it is natural to estimate the mean $\mu^0 := \mathbb{E}(Z_1) \in \mathbb{R}^K$ and quantify uncertainty in it. In the LaD setting, μ^0 has a special interpretation since

$$\mu_k^0 = \mathbb{E}(Z_{ik}) = \mathbb{E}(-\log(f_k(X_i))) = D(f_0 \| f_k) + H(f_0)$$

where $H(f_0) := -\int_{\mathcal{X}} f_0(x) \log f_0(x) \lambda(dx)$ is the entropy of the true DGP and $D(f_0 \| f_k)$ is the Kullback–Leibler (KL) divergence between the true DGP and model k,

$$D(f_0 \| f_k) := \int_{\mathcal{X}} f_0(x) \left(\log \frac{f_0(x)}{f_k(x)} \right) \lambda(dx). \tag{1}$$

Thus, $\mu^0 = (\mu_1^0, \ldots, \mu_K^0)^{\mathsf{T}}$ is a vector containing the KL divergences between the true DGP and the K models, up to a common additive constant $H(f_0)$.

Estimation and inference for μ^0 are straightforward. Let $\overline{Z}_n := \frac{1}{n} \sum_{i=1}^n Z_i$ and observe that since $Z_1, \ldots, Z_n$ are i.i.d., we have

$$\begin{aligned} \overline{Z}_n &\xrightarrow[n \to \infty]{\text{a.s.}} \mu^0, \\ \sqrt{n}(\overline{Z}_n - \mu^0) &\xrightarrow[n \to \infty]{\text{d}} \mathcal{N}(0, \Sigma^0) \end{aligned} \tag{2}$$

where $\Sigma^0 = \text{Cov}(Z_1) \in \mathbb{R}^{K \times K}$, by the strong law of large numbers and the central limit theorem (CLT), assuming Σ^0 exists and is positive definite (Condition 2.1). Thus, one could estimate μ^0 by the sample mean $\overline{Z}_n$ and construct Wald-type confidence sets for μ^0 based on an estimate of Σ^0 such as the sample covariance matrix. Instead, however, we use a Bayesian multivariate normal model since it facilitates inference for many questions of interest and tends to be more stable at small sample sizes.

Specifically, we employ a Bayesian model where

$$Z_1, \ldots, Z_n \mid \mu, \Sigma \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma),$$

$$(\mu, \Sigma) \sim \pi(\mu, \Sigma), \tag{3}$$

and $\pi(\mu, \Sigma)$ is a prior on the unknown mean μ and covariance matrix Σ; see Section 3.1. We write μ, Σ for the parameters of this Bayesian model, and μ^0, Σ^0 for the true values. Uncertainty quantification for μ enables statistical inference for the relative distances from the true DGP. Note that looking only at marginal uncertainties in each entry μ_k alone is not particularly useful, for at least two reasons: first, $H(f_0)$ is unknown and can be very difficult to estimate; second, when the log-likelihood values for models j and k are strongly correlated, it can happen that there is very little uncertainty in the difference $\mu_j - \mu_k$ even if there is high uncertainty in μ_j and μ_k. Hence, multivariate inference for $\mu = (\mu_1, \ldots, \mu_K)^{\mathsf{T}}$ is far more informative than univariate inference for each μ_k.

The LaD approach does not require any assumptions regarding the form of the models or the true DGP, except for the moment conditions needed for the CLT to apply. While the Bayesian model does assume the negative log-likelihoods $Z_1, \ldots, Z_n$ are Gaussian, which is unlikely to hold, this discrepancy has minimal impact on inferences for μ since the posterior only depends on $Z_1, \ldots, Z_n$ through the sufficient statistics, namely the sample mean and sample covariance of the Z_i's. Consequently, we find that LaD provides correctly calibrated inferences for model selection, despite the incorrectness of the Gaussian model.

2.2 Robust model selection using LaD

Using the LaD framework in Section 2.1, one can perform frequentist or Bayesian inference for various hypotheses, such as whether model k minimizes the KL divergence to the true DGP (that is, $\mu_k^0 = \min_j \mu_j^0$) or whether model k is closer than model j to the true DGP (that is, $\mu_k^0 < \mu_j^0$). To perform robust model selection, we aim to identify the simplest model that is "close enough" to having minimal KL divergence. Specifically, we consider any model whose KL divergence is within a tolerance $\delta > 0$ of the minimum to be sufficiently close. This enables more parsimonious models to be favored when their fit is

roughly comparable to that of more complex models.

Definition 1. Given $\delta > 0$, under the setup above, we say model k is *δ-optimal* if

$$D(f_0\|f_k) \leq \min_j D(f_0\|f_j) + \delta. \tag{4}$$

Note that Equation 4 is equivalent to $\mu_k^0 \leq \min_j \mu_j^0 + \delta$. While our focus in this paper is on KL divergence, Definition 1 extends naturally to other measures of discrepancy between distributions, such as Wasserstein distance or Bregman divergence.

2.2.1 Best minimal-complexity δ-optimal model

Let $M_\delta(\mu)$ denote the set of δ-optimal models as a function of μ, that is,

$$M_\delta(\mu) := \left\{ k : \mu_k \leq \min_j \mu_j + \delta \right\}. \tag{5}$$

Suppose $c : \{1, \ldots, K\} \to [0, \infty)$ assigns a complexity $c(k)$ to model k, with smaller values corresponding to simpler models. We refer to a set of models with the same $c(k)$ as a complexity class. Among δ-optimal models, the minimal complexity attained is

$$c_\delta^*(\mu) := \min\{c(k) : k \in M_\delta(\mu)\}. \tag{6}$$

Since there may be multiple δ-optimal models that attain this minimal complexity $c_\delta^*(\mu)$, among these we prefer the ones with best fit, that is, with minimal KL from the true DGP. Thus, we define the *best minimal-complexity δ-optimal model(s)* as

$$M_\delta^*(\mu) := \underset{k \,:\, c(k) = c_\delta^*(\mu)}{\arg\min} \mu_k. \tag{7}$$

The set $M_\delta^*(\mu)$ contains those models that (i) have the lowest possible complexity attainable by a δ-optimal model, and (ii) have the smallest KL divergence within that complexity class. Typically, $M_\delta^*(\mu)$ will consist of a single model, but it can contain multiple models if there are ties $\mu_j = \mu_k$. For robust model selection, our goal is to

quantify uncertainty about which model or models belong to $M_\delta^*(\mu^0)$ for the true μ^0.

2.2.2 Reproducible inference for $M_\delta^*(\mu)$

A natural way to infer the true set $M_\delta^*(\mu^0)$ would be to consider the posterior probability that each model k is in $M_\delta^*(\mu)$, that is, $P(k \in M_\delta^*(\mu) \mid Z_{1:n})$, where $\mu \mid Z_{1:n}$ is distributed according to the posterior under a Bayesian model as in Equation 3. However, if we use a continuous prior on μ and the true $M_\delta^*(\mu^0)$ contains multiple models, then $P(k \in M_\delta^*(\mu) \mid Z_{1:n})$ is not an accurate representation of uncertainty. For example, when there is correlation among models, some models that belong to $M_\delta^*(\mu^0)$ may receive arbitrarily small posterior probability; see Section 4. One solution would be to use a prior that allows $\mu_j = \mu_k$ with positive probability for any j, k, but this would make posterior inference more computationally burdensome and complicate theoretical analysis.

Instead, we propose a computationally and theoretically attractive alternative based on a smooth selection function that relaxes the hard minimum in Equation 7. For each model k and sample size n, we define a within-class soft-selection score,

$$r_{nk}(\mu) := \exp\big(-\alpha_n(\mu_k - \mu_{\min,c(k)})\big), \tag{8}$$

where $\mu_{\min,c(k)} = \min_{j : c(j)=c(k)} \mu_j$ and $\alpha_n > 0$ is a temperature parameter such that $\alpha_n \to \infty$ and $\alpha_n = o(\sqrt{n})$. In all of the examples in this paper, we use $\alpha_n = n^{0.45}$. Note that $r_{nk}(\mu) \in (0, 1]$, with $r_{nk}(\mu) = 1$ for models k having minimal μ_k within their complexity class $c(k)$. We now introduce our proposed model selection score.

Definition 2. The *smooth LaD criterion (SLC)* score is

$$w_\delta(k \mid Z_{1:n}) = P\big(c_\delta^*(\mu) = c(k) \mid Z_{1:n}\big)\, \mathbb{E}\big(r_{nk}(\mu) \mid Z_{1:n}\big) \tag{9}$$

where $\mu \mid Z_{1:n}$ is distributed according to posterior of a model as in Equation 3.

An estimator of $M_\delta^*(\mu^0)$ can be constructed as $\hat{M}_\delta^* = \{k : w_\delta(k \mid Z_{1:n}) > \omega\}$, where ω is a threshold between 0 and 1. In Section 4, we justify the SLC score by showing that

it is asymptotically consistent, that is, $w_\delta(k \mid Z_{1:n})$ concentrates on the true target set $M_\delta^*(\mu^0)$ as n grows. The SLC score $w_\delta(k \mid Z_{1:n})$ is defined as the product of (i) a between-class selection factor and (ii) a within-class selection factor. The first factor represents the probability that model k attains the minimal complexity among all δ-optimal models. The second factor quantifies model k's relative performance within its complexity class. We compute each of the two factors with a Monte Carlo approximation using posterior samples from our Bayesian model; see Algorithm 2 for details.

For interpretation, if there is only one model per complexity class, then $r_{nk}(\mu) = 1$ for all k, so Equation 9 simplifies to $w_\delta(k \mid Z_{1:n}) = P\big(\arg\min_{j \in M_\delta(\mu)} c(j) = \{k\} \mid Z_{1:n}\big)$, because $c_\delta^*(\mu) = c(k)$ if and only if $\arg\min_{j \in M_\delta(\mu)} c(j) = \{k\}$. More generally, whenever $|M_\delta^*(\mu^0)| = 1$, the SLC score $w_\delta(k \mid Z_{1:n})$ behaves similarly to the posterior probability $P(k \in M_\delta^*(\mu) \mid Z_{1:n})$. Meanwhile, when $|M_\delta^*(\mu^0)| > 1$, the SLC score $w_\delta(k \mid Z_{1:n})$ remains stable since it does not concentrate on a single model. This is because, under the posterior, $\mu_k - \mu_{\min,c(k)} = O_p(n^{-1/2})$ for the best fitting models k with complexity $c(k)$, and we choose $\alpha_n = o(\sqrt{n})$, which makes $\mathbb{E}\big(r_{nk}(\mu) \mid Z_{1:n}\big) \to 1$ as $n \to \infty$.

2.3 Interpreting the tolerance δ

The tolerance δ makes the selection procedure robust to model misspecification, since it allows one to select models that have a KL divergence within δ of being minimal. Smaller values of δ enforce stricter optimality but may only be satisfied by complex models, whereas larger values of δ encourage parsimony at the cost of fit. Rather than choosing one value of δ, we generally recommend computing the SLC score for a range of δ values, as demonstrated in the examples in Section 5. Since the units of δ may not be intuitively clear, in this section we provide a technique for putting δ on an interpretable scale.

Motivated by an idea presented by Pescador-Barrios et al. (2025) (Section 4.3), we augment the set of candidate models with a deliberately misspecified "noise" model M_{noise} with density f_{noise}, chosen to represent a baseline that does not capture any meaningful

signal in the data. Under the noise model, the expected negative log-likelihood is

$$\mu^0_{\text{noise}} := \mathbb{E}(-\log f_{\text{noise}}(X_i)) = D(f_0 \| f_{\text{noise}}) + H(f_0).$$

This noise model should satisfy: (i) full support over $\mathcal{X}$, so that $D(f_0 \| f_{\text{noise}}) < \infty$, (ii) stability under estimation, so that μ^0_{noise} can be estimated with little uncertainty, and (iii) minimal signal (for example, an intercept-only or null model). Now, we define

$$\tau := \delta / (\mu^0_{\text{noise}} - \mu^0_{\text{min}}) \tag{10}$$

where $\mu^0_{\text{min}} = \min_k \mu^0_k$. Note that, equivalently, $\tau = \delta / (D_{\text{noise}} - D_{\text{min}})$, where $D_{\text{noise}} := D(f_0 \| f_{\text{noise}})$ and $D_{\text{min}} := \min_k D(f_0 \| f_k)$. Thus, τ is a rescaled version of δ that puts it in interpretable units: Specifically, τ represents the proportion of the explainable information that we are willing to sacrifice in return for model simplicity. A value of $\tau > 1$ signifies that δ is excessively large, such that even it tolerates the noise model.

By the definition of the δ-optimal set $M_\delta(\mu^0)$ (Equation 5), for any model $k \in M_\delta(\mu^0)$,

$$D(f_0 \| f_k) \leq D_{\text{min}} + \tau \, (D_{\text{noise}} - D_{\text{min}}).$$

Rearranging this expression, we obtain

$$\frac{D_{\text{noise}} - D(f_0 \| f_k)}{D_{\text{noise}} - D_{\text{min}}} \geq 1 - \tau.$$

This yields a natural interpretation: A model is δ-optimal if it recovers at least $100(1-\tau)\%$ of the total possible improvement from the noise model to the best available model, in terms of KL divergence. Thus, the choice of δ becomes interpretable as a relative tolerance expressed on this information-theoretic scale.

Suitable choices for the noise model depend on the context. For example, for multivariate Gaussian models, we use $f_{\text{noise}} = \mathcal{N}(0, I)$. For generalized linear models, we use an intercept-only model (no covariates). For mixture models, a suitable noise baseline could be a uniform density over the observed data range or a one-component mixture.

2.4 Extending to unknown parameters and other loss functions

Up to now, we have assumed that each model consists of a single distribution with known parameters, but in practice, a model will usually have unknown parameters that need to be estimated. We now show how to handle such models in our framework.

Suppose model k is a parametric family with densities $\{f_k(x; \theta_k) : \theta_k \in \Theta_k\}$, where $\Theta_k \subseteq \mathbb{R}^{d_k}$ and d_k is the dimension of the parameter vector θ_k for model k. Furthermore, we generalize from the negative log-likelihood to an arbitrary loss function ℓ_k for model k. As before, suppose $X, X_1, \ldots, X_n \in \mathcal{X}$ i.i.d. $\sim P_0$, where P_0 has density f_0 with respect to λ. For each $k \in \{1, \ldots, K\}$, let $\ell_k : \mathcal{X} \times \Theta_k \to \mathbb{R}$ be measurable, and define

$$\theta_k^* \in \underset{\theta_k \in \Theta_k}{\arg\min} \, \mathbb{E}\big(\ell_k(X; \theta_k)\big). \tag{11}$$

Define the random vector $\ell(X; \theta^*) := [\ell_1(X; \theta_1^*), \ldots, \ell_K(X; \theta_K^*)]^{\mathsf{T}} \in \mathbb{R}^K$, and set

$$\mu^0 := \mathbb{E}\big(\ell(X; \theta^*)\big) \in \mathbb{R}^K,$$
$$\Sigma^0 := \text{Cov}\big(\ell(X; \theta^*)\big) \in \mathbb{R}^{K \times K}. \tag{12}$$

In Equations 11 and 12, the expectations are with respect to $X \sim P_0$.

Condition 2.1. *Assume:*

(i) $\mathbb{E}\big(\ell_k(X; \theta_k^*)^2\big) < \infty$, *that is,* $\ell(X; \theta^*)$ *has finite second moments, and*

(ii) Σ^0 *is a positive definite matrix.*

To estimate the parameter vector θ_k for each model k, we use the M-estimator,

$$\hat{\theta}_k := \underset{\theta_k \in \Theta_k}{\arg\min} \left(\frac{1}{n} \sum_{i=1}^{n} \ell_k(X_i; \theta_k) \right).$$

In the examples in this article, we focus on the negative log-likelihood loss, in which case the M-estimator is simply the maximum likelihood estimator (MLE), but the method extends to any loss leading to a consistent M-estimator. Plugging in the estimators

$\hat{\theta}_1, \dots, \hat{\theta}_K$ for models $1, \dots, K$, respectively, the vector of loss values for X_i becomes

$$Z_i(\hat{\theta}) := \ell(X_i; \hat{\theta}) = [\ell_1(X_i; \hat{\theta}_1), \dots, \ell_K(X_i; \hat{\theta}_K)]^{\mathsf{T}} \in \mathbb{R}^K.$$

Then, the sample mean and covariance based on $Z_1(\hat{\theta}), \dots, Z_n(\hat{\theta}) \in \mathbb{R}^K$ are

$$\begin{aligned}
\overline{Z}_n(\hat{\theta}) &:= \frac{1}{n} \sum_{i=1}^{n} \ell(X_i; \hat{\theta}), \\
S_n(\hat{\theta}) &:= \frac{1}{n} \sum_{i=1}^{n} \big(\ell(X_i; \hat{\theta}) - \overline{Z}_n(\hat{\theta})\big)\big(\ell(X_i; \hat{\theta}) - \overline{Z}_n(\hat{\theta})\big)^{\mathsf{T}}.
\end{aligned} \tag{13}$$

Under regularity conditions (Condition S2.1), including that $\hat{\theta}_k$ is a consistent estimator of θ_k^* for each k, the limits in Equation 2 also hold for $\overline{Z}_n(\hat{\theta})$ (see Theorem S2.3) generalizing from the no-parameter case to the setting of estimated parameters. Thus, frequentist inference for μ^0 could again be used in this more general setting; but as mentioned in Section 2.1, we instead recommend a Bayesian approach (Section 3).

2.5 Adjustment for overfitting bias

For large sample sizes n, plugging in a consistent estimator $\hat{\theta}_k$ as a proxy for θ^* leads to negligible error. However, for smaller sample sizes, $\overline{Z}_n(\hat{\theta})$ tends to be downward biased as an estimator of μ^0, because the parameters are fit on the same data that is used to estimate μ^0. To adjust for this overfitting, we propose a simple bias correction based on the asymptotic distribution of the likelihood-ratio test statistic.

To derive the correction, assume the loss is the negative log-likelihood $\ell_k(x; \theta_k) = -\log f_k(x; \theta_k)$. Then the likelihood-ratio test statistic is

$$\Lambda_k = 2\Big(\sum_{i=1}^{n} \ell_k(X_i; \theta_k^*) - \sum_{i=1}^{n} \ell_k(X_i; \hat{\theta}_k) \Big).$$

Rearranging and taking the expectation under P_0 yields that

$$\mu_k^0 = \mathbb{E}\Big(\frac{1}{n} \sum_{i=1}^{n} \ell_k(X_i; \theta_k^*) \Big) = \mathbb{E}\Big(\frac{1}{n} \sum_{i=1}^{n} \ell_k(X_i; \hat{\theta}_k) \Big) + \frac{1}{2n} \mathbb{E}(\Lambda_k).$$

Hence, the bias of $\frac{1}{n}\sum_{i=1}^{n}\ell_k(X_i;\hat{\theta}_k)$ as an estimator of μ_k^0 is $-(2n)^{-1}\mathbb{E}(\Lambda_k)$.

By Wilks' theorem, if model k were correctly specified, then Λ_k would be approximately χ^2-distributed with $d_k = \dim(\Theta_k)$ degrees of freedom, under regularity conditions. Since the mean of $\chi^2(d_k)$ is d_k, this suggests using the bias-corrected estimator $\hat{\mu}_k^{\mathrm{bc}} := \frac{1}{n}\sum_{i=1}^{n}\ell_k(X_i;\hat{\theta}_k) + d_k/(2n)$. Since LaD operates on vectors $Z_1,\ldots,Z_n \in \mathbb{R}^K$, we apply this adjustment entrywise to define bias-corrected LaD vectors

$$Z_i^{\mathrm{bc}}(\hat{\theta}) := Z_i(\hat{\theta}) + \frac{d}{2n} \tag{14}$$

where $d = (d_1,\ldots,d_K)^{\mathsf{T}} \in \mathbb{R}^K$. In practice, we expect all of the models to be misspecified, so Wilks' theorem does not directly apply; nonetheless, empirically we find that the approximation is sufficiently good that the bias correction in Equation 14 works well.

3   Computation

In this section, we describe our Bayesian model for the LaD framework (Section 3.1) and provide a step-by-step algorithm for our proposed methodology (Section 3.2).

3.1   Bayesian inference in the LaD framework

Let $Z_1,\ldots,Z_n \in \mathbb{R}^K$ denote the LaD vectors, $Z_i = (Z_{i1},\ldots,Z_{iK})^{\mathsf{T}}$; these may comprise negative log-likelihood values $Z_{ik} = -\log(f_k(X_i))$ (as in Section 2.1), loss values $Z_{ik} = \ell_k(X_i;\hat{\theta}_k)$ evaluated at estimated parameters $\hat{\theta}_k$ (as in Section 2.4), or bias-corrected versions of such values $Z_{ik} = \ell_k(X_i;\hat{\theta}_k)+d_k/(2n)$ (as in Equation 14). To perform Bayesian inference, we model $Z_1,\ldots,Z_n$ as i.i.d. $\sim \mathcal{N}(\mu,\Sigma)$ and place a Normal-Inverse-Wishart (NIW) prior on (μ,Σ). Specifically, the prior density of (μ,Σ) is

$$\mathrm{NIW}(\mu,\Sigma \mid \mu_0,\lambda_0,\Psi_0,\nu_0) = \mathcal{N}(\mu \mid \mu_0,\Sigma/\lambda_0)\,\mathrm{InverseWishart}(\Sigma \mid \Psi_0,\nu_0)$$

where $\mu_0 \in \mathbb{R}^K$ is the prior mean, $\lambda_0 > 0$ is a scaling factor, $\nu_0 > K-1$ is the degrees of freedom, and Ψ_0 is the (positive definite) inverse scale matrix. The posterior distribution

13

of $\mu, \Sigma \mid Z_{1:n}$ is NIW with updated parameters, specifically,

$$\mu, \Sigma \mid Z_{1:n} \sim \text{NIW}(\mu_n, \lambda_n, \Psi_n, \nu_n) \tag{15}$$

where $\lambda_n = \lambda_0 + n$, $\mu_n = (\lambda_0 \mu_0 + n \overline{Z}_n)/\lambda_n$, $\Psi_n = \Psi_0 + n S_n + (\lambda_0 n/\lambda_n)(\overline{Z}_n - \mu_0)(\overline{Z}_n - \mu_0)^\mathsf{T}$, and $\nu_n = \nu_0 + n$, where $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ and $S_n = \frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z}_n)(Z_i - \overline{Z}_n)^\mathsf{T}$.

To generate a posterior sample of $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$, all that needs to be done is to draw $\Sigma' \sim \text{InverseWishart}(\Psi_n, \nu_n)$ and then draw $\mu' \sim \mathcal{N}(\mu_n, \Sigma'/\lambda_n)$. To perform posterior inference, we generate multiple posterior samples and use Monte Carlo approximations. For the hyperparameter settings, we recommend $\mu_0 = 0$, $\nu_0 = K+2$, $\lambda_0 = 0.01$, and $\Psi_0 = I_K$ (the $K \times K$ identity matrix), since these yield a weakly informative prior that allows the data to mostly dominate.

3.2 Algorithm for LaD model selection

Algorithm 1 provides a step-by-step description of the workflow for our proposed method.

Algorithm 1   Workflow for LaD model selection

Input: Observations $X_1, \ldots, X_n$, parametrized models $k = 1, \ldots, K$ and their corresponding complexities $c(1), \ldots, c(K) \geq 0$, number of Monte Carlo samples $T > 0$, and tolerance $\delta \geq 0$.

1. Compute parameter estimates $\hat{\theta}_k$ for each model k.

2. Compute the bias-corrected LaD values $Z_{ik} = \ell_k(X_i; \hat{\theta}_k) + d_k/(2n)$ for all i and k.

3. Sample $\mu^{(t)}, \Sigma^{(t)}$ from the NIW posterior (Equation 15) for $t = 1, \ldots, T$.

4. Compute the SLC scores $\hat{w}_\delta(k)$ for each k using Algorithm 2.

5. (Optional) Compute the rescaled tolerance $\hat{\tau} = \delta/(\hat{\mu}_{\text{noise}} - \hat{\mu}_{\text{min}})$ where $\hat{\mu}_{\text{noise}}$ is an estimate of μ_{noise}^0 (Section 2.3), $\hat{\mu}_{\text{min}} = \min_k \hat{\mu}_k$, and $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n Z_{ik}$.

Output: Return the samples $\mu^{(t)}, \Sigma^{(t)}$, SLC scores $\hat{w}_\delta(k)$, and rescaled tolerance $\hat{\tau}$.

Recall that our SLC score $w_\delta(k \mid Z_{1:n})$ (Equation 9) is defined as the product of a between-class selection factor $P\big(c_\delta^*(\mu) = c(k) \mid Z_{1:n}\big)$ and a within-class selection factor $\mathbb{E}\big(r_{nk}(\mu) \mid Z_{1:n}\big)$. To compute $w_\delta(k \mid Z_{1:n})$, we generate posterior draws $\mu^{(1)}, \ldots, \mu^{(T)}$ given $Z_{1:n}$ as described in Section 3.1, then compute each of these two factors using a simple Monte Carlo approximation, and take their product. See Algorithm 2 for details.

14

Algorithm 2   Computing the smooth LaD criterion (SLC) scores

---

Input: Posterior samples $\mu^{(1)}, \ldots, \mu^{(T)} \in \mathbb{R}^K$, complexities $c(1), \ldots, c(K) \geq 0$, tolerance $\delta \geq 0$, and temperature $\alpha_n > 0$.

for $t = 1, \ldots, T$ **do**

1. **Between-class selection:** Compute the set of δ-optimal models,

$$M_\delta(\mu^{(t)}) = \big\{k : \mu_k^{(t)} \leq \min_j \mu_j^{(t)} + \delta\big\}.$$

Then, compute the minimal complexity among δ-optimal models,

$$c_\delta^*(\mu^{(t)}) = \min\big\{c(k) : k \in M_\delta(\mu^{(t)})\big\}.$$

2. **Within-class selection:** For each model $k = 1, \ldots, K$, compute the within-class minimum $\mu_{\min,c(k)}^{(t)} = \min\big\{\mu_j^{(t)} : c(j) = c(k)\big\}$, and compute the soft-selection score

$$r_{nk}(\mu^{(t)}) = \exp\big(-\alpha_n(\mu_k^{(t)} - \mu_{\min,c(k)}^{(t)})\big).$$

for $k = 1, \ldots, K$ **do**

$$\hat{p}_\delta(k) := \frac{1}{T}\sum_{t=1}^T \mathbb{1}\big(c_\delta^*(\mu^{(t)}) = c(k)\big) \quad \text{and} \quad \hat{r}(k) := \frac{1}{T}\sum_{t=1}^T r_{nk}(\mu^{(t)}).$$

Output: Return the estimated SLC scores $\hat{w}_\delta(k) := \hat{p}_\delta(k)\,\hat{r}(k)$ for $k = 1, \ldots, K$.

---

4   Theory

In this section, we prove that our proposed method is asymptotically consistent (Section 4.1) and we show that alternative approaches based on a plug-in posterior or using a hard minimum exhibit instability (Section 4.2).

4.1   Consistency of the SLC score

Our first result establishes that the SLC score $w_\delta(k \mid Z_{1:n})$ concentrates on the true target set $M_\delta^*(\mu^0)$, asymptotically.

Theorem 4.1. *Assume Conditions 2.1 and S2.1, and suppose $\delta \neq \mu_k^0 - \mu_{\min}^0$ for all $k = 1, \ldots, K$. Let $\alpha_n \geq 0$ such that $\alpha_n \to \infty$ with $\alpha_n = o(\sqrt{n})$. Define $w_\delta(k \mid Z_{1:n})$ as in Equation 9 based on the NIW posterior in Section 3.1. Then, for all $k \in \{1, \ldots, K\}$,*

$$w_\delta(k \mid Z_{1:n}) \xrightarrow[n \to \infty]{\text{p}} \mathbb{1}\big(k \in M_\delta^*(\mu^0)\big).$$

All proofs are collected in Section S2. The next result shows that the Monte Carlo estimates $\hat{w}_\delta(k)$ produced by Algorithm 2 also concentrate on $M_\delta^*(\mu^0)$.

Theorem 4.2. *Under the same assumptions as Theorem 4.1, for all $k \in \{1, \ldots, K\}$,*

$$\hat{w}_\delta(k) \xrightarrow[n \to \infty]{\mathrm{p}} \mathbb{1}\big(k \in M_\delta^*(\mu^0)\big).$$

Theorem 4.3. *Under the same assumptions as Theorem 4.1, for any fixed $\omega \in (0, 1)$,*

$$P\big(\hat{M}_\delta^* = M_\delta^*(\mu^0)\big) \xrightarrow[n \to \infty]{} 1,$$

where $\hat{M}_\delta^* := \{k : w_\delta(k \mid Z_{1:n}) > \omega\}$.

Theorem 4.3 also holds when using $\hat{w}_\delta(k)$ instead of $w_\delta(k \mid Z_{1:n})$.

4.2 Instability of alternative approaches at ties

The results in this section provide the rationale for using our soft-selection score for within-class selection (Equation 8), rather than the Bayesian plug-in posterior $P(k \in M_\delta^*(\mu) \mid Z_{1:n})$ or a hard-minimum selection score. When multiple models of equal complexity are tied for having minimal KL divergence, these naive alternatives randomly allocate probability mass among tied models, and not necessarily in a uniform way.

To understand why the issue arises, suppose there are $K = 3$ models of equal complexity and all three are equally close to the true DGP, that is, $\mu^0 = (\mu_1^0, \mu_2^0, \mu_3^0)^{\mathsf{T}}$ where $\mu_1^0 = \mu_2^0 = \mu_3^0$. Furthermore, suppose models 1 and 2 are strongly negatively correlated and have much higher variance than model 3, for example,

$$\Sigma^0 = \begin{bmatrix} 1 & -0.99 & 0 \\ -0.99 & 1 & 0 \\ 0 & 0 & 0.01 \end{bmatrix}.$$

Under the posterior, suppose for simplicity that $\mu \sim \mathcal{N}(\mu^0, \Sigma^0/n)$ given $Z_{1:n}$, which is the expected asymptotic behavior. Then with fairly high probability, $\min\{\mu_1, \mu_2\} <$
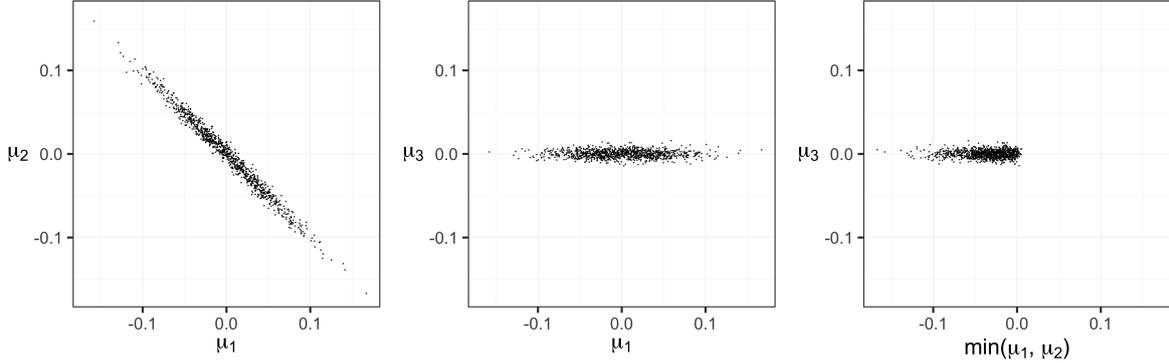
Figure 2: Instability under ties with $K = 3$. The figure shows 1,000 posterior draws from $\mu \sim \mathcal{N}(\mu^0, \Sigma^0/n)$ with $n = 500$, $\mu^0 = (0,0,0)^{\mathsf{T}}$, and Σ^0 as specified in the text. The left panel shows the strong negative correlation between μ_1 and μ_2, the middle panel shows that μ_3 is much more concentrated than μ_1 and μ_2, and the right panel makes clear that μ_3 is rarely the minimum.

μ_3 because μ_3 has much smaller variance than the other two, and the strong negative correlation makes $\mu_1 \approx -\mu_2$. Specifically, $P(\mu_k = \min_j \mu_j \mid Z_{1:n})$ is ≈ 0.48 for $k \in \{1, 2\}$ and ≈ 0.04 for $k = 3$. Thus, model 3 is rarely selected in this scenario, even though all three models are equally good. Figure 2 visualizes this pattern.

We make this phenomenon precise in the following results. For $x \in \mathbb{R}^K$, define

$$\pi_k(x) = \begin{cases} P\Big(W_j - W_k \le x_j - x_k \text{ for all } j \in M_\delta^*(\mu^0) \setminus \{k\}\Big) & \text{if } k \in M_\delta^*(\mu^0) \\ 0 & \text{if } k \notin M_\delta^*(\mu^0) \end{cases} \tag{16}$$

where $W \sim \mathcal{N}(0, \Sigma^0)$.

In Theorem 4.4, parts (a) and (b) show the instability of the plug-in posterior and the hard-minimum rule, respectively. When $|M_\delta^*(\mu^0)| > 1$, the choice among tied models is randomly distributed in a way that depends on the covariance of the limiting Gaussian.

Theorem 4.4. *Assume Conditions 2.1 and S2.1, let $\mu, \Sigma \mid Z_{1:n}$ be distributed according to Equation 15, and suppose $\delta \ne \mu_k^0 - \mu_{\min}^0$ for all $k = 1, \ldots, K$. Let $W \sim \mathcal{N}(0, \Sigma^0)$.*

(a) $\big(P(k \in M_\delta^*(\mu) \mid Z_{1:n})\big)_{k=1}^K \xrightarrow[n\to\infty]{\mathrm{d}} \big(\pi_k(W)\big)_{k=1}^K.$

(b) *Consider a modified score $\tilde{w}_\delta(k \mid Z_{1:n})$ defined as in Equation 9 but with $\tilde{r}_{nk}(\mu) = \mathbb{1}(\mu_k = \mu_{\min, c(k)})$ in place of $r_{nk}(\mu)$. Then $\big(\tilde{w}_\delta(k \mid Z_{1:n})\big)_{k=1}^K \xrightarrow[n\to\infty]{\mathrm{d}} \big(\pi_k(W)\big)_{k=1}^K.$*

17

Interestingly, the limiting behavior of $\tilde{w}_\delta(k \mid Z_{1:n})$ coincides with the asymptotic BayesBag posterior with $M = N$ (Huggins and Miller, 2023, Theorem 3.2). Corollary 4.5 shows that in the case of two tied models, the scores of the two models are asymptotically distributed as $(U, 1 - U)$ where $U \sim \text{Uniform}(0, 1)$.

Corollary 4.5. *Assume the conditions of Theorem 4.4. If $M_\delta^*(\mu^0) = \{k_1, k_2\}$ then*

$$\left(\tilde{w}_\delta(k_1 \mid Z_{1:n}), \tilde{w}_\delta(k_2 \mid Z_{1:n}) \right) \xrightarrow[n \to \infty]{\text{d}} (U, 1 - U)$$

where $U \sim \text{Uniform}(0, 1)$.

5 Examples

5.1 Gaussian mixture models

We revisit the motivating example from Section 1 using finite Gaussian mixture models on the Shapley galaxy data. We implement the LaD workflow in Algorithm 1 as follows. First, we fit a k-component Gaussian mixture for each $k = 1, \ldots, 10$ using the expectation-maximization (EM) algorithm to estimate the *maximum a posteriori* (MAP) solution (Fraley and Raftery, 2007). For this, we employ the `mclust` package in R (Scrucca et al., 2016; R Core Team, 2021) with random initialization, using a moderately informative normal-inverse-gamma prior on the mean and variance of each mixture component, independently. For each k, we run the algorithm on 50 independent restarts and keep the run with the highest observed log-likelihood. See Section S3.1 for model details.

We then compute the bias-corrected LaD values $Z_{ik} = \ell_k(X_i; \hat{\theta}_k) + d_k/(2n)$ (Equation 14) where $\hat{\theta}_k$ is the MAP estimate and $d_k = 3k - 1$ since a univariate k-component mixture with unconstrained means, variances, and mixture weights has a total of $k + k + (k - 1) = 3k - 1$ parameters. While mixtures do not perfectly satisfy the LRT conditions, we find that this bias correction is an effective approximation in practice. We draw $T = 1000$ samples of $\mu \mid Z_{1:n}$ from the NIW posterior (Equation 15), and then run Algorithm 2 to obtain the SLC scores $\hat{w}_\delta(k)$ for a range of values of the tolerance $\delta > 0$. We use

$\alpha_n = n^{0.45}$ for the temperature and $c(k) = k$ for the model complexities. To compute the rescaled tolerance $\hat{\tau}$ as described in Algorithm 1, we let $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} Z_{ik}$ be the mean of the bias-adjusted LaD values, and we define the noise baseline to be the uniform distribution over the observed range $[\min_i x_i, \max_i x_i]$, which yields $\hat{\mu}_{\text{noise}} = \log(\max_i x_i - \min_i x_i)$. This puts δ on an interpretable scale, namely, $\hat{\tau} \in [0,1]$ is the fraction of explainable KL improvement one is willing to give up for model simplicity.

Figure 3 displays the results. For comparison, the top row shows the usual Bayesian posterior on k, with AIC and BIC overlaid; all three increasingly favor larger k as n grows, exhibiting unstable drift toward complexity under misspecification. The second row shows boxplots of the LaD posterior on μ_k for each k, which provide estimates (with uncertainty quantification) of the KL divergence between the true DGP and each model k, up to a common additive constant. As n increases, we see that the uncertainty in these estimates decreases. As k increases, the estimated KL decreases and gradually levels off, indicating a diminishing improvement in fit as the model complexity grows. The third row plots the SLC scores $\hat{w}_\delta(k)$ for $\delta \in \{0.30, 0.12, 0.06\}$. Larger δ values are more tolerant, resulting in selection of simpler models; meanwhile, smaller δ values enforce a tighter tolerance, requiring more complex models in order to achieve a KL within δ of being minimal. As n increases, the $\hat{w}_\delta(k)$ values concentrate on a single k for each δ, illustrating the large-sample concentration guaranteed by our theory when no two models are tied in terms of both complexity and KL. Notably, in contrast with AIC, BIC, and the usual Bayesian posterior results in Figure 1, the LaD results stabilize as n increases.

This is illustrated further in the bottom row, which shows the *posterior path* of $\hat{w}_\delta(k)$ as $\hat{\tau}$ varies from 0 to 1. Here, we see a continuous transition from selecting more complex to simpler models as the proportion of explainable KL that we are willing to sacrifice increases from 0 to 1. For instance, while the usual Bayesian posterior selects values of k between 10 and 25 when $n = 4000$ (Figure 1), the LaD posterior path shows that $k > 8$ is only needed for very tight tolerances of $\hat{\tau} < 0.01$, that is, $< 1\%$ of explainable KL sacrificed. This plot encapsulates the key features of LaD: It operates without assuming any candidate model is correctly specified, quantifies uncertainty in their relative KL
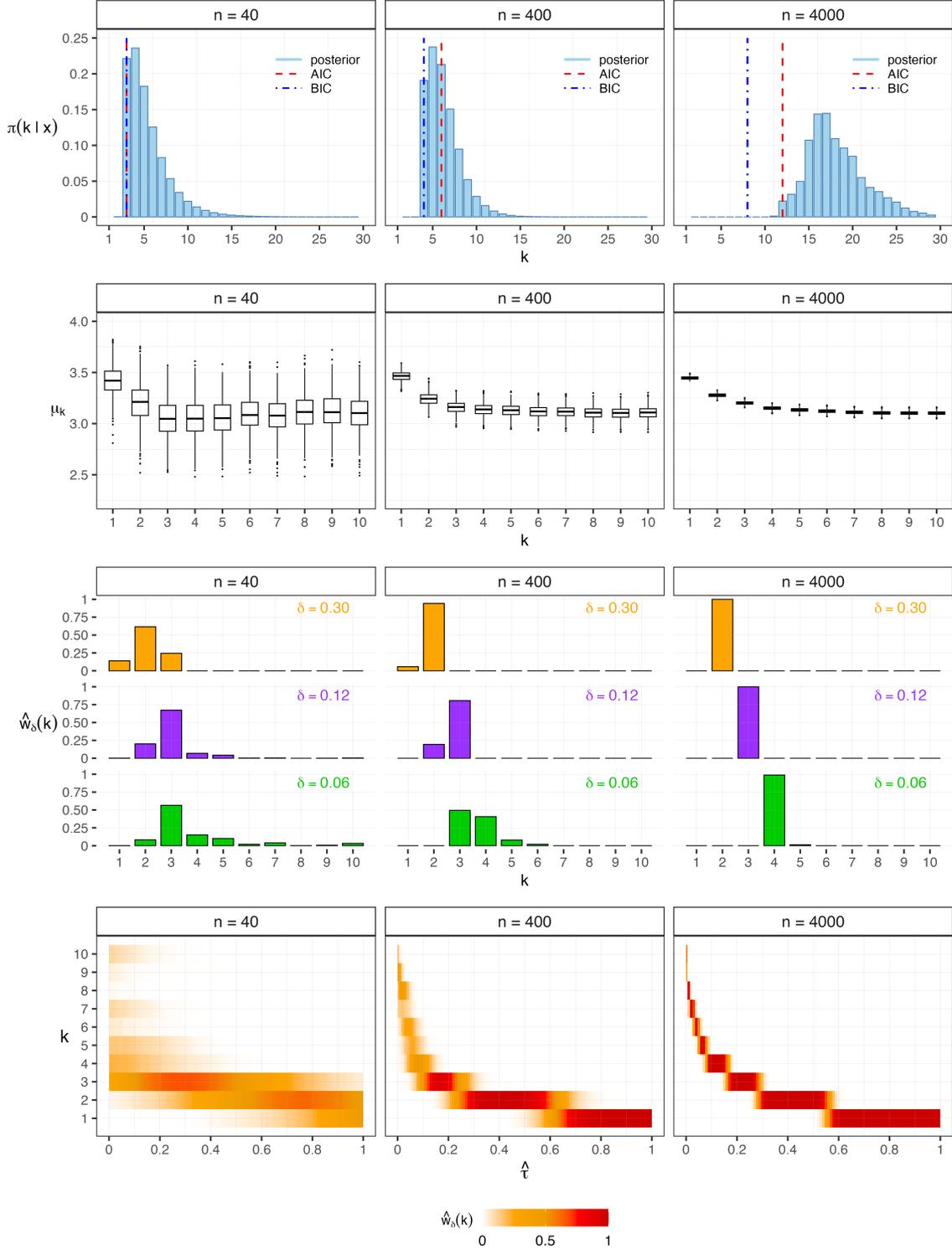
19

Figure 3: Gaussian mixture model results on the Shapley galaxy dataset. *(Top row)* Usual Bayesian posterior on k (bars) with AIC (red dashed) and BIC (blue dot-dash) overlaid; as n increases, all three trend toward larger k. *(Second row)* LaD posteriors on μ_k for each k show the fit to the true DGP as a function k, and concentrate as n grows. *(Third row)* SLC score $\hat{w}_\delta(k)$ for tolerances $\delta \in \{0.3, 0.12, 0.06\}$. Larger δ selects simpler models; increasing n concentrates the scores onto a single k for each δ. *(Bottom row)* Heatmap of "posterior path" of SLC score $\hat{w}_\delta(k)$ shows the transition from selecting more complex to simpler models as a function of the rescaled tolerance $\hat{\tau} = \delta / (\hat{\mu}_{\text{noise}} - \hat{\mu}_{\text{min}}) \in [0, 1]$.

20

divergences from the true DGP, and places these divergences on an interpretable scale.

5.2   Sparse multivariate normal mean models

In this simulation, we consider multivariate normal (MVN) models with sparse mean vectors. We generate n i.i.d. samples $x_1, \ldots, x_n \sim \mathcal{N}(\theta_0, I)$, where the true mean vector is $\theta_0 = (1, 1, 0.5, 0.5, 0.4, 0)^{\mathsf{T}}$, and I is the 6×6 identity matrix. We compare $K = 7$ candidate models, each of which is $\mathcal{N}(\theta, I)$ but with a different sparsity pattern imposed on the mean vector θ, and we define the complexity $c(k)$ to be the number of nonzero entries of θ. The models vary in complexity and KL divergence from the truth, $\min_{\theta_k} D\big(\mathcal{N}(\theta_0, I) \,\|\, \mathcal{N}(\theta_k, I)\big)$ subject to the sparsity pattern in Table 1; see Section S3.2.

Table 1: Candidate models for sparse multivariate normal simulation example.

k	θ_k	$c(k)$	$\min D$
1	$(\theta_1, 0, 0, \theta_4, 0, 0)^{\mathsf{T}}$	2	0.705
2	$(\theta_1, \theta_2, 0, 0, 0, 0)^{\mathsf{T}}$	2	0.33
3	$(\theta_1, \theta_2, 0, 0, \theta_5, 0)^{\mathsf{T}}$	3	0.25
4	$(\theta_1, \theta_2, 0, \theta_4, 0, 0)^{\mathsf{T}}$	3	0.205
5	$(\theta_1, \theta_2, \theta_3, 0, 0, 0)^{\mathsf{T}}$	3	0.205
6	$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, 0)^{\mathsf{T}}$	5	0
7	$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)^{\mathsf{T}}$	6	0

Models 1 and 2 are the most parsimonious, but differ greatly in their distance from the truth. Models 4 and 5 are equally close to the truth and have the same complexity, while model 3 is slightly farther in KL divergence but has the same complexity as 4 and 5. Models 6 and 7 both recover the true DGP exactly, but model 6 is simpler.

We implement the LaD approach in Algorithm 1 as follows. We use the MLE to estimate the parameters for model k, which is the vector $\hat{\theta}_k \in \mathbb{R}^6$ containing the sample means in the coordinates that are free in that model, and zeros elsewhere. That is, for coordinate $j \in \{1, \ldots, 6\}$, $\hat{\theta}_{kj} = \bar{x}_j \mathbb{1}(j \in J_k)$ where $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ and $J_k \subseteq \{1, \ldots, 6\}$ is the set where model k allows θ_{kj} to be nonzero. The bias-corrected LaD values are then $Z_{ik} = \ell_k(x_i; \hat{\theta}_k) + d_k/(2n)$ where $\ell_k(x_i; \hat{\theta}_k) = \frac{6}{2} \log(2\pi) + \frac{1}{2}\|x_i - \hat{\theta}_k\|^2$ and $d_k = |J_k|$. We draw $T = 1000$ samples of $\mu \mid Z_{1:n}$ (Equation 15) and compute the SLC scores $\hat{w}_\delta(k)$ using Algorithm 2 with a temperature of $\alpha_n = n^{0.45}$. We consider three tolerance levels: (i)
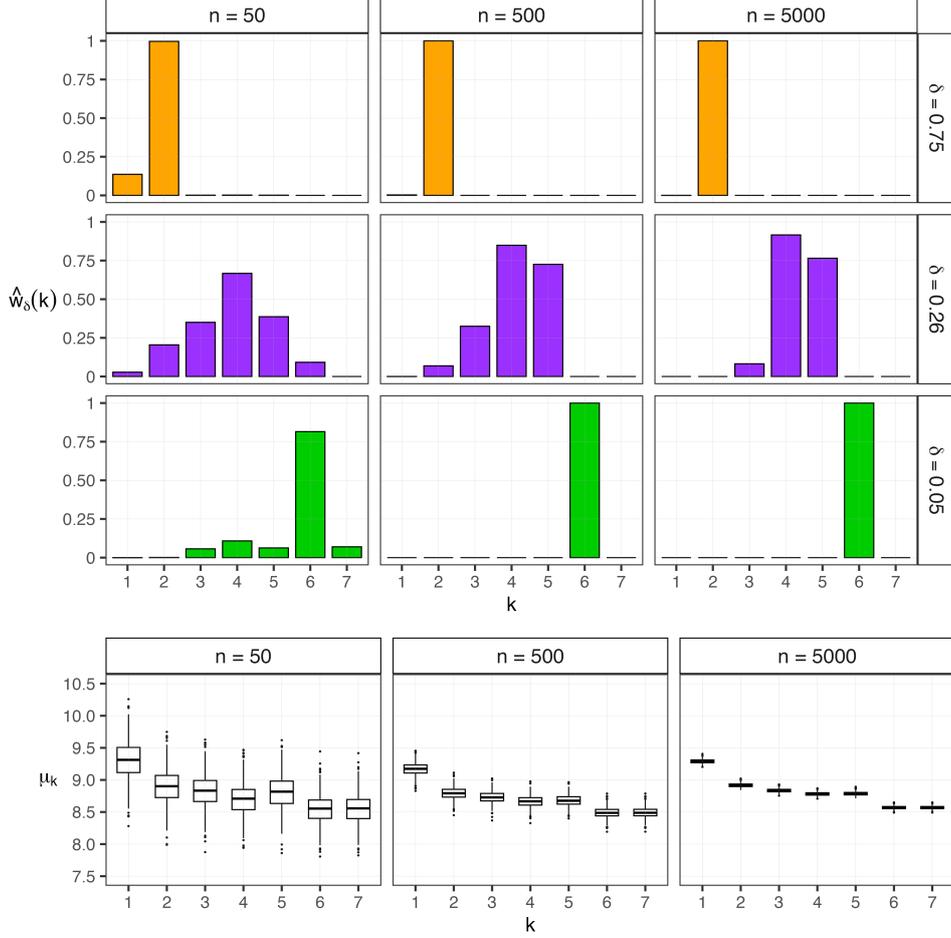
Figure 4: Sparse multivariate normal model results. *(Top)* SLC scores $\hat{w}_\delta(k)$ for sample sizes $n \in \{50, 500, 5000\}$ (columns) and tolerances $\delta \in \{0.75, 0.26, 0.05\}$ (rows). *(Bottom)* Boxplots of the LaD posteriors on μ_k for $k \in \{1, \ldots, 7\}$.

$\delta = 0.75$ is large enough to tolerate all of the models, (ii) $\delta = 0.26$ excludes models 1 and 2 but still permits models 3–7, and (iii) $\delta = 0.05$ is small enough that only models 6 and 7 are within δ of the true DGP; see Table 1. To interpret the value of δ, we compute the rescaled tolerances τ as in Algorithm 1, using $\mathcal{N}(0, I)$ for the noise model. For instance, the δ values of 0.75, 0.26, and 0.05 correspond to τ values of approximately 55.3%, 18.8%, and 3.7%, respectively.

Figure 4 summarizes the LaD results across the δ settings, for sample sizes $n \in \{50, 500, 5000\}$. Here, we show results for one representative data set of each size. In the first setting ($\delta = 0.75$), both models 1 and 2 are given a nonnegligible SLC score when n is small, but the score of model 1 quickly goes to zero and the score of model 2 goes to 1 as n grows. This is because all candidate models are δ-optimal at this tolerance level, but

models 3–7 have greater complexity, and model 1 has worse fit than model 2. The SLC score concentrates at model 2 since it is the best minimal-complexity δ-optimal model at this tolerance; see Table 1.

In the second setting ($\delta = 0.26$), the SLC scores are spread more evenly across models when n is small, due to posterior uncertainty. As n increases, the SLC scores concentrate on both models 3 and 4, which have equal complexity and equal KL divergence from the true DGP. This occurs because models 1 and 2, while simpler, are not δ-optimal at this tolerance level; models 6 and 7 have higher complexity; and model 3 has larger KL from the true DGP than models 4 and 5. Notably, the SLC score is stable in the sense that it does not randomly select one of models 4 and 5, but instead gives them roughly equal score. Thus, the method correctly treats these two models as equally good with respect to this true DGP.

In the third setting ($\delta = 0.05$), the SLC scores concentrate on model 6. This is because models 1–5 are not δ-optimal at this tolerance level, and model 7 is more complex than model 6. Thus, the method correctly chooses model 6 at this value of δ.

5.2.1 Performance comparison with other methods

Up to this point, we have considered qualitative aspects of performance in our comparisons with existing methods. In this section, we quantitatively compare model selection performance with several alternative approaches. For these comparisons, we consider the following methods, including three versions of LaD:

(i) (LaD-soft) - Our proposed LaD method (see Algorithm 1).

(ii) (LaD-hard) - A variant of LaD using hard-minimum selection (see Theorem 4.4(b)).

(iii) (LaD-diag) - A variant of LaD using diagonal covariance Σ (see Section S4).

(iv) (c-posterior) - Coarsened posterior with powers $\alpha \in \{10, 100\}$ (see Section S3.2).

(v) (Bayes) - Usual Bayesian posterior probabilities on models.

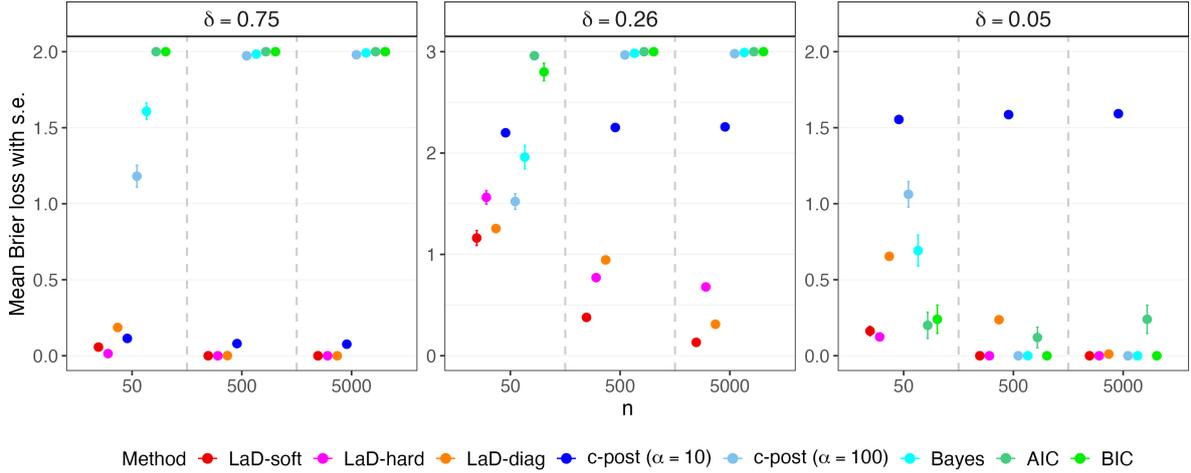(vi) (AIC, BIC) - Standard model selection criteria that select a single model.

Figure 5: Performance comparisons using Brier score. Points show the mean Brier loss, and lines show $\pm$ standard error across 50 datasets for sample sizes $n \in \{50, 500, 5000\}$ (x-axis; vertical dashed lines separate values of n). Each panel corresponds to tolerance settings: left $\delta = 0.75$, middle $\delta = 0.26$, and right $\delta = 0.05$. Colors indicate the competing methods.

For each $n \in \{50, 500, 5000\}$, we generate 50 simulated data sets from $\mathcal{N}(\theta_0, I)$ as above and, for each data set, we apply each method to select from the set of candidate models in Table 1. To quantify performance for each δ, consider the Brier loss $S_\text{B} := \sum_{k=1}^{K}(w_k - w_k^0)^2$ where w_k is the weight assigned to model k by a given method and w_k^0 is an indicator of whether model k is in the true target set, that is, $w_k^0 = \mathbb{1}(k \in M_\delta^*(\mu^0))$.

Figure 5 displays the means and standard errors of the Brier loss S_B for each tolerance setting, $\delta \in \{0.75, 0.26, 0.05\}$. LaD-soft attains the smallest Brier loss on average across all n and δ. LaD-hard performs comparably to LaD-soft except when there are ties, such as between models 4 and 5 in the case of $\delta = 0.26$, in which case LaD-soft performs better. LaD-diag works reasonably well but is generally worse than LaD-soft, particularly when $\delta = 0.05$, demonstrating that modeling the full covariance matrix is important. The c-posterior with $\alpha = 10$ is competitive at tolerance $\delta = 0.75$, but performs poorly at smaller δ. Meanwhile, the c-posterior with $\alpha = 100$ performs well at $\delta = 0.05$, but poorly at larger δ. This illustrates that the choice of α for the c-posterior needs to be selected in a δ-dependent way, but this is challenging because α does not have a direct interpretation in terms of distance from the true DGP. Finally, Bayes, AIC, and BIC are only competitive at $\delta = 0.05$ since $M_\delta^*(\mu^0)$ only contains the true DGP in this case. Overall, LaD-soft reliably provides excellent performance across all settings.

5.3 Thermal performance curves in microbial ecology

Thermal performance curves (TPCs)—unimodal functions relating performance (such as growth rate) to temperature—are widely used in microbial ecology for understanding how populations respond to environmental conditions (Sinclair et al., 2016). For example, *Prorocentrum minimum* is a mixotrophic dinoflagellate that can form harmful blooms; studying its temperature–growth relationship helps predict these events and mitigate their impact on marine ecosystems and human health (Grzebyk and Berland, 1996). There are many mechanistic and phenomenological TPC models, none of which is universally optimal across traits or taxa (Kontopoulos et al., 2024; Kellermann et al., 2019; Angilletta Jr, 2006). We analyze the *P. minimum* data of Grzebyk and Berland (1996) consisting of $n = 148$ growth-rate observations at 10 temperatures from 10°C to 31°C.

We fit $K = 9$ widely used TPC models of varying complexity: three-parameter models (Gaussian, Mitchell–Angilletta, Brière I, Eubank); four-parameter models (Weibull, Modified Gaussian); a five-parameter model (extended Brière); a six-parameter model (Poly-5); and a seven-parameter model (Sharpe–Schoolfield). These represent a diverse mix of TPC models that are commonly recommended for fitting microbial growth data (Kontopoulos et al., 2024; Grimaud et al., 2017; Thomas et al., 2012); see Section S3.3 for their functional forms.

The data consist of pairs (y_i, t_i) where y_i is the outcome (growth rate) and t_i is the temperature, for observations $i = 1, \ldots, n$. We implement LaD as follows (Algorithm 1). For each candidate model k with parameter θ_k and mean curve $g_k(t; \theta_k)$, we fit $\hat{\theta}_k$ by nonlinear least squares (NLS), modeling $y_i \sim \mathcal{N}(g_k(t_i; \theta_k), \sigma_k^2)$, a standard approach for these models implemented in R (Padfield et al., 2021). Figure 6(A) shows the fitted curve for each model, overlaid on a scatterplot of the data. The bias-corrected LaD values are $Z_{ik} = \ell_k((y_i, t_i); \hat{\theta}_k) + d_k/(2n)$ where $\ell_k((y_i, t_i); \hat{\theta}_k) = \frac{1}{2} \log(2\pi \hat{\sigma}_k^2) + (y_i - g_k(t_i; \hat{\theta}_k))^2 / (2\hat{\sigma}_k^2)$, $\hat{\sigma}_k^2 = \frac{1}{n} \sum_i (y_i - m_k(T_i; \hat{\theta}_k))^2$, and d_k is the dimension of θ_k plus 1 for the variance. We draw $T = 1000$ samples of $\mu \mid Z_{1:n}$ (Equation 15) and compute the SLC scores $\hat{w}_\delta(k)$ using Algorithm 2 with complexities $c(k) = d_k$ and temperature $\alpha_n = n^{0.45}$.

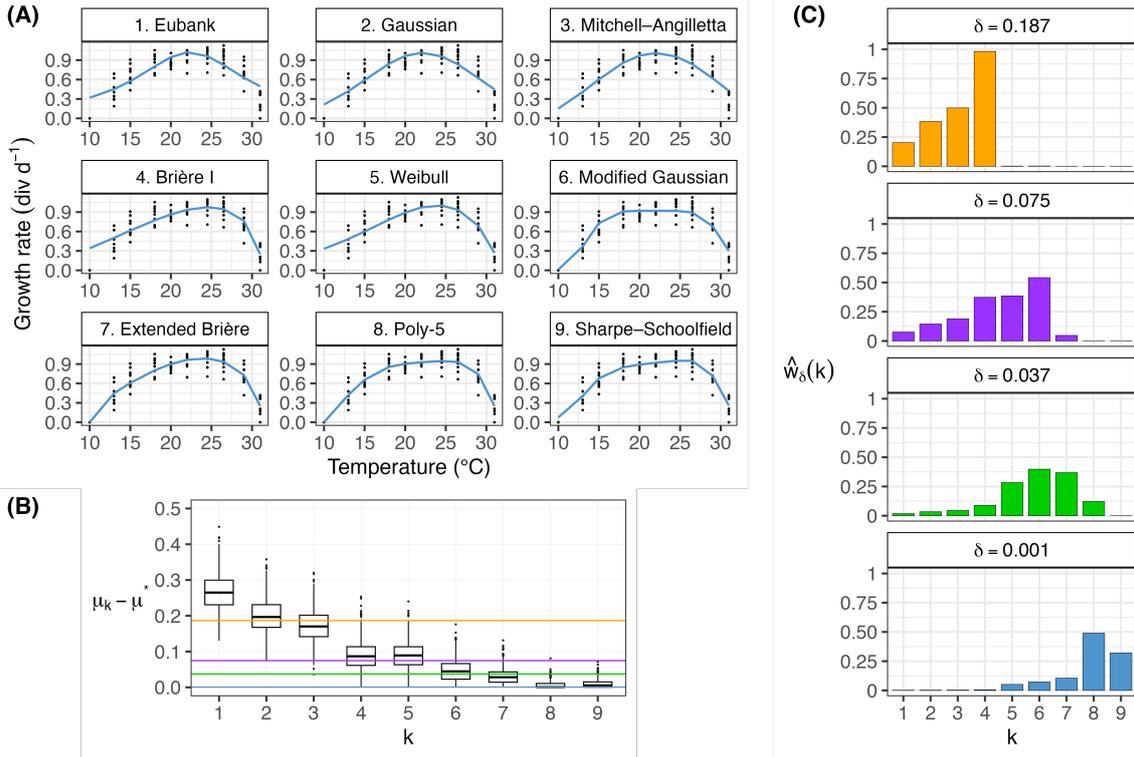To define the rescaled tolerances $\hat{\tau}$, we introduce both a noise model and a flexible

Figure 6: (A) NLS fits (blue lines) of the $K = 9$ candidate TPC models to the $n = 148$ growth-rate observations (black points) across temperatures 10°C – 31°C. Growth rates are reported as divisions per day ($\mathrm{div\,d^{-1}}$), converted from per-second rates ($\mathrm{s^{-1}}$) via $\mathrm{div\,d^{-1}} = 86{,}400/\ln 2 \times \mathrm{s^{-1}}$, as in Grzebyk and Berland (1996). (B) Posterior distributions of $\mu_k - \min_j \mu_j$ for each model k. Horizontal lines indicate selected δ thresholds (orange: $\delta = 0.187$, purple: $\delta = 0.075$, green: $\delta = 0.037$, and blue: $\delta = 0.001$). (C) SLC scores $\hat{w}_\delta(k)$ across models k for each δ, using the same color for each δ as in (B).

nonparametric model to represent upper and lower bounds, respectively, on the KL divergences from the true DGP. The noise model is defined as $y_i \sim \mathcal{N}(\varphi, \sigma_{\text{noise}}^2)$, and the flexible model is $y_i \sim \mathcal{N}(\beta_{t_i}, \sigma_{\text{flex}}^2)$ where $\beta = (\beta_t : t \in \mathcal{T})$ is defined at the set of 10 observed temperatures, $\mathcal{T}$. For the noise model and the flexible model, we compute the MLEs of the parameters and take the average negative log-likelihood to obtain $\hat{\mu}_{\text{noise}}$ and $\hat{\mu}_{\text{flex}}$, respectively; see Section S3.3 for details. We set $\hat{\tau} = \delta/(\hat{\mu}_{\text{noise}} - \hat{\mu}_{\text{flex}})$.

Figure 6(B) displays the distribution of $\mu_k - \min_j \mu_j$ for each k under the posterior $\mu \mid Z_{1:n}$, and we overlay lines indicating the δ tolerance values used in panel (C). In Figure 6(C), we show the SLC scores $\hat{w}_\delta(k)$ for $\delta = 0.187$ (orange), 0.075 (purple), 0.037 (green), and 0.001 (blue). The SLC scores behave as follows: (i) small tolerance ($\delta = 0.001$, $\hat{\tau} = 0.1\%$) put higher weight on the most flexible models such as Sharpe-Schoolfield and Poly-5; (ii) moderate tolerance ($\delta \in \{0.037, 0.075\}$, $\hat{\tau} \in \{5\%, 10\%\}$) shifts selection to simpler 4–5 parameter models; (iii) large tolerance ($\delta = 0.187$, $\hat{\tau} = 25\%$) shifts to the 3-parameter models, with Brière I receiving the highest score. Notably, the Eubank and Gaussian models receive relatively low weight in panel (C) across all δ values considered, indicating that these two models are not appropriate for this data set, regardless of the tolerance level.

5.4 Population structure admixture models

Inferring latent population structure from multilocus genotype data is a central task in population genetics (Pritchard et al., 2000; Rosenberg et al., 2002). In many realistic settings, individuals are admixed — that is, their genomes originate from multiple ancestral populations. Modeling this admixture requires decomposing each individual's genotype into proportions contributed by k unknown source populations, while estimating allele frequencies in each population. Bayesian admixture models provide a formal probabilistic framework for this task.

A widely used software program for admixture models is STRUCTURE (Pritchard et al., 2000), which performs posterior inference via Markov chain Monte Carlo. In the model, each individual is associated with a vector of ancestry proportions over k popula-

tions, however, STRUCTURE requires the user to specify k in advance. This introduces a model selection problem: The true number of populations is unknown, and must be chosen to balance fit and complexity. If k is too small, genetically distinct groups may be merged, obscuring meaningful structure. If k is too large, the model may chase noise and infer spurious populations.

In practice, users often choose k by comparing in-sample log-likelihoods or by applying *ad hoc* heuristics such as Evanno's method (Evanno et al., 2005), but these procedures lack theoretical justification. We address this problem by applying the LaD framework to perform principled, robust model selection in admixture models. The aim is to capture the explainable structure while avoiding unnecessary complexity.

We apply our methodology to the brook trout data set of Erdman et al. (2022), consisting of multilocus genotypes for $n = 8454$ fish sampled from multiple streams in the Eastern United States. Each individual is genotyped at $L = 5$ loci with two allele copies per locus. Specifically, the genotype for individual i is $x_i = [(x_{il1}, x_{il2})]_{l=1}^{L}$ where $x_{ila} \in \{1, \ldots, V_l\}$ is the allele observed at locus l, allele copy $a \in \{1, 2\}$. Following Pritchard et al. (2000), the admixture model assumes alleles are drawn from one of k latent populations with allele frequencies $\phi_{jl} = (\phi_{jl1}, \ldots, \phi_{jlV_l}) \sim \text{Dirichlet}(1, \ldots, 1)$ for $j = 1, \ldots, k$, according to ancestry proportions $q_i = (q_{i1}, \ldots, q_{ik}) \sim \text{Dirichlet}(\psi)$, where $\psi = (\psi_1, \ldots, \psi_k)$ is given a hyperprior, $\psi_j \sim \text{Uniform}(0, 10)$ independently. Specifically,

$$s_{ila} \sim \text{Categorical}(q_i),$$

$$x_{ila} \mid s_{ila} = j \sim \text{Categorical}(\phi_{jl}).$$

The global parameters are $\theta := (\phi, \psi)$, whereas q_i and s_{ila} are individual-specific.

We implement the LaD approach in Algorithm 1 as follows. For each $k = 1, \ldots, 10$, we do the following. For a given k, we run STRUCTURE 20 times and keep the run with the largest estimated log probability; see Section S3.4. We take the posterior means $\hat{\phi}_{jl}$ and $\hat{\psi}$ from that run, and compute per-individual negative log-likelihoods $\ell_k(x_i; \hat{\theta}) = -\log p(x_i \mid \hat{\phi}, \hat{\psi})$ by integrating out $q_i \mid \hat{\psi}$ and $s_{ila} \mid q_i$ using Monte Carlo; see Section S3.4 for details. It is necessary to integrate out q_i and s_{ila} when defining the negative log-likelihood
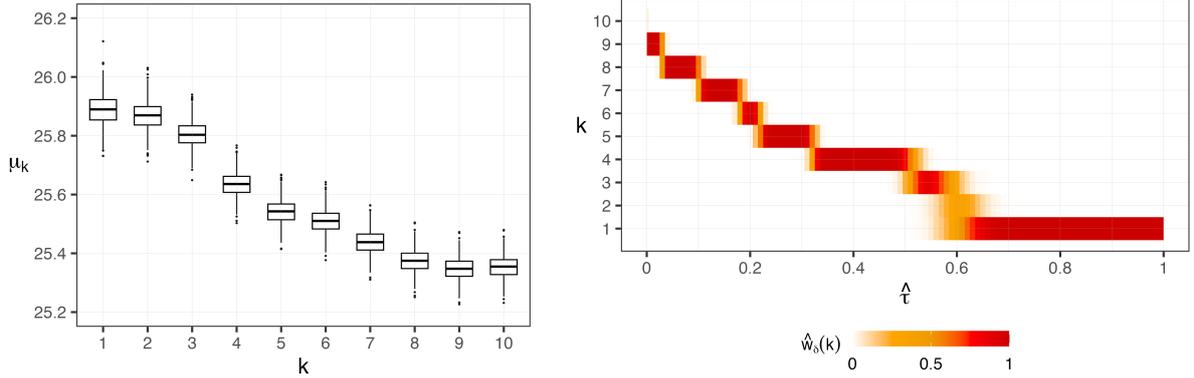
Figure 7: *(Left)* Boxplot of the posterior of μ_k for each candidate number of populations $k = 1, \ldots, 10$. *(Right)* Posterior path showing the SLC scores for $k = 1, \ldots, 10$ (y-axis) versus the rescaled tolerance $\hat{\tau} \in [0, 1]$ (x-axis), where $\hat{\tau} = \delta/(\hat{\mu}_{\text{noise}} - \min_k \hat{\mu}_k)$. Darker red indicates a higher SLC score.

values because these are observation-specific latent variables. The bias-corrected LaD values are then $Z_{ik} = \ell_k(x_i; \hat{\theta}) + d_k/(2n)$ where $d_k = k \sum_{l=1}^{L}(V_l - 1) + k$, since each ϕ_{jl} contributes $V_l - 1$ degrees of freedom and ψ contributes k. We then apply Algorithm 2 with $\alpha_n = n^{0.45}$ and complexity $c(k) = k$ to compute the SLC scores $\hat{w}_\delta(k)$. To compute the rescaled tolerances $\hat{\tau}$, we take the noise model to be discrete uniform at each locus: $x_{ila} \sim \text{Uniform}(\{1, \ldots, V_l\})$ independently for all i, l, a, which yields $\hat{\mu}_{\text{noise}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} N_{il} \log V_l$, where $N_{il} \in \{0, 1, 2\}$ is the number of observed copies at locus l for individual i.

Figure 7 shows the LaD results for the brook trout data. The boxplots (left) show the posterior distribution of μ_k for each candidate number of populations k, and the heatmap (right) shows the SLC score $\hat{w}_\delta(k)$ as a function of the rescaled tolerance $\hat{\tau}$. As we increase $\hat{\tau}$ from 0 to 1, the selected number of populations k gradually transitions from higher (more complex models) to lower (simpler models). This provides clear insights into the tradeoff between model complexity and the cost in terms of KL divergence. For instance, the $k = 4$ population model can capture 65% of the improvement in fit (that is, $\hat{\tau} = 0.35$) from the noise model to the most complex candidate model.

For comparison, Figure 8 shows the results for Evanno's method (Evanno et al., 2005), which is widely used to select k for STRUCTURE. Evanno's method is based on finding a sharp elbow in the mean log-likelihood $L(k)$ (see Section S3.4), or more precisely, a
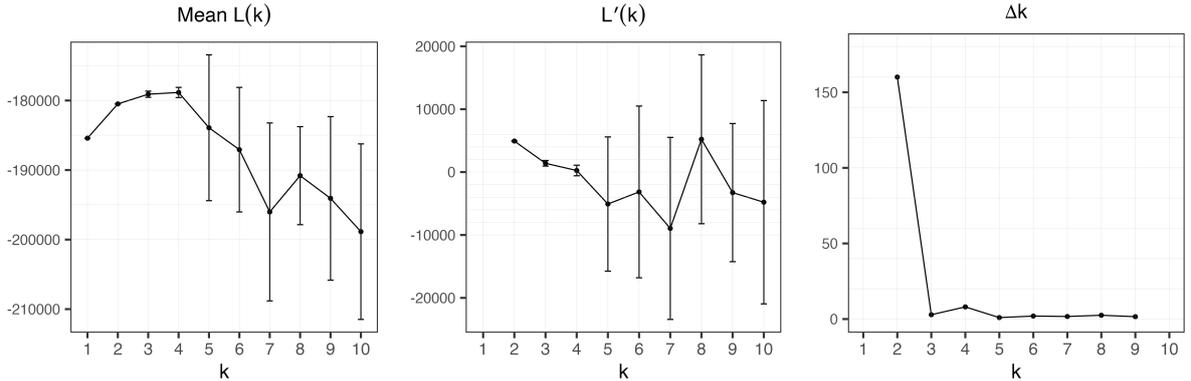
Figure 8: Evanno's method. *(Left)* Mean $L(k)$ increases with k; error bars show standard deviation across replicates. *(Middle)* First difference $L'(k)$. *(Right)* $\Delta k = \mathbb{E}|L''(k)|/\mathrm{sd}\big(L(k)\big)$ peaks at $k = 2$. Here, $|L''(k)|$ is the absolute second difference.

large value of Δk, defined as the second-order difference of $L(k)$ divided by the standard deviation of $L(k)$ across runs. Evanno's method strongly suggests that $k = 2$ should be selected on these data; see Figure 8. Interestingly, however, our LaD results in Figure 7 show that $k = 2$ is not a particularly useful choice. Using $k = 2$ captures around 40% of the possible improvement in fit ($\hat{\tau} = 0.6$), but Figure 7 shows that using $k = 1$ is nearly as good as $k = 2$ on these data in terms of fit. It is necessary to jump up to $k = 3$ or 4 to obtain more substantial gains in fit to the true DGP.

6 Discussion

The Likelihood as Data (LaD) framework is a novel approach to comparing models based on treating per-observation negative log-likelihoods as data to be analyzed. LaD quantifies uncertainty in the KL divergence between the true DGP and each candidate model, enabling robust model selection under misspecification by quantitatively characterizing the tradeoff between simplicity and fit. The LaD approach is simple to compute, asymptotically consistent, and stable in the presence of ties.

Although LaD is a versatile method, its performance hinges on several assumptions. These include independent observations, finite second moments of the per-observation negative log-likelihoods, and accurate point estimators for the model parameters. If these assumptions are not satisfied, LaD may be unstable or exhibit overconfidence. Further-

more, when there are a large number of candidate models, LaD may encounter difficulties due to the computational burden of exhaustively evaluating all models and estimating the full $K \times K$ covariance Σ.

There are several interesting directions for future work. Placing a prior on $\mu_1, \ldots, \mu_K$ that allows $\mu_k = \mu_j$ with positive probability may improve the stability of using the plug-in posterior $P(k \in M_\delta^*(\mu) \mid Z_{1:n})$ rather than our SLC score. Extending LaD to general loss functions beyond negative log-likelihood and developing a frequentist analogue based on sampling distributions are also promising directions. Large K could be handled by clustering the columns of Z and working at the cluster level rather than over all K. Finally, the LaD idea can be used more broadly beyond model selection. For example, one could explore the relationships among models and among data points via dimensionality reduction of the Z matrix.

Acknowledgments

References

H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, AC-19(6):716–723, 1974.

M. J. Angilletta Jr. Estimating and comparing thermal performance curves. *Journal of Thermal Biology*, 31(7):541–545, 2006.

R. H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, Chichester, 2000.

G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.

M. J. Drinkwater, Q. A. Parker, D. Proust, E. Slezak, and H. Quintana. The large scale distribution of galaxies in the Shapley supercluster. *Publications of the Astronomical Society of Australia*, 21(1):89–96, 2004.

B. Erdman, M. G. Mitro, J. D. Griffin, D. Rowe, D. C. Kazyak, K. Turnquist, M. Siepker, L. Miller, W. Stott, M. Hughes, et al. Broadscale population structure and hatchery introgression of midwestern brook trout: Midwestern brook trout population genetics. *Transactions of the American Fisheries Society*, 151(1):81–99, 2022.

G. Evanno, S. Regnaut, and J. Goudet. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, 14(8):2611–2620, 2005.

C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007.

G. M. Grimaud, F. Mairet, A. Sciandra, and O. Bernard. Modeling the temperature effect on the specific growth rate of phytoplankton: a review. *Reviews in Environmental Science and Bio/Technology*, 16(4):625–645, 2017.

D. Grzebyk and B. Berland. Influences of temperature, salinity and irradiance on growth of *Prorocentrum minimum* (Dinophyceae) from the Mediterranean sea. *Journal of Plankton Research*, 18(10):1837–1849, 1996.

J. H. Huggins and J. W. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79, 2023.

V. Kellermann, S. L. Chown, M. F. Schou, I. Aitkenhead, C. Janion-Scheepers, A. Clemson, M. T. Scott, and C. M. Sgrò. Comparing thermal performance curves across traits: how consistent are they? *Journal of Experimental Biology*, 222(11):jeb193433, 2019.

D.-G. Kontopoulos, A. Sentis, M. Daufresne, N. Glazman, A. I. Dell, and S. Pawar. No universal mathematical model for thermal performance curves across traits and taxonomic groups. *Nature Communications*, 15(1):8855, 2024.

D. Padfield, H. O'Sullivan, and S. Pawar. rTPC and nls.multstart: A new pipeline to fit thermal performance curves in R. *Methods in Ecology and Evolution*, 12(6):1138–1143, 2021.

G. Pescador-Barrios, S. L. Filippi, and M. Van Der Wilk. Adjusting model size in continual Gaussian processes: How big is big enough? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48974–49000. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pescador-barrios25a.html.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298(5602): 2381–2385, 2002.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289, 2016.

B. J. Sinclair, K. E. Marshall, M. A. Sewell, D. L. Levesque, C. S. Willett, S. Slotsbo, Y. Dong, C. D. Harley, D. J. Marshall, B. S. Helmuth, et al. Can we predict ectotherm responses to climate change using thermal performance curves and body temperatures? *Ecology Letters*, 19(11):1372–1385, 2016.

M. K. Thomas, C. T. Kremer, C. A. Klausmeier, and E. Litchman. A global pattern of thermal adaptation in marine phytoplankton. *Science*, 338(6110):1085–1088, 2012.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

Supplementary material for "Robust model selection using likelihood as data"

S1 Previous work

To the best of our knowledge, no prior work proposes LaD in the form we use or develops the theoretical motivation we rely on. This section briefly summarizes the most closely related previous work.

The core idea behind LaD is related to Vuong's test (Vuong, 1989), which compares two parametric models using the average per-observation log-likelihood difference and tests the null that the models are equally close to the true DGP, where "closeness" is defined in terms of KL divergence. Similarly, Diebold and Mariano (1995) propose a test for equal predictive accuracy of two competing forecasts, again yielding a pairwise testing tool rather than a joint analysis of multiple models. Several extensions of Vuong's test and other related model selection methods have been proposed (see, for example, Shimodaira, 2001; Royall and Tsou, 2003; Royall, 2004; Clarke, 2007; Barmalzan and Payandeh Najafabad, 2013; Diebold, 2015; Schennach and Wilhelm, 2017; Sayyareh, 2022; Brück et al., 2023), but none of these are close to the LaD method, which extends to more than two models and jointly quantifies uncertainty in how well each model approximates the DGP.

Several methods have been proposed for finding a parsimonious model whose fit is nearly as good as the best available model. Widely used heuristics such as the one-standard-error rule (Hastie et al., 2009; Chen and Yang, 2021) select the most parsimonious model whose cross-validated error is within one estimated standard error of the minimum. Hansen et al. (2011) introduce the model confidence set (MCS), which compares multiple candidate models via sequential testing and elimination: clearly inferior models are removed until the remaining set contains the best model(s) with a chosen confidence level. In Bayesian work, Goutis and Robert (1998) and Dupuis and Robert (2003) quantify the loss of explanatory power incurred by KL projection from a full model

to a submodel and retain only those submodels whose estimated loss stays below a pre-specified threshold; see also Vehtari and Lampinen (2004) and Lindsay and Liu (2009). Our LaD method also provides a way of tackling this goal, but operates very differently: it works directly with expected log-likelihood (equivalently, KL distance), incorporates model complexity in a transparent way, and yields a Bayesian posterior over the full vector of KL distances.

Another line of work addresses misspecification by modifying the posterior update itself. In contrast with classical robust Bayesian analysis, which focuses on sensitivity to the prior (Berger, 1994), these more recent approaches robustify likelihood-based inference by modifying the likelihood, for instance, using the coarsened posterior (Miller and Dunson, 2019) or generalized Bayes updates based on loss functions (Bissiri et al., 2016). There are numerous advances in this direction for parametric inference (Knoblauch et al., 2022; Fong et al., 2023; Huggins and Miller, 2024; Frazier et al., 2025; Dewaskar et al., 2025), but not primarily for model selection.

Meanwhile, for robust model selection, there have also been recent advances in Bayesian methodology. Shao et al. (2019) introduce a method that replaces the log marginal likelihood with the Hyvärinen score, a proper scoring rule to obtain Bayes-factor-like comparisons that remain well defined under diffuse or improper priors. Li and Dunson (2020) propose D-probabilities that compare parametric models to a flexible nonparametric reference via KL divergence, yielding complexity-penalized weights that are less sensitive to prior specification than standard posterior model probabilities. Huggins and Miller (2023) consider BayesBag, a robust Bayesian model selection method that applies bagging to the posterior by averaging posterior model probabilities over bootstrap-resampled versions of the data. More recently, Shirvaikar et al. (2024) propose a probabilistic framework for model uncertainty based on forward sampling of missing observations using one-step-ahead prediction.

There is a mature literature on predictive Bayesian approaches to model assessment, selection, and averaging—including cross-validation and predictive aggregation methods (Vehtari and Ojanen, 2012; Yao et al., 2018; Sivula et al., 2025; McLatchie et al., 2025).

These approaches are appealing for predictive performance, but their target is typically a predictive utility rather than uncertainty quantification for the KL divergence between the true DGP and each candidate model.

S2 Proofs

In this section, we provide proofs of the results in Section 4. For $x \in \mathbb{R}^d$, $\|x\|$ denotes the Euclidean norm and for $A \in \mathbb{R}^{d \times d}$, $\|A\|$ denotes the Frobenius norm. For $r > 0$ and $x_0 \in \mathbb{R}^d$, define the open ball $B_r(x_0) := \{x \in \mathbb{R}^d : \|x - x_0\| < r\}$.

Condition S2.1. *Suppose $X, X_1, X_2, \ldots, X_n$ are i.i.d. random variables defined on a probability space $(\mathcal{X}, \mathcal{F}, P)$. Assume the following conditions for all models $k = 1, \ldots, K$. Suppose $\Theta_k \subseteq \mathbb{R}^{d_k}$ is open and $\ell_k : \mathcal{X} \times \Theta_k \to \mathbb{R}$ is measurable. Throughout, expectations are taken with respect to P, and Θ_k is endowed with the Borel sigma-algebra. Let $\Theta_k^* = \arg\min_{\theta_k \in \Theta_k} \mathbb{E}\big(\ell_k(X; \theta_k)\big)$, assumed to be nonempty. Assume the following conditions hold for some fixed $\theta_k^* \in \Theta_k^*$ and some $r > 0$ such that $B_r(\theta_k^*) \subseteq \Theta_k$.*

(a) $\mathbb{E}\big(|\ell_k(X; \theta_k)|\big) < \infty$ for all $\theta_k \in B_r(\theta_k^*)$.

(b) For all $x \in \mathcal{X}$, the map $\theta_k \mapsto \ell_k(x; \theta_k)$ is twice continuously differentiable on $B_r(\theta_k^*)$.

(c) There exists a function $H : \mathcal{X} \to [0, \infty)$ such that $\sup_{\theta_k \in B_r(\theta_k^*)} \|\nabla_{\theta_k}^2 \ell_k(x; \theta_k)\| \leq H(x)$ for all $x \in \mathcal{X}$, and $\mathbb{E}\big(H(X)\big) < \infty$.

(d) $\mathbb{E}\left(\big\|\nabla_{\theta_k} \ell_k(X; \theta_k^*)\big\|^2\right) < \infty$.

(e) There exists an estimator $\hat{\theta}_k = \hat{\theta}_k(X_1, \ldots, X_n) \in \Theta_k$ such that $\|\hat{\theta}_k - \theta_k^*\| = O_p(1/\sqrt{n})$.

We use the following lemma to combine local bounds with high probability events.

Lemma S2.2. *Let $A_1, A_2, \ldots$ be events such that $P(A_n) \to 1$ as $n \to \infty$. Suppose X_n and Y_n are random variables such that $|X_n| \leq Y_n$ on A_n, and $Y_n \geq 0$ almost surely. If $Y_n = o_p(1)$, then $X_n = o_p(1)$. If $Y_n = O_p(1)$, then $X_n = O_p(1)$.*

Proof of Lemma S2.2. First, assume $Y_n = o_p(1)$. Let $\epsilon > 0$. Then

$$P\big(|X_n| > \epsilon\big) \le P(A_n^c) + P\big(|X_n| > \epsilon, A_n\big) \le P(A_n^c) + P(Y_n > \epsilon) \longrightarrow 0 \qquad (17)$$

because $Y_n = o_p(1)$ and $P(A_n) \to 1$. Since $\epsilon > 0$ is arbitrary, $X_n = o_p(1)$. Next, assume $Y_n = O_p(1)$. Let $\delta > 0$. Since $Y_n = O_p(1)$, there exists $M < \infty$ such that $P\big(|Y_n| > M\big) < \delta/2$ for all sufficiently large n. Since also $P(A_n^c) \to 0$, then for all sufficiently large n,

$$P\big(|X_n| > M\big) \le P(A_n^c) + P\big(|X_n| > M, A_n\big) \le P(A_n^c) + P\big(|Y_n| > M\big) < \delta. \qquad (18)$$

This shows that $X_n = O_p(1)$. $\qquad\qquad\square$

S2.1   Plug-in approximation error

Theorem S2.3 establishes that replacing the unknown optimal parameters θ_k^* with their estimates $\hat{\theta}_k$ introduces error $O_p(1/n)$ in the average log-likelihood, which is negligible compared to the empirical approximation error of order $O_p(1/\sqrt{n})$.

Theorem S2.3. *Assume Condition S2.1 holds for each $k = 1, \ldots, K$. Define*

$$\overline{Z}_{nk}(\theta_k) = \frac{1}{n} \sum_{i=1}^{n} \ell_k(X_i; \theta_k),$$

for $\theta_k \in \Theta_k$. Let $\overline{Z}_n(\theta) = \big(\overline{Z}_{n1}(\theta_1), \ldots, \overline{Z}_{nK}(\theta_K)\big)^{\mathsf{T}}$ for $\theta = (\theta_1, \ldots, \theta_K)$. Then

$$\overline{Z}_n(\hat{\theta}) = \overline{Z}_n(\theta^*) + O_p(1/n). \qquad (19)$$

Proof of Theorem S2.3. We fix k and show Equation 19 holds in the kth component. For notational clarity, we suppress the subscript k in the proof, and we define the function $R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i; \theta)$ to represent $\overline{Z}_{nk}(\theta_k)$. Define the event $A_n = \{\|\hat{\theta} - \theta^*\| < r\}$, that is, $A_n = \{\hat{\theta} \in B_r(\theta^*)\}$. Then $P(A_n) \to 1$ as $n \to \infty$, since $\|\hat{\theta} - \theta^*\| = o_p(1)$ by Condition S2.1(e). On the event A_n, since $B_r(\theta^*)$ is convex, the line segment between θ^* and $\hat{\theta}$ is contained in $B_r(\theta^*)$. Furthermore, for all $x \in \mathcal{X}$, the map $\theta \mapsto \ell(x; \theta)$ is twice

continuously differentiable on $B_r(\theta^*)$ by Condition S2.1(b). Thus, on event A_n, Taylor's theorem implies that there exists some $\tilde{\theta}$ on the line segment between θ^* and $\hat{\theta}$ such that

$$R_n(\hat{\theta}) = R_n(\theta^*) + \nabla_\theta R_n(\theta^*)^\mathsf{T}(\hat{\theta} - \theta^*) + \frac{1}{2}(\hat{\theta} - \theta^*)^\mathsf{T}\nabla_\theta^2 R_n(\tilde{\theta})(\hat{\theta} - \theta^*). \qquad (20)$$

To make $\tilde{\theta}$ well-defined on the whole sample space, we define $\tilde{\theta} = \theta^*$ on A_n^c. Then, by Lemma S2.2, it suffices to show that the first- and second-order terms in Equation 20 are $O_p(1/n)$. Since $P(A_n) \to 1$, this will imply that $|R_n(\hat{\theta}) - R_n(\theta^*)| = O_p(1/n)$ by Lemma S2.2, proving the result (Equation 19).

We claim the first-order term of Equation 20 is $O_p(1/n)$. To establish this, we will show that $\|\nabla_\theta R_n(\theta^*)\| = O_p(1/\sqrt{n})$ and then multiply by $\|\hat{\theta} - \theta^*\| = O_p(1/\sqrt{n})$, using Condition S2.1(e). Fix $x \in \mathcal{X}$. By Condition S2.1(b), for any $\theta \in B_r(\theta^*)$ we can write

$$\nabla_\theta \ell(x; \theta) - \nabla_\theta \ell(x; \theta^*) = \int_0^1 \nabla_\theta^2 \ell\big(x; \theta^* + t(\theta - \theta^*)\big)(\theta - \theta^*)\, dt \qquad (21)$$

by applying the fundamental theorem of calculus to each entry of the function $f(t) = \nabla_\theta \ell(x; \theta^* + t(\theta - \theta^*))$ for $t \in [0, 1]$. Taking norms and using Condition S2.1(c),

$$\left\|\nabla_\theta \ell(x; \theta) - \nabla_\theta \ell(x; \theta^*)\right\| \leq \int_0^1 \left\|\nabla_\theta^2 \ell\big(x; \theta^* + t(\theta - \theta^*)\big)\right\| \|\theta - \theta^*\|\, dt \qquad (22)$$

$$\leq \sup_{\theta' \in B_r(\theta^*)} \|\nabla_\theta^2 \ell(x; \theta')\| \|\theta - \theta^*\| \qquad (23)$$

$$\leq H(x) \|\theta - \theta^*\|. \qquad (24)$$

In particular, for each fixed x, $\nabla_\theta \ell(x; \theta) \to \nabla_\theta \ell(x; \theta^*)$ as $\theta \to \theta^*$. Moreover, by the reverse triangle inequality, for all $\theta \in B_r(\theta^*)$,

$$\|\nabla_\theta \ell(x; \theta)\| \leq \|\nabla_\theta \ell(x; \theta^*)\| + H(x)\|\theta - \theta^*\| \qquad (25)$$

$$\leq \|\nabla_\theta \ell(x; \theta^*)\| + rH(x). \qquad (26)$$

For each coordinate j and each $\theta \in B_r(\theta^*)$,

$$\left|\frac{\partial}{\partial \theta_j}\ell(X;\theta)\right| \leq \left\|\nabla_\theta \ell(X;\theta)\right\| \leq \left\|\nabla_\theta \ell(X;\theta^*)\right\| + rH(X), \tag{27}$$

and the right-hand side is integrable by Condition S2.1(c,d). Therefore, we may interchange differentiation and expectation componentwise (Folland, 1999, Theorem 2.27(b)), which implies that $R(\theta) = \mathbb{E}\big(\ell(X;\theta)\big)$ is differentiable at θ^* and $\nabla_\theta R(\theta^*) = \mathbb{E}\big(\nabla_\theta \ell(X;\theta^*)\big)$. Since Θ is open and θ^* is a minimizer of $R(\theta)$, it follows that $\nabla_\theta R(\theta^*) = 0$, and hence, $\mathbb{E}\big(\nabla_\theta \ell(X;\theta^*)\big) = 0$. Hence, the central limit theorem implies

$$\sqrt{n}\nabla_\theta R_n(\theta^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta \ell(X_i;\theta^*) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\Big(0,\, \mathrm{Cov}\big(\nabla_\theta \ell(X;\theta^*)\big)\Big) \tag{28}$$

since the covariance matrix $\mathrm{Cov}(\nabla_\theta \ell(X;\theta^*))$ exists and is finite due to Condition S2.1(d). This shows that $\|\nabla_\theta R_n(\theta^*)\| = O_p(1/\sqrt{n})$. Thus, by Cauchy–Schwarz,

$$|\nabla_\theta R_n(\theta^*)^{\mathrm{T}}(\hat{\theta} - \theta^*)| = O_p(1/\sqrt{n}) \times O_p(1/\sqrt{n}) = O_p(1/n). \tag{29}$$

Next, consider the second-order term of Equation 20. By definition, we always have $\tilde{\theta} \in B_r(\theta^*)$ since $\hat{\theta} \in B_r(\theta^*)$ on A_n and $\tilde{\theta} = \theta^*$ on A_n^c. Hence,

$$\big\|\nabla_\theta^2 R_n(\tilde{\theta})\big\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 \ell(X_i;\tilde{\theta})\right\| \tag{30}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\theta \in B_r(\theta^*)}\big\|\nabla_\theta^2 \ell(X_i;\theta)\big\| \tag{31}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}H(X_i) \tag{32}$$

by Condition S2.1(c). Therefore,

$$\left|\frac{1}{2}(\hat{\theta} - \theta^*)^{\mathrm{T}}\nabla_\theta^2 R_n(\tilde{\theta})(\hat{\theta} - \theta^*)\right| \leq \frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \frac{1}{n}\sum_{i=1}^{n}H(X_i) = O_p(1/n) \tag{33}$$

since $\big\|\hat{\theta} - \theta^*\big\| = O_p(1/\sqrt{n})$ by Condition S2.1(e) and $\frac{1}{n}\sum_{i=1}^{n}H(X_i) \xrightarrow{\mathrm{a.s.}} \mathbb{E}\big(H(X)\big) < \infty$

by the strong law of large numbers. □

S2.2 Bias correction is asymptotically negligible

While the bias correction in Section 2.5 is useful in smaller sample settings, it is simpler to work with the uncorrected LaD values in theoretical analyses. To this end, our next result shows that, asymptotically, the bias-corrected negative log-likelihood $\overline{Z}_n^{\mathrm{bc}}(\hat{\theta})$ behaves like the uncorrected version $\overline{Z}_n(\hat{\theta})$; compare Equation 35 to Equation 19. This means that asymptotically, the validity of frequentist inference for μ^0 based on $\overline{Z}_n(\hat{\theta})$ (Section 2.4) is unaffected by the bias correction.

Corollary S2.4. *Assume Condition S2.1 holds and the loss is the negative log-likelihood,* $\ell_k(x;\theta_k) = -\log f_k(x;\theta_k)$, *with parameter dimension* $d_k = \dim(\Theta_k)$. *Define*

$$\overline{Z}_n^{\mathrm{bc}}(\hat{\theta}) := \overline{Z}_n(\hat{\theta}) + \frac{d}{2n} \tag{34}$$

where $d = (d_1, \ldots, d_K)^{\mathrm{T}} \in \mathbb{R}^K$ *and* $\overline{Z}_n(\theta)$ *is defined as in Theorem S2.3. Then*

$$\overline{Z}_n^{\mathrm{bc}}(\hat{\theta}) = \overline{Z}_n(\theta^*) + O_p(1/n). \tag{35}$$

In particular, if $\sqrt{n}\big(\overline{Z}_n(\hat{\theta}) - \mu^0\big)$ *converges in distribution, then* $\sqrt{n}\big(\overline{Z}_n^{\mathrm{bc}}(\hat{\theta}) - \mu^0\big)$ *converges in distribution to the same limit.*

Proof of Corollary S2.4. By Theorem S2.3,

$$\overline{Z}_n^{\mathrm{bc}}(\hat{\theta}) = \overline{Z}_n(\hat{\theta}) + \frac{d}{2n} = \overline{Z}_n(\theta^*) + O_p(1/n) + \frac{d}{2n} = \overline{Z}_n(\theta^*) + O_p(1/n),$$

which is Equation 35. Moreover,

$$\sqrt{n}\big(\overline{Z}_n^{\mathrm{bc}}(\hat{\theta}) - \mu^0\big) = \sqrt{n}\big(\overline{Z}_n(\hat{\theta}) - \mu^0\big) + \frac{d}{2\sqrt{n}},$$

and $d/(2\sqrt{n}) \to 0$, so the two sequences have the same weak limit by Slutsky's theorem.

□

S2.3 Asymptotics of the data-averaged LaD posterior

We now return to considering all K models, and focus on the Bayesian posterior in Section 3.1. Here, we set $Z_{ik} = \ell_k(X_i; \hat{\theta}_k)$, and as in Section 3.1, $Z_i = (Z_{i1}, \ldots, Z_{iK})^{\mathsf{T}}$, $\overline{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$, $S_n = \frac{1}{n} \sum_{i=1}^{n} (Z_i - \overline{Z}_n)(Z_i - \overline{Z}_n)^{\mathsf{T}}$, and $\mu_n = (\lambda_0 \mu_0 + n\overline{Z}_n)/\lambda_n$. We use the notation $Z_i(\theta^*)$, $\overline{Z}_n(\theta^*)$, $S_n(\theta^*)$, and $\mu_n(\theta^*)$, to denote the corresponding values with θ_k^* in place of $\hat{\theta}_k$ for all k. Also, recall the definitions of μ^0 and Σ^0 in Equation 12.

Theorem S2.5. *Assume Conditions 2.1 and S2.1 hold. Then*

$$\sqrt{n}(\mu_n - \mu^0) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, \Sigma^0).$$

Proof of Theorem S2.5. By Theorem S2.3 applied coordinate-wise to the vector $\overline{Z}_n$, we have $\overline{Z}_n = \overline{Z}_n(\theta^*) + O_p(1/n)$. It follows that

$$\sqrt{n}(\mu_n - \mu_n(\theta^*)) = \sqrt{n} \frac{n}{\lambda_n}(\overline{Z}_n - \overline{Z}_n(\theta^*)) = \sqrt{n} \, O_p(1/n) = o_p(1), \tag{36}$$

because $n/\lambda_n = n/(\lambda_0 + n) \to 1$ as $n \to \infty$. Thus,

$$\sqrt{n}(\mu_n - \mu^0) = \sqrt{n}(\mu_n(\theta^*) - \mu^0) + o_p(1). \tag{37}$$

Using $\lambda_n = \lambda_0 + n$, we have

$$\sqrt{n}(\mu_n(\theta^*) - \mu^0) = \frac{n}{\lambda_n} \sqrt{n}(\overline{Z}_n(\theta^*) - \mu^0) + \sqrt{n} \frac{\lambda_0}{\lambda_n}(\mu_0 - \mu^0). \tag{38}$$

Since $n/\lambda_n \to 1$, $\sqrt{n}(\overline{Z}_n(\theta^*) - \mu^0) \xrightarrow{d} \mathcal{N}(0, \Sigma^0)$ by the CLT, and the second term of Equation 38 is $O(1/\sqrt{n})$, Slutsky's theorem gives

$$\sqrt{n}(\mu_n(\theta^*) - \mu^0) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, \Sigma^0). \tag{39}$$

Hence, with Equation 37, another application of Slutsky's theorem yields

$$\sqrt{n}\big(\mu_n - \mu^0\big) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}(0, \Sigma^0).\tag{40}$$

This completes the proof. $\qquad\square$

Theorem S2.6. *Assume Conditions 2.1 and S2.1 hold. Then*

$$S_n \xrightarrow[n\to\infty]{\mathrm{p}} \Sigma^0.$$

Proof of Theorem S2.6. Under Condition 2.1, the strong law of large numbers implies that $S_n(\theta^*) \xrightarrow[n\to\infty]{\text{a.s.}} \Sigma^0$. Define $Y_i := Z_i - \overline{Z}_n$ and $Y_i(\theta^*) := Z_i(\theta^*) - \overline{Z}_n(\theta^*)$. Then

$$S_n - S_n(\theta^*) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_iY_i^{\mathrm{T}} - Y_i(\theta^*)Y_i(\theta^*)^{\mathrm{T}}\big)\tag{41}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big(\big(Y_i - Y_i(\theta^*)\big)Y_i^{\mathrm{T}} + Y_i(\theta^*)\big(Y_i - Y_i(\theta^*)\big)^{\mathrm{T}}\Big).\tag{42}$$

By Equation 41 and the fact that $\|uv^{\mathrm{T}}\| = \|u\|\|v\|$ for the Frobenius norm,

$$\|S_n - S_n(\theta^*)\| \le \frac{1}{n}\sum_{i=1}^{n}\|Y_i - Y_i(\theta^*)\|\big(\|Y_i\| + \|Y_i(\theta^*)\|\big).\tag{43}$$

By Cauchy–Schwarz,

$$\|S_n - S_n(\theta^*)\| \le \Big(\frac{1}{n}\sum_{i=1}^{n}\|Y_i - Y_i(\theta^*)\|^2\Big)^{1/2}\Big(\frac{1}{n}\sum_{i=1}^{n}\big(\|Y_i\| + \|Y_i(\theta^*)\|\big)^2\Big)^{1/2}.\tag{44}$$

We are working with all K models, so $\theta = (\theta_1, \ldots, \theta_K)$ denotes the concatenated parameter vector in $\mathbb{R}^D$ with $D = \sum_{k=1}^{K} d_k$. Also, $Z_i(\theta) \in \mathbb{R}^K$, so $\nabla_\theta Z_i(\theta)$ denotes the $K \times D$ Jacobian matrix of the map $\theta \mapsto Z_i(\theta)$, and $\|\nabla_\theta Z_i(\theta)\|$ denotes its Frobenius norm.

We next bound the two factors in Equation 44 on the event $A_n := \{\|\hat{\theta} - \theta^*\| < r\}$. First, consider the first factor of Equation 44. Since $Y_i - Y_i(\theta^*) = Z_i - Z_i(\theta^*) - \big(\overline{Z}_n - $

$\overline{Z}_n(\theta^*))$ and $\|u - v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$,

$$\frac{1}{n}\sum_{i=1}^{n}\|Y_i - Y_i(\theta^*)\|^2 \leq \frac{2}{n}\sum_{i=1}^{n}\|Z_i - Z_i(\theta^*)\|^2 + 2\|\overline{Z}_n - \overline{Z}_n(\theta^*)\|^2. \qquad (45)$$

By Theorem S2.3, $\|\overline{Z}_n - \overline{Z}_n(\theta^*)\| = O_p(1/n)$, hence the second term of Equation 45 is $O_p(1/n^2)$. For the first term, on A_n, apply Taylor expansion of $\theta \mapsto Z_i(\theta)$ to each coordinate of $Z_i(\theta)$ around θ^*: for each i we have

$$Z_i - Z_i(\theta^*) = \nabla_\theta Z_i(\theta^*)(\hat{\theta} - \theta^*) + R_i, \qquad \|R_i\| \leq \frac{1}{2}H(X_i)\|\hat{\theta} - \theta^*\|^2, \qquad (46)$$

where the remainder bound uses Condition S2.1(c) coordinate-wise. Therefore, since $\|u + v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$,

$$\|Z_i - Z_i(\theta^*)\|^2 \leq 2\|\nabla_\theta Z_i(\theta^*)\|^2\|\hat{\theta} - \theta^*\|^2 + \frac{1}{2}H(X_i)^2\|\hat{\theta} - \theta^*\|^4. \qquad (47)$$

Averaging over i yields

$$\frac{1}{n}\sum_{i=1}^{n}\|Z_i - Z_i(\theta^*)\|^2 \leq 2\|\hat{\theta} - \theta^*\|^2\frac{1}{n}\sum_{i=1}^{n}\|\nabla_\theta Z_i(\theta^*)\|^2 + \frac{1}{2}\|\hat{\theta} - \theta^*\|^4\frac{1}{n}\sum_{i=1}^{n}H(X_i)^2. \qquad (48)$$

By Condition S2.1(d), since $\|\nabla_\theta Z_i(\theta^*)\|^2 = \sum_{k=1}^{K}\|\nabla_{\theta_k}Z_{ik}(\theta_k^*)\|^2$ and K is fixed, we have $\frac{1}{n}\sum_{i=1}^{n}\|\nabla_\theta Z_i(\theta^*)\|^2 = O_p(1)$. By Condition S2.1(e), $\|\hat{\theta} - \theta^*\|^2 = \sum_{k=1}^{K}\|\hat{\theta}_k - \theta_k^*\|^2 = O_p(1/n)$; thus, the first term of Equation 48 is $O_p(1/n)$. Also, since $\mathbb{E}(H(X)) < \infty$, we have $\frac{1}{n}\sum_{i=1}^{n}H(X_i) = O_p(1)$ and $\max_{1\leq i\leq n}H(X_i) = o_p(n)$, since for any $\epsilon > 0$,

$$P\left(\max_{1\leq i\leq n}H(X_i) > \epsilon n\right) \leq nP\left(H(X) > \epsilon n\right) \leq \frac{\mathbb{E}\left(H(X)\mathbb{1}(H(X) > \epsilon n)\right)}{\epsilon} \xrightarrow[n\to\infty]{} 0. \qquad (49)$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}H(X_i)^2 \leq \left(\max_{1\leq i\leq n}H(X_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n}H(X_i)\right) = o_p(n). \qquad (50)$$

Combining with $\|\hat{\theta} - \theta^*\|^4 = O_p(1/n^2)$ shows the second term of Equation 48 is $o_p(1/n)$. Hence, combining Equations 45 and 48, there exists a random variable U_n such that (i)

$U_n = O_p(1/n)$ and (ii) $\frac{1}{n} \sum_{i=1}^{n} \|Y_i - Y_i(\theta^*)\|^2 \leq U_n$ on A_n.

For the second factor of Equation 44, note that

$$\frac{1}{n} \sum_{i=1}^{n} \left(\|Y_i\| + \|Y_i(\theta^*)\| \right)^2 \leq \frac{2}{n} \sum_{i=1}^{n} \|Y_i\|^2 + \frac{2}{n} \sum_{i=1}^{n} \|Y_i(\theta^*)\|^2. \tag{51}$$

The second term in Equation 51 is $O_p(1)$ since $S_n(\theta^*) \xrightarrow{\text{a.s.}} \Sigma^0$ implies $\frac{1}{n} \sum_{i=1}^{n} \|Y_i(\theta^*)\|^2 = \text{tr}\left(S_n(\theta^*)\right) = O_p(1)$. Meanwhile, for the first term in Equation 51,

$$\frac{1}{n} \sum_{i=1}^{n} \|Y_i\|^2 \leq \frac{2}{n} \sum_{i=1}^{n} \|Y_i(\theta^*)\|^2 + \frac{2}{n} \sum_{i=1}^{n} \|Y_i - Y_i(\theta^*)\|^2 \tag{52}$$

and we just showed that $\frac{1}{n} \sum_{i=1}^{n} \|Y_i - Y_i(\theta^*)\|^2 \leq U_n$ on A_n, where $U_n = O_p(1/n)$. Hence, there exists V_n such that (i) $V_n = O_p(1)$ and (ii) $\frac{1}{n} \sum_{i=1}^{n} \left(\|Y_i\| + \|Y_i(\theta^*)\| \right)^2 \leq V_n$ on A_n.

From the bounds above, the right-hand side of Equation 44 is upper bounded by $U_n^{1/2} V_n^{1/2}$ on A_n, and $U_n^{1/2} V_n^{1/2} = O_p(1/\sqrt{n}) O_p(1) = o_p(1)$. Since $P(A_n) \to 1$ by Condition S2.1(e), Lemma S2.2 implies $\|S_n - S_n(\theta^*)\| = o_p(1)$. Therefore, $\|S_n - S_n(\theta^*)\| \xrightarrow{\text{P}} 0$. Along with $S_n(\theta^*) \xrightarrow{\text{a.s.}} \Sigma^0$, this shows that $S_n \xrightarrow{\text{P}} \Sigma^0$ as $n \to \infty$.

$\square$

Theorems S2.7 and S2.8 pertain to the marginal distributions of μ' and Σ', integrating out the data; this is sometimes referred to as the data-averaged posterior. In other words, we consider the marginal distributions of μ' and Σ' when $Z_1, \ldots, Z_n$ are distributed according to the true data distribution and $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ according to the Bayesian model. See the beginning of Section S2.3 for a summary of the notation.

Theorem S2.7. *Assume Conditions 2.1 and S2.1 hold, and $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15. Then*

$$\Sigma' \xrightarrow[n \to \infty]{\text{P}} \Sigma^0.$$

Proof of Theorem S2.7. By Theorem S2.6, we have $S_n \xrightarrow{p} \Sigma^0$ as $n \to \infty$. Thus, since Ψ_0 is fixed, $\lambda_0/\lambda_n \to 0$, and $\overline{Z}_n = O_p(1)$ by Theorem S2.3 and the law of large numbers,

$$\frac{1}{n} \Psi_n = \frac{1}{n} \Psi_0 + S_n + \frac{\lambda_0}{\lambda_n} (\overline{Z}_n - \mu_0)(\overline{Z}_n - \mu_0)^{\text{T}} \xrightarrow[n \to \infty]{\text{P}} \Sigma^0, \tag{53}$$

recalling the definitions of Ψ_0 and Ψ_n from Equation 15. Since $\nu_n = \nu_0 + n$, for n large enough that $\nu_n > K+1$, standard results on the moment properties of the inverse-Wishart distribution (Press, 2005) give that

$$\mathbb{E}(\Sigma' \mid \Psi_n, \nu_n) = \frac{\Psi_n}{\nu_n - K - 1} = \frac{\Psi_n}{n} \frac{n}{\nu_0 + n - K - 1} \xrightarrow[n\to\infty]{\text{p}} \Sigma^0. \tag{54}$$

For n large enough that $\nu_n > K + 3$, the variance of the (i,j)-th entry is

$$\mathrm{Var}\big((\Sigma')_{ij} \mid \Psi_n, \nu_n\big) = \frac{(\nu_n - K + 1)(\Psi_n)_{ij}^2 + (\nu_n - K - 1)(\Psi_n)_{ii}(\Psi_n)_{jj}}{(\nu_n - K)(\nu_n - K - 1)^2(\nu_n - K - 3)}. \tag{55}$$

Recall that $\Psi_n/n \xrightarrow{\text{p}} \Sigma^0$ by Equation 53. Fix any $M > \max_{a,b}\big|(\Sigma^0)_{ab}\big|$. Then,

$$P\left(\max_{a,b}\big|(\Psi_n)_{ab}\big| \leq Mn\right) \xrightarrow[n\to\infty]{} 1. \tag{56}$$

On the event $\max_{a,b}\big|(\Psi_n)_{ab}\big| \leq Mn$, the numerator of Equation 55 is $O(n^3)$ and denominator is $\Omega(n^4)$, so there exists a constant $C_{ij}(M) > 0$ such that, on this event,

$$\mathrm{Var}\big((\Sigma')_{ij} \mid \Psi_n, \nu_n\big) \leq \frac{C_{ij}(M)}{n}. \tag{57}$$

Then, for all $\epsilon > 0$, by the law of total expectation and Chebyshev's inequality,

$$P\Big(\big|(\Sigma')_{ij} - \mathbb{E}\big((\Sigma')_{ij} \mid \Psi_n, \nu_n\big)\big| > \epsilon\Big) \tag{58}$$

$$= \mathbb{E}\Big(P\Big(\big|(\Sigma')_{ij} - \mathbb{E}\big((\Sigma')_{ij} \mid \Psi_n, \nu_n\big)\big| > \epsilon \ \Big|\ \Psi_n, \nu_n\Big)\mathbb{1}\Big(\max_{a,b}\big|(\Psi_n)_{ab}\big| \leq Mn\Big)\Big)+ \tag{59}$$

$$\mathbb{E}\Big(P\Big(\big|(\Sigma')_{ij} - \mathbb{E}\big((\Sigma')_{ij} \mid \Psi_n, \nu_n\big)\big| > \epsilon \ \Big|\ \Psi_n, \nu_n\Big)\mathbb{1}\Big(\max_{a,b}\big|(\Psi_n)_{ab}\big| > Mn\Big)\Big)$$

$$\leq \frac{C_{ij}(M)}{n\epsilon^2} + P\Big(\max_{a,b}\big|(\Psi_n)_{ab}\big| > Mn\Big) \xrightarrow[n\to\infty]{} 0. \tag{60}$$

Now, we control the whole matrix using the max norm and a union bound:

$$P\Big(\big\|\Sigma' - \mathbb{E}(\Sigma' \mid \Psi_n, \nu_n)\big\|_{\max} > \epsilon\Big) \leq \sum_{i,j=1}^{K} P\Big(\big|(\Sigma')_{ij} - \mathbb{E}\big((\Sigma')_{ij} \mid \Psi_n, \nu_n\big)\big| > \epsilon\Big) \xrightarrow[n\to\infty]{} 0. \tag{61}$$

Thus,

$$\left\|\Sigma' - \mathbb{E}\big(\Sigma' \mid \Psi_n, \nu_n\big)\right\|_{\max} \xrightarrow[n\to\infty]{\text{p}} 0. \tag{62}$$

Finally, by the triangle inequality,

$$\left\|\Sigma' - \Sigma^0\right\|_{\max} \leq \left\|\Sigma' - \mathbb{E}\big(\Sigma' \mid \Psi_n, \nu_n\big)\right\|_{\max} + \left\|\mathbb{E}\big(\Sigma' \mid \Psi_n, \nu_n\big) - \Sigma^0\right\|_{\max} \xrightarrow[n\to\infty]{\text{p}} 0, \tag{63}$$

since the second term is $o_p(1)$ by Equation 54. Therefore, $\Sigma' \xrightarrow{\text{p}} \Sigma^0$, proving the result. $\square$

Theorem S2.8 establishes the limiting behavior of μ' under the data-averaged LaD posterior, that is, integrating out the data. The factor of 2 in the covariance $2\Sigma^0$ reflects the combined contribution of (i) the sampling variability of the data and (ii) the posterior uncertainty given the data.

Theorem S2.8. *Assume Conditions 2.1 and S2.1 hold, and $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15. Then*

$$\sqrt{n}(\mu' - \mu^0) \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}(0, 2\Sigma^0).$$

Proof of Theorem S2.8. Write

$$\sqrt{n}(\mu' - \mu^0) = \sqrt{n}(\mu' - \mu_n) + \sqrt{n}(\mu_n - \mu^0). \tag{64}$$

By Theorem S2.7, we have $\Sigma' \xrightarrow{\text{p}} \Sigma^0$. Since the matrix square root and inverse are continuous on the space of positive definite matrices, the continuous mapping theorem yields $(\Sigma')^{-1/2} \xrightarrow{\text{p}} (\Sigma^0)^{-1/2}$. Define the standardized quantities:

$$V_n := \sqrt{n}(\Sigma')^{-1/2}\big(\mu' - \mu_n\big), \qquad W_n := \sqrt{n}(\Sigma')^{-1/2}\big(\mu_n - \mu^0\big).$$

We characterize the limiting distributions of V_n and W_n. For V_n, since $\mu' \mid \Sigma', Z_{1:n} \sim \mathcal{N}\big(\mu_n, \Sigma'/\lambda_n\big)$, we have

$$V_n \mid \Sigma', Z_{1:n} \sim \mathcal{N}(0, nI/\lambda_n), \tag{65}$$

Thus the conditional distribution V_n does not depend on Σ' or $Z_{1:n}$, and hence $V_n \sim$

S13

$\mathcal{N}(0, nI/\lambda_n)$ marginally for all n. Since $n/\lambda_n \to 1$, it follows from Slutsky's theorem that $V_n \xrightarrow{d} \mathcal{N}(0, I)$. For W_n, from Theorem S2.5 and $(\Sigma')^{-1/2} \xrightarrow{P} (\Sigma^0)^{-1/2}$, Slutsky's theorem yields $W_n \xrightarrow{d} \mathcal{N}(0, I)$. Moreover, since we can represent $\mu' = \mu_n + \Sigma'^{1/2}\xi/\sqrt{\lambda_n}$ with $\xi \sim \mathcal{N}(0, I)$ independent of $(Z_{1:n}, \Sigma')$, it follows that $V_n = \xi\sqrt{n/\lambda_n}$ depends only on ξ, whereas W_n depends only on $(Z_{1:n}, \Sigma')$. Thus, V_n and W_n are independent. Therefore, $V_n + W_n \xrightarrow{d} \mathcal{N}(0, 2I)$. Finally, Slutsky's theorem yields

$$\sqrt{n}(\mu' - \mu^0) = (\Sigma')^{1/2}(V_n + W_n) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 2\Sigma^0), \tag{66}$$

which proves the result. $\qquad\square$

S2.4 Posterior consistency of hard min for the minimal KL

Recall from Equation 12 that $\mu^0 = (\mu_1^0, \ldots, \mu_K^0)^{\mathsf{T}}$ where $\mu_k^0 = \mathbb{E}(\ell_k(X; \theta_k^*))$.

Corollary S2.9. *Assume Conditions 2.1 and S2.1 hold, and $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15. Then*

$$\mu' \xrightarrow[n\to\infty]{P} \mu^0.$$

Proof of Corollary S2.9. By Lemma S2.8, we have $\sqrt{n}(\mu' - \mu^0) \xrightarrow{d} \mathcal{N}(0, 2\Sigma^0)$ as $n \to \infty$. Consequently, $\mu' - \mu^0 = O_p(1/\sqrt{n})$, so $\mu' \xrightarrow{P} \mu^0$ in $\mathbb{R}^K$. $\qquad\square$

Recall from Equation 5 that we define $M_\delta(\mu) = \{k : \mu_k \le \min_j \mu_j + \delta\}$ for $\delta \ge 0$.

Theorem S2.10. *Assume Conditions 2.1 and S2.1 hold, and $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15. Then*

$$P\Big(\arg\min_{1\le k\le K} \mu_k' \in M_0(\mu^0)\Big) \xrightarrow[n\to\infty]{} 1.$$

Proof of Theorem S2.10. If $M_0(\mu^0) = \{1, \ldots, K\}$, then the result is immediate. Suppose $M_0(\mu^0) \ne \{1, \ldots, K\}$. Define $m := \min\{\mu_k^0 - \mu_{\min}^0 : k \notin M_0(\mu^0)\}$ where $\mu_{\min}^0 = \min_k \mu_k^0$, and note that $m > 0$. Choose any $\eta \in (0, m/2)$. By Corollary S2.9, $P(E_n(\eta)) \to 1$ where $E_n(\eta) = \{\max_{1\le k\le K} |\mu_k' - \mu_k^0| < \eta\}$. On the event $E_n(\eta)$, for all $k \in M_0(\mu^0)$, we have

$$\mu_k' < \mu_k^0 + \eta = \mu_{\min}^0 + \eta. \tag{67}$$

Meanwhile, for all $k \notin M_0(\mu^0)$, we have $\mu_k^0 - \mu_{\min}^0 \geq m > 0$, and so since $\eta < m/2$,

$$\mu_k' > \mu_k^0 - \eta \geq \mu_{\min}^0 + m - \eta > \mu_{\min}^0 + \eta. \tag{68}$$

Consequently, on event $E_n(\eta)$, we have $\max_{k \in M_0(\mu^0)} \mu_k' < \min_{k \notin M_0(\mu^0)} \mu_k'$, and hence, $\arg \min_{1 \leq k \leq K} \mu_k' \in M_0(\mu^0)$. Therefore,

$$P\Big(\arg \min_{1 \leq k \leq K} \mu_k' \in M_0(\mu^0) \Big) \geq P\big(E_n(\eta) \big) \longrightarrow 1. \tag{69}$$

$\square$

S2.5   Posterior consistency for the complexity class

Recall from Equation 6 that $c_\delta^*(\mu) = \min \big\{ c(k) : k \in M_\delta(\mu) \big\}$.

Lemma S2.11. *The function $\mu \mapsto c_\delta^*(\mu)$ is continuous at any point $\mu \in \mathbb{R}^K$ such that $\delta \neq \mu_k - \min_j \mu_j$ for all k.*

Proof of Lemma S2.11. Since $c_\delta^*(\mu)$ is integer valued, the claim is that $\mu \mapsto c_\delta^*(\mu)$ is continuous at any point $\mu \in \mathbb{R}^K$ such that $\delta \neq \mu_k - \min_j \mu_j$ for all k. To show this, define $g : \mathbb{R}^K \to \mathbb{R}^K$ by $g(\mu) = \mu - \min_k \mu_k$ and define $f : \mathbb{R}^K \to 2^{\{1,\dots,K\}}$ by $f(x) = \{k : x_k \leq \delta\}$. We give $\mathbb{R}^K$ the usual Euclidean topology and give $2^{\{1,\dots,K\}}$ the discrete topology. Observe that $f(g(\mu)) = M_\delta(\mu)$, so if $f \circ g$ is continuous at some point μ, then $c_\delta^*(\mu)$ is also continuous at μ, because it depends on μ only through the discrete quantity $M_\delta(\mu)$.

Note that g is continuous everywhere, since $\mu \mapsto \min_k \mu_k$ is continuous. Next, note that f is continuous at any point x such that $x_k \neq \delta$ for all k. Now, let $\mu \in \mathbb{R}^K$ such that $\delta \neq \mu_k - \min_j \mu_j$ for all k. Then, letting $x = g(\mu)$, we have $x_k \neq \delta$ for all k, so f is continuous at x. Therefore, $f \circ g$ is continuous at μ. It follows that $c_\delta^*(\mu)$ is continuous at μ, proving the claim. $\square$

Theorem S2.12 shows that the data-averaged posterior is consistent for the minimal δ-optimal complexity class. More precisely, with probability tending to 1, a sample of $c_\delta^*(\mu')$ from the data-averaged posterior will be equal the true value $c_\delta^*(\mu^0)$.

Theorem S2.12. *Assume Conditions 2.1 and S2.1, and let $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15. Suppose $\delta \neq \mu_k^0 - \min_j \mu_j^0$ for all $k = 1, \ldots, K$. Then*

$$P\big(c_\delta^*(\mu') = c_\delta^*(\mu^0)\big) \xrightarrow[n \to \infty]{} 1.$$

Proof of Theorem S2.12. By Lemma S2.11, $\mu \mapsto c_\delta^*(\mu)$ is continuous at μ^0. By Corollary S2.9, $\mu' \xrightarrow{\text{P}} \mu^0$. Therefore, by the continuous mapping theorem, $c_\delta^*(\mu') \xrightarrow{\text{P}} c_\delta^*(\mu^0)$. Since $c_\delta^*(\mu)$ is integer-valued, this implies $P\big(c_\delta^*(\mu') = c_\delta^*(\mu^0)\big) \longrightarrow 1$. $\qquad\square$

Theorem S2.13 shows that given $Z_{1:n}$, the posterior is consistent for the minimal δ-optimal complexity class. This establishes consistency of the posterior, whereas Theorem S2.12 shows consistency of the data-averaged posterior.

Theorem S2.13. *Under the same assumptions as Theorem S2.12,*

$$P\big(c_\delta^*(\mu') = c_\delta^*(\mu^0) \,\big|\, Z_{1:n}\big) \xrightarrow[n \to \infty]{\text{P}} 1.$$

Proof of Theorem S2.13. Define the event $A_n = \{c_\delta^*(\mu') \neq c_\delta^*(\mu^0)\}$. By Theorem S2.12, $P(A_n) \to 0$ as $n \to \infty$. Consider the random variable $P(A_n \mid Z_{1:n})$, which is the posterior probability of A_n given the data. By Markov's inequality, for all $\epsilon > 0$,

$$P\big(P(A_n \mid Z_{1:n}) > \epsilon\big) \leq \frac{\mathbb{E}\big(P(A_n \mid Z_{1:n})\big)}{\epsilon} = \frac{P(A_n)}{\epsilon} \longrightarrow 0. \tag{70}$$

Therefore, $P(A_n \mid Z_{1:n}) \xrightarrow{\text{P}} 0$, and hence,

$$P\big(c_\delta^*(\mu') = c_\delta^*(\mu^0) \,\big|\, Z_{1:n}\big) = 1 - P(A_n \mid Z_{1:n}) \xrightarrow[n \to \infty]{\text{P}} 1.$$

$\qquad\square$

S2.6   Asymptotics of the soft selection score

Lemma S2.15 establishes the asymptotics of the soft selection score in Equation 8. First, we show a simple lemma that is used in the proof.

Lemma S2.14. *If $X_n = O_p(1)$ and $Y_n \overset{P}{\rightarrow} \infty$ then $X_n + Y_n \overset{P}{\rightarrow} \infty$ and $\min\{X_n, Y_n\} = O_p(1)$.*

Proof. Let $M \in \mathbb{R}$ and $\epsilon > 0$. Choose $B > 0$ such that $P(|X_n| > B) < \epsilon$ for all n sufficiently large. Since $P(X_n + Y_n < M) \leq P(|X_n| > B) + P(Y_n < M + B)$, we have

$$\limsup_n P(X_n + Y_n < M) \leq \limsup_n P(|X_n| > B) + \limsup_n P(Y_n < M + B) \leq \epsilon$$

because $Y_n \overset{P}{\rightarrow} \infty$. Since $M \in \mathbb{R}$ and $\epsilon > 0$ are arbitrary, this shows that $X_n + Y_n \overset{P}{\rightarrow} \infty$. Furthermore, since $P(|\min\{X_n, Y_n\}| > B) \leq P(|X_n| > B) + P(Y_n < -B)$,

$$\limsup_n P(|\min\{X_n, Y_n\}| > B) \leq \limsup_n P(|X_n| > B) + \limsup_n P(Y_n < -B) \leq \epsilon.$$

This shows that $\min\{X_n, Y_n\} = O_p(1)$. $\qquad\square$

Lemma S2.15. *Suppose $Y_1, Y_2, \ldots \in \mathbb{R}^K$ are random vectors such that $\sqrt{n}(Y_n - \mu) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma)$ for some $\mu \in \mathbb{R}^K$ and a positive definite matrix $\Sigma \in \mathbb{R}^{K \times K}$. Suppose $\alpha_n \to \infty$ such that $\alpha_n/\sqrt{n} \to 0$ as $n \to \infty$. Then*

$$\exp\left(-\alpha_n(Y_{nk} - Y_n^*)\right) \xrightarrow[n\to\infty]{P} \mathbb{1}(\mu_k = \mu^*) \tag{71}$$

for all $k = 1, \ldots, K$, where $Y_n^* = \min_k Y_{nk}$ and $\mu^* = \min_k \mu_k$.

Proof of Lemma S2.15. For each k, decompose

$$\alpha_n(Y_{nk} - Y_n^*) = \alpha_n(Y_{nk} - \mu_k) - \alpha_n(Y_n^* - \mu^*) + \alpha_n(\mu_k - \mu^*). \tag{72}$$

Let $S = \{k : \mu_k = \mu^*\}$. For the third term in Equation 72, since $\alpha_n \to \infty$, we have

$$\alpha_n(\mu_k - \mu^*) \longrightarrow \begin{cases} 0 & \text{if } k \in S \\ \infty & \text{if } k \notin S \end{cases} \tag{73}$$

S17

as $n \to \infty$. For the first term in Equation 72, we have

$$\alpha_n(Y_{nk} - \mu_k) = \frac{\alpha_n}{\sqrt{n}}\sqrt{n}(Y_{nk} - \mu_k) \xrightarrow[n \to \infty]{\text{p}} 0 \tag{74}$$

since $\alpha_n/\sqrt{n} \to 0$ and $\sqrt{n}(Y_{nk} - \mu_k) = O_p(1)$, because $\sqrt{n}(Y_n - \mu)$ converges in distribution by assumption. To study the second term, we write

$$\sqrt{n}(Y_n^* - \mu^*) = \min\left\{ \min_{k \in S} \sqrt{n}(Y_{nk} - \mu^*), \min_{k \in S^c} \sqrt{n}(Y_{nk} - \mu^*) \right\}. \tag{75}$$

Letting $W \sim \mathcal{N}(0, \Sigma)$, the first argument of Equation 75 is

$$\min_{k \in S} \sqrt{n}(Y_{nk} - \mu^*) \overset{\text{(a)}}{=} \min_{k \in S} \sqrt{n}(Y_{nk} - \mu_k) \xrightarrow[n \to \infty]{\text{d}} \min_{k \in S} W_k = O_p(1) \tag{76}$$

where (a) holds since $\mu_k = \mu^*$ for $k \in S$, and the limit holds by the continuous mapping theorem since $x \mapsto \min_{k \in S} x_k$ is a continuous function and $\sqrt{n}(Y_n - \mu) \overset{\text{d}}{\to} W$ by assumption. Meanwhile, the second argument of Equation 75 is

$$\min_{k \in S^c} \sqrt{n}(Y_{nk} - \mu^*) = \min_{k \in S^c} \left(\sqrt{n}(Y_{nk} - \mu_k) + \sqrt{n}(\mu_k - \mu^*) \right)$$
$$\geq \min_{k \in S^c} \sqrt{n}(Y_{nk} - \mu_k) + \min_{k \in S^c} \sqrt{n}(\mu_k - \mu^*).$$

As in Equation 76, $\min_{k \in S^c} \sqrt{n}(Y_{nk} - \mu_k) \overset{\text{d}}{\to} \min_{k \in S^c} W_k$. Meanwhile, the second term satisfies $\min_{k \in S^c} \sqrt{n}(\mu_k - \mu^*) \to \infty$ since $\min_{k \in S^c} \mu_k - \mu^* > 0$ by the definition of S and μ^*. It follows that

$$\min_{k \in S^c} \sqrt{n}(Y_{nk} - \mu^*) \xrightarrow[n \to \infty]{\text{p}} \infty, \tag{77}$$

by Lemma S2.14. Combining Equations 76 and 77 with Equation 75, we have that $\sqrt{n}(Y_n^* - \mu^*) = O_p(1)$ by Lemma S2.14. Therefore,

$$\alpha_n(Y_n^* - \mu^*) = \frac{\alpha_n}{\sqrt{n}}\sqrt{n}(Y_n^* - \mu^*) \xrightarrow[n \to \infty]{\text{p}} 0, \tag{78}$$

since $\alpha_n/\sqrt{n} \to 0$. Plugging Equations 73, 74, and 78 into Equation 72, we have

$$\alpha_n(Y_{nk} - Y_n^*) \xrightarrow[n\to\infty]{\mathrm{p}} \begin{cases} 0 & \text{if } k \in S \\[2mm] \infty & \text{if } k \notin S. \end{cases} \tag{79}$$

For $k \in S$, we obtain Equation 71 by the continuous mapping theorem since $x \mapsto \exp(-x)$ is continuous. For $k \in S^c$, Equation 79 implies Equation 71 since for all $\epsilon > 0$,

$$P\Big(\exp\big(-\alpha_n(Y_{nk} - Y_n^*) \big) > \epsilon \Big) = P\Big(\alpha_n(Y_{nk} - Y_n^*) < -\log \epsilon \Big) \xrightarrow[n\to\infty]{} 0.$$

This completes the proof. $\qquad\square$

S2.7 Posterior consistency of the SLC score

Before proving our main consistency result, Theorem 4.1, we establish the following lemma. Lemma S2.16 shows that, under the data-averaged posterior, the within-class soft-selection score (Equation 8) concentrates on models with the smallest KL among the set of minimal-complexity δ-optimal models. Recall the definitions of $c_\delta^*(\mu)$, $r_{nk}(\mu)$, and μ^0 from Equations 6, 8, and 12, respectively. Also recall that $\mu_{\min,c(k)} = \min_{j\,:\,c(j)=c(k)} \mu_j$.

Lemma S2.16. *Assume Conditions 2.1 and S2.1, and $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15. Suppose $\alpha_n \to \infty$ with $\alpha_n = o(\sqrt{n})$. Then, for all k such that $c(k) = c_\delta^*(\mu^0)$,*

$$r_{nk}(\mu') = \exp\big(-\alpha_n(\mu'_k - \mu'_{\min,c(k)}) \big) \xrightarrow[n\to\infty]{\mathrm{p}} \mathbb{1}\big(\mu_k^0 = \mu_{\min,c(k)}^0 \big). \tag{80}$$

This result pertains to the data-averaged posterior, that is, the marginal distribution of μ' integrating out data $Z_{1:n}$ from the true distribution.

Proof of Lemma S2.16. Let $C := \{k \,:\, c(k) = c_\delta^*(\mu^0)\}$. By Theorem S2.8, we have $\sqrt{n}(\mu' - \mu^0) \xrightarrow{d} \mathcal{N}(0, 2\Sigma^0)$ as $n \to \infty$. By the continuous mapping theorem, this convergence holds when restricted to the coordinates in C, that is, $\sqrt{n}(\mu'_C - \mu_C^0) \xrightarrow{d} \mathcal{N}(0, 2\Sigma_{C,C}^0)$ where $\mu'_C = (\mu'_j : j \in C)$, $\mu_C^0 = (\mu_j^0 : j \in C)$, and $\Sigma_{C,C}^0 = (\Sigma_{jk}^0 : j, k \in C)$. We apply Lemma S2.15 with μ'_C, μ_C^0, and $2\Sigma_{C,C}^0$ playing the roles of Y_n, μ, and Σ, respectively. This

shows that for all $k \in C$,

$$\exp\left(-\alpha_n(\mu_k' - \min_{j \in C} \mu_j')\right) \xrightarrow[n \to \infty]{\text{P}} \mathbb{1}\left(\mu_k^0 = \min_{j \in C} \mu_j^0\right).$$

The result in Equation 80 follows since $k \in C$ implies $c(k) = c_\delta^*(\mu^0)$, and therefore, $\mu_{\min,c(k)}' = \min_{j \in C} \mu_j'$ and $\mu_{\min,c(k)}^0 = \min_{j \in C} \mu_j^0$. $\qquad\square$

Lemma S2.17. *If X_n and Y_n are random variables such that X_n is bounded and $X_n \xrightarrow{\text{P}} c$ for some constant $c \in \mathbb{R}$, then $\mathbb{E}(X_n \mid Y_n) \xrightarrow{\text{P}} c$.*

Proof. These assumptions imply $X_n \to c$ in L^1, that is, $\mathbb{E}|X_n - c| \to 0$ (Jacod and Protter, 2004, Theorem 17.4). Thus,

$$\left|\mathbb{E}(X_n \mid Y_n) - c\right| = \left|\mathbb{E}(X_n - c \mid Y_n)\right| \le \mathbb{E}\left(|X_n - c| \mid Y_n\right)$$

almost surely. Taking expectations, this yields

$$\mathbb{E}\left(\left|\mathbb{E}(X_n \mid Y_n) - c\right|\right) \le \mathbb{E}|X_n - c| \to 0$$

by the law of total expectation. Hence, $\mathbb{E}(X_n \mid Y_n) \to c$ in L^1, and thus $\mathbb{E}(X_n \mid Y_n) \xrightarrow{\text{P}} c$ (Jacod and Protter, 2004, Theorem 17.2). $\qquad\square$

Proof of Theorem 4.1. Recall from Equation 9 that we define the SLC score as

$$w_\delta(k \mid Z_{1:n}) = P\left(c_\delta^*(\mu') = c(k) \mid Z_{1:n}\right) \mathbb{E}\left(r_{nk}(\mu') \mid Z_{1:n}\right) \tag{81}$$

where $r_{nk}(\mu) = \exp(-\alpha_n(\mu_k - \mu_{\min,c(k)}))$ and $\mu_{\min,c(k)} = \min_{j:c(j)=c(k)} \mu_j$ for $\mu \in \mathbb{R}^K$. By assumption, $\alpha_n \to \infty$ with $\alpha_n = o(\sqrt{n})$, and $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15. Here, we write μ' and Σ' to emphasize that these are random variables distributed according to the posterior, rather than arbitrary values of μ and Σ.

First, consider the first factor in Equation 81. By Theorem S2.13, for all k,

$$P\left(c_\delta^*(\mu') = c(k) \mid Z_{1:n}\right) \xrightarrow[n \to \infty]{\text{P}} \mathbb{1}\left(c_\delta^*(\mu^0) = c(k)\right). \tag{82}$$

S20

Now, consider the second factor in Equation 81. For all k such that $c(k) = c_\delta^*(\mu^0)$, we have $r_{nk}(\mu') \overset{\text{P}}{\to} \mathbb{1}\left(\mu_k^0 = \mu_{\min,c(k)}^0\right)$ by Lemma S2.16, and therefore,

$$\mathbb{E}\left(r_{nk}(\mu') \mid Z_{1:n}\right) \xrightarrow[n\to\infty]{\text{P}} \mathbb{1}\left(\mu_k^0 = \mu_{\min,c(k)}^0\right) \tag{83}$$

by Lemma S2.17, since $r_{nk}(\mu') \in [0, 1]$ is bounded.

We combine the two factors as follows. If $c(k) \neq c_\delta^*(\mu^0)$, then by Equations 81 and 82,

$$0 \leq w_\delta(k \mid Z_{1:n}) \leq P\left(c_\delta^*(\mu') = c(k) \mid Z_{1:n}\right) \xrightarrow[n\to\infty]{\text{P}} 0, \tag{84}$$

and hence, $w_\delta(k \mid Z_{1:n}) \overset{\text{P}}{\to} 0$. Meanwhile, if $c(k) = c_\delta^*(\mu^0)$, then applying Slutsky's theorem to Equation 81 along with Equations 82 and 83, we have

$$w_\delta(k \mid Z_{1:n}) \xrightarrow[n\to\infty]{\text{P}} \mathbb{1}\left(\mu_k^0 = \mu_{\min,c(k)}^0\right). \tag{85}$$

Hence, for all k,

$$w_\delta(k \mid Z_{1:n}) \xrightarrow[n\to\infty]{\text{P}} \mathbb{1}\left(c_\delta^*(\mu^0) = c(k)\right) \mathbb{1}\left(\mu_k^0 = \mu_{\min,c(k)}^0\right), \tag{86}$$

by combining Equations 84 and 85. Recalling the definition of $M_\delta^*(\mu^0)$, observe that

$$M_\delta^*(\mu^0) = \underset{k\,:\,c(k)=c_\delta^*(\mu^0)}{\arg\min} \mu_k^0 = \left\{k \,:\, c_\delta^*(\mu^0) = c(k),\ \mu_k^0 = \mu_{\min,c(k)}^0\right\}. \tag{87}$$

Therefore, by Equations 86 and 87, $w_\delta(k \mid Z_{1:n}) \overset{\text{P}}{\to} \mathbb{1}(k \in M_\delta^*(\mu^0))$ as $n \to \infty$.

$\square$

S2.8 Posterior consistency of the approximate SLC score

Proof of Theorem 4.2. Recall Algorithm 2 defines $\hat{w}_\delta(k)$ as the product of the between-class factor $\hat{p}_\delta(k)$ and the within-class factor $\hat{r}(k)$, where

$$\hat{p}_\delta(k) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\big(c_\delta^*(\mu^{(t)}) = c(k)\big) \quad \text{and} \quad \hat{r}(k) := \frac{1}{T} \sum_{t=1}^{T} r_{nk}(\mu^{(t)}).$$

First, consider the between-class factor $\hat{p}_\delta(k)$. Since $\mu' \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15, by Theorem S2.12,

$$P\big(c_\delta^*(\mu') = c(k)\big) \xrightarrow[n \to \infty]{} \mathbb{1}\big(c_\delta^*(\mu^0) = c(k)\big) \tag{88}$$

for all k. Each $\mu^{(t)}$ has the same distribution as μ', so for all $t = 1, \ldots, T$,

$$\mathbb{1}\big(c_\delta^*(\mu^{(t)}) = c(k)\big) \xrightarrow[n \to \infty]{\mathrm{P}} \mathbb{1}\big(c_\delta^*(\mu^0) = c(k)\big). \tag{89}$$

Since T is fixed, averaging preserves the limit in probability, hence

$$\hat{p}_\delta(k) \xrightarrow[n \to \infty]{\mathrm{P}} \mathbb{1}\big(c_\delta^*(\mu^0) = c(k)\big). \tag{90}$$

Next, consider the within-class factor $\hat{r}(k)$. Fix any k such that $c(k) = c_\delta^*(\mu^0)$. For a single posterior draw μ', we have $r_{nk}(\mu') \xrightarrow{\mathrm{P}} \mathbb{1}\big(\mu_k^0 = \mu_{\min,c(k)}^0\big)$ by Lemma S2.16. Again, each $\mu^{(t)}$ has the same distribution as μ', so for all $t = 1, \ldots, T$,

$$r_{nk}(\mu^{(t)}) \xrightarrow[n \to \infty]{\mathrm{P}} \mathbb{1}\big(\mu_k^0 = \mu_{\min,c(k)}^0\big). \tag{91}$$

Since T is fixed, averaging preserves the limit in probability, so

$$\hat{r}(k) \xrightarrow[n \to \infty]{\mathrm{P}} \mathbb{1}\big(\mu_k^0 = \mu_{\min,c(k)}^0\big). \tag{92}$$

By Slutsky's theorem, combining Equations 90 and 92 yields

$$\hat{w}_\delta(k) = \hat{p}_\delta(k)\,\hat{r}(k) \xrightarrow[n\to\infty]{\mathrm{p}} \mathbb{1}\left(c_\delta^*(\mu^0) = c(k)\right) \mathbb{1}\left(\mu_k^0 = \mu_{\min,c(k)}^0\right) = \mathbb{1}\left(k \in M_\delta^*(\mu^0)\right)$$

where the last equality holds by the same argument as in the proof of Theorem 4.1.

□

S2.9   Consistency of the estimated target set of models

Proof of Theorem 4.3. By Theorem 4.1, for all k,

$$w_\delta(k \mid Z_{1:n}) \xrightarrow[n\to\infty]{\mathrm{p}} \mathbb{1}\left(k \in M_\delta^*(\mu^0)\right). \tag{93}$$

Fix $\omega \in (0,1)$. The function $x \mapsto \mathbb{1}(x > \omega)$ is continuous at 0 and 1, so by the continuous mapping theorem,

$$\mathbb{1}\left(w_\delta(k \mid Z_{1:n}) > \omega\right) \xrightarrow[n\to\infty]{\mathrm{p}} \mathbb{1}\left(k \in M_\delta^*(\mu^0)\right). \tag{94}$$

for all k. Equivalently, $P(A_{nk}) \to 0$ as $n \to \infty$ where

$$A_{nk} := \left\{ \mathbb{1}\left(w_\delta(k \mid Z_{1:n}) > \omega\right) \neq \mathbb{1}\left(k \in M_\delta^*(\mu^0)\right) \right\}. \tag{95}$$

Since $\hat{M}_\delta^* = \{k : w_\delta(k \mid Z_{1:n}) > \omega\}$,

$$P\left(\hat{M}_\delta^* \neq M_\delta^*(\mu^0)\right) = P\left(\bigcup_{k=1}^K A_{nk}\right) \leq \sum_{k=1}^K P(A_{nk}) \xrightarrow[n\to\infty]{} 0 \tag{96}$$

since $P(A_{nk}) \to 0$ and K is finite. Therefore, $P\left(\hat{M}_\delta^* = M_\delta^*(\mu^0)\right) \to 1$ as $n \to \infty$. □

S2.10   Instability of alternative approaches at ties

Proof of Theorem 4.4. Part (a). For brevity, let us denote $M := M_\delta^*(\mu^0)$. Define $C = \{k : c(k) = c_\delta^*(\mu^0)\}$, and note that $M \subseteq C$. Let $(\mu', \Sigma') \sim p(\mu, \Sigma \mid Z_{1:n})$ as in Equation 15; we use μ' and Σ' to emphasize that these are random variables distributed according to

the posterior. By the definition of $M_\delta^*(\mu)$, for all k,

$$\mathbb{1}\left(k \in M_\delta^*(\mu')\right) = \mathbb{1}\left(c_\delta^*(\mu') = c(k), \ \mu_k' = \mu_{\min,c(k)}'\right). \tag{97}$$

Taking posterior expectations given $Z_{1:n}$ and factoring,

$$P(k \in M_\delta^*(\mu') \mid Z_{1:n}) = P(c_\delta^*(\mu') = c(k) \mid Z_{1:n}) \, P(\mu_k' = \mu_{\min,c(k)}' \mid Z_{1:n}, \ c_\delta^*(\mu') = c(k)). \tag{98}$$

We analyze four cases: (1) $k \notin C$, (2) $k \in C$, (3) $k \in C \setminus M$, and (4) $k \in M$.

(Case 1: $k \notin C$.) By Theorem S2.13, for all k,

$$P(c_\delta^*(\mu') = c(k) \mid Z_{1:n}) \xrightarrow[n\to\infty]{\mathrm{P}} \mathbb{1}\left(c_\delta^*(\mu^0) = c(k)\right) = \mathbb{1}\left(k \in C\right). \tag{99}$$

Hence, for all $k \notin C$, $P(k \in M_\delta^*(\mu') \mid Z_{1:n}) \xrightarrow[n\to\infty]{\mathrm{P}} 0$ by Equations 98 and 99.

(Case 2: $k \in C$.) Fix $k \in C$. Define the events $A := \{\mu_k' = \mu_{\min,c(k)}'\}$ and $B := \{c_\delta^*(\mu') = c(k)\}$. Then Equation 98 can be written more succinctly as

$$P(k \in M_\delta^*(\mu') \mid Z_{1:n}) = P(B \mid Z_{1:n}) \, P(A \mid Z_{1:n}, B). \tag{100}$$

By the law of total probability,

$$P(A \mid Z_{1:n}) = P(A \mid Z_{1:n}, B) \, P(B \mid Z_{1:n}) + P(A \mid Z_{1:n}, B^c) \, P(B^c \mid Z_{1:n}), \tag{101}$$

so

$$P(A \mid Z_{1:n}) - P(A \mid Z_{1:n}, B) = P(B^c \mid Z_{1:n}) \left(P(A \mid Z_{1:n}, B^c) - P(A \mid Z_{1:n}, B)\right), \tag{102}$$

hence

$$\left|P(A \mid Z_{1:n}) - P(A \mid Z_{1:n}, B)\right| \le P(B^c \mid Z_{1:n}). \tag{103}$$

By Theorem S2.13, $P(B^c \mid Z_{1:n}) \xrightarrow{\text{P}} 0$ since $k \in C$, and thus,

$$P(A \mid Z_{1:n}, B) = P(A \mid Z_{1:n}) + o_p(1). \tag{104}$$

Combining Equations 100 and 104 and the fact that $P(B \mid Z_{1:n}) \xrightarrow{\text{P}} 1$, we obtain

$$P(k \in M_\delta^*(\mu') \mid Z_{1:n}) = P(B \mid Z_{1:n})P(A \mid Z_{1:n}) + o_p(1) = P(A \mid Z_{1:n}) + o_p(1), \tag{105}$$

that is,

$$P(k \in M_\delta^*(\mu') \mid Z_{1:n}) = P\big(\mu_k' = \mu_{\min,c(k)}' \mid Z_{1:n}\big) + o_p(1). \tag{106}$$

Note that $\mu_{\min,c(k)}' = \min_{j \in C} \mu_j'$ since $k \in C$. We reexpress Equation 106 in terms of an argmin over M, rather than the min over C. First, we show that $\min_{j \in C} \mu_j' = \min_{j \in M} \mu_j'$. If $C = M$, then this is trivial; otherwise, suppose $C \setminus M \neq \varnothing$. Define the within-class gap

$$m := \min_{j \in C \setminus M} \big(\mu_j^0 - \min_{i \in M} \mu_i^0\big) > 0,$$

which is strictly positive since $M \subseteq C$ and $C \setminus M \neq \varnothing$. For $\eta > 0$, define the event $E_n(\eta) := \big\{ \max_{1 \leq k \leq K} |\mu_k' - \mu_k^0| < \eta \big\}$. By Corollary S2.9, for any fixed $\eta > 0$, we have $P\big(E_n(\eta)\big) \to 1$ as $n \to \infty$. Choose $\eta \in (0, m/2)$. On $E_n(\eta)$, for all $j \in M$,

$$\mu_j' \leq \mu_j^0 + \eta = \min_{i \in M} \mu_i^0 + \eta, \tag{107}$$

while for all $j \in C \setminus M$,

$$\mu_j' \geq \mu_j^0 - \eta \geq \min_{i \in M} \mu_i^0 + m - \eta > \min_{i \in M} \mu_i^0 + \eta. \tag{108}$$

Hence, $\min_{j \in C \setminus M} \mu_j' > \min_{j \in M} \mu_j'$, and thus, on event $E_n(\eta)$,

$$\min_{j \in C} \mu_j' = \min_{j \in M} \mu_j'. \tag{109}$$

Under the NIW posterior (Section 3.1), the conditional distribution $\mu' \mid Z_{1:n}, \Sigma'$ is

multivariate normal with positive definite covariance matrix Σ'/λ_n. Consequently, the posterior of $\mu' \mid Z_{1:n}$ has a density with respect to Lebesgue measure on $\mathbb{R}^K$, so for all $i \neq j$, we have $P(\mu'_i = \mu'_j \mid Z_{1:n}) = 0$. Hence, the minimum $\min_{j \in M} \mu'_j$ is attained at only one index, almost surely, under the posterior. Along with Equation 109, this implies

$$P\big(\{\mu'_k = \min_{j \in C} \mu'_j\} \cap E_n(\eta) \,\big|\, Z_{1:n}\big) = P\big(\{\{k\} = \arg\min_{j \in M} \mu'_j\} \cap E_n(\eta) \,\big|\, Z_{1:n}\big).$$

For any events A, B, E, if $P(A \cap E) = P(B \cap E)$, then $|P(A) - P(B)| \leq P(E^c)$. Thus,

$$\left| P\big(\mu'_k = \min_{j \in C} \mu'_j \mid Z_{1:n}\big) - P\big(\{k\} = \arg\min_{j \in M} \mu'_j \mid Z_{1:n}\big) \right| \leq P(E_n(\eta)^c \mid Z_{1:n}). \qquad (110)$$

Since $0 \leq P(E_n(\eta)^c \mid Z_{1:n}) \leq 1$, Markov's inequality yields, for all $\epsilon > 0$,

$$P\Big(P(E_n(\eta)^c \mid Z_{1:n}) > \epsilon \Big) \leq \frac{\mathbb{E}\big(P(E_n(\eta)^c \mid Z_{1:n})\big)}{\epsilon} \xrightarrow[n \to \infty]{} 0, \qquad (111)$$

so $P(E_n(\eta)^c \mid Z_{1:n}) \xrightarrow{\mathrm{P}} 0$. Combining this with Equations 106 and 110, along with the fact that $\mu'_{\min,c(k)} = \min_{j \in C} \mu'_j$ since $k \in C$,

$$P(k \in M^*_\delta(\mu') \mid Z_{1:n}) = P\Big(\{k\} = \arg\min_{j \in M} \mu'_j \,\Big|\, Z_{1:n}\Big) + o_p(1). \qquad (112)$$

(Case 3: $k \in C \setminus M$.) If $k \in C \setminus M$, then $\{\{k\} = \arg\min_{j \in M} \mu'_j\} = \varnothing$, so in this case, Equation 112 implies $P(k \in M^*_\delta(\mu') \mid Z_{1:n}) \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$.

(Case 4: $k \in M$.) Finally, we analyze the asymptotics of the argmin over M in Equation 112. Define

$$V_n := \sqrt{n}\big(\mu' - \mu_n\big), \quad W_n := \sqrt{n}\big(\mu_n - \mu^0\big),$$

where μ_n is the posterior mean defined in Section 3.1, namely, $\mu_n = \big(\lambda_0 \mu_0 + n \overline{Z}_n\big)/\lambda_n$, where the Z values are evaluated at $\hat{\theta}_k$ as in Section S2.3. By the form of the NIW

posterior in Section 3.1, we can represent $\mu' \mid Z_{1:n}, \Sigma'$ as

$$\mu' = \mu_n + (\Sigma')^{1/2}\xi/\sqrt{\lambda_n}, \tag{113}$$

with $\xi \sim \mathcal{N}(0, I)$ independent of $(Z_{1:n}, \Sigma')$. Thus, $V_n \mid Z_{1:n}, \Sigma' \sim \mathcal{N}(0, n\Sigma'/\lambda_n)$, and $W_n \mid Z_{1:n}$ is deterministic. In addition, we have $n/\lambda_n \to 1$, $\Sigma' \xrightarrow{\mathrm{P}} \Sigma^0$ (Theorem S2.7), and $W_n \xrightarrow{d} \mathcal{N}(0, \Sigma^0)$ (Theorem S2.5). Let $k \in M$ and define

$$Q_{nk} := P\Big(\{k\} = \arg\min_{j \in M} \mu'_j \;\Big|\; Z_{1:n}\Big). \tag{114}$$

For all $j \in M$, since $\mu_j^0 = \mu_k^0$,

$$\sqrt{n}(\mu'_j - \mu'_k) = (V_{nj} - V_{nk}) + (W_{nj} - W_{nk}). \tag{115}$$

Note that $\{k\} = \arg\min_{j \in M} \mu'_j$ if and only if $\mu'_j - \mu'_k > 0$ for all $j \in M \setminus \{k\}$. Thus,

$$Q_{nk} = P\big(L_k V_n + L_k W_n > 0 \mid Z_{1:n}\big), \tag{116}$$

where we define the linear difference map $L_k : \mathbb{R}^K \to \mathbb{R}^{|M|-1}$ by $L_k x := (x_j - x_k)_{j \in M \setminus \{k\}}$, and $>$ indicates componentwise inequality. We have

$$L_k V_n \mid Z_{1:n}, \Sigma' \sim \mathcal{N}\big(0, \Gamma_{nk}(\Sigma')\big), \qquad \Gamma_{nk}(\Sigma') := L_k\big(n\Sigma'/\lambda_n\big)L_k^{\mathsf{T}}.$$

Since $n/\lambda_n \to 1$ and $\Sigma' \xrightarrow{\mathrm{P}} \Sigma^0$,

$$\Gamma_{nk}(\Sigma') \xrightarrow[n\to\infty]{\mathrm{P}} \Gamma_k := L_k \Sigma^0 L_k^{\mathsf{T}}. \tag{117}$$

Let $\Phi_\Gamma(\cdot)$ denote the CDF of $\mathcal{N}(0, \Gamma)$. By the symmetry of Gaussian distributions with

mean zero, along with the fact that $W_n \mid Z_{1:n}$ is deterministic,

$$
\begin{aligned}
Q_{nk} &= P\big(- L_k V_n < L_k W_n \mid Z_{1:n}\big) \\
&= \mathbb{E}\big(P\big(- L_k V_n < L_k W_n \mid Z_{1:n}, \Sigma'\big) \mid Z_{1:n}\big) \\
&= \mathbb{E}\big(\Phi_{\Gamma_{nk}(\Sigma')}(L_k W_n) \mid Z_{1:n}\big).
\end{aligned}
\tag{118}
$$

Observe that since $k \in M$, the function $\pi_k(x)$ defined in Equation 16 can equivalently be written as

$$
\pi_k(x) = \Phi_{\Gamma_k}(L_k x)
$$

for $x \in \mathbb{R}^K$. The map $(\Gamma, x) \mapsto \Phi_\Gamma(L_k x)$ is continuous on the space of positive definite covariance matrices and vectors $x \in \mathbb{R}^{|M|-1}$, and is bounded in $[0, 1]$. Let $W \sim \mathcal{N}(0, \Sigma^0)$. Since $\Gamma_{nk}(\Sigma') \xrightarrow{\text{p}} \Gamma_k$ and $W_n \xrightarrow{\text{d}} W$, we have joint convergence

$$
\big(\Gamma_{nk}(\Sigma'), W_n\big) \xrightarrow{\text{d}} \big(\Gamma_k, W\big).
\tag{119}
$$

By the continuous mapping theorem,

$$
D_n := \Phi_{\Gamma_{nk}(\Sigma')}\big(L_k W_n\big) - \Phi_{\Gamma_k}\big(L_k W_n\big) \xrightarrow[n \to \infty]{\text{d}} 0.
\tag{120}
$$

Since the limit is constant, $D_n \xrightarrow{\text{p}} 0$, and $|D_n| \le 1$ for all n. Thus, $\mathbb{E}|D_n| \to 0$ (Jacod and Protter, 2004, Theorem 17.4). By Markov's inequality, for all $\epsilon > 0$,

$$
P\Big(\mathbb{E}\big(|D_n| \big| Z_{1:n}\big) > \epsilon\Big) \le \frac{\mathbb{E}|D_n|}{\epsilon} \xrightarrow[n \to \infty]{} 0.
\tag{121}
$$

Therefore, $\left|\mathbb{E}(D_n|Z_{1:n})\right| \leq \mathbb{E}\big(|D_n|\big|Z_{1:n}\big) = o_p(1)$. Thus, for $k \in M$,

$$
\begin{aligned}
Q_{nk} &= \mathbb{E}\Big(\Phi_{\Gamma_{nk}(\Sigma')}\big(L_k W_n\big) \,\Big|\, Z_{1:n}\Big) \\
&= \Phi_{\Gamma_k}\big(L_k W_n\big) + o_p(1) \\
&= \pi_k(W_n) + o_p(1).
\end{aligned}
\tag{122}
$$

We now combine the cases to obtain the result. Since Equations 112, 114, and 122 hold coordinate-wise for each $k \in M$, they also hold jointly for $k \in M$, because coordinate-wise convergence in probability implies joint convergence in probability. Hence,

$$
\big(P(k \in M_\delta^*(\mu') \mid Z_{1:n})\big)_{k \in M} = \big(\pi_k(W_n)\big)_{k \in M} + o_p(1).
\tag{123}
$$

Meanwhile, by cases (1) and (3) above, for $k \notin M$ we know $P(k \in M_\delta^*(\mu') \mid Z_{1:n}) \xrightarrow{\text{p}} 0$, and by definition $\pi_k(x) = 0$ for all x when $k \notin M$ (Equation 16). Thus, in fact, Equation 123 holds jointly over all $k \in \{1, \ldots, K\}$, not just $k \in M$. Since $W_n \xrightarrow{\text{d}} W \sim \mathcal{N}(0, \Sigma^0)$ and the map $x \mapsto \big(\pi_k(x)\big)_{k=1}^K$ is continuous, the continuous mapping theorem implies $\big(\pi_k(W_n)\big)_{k=1}^K \xrightarrow{\text{d}} \big(\pi_k(W)\big)_{k=1}^K$. Therefore, by Slutsky's theorem,

$$
\big(P(k \in M_\delta^*(\mu') \mid Z_{1:n})\big)_{k=1}^K \xrightarrow[n \to \infty]{\text{d}} \big(\pi_k(W)\big)_{k=1}^K.
\tag{124}
$$

Part (b). By the definition $\tilde{r}_{nk}(\mu) = \mathbb{1}(\mu_k = \mu_{\min, c(k)})$, we have

$$
\tilde{w}_\delta(k \mid Z_{1:n}) = P\big(c_\delta^*(\mu') = c(k) \mid Z_{1:n}\big) P\big(\mu_k' = \mu_{\min, c(k)}' \mid Z_{1:n}\big) = P(B \mid Z_{1:n}) P(A \mid Z_{1:n}),
\tag{125}
$$

where $A = \{\mu_k' = \mu_{\min, c(k)}'\}$ and $B = \{c_\delta^*(\mu') = c(k)\}$ as before. For $k \in C$, subtracting Equation 100 from Equation 125 and using Equation 104, we obtain

$$
\tilde{w}_\delta(k \mid Z_{1:n}) - P\big(k \in M_\delta^*(\mu') \mid Z_{1:n}\big) = P(B \mid Z_{1:n})\big(P(A \mid Z_{1:n}) - P(A \mid Z_{1:n}, B)\big) = o_p(1),
\tag{126}
$$

since $|P(B \mid Z_{1:n})| \leq 1$. Meanwhile, for $k \notin C$, Equation 99 implies that both $\tilde{w}_\delta(k \mid Z_{1:n})$

and $P(k \in M_\delta^*(\mu') \mid Z_{1:n})$ converge to 0 in probability, as in case (1); thus, their difference also converges to 0 in probability. Therefore, for all k, regardless of whether $k \in C$,

$$\tilde{w}_\delta(k \mid Z_{1:n}) - P(k \in M_\delta^*(\mu') \mid Z_{1:n}) \xrightarrow[n\to\infty]{\mathrm{P}} 0. \tag{127}$$

Therefore,

$$\left(\tilde{w}_\delta(k \mid Z_{1:n})\right)_{k=1}^{K} = \left(P(k \in M_\delta^*(\mu') \mid Z_{1:n})\right)_{k=1}^{K} + o_p(1). \tag{128}$$

Along with the result of part (a), this implies that $\left(\tilde{w}_\delta(k \mid Z_{1:n})\right)_{k=1}^{K} \xrightarrow{\mathrm{d}} \left(\pi_k(W)\right)_{k=1}^{K}$ by Slutsky's theorem. $\qquad\square$

Proof of Corollary 4.5. By Theorem 4.4(b), it suffices to show that

$$\left(\pi_{k_1}(W), \pi_{k_2}(W)\right) \overset{d}{=} (U, 1 - U) \tag{129}$$

where $W \sim \mathcal{N}(0, \Sigma^0)$ and $U \sim \mathrm{Uniform}(0, 1)$. Since $M_\delta^*(\mu^0) = \{k_1, k_2\}$, the definition of $\pi_k(x)$ reduces to

$$\pi_{k_1}(x) = P(W_{k_2} - W_{k_1} \le x_{k_2} - x_{k_1}), \quad \pi_{k_2}(x) = P(W_{k_1} - W_{k_2} \le x_{k_1} - x_{k_2}) \tag{130}$$

for all $x \in \mathbb{R}^K$. Let F be the cumulative distribution function of the Gaussian random variable $W_{k_2} - W_{k_1}$. Since Σ^0 is positive definite, $\mathrm{Var}(W_{k_2} - W_{k_1}) > 0$, so F is continuous and strictly increasing. Then, $\pi_{k_1}(x) = F(x_{k_2} - x_{k_1})$ and

$$\pi_{k_2}(x) = P\left(-(W_{k_2} - W_{k_1}) \le -(x_{k_2} - x_{k_1})\right) = 1 - F(x_{k_2} - x_{k_1}). \tag{131}$$

Hence, for all x,

$$\pi_{k_1}(x) + \pi_{k_2}(x) = 1. \tag{132}$$

Now, we consider evaluating these functions at $x = W$. By the probability integral transform,

$$\pi_{k_1}(W) = F(W_{k_2} - W_{k_1}) \sim \mathrm{Uniform}(0, 1). \tag{133}$$

S30

By Equation 132, $\pi_{k_2}(W) = 1 - \pi_{k_1}(W)$. Combined with Equation 133, this implies

$$\big(\pi_{k_1}(W), \pi_{k_2}(W)\big) \overset{\mathrm{d}}{=} (U, 1 - U), \quad U \sim \mathrm{Uniform}(0, 1). \tag{134}$$

This completes the proof. $\qquad\square$

S3 Additional details on the examples

S3.1 Gaussian mixture example details

This section provides additional details on the Gaussian mixture example in Section 5.1. For a given number of mixture components k, we model the observations $x_1, \ldots, x_n \in \mathbb{R}$ by a k-component Gaussian mixture

$$f_k(x) = \sum_{j=1}^{k} \phi_j \, \mathcal{N}\big(x \,\big|\, m_j, s_j^2\big)$$

with parameters $\theta = \big\{(\phi_j, m_j, s_j^2)\big\}_{j=1}^{k}$, where $\phi_1, \ldots, \phi_k \geq 0$ are mixture weights such that $\sum_{j=1}^{k} \phi_j = 1$, while $m_j \in \mathbb{R}$ and $s_j^2 > 0$ are the mean and variance, respectively, of the jth component.

Prior distributions. Following Fraley and Raftery (2007), we use a normal-inverse-gamma prior, which is a conjugate prior for the means and variances:

$$m_j \mid s_j^2 \sim \mathcal{N}(m_0, s_j^2/\kappa_0), \quad s_j^2 \sim \mathrm{InverseGamma}(\nu_0/2, s_0^2/2),$$

and we place a uniform prior on the vector of mixing proportions $\phi = (\phi_1, \ldots, \phi_k)$, that is, $\phi \sim \mathrm{Dirichlet}(1, \ldots, 1)$, which is also conjugate. In each run (for a given n and k), we set $m_0 = \bar{x}$, $\kappa_0 = 0.01$, and $s_0^2 = \widehat{\mathrm{Var}}(x)/k^2$, following the default prior choices in Fraley and Raftery (2007), and we set the degrees of freedom to $\nu_0 = 10$.

Expectation-maximization (EM) algorithm. We estimate $\theta = \left\{ (\phi_j, m_j, s_j^2) \right\}_{j=1}^k$ via the EM algorithm, using *maximum a posteriori* (MAP) updates to prevent likelihood singularities and improve stability (Bishop, 2006). The E-step consists of computing the responsibilities

$$r_{ij} = \frac{\phi_j \, \mathcal{N}(x_i \mid m_j, s_j^2)}{\sum_{l=1}^k \phi_l \, \mathcal{N}(x_i \mid m_l, s_l^2)}.$$

For the M-step, define $n_j := \sum_{i=1}^n r_{ij}$ and define the weighted mean $\bar{x}_j := \frac{1}{n_j} \sum_{i=1}^n r_{ij} x_i$. The MAP M-step updates are then

$$\phi_j^{\text{new}} = \frac{n_j}{n}, \qquad m_j^{\text{new}} = \frac{\kappa_0 m_0 + n_j \bar{x}_j}{\kappa_0 + n_j},$$

$$s_j^{2,\text{new}} = \frac{s_0^2 + \frac{\kappa_0 n_j}{\kappa_0 + n_j}(\bar{x}_j - m_0)^2 + \sum_{i=1}^n r_{ij}(x_i - \bar{x}_j)^2}{\nu_0 + n_j + 3}.$$

We run this EM algorithm using the `mclust` package in R (Scrucca et al., 2016; R Core Team, 2021) with random initialization. For each k, we run the algorithm on 50 independent restarts and keep the run with the highest observed log-likelihood.

Bayesian posterior on the number of components. In this example, the model index k represents the number of mixture components. As a comparison method, we obtain a posterior over the (unknown) number of components k by placing a prior on k and computing its posterior $\pi(k \mid x_{1:n})$ under a mixture of finite mixtures model (MFM; Miller and Harrison, 2018). Specifically, we take $K \sim \text{Uniform}\{1, 2, \ldots, 30\}$; so in particular, $P(K > 30) = 0$. For the mixture weights, we take $(\phi_1, \ldots, \phi_k) \mid K = k \sim \text{Dirichlet}(1, \ldots, 1)$. For the component parameters, we use independent conditionally conjugate priors on the mean and variance; specifically, given $K = k$, we take

$$m_j \sim \mathcal{N}(m_0, \sigma_0^2), \quad s_j^2 \sim \text{InverseGamma}(2, b),$$

independently for $j = 1, \ldots, k$, where $m_0 = (\min_i x_i + \max_i x_i)/2$ and $\sigma_0^2 = (\max_i x_i - \min_i x_i)^2$, and we further place a $b \sim \text{Gamma}\big(0.2, 10/\sigma_0^2\big)$ hyperprior on b.

We run an MCMC algorithm implemented in the Julia package BayesianMixtures.jl (https://github.com/jwmi/BayesianMixtures.jl) targeting the MFM posterior for 200,000 iterations, discard the first 100,000 as burn-in, and use the remaining 100,000 draws for posterior inference. The MCMC algorithm samples from the joint posterior over assignments of observations to components, component parameters, and the hyperparameter b. We obtain posterior draws of k using the technique of Miller and Harrison (2018, Equation 3.7), specifically, we sample from $k|t$ for each MCMC sample of t, where t denotes the number of occupied clusters in a posterior draw.

S3.2 Sparse MVN example details

In this section, we provide additional details on the sparse multivariate normal (MVN) model example in Section 5.2.

Kullback–Leibler (KL) divergence calculation. The KL divergence between two multivariate normal distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ on $\mathbb{R}^k$ is

$$D\big(\mathcal{N}(\mu_1, \Sigma_1) \,\|\, \mathcal{N}(\mu_2, \Sigma_2)\big) = \frac{1}{2}\bigg(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^{\mathsf{T}}\Sigma_2^{-1}(\mu_1 - \mu_2) - k + \log\frac{|\det\Sigma_2|}{|\det\Sigma_1|}\bigg).$$

In our example, $\Sigma_1 = \Sigma_2 = I$, so the KL divergence simplifies to $D(\mathcal{N}(\mu_1, \Sigma_1) \,\|\, \mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2}\|\mu_1 - \mu_2\|^2$. Thus, for a candidate MVN model k whose mean is constrained to a linear subspace Θ_k, the minimal KL divergence from the true model $\mathcal{N}(\theta_0, I)$ is

$$\min_{\theta_k \in \Theta_k} D\big(\mathcal{N}(\theta_0, I) \,\|\, \mathcal{N}(\theta_k, I)\big) = \min_{\theta_k \in \Theta_k} \frac{1}{2}\|\theta_0 - \theta_k\|^2 = \frac{1}{2}\|\theta_0 - \theta_k^*\|^2,$$

where $\theta_k^* := \arg\min_{\theta \in \Theta_k} \|\theta_0 - \theta\|^2$, that is, θ_k^* is the Euclidean orthogonal projection of θ_0 onto the parameter subspace for model k.

Coarsened posterior (c-posterior). In Section 5.2.1, we compare with the c-posterior of Miller and Dunson (2019); here we provide details on this method. For robustness to misspecification, the c-posterior is obtained by conditioning on the event that the empir-

ical distribution generated by the model is "close" to that of the observed data. When closeness is measured by KL divergence (also known as relative entropy), the c-posterior can be approximated by a power posterior, obtained by raising the likelihood to a fractional power ζ_n (Miller and Dunson, 2019). Concretely, if $\pi_k(\theta)$ is the prior under model k, then the power posterior for model k is

$$\pi_{\alpha,k}(\theta; x_{1:n}) \propto \pi_k(\theta) \prod_{i=1}^{n} f_k(x_i; \theta)^{\zeta_n},$$

where $\zeta_n := \alpha/(\alpha + n) \in (0, 1]$ and $\alpha > 0$ is a setting that control the precision of the power posterior. In particular, the power posterior recovers the standard posterior as a special case in the limit as $\alpha \to \infty$ ($\zeta_n \to 1$). The marginal power likelihood is defined as

$$m_{\alpha,k}(x_{1:n}) := \int \pi_k(\theta) \prod_{i=1}^{n} f_k(x_i; \theta)^{\zeta_n} d\theta,$$

and can be used to construct a posterior on models k that provides robust model inference. For exponential families, the power likelihood has the attractive property that it preserves conjugacy. Thus, when using conjugate priors with exponential families, the power posterior and power marginal likelihood take closed form expressions.

We specialize the c-posterior framework to the sparse MVN mean model in Section 5.2. Recall that we observe $x_1, \ldots x_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta_0, I)$ where $\theta_0 \in \mathbb{R}^6$ is the true mean, which is treated as unknown. Consider the $K = 7$ models in Table 1, with a uniform prior over models, $\pi(k) = 1/K$. For each model $k = 1, \ldots, K$, let $J_k \subseteq \{1, \ldots, 6\}$ be the set of free coordinates and define $d_k = |J_k|$; coordinates in J_k^c are fixed at 0. Denote the sample mean by $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_6)^{\mathsf{T}}$. Since the covariance matrix is I, the power likelihood for each model is

$$\prod_{i=1}^{n} f_k(x_i; \theta)^{\zeta_n} \propto \prod_{i=1}^{n} \exp\left(-\frac{\zeta_n}{2}\|x_i - \theta\|^2\right) \propto \exp\left(-\frac{\zeta_n n}{2}\|\bar{x} - \theta\|^2\right). \tag{135}$$

Denote $\theta_{J_k} = (\theta_j : j \in J_k)$ and likewise for $\theta_{J_k^c}$, as well as $\bar{x}_{J_k} = (\bar{x}_j : j \in J_k)$. Under

model k, we have $\theta_{J_k^c} = 0$, and hence, $\|\bar{x} - \theta\|^2 = \|\bar{x}_{J_k} - \theta_{J_k}\|^2 + \|\bar{x}_{J_k^c}\|^2$. Therefore,

$$\prod_{i=1}^{n} f_k(x_i; \theta)^{\zeta_n} \propto \exp\left(-\frac{\zeta_n n}{2}\|\bar{x}_{J_k} - \theta_{J_k}\|^2\right) \exp\left(-\frac{\zeta_n n}{2}\|\bar{x}_{J_k^c}\|^2\right). \tag{136}$$

For each model k, we place a Gaussian prior on the free coordinates and fix the rest to zero:

$$\theta_{J_k} \sim \mathcal{N}\left(\theta_{0J_k}, \kappa_0^{-1} I_{d_k}\right), \qquad \theta_{J_k^c} = 0,$$

where $\theta_{0J_k} \in \mathbb{R}^{d_k}$ is a fixed prior mean and $\kappa_0 > 0$ is a fixed prior precision. Then, by conjugacy, the power posterior on θ_{J_k} under model k is

$$\theta_{J_k} \mid x_{1:n}, \alpha \sim \mathcal{N}\left(\frac{\kappa_0 \theta_{0J_k} + \zeta_n n \bar{x}_{J_k}}{\kappa_0 + \zeta_n n}, (\kappa_0 + \zeta_n n)^{-1} I_{d_k}\right).$$

Observe that the constant of proportionality does not depend on k in Equation 135, and hence also in Equation 136. Thus, writing the first factor of Equation 136 in the form of a Gaussian density, we have

$$\prod_{i=1}^{n} f_k(x_i; \theta)^{\zeta_n} = \frac{C}{(\zeta_n n)^{d_k/2}} \mathcal{N}(\bar{x}_{Jk} \mid \theta_{Jk}, (\zeta_n n)^{-1} I_{d_k}) \exp\left(-\frac{\zeta_n n}{2}\|\bar{x}_{J_k^c}\|^2\right) \tag{137}$$

where C does not depend on θ or k. Then, the marginal power likelihood is

$$\begin{aligned}
m_{\alpha,k}(x_{1:n}) &= \int \pi(\theta_{J_k}) \prod_{i=1}^{n} f_k(x_i; \theta)^{\zeta_n} d\theta_{J_k} \\
&= \frac{C \exp(-\frac{\zeta_n n}{2}\|\bar{x}_{J_k^c}\|^2)}{(\zeta_n n)^{d_k/2}} \int \mathcal{N}(\bar{x}_{J_k} \mid \theta_{Jk}, (\zeta_n n)^{-1} I_{d_k}) \mathcal{N}(\theta_{J_k} \mid \theta_{0J_k}, \kappa_0^{-1} I_{d_k}) d\theta_{J_k} \\
&= \frac{C \exp(-\frac{\zeta_n n}{2}\|\bar{x}_{J_k^c}\|^2)}{(\zeta_n n)^{d_k/2}} \mathcal{N}(\bar{x}_{J_k} \mid \theta_{0J_k}, ((\zeta_n n)^{-1} + \kappa_0^{-1}) I_{d_k}).
\end{aligned}$$

Therefore, the log marginal power likelihood is

$$\log m_{\alpha,k}(x_{1:n}) = c + \frac{d_k}{2}\log\left(\frac{\kappa_0}{\kappa_0 + \zeta_n n}\right) - \frac{\kappa_0 \zeta_n n}{2(\kappa_0 + \zeta_n n)}\|\bar{x}_{J_k} - \theta_{0J_k}\|^2 - \frac{\zeta_n n}{2}\|\bar{x}_{J_k^c}\|^2 \tag{138}$$

where c is a constant that does not depend on k. Since we assume a uniform prior over

models, $\pi(k) = 1/K$, the coarsened posterior on models is

$$\pi_\alpha(k; x_{1:n}) = \frac{m_{\alpha,k}(x_{1:n})}{\sum_{j=1}^K m_{\alpha,j}(x_{1:n})}. \tag{139}$$

Equations 138 and 139 yield a closed-form expression that enables exact computation of the c-posterior probabilities for the K models, which we use in Section 5.2.1.

S3.3   Thermal performance curve example details

This section provides additional details on the example from Section 5.3.

Data source.   The dataset used for fitting was obtained from Kontopoulos et al. (2024a), available at `https://doi.org/10.6084/m9.figshare.24106161.v2`.

MLEs for noise model and flexible model.   The MLE for the noise model is $\hat{\varphi} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\hat{\sigma}_{\text{noise}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\varphi})^2$. The MLE for the flexible model is $\hat{\beta}_t = \frac{1}{n_t} \sum_{i=1}^n y_i \mathbb{1}(t_i = t)$ where $n_t = \sum_{i=1}^n \mathbb{1}(t_i = t)$, and $\hat{\sigma}_{\text{flex}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_{t_i})^2$.

Functional forms of the TPC candidates.   For completeness, we state the functional forms of the TPC candidate models here. All formulas are taken from the supplementary material of Kontopoulos et al. (2024b). Temperature parameters are real-valued, and in nonlinear least squares fitting we impose the constraints $t_{pk} \in [0, 150]$ and $t_{\min} \leq t_{pk} \leq t_{\max}$.

1. Eubank (3 parameters)

$$g_1(t; a, b, t_{pk}) = \frac{a}{(t - t_{pk})^2 + b}, \qquad a > 0, \ b > 0.$$

2. Gaussian (3 parameters)

$$g_2(t; a, b_{pk}, t_{pk}) = b_{pk} \exp\left(-\frac{1}{2a^2}(t - t_{pk})^2\right), \qquad a > 0, \ b_{pk} > 0.$$

3. Mitchell–Angilletta (3 parameters)

$$g_3(t;\, a, b, t_{pk}) = \frac{a}{2b}\left(1 + \cos\left(\frac{t - t_{pk}}{b}\cdot \pi\right)\right), \qquad a > 0,\ b > 0.$$

4. Brière I (3 parameters)

$$g_4(t;\, a, t_{\min}, t_{\max}) = a\, t\, (t - t_{\min})\sqrt{t_{\max} - t}, \qquad a > 0,\ t_{\min} < t < t_{\max}.$$

5. Weibull (4 parameters)

$$g_5(t;\, b, c, b_{pk}, t_{pk}) = b_{pk}\left(\frac{c-1}{c}\right)^{(1-c)/c}\left(\frac{t - t_{pk}}{b} + \left(\frac{c-1}{c}\right)^{1/c}\right)^{c-1}$$
$$\times \exp\left(-\left(\frac{t - t_{pk}}{b} + \left(\frac{c-1}{c}\right)^{1/c}\right)^{c} + \frac{c-1}{c}\right),$$

$b > 0,\ c > 0,\ b_{pk} > 0.$

6. Modified Gaussian (4 parameters)

$$g_6(t;\, a, b, b_{pk}, t_{pk}) = b_{pk}\exp\left(-\frac{1}{2}\left(\frac{|t - t_{pk}|}{a}\right)^{b}\right), \qquad a > 0,\ b > 0,\ b_{pk} > 0.$$

7. Extended Brière (5 parameters)

$$g_7(t;\, a, b, c, t_{\min}, t_{\max}) = a\, t\, (t - t_{\min})^{b}\, (t_{\max} - t)^{c},$$

$a > 0,\ b > 0,\ c > 0,\ t_{\min} < t < t_{\max}.$

8. Fifth-order polynomial (Poly-5; 6 parameters)

$$g_8(t;\, a, b, c, d, f, g) = a + bt + ct^2 + dt^3 + ft^4 + gt^5.$$

9. (Extended) Sharpe–Schoolfield (7 parameters)

$$g_9(T;\, a, b_0, t_{L50}, t_{H50}, \Delta H_L, \Delta H_A^{\ddagger}, \Delta H_H)$$

$$= \frac{b_0 \dfrac{T}{T_{\text{ref}}} \exp\left(\dfrac{\Delta H_A^{\ddagger}}{R} \left(\dfrac{1}{T_{\text{ref}}} - \dfrac{1}{T} \right) \right)}{a + \exp\left(\dfrac{\Delta H_L}{R} \left(\dfrac{1}{T_{L50}} - \dfrac{1}{T} \right) \right) + \exp\left(\dfrac{\Delta H_H}{R} \left(\dfrac{1}{T_{H50}} - \dfrac{1}{T} \right) \right)},$$

$a > 0,\ b_0 > 0,\ \Delta H_L > 0,\ \Delta H_A^{\ddagger} > 0,\ \Delta H_H > 0$. Here, $T_{\text{ref}} = 273.5$ (temperatures in Kelvin) and $R = 1.987$.

S3.4 Population structure admixture example details

Here, we provide additional details on the admixture model example in Section 5.4.

STRUCTURE output: Estimated Ln Prob of Data   STRUCTURE performs posterior inference for the admixture model using MCMC, and reports an "Estimated Ln Prob of Data" (ELP), which is intended as a rough approximation to the log marginal likelihood $\log p(x \mid k)$. For LaD, we use this quantity as a fit diagnostic to filter out failed MCMC runs; specifically, we run the STRUCTURE algorithm 20 times with random restarts, and keep the run with the highest ELP. This quantity is also used for implementing Evanno's method, in which the ELP is used to compute $L(k)$.

As described in Pritchard et al. (2000), the ELP is constructed from the Bayesian deviance

$$\text{dev}(\phi, q, s) := -2 \log p(x \mid \phi, q, s)$$

evaluated on the MCMC output. Specifically, Pritchard et al. (2000) use a normal approximation to show that

$$-2 \log p(x \mid k) \approx \mathbb{E}\big(\text{dev}(\phi, q, s) \mid x, k \big) + \frac{1}{4} \text{Var}\big(\text{dev}(\phi, q, s) \mid x, k \big).$$

In practice, $\mathbb{E}\big(\text{dev}(\phi, q, s) \mid x, k \big)$ and $\text{Var}\big(\text{dev}(\phi, q, s) \mid x, k \big)$ are approximated by Monte Carlo averages over the post–burn-in MCMC draws. The resulting approximation of

$\log p(x \mid k)$ is what STRUCTURE returns as "Estimated Ln Prob of Data" (ELP).

Per-individual negative log-likelihoods. For the LaD technique, we need to compute the per-individual negative log-likelihood values $\ell_k(x_i; \hat{\phi}, \hat{\psi}) = -\log p(x_i \mid \hat{\phi}, \hat{\psi})$, integrating out q_i and s_{ila} for all l, a, given the point estimate $(\hat{\phi}, \hat{\psi})$ for model k. We compute these using Monte Carlo approximations, as follows. Fix a given model k, and let $\hat{\phi}$ and $\hat{\psi}$ denote the posterior means obtained from MCMC. First, observe that

$$p(x_{ila} = v \mid \phi, q_i) = \sum_{j=1}^{k} p(s_{ila} = j \mid q_i)\, p(x_{ila} = v \mid s_{ila} = j, \, \phi) = \sum_{j=1}^{k} q_{ij}\phi_{jl}(v).$$

We write $\phi_{jl}(v)$ instead of ϕ_{jlv} for visual clarity of the notation in what follows. Thus, for the whole genotype of individual i,

$$p(x_i \mid \phi, \psi) = \int_{\Delta^{k-1}} \left(\prod_{l=1}^{L} \prod_{a=1}^{2} \sum_{j=1}^{k} q_{ij}\phi_{jl}(x_{ila}) \right) \mathrm{Dirichlet}(q_i \mid \psi)\, dq_i,$$

where the integral is over the simplex $\Delta^{k-1} = \{q \in \mathbb{R}_+^k : \sum_{j=1}^{k} q_j = 1\}$ and $\mathrm{Dirichlet}(q \mid \psi) \propto \prod_{j=1}^{k} q_j^{\psi_j - 1}$. For each individual i, draw $q_i^{(b)} \sim \mathrm{Dirichlet}(\hat{\psi})$ for $b = 1, \ldots, B$. Then, for each sample b, compute

$$\log p(x_i \mid \hat{\phi}, q_i^{(b)}) = \sum_{l=1}^{L} \sum_{a=1}^{2} \log \left(\sum_{j=1}^{k} q_{ij}^{(b)}\hat{\phi}_{jl}(x_{ila}) \right).$$

We would like to form a Monte Carlo approximation $\tilde{p}(x_i \mid \hat{\phi}, \hat{\psi}) = \frac{1}{B} \sum_{b=1}^{B} p(x_i \mid \hat{\phi}, q_i^{(b)})$, but we need to do so in a numerically stable way. To this end, we use the log-sum-exp trick; that is, we compute $C := \max_b \log p(x_i \mid \hat{\phi}, q_i^{(b)})$ and then

$$\log \tilde{p}(x_i \mid \hat{\phi}, \hat{\psi}) = \log \left(\frac{1}{B} \sum_{b=1}^{B} \exp \left(\log p(x_i \mid \hat{\phi}, q_i^{(b)}) - C \right) \right) + C.$$

This yields Monte Carlo approximations of the per-individual negative log-likelihood values via $\tilde{\ell}_k(x_i; \hat{\phi}, \hat{\psi}) = -\log \tilde{p}(x_i \mid \hat{\phi}, \hat{\psi})$. The additional computational burden of this procedure is not significant relative to the time required for posterior sampling with the

MCMC algorithm.

S4 Alternative LaD model with diagonal covariance

Our proposed LaD model (Section 3.1) places a NIW prior on (μ, Σ), which uses a full covariance matrix for Σ. In Section 5.2.1, we compare with an alternative LaD model that employs a diagonal covariance matrix. We provide the details of this diagonal covariance approach here.

Assume $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_K^2)$, and model each coordinate k independently with a normal-inverse-gamma prior. Specifically, for each k, consider modeling

$$Z_{ik} \mid \mu_k, \sigma_k^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k, \sigma_k^2), \quad \sigma_k^2 \sim \text{InverseGamma}(a_{0k}, b_{0k}), \quad \mu_k \mid \sigma_k^2 \sim \mathcal{N}(\mu_{0k}, \sigma_k^2/\lambda_{0k}),$$

where $\text{InverseGamma}(a, b)$ denotes an inverse-gamma distribution with shape a and scale b. Let $\overline{Z}_{\cdot k} = \frac{1}{n} \sum_{i=1}^n Z_{ik}$. Then the posterior is, independently for $k = 1, \ldots, K$,

$$\sigma_k^2 \mid Z_{1:n} \sim \text{InverseGamma}(a_{nk}, b_{nk}), \quad \mu_k \mid \sigma_k^2, Z_{1:n} \sim \mathcal{N}(\mu_{nk}, \sigma_k^2/\lambda_{nk}),$$

where

$$\lambda_{nk} = \lambda_{0k} + n, \quad \mu_{nk} = \frac{\lambda_{0k}\mu_{0k} + n\overline{Z}_{\cdot k}}{\lambda_{nk}}, \quad a_{nk} = a_{0k} + \frac{n}{2},$$

$$b_{nk} = b_{0k} + \frac{1}{2}\sum_{i=1}^n (Z_{ik} - \overline{Z}_{\cdot k})^2 + \frac{1}{2}\frac{\lambda_{0k}n}{\lambda_{nk}}(\overline{Z}_{\cdot k} - \mu_{0k})^2.$$

We set the hyperparameters to match the NIW prior $\Sigma \sim \text{InverseWishart}(\Psi_0, \nu_0)$ and $\mu \mid \Sigma \sim \mathcal{N}(\mu_0, \Sigma/\lambda_0)$, by choosing

$$a_{0k} = \frac{\nu_0 - K + 1}{2}, \quad b_{0k} = \frac{\Psi_{0,kk}}{2}, \quad \mu_{0k} = (\mu_0)_k, \quad \lambda_{0k} = \lambda_0,$$

for $k = 1, \ldots, K$. With this choice, the NIW marginal for Σ_{kk} match $\text{IG}(a_{nk}, b_{nk})$ and yields $\mu_k \mid \sigma_k^2 \sim \mathcal{N}(\mu_{0k}, \sigma_k^2/\lambda_{0k})$ under independence.

References

G. Barmalzan and T. Payandeh Najafabad. Model confidence set based on Kullback–Leibler divergence distance. *Journal of Statistical Research of Iran JSRI*, 9(2):179–193, 2013.

J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

F. Brück, J.-D. Fermanian, and A. Min. A corrected Clarke test for model selection and beyond. *Journal of Econometrics*, 235(1):105–132, 2023.

Y. Chen and Y. Yang. The one standard error rule for model selection: Does it work? *Stats*, 4(4):868–892, 2021.

K. A. Clarke. A simple distribution-free test for nonnested model selection. *Political Analysis*, 15(3):347–363, 2007.

M. Dewaskar, C. Tosh, J. Knoblauch, and D. B. Dunson. Robustifying likelihoods by optimistically re-weighting data. *Journal of the American Statistical Association*, pages 1–12, 2025.

F. X. Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1, 2015.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

J. A. Dupuis and C. P. Robert. Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111(1-2):77–94, 2003.

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

E. Fong, C. Holmes, and S. G. Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1357–1391, 2023.

C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007.

D. T. Frazier, D. J. Nott, and C. Drovandi. Synthetic likelihood in misspecified models. *Journal of the American Statistical Association*, 120(550):884–895, 2025.

C. Goutis and C. P. Robert. Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika*, 85(1):29–37, 1998.

P. R. Hansen, A. Lunde, and J. M. Nason. The model confidence set. *Econometrica*, 79 (2):453–497, 2011.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

J. H. Huggins and J. W. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79, 2023.

J. H. Huggins and J. W. Miller. Reproducible parameter inference using bagged posteriors. *Electronic Journal of Statistics*, 18(1):1549–1585, 2024.

J. Jacod and P. Protter. *Probability Essentials*. Springer Science & Business Media, 2004.

J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

D. Kontopoulos, L. Pembrey, A. Gibson, S. Humphries, and A. P. Allen. Data from "No model to rule them all: A systematic comparison of 83 thermal performance curve

models across traits and taxonomic groups"', 2024a. URL https://doi.org/10.6084/m9.figshare.24106161.v2. CC-BY-4.0 license.

D.-G. Kontopoulos, A. Sentis, M. Daufresne, N. Glazman, A. I. Dell, and S. Pawar. No universal mathematical model for thermal performance curves across traits and taxonomic groups. *Nature Communications*, 15(1):8855, 2024b.

M. Li and D. B. Dunson. Comparing and weighting imperfect models using D-probabilities. *Journal of the American Statistical Association*, 115(531):1349–1360, 2020.

B. Lindsay and J. Liu. Model assessment tools for a model false world. *Statistical Science*, 24(3):303–318, 2009.

Y. McLatchie, S. Rögnvaldsson, F. Weber, and A. Vehtari. Advances in projection predictive inference. *Statistical Science*, 40(1):128–147, 2025.

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

J. W. Miller and M. T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

S. J. Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Courier Corporation, 2005.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

R. Royall. *The Likelihood Paradigm for Statistical Evidence*, chapter 5, pages 119–152. University of Chicago Press, 2004.

R. Royall and T.-S. Tsou. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):391–404, 2003.

A. Sayyareh. Testing several rival models using the extension of Vuong's test and quasi clustering. *Journal of the Iranian Statistical Society*, 20(2):43–63, 2022.

S. M. Schennach and D. Wilhelm. A simple parametric model selection test. *Journal of the American Statistical Association*, 112(520):1663–1674, 2017.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289, 2016.

S. Shao, P. E. Jacob, J. Ding, and V. Tarokh. Bayesian model comparison with the Hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, 2019.

H. Shimodaira. Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. *Communications in Statistics - Theory and Methods*, 30(8-9):1751–1772, 2001.

V. Shirvaikar, S. G. Walker, and C. Holmes. A general framework for probabilistic model uncertainty. *arXiv preprint arXiv:2410.17108*, 2024.

T. Sivula, M. Magnusson, A. A. Matamoros, and A. Vehtari. Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *Bayesian Analysis*, 1(1):1–31, 2025.

A. Vehtari and J. Lampinen. Model selection via predictive explanatory power. *Report B38, Laboratory of Computational Engineering, Helsinki University of Technology*, 2004.

A. Vehtari and J. Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1003, 2018.