Relational Dose-Response Modeling for Cancer Drug Studies

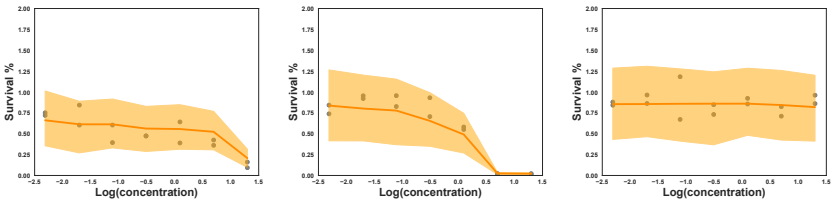
Wesley Tansey*[1,2], Christopher Tosh[1], and David M. Blei[1,3,4]

[1]Data Science Institute, Columbia University, New York, NY, USA
[2]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA
[4]Department of Statistics, Columbia University, New York, NY, USA
[5]Department of Computer Science, Columbia University, New York, NY, USA

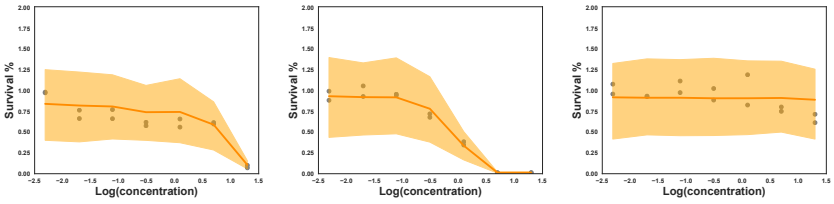
Abstract

Exploratory cancer drug studies test multiple tumor cell lines against multiple candidate drugs. The goal in each paired (cell line, drug) experiment is to map out the dose-response curve of the cell line as the dose level of the drug increases. The level of natural variation and technical noise in these experiments is high, even when multiple replicates are run. Further, running all possible combinations of cell lines and drugs may be prohibitively expensive, leading to missing data. Thus, estimating the dose-response curve is a denoising and imputation task. We cast this task as a functional matrix factorization problem: finding low-dimensional structure in a matrix where every entry is a noisy function evaluated at a set of discrete points. We propose Bayesian Tensor Filtering (BTF), a hierarchical Bayesian model of matrices of functions. BTF captures the smoothness in each individual function while also being locally adaptive to sharp discontinuities. The BTF model can incorporate many types of likelihoods, making it flexible enough to handle a wide variety of data. We derive efficient Gibbs samplers for three classes of likelihoods: (i) Gaussian, for which updates are fully conjugate; (ii) binomial and related likelihoods, for which updates are conditionally conjugate through Pólya–Gamma augmentation; and (iii) non-conjugate likelihoods, for which we develop an analytic truncated elliptical slice sampling routine. We compare BTF against a state-of-the-art method for dynamic Poisson matrix factorization, showing BTF better reconstructs held out data in synthetic experiments. Finally, we build a dose-response model around BTF and apply it to real data from two multi-sample, multi-drug cancer studies. We show that the BTF-based dose-response model outperforms the current standard approach in biology. Code is available at `https://github.com/tansey/functionalmf`.

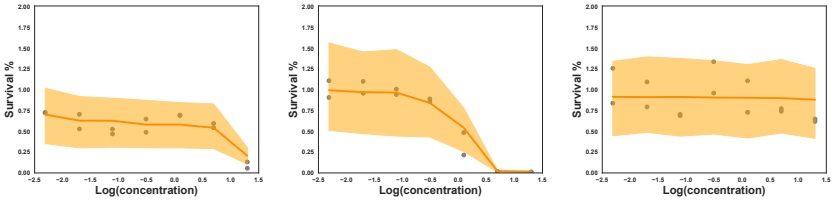
*`wesley.tansey@columbia.edu` (corresponding author)



(a) Tumor 1, Drug 1 (b) Tumor 1, Drug 2 (c) Tumor 1, Drug 3



(d) Tumor 2, Drug 1 (e) Tumor 2, Drug 2 (f) Tumor 2, Drug 3



(g) Tumor 3, Drug 1 (h) Tumor 3, Drug 2 (i) Tumor 3, Drug 3

Figure 1: Sample of data from an organoid cancer drug experiment. Gray dots are observed outcomes; the orange line is the mean predicted response; bands represent 50% posterior predictive credible intervals for observations. All nine experiments were held out from the model at training time.

1 Introduction

To search for new therapeutics, biologists carry out exploratory studies of drugs. They test multiple drugs, at different doses, against multiple biological samples (e.g., different tumor cell lines). The goal is to trace the dose-response curves, and to understand the efficacy of each drug. The experiments in such dose-response studies are costly; each one can take weeks or months to conduct in the lab. Consequently, the exhaustive set of all (sample, drug) combinations is often prohibitively expensive. This leaves missing data, dose-response curves for which no data is available, that must be imputed. Moreover, even with an exhaustive set of experiments, the noise level in the data is high and estimating the true expected response curves is challenging.

Figure 1 shows data from a cancer drug discovery study [12]. Each panel illustrates the interaction of one type of drug with one type of tumor (organoid cell line). The gray points are the results of a set of experiments, each set with 2 replicates measured at 7 different doses. The inferential goal is to use

the observations (gray points) to predict the true dose-response curves. The predictions—the orange lines and uncertainty bands—come from a dose-response model built on top of Bayesian tensor filtering (BTF), the method we propose in this paper.

Notice there is relational structure in the outcomes: each drug has similar effects on each tumor. Thus, we treat predictive modeling of dose-response as a factorization problem. The relational structure in the data arises because tumors share latent molecular attributes, such as genomic mutations, and drugs share latent pharmaceutical attributes, such as chemical structures. In each experiment, tumor and drug attributes interact, creating the shared patterns of dose-response.

While traditional factorization considers a matrix of scalars, the entries of this matrix are latent dose-response curves subsampled at different doses. To model such curves, we model drug attributes as changing functions of dose. While the effects usually vary smoothly, there are occasional sharp jumps, such as between the sixth and seventh dose levels of drug 1. Capturing latent structure in dose-response curves requires handling this type of non-stationarity.

As a final wrinkle, depending on the drugs being tested, equipment being used, or samples under study, the types of observations will change. Some studies may generate count data for the number of cells surviving; others measure a real-valued metric of cell health or survival. In the case of cancer, drugs are chosen that will only kill cells, not facilitate growth; effects in these curves are therefore upper bounded. All of these complexities mean that good models of dose-response curves must handle many different likelihoods and allow the scientist to encode biological constraints on parameter values.

Bayesian tensor filtering (BTF) is a probabilistic method for functional matrix factorization that handles these special properties of data about dose-response curves. BTF uses structured shrinkage priors that encourage smoothness in the estimated functions; it is locally adaptive, enabling the functions to make sharp jumps when the data calls for it; and it can accommodate any likelihood function. We derive efficient MCMC inference methods for BTF: specialized inference for Gaussian and binomial likelihoods, and a new inference method called generalized analytic slice sampling (GASS), for more general likelihoods. BTF enables us to develop a new state-of-the-art method for dose-response modeling in cancer drug studies.

1.1 Outline

The remainder of this paper is as follows. Section 3 presents Bayesian tensor filtering, a flexible model for functional matrix factorization. In section 4, we detail generalized analytic slice sampling, a procedure for sampling from posteriors with constrained multivariate normal priors and non-conjugate likelihoods. These two procedures are combined in section 5 to create a new Bayesian dose-response model for multi-drug, multi-sample cancer studies. Section 6 empirically studies GASS, BTF, and the proposed dose-respose model. We benchmark GASS on a constrained, non-conjugate Gaussian process posterior inference task, showing

it mixes faster and has better coverage than a set of alternative inference approaches. We compare BTF to a state-of-the-art method for functional Poisson matrix factorization and find it better models non-stationary functional matrices. Finally, we study the dose-response model on two real cancer datasets and find it has better reconstruction performance than standard approaches used in the dose-response literature.

2 Related work

We survey a collection of the most relevant work to the proposed dose-response model. Much more work has been done on many of the components involved. We refer the reader to Bhadra et al. [2] for a more complete survey on horseshoe shrinkage in complex models. For an overview of trend filtering, see Tibshirani [23]; see Faulkner and Minin [7] for a Bayesian extension.

2.1 Bayesian factor modeling

Many models have been developed for Bayesian factor analysis with smooth structure. Zhang and Paisley [29] apply a group lasso penalty to the rows and columns of a matrix then derive a variational EM algorithm for inference. Hahn et al. [9] use horseshoe priors for sparse Bayesian factor analysis in causal inference scenarios with many instrumental variables. Kowal et al. [11] develop a time series factor model using a Bayesian trend filtering prior on top of a linear dynamical system with Pólya–Gamma augmentation for binomial observations. Schein et al. [20] develop Poisson-Gamma dynamical systems (PGDS), a dynamic matrix factorization model specifically for Poisson-distributed observations; we compare BTF with a tensor extension of PGDS in Section 6. Unlike the above models, BTF is likelihood-agnostic through GASS inference and enables modeling of independently-evolving columns rather than a common time dimension.

2.2 Dose-response modeling

Different variants on the dose-response task have motivated recent work in statistics. Most of these methods consider the case of learning unconstrained responses as a function of molecular features of the drugs and samples. For instance, Low-Kam et al. [14] proposed a Bayesian regression tree model for estimating toxicity at different levels of certain nanoparticles. Wheeler [27] modeled dose-response with molecular descriptors via additive Bayesian splines. Lin and Dunson [13] consider a posterior projection approach for the shape-constrained case, where effects are known to be a monotonic function of dose; we compare to posterior projections in section 6.

In multi-sample, multi-drug studies, the dose-response modeling task is a relational one. Tumors with similar molecular profiles will respond similarly and drugs that target the same pathway will effect tumors similarly. The state of the art in bioinformatics for relational dose-response modeling is a logistic factor

model [25]. This model is fast, but places strong parametric assumptions on the shape of the response curve, and fails to provide any uncertainty quantification. The BTF dose-response model is designed for pilot studies, namely tumor organoids [5], where experiments take weeks or months. In these studies, better modeling with uncertainty quantification is preferable over fast computation. We compare to a logistic factor model in section 6.

3 Bayesian tensor filtering for functional matrix factorization

Let $Y \in \mathbb{R}^{N \times M \times T \times R}$ be an $N \times M$ matrix of noisy functions evaluated at T points, with each point observed R times. The goal in functional matrix factorization is to leverage the relational structure between entries to denoise the observations and predict missing functions. We develop Bayesian tensor filtering (BTF), a hierarchical model of functional matrices. Since our main application is dose-response modeling, we describe BTF in terms of Y being a matrix of N biological samples tested against M drugs, each at T doses with R replicates.

3.1 Latent attributes for biological samples

Biological samples in a study will share molecular attributes. In cancer, different tumor samples will contain similar patterns of genomic mutations, copy number alterations, and gene expression [26]. In mixed tissue experiments, cells that have differentiated into the same type will often respond similarly [e.g., 10]. These attributes are captured in BTF with a latent vector, $w_i \in \mathbb{R}^D$ for the i^{th} sample, as in standard matrix factorization,

$$w_i \sim \text{MVN}(\mathbf{0}, \sigma^2 I)\,, \qquad \sigma^{-2} \sim \text{Gamma}(0.1, 0.1)\,. \tag{1}$$

The choice of the embedding size, D, is a hyperparameter.

3.2 Latent dose-specific attributes for drugs

For the j^{th} column in the functional matrix, BTF models an entire curve $V_j \in \mathbb{R}^{T \times D}$. Intuitively, we expect the effects to mostly vary smoothly with dose. In BTF, this translates to the prior belief that V_{jt} and $V_{j(t+1)}$ should be similar. To encode this, we place (improper) priors on the differences between dose-specific drug embeddings, rather than on the embeddings themselves,

$$(\Delta^{(k)} V_j)_\ell \propto \quad \text{MVN}(\mathbf{0}, \rho^2 \tau_{j\ell}^2 I)\,. \tag{2}$$

We call $\Delta^{(k)} \in \mathbb{R}^{L \times T}$ the composite trend filtering matrix; it contains all $(0, \ldots, k)$ trend filtering [23] matrices. The ordinary trend filtering matrix encodes only the $(k+1)^{\text{th}}$-order differences, implicitly assuming all lower-order differences are not smooth. The composite trend filtering matrix encodes all $(q+1)^{\text{th}}$-order

differences for $q = 1, \ldots, k$. For example, the $k = 1$ case yields a prior on the first and second order differences,

$$\Delta^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & \ldots & 0 & 0 & 0 \\ & & & & \ldots & & & \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -1 \\ 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 & 0 \\ & & & & \ldots & & & \\ 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 \end{bmatrix} . \tag{3}$$

The first line of eq. (3) places an independent prior on the first dose level in each drug, v_{j1}, to make the matrix $(\Delta^\top \mathcal{T} \Delta)$ non-singular, where $\mathcal{T} = \text{diag}(1/(\rho^2 \tau_j^2))$. This ensures the resulting prior on V_j is proper; see the supplementary material for details.

Column independence distinguishes functional matrix factorization from time-series tensor factorization [28, 21, 8, 22] where all columns are progressing through time together. In BTF, columns are evolving independently, though potentially with similar latent attributes. This independent evolution captures the notion that two drugs treated at the same concentration may have totally different effects due to the molecular size of the drug, its targeting receptor, and its chemical structure.

3.3 Global-local shrinkage priors

The variance parameters in eq. (2) control the smoothness of each curve. Small values of ρ^2 and $\tau_{j\ell}^2$ will shrink the $(j, \ell)^{\text{th}}$ difference to nearly zero, resulting in the curve being smoother; larger values enable the curve to jump in response to the data. BTF uses a global-local shrinkage model [18] where ρ^2 controls the smoothness of the entire matrix and $\tau_{j\ell}^2$ is a local shrinkage term for a specific drug at a specific dose. BTF places a horseshoe+ (HS+) prior [3, 1] on the shrinkage parameters,

$$\tau_{j\ell} \sim \text{C}^+(0, \phi_{j\ell}) \qquad \phi_{j\ell} \sim \text{C}^+(0, 1) \qquad \rho \sim P(\rho)\,. \tag{4}$$

The HS+ prior is asymptotic at zero and consequently shrinks most $\tau_{j\ell}$ values to nearly zero. However, it has heavy tails that decay very slowly. Thus, when the data suggests that a change in dose results in a sharp change in effect, the drug attributes are able to make sharp jumps. As noted by Bhadra et al. [1], a full Bayesian specification could choose a reasonable prior for ρ, such as a standard Cauchy or Uniform$(0, 1)$. If an estimate of the number of non-zero entries is available, Van Der Pas et al. [24] make an asymptotic argument for setting $\hat{\rho}$ to the expected number of non-zeros. In practice, we find BTF is robust to the choice of global shrinkage parameter and instead perform a grid search over a handful of ρ values.

3.4 Posterior inference

Inference in BTF is performed through an efficient Gibbs sampler. The updates for the latent attributes depend on the form of $P(y_{ijtr}; w_i^\top v_{jt})$, the likelihood function for replicate r of sample i, treated with drug j, at dose t. Specifically, likelihoods fall into three categories: (i) Gaussian, for which updates are fully conjugate; (ii) binomial and related likelihoods, for which updates are conditionally conjugate through Pólya–Gamma augmentation; and (iii) black-box likelihoods, for which updates are non-conjugate. The derivations for the Gaussian and binomial categories, as well as the horseshoe parameter updates, are in the supplementary material. In the remainder of the paper, we focus on inference for non-conjugate likelihoods, as this is the category of observations for the Bayesian dose-response model in Section 5.

4 Generalized analytic slice sampling for black-box likelihoods

Posterior inference in BTF is done via Gibbs sampling. However, this requires us to be able sample from the conditional distributions for each of the parameters. For generic likelihoods, the conditional distributions of the latent attributes w_i and v_{jt} are typically not available in closed form. Moreover, the likelihood may impose hard constraints on the values of matrix entries. For instance, in Poisson factorization, the $w_i^\top v_{jt}$ corresponds to the Poisson rate of observed entries in the tensor, which must always be positive. In other cases, such as the dose-response model in Section 5, the inner product may parameterize a probability or percentile, requiring $w_i^\top v_{jt} \in [0, 1]$. A naive MCMC-within-Gibbs step with rejection sampling for invalid proposals will have a high rejection rate and lead to poor mixing. In this section, we present an exact approach for generic likelihoods that handles arbitrary linear constraints.

Sampling from the conditional distributions of the latent attributes can be reduced to the problem of sampling from the posterior of a vector x with a multivariate normal prior constrained by a set of linear inequalities,

$$x \sim P(y; x)\mathrm{MVN}(x; \mu, \Sigma)\mathbb{I}[Dx \geq \gamma]\,. \tag{5}$$

Slice sampling [17] samples from $P(x|y)$ by sampling over the augmented joint distribution $P(x, \epsilon \mid y)$ where $\epsilon = P(y \mid x)$. When the prior is an unconstrained multivariate normal, the augmentation can be done by noting that a multivariate normal forms an ellipse of equal probability. Elliptical slice sampling [16] samples from the posterior on x by sampling a candidate ellipse v from the prior and sampling an angle $\theta \in [-\pi, \pi]$ such that $x' = x\cos(\theta) + v\sin(\theta)$ and $P(y; x') \geq P(y; x) \times u\,, u \sim U(0, 1)$. Adding constraints as in eq. (5) could be handled by pushing the constraints into the likelihood, but would result in high rejection rates.

We extend elliptical slice sampling to directly handle constrained multivariate normal priors. Our approach is a generalization of the analytic slice sampling

Algorithm 1: Generalized analytic slice sampling (GASS) for constrained MVN priors

Data: Valid current point x, mean μ, covariance Σ, log-likelihood $\mathcal{L}$, constraints (D, γ)
Result: MCMC sample from $P(x') \propto \exp(\mathcal{L}(x'))\text{MVN}(x'; \mu, \Sigma)\mathbb{I}[Dx' \geq \gamma]$
$t = \mathcal{L}(x) + \log \epsilon, \qquad \epsilon \sim U(0, 1)$;
Sample proposal $v \sim \text{MVN}(v; \mathbf{0}, \Sigma)$;
Grid approximation $\mathcal{G} = \text{grid}(-\pi, \pi)$;
foreach *constraint* $(d_i, \gamma_i) \in (D, \gamma)$ **do**
  $a = d_i^\top (x - \mu)$, $b = d_i^\top v$, $c = \gamma_i - d_i^\top \mu$;
  if $a^2 + b^2 - c^2 \geq 0$ *and* $a \neq -c$ **then**
    Get θ_1, θ_2 as in eq. (6);
    if $a^2 > c^2$ **then**
      $\mathcal{G} = \mathcal{G} \bigcap [\theta_1, \theta_2]$;
    else
      $\mathcal{G} = \mathcal{G} \bigcap ([-\pi, \theta_1] \bigcup [\theta_2, \pi])$;
    end
  end
end
Generate candidate samples $\mathcal{X} = \{x' \colon x\cos(\theta_g) + v\sin(\theta_g) + \mu, \theta_g \in \mathcal{G}\}$;
Select uniformly from sufficiently likely candidates $\{x' \colon \mathcal{L}(x') \geq t, x' \in \mathcal{X}\}$.

procedure of Fagan et al. [6] for truncated multivariate normals. The key difference is that the original analytic slice sampler only considered centered truncated multivariate normals with no likelihood component. Generalizing this procedure to handle the more general case in eq. (5) introduces several edge cases.

4.1 Algorithm

The full GASS procedure is presented in Algorithm 1. The idea of GASS is to note that the constraints can be pushed inside the proposal update. Given a single constraint requiring that the output point satisfies $d^\top x' \geq \gamma$, a valid angle θ must satisfy $a\cos\theta + b\sin\theta - c \geq 0$, where $a = d^\top(x - \mu)$, $b = d^\top(v - \mu)$, and $c = \gamma - d^\top \mu$. Basic trigonometry implies that the feasible range of θ is a subset of $[-\pi, \pi]$ whose boundary points are

$$\theta_1, \theta_2 \;=\; 2\text{arctan}\left(\frac{b \pm \sqrt{a^2 + b^2 - c^2}}{a + c}\right). \tag{6}$$

Two cases cause the entire ellipse to be valid: (i) $(a^2 + b^2 - c^2) < 0$ and (ii) $a = -c$. In the first case, $a^2 + b^2 < c^2 \Rightarrow a\cos\theta + b\sin\theta > c$, for all θ. In the second case, the only place the constraint touches the ellipse is on the extremal

point of the ellipse and thus its selection has probability zero. For all other cases, the subset is determined based on the sign of $a^2 - c^2$. A positive sign indicates the quadratic in the inequality is concave and eq. (6) defines the boundaries of a contiguous region; a negative sign indicates convexity and thus the complement of the interval. As the output sample may need to satisfy many such constraints, we can simply repeat the above process to find all the valid regions and take their intersection. We then numerically approximate the valid θ regions with a fine-grained 1D grid. Sampling is performed in a quasi-Monte Carlo fashion, uniformly over the valid grid points.

4.2 Conditioning heuristic

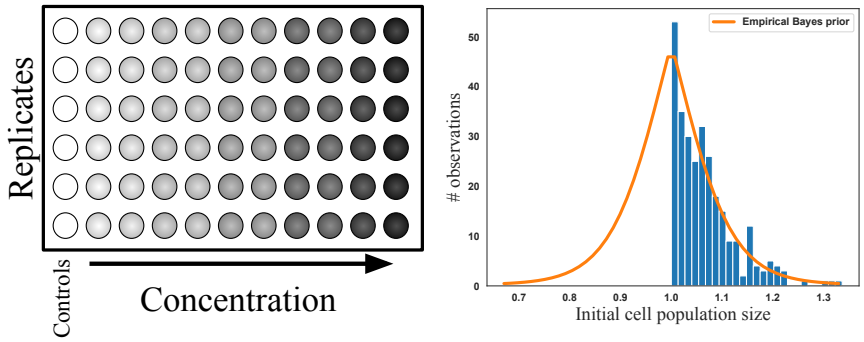
Elliptical slice sampling schemes like GASS can suffer from poor mixing when the likelihood overwhelms the prior. In this case, the angle of the ellipse will be very sharp, causing the sampler to have a small region of the posterior that it can jump to with non-negligible probability at each step. Fagan et al. [6] suggest using an expectation propagation for generic truncated normals. In the case of BTF, the prior parameters for W are a function of V, and vice versa. Thus, expectation propagation would need to be performed every iteration of the Gibbs sampler, which would considerably increase the computational cost of inference.

We instead approximate the entire functional matrix once at the start, by a constrained matrix factorization. This is fast as it only requires alternating between solving linear programs for the rows and columns. After fitting the rows and columns, we calculate an over-estimate of the variance, analogous to an EP approximation, as a multiple of the empirical squared error in the estimate for each column and row. BTF uses the pseudo-EP approximation at every step in the Gibbs sampler to calculate an adjusted prior, following the same updates as in the Gaussian likelihood case (see the supplementary material for details). The log-likelihood used in the GASS procedure is then the original log-likelihood minus the log-pseudo-EP likelihood, leaving the resulting distribution equivalent but increasing the range of admissible angles θ.

5 Bayesian dose-response model for cancer drug studies

The BTF model is a general framework for functional matrix factorization. To apply this framework to dose-response data requires an extra set of modeling steps on top of BTF. Here we describe the cancer drug experiments in detail and an empirical Bayes procedure to estimate the observation likelihood in the face of technical error.

We describe the dose-response model in the context of the specific protocol used for an internal study at Columbia University Medical Center. The second study we analyze uses a different protocol with a different number of plates, wells per plate, concentrations, replications, and other experimental designs.



(a) Experimental design

(b) Empirical Bayes estimate

Figure 2: Left: The layout of each microwell plate experiment used to generate a single functional matrix entry. Cells are pipetted one column at a time, leading to correlated errors. Right: Estimate of the prior distribution of mean cell counts in each column, relative to the control column mean. The prior is estimated empirically assuming the lowest concentration had no effect if it had a higher mean.

These distinctions are changes to the dimensions of the resulting tensor, but the fundamental statistical inference technique remains the same. We use the same method to analyze and compare both datasets in section 6.

5.1 Experimental design and technical error

Each functional matrix entry is derived from a microwell plate experiment where a drug is tested against a biological sample. Each experiment measures cell counts after applying the drug at 9 different concentrations. The cell counts are measured relative to a baseline control population where no drug was applied. For the control and each concentration level, 6 replicates are tested. Figure 2a shows the design of each 60-well plate experiment. All experiments are normalized by dividing the population size estimates at each concentration by the control mean for the plate.

The first step in each experiment is to pipette an initial population of cells into each of the 60 microwells on the plate. This is a time consuming process for the biologist, often taking hours to pipette a single plate. To speed up the plating process, biologists use a multi-headed pipette that enables them to simultaneously fill an entire column of each plate. This reduces the burden on the biologist, but comes at a cost: correlated errors.

When a biologist fills a microwell, they first draw a pool of cells into the pipette. Given the small volumes involved in laboratory experiments, the actual number of cells drawn can vary substantially on a relative basis. Using a multi-headed pipette transforms this variation into a hierarchical model: first a pool of cells is drawn into the pipette, then it is split among all the heads. The majority

of the variation comes in the initial sampling, with small noise added in the splitting process. This has the unintended side effect of creating correlated errors between all microwells in a single column.

5.2 Empirical Bayes likelihood estimation

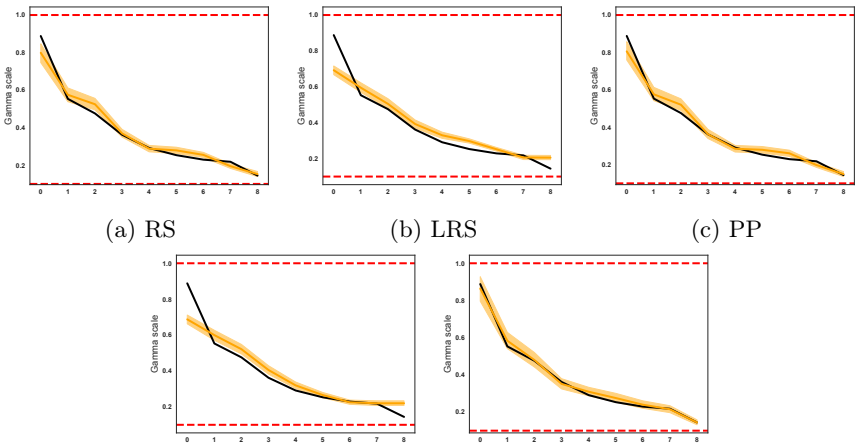
The correlated errors in the columns render the exact effects unidentifiable. Each column has two latent variables affecting the final population size of cells: a dose-level effect from the drug and an initial population size from the pipetting. Since both of these variables affect all replicates in a column, disentangling them precisely is impossible. Nevertheless, an estimate of dose-response must be provided.

We take an empirical Bayes approach to disentangling the variation in drug effects from the technical error in pipetting. In most experiments, the lowest concentration tested is too small to have any effect on cell survival. We therefore make the assumption that any experiment where the mean of the control replicates is lower than the mean of the replicates treated at the lowest concentration has effectively two sets of control columns. This enables estimation of the variation between means and form an empirical Bayes prior for the pipetting error.

Specifically, we form a histogram of all lowest-concentration means greater than the control mean on the same plate. We then fit a Poisson GLM with 3 degrees of freedom to the histogram to estimate the prior probability that the mean of the initial population of cells was higher than the control mean. We then assume the true distribution is symmetric and obtain an empirical Bayes prior on the means. Figure 2b shows an example histogram and empirical Bayes prior estimate. The within-column variance is identifiable and estimated using the controls. The empirical Bayes likelihood is then a gamma mixture model that integrates out the uncertainty in the initial population mean,

$$P(y_{ijt} \mid w_i^\top v_{jt}) = \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \hat{m}_k Ga(y_{ijtr}; \hat{a}_k, \hat{b}_k w_i^\top v_{jt}) \right) \mathbb{1}[0 \leq w_i^\top v_{jt} \leq 1], \quad (7)$$

where $(\hat{m}, \hat{a}, \hat{b})_k$ are the weights derived from the empirical Bayes procedure. The scale regression form of the inner product is due to the property that the gamma random variable is being multiplied by the effect of the drug. That is, the population of cells is being killed at some latent rate. The inner product is constrained to be a proportion, as the drugs are known not to help any cells grow (i.e., the proportion must be at most 1) and a drug cannot kill more than all of the cells. The likelihood in eq. (7) is non-conjugate to the BTF hierarchical model and thus we use the generalized elliptical slice sampling routine from Section 4 for inference.



(a) RS (b) LRS (c) PP

(d) LPP (e) GASS

Figure 3: Sample fits for different methods on the gamma scale estimation benchmark; black is the true scale, orange is the estimated scale, bands are 90% credible intervals, dashed lines are constraint boundaries. GASS captures the shape of the curve and has good coverage after $10K$ Gibbs iterations.

6 Results

We study the performance of the proposed dose-response model and its components, BTF and GASS. We first benchmark GASS against different alternative methods for nonconjugate inference, where GASS mixes faster and has lower error. Then we study BTF on a dynamic matrix factorization problem with non-conjugate Poisson observations; BTF outperforms a recent Bayesian tensor decomposition approach designed for functional count matrices. Finally, we apply the dose-response model to a real cancer drug study. We run 5 independent trials, holding out a different subset of entire dose-response curves and report averages over all trials; the BTF-based dose-response model outperforms all baselines in terms of log probability on held out data.

MSE ($\times 10^3$)

Sampler	$m = 100$	$m = 500$	$m = 1000$	$m = 5000$	$m = 10000$
RS	1.29 ± 0.09	1.27 ± 0.08	1.25 ± 0.08	1.23 ± 0.08	1.24 ± 0.08
LRS	5.18 ± 0.24	5.05 ± 0.22	5.03 ± 0.22	5.03 ± 0.22	5.05 ± 0.22
PP	1.16 ± 0.07	1.08 ± 0.07	1.07 ± 0.07	1.06 ± 0.07	1.06 ± 0.07
LPP	5.03 ± 0.22	5.05 ± 0.22	5.03 ± 0.22	5.04 ± 0.22	5.04 ± 0.22
GASS	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.55 \pm 0.03}$	$\mathbf{0.52 \pm 0.03}$	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.47 \pm 0.03}$

90% Credible Interval Coverage

Sampler	$m = 100$	$m = 500$	$m = 1000$	$m = 5000$	$m = 10000$
RS	0.49 ± 0.02	0.56 ± 0.02	0.58 ± 0.02	0.60 ± 0.02	0.60 ± 0.02
LRS	0.19 ± 0.01	0.25 ± 0.02	0.27 ± 0.02	0.27 ± 0.02	0.28 ± 0.02
PP	0.58 ± 0.02	0.64 ± 0.02	0.65 ± 0.02	0.67 ± 0.02	0.66 ± 0.02
LPP	0.24 ± 0.01	0.29 ± 0.02	0.30 ± 0.01	0.31 ± 0.01	0.30 ± 0.01
GASS	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$

Table 1: Benchmark performance for GASS versus alternative non-conjugate elliptical sampling approaches. Results are averages over 100 independent trials $\pm$ standard error.

6.1 GASS benchmarks

We benchmark GASS on a simulation study with a constrained multivariate normal prior with a non-conjugate gamma scale likelihood,

$$\begin{aligned} y_i^{(r)} &\sim \text{Gamma}(y_i; a, \theta_i) \\ \boldsymbol{\theta} &\sim \text{MVN}(\boldsymbol{\theta}; \boldsymbol{\mu}, \Sigma) \mathbb{1}[0.1 \leq \boldsymbol{\theta} \leq 1] \prod_{i=1}^{n-1} \mathbb{1}[\theta_i \geq \theta_{i+1}] \\ \boldsymbol{\mu} &= [0.95, 0.8, 0.75, 0.5, 0.29, 0.2, 0.17, 0.15, 0.15] \\ \Sigma_{ij} &= \tau \exp(-\frac{1}{2b}(i-j)^2)\,. \end{aligned} \tag{8}$$

The covariance matrix in the unconstrained prior corresponds to a squared exponential kernel. We set the hyperparameters $a = 100$, $\tau = 0.1$, and $b = 3$; all hyperparameters are assumed known. We use $R = 3$ replicates for $\mathbf{y}$. We compare GASS against four different variants of elliptical slice sampling (ESS),

- **Rejection sampling (RS).** Samples are drawn using the unconstrained ESS model, with the constraints pushed into the likelihood. Any violated constraint generates a zero probability and corresponds to a rejection sampler.
- **Logistic rejection sampling (LRS).** The ESS is used to model logits, which are then passed through the logistic transform to satisfy the $[0, 1]$ constraint; rejection sampling again handles the monotonicity constraint.

Poisson Dynamical System			
	Observations	True rate	
Model	NLL	MAE	RMSE
NMF	437.32 ± 31.73	1.46 ± 0.32	2.26 ± 0.57
PGDS	396.98 ± 11.86	1.24 ± 0.22	1.98 ± 0.40
BTF	$\mathbf{369.91 \pm 7.66}$	$\mathbf{0.87 \pm 0.18}$	$\mathbf{1.24 \pm 0.28}$

Table 2: Mean results ± standard error on held out data in the dynamical system benchmark; smaller is better for all metrics. NMF: nonnegative matrix factorization; PGDS: Poisson-gamma dynamical system; BTF: Bayesian tensor filtering; NLL: negative log-likelihood; MAE: mean absolute deviation from truth; RMSE: root mean-squared error from truth.

- **Posterior projections (PP).** No constraints are imposed on the model during posterior inference. Instead, we use the posterior projection approach of Lin and Dunson [13] to post-hoc enforce the constraints.
- **Logistic posterior projections (LPP).** A hybrid of the previous two approaches combined: modeling the logits for $[0, 1]$ constraints but projecting the posterior samples to the monotone surface.

We compare performance with $2m$ MCMC steps, where the first m are a burn-in phase and the last m are used for posterior approximation; we consider $m = [100, 500, 1000, 5000, 10000]$. Performance is measured in terms of mean squared error (MSE) and coverage rate of the 90% credible intervals for every θ_i point. Results are averaged over 100 independent trials with $(\boldsymbol{\theta}, \mathbf{y})$ resampled from eq. (8) at the start of each trial.

Figure 3 shows examples of the fits for each method, with 90% credible intervals. GASS is the only procedure that results in good covarate of the true mean and captures the shape of the overall curve; the other methods tend to underestimate the extremal points and overly-smooth the curve.

Table 1 shows the aggregate results of the benchmarks. GASS outperforms all four comparison methods in terms of both error and coverage. After $m = 100$ samples the MSE for GASS is lower and coverage is higher than any of the other strategies after $m = 10000$ samples. Further, the model appears to have fully mixed after 5000 samples, with the coverage rate close to exactly 90%.

6.2 Non-stationary Poisson dynamical systems

We benchmark BTF on a synthetic functional Poisson matrix dataset where the observations are Poisson distributed with a latent rate curve for each function. The rate at every point in the curve is the inner product of two gamma random

Cancer Drug Studies

	Pilot Study	Landscape Study
Model	NLL	NLL
NMF	262.75 ± 308.12	25573.14
LMF	589.17 ± 582.29	$> 10^6$
BTF	$\mathbf{-80.22 \pm 9.67}$	$\mathbf{-3268.11}$

Table 3: Left: mean results ± standard error on held out data for the pilot cancer drug studies. Right: Results on a single test set of 1000 curves for the landscape study. NMF: nonnegative matrix factorization; LFM: logistic factor model; BTF: Bayesian tensor filtering; NLL: negative log-likelihood.

vectors,

$$h_{j\ell} \sim \text{Bern}(0.2)\,, \quad u_{j\ell d} \sim (1 - h_{j\ell})\delta_0 + h_{j\ell}\text{Ga}(1,1)\,, \quad v_{jtd} = \sum_{\ell=1}^{t} u_{j\ell d}\,,$$
$$w_{id} \sim \text{Ga}(1,1)\,, \qquad y_{ijt} \sim \text{Pois}(\langle w_i, v_{jt}\rangle)\,.$$

The resulting true rates form a monotonic curve of constant plateaus with occasional jumps. As in the dose-response data, the columns evolve independently of each other, rather than through a common time parameter. We set the latent factor dimension to 3.

We compare BTF to nonnegative matrix factorization (NMF) and the Poisson-gamma dynamical system (PGDS) model of Schein et al. [20]. We use the default parameters for PGDS; for BTF, we set $\rho^2 = 0.1$; both models use the true factor dimension 3. We run both BTF and PGDS for 2000 burn-in iterations and collect 2000 samples on an $11 \times 12 \times 20$ tensor with the upper left $3 \times 3 \times 20$ corner held out. We conduct 5 independent trials, regenerating new data each time and evaluating the models on the held out data. We measure performance in three metrics: mean absolute error (MAE) on the true rate, root mean squared error (RMSE) on the true rate, and negative log-likelihood (NLL) on held out observations. Table 2 presents the results.

The PGDS model outperforms the NMF baseline, and BTF outperforms both methods. There are two possible reasons for the better performance of BTF relative to PGDS. First, the PGDS model uses a common "time" factor for all columns, but in this simulation columns evolve independently. Second, the large discrete jumps are not well-modeled by PGDS. In follow-up experiments, we found no improvement for PGDS from using larger factor sizes to potentially account for the first issue. This suggests the local adaptivity of BTF accounts for the better performance.

6.3 Cancer drug study

We evaluate the proposed empirical Bayes dose-response model, built on top of BTF, on two cancer drug studies. First, we use a small internal pilot study conducted at Columbia University Medical Center. The pilot study tested 35 drugs against 28 tumor organoids, each at 9 different concentrations with 6 replicates. Second, we run on a large-scale, "landscape" study [12] that tested 67 drugs against 284 tumor organoids, each at 7 different concentrations with 2 replicates. For the pilot study, we run 5 independent trials, holding out 30 curves at random, subject to the constraint that no column or row is left without any observations in the training set. For the landscape study, we hold out a single test set of 1000 curves ($\approx 5\%$ of the total entries). Since this is real data, MAE and RMSE from the truth are not available; we measure performance solely in terms of negative log-likelihood on the held out data.

The standard dose-response modeling approach in cancer datasets is a log-linear logistic model [25]. For a baseline, we extend that model to a logistic factor model (LFM), using the same preprocessing strategy. We also compare to NMF as a second baseline. To ensure the monotonicity, we project the NMF results to be monotone curves using the PAV algorithm as in Lin and Dunson [13]. We choose the factor size in both models by 5-fold cross-validation on the training set.

For BTF, we perform a grid search over hyperparameters: $\rho^2 = \{0.001, 0.01, 0.1\}$, factor size $D = \{1, 3, 5, 8\}$, and the order of the trend filtering matrix $k = \{0, 1\}$; we select the best model using the deviance information criterion [4]. We evaluate the BTF model using the average of the posterior draws, rather than the full Bayes estimate; this enables us to fairly compare with the NMF and LFM point estimates of the latent mean. We run 10000 Gibbs sampling steps in both studies, discarding the first 5000 as burn-in.

Table 3 present the results. The BTF dose-response model outperforms both baselines in terms of negative log-likelihood of the held out data in both studies. Furthermore, the BTF procedure is also more stable in the pilot study cross-validation, with a much lower reconstruction variance than either baseline. This suggests BTF not only forms a more accurate basis for a dose-response model, but is also more reliable.

Qualitative results on the held out predictions are in fig. 1 (orange). All 9 plots are for real data from the landscape study, with the gray observations held out. The orange line shows the posterior mean of the predicted curves. The curves have all of the desired properties: monotonicity with dose, bounded between zero and one, mostly smooth, locally adaptive to sharp jumps in the data, and highly predictive of the outcomes of the experiments. The orange bands show the 50% approximate posterior credible intervals using the empirical Bayes likelihood model. The credible intervals are conservative, estimating a larger variance than is actually observed in the outcomes. Even still, the NMF and LMF models far exceed these bands in certain points in the curve. This is due to the heteroskedastic nature of the likelihood and the misspecification of the NMF and LMF loss functions. Both competing models optimize for squared

error, effectively making a heteroskedastic assumption on the model. In RMSE terms, all three models perform nearly identically, within ± 0.01 of each other on both datasets. Judging the models by RMSE would be misleading, since the high degree of noise in the first three dose levels dominates the overall loss and obscures the real fit of the model.

7 Discussion

Multi-sample, multi-drug cancer studies are time and resource intensive. The outcomes from these studies are noisy, often incomplete, observations of biological responses to candidate therapies. Denoising observations and imputing missing experiments is an important step in the scientific analysis and drug discovery pipeline. The Bayesian tensor filtering model we presented in this paper enables scientists to flexibly model dose-response curves with consideration for measurement error and biological constraints on the shape of the curve. While the BTF model is an improvement over the state of the art, we believe there are several improvements that could be made.

In many studies, side information about both the samples and the drugs are available. Cell lines and organoids are often sequenced to identify genomic mutations, gene expression levels, DNA methylation, and other molecular profiling information. The chemical structure of the candidate drugs are often known and in some cases the mechanisms of action have been identified. Utilizing this side information could help to improve denoising and also reveal important drivers of therepeutic response.

Finally, the current BTF model is computationally intensive. For small scale studies like our pilot study, the model runs in a few hours on a laptop. The landscape study required several days on a compute cluster to perform the hyperparameter search. Relative to the years required for the landscape experiments, the run time is negligible. Nevertheless, offering an alternative inference approach that can scale more efficiently, such as variational inference, may make the BTF model useful for a broader group of scientists.

References

[1] Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017.

[2] Anindya Bhadra, Jyotishka Datta, Yunfan Li, and Nicholas G Polson. Horseshoe regularization for machine learning in complex and deep models. *arXiv preprint arXiv:1904.10939*, 2019.

[3] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[4] Gilles Celeux, Florence Forbes, Christian P. Robert, and D. Michael Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4): 651–673, 2006.

[5] Jarno Drost and Hans Clevers. Organoids in cancer research. *Nature Reviews Cancer*, 2018.

[6] Francois Fagan, Jalaj Bhandari, and John Cunningham. Elliptical slice sampling with expectation propagation. In *Uncertainty in Artificial Intelligence*, 2016.

[7] James R. Faulkner and Vladimir N. Minin. Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis*, 13(1):225, 2018.

[8] Laetitia Gauvin, André Panisson, and Ciro Cattuto. Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach. *PLoS one*, 9(1), 2014.

[9] P. Richard Hahn, Jingyu He, and Hedibert Lopes. Bayesian factor model shrinkage for linear IV regression with many instruments. *Journal of Business & Economic Statistics*, 36(2):278–287, 2018.

[10] Lei Huang, Shengnan Wu, and Da Xing. High fluence low-power laser irradiation induces apoptosis via inactivation of Akt/GSK3β signaling pathway. *Journal of Cellular Physiology*, 226(3):588–601, 2011.

[11] D. R. Kowal, D. S. Matteson, , and D. Ruppert. Dynamic shrinkage processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

[12] Jin-Ku Lee, Zhaoqi Liu, Jason K. Sa, Sang Shin, Jiguang Wang, Mykola Bordyuh, Hee Jin Cho, Oliver Elliott, Timothy Chu, Seung Won Choi, Daniel I. S. Rosenbloom, In-Hee Lee, Yong Jae Shin, Hyun Ju Kang, Donggeon Kim, Sun Young Kim, Moon-Hee Sim, Jusun Kim, Taehyang Lee, Yun Jee Seo, Hyemi Shin, Mijeong Lee, Sung Heon Kim, Yong-Jun Kwon, Jeong-Woo Oh, Minsuk Song, Misuk Kim, Doo-Sik Kong, Jung Won Choi, Ho Jun Seol, Jung-Il Lee, Seung Tae Kim, Joon Oh Park, Kyoung-Mee Kim, Sang-Yong Song, Jeong-Won Lee, Hee-Cheol Kim, Jeong Eon Lee, Min Gew Choi, Sung Wook Seo, Young Mog Shim, Jae Ill Zo, Byong Chang Jeong, Yeup Yoon, Gyu Ha Ryu, Nayoung K. D. Kim, Joon Seol Bae, Woong-Yang Park, Jeongwu Lee, Roel G. W. Verhaak, Antonio Iavarone, Jeeyun Lee, Raul Rabadan, and Do-Hyun Nam. Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. *Nature Genetics*, 50(10):1399–1411, 2018.

[13] Lizhen Lin and David B Dunson. Bayesian monotone regression using Gaussian process projection. *Biometrika*, 101(2):303–317, 2014.

[14] Cecile Low-Kam, Donatello Telesca, Zhaoxia Ji, Haiyuan Zhang, Tian Xia, Jeffrey I Zink, and Andre E Nel. A Bayesian regression tree approach to identify the effect of nanoparticles' properties on toxicity profiles. *The Annals of Applied Statistics*, 9(1):383–401, 2015.

[15] Enes Makalic and Daniel F. Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2015.

[16] Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *Artificial Intelligence and Statistics*, 2010.

[17] Radford M Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

[18] Nicholas G. Polson and James G. Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.

[19] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

[20] Aaron Schein, Hanna Wallach, and Mingyuan Zhou. Poisson-gamma dynamical systems. In *Advances in Neural Information Processing Systems*, 2016.

[21] Stephan Spiegel, Jan Clausen, Sahin Albayrak, and Jérôme Kunegis. Link prediction on evolving data using tensor factorization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2011.

[22] Koh Takeuchi, Hisashi Kashima, and Naonori Ueda. Autoregressive tensor factorization for spatio-temporal predictions. In *International Conference on Data Mining*, 2017.

[23] Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

[24] Stéphanie L. Van Der Pas, Bas J.K. Kleijn, and Aad W. Van Der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618, 2014.

[25] Daniel J. Vis, Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew J. Garnett, and Lodewyk FA Wessels. Multilevel models improve precision and speed of IC50 estimates. *Pharmacogenomics*, 17(7):691–700, 2016.

[26] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart, and Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113, 2013.

[27] Matthew W Wheeler. Bayesian additive adaptive basis tensor product models for modeling high dimensional surfaces: an application to high-throughput toxicity testing. *Biometrics*, 75(1):193–201, 2019.

[28] Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, and Jaime G. Carbonell. Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In *International Conference on Data Mining*, 2010.

[29] Aonan Zhang and John Paisley. Deep Bayesian nonparametric tracking. In *International Conference on Machine Learning*, pages 5828–5836, 2018.

A Gaussian likelihood

When the likelihood is normal, $y_{ijt} \sim \mathcal{N}(w_i^\top v_{jt}, \nu^2)$, where ν^2 is a nuisance parameter, the factor and loading updates are conjugate. Let $\tilde{V} = (v_{1,1}, v_{1,2}, \ldots, v_{1,T}, v_{2,1}, \ldots, v_{M,T})$, and $\Omega^{-1} = \text{diag}\{1/\nu^2\}$, then the updates are multivariate normal,

$$
\begin{aligned}
Q^{(i)} &= (\tilde{V}^\top \Omega^{-1} \tilde{V} + \text{diag}(\sigma^{-2}))^{-1} \\
(w_i \mid -) &\sim \text{MVN}\left(Q^{(i)} \tilde{V}^\top \Omega^{-1} \text{vec}(Y_i^\top), Q^{(i)}\right) \\
\mathcal{T}^{(j)} &= \text{diag}(1/(\rho^2 \tau_j^2)) \\
\Sigma^{(j)} &= (I_D \otimes \Delta^\top \mathcal{T} \Delta) + (W \otimes I_T)^\top \Omega^{-1} (W \otimes I_T) \\
(\text{vec}(V_j) \mid -) &\sim \text{MVN}(\Sigma^{(j)} (W \otimes I_T) \Omega^{-1} \text{vec}(Y_{\cdot j}^\top), \Sigma^{(j)}) \,,
\end{aligned}
\tag{9}
$$

where `diag` diagonalizes the given vector, `vec` is the vectorization operator, and $\otimes$ is the Kronecker product. In both the w_i and V_j updates the precision matrices will be sparse, making sampling from the conditionals computationally tractable.

B Binomial and related likelihoods via Pólya–Gamma augmentation

When the likelihood is binomial, $y_{ijt} \sim \mathrm{Bin}(n_{ijt}, 1/\{1 + e^{w_i^\top v_{jt}}\})$, where n_{ijt} is a nuisance parameter, the updates are conditionally conjugate given a Pólya–Gamma (PG) latent variable sample [19],

$$(\psi_{ijt} \mid -) \sim \mathrm{PG}(n_{ijt}, w_i^\top v_{jt}), \qquad (w_i \mid -) \sim N(m_{\psi_i}, \Sigma_{\psi_i}), \tag{10}$$

where $\Sigma_{\psi_i} = (\tilde{V}^\top \Psi_i \tilde{V} + \sigma^{-2} I)^{-1}$, $m_{\psi_i} = \Sigma_{\psi_i} \tilde{V}^\top \kappa$, $\Psi_i = \mathrm{diag}(\psi_{(i,1,1)}, \ldots, \psi_{(i,M,T)})$, and $\kappa = (y_{(i,1,1)} - n_{(i,1,1)}/2, \ldots, y_{(i,M,T)} - n_{i,M,T}/2)$. The updates for V follow analogously. PG augmentation can be applied to binomial, Bernoulli, negative binomial, and multinomial likelihoods, among others.

C Local shrinkage updates

The local shrinkage parameters $\tau_{j\ell}$ can be updated through a double latent variable augmentation trick,

$$\begin{aligned}
(\tau_{j\ell} \mid -) \sim &\quad \text{InvGamma}(D+1, \left|\left|\Delta^{(k)} V_j\right|\right|_2^2 / 2 + 1/c_{j\ell}) \\
(c_{j\ell} \mid -) \sim &\quad \text{InvGamma}(1, 1/\tau_{j\ell}^2 + 1/\phi_{j\ell}) \\
(\phi_{j\ell} \mid -) \sim &\quad \text{InvGamma}(1, 1/c_{j\ell} + 1/\eta_{j\ell}) \\
(\eta_{j\ell} \mid -) \sim &\quad \text{InvGamma}(1, 1/\phi_{j\ell} + 1)\,.
\end{aligned} \tag{11}$$

The updates in eq. (11) come from the HS+ prior being a two-level horseshoe prior. The inverse-gamma latent variable augmentation for the horseshoe is fast and typically mixes quickly [15].