Compressive Bayesian non-negative matrix factorization for mutational signatures analysis

Alessandro Zito[1] and Jeffrey W. Miller[1]

[1]Department of Biostatistics, Harvard University, Boston, MA, 02115, U.S.A.

Abstract

Non-negative matrix factorization (NMF) is widely used in many applications for dimensionality reduction. Inferring an appropriate number of factors for NMF is a challenging problem, and several approaches based on information criteria or sparsity-inducing priors have been proposed. However, inference in these models is often complicated and computationally challenging. In this paper, we introduce a novel methodology for overfitted Bayesian NMF models using "compressive hyperpriors" that force unneeded factors down to negligible values while only imposing mild shrinkage on needed factors. The method is based on using simple semi-conjugate priors to facilitate inference, while setting the strength of the hyperprior in a data-dependent way to achieve this compressive property. We apply our method to mutational signatures analysis in cancer genomics, where we find that it outperforms state-of-the-art alternatives. In particular, we illustrate how our compressive hyperprior enables the use of biologically informed priors on the signatures, yielding significantly improved accuracy. We provide theoretical results characterizing the posterior and its concentration, and we demonstrate the method in simulations and on real data from cancer applications.

1 Introduction

Non-negative matrix factorization (NMF) is a dimensionality reduction technique that decomposes a non-negative matrix into the product of two lower-dimensional non-negative matrices of a desired rank by minimizing a given loss function, such as the squared-error loss or Kullback–Leibler divergence (Lee and Seung, 1999; Gillis, 2020). The literature on Bayesian NMF is a rich one, comprising parametric (Schmidt et al., 2009; Cemgil, 2009; Lu and Ye, 2022; Lu and Chai, 2022; Rahiche and Cheriet, 2022) and non-parametric models (Hoffman et al., 2010; Gopalan et al., 2014; Zhou, 2018, 2015; Ayed and Caron, 2021), multi-study models (Grabski et al., 2023), and spatially dependent structures (Townes and Engelhardt, 2023), to mention a few. Refer to Zhou and Carin (2015) for a summary of Bayesian factorization methods for discrete outcomes.

In cancer genomics, NMF has been successfully used to discover a wide range of mutational signatures corresponding to distinct processes, such as damaged DNA repair mechanisms and environmental mutagens like tobacco smoking or metabolic byproducts (Alexandrov et al., 2013). These signatures, defined as vectors of the frequencies with which different types of point mutations occur, are inferred from mutation counts in whole-genome or whole-exome sequencing using NMF algorithms (Nik-Zainal et al., 2012; Alexandrov et al., 2013, 2020). Identifying these patterns in the DNA of cancer patients is a significant advance toward understanding the etiology of cancer (Koh et al., 2021) and improving the effectiveness of precision therapies (Aguirre et al., 2018; Gulhan et al., 2019).

A number of NMF-based methods have been proposed for mutational signatures analysis; see Islam et al. (2022) and references therein. However, a difficult aspect of NMF is choosing an appropriate number of latent factors, which corresponds to the number of mutational signatures present in the data. Choosing too many factors can lead to the inference of spurious signatures, while choosing too few factors can lead to incorrectly merging distinct signatures. Existing techniques for selecting the number of factors include the Bayesian information criterion (Rosales et al., 2016; Fischer et al., 2013), cross-validation (Lal et al., 2021), or even neural networks (Nebgen et al., 2021; Islam et al., 2022). The disadvantage of such approaches is that they require estimating a separate model for each choice of rank or regularization parameter, making them computationally intensive. Alternatively, sparsity-inducing Bayesian nonparametric models can capture the appropriate latent dimensionality of the data (Bhattacharya and Dunson, 2011; Legramanti et al., 2020; Gopalan et al., 2014) but are highly sensitive to model misspecification. Another popular solution is to use automatic relevance determination (ARD); however, to our knowledge, Bayesian models with ARD that provide a full uncertainty quantification are limited to continuous outcomes (Brouwer et al., 2017), while state-of-the-art algorithms for discrete data (Tan and Févotte, 2013; Kim et al., 2016) are fast but only provide point estimates. Moreover, the theoretical properties of Poisson NMF and ARD are still largely unexplored.

In this article, we introduce a novel Bayesian NMF method that yields accurate and reliable inference in a computationally simple way by using *compressive hyperpriors* to drive the weights of unneeded factors to zero. Specifically, we use a Poisson factorization model with semi-conjugate Dirichlet and gamma priors over the signatures and the loadings, respectively, and we induce sparsity by using a shrinkage hyperprior that strengthens with the amount of data in such a way that (a) unneeded factors are given negligible weight with probability tending to one, while (b) needed factors are given weights that are only mildly shrunk. This compressive property enables the method to select an appropriate number of latent factors in a continuous way, without the need to fit multiple models or discretely jump between models in Markov chain Monte Carlo samplers. Further, posterior inference can be carried out using a simple auxiliary variable Gibbs sampling algorithm as in Dunson and

Herring (2005) and Cemgil (2009), making the method easy to implement. We establish theoretical properties justifying the use of the proposed model, including finite-sample and asymptotic results characterizing the sparsity-inducing effect of the compressive hyperprior.

Additionally, the compressive hyperprior makes it straightforward to take advantage of existing information about the latent factors. This is especially useful for mutational signatures analysis, since the *Catalogue Of Somatic Mutations In Cancer* (COSMIC) database provides a curated set of mutational signatures and their putative etiologies (Alexandrov et al., 2020). By using an informative prior centered at the COSMIC signatures, our model obtains improved sensitivity to detect the presence of signatures and infers unambiguous matches to the original database.

The paper is organized as follows. Section 2 defines the model, introduces our compressive hyperprior, and covers posterior inference. In Section 3, we provide theoretical results characterizing the posterior and its concentration. Section 4 provides background on mutational signatures analysis. Section 5 contains a simulation study comparing to leading methods, and Section 6 presents applications to mutational signatures analysis on breast cancer and pancreatic cancer datasets. We conclude with a brief discussion in Section 7.

2 Methodology

2.1 Poisson non-negative matrix factorization model

We describe the model in the context of mutational signatures, our primary application of interest. Let X_{ij} represent the number of mutations for channel i in sample j, where $i = 1, \ldots, I$ and $j = 1, \ldots, J$, and let $X \in \mathbb{N}^{I\times J}$ denote the matrix with entries X_{ij}, where $\mathbb{N} = \{0, 1, 2, \ldots\}$ is the nonnegative integers. Typically, the channels would consist of the 96 single-base substitution (SBS) types; see Sections 4 and S5 for details. Non-negative matrix factorization (NMF) consists of finding two non-negative matrices $R \in \mathbb{R}_+^{I\times K}$ and $\Theta \in \mathbb{R}_+^{K\times J}$ such that $X \approx R\,\Theta$, with the *rank* K typically chosen so that $K \leq \min\{I, J\}$. The kth column of R, denoted $r_k = (r_{1k}, \ldots, r_{Ik})$, is referred to as the kth *mutational signature*. The kth row of Θ, denoted $\theta_k = (\theta_{k1}, \ldots, \theta_{kJ})$, is the vector of weights representing the *loading* of signature k in each of the J samples, sometimes referred as *activity*.

From a probabilistic perspective, it is natural to model the mutation counts as

$$X_{ij} \sim \text{Poisson}\Bigg(\sum_{k=1}^{K} r_{ik}\theta_{kj}\Bigg) \tag{1}$$

independently for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where $\text{Poisson}(\lambda)$ denotes the Poisson distribution with mean λ. In Section S5, we show that Equation (1) can be derived from first principles by modeling the occurrences of mutations as continuous-time Markov processes across the genome.

In mutational signatures analysis, it is common to impose the constraint that $\sum_{i=1}^{I} r_{ik} = 1$ for all $k = 1, \ldots, K$. This avoids scaling ambiguities in both the signature vectors $r_k = (r_{1k}, \ldots, r_{Ik})$ and their loadings $\theta_k = (\theta_{k1}, \ldots, \theta_{kJ})$. Most methods do not impose such a constraint during inference, opting to enforce it as a post-processing step (Tan and Févotte, 2013; Drummond et al., 2023). However, we find that building $\sum_{i=1}^{I} r_{ik} = 1$ into our model has the additional benefits of simplifying the inference algorithm and enabling direct use of COSMIC signatures for constructing informative priors.

2.2 Prior

For the prior distribution on the signatures r_k and loadings θ_k, we take

$$r_k = (r_{1k}, \ldots, r_{Ik}) \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \tag{2}$$

$$\theta_{k1}, \ldots, \theta_{kJ} \mid \mu_k \;\sim\; \text{Gamma}(a,\, a/\mu_k), \tag{3}$$

independently for $k = 1, \ldots, K$, where $\alpha > 0$, $a > 0$, and $\mu_k > 0$ is given a hyperprior $\mu_k \sim \pi(\mu_k)$. The Dirichlet prior in Equation (2) automatically enforces the constraint that $\sum_{i=1}^{I} r_{ik} = 1$. Here, Gamma$(a, b)$ denotes the gamma distribution with mean a/b and variance a/b^2. Thus, the prior mean of the loadings is $\mathbb{E}(\theta_{kj} \mid \mu_k) = \mu_k$, implying that μ_k controls the overall contribution of signature r_k to the factorization and, in turn, to the total number of mutations generated by process k.

We refer to $\mu_1, \ldots, \mu_K$ as *relevance weights*, following the usage of this type of prior structure in automatic relevance determination (ARD) for shrinking the weights of unneeded factors to near-zero values (Tan and Févotte, 2013). However, unlike Tan and Févotte (2013), we take a fully Bayesian approach, quantifying uncertainty in r_k and θ_k rather than just optimizing them. Also, unlike typical uses of ARD in Bayesian neural networks and Gaussian processes (Neal, 1996; Bishop, 2006), the marginal likelihood in the Poisson NMF model does not have a closed-form expression and thus is not amenable to direct optimization of μ_k. Nonetheless, it turns out that with a certain choice of data-dependent hyperprior on μ_k, we can obtain appealing computational properties similar to ARD, while performing inference with simple Gibbs sampling updates; we discuss this next. See Section S1 for detailed discussion of the differences compared to previous methods.

2.3 Compressive hyperprior

The hyperprior on the relevance weights $\pi(\mu_k)$ plays a crucial role in inferring the number of factors. Several approaches have been developed to provide sparsity in Gaussian factorization models, such as spike-and-slab priors that introduce exact zeros in the loadings (Carvalho et al., 2008; Ročková and George, 2016) or induce cumulative shrinkage to near-zero values for redundant factors (Legramanti et al., 2020; Frühwirth-Schnatter, 2023); also see Liu et al. (2019) for an extension to NMF settings.

However, in our model, we found that spike-and-slab priors over μ_k tend to make posterior inference difficult, likely due to the strong multimodal nature of the resulting posterior.

Instead, we propose a simpler alternative inspired by ARD methods, letting

$$\mu_k \sim \mathrm{InvGamma}(aJ + 1,\, \varepsilon a J) \tag{4}$$

independently for $k = 1, \ldots, K$, where $\mathrm{InvGamma}(a_0, b_0)$ denotes the inverse-gamma distribution with mean $b_0/(a_0 - 1)$ when $a_0 > 1$. Here, a is the shape parameter from the prior on θ_{kj} in Equation (3), and we set $\varepsilon > 0$ to be a small constant, such as $\varepsilon = 0.001$. Note that this makes μ_k small *a priori*, since $\mathbb{E}(\mu_k) = \varepsilon$. Further, the full conditional mean is $\mathbb{E}(\mu_k \mid \Theta) = \varepsilon/2 + \bar{\theta}_k/2$, where $\bar{\theta}_k = \frac{1}{J}\sum_{j=1}^{J} \theta_{kj}$, implying that the prior mean ε and the average loading $\bar{\theta}_k$ for signature k have equal influence on the posterior for μ_k. This *strength-matching* property of the hyperprior remains stable as the sample size J increases. We refer to Equation (4) as a *compressive hyperprior*. We also note that this should not be interpreted as a meaningful representation of prior uncertainty regarding μ_k. Instead, it is designed to yield a posterior with good properties in terms of computation and accuracy.

This deceptively simple choice has some key features. First, it favors sparse solutions, since $\mathbb{E}(\mu_k) = \varepsilon$. This makes it so that for any extra unneeded signatures, Equation (4) encourages the corresponding relevance weights μ_k to be small, on the order of ε. Second, despite its growing strength with J, this hyperprior does not overly shrink the loadings θ_{kj} for factors that are needed to fit the data. We establish these properties rigorously in Section 3, particularly Theorems 3 and 4. Another important feature of the strength-matching hyperprior in Equation (4) is that small departures from the assumed Poisson NMF model do not strongly affect the number of factors used by the model. As J grows, a fixed-strength hyperprior on μ_k would be overwhelmed by the likelihood since the number of parameters θ_{kj} grows with J, leading to the inclusion of spurious extra signatures when the model is slightly misspecified. Meanwhile, the strength-matching drives out spurious extra signatures by maintaining a balance between the contribution from the loadings and from the hyperprior; see Section S7.1. Finally, the resulting posterior density is highly tractable due to the use of semi-conjugate distributions. This improves the performance of the sampling algorithm that we employ for inference, providing expeditious convergence to NMF solutions along with Bayesian uncertainty quantification.

2.4 Choice of model settings

We now discuss the role of the values of K, ε, a, and α in our proposed model. It turns out that, while the posterior distribution is significantly affected by a, the precise values of K, ε, and α are not important; see the sensitivity analyses in Sections S7.2 and S8. We find the following choices to work well as defaults: $K = 20$, $\varepsilon = 0.001$, $a = 1$, $\alpha = 0.5$.

The hyperprior mean, ε, is the value that the relevance weights μ_k and loadings θ_{kj} for unused signatures will be driven down to. Thus, it is not critical to choose a particular value of ε, as long as it is considerably less than the smallest true nonzero loadings.

In our compressive model, K represents the maximum number of signatures that might be encountered in the data. When the true number of signatures K^0 is less than K and ε is small, the compressive hyperprior drives the loadings θ_{kj} for the $K - K^0$ redundant factors down to ε. In Section S7.2, we observe that any value of $K \geq K^0$ works equally well. Hence, this shrinkage mechanism mirrors the automatic selection of the number of active components in overfitted mixture models (Rousseau and Mengersen, 2011). We find that $K = 20$ represents a good compromise between the average number of signatures found across cancer types (Alexandrov et al., 2020) and the additional computational cost of using a larger K; see Section 2.5. If all signatures are included by the model, we suggest raising K to a larger value and re-running the model.

The gamma shape parameter a has an important and non-obvious role, which is that it serves as a threshold for inclusion of signatures in the model. More precisely, we show that a signature will be included only if the average number of mutations due to that signature is above a; see Figure 1 and Section 3.3. We find that $a = 1$ is generally reliable.

The Dirichlet concentration α in Equation (2) controls the prior entropy of the signatures. Small values of α lead to low-entropy signatures in which only a few channels have nonnegligible probability, whereas larger values of α lead to flatter, more uniform signatures. We find $\alpha = 0.5$ to adapt well to both low- and high-entropy signatures without being excessively informative, but similar results were obtained with $\alpha = 0.25$ and $\alpha = 1$ in our simulations. See Section 3.3 for further discussion in connection with the theory.

2.5 Posterior inference

Posterior inference for the hierarchical model defined by Equations (1) to (4) can be efficiently performed via Gibbs sampling. Since the sum of independent Poisson random variables is Poisson, we can equivalently write the hierarchical model as

$$\begin{aligned} X_{ij} &= \textstyle\sum_{k=1}^{K} Y_{ijk}, \\ Y_{ijk} \mid \mu_k, r_k, \theta_k &\sim \text{Poisson}(r_{ik}\theta_{kj}), \\ (r_{1k}, \ldots, r_{Ik}) &\sim \text{Dirichlet}(\alpha, \ldots, \alpha), \\ \theta_{k1}, \ldots, \theta_{kJ} \mid \mu_k &\sim \text{Gamma}(a,\, a/\mu_k), \\ \mu_k &\sim \text{InvGamma}(aJ + 1,\, \varepsilon a J). \end{aligned} \tag{5}$$

Each auxiliary variable $Y_{ijk} \in \mathbb{N}$ can be interpreted as the number of mutations due to signature k in channel i for sample j. Defining the vector $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijK})$, it turns out that $Y_{ij} \mid X_{ij}, R, \Theta$ follows a Multinomial$\big(X_{ij}, (q_{ij1}, \ldots, q_{ijK})\big)$ distribution, where $q_{ijk} = r_{ik}\theta_{kj}/Q_{ij}$ and $Q_{ij} = \sum_{k=1}^{K} r_{ik}\theta_{kj}$. This auxiliary variable decomposition has been used in several previous methods (Dunson and Herring, 2005; Cemgil, 2009; Rosales et al., 2016; Zhou and Carin, 2015). The rest of the sampler relies on standard semi-conjugate updates, which are straightforward to derive. Specifically, we iterate the following steps.

1. For each i and j, update $(Y_{ij} \mid X, R, \Theta) \sim \text{Multinomial}\big(X_{ij}, (q_{ij1}, \ldots, q_{ijK})\big)$ where $q_{ijk} = r_{ik}\theta_{kj}/Q_{ij}$ and $Q_{ij} = \sum_{k=1}^{K} r_{ik}\theta_{kj}$.
2. For each k, update $(r_k \mid Y) \sim \text{Dirichlet}\big(\alpha + \sum_{j=1}^{J} Y_{1jk},\ \ldots,\ \alpha + \sum_{j=1}^{J} Y_{Ijk}\big)$.
3. For each j and k, update $(\theta_{kj} \mid Y, \mu) \sim \text{Gamma}\big(a + \sum_{i=1}^{I} Y_{ijk},\ a/\mu_k + 1\big)$.
4. For each k, update $(\mu_k \mid \Theta) \sim \text{InvGamma}\big(2aJ + 1,\ \varepsilon aJ + a\sum_{j=1}^{J} \theta_{kj}\big)$.

The model is symmetric with respect to the order of the factors, in the sense that the priors and likelihood are invariant to permutations of $k = 1, \ldots, K$. While attractive from a modeling standpoint, this symmetry could potentially lead to label switching when running the Gibbs sampler, complicating the calculation of posterior expectations. However, we have not encountered any label switching, so this has not been an issue in practice.

2.6 Informative priors based on known signatures

A favorable aspect of mutational signatures analysis is the abundance of historical data on signatures across many cancer types. The COSMIC database contains a curated collection of signatures, annotated with associated cancer types and inferred etiologies (Alexandrov et al., 2020). It is natural to leverage such prior information as follows. Suppose $s_k = (s_{1k}, \ldots, s_{Ik})$, for $k = 1, \ldots, K_{\text{pre}}$, are pre-defined mutational signatures known to occur in cancer. To allow for variation in signatures across studies, we let $\rho_k = (\rho_{1k}, \ldots, \rho_{Ik})$ denote a study-specific version of s_k. We then generalize Equation (1) by independently modeling

$$X_{ij} \sim \text{Poisson}\Bigg(\sum_{k=1}^{K_{\text{pre}}} \rho_{ik}\omega_{kj} + \sum_{k=1}^{K_{\text{new}}} r_{ik}\theta_{kj}\Bigg), \tag{6}$$

where r_{ik} and θ_{ik} are given the priors in Equations (2) and (3), respectively, and

$$\begin{aligned} \rho_k &\sim \text{Dirichlet}(\beta_k s_{1k}, \ldots, \beta_k s_{Ik}), \\ \omega_{kj} \mid \tau_k &\sim \text{Gamma}(b,\ b/\tau_k), \\ \tau_k &\sim \text{InvGamma}(bJ + 1,\ \varepsilon bJ). \end{aligned} \tag{7}$$

Thus, the prior on ρ_k is centered at s_k, with concentration parameter β_k. The loadings ω_{kj} and corresponding relevance weights τ_k are given the same prior and compressive hyperprior as θ_{kj} and μ_k, respectively, but with b in place of a.

The model in Equation (6) is reminiscent of the recovery-discovery model discussed by Grabski et al. (2023), when only a single study is taken into consideration. In such a framework, the prior rank for the lower-dimensional matrices, $K_{\text{pre}} + K_{\text{new}}$, is often greater than J. This is at odds with the classic approach to NMF, where the factorization rank is typically smaller than the rank of X. However, the compressive mechanism behind our priors in Equation (7) still ensures a parsimonious representation in the posterior, such that only the active signatures have a nonnegligible relevance weight τ_k. Posterior inference can be performed using the same steps as in Section 2.5, with minor adjustments for handling ρ_k, ω_{kj}, and τ_k; see Section S4 for details.

3 Theory

We establish closed-form expressions for the posteriors of $\mu_k \mid Y$ (Theorem 1) and $\theta_{kj} \mid Y$ (Theorem 2) and we characterize the asymptotic distributions of $\mu_k \mid Y$ (Theorem 3) and $\theta_{kj} \mid Y$ (Theorem 4). These results give insight into the behavior of the model and the choice of model settings; see Section 3.3 for a discussion of the interpretation of the theory.

3.1 Distributional properties of the posterior

We show that the distribution of $\mu_k \mid Y$ under the model in Equation (5), where Y is the tensor $(Y_{ijk}) \in \mathbb{R}^{I\times J\times K}$, has a closed-form expression in terms of confluent hypergeometric functions. We refer to the resulting novel family of distributions as *inverse Kummer*.

Definition 1. The *inverse Kummer* distribution with parameters $\lambda > 0$, $\beta > 0$, $\delta > 0$, and $\gamma \in \mathbb{R}$ is a continuous distribution on $(0, \infty)$ with probability density function

$$\pi(\mu) = \frac{\mu^{-(\lambda-\gamma)-1}(1+\mu/\delta)^{-\gamma}e^{-\beta/\mu}}{\delta^{\gamma-\lambda}\,\Gamma(\lambda)\,U(\lambda,\lambda+1-\gamma,\beta/\delta)}. \tag{8}$$

We write $\mu \sim \text{InvKummer}(\lambda, \beta, \gamma, \delta)$ to denote that μ has the density in Equation (8).

Here, $U(a, b, z)$ denotes the confluent hypergeometric function of the second kind,

$$U(a,b,z) = \frac{1}{\Gamma(a)}\int_0^\infty t^{a-1}(1+t)^{b-a-1}e^{-zt}\mathrm{d}t,$$

with $z > 0$ (Abramowitz and Stegun, 1972). We call this an inverse Kummer distribution since if $\mu \sim \text{InvKummer}(\lambda, \beta, \gamma, \delta)$ then $1/\mu$ follows a Kummer distribution, which was introduced by Armero and Bayarri (1997) when studying a M/M/∞ queuing problem. The moments have closed-form expressions in terms of the hypergeometric function, following Equation (6.7) in Armero and Bayarri (1997): for $m < \lambda$, we write

$$\mathbb{E}(\mu^m) = \delta^m \frac{\Gamma(\lambda - m)}{\Gamma(\lambda)} \frac{U(\lambda - m,\, \lambda - m + 1 - \gamma,\, \beta/\delta)}{U(\lambda,\, \lambda + 1 - \gamma,\, \beta/\delta)}. \tag{9}$$

The inverse Kummer is a generalization of the inverse gamma, since $\mathrm{InvKummer}(\lambda, \beta, 0, \delta) = \mathrm{InvGamma}(\lambda, \beta)$. Moreover, in Section S3.1 we show that when $\lambda > 2$, the mean $\mathbb{E}(\mu)$ is monotonically increasing as a function of γ. In the compressive NMF model in Equation (5), the inverse Kummer arises as the posterior distribution of the relevance weights given the latent counts, integrating out the signatures and loadings, as shown below.

Theorem 1. *Let $Y = (Y_{ijk}) \in \mathbb{N}^{I\times J\times K}$ denote the tensor of latent counts. Under the hierarchical model in Equation (5), we have*

$$(\mu_k \mid Y) \sim \mathrm{InvKummer}\big(2aJ + 1,\, \varepsilon a J,\, J\bar{Y}_k + aJ,\, a\big),$$

where $\bar{Y}_k = \frac{1}{J}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ijk}$ is the average number of mutations assigned to signature k.

See Section S2 for the proof. Figure 1(A) shows the density of $\mu_k \mid Y$ for various $\bar{Y}_k$ and J values. Similarly, we show that $\theta_{kj} \mid Y$ has a closed-form density in Theorem S3.1, and we provide its expectation in the following result.

Theorem 2. *In the same setting as Theorem 1, let $Y_{jk} = \sum_{i=1}^{I} Y_{ijk}$. Then*

$$\mathbb{E}(\theta_{kj} \mid Y) = (a + Y_{jk}) \frac{U\big(2aJ + 1, J(a - \bar{Y}_k) + 1, \varepsilon J\big)}{U\big(2aJ + 1, J(a - \bar{Y}_k) + 2, \varepsilon J\big)}.$$

Despite its complicated-looking form, $\mathbb{E}(\theta_{kj} \mid Y)$ has a simple approximation as $J \to \infty$. We show this in Equation (11) and use it to interpret the posterior behavior of the loadings.

3.2 Asymptotic results and compressive property

The essence of the compressive hyperprior is that for unneeded factors, the relevance weights are shrunk to $\approx \varepsilon$, while for needed factors they are only partly shrunk towards ε. We characterize this behavior by studying the concentration of $\mu_k \mid Y$ as $J \to \infty$.

Theorem 3. *Consider the hierarchical model in Equation (5). Suppose $\bar{Y}_k \to y$ as $J \to \infty$ for some $y \geq 0$. For any $c_1, c_2, \ldots \in [0, \infty)$ such that $c_J \to \infty$, $c_J/\sqrt{J} \to 0$, and $|\bar{Y}_k - y| = o(c_J/\sqrt{J})$, we have*

$$\mathbb{P}\big(|\mu_k - \mu_*| \leq c_J/\sqrt{J} \mid Y\big) \xrightarrow[J\to\infty]{} 1,$$

where $\mu_* = 2a\varepsilon / \big(\sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon} - (y-a+\varepsilon)\big)$. *Moreover, there exist constants* $D_1, D_2, \ldots \in \mathbb{R}$ *and* $v_1, v_2, \ldots \in \mathbb{R}$ *such that*

$$\big(\sqrt{J}(\mu_k - \mu_*) - \Delta_J\big) \mid Y \xrightarrow[J\to\infty]{\mathrm{d}} \mathcal{N}\bigg(0,\ \frac{\mu_*^3(\mu_* + a)}{2a\mu_*^2 + a^2\varepsilon}\bigg),$$

where $\Delta_J = D_J \sqrt{J}(\bar{Y}_k - y) + v_J$, $\lim_{J\to\infty} D_J = \mu_*^2 / \left(2a\varepsilon + \mu_*(y-a+\varepsilon)\right)$, *and* $v_J \to 0$.

Here, $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution, and $\mathcal{N}(m, s^2)$ is the normal distribution with mean m and variance s^2. This result shows that $\mu_k \mid Y$ concentrates at μ_* at a $1/\sqrt{J}$ rate whenever $\bar{Y}_k$ converges to y at a $1/\sqrt{J}$ rate. Moreover, $\sqrt{J}(\mu_k - \mu_*)$ is asymptotically normal with mean Δ_J; see Theorem S2.2 for the analytic form of D_J and v_J. As a function of y, the point μ_* at which concentration occurs follows an elbow-shaped curve (Figure 1). This illuminating fact helps understand the sparsity-inducing effect; see Section 3.3. The following corollary provides a criterion for thresholding the relevance weights.

Corollary 1. *Consider the hierarchical model in Equation (5). If* $\bar{Y}_k = o(c_J/\sqrt{J})$ *for some* $c_J \geq 0$ *such that* $c_J \to \infty$ *and* $c_J/\sqrt{J} \to 0$, *then for all* $C > 1$,

$$\mathbb{P}(\mu_k > C\varepsilon \mid Y) \xrightarrow[J\to\infty]{} 0.$$

Hence, when signature k is not being used by the model, Corollary 1 shows that the posterior for μ_k concentrates on the interval $(0, C\varepsilon)$, for any $C > 1$. We refer to this as the *compressive property* of the model. This provides a natural criterion for selecting signatures for inclusion in the model, by using a threshold of $\mu_k > C\varepsilon$ to decide which signatures to keep and which to discard. Inspection of the proofs of Theorems 1 and 3 shows that they hold for any prior on signatures r_k such that $\sum_i r_{ik} = 1$. Consequently, the concentration results in Theorem 3 and Corollary 1 also hold for the relevance weights τ_k of the augmented model in Equation (7), which employs informative priors on r_k.

Next, using Theorem 3, we establish the asymptotic distribution of $\theta_{kj} \mid Y$, both in the case of our compressive hyperprior and a fixed-strength hyperprior, for comparison.

Theorem 4. *Under the assumptions of Theorem 3, it holds that*

$$(\theta_{kj} \mid Y) \xrightarrow[J\to\infty]{\mathrm{d}} \mathrm{Gamma}\big(a + Y_{jk},\ a/\mu_* + 1\big),$$

where μ_* *is defined as in Theorem 3. In the fixed-strength case where* $\mu_k \sim \mathrm{InvGamma}(a_0, b_0)$ *for fixed* $a_0, b_0 > 0$, *we have*

$$(\theta_{kj} \mid Y) \xrightarrow[J\to\infty]{\mathrm{d}} \mathrm{Gamma}\big(a + Y_{jk},\ a/y + 1\big).$$

For the fixed-strength hyperprior, this result naturally arises from the fact that $\mu_k \mid Y$ concentrates at y (see Theorem S3.2). Note that θ_{kj} and Y_{jk} do not depend on J, since they are specific to patient j and signature k.

3.3 Interpretation of the theoretical results

The goal of our theoretical analysis is to show that the model correctly infers which signatures to include or exclude, without introducing strong biases. The results also elucidate the role of the model settings (K, ε, a, and α) in Equation (5) in relation to these effects.

Inclusion or exclusion of signature k is represented through the relevance weights μ_k and loadings θ_{kj}. Thus, our first results focus on the posterior distributions of $\mu_k \mid Y$ and $\theta_{kj} \mid Y$, showing they have closed-form expressions which facilitate further analysis (Theorems 1 and S3.1). Furthermore, Theorem 2 establishes an analytic form for $\mathbb{E}(\theta_{kj} \mid Y)$. Recall that $Y = (Y_{ijk})$ where Y_{ijk} is the number of mutations in channel i due to signature k for sample j. Letting $Y_{jk} = \sum_{i=1}^{I} Y_{ijk}$, we write $\bar{Y}_k = \frac{1}{J}\sum_{j=1}^{J} Y_{jk}$ to denote the average number of mutations due to process k across samples.

Our next result, Theorem 3, considers the asymptotics of $\mu_k \mid Y$ using Theorem 1. As $J \to \infty$ and $\bar{Y}_k \to y$, this result (i) establishes that $\mu_k \mid Y$ concentrates at a $1/\sqrt{J}$ rate whenever $\bar{Y}_k \to y$ at a $1/\sqrt{J}$ rate, (ii) provides a simple closed-form expression for the point μ^* at which $\mu_k \mid Y$ concentrates, and (iii) shows that $\mu_k \mid Y$ is asymptotically normal.

While convoluted at first sight, the relationship between μ_* and y has a simple interpretation: by a first-order Taylor approximation to the denominator, when $\varepsilon \ll |y - a|$,

$$\mu_* \approx \begin{cases} \dfrac{y-a}{2} & \text{if } y > a \\ \dfrac{\varepsilon a(a-y)}{(a-y)^2 + (a+y)\varepsilon} & \text{if } 0 \leq y < a. \end{cases} \tag{10}$$

See Section S3.2 for the derivation. Equation (10) shows that μ_* grows linearly with y when $y > a$, and μ_* is close to ε when $0 < y < a$; in particular, $\mu_* \approx \varepsilon a/(a + \varepsilon)$ if $y \approx 0$. This can be seen in the elbow-shaped curves seen in Figure 1(B), which displays the posterior relationship between μ_k and $\bar{Y}_k$ for a range of J and a values. We see that when $\bar{Y}_k$ is less than a, the inferred μ_k is negligible due to the compressive hyperprior. Since μ_k jointly controls all loadings θ_{kj}, this effectively excludes signature k from the decomposition. Corollary 1 suggests a simple thresholding rule to determine which signatures should be excluded. Note that in Figure 1(B), the distribution of $\mu_k \mid Y$ contracts as J grows, but the mean $\mathbb{E}(\mu_k \mid Y)$ is unaffected by J; see also Figure S7.1. Hence, a represents a cutoff below which signatures are shrunk to ε, regardless of J.

Since the compressive hyperprior induces significant shrinkage on μ_k, we next analyze whether this induces an excessive bias in the loadings θ_{kj}. To this end, Theorem 4 uses Theorem S3.1 to show that

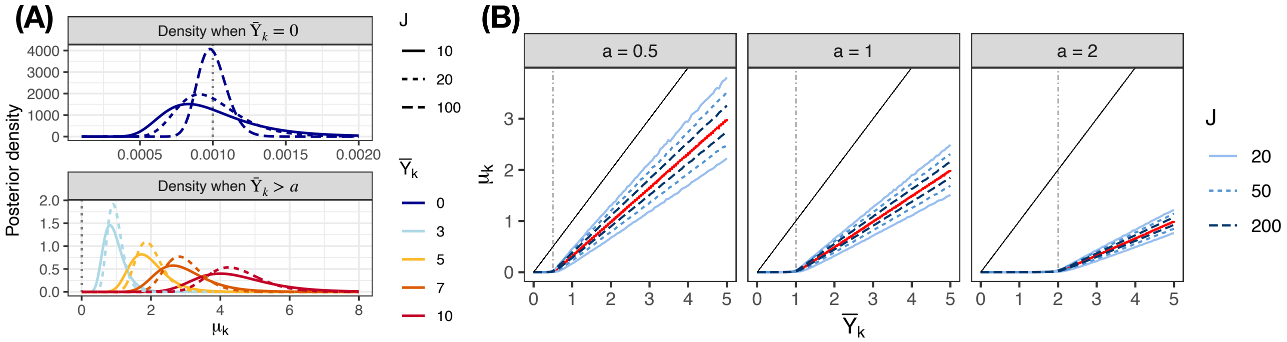


Figure 1: (A) Density of $\mu_k \mid Y$ for a range of values of $\bar{Y}_k$ and sample size J. Here, $a = 1$ and $\varepsilon = 0.001$. (B) Posterior of $\mu_k \mid Y$ as a function of $\bar{Y}_k$, for varying J and a (vertical line). Blue lines are 10th and 90th percentiles. The solid black line is $\mu_k = \bar{Y}_k$ and the red line is $\mathbb{E}(\mu_k \mid \bar{Y}_k)$.

the asymptotic distribution of $\theta_{kj} \mid Y$ is a gamma distribution under both the compressive hyperprior and a fixed-strength hyperprior. Let θ_{kj}^{C} and $\theta_{kj}^{\mathrm{FS}}$ denote gamma random variables with these limiting distributions, where the superscripts are short for *compressive* and *fixed-strength*, respectively. By Equation (10),

$$\mathbb{E}(\theta_{kj}^{\mathrm{C}}) \approx \frac{y-a}{y+a}(a + Y_{jk}), \qquad \mathbb{E}(\theta_{kj}^{\mathrm{FS}}) = \frac{y}{y+a}(a + Y_{jk}), \tag{11}$$

when $y > a$ and $\varepsilon \ll |y - a|$. Equation (11) justifies the claim that, for signatures having a significant contribution in terms of mutation count (in the sense that $y = \lim_{J\to\infty} \bar{Y}_k$ is large relative to a), the loadings are not strongly biased due to using the strength-matching compressive hyperprior rather than a fixed-strength hyperprior. In the fixed-strength case, the sparsity-inducing elbow shape of the curves goes away as J grows since $\mu_k \mid Y$ concentrates at y; see Theorem S3.2 and Figure S7.1. This aggravates overestimation of the rank when the model is misspecified in the fixed-strength case; see Section S7.1.

4 Background on mutational signatures analysis

Cancer development in humans is connected to the accumulation of mutations in the DNA of somatic cells. When considering single-base substitutions (SBS), mutations are classified according to which of the four nucleotide bases was present before and after the mutation, on the strand containing the pyrimidine before the mutation occurred. Recalling that adenine (A) and guanine (G) are *purines* while cytosine (C) and thymine (T) are *pyrimidines* and that C always binds with G and T with A, there are six possible types of substitutions, namely, C>T, C>G, C>A, T>A, T>C and T>G. To account for context-specific variability due to adjacent bases, mutations are further classified according to which bases (A, G, C, or T) occur on the 5' and the 3' sides on the strand containing the pre-substitution pyrimidine. This makes for a total of 6 × 4 × 4 = 96 types of single-base substitutions, referred to as *mutational channels* (Alexandrov et al., 2013); see Section S5 for details. Less frequent than SBS mutations are *in*sertions and *del*etions (indels), which consist of the removal or addition of one or more nucleotides at a given position. Indels are commonly categorized into $I = 83$ types, described in Alexandrov et al. (2020).

It has been observed that many mutation-causing processes consistently produce each type of mutation at a particular rate: for instance, ultraviolet radiation produces a large number of C>T substitutions in melanoma and glioma (Greenman et al., 2007). Due to this, the mutational processes acting on somatic cells can be characterized using *mutational signatures* (Nik-Zainal et al., 2012; Alexandrov et al., 2013), which consist of vectors containing the probability of each type of mutation under consideration.

A curated set of known signatures is maintained in the COSMIC database[1] (Alexandrov et al., 2020), which currently lists 86 SBS and 23 indel signatures. Many signatures can be attributed to a specific etiology that has been experimentally validated. For example, signatures SBS7a, b, c, and d are all linked to ultraviolet light exposure, while ID6 is associated with homologous recombination deficiencies. Other signatures, such as SBS60, require further investigation to understand whether they arise from true biological processes or are due to technical artifacts. See Koh et al. (2021) for and extensive overview and further details.

The Poisson NMF model commonly used in mutational signatures analysis can be derived from first principles by modeling the occurrence of nucleotide substitutions at each base in the genome as a continuous-time Markov process. Aggregating the resulting substitution counts across the entire genome, and modeling each mutational process as independent, we show in Section S5 that the counts are well approximated by the Poisson NMF model in Equation (1). This provides a compelling biological justification for the model.

5 Simulations

In this section, we conduct a simulation study to evaluate the performance of our compressive Poisson NMF method in terms of (a) detecting the true number of signatures that are active in the data and (b) accurately recovering the true signatures and their associated loadings. We also conduct a sensitivity analysis, presented in Section S7.

5.1 Setup of the simulations

We simulate data and true parameters as follows. The mutation counts are generated as $X_{ij} \sim \mathrm{NegBin}\big(1/\tau,\, 1/(1+\tau\lambda_{ij}^0)\big)$, where $\lambda_{ij}^0 = \sum_{k=1}^{K_{\mathrm{pre}}^0} \rho_{ik}^0 \omega_{kj}^0 + \sum_{k=1}^{K_{\mathrm{new}}^0} r_{ik}^0 \theta_{kj}^0$. We parametrize the negative binomial such that the mean and variance of X_{ij} are λ_{ij}^0 and $\lambda_{ij}^0(1+\tau\lambda_{ij}^0)$, respectively, where $\tau > 0$

[1] https://cancer.sanger.ac.uk/signatures/

is a parameter controlling overdispersion. We set $K^0_{\text{pre}} = 4$, and for $k = 1, \dots, 4$, we define $\rho^0_k = (\rho^0_{1k}, \dots, \rho^0_{Ik})$ to be COSMIC signatures SBS1, SBS2, SBS5, and SBS13, respectively. SBS1 is a sparse signature present in every cancer type, arising from the spontaneous deamination of 5-methylcytosine. SBS5 is a rather flat signature that has been shown to appear in every cancer type. SBS2 and SBS13 are commonly occurring signatures associated with APOBEC activity. Meanwhile, we randomly generate $r^0_k = (r^0_{1k}, \dots, r^0_{Ik})$ as $r^0_k \sim \text{Dirichlet}(0.25, \dots, 0.25)$, independently for $k = 1, \dots, K^0_{\text{new}}$. We generate loadings by setting $\omega^0_{kj} = w_k \xi_{kj}$ where $w_k \sim \text{Gamma}(100, 1)$ and $\xi_{kj} \sim \text{Gamma}(0.5, 0.5)$ independently, and θ^0_{kj} in the same way as ω^0_{kj}.

We consider two overdispersion settings: $\tau = 0$, in which case the negative binomial reduces to a Poisson (so the Poisson NMF model is correct), and $\tau = 0.15$, resulting in mild misspecification. For the number of non-COSMIC signatures K^0_{new}, we consider $K^0_{\text{new}} = 2$ and $K^0_{\text{new}} = 6$, so that the total number $K^0 = K^0_{\text{pre}} + K^0_{\text{new}}$ is either 6 or 10. We consider a range of sample sizes $J \in \{50, 100, 200\}$. For each combination of τ, J, and K^0, we generate 20 replicate sets of parameters and data matrices. On each simulated data matrix, we run:

(i) CompNMF: our compressive NMF model in Equation (5) with $K = 20$ and $\varepsilon = 0.001$,

(ii) CompNMF+cosmic: our enhanced model in Equation (6) with $K_{\text{new}} = 15$ *de novo* signatures and the $K_{\text{pre}} = 67$ COSMIC v3.4 signatures that are not regarded as "possible sequencing artifacts", and $\varepsilon = 0.001$,

(iii) PoissonCUSP: the CUSP infinite spike-and-slab model of Legramanti et al. (2020) using their Algorithm 2 to adaptively tune K,

(iv) signeR: the SIGNER model (Rosales et al., 2016; Drummond et al., 2023) with default parameters (`estimate_hyper = FALSE`) and with K ranging from 2 to 20,

(v) SignatureAnalyzer: as implemented in the `sig_auto_extract` function of the `sigminer` package (Wang et al., 2020), with selection method set to `L1KL` and $K = 20$,

(vi) SigProfiler: SIGPROFILEREXTRACTOR v1.1.23 (Islam et al., 2022) with 5 replicates for each $K \in \{2, \dots, 20\}$, using the `sigprofiler_extract` wrapper in `sigminer`.

(vii) BayesNMF: the Bayesian NMF model based on Gaussian likelihood and exponential ARD priors from Brouwer et al. (2017), with $K = 20$ and default settings.

We provide a description of each of these previous methods in Section S6.1. For methods (i), (ii), and (iii), we set $a = 1$ and $\alpha = 0.5$, and run the sampler for 5000 iterations, discarding the first 4000 as burn-in. In method (iii), adaptation of K was started after 500 iterations. In method (ii), we set the parameter β_k such that under the Dirichlet prior in Equation (7), the median cosine similarity between the prior mean and a sample from the prior is approximately 0.975. This makes β_k depend on the sparsity of the signature s_k. For example, we set $\beta_k = 17.29$ for the sparse signature SBS2, while $\beta_k = 1337.26$ for the rather flat SBS3. None of the previous methods allow one to simultaneously use informative COSMIC priors and vague priors for *de novo* analysis. Some could potentially be adapted to Equation (6), but others cannot – for instance, the CUSP prior would result in an unnatural asymmetry across the known signatures.

5.2 Simulation results

Figure 2 shows the main results of the simulation study. Figure 2(A) reports the estimated number of signatures for each method, for each combination of τ, J, and K^0; the boxplots summarize the distribution of estimated values across the 20 replicates. The estimated number of signatures K^* is defined as follows for each method: For (i) and (ii), K^* is the number of signatures for which the posterior mean of μ_k is greater than $5\varepsilon = 0.005$; for (iii), K^* is the number of signatures for which the posterior probability of being assigned to the spike is less than 0.05; and for (iv), (v), and (vi), K^* is the suggested solution returned by the corresponding package. For (vii), since no automatic selection method is provided by the software, we include signatures for which the posterior mean cosine similarity between the signature and the uniform distribution is less than 0.95.

As expected, all methods but BayesNMF accurately estimate K^0 when $\tau = 0$ (zero overdispersion), since the Poisson NMF model is correct in this case. Meanwhile, when $\tau = 0.15$ (mild overdispersion), we observe a noticeably different pattern of results. Our CompNMF+cosmic method, which leverages informative priors based on all 67 COSMIC signatures, correctly recovers K^0 more often than all other methods, both when $K^0 = 6$ and $K^0 = 10$. CompNMF, which does not rely on COSMIC, tends to slightly overestimate K^0. This is likely due to the introduction of spurious signatures used by the model to accommodate the overdispersion. Interestingly, signeR works well even when the model is mildly misspecified, as it tends to be conservative (Drummond et al., 2023). SigProfiler is even more conservative, tending to slightly underestimate K^0 in these simulations. In contrast, SignatureAnalyzer and PoissonCUSP strongly overestimate K^0 when there is overdispersion. BayesNMF performs poorly since it incorrectly uses a Gaussian likelihood.

Figure 2(B) displays the precision and sensitivity, averaged over the 20 replicates, for each model in each setting. Here, we define *precision* as the proportion of estimated signatures that have a cosine similarity ≥ 0.9 with at least one of the ground truth signatures, and *sensitivity* is the proportion of ground truth signatures for which there is an estimated signature with cosine similarity ≥ 0.9. Since this is a simulation study, the ground truth signatures ρ_k^0 and r_k^0 are known. The cutoff value of 0.9 was chosen following Islam et al. (2022), but we also vary it as described below. The contour lines in Figure 2(B) represent the F_1 score, defined as $2 \times \text{precision} \times \text{sensitivity}/(\text{precision} + \text{sensitivity})$.

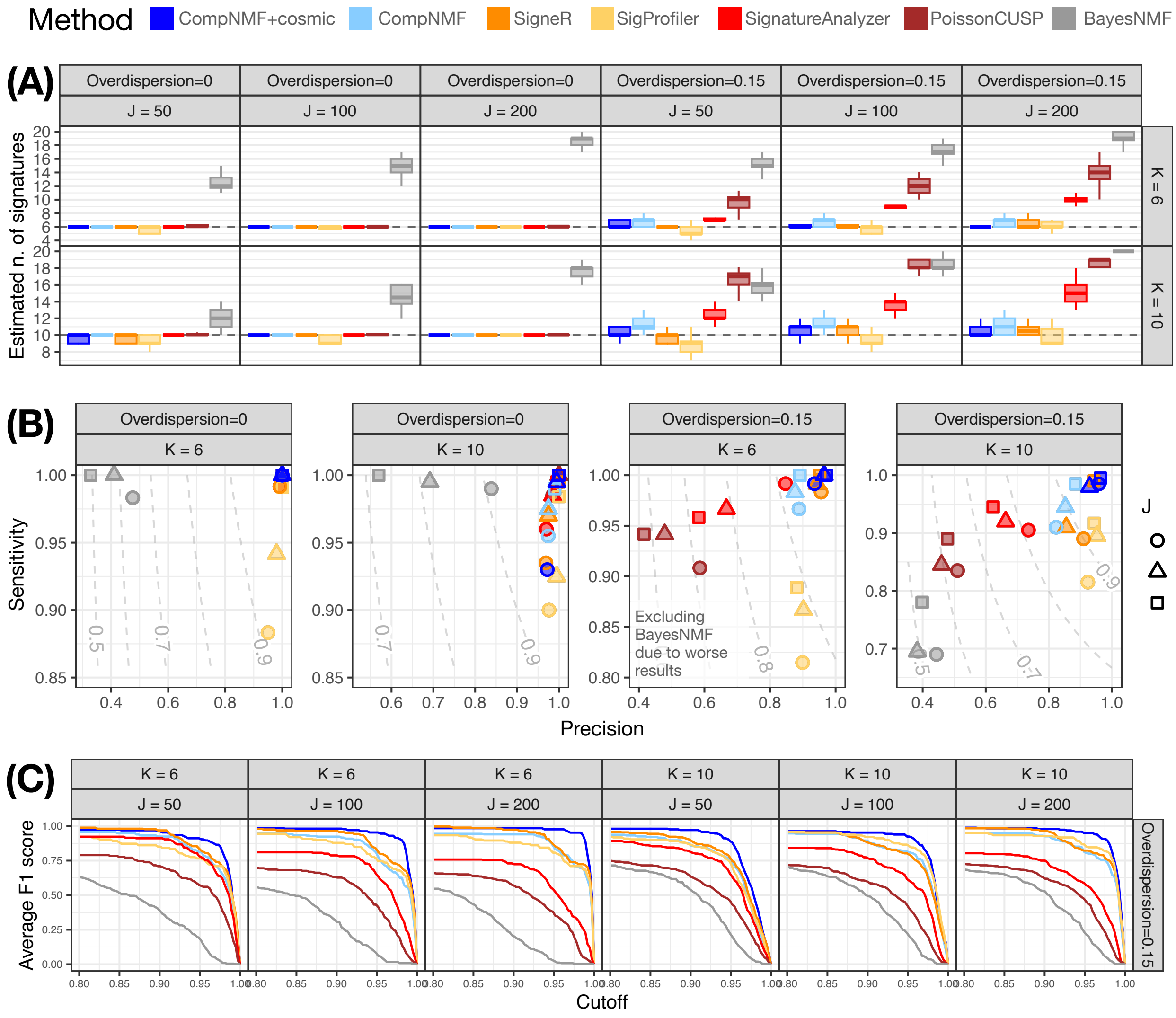


Figure 2: (A) Estimated number of signatures by each model, for varying J, overdispersion, and true number of signatures, across 20 replicated data set in each scenario. The horizontal grey line indicates the true number ($K^0 = 6$ or $K^0 = 10$). (B) Average precision and sensitivity across 20 replicate data sets in each scenario, with a 0.9 cutoff for the cosine similarity. The dashed contour lines in the background indicate the F_1 score. (C) Average F_1 score as a function of the cosine similarity cutoff, across 20 replicates in each scenario, when the overdispersion is set to 0.15.

Points in the top-right corner of the plot in Figure 2(B) indicate better performance. When the Poisson model is correct (no overdispersion), all of the methods perform well, except that SigProfiler has lower sensitivity for smaller sample sizes. CompNMF+cosmic consistently performs the best, with precision and sensitivity close to one in all of these settings. CompNMF and signeR also perform well, but with somewhat reduced precision and sensitivity when K^0 is larger and when there is overdispersion. SigProfiler has good precision but much lower sensitivity, particularly in the presence of overdispersion. Meanwhile, SignatureAnalyzer and PoissonCUSP struggle in the overdispersed scenarios, exhibiting severely degraded precision as well as low sensitivity. Finally, BayesNMF shows good sensitivity under no overdispersion but suffers the most when there is overdispersion.

To see the range of performance exhibited by each method as the cosine similarity cutoff varies, Figure 2(C) shows the average F_1 score versus the cutoff value. As before, we see that CompNMF+cosmic is a clear standout, generally exhibiting the best performance overall, and particularly excelling at cutoff values between 0.9 and 0.98. CompNMF, signeR, and SigProfiler are roughly comparable by this metric, whereas SignatureAnalyzer and especially PoissonCUSP and BayesNMF suffer from significantly lower F_1 scores.

The dominant performance of CompNMF+cosmic demonstrates the benefits of using our informative prior. Notably, since CompNMF+cosmic includes all $K_{\text{pre}} = 67$ COSMIC signatures—whereas only $K^0_{\text{pre}} = 4$ COSMIC signatures are used to generate the simulated data—this implies that 63 of these signatures are correctly compressed out of the model, illustrating the effectiveness of our compressive hyperprior. signeR performs impressively well in these simulations, tending to fall between CompNMF and CompNMF+cosmic in nearly all metrics. However, signeR tends to be computationally slow; see Section S6.3.

The poor performance of SignatureAnalyzer and PoissonCUSP under misspecification appears to be due to overfitting. Indeed, the overestimation of K^0 seen in Figure 2(A), along with the fact that SignatureAnalyzer and PoissonCUSP yield the lowest root-mean-square error for the count matrices (see Section S6), is a clear sign of overfitting. This is not surprising for PoissonCUSP, as its flexible nonparametric nature allows it to fit the data closely. Unfortunately, in this case, it is fitting noise rather than signal.

6 Application

6.1 Benchmark comparison on the 21 breast cancer dataset

In this section, we apply our method to the 21 breast cancer dataset used by Nik-Zainal et al. (2012) in their landmark paper originating the study of mutational signatures in cancer. In particular, we aim to assess the effect of our compressive hyperprior for small sample sizes and evaluate whether our informative prior provides an advantage in practice.

The dataset is based on whole-genome sequencing of $J = 21$ patients, and consists of mutation counts for each of the $I = 96$ channels for each patient. We obtained the data from the signeR package (Drummond et al., 2023). The total mutation count is fairly homogeneous across patients, except for patient PD4120a, for whom a large number of mutations was detected $(70,690)$ compared to the others $(18,871$ on average). This is attributable to the fact that PD4120a was sequenced at nearly 200x coverage compared to 30x for the others. We evaluate the performance of CompNMF and CompNMF+cosmic compared to the same alternatives as in Section 5, excluding BayesNMF due to its worse performance overall. Details of the settings used in all methods are reported in Section S8.

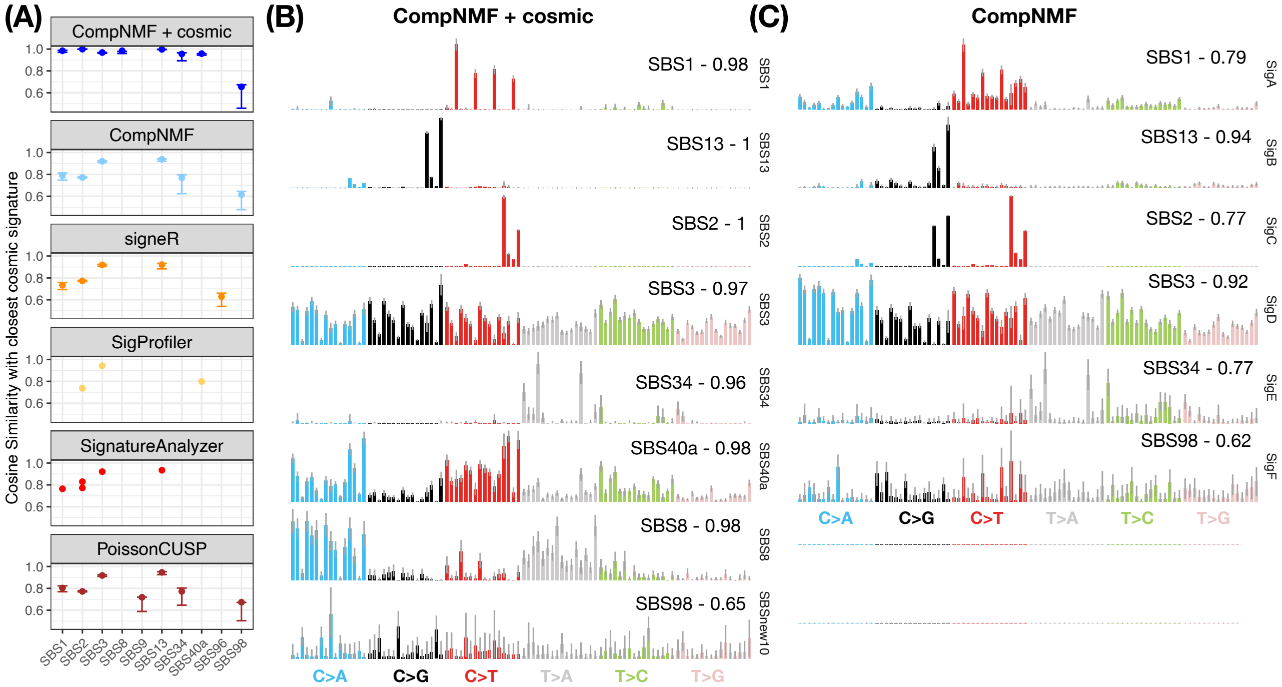


Figure 3: (A) Cosine similarity between the signatures inferred by each method and the best matching signature in COSMIC. Vertical error bars indicate 90% posterior credible intervals. (B) and (C) Signatures inferred by CompNMF+cosmic and CompNMF. The vertical grey lines indicate univariate 90% posterior credible intervals for each mutational channel. The numbers on the top-right corners denote the cosine similarity with the best matching COSMIC signature.

All of the methods fit the count matrix roughly equally well, with the exception of SigProfiler. Specifically, the root-mean-squared error (RMSE) between the mutation count matrix X and $\hat{R}\hat{\Theta}$ was 9.51 for CompNMF, 9.57 for CompNMF+cosmic, 9.81 for signeR, 10.07 for PoissonCUSP, 10.08 for SignatureAnalyzer, and 37.08 for SigProfiler.

However, major differences were observed in the sets of estimated signatures. For each method, Figure 3(A) shows the cosine similarity between each estimated signature and the best matching COSMIC signature. Overall, CompNMF+cosmic recovers the most signatures with the highest cosine similarities, which is not surprising since COSMIC is used to construct the informative prior. Nonetheless, this compellingly demonstrates that the informative prior enables more signal to be extracted from the data. For instance, CompNMF+cosmic finds evidence of the presence of SBS8 and SBS40a, which are not recovered by any other method except for SigProfiler, though with moderate cosine similarity. SBS8 appears in other breast cancers and may be associated with homologous recombination deficiency. SBS40a is a flat signature that appears in every cancer type (Alexandrov et al., 2020).

CompNMF recovers all but two of the signatures found by CompNMF+cosmic, although with somewhat lower cosine similarities. PoissonCUSP yields similar results to CompNMF, but includes what appears to be a spurious match to SBS9, since it has low cosine similarity and SBS9 is not found by any other method. Next, signeR is also similar to CompNMF, but does not recover SBS34, which

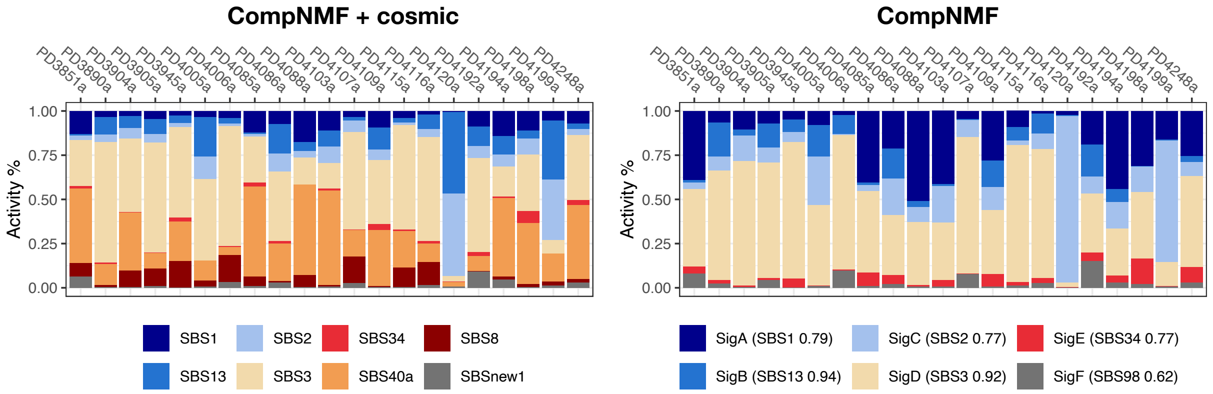


Figure 4: Average posterior loadings (%) by patient in CompNMF+cosmic and CompNMF.

does appear to be truly present since CompNMF+cosmic detects it in many samples. SignatureAnalyzer recovers four of the same signatures as CompNMF; however, it misses SBS34 and produces two signatures that are both matched to SBS2. Lastly, SigProfiler recovers only three signatures. The strong performance of the CompNMF methods demonstrates the utility of our compressive hyperprior.

Figures 3(B) and (C) show the estimated signature vectors r_k for our two methods (CompNMF+cosmic and CompNMF); corresponding figures for all other methods are in Section S8. The informative prior enables CompNMF+cosmic to obtain clean signature estimates that are unambiguously matched to known COSMIC signatures, while still allowing for dataset-specific departures. In contrast, due to the small size of this dataset, all of the other methods produce "merged" signatures that are combinations of two or more COSMIC signatures. For instance, CompNMF+cosmic perfectly distinguishes SBS2 and SBS13 —two APOBEC related dignatures— whereas these are merged into a single signature by all other methods; compare SBS2 and SBS13 in Figure 3(B) to CompNMF SigC in Figure 3(C). Similarly, CompNMF+cosmic perfectly separates SBS1 and SBS40a, whereas these two are combined by all the other methods; see SigA in Figure 3(C) and see Section S8.

The vertical bars in Figure 3(A,B,C) indicate model-based uncertainty in the form of 90% credible intervals, if provided. This uncertainty quantification is particularly useful for identifying signatures that may be spurious. For instance, for both CompNMF+cosmic and CompNMF, the estimated signature matched to SBS98 has high uncertainty in the cosine similarity and in the signature vector itself, indicating that one should be skeptical about the validity of the estimated signature and whether it should be matched to SBS98.

Figure 4 shows the loadings θ_{kj} estimated by our models. The CompNMF+cosmic loadings provide insight into the interpretation of the other methods' results; see Section S8. For instance, in Figure 4, the loadings for SBS40a appear to have been split up and added to the loadings for SBS1 and SBS3 to make the CompNMF loadings for SigA and SigD, respectively. Indeed, it appears that SBS40a and SBS1 (Figure 3(B)) were agglomerated to make SigA (Figure 3(C)), and that SBS40a and SBS3 were combined in SigD. Similarly, CompNMF SigC is a combination of SBS2 and SBS13. This is apparent in patient PD4120a, for whom almost all mutations are attributed to SigC by CompNMF, whereas CompNMF+cosmic splits the loading into equal contributions from SBS2 and SBS13.

6.2 Indels in pancreatic adenocarcinoma

We further test our compressive NMF approach on indel data, which tends to exhibit greater sparsity than SBS data. Indels are the second most common type of mutation and are typically around 10x less frequent than SBS mutations (Alexandrov et al., 2020). We consider a dataset from a group of pancreatic cancer patients in the ICGC cohort (`Panc-AdenoCA`, Synapse repository syn7364923).[2] The data comprise $J = 241$ individuals with a total of 233,175 indel mutations across the $I = 83$ channels. The patients are fairly heterogeneous in terms of the number of mutations, with a sample mean of 967 indels, standard deviation of 3,415, and a maximum of 48,137. The mutation count matrix is sparse, with 41% of the entries being equal to zero.

We compare CompNMF and CompNMF+cosmic with SigProfiler and SignatureAnalyzer. We exclude signeR and PoissonCUSP, since signeR is not designed for indels while PoissonCUSP showed somewhat worse performance in our simulations. The 23 indel signatures in COSMICv3.4 are used for the informative prior in CompNMF+cosmic and for interpretation of the results. For both compressive NMF methods, we run four chains of the MCMC algorithm, randomly initializing from the prior and running for 12,000 iterations, keeping the last 2,000 iterations for posterior inference. We use only the chain with the highest log-posterior. For R, Θ, and μ, respectively, the average effective sample sizes are 1,438, 1,549, and 548 for CompNMF and 1,425, 1,698 and 578 for CompNMF+cosmic, indicating good convergence of the MCMC chains. See Section S8 for further details.

Figure 5(A) summarizes the results in terms of cosine similarity to the nearest COSMIC signature and estimated loading. CompNMF and CompNMF+cosmic yield similar results, both estimating six signatures and assigning them similar loadings; in particular, ID6, ID12, and ID2 have high cosine similarity with their matching estimates. SigProfiler finds five signatures, with a match to ID9 instead of ID12 and ID23, although the cosine similarity is not high. SignatureAnalyzer finds 10 signatures, but the ones matching to ID8, ID18, ID19, and ID21 have very small loadings and relatively low cosine similarity. This corroborates the finding from our simulations that SignatureAnalyzer tends to return a larger number of signatures, but not necessarily with high quality. The RMSE between $\hat{R}\hat{\Theta}$ and X is 5.24 for CompNMF, 5.28 for CompNMF+cosmic, 5.35 for SignatureAnalyzer, and a much larger 20.57 for SigProfiler, following the same trend as on the breast cancer data.

Figure 5(B) shows the set of signatures inferred by CompNMF+cosmic. The posterior credible intervals are tight, suggesting high confidence about the solution. Based on the cosine similarities, it

[2] `https://www.synapse.org/Synapse:syn7364923`, `finalconsensuspassonly.snvmnvindel.icgc.public.maf.gz`

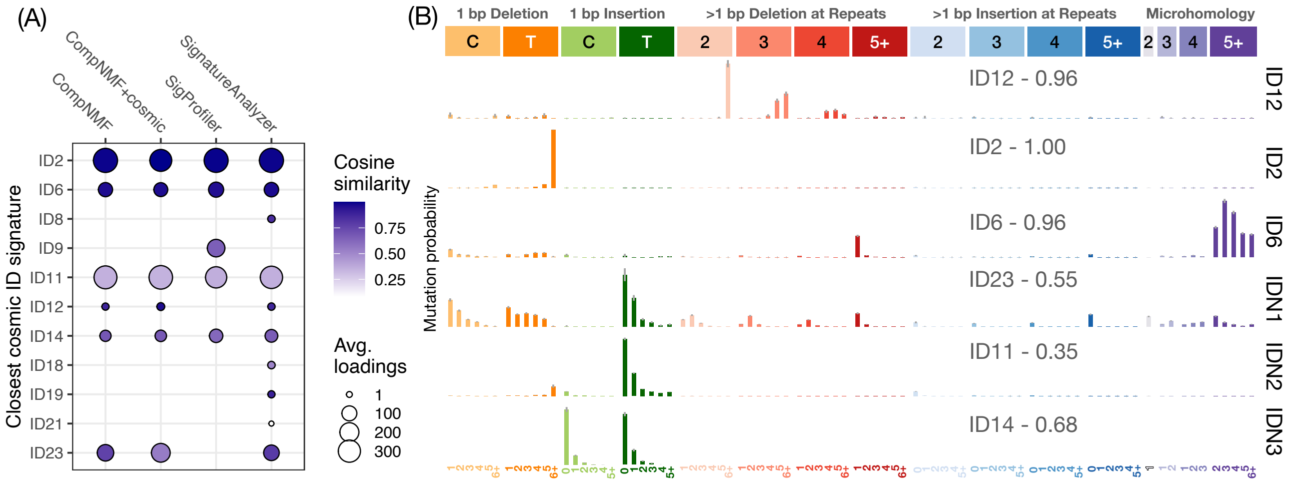


Figure 5: (A) NMF solution from each model on the Panc-AdenoCA dataset. Points indicate the closest indel signature from COSMIC v3.4 and the color intensity denotes the cosine similarity. Size of points depicts the average loading associated to the signature across patients. (B) Indel signatures found by CompNMF+cosmic. Grey bars indicate 90% posterior credible intervals.

is clear that the model has recovered COSMIC signatures ID2, ID12, and ID6, while the other three estimated signatures—labelled IDN1, 2, and 3—appear to be novel. The full results for all methods are reported in Section S8.

7 Discussion

This article introduces a novel Bayesian NMF method that obtains state-of-the-art performance for mutational signatures analysis. In particular, our compressive hyperprior provides a simple but effective—and theoretically well-justified—technique for determining the subset of active factors. This enables the use of rich informative priors based on the COSMIC database of known signatures, significantly boosting the method's precision and sensitivity for recovering true signatures. The informative prior also disambiguates the allocation of loadings to signatures, resulting in more accurate estimation of the contribution of each signature to each sample, as well as clarifying the results of other methods. Furthermore, the method provides posterior uncertainty quantification, which helps distinguish real from spurious signatures and can be used for downstream analyses.

Our framework removes scaling ambiguities by normalizing the signatures via Dirichlet priors. However, strict identifiability of the parameters requires further investigation. Current approaches in Bayesian factor analysis obtain identifiability by enforcing sparsity in the loadings through shrinkage priors (Carvalho et al., 2008; Frühwirth-Schnatter et al., 2024), and strict sparsity constraints over the loadings are also needed to guarantee the uniqueness of an NMF solution in non-Bayesian settings (Donoho and Stodden, 2003). Alternatively, near-identifiability can be obtained using volumeregularized NMF solutions (Ang and Gillis, 2019), but such methods still lack a fully Bayesian counterpart for count data. This suggests a possible direction for future research.

There are several other interesting directions for future work as well. First, one can envision several extensions of the model. Following Grabski et al. (2023), it would be interesting to jointly model multiple studies or multiple cancer types using a hierarchical model with study-specific or cancer type-specific parameters. Another useful extension of the model would be to include sample-specific covariates, which could be helpful in improving targeted therapies (Aguirre et al., 2018). Additionally, the scope of applicability of the compressive hyperprior technique is potentially broader than Poisson NMF models, and might prove useful in other latent factorization models such as Gaussian factor models (Bhattacharya and Dunson, 2011; Legramanti et al., 2020) or in user-item recommendation (Gopalan et al., 2014), especially when prior information on the factors is available.

Finally, while our method can handle mild misspecification in the form of small overdispersion, it is fundamentally based on the assumption that the counts are Poisson distributed – like all of the leading methods considered in our empirical results (signeR, SigProfiler, SignatureAnalyzer, PoissonCUSP). Consequently, larger departures from the assumed Poisson NMF model can be expected to negatively impact the performance of all of these methods. Of course, overdispersion can simply be handled by modeling the data as negative binomial rather than Poisson (Lyu et al., 2020). However, there are many other plausible sources of misspecification, and any parametric elaboration of the model will inevitably be misspecified in some way. Thus, an important area for future work is providing improved robustness to misspecification for mutational signatures analysis and NMF more generally.

Acknowledgments

We would like to thank Scott Carter for suggesting this research direction. We also thank Jonathan Huggins, Catherine Xue, and Giovanni Parmigiani for many helpful discussions. J.W.M. was supported by the Department of Defense grant W81XWH-18-1-0357 and the National Institutes of Health grant 5R01CA240299.

Code availability

Code is publicly accessible via GitHub at https://github.com/alessandrozito/CompressiveNMF.

Supplementary material

The supplementary material contains proofs of the theoretical results, the rationale for the Poisson NMF model likelihood, simulation details, and additional empirical results.

References

Abramowitz, M. and I. A. Stegun (1972). *Handbook of Mathematical Functions with formulas, graphs, and mathematical tables.*

Aguirre, A. J., J. A. Nowak, N. D. Camarda, et al. (2018). Real-time genomic characterization of advanced pancreatic cancer to enable precision medicine. *Cancer Discovery 8*(9), 1096–1111.

Alexandrov, L. B., J. Kim, N. J. Haradhvala, et al. (2020). The repertoire of mutational signatures in human cancer. *Nature 578*(7793), 94–101.

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, et al. (2013). Signatures of mutational processes in human cancer. *Nature 500*(7463), 415–421.

Ang, A. M. S. and N. Gillis (2019). Algorithms and comparisons of nonnegative matrix factorizations with volume regularization for hyperspectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12*(12), 4843–4853.

Armero, C. and M. Bayarri (1997). A Bayesian analysis of a queueing system with unlimited service. *Journal of Statistical Planning and Inference 58*(2), 241–261.

Ayed, F. and F. Caron (2021). Nonnegative Bayesian nonparametric factor models with completely random measures. *Statistics and Computing 31*(5), 63.

Bhattacharya, A. and D. B. Dunson (2011). Sparse Bayesian infinite factor models. *Biometrika 98*(2), 291–306.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Brouwer, T., J. Frellsen, and P. Lió (2017). Comparative Study of Inference Methods for Bayesian Nonnegative Matrix Factorisation. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.*

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association 103*(484), 1438–1456.

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience 2009*, 785152.

Donoho, D. and V. Stodden (2003). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, Volume 16.

Drummond, R. D., A. Defelicibus, M. Meyenberg, R. Valieris, E. Dias-Neto, R. A. Rosales, and I. T. da Silva (2023). Relating mutational signature exposures to clinical data in cancers via signeR 2.0. *BMC Bioinformatics 24*(1), 439.

Dunson, D. B. and A. H. Herring (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics 6*(1), 11–25.

Fischer, A., C. J. Illingworth, P. J. Campbell, and V. Mustonen (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology 14*(4), R39.

Frühwirth-Schnatter, S. (2023). Generalized cumulative shrinkage process priors with applications to sparse Bayesian factor analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 381*(2247), 20220148.

Frühwirth-Schnatter, S., D. Hosszejni, and H. F. Lopes (2024). Sparse Bayesian factor analysis when the number of factors is unknown. *Bayesian Analysis*, 1 – 48.

Gillis, N. (2020). *Nonnegative Matrix Factorization.*

Gopalan, P., F. Ruiz, R. Ranganath, and D. Blei (2014). Bayesian nonparametric Poisson factorization for recommendation systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Volume 33, pp. 275–283.

Grabski, I., L. Trippa, and G. Parmigiani (2023). Bayesian multi-study non-negative matrix factorization for mutational signatures. *bioRxiv:10.1101/2023.03.28.534619*.

Greenman, C., P. Stephens, R. Smith, et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature 446*(7132), 153–158.

Gulhan, D. C., J. J.-K. Lee, G. E. M. Melloni, I. Cortés-Ciriano, and P. J. Park (2019). Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nature Genetics 51*(5), 912–919.

Hoffman, M. D., D. M. Blei, and P. R. Cook (2010). Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 439–446.

Islam, S. A., M. Díaz-Gay, Y. Wu, et al. (2022). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics 2*(11), 100179.

Kim, J., K. W. Mouw, P. Polak, et al. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics 48*(6), 600–606.

Koh, G., A. Degasperi, X. Zou, S. Momen, and S. Nik-Zainal (2021). Mutational signatures: emerging concepts, caveats and clinical applications. *Nature Reviews Cancer 21*(10), 619–637.

Lal, A., K. Liu, R. Tibshirani, A. Sidow, and D. Ramazzotti (2021). De novo mutational signature discovery in tumor genomes using SparseSignatures. *PLOS Computational Biology 17*(6), 1–24.

Lee, D. D. and H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

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

Liu, Y., W. Dong, W. Song, and L. Zhang (2019). Bayesian nonnegative matrix factorization with a truncated spike-and-slab prior. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1450–1455.

Lu, J. and C. P. Chai (2022). Robust Bayesian nonnegative matrix factorization with implicit regularizers. *ArXiv preprint arXiv:2208.10053*.

Lu, J. and X. Ye (2022). Flexible and hierarchical prior for Bayesian nonnegative matrix factorization. *ArXiv preprint arXiv:2205.11025*.

Lyu, X., J. Garret, G. Rätsch, and K.-V. Lehmann (2020). Mutational signature learning with supervised negative binomial non-negative matrix factorization. *Bioinformatics 36*(Issue Supplement_1), i154–i160.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks.* Number 118. Springer. Lecture Notes in Statistics.

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

Nik-Zainal, S., L. B. Alexandrov, D. C. Wedge, et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell 149*(5), 979–993.

Rahiche, A. and M. Cheriet (2022). Variational Bayesian orthogonal nonnegative matrix factorization over the Stiefel manifold. *IEEE Transactions on Image Processing 31*, 5543–5558.

Ročková, V. and E. I. George (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*(516), 1608–1622.

Rosales, R. A., R. D. Drummond, R. Valieris, E. Dias-Neto, and I. T. da Silva (2016). signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics 33*(1), 8–16.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 689–710.

Schmidt, M. N., O. Winther, and L. K. Hansen (2009). Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation*, pp. 540–547.

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

Townes, F. W. and B. E. Engelhardt (2023). Nonnegative spatial factorization applied to spatial genomics. *Nature Methods 20*(2), 229–238.

Wang, S., Z. Tao, T. Wu, and X.-S. Liu (2020). Sigflow: an automated and comprehensive pipeline for cancer genome mutational signature analysis. *Bioinformatics 37*(11), 1590–1592.

Xue, C., A. Zito, and J. W. Miller (2024). Improved control of dirichlet location and scale near the boundary.

Zhou, M. (2015). Infinite edge partition models for overlapping community detection and link prediction. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Volume 38, pp. 1135–1143.

Zhou, M. (2018). Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis 13*(4), 1065–1093.

Zhou, M. and L. Carin (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 307–320.

Supplementary Material for “Compressive Bayesian non-negative matrix factorization for mutational signatures analysis”

Section S1 presents an in-depth comparison of our method and existing Bayesian NMF models in the literature. Section S2 contains the proofs of the theoretical results. Section S3 provides additional results related to the inverse Kummer distribution, the marginal distribution of the latent counts, and the behavior of the posterior under a fixed-strength hyperprior. Section S4 gives a detailed description of the sampler for our NMF model with informative priors. Section S5 provides additional background on mutational signatures, including a first-principles derivation of the model. Section S6 contains additional details of the simulations in Section 5, including an adaptation of the CUSP model of Legramanti et al. (2020) to the setting of Poisson NMF, and an additional simulation under sparser data. Section S7 presents sensitivity analyses for the results in the simulation, including a comparison between the compressive and the fixed-strength hyperprior, and sensitivity to the choice of K, ε α, and a. Finally, Section S8 provides additional results on the application, including sensitivity analysis for the 21 breast cancer dataset and details of the pancreatic cancer application.

S1 Differences with previous work

In this section, we discuss the differences between our compressive NMF approach and other related methods in the literature. Specifically, Section S1.1 highlights the differences with Tan and Févotte (2013); Section S1.2 discusses other ARD methods in the literature; Section S1.3 focuses on Bayesian Poisson factorization methods for recommender systems; Section S1.4 presents an overview of Bayesian nonparametric approaches; Section S1.5 discusses other parametric Bayesian NMF models; and Section S1.6 discusses the relationship between the compressive hyperprior and continuous global-local shrinkage priors.

S1.1 Tan and Févotte’s ARD-NMF algorithm

Tan and Févotte (2013) introduce an ARD-based NMF algorithm that is used by SignatureAnalyzer (Kim et al., 2016; Alexandrov et al., 2020), one of the most prominent methods for mutational signatures analysis. They present a general majorization-minimization algorithm for NMF under the penalized β-divergence loss, solving the following optimization problem:

$$\min_{R\geq\mathbb{R}_+^{I\times K},\Theta\in\mathbb{R}_+^{K\times J},\lambda_k\in\mathbb{R}_+} \frac{1}{\phi}\sum_{i=1}^{I}\sum_{j=1}^{J} d_\beta\big(X_{ij}\mid r_k^\top\theta_k\big) + \sum_{k=1}^{K}\frac{1}{\lambda_k}\big(f(r_k)+f(\theta_k)+b_0\big) + c_0\log\lambda_k,$$

where $d_\beta(x \mid y)$ is the β-divergence, f is a penalty on r_{ik} and θ_{kj}, λ_k are relevance weights needed for the selection, and b and c are constants. When $\phi = 1$ and $\beta = 1$, $d_1(x \mid y) = x\log(x/y) - x + y$ is the KL divergence, which is equivalent to the Poisson likelihood in Equation (1). Moreover, if f is an ℓ_1 penalty, then the penalty term corresponds to using exponential priors on θ_{kj} and r_{ik}, specifically, $r_{ik} \mid \lambda_k \sim \text{Exp}(1/\lambda_k)$ and $\theta_{kj} \mid \lambda_k \sim \text{Exp}(1/\lambda_k)$, where $\mathbb{E}(r_{ik}) = \mathbb{E}(\theta_{kj}) = \lambda_k$, and $\lambda_k \sim \text{InvGamma}(a_0, b_0)$. This is an ARD structure in which λ_k controls the mean of loadings and the signatures jointly. In addition to developing an algorithm to solve the optimization above, they provide a data-driven approach for choosing a_0 and b_0 at every iteration. The rank K^* is selected via a thresholding rule after optimization.

There are several important differences between this method and ours. First, Tan and Févotte (2013) use *maximum a posteriori* (MAP) estimation, which only provides a point estimate of the parameters, without any uncertainty quantification. In contrast, we employ a Gibbs sampler to draw samples from the full posterior, providing uncertainty quantification for the signatures and the associated loadings. Second, the relevance weights λ_k in the Tan and Févotte (2013) model simultaneously control r_{ik} and θ_{kj}, which are both unnormalized. This implies that the solution of Tan and Févotte (2013) suffers from scaling ambiguities. Meanwhile, in our model, μ_k only controls the loadings (Equation (3)) since the signatures are normalized to sum to one (Equation (2)). This solves the scaling issue and allows us to easily incorporate prior knowledge on the COSMIC signatures in our framework, which are themselves normalized. It is not clear whether the Tan and Févotte (2013) approach can be extended to use such prior information, especially since their exponential priors for the signatures depend on λ_k alone. Hence, while their framework exclusively considers an unsupervised setting, ours can have both unsupervised components and supervised components akin to variable selection, thanks to the theoretical results in Theorem 3 and Corollary 1. Such theoretical derivations are unique to our framework; for instance, their λ_k does not have a closed-form posterior in the same way as our μ_k, to our knowledge. For these reasons, we argue that the compressive hyperprior in Equation (4) has a clearer interpretation than the prior over λ_k in Tan and Févotte (2013), who need to tune a_0 and b_0 at every iteration of the optimization to perform their selection. Finally, the Tan and Févotte (2013) estimates degrade rapidly when the model is not exactly correct (see the SignatureAnalyzer results in Figures 2, S6.2 and S6.4), whereas our framework provides robustness to small departures from the model.

S1.2 Other ARD-based NMF methods

Brouwer et al. (2017) consider a class of Bayesian NMF models with ARD priors that we compare with in simulations; see the BayesNMF results in Figures 2, S6.2 and S6.4. While their relevance weights have gamma priors, their model is based on a Gaussian likelihood, which is not well suited for our application since mutation count data is nonnegative integer valued and may be sparse. Consequently, BayesNMF does not perform well for mutational signatures analysis. Moreover, Brouwer et al. (2017) do not provide an explicit rule to prune the unneeded signatures in their paper and their code. For this reason, in our simulations, we exclude the signatures that appear relatively flat in the posterior, since empirically we observe that any extra signatures tend to be flat.

Mørup and Hansen (2009) present an NMF algorithm that normalizes the sum of squares of the signature entries, that is, $\sum_{i=1}^{I} r_{ik}^2 = 1$, but does not impose a non-negativity constraint $r_{ik} \geq 0$. Hence, this model is not suitable for mutational signature analysis, since the true mutation probabilities cannot be negative.

Relatedly, Rahiche and Cheriet (2022) present a Bayesian orthogonal NMF model for continuous data, using a Gaussian likelihood. Specifically, the prior distribution on the signature matrix is uniform on the space of orthogonal matrices, so that $R^\top R = I$ almost surely, and the model dimension is regulated via gamma ARD priors over the loadings. This framework is not suitable for our application, since there is no biological reason why mutational signatures should be orthogonal. Moreover, the prior on R may take on negative values, which is dealt with by setting them to zero in their estimating algorithm.

S1.3 Bayesian Poisson factorization for recommender systems

Gopalan et al. (2015) propose a model for recommendation systems that employs the same Poisson NMF likelihood as our model (Equation (1)). However, they use a different hierarchical prior structure in which $r_{ik} \mid \xi_i \sim \text{Gamma}(a, \xi_i)$ with $\xi_i \sim \text{Gamma}(a', a'/b')$, and $\theta_{kj} \mid \eta_j \sim \text{Gamma}(c, \eta_j)$ with $\eta_j \sim \text{Gamma}(c', c'/d')$, where $i = 1, \dots, I$ are *users* and $j = 1, \dots, J$ are *items* in a recommendation system. Here, ξ_i and η_j are global weights specific to user i and item j, respectively, and there is no weight controlling the global contribution of each k. See also Gopalan et al. (2014) for a closely related model. While they mention that their hierarchical gamma priors are helpful in capturing sparse factors when the shape parameters are set to small values, their model is not specifically designed for rank selection. Rather, their contribution is centered around obtaining fast inferences for the model via variational Bayes methods, and they do not explore the distributional properties of r_{ik} and θ_{kj}.

Gopalan et al. (2014) introduce a nonparametric version of the model presented by Gopalan et al. (2015). This version is designed to infer the rank K^* from the data by setting $K = \infty$ in Equation (1) and employing a Bayesian nonparametric prior. Specifically, they adopt a gamma process prior: $r_{ik} = s_i \cdot v_{ik} \prod_{\ell=1}^{k-1}(1 - v_{\ell k})$ where $v_{ik} \sim \text{Beta}(1, \alpha)$, $s_i \sim \text{Gamma}(\alpha, c)$, and $\theta_{kj} \sim \text{Gamma}(a, b)$ for $k = 1, \dots, \infty$. Notice that this structure could be equivalently imposed on the loadings θ_{kj}, leaving r_{ik} unconstrained. Again, Gopalan et al. (2014) rely on variational inference for posterior computation, using a finite truncation to the stick-breaking construction. Neither of these models (Gopalan et al., 2014, 2015) has an ARD structure, and they do not consider any normalization across mutational channels. Consequently, they exhibit the same scaling ambiguities as the method of Tan and Févotte (2013) and do not facilitate the use of informative priors based on COSMIC signatures.

S1.4 Bayesian nonparametric factorization approaches

Along the same lines as Gopalan et al. (2014), several methods employ a finite truncation of the gamma process to develop Bayesian nonparametric priors and select the number of factors in an NMF model. For example, Hoffman et al. (2010) consider $X_{ij} \sim \text{Exp}(\sum_{k=1}^{K} \vartheta_k r_{ik} \theta_{kj})$ with $r_{ik} \sim \text{Gamma}(a, a)$, $\theta_{kj} \sim \text{Gamma}(b, b)$, and $\vartheta_k \sim \text{Gamma}(\alpha/K, \alpha c)$. In this case, K is a truncation level for the gamma process prior, which is obtained when $K \to \infty$ (Kingman, 1993). They mention briefly that only a few ϑ_k are large when K is large, but no theoretical justification is provided. Moreover, r_{ik} and θ_{kj} are all independent and identically distributed *a priori*, so in particular, they do not normalize the signatures. Similar finite approximations of the gamma process for count data are considered by Zhou (2018) in the context of negative binomial factorizations, and by Zhou (2015) in an infinite community detection problem where communities are overlapping. Additionally, Ayed and Caron (2021) extend the model of Zhou (2015) by using priors based on completely random measures to perform similar community detection inference in count networks. Refer to Zhou and Carin (2015) for a general account.

In a separate thread, several Bayesian nonparametric priors based on multiplicative processes have been developed for factor analysis (Bhattacharya and Dunson, 2011; Durante, 2017; Legramanti et al., 2020; Schiavon et al., 2022; Frühwirth-Schnatter, 2023). These infer the appropriate rank of the factorization by increasingly penalizing the number of latent dimensions through the prior. A member of this family is the CUSP prior (Legramanti et al., 2020), which we compare with in our simulations; see Section S6.2 for details.

It is worth stressing that the approaches discussed above are all finite approximations of an infinite stochastic process. Instead, our framework is a purely parametric one where K is a large but fixed upper bound that does not influence the prior on θ_{kj} or μ_k. A closer analogy to our compressive approach would be the mechanisms underlying the emptying out of extra components in overfitted mixture models (Rousseau and Mengersen, 2011). Furthermore, the introduction of informative priors based on COSMIC would result in an unnatural asymmetry in the nonparametric models described above, given the required exchangeability of the mechanism governing the rank selection.

In a separate instance, a recent contribution from Townes and Engelhardt (2023) describes a Poisson NMF model with Gaussian process priors over the loadings to account for spatial covariates in a spatial transcriptomics application. An interesting direction would be to combine this spatial model with our ARD-based approach, especially since ARD has long been used with Gaussian processes.

S1.5 Other Bayesian NMF models

The original formulation of Poisson NMF was described in Cemgil (2009), and later used in Rosales et al. (2016) and Drummond et al. (2023) for mutational signature analysis in single-study settings. These papers treat K as a tuning hyperparameter and run a separate NMF decomposition for each of a range of choices of K, choosing the final decomposition that minimizes some information criteria. Related to this, Grabski et al. (2023) consider a multi-study version of Rosales et al. (2016) where signatures are excluded from the model through Bernoulli random variables. Similar to our framework, they also incorporate COSMIC priors in their framework. See the recent contribution of Hansen et al. (2025) for a further extension using probit-hyperprior structures.

More recently, Lu and Chai (2022) and Lu and Ye (2022) study the general NMF (both Gaussian or exponential) problem under various types of penalization and choices of hierarchical priors, respectively. Specifically, Lu and Chai (2022) presents a general class of priors that enforce a desired L_p penalization for any $p \geq 1$, as protection against overfitting. Instead, Lu and Ye (2022) considers a rectified-normal prior over (unnormalized) signatures and loadings that ensure sparsity in the final solution. Both contributions do not consider count data.

S1.6 Global-local shrinkage priors

Global-local shrinkage priors such as the horseshoe (Polson and Scott, 2011) bear a resemblence to our compressive hierarchical model, but there are some fundamental differences. In such models, there is a local parameter for each effect of interest and global parameters that govern a set of effects as a group. Sparsity is induced by giving each global and local parameter a heavy-tailed prior with considerable mass near zero, effectively approximating spike-and-slab mixtures in a continuous way (Polson and Scott, 2011; Bhadra et al., 2019). In this way, sparsity can occur at either the individual (local) or group (global) level. See Datta and Dunson (2016) and Hamura et al. (2022) for examples of horseshoe-type priors applied to Poisson likelihoods.

Similarly, our compressive hierarchical model features both global and local parameters, in the form of relevance weights μ_k and loadings θ_{kj}, respectively. However, we do not employ spike-and-slab priors or continuous approximations thereof. Instead, sparsity is induced via our strength-matching hyperprior on μ_k, which shrinks the relevance weights (global parameters) and, in turn, the loadings (local parameters) for unneeded factors down to ε; see Section 3.3. Finally, unlike global-local approaches, our model does not require further hierarchical structure among the μ_k's, since the values for the hyperparameters in Equation (4) are sufficient to drive the selection.

S2 Proofs

Proof of Theorem 1. The proof proceeds by directly marginalizing r_k and θ_k from the joint distribution. For notational simplicity, we first handle the general case of $\mu_k \sim \text{InvGamma}(a_0, b_0)$, and then plug in the values of a_0 and b_0 for the compressive hyperprior. Under the model in Equation (5), the joint density of $Y = (Y_{ijk})$, $\Theta = (\theta_{kj})$, $R = (r_{ik})$, and $\mu = (\mu_k)$ is

$$\begin{aligned}
\pi(Y, \Theta, R, \mu) \propto & \Big\{ \prod_{i,j,k} e^{-r_{ik}\theta_{kj}} \frac{(r_{ik}\theta_{kj})^{Y_{ijk}}}{Y_{ijk}!} \Big\} \times && \text{(Latent Poisson counts)} \\
& \Big\{ \prod_{i,k} r_{ik}^{\alpha-1} \Big\} \times && \text{(Dirichlet prior on } r_k\text{)} \\
& \Big\{ \prod_{j,k} \Big(\frac{a}{\mu_k}\Big)^{a} \theta_{kj}^{a-1} e^{-a\theta_{kj}/\mu_k} \Big\} \times && \text{(Gamma prior on } \theta_k\text{)} \\
& \Big\{ \prod_{k} \mu_k^{-a_0-1} e^{-b_0/\mu_k} \Big\}, && \text{(InvGamma prior on } \mu_k\text{)}
\end{aligned} \tag{S1}$$

dropping constants of proportionality. Since $\sum_i r_{ik} = 1$ for all k, we have $\prod_i e^{-r_{ik}\theta_{kj}} = e^{-\theta_{kj}}$. Thus, by Equation (S1), we have

$$\pi(Y, \Theta, R, \mu) = f(Y, R) \Big\{ \prod_{j,k} \theta_{kj}^{\sum_i Y_{ijk}+a-1} e^{-\theta_{kj} - a\theta_{kj}/\mu_k} \Big\} \Big\{ \prod_k \mu_k^{-Ja-a_0-1} e^{-b_0/\mu_k} \Big\} \tag{S2}$$

where $f(Y, R)$ is a function that does not depend on Θ or μ. Hence,

$$\begin{aligned}
\pi(\mu \mid Y) \underset{\mu}{\propto} \pi(Y, \mu) &= \int\!\!\int \pi(Y, \Theta, R, \mu)\, d\Theta\, dR \\
&= \Big\{ \int f(Y, R)\, dR \Big\} \Big\{ \prod_k \mu_k^{-Ja-a_0-1} e^{-b_0/\mu_k} \Big\} \Big\{ \prod_{j,k} \int \theta_{kj}^{\sum_i Y_{ijk}+a-1} e^{-\theta_{kj} - a\theta_{kj}/\mu_k} d\theta_{kj} \Big\} \\
&\underset{\mu}{\propto} \Big\{ \prod_k \mu_k^{-Ja-a_0-1} e^{-b_0/\mu_k} \Big\} \Big\{ \prod_{j,k} \frac{\Gamma(a + \sum_i Y_{ijk})}{(1 + a/\mu_k)^{a+\sum_i Y_{ijk}}} \Big\}.
\end{aligned}$$

Since this factors over k into products that depend on $\mu_1, \ldots, \mu_K$, respectively, it follows that

$$\begin{aligned}
\pi(\mu_k \mid Y) &\propto \mu_k^{-Ja-a_0-1} e^{-b_0/\mu_k} \Big(1 + \frac{a}{\mu_k}\Big)^{-Ja-\sum_{i,j} Y_{ijk}} \\
&\propto \mu_k^{-(a_0 - \sum_{i,j} Y_{ijk})-1} \Big(\frac{\mu_k}{a} + 1\Big)^{-Ja-\sum_{i,j} Y_{ijk}} e^{-b_0/\mu_k} \\
&\propto \text{InvKummer}\Big(\mu_k \;\Big|\; a_0 + Ja,\, b_0,\, Ja + \sum_{i,j} Y_{ijk},\, a\Big)
\end{aligned} \tag{S3}$$

by Definition 1. The proof is completed by letting $a_0 = aJ + 1$ and $b_0 = \varepsilon a J$. □

Proof of Theorem 2. Recall that $(\theta_{kj} \mid Y, \mu_k) \sim \operatorname{Gamma}(a + Y_{jk}, a/\mu_k + 1)$. Therefore, $\mathbb{E}(\theta_{kj} \mid Y, \mu_k) = (a + Y_{jk})\mu_k/(\mu_k + a)$. By the law of iterated expectation,

$$\mathbb{E}(\theta_{kj} \mid Y) = \mathbb{E}\big(\mathbb{E}(\theta_{kj} \mid Y, \mu_k) \,\big|\, Y\big) = (a + Y_{jk})\mathbb{E}\bigg(\frac{\mu_k}{\mu_k + a} \,\bigg|\, Y\bigg).$$

This last expectation is available analytically in terms of Kummer hypergeometric functions, as follows. As in the proof of Theorem 1, we first consider $\mu_k \sim \operatorname{InvGamma}(a_0, b_0)$ for generic a_0 and b_0. Define $C = a^{J\bar{Y}_k - a_0}\Gamma(a_0 + aJ)U(a_0 + aJ, a_0 + 1 - J\bar{Y}_k, b_0/a)$, noting that by Equation (8), C can be interpreted as the normalizing constant of the density of $\mu_k \mid Y$, that is, the constant of proportionality in Equation (S3). Then

$$\begin{aligned}
\mathbb{E}(\theta_{kj} \mid Y) &= (a + Y_{jk})\mathbb{E}\bigg(\frac{\mu_k/a}{\mu_k/a + 1} \,\bigg|\, Y\bigg) \\
&= \frac{a + Y_{jk}}{aC} \int_0^\infty \mu_k^{-(a_0 - J\bar{Y}_k - 1) - 1} \bigg(\frac{\mu_k}{a} + 1\bigg)^{-aJ - J\bar{Y}_k - 1} e^{-b_0/\mu_k} \mathrm{d}\mu_k \\
&= \frac{a + Y_{jk}}{aC} a^{J\bar{Y}_k + 1 - a_0}\Gamma(a_0 + aJ)U(a_0 + aJ, a_0 - J\bar{Y}_k, b_0/a) \\
&= (a + Y_{jk})\frac{U(a_0 + aJ, a_0 - J\bar{Y}_k, b_0/a)}{U(a_0 + aJ, a_0 + 1 - J\bar{Y}_k, b_0/a)}.
\end{aligned}$$

The proof is completed by letting $a_0 = aJ + 1$ and $b_0 = \varepsilon aJ$. □

The following preliminary results are used in the proof of Theorem 3.

Lemma 1. *Let $\varepsilon > 0$, $a > 0$, and $y_n \geq 0$ such that $y_n \to y$ for some $y \in [0, \infty)$. Let T_n be the random variable on $(0, \infty)$ with probability density function*

$$f_n(t) \propto t^{2an}(1 + t)^{-ny_n - an}e^{-\varepsilon nt} \tag{S4}$$

for $t > 0$. Then

$$\sqrt{n}(T_n - t_n^*) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\bigg(0, \frac{1 + t^*}{2a/(t^*)^2 + \varepsilon}\bigg),$$

where $t_n^* = \big(\sqrt{(y_n - a + \varepsilon_n)^2 + 8a\varepsilon_n} - (y_n - a + \varepsilon_n)\big)/(2\varepsilon_n)$, $\varepsilon_n = \varepsilon - 1/n$, and $t^* = \lim_{n\to\infty} t_n^*$.

Proof. The proof follows by direct application of the results in Miller (2021). We write $f_n(t) \propto \exp(-ng_n(t))e^{-t}$, where $g_n(t) = -2a\log(t) + (y_n + a)\log(1 + t) + (\varepsilon - 1/n)t$, to put the density of T_n in the form considered by Miller (2021), with $\Theta = (0, \infty)$ and $\pi(t) = e^{-t}$. In Lemma 2, we show that the functions $g_n(t)$ and the point t^* satisfy the assumptions of Theorem 5 in Miller (2021). Thus, the density of $\sqrt{n}(T_n - t_n^*)$ converges in total variation to $\mathcal{N}(0, 1/g''(t^*))$, where $g''(t^*) = (2a/(t^*)^2 + \varepsilon)/(1 + t^*)$ by Lemma 2, part 5. Convergence in total variation implies convergence in distribution, by the portmanteau lemma. □

For $x \in \mathbb{R}$ and $\delta \geq 0$, define $B_\delta(x) = \{t \in \mathbb{R} : |t - x| < \delta\}$.

Lemma 2. *Let $\varepsilon > 0$, $a > 0$, and $y_n \geq 0$ such that $y_n \to y$ for some $y \in [0, \infty)$, and define*

$$g_n(t) = -2a\log(t) + (y_n + a)\log(1+t) + (\varepsilon - 1/n)t, \tag{S5}$$

$$g(t) = -2a\log(t) + (y + a)\log(1+t) + \varepsilon t, \tag{S6}$$

for $t > 0$. Define $t^* = \big(\sqrt{(y - a + \varepsilon)^2 + 8a\varepsilon} - (y - a + \varepsilon)\big)/(2\varepsilon)$ and let $\delta \in (0, t^*)$. Then

(1) *g_n has continuous third derivatives and (g_n''') is uniformly bounded on $B_\delta(t^*)$,*

(2) *$g_n \to g$ pointwise,*

(3) *$g(t) > g(t^*)$ for all $t \in (0,\infty) \setminus \{t^*\}$,*

(4) *$\liminf_{n\to\infty} \inf_{t\in(0,\infty)\setminus K} g_n(t) > g(t^*)$ where $K = \{t \in \mathbb{R} : |t - t^*| \leq \delta/2\}$,*

(5) *$g''(t^*) = (2a/(t^*)^2 + \varepsilon)/(1 + t^*) > 0$.*

Proof. (1) The derivatives of g_n are

$$g_n'(t) = -\frac{2a}{t} + \frac{y_n + a}{1+t} + \varepsilon - \frac{1}{n}, \tag{S7}$$

$$g_n''(t) = \frac{2a}{t^2} - \frac{y_n + a}{(1+t)^2}, \tag{S8}$$

$$g_n'''(t) = -\frac{4a}{t^3} + \frac{2(y_n + a)}{(1+t)^3}. \tag{S9}$$

Thus, $g_n'''(t)$ is continuous at all $t \in (0, \infty)$ and is uniformly bounded on $B_\delta(t^*)$ since

$$|g_n'''(t)| \leq \left|-\frac{4a}{t^3}\right| + \left|\frac{2(y_n + a)}{(1+t)^3}\right| \leq \frac{4a}{t^3} + \frac{2(y_n + a)}{t^3} \leq \frac{2\sup_n y_n + 6a}{(t^* - \delta)^3} < \infty$$

for all $n \geq 1$ and all $t \in B_\delta(t^*)$, since $y_n \to y$ and $\delta < t^*$ by assumption.

(2) This follows directly from the assumption that $y_n \to y$ as $n \to \infty$.

(3) Similarly to Equation (S7),

$$g'(t) = -2a/t + (y + a)/(1 + t) + \varepsilon.$$

Setting $g'(t) = 0$ and solving via the quadratic formula, we find that any critical point must occur at $\hat{t} = (-b \pm \sqrt{b^2 + 8a\varepsilon})/(2\varepsilon)$, where $b = y - a + \varepsilon$. Since $|-b| = \sqrt{b^2} < \sqrt{b^2 + 8a\varepsilon}$, the only solution on $(0, \infty)$ is $\hat{t} = t^*$, where t^* is defined in the statement of the lemma. Thus, $g'(t)$ has a unique zero on $(0, \infty)$. Since $g'(t) \to -\infty$ as $t \to 0$, and $g'(t) \to \varepsilon$ as $t \to \infty$, the intermediate value theorem implies that $g'(t) < 0$ for all $t \in (0, t^*)$ and $g'(t) > 0$ for all $t \in (t^*, \infty)$. Therefore, by the fundamental theorem of calculus, $g(t)$ is strictly monotone decreasing on $(0, t^*]$ and strictly monotone increasing on $[t^*, \infty)$. This implies that $g(t)$ is uniquely minimized at t^*, as claimed.

(4) Let $\varepsilon_n = \varepsilon - 1/n$, and suppose n is large enough that $\varepsilon_n > 0$. By applying the argument in (3) to g_n instead of g, we have that g_n is strictly monotone decreasing on $(0, t_n^*)$ and strictly monotone increasing on (t_n^*, ∞), where $t_n^* = \big(\sqrt{(y_n - a + \varepsilon_n)^2 + 8a\varepsilon_n} - (y_n - a + \varepsilon_n)\big)/(2\varepsilon_n)$. Note that $t_n^* \to t^*$ as $n \to \infty$, since $y_n \to y$ and $\varepsilon_n \to \varepsilon$. Let $k_1 = t^* - \delta/2$ and $k_2 = t^* + \delta/2$, so that $K = [k_1, k_2]$. Choose N such that for all $n \geq N$, we have (i) $\varepsilon_n > 0$, (ii) $t_n^* \in K$, (iii) $g_n(k_1) \geq (g(k_1) + g(t^*))/2$, and (iv) $g_n(k_2) \geq (g(k_2) + g(t^*))/2$, using (3) and the fact that $g_n \to g$ pointwise. Let $c = (\min\{g(k_1), g(k_2)\} + g(t^*))/2$, noting that $c > g(t^*)$. Then for all $n \geq N$, we have (a) if $0 < t < k_1$ then $g_n(t) \geq g_n(k_1) \geq c$ since g_n is monotone decreasing on $(0, t_n^*)$, and (b) if $t > k_2$ then $g_n(t) \geq g_n(k_2) \geq c$ since g_n is monotone increasing on (t_n^*, ∞). Therefore,

$$\liminf_{n\to\infty} \inf_{t \in (0,\infty)\setminus K} g_n(t) \geq c > g(t^*).$$

(5) From the proof of (3), we have $g'(t^*) = 0$, and thus, $(y + a)/(1 + t^*) = 2a/t^* - \varepsilon$. Therefore,

$$g''(t^*) = \frac{2a}{(t^*)^2} - \frac{y+a}{(1+t^*)^2} = \frac{2a}{(t^*)^2} - \frac{2a/t^* - \varepsilon}{1+t^*} = \frac{2a/(t^*)^2 + \varepsilon}{1+t^*} > 0.$$

□

Theorem S2.1. *In the setting of Lemma 1, for any sequence $d_1, d_2, \ldots \in [0, \infty)$ such that $d_n \to \infty$ and $|y_n - y| = o(d_n/\sqrt{n})$, we have*

$$\mathbb{P}(|T_n - t^*| \geq d_n/\sqrt{n}) \xrightarrow[n\to\infty]{} 0$$

where t^* is defined as in Lemma 1.

Proof. By Lemma 3, $t_n^* - t^* = C_n(y_n - y) + q_n$ where C_n converges to a constant and $q_n = \mathcal{O}(1/n)$. Thus,

$$\frac{\sqrt{n}}{d_n}|t_n^* - t^*| \leq |C_n|\frac{\sqrt{n}}{d_n}|y_n - y| + \frac{\sqrt{n}}{d_n}|q_n| \longrightarrow 0$$

as $n \to \infty$, since $|y_n - y| = o(d_n/\sqrt{n})$ and $d_n \to \infty$ by assumption. This implies that $d_n - \sqrt{n}|t_n^* - t^*| = d_n(1 - (\sqrt{n}/d_n)|t_n^* - t^*|) \to \infty$, since $d_n \to \infty$. By the triangle inequality, $|T_n - t^*| \leq |T_n - t_n^*| + |t_n^* - t^*|$.

Thus,

$$\begin{aligned}\mathbb{P}(|T_n - t^*| \geq d_n/\sqrt{n}) &\leq \mathbb{P}(|T_n - t_n^*| + |t_n^* - t^*| \geq d_n/\sqrt{n}) \\ &= \mathbb{P}(|\sqrt{n}(T_n - t_n^*)| \geq d_n - \sqrt{n}|t_n^* - t^*|) \xrightarrow[n\to\infty]{} 0,\end{aligned}$$

because $\sqrt{n}(T_n - t_n^*)$ converges in distribution to a normal distribution. □

Lemma 3. *In the setting of Lemma 1, we have*

$$t_n^* - t^* = C_n(y_n - y) + q_n \tag{S10}$$

where $C_n \to -t^*/\sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon}$ and $q_n = \mathcal{O}(1/n)$.

Proof. Define $x_n = y_n - a + \varepsilon$ and $x = y - a + \varepsilon$. Then $x_n \to x$ and $x_n - x = y_n - y$. In terms of x_n and x, we have

$$\begin{aligned} t_n^* &= \frac{\sqrt{(x_n - 1/n)^2 + 8a(\varepsilon - 1/n)} - x_n + 1/n}{2(\varepsilon - 1/n)}, \\ t^* &= \frac{\sqrt{x^2 + 8a\varepsilon} - x}{2\varepsilon}. \end{aligned}$$

For any $z_1, z_2 \geq 0$, we have $(z_1 - z_2)(z_1 + z_2) = z_1^2 - z_2^2$ and $(\sqrt{z_1} - \sqrt{z_2})(\sqrt{z_1} + \sqrt{z_2}) = z_1 - z_2$. Letting $k_n = \sqrt{(x_n - 1/n)^2 + 8a(\varepsilon - 1/n)} + \sqrt{x^2 + 8a\varepsilon}$, we manipulate $t_n^* - t^*$ as follows:

$$\begin{aligned} t_n^* - t^* &= \frac{1}{2(\varepsilon - 1/n)}\Big(\sqrt{(x_n - 1/n)^2 + 8a(\varepsilon - 1/n)} - x_n + 1/n - 2(\varepsilon - 1/n)t^*\Big) \\ &= \frac{1}{2(\varepsilon - 1/n)}\Big(\sqrt{(x_n - 1/n)^2 + 8a(\varepsilon - 1/n)} - \sqrt{x^2 + 8a\varepsilon} - (x_n - x) + \frac{2t^* + 1}{n}\Big) \\ &= \frac{1}{2(\varepsilon - 1/n)}\Big(\frac{(x_n - 1/n)^2 - x^2 - 8a/n}{k_n} - (x_n - x) + \frac{2t^* + 1}{n}\Big) \\ &= \frac{1}{2(\varepsilon - 1/n)}\Big(\frac{x_n^2 - x^2 - 2x_n/n + 1/n^2 - 8a/n}{k_n} - (x_n - x) + \frac{2t^* + 1}{n}\Big) \\ &= \frac{1}{2(\varepsilon - 1/n)}\Big(\frac{(x_n - x)(x_n + x) - 2x_n/n + 1/n^2 - 8a/n}{k_n} - (x_n - x) + \frac{2t^* + 1}{n}\Big) \\ &= \frac{1}{2(\varepsilon - 1/n)}\Big(\Big(\frac{x_n + x}{k_n} - 1\Big)(x_n - x) - \frac{2x_n + 8a}{k_n n} + \frac{1}{k_n n^2} + \frac{2t^* + 1}{n}\Big) \\ &= \frac{1}{2(\varepsilon - 1/n)}\Big(\frac{x_n + x}{k_n} - 1\Big)(y_n - y) + q_n \\ &= C_n(y_n - y) + q_n \end{aligned}$$

where

$$C_n = \frac{1}{2(\varepsilon - 1/n)}\Big(\frac{x_n + x}{k_n} - 1\Big)$$

and

$$q_n = \frac{1}{2(\varepsilon - 1/n)} \left(-\frac{2x_n + 8a}{k_n n} + \frac{1}{k_n n^2} + \frac{2t^* + 1}{n} \right). \tag{S11}$$

Since $x_n \to x$, it follows that $k_n \to 2\sqrt{x^2 + 8a\varepsilon} > 0$. Therefore, $C_n \to -t^*/\sqrt{x^2 + 8a\varepsilon}$ and $q_n = \mathcal{O}(1/n)$. □

Theorem S2.2. *Let $\varepsilon > 0$, $a > 0$, and $y_n \geq 0$ such that $y_n \to y$ for some $y \in [0, \infty)$. If $\mu_n \sim \mathrm{InvKummer}(2an + 1, \varepsilon an, ny_n + an, a)$ then we have the following results.*

(1) *For any sequence $c_1, c_2, \ldots \in [0, \infty)$ such that $c_n \to \infty$, $c_n/\sqrt{n} \to 0$, and $|y_n - y| = o(c_n/\sqrt{n})$, we have*

$$\mathbb{P}(|\mu_n - \mu_*| \geq c_n/\sqrt{n}) \xrightarrow[n\to\infty]{} 0$$

where $\mu_* = 2a\varepsilon/(\sqrt{(y - a + \varepsilon)^2 + 8a\varepsilon} - (y - a + \varepsilon))$.

(2) *There exist constants $D_1, D_2, \ldots \in \mathbb{R}$ and $v_1, v_2, \ldots \in \mathbb{R}$ such that*

$$\sqrt{n}(\mu_n - \mu_*) - \Delta_n \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0, \ \frac{\mu_*^3(\mu_* + a)}{2a\mu_*^2 + a^2\varepsilon}\right),$$

where $\Delta_n = D_n\sqrt{n}(y_n - y) + v_n$, $\lim_{n\to\infty} D_n = \mu_*^2/(2a\varepsilon + \mu_*(y - a + \varepsilon))$, and $v_n \to 0$.

Proof. Let $\mu_n \sim \mathrm{InvKummer}(2an + 1, \varepsilon an, ny_n + an, a)$ and define $T_n = a/\mu_n$. Then T_n has the density $f_n(t)$ studied in Lemma 1 and Theorem S2.1; see Equation (S4).

(1) Pick n large enough so that $\mu_* - c_n/\sqrt{n} > 0$. Then, $|\mu_n - \mu_*| \leq c_n/\sqrt{n}$ if and only if $\mu_* - c_n/\sqrt{n} \leq \mu_n \leq \mu_* + c_n/\sqrt{n}$, or equivalently, $a/(\mu_* + c_n/\sqrt{n}) \leq a/\mu_n \leq a/(\mu_* - c_n/\sqrt{n})$. Define $d_1, d_2, \ldots$ such that

$$d_n = \left(\frac{a/\mu_*}{\mu_* + c_n/\sqrt{n}}\right) c_n,$$

and observe that $d_n \to \infty$ and $d_n/\sqrt{n} \to 0$. Moreover, $|y_n - y| = o(d_n/\sqrt{n})$ since

$$\frac{|y_n - y|}{d_n/\sqrt{n}} = \frac{\mu_* + c_n/\sqrt{n}}{a/\mu_*} \frac{|y_n - y|}{c_n/\sqrt{n}} \xrightarrow[n\to\infty]{} 0 \tag{S12}$$

since $c_n/\sqrt{n} \to 0$ and $|y_n - y| = o(c_n/\sqrt{n})$ by assumption. Then for all n sufficiently large,

$$\frac{a}{\mu_* + c_n/\sqrt{n}} \overset{\text{(a)}}{=} \frac{a}{\mu_*} - \frac{d_n}{\sqrt{n}} \overset{\text{(b)}}{=} t^* - \frac{d_n}{\sqrt{n}} \leq t^* + \frac{d_n}{\sqrt{n}} \overset{\text{(c)}}{=} \frac{a}{\mu_*} + \frac{d_n}{\sqrt{n}} \overset{\text{(d)}}{\leq} \frac{a}{\mu_* - c_n/\sqrt{n}} \tag{S13}$$

where t^* is defined as in Lemma 1. Step (a) holds since $c_n = \mu_* d_n/(a/\mu_* - d_n/\sqrt{n})$. Steps (b) and (c) hold since $t^* = a/\mu_*$. To justify step (d), first note that for any $x \in (0, 1/2)$, we have $1 + x \leq (1 - x)/(1 - 2x)$ because $(1 + x)(1 - 2x) = 1 - x - 2x^2 \leq 1 - x$. Choosing $x = \mu_* d_n/(a\sqrt{n})$, we have $x \in (0, 1/2)$ for all n sufficiently large because $d_n/\sqrt{n} \to 0$, and thus,

$$\frac{a}{\mu_*} + \frac{d_n}{\sqrt{n}} = \frac{a}{\mu_*}\Big(1 + \frac{\mu_* d_n}{a\sqrt{n}}\Big) \le \frac{a}{\mu_*}\Big(\frac{1 - \mu_* d_n/(a\sqrt{n})}{1 - 2\mu_* d_n/(a\sqrt{n})}\Big) = \frac{a}{\mu_* - c_n/\sqrt{n}}.$$

Theorem S2.1 applies since $y_n \to y$, $d_n \to \infty$, and $|y_n - y| = o(d_n/\sqrt{n})$ by Equation (S12). Together, Theorem S2.1 and Equation (S13) imply that

$$1 = \lim_{n\to\infty} \mathbb{P}(|T_n - t^*| \le d_n/\sqrt{n}) \le \lim_{n\to\infty} \mathbb{P}(|\mu_n - \mu_*| \le c_n/\sqrt{n})$$

concluding the proof of (1).

(2) By Lemma 1,

$$\sqrt{n}(T_n - t_n^*) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0,\ \frac{1+t^*}{2a/(t^*)^2 + \varepsilon}\right)$$

and $t_n^* \to t^*$. Hence, by the delta method (van der Vaart, 2000, Theorem 3.8),

$$\sqrt{n}(h(T_n) - h(t_n^*)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0,\ h'(t^*)^2 \frac{1+t^*}{2a/(t^*)^2 + \varepsilon}\right) \tag{S14}$$

where $h(t) = a/t$ for $t > 0$. Since $\mu_n = a/T_n = h(T_n)$ and $\mu_* = a/t^* = h(t^*)$,

$$\sqrt{n}(\mu_n - \mu_*) = \sqrt{n}(h(T_n) - h(t^*)) = \sqrt{n}(h(T_n) - h(t_n^*)) + \sqrt{n}(h(t_n^*) - h(t^*)). \tag{S15}$$

Therefore, combining Equations (S14) and (S15) along with $h'(t) = -a/t^2$ and writing the variance in terms of μ_* yields

$$\sqrt{n}(\mu_n - \mu_*) - \Delta_n \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0,\ \frac{\mu_*^3(\mu_* + a)}{2a\mu_*^2 + a^2\varepsilon}\right) \tag{S16}$$

where $\Delta_n = \sqrt{n}(h(t_n^*) - h(t^*)) = \sqrt{n}(a/t_n^* - a/t^*) = -a\sqrt{n}(t_n^* - t^*)/(t_n^* t^*)$. We can further characterize Δ_n using the proof of Lemma 3. Specifically,

$$t_n^* - t^* = \frac{1}{2(\varepsilon - 1/n)}\Big(\frac{y_n + y - 2a + 2\varepsilon}{k_n} - 1\Big)(y_n - y) + q_n,$$

where q_n is defined in Equation (S11) and satisfies $q_n = \mathcal{O}(1/n)$. Hence,

$$\Delta_n = \sqrt{n}(y_n - y)\underbrace{\frac{-a}{2t_n^* t^*(\varepsilon - 1/n)}\left(\frac{y_n + y - 2a + 2\varepsilon}{k_n} - 1\right)}_{D_n} + \underbrace{\frac{-a\sqrt{n}\,q_n}{t_n^* t^*}}_{v_n}.$$

Note that $v_n \to 0$ since $q_n = \mathcal{O}(1/n)$ and $t_n^* \to t^*$. Finally, since $a/t_n^* \to a/t^* = \mu_*$ and $k_n \to 2\sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon}$,

$$\begin{aligned}\lim_{n\to\infty} D_n &= -\frac{\mu_*^2}{2a\varepsilon}\Big(\frac{y-a+\varepsilon}{\sqrt{(y-a+\varepsilon)^2+8a\varepsilon}} - 1\Big)\\ &= \frac{\mu_*}{\sqrt{(y-a+\varepsilon)^2+8a\varepsilon}}\\ &= \frac{\mu_*^2}{2a\varepsilon + \mu_*(y-a+\varepsilon)}.\end{aligned}$$

□

Proof of Theorem 3. Theorem S2.2 is simply a restatement of Theorem 3 in more familiar notation, where $n = J$, $y_n = \bar{Y}_k$, and μ_n is distributed according to $\mu_k \mid Y$. □

Proof of Corollary 1. Suppose $\bar{Y}_k = o(c_J/\sqrt{J})$ where $c_J \to \infty$ and $c_J/\sqrt{J} \to 0$. Then $y = 0$ in Theorem 3, and therefore the concentration point is $\mu^* = 2a\varepsilon/(\sqrt{(\varepsilon-a)^2+8a\varepsilon} - (\varepsilon - a))$. It holds that $\mu^* < \varepsilon$, since $(\varepsilon-a)^2+8a\varepsilon > (\varepsilon-a)^2+4a\varepsilon = (\varepsilon+a)^2$, and hence, $\sqrt{(\varepsilon-a)^2+8a\varepsilon}-(\varepsilon-a) > 2a$. Let $d = C\varepsilon - \mu^*$, noting that $d > 0$ since $C > 1$ and $\mu^* < \varepsilon$. Then by Theorem 3, for all J sufficiently large that $c_J/\sqrt{J} \leq d$,

$$\mathbb{P}(\mu_k > C\varepsilon \mid Y) \leq \mathbb{P}(|\mu_k - \mu^*| > d \mid Y) \leq \mathbb{P}(|\mu_k - \mu^*| \geq c_J/\sqrt{J} \mid Y) \longrightarrow 0$$

as $J \to \infty$. □

Proof of Theorem 4. From Section 2.5, we have $(\theta_{kj} \mid Y, \mu_k) \sim \mathrm{Gamma}(a + Y_{jk}, a/\mu_k + 1)$ where $Y_{jk} = \sum_{i=1}^I Y_{ijk}$. We use the Laplace transform of $(\theta_{kj} \mid Y)$ to characterize its limit as $J \to \infty$. For all $t > 0$,

$$\begin{aligned}\mathbb{E}(e^{-t\theta_{kj}} \mid Y) &= \mathbb{E}\big(\mathbb{E}(e^{-t\theta_{kj}} \mid Y, \mu_k)\,\big|\, Y\big)\\ &= \mathbb{E}\Big(\Big(\frac{a/\mu_k + 1}{t + a/\mu_k + 1}\Big)^{a+Y_{jk}}\,\Big|\, Y\Big)\\ &\xrightarrow[J\to\infty]{} \Big(\frac{a/\mu_* + 1}{t + a/\mu_* + 1}\Big)^{a+Y_{jk}}.\end{aligned}$$

The first equality holds by the law of iterated expectations. The second equality follows from the Laplace transform of a gamma distribution. The limit is a consequence of the portmanteau lemma (since $x \mapsto (a/x+1)/(t+a/x+1)$ is a bounded continuous function for $x > 0$) and the fact that $(\mu_k \mid Y) \xrightarrow{\mathrm{d}} \mu_*$ by Theorem 3. Notice that the limit is the Laplace transform of $\mathrm{Gamma}(a+Y_{jk},\, a/\mu_* + 1)$. Thus, by the continuity theorem for Laplace transforms (Feller, 1971, Chapter 13, Theorem 2), $(\theta_{kj} \mid Y) \xrightarrow{d} \mathrm{Gamma}(a + Y_{jk},\, a/\mu_* + 1)$, as desired. The second part of the statement concerning the fixed-strength hyperprior can be proved in the same manner, but using the fact that $(\mu_k \mid Y) \xrightarrow{\mathrm{d}} y$ by Theorem S3.2. □

S3 Additional theoretical results

S3.1 Further properties of the Inverse Kummer

We study the relationship between the hyperparameter $\gamma \in \mathbb{R}$ of the inverse Kummer distribution (Definition 1), and the first moment of the distribution. In particular, the following proposition implies that the mean of an inverse Kummer with $\gamma > 0$ is larger than the mean of the corresponding inverse gamma distribution, when $\lambda > 2$.

Proposition S3.1. *Let $\mu \sim \mathrm{InvKummer}(\lambda, \beta, \gamma, \delta)$. If $\lambda > 2$, then $\mathbb{E}(\mu)$ is monotone increasing as a function of γ, for $\gamma \in (0, \infty)$.*

Proof. Fix $\epsilon \in (0, 1)$. Let $\mu_\gamma \sim \mathrm{InvKummer}(\lambda, \beta, \gamma, \delta)$ and $\mu_{\gamma+\epsilon} \sim \mathrm{InvKummer}(\lambda, \beta, \gamma + \epsilon, \delta)$, and define $f(t) = t^{\lambda-1}(1 + t)^{-\gamma} e^{-\beta t/\delta}$. Our aim is to show that $\mathbb{E}(\mu_{\gamma+\epsilon}) \geq \mathbb{E}(\mu_\gamma)$. By Equation (9), this is true if and only if

$$\frac{\int_0^\infty \frac{1}{t(1+t)^\epsilon} f(t)\mathrm{d}t}{\int_0^\infty \frac{1}{(1+t)^\epsilon} f(t)\mathrm{d}t} \geq \frac{\int_0^\infty \frac{1}{t} f(t)\mathrm{d}t}{\int_0^\infty f(t)\mathrm{d}t}. \tag{S17}$$

Let X be the continuous random variable on $(0, \infty)$ with probability density function $p(x) = f(x)/\int_0^\infty f(t)\mathrm{d}t$. Then, after multiplying and dividing the left-hand side by $\int_0^\infty f(t)\mathrm{d}t$, Equation (S17) can be written in terms of expectations as

$$\frac{\mathbb{E}\left(\frac{1}{X(1+X)^\epsilon}\right)}{\mathbb{E}\left(\frac{1}{(1+X)^\epsilon}\right)} \geq \mathbb{E}\left(\frac{1}{X}\right). \tag{S18}$$

or equivalently,

$$\mathrm{Cov}\left(\frac{1}{X}, \frac{1}{(1+X)^\epsilon}\right) \geq 0. \tag{S19}$$

Define $g(x) = -1/x$ and $h(x) = -1/(1 + x)^\epsilon$ for $x \in (0, \infty)$. Observe that, by transformation of random variables, $1/X \sim \mathrm{InvKummer}(\lambda, \beta, \gamma, \delta)$. Thus, by Equation (9), since $\lambda > 2$ by assumption,

$$\mathbb{E}|g(X)|^2 = \mathbb{E}\left(\frac{1}{X^2}\right) < \infty,$$
$$\mathbb{E}|h(X)|^2 = \mathbb{E}\left(\frac{1}{(1+X)^{2\epsilon}}\right) \leq \mathbb{E}\left(\frac{1}{X^{2\epsilon}}\right) < \infty.$$

Therefore, since $g(x)$ and $h(x)$ are monotone increasing and have finite second moments, we have $\mathrm{Cov}(g(X), h(X)) \geq 0$ by Schmidt (2003), which is equivalent to Equation (S19). This completes the proof. □

S3.2 Characterizing the concentration point of inverse Kummer

We characterize the point at which the inverse Kummer concentrates, μ_*, in Theorem 3. Define

$$\mu_*(\varepsilon, y, a) = \frac{2a\varepsilon}{\sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon} - (y-a+\varepsilon)} \tag{S20}$$

for $\varepsilon > 0$, $y \geq 0$, and $a > 0$.

Proposition S3.2. *For all $\varepsilon > 0$ and $a > 0$, $\mu_*(\varepsilon, y, a)$ is monotone increasing as a function of y.*

Proof. Fix $\varepsilon > 0$ and define

$$g(y) = \sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon},$$

so that $\mu_*(\varepsilon, y, a) = 2a\varepsilon/(g(y) - (y-a+\varepsilon))$. Then

$$\frac{\partial \mu_*}{\partial y} = \frac{-2a\varepsilon\big(g'(y) - 1\big)}{\big(g(y) - (y-a+\varepsilon)\big)^2}.$$

Differentiating g, we find that

$$g'(y) = \frac{y-a+\varepsilon}{\sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon}} < 1.$$

Therefore, $\partial\mu_*/\partial y > 0$, showing that μ_* is monotone increasing as a function of y. □

Next, we derive Equation (10) using a first-order Taylor approximation. Fix $y \geq 0$ and $a > 0$, and define

$$h(\varepsilon) = \sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon} \tag{S21}$$

for $\varepsilon > 0$. Differentiating and simplifying, we find that

$$h'(\varepsilon) = \frac{y+\varepsilon+3a}{h(\varepsilon)}. \tag{S22}$$

Thus, $h(0) = |y-a|$ and $h'(0) = (y+3a)/|y-a|$. Hence, a first-order Taylor approximation to h at $\varepsilon = 0$ yields

$$h(\varepsilon) \approx h(0) + h'(0)\varepsilon = |y-a| + \frac{y+3a}{|y-a|}\varepsilon \tag{S23}$$

when ε is small relative to $|y-a|$. Plugging this into the definition of μ_* in Equation (S20), we obtain

$$\mu_*(\varepsilon, y, a) \approx \frac{2a\varepsilon}{|y-a| + \frac{y+3a}{|y-a|}\varepsilon - (y-a+\varepsilon)}. \tag{S24}$$

When $y > a$, we have $|y - a| = y - a$, so in this case Equation (S24) becomes

$$\mu_*(\varepsilon, y, a) \approx \frac{y-a}{2}.$$

Meanwhile, when $0 \leq y < 1$, we have $|y - a| = a - y$, so in this case

$$\mu_*(\varepsilon, y, a) \approx \frac{\varepsilon a(a-y)}{(a-y)^2 + (a+y)\varepsilon}$$

by collecting and rearranging terms. Therefore, the first-order Taylor approximation is

$$\mu_*(\varepsilon, y, a) \approx \begin{cases} \dfrac{y-a}{2} & \text{if } y > a \\ \dfrac{\varepsilon a(a-y)}{(a-y)^2 + (a+y)\varepsilon} & \text{if } 0 \leq y < a \end{cases} \tag{S25}$$

as claimed in Equation (10). Plugging the above formula into the mean of the gamma distributions of Theorem 4 yields the approximation in Equation (11).

S3.3 Relationship between the relevance weights and latent counts

We further characterize the relationship between the latent counts Y_{ijk} and relevance weights μ_k. In particular, we derive the distribution of $Y_{ijk} \mid \mu_k$, integrating out r_k and θ_k in Equation (5). We show that, appealingly, this distribution has a closed-form expression in terms of hypergeometric functions. First, the distribution of $Y_{ijk} \mid \mu_k, r_k$, integrating out θ_{kj} in Equation (5), is easily seen to be $Y_{ijk} \mid \mu_k, r_k \sim \text{NegBin}\big(a, a/(a + r_{ik}\mu_k)\big)$, where the negative binomial is parametrized such that the mean and variance are $\mathbb{E}(Y_{ijk} \mid \mu_k, r_k) = \mu_k r_{ik}$ and $\text{Var}(Y_{ijk} \mid \mu_k, r_k) = \mu_k r_{ik}(a + \mu_k r_{ik})$. When both r_k and θ_k are integrated out, we obtain the following result, where ${}_2F_1(a, b, ; c, z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}$ denotes the Gauss-hypergeometric function and $(a)_n = \Gamma(a + n)/\Gamma(a)$ is the ascending factorial (Abramowitz and Stegun, 1972).

Proposition S3.3. *The probability mass function of $Y_{ijk} \mid \mu_k$ under the model in Equation (5) is*

$$\mathbb{P}(Y_{ijk} = y \mid \mu_k) = \left(\frac{\mu_k}{a}\right)^y \frac{(a)_y (\alpha)_y}{y!(\alpha I)_y} \, {}_2F_1\Big(y + a,\ y + \alpha,\ y + \alpha I,\ -\frac{\mu_k}{a}\Big),$$

for $y \in \{0, 1, 2, \ldots\}$. Furthermore, the mean of this distribution is $\mathbb{E}(Y_{ijk} \mid \mu_k) = \mu_k / I$.

Proof. Recall that $Y_{ijk} \mid \mu_k, r_k, \theta_k \sim \text{Poisson}(r_{ik}\theta_{kj})$ and $\theta_{kj} \mid \mu_k \sim \text{Gamma}(a, a/\mu_k)$. Also, $r_{ik} \sim \text{Beta}(\alpha,\ \alpha I - \alpha)$ by marginalization property of the Dirichlet distribution. Hence,

$$\pi(Y_{ijk} = y \mid \mu_k) = \int_0^1 \left\{ \int_0^\infty \mathbb{P}(Y_{ijk} = y \mid \mu_k, r_k, \theta_k) p(\theta_{kj} \mid \mu_k) \mathrm{d}\theta_{kj} \right\} p(r_{ik}) \mathrm{d}r_{ik}$$

$$
\begin{aligned}
&= \int_0^1 \left\{ \int_0^\infty (r_{ik}\theta_{kj})^y \frac{e^{-r_{ik}\theta_{kj}}}{y!} \frac{(a/\mu_k)^a}{\Gamma(a)} \theta_{kj}^{a-1} e^{-(a/\mu_k)\theta_{kj}} d\theta_{kj} \right\} \frac{\Gamma(\alpha I)}{\Gamma(\alpha)\Gamma(\alpha I - \alpha)} r_{ik}^{\alpha-1}(1-r_{ik})^{\alpha I-\alpha-1} \mathrm{d}r_{ik} \\
&= \frac{(a/\mu_k)^a}{y!\,\Gamma(a)} \frac{\Gamma(\alpha I)}{\Gamma(\alpha)\Gamma(\alpha I - \alpha)} \int_0^1 \left\{ \int_0^\infty \theta_{kj}^{y+a-1} e^{-(r_{ik}+a/\mu_k)\theta_{kj}} d\theta_{kj} \right\} r_{ik}^{y+\alpha-1}(1-r_{ik})^{\alpha I-\alpha-1} \mathrm{d}r_{ik} \\
&= \frac{(a/\mu_k)^a}{y!\,\Gamma(a)} \frac{\Gamma(\alpha I)}{\Gamma(\alpha)\Gamma(\alpha I - \alpha)} \int_0^1 \frac{\Gamma(y+a)}{(r_{ik}+a/\mu_k)^{y+a}} r_{ik}^{y+\alpha-1}(1-r_{ik})^{\alpha I-\alpha-1} \mathrm{d}r_{ik} \\
&= \frac{(a)_y(\mu_k/a)^y}{y!} \frac{\Gamma(\alpha I)}{\Gamma(\alpha)\Gamma(\alpha I - \alpha)} \int_0^1 (t\mu_k/a+1)^{-(y+a)} t^{y+\alpha-1}(1-t)^{(y+\alpha I)-(y+\alpha)-1} \mathrm{d}t \\
&= \left(\frac{\mu_k}{a}\right)^y \frac{(a)_y(\alpha)_y}{y!(\alpha I)_y} \, {}_2F_1\big(y+a,\, y+\alpha,\, y+\alpha I,\, -\mu_k/a\big),
\end{aligned}
$$

where $(x)_n = \Gamma(x+n)/\Gamma(x)$ is the ascending factorial, and

$$
{}_2F_1(a,b,c,z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\mathrm{d}t,
$$

with $c > b > 0$ and $a, z \in \mathbb{R}$ is an alternative representation of the Gauss-hypergeometric function; see Abramowitz and Stegun (1972). □

This result further explains the role of each μ_k in the mutational process: it directly controls the contribution of signature k in determining the number of mutations X_{ij} in channel i for patient j, for given values of hyperparameters a and α.

S3.4 Posterior distribution of the loadings

We now characterize the posterior density of the loadings θ_{kj} given the tensor of latent counts Y. Interestingly, the density has an analytical form in terms of Kummer hypergeometric functions and belongs to the class of gamma-Kummer continuous mixtures. Note that the expected value of this distribution is presented in Theorem 2 in the main manuscript.

Theorem S3.1. *In the setting of Theorem 1, let $Y_{jk} = \sum_{i=1}^I Y_{ijk}$ and $\bar{Y}_k = \frac{1}{J}\sum_{j=1}^J Y_{jk}$. Then, the density of the loading θ_{kj} is*

$$
\pi(\theta_{kj} \mid Y) = \theta_{kj}^{a+Y_{jk}-1} e^{-\theta_{kj}} \frac{U\big(2aJ+1, J(a-\bar{Y}_k)+a+2+Y_{jk}, \varepsilon J + \theta_{kj}\big)}{\Gamma(a+Y_{jk}) U\big(2aJ+1, J(a-\bar{Y}_k)+2, \varepsilon J\big)}.
$$

Proof. From Section 2.5, recall that $(\theta_{kj} \mid Y, \mu_k) \sim \mathrm{Gamma}(a + Y_{jk}, a/\mu_k + 1)$. For notational simplicity, we first consider the general case where the hyperprior is $\mu_k \sim \mathrm{InvGamma}(a_0, b_0)$, and then we plug in the values of a_0 and b_0 for the compressive hyperprior. In the proof of Theorem 1, we showed that $(\mu_k \mid Y) \sim \mathrm{InvKummer}\big(a_0 + Ja,\, b_0,\, Ja + J\bar{Y}_k,\, a\big)$. As in the proof of Theorem 2, let $C = a^{J\bar{Y}_k - a_0}\Gamma(a_0 + aJ)U(a_0 + aJ, a_0 + 1 - J\bar{Y}_k, b_0/a)$ denote the normalizing constant of the density of $\mu_k \mid Y$, that is, the constant of proportionality in Equation (S3). Then, using the inverse Kummer density in Equation (8) to integrate,

$$
\begin{aligned}
\pi(\theta_{kj} \mid Y) &= \int_0^\infty \pi(\theta_{kj} \mid Y, \mu_k)\pi(\mu_k \mid Y)\mathrm{d}\mu_k \\
&= \frac{a^{a+Y_{jk}}}{C\Gamma(a+Y_{jk})}\theta_{kj}^{a+Y_{jk}-1}e^{-\theta_{kj}}\int_0^\infty \mu_k^{-(a_0-J\bar{Y}_k+a+Y_{kj})-1}\left(1+\frac{\mu_k}{a}\right)^{-aJ-J\bar{Y}_k+a+Y_{jk}}e^{-(b_0+a\theta_{kj})/\mu_k}\mathrm{d}\mu_k \\
&= \frac{a^{a+Y_{jk}}}{C\Gamma(a+Y_{jk})}\theta_{kj}^{a+Y_{jk}-1}e^{-\theta_{kj}}a^{J\bar{Y}_k-a-Y_{jk}-a_0}\Gamma(a_0+aJ)U(a_0+aJ, a_0+1-J\bar{Y}_k+a+Y_{jk}, b_0/a+\theta_{kj}) \\
&= \theta_{kj}^{a+Y_{jk}-1}e^{-\theta_{kj}}\frac{U(a_0+aJ, a_0+1-J\bar{Y}_k+a+Y_{jk}, b_0/a+\theta_{kj})}{\Gamma(a+Y_{jk})U(a_0+aJ, a_0+1-J\bar{Y}_k, b_0/a)}
\end{aligned}
$$

Substituting $a_0 = aJ + 1$ and $b_0 = \varepsilon aJ$ completes the proof. □

S3.5 Theoretical results for the fixed-strength hyperprior

In this section, we derive concentration and asymptotic normality results when using the fixed-strength hyperprior, that is, $\mu_k \sim \mathrm{InvGamma}(a_0, b_0)$ for fixed choices of a_0 and b_0. In our model, the resulting posterior for is $\mu_k \mid Y \sim \mathrm{InvKummer}(a_0 + Ja,\ b_0,\ Ja + J\bar{Y}_k,\ a)$. As before, we switch notation to ease readability, using n, y_n, and μ_n in place of J, $\bar{Y}_k$, and μ_k, respectively. We discuss the interpretation of this result in comparison with the compressive case in Section S7.1.

Theorem S3.2. *Let $a > 0$, $a_0 > 0$, $b_0 > 0$, and $y_n > 0$ such that $y_n \to y$ for some $y > 0$. If $\mu_n \sim \mathrm{InvKummer}(a_0 + na,\ b_0,\ na + ny_n,\ a)$, then:*

(1) *for any sequence $c_1, c_2, \ldots \in [0, \infty)$ such that $c_n \to \infty$ with $c_n/\sqrt{n} \to 0$ and $|y_n - y| = o(c_n/\sqrt{n})$, we have*

$$\mathbb{P}(|\mu_n - y| \geq c_n/\sqrt{n}) \xrightarrow[n\to\infty]{} 0,$$

(2) *defining $\Delta_n = \sqrt{n}(y_n - y)$, we have*

$$\sqrt{n}(\mu_n - y) - \Delta_n \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0,\ \frac{y(y+a)}{a}\right).$$

Proof. Letting $\mu_n \sim \mathrm{InvKummer}(a_0 + na,\ b_0,\ na + ny_n,\ a)$ and $T_n = a/\mu_n$, the density of T_n is $f_n(T_n) \propto t^{a_0+an-1}(1+t)^{-ny_n-an}e^{-b_0t}$. The rest of the proof follows the same steps used in the proofs of Theorem S2.1 and Theorem S2.2.

(1) Similarly to Lemma 1, we have $f_n(t) \propto \exp(-ng_n(t))\,\pi(t)$, where $\pi(t) = t^{a_0-1}e^{-b_0t}$, and

$$
\begin{aligned}
g_n(t) &= -a\log(t) + (y_n + a)\log(1+t), \\
g(t) &= -a\log(t) + (y + a)\log(1+t).
\end{aligned}
$$

Call $t_n^* = a/y_n$ and $t^* = a/y$. It is easy to verify that $g_n'(t_n^*) = 0$ and $g'(t^*) = 0$. Moreover, the conditions of Theorem 5 in Miller (2021) are met for $g_n(t)$ and $g(t)$, by the same arguments used in Lemma 1. In particular, since $g''(t) = a/t^2 - (y+a)/(1+t)^2$, we have $g''(t^*) = g''(a/y) = a(a+y)/y^3 > 0$ and so

$$\sqrt{n}(T_n - t_n^*) = \sqrt{n}(T_n - a/y_n) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0, \frac{a(y+a)}{y^3}\right). \tag{S26}$$

Then, by the same reasoning as Theorem S2.1, it holds that $\mathbb{P}(|T_n - t_n^*| \geq d_n/\sqrt{n}) \to 0$ for any $d_n \to \infty$ and $d_n/\sqrt{n} \to 0$. The proof is completed by mirroring the steps detailed in point (1) of Theorem S2.2, defining

$$d_n = \left(\frac{a/y}{y + c_n/\sqrt{n}}\right) c_n.$$

(2) As in Theorem S2.2, part (2), we can apply the transformation $h(t) = a/t$ to Equation (S26). The delta method (van der Vaart, 2000) then yields

$$\sqrt{n}(h(T_n) - h(a/y_n)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0, h'(a/y)^2 \frac{a(y+a)}{y^3}\right).$$

Hence, we write

$$\sqrt{n}(\mu_n - y) = \sqrt{n}(h(T_n) - h(a/y)) = \sqrt{n}(h(T_n) - h(a/y_n)) + \sqrt{n}(h(a/y_n) - h(a/y)).$$

Simplifying the variance and recalling that $h'(a/y)^2 = y^4/a^2$, we obtain

$$\sqrt{n}(\mu_n - y) - \Delta_n \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}\left(0, \frac{y(y+a)}{a}\right),$$

with $\Delta_n = \sqrt{n}(y_n - y)$. This concludes the proof. □

S4 Gibbs sampler for NMF with informative priors

In this section, we present the general Gibbs sampler for the model in Equation (6) with priors as in Equation (7). Each step follows from simple semi-conjugate prior updates, so we omit the derivations. Note that we set the value of β_k depending on the level of sparsity of each COSMIC signature. Specifically, for a given s_k, we calculate β_k by drawing 1000 samples $\rho_k \sim \text{Dirichlet}(\beta_k s_{1k}, \dots, \beta_k s_{Ik})$ for a range of plausible β_k values (from 10 to 5000, evenly spaced on a log scale), and we select the value for which the median cosine similarity between s_k and the sampled ρ_k vectors is closest to 0.975. This ensures that all signatures have approximately equal variance under the prior. All values are available at https://github.com/alessandrozito/CompressiveNMF. As alternative strategy, one can also leverage the recent approach of Xue et al. (2024), which elicits the hyperparameters of a Dirichlet distribution based on a target location and a notion of distance, such as consine similarity.

Inference in the CompNMF+cosmic model is performed by iterating the following steps.

1. For $i = 1, \dots, I$ and $j = 1, \dots, J$, update the latent mutation counts by drawing
$$(Y_{ij} \mid -) \sim \text{Multinomial}\big(X_{ij}, (\tilde{q}_{ij1}, \dots, \tilde{q}_{ijK_{\text{pre}}}, q_{ij1}, \dots, q_{ijK_{\text{new}}})\big)$$
where $\tilde{q}_{ijk} = \rho_{ik}\omega_{kj}/Q_{ij}$ and $q_{ijk} = r_{ik}\theta_{kj}/Q_{ij}$, with $Q_{ij} = \sum_{k=1}^{K_{\text{pre}}} \rho_{ik}\omega_{kj} + \sum_{k=1}^{K_{\text{new}}} r_{ik}\theta_{kj}$.

2. For $k = 1, \dots, K_{\text{pre}}$, update the COSMIC signatures by drawing
$$(\rho_k \mid -) \sim \text{Dirichlet}\left(\beta_k s_{1k} + \sum_{j=1}^{J} Y_{1jk}, \dots, \beta_k s_{Ik} + \sum_{j=1}^{J} Y_{Ijk}\right).$$

3. For $k = K_{\text{new}} + 1, \dots, K_{\text{pre}} + K_{\text{new}}$, update the *de novo* signatures by drawing
$$(r_k \mid -) \sim \text{Dirichlet}\left(\alpha + \sum_{j=1}^{J} Y_{1jk}, \, \dots, \, \alpha + \sum_{j=1}^{J} Y_{Ijk}\right).$$

4. For $k = 1, \dots, K_{\text{pre}}$ and $j = 1, \dots, J$, update the loadings associated to the COSMIC signatures by drawing
$$(\omega_{kj} \mid -) \sim \text{Gamma}\left(b + \sum_{i=1}^{I} Y_{ijk}, \, \frac{b}{\tau_k} + 1\right).$$

5. For $k = K_{\text{new}} + 1, \dots, K_{\text{pre}} + K_{\text{new}}$ and $j = 1, \dots, J$, update the loadings associated to the *de novo* signatures by drawing
$$(\theta_{kj} \mid -) \sim \text{Gamma}\left(a + \sum_{i=1}^{I} Y_{ijk}, \, \frac{a}{\mu_k} + 1\right).$$

6. For $k = 1, \dots, K_{\text{pre}}$, update the relevance weights associated to the COSMIC signatures by drawing
$$(\tau_k \mid -) \sim \text{InvGamma}\left(2bJ + 1, \, \varepsilon bJ + b\sum_{j=1}^{J} \omega_{kj}\right).$$

7. For $k = K_{\text{new}} + 1, \dots, K_{\text{pre}} + K_{\text{new}}$, update the relevance weights associated to the *de novo* signatures by drawing
$$(\mu_k \mid -) \sim \text{InvGamma}\left(2aJ + 1, \, \varepsilon aJ + a\sum_{j=1}^{J} \theta_{kj}\right).$$

One important behavior we noticed is that, occasionally, the sampler either (i) morphs a novel signature into a COSMIC one even if that COSMIC signature has been specified in the prior, or (ii) morphs an existing COSMIC signature into another COSMIC one. This is due to the multi-modal nature of the NMF model and can be influenced by initialization. While carefully eliciting each β_k as above does help, such incoherence sometimes can hold for relatively flat COSMIC signatures, such as SBS3, SBS5, or SBS40a,b. To solve the issue, we apply a label-switching step at 2/3 of the burn-in phase, where the signatures that have not been compressed out of the model are re-matched to the COSMIC signatures via the Hungarian algorithm. This does not invalidate the MCMC algorithm since it is only performed in the burn-in phase.

S5 Rationale for the model likelihood

S5.1 Types of base pair substitutions

In DNA, there are four bases: cytosine (C), thymine (T), adenine (A), and guanine (G). Considering both strands of the double helix, cytosine always pairs with guanine, and thymine always pairs with adenine. Thus, if we distinguish one of the two strands of a given DNA molecule, there are four possible base pairs at each point: C-G, G-C, T-A, and A-T.

When considering base pair substitutions at a given point, the convention is to distinguish the strand containing the pyrimidine (C or T) before the substitution has been made. Recall that cytosine (C) and thymine (T) are *pyrimidines*, whereas adenine (A) and guanine (G) are *purines*. With this convention, there are six possible types of substitutions at any given point:

	before	after	abbreviation
1	C-G	A-T	C>A
2	C-G	G-C	C>G
3	C-G	T-A	C>T
4	T-A	A-T	T>A
5	T-A	C-G	T>C
6	T-A	G-C	T>G

Sometimes, these are abbreviated denoting only the pre-substitution pyrimidine and what it changes to, as seen above.

These six classes can be further divided by considering the trinucleotide context, that is, the bases directly adjacent to the base undergoing substitution. The convention is to label the context in terms of the bases (C, T, A, or G) on the 5' and 3' sides on the strand containing the pre-substitution pyrimidine. For instance, in a substitution C>A, the C may be flanked by a T on the 5' side and a G on the 3' side:

before	after	abbreviation
TCG	TAG	T[C>A]G
5’ 3’	5’ 3’	

There are $4 \times 4 = 16$ different contexts for each of the original six substitution types. Therefore, there are $16 \times 6 = 96$ single-base substitution types when the trinucleotide context is taken into account.At each position in the genome, one of the two strands contains a pyrimidine C or T, flanked by bases on the 5’ and 3’ sides, say, X and Y, respectively: so the trinucleotide context is either XCY or XTY. Thus, each position in the genome can be in one of $2 \times 4 \times 4 = 32$ possible states. Since there are three possible single-base substitutions at every position, we arrive at a total of $32 \times 3 = 96$ *mutational channels*.

S5.2 Continuous-time Markov process for substitutions

Focusing on one position ℓ in the genome, let us assume mutations at ℓ occur as a time-homogeneous continuous-time Markov process, holding the neighboring bases fixed. More precisely, when the current state is a, it remains a for an Exponential($|\Lambda_{aa}|$) amount of time and then transitions to $b \neq a$ with probability $\Lambda_{ab}/|\Lambda_{aa}|$, where Λ is a 32×32 matrix such that (i) $\Lambda_{ab} \geq 0$ for $a \neq b$, and (ii) $\sum_b \Lambda_{ab} = 0$. This is equivalent to saying that transitions from a to b occur with rate Λ_{ab}; thus, Λ is called the transition rate matrix. See Lawler (2018) for background.

Let S_ℓ^t denote the state at locus ℓ at time t, and let S_ℓ^0 be the state at ℓ for the normal (germline) genome of the individual under consideration. Let $P_{ab}^t = \mathbb{P}(S_\ell^t = b \mid S_\ell^0 = a)$ be the probability that the state is b at time t given that the state is a at time 0. From the theory of continuous-time Markov processes, we have that

$$P^t = \exp(t\Lambda) = \sum_{k=0}^{\infty} \frac{(t\Lambda)^k}{k!}$$

where $\exp(\cdot)$ denotes the matrix exponential. Since the mutation rates Λ_{ab} are very small, it is reasonable to use a first-order Taylor approximation, $P^t \approx I + t\Lambda$.

S5.3 Substitution counts are approximately Poisson distributed

Let a_i and b_i denote the starting and ending states, respectively, for each of the substitution types $i = 1, \ldots, 96$. Let $\lambda_i = \Lambda_{a_i b_i}$, and define $X_i^t = \#\{\ell : S_\ell^0 = a_i, S_\ell^t = b_i\}$, that is, X_i^t is the number of positions in the genome that undergo substitution i, starting at state a_i at time 0 and ending at state b_i at time t.

Now, consider all of the positions ℓ that are in state a at time 0, and to simplify the math, let us assume that (a) no two of these positions are adjacent, and (b) that substitions occur independently across positions. Of the 32 states, only four of them can be reached from a: the state can remain at a, or one of three substitutions can occur. Suppose these three substitutions are $i = 1, 2, 3$, so that the starting states are $a_1 = a_2 = a_3 = a$ and the ending states are b_1, b_2, b_3, respectively. Let $s^0 = (s^0_\ell : \ell = 1, \ldots, L)$ be a fixed vector of starting states for all positions ℓ, and let $n = \#\{\ell : s^0_\ell = a\}$ be the number of positions starting in state a at time 0. Let $\lambda_0 = \lambda_1 + \lambda_2 + \lambda_3$ be the sum of the rates for substitution types $1, 2, 3$. By the definition of P^t_{ab} and the assumption of independence across positions, letting $X^t_0 = X^t_1 + X^t_2 + X^t_3$, the vector $(X^t_1, X^t_2, X^t_3, n - X^t_0)$ follows a multinomial distribution. Specifically, for non-negative integers x_1, x_2, x_3 such that $x_0 := x_1 + x_2 + x_3 \leq n$, we have

$$\begin{aligned} \mathbb{P}(X^t_{1:3} = x_{1:3} \mid S^0 = s^0) &= \frac{n!}{(n - x_0)! x_1! x_2! x_3!} (P^t_{aa})^{n-x_0} \prod_{i=1}^{3} (P^t_{a_i b_i})^{x_i} \\ &\approx \frac{n!}{(n - x_0)! x_1! x_2! x_3!} (1 - t\lambda_0)^{n-x_0} (t\lambda_1)^{x_1} (t\lambda_2)^{x_2} (t\lambda_3)^{x_3} \end{aligned} \tag{S27}$$

by the first-order Taylor approximation, $P^t \approx I + t\Lambda$. Since the genome is large and mutation rates are small, it is natural to assume that n is large and $t\lambda_0 = O(1/n)$. Thus, letting $c = nt\lambda_0$ we have $(1 - t\lambda_0)^n = (1 - c/n)^n \approx e^{-c} = \exp(-nt\lambda_0)$ and $(1 - t\lambda_0)^{-x_0} = (1 - c/n)^{-x_0} \approx 1$ when $x_0 \ll n$, which is the case with high probability. Plugging these approximations into Equation (S27) yields

$$\begin{aligned} &\approx \frac{n!}{(n - x_0)! x_1! x_2! x_3!} \exp(-nt\lambda_0) (t\lambda_1)^{x_1} (t\lambda_2)^{x_2} (t\lambda_3)^{x_3} \\ &= \frac{n!\, n^{-x_0}}{(n - x_0)!} \prod_{i=1}^{3} \exp(-nt\lambda_i) \frac{(nt\lambda_i)^{x_i}}{x_i!} \end{aligned}$$

since $x_0 = x_1 + x_2 + x_3$ and $\lambda_0 = \lambda_1 + \lambda_2 + \lambda_3$ by definition. By Stirling's approximation,

$$\frac{n!\, n^{-x_0}}{(n - x_0)!} \sim \frac{\sqrt{2\pi n}\,(n/e)^n n^{-x_0}}{\sqrt{2\pi(n - x_0)}\,((n - x_0)/e)^{n-x_0}} = \sqrt{\frac{n}{n - x_0}} \frac{e^{-x_0} n^n}{(n - x_0)^n} \frac{(n - x_0)^{x_0}}{n^{x_0}} \longrightarrow 1$$

as $n \to \infty$ with x_0 fixed, since $(1 - x_0/n)^n \to e^{-x_0}$. Hence, we have

$$\mathbb{P}(X^t_{1:3} = x_{1:3} \mid S^0 = s^0) \approx \prod_{i=1}^{3} \text{Poisson}(x_i \mid nt\lambda_i)$$

when n is large, $x_0 \ll n$, and $t\lambda_0 = O(1/n)$.

For each of the 32 distinct possible starting states a, the same approximation applies to the set of positions starting in state a. Modeling these 32 sets of positions independently, we have

$$\mathbb{P}(X^t_{1:96} = x_{1:96} \mid S^0 = s^0) \approx \prod_{i=1}^{96} \text{Poisson}(x_i \mid n_i t\lambda_i)$$

where $n_i = \#\{\ell : s^0_\ell = a_i\}$. In other words, the counts of the 96 substitution types are approximately distributed as independent Poisson random variables with rates $n_i t \lambda_i$.

The preceding derivation ignores the fact that a substitution at one position changes the context of the two adjacent positions. However, since it is rare for single-base substitutions to occur at two adjacent positions, the effect of ignoring this should be negligible.

S5.4 Multiple mutational processes

Suppose x_{ij} is the number of mutations of substitution type i for subject j, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where $I = 96$. The derivation above justifies modeling these mutation counts as

$$X_{ij} \sim \text{Poisson}(n_{ij} \lambda_{ij} t_j)$$

independently, where t_j is the age or exposure time of subject j, λ_{ij} is the mutation rate for substitution type i in subject j, and n_{ij} is the number of positions that are in state a_i in the normal genome of subject j, out of all positions that were measured. The positions measured may be a subset of the genome due to whole-exome/targeted sequencing or low sequencing depth, for example.

From birth, each subject is exposed to many mutational processes, such as environmental exposures, replication errors, defective DNA repair mechanisms, and so on. Each mutational process causes each substitution type to occur at a given rate, and the profile of rates across the 96 substitution types can be expected to vary depending on the mutational process. Since rates are additive in a continuous-time Markov process, it is natural to model the subject-specific mutation rates λ_{ij} as linear combinations of these mutational process rate profiles, with non-negative weights depending on the exposure of the subject to each process. Further, assuming the opportunity counts n_{ij} are constant (or nearly constant) across all subjects j, one can absorb n_{ij} into λ_{ij}, which changes the interpretation of λ_{ij} by reparametrizing it. This leads to using a representation of the form

$$n_{ij} \lambda_{ij} t_j = \sum_{k=1}^{K} r_{ik} \theta_{kj}$$

where the weight $\theta_{kj} \geq 0$ is the exposure of subject j to process k, and $(r_{1k}, \ldots, r_{Ik})$ is the mutation rate profile for mutational process k, which is referred to as its mutational signature. Thus, we arrive at the Poisson non-negative matrix factorization model in Equation (1),

$$X_{ij} \sim \text{Poisson}\Big(\sum_{k=1}^{K} r_{ik} \theta_{kj}\Big).$$

A statistical issue with this representation is that there is a non-identifiability between the r_{ik}'s and θ_{kj}'s, since arbitrary multiplicative constants c_k can be moved between them. We deal with this by normalizing the mutational signatures to sum to 1, that is, by enforcing the constraint $\sum_{i=1}^{I} r_{ik} = 1$ for all k.

S6 Extended simulation results

We now present additional results for the simulations in Section 5. Specifically, Section S6.1 describes the methods that we compare with; Section S6.2 details the MCMC sampler for PoissonCUSP; Section S6.3 presents further assessment of the reconstruction errors for signatures and loadings as well as details on computation time and effective sample size of the MCMC algorithms; and Section S6.4 presents an additional simulation under sparse data representing indel counts.

S6.1 Description of competing methods

We now describe the NMF mutational signatures methods that we use for comparisons in our simulation, most of which appear in the recent review by Islam et al. (2022). Currently, the most prominent method is SigProfiler (Alexandrov et al., 2020) and its successor, SigProfilerExtractor (Islam et al., 2022). SigProfilerExtractor uses a bootstrap-like procedure to resample the mutation count matrix from a Poisson model, and fits an NMF model to each resampled data matrix by minimizing the Poisson Kullback–Leibler divergence. The number of signatures is selected by applying the algorithm for a range of K values and using a neural network to choose K.

A leading Bayesian method is signeR (Rosales et al., 2016; Drummond et al., 2023) which is based on the Poisson NMF model in Equation (1), but employs independent hierarchical gamma priors over both the signatures and the loadings. Selection of K is performed using the Bayesian information criteria (BIC) after running a separate model for each K. Hence, both SigProfilerExtractor and signeR are particularly slow when the range of possible K values is moderate to large.

A faster alternative is offered by SignatureAnalyzer (Kim et al., 2016), which fits the Poisson NMF model in Equation (1) using a *maximum a posteriori* estimation algorithm (Tan and Févotte, 2013). Similarly to our approach, SignatureAnalyzer uses ARD with inverse-gamma hyperpriors on the relevance weights to determine the number of signatures. While the method is efficient and flexible in terms of the choice of the objective function (Kullback–Leibler or squared error) and prior (exponential or half-normal), it does not provide any uncertainty quantification.

A natural Bayesian approach to selecting the number of latent factors is to use spike-and-slab priors, where the relevance weight of each of the K signatures has some probability of being sampled from a spike close to zero. This is the approach taken in the elegant nonparametric factorization model proposed by Legramanti et al. (2020), which infers the number of factors using a cumulative shrinkage spike-and-slab process prior (CUSP). This model assumes an infinite number of factors *a priori*, but the probability that μ_k comes from the spike—effectively removing that factor from the model—increases with k; Posterior inference is performed via an adaptive Metropolis algorithm; we defer to Section S6.2 for details. Finally, a fully Bayesian approach to ARD using a Gaussian likelihood, referred to as BayesNMF, is presented by Brouwer et al. (2017).

S6.2 Details of PoissonCUSP

The CUSP model described by Legramanti et al. (2020) is a spike-and-slab shrinkage prior that enables automatic selection of the number of latent factors in Gaussian factorization models. We adapt it to the Poisson factorization model as follows. We begin by specifying the following prior structure for the signatures and the loadings:

$$(r_{1k}, \dots, r_{Ik}) \sim \text{Dirichlet}(\alpha, \dots, \alpha), \quad \theta_{kj} = \vartheta_{kj}\mu_k, \quad \vartheta_{kj} \sim \text{Gamma}(a, a).$$

Then, we let

$$\mu_k \sim (1-\pi_k)\text{Gamma}(a_0, b_0) + \pi_k \delta_{\mu_\infty},$$

where δ_x denotes the point mass at x and π_k is the prior probability of sampling the spike $\mu_\infty = 0.01$, which is modeled as

$$\pi_k = \sum_{\ell=1}^{k} \phi_\ell, \quad \phi_\ell = v_\ell \prod_{m=1}^{\ell-1}(1-v_m), \quad v_\ell \sim \text{Beta}(1, \alpha).$$

Hence, π_k increases as k increases. As a spike-and-slab prior, this enables automatic selection of the number of signatures. Each iteration of the Gibbs sampler for the basic PoissonCUSP model consists of the following steps:

1. For $i = 1, \dots, I$ and $j = 1, \dots, J$, sample auxiliary variables $Y_{ij} = (Y_{ij1}, \dots, Y_{ijK})$ according to $Y_{ij} \sim \text{Multinomial}\big(X_{ij}, (q_{ij1}, \dots, q_{ijK})\big)$, where $q_{ijk} = r_{ik}\theta_{kj} / \sum_{\kappa=1}^{K} r_{i\kappa}\theta_{\kappa,j}$.

2. Sample the individual loadings ϑ_{kj} from $\vartheta_{kj} \sim \text{Gamma}\big(a + \sum_{i=1}^{I} Y_{ijk},\, a + \mu_k\big)$.

3. Sample the signatures $r_k = (r_{1k}, \dots, r_{ik})$ from $r_k \sim \text{Dirichlet}\big(\alpha + \sum_{j=1}^{J} Y_{1jk}, \dots, \alpha + \sum_{j=1}^{J} Y_{ijk}\big)$.

4. For each $k = 1, \dots, K$, sample the categorical auxiliary variables Z_k as follows

$$\mathbb{P}(Z_k = \ell \mid -) = \begin{cases} \phi_\ell \, \mu_\infty^{\sum_{i,j} Y_{ijk}} \exp(-\mu_\infty \sum_j \vartheta_{kj}) & \text{if } 1 \leq \ell \leq k \\ \phi_\ell \, \dfrac{b_0^{a_0}}{\Gamma(a_0)} \dfrac{\Gamma(a_0 + \sum_{ij} Y_{ijk})}{(b_0 + \sum_j \vartheta_{kj})^{a_0 + \sum_{ij} Y_{ijk}}} & \text{if } k < \ell \leq K \end{cases}$$

where $\phi_\ell = \mathbb{P}(Z_k = \ell)$ is the prior probability of the auxiliary variables. These probabilities are derived by integrating out the parameters μ_k, relying on conjugacy.

5. For $\ell = 1, \dots, K-1$, sample the sticks from

$$v_\ell \sim \text{Beta}\bigg(1 + \sum_{k=1}^{K} \mathbb{1}(Z_k = \ell),\ \alpha_\pi + \sum_{k=1}^{K} \mathbb{1}(Z_k > \ell)\bigg),$$

and fix $v_K = 1$.

6. Calculate $\phi_1, \dots, \phi_K$ via the stick-breaking construction, namely $\phi_\ell = v_\ell \prod_{m=1}^{\ell-1}(1 - v_m)$.

7. For $k = 1, \dots, K$, if $Z_k \leq k$ then set $\mu_k = \delta_{\mu_\infty}$, otherwise sample

$$\mu_k \sim \text{Gamma}\bigg(a_0 + \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ijk},\ b_0 + \sum_{j=1}^{J} \vartheta_{kj}\bigg).$$

Notice that here we assume a multiplicative structure for θ_{kj}, while our compressive hyperprior approach assumes a hierarchical one. In principle, one could specify $\theta_{kj} \sim \text{Gamma}(a, a/\mu_k)$ and $\mu_k \sim (1-\pi_k)\text{InvGamma}(a_0, b_0) + \pi_k \delta_{\mu_\infty}$ to mimic our model. However, in practice, we found that this approach did not yield the desired shrinkage effect, since the Gibbs sampler preferred to sample from the prior slab rather than allocating the signatures to the spike. This happened both when $a_0 = 1$ and $b_0 = 1$ were fixed and when they were chosen according to our compressive model as $a_0 = aJ + 1$ and $b_0 = a\mu_\infty J$ with $\mu_\infty = 0.01$. This is likely due to a mixing issue and to the strong multimodal nature of the resulting posterior. Instead, we found that the sampler described above worked better.

Inference for the number of signatures K^* in PoissonCUSP can be performed using the same adaptive Metropolis sampler as in Algorithm 2 of Legramanti et al. (2020). This automatically tunes the number of columns in the factorization in a random manner, by eliminating the columns that in a given iteration fall within the spike (that is, the signatures for which $Z_k \leq k$) and potentially adding novel ones. Refer to Legramanti et al. (2020) for a description. We implement their algorithm with their choice of tuning hyperparameters. In particular, in our simulations we use a spike location of $\mu_\infty = 0.01$, we use slab parameters $a_0 = b_0 = 1$, and we start at $K = 20$.

S6.3 Additional simulation results

S6.3.1 Reconstruction errors

We compare the models by assessing the root mean squared error (RMSE) for (i) the observed counts X, (ii) the true mean matrix $\Lambda^0 = (\lambda^0_{ij})$, (iii) the true matrix of signatures R^0, and (iv) the true loadings matrix Θ^0.

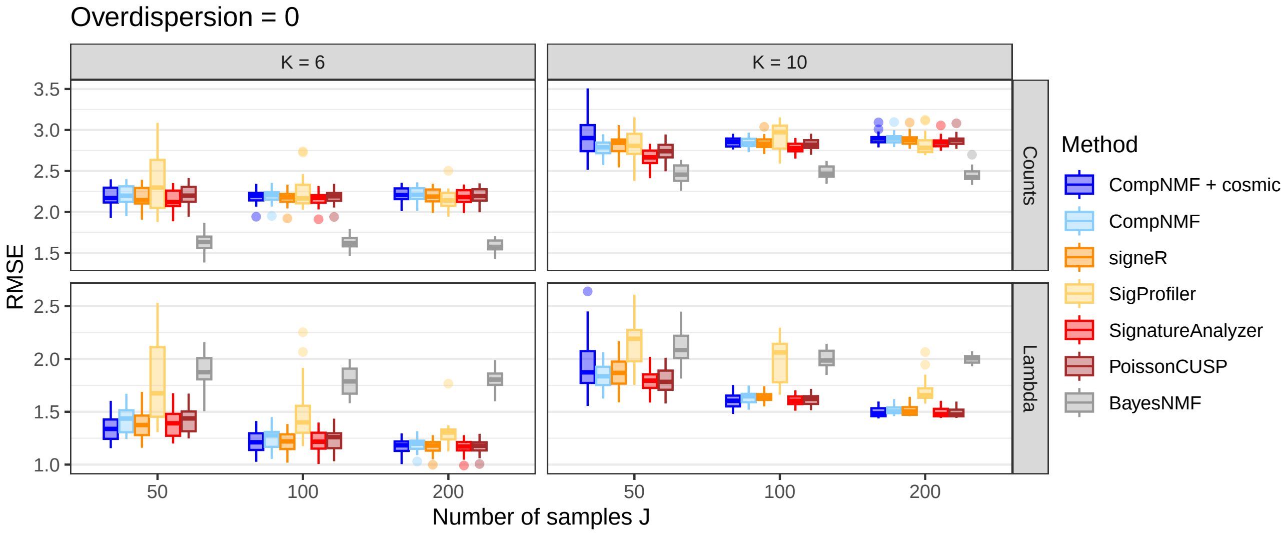


Figure S6.1: RMSE between (top) $\hat{\Lambda}$ and the count matrix X, and (bottom) $\hat{\Lambda}$ and the true mean matrix Λ^0, for each method across 20 replicates, when data are generated with overdispersion $\tau = 0$, i.e., the Poisson model.

First, the RMSE for X and Λ^0 is calculated as $\textsc{rmse}(X, \hat{\Lambda}) = \big(\sum_{ij}(X_{ij} - \hat{\lambda}_{ij})^2/IJ\big)^{1/2}$ and $\textsc{rmse}(\Lambda^0, \hat{\Lambda}) = \big(\sum_{ij}(\lambda^0_{ij} - \hat{\lambda}_{ij})^2/IJ\big)^{1/2}$, respectively, where $\hat{\lambda}_{ij} = \sum_{k=1}^{K^*} \hat{r}_{ik}\hat{\theta}_{kj}$ and K^* is the estimated number of signatures for each method. Figures S6.1 and S6.2 display the comparison across all models and replicate data sets for overdispersion $\tau = 0$ and $\tau = 0.15$, respectively. In the correctly specified case ($\tau = 0$), no method performs overwhelmingly better than all the others. This is expected since all models correctly estimate the true number of signatures in this case; see Figure 2(A). One exception is SigProfiler, which performs noticeably worse than the others, particularly in terms of the RMSE for Λ^0. The performance of SigProfiler does improve somewhat as the sample size increases, suggesting that the algorithm might struggle in small dimensions or on relatively low mutation counts.

The situation changes, however, in the overdispersed (negative binomial) case with $\tau = 0.15$, displayed in Figure S6.2. Here, SignatureAnalyzer and PoissonCUSP obtain the lowest $\textsc{rmse}(X, \hat{\Lambda})$ across all values of J, however, their corresponding $\textsc{rmse}(\Lambda^0, \hat{\Lambda})$ shows the opposite trend. This is a clear indication of overfitting, which is reinforced by the fact that both models overestimate the number of signatures; see Figure 2(A).

To enable a direct comparison of performance in terms of estimating the true signature and loadings matrices R^0 and Θ^0, we first perform a matching step to maximize the total pairwise cosine similarity between the estimated $\hat{R}$ and true R^0 using the Hungarian algorithm. If the true number of signatures and the estimated number differ, we also pad the smaller matrix with zeros for the non-matched signatures, which penalizes incorrect estimation of K. The rows of $\hat{\Theta}$ and Θ^0 are also permuted accordingly, and are also padded with zeros to make the dimensions the same. We then calculate $\textsc{rmse}(R^0, \hat{R}) = \big(\sum_{ik}(r^0_{ik} - \hat{r}_{ik})^2/IK\big)^{1/2}$ and $\textsc{rmse}(\Theta^0, \hat{\Theta}) = \big(\sum_{kj}(\theta^0_{kj} - \hat{\theta}_{kj})^2/KJ\big)^{1/2}$, where $\hat{r}_{ik}$ and

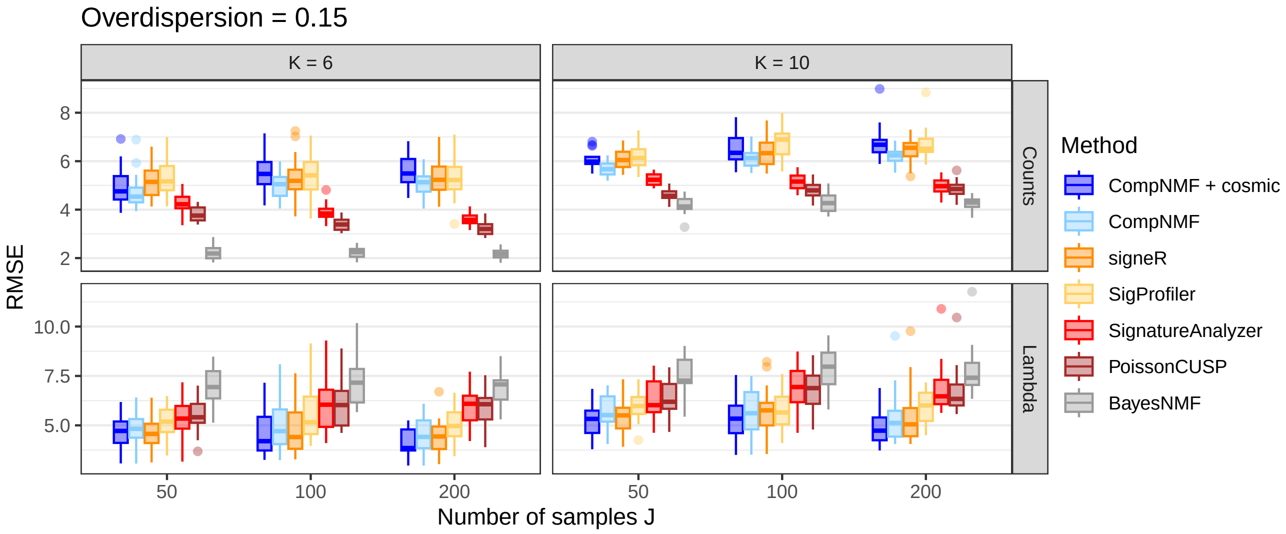


Figure S6.2: RMSE between (top) $\hat{\Lambda}$ and the count matrix X, and (bottom) $\hat{\Lambda}$ and the true mean matrix Λ^0, for each method across 20 replicates, when data are generated with overdispersion be $\tau = 0.15$, i.e., negative binomial.

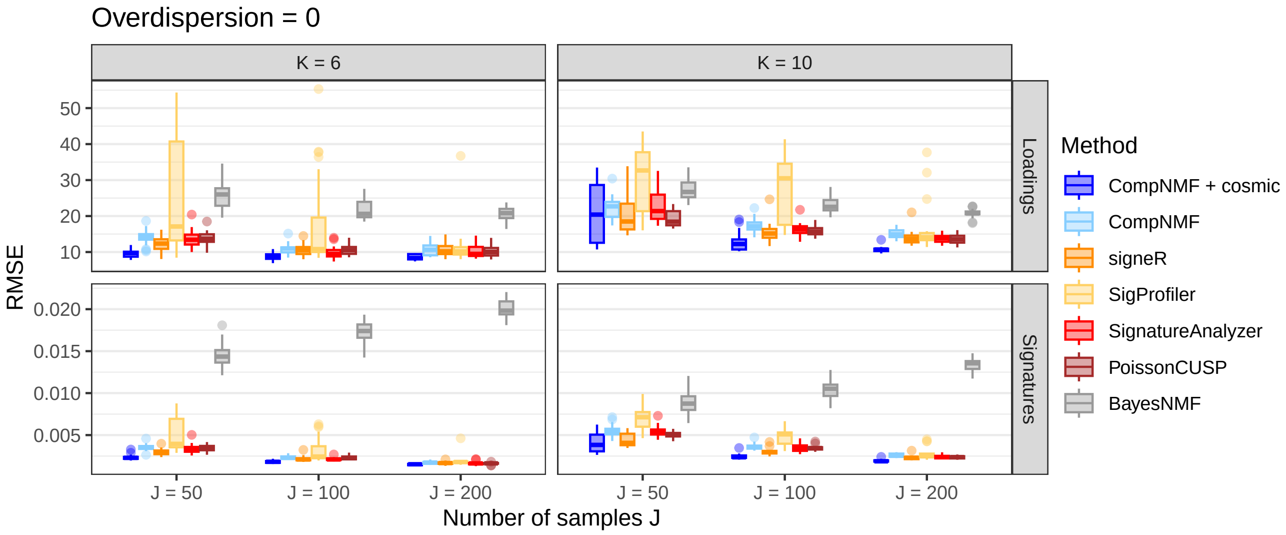


Figure S6.3: RMSE between (top) the estimated loadings $\hat{\Theta}$ and true loadings Θ^0, and (bottom) the estimated signatures $\hat{R}$ and true signatures R^0, over 20 replicate datasets, when data are generated with overdispersion $\tau = 0$, i.e., the Poisson model.

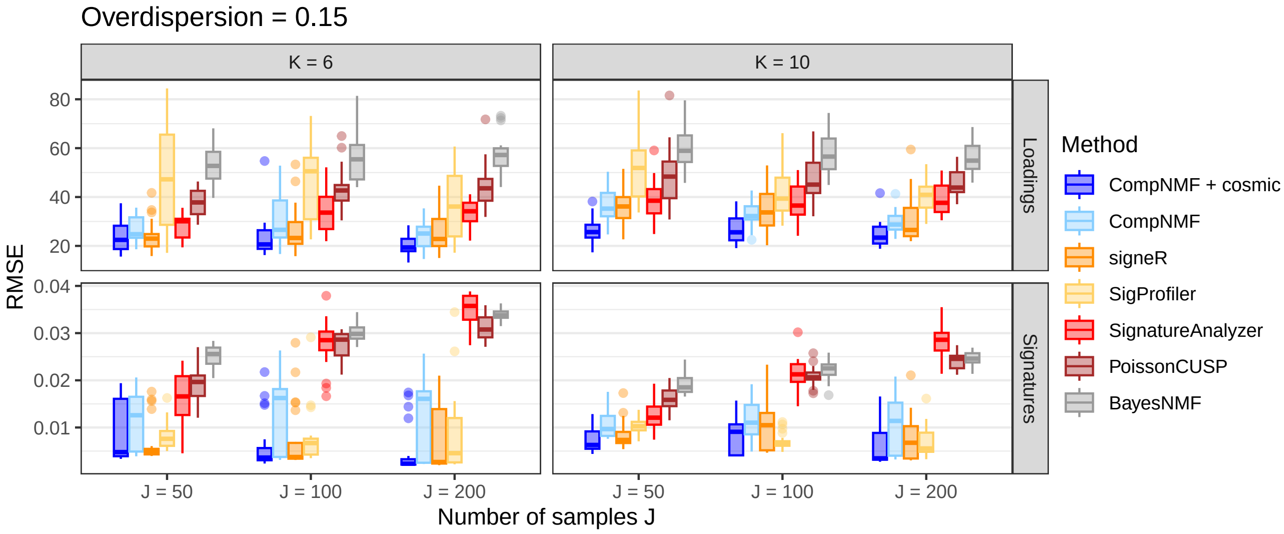


Figure S6.4: RMSE between (top) the estimated loadings $\hat{\Theta}$ and true loadings Θ^0, and (bottom) the estimated signatures $\hat{R}$ and true signatures R^0, over 20 replicate datasets, when data are generated with overdispersion $\tau = 0.15$, i.e., negative binomial.

$\hat{\theta}_{kj}$ are the estimated signatures and loadings, obtained by averaging posterior samples in the case of the Bayesian models. The results for $\tau = 0$ and $\tau = 0.15$ are shown in Figures S6.3 and S6.4, respectively. The best performance is attained by CompNMF+cosmic. While this model uses information from the true COSMIC signatures in the prior, it correctly filters out the signatures that are not needed. Moreover, the improved estimation of the COSMIC signatures improves the estimation of the associated loadings. The performance of all the other models, with the exception of SigProfiler, is virtually identical when $\tau = 0$. Finally, when $\tau = 0.15$, PoissonCUSP, SignatureAnalyzer and BayesNMF perform poorly due to overfitting, as before.

S6.3.2 Computation time and effective sample size

Figure S6.5 shows the total computation time required by each method, as a function of the number of samples J. All computations were performed on an AMD Ryzen 3900-based dedicated server with 128GB of memory, running Ubuntu 20.04, R version 4.3.1 linked to Intel MKL 2019.5-075. Calculations were split across 20 cores via the `foreach` package, allocating one dataset per core for each combination of τ, J, and K^0, and running each method sequentially. Not surprisingly, SignatureAnalyzer is the fastest method by an order of magnitude, always taking under one minute to complete. BayesNMF is the second fastest method since inference under the Gaussian NMF model does not require the inclusion of additional latent variables at each iteration of the Gibbs sampler. PoissonCUSP is the third fastest method, because the number of signatures in the model varies adaptively within the Gibbs sampler, preventing unnecessary computations on inactive factors. CompNMF is slightly slower than PoissonCUSP since no adaptation in performed, but it is faster than signeR. SigProfiler

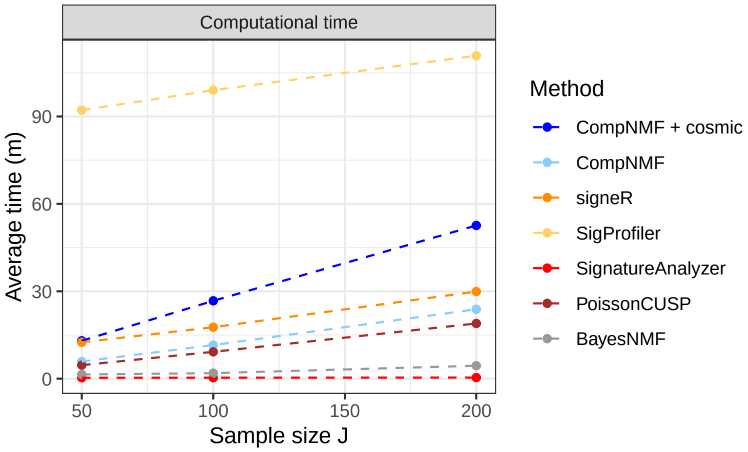


Figure S6.5: Average computation time (in minutes) for each method. Times shown are the average across all combinations of overdispersion τ and number of signatures K^0 used in the simulation study.

was the slowest among all methods by a wide margin. Thus, in terms of computation time, both CompNMF and CompNMF+cosmic are competitive with the next best-performing existing method, signeR. Furthermore, signeR is written in `C++`, whereas CompNMF and CompNMF+cosmic are written in `R` code, so they could be significantly faster if implemented in `C++` or similar. Additionally, they could be sped up by updating only the non-compressed factors when running the Gibbs sampler, following the adaptive approach of PoissonCUSP.

It is also interesting to compare the performance across models in terms of effective sample size (ESS). These are displayed in Figure S6.6. ESS values were calculated using the last 1000 posterior samples for each model to ensure a fair comparison. Each boxplots reports the average univariate ESS for R and Θ across 20 replicates. The ESS for the relevance weights μ_k (or a comparable quantity) are also reported wherever present. We observe a limited difference in the quality of the samples for the models using a Poisson likelihood. This is because the underlying sampler, which is based on multinomial data augmentation, is very similar for all methods. Interestingly, PoissonCUSP performs considerably worse when looking at the relevance weights. BayesNMF, instead, relies on a Gaussian likelihood and has the advantage of not requiring further data augmentation; this leads to considerably better ESS for the signature matrix, though no normalization constraint is imposed in the model. Overall, our CompNMF models exhibit comparable performance in terms of ESS relative to the other Poisson NMF models.

S6.4 Simulation with sparse indel mutations

In this section, we consider a simulation setting in which the mutation counts are considerably sparser. In practice, sparsity occurs naturally in count data for insertion-deletion mutations (indels), which consist of the addition or removal of one or more nucleotides at a given position in the genome.

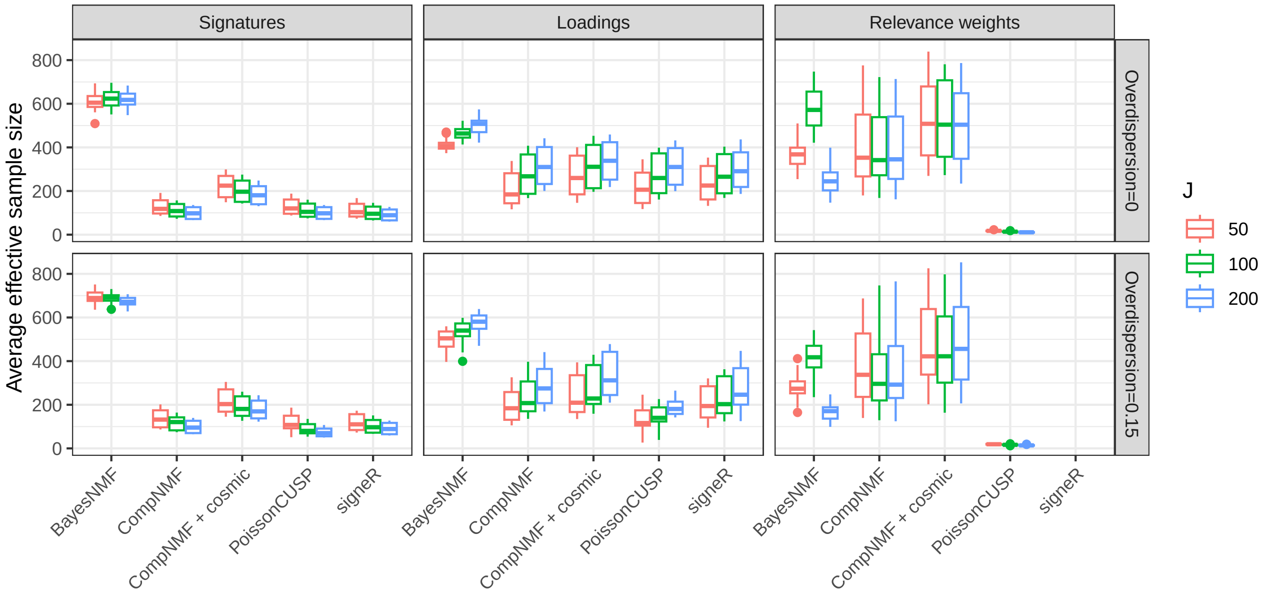


Figure S6.6: Average effective sample sizes for MCMC algorithms drawing from the posterior of the signatures and the loadings for the Bayesian methods in the simulation.

Alexandrov et al. (2013) categorize indels into 83 mutational channels, and the resulting indel signatures tend to be lower entropy than SBS signatures. In COSMIC v3.4, there are 23 recognized indel mutational signatures with a putative etiology, which we can use to define informative priors in our framework.

Similarly to the SBS simulation study in Section 5, we simulate indel counts as $X_{ij} \sim \text{NegBin}\big(1/\tau,\, 1/(1+\tau\lambda_{ij}^0)\big)$, where $\lambda_{ij}^0 = \sum_{k=1}^{K_{\text{pre}}^0} \rho_{ik}^0 \omega_{kj}^0 + \sum_{k=1}^{K_{\text{new}}^0} r_{ik}^0 \theta_{kj}^0$. We again set $K_{\text{pre}}^0 = 4$, and for $k = 1, \ldots, 4$, we define $\rho_k^0 = (\rho_{1k}^0, \ldots, \rho_{Ik}^0)$ to be COSMIC indel signatures ID1, ID2, ID8, and ID9, respectively. ID1 and ID2 are both related to replication slippage and defective DNA mismatch repair mechanisms. ID8 is a clock-like signature and ID9 does not have an assigned etiology but has been shown to appear in many cancer types. We also randomly generate $r_k^0 = (r_{1k}^0, \ldots, r_{Ik}^0)$ as $r_k^0 \sim \text{Dirichlet}(0.05, \ldots, 0.05)$, independently for $k = 1, \ldots, K_{\text{new}}^0$, and choose $K_{\text{new}}^0 = 2$ for simplicity, yielding a total of $K = 6$ true signatures. We generate loadings by setting $\omega_{kj}^0 = w_k \xi_{kj}$ where $w_k \sim \text{Gamma}(50, 1)$ and $\xi_{kj} \sim \text{Gamma}(0.5, 0.5)$ independently, and θ_{kj}^0 in the same way as ω_{kj}^0. These simulation settings were chosen to produce data similar to observed indel count data.

We generate 20 simulated datasets with $J = 100$ for each $\tau \in \{0, 0.15\}$, representing the correctly specified ($\tau = 0$) and misspecified cases ($\tau = 0.15$). The resulting mutation count matrices are sparse, with approximately 65% of entries in the X matrix being equal to zero; for comparison, the SBS simulation in Section 5 produces around 30% zeros. This sparsity is a consequence of the sparser signatures and the smaller loadings (w_k has a mean of 50 instead of 100). We run the following models:

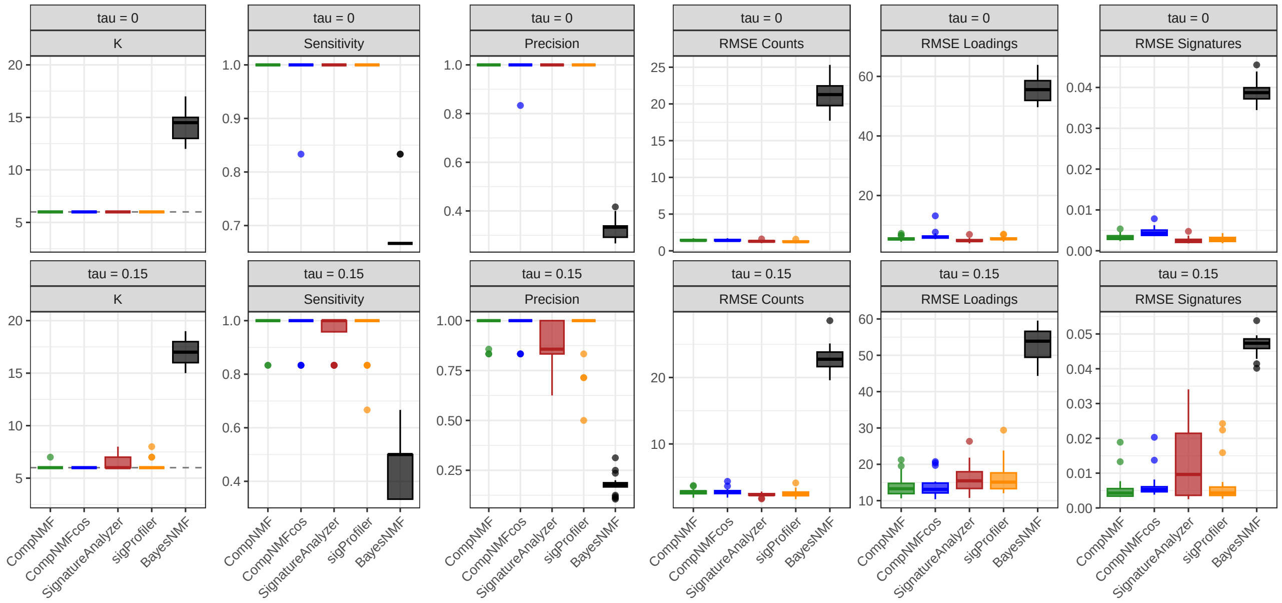


Figure S6.7: Results from the sparse indel simulation. Boxplots summarize the results over 20 replicated datasets in each case. Top row: results from the correctly specified case ($\tau = 0$). Bottom row: results from the misspecified case ($\tau = 0.15$). From left to right, the panels display: (1) number of estimated signatures by each model, (2) sensitivity and (3) precision with respect to the estimated signature profile $\hat{R}$, (4) RMSE between the mutation matrix and the product $\hat{R}\hat{\Theta}$, (5) RMSE between R^0 and $\hat{R}$, and (6) RMSE between Θ^0 and $\hat{\Theta}$.

1. CompNMF with default values $K = 20$, $\varepsilon = 0.001$, $\alpha = 0.5$ and $a = 1$.

2. CompNMF+cosmic with $K_{\text{new}} = 10$ *de novo* signatures and $K_{\text{pre}} = 23$ COSMIC v3.4 ID signatures, and with $\varepsilon = 0.001$.

3. SignatureAnalyzer with $K = 25$ and selection method set to `L1KL` .

4. SigProfiler with K ranging from 2 to 20, with 10 NMF replicates and random initialization.

5. BayesNMF with $K = 20$ and default hyperparameters.

Note that signeR is not included since the signeR software is only designed to handle SBS mutations. We also exclude PoissonCUSP from these comparisons due to its lower performance in our other experiments.

Figure S6.7 displays the results. When the model is correct ($\tau = 0$), all of the models perform extremely well, with the exception of BayesNMF. The lower performance of BayesNMF is presumably due to its usage of a Gaussian likelihood, which is not well suited for sparse count data. In the misspecified setting ($\tau = 0.15$), CompNMF, CompNMF+cosmic, and SigProfiler perform nearly as well as they do when the model is correct. Meanwhile, SignatureAnalyzer has somewhat reduced performance, and the performance of BayesNMF is degraded even further. These results indicate that the good performance of our compressive NMF models extends to sparse settings as well.

S7 Sensitivity analyses

In this section, we conduct a thorough analysis of the sensitivity of our model with respect to the choice of its settings. Section S7.1 presents a comparison between our proposed compressive hyperprior and a fixed-strength hyperprior. Section S7.2 presents sensitivity analyses with respect to the choice of K, ε, a, and α in the compressive case.

S7.1 Fixed-strength versus compressive hyperprior

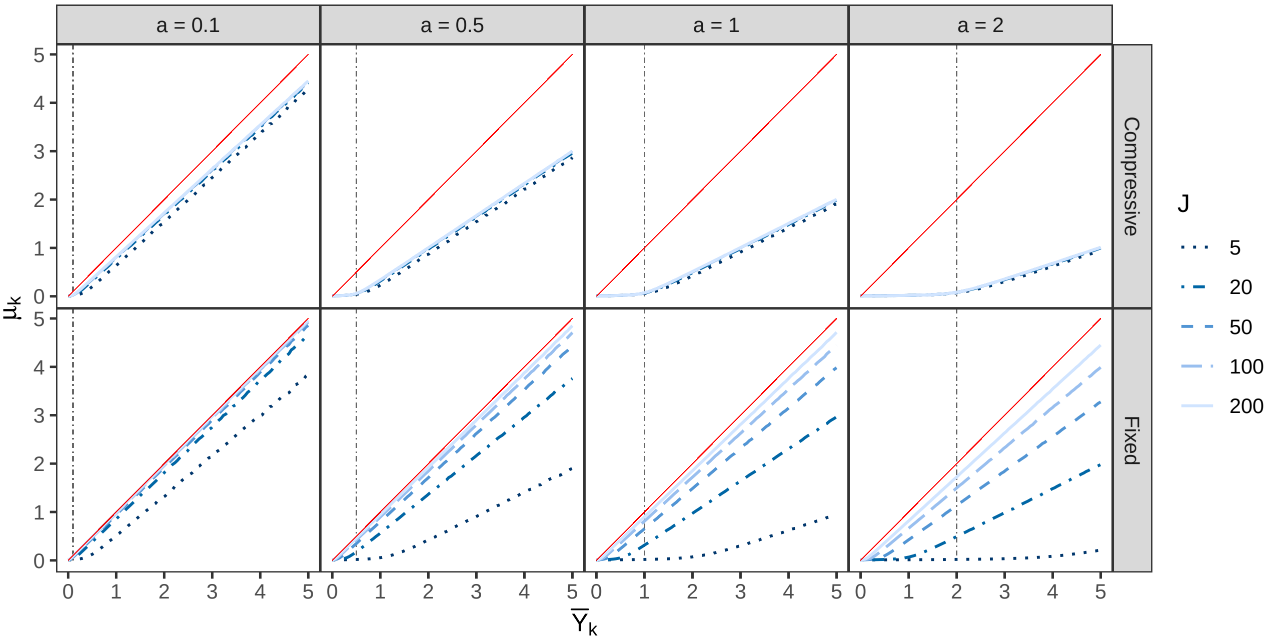


Figure S7.1: Comparison between the posterior mean of $\mu_k \mid Y$ under our compressive hyperprior $\mu_k^{\mathrm{C}} \sim \mathrm{InvGamma}(aJ+1, \varepsilon aJ)$ and the fixed-strength hyperprior $\mu_k^{\mathrm{FS}} \sim \mathrm{InvGamma}(aN+1, \varepsilon aN)$ with $N = 10$ and $\varepsilon = 0.001$, for varying values of a and sample size J. The red line indicates where $\bar{Y}_k$ is equal to μ_k, while the vertical dashed line shows the value of a.

The form of the hyperprior on the relevance weights μ_k is essential for automatic relevance determination (ARD) in NMF to work properly. Recall that our approach involves letting μ_k have a small prior mean, like 0.01 or 0.001, which makes unneeded factors shrink to small values and effectively removes them from the decomposition (Tan and Févotte, 2013; Brouwer et al., 2017). However, the strength of this hyperprior on μ_k can significantly impact the results. Here, we empirically compare our strength-matching compressive approach, in which $\mu_k \sim \mathrm{InvGamma}(aJ+1, \varepsilon aJ)$, and the fixed-strength hyperprior $\mu_k \sim \mathrm{InvGamma}(a_0, b_0)$ where a_0 and b_0 are fixed. To make a direct comparison, we set $a_0 = aN + 1$ and $b_0 = \varepsilon aN$, where $N = 10$ is fixed. This ensures that $\mathbb{E}(\mu_k) = \varepsilon$ in both cases.

By the proof of Theorem 1,

$$(\mu_k \mid Y) \sim \mathrm{InvKummer}\big(\mu_k \bigm| a_0 + Ja,\, b_0,\, Ja + J\bar{Y}_k,\, a\big) \tag{S28}$$

in the fixed-strength case. We use superscripts to distinguish the relevance weights in the compressive and fixed-strength cases, denoting them μ_k^{C} and μ_k^{FS}, respectively.

Effect on relevance weight posteriors. We first compare the posteriors of μ_k^{C} and μ_k^{FS} given the latent counts Y. Figure S7.1 shows the posterior means $\mathbb{E}(\mu_k^{\mathrm{C}} \mid Y)$ and $\mathbb{E}(\mu_k^{\mathrm{FS}} \mid Y)$ for varying values of $\bar{Y}_k$, a, and sample size J. Suppose $\bar{Y}_k \to y$ as $J \to \infty$. From Theorem 3 and Theorem S3.2, we can fully characterize the concentration point of both posteriors as $J \to \infty$, specifically, $\mu_k^{\mathrm{C}} \mid Y$ concentrates at $\mu_* = 2a\varepsilon/\big(\sqrt{(y-a+\varepsilon)^2 + 8a\varepsilon} - (y-a+\varepsilon)\big)$ and $\mu_k^{\mathrm{FS}} \mid Y$ concentrates at y. This is evident from Figure S7.1: In the compressive case, the posterior mean of μ_k^{C} is insensitive to the value of J, being very close to μ_* even when J is very small. Furthermore, a controls the location of the "elbow" of the curves in Figure S7.1, which leads to the sparsity inducing effect described in Section 3.3. Meanwhile, in the fixed-strength case, the posterior mean of μ_k^{FS} approaches the 45° line where $\mu_k = \bar{Y}_k$. Consequently, the elbow goes away as $J \to \infty$, and a does not play the role of a threshold in the fixed-strength case. Also see Section S7.2 for more on the effect of a.

Effect on signature recovery performance. To see the effect on performance for recovering the signatures, we compare the compressive and fixed-strength hyperpriors in the simulation setup of Section 5. We set $K_{\mathrm{pre}}^0 = 4$ and $K_{\mathrm{new}}^0 = 6$, so that there are 10 true signatures in the data. Counts are generated in (i) the correctly specified case (Poisson, $\tau = 0$) and (ii) the misspecified case (negative binomial with overdispersion $\tau = 0.15$). We generate 20 datasets for each combination of sample size $J \in \{20, 50, 100, 200, 300, 400, 500\}$ and overdispersion $\tau \in \{0, 0.15\}$, and run the models using $K = 20$, $\varepsilon = 0.01$, $a = 1$, and $\alpha = 0.5$ for 4,000 MCMC iterations, using the samples from the last 1,000 iterations for inference. Here, for simplicity, we do not use the informative COSMIC priors.

Figure S7.2 shows the estimated number of signatures K^*, precision, and sensitivity. In the correctly specified case ($\tau = 0$), the compressive and fixed-strength models behave very similarly – both perform nearly perfectly except for a slight drop in performance when J is small. However, in the misspecified case ($\tau = 0.15$), there is a stark difference in performance. We see that the fixed-strength model dramatically overestimates the number of active signatures, exhibits very low precision, and has decreased sensitivity. The compressive model is much better able to mitigate the effects of misspecification, incurring only a small loss in performance relative to the correctly specified case.

Estimated signatures. To further examine these differences, Figure S7.3 shows the posterior mean for each of the $K = 20$ signatures and their associated relevance weights, for one dataset with $J = 300$ and $\tau = 0.15$. Inferred signatures have been matched and arranged using the Hungarian algorithm to aid visualization so that signatures A, B, C, … in the compressive case roughly

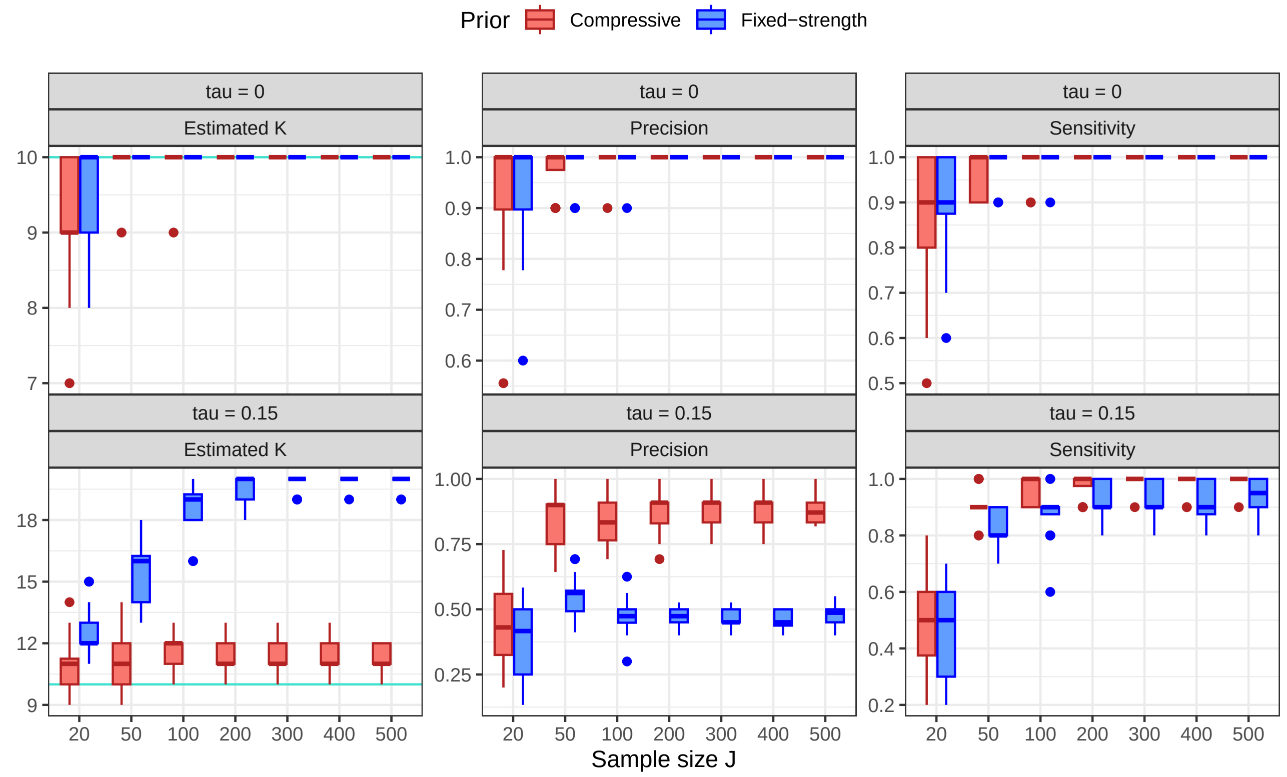


Figure S7.2: Simulation results: Boxplots of the number of active signatures K^* in the posterior (left column), precision (center column) and sensitivity (right column) across the 20 simulated datasets under compressive and fixed-strength hyperpriors, for $\tau = 0$ (top row) and $\tau = 0.15$ (bottom row). Signatures are considered active if the average posterior relevance weight is such that $\hat{\mu}_k > 5\varepsilon$. The horizontal turquoise line indicates the true number of signatures in the data.

match signatures A, B, C, ... in the fixed-strength counterpart, in terms of cosine similarity. We see that signatures A through K look very similar in the two cases. On the other hand, signatures L through T are very different. In the compressive model, signatures L–T are approximately uniform because they correspond to signatures for which $\bar{Y}_k \approx 0$ and $\mu_k \approx \varepsilon$; therefore, the full conditional is $(r_k \mid Y) \sim \text{Dirichlet}(\alpha + \sum_{j=1}^{J} Y_{1jk}, \ldots, \alpha + \sum_{j=1}^{J} Y_{Ijk}) \approx \text{Dirichlet}(\alpha, \ldots, \alpha)$, for which the mean is $(1/I, \ldots, 1/I)$. Meanwhile, in the fixed-strength model, signatures L–T are low-entropy signatures and the corresponding values of μ_k are far from ε. Note that the traceplots for μ_k are fairly stable, indicating that the MCMC samplers appear to be performing well.

Effect on estimation performance. Figure S7.4 shows the root mean squared error (RMSE) between the estimated and true values of the signatures, loadings, and count data matrix across the 20 datasets. We observe a similar trend as before, with the compressive model outperforming the fixed-strength model in terms of RMSE on the loadings and signatures, especially for larger values of J. While the RMSE of the counts is higher for compressive than fixed-strength, this appears to be due to over-fitting by the fixed-strength model, as evidenced by (i) its numerous spurious signatures in

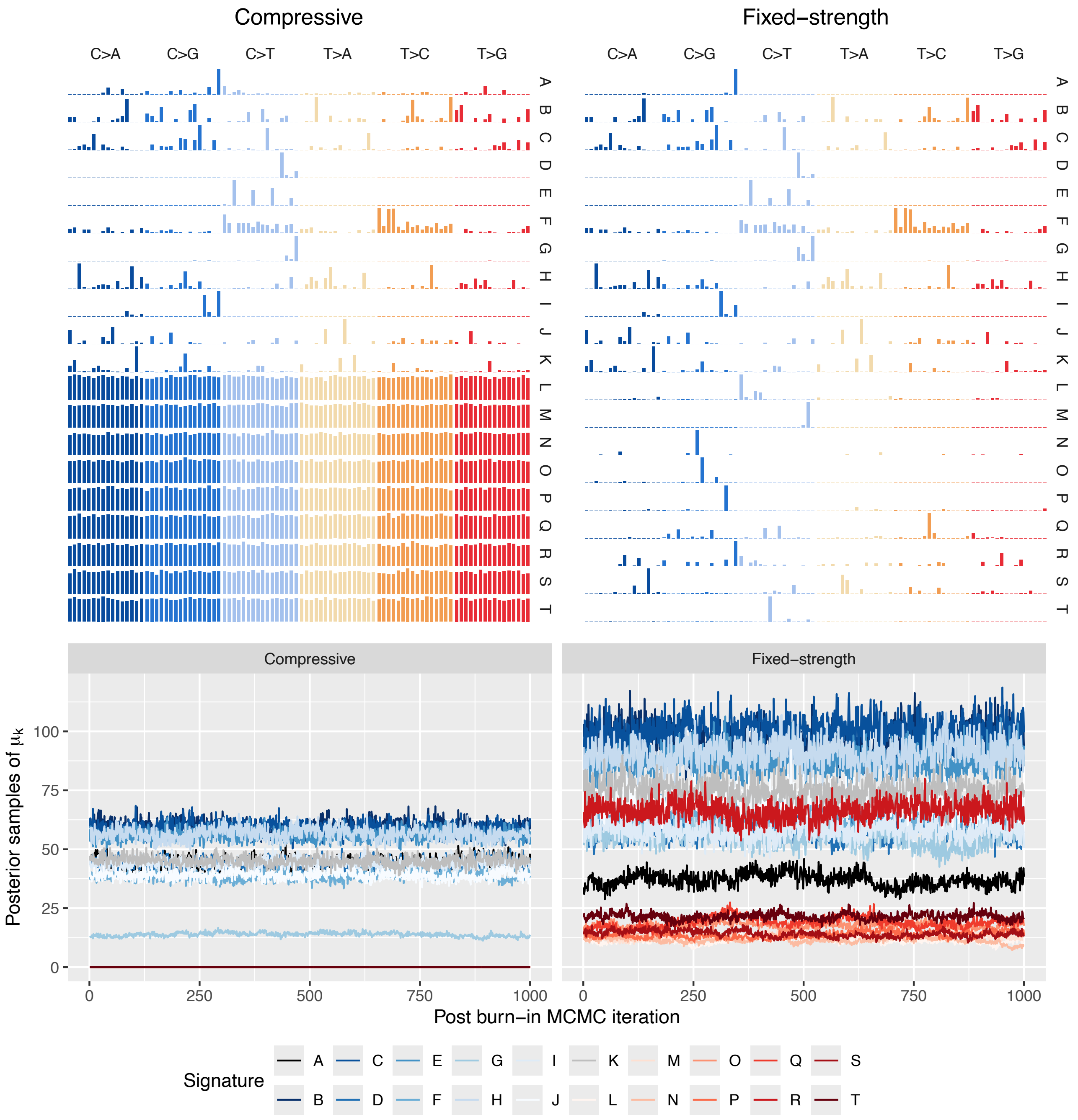


Figure S7.3: Posterior mean of R and traceplot for μ in one of the 20 simulated datasets with $J = 300$ and $\tau = 0.15$.

Figure S7.3 and (ii) the fact that the true parameters are more accurately estimated by the compressive model.

S7.2 Sensitivity to the choice of model settings

In this section, we evaluate the sensitivity of our proposed model to the choice of settings (K, ε, a, and α). We find that the model's results are fairly robust to these settings over a wide range of

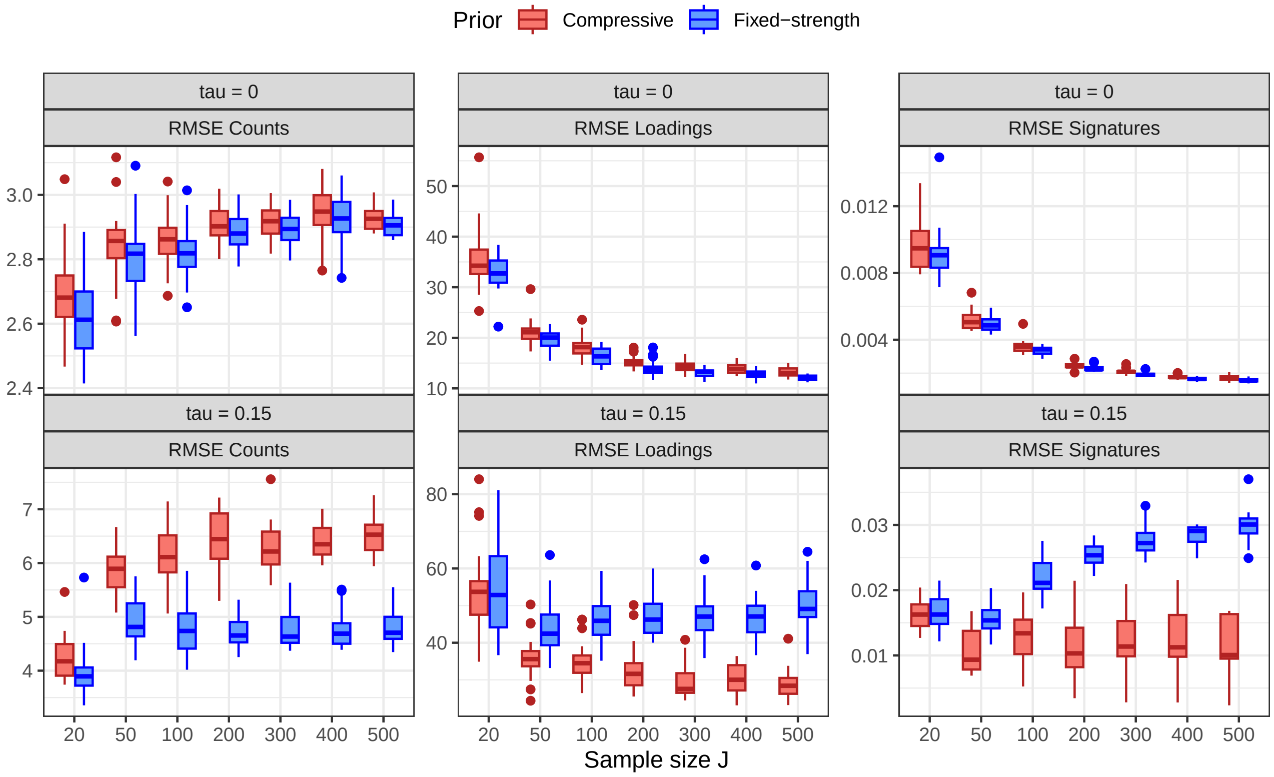


Figure S7.4: RMSE between the count matrix X and the product $\hat{R}\hat{\Theta}$ (left), RMSE between the true loadings Θ^0 and estimated $\hat{\Theta}$ (center), and RMSE between the true signatures R^0 and $\hat{R}$ (right) for both models under $\tau = 0$ (top row) and $\tau = 0.15$ (bottom row).

values. The a value has the most important effect, controlling the threshold above which signatures are included in the model. We use the same simulation setup as in Section 5, and generate 40 simulated datasets for each value of overdispersion $\tau \in \{0, 0.15\}$, with $J = 100$, $K^0_{\text{pre}} = 4$, and $K^0_{\text{new}} = 2$, so that there are 6 true signatures.

S7.2.1 Sensitivity to K and ε

We first evaluate sensitivity with respect to the maximum number of signatures in the model, K, and the prior mean of the relevance weights, $\varepsilon = \mathbb{E}(\mu_k)$. For each dataset, for each combination of $K \in \{5, 10, 20, 30, 40, 50\}$ and $\varepsilon \in \{0.001, 0.01, 0.1, 0.25, 1\}$, we run our compressive model with $a = 1$ and $\alpha = 0.5$ for 4,000 MCMC iterations, retaining the samples from the last 1,000 iterations for inference. Notice that when $K = 5$, the model necessarily has fewer signatures than the true number.

Figure S7.5 displays the results, quantified in terms of (i) estimated number of signatures, (ii) RMSE between the true loadings matrix Θ and the model estimate $\hat{\Theta}$, and (iii) RMSE between the true signature matrix R and the model estimate $\hat{R}$. When $K = 5$, the performance suffers since there are 6 true signatures. However, the performance is relatively constant for all values of K greater than

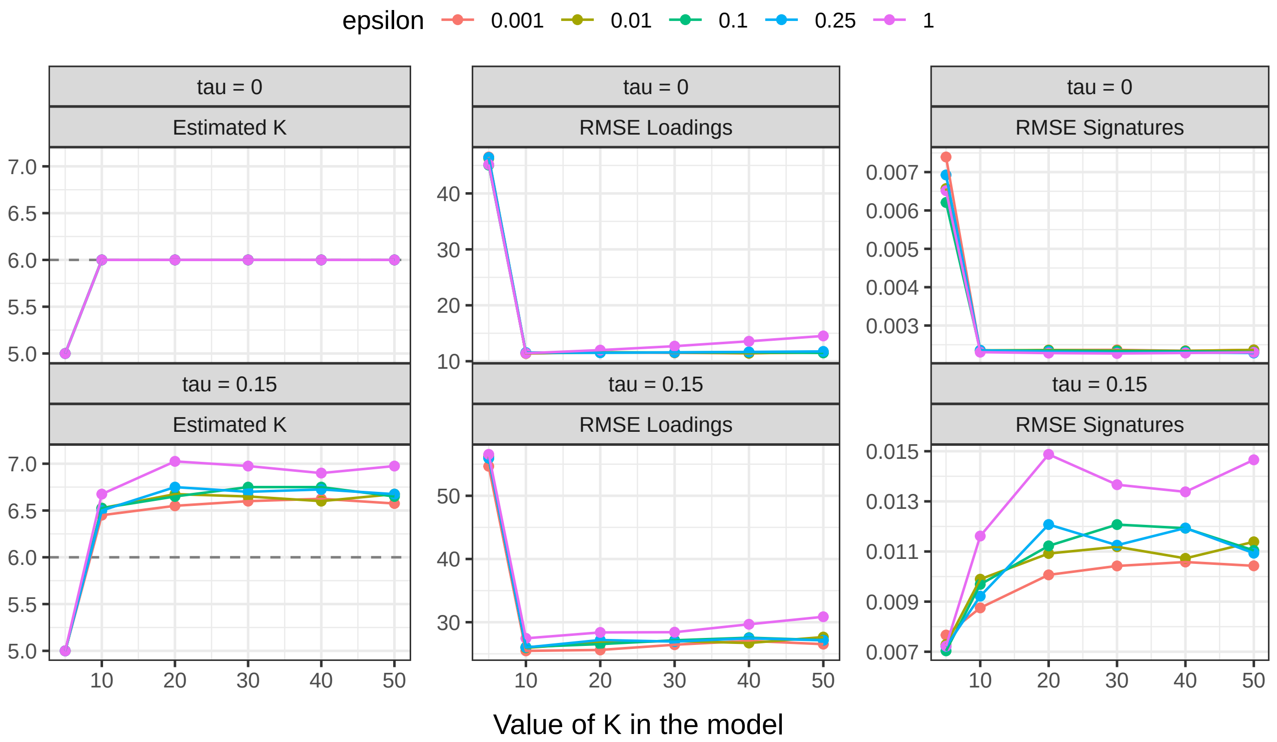


Figure S7.5: Left: estimated number of signatures. Center: RMSE between Θ and $\hat{\Theta}$. Right: RMSE between R and $\hat{R}$. The dashed horizontal gray line on the left panel indicates the true number of signatures in the data. Each point represents the average over 40 simulated datasets, for each combination of K, ε, and overdispersion τ, holding $a = 1$ and $\alpha = 0.5$ fixed. Each dataset was simulated with $J = 100$.

the true number. Similarly, the performance is relatively constant as a function of ε, for small values of ε. Performance degrades for values of $\varepsilon \geq 1$, which is well outside the range of recommended values. Overall, the posterior is robust to the choice of K and ε, as long as K is larger than the true number and ε is small, such as 0.01 or 0.001.

S7.2.2 Sensitivity to a and α

For each dataset, for each combination of $a \in \{0.1, 0.5, 1, 2\}$ and $\alpha \in \{0.01, 0.1, 0.5, 1, 2\}$, we run our model with $\varepsilon = 0.001$ and $K = 20$ for 4,000 MCMC iterations and retain the samples from last 1,000 iterations for inference. Figure S7.6 displays the results. We observe that smaller values of a produce a larger number of active signatures. This is consistent with the interpretation of a discussed in Section 3.3: it serves as a threshold such that signature k is included if $\bar{Y}_k > a$, where $\bar{Y}_k$ is the average number of mutations due to signature k. In turn, smaller a produces larger RMSE for the both R and Θ. This is likely due to overfitting in the cases where $a < 1$, as a result of introducing spurious signatures (Figure S7.6). Reconstruction of signatures and loadings appears more favorable when $a = 2$ in the misspecified case ($\tau = 0.15$), though $a = 1$ performs better than $a = 2$ in the correctly specified setting ($\tau = 0$). Regarding α, we observe that when a is not too small, the choice of

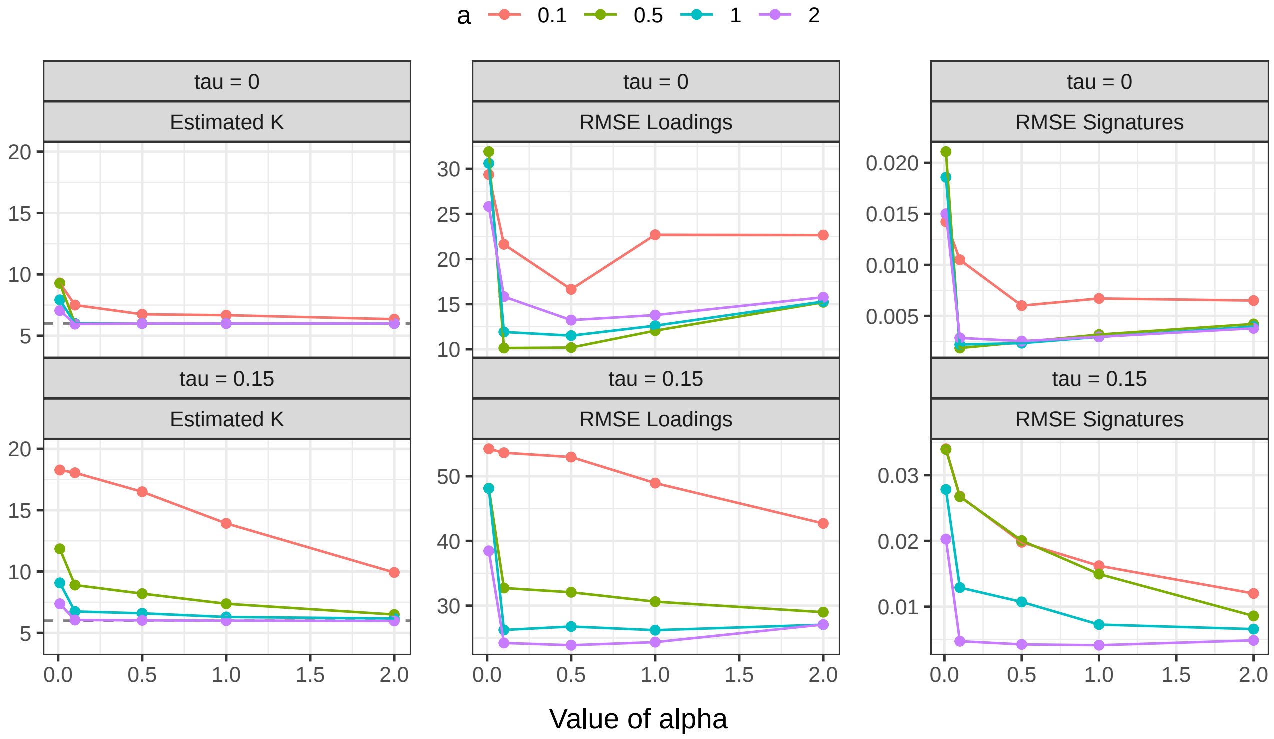


Figure S7.6: Left: estimated number of signatures. Center: RMSE between Θ and $\hat{\Theta}$. Right: RMSE between R and $\hat{R}$. The dashed horizontal gray line on the left panel indicates the true number of signatures in the data ($K^0 = 6$). Each point represents the average over 40 simulated datasets, for each combination of a, α, and overdispersion τ, holding $\varepsilon = 0.001$ and $K = 20$ fixed. Each dataset was simulated with $J = 100$.

α has little effect on the estimated number of signatures. One exception is that when α is very small ($\alpha = 0.01$), the model introduces a few extra signatures, presumably because this creates a strong prior preference for sparse signatures.

S8 Additional application results

This section contains additional details and results for the applications to the 21 breast cancer data and pancreatic adenocarcinoma indel data in Section 6.

S8.1 Additional results for the 21 breast cancer data

We provide further details on the inputs and outputs of the methods, and perform a sensitivity analysis for our CompNMF model.

S8.1.1 Details of each method

The methods we run are the same as in Section 5, with only the following changes:

(i) CompNMF: $K = 15$, $\varepsilon = 0.01$, $a = 1$, and $\alpha = 0.5$.

Table S8.1: Effective sample sizes of the Bayesian methods on the 21 breast cancer data. Numbers denote the average univariate effective sample sizes across all entries of R, Θ, and μ, with standard deviations in parenthesis, and are calculated using 2,000 posterior samples for each method.

	R	Θ	μ
CompNMF	1,229 (±953)	1,065 (±866)	670 (±531)
CompNMF+cosmic	219 (±161)	65 (±27)	697 (±867)
signeR	98 (±70)	36 (±23)	
PoissonCUSP	123 (±94)	57 (±12)	20 (±8)

(ii) CompNMF+cosmic: $K_{\text{new}} = 10$ *de novo* signatures and $K_{\text{pre}} = 67$ COSMIC v3.4 signatures, with $\varepsilon = 0.01$.

(iii) PoissonCUSP: starting at $K = 15$, with spike at $\mu_\infty = 0.01$ and $a_0 = b_0 = 1$.

(iv) signeR: we set `estimate_hyper = TRUE` as in Rosales et al. (2016), and let K range from 2 to 15; however, unlike Rosales et al. (2016), we do not include the opportunity counts in the model.

(v) SignatureAnalyzer: $K = 15$, selection method set to `L1W.L2H` (same as Alexandrov et al., 2020).

(vi) SigProfiler: K ranging from 2 to 15.

For both compressive methods, we randomly initialize by sampling from the prior, and run the Gibbs sampler for 12,000 iterations, discarding the first 10,000 as burn-in. This is repeated four times, and we select the run yielding the highest average log-posterior. PoissonCUSP is run for 12,000 iterations for a single chain, discarding the first 10,000 as burn-in.

The effective sample sizes (ESSs) of each Bayesian method are displayed in Table S8.1. We see that our unsupervised compressive NMF achieves the best performance and has good mixing overall. Adding the informative prior leads to lower effective sample sizes, but still higher than both signeR and PoissonCUSP.

S8.1.2 Signatures and loadings for each method

The complete sets of signatures inferred by each method on the 21 breast cancer data (Section 6) are shown in Figures S8.1 to S8.6 and are discussed below.

SigProfiler estimates only three signatures (matched to SBS2, SBS3, and SBS40a), and the cosine similarity is high only for SBS3; see Figure S8.1. This is likely a consequence of the small sample size, which can lead to the merging of two or more signatures due to insufficient signal to distinguish them. Interestingly, SignatureAnalyzer estimates five signatures (matched to SBS1, SBS2, SBS2, SBS3, and SBS13), but there is duplication since one of them appears to be a combination of SBS2 and SBS13, and ends up being matched to SBS2; see Figure S8.2.

We find that signeR infers five signatures (matched to SBS1, SBS2, SBS3, SBS13, and SBS96), similar to the results in the signeR paper (Rosales et al., 2016) but with slight differences since we do not account for the opportunity count matrix, in order to provide a consistent comparison between methods; see Figure S8.3. In the signeR results, SBS1 (which is sparse) appears to have been merged with a flatter signature and thus is retrieved with lower cosine similarity. The method also infers a signature that is matched to SBS96, but with low cosine similarity.

PoissonCUSP and CompNMF estimate a larger number of signatures (seven and six, respectively); see Figures S8.4 and S8.5. In both cases, SBS1, SBS2, SBS3, and SBS13 are inferred with cosine similarities comparable to the other methods. They also both infer SBS34 and SBS98, but these have larger credible intervals, indicating greater uncertainty. SBS98 also has particularly low cosine similarity in both cases, suggesting that it may be spurious. PoissonCUSP also estimates a signature matched to SBS9, but again with relatively high uncertainty and low cosine similarity. SBS9 has been found in other breast cancer types (Alexandrov et al., 2020), but its current hypothesized etiology (polymerase eta somatic hypermutation in lymphoid cells) has not been validated.

Finally, our CompNMF+cosmic model estimates eight signatures, all with cosine similarity near 1, except for SBS98; see Figure S8.6. The estimated signature matched to SBS98 has very high uncertainty and low cosine similarity, suggesting it is probably a spurious match. As in the simulations, the informative prior appears to provide significantly improved sensitivity to detect the presence of signatures, while still allowing for departures from the COSMIC signatures.

The loadings estimated by PoissonCUSP, signeR, SignatureAnalyzer, and SigProfiler are displayed in Figure S8.7 as percentages.

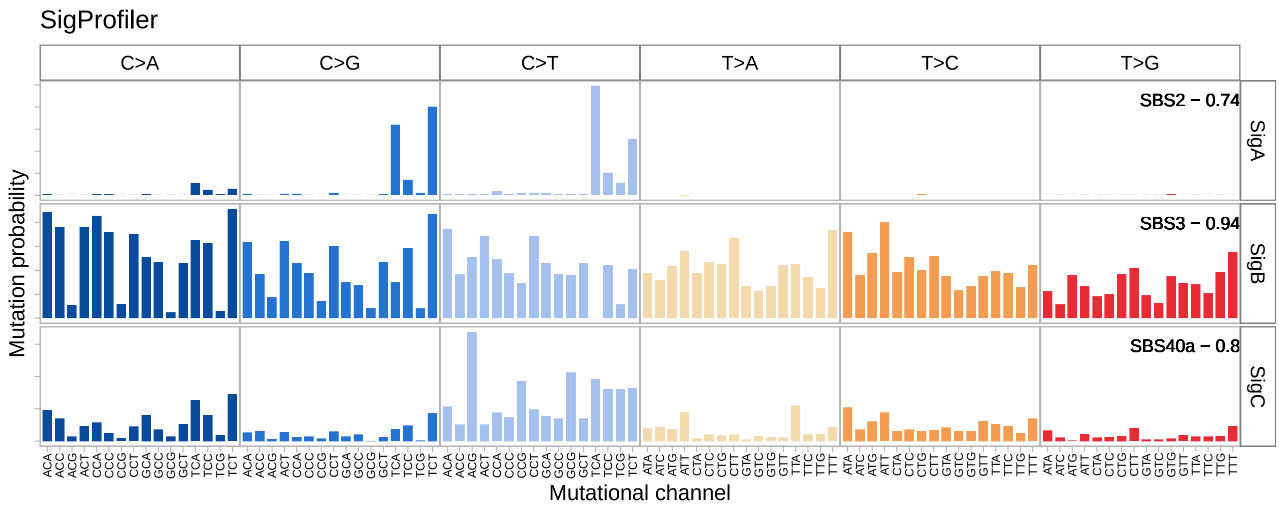


Figure S8.1: Mutational signatures inferred by SigProfiler on the 21 breast cancer dataset.

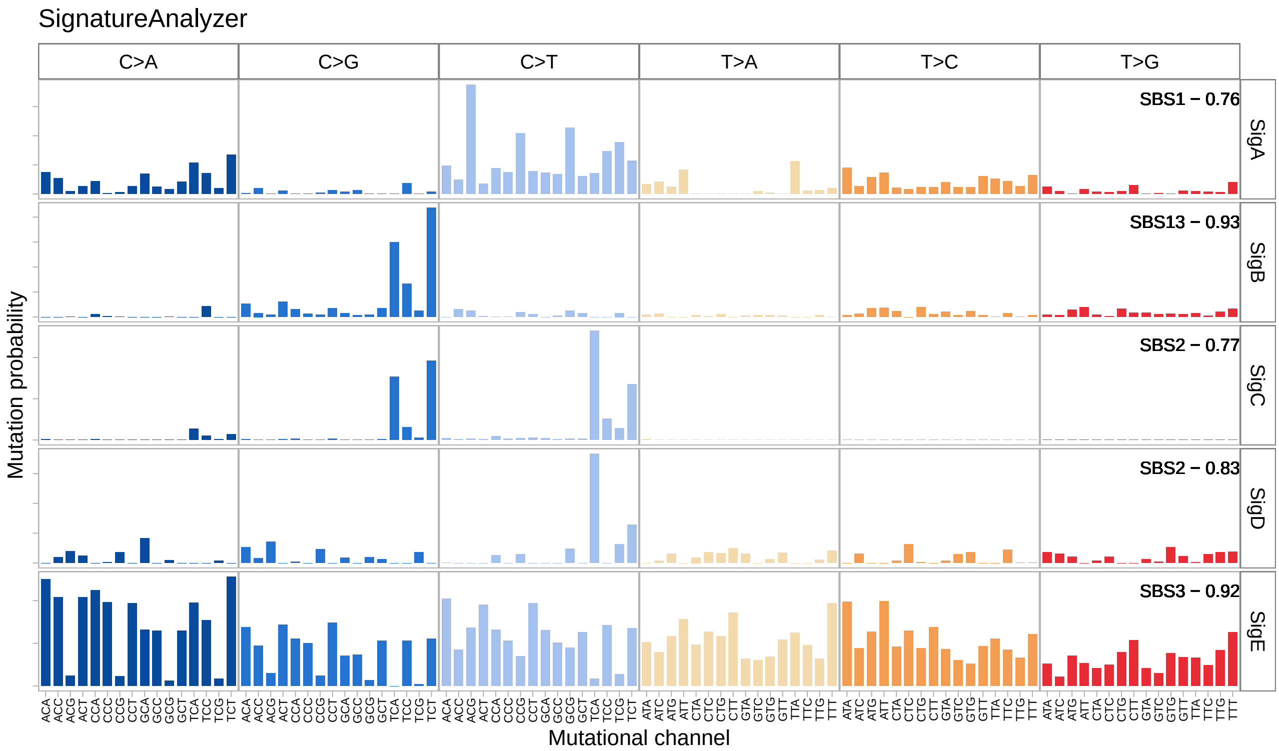


Figure S8.2: Mutational signatures inferred by SignatureAnalyzer on the 21 breast cancer dataset.

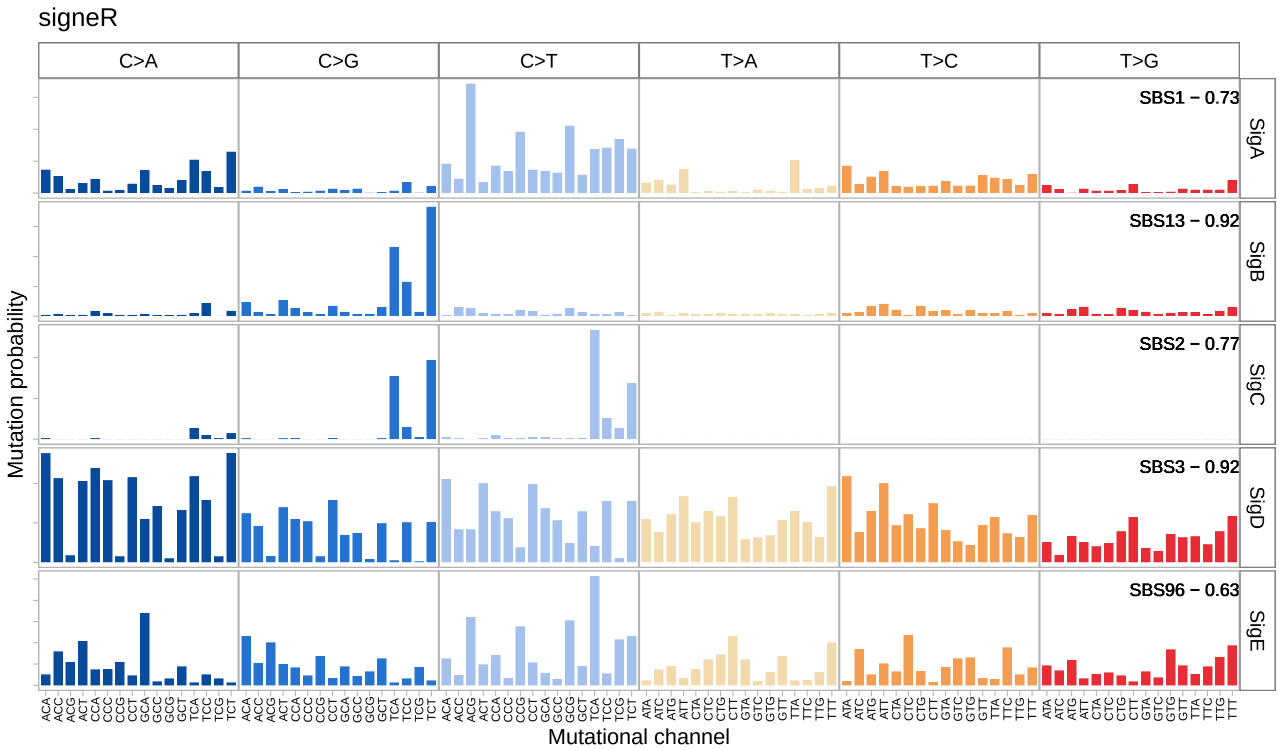


Figure S8.3: Mutational signatures inferred by signeR on the 21 breast cancer dataset.

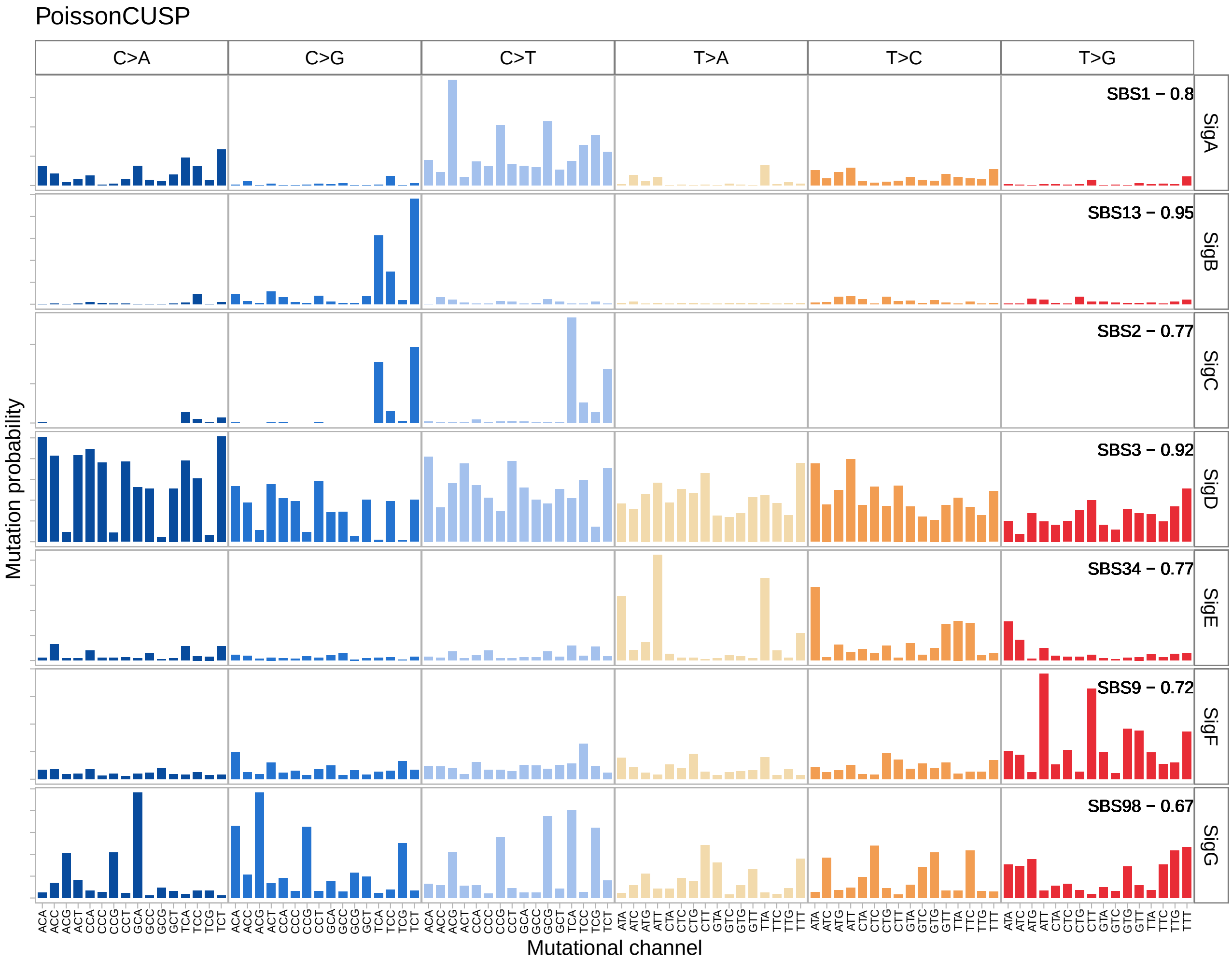


Figure S8.4: Mutational signatures inferred by PoissonCUSP on the 21 breast cancer dataset.

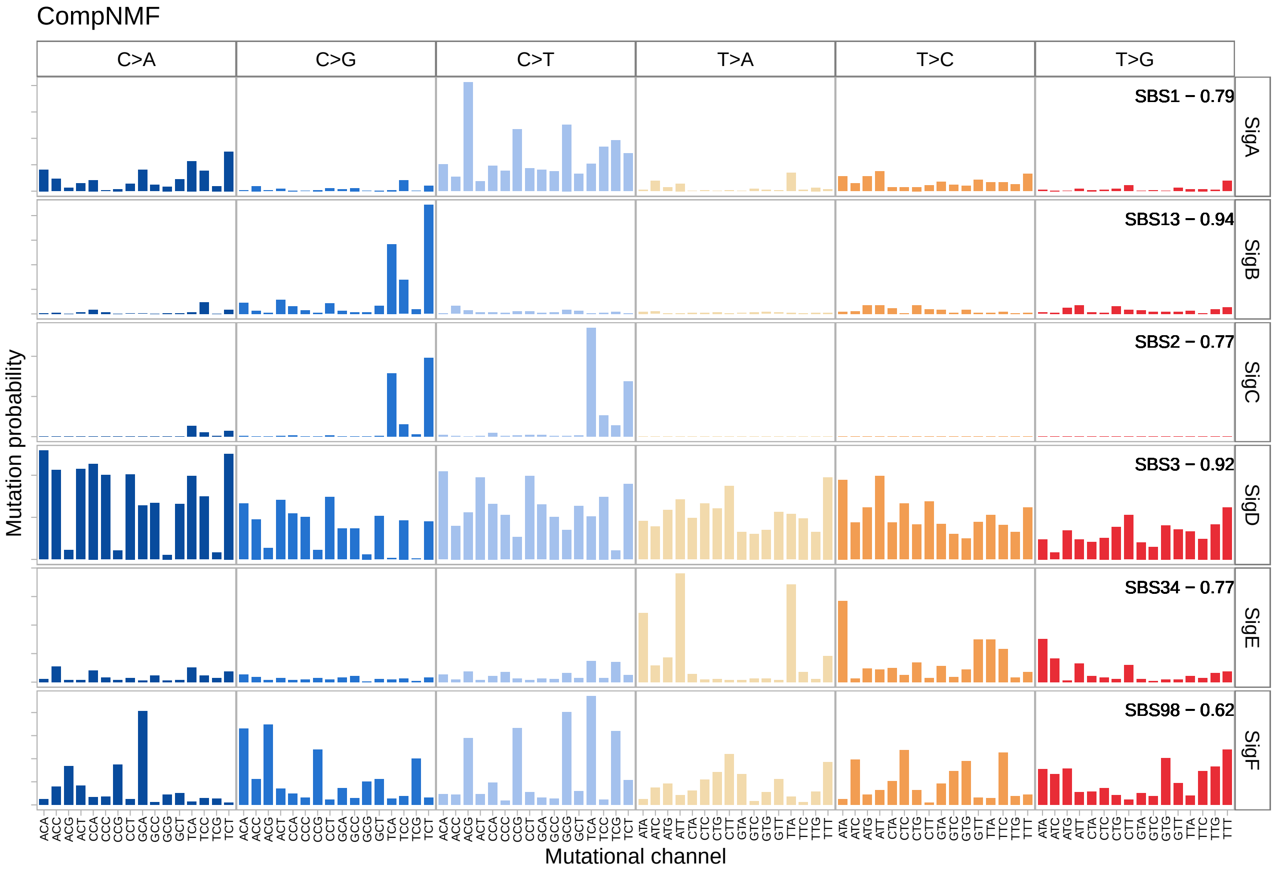


Figure S8.5: Mutational signatures inferred by CompNMF on the 21 breast cancer dataset.

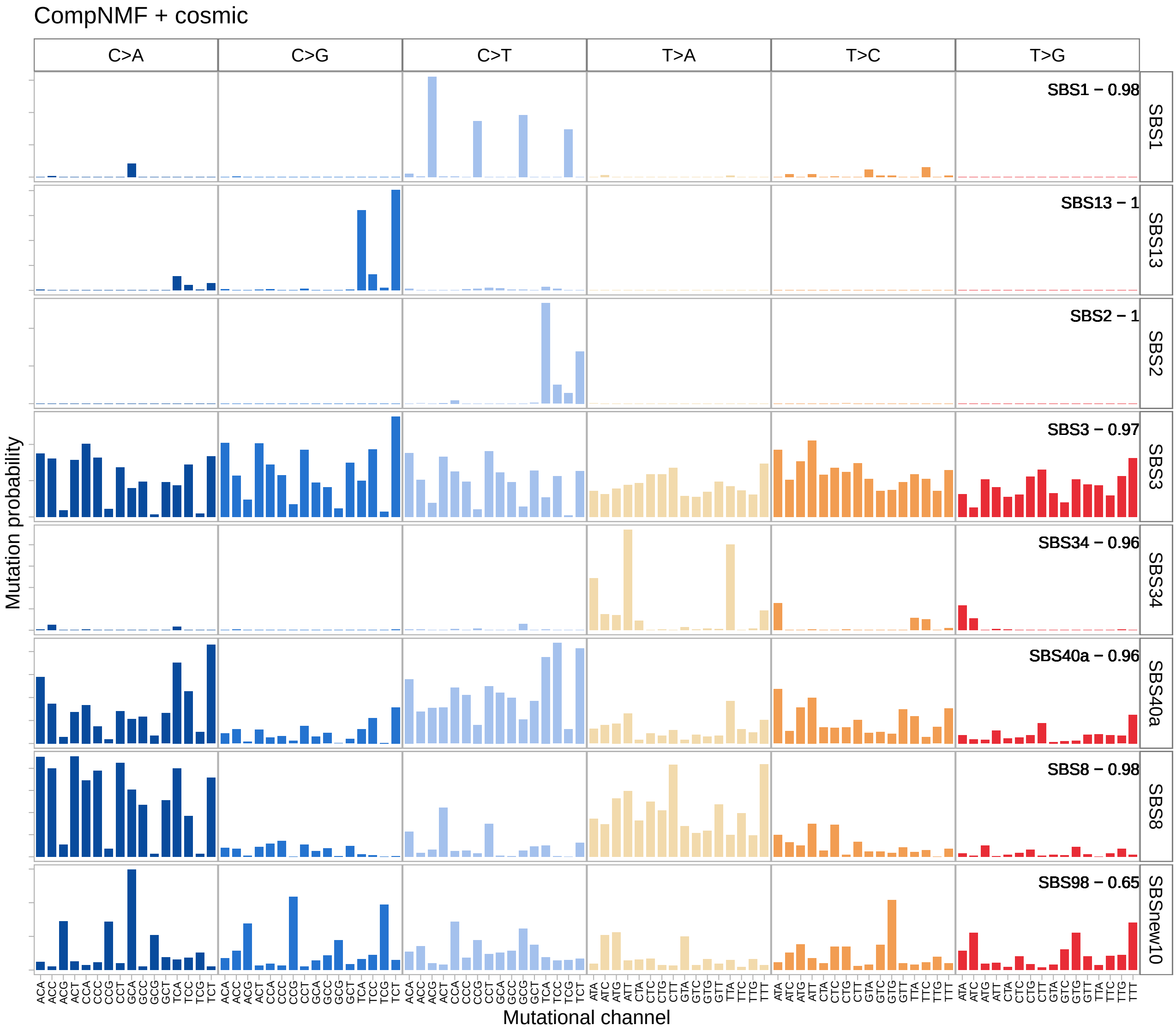


Figure S8.6: Mutational signatures inferred by CompNMF+cosmic on the 21 breast cancer dataset.

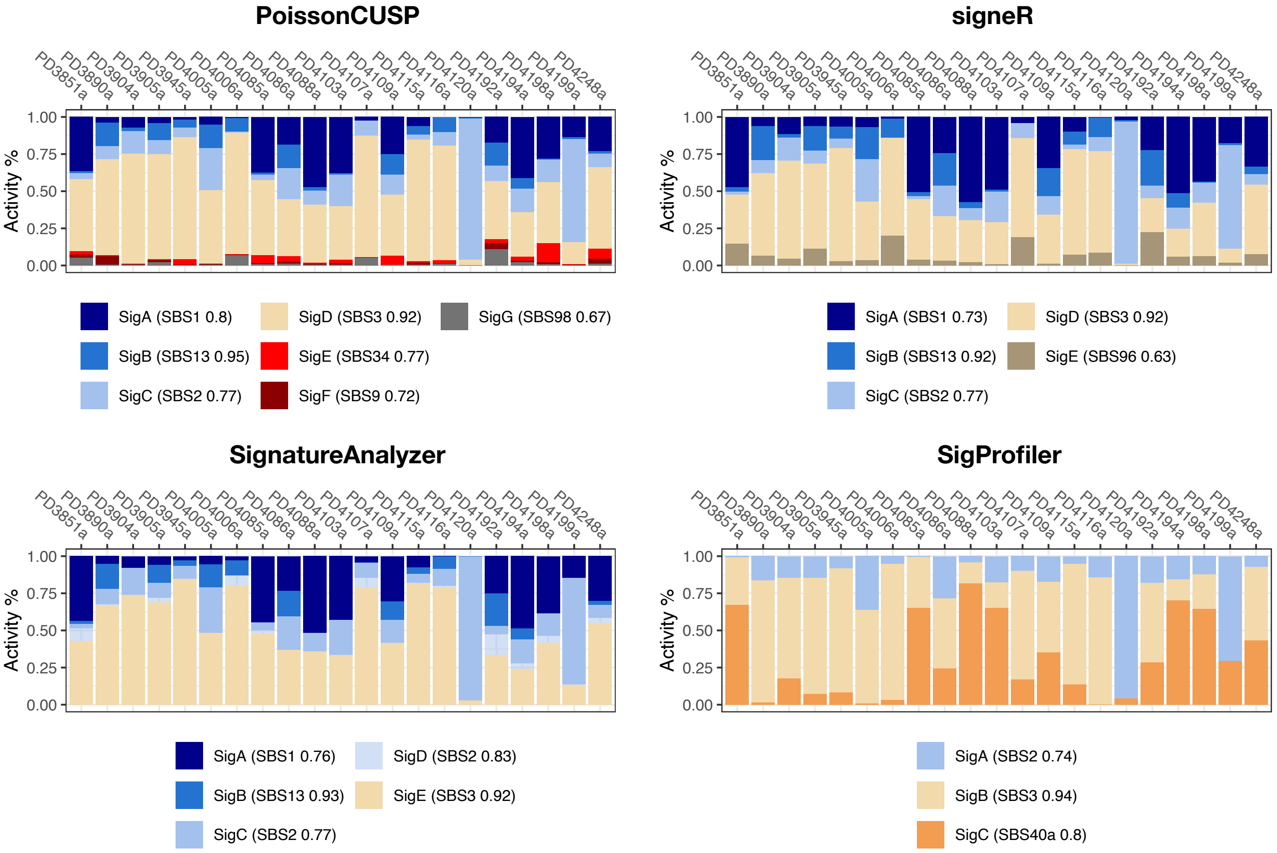


Figure S8.7: Loading of each signature (as a percentage of the total loading) for each patient, for PoissonCUSP, signeR, SignatureAnalyzer, and SigProfiler on the 21 breast cancer dataset.

S8.1.3 Sensitivity analysis for CompNMF on the 21 breast cancer dataset

We now assess the effect of the choice of K, ε, a, and α on the posterior inferred by the unsupervised compressive NMF model (that is, without the informative prior) applied to the 21 breast cancer dataset. We consider the following ranges of values:

- $K = \{15, 20, 25, 40\}$,
- $\varepsilon = \{0.01, 0.1, 0.25\}$,
- $a = \{1, 2, 0.5\}$,
- $\alpha = \{0.5, 1, 2\}$.

For each combination of K, ε, a, and α, we run the compressive NMF model for one randomly initialized MCMC chain for 12,000 iterations and keep the last 2,000 for inference.

Figure S8.8 displays the estimated value of K^* across all combinations of model settings. The number of inferred signatures is only slightly affected by the values of K, ε, and α. This is consistent with the sensitivity analysis in Section S7: these model settings have minimal impact as long as K is large enough, ε is small, and α is not too extreme. The most impactful setting is the choice of a: on average, $a = 0.5$ yields 8 signatures, $a = 1$ yields 6 signatures, and $a = 2$ yields 4 signatures. This is expected, since a determines the location of the elbow of the compression curve, as described by Theorem 3 and Figure S7.1. In other words, the larger a is, the more mutations need to be caused by a signature in order for it to be included.

To evaluate the sensitivity of the signatures R and loadings Θ to the choice of model settings, since the ground truth is not known for this real data, we define R° and Θ° to be the estimated values when using the model settings presented in Section S8.1.1: $K = 15$, $\varepsilon = 0.01$, $a = 1$, and $\alpha = 0.5$. Then for each combination of K, ε, a, and α, we compute the RMSE between R° and Θ° and the inferred values of R and Θ, respectively; see Figure S8.9. As before, there is not much variation in the results across values of K, ε, and α. We observe some variation in the RMSE results as a varies, with our recommended value of $a = 1$ performing best overall in these experiments.

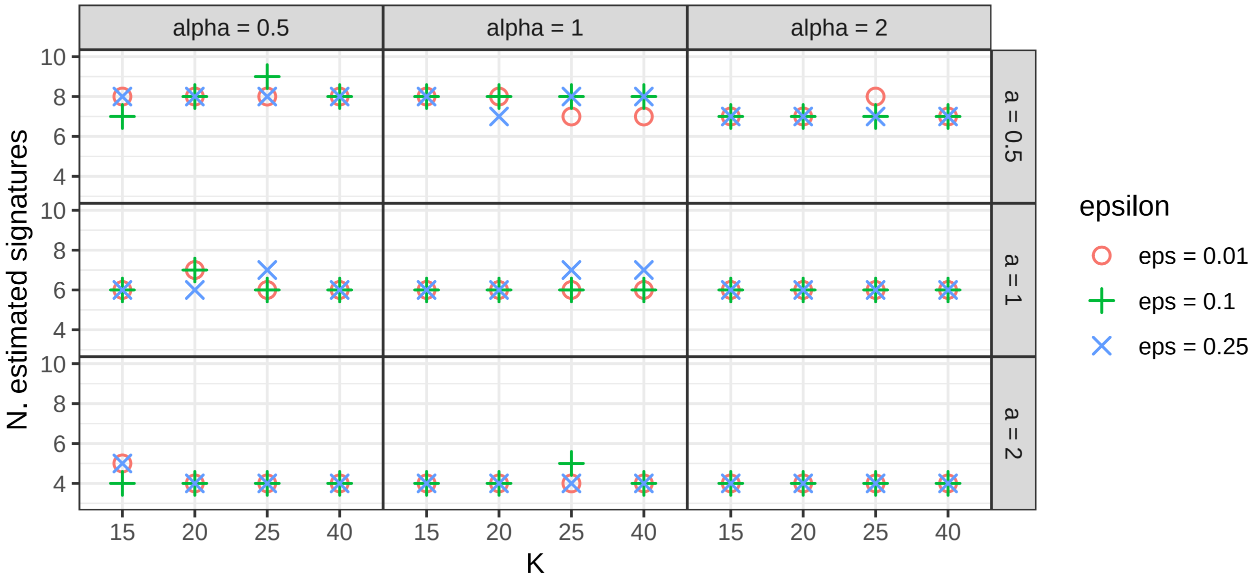


Figure S8.8: Number of estimated signatures in the 21 breast cancer data for varying K, ε, a, α.

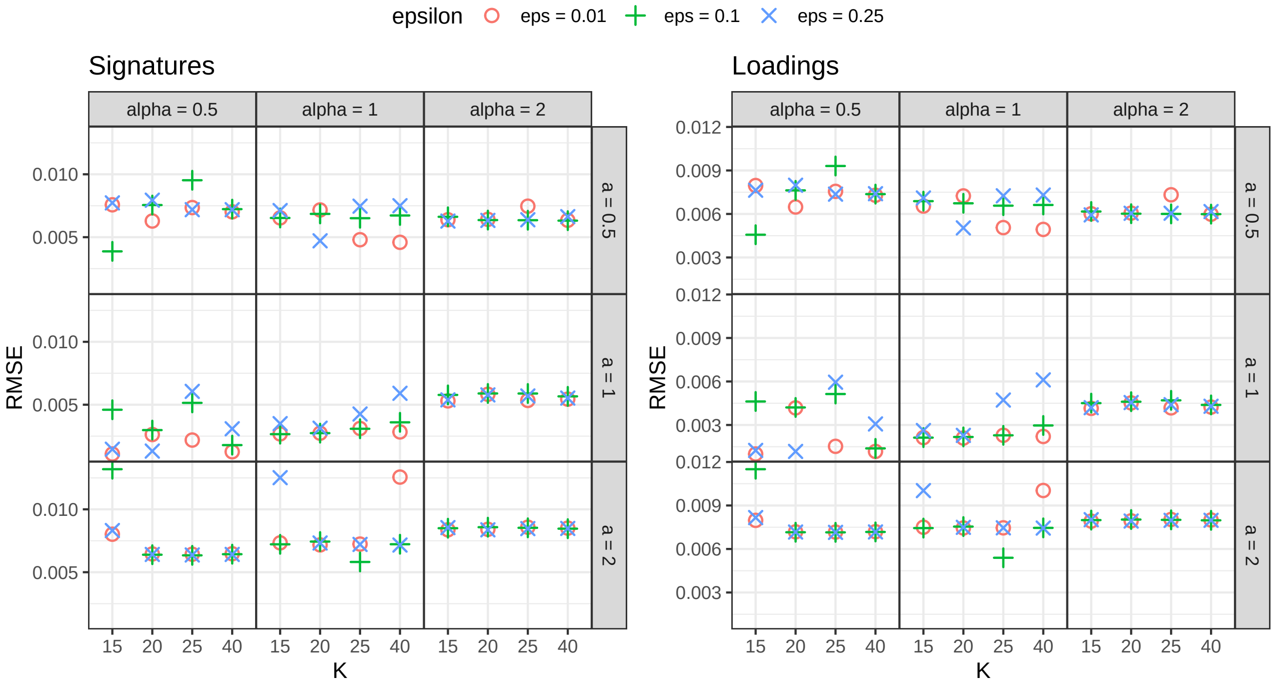


Figure S8.9: RMSE of R and Θ for varying values of the model settings K, ε, a, and α. Here, RMSE is defined as the root mean squared distance from the estimates obtained when using the settings from the main text: $K = 15$, $\varepsilon = 0.01$, $a = 1$, and $\alpha = 0.5$.

S8.2 Details for the pancreatic adenocarcinoma application

We now report the details for the ICGC `Panc-AdenoCA` data analysis. We run the following methods:

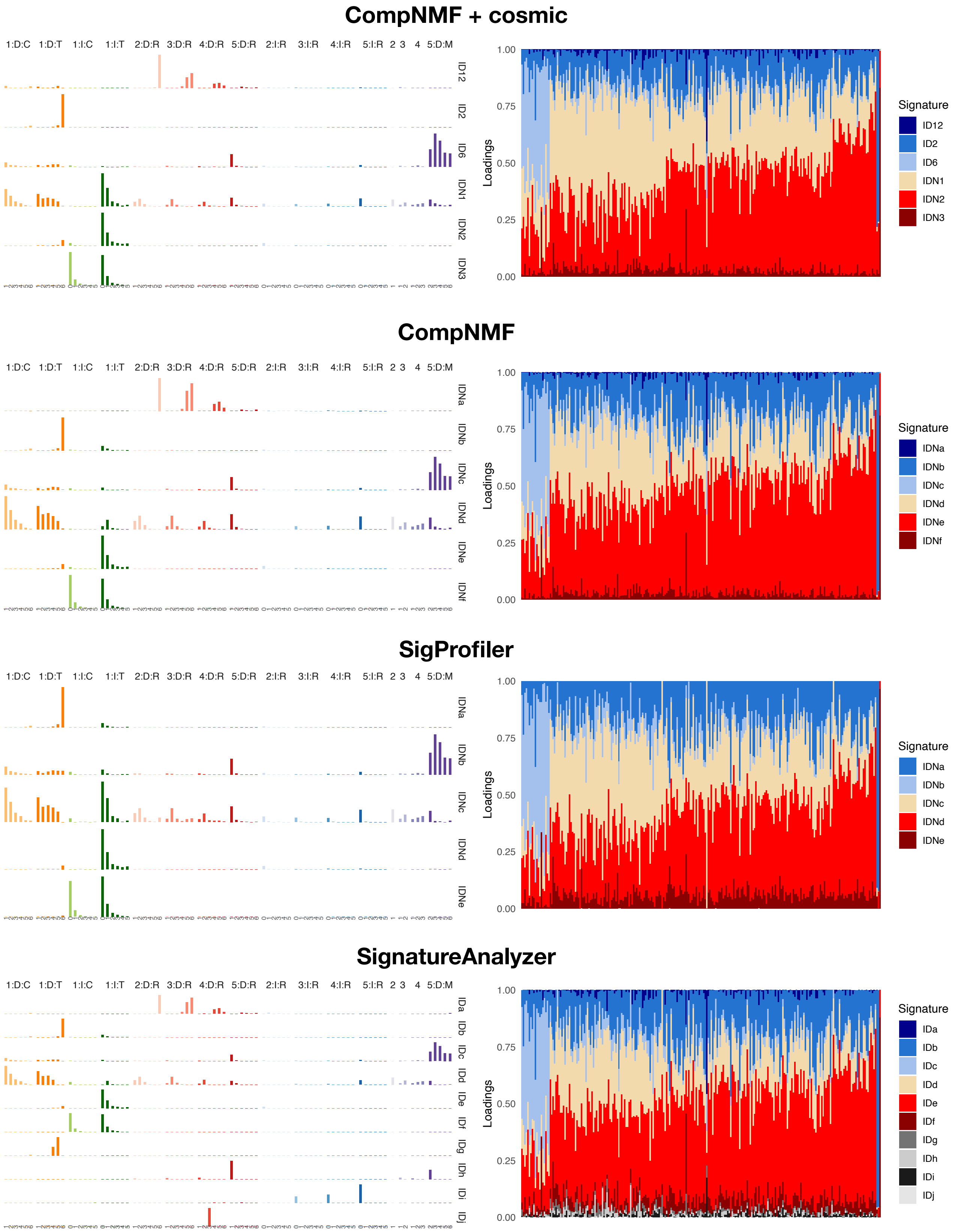
(i) CompNMF with $K = 20$, $a = 1$, $\varepsilon = 0.001$, and $\alpha = 0.5$,

(ii) CompNMF + cosmic with $K^{\text{pre}} = 23$ and $K^{\text{new}} = 10$, β_k obtained so that prior samples from the Dirichlet have 0.95 cosine similarity with the associated COSMIC signature,

(iii) SignatureAnalyzer with $K = 25$ and `method = L1W.L2H` (same as Alexandrov et al., 2020),

(iv) SigProfilerExtractor, with K ranging from 2 to 15 and 25 replicates each.

Figure S8.10 shows the inferred signatures and loadings in each case.



CompNMF + cosmic

CompNMF

SigProfiler

SignatureAnalyzer

Figure S8.10: Signatures and normalized loadings in the four methods.

References

Abramowitz, M. and I. A. Stegun (1972). *Handbook of Mathematical Functions with formulas, graphs, and mathematical tables.*

Alexandrov, L. B., J. Kim, N. J. Haradhvala, et al. (2020). The repertoire of mutational signatures in human cancer. *Nature 578*(7793), 94–101.

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, et al. (2013). Signatures of mutational processes in human cancer. *Nature 500*(7463), 415–421.

Ayed, F. and F. Caron (2021). Nonnegative Bayesian nonparametric factor models with completely random measures. *Statistics and Computing 31*(5), 63.

Bhadra, A., J. Datta, N. G. Polson, and B. Willard (2019). Lasso meets horseshoe: A survey. *Statistical Science 34*(3), 405 – 427.

Bhattacharya, A. and D. B. Dunson (2011). Sparse Bayesian infinite factor models. *Biometrika 98*(2), 291–306.

Brouwer, T., J. Frellsen, and P. Lió (2017). Comparative Study of Inference Methods for Bayesian Nonnegative Matrix Factorisation. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.*

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience 2009*, 785152.

Datta, J. and D. B. Dunson (2016, 12). Bayesian inference on quasi-sparse count data. *Biometrika 103*(4), 971–983.

Drummond, R. D., A. Defelicibus, M. Meyenberg, R. Valieris, E. Dias-Neto, R. A. Rosales, and I. T. da Silva (2023). Relating mutational signature exposures to clinical data in cancers via signeR 2.0. *BMC Bioinformatics 24*(1), 439.

Durante, D. (2017). A note on the multiplicative gamma process. *Statistics & Probability Letters 122*, 198–204.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. New York: John Wiley & Sons Inc.

Frühwirth-Schnatter, S. (2023). Generalized cumulative shrinkage process priors with applications to sparse Bayesian factor analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 381*(2247), 20220148.

Gopalan, P., J. M. Hofman, and D. M. Blei (2015). Scalable recommendation with hierarchical Poisson factorization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, Arlington, Virginia, USA, pp. 326–335. AUAI Press.

Gopalan, P., F. Ruiz, R. Ranganath, and D. Blei (2014). Bayesian nonparametric Poisson factorization for recommendation systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Volume 33, pp. 275–283.

Gopalan, P. K., L. Charlin, and D. Blei (2014). Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems*, Volume 27.

Grabski, I., L. Trippa, and G. Parmigiani (2023). Bayesian multi-study non-negative matrix factorization for mutational signatures. *bioRxiv:10.1101/2023.03.28.534619*.

Hamura, Y., K. Irie, and S. Sugasawa (2022). On global-local shrinkage priors for count data. *Bayesian Analysis 17*(2), 545 – 564.

Hansen, B., I. N. Grabski, G. Parmigiani, and R. D. Vito (2025). Bayesian probit multi-study nonnegative matrix factorization for mutational signatures.

Hoffman, M. D., D. M. Blei, and P. R. Cook (2010). Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 439–446.

Islam, S. A., M. Díaz-Gay, Y. Wu, et al. (2022). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics 2*(11), 100179.

Kim, J., K. W. Mouw, P. Polak, et al. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics 48*(6), 600–606.

Kingman, J. F. C. (1993). *Poisson processes*, Volume 3 of *Oxford Studies in Probability.* New York: The Clarendon Press Oxford University Press. Oxford Science Publications.

Lawler, G. F. (2018). *Introduction to Stochastic Processes.* CRC Press.

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

Lu, J. and C. P. Chai (2022). Robust Bayesian nonnegative matrix factorization with implicit regularizers. *ArXiv preprint arXiv:2208.10053*.

Lu, J. and X. Ye (2022). Flexible and hierarchical prior for Bayesian nonnegative matrix factorization. *ArXiv preprint arXiv:2205.11025*.

Miller, J. W. (2021, jan). Asymptotic normality, concentration, and coverage of generalized posteriors. *J. Mach. Learn. Res. 22*(1).

Mørup, M. and L. K. Hansen (2009). Tuning pruning in sparse non-negative matrix factorization. In *2009 17th European Signal Processing Conference*, pp. 1923–1927.

Polson, N. G. and J. G. Scott (2011, 10). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9*. Oxford University Press.

Rahiche, A. and M. Cheriet (2022). Variational Bayesian orthogonal nonnegative matrix factorization over the Stiefel manifold. *IEEE Transactions on Image Processing 31*, 5543–5558.

Rosales, R. A., R. D. Drummond, R. Valieris, E. Dias-Neto, and I. T. da Silva (2016). signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics 33*(1), 8–16.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 689–710.

Schiavon, L., A. Canale, and D. B. Dunson (2022). Generalized infinite factorization models. *Biometrika 109*(3), 817–835.

Schmidt, K. D. (2003). On the covariance of monotone functions of a random variable. Technical report. Inst. für Math. Stochastik.

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

Townes, F. W. and B. E. Engelhardt (2023). Nonnegative spatial factorization applied to spatial genomics. *Nature Methods 20*(2), 229–238.

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

Zhou, M. (2015). Infinite edge partition models for overlapping community detection and link prediction. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Volume 38, pp. 1135–1143.

Zhou, M. (2018). Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis 13*(4), 1065–1093.

Zhou, M. and L. Carin (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 307–320.