Tensor-on-tensor regression

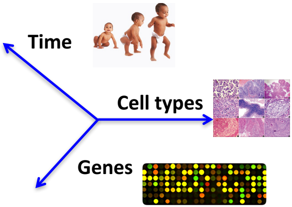
Eric F. Lock
(elock@umn.edu)

University of Minnesota
Division of Biostatistics

Texas A&M, 03/09/2018

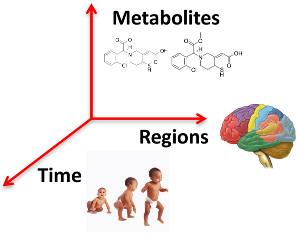
Multi-way (tensor) data

- Two-way data matrix $\mathbf{Y} : N \times Q$
  - N cases, Q features (genes, metabolites, voxels,...)
- Multi-way data $\mathbb{Y} : N \times Q_1 \times Q_2 \times \cdots \times Q_L$
  - E.g., N subjects, Q_1 genes, Q_2 cell types, Q_3 time points



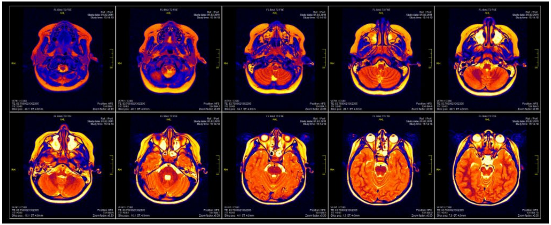
Multi-way (tensor) data

- Two-way data matrix $\mathbf{Y} : N \times Q$
  - N cases, Q features (genes, metabolites, voxels,...)
- Multi-way data $\mathbb{Y} : N \times Q_1 \times Q_2 \times \cdots \times Q_L$
  - E.g., N subjects, Q_1 metabolites, Q_2 brain regions, Q_3 time points



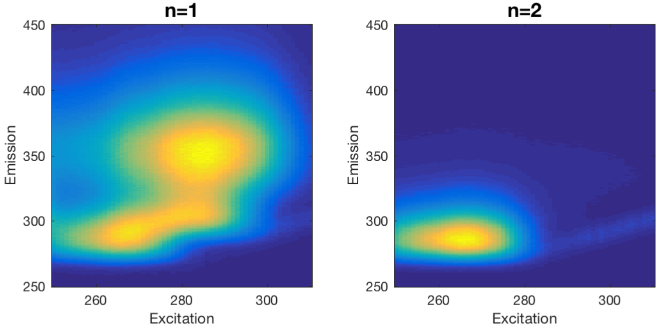
Multi-way (tensor) data

- Two-way data matrix $\mathbf{Y} : N \times Q$
  - N cases, Q features (genes, metabolites, voxels,...)
- Multi-way data $\mathbb{Y} : N \times Q_1 \times Q_2 \times \cdots \times Q_L$
  - E.g., N subjects, Q_1 time points, $Q_2 \times Q_3 \times Q_4$ voxel image



Application: Fluorescence data

- Spectrofluorometer applied to 5 chemical samples [1]
- Intensity measured for
  - 61 excitation wavelengths (250nm-310nm)
  - 201 emission wavelengths (250nm-450nm)
- $\mathbb{Y} : 5 \times 61 \times 201$



[1] http://www.models.life.ku.dk/nwaydata

Fluorescence: matrix dimension reduction

- Matricized data $\mathbf{Y} : 5 \times (61 \cdot 201) = 5 \times 12261$
- Principal components analysis / SVD factorization:

$$\mathbf{Y} \approx \mathbf{U}\mathbf{V}^T$$

where

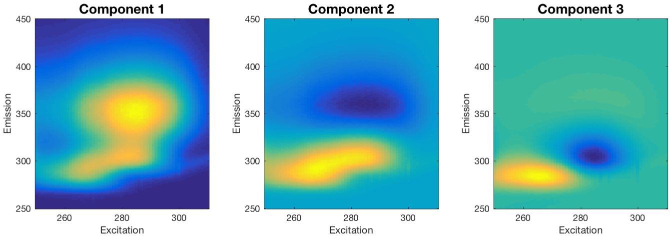
  - $\mathbf{U} : 5 \times R$ sample scores,
  - $\mathbf{V} : 12261 \times R$ feature loadings ($\mathbf{V}^T\mathbf{V} = \mathbf{I}$)

$$\mathbf{Y}[n, p] \approx \sum_{r=1}^{R} \mathbf{U}[n, r]\mathbf{V}[p, r]$$

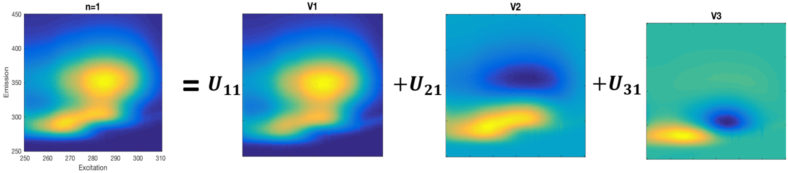
- 99% of variation in $\mathbf{Y}$ explained with $R = 3$ components

Fluorescence: matrix dimension reduction

- Components $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3$:



- Approximation for sample $n = 1$:



Candecomp/Parafac (CP) factorization

- Multi-way data $\mathbb{Y}: N \times Q_1 \times Q_2 \times \cdots \times Q_L$
- Rank-R CP factorization [RA Harshman, 1970]:

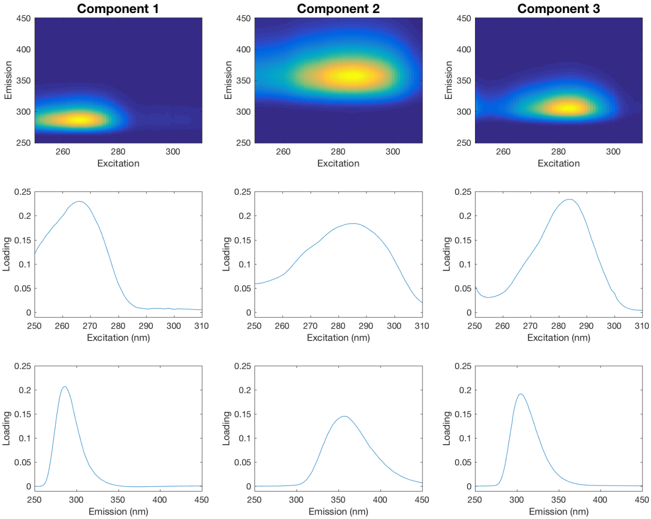
$$\mathbb{X} \approx [\![\mathbf{U}, \mathbf{V}_1, \ldots, \mathbf{V}_L]\!]$$

  - $\mathbf{U}: N \times R$
  - $\mathbf{V}_l : Q_l \times R$ for $l = 1, \ldots, L$.

$$\mathbb{X}[n, q_1, q_2, \ldots, q_L] \approx \sum_{r=1}^{R} \mathbf{U}[n, r] \prod_{l=1}^{L} \mathbf{V}_l[q_l, r]$$

- $||V_{lr}|| = 1$ for all $l = 1, \ldots L$ and $r = 1, \ldots, R$.

- 99% of variation explained with $R = 3$ CP components



Fluorescence data

- Each sample composed of 3 amino acids
  - Tryptophan (Trp), tyrosine (Tyr), phenylalanine (Phe)
- $\mathbf{X}: 5 \times 3$: concentration of each amino acid in Mole/L
- How does amino acid composition relate to fluorescence components?

SupCP

- Supervised CP factorization (SupCP):

$$\mathbb{Y} = [\![\mathbf{U}, \mathbf{V}_1, \ldots, \mathbf{V}_L]\!] + \mathbb{E}$$
$$\mathbf{U} = \mathbf{XB} + \mathbf{F}$$

where

  - $\mathbf{U} : N \times R$ is a latent score matrix for samples
  - $\{\mathbf{V}_l : Q_l \times R\}_{l=1}^{L}$ are loading matrices for each dimension
  - $\mathbb{E} : N \times Q_1 \times \cdots \times Q_L$ is error array with iid $N(0, \sigma_e^2)$ entries
  - $\mathbf{B} : P \times R$ are regression coefficients for $\mathbf{Y}$ on $\mathbf{U}$
  - $\mathbf{F} : N \times R$ has iid rows $\text{MVN}(\mathbf{0}, \Sigma_f)$

SupCP

- Maximize likelihood over $\{\mathbf{V}_1, \ldots, \mathbf{V}_L, \mathbf{B}, \Sigma_f, \sigma_e^2\}$
- Expectation Maximization (EM) algorithm
  - $\hat{L}(\{\mathbf{V}_l\}_{l=1}^L, \mathbf{B}, \Sigma_f, \sigma_e^2) = E_{\mathbf{U}}\ L(\mathbf{U}, \mathbb{X}, \mathbf{Y}, \{\mathbf{V}_l\}_{l=1}^L, \mathbf{B}, \Sigma_f, \sigma_e^2)$
  - Update $\{\{\mathbf{V}_l\}_{l=1}^L, \mathbf{B}, \Sigma_f, \sigma_e^2\}$ to maximize $\hat{L}(\{\mathbf{V}_l\}_{l=1}^L, \mathbf{B}, \Sigma_f, \sigma_e^2)$
- Predictive model of $\mathbb{Y}$ from $\mathbf{X}$:

$$\mathbb{Y} = [\![\mathbf{XB}, \mathbf{V}_1, \ldots, \mathbf{V}_L]\!] + [\![\mathbf{F}, \mathbf{V}_1, \ldots, \mathbf{V}_L]\!] + \mathbb{E}$$

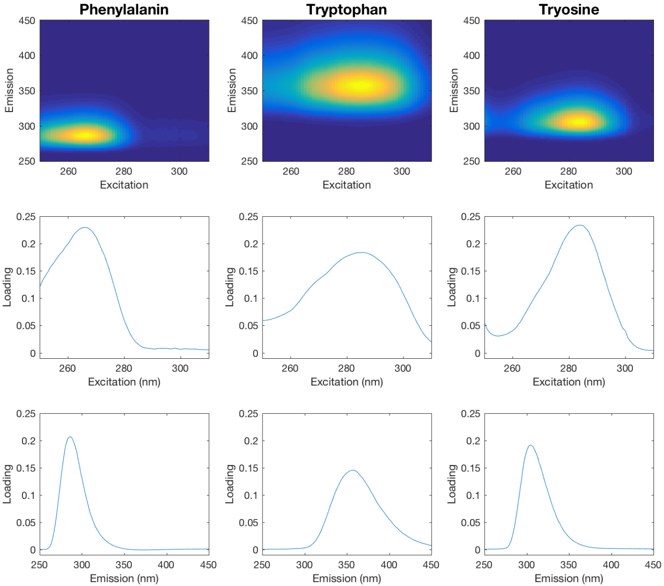
  - Prediction of $\mathbb{Y}$ given $\mathbf{X}$ $E(\mathbb{Y} \mid \mathbf{X})$
  - Structured residual variation in $\mathbb{Y}$
  - IID error in $\mathbb{Y}$

Fluorescence: SupCP

- Likelihood cross-validation selects $R = 3$
- Scaled coefficients for rank-3 fluorescence model:

	Component 1	Component 2	Component 3
$B_{\text{Phe}} * \text{sd}(X_{\text{Phe}})$	**7638**	138	130
$B_{\text{Trp}} * \text{sd}(X_{\text{Trp}})$	140	**11734**	-5
$B_{\text{Tyr}} * \text{sd}(X_{\text{Tyr}})$	87	-72	**8514**

Fluorescence: SupCP



SupCP: Special cases

- For a matrix $\mathbf{Y} : N \times Q$, SupCP reduces to SupSVD (G Li et al, JMVA, 2016)

- Reduces to a least-squares CP factorization as

$$\min_r \Sigma_f[r, r]/\sigma_e^2 \to \infty$$

- Reduces to least-squares tensor response regression as

$$||\Sigma_f||_F^2/\sigma_e^2 \to 0$$

  - Predict $\mathbb{Y} : N \times Q_1 \times \cdots \times Q_L$ from $\mathbf{X} : N \times P$.

Tensor-on-Tensor regression

- Predict $\mathbb{Y} : N \times Q_1 \times \cdots \times Q_M$ from $\mathbb{X} : N \times P_1 \times \cdots \times P_L$
- Special case: $\mathbf{Y} : N \times Q$, $\mathbf{X} : N \times P$
  - Partial least squares, reduced rank regression
- Special case: $Y : N \times 1$, $\mathbb{X} : N \times P_1 \times \cdots \times P_L$
  - Tensor regression ([H Zhou, L Li, & H Zhu; 2013] & others)
- Special case: $\mathbb{Y} : N \times Q_1 \times \cdots \times Q_M$, $\mathbf{X} : N \times P$
  - Tensor response regression ([L Li & X Zhang; 2016] & others)

Application: facial images

- 2000 frontalized facial images from different individuals [2]
- Each image 90×90 pixels, over 3 colors

- Data $\mathbb{X}$: Individuals $\times$ X $\times$ Y $\times$ Color

[2] Labeled Face in the Wild: `http://vis-www.cs.umass.edu/lfw/`

Application: facial images

- 72 describable attributes measured for each face [3]
  - Smiling / not smiling
  - male / female
  - beard/ no beard
- Measured on continuous scale
  - Smiling: positive values, not smiling: negative values.
- Predict $\mathbf{Y} : N \times Q$ from $\mathbb{X} : N \times P_1 \times P_2 \times P_3$
  - Facial attributes $\mathbf{Y}$: Individual $\times$ Attribute, from
  - Facial images $\mathbb{X}$: Individual $\times$ X $\times$ Y $\times$ Color

[3] [N Kumar, AC Berg, PN Belhumeur, & SK Nayar; 2009]

Contracted tensor product

- $\mathbb{A} : I_1 \times \cdots \times I_K \times P_1 \times \cdots \times P_L$ and $\mathbb{B} : P_1 \times \cdots \times P_L \times J_1 \times \cdots \times J_M$
- Define the *contracted tensor product*

$$\langle \mathbb{A}, \mathbb{B} \rangle_L : I_1 \times \cdots \times I_K \times J_1 \times \cdots \times J_M$$

by $\langle \mathbb{A}, \mathbb{B} \rangle_L [i_1, \dots i_K, j_1, \dots, j_M]$

$$= \sum_{p_1=1}^{P_1} \cdots \sum_{p_L=1}^{P_L} \mathbb{A}[i_1, \dots i_K, p_1, \dots, p_L] \mathbb{B}[p_1, \dots p_L, j_1, \dots, j_M].$$

- For matrices $\mathbf{A} : I \times P$ and $\mathbf{B} : P \times Q$,

$$\langle \mathbf{A}, \mathbf{B} \rangle_1 = \mathbf{AB},$$

Tensor-on-tensor regression

- Predict $\mathbb{Y} : N \times Q_1 \times \cdots \times Q_M$ from $\mathbb{X} : N \times P_1 \times \cdots \times P_L$:

$$\mathbb{Y} = \langle \mathbb{X}, \mathbb{B} \rangle_L + \mathbb{E}$$

  - $\mathbb{B} : P_1 \times \cdots \times P_L \times Q_1 \times \cdots \times Q_M$ is a coefficient array
  - $\mathbb{E} : N \times Q_1 \times \cdots \times Q_M$ is an error array

$$\mathbb{Y}[n, q_1, \ldots, q_M] \approx \sum_{p_1}^{P_1} \cdots \sum_{p_L}^{P_L} \mathbb{X}[N, p_1, \ldots, p_L] \mathbb{B}[p_1, \ldots, p_L, q_1, \ldots, q_M]$$

- Supports all valid linear relations between $\mathbb{X}$ and $\mathbb{Y}$

Candecomp/Parafac (CP) factorization

- Assume $\mathbb{B}$ has a rank R CP factorization:

$$\mathbb{B} = [\![\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M]\!]$$

  - $\mathbf{U}_l : P_l \times R$
  - $\mathbf{V}_m : Q_m \times R$

$$\mathbb{B}[p_1, \ldots, p_L, q_1, \ldots, q_M] = \sum_{r=1}^{R} \prod_{l=1}^{L} \mathbf{U}_l[p_l, r] \prod_{m=1}^{M} \mathbf{V}_m[q_m, r]$$

- Full dimension of $\mathbb{B}$: $\prod_{l=1}^{L} P_l \prod_{m=1}^{M} Q_m$
- Reduced dimension of $\mathbb{B}$: $R\,(P_1 + \cdots + P_L + Q_1 + \cdots + Q_M)$

Penalized least squares estimation

- 1.) Unpenalized criterion:

$$\underset{\text{rank}(\mathbb{B}) \leq R}{\arg\min} ||\mathbb{Y} - \langle \mathbb{X}, \mathbb{B} \rangle_L||_F^2.$$

- 2.) Separable L_p penalty. E.g., L_2:

$$\underset{\text{rank}(\mathbb{B}) \leq R}{\arg\min} ||\mathbb{Y} - \langle \mathbb{X}, \mathbb{B} \rangle_L||_F^2 + \lambda \sum_{l=1}^{L} ||\mathbf{U}_l||_F^2 \sum_{m=1}^{M} ||\mathbf{V}_l||_F^2$$

- 3.) Global L_2 penalty:

$$\underset{\text{rank}(\mathbb{B}) \leq R}{\arg\min} ||\mathbb{Y} - \langle \mathbb{X}, \mathbb{B} \rangle_L||_F^2 + \lambda ||\mathbb{B}||_F^2$$

- Iteratively update $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M$ to optimize objective.

Special cases

- For $\mathbf{X} = N \times P$ and $\mathbf{Y} : N \times Q$:
- No penalty (1) corresponds to reduced rank regression [AJ Izenman, 1975]

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

- $\mathbf{B} : P \times Q$ has a rank R factorization

$$\mathbf{B} = \mathbf{U}\mathbf{V}^T$$

  - $\mathbf{U} : P \times R$
  - $\mathbf{V} : Q \times R$

Special cases

- For $\mathbb{X} : N \times P$, $\mathbb{Y} : N \times 1$, global penalty (3) corresponds to ridge regression
- For $\mathbb{X} : N \times P$, $\mathbb{Y} : N \times Q$, global penalty (3) corresponds to reduced rank ridge regression [A Mukherjee and Ji Zhu, 2011]
- For $\mathbb{X} : N \times P_1 \times \cdots \times P_L$, $\mathbb{Y} : N \times 1$, separable penalty (2) corresponds to tensor regression [H Zhou, L Li, and H Zhu, 2013]
- For $\mathbb{X} : N \times P_1 \times \cdots \times P_L$, $\mathbb{Y} : N \times 1$, separable penalty (3) corresponds to tensor ridge regression [W Guo, I Kotsia, I Patras, 2012]
- Whenever $\mathbb{B}$ is a matrix ($P \times Q$, $P_1 \times P_2$,$Q_1 \times Q_2$) separable L_2 penalty (2) corresponds to a nuclear norm penalty on $\mathbb{B}$.

Convergence

- Iterative algorithms prone to local optima for objectives 1,2,3.
- Remedies:
  - **Tempered regularization**: start with larger λ that gradually decreases to desired regularization
  - **Simulated annealing**: add decreasing level of random noise at each update

Consistency

- Assumptions:
  - Error array $\mathbb{E}$ has mean 0 and finite second moment
  - True coefficient array $\mathbb{B}_0$ has rank R
  - $\theta_0 = \{\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M\}$ is identifiable and in interior of compact space Θ
  - $\mathbb{X}$ is bounded
- Result:
  - For $R = R_0$ and $\lambda \geq 0$, $\hat{\mathbb{B}}_N \xrightarrow{p} \mathbb{B}_0$ as $n \to \infty$ under objectives 1.), 2.), or 3.).

MCMC Inference

- Global L_2 penalty (3.) gives mode of posterior for Bayesian model
  - Errors $\mathbb{E}$ are iid Normal$(0, \sigma^2)$.
  - Prior $\mathbb{B}$ is spherical Gaussian over rank R tensors:

$$\text{pr}(\mathbb{B}) \propto \begin{cases} \exp\left(-\frac{\lambda}{2\sigma^2}||\mathbb{B}||_F^2\right) \text{ if } \text{rank}(\mathbb{B}) \leq R. \\ 0 \text{ otherwise,} \end{cases}$$

- Choose prior for σ^2: $\text{pr}(\sigma^2) \propto 1/\sigma^2$

MCMC Inference

- Gibbs sample from full conditional distributions of

$$\{\sigma^2, \mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M.\}$$

- For t'th sample,

$$\mathbb{B}^{(t)} = [\![\mathbf{U}_1^{(t)}, \ldots, \mathbf{U}_L^{(t)}, \mathbf{V}_1^{(t)}, \ldots, \mathbf{V}_M^{(t)}]\!].$$

- For new data $\mathbb{X}_{\text{new}} : \tilde{N} \times P_1 \times \cdots \times P_L$, simulate outcomes

$$\mathbb{Y}_{\text{new}}^{(t)} = \langle \mathbb{X}_{\text{new}}, \mathbb{B}^{(t)} \rangle_L + \mathbb{E}_{\text{new}}^{(t)},$$

  where $\mathbb{E}_{\text{new}}^{(t)}$ is generated with independent $\text{N}(0, \sigma^{2(t)})$ entries.

Application: facial images

- Predict $\mathbf{Y}$: Individual $\times$ Attribute from $\mathbb{X}$: Individual $\times$ X $\times$ Y $\times$ Color

- Split into training and test set of size 1000

- For estimate $\hat{\mathbb{B}}$, consider RPE for test data:

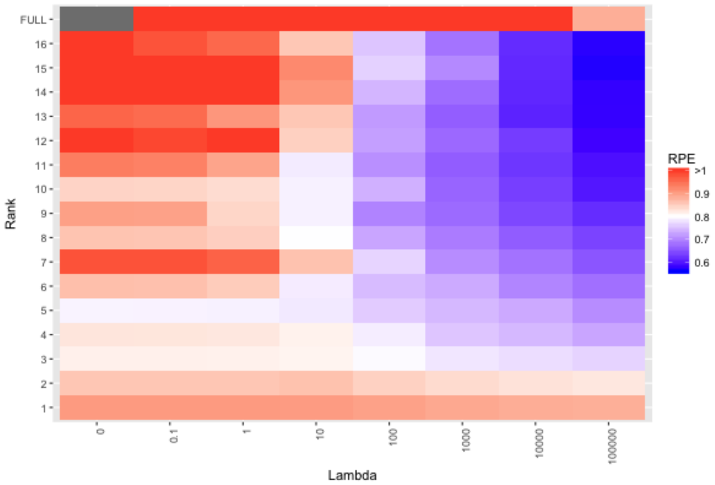
$$\text{RPE} = \frac{||\mathbb{Y}\text{new} - \langle \mathbb{X}\text{new}, \hat{\mathbb{B}} \rangle_L||_F^2}{||\mathbb{Y}\text{new}||_F^2}.$$

- Choose λ and R to minimize test RPE

- Inference via 5000 Gibbs samples

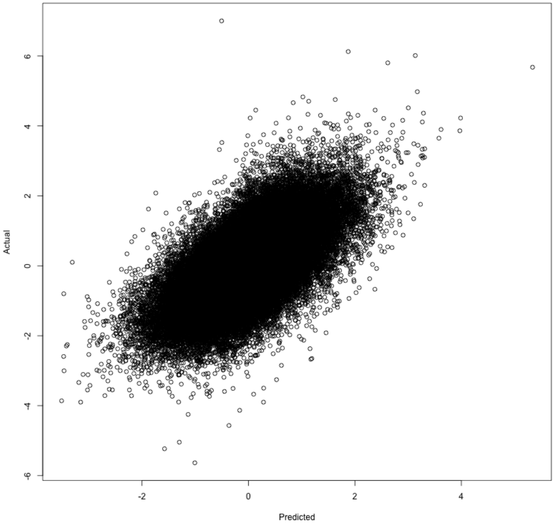
Application: facial images

Relative prediction error (RPE):



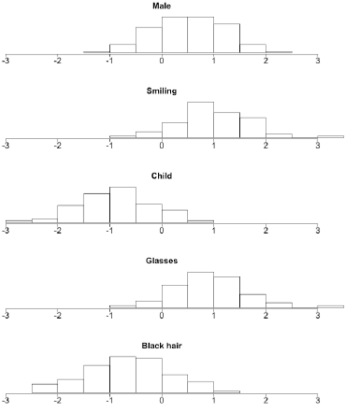
Application: facial images

Predicted vs. actual outcome for test data:



Application: facial images

- Bayesian posterior densities
  - 90% credible interval coverage rate: 0.887
  - 95% credible interval coverage rate: 0.934



Application: facial images

- Alternatively, predict

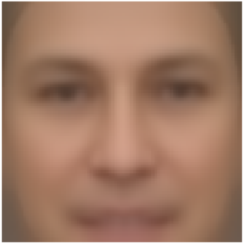
$$\mathbb{Y} : \text{Individual} \times \text{X} \times \text{Y} \times \text{Color}$$

from

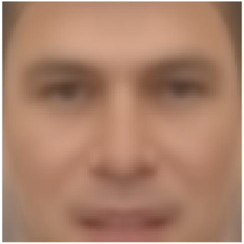
$$\mathbf{X} : \text{Individual} \times \text{Attribute}$$

- Use SupCP model with $R = 200$

Application: facial images

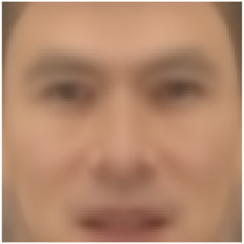


(a) Mean

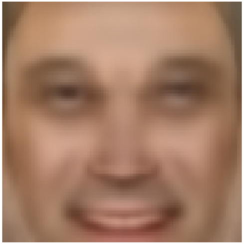


(b) Male

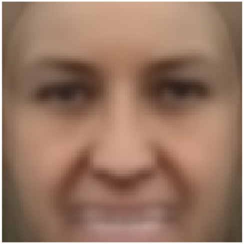
(c) Male; moustache



(d) Male; Asian



(e) Male; smiling



(f) Female; black; lipstick

Thank you!

- Email: elock@umn.edu
- SupCP
  - **Article**: Lock, EF and Li, G, "Supervised multiway factorization", *arXiv*: 1701.01037, 2016.
  - **Code**: `https://github.com/lockEF/SupCP`
- Tensor-on-tensor regression
  - **Article**: Lock, EF, "Tensor-on-tensor regression", *JCGS*, doi:10618600.2017.1401544, 2017.
  - **Code**: `https://github.com/lockEF/MultiwayRegression`
- Slides: `http://ericfrazerlock.com/Talks.html`