ROBUST DISCOVERY OF MUTATIONAL SIGNATURES USING POWER POSTERIORS

CATHERINE XUE, JEFFREY W. MILLER, SCOTT L. CARTER, AND JONATHAN H. HUGGINS

ABSTRACT. Mutational processes, such as the molecular effects of carcinogenic agents or defective DNA repair mechanisms, are known to produce different mutation types with characteristic frequency profiles, referred to as mutational signatures. Non-negative matrix factorization (NMF) has successfully been used to discover many mutational signatures, yielding novel insights into cancer etiology and targeted therapies. However, the NMF model is only a rough approximation to reality, and even small departures from this assumed model can have large negative effects on the accuracy and reliability of the results. We propose a new approach to mutational signatures analysis that improves robustness to misspecification by using a power posterior for a fully Bayesian NMF model, while employing a sparsity-inducing prior to automatically infer the number of active signatures. In extensive simulation studies, we find that our proposed approach recovers more true signatures with greater accuracy than current leading methods. On whole-genome sequencing data for six cancer types from the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, we find that our method is able to accurately recover more signatures than the current state-of-the-art.

1. INTRODUCTION

Carcinogenic processes such as UV radiation, smoking, defective DNA repair mechanisms, and naturally occuring biochemical reactions generate characteristic patterns of somatic mutations known as mutational signatures (Helleday et al., 2014; Koh et al., 2021; Nik-Zainal et al., 2012b). While these processes cannot be observed directly in patients, the cumulative effect of multiple processes on an individual tumor can be quantified using genome sequencing, and their distinct mutational signatures can be inferred using statistical modeling. Non-negative matrix factorization (NMF) models have proven effective in estimating mutational signatures as well as the mutational load due to each signature in each tumor sample (Alexandrov et al., 2013b; Nik-Zainal et al., 2012a). Mutational signatures analysis has contributed to novel insights in a variety of areas of cancer research (Alexandrov et al., 2013a, 2015, 2020; Li et al., 2020; Nik-Zainal et al., 2012b) and has emerging translations to clinical outcomes (Chakravarty and Solit, 2021).

Existing methods for mutational signatures analysis are fundamentally limited by the fact that they assume—either explicitly or implicitly—a particular probabilistic model for how mutations arise. However, any assumed model will only be a rough approximation to reality. Unfortunately, using an incorrect model—known as *model misspecification*—can lead to spurious inferences (Miller and Dunson, 2018). In particular, a key challenge in mutational signature discovery is determining the number of active signatures (Alexandrov et al., 2013a,b; Kim et al., 2016; Rosales et al., 2017), and standard statistical methods for this type of model selection problem tend to be extremely sensitive to misspecification (Cai et al.,

2021; Miller and Dunson, 2018). For instance, methods based on automatic relevance determination (ARD) (Kim et al., 2016; Tan and Févotte, 2013) and Bayesian spike-and-slab models (Legramanti et al., 2020) tend to overestimate the number of signatures when there is mild overdispersion (Zito and Miller, 2024). Of course, robustness to outliers can be obtained using overdispersed models such as negative binomial (Lyu et al., 2020), but this does not handle any other types of misspecification. Perhaps due to this sensitivity, the current leading methods rely heavily on *ad hoc* techniques such as manual filtering (Alexandrov et al., 2020) or neural networks (Nebgen et al., 2021) for determining the number of signatures.

In this paper, we show that leading methods for mutational signatures analysis are not robust to model misspecification, and we introduce a novel method that exhibits better performance under misspecification. Specifically, we investigate the degree to which misspecification causes existing methods to (1) fail to find important processes or (2) infer spurious processes that do not actually exist. Our findings indicate a lack of robustness in two widely used methods: SigProfilerExtractor (Alexandrov et al., 2013b, 2020; Islam et al., 2022) and SignatureAnalyzer (Alexandrov et al., 2020; Kim et al., 2016; Tan and Févotte, 2013); see Figure 1 for examples. To address these limitations, we propose a new method called *BayesPowerNMF*. By leveraging a power posterior (Miller and Dunson, 2018) and a sparsity-inducing prior (Zito and Miller, 2024), BayesPowerNMF provides nonparametric robustness to model misspecification and automated, principled selection of the number of latent processes. Through simulation studies and a comparison to existing methods on whole-genome data for six cancer types (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), we show that our BayesPowerNMF method finds more true processes and fewer spurious processes than competitors; see Figure 1 for two illustrations.

The article is organized as follows. In Section 2, we introduce our proposed methodology. We present simulation results comparing our method to leading existing methods in Section 3. Then, in Section 4, we compare the methods in an application to whole-genome data from the Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG). We conclude with a discussion, including limitations and possible directions for future work, in Section 5.

2. Methodology

In this section, we describe the standard Poisson NMF model, our sparsity-inducing prior, the power posterior, our technique for choosing the power, and the overall workflow of our BayesPowerNMF method. NMF is based on the assumption that the data matrix $\boldsymbol{X} = (X_{nm}) \in \mathbb{R}_+^{N \times M}$ can be approximated by a low-rank factorization of the form

$$\boldsymbol{X} \approx \boldsymbol{LR},$$

where the (unknown) matrices $\boldsymbol{L} = (L_{nk}) \in \mathbb{R}_+^{N \times K}$ and $\boldsymbol{R} = (R_{km}) \in \mathbb{R}_+^{K \times M}$ have nonnegative entries (Fu et al., 2019; Lee and Seung, 2000). Here, $\mathbb{R}_+ = [0, \infty)$ and we use bold font to indicate matrices. Since arbitrary multiplicative constants can be moved between $\boldsymbol{L}$ and $\boldsymbol{R}$ without affecting the product $\boldsymbol{LR}$, we constrain the rows of $\boldsymbol{R}$ to sum to 1, that is, $\sum_{m=1}^M R_{km} = 1$ for all $k = 1, \ldots, K$. Low-rank factorizations such as this often provide valuable insights into the latent structure giving rise to the data.

In mutational signatures analysis, the rows of $\boldsymbol{X}$ correspond to samples $n = 1, \ldots, N$, the columns of $\boldsymbol{X}$ correspond to a pre-defined set of M non-overlapping mutation types, and each entry $X_{nm} \in \{0, 1, 2, \ldots\}$ represents the number of times that mutation type m is observed
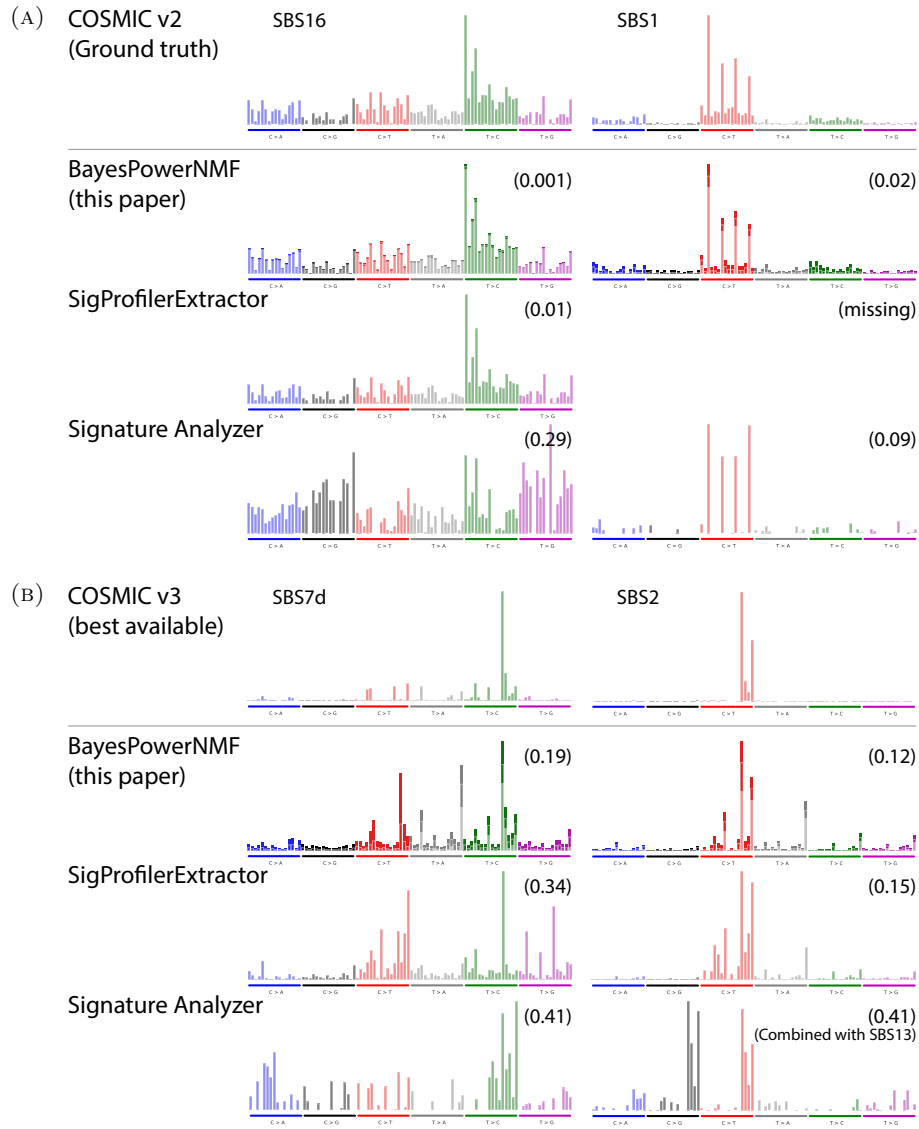
FIGURE 1. **(A)** Examples from simulated misspecified liver cancer data. For two ground truth signatures used to generate the data (SBS16 and SBS1 from COSMIC v2, shown in the top row), we show the best-matching signatures inferred by two leading methods (SigProfilerExtractor and SignatureAnalyzer) and our proposed method (BayesPowerNMF). The cosine errors between the inferred and true signatures are in parentheses. When the model is misspecified, SigProfilerExtractor tends to miss more signatures (such as SBS1 in this case) and SignatureAnalyzer tends to have significantly higher cosine errors compared to BayesPowerNMF. **(B)** Examples from real melanoma data. Same as in (A), but comparing with the closest match to SBS7d and SBS2 from COSMIC v3. For BayesPowerNMF, the bolded section of each bar indicates a 95% credible interval, quantifying the uncertainty in each signature. Meanwhile, SigProfilerExtractor and SignatureAnalyzer only provide point estimates that may be misleading in cases such as SBS7d for which there is high uncertainty.

in sample n. The idea of the NMF model is that mutations arise due to multiple unknown mutational processes, and each such process generates mutations according to a distinct profile of frequencies. The kth row of $\boldsymbol{R}$, say $R_k = (R_{k1}, \ldots, R_{kM})$, represents the frequencies with which mutation types $m = 1, \ldots, M$ occur under the kth mutational process, and R_k is referred to as the *mutational signature* of this process. Thus, K represents the number of mutational processes represented in the model, (that is, the number of latent factors). The loading L_{nk} represents the activity of process k in sample n.

2.1. **Model and sparsity-inducing prior.** Our methodology starts with the *Poisson NMF model*

$$X_{nm} \sim \mathsf{Poisson}\bigg(\sum_{k=1}^{K} L_{nk} R_{km} \bigg), \qquad n = 1, \ldots, N, \ m = 1, \ldots, M,$$

independently; or in more compact matrix notation,

$$\boldsymbol{X} \sim \mathsf{Poisson}(\boldsymbol{LR}). \tag{1}$$

This model is standard in mutational signatures analysis, and it can be justified from first principles (Zito and Miller, 2024). The parameters of the model are $(\boldsymbol{L}, \boldsymbol{R})$ and the likelihood is

$$\mathsf{Poisson}(\boldsymbol{X} \mid \boldsymbol{LR}) = \prod_{n=1}^{N} \prod_{m=1}^{M} \mathsf{Poisson}\big(X_{nm} \mid \textstyle\sum_{k=1}^{K} L_{nk} R_{km} \big).$$

For the prior on $\boldsymbol{R}$, we use independent M-dimensional Dirichlet priors on the rows,

$$R_k \sim \mathsf{Dirichlet}_M(\alpha_0, \ldots, \alpha_0), \qquad k = 1, \ldots, K,$$

where $\alpha_0 > 0$. For the prior on $\boldsymbol{L}$, since the number of active mutational processes is not known *a priori*, we employ a sparsity-inducing prior introduced by Zito and Miller (2024). Specifically, we place independent gamma priors on the loadings,

$$L_{nk} \sim \mathsf{Gamma}(a, \, a/\mu_k) \qquad n = 1, \ldots, N, \ k = 1, \ldots, K,$$

where $a > 0$ is the shape and a/μ_k is the rate, so that $\mathbb{E}(L_{nk}) = \mu_k$ and hence μ_k is the mean prior loading for the kth mutational process. Finally, we give μ_k an inverse gamma hyperprior,

$$\mu_k \sim \mathsf{InverseGamma}(a_0, b_0), \qquad k = 1, \ldots, K,$$

independently, where $a_0 = N_0 a + 1$, $b_0 = \varepsilon(a_0 - 1)$, and $\varepsilon > 0$ is a small nonzero constant. With these choices, the prior mean of μ_k becomes $\mathbb{E}(\mu_k) = \varepsilon$ and its full conditional mean is

$$\mathbb{E}(\mu_k \mid \boldsymbol{L}, \boldsymbol{R}, \boldsymbol{X}, \mu_1, \ldots, \mu_{k-1}, \mu_{k+1}, \ldots, \mu_K) = \frac{N_0}{N_0 + N} \varepsilon + \frac{N}{N_0 + N} \bigg(\frac{1}{N} \sum_{n=1}^{N} L_{nk} \bigg).$$

This choice of hyperprior has a compressive property that induces column-wise sparsity in the loadings matrix under the posterior distribution. More precisely, given the data $\boldsymbol{X}$, it shrinks the μ_k value for inactive processes down to $\mu_k \approx \varepsilon$, while having a moderate shrinkage effect on the μ_k value for active processes. Consequently, the loadings L_{nk} for inactive processes shrink to small values $\approx \varepsilon$, while the loadings for active processes are only slightly affected by the hyperprior (Zito and Miller, 2024).

In our experiments, we set $\alpha_0 = 0.5$, $a = 0.5$, $\varepsilon = 0.001$, and $N_0 = 10$. Choosing $\alpha_0 = 0.5$ and $a = 0.5$ makes the priors on R and L weakly informative. Specifically, the prior on R_k is the Jeffreys prior for a multinomial likelihood and the conditional prior variance of L_{nk} is $\text{Var}(L_{nk} \mid \mu_k) = 2\mu_k^2$. The value of ε does not matter much as long as it is significantly smaller than the smallest true loadings. The choice of $N_0 = 10$ was found to work well empirically, providing sparsity without shrinking the μ_k values too strongly.

2.2. **Power posterior.** The standard Poisson NMF model works well on data generated from the model itself, but its performance suffers when the model is not exactly correct, as we demonstrate in Section 3. To address this issue, we employ the power posterior technique for improving robustness to misspecification (Miller and Dunson, 2018). Letting $\pi_0(\theta)$ denote the prior density on $\theta = (\boldsymbol{L}, \boldsymbol{R}, \mu)$, the standard Bayesian posterior is

$$\pi(\theta \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \theta)\, \pi_0(\theta)}{Z(\boldsymbol{X})},$$

where $p(\boldsymbol{X} \mid \theta) = \mathsf{Poisson}(\boldsymbol{X} \mid \boldsymbol{L}\boldsymbol{R})$ and $Z(\boldsymbol{X}) = \int p(\boldsymbol{X} \mid \theta)\, \pi_0(\theta)\, d\theta$. For $\xi \in [0, 1]$, the *power posterior* is defined as

$$\pi_\xi(\theta \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \theta)^\xi\, \pi_0(\theta)}{Z_\xi(\boldsymbol{X})},$$

where $Z_\xi(\boldsymbol{X}) = \int p(\boldsymbol{X} \mid \theta)^\xi\, \pi_0(\theta)\, d\theta$ is the normalization constant. In particular, when $\xi = 0$ or $\xi = 1$, we recover the prior or the standard posterior, respectively; that is, $\pi_0(\theta \mid \boldsymbol{X}) = \pi_0(\theta)$ and $\pi_1(\theta \mid \boldsymbol{X}) = \pi(\theta \mid \boldsymbol{X})$. As shown by Miller and Dunson (2018), using a power posterior with $0 < \xi < 1$ provides nonparametric robustness for misspecified latent variable models – particularly when the latent dimensionality is unknown, such as in mutational signatures analysis. Furthermore, Medina et al. (2022) showed that—compared to using the standard posterior—using the power posterior of a misspecified model is closer to the standard posterior of a more flexible well-specified model, in terms of Kullback–Leibler divergence.

To perform inference, we use the Stan probabilistic programming system (Carpenter et al., 2017) to draw Markov chain Monte Carlo (MCMC) samples from the power posterior of the NMF model. Specifically, we define the log target density to be

$$\log \pi_\xi(\theta \mid \boldsymbol{X}) = \text{const} + \log \pi_0(\theta) + \xi \sum_{n=1}^{N} \sum_{m=1}^{M} \log \mathsf{Poisson}\big(X_{nm} \mid \textstyle\sum_{k=1}^{K} L_{nk} R_{km}\big).$$

Using Stan makes it a simple matter of multiplying the log-likelihood part of this function by ξ in order to target the power posterior rather than the standard posterior. For the MCMC sampler settings, we use four chains, each with 10,000 samples after a burn-in of 10,000 iterations. We choose the best of these four chains, defined as the one with the largest approximate marginal likelihood, using the approximation of Pritchard et al. (2000).

2.3. **Choosing the power.** The power ξ should be selected to reflect the severity of the model misspecification (Medina et al., 2022; Miller and Dunson, 2018). When there is no misspecification—that is, when the data are generated by the assumed model—we would ideally take $\xi = 1$, which corresponds to the standard Bayesian posterior. When there is misspecification, smaller values of ξ are preferable. However, it can be difficult to choose ξ in a data-driven way.

We propose the following simulation-driven approach to choosing ξ. First, select a set of *pilot signatures* (for example, we use a relevant subset of the COSMIC v2 signatures[1]; Alexandrov et al., 2015; Nik-Zainal et al., 2016) and estimate loadings to fit the input data using these pilot signatures. Next, generate pilot data from several plausible data generating processes (DGPs) based on the pilot signatures and corresponding estimated loadings. Then, fit each pilot data set using a range of ξ values, and select a value of ξ that performs well across all of the pilot data sets.

By exploring simulated data from DGPs that are not equal to the assumed model, but are anchored at parameters fit to the input data, we are able to select a ξ that yields robustness to plausible perturbations within an appropriate data-driven neighborhood of the model. This yields robustness to any DGPs that would require a similar power ξ, not just to the specific forms of DGP used to generate pilot data. Thus, it is not critical to get the DGPs exactly right – just to make them sufficiently rich and plausible.

2.4. **Simulation data generating processes.** Given a matrix of pilot signatures $\boldsymbol{R}^\star = (R_{km}^\star) \in \mathbb{R}_+^{K \times M}$ and corresponding matrix of loadings $\boldsymbol{L}^\star = (L_{nk}^\star) \in \mathbb{R}_+^{N \times K}$, we generate pilot data sets using the following four DGPs based on $\boldsymbol{R}^\star$ and $\boldsymbol{L}^\star$:

(1) The well-specified DGP generates data from the assumed Poisson NMF model, as in Equation (1):

$$\boldsymbol{X} \sim \mathsf{Poisson}(\boldsymbol{L}^\star \boldsymbol{R}^\star).$$

(2) The α-contaminated DGP generates data from the Poisson NMF model, but with an additional "contamination signature" $\tilde{R}_n$: for $\alpha \in (0,1)$,

$$\tilde{R}_n \sim \mathsf{Dirichlet}_M(1,\ldots,1), \qquad\qquad\qquad n = 1,\ldots,N,$$
$$\boldsymbol{X} \sim \mathsf{Poisson}\Big((1-\alpha)\boldsymbol{L}^\star \boldsymbol{R}^\star + \alpha(\boldsymbol{L}^\star \mathbf{1}_{K \times M}) \odot \tilde{\boldsymbol{R}}\Big), \qquad (2)$$

where $\mathbf{1}_{K \times M}$ denotes the $K \times M$ matrix of ones, $\odot$ denotes entry-wise multiplication, and $\tilde{\boldsymbol{R}}$ is the matrix with rows $\tilde{R}_1,\ldots,\tilde{R}_N$. The contamination signature $\tilde{R}_n$ is different for each sample, to reflect the possibility of tumor-specific variability arising from microenvironment effects or rare mutational processes. Equation (2) gives $\tilde{R}_n$ a loading of $\alpha \sum_{k=1}^K L_{nk}^\star$, and the existing loadings $L_{nk}^\star$ are downweighted by a factor of $1-\alpha$ to keep the total loading invariant.

(3) The γ-perturbed DGP generates each sample using slightly perturbed versions of each signature, with the perturbations being sample-specific: given $\gamma > 0$ and $\beta_1,\ldots,\beta_K > 0$,

$$R_k^{(n)} \sim \mathsf{Dirichlet}_M(\beta_k R_{k1}^\star,\ldots,\beta_k R_{kM}^\star), \qquad n = 1,\ldots,N, \ k = 1,\ldots,K,$$
$$X_n \sim \mathsf{Poisson}(L_n^\star \boldsymbol{R}^{(n)}), \qquad\qquad n = 1,\ldots,N,$$

where $L_n^\star$ is the nth row of $\boldsymbol{L}^\star$, $\boldsymbol{R}^{(n)}$ is the matrix with rows $R_1^{(n)},\ldots,R_K^{(n)}$, and X_n is the nth row of $\boldsymbol{X}$. The constants β_k control the size of perturbations, and are set such that the expected *cosine error* between $R_k^{(n)}$ and $R_k^\star$ is approximately γ, where the cosine error is defined as $1 - R_k^\mathsf{T} R_\ell^\star / (\|R_k\|\|R_\ell^\star\|)$; see Appendix B for details. This DGP reflects the possibility that each mutational process might behave slightly differently

---

across tumors, for instance due to the tumor microenvironment, different failure modes of the same DNA repair mechanism, or smoking different brands of cigarette.

(4) The κ-overdispersed DGP generates data with higher variance than the well-specified data, to reflect the fact that there is often additional variability due to sample- and mutation-type specific effects that may be biological or technical: given $\kappa > 1$,

$$\boldsymbol{X} \sim \mathsf{NegativeBinomial}(\boldsymbol{L}^\star \boldsymbol{R}^\star, \kappa),$$

where $\mathsf{NegativeBinomial}(\mu, \kappa)$ denotes the negative binomial distribution with mean μ and variance $\kappa\mu$.

In our experiments, we use $\alpha = 0.02$, $\gamma = 0.0025$, and $\kappa = 2$ for the settings of these DGPs. These settings are chosen to reflect plausible levels of biological and technical perturbations from the Poisson model. Specifically, $\alpha = 0.02$ equates to 2% of observed mutations originated from some source of contamination and $\gamma = 0.0025$ is a threshold where the simulated signatures are almost always still recognizable as coming from the original reference signature (99% of simulated signatures have cosine error < 0.1). We set $\kappa = 2$ to reflect the empirical overdispersion seen in loadings through exploratory analysis of real count data.

2.5. **Overall workflow.** To summarize, the overall workflow of our proposed method, *BayesPowerNMF*, consists of the following five stages (which are depicted in Figure 2):

(1) *Define pilot signatures and loadings.* To tailor the pilot data sets to the input data set, we define the pilot signatures $\boldsymbol{R}^\star$ and loadings $\boldsymbol{L}^\star$ as follows. Using non-negative least squares (Lawson and Hanson, 1987), we estimate loadings for the 30 COSMIC v2 reference signatures (Alexandrov et al., 2015; Nik-Zainal et al., 2016) for each individual in the cohort. We then drop any reference signatures that, based on the estimated loadings, do not contribute at least 2% of the total mutation count in the cohort or at least 10% of the mutation count of any individual, so that each remaining signature is well represented. The remaining loadings are not re-fit, to maintain the accuracy of the loadings.

(2) *Generate pilot data sets.* For each of four DGPs, generate a pilot data set(s) based on parameters $\boldsymbol{L}^\star$ and $\boldsymbol{R}^\star$, which results in J total simulated data sets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J$. As described in Section 2.4, we use $J = 4$: the well-specified DGP plus the three misspecified DGPs with parameters $\alpha = 0.02$, $\gamma = 0.0025$, and $\kappa = 2$.

(3) *Estimate parameters for a range of ξ values.* For each power ξ in a range of candidate values, for each pilot data set, we estimate $\boldsymbol{L}$ and $\boldsymbol{R}$ using the mean of the power posterior (Section 2.2) and match the estimated signatures R_k to the pilot signatures $R_k^\star$. Matching R_k to $R_k^\star$ is done using the Hungarian algorithm for optimal bipartite matching (Kuhn, 1955) with cosine error as the assignment cost.

(4) *Choose ξ to maximize accuracy across pilot data sets.* We choose the largest power ξ that accurately recovers $\boldsymbol{L}^\star$ and $\boldsymbol{R}^\star$ according to the following metrics (see Figure S4): *(i)* recovering as many true signatures as possible with cosine error < 0.3, and *(ii)* not inferring any *spurious signatures* that fail to match any true signature.

(5) *Apply power posterior to the input data.* Using the selected power ξ, we apply the power posterior to the original input data $\boldsymbol{X}$ to estimate $\boldsymbol{L}$ and $\boldsymbol{R}$. This yields the final estimates of $\boldsymbol{L}$ and $\boldsymbol{R}$ produced by the workflow.
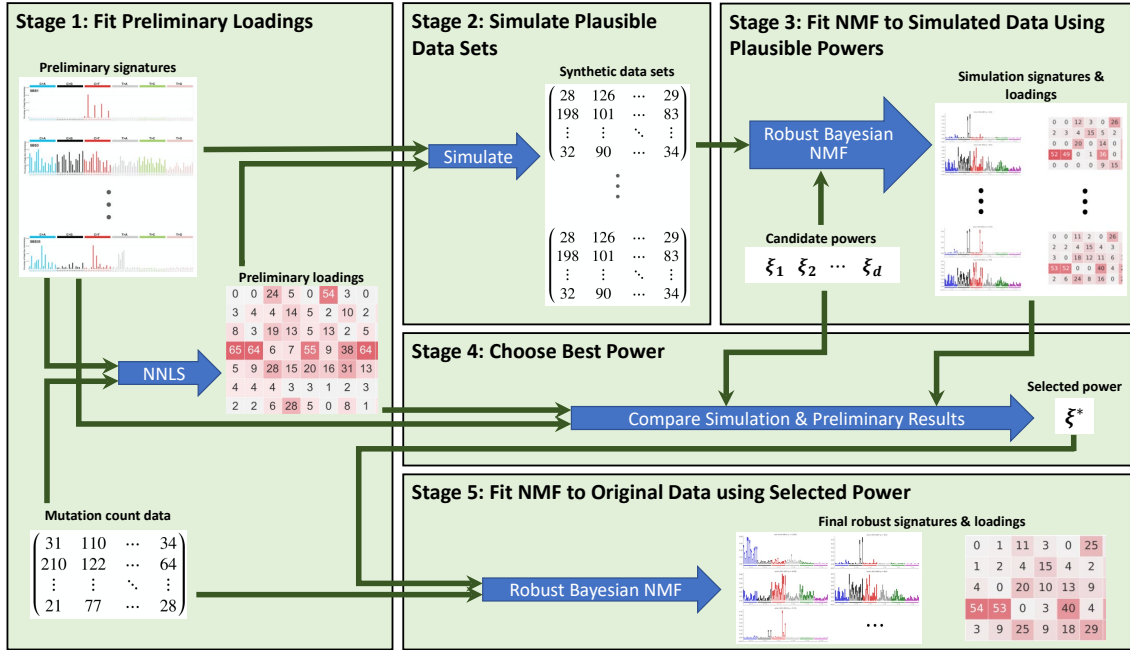
FIGURE 2. *BayesPowerNMF workflow for robust mutational signature discovery.* **(1)** Pilot signatures from a reference (such as COSMIC) are used to estimate loadings to fit the input data. **(2)** These signatures and loadings are used to simulate pilot data sets from DGPs representing plausible perturbations of the model. **(3)** The power posterior for the Bayesian NMF model is applied to each combination of candidate power ξ_i and simulated data set $\boldsymbol{X}_j$. **(4)** We identify the power that best recovers the pilot signatures and loadings across all of the pilot data sets. **(5)** Using the selected power, we apply the power posterior for the Bayesian NMF model to the original input data.

3. SIMULATIONS

In this section, we compare performance on synthetic data simulating six cancer types, using the four data generating processes described in Section 2.4. We compare three NMF methods for inferring mutational signatures: (1) SigProfilerExtractor, a version of the algorithm used to define the COSMIC signatures (Alexandrov et al., 2013b; Islam et al., 2022), (2) SignatureAnalyzer, based on a Bayesian point estimation algorithm (Tan and Févotte, 2013) used in several previous publications (Haradhvala et al., 2018; Kasar et al., 2015; Kim et al., 2016), and (3) BayesPowerNMF, our proposed method.

3.1. **Simulated data.** We generate synthetic data sets in the well-specified case for six cancer types (lung adenocarcinoma, stomach, melanoma, ovary, breast, and liver) and in the three misspecified cases (α-contaminated, γ-perturbed, and κ-overdispersed) for three cancer types (lung adenocarcinoma, ovary, and liver). These simulated data sets are generated as in stages 1 and 2 of the BayesPowerNMF workflow in Section 2.5, with the input data set being a mutation count matrix for the corresponding cancer type from the PCAWG project (Alexandrov et al.,

2020). In addition, we consider randomly subsampled data sets of sizes $20, 30, 50, 80, 120, 170$, and 230 from the simulated well-specified liver data set of 326 samples. Thus, altogether, starting from 6 real data sets (one for each cancer type), we generate 15 complete synthetic data sets (6 well-specified and 3×3 misspecified) plus 104 data sets subsampled from the complete well-specified liver data set (32 of size $N = 20$, 20 of size $N = 30$, 12 of size $N = 50$, and 10 each of the remaining sizes). Using each of these synthetic data sets as input, we then apply each of the three competing methods listed above: SigProfilerExtractor, SignatureAnalyzer, and BayesPowerNMF (following our workflow in Section 2.5).

3.2. **Performance evaluations.** We evaluate performance by measuring how well the ground-truth signatures and loadings are recovered in terms of several metrics. Here, ground truth is known since the data is simulated. First, we compute the optimal matching (in terms of cosine error) between estimated and true signatures. We then quantify how accurately each signature is recovered by computing the cosine error between each pair of matched signatures. We compute the precision and recall for recovering the true signatures by setting a threshold on cosine error and considering a true signature to be correctly recovered if the cosine error of its matching estimated signature is within this threshold. Precision is defined as (# correctly recovered) / (# estimated signatures) and recall is (# correctly recovered) / (# true signatures). Plots of precision and recall are then generated by varying the threshold. Additionally, we use bubble plots to display all of the matched pairs of signatures, along with their cosine errors and loadings, for each method and each data set. This provides a convenient visual summary of the results in a single plot.

3.3. **Simulation results.**

3.3.1. *Overall precision and recall.* Figure 3 shows the precision and recall for each of the three methods, over a range of cosine error thresholds from 0 to 0.2. These curves are averaged over all of the well-specified data sets (left side) and misspecified data sets (right side). Our BayesPowerNMF method provides the highest precision and recall at nearly every threshold. In particular, BayesPowerNMF has much higher recall in the misspecified settings. SigProfilerExtractor performs only somewhat worse than BayesPowerNMF in the well-specified settings, but its recall suffers under misspecification.

The signatures missed by SigProfilerExtractor tend to be ones with smaller true loading; see Figure S2. One reason for this may be because SigProfilerExtractor employs a consensus bootstrap approach that requires agreement across the estimates based on different bootstrap data sets. To compare the performance for each signature individually, we plot the cosine error from the ground truth for SigProfilerExtractor versus BayesPowerNMF in Figure S3; this illustrates that BayesPowerNMF tends to exhibit better performance on a signature-by-signature basis.

We find that SignatureAnalyzer exhibits worse performance across all scenarios, both in terms of precision and recall. The reason for this is that SignatureAnalyzer estimates many signatures, but the estimated signatures have a high cosine error relative to the matching true signatures. Although SignatureAnalyzer is based on a Bayesian model, it employs *maximum a posteriori* estimation rather than posterior uncertainty quantification, and does not benefit from the robustness properties of the power posterior.
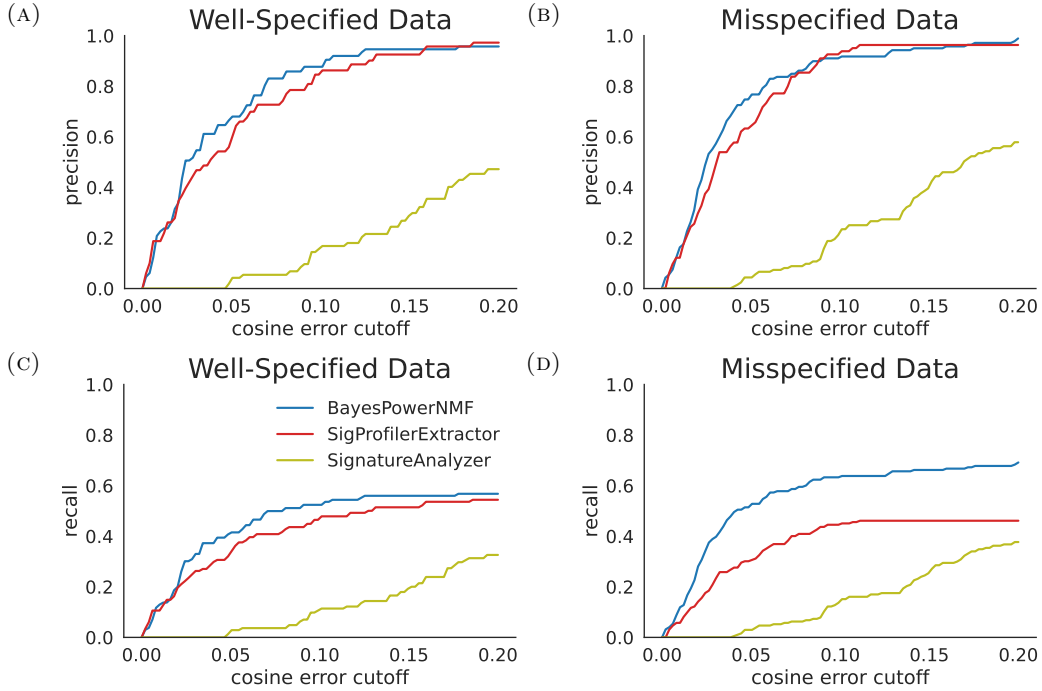
FIGURE 3. (Top) Precision curves for each method as a function of cosine error threshold, averaged across the simulated well-specified data sets (left) and misspecified data sets (right). (Bottom) Same for the recall curves. BayesPowerNMF has the highest precision and recall overall. SigProfilerExtractor performs nearly as well as BayesPowerNMF in the well-specified settings, but its recall is lower under misspecification. SignatureAnalyzer exhibits low precision and recall in these simulations.

3.3.2. *Lung adenocarcinoma results.* To understand what each method is doing at a more granular level, Figure 4 shows the cosine error and loading for each signature. The true COSMIC signatures are listed on the left, and each column summarizes the results from a given method on a given data set. The size of each bubble represents the loading given by that method to the estimated signature that was matched to the true signature listed on the left. The shade of the bubble represents the cosine error between the estimated signature and the matching true signature. The left-most column (labelled GT) shows the ground truth loadings used to generate the simulated data, and we order the rows (signatures) according to these ground truth loadings. Therefore, a method is performing well on a given data set if *(i)* it has bubbles for the same signatures as GT, *(ii)* the sizes of the bubbles are similar to GT, and *(iii)* the shade of the bubble is white or light gray. Methods with bubbles above the red line estimated a spurious signature that does not match to any of the true signatures used to generate the data.

For the simulated lung adenocarcinoma data sets, which have $N = 38$ samples, Figure 4 shows that BayesPowerNMF tends to recover more true signatures with lower cosine error than the other methods. SigProfilerExtractor performs second best, but misses some signatures
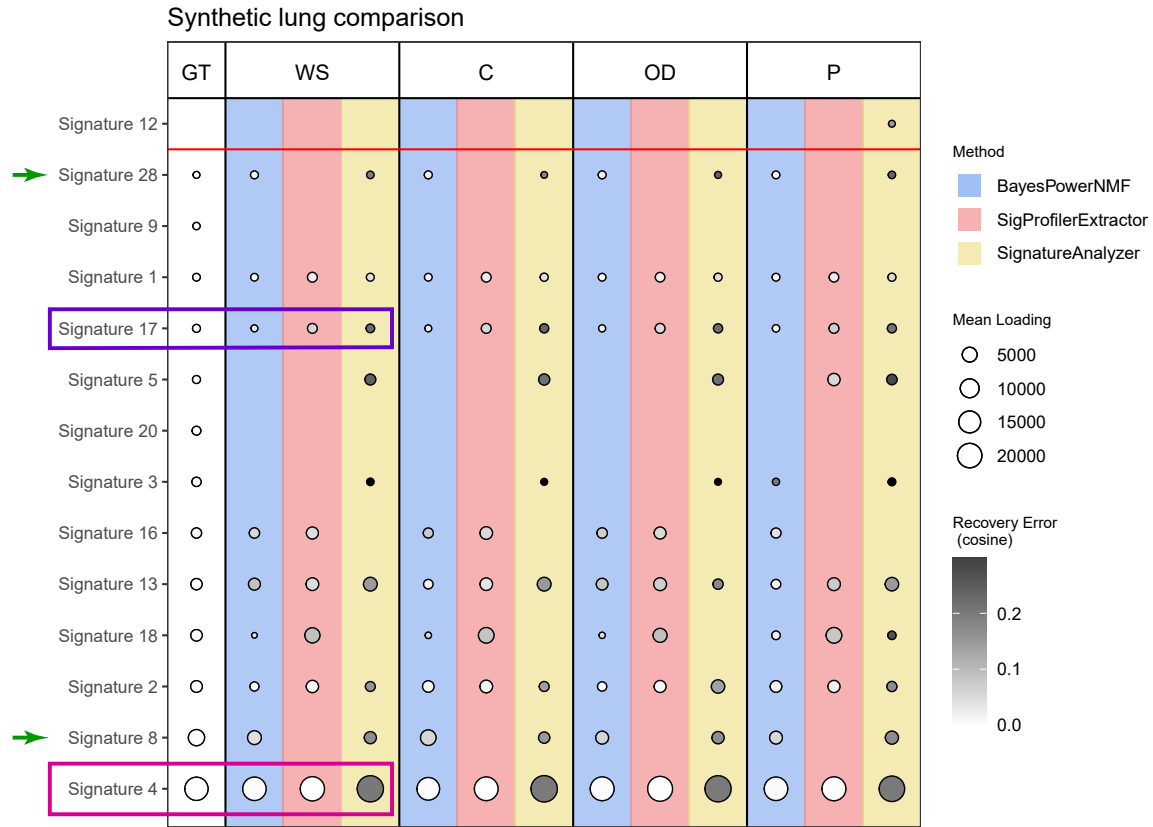
FIGURE 4. Comparison of results on simulated lung adenocarcinoma data sets, via a bubble plot showing the loading and cosine error of the estimated signatures. Rows represent the true COSMIC v2 signatures, ordered by the ground truth (GT) loading. Each column shows the results for a given method (blue = BayesPowerNMF, red = SigProfilerExtractor, yellow = SignatureAnalyzer) on a given data set (WS = well-specified, C = α-contaminated, OD = κ-overdispersed, P = γ-perturbed). Presence of a bubble means an estimated signature matched that true signature. Bubble size = estimated loading, bubble shade = cosine error between estimated and true, and a bubble above the red line represents an estimated signature that did not match any true signature used to generate the data. Overall, BayesPowerNMF consistently finds the most signatures and has the lowest cosine errors.

(in particular, Signatures 8 and 28, indicated by green arrows) and tends to have slightly higher cosine error. Note that Signature 8 was not recovered by SigProfilerExtractor in any of the cases, even though this signature has a large true loading (second largest out of all the signatures). SignatureAnalyzer does recover many signatures that match to the true signatures, but they have very high cosine error, even for signatures with high true loading such as Signature 4. On the γ-perturbed data, SignatureAnalyzer also produces a spurious signature that matches to COSMIC Signature 12, which was not used to generate the data.

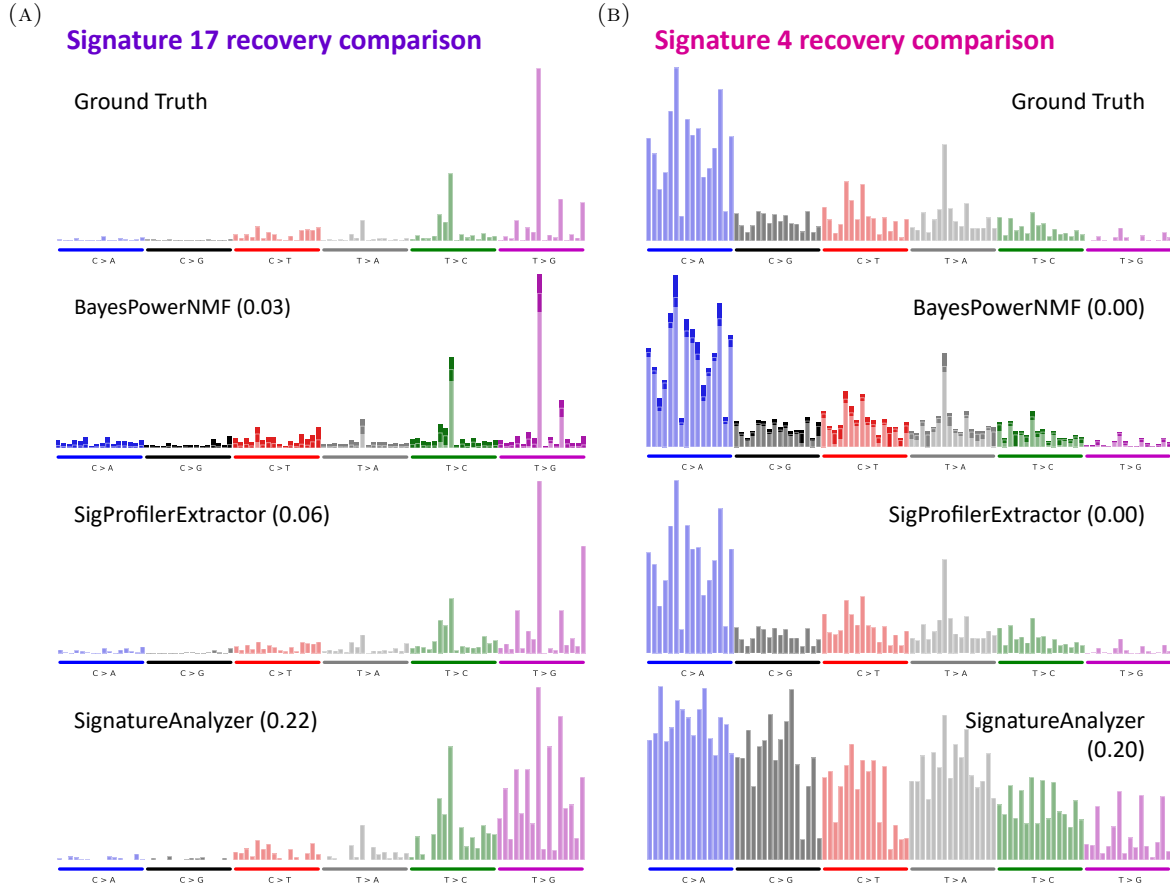(A)   (B)

FIGURE 5. Comparison of estimated signatures on simulated lung adenocarcinoma data sets. **(A)** True Signature 17 and matching estimated signatures (along with their cosine error, in parentheses) for each method on the WS data. For BayesPowerNMF, the bolded section of each bar for represents 95% credible interval and the thin white line in the middle represents the posterior mean. **(B)** Same panel A but for Signature 4.

Figure 5A shows true Signature 17 and the matching estimated signatures for each method on the well-specified (WS) data (corresponding to the purple box in Figure 4). This is an example of a signature with relatively small true mean loading and, in comparison with Signature 4, it also has many more near-zero values in the signature itself. BayesPowerNMF recovers the signature with the lowest cosine error (0.03), followed by SigProfilerExtractor (0.06), and then SignatureAnalyzer (0.22). BayesPowerNMF also provides uncertainty quantification in the signature vector, indicated in the plot by the darker region at the top of each bar, which represents a 95% credible interval.

Similarly, Figure 5B shows true Signature 4 and the matching estimated signatures on the WS data (corresponding to the magenta box in Figure 4). This is an example of a signature with large true mean loading (the largest in this simulated data). All three methods estimate a signature that matches to Signature 4, with estimated loading similar to the true loading.
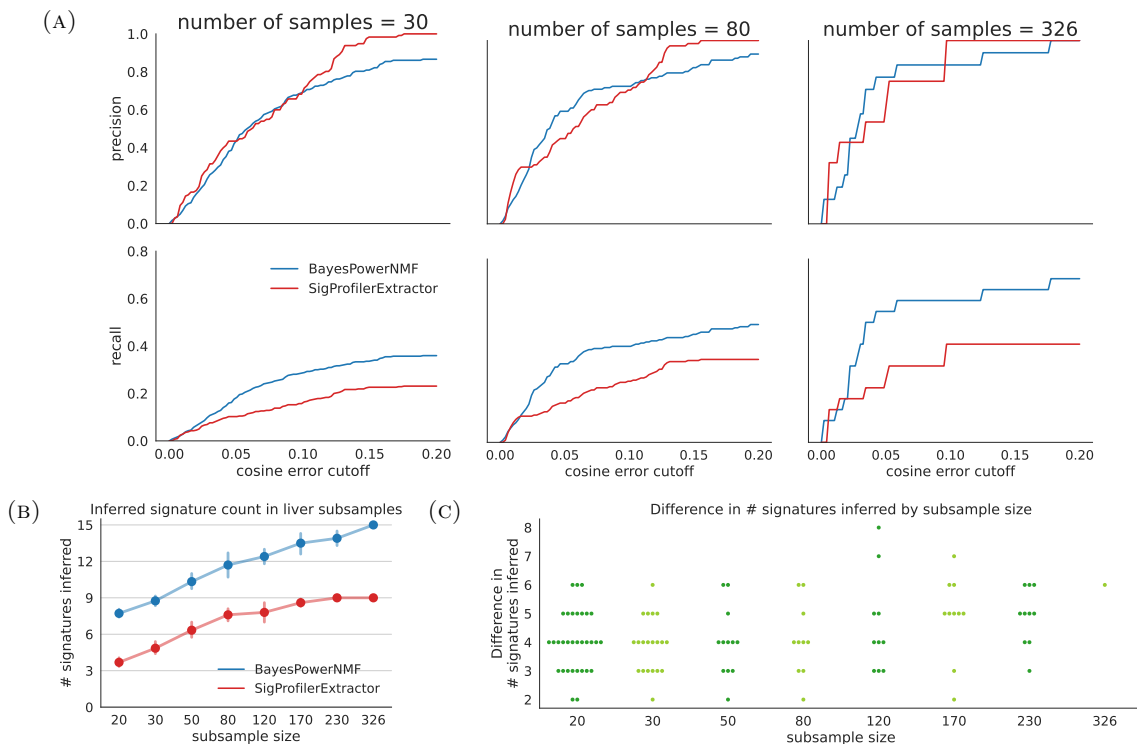
FIGURE 6. Results for subsampled liver cancer data sets of increasing size. **(A)** Precision and recall of matched signatures, as a function of the cosine error threshold for defining a correct match. BayesPowerNMF almost always strictly dominates SigProfilerExtractor in recall at all sample sizes, and has comparable or better precision for cosine error thresholds values of practical utility (< 0.1). **(B)** Number of signatures estimated (with whiskers representing 95% confidence intervals) at each sample size. **(C)** Difference between number of signatures estimated by BayesPowerNMF and SigProfilerExtractor for each subsampled data set. BayesPowerNMF recovers more signatures than SigProfilerExtractor at all sample sizes. The true number of signatures is 21.

Furthermore, BayesPowerNMF and SigProfilerExtractor both recover the true signature with very low cosine error (< 0.005). However, SignatureAnalyzer's matching signature has a high cosine error (0.2). For BayesPowerNMF, the 95% credible regions (bolded regions of the bars) tend to indicate less uncertainty in the estimates for Signature 4 compared to 17, especially in the mutation types with smaller rates; this makes sense since Signature 4 has higher true mean loading and thus there is more information about it in the data.

3.3.3. *Subsampled liver cancer data sets of varying size.* Next, we evaluate the effects of sample size and sampling variability by considering the subsampled data sets of varying size taken from the complete simulated well-specified liver cancer data set of $N = 326$ samples. The full cohort of $N = 326$ samples was generated using 21 ground truth signatures. From this, we repeatedly subsampled data sets of various sizes and ran each method on each subsampled data set. Figure 6 summarizes the results for BayesPowerNMF and SigProfilerExtractor on

these data; SignatureAnalyzer is excluded from this analyses due to its overall poor performance. Figure 6A plots the precision and recall curves as a function of cosine error threshold, like in Figure 3. As before, we see that BayesPowerNMF outperforms SigProfilerExtractor, particularly in terms of recall. SigProfilerExtractor does exhibit slightly higher precision at higher cosine error thresholds, but this comes at the expense of much lower recall. Figure 6B shows the number of signatures inferred by each method as a function of the size of the data set. This shows that BayesPowerNMF requires a much smaller sample size to recover the same number of signatures as SigProfilerExtractor. For instance, BayesPowerNMF recovers an average of 9 signatures on a data set of size $N = 30$, whereas SigProfilerExtractor requires $N = 230$ to recover this many signatures on average. Thus, BayesPowerNMF accurately recovers the same number of signatures with a fraction of the sample size.

To examine the effects of sampling variability, Figure 6C shows the difference in the number of signatures inferred by BayesPowerNMF and SigProfilerExtractor for each individual subsampled data set. On every individual data set, BayesPowerNMF always recovers at least 2 more signatures than SigProfilerExtractor, and sometimes as many as 7 or 8 more.

3.4. **Summary of simulation results.** Our simulation results demonstrate that BayesPowerNMF has improved performance compared to leading methods—enabling the discovery of a greater number of meaningful signatures—while also being robust to plausible model misspecification. One downside of BayesPowerNMF is that the additional robustness obtained in misspecified settings comes at a cost of reduced performance in well-specified settings compared to using the standard Bayesian posterior; see Figure S1. Overall, we find that BayesPowerNMF has the following specific advantages.

Higher precision. Compared to SigProfilerExtractor and SignatureAnalyzer, BayesPowerNMF tends to have higher precision. This holds across different DGPs (Figures 3 and S3A) and across sample sizes (Figure 6A), especially for smaller cosine error cutoffs that would produce recognizable signature matches (for instance, 0.05 or 0.1).

Higher recall. BayesPowerNMF consistently outperforms the competitors in terms of recall, particularly in misspecified cases (Figures 3 and 6A and Table S1). In the misspecified cases, SigProfilerExtractor tends to be too conservative, inferring too few signatures (Figure 3 and Table S2). SignatureAnalyzer sometimes infers more signatures, but many of the estimated signatures either have very large cosine error (> 0.2) or do not even get matched to one of the true signatures used to generate the data.

Uncertainty quantification. As a fully Bayesian method, BayesPowerNMF quantifies uncertainty in the signatures as well as the loadings via posterior samples produced by MCMC (Figures 5A and 5B). We find that the model's reported uncertainty in the signatures is correlated with the actual recovery error (Figure S3B), indicating that the uncertainty quantification is meaningful. In contrast, SigProfilerExtractor or SignatureAnalyzer do not provide uncertainty quantification in either the signatures or the loadings.

Few spurious signatures. We consider an estimated signature to be "spurious" if it is matched to a COSMIC signature that was not used to generate the simulated data set. Out of the 15 complete simulated data sets we consider, BayesPowerNMF infers a spurious signature in just one data set (see Tables S1 and S2). SignatureAnalyzer inferred spurious signatures in several synthetic data sets, even in the well-specified case (see Table S1).

4. Application

We compare the results of BayesPowerNMF and SigProfilerExtractor on whole-genome sequencing (WGS) data from the PCAWG project (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). SignatureAnalyzer is excluded from this comparison due to its overall poor performance. We consider the PCAWG WGS data for the same six cancer types as in the simulation study: lung adenocarcinoma ($N = 38$), stomach ($N = 75$), melanoma ($N = 107$), ovary ($N = 113$), combined breast cancers ($N = 214$), and liver ($N = 326$). While any set of mutation types could be used to define the input count data matrices, we consider the 96 types of single-base substitutions (SBSs) since these are the most commonly used set for mutational signatures analyses. We run BayesPowerNMF and SigProfilerExtractor on the $N \times 96$ count data matrix for each cancer type separately.

To evaluate performance, we use the COSMIC v3 signatures[2] as a proxy for the true signatures. (Note, however, that we still use COSMIC v2 in the BayesPowerNMF workflo when defining pilot signatures for calibration.) The COSMIC signature database is a curated reference set of mutational signatures that is commonly used in cancer genomics studies. The current versions of the COSMIC signatures are based on 2,780 cancer genomes from the PCAWG project—along with a number of other available cancer genomes—spanning a wide range of cancer types (Alexandrov et al., 2020; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). The COSMIC v3 signatures were defined by running a previous iteration of SigProfilerExtractor on each cancer type separately, and combining the results to construct a set of consensus signatures. Thus, one would expect SigProfilerExtractor to perform very well in these comparisons, since it was instrumental in defining our proxy for ground truth.

Since the actual ground truth is unknown on real data, we quantify method accuracy using using three proxies for quality. First, we compare the number of signatures inferred by each method on each cancer type. For a given amount of data, it is desirable to infer more signatures since this indicates improved recall. However, it is also important that the inferred signatures correspond to true signatures and are accurately estimated, in other words, it is important to have high precision. Thus, second, we compute the error with which the COSMIC signatures are reconstructed. More precisely, we assign the inferred signatures to the best-matching COSMIC v3 signatures using the Hungarian algorithm, and compute the cosine error between each matched pair. It is desirable to have smaller error, since this indicates greater accuracy in recovering the current best understanding of ground truth. Third, we compare the estimated loadings for each method to the loadings presented by the PCAWG consortium, which they obtained by running non-negative least squares (NNLS) using the complete set of COSMIC signatures, rather than using NMF. We refer to these as the "COSMIC+NNLS" loadings.

4.1. **Application results.** Figure 7 summarizes the results; see the description of Figure 4 and its caption for the interpretation of this plot.

Number of signatures inferred. In four of the six cancer types, BayesPowerNMF estimates more signatures than SigProfilerExtractor (Figure 7 and Table S3). Furthermore, in one of the two cancer types where SigProfilerExtractor infers more signatures (melanoma), some of the estimated signatures appear to be near duplicates rather than authentically different signatures (Figure S5). In many cases, BayesPowerNMF recovers almost as many signatures

---

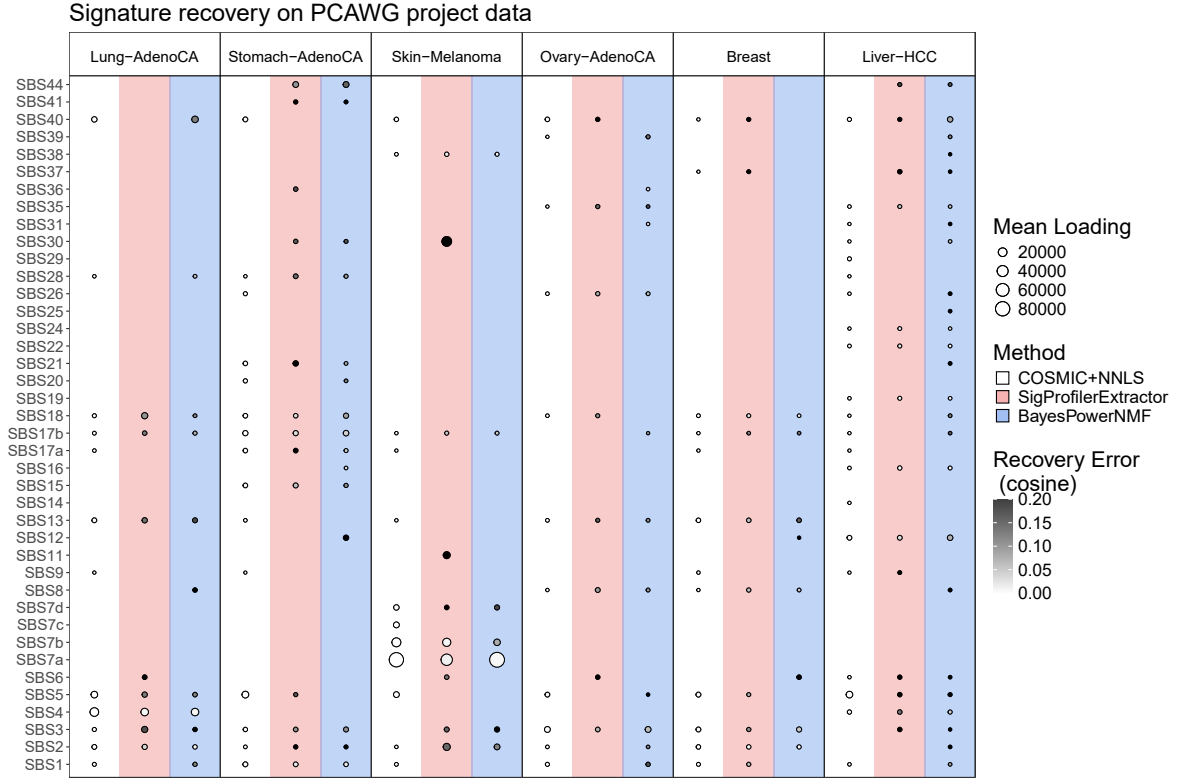[2]https://cancer.sanger.ac.uk/signatures/downloads/

FIGURE 7. Mean loading and cosine error between each estimated signature and its matching COSMIC v3 signature. Results are shown for each method applied to the PCAWG data for each of six cancer types. See the description of Figure 4 for explanation of the plot. COSMIC+NNLS uses non-negative least squares to estimate loadings for the COSMIC v3 signatures.

as COSMIC+NNLS (Figure 7 and Table S3), even though BayesPowerNMF is only using a small subset of data from one cancer type whereas COSMIC+NNLS is effectively using ≈ 3,000 samples.

Reconstruction error. Both methods perform roughly equally well in terms of the accuracy with which they reconstruct COSMIC signatures. Indeed, the difference between the average cosine errors for BayesPowerNMF and SigProfilerExtractor (0.146 and 0.133, respectively) is not statistically significant ($p = 0.9265$, Mann–Whitney rank test). This is despite the fact that (1) SigProfilerExtractor is at an advantage since the COSMIC signatures were constructed using its algorithm and (2) BayesPowerNMF recovers a greater proportion of signatures (Table S3).

Loadings. In most cases, the loadings estimated by BayesPowerNMF and SigProfilerExtractor are of similar magnitude to one another and to the COSMIC+NNLS loadings, at least for the signatures that they have in common (Figure 7). However, on the melanoma data, SigProfilerExtractor gives large loadings to two apparently spurious signatures that have high cosine error with their nearest COSMIC v3 matches (SBS30 and SBS11).

Overall, the real data results lead to similar conclusions as the simulation study results. These findings indicate that our proposed method provides improved accuracy and robustness to model misspecification compared to the state-of-the-art, both on synthetic examples and real biological data. It is notable that SigProfilerExtractor appears to perform more poorly on the real PCAWG data than on our simulated misspecified counts data, compared to BayesPowerNMF. This suggests that the modes of misspecification in real mutation counts data are more complex than is captured in each of our DGPs for misspecified counts, but this is accounted for in BayesPowerNMF due to the flexibility of the power posterior.

Finally, we note that the BayesPowerNMF workflow uses COSMIC v2 for the pilot signatures even on this PCAWG data, where we ultimately compare to COSMIC v3. BayesPowerNMF still successfully infers v3 signatures on this data, including ones with no corresponding signature in v2 (specifically, SBS31 and above in Figure 7). This suggests that the precise choice of pilot signatures does not strongly affect the final results from BayesPowerNMF.

5. Discussion

Our results elucidate specific shortcomings of two of the standard methods for mutational signature discovery, SigProfilerExtractor and SignatureAnalyzer, and highlight the benefits of carrying out thorough simulation studies to better understand the failure modes of any given method before proceeding to analysis of real data. SigProfilerExtractor uses a consensus bootstrap approach, which may explain why it tends to miss more signatures with small loadings (Figure S2). Specifically, since each bootstrap sample uses only $\approx 63\%$ of the data on average, it makes sense that SigProfilerExtractor would have lower power to discover mutational processes, particularly when they have smaller loading (Figure S2). BayesPowerNMF does not have this issue since it uses all of the available data. This suggests that the SigProfilerExtractor results might be improved by using a version of bootstrap with continuous weights, such as the Bayesian bootstrap, rather than multinomial bootstrap.

Furthermore, while SigProfilerExtractor is generally more conservative in the sense that it recovers fewer signatures than BayesPowerNMF, it can still overfit the data in cancer types with high mutational burden (such as melanoma, see Figure S5) or many samples (such as liver). This makes sense, since these are the cases in which we would expect the negative effects of misspecification to be most pronounced (Miller and Dunson, 2018). BayesPowerNMF avoids this issue by employing a power posterior to improve robustness to misspecification.

Similarly, a disadvantage of SignatureAnalyzer is that it relies heavily on the correctness of the Poisson NMF model (Zito and Miller, 2024). While SigProfiler's use of bootstrapping can somewhat mitigate the effects of misspecification (Huggins and Miller, 2023, 2024) and BayesPowerNMF's use of the power posterior provides robustness, SignatureAnalyzer has no mechanism to compensate for model misspecification. As a result, it appears that SignatureAnalyzer is particularly negatively affected by misspecification. Furthermore, SignatureAnalyzer uses an estimation algorithm based on several heuristic approximations to the objective function (Tan and Févotte, 2013). The use of these heuristics may explain why SignatureAnalyzer was not able to accurately recover the true signatures even on the simulated well-specified data sets, a setting in which we would expect a model-based method to perform well. In contrast, BayesPowerNMF does not suffer from this issue, since it does not employ any approximations other than MCMC sampling. Consequently, BayesPowerNMF exhibits good performance on the well-specified data, as expected.

The main disadvantage of BayesPowerNMF is that, with currently available inference techniques, it becomes computationally prohibitive for larger data sets. Improving the scalability of BayesPowerNMF is an important direction for future work. A related disadvantage is the somewhat complicated workflow for selecting the power ξ to be used in the power posterior. It would be preferable to have a simpler method of determining an appropriate power for a given set of plausible data generating processes.

Overall, BayesPowerNMF yields superior performance by *(i)* using all the available data and employing a full Bayesian model to extract as much information as possible from the data, while *(ii)* using a power posterior to obtain robustness to model misspecification without assuming a particular form of misspecification. More generally, our methodology can be used to perform non-negative matrix factorization for other applications as well, since it is not limited to mutational signatures analysis. It would be interesting to explore other applications of the method in future work.

Acknowledgments

References

L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, A. P. C. G. Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500: 415–421, Aug. 2013a. ISSN 1476-4687. doi: 10.1038/nature12477.

L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3:246–259, Jan. 2013b. ISSN 2211-1247. doi: 10.1016/j.celrep.2012.12.008.

L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton. Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47:1402–1407, Dec. 2015. ISSN 1546-1718. doi: 10.1038/ng.3441.

L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, PCAWG Mutational Signatures Working Group, G. Getz, S. G. Rozen, M. R. Stratton, and PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature*, 578:94–101, Feb. 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-1943-3.

D. Cai, T. Campbell, and T. Broderick. Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, pages 1158–1169. PMLR, 2021.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 2017.

D. Chakravarty and D. B. Solit. Clinical cancer genomic profiling. *Nature Reviews. Genetics*, 22:483–501, Aug. 2021. ISSN 1471-0064. doi: 10.1038/s41576-021-00338-8.

X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma. Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications. *IEEE Signal Processing Magazine*, 36:59–80, 2019. ISSN 1558-0792. doi: 10.1109/MSP.2018.2877582.

N. J. Haradhvala, J. Kim, Y. E. Maruvka, P. Polak, D. Rosebrock, D. Livitz, J. M. Hess, I. Leshchiner, A. Kamburov, K. W. Mouw, M. S. Lawrence, and G. Getz. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nature Communications*, 9(1):1–9, Apr. 2018.

T. Helleday, S. Eshtad, and S. Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews. Genetics*, 15:585–598, Sept. 2014. ISSN 1471-0064. doi: 10.1038/nrg3729.

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

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

S. M. A. Islam, M. Díaz-Gay, Y. Wu, M. Barnes, R. Vangara, E. N. Bergstrom, Y. He, M. Vella, J. Wang, J. W. Teague, P. Clapham, S. Moody, S. Senkin, Y. R. Li, L. Riva, T. Zhang, A. J. Gruber, C. D. Steele, B. Otlu, A. Khandekar, A. Abbasi, L. Humphreys, N. Syulyukina, S. W. Brady, B. S. Alexandrov, N. Pillay, J. Zhang, D. J. Adams, I. Martincorena, D. C. Wedge, M. T. Landi, P. Brennan, M. R. Stratton, S. G. Rozen, and L. B. Alexandrov. Uncovering novel mutational signatures by *de novo* extraction with SigProfilerExtractor. *Cell Genomics*, 2:100179, Nov. 2022. ISSN 2666-979X. doi: 10.1016/j.xgen.2022.100179.

S. Kasar, J. Kim, R. Improgo, G. Tiao, P. Polak, N. Haradhvala, M. S. Lawrence, A. Kiezun, S. M. Fernandes, S. Bahl, C. Sougnez, S. Gabriel, E. S. Lander, H. T. Kim, G. Getz, and J. R. Brown. Whole-genome sequencing reveals activation- induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications*, 6(8866):5446, Nov. 2015.

J. Kim, K. W. Mouw, P. Polak, L. Z. Braunstein, A. Kamburov, D. J. Kwiatkowski, J. E. Rosenberg, E. M. Van Allen, A. D'Andrea, and G. Getz. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*, 48: 600–606, June 2016. ISSN 1546-1718. doi: 10.1038/ng.3557.

G. Koh, A. Degasperi, X. Zou, S. Momen, and S. Nik-Zainal. Mutational signatures: emerging concepts, caveats and clinical applications. *Nature Reviews. Cancer*, 21:619–637, Oct. 2021. ISSN 1474-1768. doi: 10.1038/s41568-021-00377-7.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

C. L. Lawson and R. J. Hanson. *Solving least squares problems*. Society for Industrial and Applied Mathematics, 1987. ISBN 9780898713565. doi: 10.1137/1.9781611971217.

D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13, 2000.

S. Legramanti, D. Durante, and D. B. Dunson. Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752, 2020.

Y. Li, N. D. Roberts, J. A. Wala, O. Shapira, S. E. Schumacher, K. Kumar, E. Khurana, S. Waszak, J. O. Korbel, J. E. Haber, M. Imielinski, PCAWG Structural Variation Working Group, J. Weischenfeldt, R. Beroukhim, P. J. Campbell, and PCAWG Consortium. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578:112–121, Feb. 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1913-9.

X. Lyu, J. Garret, G. Rätsch, and K.-V. Lehmann. Mutational signature learning with supervised negative binomial non-negative matrix factorization. *Bioinformatics (Oxford, England)*, 36:i154–i160, July 2020. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa473.

M. A. Medina, J. L. M. Olea, C. Rush, and A. Velez. On the robustness to misspecification of α-posteriors and their variational approximations. *The Journal of Machine Learning Research*, 23(1):6579–6629, 2022.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2018.

B. T. Nebgen, R. Vangara, M. A. Hombrados-Herrera, S. Kuksova, and B. S. Alexandrov. A neural network for determination of latent dimensionality in non-negative matrix factorization. *Machine Learning: Science and Technology*, 2(2):025012, 2021.

S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jönsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. Langerød, A. Tutt, J. W. M. Martens, S. A. J. R. Aparicio, Å. Borg, A. V. Salomon, G. Thomas, A.-L. Børresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, M. R. Stratton, and B. C. W. G. of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149:979–993, May 2012a. ISSN 1097-4172. doi: 10.1016/j.cell.2012.04.024.

S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton, A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R. Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Jönsson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerød, S. A. J. R. Aparicio, A. Tutt,

A. M. Sieuwerts, Å. Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A.-L. Børresen-Dale, P. A. Futreal, M. R. Stratton, P. J. Campbell, and Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers. *Cell*, 149:994–1007, May 2012b. ISSN 1097-4172. doi: 10.1016/j.cell.2012.04.023.

S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjærde, A. Langerød, M. Ringnér, S.-M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. J. Hooijer, S. J. Jang, D. R. Jones, H.-Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J.-Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporté, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodríguez-González, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. v. Veer, A. Tutt, S. Knappskog, B. K. T. Tan, J. Jonkers, A. Borg, N. T. Ueno, C. Sotiriou, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. Van De Vijver, J. W. M. Martens, A.-L. Børresen-Dale, A. Viari, A. L. Richardson, G. Kong, G. Thomas, and M. R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, June 2016.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, June 2000. ISSN 0016-6731. doi: 10.1093/genetics/155.2.945.

R. A. Rosales, R. D. Drummond, R. Valieris, E. Dias-Neto, and I. T. da Silva. signeR: An empirical Bayesian approach to mutational signature discovery. *Bioinformatics*, 33(1):8–16, Jan. 2017.

V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1592–1605, May 2013.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, Feb. 2020.

A. Zito and J. W. Miller. Compressive Bayesian non-negative matrix factorization for mutational signatures analysis. *arXiv preprint arXiv:2404.10974*, 2024.

Appendix A. Additional empirical results

	Lung (13)	Stomach (21)	Skin (14)	Ovary (12)	Breast (12)	Liver (21)
SPE	7	11	6	7	**10**	9
SA	5 +4	5 +8 +1	6 +4 +4	5 +2	3 +5 +2	8 +6
BPN	**9**	10 +1	5 +1	7	7	15
$BPN_{\xi=1}$	9 +2	**14** +1	**13**	**10**	12 +3	**16** +1

TABLE S1. Mutational signature discovery results for simulated data, comparing SigProfilerExtractor, SignatureAnalyzer, and BayesPowerNMF on well-specified data for six cancer types. The number of COSMIC v2 signatures used to generate the data is in parentheses after each cancer type. In each entry of the form X +Y +Z in the table, X = the number of estimated signatures that were matched to a ground truth signature with a cosine error of < 0.2, Y = the number of estimated signatures that were matched to a ground truth signature with a cosine error of ≥ 0.2, and Z = the number of estimated signatures that were matched to a COSMIC v2 signature that was not used to generate the data. Bold indicates the best performing method in each case. (SPE = SigProfilerExtractor; SA = SignatureAnalyzer; BPN = BayesPowerNMF; $BPN_{\xi=1}$ = BayesPowerNMF with power of 1, which corresponds to the standard posterior.)

	Lung (13)			Ovary (12)			Liver (21)		
	SPE	SA	BPN	SPE	SA	BPN	SPE	SA	BPN
well-specified	7	5 +4	**9**	7	5 +2	7	9	8 +6	**15**
α-contaminated	7	5 +4	**9**	5 +1	5 +2	5	9	7 +5	**15**
γ-perturbed	7	3 +7 +1	**10**	5 +1	5 +2	8	9	10 +5	**17**
κ-overdispersed	7	4 +5	**9**	5	5 +2	**9** +1	9	8 +7	**15**

TABLE S2. Same as Table S1 but for both the well-specified and misspecified settings, for three cancer types.
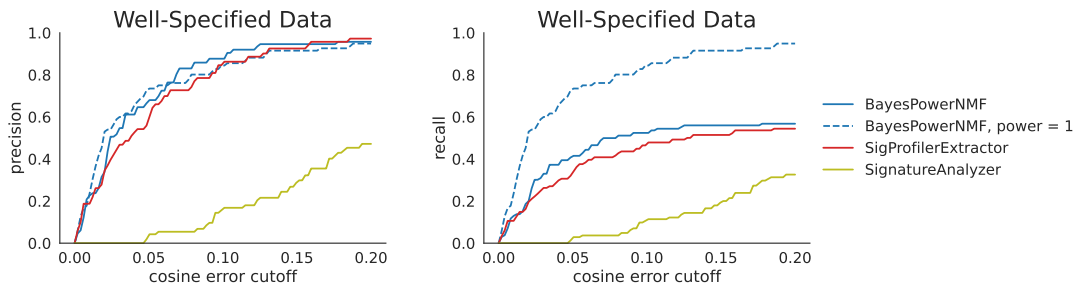
FIGURE S1. Same as Figure 3 but also showing the precision and recall for the standard Bayesian posterior (blue dashed lines) for the same NMF model. This corresponds to BayesPowerNMF with a power of $\xi = 1$. When the model is correct (that is, in the well-specified case), the standard posterior exhibits higher recall than the power posterior, while maintaining comparable precision.
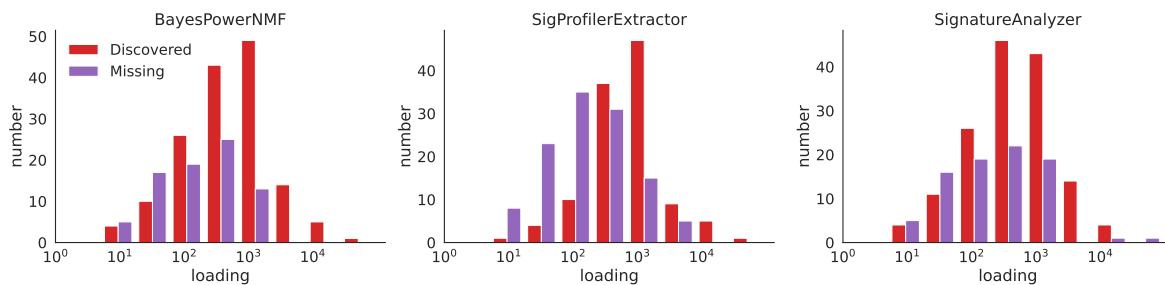

FIGURE S2. Histograms of ground-truth loadings for pilot signatures discovered and missed by BayesPowerNMF, SigProfilerExtractor, and SignatureAnalyzer across all simulated data sets. SigProfilerExtractor systematically misses many more signatures with small mean loading in the pilot data set.

	Lung		Stomach			Skin			Ovary			Breast			Liver		
BPN	10	+1	9	+1	+5	5	+1		8	+1	+3	6		+2	12	+4	+9
SPE	7	+1	7	+3	+4	5	+1	+4	6	+1	+1	8	+2		7	+4	+3
COSMIC	12		15			12			11			12			21		

TABLE S3. Number of signatures inferred by each method on the PCAWG data for six cancer types, treating COSMIC+NNLS as ground truth. (SPE = SigProfilerExtractor; BPN = BayesPowerNMF; COSMIC = COSMIC+NNLS.)
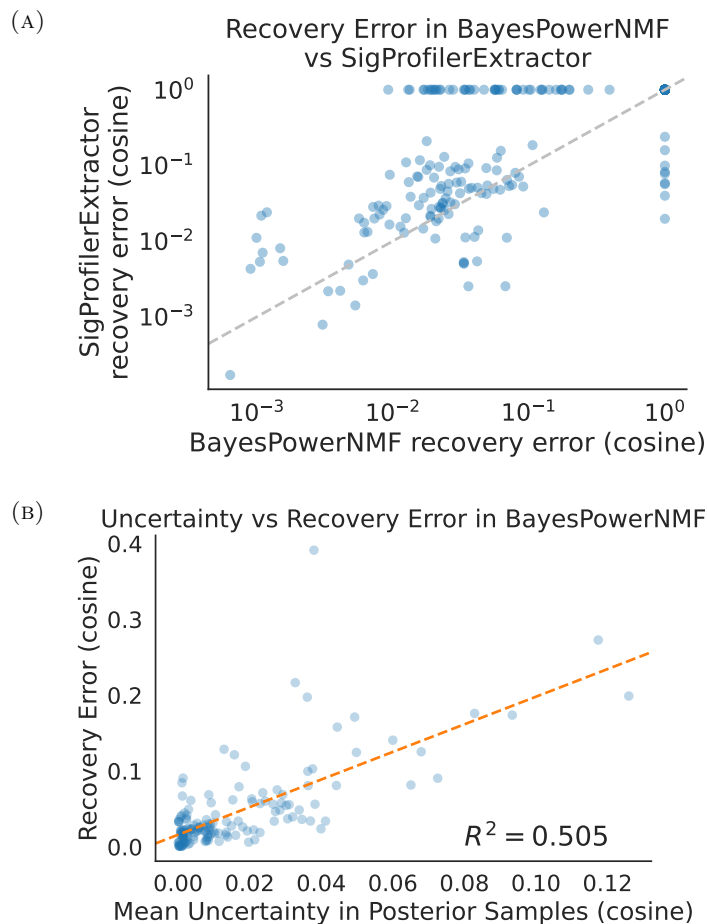
(A)

(B)

FIGURE S3. **(A)** Recovery error for SigProfilerExtractor versus BayesPowerNMF. For each of the 15 complete synthetic data sets (see Section 3), for each ground truth signature, we plot the cosine error between the true signature and the signatures estimated by SigProfilerExtractor (y-axis) and BayesPowerNMF (x-axis); a cosine error of 1 indicates that no matching signature was inferred by that method. SigProfilerExtractor misses many of the signatures recovered by BayesPowerNMF (see dots along the top), and has higher cosine error for most of the signatures recovered by both methods (see dots above the diagonal). **(B)** In simulations with BayesPowerNMF, the posterior uncertainty in each signature is correlated with the cosine error between the estimated signature and the ground truth signature. This indicates that the uncertainty quantification is providing meaningful information about the actual error. Here, uncertainty is defined as the mean cosine error between posterior samples and the posterior mean, for each signature.
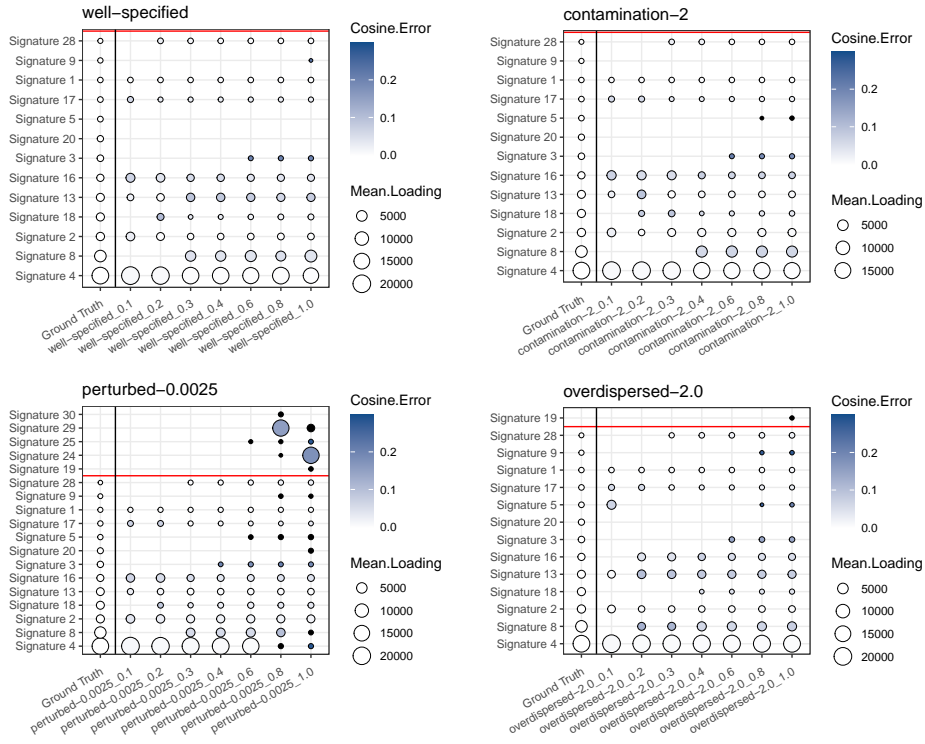
FIGURE S4. Bubble plots used to select the power ξ in stage 4 of the BayesPowerNMF workflow. Each panel corresponds to one simulated data set. See the description of Figure 4 for the interpretion of this type of plot.
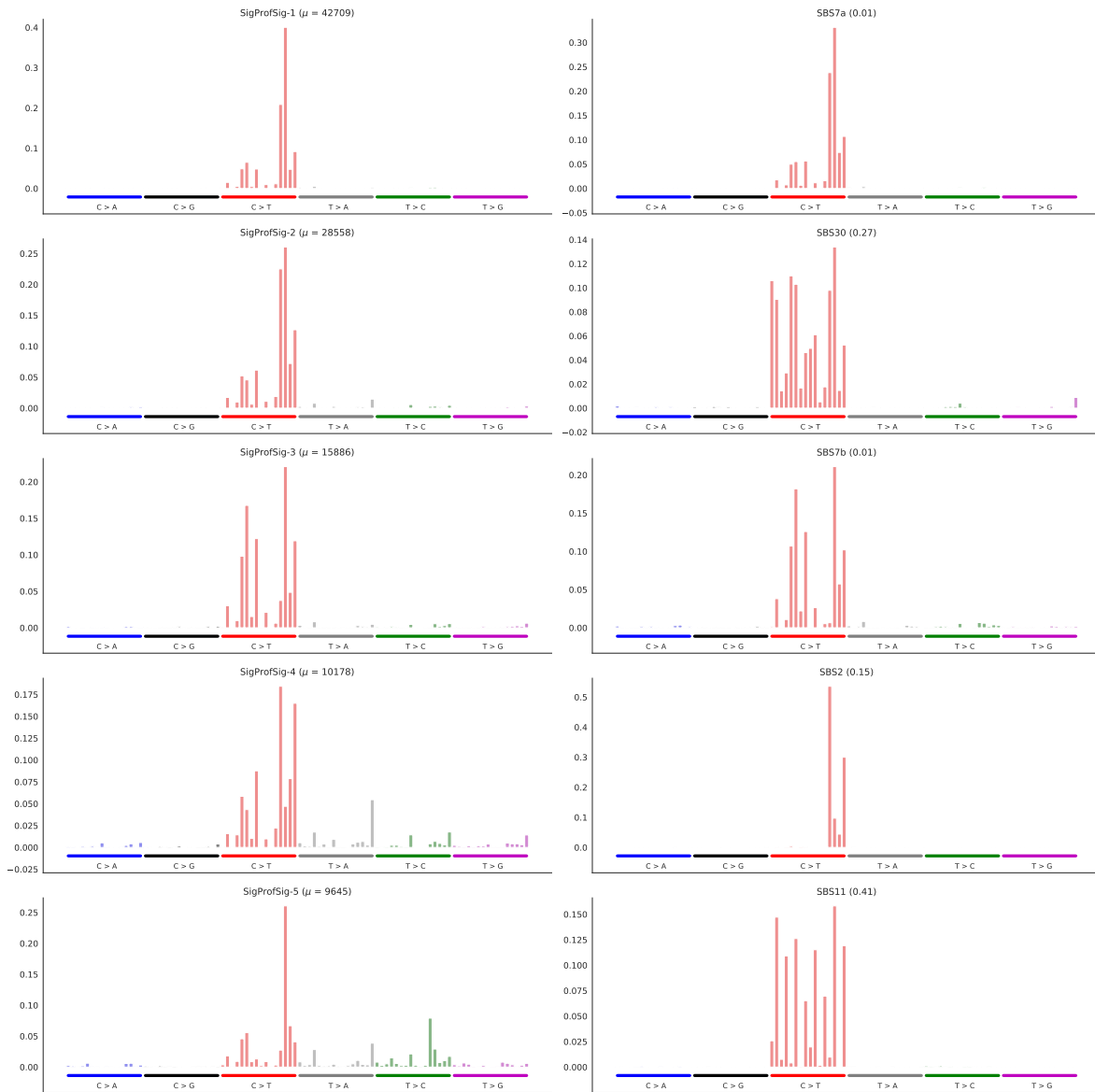
FIGURE S5. The top 5 out of 10 signatures inferred by SigProfilerExtractor from melanoma mutation counts from the PCAWG project (left) and their best match reference signatures from COSMIC v3 (right). There appears to be potential duplication between (a) SigProfSig-1, SigProfSig-2, and SigProfSig-5, and (b) SigProfSig-3 and SigProfSig-4, in the sense that these may be slightly perturbed versions of the same "true" signature.

Appendix B. Details on γ-perturbed data-generation process

Naïvely, one might try to simulate perturbed signatures by sampling from $\text{Dirichlet}(\beta R_k^\star)$ using the same concentration parameter $\beta > 0$ for all pilot signatures $R_k^\star$. However, in terms of cosine error, this leads to very different magnitudes of perturbation depending on the sparsity or flatness of $R_k^\star$; see Figure S6.
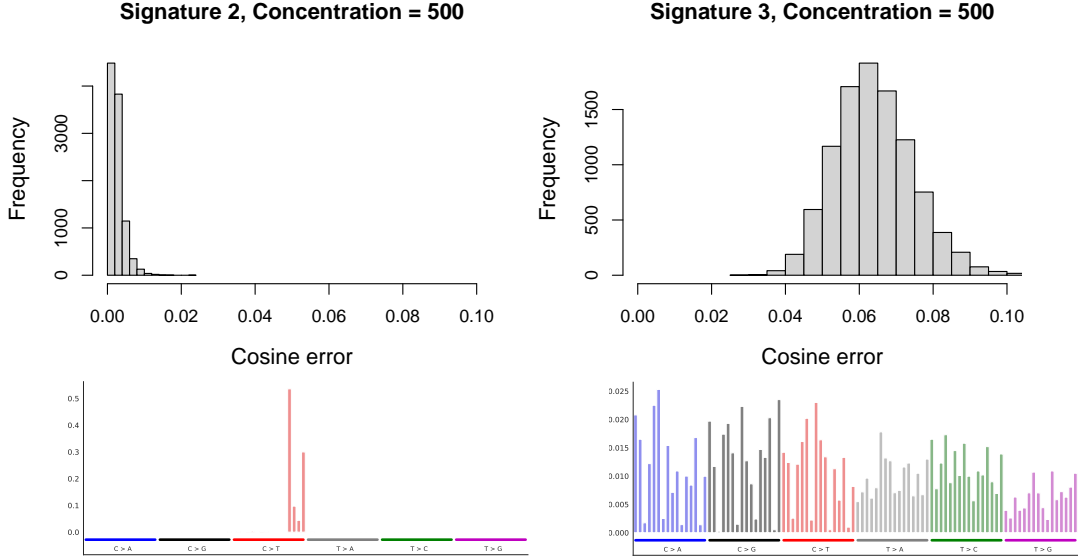

FIGURE S6. (Top) Distribution of cosine errors between $R_k^\star$ and 10,000 samples from $\text{Dirichlet}(\beta R_k^\star)$, where $\beta = 500$ and $R_k^\star$ is Signature 2 (left) or Signature 3 (right) from COSMIC v2. (Bottom) Mutation frequency profile for Signature 2 (left) and Signature 3 (right).

We quantify the "flatness" of a signature $R_k^\star$ via

$$\text{flatness}(R_k^\star) = \frac{1}{96 \left\| R_k^\star \right\|^2}.$$

Larger values indicate a flatter signature and smaller values indicate a spikier signature. This expression is specifically inspired by the form of the cosine error formula, and takes values in $[0, 1]$. Through simulations using each $R_k^\star$ in the set of COSMIC v2 signatures, we observe that for a range of concentration parameters β ranging from 500 to 10,000, the mean cosine error for $\text{Dirichlet}(\beta R_k^\star)$ has a very strong positive linear correlation with $\text{flatness}(R_k^\star)$ (Figure S7A). Further, the slopes of the best-fit line obtained with ordinary least squares (OLS) are strongly linearly correlated with β on a log-log scale (Figure S7B).

Based on these observations, we use OLS to fit the empirical relationship between $\text{flatness}(R_k^\star)$, the desired mean cosine error $\gamma \in (0, 1)$, and the concentration parameter β, specifically,

$$\beta(R_k^\star) = \exp\left(\frac{3.6641 - \log\left(96\gamma \|R_k^\star\|^2\right)}{0.9820} \right).$$
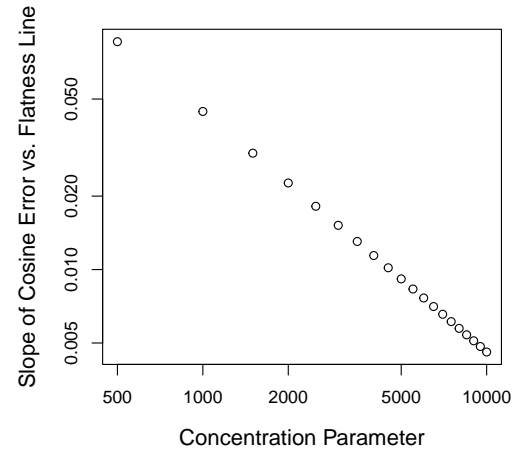
(A)
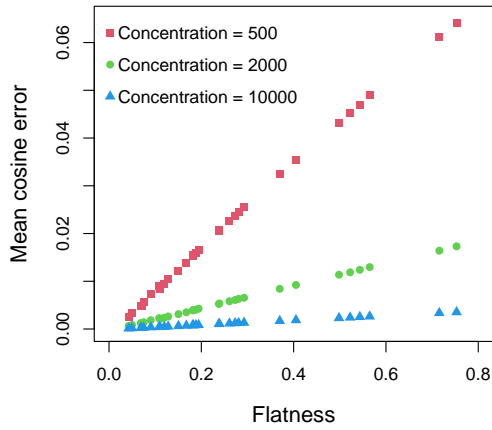
(B)

FIGURE S7. **(A)** Strong linear relationship between flatness($R_k^\star$) and the mean cosine error between a sample from Dirichlet($\beta R_k^\star$) and $R_k^\star$ itself for several different values of β. **(B)** Strong linear relationship between the slope of the lines in Figure S7A and the concentration parameter β on a log scale.

This relationship allows us to choose a signature-specific concentration parameter to simulate new signatures, while controlling the resulting mean cosine errors across all signatures. We use the function $\beta(R_k^\star)$ to set the values of β_k for the γ-perturbed DGP in Section 2.4.