Under Review at ICML 2024 AI for Science Workshop

Anonymous Authors[1]

Abstract

This paper introduces a novel inference scheme for a class of hurdle priors that exploits sparsity to scale large machine learning models with convolution-closed likelihood distributions, such as the Gaussian and Poisson. We call this the convolution-closed hurdle motif, and focus on the non-negative Tucker decomposition, a tool popular in the literature for modeling multi-way relational data. We apply an instance of the class of hurdle priors, the hurdle gamma prior, to a probabilistic non-negative Tucker method and derive an inference scheme that scales with only the non-zero latent parameters in the core tensor. This scheme avoids the typical exponential blowup in computational cost present in Tucker decomposition, efficiently fitting the data to a high-dimensional latent space. We derive and implement a closed-form Gibbs sampler for full posterior inference and fit our model to longitudinal microbiome data. Using this hurdle motif to quickly train our model, we reveal interpretable qualitative structure and encouraging classification results.

1. Introduction

Sparse data, often relational, are frequently stored as matrices or tensors. Practitioners often require methods that scale appropriately to model high-dimensional latent structure, such as regularization approaches (Ishwaran & Rao, 2005; Zou, 2006) and gradient-based methods (Hoffman et al., 2013; Kingma & Ba, 2014; Ranganath et al., 2014). This high-dimensional latent structure can be computationally expensive to model, as generally, computation scales with the number of latent parameters. However, much high-dimensional latent structure is sparse. A canonical setting with this phenomena is that of training neural networks, as the weight matrices of trained neural networks are often high-dimensional and sparse, and with careful manipulation, we may exploit that sparsity for computational benefit (Louizos et al., 2017).

This kind of latent structure is ubiquitous, and we turn our attention to scientifically interesting settings such as longitudinal microbiome data (Ma & Li, 2023; Shi et al., 2023) and dynamic networks (Aguiar et al., 2023). Tensor decomposition methods are natural ways to model such highly structured data without compromising the data structure (Tucker, 1966; Kolda & Bader, 2009). Much of the literature assumes low-rank structure in these sparse, high-dimensional data. However, recent work calls for *also* modeling the latent structure as sparse and high-dimensional (Hood & Schein, 2024). Guided by established schemes for modeling sparsity, we aim to develop computationally scalable probabilistic generative models for large-scale scientific applications.

This paper builds on classical statistical motifs, specifically hurdle (Cragg, 1971) and conditionally conjugate models, for modeling sparse data to estimate sparse latent spaces. We first review the hurdle model as a tool for modeling sparsity and define hurdle conjugate priors. Tailoring our method to sparse count data by building on previous work that leverages the Poisson likelihood's scalability, we propose a convolution-closed data augmentation scheme that significantly reduces the computational cost typically present multi-linear tensor decomposition methods. We apply a specific instance, the hurdle gamma prior, to fit a large-core probabilistic Tucker decomposition, demonstrating the advantage of our method using a fast Gibbs sampler for efficient posterior inference.

Contributions. Our contributions are as follows:

- We define a class of hurdle priors to impose sparsity in a latent parameter space.
- We combine these hurdle priors with a convolution-closed likelihood to develop a novel data augmentation scheme for efficient complete conditional updates.
- We incorporate a hurdle conjugate prior into a probabilistic Tucker decomposition model which allows the size of its core tensor to increase without suffering the typical exponential blowup in computational cost.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

- We derive and implement a closed-form Gibbs sampler for efficient posterior inference under the tailored probabilistic Tucker model and fit models to longitudinal microbiome data, revealing interpretable qualitative structure and encouraging quantitative results.

2. A Family of Hurdle Conjugate Priors

The hurdle model is defined by its sampling scheme,

$$b \sim \text{Bernoulli}(\rho), \tag{1}$$

$$\lambda \mid b \sim \begin{cases} \delta_0 & \text{if } b = 0 \\ g_\theta(\lambda) & \text{otherwise} \end{cases} \tag{2}$$

where $0 \notin \text{supp}(g_\theta)$ and $\rho \in (0, 1)$ is the hurdle parameter. We use $F_\lambda(\cdot)$ to define the likelihood function parameterized by λ (i.e. $F_\lambda(y) = \text{Poisson}(y; \lambda)$ or $F_{(\lambda,\sigma^2)}(y) = N(y; \lambda, \sigma^2)$, for known σ^2). Suppose we have a conditionally-conjugate pair, as in (3-5), where (3) specifies the conditionally conjugate prior, (4) specifies the likelihood, and (5) specifies the conditional posterior (where $\{c_k\}_{k=1}^K$ are scaling constants). For $k = 1, \ldots, K$,

$$\lambda_k \overset{\text{iid}}{\sim} g_\theta(\lambda_k), \tag{3}$$

$$y_k \mid \lambda_k \overset{\text{ind.}}{\sim} F_{c_k \cdot \lambda_k}(\cdot), \tag{4}$$

$$(\lambda_k \mid y_k, c_k) \overset{\text{ind.}}{\sim} g_{\theta'}(\lambda_k), \tag{5}$$

for some θ' that depends on c_k and y_k. We consider the setting where F_λ is convolution-closed.

Definition 2.1. A distribution F_λ is *convolution-closed* if for independently sampled $X_1 \sim F_{\lambda_1}, X_2 \sim F_{\lambda_2}$, the sum $X_1 + X_2 \sim F_{\lambda_1+\lambda_2}$.

Examples of convolution-closed distributions include the Gaussian, Poisson, binomial, negative binomial, gamma, multivariate Gaussian, inverse Gaussian, generalized Poisson, Tweedie, and multinomial, many of which have conjugate priors. We propose the following data augmentation scheme. Let $\bar{c} = \max_k c_k$. For each $y_k \sim F_{c_k \lambda_k}$, we wish to sample an auxiliary $\tilde{y}_k$, such that $\bar{y}_k \equiv y_k + \tilde{y}_k \sim F_{\bar{c}\lambda_k}(\cdot)$ which does not depend on c_k. If F_λ is convolution-closed, it follows that

$$y_k \sim F_{c_k \lambda_k}, \tilde{y}_k \sim F_{(\bar{c}-c_k)\lambda_k} \text{ independently}, \tag{6}$$

$$\bar{y}_k \equiv y_k + \tilde{y}_k \sim F_{\bar{c}\lambda_k}. \tag{7}$$

Under the conditionally-conjugate prior, we consider

$$b_k \sim \text{Bernoulli}(p), \tag{8}$$

$$\lambda_k \mid b_k \sim \begin{cases} \delta_0 & \text{if } b_k = 0 \\ g(\lambda_k) & \text{otherwise} \end{cases} \tag{9}$$

$$\bar{y}_k \mid \lambda_k \sim F_{\bar{c}\lambda_k}(\bar{y}_k). \tag{10}$$

Given the stated model, we are interested in posterior inference, and would like to approximate the true posterior using methods that leverage conditional-conjugacy, such as Gibbs sampling or variational inference (Casella & George, 1992; Wainwright et al., 2008; Blei et al., 2016). Inference requires iteratively updating $(b_k \mid \lambda_k, y_k)$, $(\lambda_k \mid y_k, b_k)$, and $(y_k \mid b_k, \lambda_k)$, conditional on all other parameters. Note that auxiliary sampling scales with the number of nonzero b_k. When $b_k = 0$, then $\lambda_k = 0$ and $\bar{c}\lambda_k = c_k\lambda_k = 0$. We avoid sampling auxiliary $\tilde{y}_k$, as y_k is already distributed as $F_{\bar{c}\lambda_k} = F_0$. Conditional on $\bar{y}_k, \bar{c}, b$, we have a closed-form conjugate update for λ_k. It is evident that

$$(b_k \mid \lambda_k > 0, \bar{c}, \bar{y}_k) = 1, \tag{11}$$

$$(b_k \mid \lambda_k = 0, \bar{c}, \bar{y}_k) = 0. \tag{12}$$

Updates to b_k are immediate and violate detailed balance (i.e. the Markov chain will get stuck in this state and not explore the full posterior). As such, we collapse out λ_k to derive an update

$$P(b_k = 1 \mid \bar{c}, \bar{y}_k) \propto P(\bar{y}_k \mid b_k = 1, \bar{c}) P(b_k = 1) \tag{13}$$

$$\propto p \cdot \underbrace{\int \underbrace{P(\bar{y}_k \mid \lambda, \bar{c})}_{F_{\bar{c}\lambda}(\bar{y}_k)} \underbrace{P(\lambda \mid b_k = 1)}_{g(\lambda)} d\lambda}_{f(\bar{c}, \bar{y}_k)} \tag{14}$$

which is a function of $\bar{y}_k$ and $\bar{c}$ independent of c_k. Then for $\bar{y}_k = \bar{y}_{k'}$,

$$P(b_k = 1 \mid \bar{y}_k, \bar{c}) = P(b_{k'} = 1 \mid \bar{y}_{k'}, \bar{c}). \tag{15}$$

In particular, this implies that

$$P(b_k = 1 \mid \bar{y}_k = 0, \bar{c}) = \frac{p \cdot f(\bar{c}, 0)}{p \cdot f(\bar{c}, 0) + (1 - p)} \tag{16}$$

$$= P(b_{k'} = 1 \mid \bar{y}_{k'} = 0, \bar{c}) \tag{17}$$

for $k \neq k'$. Let $\tilde{p} = P(b_k = 1 \mid \bar{y}_k = 0, \bar{c})$. Then we may update b_k by sampling

$$n_0 \sim \text{Binomial}\Big(\sum_k 1\{\bar{y}_k = 0\}, \tilde{p}\Big), \tag{18}$$

and then sampling n_0 of the k classes such that $\bar{y}_k = 0$ (without replacement). We use this convolution-closed augmentation scheme to sample $\{(b_k \mid \bar{y}_k = 0, \bar{c})\}$ jointly.

Leveraging sparsity to compute the evidence. Note that collapsing out λ_k relies on the ability to compute the evidence term $P(\bar{y}_k \mid b_k = 1) = \int F_{\bar{c}\lambda_k}(\bar{y}_k) g(\lambda_k) d\lambda_k$. Generally, computing the evidence term is intractable. Since $\bar{y}_k > 0$ implies $b_k > 0$, we only need to resample b_k (and thus compute the evidence term) when $\bar{y}_k = 0$. In many cases, the likelihood $F_{\bar{c}\lambda}(0)$ simplifies when $y = 0$, and so P(0) is tractable and cheap to compute. This yields cheap updates for $b_k \mid \bar{c}, \bar{y}_k = 0$.

Computational benefit in sparse regimes. Consider the setting where $\lambda_k = 0 \implies \bar{y}_k = 0$, then when $\bar{y}_k > 0$, $b_k = 1$ almost surely. When $b_k = 0$, then $\lambda_k = 0$ and so we avoid computation for λ_k. In the dense setting, however, we must re-sample λ_k for all K latent classes, a significantly more expensive procedure compared to the case when most $b_k = 0$.

3. Sparse Non-Negative Tucker Decomposition

We pair a hurdle gamma prior with the Poisson likelihood to model sparse count tensor data using the hurdle motif. The gamma hurdle (Jacobs, 2022) is an established tool for statistical modeling and sparsity in Tucker is an established idea. Sparse alternatives have been applied to the Tucker decomposition, including spike-and-slab priors on the individual core elements (Fang et al., 2021; Park et al., 2021; Zhang & Ng, 2022). However, these methods do not exploit sparsity for computational benefit, but instead for interpretability and generalization. As such, these methods scale poorly with the size of the core tensor. We exploit sparsity for computation and interpretability.

3.1. Tucker Decomposition

The Tucker decomposition of a tensor $\text{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_M}$ decomposes Y into a core tensor $\Lambda \in \mathbb{R}^{J_1 \times \cdots \times J_M}$ and M factor matrices $\theta^{(m)} \in \mathbb{R}^{I_m \times J_m}$. Tucker reconstructs Y as a sum of $|\Lambda| = \prod_{m=1}^{M} J_m$ weighted outer products:

$$\widehat{\text{Y}} \equiv \sum_{j_1=1}^{J_1} \cdots \sum_{j_M=1}^{J_M} \lambda_{j_1,\ldots,j_M} \theta^{(1)}_{j_1} \otimes \cdots \otimes \theta^{(M)}_{j_M}, \tag{19}$$

where each element of the core tensor $\lambda_{j_1,\ldots j_M}$ corresponds to a *weight* assigned to each outer product. We use $\boldsymbol{i} = (i_1, \ldots i_M)$ to index the observed tensor Y and $\boldsymbol{j} = (j_1, \ldots, j_M)$ to index the core tensor.

Recent work advocates for modeling complex network data using the non-negative Tucker decomposition (Schein et al., 2016; De Bacco et al., 2017; Aguiar et al., 2023). Tucker yields expressive, rich latent structure, embedding individuals into clusters, with distinct modalities to capture latent structure (i.e. temporal and spatial). However, its usefulness is limited: computation generically scales linearly with the size of the core tensor, which makes fitting Tucker difficult in practice.

For high-dimensional, sparse count data, it is natural to adopt a conditionally Poisson likelihood, so that $\text{Y} \sim \text{Poisson}(\widehat{\text{Y}})$ and constrain the parameters of $\theta^{(m)}$ and Λ to be non-negative. We refer to this adaptation as the Poisson Tucker decomposition (Schein et al., 2016). Previous work argues for using the Poisson likelihood to model sparse data for its interpretable appeal near zero and computational tractability (Chi & Kolda, 2012). Evaluating the log-likelihood of a matrix or tensor Y requires evaluating the log-likelihood at the non-zero values of Y only:

$$\log(\text{P}(\text{Y} \mid \widehat{\text{Y}}(\Theta))) = \sum_{\boldsymbol{i}} \log(\text{Poisson}(y_{\boldsymbol{i}}; \hat{y}_{\boldsymbol{i}})) \tag{20}$$

$$\propto_{\widehat{\text{Y}}} \sum_{\boldsymbol{i}} y_{\boldsymbol{i}} \log(\hat{y}_{\boldsymbol{i}}) - \hat{y}_{\boldsymbol{i}}, \tag{21}$$

a significant reduction in computational complexity when Y is sparse (i.e. $||\text{Y}||_0 << |\text{Y}|$). The Poisson additivity property gives a latent *parts-based* representation (Lee & Seung, 1999) of the observed data,

$$y_{\boldsymbol{i}k} \sim \text{Poisson}(\mu_{\boldsymbol{i}k}), \qquad y_{\boldsymbol{i}} = \sum_k y_{\boldsymbol{i}k}, \tag{22}$$

where each observed count $y_{\boldsymbol{i}}$ is the sum of K latent Poisson random variables $\{y_{\boldsymbol{i}k}\}_{k=1}^{K}$. EM, Gibbs sampling, and variational inference methods built around this scheme scale as O $(K||\text{Y}||_0)$(Gopalan et al., 2013; 2014; Schein et al., 2015). Inference involves allocating observed counts $y_{\boldsymbol{i}}$ across K latent classes through multinomial thinning,

$$y_{\boldsymbol{i}} = \sum_{k=1}^{K} y_{\boldsymbol{i}k}, \tag{23}$$

$$\{y_{\boldsymbol{i}k}\}_k \mid y_{\bullet} \sim \text{Multinomial}(y_{\bullet}, \frac{\lambda_{\boldsymbol{i}k}}{\sum_{k=1}^{K} \lambda_{\boldsymbol{i}k}})), \tag{24}$$

and updating parameters conditional on these *latent sub-counts*. In Tucker decomposition, $K = |\Lambda|$.

Bayesian Poisson Tucker decomposition. We assume priors on the parameters $\lambda_{j_1,\ldots,j_M}$, $\theta^{(m)}_{i_m j_m}$ and infer them through posterior estimation. For instance, $\lambda_{\boldsymbol{j}} \sim \text{P}(\lambda_{\boldsymbol{j}} \mid \phi)$, where ϕ parameterizes the prior distribution over $\lambda_{\boldsymbol{j}}$. Under this formulation, computation generally scales with the size of the core tensor, which experiences an exponential blowup in parameters (exponential, since size of the core tensor is exponential in M). Workarounds such as ALℓ_0CORE (Hood & Schein, 2024) have been proposed, which places an ℓ_0 constraint on Λ and infers the nonzero locations and values in the core tensor. The authors derive an inference scheme that scales computationally as $\text{O}(||\Lambda||_0 \cdot ||\text{Y}||_0)$ and enforce a strict upper bound on $||\Lambda||_0$. Inspired by the explicit l_0-constraint on the core tensor, this work places a hurdle prior on each element of the core tensor to *implicitly* provide ℓ_0 regularization. We apply the hurdle prior to the factor matrices in addition to the core tensor, incorporating sparsity across all of Tucker's latent components.

The hurdle gamma prior. A natural choice of prior for the Poisson likelihood is its conjugate prior, the gamma distribution, defined as

$$\text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^\alpha e^{-\beta\lambda}. \tag{25}$$

In Bayesian Poisson matrix and tensor factorization models (including Poisson Tucker) the gamma prior yields easy-to-compute complete conditionals, $\text{P}(\lambda \mid -)$, conducive to efficient posterior estimation via MCMC and variational inference methods. Since the gamma distribution places zero density at $\lambda = 0$, gamma priors restrict parameters to be positive dense solutions. We model the elements of the factor matrices and core tensor with hurdle gamma priors, and under the Poisson (a convolution-closed likelihood), apply the hurdle motif to speed up inference.

We alternate between allocating counts to latent sub-counts, as described above, and updating parameters conditional on these latent sub-counts. Upon allocating, inference simplifies to computing the complete conditionals under the following model:

$$b \sim \text{Bernoulli}(\rho), \tag{26}$$

$$\lambda \mid b \sim \begin{cases} \delta_0 & \text{if } b = 0 \\ \text{Gamma}(\alpha, \beta) & \text{otherwise} \end{cases} \tag{27}$$

$$y \mid \lambda \sim \text{Poisson}(c\lambda), \quad c \in \mathbb{R}_{\geq 0} \tag{28}$$

which yields closed-form complete conditional updates for each of the parameters b and λ.

Updating b. As in (11) and (12), conditioning on λ determines b so we marginalize out λ. When $y > 0$, then $b = 1$. Otherwise, $b \mid c, y = 0 \sim \text{Bernoulli}(\tilde{p})$, where

$$\tilde{p} = \frac{\rho\beta^\alpha}{(1-\rho)(\beta + c)^\alpha + \rho\beta^\alpha}. \tag{29}$$

Updating $\boldsymbol{\lambda}$. By the hurdle and conditional conjugacy,

$$(\lambda \mid b, y) \sim \begin{cases} 0 & \text{if } b = 0 \\ \text{Gamma}(\alpha + y, \beta + c) & \text{otherwise.} \end{cases} \tag{30}$$

Convolution-closed Poisson augmentation. We introduce auxiliary Poisson random variables to sample $b_{\boldsymbol{j}}$ jointly. Updates to $\lambda_{\boldsymbol{j}}$ condition on the latent sub-counts $y_{\boldsymbol{j}}$. The update to $b_{\boldsymbol{j}} \mid y_{\boldsymbol{j}}, c_{\boldsymbol{j}}$ yields different Bernoulli success parameters $\tilde{p}_j$ for each $b_{\boldsymbol{j}}$. Naively, each Gibbs update iterates over all $b_{\boldsymbol{j}}$, scaling computationally with the size of the core. We apply the convolution-closed hurdle motif to work around this problem, as the Poisson is a convolution-closed likelihood and the gamma is its conjugate prior.

Letting $\bar{c} = \max_{\boldsymbol{j}}(c_{\boldsymbol{j}})$, we sample auxiliary counts $\tilde{y}_{\boldsymbol{j}} \sim \text{Poisson}((\bar{c} - c_{\boldsymbol{j}})\lambda_{\boldsymbol{j}})$, such that

$$\bar{y}_{\boldsymbol{j}} \equiv y_{\boldsymbol{j}} + \tilde{y}_{\boldsymbol{j}} \sim \text{Poisson}(\bar{c}\lambda_{\boldsymbol{j}}). \tag{31}$$

Conditional on $\bar{y}_{\boldsymbol{j}}$, $\tilde{p} = \tilde{p}_{\boldsymbol{j}} = \tilde{p}_{\boldsymbol{j}'}$ for all $\boldsymbol{j} \neq \boldsymbol{j}'$. As such, we sample the $b_{\boldsymbol{j}}$ jointly, sampling $n \sim \text{Binomial}(|\Lambda| - ||\text{Y}^\Lambda||_0, \tilde{p})$ and then sampling n new multi-indices in the core at random without replacement. Y^Λ is the tensor of size $|\Lambda|$ containing the latent sub-counts $y_{\boldsymbol{j}}$. The greater the difference between n and $|\Lambda|$, the greater the reduction in computational cost. Our method only requires sampling $\tilde{y}_{\boldsymbol{j}}$ for those $\lambda_{\boldsymbol{j}} > 0$, which scales as $\text{O}(||\Lambda||_0)$ instead of $\text{O}(|\Lambda|)$. As above, if $b = 0$ then $\lambda = 0$. Otherwise, $(\lambda \mid b = 1, \bar{y}, \bar{c}) \sim \text{Gamma}(\alpha + \bar{y}, \beta + \bar{c})$ by conditional conjugacy. Iterating between these updates (on the factor matrices and core tensor) and allocating observed counts to latent sub-counts, which scales $\text{O}(||\Lambda||_0||Y||_0)$, forms a Gibbs sampler with stationary distribution equal to the exact posterior distribution, the target of interest.

4. Experimental Results

Data. We demonstrate our method's effectiveness by fitting Tucker to data from a microbiome longitudinal study, where Y_{ijt} denotes the gene count of gene i of subject j at time t. We qualitative and quantitatively evaluate our method on the FARMM cohort (Tanes et al., 2021), where $\text{Y} \in \mathbb{N}_0^{343 \times 30 \times 16}$. A description of the FARMM dataset may be found at (Tanes et al., 2021). We consider 343 genes, 30 subjects, each with phenotype of (vegan, omnivore, or EEN), and 16 days. Subjects receive antibiotic treatment on days 6-8 of the study. We omit day 1 observations to be consistent with previous methods (none of the vegan subjects record observations on the first day).The tensor is 76% sparse and contains $\approx 11\%$ missing values. Our approach handles missing data naturally as latent variables, imputing them during inference.

Hyperparameter selection. For simplicity, we use hurdle parameter $\rho = 0.9$ and $\text{Gamma}(1, 1)$ priors for each element of the core tensor and hurdle $\text{Gamma}(1, 10)$ priors on the elements of the factor matrices. To let sparsity levels differ across latent factors, we use Beta priors (conjugate to the binomial) $\rho_{j_m} \sim \text{Beta}(1, 1)$, where

$$\text{Beta}(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{32}$$

The sparse factor matrices distinguish subjects by phenotype in the posterior even though our model does not have access to labels while training. We fit our model for a variety of core sizes, ranging from $(3, 3, 3)$ to $(25, 3, 3)$.

Qualitative evaluation. We identify temporal structure in the time-specific factor matrix and groups of genes through the gene-specific factor matrix. The core tensor allows for all possible multi-linear latent gene-subject-time interactions. We examine the inferred values in the core tensor to evaluate inferred latent interactions, plot each subject's loading onto each latent factor, and plot the time series for each temporal latent factor. Figures 1-3 show interpretable inferred latent structure. The qualitative and quantitative results are taken from observing one sample from the posterior at random (we simply use the last saved sample). We discard the first 500 samples and save the last 1,000.

Quantitative evaluation. Our model outputs a set of posterior samples, each which contains a sample core tensor and factor matrices. We train a logistic regression classifier on the subject factor matrix to classify subjects by phenotype (EEN, or not EEN) and predict each subject's phenotype using a leave-one-out procedure. We report the area under the precision-recall curve (AUC-PR) error and compare our method to baselines from (Shi et al., 2023). We repeat this procedure for 20, 30, 40, and 50% missing data points, holding out samples from different time points at random.

Our method quantifies uncertainty around parameter estimates via Gibbs sampling, which samples parameters from the exact posterior distribution $\Theta \sim \mathrm{P}(\Theta \mid \mathrm{Y})$. One drawback of Gibbs sampling is its runtime, as sampling $1,000$ samples takes longer than a typical optimization procedure such as EM or VI. However, we note that on the FARMM dataset, for a $(15, 3, 3)$ core tensor with approximately 100 nonzero elements, one Gibbs iteration takes about 0.1 seconds on a laptop. We suspect that as the size of the observed tensor and size of the core increases, our relative computational advantage grows.

4.1. Qualitative Results

We fit a model with core tensor $\Lambda \in \mathbb{R}^{15\times3\times3}$. Our method yields a 5.9% sparse core tensor, with 8 out of 135 core elements exactly zero. The estimated 343×15 gene factor matrix is 42% sparse, while the 30×3 subject and 15×3 time factor matrices are 11% sparse.

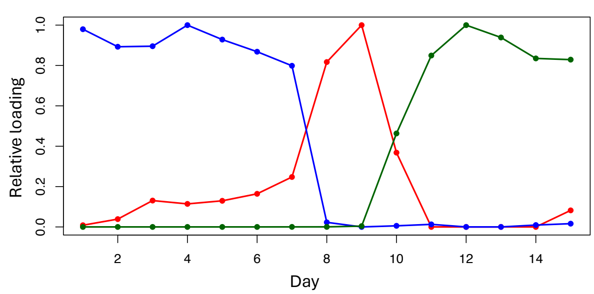

Figure 1. Time series for each latent time factor. Factor 1, in red, is most active in days 8-10. Fact 2, in blue, is active in days 1-7. Factor 3, in green, is active day 10 through the end of the study.

Factor matrices. Figure 1 shows distinct latent factors corresponding to different temporal pattern. Factor 1 (red) captures temporal structure before antibiotic treatment, while factors 2 (green) and 3 (blue) capture relatively acute and chronic responses to treatment, respectively. Figure 2 shows each subject's loading onto the latent factors. Each subject is colored by phenotype and the subject-specific latent factors delineate between phenotypes. The vegan phenotype (green) corresponds mostly to factor 1, while the EEN (black) corresponds mostly to factor 2 and omnivore (red) to factor 3.

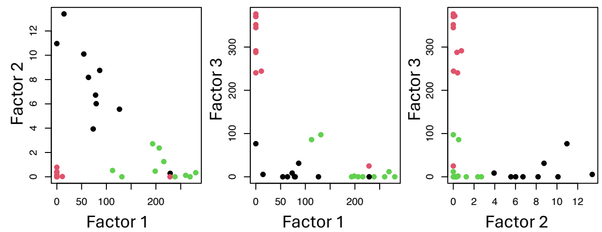

Figure 2. Tucker's learned latent factors separate subjects by phenotype. Subjects are colored by phenotype, according to EEN (black), omnivore (red), and vegan (green).

Core slices. The Tucker decomposition allows for all possible multi-linear interactions between latent gene, subject, and time factors. We explore the core to find the strongest interactions, determined by core value. We identify heterogeneous responses to antibiotic treatment by latent subject component that corresponds to known phenotype groupings. Darker colors represent higher values in the core.

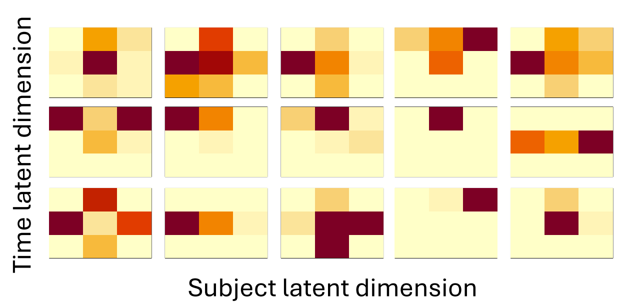

Figure 3. The core tensor. Each subfigure corresponds to a different gene-specific latent factor of the core tensor, where each slice shows different interactions between time and subject latent factors, organized by gene factor. We show all 15 latent factors to demonstrate that all 15 latent gene factors interact with the latent subject and time components differently.

4.2. Quantitative Results

We find an interesting relationship between classification accuracy and core size. As the number of gene specific latent components grows, our classifier achieves lower error,

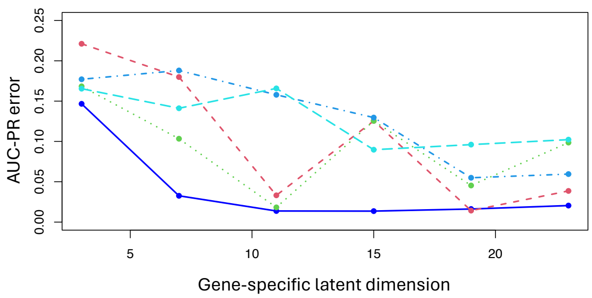

Figure 4. AUC-PR error (median over 10 masks) as a function of the gene-specific latent dimension, keeping the subject-specific latent dimension fixed at $K = 3$, for missing data proportions ranging from 11-50%. Missing proportions are 11% (dark blue), 20% (red), 30% (green), 40% (teal), and 50% (light blue).

even though the subject factor matrix is fixed. Increasing the core tensor size along one dimension yields more precise latent structure in other modalities, despite fixing the latent dimension specific to that modality.

After fitting a $(25, 3, 3)$ instance of our model, we run leave-one-out logistic regression to classify subjects by phenotype, as outlined above. Since our method does not rank principle components, like that of existing methods, we use all 3 components for logistic regression instead of 2, as done in previous studies. While the extra parameter likely inflates our model's relative performance, we consider the identification of 3 distinct, informative latent factors an advantage of our model. Our AUC-PR error (median across 10 random masks) is lower than that of existing methods, as shown Table 1, and we see this as a promising sign.

Table 1. AUC-PR error for subject-phenotype classification task.

% missing data	20%	30%	40%	50%
our method	**0.029**	**0.12**	**0.06**	**0.11**
TEMPTED	> 0.1	> 0.12	> 0.14	> 0.15
MicroTensor	> 0.25	> 0.3	> 0.3	> 0.3
CTF	> 0.4	> 0.4	> 0.4	> 0.4

5. Conclusion

We demonstrate the interpretable and computational benefits of imposing a sparse, high-dimensional latent space on non-negative Tucker decomposition. We provide a class of hurdle priors and corresponding inference scheme with this capability and see the general motif of exploiting sparsity for computational savings as a promising future direction.

References

Aguiar, I., Taylor, D., and Ugander, J. A tensor factorization model of multilayer network interdependence, February 2023. URL http://arxiv.org/abs/2206.01804. arXiv:2206.01804 [cs].

Blei, D. M., Ranganath, R., and Mohamed, S. Variational Inference: Foundations and Modern Methods (NeurIPS Tutorial), 2016.

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

Chi, E. C. and Kolda, T. G. On Tensors, Sparsity, and Nonnegative Factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, January 2012. ISSN 0895-4798. doi: 10.1137/110859063. URL https://epubs.siam.org/doi/10.1137/110859063. Publisher: Society for Industrial and Applied Mathematics.

Cragg, J. G. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: journal of the Econometric Society*, pp. 829–844, 1971.

De Bacco, C., Power, E. A., Larremore, D. B., and Moore, C. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317, April 2017. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.95.042317. URL http://link.aps.org/doi/10.1103/PhysRevE.95.042317.

Fang, S., Kirby, R. M., and Zhe, S. Bayesian streaming sparse Tucker decomposition. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 558–567. PMLR, December 2021. URL https://proceedings.mlr.press/v161/fang21b.html. ISSN: 2640-3498.

Gopalan, P., Hofman, J. M., and Blei, D. M. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.

Gopalan, P. K., Charlin, L., and Blei, D. Content-based recommendations with poisson factorization. *Advances in neural information processing systems*, 27, 2014.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Hood, J. and Schein, A. J. The ALℓ_0CORE tensor decomposition for sparse count data. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine*

Learning Research, pp. 4654–4662. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/hood24a.html.

Ishwaran, H. and Rao, J. S. Spike and slab variable selection: frequentist and bayesian strategies. 2005.

Jacobs, A. Investigaion of the gamma hurdle model for a single population mean. 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kolda, T. G. and Bader, B. W. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, August 2009. ISSN 0036-1445. doi: 10.1137/07070111X. URL https://epubs.siam.org/doi/10.1137/07070111X. Publisher: Society for Industrial and Applied Mathematics.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755):788–791, October 1999. ISSN 1476-4687. doi: 10.1038/44565. URL https://www.nature.com/articles/44565. Number: 6755 Publisher: Nature Publishing Group.

Louizos, C., Welling, M., and Kingma, D. P. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

Ma, S. and Li, H. A tensor decomposition model for longitudinal microbiome studies. *The Annals of Applied Statistics*, 17(2):1105–1126, 2023.

Park, M., Jang, J.-G., and Sael, L. VEST: Very Sparse Tucker Factorization of Large-Scale Tensors. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 172–179, January 2021. doi: 10.1109/BigComp51126.2021.00041. URL https://ieeexplore.ieee.org/document/9373235. ISSN: 2375-9356.

Ranganath, R., Gerrish, S., and Blei, D. M. Black Box Variational Inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014. URL http://arxiv.org/abs/1401.0118. arXiv:1401.0118 [cs, stat].

Schein, A., Paisley, J., Blei, D. M., and Wallach, H. Bayesian Poisson Tensor Factorization for Inferring Multilateral Relations from Sparse Dyadic Event Counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1045–1054, New York, NY, USA, August 2015. Association for Computing Machinery. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783414. URL https://dl.acm.org/doi/10.1145/2783258.2783414.

Schein, A., Zhou, M., Blei, D. M., and Wallach, H. Bayesian Poisson tucker decomposition for learning the structure of international relations. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 2810–2819, New York, NY, USA, June 2016. JMLR.org.

Shi, P., Martino, C., Han, R., Janssen, S., Buck, G., Serrano, M., Owzar, K., Knight, R., Shenhav, L., and Zhang, A. Time-informed dimensionality reduction for longitudinal microbiome studies. 07 2023.

Tanes, C., Bittinger, K., Gao, Y., Friedman, E. S., Nessel, L., Paladhi, U. R., Chau, L., Panfen, E., Fischbach, M. A., Braun, J., et al. Role of dietary fiber in the recovery of the human gut microbiome and its metabolome. *Cell host & microbe*, 29(3):394–407, 2021.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Zhang, X. and Ng, M. K. Sparse nonnegative tucker decomposition and completion under noisy observations. *arXiv preprint arXiv:2208.08287*, 2022.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.