Conditioning on posterior samples for flexible frequentist goodness-of-fit testing

Ritwik Bhaduri* Aabesh Bhattacharyya† Rina Foygel Barber†
Lucas Janson*

Abstract

Tests of *goodness of fit* are used in nearly every domain where statistics is applied. One powerful and flexible approach is to sample artificial data sets that are exchangeable with the real data under the null hypothesis (but not under the alternative), as this allows the analyst to conduct a valid test using *any* test statistic they desire. Such sampling is typically done by conditioning on either an exact or approximate sufficient statistic, but existing methods for doing so have significant limitations, which either preclude their use or substantially reduce their power or computational tractability for many important models. In this paper, we propose to condition on samples from a Bayesian posterior distribution, which constitute a very different type of approximate sufficient statistic than those considered in prior work. Our approach, *approximately co-sufficient sampling via Bayes* (aCSS-B), considerably expands the scope of this flexible type of goodness-of-fit testing. We prove the approximate validity of the resulting test, and demonstrate its utility on three common null models where no existing methods apply, as well as its outperformance on models where existing methods do apply.

1 Introduction

Goodness-of-fit (GoF) testing refers to the problem of testing whether a particular family of distributions (the "model") is consistent with the observed data: for instance, does the data follow a Gaussian distribution? GoF testing is heavily studied in statistics and has applications across domains, including biology (Guo and Thompson, 1992), economics (Cowell et al., 2009), astronomy (Acharya and Kashyap, 2024), and finance (Frezza, 2014). In this paper we will consider *parametric* null hypotheses of the form

$$H_0 : X \sim f_\theta \quad \text{for some } \theta \in \Theta, \tag{1}$$

where $\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ represents a parametric family of densities. This is hypothesis testing with a *composite* null hypothesis space (i.e., the model whose goodness-of-fit is being tested). In GoF testing, it is common to leave the alternative hypothesis unspecified in general: we can interpret the test as asking whether the model appears to fit the data,

*Department of Statistics, Harvard University
†Department of Statistics, University of Chicago

or not. However, in specific settings, it may be the case that we design a test with a particular alternative in mind, as we will see in the examples developed later on. A key challenge of GoF testing problems is that often, any alternative hypothesis of interest would typically be very high-dimensional or even infinite-dimensional: for instance, if the data does not follow a Gaussian distribution (the null), then perhaps it instead follows some heavier-tailed distribution (a nonparametric alternative). In such settings, any powerful test statistic is often too complex to permit any theoretical calculation of its null distribution, even asymptotically.

If the null hypothesis were simple, i.e., if $\Theta = \{\theta_0\}$, then any function T of the data X could be used as a test statistic and converted to a valid p-value by fixing a positive integer M, generating i.i.d. samples $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ from f_{θ_0}, and computing

$$\text{pval} = \frac{1 + \sum_{m=1}^{M} \mathbf{1}\{T(\widetilde{X}^{(m)}) \geq T(X)\}}{M+1}. \tag{2}$$

To see why the above p-value is valid, note that the numerator of Equation (2) is the rank of $T(X)$ among $T(X), T(\widetilde{X}^{(1)}), \dots, T(\widetilde{X}^{(M)})$, and under H_0, these $M+1$ random variables are exchangeable. Unlike standard parametric tests like the likelihood ratio, Wald, and score tests, which each prescribe a specific test statistic based on an alternative hypothesis space that must satisfy certain strong regularity conditions, the absolute flexibility in the choice of test statistic function T in (2) allows the user to leverage any domain knowledge, prior information, and believed structure under the alternative (no matter what it is) to make the test as powerful as possible.

The appealing strategy of the previous paragraph works because H_0 is a simple null—it contains just one distribution, and so the analyst knows what distribution to sample from in order to get exchangeable copies of the data. The idea of *co-sufficient sampling* extends this strategy to *composite* null hypotheses by conditioning on a sufficient statistic for the null model, rendering the data distribution conditionally parameter-free under H_0 so the analyst knows once again what *conditional* distribution to sample from in order to get exchangeable copies of the data. Exact co-sufficient sampling (Bartlett, 1937; Engen and Lillegård, 1997; Agresti, 1992; Stephens, 2012) can only be fruitfully applied for a very narrow class of parametric null models, motivating approximate versions (Barber and Janson, 2022; Zhu and Barber, 2023; Xie and Huang, 2025; Awan and Cai, 2020) that apply to a much broader class of models. Yet these existing works' limitations in scope, power, and computational efficiency still prevent the idea of co-sufficient sampling from realizing its full methodological potential in terms of generality and performance. Section 2 will review in more detail the landscape of existing methods based on the idea of co-sufficient sampling, as well as their shortcomings and other related work.

Our contributions. This work presents a novel approach to approximate co-sufficient sampling that uses draws from a Bayesian posterior distribution as the approximate sufficient statistic that is conditioned on; we refer to our method as *approximate co-sufficient sampling via Bayes* (aCSS-B). We prove approximate exchangeability for aCSS-B's samples, and as a corollary, prove approximate type-I error control when the p-value (2) is constructed with its samples. Note that while our method uses Bayesian sampling as a tool, these results provide frequentist guarantees that do not rely on an assumed prior. We demonstrate its performance on a suite of examples. The main advantages of aCSS-B over prior work are that it applies considerably more broadly than previous methods and, even when previous methods apply, aCSS-B is often more powerful, and may be more straightforward to implement and more computationally efficient. We demonstrate aCSS-B's improved generality for three canonical parametric null models in which no prior methods apply: a group-sparse linear model, a low-rank matrix model, and a linear spline model (and to complement these findings, additional experiments show that aCSS-B performs well relative to existing methods on examples where prior methods do apply).

2 Background

Our objective is to construct a valid and powerful test of the composite parametric null hypothesis (1), and we reduce this problem to one of sampling (approximately) exchangeable copies $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ of the data X under H_0, as once this is accomplished Equation (2) provides a valid p-value (for any test statistic function T). Note that for such a p-value to also provide *powerful* inference, we need sufficient "diversity" among the sampled copies; for instance, if we set $\widetilde{X}^{(m)} = X$ for all m, then trivially the copies would be exchangeable, but the p-value (2) would be deterministically equal to 1 for any choice of test statistic (i.e., valid but powerless).

2.1 Co-sufficient sampling

Co-sufficient sampling (CSS) (Bartlett, 1937; Engen and Lillegård, 1997; Agresti, 1992; Stephens, 2012) identifies a sufficient statistic $S(X)$ for the null model H_0 and then samples the copies $\widetilde{X}^{(m)}$ i.i.d. (and independent of X) from the conditional distribution of $X \mid S(X)$, denoted $f_\theta(\cdot \mid S(X))$, which by definition of sufficiency is a known conditional distribution (does not depend on θ) under H_0. It is then immediate that $X, \widetilde{X}^{(1)}, ..., \widetilde{X}^{(M)}$ are conditionally i.i.d., and hence exchangeable, under H_0.

However, for this method to be powerful, $S(X)$ must be information-theoretically "compact" (in the sense that it does not contain too much information about X), otherwise conditioning on it will force all the sampled copies to be so similar to X that the p-value (2) has little or no power under the alternative (with the extreme worst case being $S(X) = X$). Unfortunately, for all but the absolute simplest models, there do not exist any such "compact" sufficient statistics; see Barber and Janson (2022) for examples where this issue arises, which include logistic regression, curved exponential families (even as simple as two independent Gaussians with equal means and unequal variances), heavy-tailed distributions, and latent variable models.

2.2 Approximate co-sufficient sampling

Approximate co-sufficient sampling (aCSS) was introduced in Barber and Janson (2022) to address the limitations of CSS testing. The key idea is to identify a statistic $S(X)$ that is *nearly* sufficient, in the sense that $f_\theta(\cdot \mid S(X))$ depends very weakly on θ, while still being quite information-theoretically compact. Then using copies $\widetilde{X}^{(m)}$ sampled from $f_{\hat{\theta}}(\cdot \mid S(X))$, for some estimator $\hat{\theta}$, results in approximate validity due to approximate sufficiency of $S(X)$ and high power due to the "compactness" of $S(X)$. In particular, Barber and Janson (2022) propose using, for both $S(X)$ and $\hat{\theta}$, the maximizer of the null log likelihood plus a random linear perturbation. They prove approximate validity under conditions resembling those for asymptotic normality of the maximum likelihood estimator (MLE), and empirically demonstrate aCSS's high power on a range of examples.

Several extensions of the aCSS method have been proposed to handle more complex settings—in particular, settings where due to high dimensionality or non-regularity of the null model, the MLE is inconsistent. These extensions enable adding regularization, via either constraints or a penalty on θ, in addition to the random perturbation to the log likelihood in defining $S(X)$ and $\hat{\theta}$ in aCSS. Concretely, Zhu and Barber (2023) develop a version of aCSS that allows for linear constraints $a_i^\top \theta \leq b_i$ (or regularization via a corresponding penalty function, $\max_i a_i^\top \theta$, which includes settings such as the Lasso). Xie and Huang (2025) extend further to allow for a group-wise penalty of the form $\sum_j \rho(\|\theta_{G_j}\|_2)$, where ρ is smooth (e.g., the group Lasso); their work also allows regularization via nonlinear constraints, $G_i(\theta) \leq b_i$, for functions G_i that are either smooth or are an ℓ_p norm.

Limitations of existing aCSS methods. The existing methods in the aCSS family—the original method proposed by Barber and Janson (2022), and the extensions to regularized versions of aCSS developed by Zhu and Barber (2023) and Xie and Huang (2025)—all suffer from certain limitations in scope. A key limitation is that these methods all require the null model to be open and convex (i.e., $\Theta \subseteq \mathbb{R}^d$ is open and convex), excluding models with any kind of structural constraint such as sparsity or low rank. While structures such as sparsity can sometimes be encoded via regularization, this is a special case and is not true for many other types of structure: for instance, none of the existing variants can encode a low-rank matrix constraint (we will consider such an example further, below). Finally, in practice even when one of these methods can be applied to a given problem, it can be computationally expensive and sensitive to tuning parameters, and can have low power.

2.3 Additional related work

There is an enormous literature on GoF testing dating back many decades, with classical examples including the χ^2, score, likelihood ratio, and Wald tests (see, e.g., GoF textbooks such as D'Agostino, 2017). GoF testing continues to be a subject of contemporary research, with recent papers considering challenging parametric or nonparametric null hypotheses and innovative tests (see, e.g., Candès et al., 2018; Berrett et al., 2020; Lundborg et al., 2022; Ramdas et al., 2022; Gangrade et al., 2023; Sen and Sen, 2014; Saha and Ramdas, 2024; Chwialkowski et al., 2016). What sets CSS and aCSS type methods, including the method proposed in this paper, apart from the rest of the GoF testing literature is their flexibility in the choice of test statistic: CSS and aCSS methods are *wrapper* methods in the sense that they can wrap around *any* test statistic, in principle enabling them to achieve high power for many different alternative distributions via unrestricted alternative-specific choices of test statistic.

One work closely related to the aCSS literature is Awan and Cai (2020), which proposes a sampling method that approximates co-sufficient sampling. However, they do not prove their sampling can be used for approximately valid testing, and it is unclear how to use their approximation error bounds to do so. A final point is that the method we propose in this paper relies heavily on ideas from Bayesian sampling such as Markov chain Monte Carlo (MCMC) (Chib and Greenberg, 1995; Casella and George, 1992) and the Laplace approximation (Shun and McCullagh, 1995), though we emphasize that our problem statement, method, and guarantees remain purely frequentist in nature.

3 Main results

In this section, we formally propose our method, aCSS-B, and prove theoretical guarantees on the excess type-I error of the resulting test. Similar to CSS, aCSS, and its extensions, aCSS-B will address the goodness of fit hypothesis testing problem as stated in Section 2.

3.1 Method

At a high level, our aim is to sample copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ that are approximately exchangeable with the data X, so that the quantity pval defined as in (2) is approximately valid as a p-value to test the null hypothesis of goodness-of-fit (i.e., the hypothesis that the distribution of X lies in the parametric family $\{f_\theta\}_{\theta\in\Theta}$).

To do so, the aCSS-B procedure operates as follows: after sampling the data X (assumed to be drawn from the density f_{θ_0}, for some unknown $\theta_0 \in \Theta$), we first define a prior density π on Θ and generate B draws from the corresponding posterior, denoted by $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$. We then estimate the distribution of $X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$ (note that we cannot compute this conditional distribution exactly, since θ_0 is unknown), and sample the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ from this estimated distribution.

Now we turn to calculating the required components of the procedure. Formally, we will assume that each density f_θ in our parametric family is a density with respect to some common base measure $\nu_{\mathcal{X}}$ on $\mathcal{X}$, and will choose a prior with density π with respect to a base measure ν_Θ on Θ. The posterior distribution of $\theta \mid X$ is then defined by the density

$$\pi(\theta \mid x) = \frac{\pi(\theta) f_\theta(x)}{\bar{f}_\pi(x)}, \tag{3}$$

which is again a density with respect to ν_Θ, where

$$\bar{f}_\pi(x) = \int_\Theta \pi(\theta) f_\theta(x) \, \mathsf{d}\nu_\Theta(\theta)$$

denotes the density of the marginal likelihood of X, under the prior $\theta \sim \pi$. (Formally, to ensure that future quantities will be well-defined, we will assume from this point on that the support of $f_\theta(x)$ is contained in the support of $\bar{f}_\pi(x)$, for every θ—that is, $\bar{f}_\pi(x)$ is positive for any value x we might observe. For instance, if f_θ has the same support for every θ, then this assumption is satisfied.)

Next, a straightforward calculation shows that the conditional distribution of $X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$ has density

$$\propto f_{\theta_0}(x) \cdot \prod_{b=1}^{B} \frac{f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)} \tag{4}$$

with respect to $\nu_{\mathcal{X}}$. Since θ_0 is unknown, however, we will replace $f_{\theta_0}(x)$ with $\bar{f}_\pi(x)$, so that the copies $\widetilde{X}^{(m)}$ are sampled according to density

$$g_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}}. \tag{5}$$

These steps describe the process of generating the copies $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ for the aCSS-B procedure; see Algorithm 1 for a complete definition of the method.

Algorithm 1: aCSS-B method

```
1: Given: prior density $\pi$ on $\Theta$, and test statistic $T : \mathcal{X} \to \mathbb{R}$.
2: Observe data $X \sim f_{\theta_0}$.
3: Generate $B$ posterior samples,
       $\widehat{\theta}_1, \dots, \widehat{\theta}_B \mid X \overset{\text{i.i.d.}}{\sim} \pi(\cdot \mid X)$.
4: Generate $M$ copies of the data,
       $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)} \mid X, \widehat{\theta}_{1:B} \overset{\text{i.i.d.}}{\sim} g_\pi(\cdot \mid \widehat{\theta}_{1:B})$,
   where the density $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is defined as
       $g_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}}$.
5: Compute the p-value
       $\text{pval} = \frac{1 + \sum_{m=1}^M \mathbf{1}\{T(\widetilde{X}^{(m)}) \geq T(X)\}}{M+1}$.
```

Before presenting our finite-sample theoretical results, we first give some intuition for why the aCSS-B procedure can be expected to provide approximately exchangeable copies of the data X.

Intuition: the Bayesian setting. Suppose that we are in a true Bayesian setting, where the unknown parameter θ_0 was in fact drawn from the prior density π. In that case, the expression (4) gives the conditional density of $X \mid (\theta_0, \widehat{\theta}_1, \dots, \widehat{\theta}_B)$. But after marginalizing over θ_0, the expression (5) is the conditional density of $X \mid (\widehat{\theta}_1, \dots, \widehat{\theta}_B)$—this is the exact, rather than approximate, conditional distribution for X, and thus drawing the copies $\widetilde{X}^{(m)}$ from this density leads to exact exchangeability, and validity of the test.

Intuition: sufficiency of the posterior. We will now consider an alternative viewpoint on the intuition behind the method, without assuming a Bayesian setting—that is, we return to the setting of a fixed θ_0. After observing B draws from the posterior (for a large B), we have approximately observed the posterior density of $\theta \mid X$ with respect to the base measure ν_Θ, which is computed in (3). The following standard result, proved for completeness in Appendix B.1, tells us that $\pi(\,\cdot \mid X)$, which we view as statistic of the data (i.e., a map from the data X to this density function), is in fact a minimal sufficient statistic.

Proposition 3.1. *Let $X \sim f_\theta$ for the parametric family $\{f_\theta : \theta \in \Theta\}$. Fix any prior on Θ, with a positive density π (relative to some base measure ν_Θ). Then the posterior density $\pi(\,\cdot \mid X)$ is a minimal sufficient statistic for X.*

In other words, by conditioning on $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$, we are conditioning on an approximation to the posterior distribution, which can be approximated arbitrarily well by the empirical measure of its samples (Vapnik and Chervonenkis, 2013). Since the posterior distribution is a sufficient statistic, this means that the copies $\widetilde{X}^{(m)}$ will be approximately exchangeable with X (for large B), since we have nearly removed the effect of the unknown parameter θ_0.

Comparison to aCSS and its extensions. The core idea of our method is similar to the aCSS method of Barber and Janson (2022) (and the regularized extensions of aCSS proposed by Zhu and Barber (2023) and Xie and Huang (2025)), since our idea is to condition on an approximately sufficient statistic for X in order to generate the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$. However, our method makes a very different choice of information to condition on: while aCSS and its regularized variants each condition on a noisy version of the MLE for θ_0 given X (or, a noisy version of the regularized MLE), we instead condition on a large number B of draws from the posterior distribution of $\theta \mid X$. This different approximate sufficient statistic only requires us to sample from the posterior distribution of $\theta \mid X$, as opposed to requiring optimization of the likelihood, which can be computationally more expensive in many problems (Ma et al., 2019). We will also see in our empirical examples below that aCSS-B offers greater flexibility and is applicable to a wider range of problems.

3.2 Theoretical Guarantees

As we have seen in Proposition 3.1, the posterior density $\pi(\,\cdot \mid X)$ is a minimal sufficient statistic for the data X; this explains why, as $B \to \infty$, we can expect that the aCSS-B method should be valid. In practice, of course, we run aCSS-B with a finite set of posterior samples—thereby yielding an approximately valid test.

In this section, we examine the behavior of the method when B is finite, to understand the role of B in the validity of the aCSS-B approach. In order to establish theoretical guarantees, we will first need a few definitions. Below, let π_0 denote any prior density on Θ, with respect to the same base measure ν_Θ. For intuition, we should think of π_0 as being concentrated near the unknown true parameter value θ_0.

Definition 3.1 (Prior concentration). *For any prior with density π_0, define*

$$\epsilon(\pi_0) = \mathrm{d}_{\mathrm{TV}}\left(f_{\theta_0}, \bar{f}_{\pi_0}\right), \tag{6}$$

where as before, $\bar{f}_{\pi_0}$ denotes the density of the marginal likelihood of X when we draw $\theta \sim \pi_0$, i.e.,

$$\bar{f}_{\pi_0}(x) = \int_\Theta f_\theta(x)\pi_0(\theta)\ \mathsf{d}\nu_\Theta(\theta).$$

Definition 3.2 (Posterior sensitivity)**.** *Given observed data $X \in \mathcal{X}$, let $\pi(\cdot \mid X)$ (respectively, $\pi_0(\cdot \mid X)$) denote the posterior distribution of θ under the prior $\theta \sim \pi$ (respectively, $\theta \sim \pi_0$). Define*

$$\Delta(\pi_0) = \mathbb{E}_{\theta_0}\left[\mathrm{d}_{\chi^2}\left(\pi_0(\cdot \mid X) \,\|\, \pi(\cdot \mid X)\right)^{1/2}\right], \tag{7}$$

where d_{χ^2} denotes the χ^2 divergence between distributions, and where the expected value is taken with respect to $X \sim f_{\theta_0}$.

Note that, if the data X carries strong information for inferring the parameter θ, we might expect that the posterior is not affected too much by the choice of prior—and therefore $\Delta(\pi_0)$ would not be too large, even for some π_0 that is strongly concentrated near θ_0 (so that $\epsilon(\pi_0) \approx 0$). We emphasize that this prior π_0 will appear in the upper bound of the theoretical guarantee, but does *not* need to be specified for running the algorithm.

With the above definitions, we state the main result, which bounds the distance to exchangeability—and therefore, the type-I error of the aCSS-B procedure.

Theorem 3.1. *After observing the data X, let $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ be sampled as in Algorithm 1, for some positive prior density π on Θ. Then, if $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$,*

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}\right) \le \inf_{\pi_0}\left\{\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}\right\},$$

where the infimum is taken over all densities π_0 on Θ with respect to base measure ν_Θ such that the support of $\bar{f}_{\pi_0}(x)$ contains the support of $f_\theta(x)$ for all θ.

In other words, we are showing that the copies $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ are approximately exchangeable with X. In particular, this implies that for any predefined test statistic $T : \mathcal{X} \to \mathbb{R}$ and rejection threshold $\alpha \in [0, 1]$, the p-value satisfies

$$\mathbb{P}\left(\mathrm{pval}_T\left(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}\right) \le \alpha\right) \le \alpha + \inf_{\pi_0}\left\{\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}\right\}.$$

The proof of this theorem is given in Appendix B.2.

3.3 Challenges: sampling from the posterior, and sampling the copies

As described in Algorithm 1, the application of aCSS-B to any hypothesis testing problem requires two sampling steps—first, drawing B independent samples $\widehat{\theta}_b$ from the posterior distribution, and second, sampling M independent copies $\widetilde{X}^{(m)}$ from $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. Both of these steps may be computationally challenging in practice, and in this section we briefly describe some solutions.

First we consider sampling the posterior draws $\widehat{\theta}_b$. In practice, it is often only possible to sample from the posterior with MCMC techniques, and sampling exactly i.i.d. draws is computationally infeasible; this is a standard challenge in Bayesian statistics. In our implementation, we employ MCMC techniques such as Gibbs sampling or Metropolis–Hastings (depending on the specific setting), along with adequate thinning to reduce autocorrelation between samples (Riabiz et al., 2022) and burn-in (Stewart and Johnson, 2009) so that the samples are more representative of the target distribution. These techniques, and the extent to which they approximate i.i.d. sampling from the posterior sampling, are well-studied in the Bayesian literature (Gelman et al., 1995; Gagniuc, 2017; Barber, 2012), so we do not explore their theoretical properties further here.

Our second challenge is the problem of sampling the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ i.i.d. from $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. This is again often infeasible to perform exactly—and unlike the problem of sampling from the posterior, this is no longer a standard challenge in Bayesian statistics, since it is not typical to condition on more than one draw from the posterior. Therefore, this requires careful treatment, to ensure that any approximations we take do not invalidate our finite-sample type-I error guarantees. In Appendix A, we provide a theoretical analysis of this challenge. Specifically, we treat two key issues that arise: first, that MCMC sampling techniques introduce dependence among the copies, and second, that the marginal density $\bar{f}_\pi(x) = \int_\Theta \pi(\theta) f(x; \theta)\,\mathrm{d}\theta$ (which appears in the denominator of $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$) may be hard to calculate exactly in some problem settings. Our results in Appendix A establish type-I error control even when we use approximate strategies for sampling the copies, and bound the increase in type-I error when we use an approximation of $\bar{f}_\pi(x)$.

4 Experiments

In this section, we examine the performance of aCSS-B in five simulated examples.[1] The first two examples are in settings where aCSS or its regularized extensions can be applied, and we will compare aCSS-B to these methods: specifically, we compare to the original aCSS method of Barber and Janson (2022) in the setting of logistic regression (Section 4.1), and to the regularized aCSS method of Zhu and Barber (2023) in the setting of a mixture of Gaussians (Section 4.2). We then follow with three examples where neither aCSS nor its regularized extensions can be applied due to the nature of the statistical problem: low-rank matrix estimation (Section 4.3), group-sparse regression (Section 4.4), and a linear spline model (Section 4.5).

Across all five examples, we implement aCSS-B with $B = 25$ posterior draws, and with $M = 300$ copies $\widetilde{X}^{(m)}$ sampled for running the test. Informally, we find aCSS-B to be easy to tune—we (successfully) use the same default value of B for all five simulations, and observe that standard, default priors work well out-of-the-box in each setting. For each example in the following subsections, details for the process of sampling the posterior draws $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$, and for sampling the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$, can be found in the corresponding subsection of Appendix C.

For each experiment, after computing a p-value, we test the null hypothesis at the level $\alpha = 0.05$. We compute the type-I error rate for the setting where the null hypothesis is true (that is, $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$), and compute power for settings where the

[1] Code for reproducing all experiments is available at `https://github.com/Ritwik-Bhaduri/aCSS-B/`.

null is false, across a range of different signal strengths. In each example we compare this power against an oracle, which is given access to θ_0—that is, the oracle can calculate the null distribution of the test statistic $T(X)$ by simply sampling from f_{θ_0}. Results are reported after averaging over 500 independent trials, with standard error bars shown in the figures.

4.1 Logistic Regression

Our first example is in a setting where aCSS can be applied, so that we can compare aCSS-B against aCSS. This example reproduces the setting of Barber and Janson (2022, Example 1, Section 4.5.2).

The model. We consider a logistic regression model in dimension $d = 5$, with $n = 100$ independent observations. Here $X_i \in \{0, 1\}$ is binary and the covariate $Z_i \in \mathbb{R}^d$ is treated as fixed; the likelihood function is

$$f(x; \theta) = \prod_{i=1}^{n} \left(\frac{e^{Z_i^\top \theta}}{1 + e^{Z_i^\top \theta}} \right)^{x_i} \cdot \left(\frac{1}{1 + e^{Z_i^\top \theta}} \right)^{1-x_i} \tag{8}$$

with parameter $\theta \in \Theta = \mathbb{R}^d$. (We can interpret $f(x; \theta)$ as a density with respect to the base measure $\nu_{\mathcal{X}}$ on $\mathcal{X} = \{0, 1\}^n$ that places mass 1 on each point $x \in \mathcal{X}$, i.e., the counting measure.) The true parameter vector is given by $\theta_0 = 0.2 \cdot \mathbf{1}_d$. Following the same setup as in the experiment presented in Barber and Janson (2022), we test a conditional independence hypothesis by considering a variable $Y_i \in \mathbb{R}$ whose conditional distribution given Z_i is independent of X_i under the null hypothesis, but is dependent under the alternative:

$$Y \mid (X, Z) \sim a\left(b(Z) + \beta_0^\top Z \cdot \mathbf{1}_{X=0} + \beta_1^\top Z \cdot \mathbf{1}_{X=1}\right),$$

where $a(t) = t + 0.5t^3$, $b(z) = 0.5 \sum_{j=1}^{5} (z_j)_+$, $\beta_0 = c \cdot \mathbf{e}_1$, and $\beta_1 = c \cdot \mathbf{e}_5$, where $\mathbf{e}_j$ is the jth basis vector and where $c \in \{0, 0.1, 0.2, \ldots, 1\}$ indicates the signal strength (with $c = 0$ corresponding to the null hypothesis). As in Barber and Janson (2022, Example 1, Section 4.5.2), we use a test statistic based on sliced inverse regression; see that paper for more details.

Can we apply existing aCSS methods? Barber and Janson (2022) apply aCSS to this problem, and we will compare aCSS-B against the exact same implementation of aCSS as used for this problem in that paper.

Choice of prior for implementing aCSS-B. We choose the prior density π as

$$\pi(\theta) = \prod_{j=1}^{d} \phi(\theta_j; 0, 1),$$

where $\phi\left(\cdot; \mu_j, \sigma^2\right)$ is the density of the normal distribution with mean μ and variance σ^2—that is, π is a standard Gaussian prior on the parameter vector $\theta \in \mathbb{R}^d$.

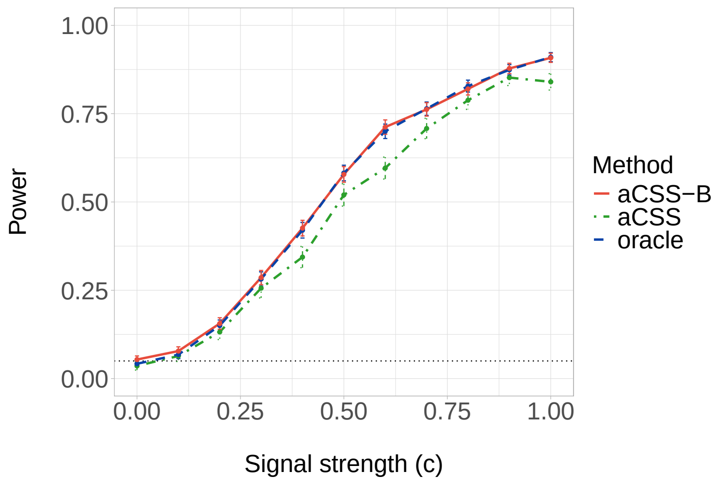

Figure 1: Power comparison between aCSS-B, aCSS, and an oracle for the logistic regression model of Section 4.1.

Results. Figure 1 compares the performance of aCSS-B to that of aCSS and the oracle. For this example, the oracle consists of sampling the copies $\widetilde{X}^{(m)}$ from the logistic model specified in (8), independently of Y. First, we see that all three methods result in a type-I error level of 5% under the null (i.e., signal strength $c = 0$). Under the alternative (signal strength $c > 0$), the methods show similar power, but we can see that aCSS has slightly lower power than the oracle, while aCSS-B appears to have power equal to that of the oracle.

4.2 Mixture of Gaussians

Next, we consider an example where a regularized extension of aCSS can be applied: the setting of a mixture of Gaussians. This example reproduces an experiment from Zhu and Barber (2023, Section 6.1), and we will now compare aCSS-B against their regularized aCSS method (which we will refer to as reg-aCSS).

The model. The null data distribution is given by sampling n i.i.d. draws from a mixture of two Gaussians, so that the likelihood function for the data $X = (X_1, \ldots, X_n)$ is given by

$$f(x;\theta) = \prod_{i=1}^{n} \left(w_1 \phi(x_i;\mu_1,\sigma_1^2) + (1-w_1)\phi(x_i;\mu_2,\sigma_2^2) \right).$$

This family of distributions is therefore parameterized by $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \in \Theta$, where

$$\Theta = (0,1) \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+ \subseteq \mathbb{R}^5.$$

Therefore, the GoF test can be interpreted as testing the null hypothesis that the underlying model is a Gaussian mixture with 2 components. We will consider an alternative that the data is instead drawn from a Gaussian mixture with > 2 components. In our simulations, we take $n = 200$ and the data is generated from the following mixture

$$p\mathcal{N}(0, 0.01) + \frac{1-p}{2}\mathcal{N}(0.4, 0.01) + \frac{1-p}{2}\mathcal{N}(-0.4, 0.01) \tag{9}$$

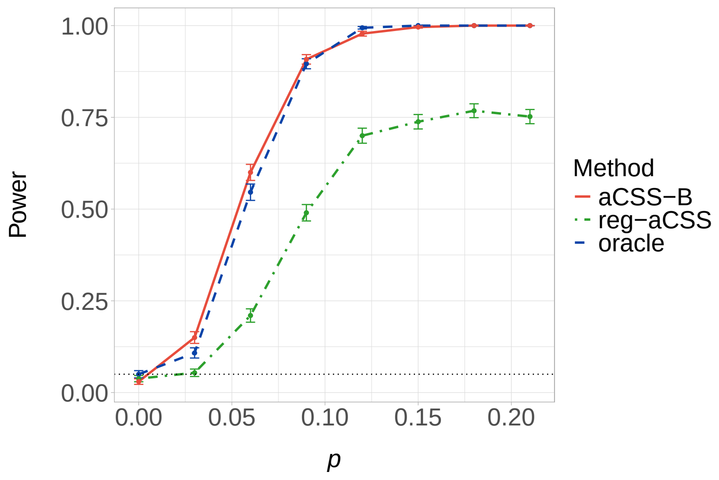

Figure 2: Power comparison between aCSS-B, reg-aCSS, and an oracle for the Gaussian mixture model of Section 4.2.

where $p = 0$ corresponds to the null hypothesis being true, while $p > 0$ corresponds to the alternative. As in Zhu and Barber (2023, Section 6.1.1), we use a test statistic based on k-means clustering with $k = 2$ versus $k = 3$; see that paper for more details.

Can we apply existing aCSS methods? For this problem, aCSS cannot be applied because the likelihood maximization problem is degenerate (since σ_1^2 or σ_2^2 can be arbitrarily close to zero, leading to a likelihood that can approach infinity). Instead, Zhu and Barber (2023) apply reg-aCSS to this problem, placing constraints that bound σ_1^2, σ_2^2 away from zero, but find low power compared to the oracle. We will compare aCSS-B against the exact same implementation of aCSS-B as used for this problem in that paper.

Choice of prior for implementing aCSS-B. We assume the following prior distributions: $w_1 \sim \text{Beta}(2, 2)$, and

$$\sigma_j^2 \sim \text{Inv-Gamma}(1, 0.5), \quad \mu_j \mid \sigma_j^2 \sim \mathcal{N}(0, \sigma_j^2)$$

independently for each $j = 1, 2$.

Results. Figure 2 compares the performance of aCSS-B to that of reg-aCSS and the oracle. For this example, the oracle consists of sampling the entries of $\widetilde{X}^{(m)}$ i.i.d. from the two-component mixture specified in (9) if we set $p = 0$. All three methods result in a type-I error level of 5% under the null (i.e., mixture weight $p = 0$). Under the alternative (mixture weight $p > 0$), aCSS-B enjoys similar power as the oracle across nearly all values of p, while reg-aCSS shows lower power.

4.3 Rank-1 matrix

Next we turn to examples that are beyond the scope of aCSS and its regularized extensions. Our first such example lies in the setting of low-rank matrix data.

The model. We assume that the data $X \in \mathbb{R}^{n \times n}$ (with $n = 10$) is generated as

$$X = A + W,$$

where A is a fixed matrix representing the underlying signal, while W is noise, with $W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.25)$. Under the null, the signal A is a rank-1 matrix. Therefore the GoF test can be represented with the family

$$f_\theta(x) = \prod_{i=1}^{n} \prod_{j=1}^{n} \phi(X_{ij} - A_{ij}; 0, 0.25),$$

parametrized by $\theta = A \in \Theta \subseteq \mathbb{R}^{n \times n}$, where Θ is the space of rank-1 matrices. Under the alternative, the rank of the underlying signal is > 1.

In order to generate the data, we define $A_0 = U_1 V_1^\top + c U_2 V_2^\top$ where $U = [U_1, U_2], V = [V_1, V_2] \in \mathbb{R}^{n \times 2}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. We vary c, with $c = 0$ corresponding to the null—note that $\text{rank}(A) = 1$ under the null and $\text{rank}(A) = 2$ under the alternative. The test statistic $T(X)$ is defined as the second largest eigenvalue of $X^\top X$.

Can we apply existing aCSS methods? In this example, the null parameter space is $\Theta = \{A \in \mathbb{R}^{n \times n} : \text{rank}(A) = 1\}$. The challenging nature of the rank-1 constraint means that none of the existing aCSS methods can be applied—specifically, Barber and Janson (2022) would require a convex and open null model space, while the regularized forms of aCSS (Zhu and Barber, 2023; Xie and Huang, 2025) allow constraints on the parameter θ but only limited types of constraints are allowed, which again cannot encompass a rank-1 restriction. (We could instead consider reparameterizing as $A = uv^\top$ and taking $\theta = (u, v)$, but the assumptions of aCSS are again violated—in this case, because the existing aCSS theory requires strong convexity of the log-likelihood at the MLE, which cannot hold under this reparametrization due to nonidentifiability.) However, aCSS-B can be applied here with a suitable choice of prior.

Choice of prior for implementing aCSS-B. To specify the prior distribution for the null, we introduce two vectors $U, V \in \mathbb{R}^n$ such that $A = UV^\top$ and assume independent multivariate standard Gaussian priors for U and V, that is, $U, V \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$.

Results. Figure 3 compares the performance of aCSS-B to that of the oracle. For this example, the oracle consists of sampling $\widetilde{X}^{(m)} = A_0 + W$ where $A_0 = U_1 V_1^\top$ and W has i.i.d standard normal entries (recall that the data is generated with mean $A = U_1 V_1^\top + c U_2 V_2^\top$, where $0 \leq c \leq 1$, so this choice of A_0 represents a rank-1 approximation to the true distribution). We can see that the aCSS-B method achieves type-I error control at level 5% under the null (i.e., when A has rank 1), as desired, and has nearly the same power as the oracle under the alternative.

4.4 Group-sparse regression

Our next example studies a group-sparse linear regression model.

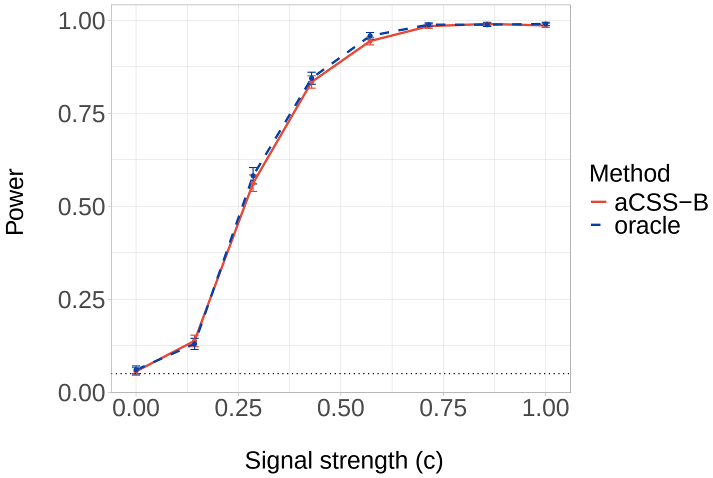

Figure 3: Power comparison between aCSS-B and an oracle for the rank-1 matrix model of Section 4.3.

The model. We consider the following model for data $X \in \mathbb{R}^n$, given covariates $Z \in \mathbb{R}^{n\times d}$ (which we treat as fixed):

$$X = Z\beta + \epsilon \text{ where } \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{10}$$

where we chose $n = 100$ and $d = 50$. Given the partition $\{1,\dots,d\} = I_1 \cup \dots \cup I_{10}$ into $G = 10$ groups of equal size, $|I_g| = 5$, our null hypothesis is that only one group is "active" in the regression: that is, the active set $A = \{g : \beta_{I_g} \neq (0,\dots,0)\}$ has size $|A| = 1$ while under the alternative $|A| > 1$.

To generate the data, we draw the entries of Z independently from $\mathcal{N}(0,1)$ and choose two group indices $g_1 \neq g_2$ uniformly at random from the 10 groups. The coefficients are generated as follows:

$$\beta_{I_{g_1}} \sim \mathcal{N}(\mathbf{0}_{|I_{g_1}|}, \mathbf{I}_{|I_{g_1}|}),\ \beta_{I_{g_2}} \sim \mathcal{N}(\mathbf{0}_{|I_{g_2}|}, c^2 \cdot \mathbf{I}_{|I_{g_2}|}),\quad \beta_j = 0\ \forall\ j \in [d]\backslash(I_{g_1} \cup I_{g_2}). \tag{11}$$

For the null, we consider $c = 0$ so that there is only one active group (namely, g_1), and for the alternative we take $c \in (0,1)$ so that there are two active groups (g_1 and g_2). We refer to c as the signal strength.

To define the test statistic, we first compute an estimate $\widehat{\beta}$ of the regression coefficients by fitting 5-fold cross-validated group LASSO from the R package `gglasso` (Wu and Lange, 2020) on the data. The test statistic is then defined as

$$T(X) := \frac{\sum_{g\neq\widehat{g}} ||\widehat{\beta}_{I_g}||_\infty}{||\widehat{\beta}_{I_{\widehat{g}}}||_\infty} \text{ where } \widehat{g} = \underset{g\in[G]}{\arg\max}\, ||\widehat{\beta}_{I_g}||_\infty,$$

which measures the sum of the largest estimated coefficients *outside* of the top selected group $\widehat{g}$ relative to the largest one within $\widehat{g}$.

Can we apply existing aCSS methods? In this example, the null parameter space is

$$\Theta = \{(\beta_1,\dots,\beta_d) \in \mathbb{R}^d : \beta_{I_g} = \mathbf{0}_{d_g} \text{ for all } g \neq g^\star,\ \text{for some } g^\star \in \{1,\dots,G\}\}.$$

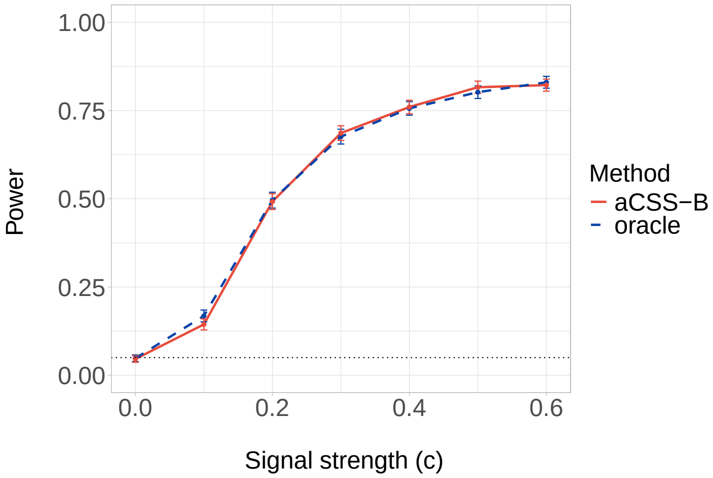

Figure 4: Power comparison between aCSS-B and an oracle for the group sparse model of Section 4.4.

As for the rank-1 constraint in the previous example, aCSS and its regularized extensions cannot be applied to this problem because of the challenging nature of the group-sparsity constraint.

Choice of prior for implementing aCSS-B. For the prior π, we consider a discrete uniform distribution for the active group and choose a standard Gaussian prior for the coefficients of the active group, while the rest of the coefficients are set to zero:

$$\begin{aligned}
&g^\star \sim \text{Unif}(\{1,\ldots,\text{G}\}),\\
&\beta_{I_{g^\star}} \sim \mathcal{N}(\mathbf{0}_{|I_{g^\star}|}, \mathbf{I}_{|I_{g^\star}|}),\\
&\beta_{I_g} = \mathbf{0}_{|I_g|} \;\forall\; g \neq g^\star.
\end{aligned}$$

Results. Figure 4 compares the performance of aCSS-B to that of the oracle. For this example, the oracle consists of sampling $\widetilde{X}^{(m)}$ from a linear model as in Equation (10) where the coefficient vector β defined in (11) is redefined to have a single active group by setting $\beta_{I_{g_2}}$ to be zero (while keeping the original coefficients in the first active group, $\beta_{I_{g_1}}$). We can see that the aCSS-B method achieves type-I error control at level 5% under the null (i.e., when the coefficient vector indeed has only one nonzero group), and the power is essentially the same as the oracle under the alternative.

4.5 Linear spline regression model

Our final example is in a nonlinear regression setting, where X follows a linear spline model given covariates Z.

The model. Consider a linear spline model with k knots $t_1 < \cdots < t_k \in \mathbb{R}$, with $n = 50$ observations,

$$X_i = \mu_i + \epsilon_i \text{ for } \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.25), \tag{12}$$

where

$$\mu_i = \beta_{0j} + \beta_{1j} Z_i \text{ if } t_{j-1} \leq Z_i < t_j.$$

(Here for convenience we define $t_0 = -\infty$ and $t_{k+1} = \infty$; as in previous examples the covariates $Z_i \in \mathbb{R}$ are treated as fixed.) The parameters $\beta_{ij} \in \mathbb{R}$ are constrained so that the mean function is continuous in $\mathbb{R}$. The null hypothesis corresponds to the case where we have exactly $k = 1$ knot, while under the alternative we have $k = 2$.

To generate data from this distribution, we generate $Z_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(-5, 5)$ and choose two knots as $t_1 = -1.67$ and $t_2 = 1.67$. The coefficients are generated as

$$\beta_{01} = 1, \beta_{11} = -1, \beta_{21} = 1, \beta_{31} = 1 - c \tag{13}$$

and the other intercepts β_{02}, β_{03} are chosen so that the mean function is continuous. We consider a sequence of equally spaced values for c from 0 to 1.8—note that $c = 0$ corresponds to the null, since the slopes of the second and third segment are the same and thus effectively we only have $k = 1$ knot. Under the alternative $c > 0$, there are $k = 2$ knots, with larger values of c denoting further deviation from the null.

The test statistic $T(X)$ is the residual sum of squares from a linear spline model with one knot, which we fit using the `segmented` (Muggeo, 2023) package in `R`.

Can we apply existing aCSS methods? In this example, the null parameter space is

$$\Theta = \Big\{ (\beta_{01}, \beta_{11}, \ldots, \beta_{0,k+1}, \beta_{1,k+1}, t_1, \ldots, t_k) \in \mathbb{R}^{3k+2} : \\ \beta_{0j} + \beta_{1j} t_j = \beta_{0(j+1)} + \beta_{1(j+1)} t_j \text{ for all } j = 1, \ldots, k, \; t_1 < \cdots < t_k \Big\}.$$

The above constraints, which stem from the continuity of the mean function at the knots in the linear spline model, cannot be accommodated by the existing aCSS methods. However, aCSS-B can still be applied, through a carefully constructed prior as we will see below.

Choice of prior for implementing aCSS-B. While we can choose priors which respect the constraints in this problem, sampling from the resulting posterior distribution would be complicated. However, we can avoid this problem by using the following reparameterization:

$$X = \gamma_0 + \gamma_1 Z + \sum_{j=1}^{k} \gamma_{1+j} b_j(Z) + \epsilon, \tag{14}$$

where $b_j(Z) = (Z - t_j)_+$ is applied elementwise to $Z = (Z_1, \ldots, Z_n)$, for all $j = 1, \ldots, k$. In this reparameterization, $\gamma = (\gamma_0, \ldots, \gamma_{k+1}) \in \mathbb{R}^{k+2}$ is unconstrained, and the knots $t_1, \ldots, t_k$ do not need to be ordered. Therefore, the parameter space becomes $\Theta = \mathbb{R}^{2k+2}$. (We note, however, that existing aCSS methods nonetheless cannot be applied, even with this reparametrization—this is because the log-likelihood is no longer differentiable with respect to the parameters t_j.)

On these unconstrained parameters, we choose the standard Gaussian priors:

$$\gamma_0, \gamma_1, \gamma_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \\ t_1 \sim \mathcal{N}(0, 1),$$

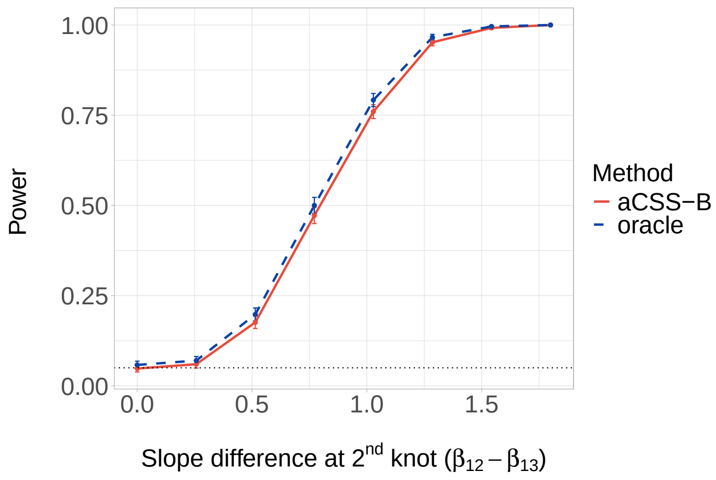

Figure 5: Power comparison between aCSS-B and an oracle for the linear spline model of Section 4.5.

for the null model with $k = 1$ knots.

Results. Figure 5 illustrates the performance aCSS-B as compared to the oracle. For this example, the oracle consists of sampling $\widetilde{X}^{(m)}$'s from the model given by Equations (12) and (13) with $c = 0$. We see that the aCSS-B method achieves type-I error control at level 5% under the null, and the power is nearly the same as the oracle under the alternative.

Acknowledgements

R.B. and L.J. were partially supported by the National Science Foundation via grant DMS-2045981. A.B. and R.F.B. were partially supported by the National Science Foundation via grant DMS-2023109. R.F.B. was partially supported by the Office of Naval Research via grant N00014-24-1-2544.

References

Acharya, A. and Kashyap, V. L. (2024). Spectral fit residuals as an indicator to increase model complexity. *Research Notes of the AAS*, 8(1):1.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical science*, 7(1):131–153.

Awan, J. and Cai, Z. (2020). Approximate co-sufficient sampling for goodness-of-fit tests and synthetic data. *arXiv preprint arXiv:2006.02397*.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barber, R. F., Drton, M., and Tan, K. M. (2016). Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, pages 15–36. Springer.

Barber, R. F. and Janson, L. (2022). Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *The Annals of Statistics*, 50(5):2514–2544.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):175–197.

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, 76(4):633–642.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:‘model-x’knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Thomson Learning.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the american statistical association*, 90(432):1313–1321.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *The american statistician*, 49(4):327–335.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American statistical association*, 96(453):270–281.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR.

Cowell, F., Flachaire, E., and Bandyopadhyay, S. (2009). Goodness-of-fit: An economic approach. *Economics Series Working Papers 444, University of Oxford, Department of Economics.*

D’Agostino, R. B. (2017). *Goodness-of-fit-techniques.* Routledge.

Engen, S. and Lillegård, M. (1997). Stochastic simulations conditioned on sufficient statistics. *Biometrika*, 84(1):235–240.

Fraser, D. (1963). On sufficiency and the exponential family. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(1):115–123.

Frezza, M. (2014). Goodness of fit assessment for a fractal model of stock markets. *Chaos, Solitons & Fractals*, 66:41–50.

Gagniuc, P. A. (2017). *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons.

Gangrade, A., Rinaldo, A., and Ramdas, A. (2023). A sequential test for log-concavity. *arXiv preprint arXiv:2301.03542*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

Guo, S. W. and Thompson, E. A. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, pages 361–372.

Lundborg, A. R., Shah, R. D., and Peters, J. (2022). Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850.

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. (2019). Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860.

Muggeo, V. M. (2023). *segmented: Regression Models with Break-Points / Change-Points Estimation*. R package version 1.6-4.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1):3–26.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. `https://artowen.su.domains/mc/`.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109.

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

Saha, A. and Ramdas, A. (2024). Robust likelihood ratio tests for composite nulls and alternatives. *arXiv preprint arXiv:2408.14015*.

Sen, A. and Sen, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika*, 101(4):927–942.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4):749–760.

Stephens, M. A. (2012). Goodness-of-fit and sufficiency: Exact and approximate tests. *Methodology and Computing in Applied Probability*, 14:785–791.

Stewart, L. T. and Johnson, J. D. (2009). Determining optimum burn-in and replacement times using Bayesian decision theory. *IEEE Transactions on Reliability*, 21(3):170–175.

Vapnik, V. N. and Chervonenkis, A. Y. (2013). *On the Uniform Convergence of the Frequencies of Occurrence of Events to Their Probabilities*, pages 7–12. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wu, C. and Lange, K. (2020). *gglasso: Group Lasso Regularization.* R package version 1.5.

Xie, J. and Huang, D. (2025). A generalized framework for approximate co-sufficient sampling. *arXiv preprint arXiv:2506.12334*.

Zhu, W. and Barber, R. F. (2023). Approximate co-sufficient sampling with regularization. *arXiv preprint arXiv:2309.08063*.

A Theoretical guarantees for sampling the copies

In this section, we return to the question raised in Section 3.3: since sampling $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ i.i.d. from the distribution $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is often computationally infeasible, can we develop approximations to this sampling step without losing finite-sample type-I error control?

Specifically, we will consider two aspects of this question: the challenge of constructing independent draws, and the challenge of sampling from the correct distribution when the marginal density $\bar{f}_\pi(x)$ is difficult to compute.

A.1 Dependence among the copies

In Algorithm 1, we assume that, after observing the posterior draws $\widehat{\theta}_1, \dots, \widehat{\theta}_B$, the copies $\widetilde{X}^{(m)}$ are then sampled i.i.d. from the distribution $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. In practice, sampling from a complex distribution is often carried out via MCMC based strategies, and the resulting samples are only approximately i.i.d.—specifically, samples obtained by running a Markov chain will have dependence. While it is common in many sampling problems to assume mixing conditions for the Markov chain, in order to ensure that the resulting samples are approximately i.i.d., here we will use a different approach in order to ensure finite-sample validity.

Formally, define the joint sampling distribution of the copies, conditional on the data X and the posterior draws $\widehat{\theta}_1, \dots, \widehat{\theta}_B$, as

$$(\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \mid X, \widehat{\theta}_{1:B} \sim \widetilde{Q}^M_\pi(\,\cdot \mid X, \widehat{\theta}_{1:B}).$$

We will assume the following property of this joint distribution:

$$\begin{aligned}&\text{For any } \theta_1, \dots, \theta_B \in \Theta,\\ &\qquad \text{if } X \sim g_\pi(\,\cdot \mid \theta_{1:B}) \text{ and } (\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \mid X \sim \widetilde{Q}^M_\pi(\,\cdot \mid X, \theta_{1:B}),\\ &\qquad\qquad \text{then the random vector } (X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \text{ is exchangeable.}\end{aligned} \tag{15}$$

For example, the i.i.d. sampling strategy of Algorithm 1 satisfies this condition—we can define $\widetilde{Q}^M_\pi(\,\cdot \mid X, \theta_{1:B})$ as follows:

$$\widetilde{Q}^M_\pi(\,\cdot \mid X, \theta_{1:B}) \text{ is the distribution with joint density } g_\pi(x_1 \mid \widehat{\theta}_{1:B}) \cdot \ldots \cdot g_\pi(x_M \mid \widehat{\theta}_{1:B}). \tag{16}$$

However, even in settings where i.i.d. sampling is infeasible, we can nonetheless use MCMC strategies—e.g., the permuted serial sampler (Besag and Clifford, 1989)—to draw the copies from a joint distribution $\widetilde{Q}^M_\pi(\,\cdot \mid X, \theta_{1:B})$ that exactly satisfies this condition (see Barber and Janson (2022, Section 2.2.3) for more details on how to implement this in the setting of aCSS).

The aCSS-B method, with more general sampling strategy, is presented in Algorithm 2. Of course, the original version of the method, given in Algorithm 1, is simply a special case obtained by choosing the i.i.d. sampling strategy as in (16).

Algorithm 2: aCSS-B method (general case)

```
Given: prior density $\pi$ on $\Theta$, and test statistic $T : \mathcal{X} \to \mathbb{R}$.
Observe data $X \sim f_{\theta_0}$.
Generate $B$ posterior samples,
    $\widehat{\theta}_1, \dots, \widehat{\theta}_B \mid X \overset{\text{i.i.d.}}{\sim} \pi(\,\cdot \mid X)$.
Generate $M$ copies of the data,
    $\left(\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}\right) \mid X, \widehat{\theta}_{1:B} \sim \widetilde{Q}^M_\pi\left(\,\cdot \mid X, \widehat{\theta}_{1:B}\right)$,
    where the distribution $\widetilde{Q}^M_\pi(\,\cdot \mid X, \widehat{\theta}_{1:B})$ is chosen to satisfy (15).
Compute the p-value
    $\text{pval} = \dfrac{1 + \sum_{m=1}^M \mathbf{1}\{T(\widetilde{X}^{(m)}) \geq T(X)\}}{M+1}$.
```

Our next result, proved in Appendix B.3, is a generalization of Theorem 3.1, showing that as long as the copies are sampled from a distribution satisfying (15), the same bound on type-I error still holds.

Theorem A.1. *After observing the data X, let $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$ be sampled as in Algorithm 2, for some positive prior density π on Θ. Then, if $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$,*

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}\right) \leq \inf_{\pi_0} \left\{ \epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}} \right\},$$

where the infimum is taken over all densities π_0 on Θ with respect to base measure ν_Θ such that the support of $\bar{f}_{\pi_0}(x)$ contains the support of $f_\theta(x)$ for all θ.

A.2 Estimating the distribution for the copies

Thus far, we have considered the setting where the density $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is computable exactly, even though drawing copies independently from this distribution may not be feasible. Next, we turn to the problem of computing this distribution itself. Recall that the density $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is defined as

$$\propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}},$$

where

$$\bar{f}_\pi(x) = \int_\Theta f_\theta(x)\pi(\theta) \, \mathsf{d}\nu_\Theta(\theta)$$

is the marginal density of X, integrated over the prior $\theta \sim \pi$. In many settings, evaluating the likelihood $f_\theta(x)$ at a single value θ is straightforward, but calculating the marginal likelihood can only be carried out approximately: the integral defining the marginal likelihood is rarely available in closed form except for simple conjugate models, and is well known to be intractable in practice in many settings (Chib, 1995; Gelman et al., 1995). A variety of numerical methods have been proposed to estimate $\bar{f}_\pi(x)$, such as the Laplace approximation (Barber et al., 2016), importance sampling (Newton and Raftery, 1994), Chib's MCMC estimator (Chib, 1995; Chib and Jeliazkov, 2001), bridge sampling, and path sampling (Meng and Wong, 1996; Gelman and Meng, 1998). These techniques can be leveraged to provide an estimated marginal density, $\widehat{f}_\pi(x)$, that we can then use in place of $\bar{f}_\pi(x)$ in the aCSS-B method. (Note that $\widehat{f}_\pi(x)$ is treated as fixed—that is, we assume this estimate of the marginal distribution is computed independently of the data used in the aCSS-B procedure.)

Specifically, we can run Algorithm 1 with density $\widehat{g}_\pi(\cdot \mid \widehat{\theta}_{1:B})$, defined as

$$\widehat{g}_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x)}{\widehat{f}_\pi(x)^{B-1}},$$

in place of $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. (As for $\bar{f}_\pi$ earlier, now we will need to assume positivity of $\widehat{f}_\pi(x)$ in order for this to be well-defined—that is, the support of $f_\theta(x)$ is contained in the support of $\widehat{f}_\pi(x)$, for every θ.) Or, more generally, if sampling i.i.d. copies is not feasible then we can run Algorithm 2 with a joint distribution $\widehat{Q}_\pi^M(\cdot \mid \widehat{\theta}_{1:B})$, satisfying

$$\begin{aligned}&\text{For any } \theta_1, \dots, \theta_B \in \Theta,\\ &\qquad \text{if } X \sim \widehat{g}_\pi(\cdot \mid \theta_{1:B}) \text{ and } (\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \mid X \sim \widehat{Q}_\pi^M(\cdot \mid X, \theta_{1:B}),\\ &\qquad\qquad \text{then the random vector } (X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \text{ is exchangeable.}\end{aligned} \tag{17}$$

Of course, this modification comes at the potential cost of additional type-I error inflation, since we are now sampling the copies from an approximation to the original distribution. The following theorem, proved in Appendix B.4, provides a bound on the type-I error inflation of aCSS-B, when we use an estimate $\widehat{f}_\pi$ in place of the exact marginal likelihood $\bar{f}_\pi$.

Theorem A.2. *After observing the data X, let $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ be sampled via Algorithm 2 except implemented with a joint distribution $\widehat{Q}^M_\pi(\,\cdot \mid X, \theta_{1:B})$ satisfying (17), for some positive prior density π on Θ. Assume that*

$$\mathbb{E}_{\theta_0}\left[\left|\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right|\right] \leq \delta_B.$$

Then, if $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$,

$$\mathrm{d}_{\text{exch}}\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \leq \inf_{\pi_0}\left\{\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}\right\} + \delta_B,$$

where the infimum is taken over all densities π_0 on Θ with respect to base measure ν_Θ such that the support of $\bar{f}_{\pi_0}(x)$ contains the support of $f_\theta(x)$ for all θ.

In other words, the additional term in the bound, δ_B, is small whenever the estimated marginal density $\widehat{f}_\pi(x)$ is a sufficiently accurate approximation to $\bar{f}_\pi(x)$. (Note that the dependence on B implies that we require a more accurate approximation $\widehat{f}_\pi$ when B is large.)

B Proofs

B.1 Proof of Proposition 3.1

Since π is assumed to be a positive density, note that $f_\theta(x) = \frac{\pi(\theta|x)}{\pi(\theta)} \cdot \bar{f}_\pi(x)$, where on the right-hand side, the first term depends on x only via the statistic $\pi(\,\cdot \mid x)$, and the second term depends on x only and not on θ. By the Fisher–Neyman factorization theorem (Casella and Berger, 2002, pg. 276), this implies that $\pi(\,\cdot \mid X)$ is a sufficient statistic of X. On the other hand, it is well known that the likelihood function $\theta \mapsto f_\theta(X)$ is a minimal sufficient statistic (see, e.g., Fraser (1963)). Since the posterior density function $\theta \mapsto \pi(\theta \mid X)$ depends on X only through the likelihood function $\theta \mapsto f_\theta(X)$, this means that the posterior is minimal sufficient.

B.2 Proof of Theorem 3.1

As explained in Appendix A.1, by defining $\widetilde{Q}^M_\pi(\,\cdot \mid X, \theta_{1:B})$ as in (16), the result of Theorem 3.1 is simply a special case of Theorem A.1—see Appendix B.3 for the proof of this more general theorem.

B.3 Proof of Theorem A.1

Let P_0 denote the joint distribution of $(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)$, which is given by

$$\begin{cases} X \sim f_{\theta_0}, \\ \{\widehat{\theta}_b\}_{b=1,\ldots,B} \mid X \overset{\text{iid}}{\sim} \pi(\,\cdot \mid X). \end{cases} \tag{18}$$

This distribution has the following joint density at $(x, \theta_1, \ldots, \theta_B)$:

$$f_{\theta_0}(x) \cdot \prod_{b=1}^{B} \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)}$$

(with respect to the base measure $\nu_\mathcal{X} \times \nu_\Theta \times \cdots \times \nu_\Theta$). The aCSS-B procedure can be equivalently characterized as

$$\begin{cases} (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim P_0, \\ (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim \widetilde{Q}^M_\pi(\,\cdot \mid X, \widehat{\theta}_{1:B}). \end{cases} \tag{19}$$

Our goal, then, is to bound the distance to exchangeability induced by this distribution.

Step 1: Defining P_1 to bound distance to exchangeability. We next define a joint distribution P_1, with joint density at $(x, \theta_1, \ldots, \theta_B)$ given by:

$$\frac{1}{B}\sum_{b=1}^{B} \frac{\pi_0(\theta_b)}{\pi(\theta_b)} \cdot \bar{f}_\pi(x) \cdot \prod_{b=1}^{B} \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)}.$$

(We can verify that this expression integrates to 1 by definition of $\bar{f}_\pi$, and therefore defines a valid joint density.) Note that, by construction, under the joint distribution P_1 we have

$$X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim g_\pi(\,\cdot \mid \widehat{\theta}_{1:B}),$$

where we recall that $g_\pi(\,\cdot \mid \widehat{\theta}_{1:B})$ is the distribution with density $\propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}}$. Therefore, if we consider the sampling distribution

$$\begin{cases} (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim P_1, \\ (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim \widetilde{Q}^M_\pi(\,\cdot \mid X, \widehat{\theta}_{1:B}), \end{cases} \tag{20}$$

then by our assumption (15) on $\widetilde{Q}^M_\pi(\,\cdot \mid X, \widehat{\theta}_{1:B})$, it holds that $(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ are exchangeable conditional on $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$, and consequently are also exchangeable after marginalizing over $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$.

Consequently, we can see that $\mathrm{d}_{\mathrm{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ is bounded by the total variation distance between the sampling distributions given in (19) and in (20), which can be simplified to

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \le \mathrm{d}_{\mathrm{TV}}(P_0, P_1),$$

since the distributions (19) and (20) differ only in their first line. From this point on, then, we only need to bound this last total variation distance.

Step 2: Defining $P_{0.5}$. First, we define an intermediate distribution, $P_{0.5}$, with joint density at $(x, \theta_1, \ldots, \theta_B)$ given by

$$\frac{1}{B}\sum_{b=1}^{B} \frac{\pi_0(\theta_b)/\bar{f}_{\pi_0}(x)}{\pi(\theta_b)/\bar{f}_{\pi}(x)} \cdot f_{\theta_0}(x) \cdot \prod_{b=1}^{B} \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_{\pi}(x)},$$

where

$$\bar{f}_{\pi_0}(x) = \int_{\Theta} f_{\theta}(x)\pi_0(\theta) \ \mathsf{d}\nu_{\Theta}(\theta)$$

is the marginal likelihood corresponding to the prior π_0. We can then write

$$\mathrm{d}_{\mathrm{TV}}(P_0, P_1) \leq \mathrm{d}_{\mathrm{TV}}(P_0, P_{0.5}) + \mathrm{d}_{\mathrm{TV}}(P_{0.5}, P_1).$$

We now bound the remaining two terms separately.

Step 3: Bounding $\mathrm{d}_{\mathrm{TV}}(P_{0.5}, P_1)$. Note that we can express the distribution $P_{0.5}$ as a mixture, $P_{0.5} = \frac{1}{B}\sum_{b=1}^{B} P_{0.5,b}$, where $P_{0.5,b}$ has joint density

$$f_{\theta_0}(x) \cdot \frac{f_{\theta_b}(x)\pi_0(\theta_b)}{\bar{f}_{\pi_0}(x)} \cdot \prod_{b' \neq b} \frac{f_{\theta_{b'}}(x)\pi(\theta_{b'})}{\bar{f}_{\pi}(x)},$$

which can be interpreted as follows:

$$\begin{cases} X \sim f_{\theta_0}, \\ \widehat{\theta}_b \mid X \ \sim \ \pi_0(\,\cdot \mid X), \\ \{\widehat{\theta}_{b'}\}_{b' \neq b} \mid X, \widehat{\theta}_b \ \overset{\text{iid}}{\sim} \ \pi(\,\cdot \mid X). \end{cases} \tag{21}$$

Similarly, we can express the distribution P_1 as a mixture, $P_1 = \frac{1}{B}\sum_{b=1}^{B} P_{1,b}$, where $P_{1,b}$ has joint density

$$\pi_0(\theta_b) \cdot f_{\theta_b}(x) \cdot \prod_{b' \neq b} \frac{f_{\theta_{b'}}(x)\pi(\theta_{b'})}{\bar{f}_{\pi}(x)},$$

which can be interpreted as follows:

$$\begin{cases} \widehat{\theta}_b \sim \pi_0, \\ X \mid \widehat{\theta}_b \ \sim \ f_{\widehat{\theta}_b}, \\ \{\widehat{\theta}_{b'}\}_{b' \neq b} \mid X, \widehat{\theta}_b \ \overset{\text{iid}}{\sim} \ \pi(\,\cdot \mid X). \end{cases}$$

Equivalently, by swapping the order in which we draw $\widehat{\theta}_b$ and X, this is the same as

$$\begin{cases} X \sim \bar{f}_{\pi_0}, \\ \widehat{\theta}_b \mid X \sim \pi_0(\,\cdot \mid X), \\ \{\widehat{\theta}_{b'}\}_{b' \neq b} \mid X, \widehat{\theta}_b \ \overset{\text{iid}}{\sim} \ \pi(\,\cdot \mid X). \end{cases} \tag{22}$$

By comparing (21) and (22), we can see that

$$\mathrm{d}_{\mathrm{TV}}(P_{0.5,b}, P_{1,b}) = \mathrm{d}_{\mathrm{TV}}(f_{\theta_0}, \bar{f}_{\pi_0}),$$

and moreover, $\mathrm{d}_{\mathrm{TV}}(f_{\theta_0}, \bar{f}_{\pi_0}) \leq \epsilon(\pi_0)$ by our definition of $\epsilon(\pi_0)$ (see Definition 3.1). Then

$$\mathrm{d}_{\mathrm{TV}}(P_{0.5}, P_1) = \mathrm{d}_{\mathrm{TV}}\left(\frac{1}{B}\sum_{b=1}^{B} P_{0.5,b}, \frac{1}{B}\sum_{b=1}^{B} P_{1,b}\right) \leq \frac{1}{B}\sum_{b=1}^{B} \mathrm{d}_{\mathrm{TV}}(P_{0.5,b}, P_{1,b}) \leq \epsilon(\pi_0).$$

Step 4: Bounding $\mathrm{d}_{\mathrm{TV}}(P_0, P_{0.5})$. By comparing (18) with (21), we can see that the marginal distribution of X is the same for both (i.e., $X \sim f_{\theta_0}$), while the conditional distribution of $(\widehat{\theta}_1, \dots, \widehat{\theta}_B) \mid X$ is different: under P_0, these B values are drawn i.i.d. from the posterior $\pi(\,\cdot \mid X)$, while under $P_{0.5}$, we instead draw one $\widehat{\theta}_b$ (with b selected uniformly at random) from $\pi_0(\,\cdot \mid X)$, and the remaining values $\{\widehat{\theta}_{b'}\}_{b' \neq b}$ are drawn i.i.d. from $\pi(\,\cdot \mid X)$. We will now need the following lemma:

Lemma B.1. *Let P and Q be probability measures on a measurable space $(\Omega, \mathcal{F})$ with $Q \ll P$ and let $B > 1$ be an integer. Let $Z = (Z_1, \dots, Z_B)$, where $Z_1, \dots, Z_B \overset{iid}{\sim} P$. Define a corrupted vector $Z' = (Z'_1, \dots, Z'_B)$, where we draw a random $K \sim$ Uniform $\{1, \dots, B\}$, and sample $Z'_K \sim Q$, and set $Z'_k = Z_k$ for all $k \neq K$. Then*

$$\mathrm{d}_{\mathrm{TV}}\left(Z, Z'\right) \leq \frac{1}{2}\sqrt{\frac{\mathrm{d}_{\chi^2}(Q \| P)}{B}}.$$

Applying this lemma, we then have

$$\mathrm{d}_{\mathrm{TV}}\left((P_0)_{\widehat{\theta}_{1:B}|X}, (P_{0.5})_{\widehat{\theta}_{1:B}|X}\right) \leq \frac{1}{2}\sqrt{\frac{\mathrm{d}_{\chi^2}\big(\pi_0(\,\cdot \mid X) \,\|\, \pi(\,\cdot \mid X)\big)}{B}},$$

where $(P_0)_{\widehat{\theta}_{1:B}|X}$ denotes the conditional distribution of $(\widehat{\theta}_1, \dots, \widehat{\theta}_B) \mid X$ under the joint distribution P_0, and same for $P_{0.5}$. Therefore,

$$\mathrm{d}_{\mathrm{TV}}(P_0, P_{0.5}) = \mathbb{E}\left[\mathrm{d}_{\mathrm{TV}}\left((P_0)_{\widehat{\theta}_{1:B}|X}, (P_{0.5})_{\widehat{\theta}_{1:B}|X}\right)\right] \leq \frac{\Delta(\pi_0)}{2\sqrt{B}},$$

by definition of $\Delta(\pi_0)$ (see Definition 3.2).

B.3.1 Proof of Lemma B.1

By construction, the distribution of Z' can be written as

$$\widetilde{Q} \;=\; \frac{1}{B}\sum_{b=1}^{B}\left(P^{b-1} \otimes Q \otimes P^{B-b}\right) \;\ll\; P^B.$$

If we denote $f = \frac{\mathsf{d}Q}{\mathsf{d}P}$, then for each b,

$$\frac{\mathsf{d}\left(P^{b-1} \times Q \times P^{B-b}\right)}{\mathsf{d}P^B}(x_1, \dots, x_B) = f(x_b),$$

for P^B-almost-every $(x_1, \dots, x_B)$. Hence we can calculate the Radon–Nikodym derivative

$$\frac{\mathsf{d}\widetilde{Q}}{\mathsf{d}P^B}(x_1, \dots, x_B) = \frac{1}{B}\sum_{b=1}^{B} f(x_b),$$

for P^B-almost-every $(x_1, \dots, x_B)$. By definition of total variation,

$$\mathrm{d}_{\mathrm{TV}}(P^B, \widetilde{Q}) = \frac{1}{2}\int_{\Omega^B}\left|1 - \frac{1}{B}\sum_{b=1}^{B} f(x_b)\right| \mathsf{d}P^B(x).$$

Now define $\Delta(x) = 1 - f(x)$. Note that

$$\int_\Omega \Delta(x) \, \mathsf{d}P(x) = \int_\Omega (1 - f(x)) \, \mathsf{d}P(x) = 0$$

and

$$\int_\Omega \Delta(x)^2 \, \mathsf{d}P(x) = \int_\Omega (1 - f(x))^2 \, \mathsf{d}P(x) = \mathrm{d}_{\chi^2}(Q\|P),$$

by definition of f. Then

$$\begin{aligned}\mathrm{d}_{\mathrm{TV}}(P^B, \widetilde{Q}) = \frac{1}{2B} \int_{\Omega^B} \left| \sum_{b=1}^B \Delta(x_b) \right| \mathsf{d}P^B(x) &\leq \frac{1}{2B} \left[\int_{\Omega^B} \left(\sum_{b=1}^B \Delta(x_b) \right)^2 \mathsf{d}P^B(x) \right]^{1/2} \\ &= \frac{1}{2B} \left[\sum_{b=1}^B \sum_{b'=1}^B \int_{\Omega^B} \Delta(x_b)\Delta(x_{b'}) \, \mathsf{d}P^B(x) \right]^{1/2},\end{aligned}$$

by Cauchy–Schwarz. But for $b \neq b'$, we have

$$\int_{\Omega^B} \Delta(x_b)\Delta(x_{b'}) \, \mathsf{d}P^B(x) = \left(\int_\Omega \Delta(x_b) \, \mathsf{d}P(x_b) \right) \cdot \left(\int_\Omega \Delta(x_{b'}) \, \mathsf{d}P(x_{b'}) \right) = 0,$$

while for the case $b = b'$ we have

$$\int_{\Omega^B} \Delta(x_b)^2 \, \mathsf{d}P^B(x) = \mathrm{d}_{\chi^2}(Q\|P).$$

Therefore,

$$\mathrm{d}_{\mathrm{TV}}(P^B, \widetilde{Q}) \leq \frac{1}{2B} \left[\sum_{b=1}^B \sum_{b'=1}^B \mathrm{d}_{\chi^2}(Q\|P) \cdot \mathbf{1}_{b=b'} \right]^{1/2} = \frac{\sqrt{B\mathrm{d}_{\chi^2}(Q\|P)}}{2B},$$

which completes the proof.

B.4 Proof of Theorem A.2

Recall from the proof of Theorem A.1, the joint distribution P_1 on $(X, \widehat{\theta}_1, \dots, \widehat{\theta}_B)$, which we defined to have the joint density

$$\frac{1}{B} \sum_{b=1}^B \frac{\pi_0(\theta_b)}{\pi(\theta_b)} \cdot \bar{f}_\pi(x) \cdot \prod_{b=1}^B \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)}.$$

Now define a distribution P_2, with joint density

$$\frac{\left(\frac{\bar{f}_\pi(x)}{\widehat{f}_\pi(x)}\right)^{B-1}}{\mathbb{E}_{\bar{f}_{\pi_0}}\left[\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1}\right]} \cdot \frac{1}{B} \sum_{b=1}^B \frac{\pi_0(\theta_b)}{\pi(\theta_b)} \cdot \bar{f}_\pi(x) \cdot \prod_{b=1}^B \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)},$$

which differs from P_1 only in the presence of the first term. We can observe that, under P_2, the conditional distribution of $X \mid (\widehat{\theta}_1, \dots, \widehat{\theta}_B)$ is given by

$$X \mid (\widehat{\theta}_1, \dots, \widehat{\theta}_B) \sim \widehat{g}_\pi(\,\cdot \mid \widehat{\theta}_{1:B}),$$

where we recall that this joint density is defined as $\widehat{g}_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x)}{\widehat{f}_\pi(x)^{B-1}}$. Therefore, if we consider the sampling distribution

$$\begin{cases} (X, \widehat{\theta}_1, \dots, \widehat{\theta}_B) \sim P_2, \\ (\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \dots, \widehat{\theta}_B) \sim \widehat{Q}_\pi^M(\,\cdot \mid X, \widehat{\theta}_{1:B}), \end{cases} \tag{23}$$

then by our assumption (17) on $\widehat{Q}_\pi^M(\,\cdot \mid X, \widehat{\theta}_{1:B})$, it holds that $(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)})$ are exchangeable conditional on $(\widehat{\theta}_1, \dots, \widehat{\theta}_B)$, and consequently are also exchangeable after marginalizing over $(\widehat{\theta}_1, \dots, \widehat{\theta}_B)$.

Now compare this to running aCSS-B with the approximated marginal, which can be characterized by the sampling distribution

$$\begin{cases} (X, \widehat{\theta}_1, \dots, \widehat{\theta}_B) \sim P_0, \\ (\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \dots, \widehat{\theta}_B) \sim \widehat{Q}_\pi^M(\,\cdot \mid X, \widehat{\theta}_{1:B}). \end{cases}$$

Comparing this with (23), we can see that

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}\right) \le \mathrm{d}_{\mathrm{TV}}(P_0, P_2),$$

similarly to the proof of Appendix B.3. Next, we have already shown in the proof of Appendix B.3 that $\mathrm{d}_{\mathrm{TV}}(P_0, P_1) \le \epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}$, so from this point on we only need to bound $\mathrm{d}_{\mathrm{TV}}(P_1, P_2)$.

By definition of the joint density for P_2 as compared to P_1, we can compute the Radon–Nikodym derivative as

$$\frac{\mathsf{d}P_2(x, \theta_1, \dots, \theta_B)}{\mathsf{d}P_1(x, \theta_1, \dots, \theta_B)} = \frac{\left(\frac{\bar{f}_\pi(x)}{\widehat{f}_\pi(x)}\right)^{B-1}}{\mathbb{E}_{\bar{f}_{\pi_0}}\left[\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1}\right]}.$$

If the denominator is ≤ 1, then we have

$$\mathrm{d}_{\mathrm{TV}}(P_1, P_2) = \mathbb{E}_{P_1}\left[\left(1 - \frac{\mathsf{d}P_2(X, \widehat{\theta}_1, \dots, \widehat{\theta}_B)}{\mathsf{d}P_1(X, \widehat{\theta}_1, \dots, \widehat{\theta}_B)}\right)_+\right] \le \mathbb{E}_{P_1}\left[\left(1 - \left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1}\right)_+\right],$$

while if instead the denominator is ≥ 1 then

$$\mathrm{d}_{\mathrm{TV}}(P_1, P_2) = \mathbb{E}_{P_1}\left[\left(\frac{\mathsf{d}P_2(X, \widehat{\theta}_1, \dots, \widehat{\theta}_B)}{\mathsf{d}P_1(X, \widehat{\theta}_1, \dots, \widehat{\theta}_B)} - 1\right)_+\right] \le \mathbb{E}_{P_1}\left[\left(\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right)_+\right].$$

In either case, then,

$$\mathrm{d}_{\mathrm{TV}}(P_1, P_2) \le \mathbb{E}_{P_1}\left[\left|\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right|\right] = \mathbb{E}_{\theta_0}\left[\left|\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right|\right],$$

where the last step holds since under P_1, the marginal distribution of X is given by f_{θ_0}. This verifies that $\mathrm{d}_{\mathrm{TV}}(P_1, P_2) \le \delta_B$, which completes the proof.

C Sampling details

In this appendix, we provide details for the implementation of aCSS-B in our five empirical examples. For each example we will specify the procedure for sampling the posterior draws $\widehat{\theta}_1, \dots, \widehat{\theta}_B$, and for sampling the copies $\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(M)}$.

C.1 Logistic Regression

This section gives the implementation details for the logistic regression experiment presented in Section 4.1. (For the comparison to aCSS, the implementation of aCSS is exactly as in Barber and Janson (2022, Example 1, Section 4.5.2).)

Sampling from the posterior: The posterior distribution is challenging to sample from directly, so we instead sample the draws $\widehat{\theta}_b$ using Metropolis–Hastings sampling.

We now give details. The (unnormalized) log-density of the posterior distribution is given by

$$\Psi(\theta; x) = \log f_\theta(x) + \log \pi(\theta),$$

where

$$f_\theta(x) = \prod_{i=1}^{n} \left(\frac{e^{Z_i^\top \theta}}{1 + e^{Z_i^\top \theta}} \right)^{x_i} \cdot \left(\frac{1}{1 + e^{Z_i^\top \theta}} \right)^{1 - x_i},$$

$$\pi(\theta) = \prod_{j=1}^{d} \phi\left(\theta_j; 0, 1\right).$$

We will use a Laplace approximation to the posterior distribution as the proposal distribution. Define

$$\widehat{\theta} = \arg\max_\theta \Psi(\theta).$$

Since this does not admit a closed-form solution, in practice we solve this optimization problem numerically via `optim` in `R`. By a second-order Taylor expansion to Ψ around $\widehat{\theta}$, we have

$$\Psi(\theta) \approx \Psi_{\text{Laplace}}(\theta) := \Psi(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^\top \mathbf{H}(\theta - \widehat{\theta}),$$

where $\mathbf{H}$ is the Hessian of the negative log posterior evaluated at $\widehat{\theta}$:

$$\mathbf{H} = -\nabla^2 \Psi(\theta)\big|_{\theta = \widehat{\theta}} = \sum_{i=1}^{n} \frac{e^{Z_i^\top \widehat{\theta}}}{\left(1 + e^{Z_i^\top \widehat{\theta}}\right)^2} Z_i Z_i^\top + \mathbf{I}_d.$$

Therefore, $\Psi(\theta)$ is approximately equal to $\Psi_{\text{Laplace}}(\theta)$, which is the (unnormalized) log-density of a multivariate Gaussian distribution, $\mathcal{N}(\widehat{\theta}, \mathbf{H}^{-1})$. To sample the posterior draws $\widehat{\theta}_b$, we therefore run the Metropolis–Hastings algorithm (Owen, 2013), with proposal distribution $\mathcal{N}(\widehat{\theta}, \mathbf{H}^{-1})$. We use a burn-in of 500 steps, and then extract samples $\widehat{\theta}_1, \dots, \widehat{\theta}_B$ at every tenth step.

Sampling the copies: Next we give details on sampling the copies $\widetilde{X}^{(m)}$. Recall that our goal is to sample the copies i.i.d. from the distribution with density defined as in (5). We will first compute an approximation to the marginal density (recall Appendix A.2). Using the same Laplace approximation as above, we replace $\bar{f}_\pi(x)$ with

$$\widehat{f}_\pi(x) \propto \int \exp(\Psi_{\text{Laplace}}(\theta))\,\mathrm{d}\theta = \int \exp(\Psi(\widehat{\theta}))\exp\left(-\frac{1}{2}(\theta-\widehat{\theta})^\top \mathbf{H}(\theta-\widehat{\theta})\right)\mathrm{d}\theta$$
$$= \exp(\Psi(\widehat{\theta}))\sqrt{\frac{(2\pi)^d}{\det(\mathbf{H})}}.$$

(Note that this right-hand side is, implicitly, a function of x, since $\widehat{\theta}$ and $\mathbf{H}$ both depend on x.) This then leads to an approximate density, $\widehat{g}_\pi(x \mid \widehat{\theta}_1,\dots,\widehat{\theta}_B)$ (as in (5)) from which to sample the copies. This density then serves as the target density in a Gibbs sampling algorithm (Owen, 2013). We use the permuted serial sampler (Besag and Clifford, 1989), as follows:

1. **Initialization.** Draw $m_0 \in \{0,\dots,M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X,$$

2. **Iterations.** For $t = m_0+1,\dots,M$, for each $i = 1,\dots,n$,

$$\text{sample } \widetilde{X}_i^{(t)} \sim \widehat{g}_\pi(\cdot \mid x_{-i}, \widehat{\theta}_{1:B}) \text{ using } x_{-i} = (\widetilde{X}_{<i}^{(t)}, \widetilde{X}_{>i}^{(t-1)}),$$

where $\widehat{g}_\pi(\cdot \mid x_{-i}, \widehat{\theta}_{1:B})$ is the conditional density induced by $\widehat{g}_\pi(x \mid \widehat{\theta}_1,\dots,\widehat{\theta}_B)$. Similarly, for $t = m_0-1,\dots,0$, for each $i = n,\dots,1$,

$$\text{sample } \widetilde{X}_i^{(t)} \sim \widehat{g}_\pi(\cdot \mid x_{-i}, \widehat{\theta}_{1:B}) \text{ using } x_{-i} = (\widetilde{X}_{<i}^{(t+1)}, \widetilde{X}_{>i}^{(t)}).$$

Since each $\widetilde{X}_i^{(t)}$ is Bernoulli, we can compute the probability explicitly as

$$\mathbb{P}(X_i^{(t)} = 0) = \frac{\widehat{g}_\pi(0 \mid x_{-i}, \widehat{\theta}_{1:B})}{\widehat{g}_\pi(0 \mid x_{-i}, \widehat{\theta}_{1:B}) + \widehat{g}_\pi(1 \mid x_{-i}, \widehat{\theta}_{1:B})}$$

which allows us to draw from the conditional distribution.

3. **Output.** Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)},\dots,\widetilde{X}^{(m_0-1)},\widetilde{X}^{(m_0+1)},\dots,\widetilde{X}^{(M)}$.[2]

C.2 Mixture of Gaussians

This section gives the implementation details for the mixture of Gaussians experiment presented in Section 4.2. (For the comparison to reg-aCSS, the implementation of reg-aCSS is exactly as in Zhu and Barber (2023, Section 6.1).)

[2] For the serial sampler to return exchangeable copies, as required in (15) or (17), formally we would also need to randomly permute the set of copies produced by this algorithm—but since the p-value is invariant to shuffling the copies, it is unnecessary for this last step.

Sampling from the posterior: Define $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. To sample from the posterior of θ we introduce the latent variable $Z \in \{1, 2\}^n$.

$$X_i \mid Z_i = j \sim \mathcal{N}(\mu_j, \sigma_j^2), \quad \mathbb{P}(Z_i = j) = w_j,$$

for each $j = 1, 2$, where for convenience we write $w_2 = 1 - w_1$. This makes sampling from the posterior of θ straightforward. The full conditional distributions are given by the following:

$$\mathbb{P}(Z_i = j \mid X, w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \frac{w_j\, \phi(x_i; \mu_j, \sigma_j^2)}{\sum_{\ell=1}^{2} w_\ell\, \phi(x_i; \mu_\ell, \sigma_\ell^2)}, \quad i = 1, \ldots, n, \tag{24}$$

$$w_1 \mid Z \;\sim\; \text{Beta}\left(2 + n_1, 2 + n_2\right), \quad n_j = \sum_{i=1}^{n} \mathbf{1}\{Z_i = j\}, \tag{25}$$

$$\mu_j \mid \sigma_j^2, Z, X \;\sim\; \mathcal{N}\left(\frac{\sum_{i:z_i=j} x_i}{1 + n_j}, \frac{\sigma_j^2}{1 + n_j}\right), \tag{26}$$

$$\sigma_j^2 \mid \mu_j, Z, X \;\sim\; \text{Inv-Gamma}\left(\frac{3}{2} + \frac{n_j}{2}, \frac{1}{2} + \frac{1}{2} \sum_{i:Z_i=j} (x_i - \mu_j)^2 + \frac{1}{2}\mu_j^2\right). \tag{27}$$

These allow for efficient sampling using the Gibbs sampler.

We begin by initializing the two–component mixture as follows:

$$w^{(0)} = \left(w_1^{(0)}, w_2^{(0)}\right) = \left(\tfrac{1}{2}, \tfrac{1}{2}\right).$$

The latent allocations are then initialized by splitting the data at its sample median:

$$Z_i^{(0)} = \begin{cases} 1, & X_i \le \text{median}(X), \\ 2, & X_i > \text{median}(X), \end{cases} \qquad i = 1, \ldots, n.$$

Conditional on these initial assignments, we set each component's parameters to the empirical moments of its cluster:

$$\mu_j^{(0)} = \frac{1}{n_j^{(0)}} \sum_{i:Z_i^{(0)}=j} X_i, \qquad \sigma_j^{2\,(0)} = \frac{1}{n_j^{(0)}} \sum_{i:Z_i^{(0)}=j} \left(X_i - \mu_j^{(0)}\right)^2, \quad n_j^{(0)} = \sum_{i=1}^{n} \mathbf{1}\{Z_i^{(0)} = j\}.$$

Thereafter, for each iteration t, we perform the following Gibbs-sampling updates:

1. Independently for each i, sample each $Z_i^{(t)} \mid X, w^{(t-1)}, \mu^{(t-1)}, \sigma^{2,(t-1)}$ according to (24).
2. Sample $w_1^{(t)} \mid Z^{(t)}$ according to (25), and set $w_2^{(t)} = 1 - w_1^{(t)}$.
3. For each $j = 1, 2$, draw $\mu_j^{(t)} \mid \sigma_j^{2,(t-1)}, Z^{(t)}, X$ and $\sigma_j^{2,(t)} \mid \mu_j^{(t)}, Z^{(t)}, X$ according to (26) and (27), respectively.

We discard the first 500 draws and extract the posterior samples $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$ at every tenth step.

Sampling the copies: To sample the copies, we again use an approximation of the marginal $\bar{f}_\pi(x)$ and sample from

$$\widehat{g}_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x)}{\widehat{f}_\pi(x)^{\,B-1}}.$$

In order to approximate the marginal, we note that

$$\bar{f}_\pi(x) = \int f_\theta(x)\,\pi(\theta)\,d\theta,$$

and define the log of the unnormalized posterior as

$$\Psi(\theta; x) = \log f_\theta(x) + \log \pi(\theta).$$

For our choice of priors it takes the following form:

$$\Psi(\theta; x) = \sum_{i=1}^{n} \log\left(w_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) + (1 - w_1) \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right) \right)$$

$$+ \log w_1 + \log(1 - w_1) + \sum_{j=1}^{2} \left[\log(0.5) - 2\log(\sigma_j^2) - \frac{0.5}{\sigma_j^2} - \tfrac{1}{2}\log(2\pi\sigma_j^2) - \frac{\mu_j^2}{2\sigma_j^2} \right].$$

In our setting since we have two components in the null mixture distribution, the posterior is invariant to the swap of the component labels and has *two* modes. A single Laplace approximation is therefore inaccurate. We approximate $\bar{f}_\pi(x)$ by a mixture of two normal distributions where the means of the two components are identified as the two local maxima of $\Psi(\theta; x)$; say $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$. Following the same idea as in the usual construction of a Laplace approximation, the variance of these individual Gaussian components would be the inverse of the determinant of the Hessian of $\Psi(\theta; x)$ at $\theta = \widetilde{\theta}_1$ and $\widetilde{\theta}_2$, respectively, where the Hessians at these two points are defined as

$$\mathbf{H}_1 = -\nabla^2 \Psi(\theta)\big|_{\theta = \widetilde{\theta}_1} \qquad \text{and} \qquad \mathbf{H}_2 = -\nabla^2 \Psi(\theta)\big|_{\theta = \widetilde{\theta}_2},$$

both computable in closed form. Our two–mode Laplace approximation to the marginal is then

$$\widehat{f}_\pi(x) \;=\; e^{L_1} + e^{L_2},$$

where

$$L_s \;=\; \Psi(\widetilde{\theta}_s; x) \;-\; \tfrac{1}{2}\operatorname{logdet}\mathbf{H}_s \;+\; \tfrac{5}{2}\log(2\pi), \qquad s = 1, 2$$

is the contribution to the integral from the s^{th} component of the normal mixture approximation to the (unnormalized) posterior. In order to obtain $\widehat{\theta}_s$, we use `optim` in `R` as follows:

1. **Reparametrization.** To convert the constrained optimization into an unconstrained one, define the following reparameterization for $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_1^2)$:

   $$\vartheta(\theta) = (z_1, \mu_1, s_1, \mu_2, s_2) \text{ where } z_1 = \log\left(\frac{w_1}{1 - w_1}\right),\ s_1 = \log(\sigma_1^2),\ \text{and } s_2 = \log(\sigma_2^2).$$

2. **Initialization.** Apply k-means with $k = 2$ on the observed data and obtain two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ such that $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \dots, n\}$ with the cluster centers c_1 and c_2 respectively. Define $\tau_1^2 = \frac{1}{|\mathcal{C}_1|-1} \sum_{i \in \mathcal{C}_1} (x_i - c_1)^2$ and $\tau_2^2 = \frac{1}{|\mathcal{C}_2|-1} \sum_{i \in \mathcal{C}_2} (x_i - c_2)^2$. Define the following two initializations:
   - $\vartheta_1^{(0)} = \vartheta\left(\frac{|\mathcal{C}_1|}{n}, c_1, \tau_1^2, c_2, \tau_2^2\right)$,
   - $\vartheta_2^{(0)} = \vartheta\left(\frac{|\mathcal{C}_2|}{n}, c_2, \tau_2^2, c_1, \tau_1^2\right)$.

3. **Optimization.** Starting from the two initial values $\vartheta_1^{(0)}$ and $\vartheta_2^{(0)}$, maximize Ψ in the unconstrained parameter space $\mathbb{R}^5$ using `optim` with the `BFGS` algorithm to obtain $\widehat{\vartheta}_1$ and $\widehat{\vartheta}_2$, respectively. Map these back to the original parameters and get $\widetilde{\theta}_s = \vartheta^{-1}(\widehat{\vartheta}_s)$ for $s \in \{1, 2\}$.

Thus $\widehat{f}_\pi(x)$ gives us an approximation to the marginal. We would like to sample the X_i's from the $\widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_\theta(x_i)}{\widehat{f}_\pi(x)}$. To do that we will be approximating $\widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B})$ using a two component normal distribution which will be used as a proposal in a Metropolis–Hastings algorithm. The rest of this section describes how to obtain this mixture of normal proposal from $\widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B})$.

Define the negative log-density as $\zeta(x_i) := -\log \widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B})$. We consider a grid $\{\xi_j\}_{j=1}^K$ with $K = 20$, spanning $[a, b]$ where $a = \min_i X_i, b = \max_i X_i$ and evaluate $\zeta(\xi_j)$ for all $j = 1, \dots, 20$. We fit a continuous piecewise–quadratic surrogate of the form

$$Q(\xi; c, \beta) \;=\; \beta_0 + \mathbf{1}\{\xi \le c\}\big(\beta_{1L}(\xi - c) + \beta_{2L}(\xi - c)^2\big) + \mathbf{1}\{\xi > c\}\big(\beta_{1R}(\xi - c) + \beta_{2R}(\xi - c)^2\big),$$

where $\beta = (\beta_0, \beta_{1L}, \beta_{2L}, \beta_{1R}, \beta_{2R})$, to these ζ evaluations by minimizing the objective function

$$\mathrm{SSE}(c) := \min_\beta \sum_{j=1}^K \{\zeta(\xi_j) - Q(\xi_j; c, \beta)\}^2.$$

For each candidate value of c, the optimum $\widehat{\beta}$ can be obtained by a least squares algorithm and the optimum value c (say $\widehat{c}$) is chosen via a grid search on 400 different values. This yields coefficients $(\widehat{\beta}_0, \widehat{\beta}_{1L}, \widehat{\beta}_{2L}, \widehat{\beta}_{1R}, \widehat{\beta}_{2R})$ at breakpoint $\widehat{c}$.

For each arm $s \in \{L, R\}$, we map the quadratic to a Gaussian as follows:

$$v_s = \max\left\{\epsilon, \frac{1}{2\widehat{\beta}_{2s}}\right\}, \qquad m_s = \widehat{c} - \widehat{\beta}_{1s} v_s$$

with $\epsilon = 10^{-8}$. To obtain the mixture weights, we match the magnitude of Q at the two means m_L and m_R with that of a two-component mixture of normal distribution with means m_L and m_R and variances v_L and v_R by solving the following equation for p and q:

$$\begin{bmatrix} \phi(m_L; m_L, v_L) & \phi(m_L; m_R, v_R) \\ \phi(m_R; m_L, v_L) & \phi(m_R; m_R, v_R) \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} \exp\left(-Q\left(m_L; \widehat{c}, \widehat{\beta}\right)\right) \\ \exp\left(-Q\left(m_R; \widehat{c}, \widehat{\beta}\right)\right) \end{bmatrix}.$$

where $\phi(\,\cdot\,; m, v)$ denotes the pdf of a Gaussian distribution with mean m and variance v. Finally, we set the weight (of the L component) as

$$w = \frac{p}{p+q}.$$

So, the proposal x_{prop} is drawn from the two–component Gaussian mixture density

$$w\,\phi(\cdot; \mu_L, v_L) + (1-w)\,\phi(\cdot; \mu_R, v_R).$$

This yields the following Metropolis–Hastings algorithm with the permuted serial sampler:

1. **Initialization.** Draw $m_0 \in \{0, \dots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X.$$

2. **Iterations.** For $t = m_0 + 1, \dots, M$, for each $i = 1, \dots, n$ draw x_{prop} from the density $w\,\phi(\cdot; \mu_L, v_L) + (1-w)\,\phi(\cdot; \mu_R, v_R)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

where

$$\alpha = \min\left\{1,\ \frac{g_\pi\Big(x_{\text{prop}} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)}{\widehat{g}_\pi\Big(X_i^{(t-1)} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)} \cdot \frac{w\,\phi(X_i^{(t-1)}; \mu_L, v_L) + (1-w)\,\phi(X_i^{(t-1)}; \mu_R, v_R)}{w\,\phi(x_{\text{prop}}; \mu_L, v_L) + (1-w)\,\phi(x_{\text{prop}}; \mu_R, v_R)}\right\}$$

and $x_{-i} = \big(\widetilde{X}_{<i}^{(t)},\ \widetilde{X}_{>i}^{(t-1)}\big)$. Similarly, for $t = m_0 - 1, \dots, 0$, for each $i = n, \dots, 1$, we draw x_{prop} from the density $w\,\phi(\cdot; \mu_L, v_L) + (1-w)\,\phi(\cdot; \mu_R, v_R)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

as before with the only difference that $x_{-i} = \big(\widetilde{X}_{<i}^{(t+1)},\ \widetilde{X}_{>i}^{(t)}\big)$.

3. **Output.** Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \dots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \dots, \widetilde{X}^{(M)}$.

C.3 Rank-1 matrix

This section gives the implementation details for the rank-1 matrix experiment presented in Section 4.3.

Sampling from the posterior: Recall that our prior is the multivariate standard normal distribution, $U, V \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$. While the posterior distribution of (U, V)—that is, the distribution of $(U, V) \mid X$—is challenging to work with, it is straightforward to compute the conditionals of the posterior distribution: we have

$$U \mid X, V \sim \mathcal{N}\left(\frac{4XV}{4\|V\|_2^2 + 1},\ \frac{1}{4\|V\|_2^2 + 1}\mathbf{I}_n\right) \tag{28}$$

and

$$V \mid X, U \sim \mathcal{N}\left(\frac{4X^\top U}{4\|U\|_2^2+1}, \frac{1}{4\|U\|_2^2+1}\mathbf{I}_n\right). \tag{29}$$

Thus, we can easily sample from the posterior using Gibbs sampling. We begin by initializing U, V via a rank-1 approximation to X: writing $u, v \in \mathbb{R}^n$ as the leading left and right singular vectors of X, respectively, we initialize with

$$U = \sqrt{n} \cdot u, \quad V = \sqrt{n} \cdot v,$$

and iterate the Gibbs sampler,

$$\begin{cases} \text{Sample } U \mid V, X \text{ according to distribution (28)}, \\ \text{Sample } V \mid U, X \text{ according to distribution (29)}. \end{cases}$$

We use a burn-in of 500 steps, and then extract samples $\widehat{\theta}_1, \dots, \widehat{\theta}_B$ at every tenth step.

Sampling the copies: As in earlier examples, we will need to use a Laplace approximation for the marginal distribution of X. However, in this specific setting, we will need to proceed a bit differently—this is because the negative log likelihood of $(U, V) \mid X$ is a nonconvex function, and indeed has issues of identifiability, as, e.g., the likelihood takes equal values at (u, v) and $(-u, -v)$. Instead, we will first marginalize over U exactly, and then use a Laplace approximation for marginalizing over V.

First, we calculate the distribution of $X \mid V$, i.e., we marginalize over U: using X_j to denote the jth row of X, we have

$$X_j \mid V \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, 0.25\,\mathbf{I}_n + VV^\top\right).$$

The distribution of $X \mid V = v$ therefore has conditional density

$$\begin{aligned} f_v(x) &= \frac{1}{(2\pi)^{n^2/2}\ \det(0.25\,\mathbf{I}_n + vv^\top)^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^n x_j^\top (0.25\,\mathbf{I}_n + vv^\top)^{-1} x_j} \\ &= \frac{1}{(2\pi)^{n^2/2}\left[(0.25)^n(1+4\|v\|_2^2)\right]^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^n x_j^\top \left(4\mathbf{I}_n - \frac{16}{1+4\|v\|_2^2} vv^\top\right) x_j} \\ &= \frac{4^{\,n^2/2}}{(2\pi)^{n^2/2}\,(1+4\|v\|_2^2)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2 + \frac{8}{1+4\|v\|_2^2}\ \|xv\|_2^2}. \end{aligned}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, and x_j now denotes the j^{th} row of $x \in \mathbb{R}^{n \times n}$. Writing $X = A\,\mathrm{diag}\{d\}\,B^\top$ as the singular value decomposition of X where $A, B \in \mathbb{R}^{n \times n}$, we can rewrite the marginal as follows:

$$
\begin{aligned}
\bar{f}_\pi(x) &= \mathbb{E}_{V\sim\mathcal{N}(0,\mathbf{I}_n)}\left[f_V(x)\right] \\
&= \mathbb{E}_{V\sim\mathcal{N}(0,\mathbf{I}_n)}\left[\frac{4^{n^2/2}}{(2\pi)^{n^2/2}\left(1+4\|V\|_2^2\right)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2+8\frac{\left\|A\cdot\mathrm{diag}\{d\}\cdot B^\top V\right\|_2^2}{1+4\|V\|_2^2}}\right] \\
&= \mathbb{E}_{V\sim\mathcal{N}(0,\mathbf{I}_n)}\left[\frac{4^{n^2/2}}{(2\pi)^{n^2/2}\left(1+4\|V\|_2^2\right)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2+8\frac{\|\mathrm{diag}\{d\}V\|_2^2}{1+4\|V\|_2^2}}\right] \\
&= \mathbb{E}_{W_1,\ldots,W_n\overset{\text{iid}}{\sim}\chi_1^2}\left[\frac{4^{n^2/2}}{(2\pi)^{n^2/2}\left(1+4\sum_{i=1}^n W_i\right)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2+8\frac{\sum_{i=1}^n W_i d_i^2}{1+4\sum_{i=1}^n W_i}}\right], \qquad (30)
\end{aligned}
$$

where the second-to-last step holds due to rotational invariance of the standard normal distribution and the ℓ_2 norm. In the last step, we reparametrize the integrand in Equation (30) in terms of W instead of V (i.e., $W_i \sim \chi_1^2$ replaces V_i^2). This reparametrization addresses the rotational and sign invariances of f in terms of V, thereby avoiding complications in the integration arising from multimodality. Before proceeding with the Laplace approximation, we again reparametrize $t_i = \log W_i$; otherwise the χ_1^2 prior for W_i (and hence the unnormalized posterior integrand) is unbounded at 0, violating the regularity conditions for a Laplace approximation. Since $W_i \sim \chi_1^2$, denoting by $\pi_{W_i}(w_i)$ the χ_1^2 density at w_i, the change of variables gives

$$
\pi(t_i) = \pi_{W_i}(e^{t_i})\left|\frac{d}{dt_i}e^{t_i}\right| = \frac{1}{\sqrt{2\pi}}\, e^{t_i/2}\, e^{-e^{t_i}/2}, \qquad t_i \in \mathbb{R},\ i \in \{1,\ldots,n\},
$$

hence

$$
\log\pi(t) = \sum_{i=1}^n \left(-\tfrac{1}{2}\log(2\pi) + \tfrac{1}{2}t_i - \tfrac{1}{2}e^{t_i}\right), \quad t = (t_1,\ldots,t_n)^\top.
$$

If we denote

$$
S_1(t) := \sum_{i=1}^n e^{t_i}, \qquad S_d(t) := \sum_{i=1}^n d_i^2 e^{t_i},
$$

from Equation (30), the conditional density $f_t(x)$ is

$$
f_t(x) = \frac{4^{n^2/2}}{(2\pi)^{n^2/2}\left(1+4S_1(t)\right)^{n/2}} \exp\!\left(-2\|x\|_F^2 + \frac{8\,S_d(t)}{1+4S_1(t)}\right).
$$

Define

$$
\begin{aligned}
\Psi(t;x) &:= \log f_t(x) + \log\pi(t) \\
&= -\frac{n}{2}\log\!\left(1+4S_1(t)\right) + \frac{8\,S_d(t)}{1+4S_1(t)} + \sum_{i=1}^n\left(\tfrac{1}{2}t_i - \tfrac{1}{2}e^{t_i}\right) + \mathrm{const}(x)
\end{aligned}
$$

where $\mathrm{const}(x) = \frac{n^2}{2}\log 4 - \frac{n^2}{2}\log(2\pi) - 2\|x\|_F^2$ does not depend on t. Introduce the following notations:

$$
c(t) = 1 + 4S_1(t), \qquad N_i(t) = c(t)\,d_i^2 - 4S_d(t).
$$

The gradient is then given by

$$\nabla_t \Psi(t; x) = \left(\frac{\partial \Psi}{\partial t_1}, \ldots, \frac{\partial \Psi}{\partial t_n}\right)^{\top}, \quad \text{where} \quad \frac{\partial \Psi}{\partial t_j} = \frac{1}{2} - \frac{1}{2}e^{t_j} - \frac{2n}{c(t)}e^{t_j} + \frac{8e^{t_j}}{c(t)^2}\big(c(t)\, d_j^2 - 4S_d(t)\big).$$

The Hessian of the negative log–posterior is $H(t; x) := -\nabla_t^2 \Psi(t; x)$ with

$$\begin{aligned} H_{jk}(t; x) &= -\Big[\frac{8n}{c(t)^2}e^{t_j}e^{t_k} + \frac{32}{c(t)^2}e^{t_j}e^{t_k}\,(d_j^2 - d_k^2) - \frac{64}{c(t)^3}e^{t_j}e^{t_k}\,N_j(t)\Big], & j \neq k, \\ H_{jj}(t; x) &= \frac{1}{2}e^{t_j} + 2n\left(\frac{e^{t_j}}{c(t)} - \frac{4e^{2t_j}}{c(t)^2}\right) - \ 8e^{t_j}\left(\frac{N_j(t)}{c(t)^2} - \frac{8e^{t_j}\,N_j(t)}{c(t)^3}\right), & j = k. \end{aligned}$$

We find $\hat{t} = \arg\max_t \Psi(t; x)$ by using `optim` in `R` and denote $\mathbf{H} := H(\hat{t}; x)$. The Laplace approximation to the marginal is

$$\widehat{f_\pi}(x) = (2\pi)^{n/2}\ \det(\mathbf{H})^{-1/2}\ \exp\big\{\Psi(\hat{t}; x)\big\}.$$

Using $\widehat{f_\pi}(x)$ we define $\widehat{g_\pi}\left(\ \cdot \mid \widehat{\theta}_{1:B}\right)$ as in Appendix A.2 which serves as the target density. Consider the density induced by $\widehat{g_\pi}(\cdot \mid \widehat{\theta}_{1:B})$ on each X_{ij}, i.e.

$$\widehat{g_\pi}(x_{ij} \mid x_{-ij}, \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x_{ij})}{\widehat{f_\pi}(x)^{B-1}},$$

where x_{-ij} denotes all entries of x except the (i, j)-th position. In order to sample each X_{ij} from that density, we use the Metropolis–Hastings algorithm. For the proposal density, we perform a Laplace approximation on this density. Consider $\zeta(x_{ij}) = \log \widehat{g_\pi}\left(x_{ij} \mid x_{-ij}, \widehat{\theta}_{1:B}\right)$, let

$$x_{ij}^{\star} = \arg\max_{x_{ij}} \zeta(x_{ij}),$$

and denote the Hessian at $x_{ij}^{\star}$ by $\zeta''(x_{ij}^{\star})$. (Note that $x_{ij}^{\star}$ and $\zeta''(x_{ij}^{\star})$ are implicitly functions of x_{-ij}). The proposal distribution is given by $\mathcal{N}\left(x_{ij}^{\star}, -\frac{1}{\zeta''(x_{ij}^{\star})}\right)$. This optimization is performed numerically via `optim` in `R` and the resulting Metropolis–Hastings algorithm has average acceptance probability very close to 1. Putting all of these steps together, we obtain the following permuted serial sampler to sample $\widetilde{X}^{(m)}$'s:

1. **Initialization.** Draw $m_0 \in \{0, \ldots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \ \leftarrow\ X,$$

2. **Iterations.** For $t = m_0 + 1, \ldots, M$, for each $i = 1, \ldots, m$ and $j = 1, \ldots, n$, draw $x_{\text{prop}} \sim \mathcal{N}\Big(x_{ij}^{\star},\ -\frac{1}{\zeta''(x_{ij}^{\star})}\Big)$, $u \sim \text{Unif}(0, 1)$ and set

$$\widetilde{X}_{ij}^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \leq \alpha, \\ \widetilde{X}_{ij}^{(t-1)}, & \text{otherwise,} \end{cases}$$

where

$$\alpha = \min\left\{ 1,\ \frac{\widehat{g}_\pi\Big(x_{\text{prop}} \mid x_{-ij}, \widehat{\theta}_{1:B}\Big)}{\widehat{g}_\pi\Big(X_{ij}^{(t-1)} \mid x_{-ij}, \widehat{\theta}_{1:B}\Big)} \cdot \frac{\phi\Big((X_{ij}^{(t-1)} - x_{ij}^\star)\sqrt{-\zeta''(x_{ij}^\star)}\Big)}{\phi\Big((x_{\text{prop}} - x_{ij}^\star)\sqrt{-\zeta''(x_{ij}^\star)}\Big)} \right\},$$
$$x_{-ij} = \big(\widetilde{X}_{<i,1:n}^{(t)},\ \widetilde{X}_{i,<j}^{(t)},\ \widetilde{X}_{i,>j}^{(t-1)},\ \widetilde{X}_{>i,1:n}^{(t-1)}\big).$$

Similarly, for $t = m_0 - 1, \ldots, 0$, for each $j = n, \ldots, 1$ and $i = m, \ldots, 1$, we draw $x_{\text{prop}} \sim \mathcal{N}\Big(x_{ij}^\star,\ -\frac{1}{\zeta''(x_{ij}^\star)}\Big)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_{ij}^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_{ij}^{(t-1)}, & \text{otherwise,} \end{cases}$$

as before with the only difference that $x_{-ij} = \big(\widetilde{X}_{<i,1:n}^{(t)},\ \widetilde{X}_{i,<j}^{(t)},\ \widetilde{X}_{i,>j}^{(t+1)},\ \widetilde{X}_{>i,1:n}^{(t+1)}\big)$.

3. **Output.** Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.

C.4 Group-sparse regression

This section gives the implementation details for the group-sparse regression experiment presented in Section 4.4.

Sampling from the posterior We restate the priors defined in Section 4.4 here with the general parameters as follows:

$$\begin{aligned} &g^\star \sim \text{Unif}(\{1,\ldots, \text{G}\}), \\ &\beta_{I_{g^\star}} \sim \mathcal{N}(\mathbf{0}_{|I_{g^\star}|}, \mathbf{I}_{|I_{g^\star}|}), \\ &\beta_{I_g} = \mathbf{0}_{|I_g|}\ \forall\ g \neq g^\star. \end{aligned}$$

Under this prior, the posterior distribution of β can be derived with the following hierarchical structure: defining

$$D_g(X) = |A_g|^{-\frac{1}{2}} \exp\left(\frac{1}{2} b_g^\top A_g^{-1} b_g - \frac{1}{2} X^\top X\right)$$

for each $g \in [G]$, where
$$A_g = Z_{I_g}^\top Z_{I_g} + I_{d_g}, \quad b_g = Z_{I_g}^\top X,$$
we first sample the active group $g^\star$ as

$$\mathbb{P}(g^\star = g \mid X) \propto D_g(X),$$

then sample $\beta \mid g^\star, X$ as

$$\begin{aligned} &\beta_{I_{g^\star}} \mid g^\star, X \sim \mathcal{N}\left(A_{g^\star}^{-1} b_{g^\star}, A_{g^\star}^{-1}\right), \\ &\beta_{I_g} \mid g^\star, X = 0\ \forall\ g \neq g^\star. \end{aligned}$$

Sampling the copies: In this case, we can evaluate the marginal of X exactly. Up to normalizing constants,

$$\bar{f}_\pi(X) = \frac{1}{G}\sum_{g=1}^{G}\int_{\mathbb{R}^{d_g}} \exp\left(-\frac{1}{2}||X - Z\beta||^2 - \frac{1}{2}||\beta_{I_g}||^2\right) \mathsf{d}\beta_{I_g} = \frac{1}{G}\sum_{g=1}^{G} D_g(X).$$

In order to sample from the final sampling density $g_\pi(\,\cdot \mid \widehat{\theta}_{1:B})$ as in Equation (5), we use a Laplace approximation as the proposal in a Metropolis–Hastings sampler. Our goal is to approximate the conditional sampling density of X_i, and we will update the X_i's one at a time. Since this conditional density is proportional to $g_\pi(X \mid \widehat{\theta}_{1:B})$, we define

$$\zeta(x_i) = -\frac{1}{2}\sum_{b=1}^{B}(x_i - Z_i^\top \widehat{\beta}_b)^2 - (B-1)\log\left(\frac{1}{G}\sum_{g=1}^{G} D_g(x)\right).$$

Taking the derivatives, we can find the maximum and also the Hessian which can then be used to perform a Laplace approximation.

$$\begin{aligned}
\zeta'(x_i) &= -\sum_{b=1}^{B}(x_i - Z_i^\top \widehat{\beta}_b) - (B-1)\frac{\sum_{g=1}^{G}\frac{\partial}{\partial x_i} D_g(x)}{\sum_{g=1}^{G} D_g(x)}, \\
\zeta''(x_i) &= \frac{B-1}{\left(\sum_{g=1}^{G} D_g(x)\right)^2}\left[\left(\sum_{g=1}^{G} D_g(x)\right)\left(\sum_{g=1}^{G}\frac{\partial^2}{\partial x_i^2} D_g(x)\right) - \left(\sum_{g=1}^{G}\frac{\partial}{\partial x_i} D_g(x)\right)^2\right],
\end{aligned}$$

where Z_i is the i^{th} row of the covariate matrix Z and

$$\begin{aligned}
\frac{\partial}{\partial x_i} D_g(x) &= \left[D_g(x)\left(Z_{I_g} A_g^{-1} b_g - x\right)\right]_i, \\
\frac{\partial^2}{\partial x_i^2} D_g(x) &= \left[D_g(x)\left(\frac{1}{\sigma^2} Z_{I_g} A_g^{-1} Z_{I_g}^\top - I_n\right) + \left(Z_{I_g} A_g^{-1} b_g - x\right) D_g'(x)^\top\right]_{ii}.
\end{aligned}$$

The optimization is carried out numerically by `optim` in `R` to find the maximum $x_i^\star$ and then the approximated Gaussian density $\mathcal{N}\left(x_i^\star, -\frac{1}{\zeta''(x_i^\star)}\right)$ is used as the proposal for our Metropolis–Hastings sampling. The acceptance probability in the Metropolis–Hastings stays close to 1. We use the permuted serial sampler, as follows:

1. **Initialization.** Draw $m_0 \in \{0, \ldots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X,$$

2. **Iterations.** For $t = m_0 + 1, \ldots, M$, for each $i = 1, \ldots, n$ draw $x_{\text{prop}} \sim \mathcal{N}\left(x_i^\star, -\frac{1}{\zeta''(x_i^\star)}\right)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \leq \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

where

$$\alpha = \min\left\{1,\ \frac{g_\pi\Big(x_{\text{prop}} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)}{g_\pi\Big(X_i^{(t-1)} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)} \cdot \frac{\phi\Big((X_i^{(t-1)} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\Big)}{\phi\Big((x_{\text{prop}} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\Big)}\right\}$$

and $x_{-i} = \big(\widetilde{X}_{<i}^{(t)},\ \widetilde{X}_{>i}^{(t-1)}\big)$. Similarly, for $t = m_0 - 1, \ldots, 0$, for each $i = n, \ldots, 1$, we draw $x_{\text{prop}} \sim \mathcal{N}\Big(x_i^\star,\ -\frac{1}{\zeta''(x_i^\star)}\Big)$, $u \sim \text{Unif}(0,1)$ and set $\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \leq \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$ as before with the only difference that $x_{-i} = \big(\widetilde{X}_{<i}^{(t+1)},\ \widetilde{X}_{>i}^{(t)}\big)$.

3. **Output.** Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.

C.5 Linear spline regression

This section gives the implementation details for the linear spline regression experiment presented in Section 4.5.

Sampling from the posterior: Following Equation (14), with $k = 1$ (i.e., one knot) the linear spline model can be represented as

$$X = \gamma_0 + \gamma_1 Z + \gamma_2 b_1(Z) + \epsilon$$

which can be further written as:

$$X = h_t(Z)\gamma + \epsilon,$$

where

$$h_t(Z) = (\mathbf{1}_n, Z, b_1(Z)) \in \mathbb{R}^{n\times 3} \quad \text{and} \quad \gamma = (\gamma_0, \gamma_1, \gamma_2)^\top.$$

Let $Z_{(i)}$ denote the i-th order statistic of $Z_1, \ldots, Z_n$ and for notational convenience we shall denote $Z_{(0)} = -\infty$ and $Z_{(n+1)} = \infty$. We shall use $X_{(i)}$ to denote the response corresponding to covariate $Z_{(i)}$. Recall that we are using the prior distribution

$$\begin{aligned} &\gamma_0, \gamma_1, \gamma_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\ &t_1 \sim \mathcal{N}(0,1). \end{aligned}$$

To draw the posterior samples, we use a Gibbs sampling algorithm where the conditional distributions are as follows. First, the conditional distribution of γ is given by

$$(\gamma \mid t, Z, X) \sim \mathcal{N}\left(\mu_\gamma, V_\gamma\right), \tag{31}$$
$$\text{where } V_\gamma = \left(4h_t(Z)^\top h_t(Z) + \mathbf{I}_3\right)^{-1} \quad \text{and} \quad \mu_\gamma = V_\gamma\left(4h_t(Z)^\top X\right).$$

Next, the conditional distribution of t_1 has density

$$\begin{aligned} P(t_1 \mid \gamma, X, Z) &\propto \phi\left(\frac{\|X - h_t(Z)\gamma\|}{\sigma}\right)\phi\left(\frac{t_1}{\tau_2}\right) \\ &\propto \exp\left(-2\sum_{i=1}^n \left(X_i - \gamma_0 - \gamma_1 Z_i - \gamma_2 b_1(Z_i)\right)^2\right)\exp\left(-\frac{t_1^2}{2}\right). \end{aligned}$$

Recalling that $b_1(z) = (z - t_1)_+$, we note that in this distribution, for any i, when $t_1 \in (Z_{(i)}, Z_{(i+1)})$, the density $P(t_1 \mid \gamma, X, Z)$ is

$$\propto \exp\left(-2 \sum_{i'=i+1}^{n} \left(X_{(i')} - \gamma_0 - \gamma_1 Z_{(i')} - \gamma_2 (Z_{(i')} - t_1)\right)^2\right) \exp\left(-\frac{t_1^2}{2}\right),$$

and is therefore proportional to the following Gaussian distribution:

$$P(t_1 \mid t_1 \in (Z_{(i)}, Z_{(i+1)}), \gamma, X, Z) \propto \mathcal{N}(\mu_{t,i}, \sigma_{t,i}^2), \tag{32}$$

$$\begin{aligned} \text{where,} \qquad \mu_{t,i} &= \left(4(n-i)\,\gamma_2^2 + 1\right)^{-1} \left(4\gamma_2 \sum_{i'=i+1}^{n} \left(X_{(i')} - \gamma_0 - \gamma_1 Z_{(i')} - \gamma_2\, Z_{(i')}\right)\right), \\ \sigma_{t,i}^2 &= \left(4(n-i)\,\gamma_2^2 + 1\right)^{-1}. \end{aligned}$$

The probability that $t_1 \in (Z_{(i)}, Z_{(i+1)})$ can be obtained by evaluating the following integral

$$\begin{aligned} &\mathbb{P}(t_1 \in (Z_{(i)}, Z_{(i+1)}) \mid \gamma, X, Z) \\ &= \int_{t_1 = Z_{(i)}}^{Z_{(i+1)}} e^{-2\sum_{i'=1}^{i}\left(X_{(i')} - \gamma_0 - \gamma_1 Z_{(i')}\right)^2 - 2\sum_{i'=i+1}^{n}\left(X_{(i')} - \gamma_0 - \gamma_1 Z_{(i')} - \gamma_2 (Z_{(i')} - t_1)\right)^2 - \frac{1}{2} t_1^2}\, \mathsf{d}t_1 \\ &=: w_i. \end{aligned} \tag{33}$$

This integration can be solved by expressing the exponent in the integrand as a quadratic in t_1 and then using the Gaussian CDF. This enables us to sample from the posterior of t_1 by first sampling the interval in which t_1 lies using the above probability weights and subsequently drawing t_1 from a truncated normal distribution. We will represent this sampling distribution concisely as follows:

$$t_1 \mid Z, X, \gamma \sim \sum_{i=0}^{n} w_i \, \mathrm{TN}(\mu_{t,i}, \sigma_{t,i}^2, Z_{(i)}, Z_{(i+1)}), \tag{34}$$

where $\mathrm{TN}(\mu, \sigma^2, a, b)$ is the truncated normal distribution—that is, the distribution $\mathcal{N}(\mu, \sigma^2)$ truncated to the interval $[a, b]$. This results in the following Gibbs sampling algorithm to sample from the posterior distribution of the parameters. We use the `R` package `segmented` (Muggeo, 2023) to find the position of the knots and initialize $t^{(0)}$ at that point. We initialize $\gamma^{(0)}$ as the coefficient vector from the regression of X on $h_{t^{(0)}}$. For $b = 1, \ldots, B$,

1. Generate $\gamma^{(b)}$ from $\gamma \mid t_1^{(b-1)}, Z, X$ as in Equation (31).

2. Generate $t_1^{(b)}$ from $t_1 \mid Z, X, \gamma^{(b)}$ as in Equation (34).

After the burn-in of $B_0 = 500$, we extract $B = 25$ posterior samples $\{t^{(b)}, \gamma^{(b)}\}$ at every tenth step.

Sampling the copies: Our first step is to approximate the marginal $\bar{f}_\pi(x)$. In this case, we first marginalize out γ from the distribution of $X \mid t, \gamma$ as follows:

$$\begin{aligned} X \mid t_1, \gamma &\sim \mathcal{N}\left(h_t(Z)\gamma, 0.25 I_n\right) \\ \implies X \mid t_1 &\sim \mathcal{N}\left(\mathbf{0}_n, h_t(Z)h_t(Z)^\top + 0.25\mathbf{I}_n\right). \end{aligned}$$

Now consider the joint density of (X, t_1). Since our prior on t_1 is $\mathcal{N}(0,1)$, the joint density is given by

$$\Psi(x, t_1) = \frac{1}{\sqrt{(2\pi)^n |h_t(Z)h_t(Z)^\top + 0.25\mathbf{I}_n|}} e^{-\frac{1}{2}x^\top (h_t(Z)h_t(Z)^\top + 0.25\mathbf{I}_n)^{-1}x} \cdot \frac{1}{\sqrt{2\pi}} e^{-t_1^2/2}$$

and the marginal is then given by $\bar{f}_\pi(x) = \int \Psi(x, t_1)\, \mathsf{d}t_1$. Note that evaluating $\Psi(x, t_1)$ at a single value t_1 is simple, but integrating this quantity is challenging. We will therefore take an approximation: first re-define $Z_{(0)} = -C$ and $Z_{(n+1)} = C$ (for a large value C chosen so that the tail mass of Ψ outside $[-C, C]$ is negligible), and let $Z_{(1)}, \dots, Z_{(n)}$ be as before. We then define a grid $z_0 \le \dots \le z_{K(n+1)}$ where $z_{Ki} = Z_{(i)}$ for each i, and the points $z_{K(i-1)}, \dots, z_{Ki}$ form an equally spaced grid from $Z_{(i-1)}$ to $Z_{(i)}$ for each i. After evaluating $\Psi(z_0), \dots, \Psi(z_{K(n+1)})$, we approximate the integral with the trapezoid rule:

$$\widehat{f}_\pi(x) = \sum_{i=1}^{n+1} \sum_{k=1}^{K} \frac{\Psi(x, z_{K(i-1)+k-1}) + \Psi(x, z_{K(i-1)} + k)}{2} \cdot \frac{Z_{(i)} - Z_{(i-1)}}{K}.$$

For our implementation we choose $C = 10$ and $K = 20$.

Using this we can define the approximate target density $\widehat{g}_\pi\left(\cdot \mid \widehat{\theta}_{1:B}\right)$ as in Appendix A.2 which serves as the target density. In order to sample from that density, we use the Metropolis–Hastings algorithm. Consider $\zeta(x_i) = \log \widehat{g}_\pi\left(x_i \mid x_{-i}, \widehat{\theta}_{1:B}\right)$,

$$x_i^\star = \arg\max_{x_i} \zeta(x_i),$$

and denote the Hessian at $x_i^\star$ by $\zeta''(x_i^\star)$. The proposal distribution is given by $\mathcal{N}\left(x_i^\star, -\frac{1}{\zeta''(x_i^\star)}\right)$. This optimization is performed numerically via `optim` in `R`. We use the permuted serial sampler, as follows:

1. **Initialization.** Draw $m_0 \in \{0, \dots, M\}$ uniformly at random, and set
$$\widetilde{X}^{(m_0)} \leftarrow X.$$

2. **Iterations.** For $t = m_0 + 1, \dots, M$, for each $i = 1, \dots, n$ draw $x_{\text{prop}} \sim \mathcal{N}\left(x_i^\star, -\frac{1}{\Psi''(x_i^\star)}\right)$, $u \sim \text{Unif}(0,1)$ and set
$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

where

$$\alpha = \min\left\{1,\ \frac{\widehat{g}_\pi\Big(x_{\text{prop}} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)}{\widehat{g}_\pi\Big(X_i^{(t-1)} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)} \cdot \frac{\phi\Big((X_i^{(t-1)} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\Big)}{\phi\Big((x_{\text{prop}} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\Big)}\right\}$$

and $x_{-i} = \big(\widetilde{X}_{<i}^{(t)},\ \widetilde{X}_{>i}^{(t-1)}\big)$. Similarly, for $t = m_0 - 1, \ldots, 0$, for each $i = n, \ldots, 1$, we draw $x_{\text{prop}} \sim \mathcal{N}\Big(x_i^\star,\ -\frac{1}{\zeta''(x_i^\star)}\Big)$, $u \sim \text{Unif}(0,1)$ and set $\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$ as before with the only difference that $x_{-i} = \big(\widetilde{X}_{<i}^{(t+1)},\ \widetilde{X}_{>i}^{(t)}\big)$.

3. **Output.** Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.