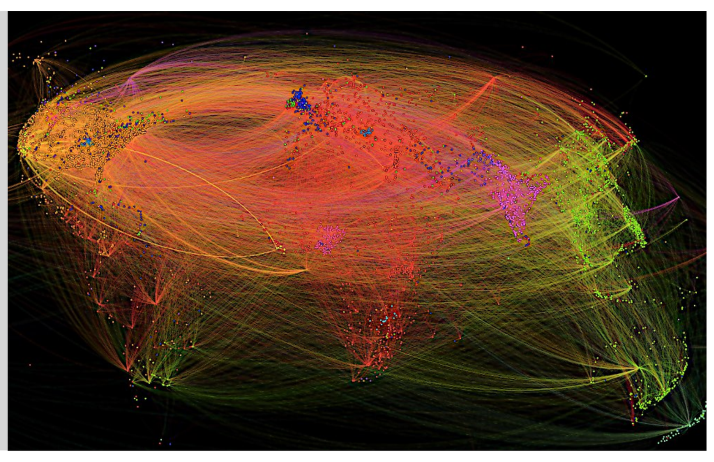
The Convolution-Closed Hurdle Motif
With an Application to Tensor Decomposition

John Hood[1] and Aaron Schein[1]

[1]Department of Statistics, University of Chicago

How to Efficiently Model High-Dimensional, Sparse Latent Spaces?

Sparse and high-dimensional data are ubiquitous in scientific applications. Practitioners often seek to build complex models for such data, whose parameters and latent variables are themselves high-dimensional.



Techniques that promote model sparsity are typically motivated either to avoid **overfitting** or to improve model **interpretability**. Although effective for both, such techniques frequently introduce **increased computational cost** to inference. In some cases, **model sparsity can improve** rather than exacerbate **the computational cost of inference**.

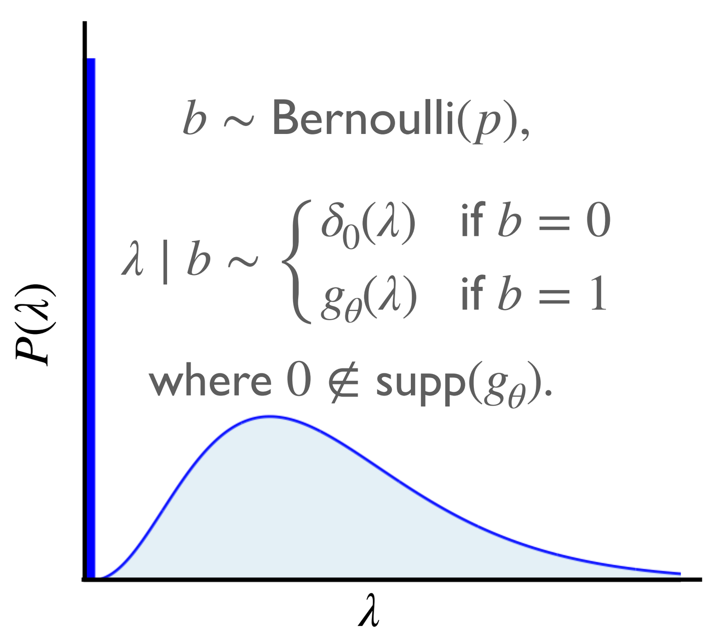
Hurdle Conjugate Priors for Modeling Sparsity

The hurdle model for λ is defined by it's generative process.

$$b \sim \text{Bernoulli}(p),$$

$$\lambda \mid b \sim \begin{cases} \delta_0(\lambda) & \text{if } b = 0 \\ g_\theta(\lambda) & \text{if } b = 1 \end{cases}$$

where $0 \notin \text{supp}(g_\theta)$.



Every member of the exponential family has a ***conjugate prior***.

$$g_\theta(\lambda) = p(\lambda \mid \theta) = h(\lambda) e^{\theta^T \lambda - \tilde{\eta}(\theta)}$$

$$F_\lambda(y) = p(y \mid \lambda) = f(y) e^{y^T \lambda - \eta(\lambda)}$$

$$\implies p(\lambda \mid y, \theta) = g_{\theta'}(\lambda)$$

In this setting, the posterior $p(\lambda \mid y)$ is in the same family as the prior. Many Gibbs sampling and variational inference methods leverage conditionally conjugate models for effficient inference.

Convolution-Closed Likelihoods

A distribution F_λ is **convolution-closed** in λ if for independently sampled $X_1 \sim F_{\lambda_1}, X_2 \sim F_{\lambda_2}$, $X_1 + X_2 \sim F_{\lambda_1+\lambda_2}$ in it's marginal.

Many closed-convolution distributions are members of the exponential family and have closed-form conjugate priors.

$$P(\lambda) = \text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$P(y \mid c, \lambda) = \text{Poisson}(y; c \cdot \lambda) = \frac{e^{-c\lambda}(c\lambda)^y}{y!}, \; c > 0$$

$$P(\lambda \mid y) = \text{Gamma}(\lambda; \alpha + y, \beta + c)$$

The posterior depends on only the sums of the sufficient statistics $\sum_i y_i, \sum_i c_i$.

For $y_i \sim \text{Poisson}(c_i \lambda)$, then marginally, $\sum_i y_i \sim \text{Poisson}(\lambda \cdot \sum_i c_i)$ and

$$P(\lambda \mid \{y_i, c_i\}) = P(\lambda \mid \sum_i y_i, \sum_i c_i)$$

$$= \text{Gamma}(\lambda; \alpha + \sum_i y_i, \beta + \sum_i c_i)$$

The Closed-Convolution Hurdle Motif

<u>**Model**</u>

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

$$\lambda_k \mid b_k \sim \begin{cases} \delta_0(\lambda_k) & \text{if } b_k = 0 \\ g_\theta(\lambda_k) & \text{if } b_k = 1 \end{cases}$$

$$y_k \mid \lambda_k \sim F_{c_k \lambda}(y_k), \quad c_k \in \mathbb{R}$$

data augmentation

$$\tilde{y}_k \mid \lambda_k \sim F_{(\bar{c}-c_k)\lambda_k}(\tilde{y}_k), \; \bar{c} = \max_k c_k$$

$$\implies \bar{y}_k \mid \lambda_k = \tilde{y}_k + y_k \sim F_{\bar{c}\lambda_k}(\bar{y})$$

<u>**Inference**</u>

$$y_k \sim p(y_k \mid -),$$

$$\lambda_k \sim p(\lambda_k \mid -),$$

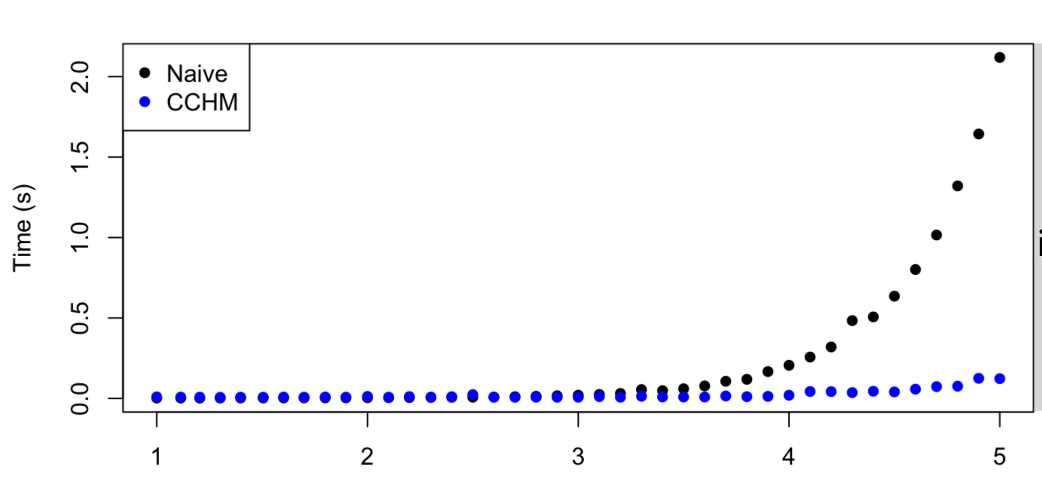
Naive	CCHM
	cheap in sparse regimes
$b_k \sim p(b_k \mid - \setminus_{\lambda_k})$	$\tilde{y}_k \sim p(\tilde{y}_k \mid -)$,
depends on c_k → expensive	$b_k \sim p(b_k \mid - \setminus_{\lambda_k})$
	does not depend on c_k

Gibbs sampling involves iteratively resampling each latent variable, conditional on all the others held fixed.

A Data Augmentation Scheme That Leads to Faster Inference

Conditioning on $\tilde{y}_k, \bar{y}_k$ drops the dependence of $P(b_k \mid - \setminus_{\lambda_k})$ on c_k. We can sample from $P((b_k)_{k=1}^K \mid \bar{y}_k = 0, - \setminus_{\lambda_k})$ jointly.

$$P(b_k \mid \bar{y}_k = 0, -) = P(b_{k'} \mid \bar{y}_{k'} = 0, -) = \tilde{p}_0 \forall k \text{ s.t. } \bar{y}_k = 0$$



$$n \sim \text{Binomial}(\sum_{k=1}^K 1\{y_k = 0\}, \tilde{p}_0),$$

$$\text{ind} \sim \text{sample}(\{k : y_k = 0\}, n, \text{replace} = \text{F})$$

$$b_{\text{ind}} \leftarrow 1$$

Applications to Tensors

Tensors can be informally understood as matrices generalized to $M \geq 2$ dimensions or modes—e.g., while a matrix Y contains observations y_{d_1,d_2}, an M-mode tensor Y contains observations $y_{d_1,\ldots,d_m}$.

What is the Tucker Decomposition?

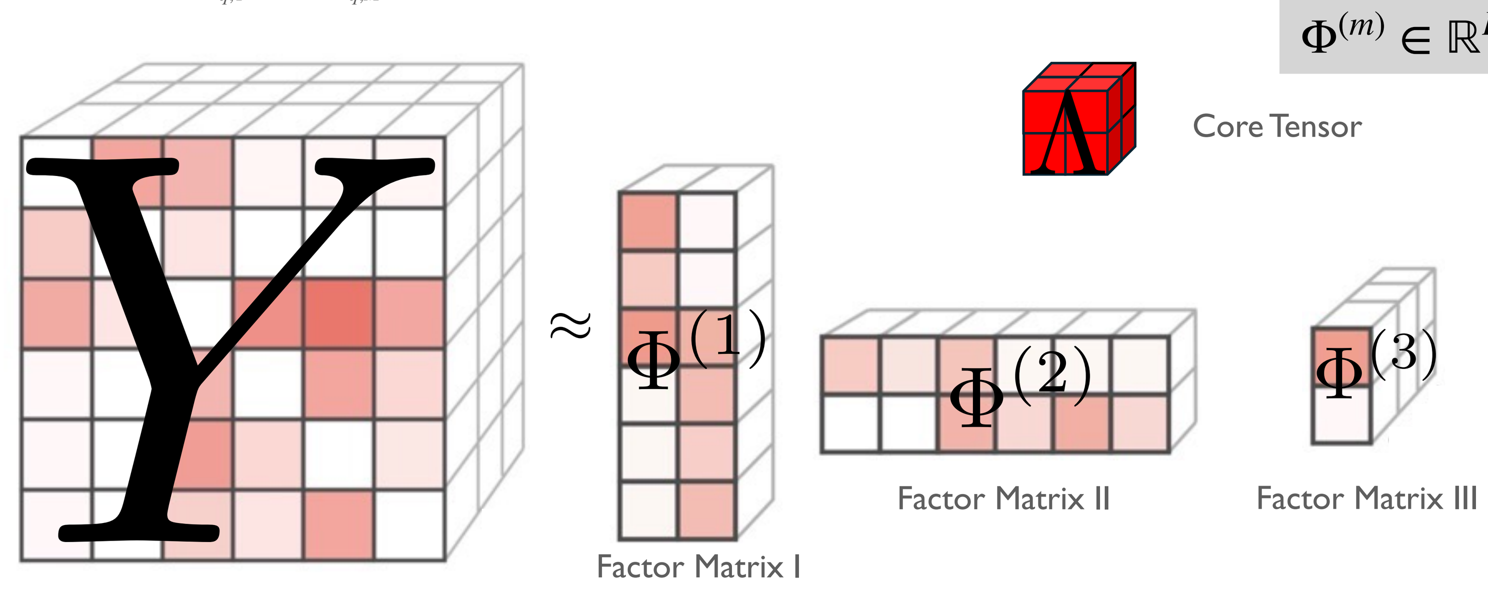
Tucker decompositions seek a multi-linear reconstruction $\hat{Y} \approx Y$.

$$\hat{Y} = \sum_{k_{q,1}=1}^{K_1} \cdots \sum_{k_{q,M}=1}^{K_M} \lambda_{k_q} \left(\phi^{(1)}_{k_{q,1}} \circ \cdots \circ \phi^{(M)}_{k_{q,M}} \right)$$

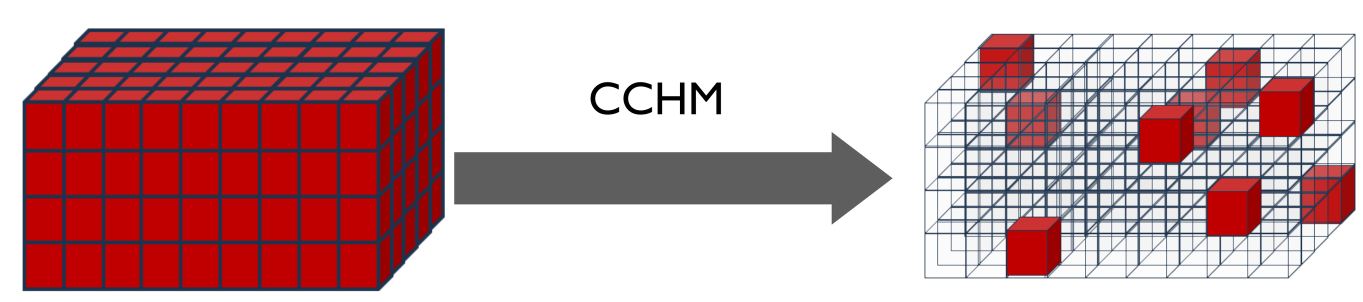
$Y \in \mathbb{R}^{D_1 \times \ldots \times D_M}$

$\Lambda \in \mathbb{R}^{K_1 \times \ldots \times K_m}$

$\Phi^{(m)} \in \mathbb{R}^{D_m \times K_m}$

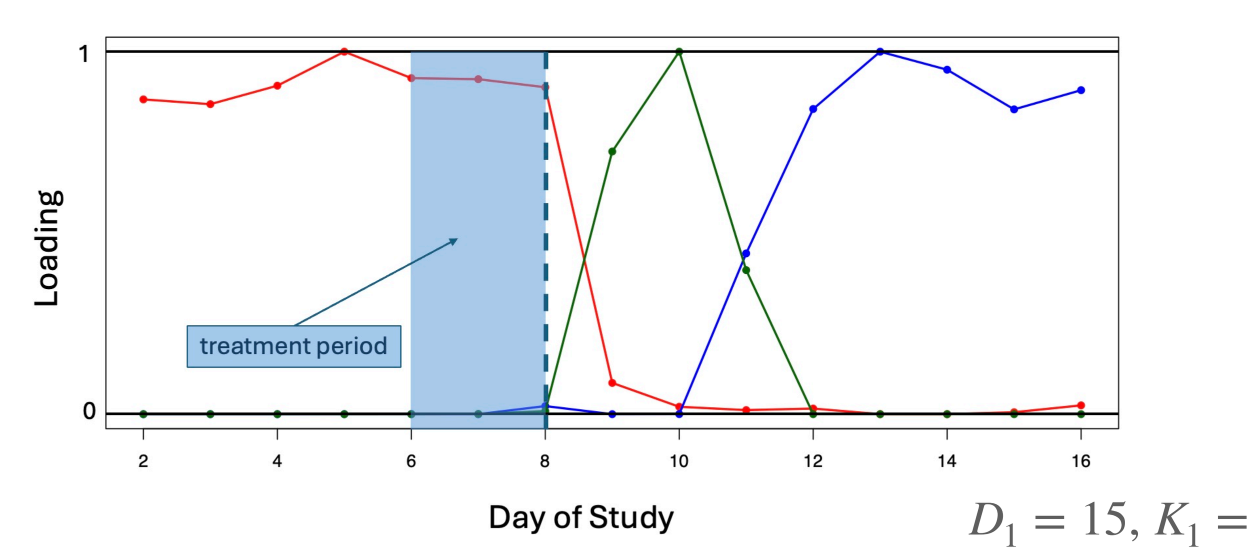


Problem: Traditional Tucker decomposition methods generally do not scale well with the ℓ_0-norm of the core tensor, $||\Lambda||_0 = \sum_k 1\{\Lambda_k > 0\}$.



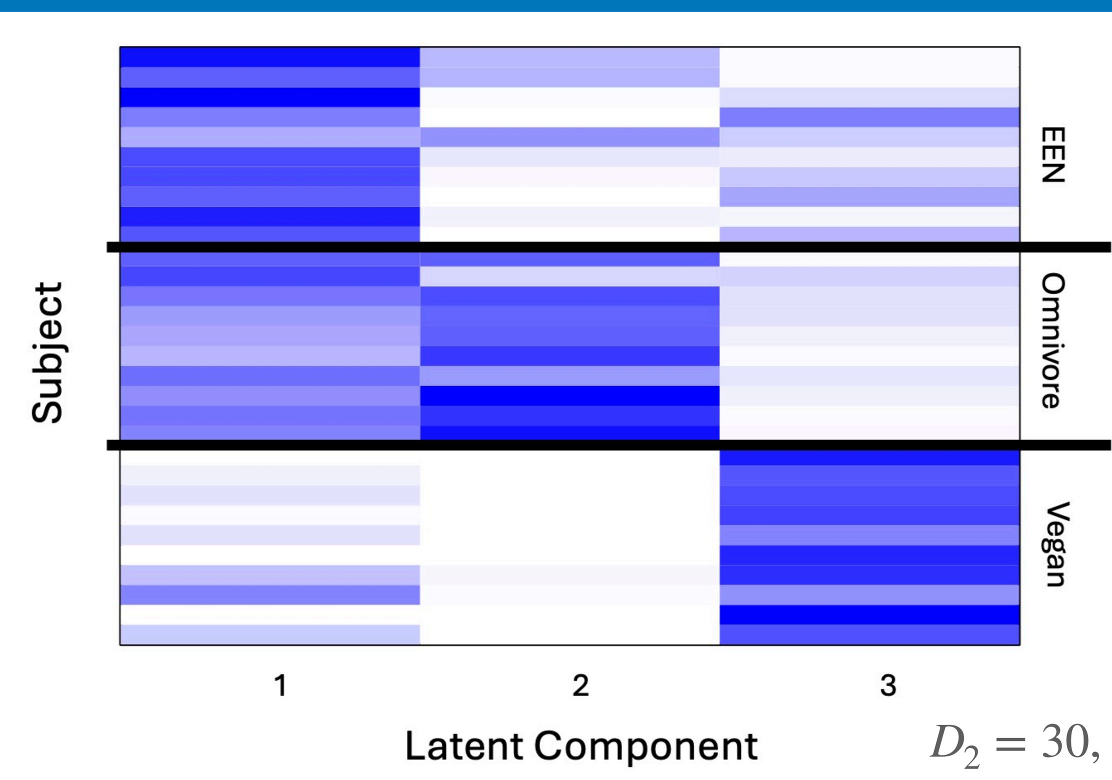
We apply the convolution-closed hurdle motif to sparsify the core tensor of a Tucker decomposition, speeding up inference.

CCHM Tucker Uncovers Hidden Temporal Structure



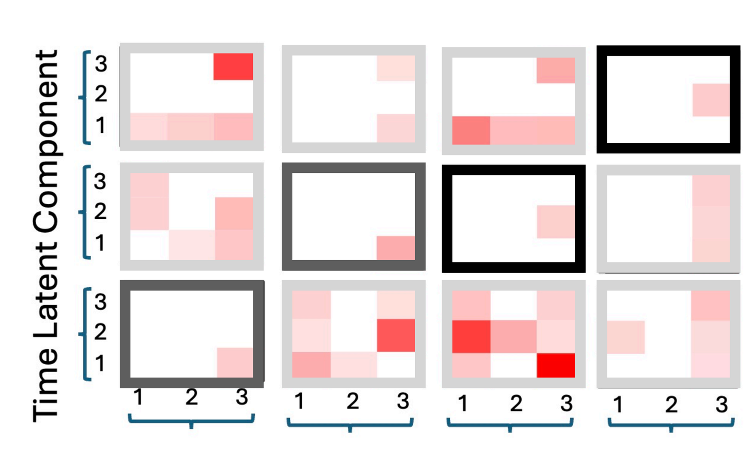
$D_1 = 15, K_1 = 3$

Separates Subjects by Phenotype



$D_2 = 30, K_2 = 3$

and Identifies Sparse Multi-linear Interactions



$\hat{\Lambda} \in \mathbb{R}^{15 \times 3 \times 3}$

71 % sparse

multi-linear structure

repetitive structure

Questions? Email: johnhood@uchicago.edu

This work is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2140001.