Beyond matrices: tensor decomposition and its applications

Miaoyan Wang

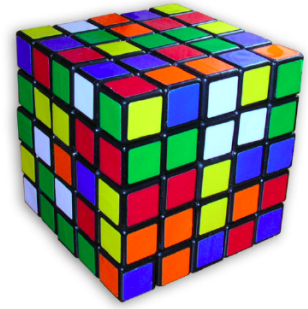
Department of Statistics

University of Wisconsin, Madison

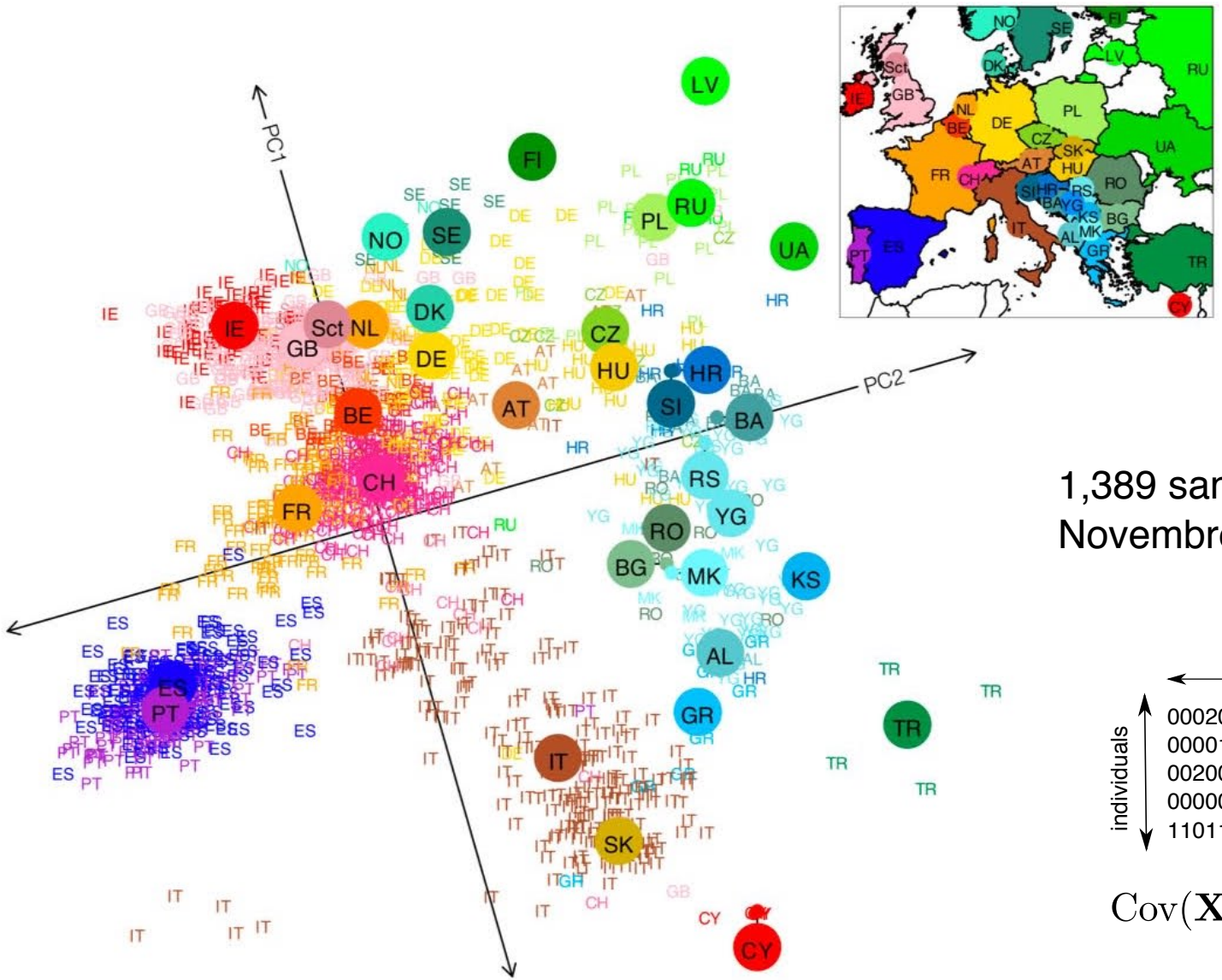
(Joint work with Yun S. Song from UC Berkeley)

Stanford Statistics Seminar

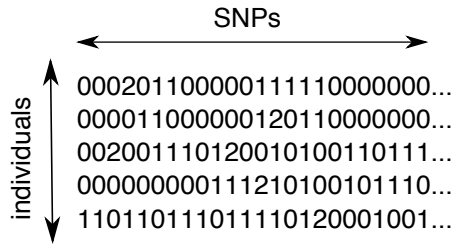
Aug 7th, 2018



A successful story: PCA of Europeans

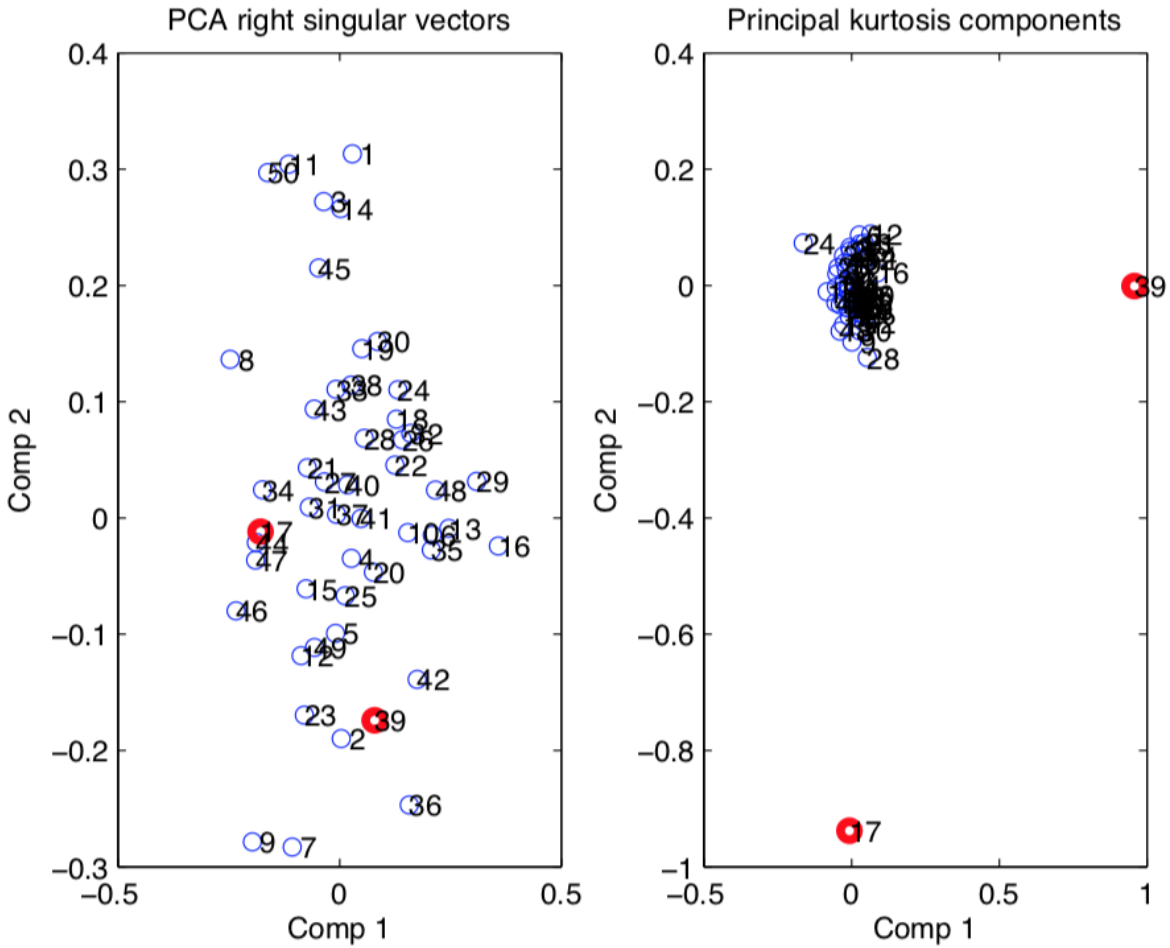


1,389 samples, ~ 200k SNPs
Novembre et al. (2008)



$$\mathrm{Cov}(\mathbf{X}) = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

Matrix methods are powerful, however...



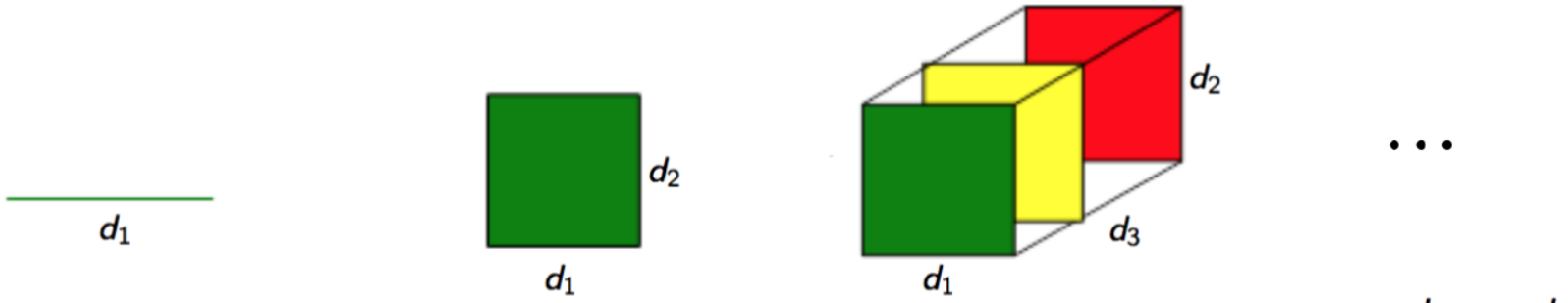
All Gaussian except points 17 and 39.

left: matrix PCA; right: principal components of kurtosis.

Figure credit: Jason Morton and Lek-Heng Lim (2009/2015).

What is a tensor?

- Tensors are generalizations of vectors and matrices:



- An order-k tensor $\mathcal{A} = [\![a_{i_1 \ldots i_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ is a hypermatrix with dimensions $(d_1, \ldots, d_k)$ and entries $a_{i_1 \ldots i_k} \in \mathbb{R}$.
- This talk will focus on tensor of order 3 or greater, also known as higher-order tensors.

Tensors in statistical modeling

"Tensors are the new matrices" that tie together a wide range of areas:

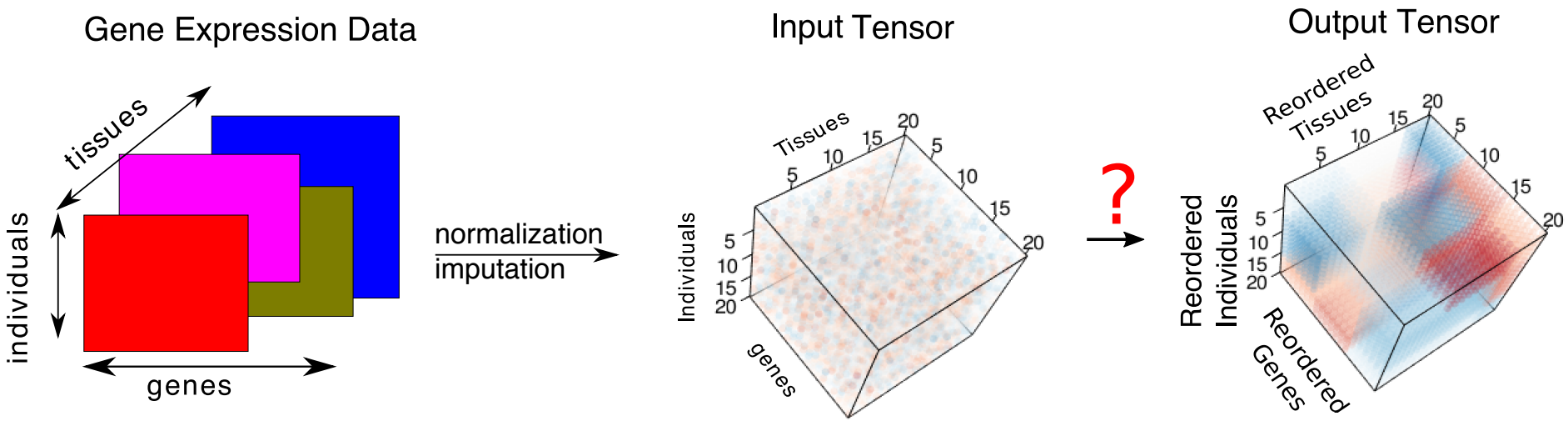
- Longitudinal social network data $\{\boldsymbol{Y}_t : t = 1, ..., n\}$
- Joint probability for multivariate categorical data $\mathbb{P}(X_1, X_2, X_3)$
- Moments of multivariate distribution $\mathbb{E}_{\boldsymbol{x} \sim D} \boldsymbol{x}^{\otimes 3}$
- Markov models for the phylogenetic tree $K_{1,3}$

P. Hoff 2015, Montanari-Richard 2014, Anandkumar et al 2014

Bhattacharya-Dunson 2012, Mossel et al 2004, P. McCullagh 1987

Tensors in genomics

- Many biomedical datasets come naturally in a multiway form.
- Multi-tissue multi-individual gene expression measures could be organized as a multiarray dataset $\mathcal{Y} = [\![Y_{git}]\!] \in \mathbb{R}^{n_G \times n_I \times n_T}$.



Multi-way Clustering

To identify subsets of genes that are similarly expressed within subsets of individuals and tissues, we seek local blocks in the expression tensor.

Talk outline

Prohibitive Computational Complexity

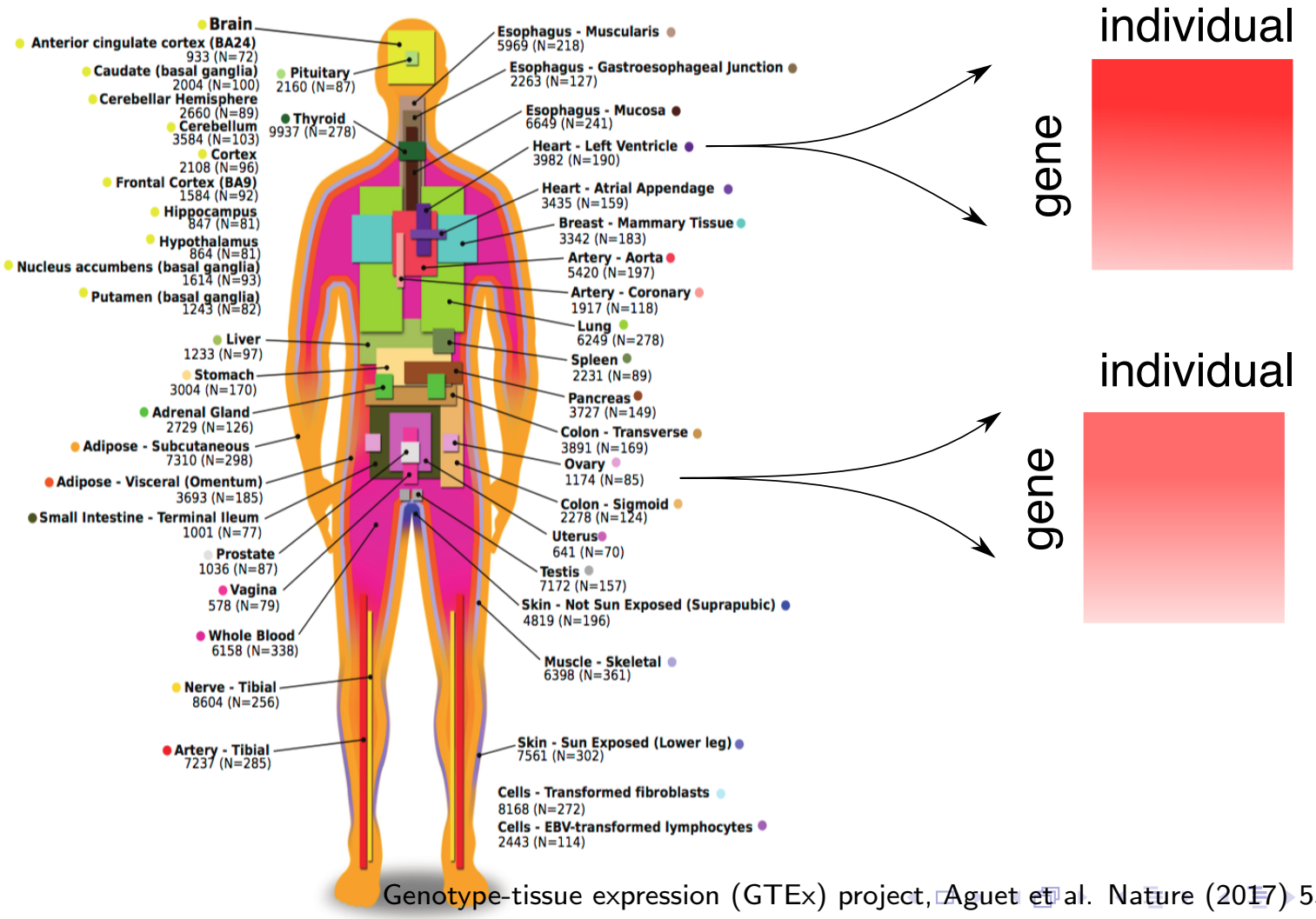
Most higher-order tensor problems are NP-hard [Hillar & Lim, 2013].

Topics I will address:

- Tensor decomposition and applications in GTEx data
- Binary tensor decomposition and its statistical optimality

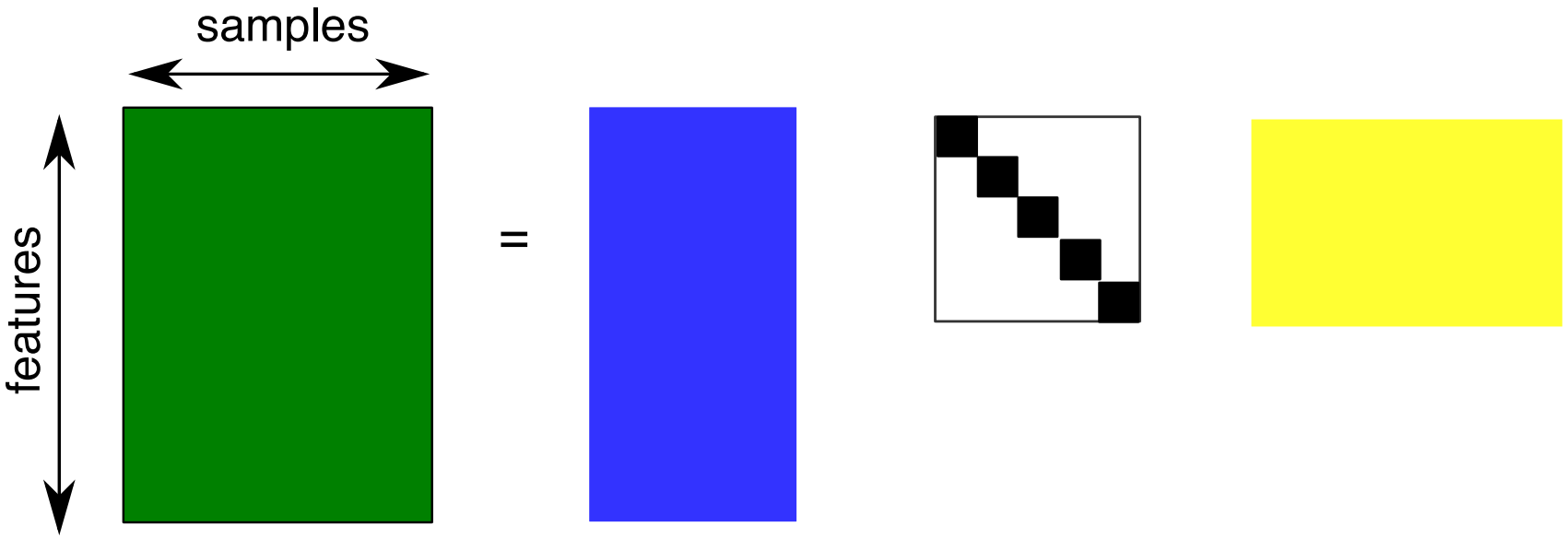
Multi-tissue gene expression analysis

- Central dogma of genetics: DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{translation}}$ protein.
- GTEx RNA-seq data: expression profiles ($\sim 25,000$ genes) measured from 544 individuals across 53 tissues.



Genotype-tissue expression (GTEx) project, Aguet et al. Nature (2017) 550, 204-213.

Review: Matrix SVD for biclustering



$$\mathbf{X} = \mathbf{U} \Lambda \mathbf{V}^T$$

$$= \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^T$$

- Columns of $\mathbf{U}$ describe patterns across samples
- Columns of $\mathbf{V}^T$ describe patterns across genes

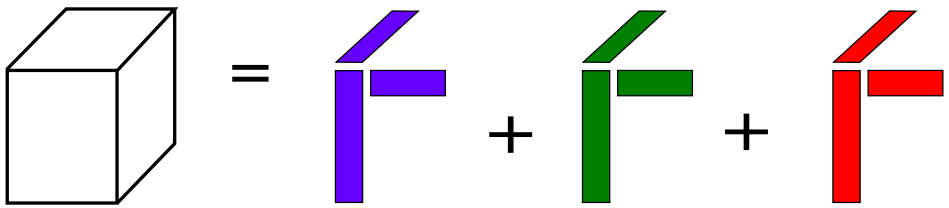
Data Science Specialization (COURSERA) by Brian Caffo, Jeff Leek, Peng Ding

Tensor rank decomposition

- Tensor analogue of matrix SVD:

$$\mathcal{Y} = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$$

- Example: a rank-3 order-3 tensor.

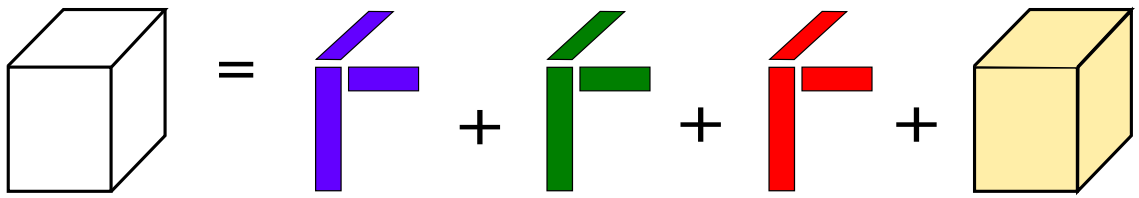


$$\mathcal{Y} = \lambda_1 \mathbf{u}_1^{\otimes 3} + \lambda_2 \mathbf{u}_2^{\otimes 3} + \lambda_3 \mathbf{u}_3^{\otimes 3}$$

Low-rank approximation

- a noisy rank-R tensor:

$$\underbrace{\mathcal{Y}}_{\text{observation}} = \underbrace{\sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r}_{\text{signal}} + \underbrace{\mathcal{E}}_{\text{noise}},$$



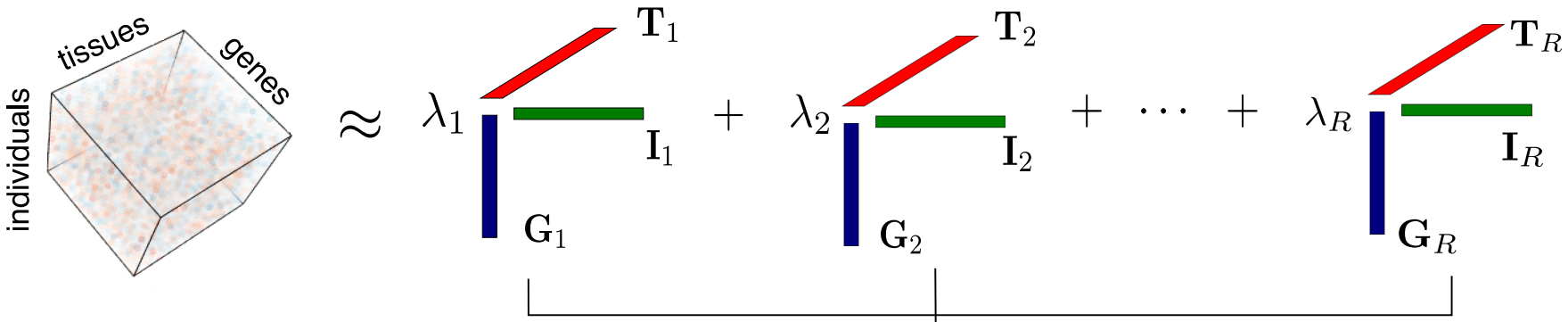
- Given $\mathcal{Y}$ and R, find the best rank-R approximation:

$$\min_{\boldsymbol{a}_r, \boldsymbol{b}_r, \boldsymbol{c}_r, \lambda_r : r \in [R]} \|\mathcal{Y} - \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r\|_F^2$$

- Minimizing the squared loss corresponds to maximizing the likelihood under the i.i.d. Gaussian error model.

Multi-tissue gene expression analysis

We have developed a semi-nonnegative tensor decomposition for three-way clustering of multi-tissue multi-individual gene expression data.



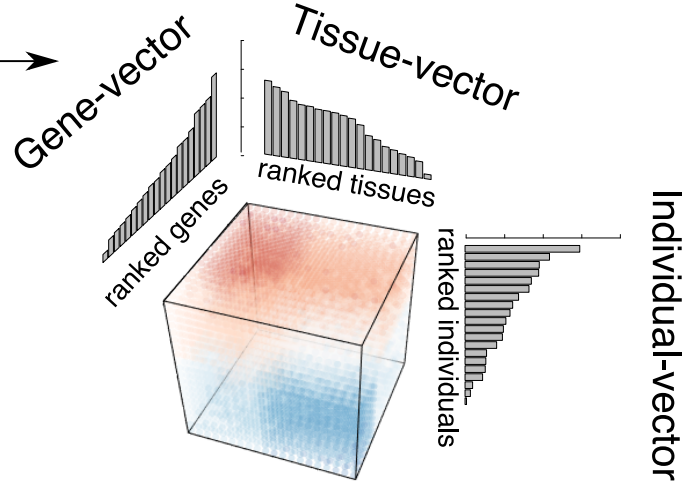
$$\mathcal{Y} = \sum_{r=1}^{R} \lambda_r \mathbf{T}_r \otimes \mathbf{G}_r \otimes \mathbf{I}_r + [e_{ijk}]$$

with $\mathbf{T}_r \geq 0$ entrywise and $e_{ijk} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

Each gene expression moduel consists of

λ_r : Singular-value $\quad$ $\mathbf{T}_r$: Tissue-vector

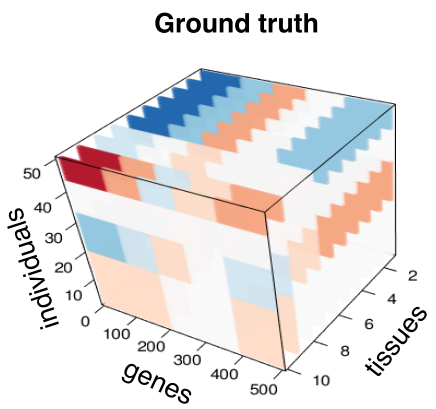
$\mathbf{G}_r$: Gene-vector $\quad$ $\mathbf{I}_r$: Individual-vector

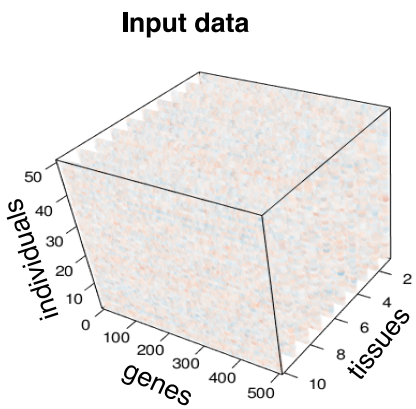


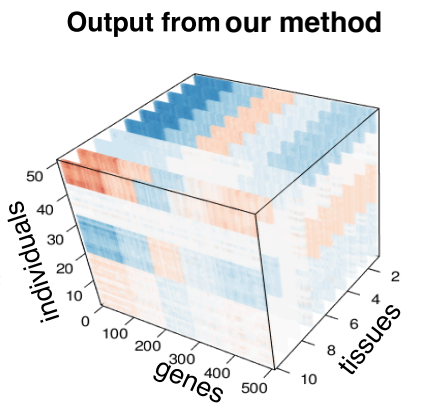
Wang et al., under review (2017). doi.org/10.1101/229245,

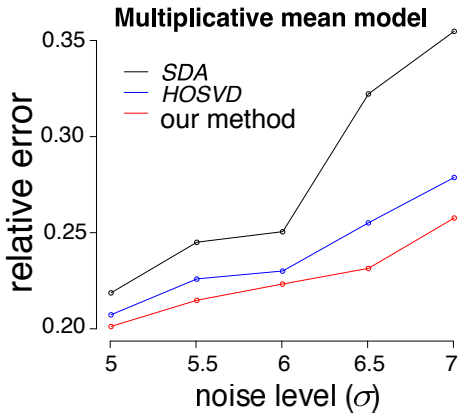
Performance comparison

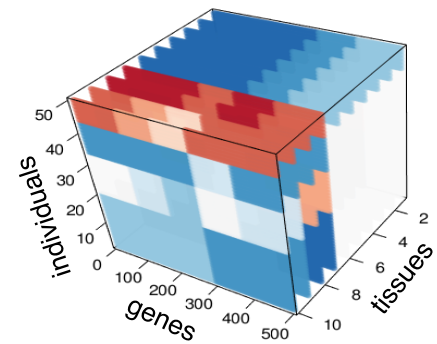
Our algorithm recovers low-rank tensor signals with higher accuracy compared to existing methods.

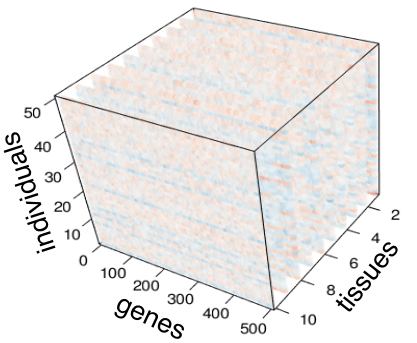


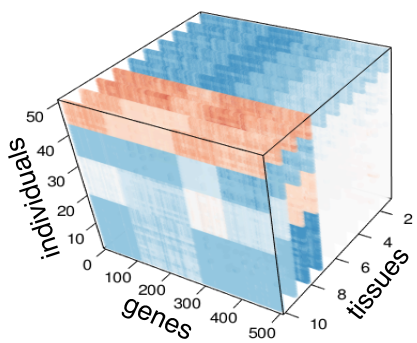


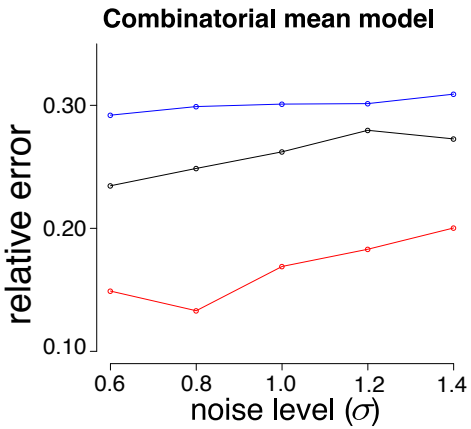






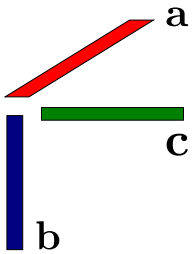




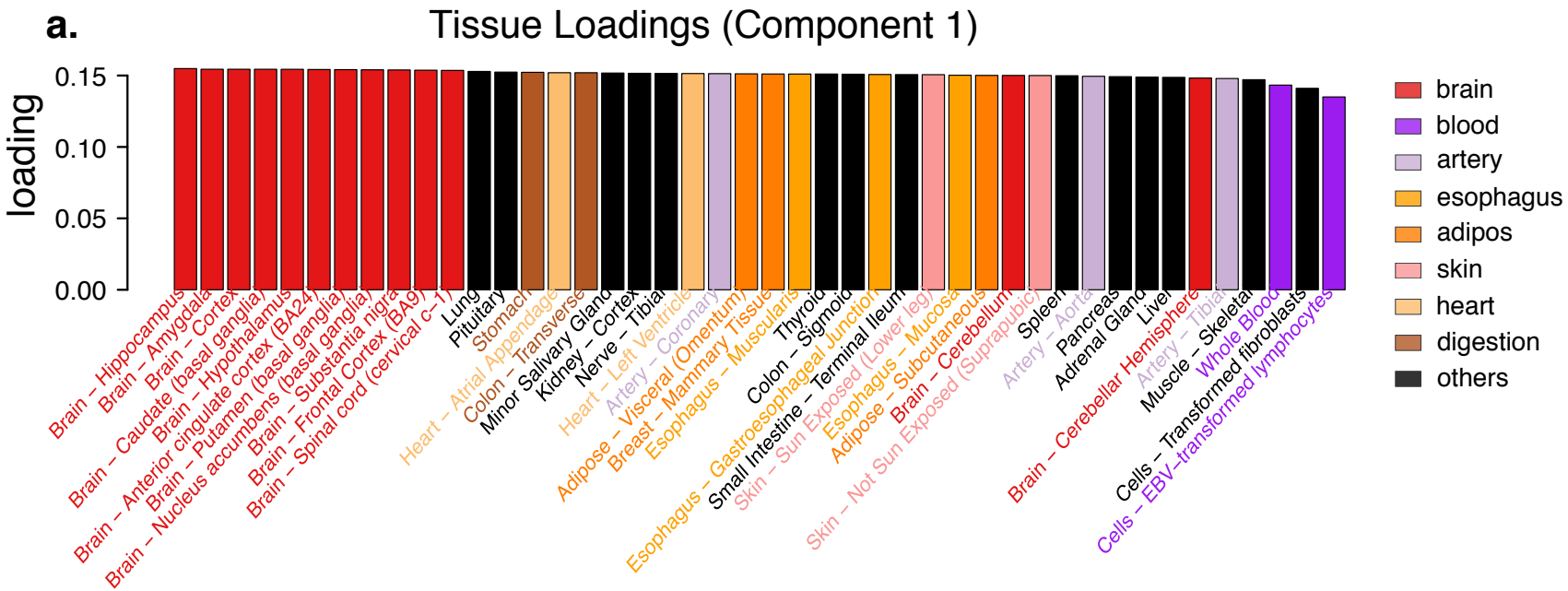


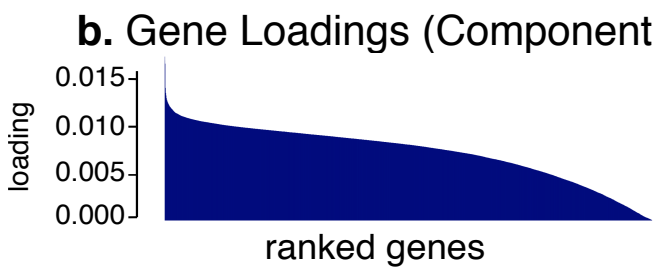
SDA: sparse decomposition of array (Hore et al. Nat. Gen. 2016);
HOSVD: higher-order singular value decomposition (Omberg et al. PNAS. 2007).

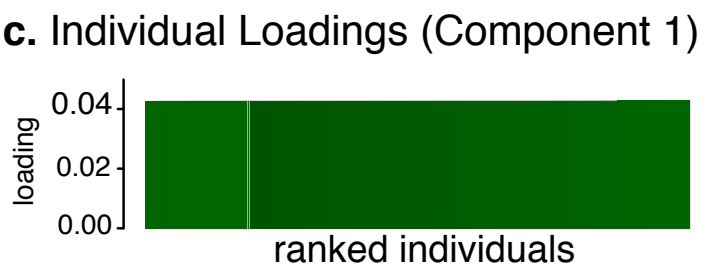
Component 1: shared, global expression



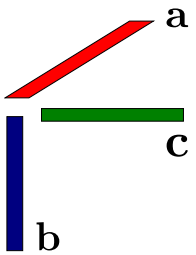
- Both tissue- and individual-loadings are essentially flat $\Rightarrow$ this component captures global expression common to all samples.



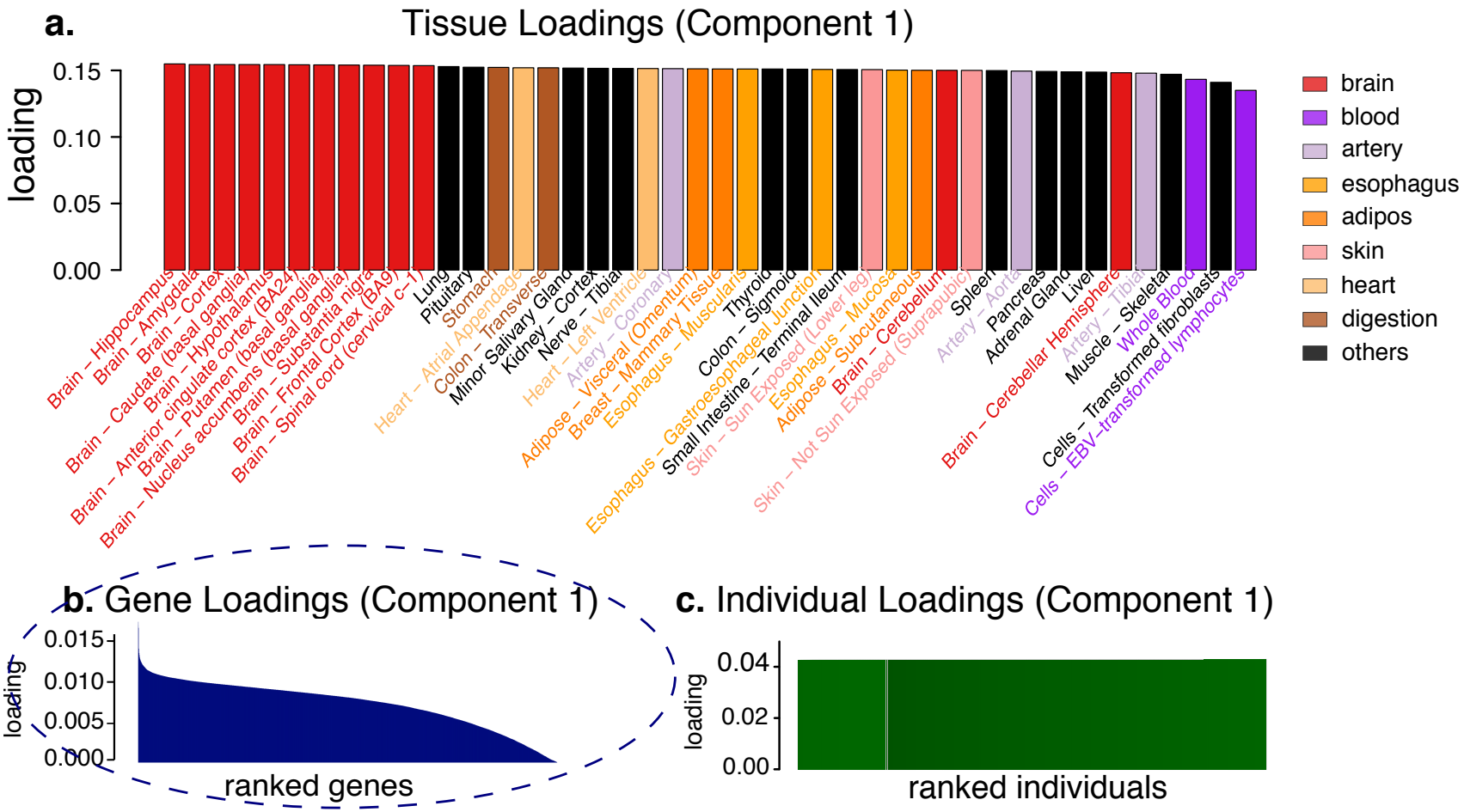




Component I: shared, global expression

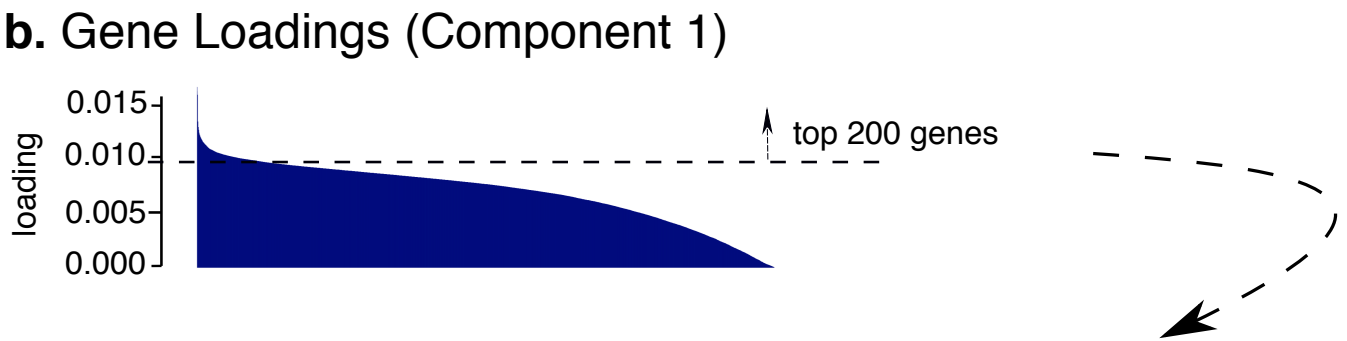


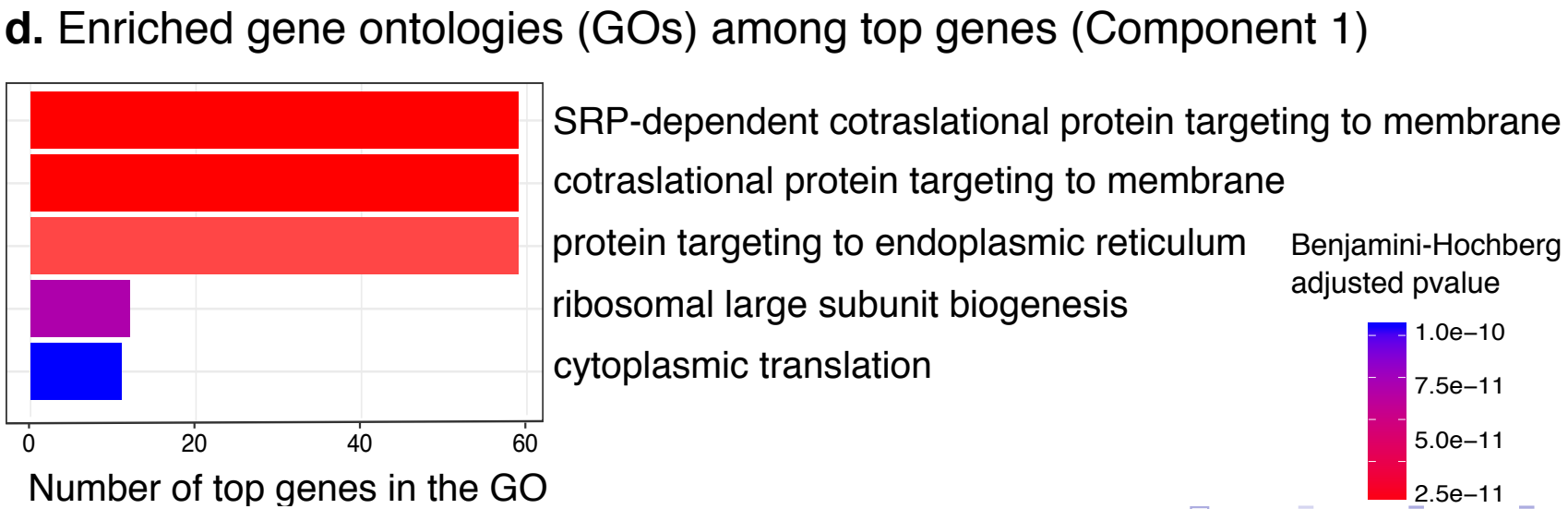
- Both tissue- and individual-loadings are essentially flat $\Rightarrow$ this component captures global expression common to all samples.



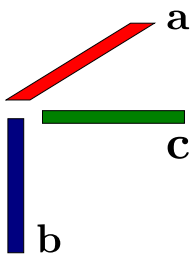
Component I: shared, global expression

- Top genes are mainly mitochondrial genes (15/20 top genes); other non-mitochondrial genes include *ACTB*, *EEF1A1*, and *EEF2*.

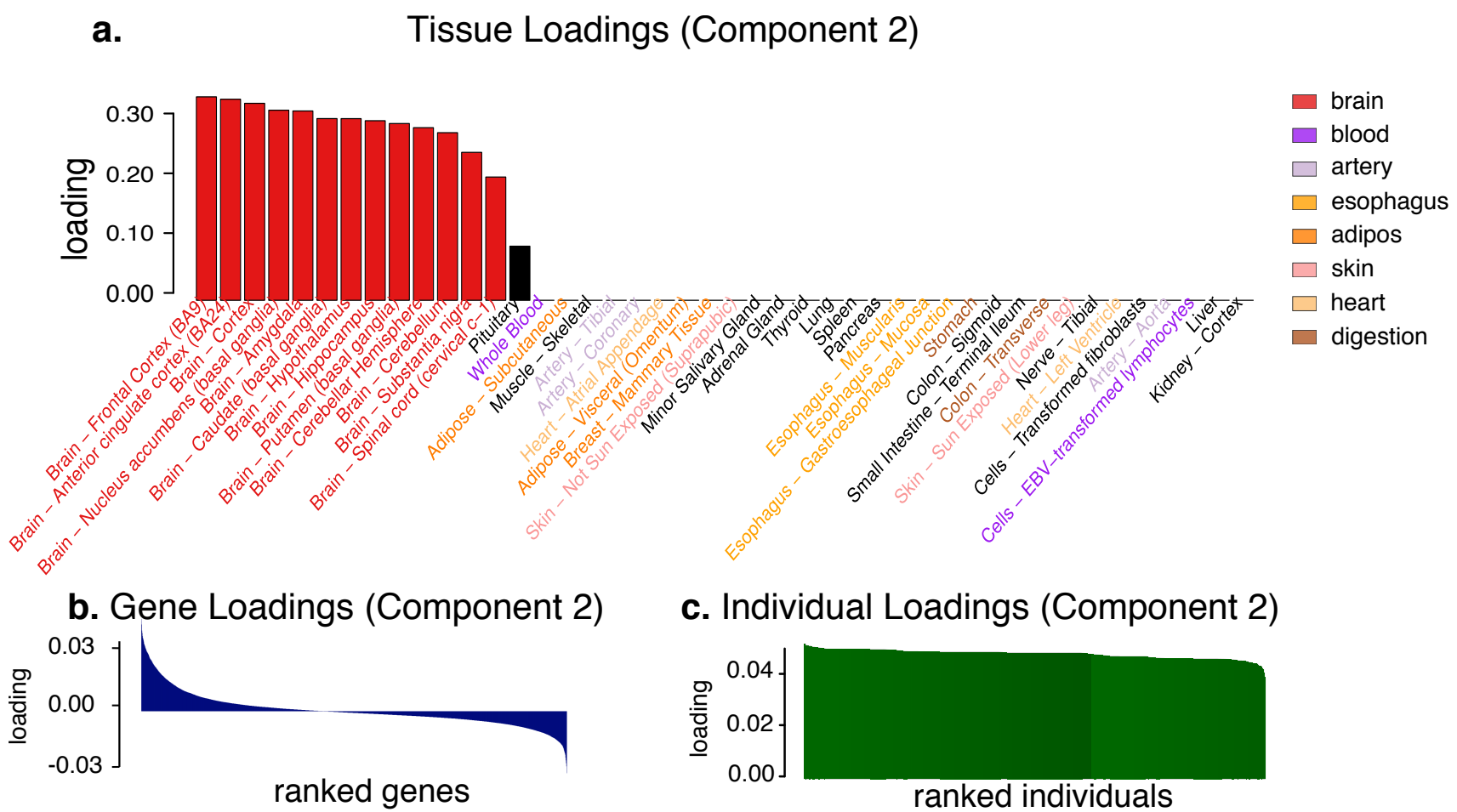




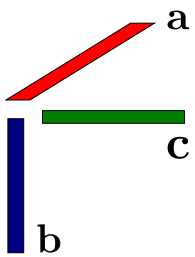
Component II: brain tissues



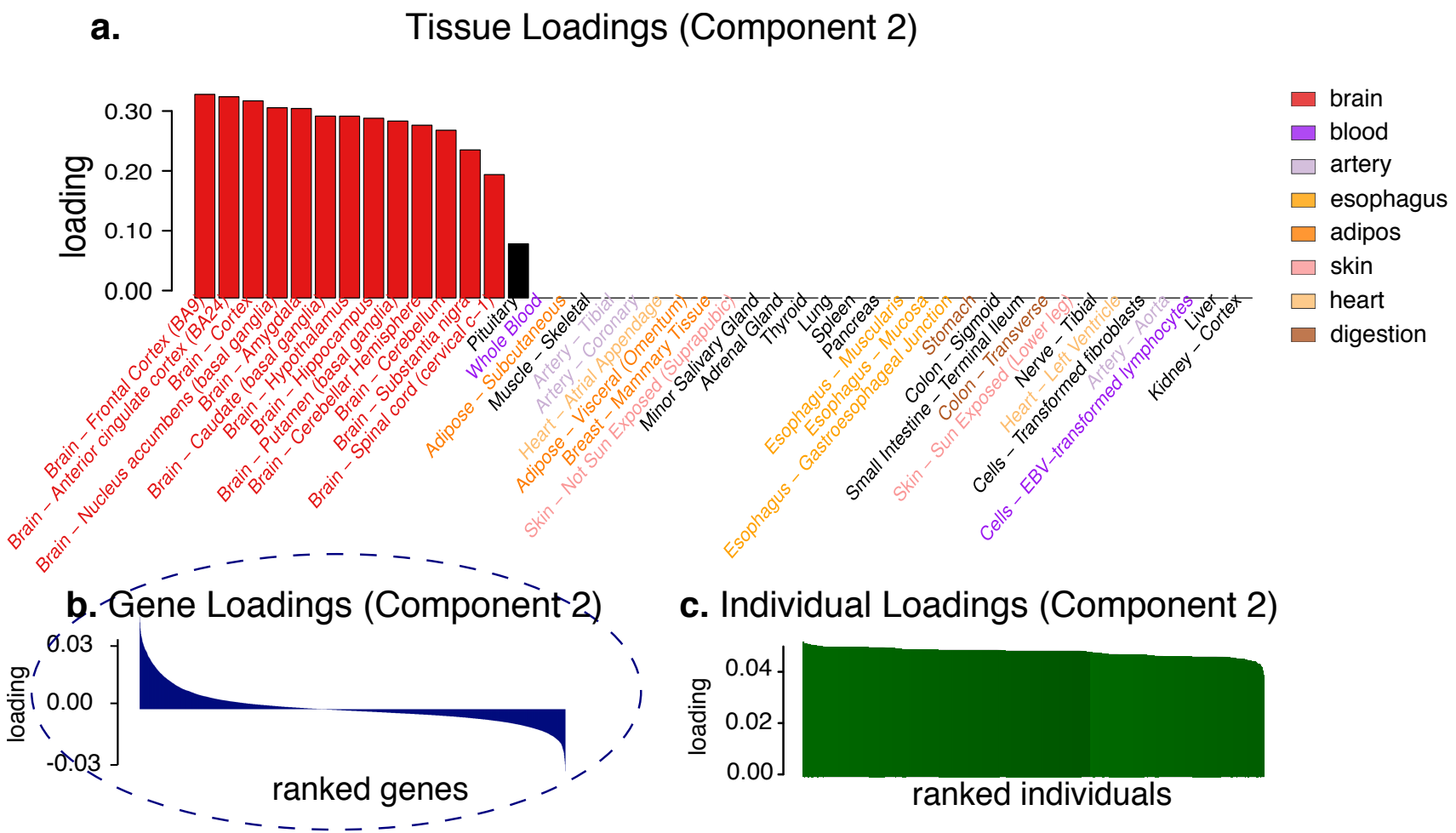
- Tissue vector clearly separates brain tissues from non-brain tissues.



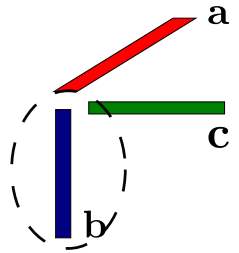
Component II: brain tissues



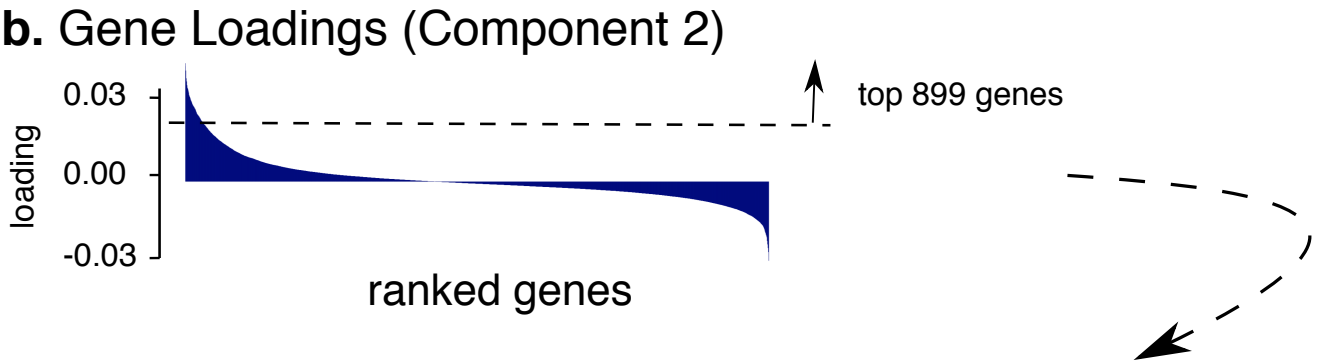
- Tissue vector clearly separates brain tissues from non-brain tissues.

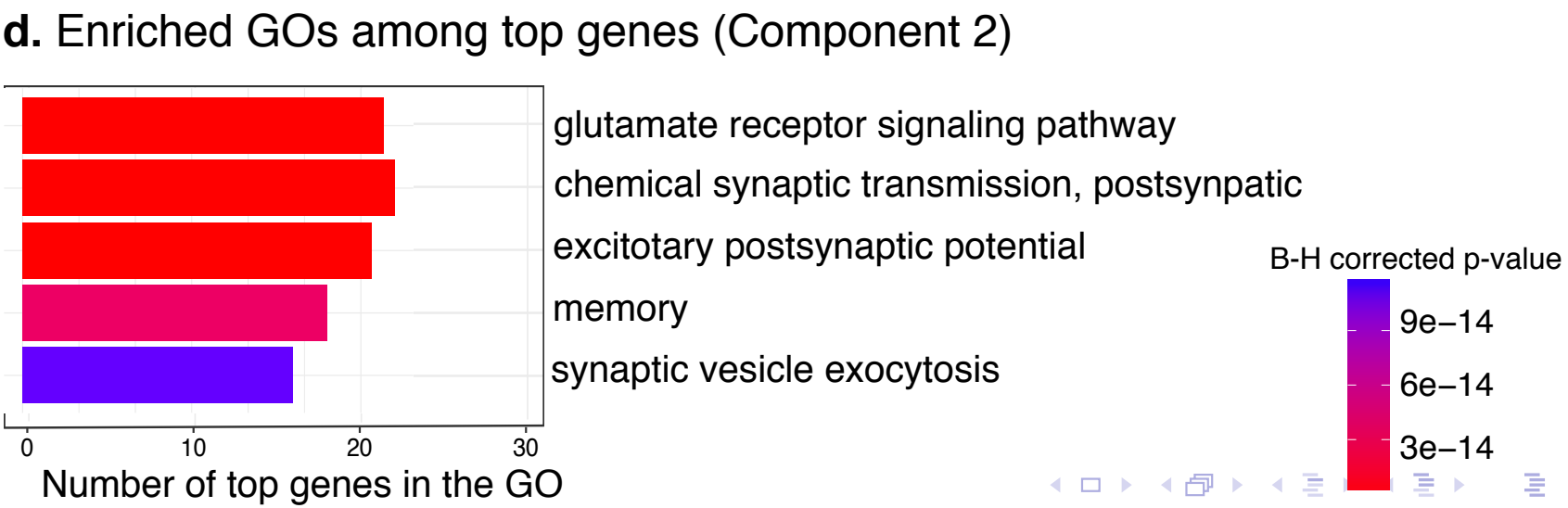


Component II: brain tissues

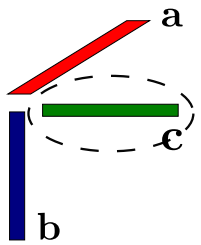


- Enriched GOs are mostly related to glutamate receptor signaling pathway, chemical synaptic transmission, and memory.

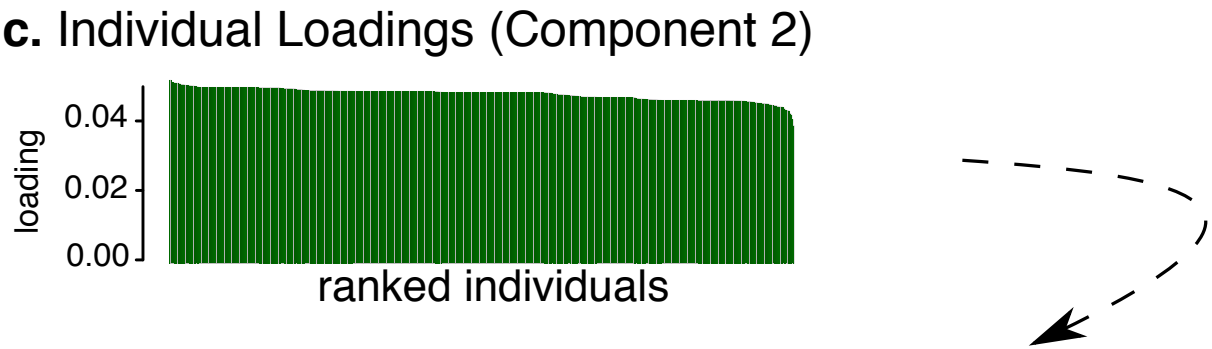




Component II: brain tissues



- Age explains more variation (24.4%) than gender (0.3%) or ethnicity (4.3%).



d.

Linear model: individual-vector 2 $\sim$ age+gender+ethnicity

Attribute	age	gender	ethnicity
Coefficient estimate	-7.3e-04	-8.6e-05	3.1e-04
P-value	**<2e-16**	0.12	2.3e-08
Proportion of variance explained	**24.4%**	0.3%	4.3%

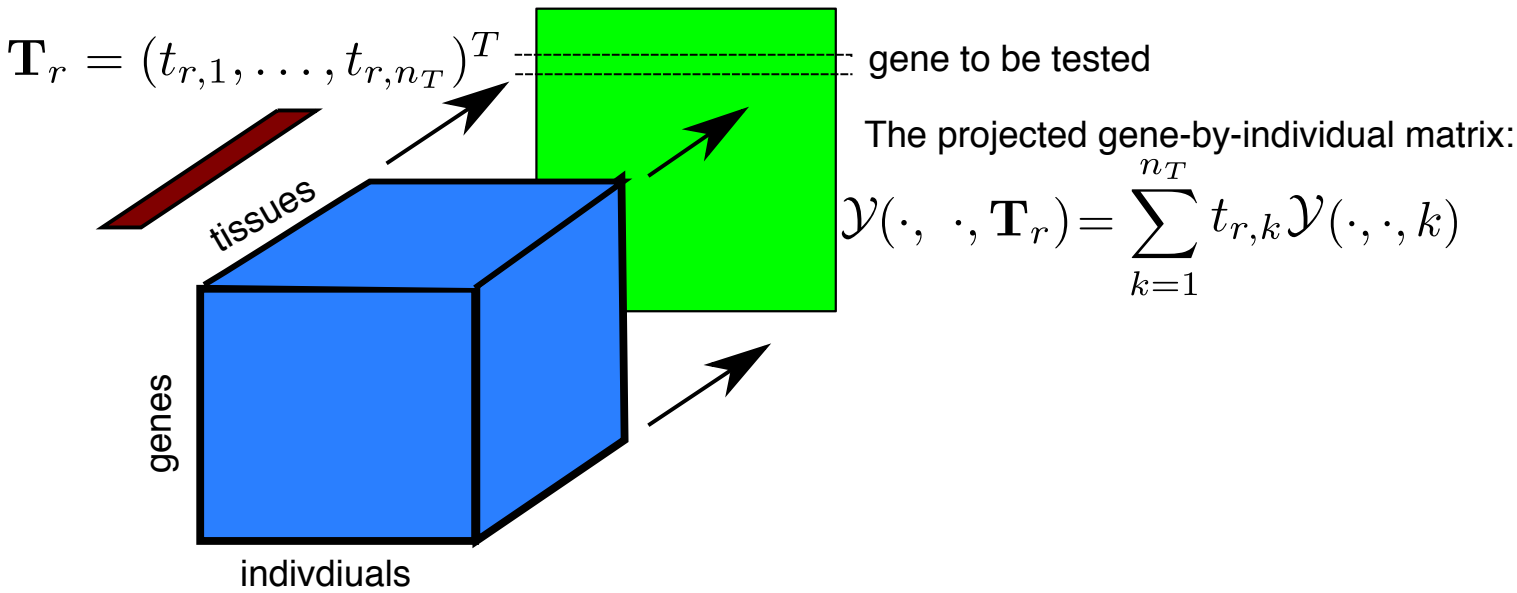
- How to pinpoint the age-related genes?

Tensor projection to detect differentially-expressed genes

- We project the expression tensor through the tissue-vector.
- For a given test gene, we perform the following regression analysis

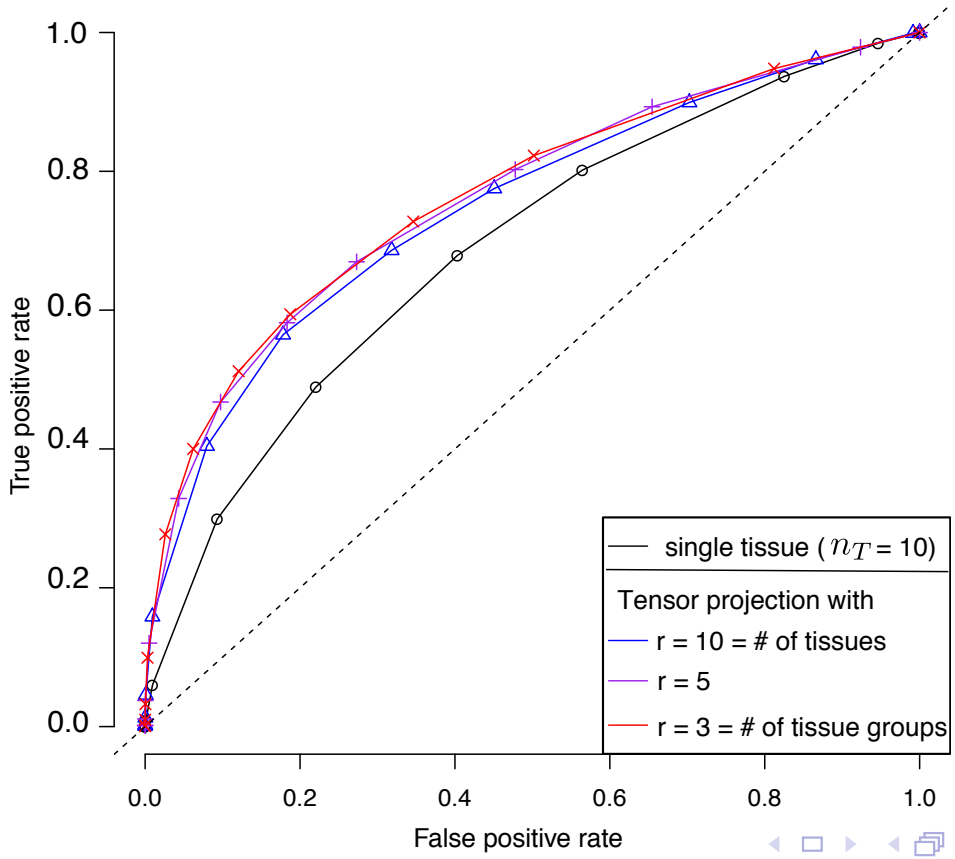
$$\mathcal{Y}(\text{test gene}, \cdot, \boldsymbol{T}_r) = \boldsymbol{\beta_0 W} + \boldsymbol{\beta_1 X} + \boldsymbol{\varepsilon},$$

where **W** is the covariate, and **X** is the primary variable of interest, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I})$. Consider $\mathcal{H}_0 : \beta_1 = 0$ vs. $\mathcal{H}_\alpha : \beta_1 \neq 0$.



Power to detect differentially-expressed (DE) genes

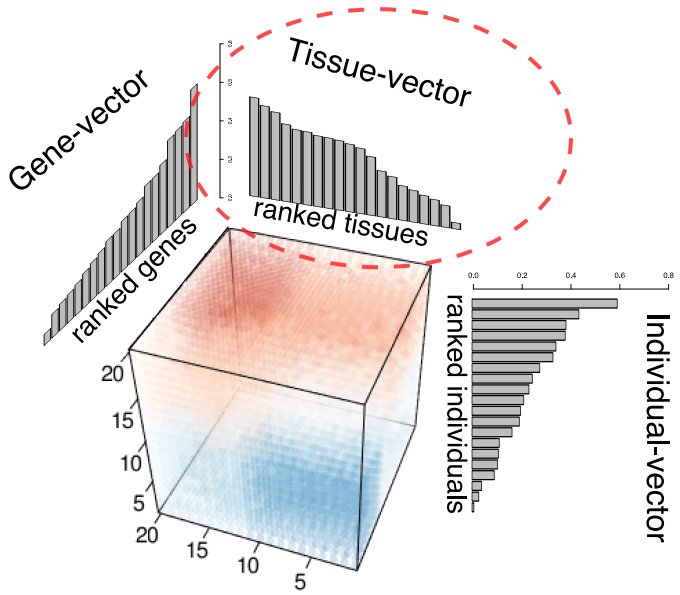
- Using tensor-projection procedures, we identified 694 age-related genes in the brain cluster. Comparatively, the classical single-tissue analysis identifies only 514 age-related genes.
- Receiver operating characteristic (ROC) plot from simulated expression data with age-related genes.



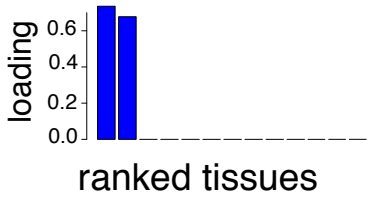
Gene expression in the brain

- We analyze 13 brain tissues ("subtensor") and apply our tensor method to this subtensor.
- Expression modules are spatially restricted to specific brain regions, such as the two cerebellum tissues, three cortex tissues, and three basal ganglia tissues.

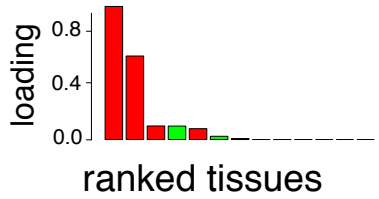
Recall: i^{th} tensor component, i = 1, 2, ...



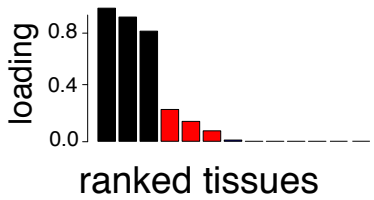
Tissue-vector (comp 2)



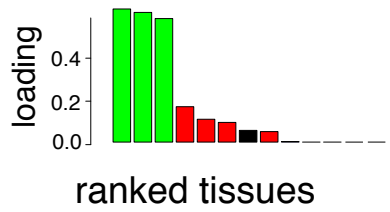
Tissue-vector (comp 3)



Tissue-vector (comp 4)



Tissue-vector (comp 5)



Cerebellum Cortex Basal ganglia
Others (Hippocampus, Amygdala, Spinal cord, etc)

Gene expression in the brain

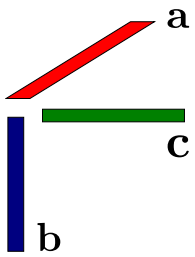
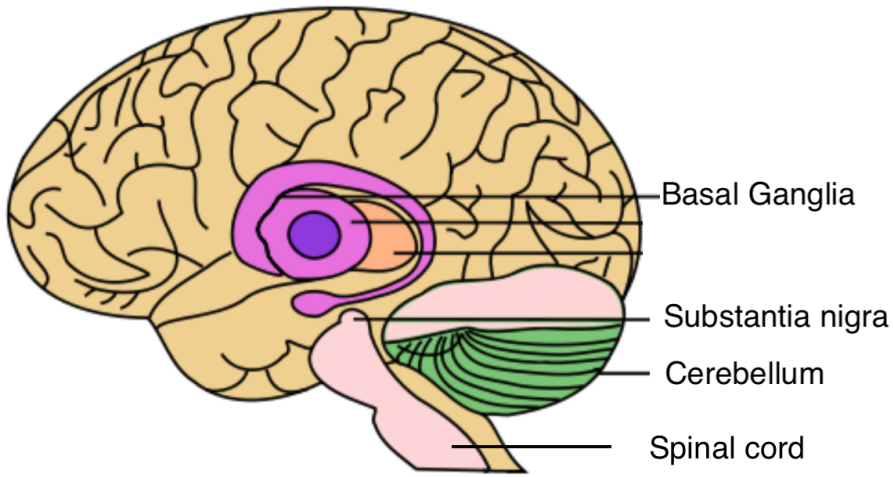


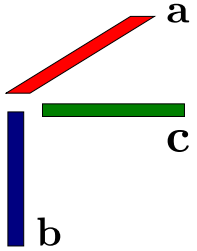
Table 1: Three-way clustering analysis of gene expression in the brain

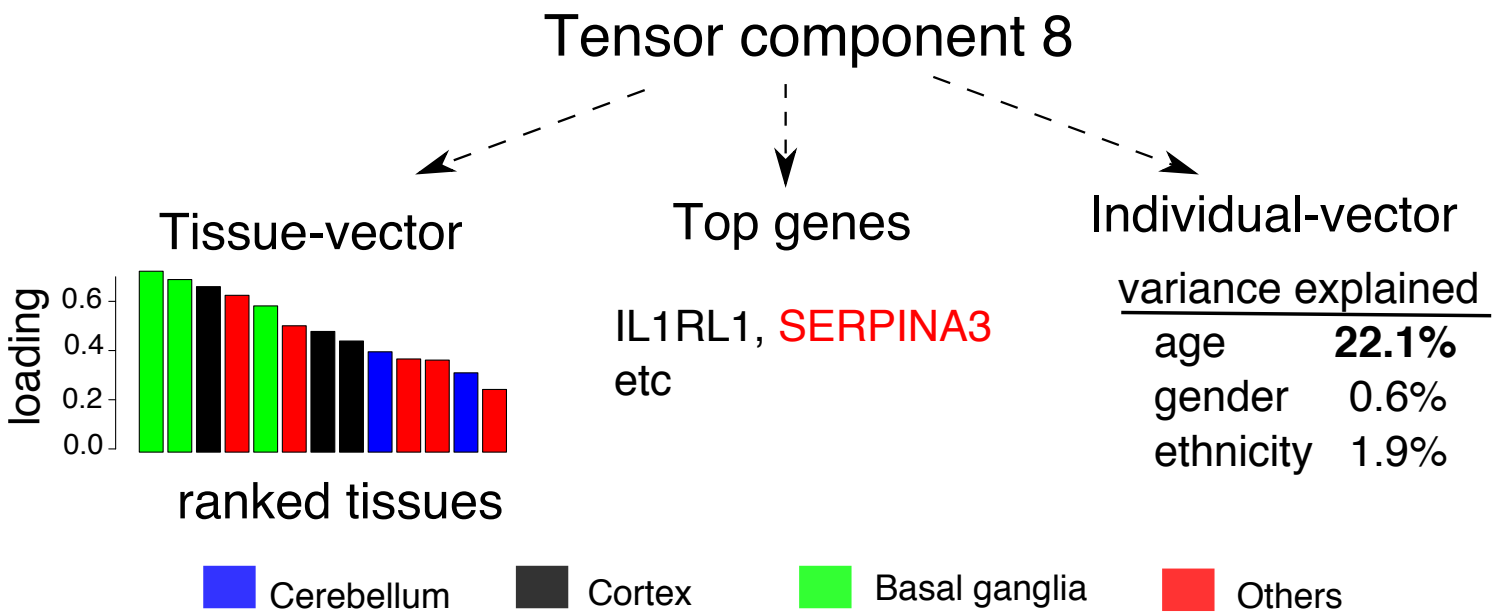
Tissue	Gene	Individual		
enriched region	enriched ontology	variance explained by		
		age	gender	ethnicity
cerebellum	dorsal spinal cord development	0.0%	8.0%	0.2%
cortex	behavior defense response	16.7%	0.6%	1.4%
basal ganglia	forebrain generation of neurons	1.3%	0.8%	1.7%
others	embryonic skeletal system morphogenesis	10.5%	0.7%	5.2%

Red number indicates p-value < 0.001.



Power to detect differentially-expressed genes





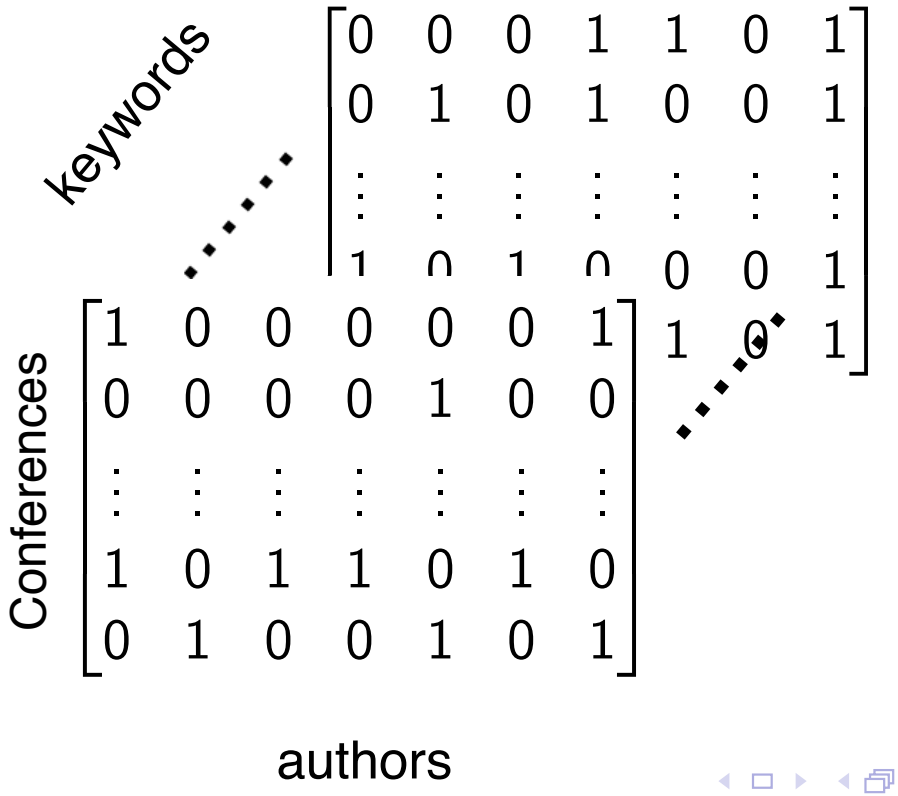
- The 2nd top gene SERPINA3 is implicated in Alzheimer's disease.
- This gene has a moderate age effect in each single tissue (p ranging from 0.05 to 1.1×10^{-6} with all effects in the same direction).
- Tensor projection method yields $p = 1.0 \times 10^{-8}$ for the age effect of this gene $\Rightarrow$ an increase of significance by two orders of magnitude.

Outline

- Tensor decomposition with applications in GTEx data
- Binary tensor decomposition and its statistical optimality

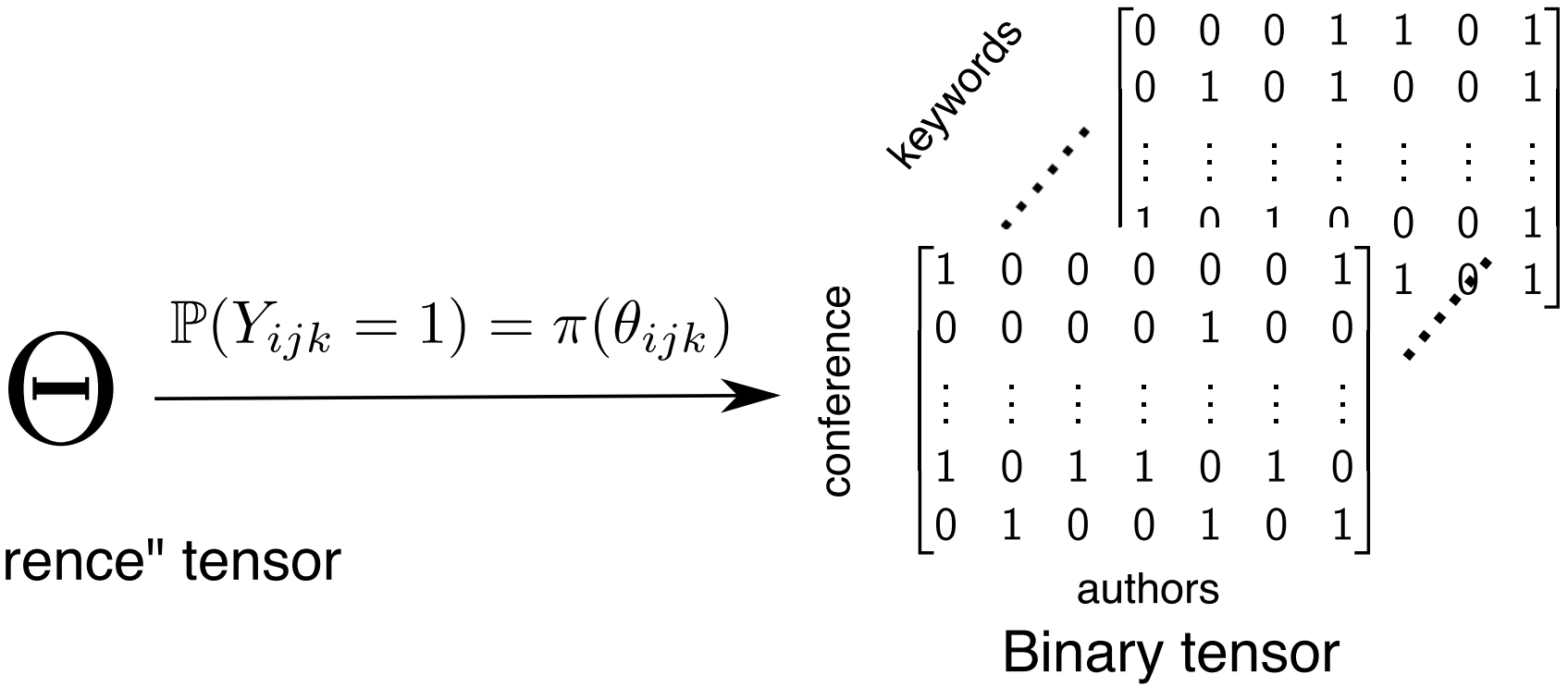
Bibliography dataset in DBLP

- Three-way bibliography relationships (author, conference, keyword) extracted from DBLP database.
- A tensor entry is 1 if the triplet co-occurs in a bibliography entry.
- Could be organized into a tensor of the form $\mathcal{Y} \in \{0,1\}^{10k \times 200 \times 10k}$ (10k authors, 200 conferences, and 10k keywords)



Probabilistic models for binary tensors

Directly applying rank-R approximation to $\mathcal{Y} = [\![Y_{ijk}]\!]$ is sub-optimal. why?
(symmetry w.r.t. $0 \leftrightarrow 1$; non-negativity; non-linearity)



"Preference" tensor

- Θ is unknown. Θ has low rank.
- $\pi : \mathbb{R} \mapsto [0, 1]$ is a known function (e.g., the logistic curve).
- $\Theta \in \mathbb{R}^{d_1 \times d_2 \times d3}$, $\mathcal{Y} \in \{0, 1\}^{d_1 \times d_2 \times d_3}$.

Probabilistic models for binary tensors

Latent variable formulation:

- Assume an underlying real-valued tensor $\mathcal{Z} = [\![Z_{ijk}]\!]$:

$$Y_{ijk}|Z_{ijk} = \begin{cases} 1 & \text{if } Z_{ijk} \geq 0, \\ 0 & \text{if } Z_{ijk} < 0. \end{cases}$$

- The latent tensor $\mathcal{Z}$ is further modeled as:

$$\mathcal{Z} = \underbrace{\Theta}_{\text{signal}} + \underbrace{\mathcal{E}}_{\text{noise}}, \quad \text{where} \quad \text{Rank}(\Theta) \leq R,$$

  and the entries in $\mathcal{E} = [\![\varepsilon_{ijk}]\!]$ follow i.i.d. $N(0, \sigma^2)$.
- Assume R is known or otherwise could be estimated (e.g., using BIC).
- We observe $\mathcal{Y} = [\![Y_{ijk}]\!] = \{0, 1\}^{d_1 \times d_2 \times d_3}$.

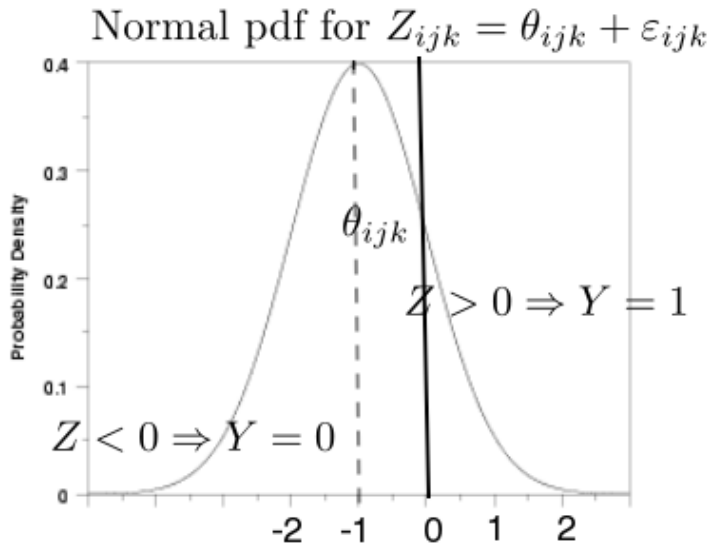
Connection with GLM

More generally, we can model the binary tensor $\mathcal{Y}$ as

$$\mathcal{Y} = [\![Y_{ijk}]\!] \sim_{\text{ind.}} \text{Bernoulli}(\pi(\Theta)), \quad \text{where} \quad \text{Rank}(\Theta) \leq R,$$

and $\pi : \mathbb{R} \mapsto [0, 1]$ is the link function (we allow π to be applied to tensors in a point-wise manner):

- Logistic link: $\pi(\theta) = \frac{e^\theta}{1+e^\theta} \iff$ latent noise $\mathcal{E}$ follows i.i.d. logistic distribution.
- Probit link: $\pi(\theta) = \Phi(\theta/\sigma)$, where $\Phi(\cdot)$ is the CDF for standard normal distribution $\iff$ latent noise $\mathcal{E}$ follows i.i.d. $N(0, \sigma^2)$.



Low-rank tensor estimation

- **Goal**: estimate low-rank $\Theta \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and/or $\pi(\Theta) \in [0,1]^{d_1 \times d_2 \times d_3}$.
- **Data**: $\mathcal{Y} \in \{0,1\}^{d_1 \times d_2 \times d_3}$.

Log-likelihood:

$$\mathcal{L}_{\mathcal{Y}}(\Theta) = \sum_{ijk} \left(\mathbb{1}_{(Y_{ijk}=1)} \log \pi(\theta_{ijk}) + \mathbb{1}_{(Y_{ijk}=0)} \log\left(1 - \pi(\theta_{ijk})\right) \right).$$

We propose to estimate Θ using the constrained MLE:

$$\widehat{\Theta}_{\text{MLE}} = \arg\max_{\Theta \in \mathcal{D}} \mathcal{L}_{\mathcal{Y}}(\Theta).$$

where $\mathcal{D}(R, \alpha) = \left\{ \Theta \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \text{rank}(\Theta) \leq R, \|\Theta\|_\infty \leq \alpha \right\}$.

Convergence rate

Define $\text{Loss}(\Theta_1, \Theta_2) = \frac{1}{\sqrt{\prod_k d_k}} \|\Theta_1 - \Theta_2\|_F$, where $\Theta_1, \Theta_2 \in \mathbb{R}^{d_1 \times \cdots \times d_K}$.

Error bound for constrained MLE (W. 2018)

Assume that $\Theta^{\text{true}} \in \mathcal{D}(R, \alpha)$ and regularity conditions on the link function π. Then with high probability,

$$\text{Loss}\left(\widehat{\Theta}_{\text{MLE}}, \Theta^{\text{true}}\right) \leq C \frac{L_\alpha}{\gamma_\alpha} \sqrt{\frac{\log(K) R^{K-1} \sum_k d_k}{\prod_k d_k}},$$

where

$$L_\alpha := \sup_{|\theta| \leq \alpha} \frac{\dot{\pi}(\theta)}{\pi(\theta)(1-\pi(\theta))}, \quad \gamma_\alpha = \inf_{|\theta| \leq \alpha} \left(\frac{\dot{\pi}^2(\theta)}{\pi^2(\theta)} - \frac{\ddot{\pi}(\theta)}{\pi(\theta)} \right) > 0.$$

Is this bound tight?

Special case: Probit link

$$L_{\alpha,\sigma} \approx \frac{\frac{\alpha}{\sigma}+1}{\sigma}, \quad \gamma_{\alpha,\sigma} \approx \frac{\alpha}{\sigma^3} e^{-\alpha^2/2\sigma^2}.$$

Theorem (W. 2018)

- Upper bound for constrained MLE:

$$\text{Loss}\left(\widehat{\Theta}_{\text{MLE}}, \Theta^{\text{true}}\right) \leq \sigma\left(\frac{\sigma}{\alpha}+1\right) e^{\alpha^2/2\sigma^2} C(R,K)\sqrt{\frac{d_{\max}}{\prod_k d_k}}.$$

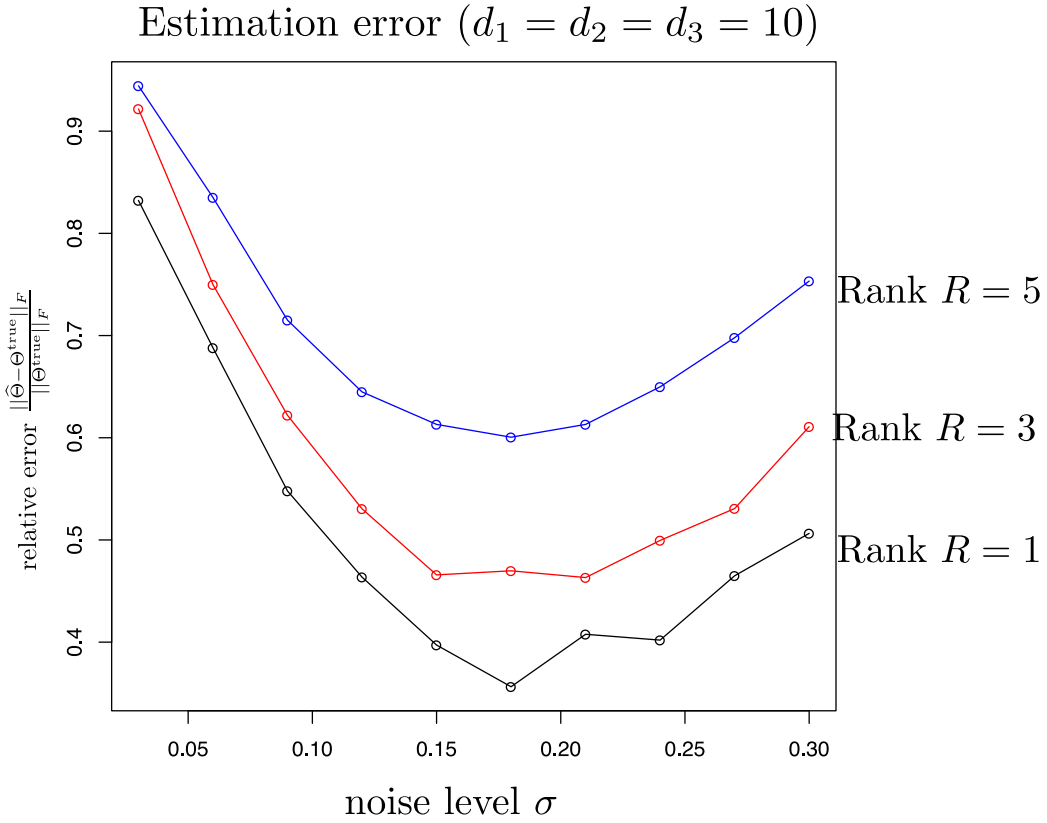
- Lower bound for any estimator:

$$\inf_{\widehat{\Theta}} \sup_{\Theta^{\text{true}} \in \mathcal{D}(R,\alpha)} \mathbb{E}\text{Loss}(\widehat{\Theta}, \Theta^{\text{true}}) \geq (\alpha \wedge \sigma) C'(R,K)\sqrt{\frac{d_{\max}}{\prod_k d_k}}.$$

	Signal-to-noise ratio (SNR)		
	$\sigma \ll \alpha$	$\sigma \approx \alpha$	$\sigma \gg \alpha$
Upper bound	$\sigma e^{1/2\sigma^2}$ ↘	σ	σ^2 ↗
Lower bound	σ	σ	α

Table: Error bound as a function of σ in three SNR regimes

Noise helps!



- When the noise and signal are comparable, our error bound for the constrained MLE is rate-optimal.
- When the noise is significantly smaller than the signal, increasing the noise improves the estimation. Counter-intuition?
- Phenomenon first observed for 1-bit matrix completion (Davenport et al, 2014, Cai-Zhou 2013).

Noise helps!

Intuition: If $\sigma = 0$, we cannot recover the magnitude of the underlying tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ from the observed binary tensor $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$.

- Consider the rank-1 probit model:

$$Y_{ijk} = \mathbb{1}_{\{(\theta_{ijk} + \varepsilon_{ijk}) > 0\}}, \text{ where } \Theta = [\![\theta_{ijk}]\!] = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}, \ \varepsilon_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

- Now remove the noise:

$$Y_{ijk} = \mathbb{1}_{\{\theta_{ijk} > 0\}}, \quad \Theta = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}.$$

- We cannot distinguish $\boldsymbol{c} = (c_1, \ldots, c_d)^T$ and $\tilde{\boldsymbol{c}} = (c_1\lambda_1, \ldots, c_d\lambda_d)^T$ if $\lambda_i > 0$ for all $i \in [d]$.

Computation

Optimization via iterative GLM

$$\max_{\Theta=[\![\theta_{ijk}]\!]} \mathcal{L}_{\mathcal{Y}}(\Theta) = \sum_{ijk} \log \pi\left(q_{ijk}\theta_{ijk}\right),$$

$$\text{subject to } \Theta = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r, \quad \|\Theta\|_\infty \leq \alpha,$$

where $q_{ijk} = 2Y_{ijk} - 1$ taking values -1 or 1 and $\pi(\cdot)$ is the link function.

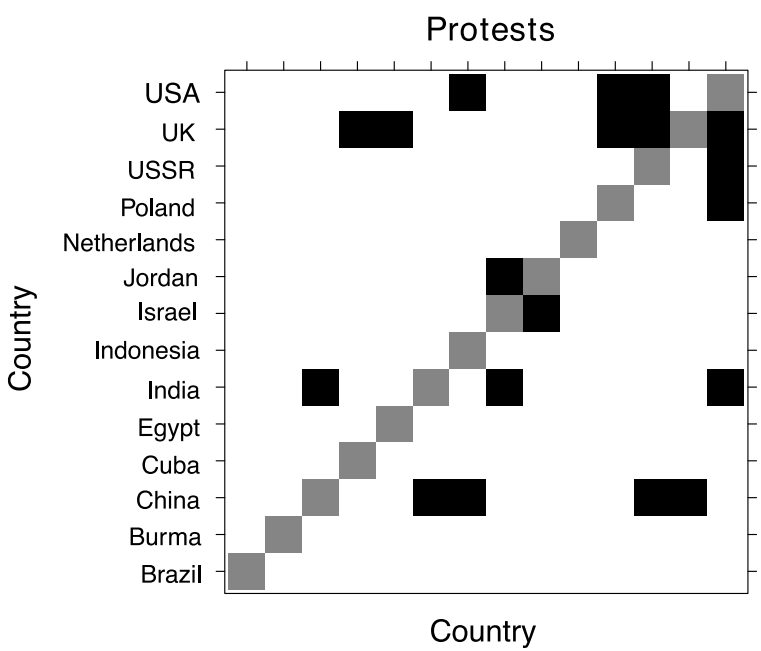
- Denote by $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_R]$, $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R]$ and $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_R]$.
- Objective function is multi-convex in $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$.
- Update $\boldsymbol{A}^{t+1} \leftarrow \max_{\boldsymbol{A}} \mathcal{L}_{\mathcal{Y}}(\boldsymbol{A}; \boldsymbol{B}^t, \boldsymbol{C}^t)$; solved by GLM.
- Modify $\boldsymbol{A}^{t+1} \leftarrow \rho\boldsymbol{A}^t + (1-\rho)\boldsymbol{A}^{t+1}$ subject to infinity-norm constraint; 1-dimension optimization over $\rho \in [0, 1]$.

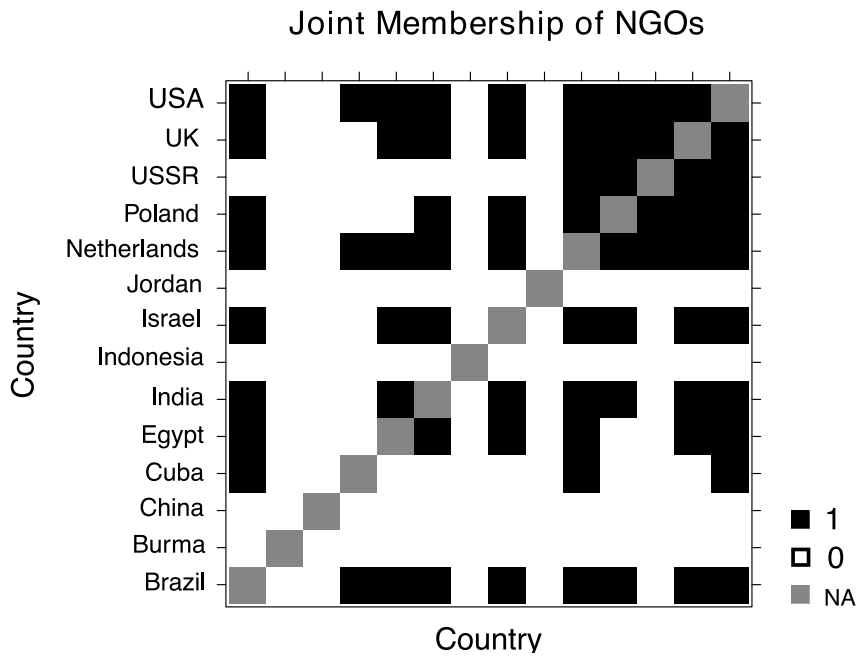
Experiments with real data

- **Nations**. $14 \times 14 \times 56$ binary tensor consisting of 56 relations among 14 countries.
- **Human connectome project (HCP)**. $68 \times 68 \times 202$ binary tensor consisting of connectivity patterns between 68 brain regions among 202 individuals.
- **Enron**. $581 \times 124 \times 48$ binary tensor depicting three-way relationships (sender, receiver, time) from the Enron email dataset.
- **Kinship**. $104 \times 104 \times 26$ binary tensor consisting of 26 types of relations among 104 individuals.

Political relationships in 1950-1965

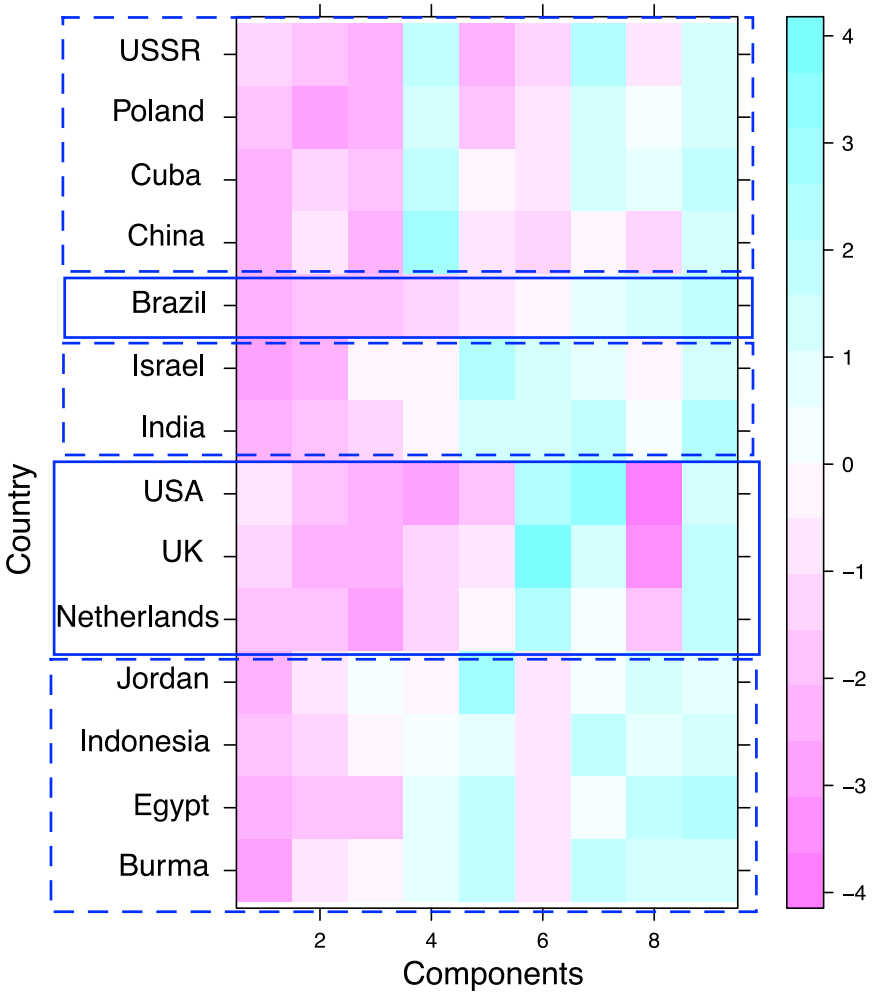
- 56 types of relationships among 14 countries between 1950-1965.
- Multi-relational data can be organized as a tensor $\mathcal{Y} \in \{0,1\}^{14\times14\times56}$. A tensor entry is 1 if the relation holds between two countries.
- Relations: "sends tourists to", "exports books to", "joint membership of NGOs", "conferences", etc.
- Countries: USSR, Poland, China, UK, Brazil, etc.



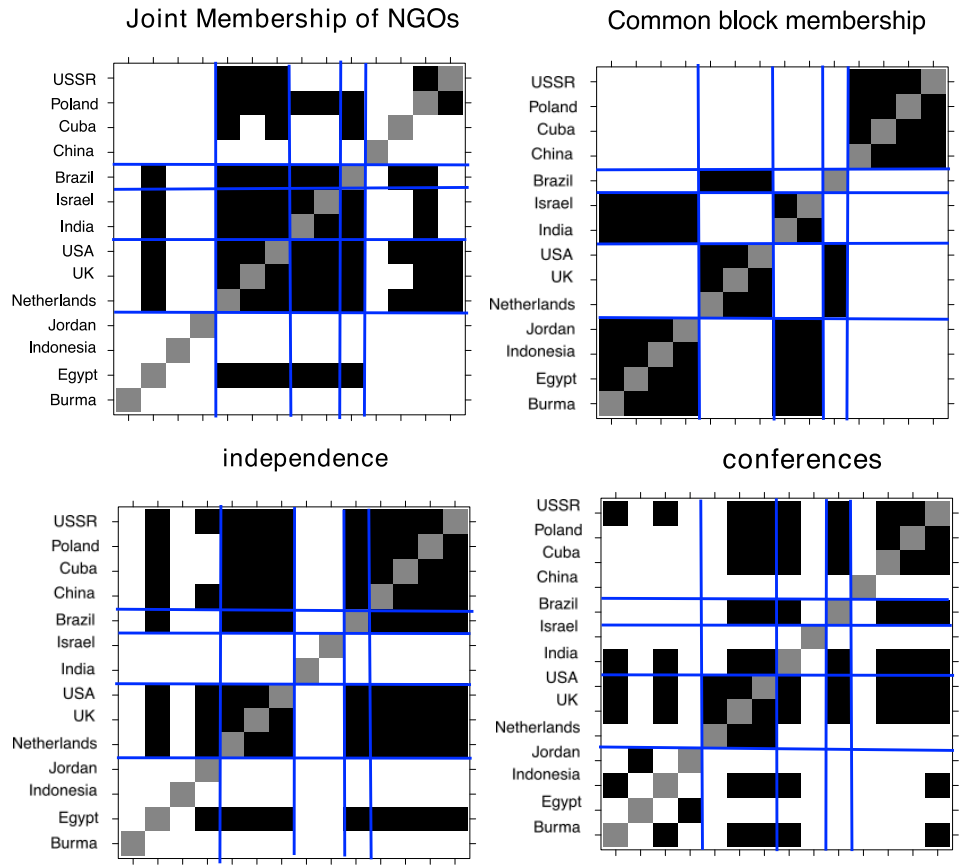


Estimates from binary tensor decomposition

Binary tensor components $\mathbf{a}_1, \ldots, \mathbf{a}_9$



Relation types with large loadings

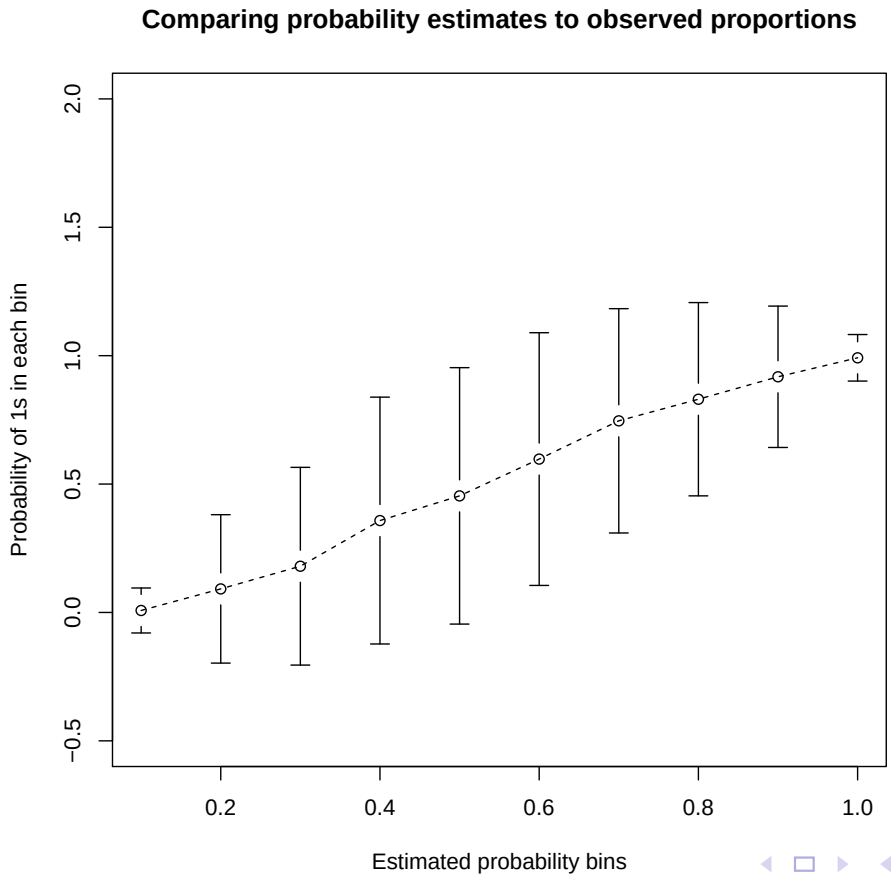


Nations dataset

The 14 countries are partitioned into

- One group containing countries from the communist bloc.
- Two groups from the western bloc; Brazil is separated from the rest.
- Two groups for the neutral bloc.

Goodness-of-fit: How well do we estimate individual probabilities?



Comparison with real-valued tensor decomposition

- Binary tensor decomposition has better out-of-sample performance compared to classical CP tensor decomposition.
- Link prediction from 5-fold cross-validation:

dataset	non-zeros	tensor decomposition method			
		binary (logistic link)		real-valued	
		AUC	MSE	AUC	MSE
HPC	35.3%	0.9860	1.3×10^{-3}	0.9314	1.4×10^{-2}
Nations	21.1%	0.9169	1.1×10^{-2}	0.8619	2.2×10^{-2}
Kinship	3.8%	0.9708	1.2×10^{-4}	0.9436	1.4×10^{-3}
Enron	0.01%	0.9432	6.4×10^{-3}	0.7956	6.3×10^{-5}

AUC: area under the receiver operating characteristic (ROC) curve;

MSE: mean squared error

References:

- M. Wang and Y. S. Song. Tensor decomposition via two-mode higher-order SVD (HOSVD). **Journal of Machine Learning Research W&CP (AISTATS)**, Vol. 54, (2017) 614-622.
- M. Wang, K. Dao Duc, J. Fischer, and Y. S. Song. Operator norm inequalities between tensor unfoldings on the partition lattice. **Linear Algebra and its Applications**, Vol. 520 (2017) 44-66.
- M. Wang, J. Fischer, and Y. S. Song. Three-way clustering of multi-tissue multi-individual gene expression data via semi-nonnegative tensor decomposition. **Under Revision in the Annals of Applied Statistics**. 2018.
- M. Wang, Leaning from binary multiway data: probabilistic tensor decomposition and its statistical optimality. *Preprint*.
- M. Wang, L. Li. Generalized linear models for tensor-response regression. *In preparation.*

Acknowledgements:

- We thank Lexin Li, Lek-Heng Lim, and Junhyong Kim for helpful discussions.
- Simons Math + X Research Grant from the Simons Foundation, a Packard Fellowship for Science and Engineering (to Y.S.S).