Algebraic models for higher-order correlations

Lek-Heng Lim and Jason Morton

U.C. Berkeley and Stanford Univ.

December 15, 2008

Tensors as hypermatrices

Up to choice of bases on U, V, W, a tensor $A \in U \otimes V \otimes W$ may be represented as a hypermatrix

$$\mathcal{A} = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$$

where $\dim(U) = l, \dim(V) = m, \dim(W) = n$ if

1. we give it coordinates;
2. we ignore covariance and contravariance.

Henceforth, tensor = hypermatrix.

Multilinear matrix multiplication

- Matrices can be multiplied on left and right: $A \in \mathbb{R}^{m \times n}$, $X \in \mathbb{R}^{p \times m}$, $Y \in \mathbb{R}^{q \times n}$,

$$C = (X, Y) \cdot A = XAY^{\top} \in \mathbb{R}^{p \times q},$$
$$c_{\alpha\beta} = \sum\nolimits_{i,j=1}^{m,n} x_{\alpha i} y_{\beta j} a_{ij}.$$

- 3-tensors can be multiplied on three sides: $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$, $X \in \mathbb{R}^{p \times l}$, $Y \in \mathbb{R}^{q \times m}$, $Z \in \mathbb{R}^{r \times n}$,

$$\mathcal{C} = (X, Y, Z) \cdot \mathcal{A} \in \mathbb{R}^{p \times q \times r},$$
$$c_{\alpha\beta\gamma} = \sum\nolimits_{i,j,k=1}^{l,m,n} x_{\alpha i} y_{\beta j} z_{\gamma k} a_{ijk}.$$

- Correspond to change-of-bases transformations for tensors.
- Define 'right' (covariant) multiplication by $(X, Y, Z) \cdot \mathcal{A} = \mathcal{A} \cdot (X^{\top}, Y^{\top}, Z^{\top})$.

Symmetric tensors

- Cubical tensor $[\![a_{ijk}]\!] \in \mathbb{R}^{n\times n\times n}$ is **symmetric** if

$$a_{ijk} = a_{ikj} = a_{jik} = a_{jki} = a_{kij} = a_{kji}.$$

- For order p, invariant under all permutations $\sigma \in \mathfrak{S}_p$ on indices.
- $\mathsf{S}^p(\mathbb{R}^n)$ denotes set of all order-p symmetric tensors.
- Symmetric multilinear matrix multiplication $\mathcal{C} = (X, X, X) \cdot \mathcal{A}$ where

$$c_{\alpha\beta\gamma} = \sum\nolimits_{i,j,k=1}^{l,m,n} x_{\alpha i} x_{\beta j} x_{\gamma k} a_{ijk}.$$

Examples of symmetric tensors

- Higher order derivatives of real-valued multivariate functions.
- Moments of a vector-valued random variable $\mathbf{x} = (x_1, \dots, x_n)$:

$$\mathcal{S}_p(\mathbf{x}) = \left[E(x_{j_1} x_{j_2} \cdots x_{j_p})\right]_{j_1,\dots,j_p=1}^{n}.$$

- Cumulants of a random vector $\mathbf{x} = (x_1, \dots, x_n)$:

$$\mathcal{K}_p(\mathbf{x}) = \left[\sum_{A_1 \sqcup \dots \sqcup A_q = \{j_1,\dots,j_p\}} (-1)^{q-1}(q-1)! E\left(\prod_{j \in A_1} x_j\right) \cdots E\left(\prod_{j \in A_q} x_j\right)\right]_{j_1,\dots,j_p=1}^{n}.$$

$\mathcal{K}_p(x)$ for $p = 1, 2, 3, 4$ are expectation, variance, skewness, and kurtosis.

Cumulants

- In terms of log characteristic and cumulant generating functions,

$$\kappa_{j_1\cdots j_p}(\mathbf{x}) = \frac{\partial^p}{\partial t_{j_1}^{\alpha_1}\cdots\partial t_{j_p}^{\alpha_p}} \log \mathbf{E}(\exp(\langle \mathbf{t}, \mathbf{x}\rangle)\Big|_{\mathbf{t}=\mathbf{0}}$$
$$= (-1)^p \frac{\partial^p}{\partial t_{j_1}^{\alpha_1}\cdots\partial t_{j_p}^{\alpha_p}} \log \mathbf{E}(\exp(i\langle \mathbf{t}, \mathbf{x}\rangle)\Big|_{\mathbf{t}=\mathbf{0}}.$$

- In terms of Edgeworth expansion,

$$\log \mathbf{E}(\exp(i\langle \mathbf{t}, \mathbf{x}\rangle) = \sum_{\alpha=0}^{\infty} i^{|\alpha|}\kappa_\alpha(\mathbf{x})\frac{\mathbf{t}^\alpha}{\alpha!}, \quad \log \mathbf{E}(\exp(\langle \mathbf{t}, \mathbf{x}\rangle) = \sum_{\alpha=0}^{\infty} \kappa_\alpha(\mathbf{x})\frac{\mathbf{t}^\alpha}{\alpha!},$$

  $\alpha = (\alpha_1, \dots, \alpha_p)$ is a multi-index, $\mathbf{t}^{\boldsymbol{\alpha}} = t_1^{\alpha_1}\cdots t_p^{\alpha_p}$, $\alpha! = \alpha_1!\cdots\alpha_p!$.
- For each $\mathbf{x}$, $\mathcal{K}_p(\mathbf{x}) = [\kappa_{j_1\cdots j_p}(\mathbf{x})] \in \mathsf{S}^p(\mathbb{R}^n)$ is a symmetric tensor.
- [Fisher, Wishart; 1932]

Properties of cumulants

Multilinearity: If $\mathbf{x}$ is a $\mathbb{R}^n$-valued random variable and $A \in \mathbb{R}^{m\times n}$

$$\mathcal{K}_p(A\mathbf{x}) = (A, \dots, A) \cdot \mathcal{K}_p(\mathbf{x}).$$

Independence:

- If $\mathbf{x}_1, \dots, \mathbf{x}_k$ are mutually independent of $\mathbf{y}_1, \dots, \mathbf{y}_k$, then

$$\mathcal{K}_p(\mathbf{x}_1 + \mathbf{y}_1, \dots, \mathbf{x}_k + \mathbf{y}_k) = \mathcal{K}_p(\mathbf{x}_1, \dots, \mathbf{x}_k) + \mathcal{K}_p(\mathbf{y}_1, \dots, \mathbf{y}_k).$$

- If I and J partition $\{j_1, \dots, j_n\}$ so that $\mathbf{x}_I$ and $\mathbf{x}_J$ are independent, then

$$\kappa_{j_1 \cdots j_n}(\mathbf{x}) = 0.$$

Gaussian: If $\mathbf{x}$ is multivariate normal, then $\mathcal{K}_p(\mathbf{x}) = 0$ for all $p \geq 3$.

Support: There are no distributions where

$$\mathcal{K}_p(\mathbf{x}) \begin{cases} \neq 0 & 3 \leq p \leq n, \\ = 0 & p > n. \end{cases}$$

Estimation of cumulants

- How do we estimate $\mathcal{K}_p(\mathbf{x})$ given multiple observations of $\mathbf{x}$?
- Central and non-central moments are

$$\hat{m}_n = \frac{1}{n}\sum\nolimits_t (x_t - \bar{x})^n, \quad \hat{s}_n = \frac{1}{n}\sum\nolimits_t x_t^n, \quad \text{etc.}$$

- Cumulant estimator $\hat{\mathcal{K}}_p(\mathbf{x})$ for $p = 1, 2, 3, 4$ given by

$$\begin{aligned}
\hat{\kappa}_i &= \hat{m}_i = \tfrac{1}{n}\hat{s}_i \\
\hat{\kappa}_{ij} &= \tfrac{n}{n-1}\hat{m}_{ij} = \tfrac{1}{n-1}(\hat{s}_{ij} - \tfrac{1}{n}\hat{s}_i\hat{s}_j) \\
\hat{\kappa}_{ijk} &= \tfrac{n^2}{(n-1)(n-2)}\hat{m}_{ijk} = \tfrac{n}{(n-1)(n-2)}[\hat{s}_{ijk} - \tfrac{1}{n}(\hat{s}_i\hat{s}_{jk} + \hat{s}_j\hat{s}_{ik} + \hat{s}_k\hat{s}_{ij}) + \tfrac{2}{n^2}\hat{s}_i\hat{s}_j\hat{s}_k] \\
\hat{\kappa}_{ijk\ell} &= \tfrac{n^2}{(n-1)(n-2)(n-3)}[(n+1)\hat{m}_{ijk\ell} - (n-1)(\hat{m}_{ij}\hat{m}_{k\ell} + \hat{m}_{ik}\hat{m}_{j\ell} + \hat{m}_{i\ell}\hat{m}_{jk})] \\
&= \tfrac{n}{(n-1)(n-2)(n-3)}[(n+1)\hat{s}_{ijk\ell} - \tfrac{n+1}{n}(\hat{s}_i\hat{s}_{jk\ell} + \hat{s}_j\hat{s}_{ik\ell} + \hat{s}_k\hat{s}_{ij\ell} + \hat{s}_\ell\hat{s}_{ijk}) \\
&\qquad - \tfrac{n-1}{n}(\hat{s}_{ij}\hat{s}_{k\ell} + \hat{s}_{ik}\hat{s}_{j\ell} + \hat{s}_{i\ell}\hat{s}_{jk}) + \hat{s}_i^2(\hat{s}_{jk} + \hat{s}_{j\ell} + \hat{s}_{k\ell}) \\
&\qquad\quad + \hat{s}_j^2(\hat{s}_{ik} + \hat{s}_{i\ell} + \hat{s}_{k\ell}) + \hat{s}_k^2(\hat{s}_{ij} + \hat{s}_{i\ell} + \hat{s}_{j\ell}) + \hat{s}_\ell^2(\hat{s}_{ij} + \hat{s}_{ik} + \hat{s}_{jk}) \\
&\qquad\qquad - \tfrac{6}{n^2}\hat{s}_i\hat{s}_j\hat{s}_k\hat{s}_\ell].
\end{aligned}$$

Factor analysis

- Linear generative model

$$\mathbf{y} = A\mathbf{s} + \boldsymbol{\varepsilon}$$

  noise $\boldsymbol{\varepsilon} \in \mathbb{R}^m$, factor loadings $A \in \mathbb{R}^{m \times r}$, hidden factors $\mathbf{s} \in \mathbb{R}^r$, observed data $\mathbf{y} \in \mathbb{R}^m$.
- Do not know A, $\mathbf{s}$, $\boldsymbol{\varepsilon}$, but need to recover $\mathbf{s}$ and sometimes A from multiple observations of $\mathbf{y}$.
- Time series of observations, get matrices $Y = [\mathbf{y}_1, \dots, \mathbf{y}_n]$, $S = [\mathbf{s}_1, \dots, \mathbf{s}_n]$, $E = [\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n]$, and

$$Y = AS + E.$$

Factor analysis: Recover A and S from Y by a low-rank matrix approximation $Y \approx AS$

Principal and independent components analysis

Principal components analysis: $\mathbf{s}$ Gaussian,

$$\hat{\mathcal{K}}_2(\mathbf{y}) = Q\Lambda_2 Q^\top = (Q, Q) \cdot \Lambda_2,$$

$\Lambda_2 \approx \hat{\mathcal{K}}_2(\mathbf{s})$ diagonal matrix, $Q \in \mathrm{O}(n, r)$, [Pearson; 1901].

Independent components analysis: $\mathbf{s}$ statistically independent entries, ε Gaussian

$$\hat{\mathcal{K}}_p(\mathbf{y}) = (Q, \dots, Q) \cdot \Lambda_p, \quad p = 2, 3, \dots,$$

$\Lambda_p \approx \hat{\mathcal{K}}_p(\mathbf{s})$ diagonal tensor, $Q \in \mathrm{O}(n, r)$, [Comon; 1994].

What if

- $\mathbf{s}$ not Gaussian, e.g. power-law distributed data in social networks.
- $\mathbf{s}$ not independent, e.g. functional components in neuroimaging.
- ε not white noise, e.g. idiosyncratic factors in financial modelling.

Principal cumulant components analysis

- Note that if $\varepsilon = \mathbf{0}$, then

$$\mathcal{K}_p(\mathbf{y}) = \mathcal{K}_p(Q\mathbf{s}) = (Q, \dots, Q) \cdot \mathcal{K}_p(\mathbf{s}).$$

- In general, want principal components that account for variation in all cumulants simultaneously

$$\min_{Q \in \mathrm{O}(n,r),\, \mathcal{C}_p \in \mathsf{S}^p(\mathbb{R}^r)} \sum\nolimits_{p=1}^{\infty} \alpha_p \|\hat{\mathcal{K}}_p(\mathbf{y}) - (Q, \dots, Q) \cdot \mathcal{C}_p\|_F^2,$$

- $\mathcal{C}_p \approx \hat{\mathcal{K}}_p(\mathbf{s})$ not necessarily diagonal.
- Appears intractable: optimization over infinite-dimensional manifold

$$\mathrm{O}(n, r) \times \prod\nolimits_{p=1}^{\infty} \mathsf{S}^p(\mathbb{R}^r).$$

- Surprising relaxation: optimization over a single Grassmannian $\mathrm{Gr}(n, r)$ of dimension $r(n - r)$,

$$\max_{Q \in \mathrm{Gr}(n,r)} \sum\nolimits_{p=1}^{\infty} \alpha_p \|\hat{\mathcal{K}}_p(\mathbf{y}) \cdot (Q, \dots, Q)\|_F^2.$$

- In practice $\infty = 3$ or 4.

Geometric insights

- Secants of Veronese in $\mathsf{S}^p(\mathbb{R}^n)$ — not closed, not irreducible, difficult to study.
- Symmetric subspace variety in $\mathsf{S}^p(\mathbb{R}^n)$ — closed, irreducible, easy to study.
- Stiefel manifold $\mathrm{O}(n,r)$: set of $n \times r$ real matrices with orthonormal columns. $\mathrm{O}(n,n) = \mathrm{O}(n)$, usual orthogonal group.
- Grassman manifold $\mathrm{Gr}(n,r)$: set of equivalence classes of $\mathrm{O}(n,r)$ under left multiplication by $\mathrm{O}(n)$.
- Parameterization of $\mathsf{S}^p(\mathbb{R}^n)$ via

$$\mathrm{Gr}(n,r) \times \mathsf{S}^p(\mathbb{R}^r) \to \mathsf{S}^p(\mathbb{R}^n).$$

- More generally

$$\mathrm{Gr}(n,r) \times \prod_{p=1}^{\infty} \mathsf{S}^p(\mathbb{R}^r) \to \prod_{p=1}^{\infty} \mathsf{S}^p(\mathbb{R}^n).$$

From Stieffel to Grassmann

- Given $\mathcal{A} \in \mathsf{S}^p(\mathbb{R}^n)$, some $r \ll n$, want
$$\min_{X \in \mathrm{O}(n,r),\, \mathcal{C} \in \mathsf{S}^p(\mathbb{R}^r)} \|\mathcal{A} - (X, \dots, X) \cdot \mathcal{C}\|_F,$$
- Unlike approximation by secants of Veronese, subspace approximation problem always has an globally optimal solution.
- Equivalent to
$$\max_{X \in \mathrm{O}(n,r)} \|(X^\top, \dots, X^\top) \cdot \mathcal{A}\|_F = \max_{X \in \mathrm{O}(n,r)} \|\mathcal{A} \cdot (X, \dots, X)\|_F.$$
- Problem defined on a Grassmannian since
$$\|\mathcal{A} \cdot (X, \dots, X)\|_F = \|\mathcal{A} \cdot (XQ, \dots, XQ)\|_F,$$
for any $Q \in \mathrm{O}(r)$. Only the subspaces spanned by X matters.
- Equivalent to
$$\max_{X \in \mathrm{Gr}(n,r)} \|\mathcal{A} \cdot (X, \dots, X)\|_F\,.$$
- Once we have optimal $X_* \in \mathrm{Gr}(n, r)$, may obtain $\mathcal{C}_* \in \mathsf{S}^p(\mathbb{R}^r)$ up to $\mathrm{O}(n)$-equivalence,
$$\mathcal{C}_* = (X_*^\top, \dots, X_*^\top) \cdot \mathcal{A}.$$

Coordinate-cycling heuristics

- Alternating Least Squares (i.e. Gauss-Seidel) is commonly used for minimizing
$$\Psi(X, Y, Z) = \|\mathcal{A} \cdot (X, Y, Z)\|_F^2$$
for $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ cycling between X, Y, Z and solving a least squares problem at each iteration.
- What if $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$ and
$$\Phi(X) = \|\mathcal{A} \cdot (X, X, X)\|_F^2?$$
- Present approach: disregard symmetry of $\mathcal{A}$, solve $\Psi(X, Y, Z)$, set
$$X_* = Y_* = Z_* = (X_* + Y_* + Z_*)/3$$
upon final iteration.
- Better: L-BFGS on Grassmannian.

Newton/quasi-Newton on a Grassmannian

- Objective $\Phi : \mathrm{Gr}(n, r) \to \mathbb{R}$, $\Phi(X) = \|\mathcal{A} \cdot (X, X, X)\|_F^2$.
- $\mathbf{T}_X$ tangent space at $X \in \mathrm{Gr}(n, r)$

$$\mathbb{R}^{n\times r} \ni \Delta \in \mathbf{T}_X \qquad \Longleftrightarrow \qquad \Delta^\top X = 0$$

1. Compute Grassmann gradient $\nabla\Phi \in \mathbf{T}_X$.
2. Compute Hessian or update Hessian approximation

$$H : \Delta \in \mathbf{T}_X \to H\Delta \in \mathbf{T}_X.$$

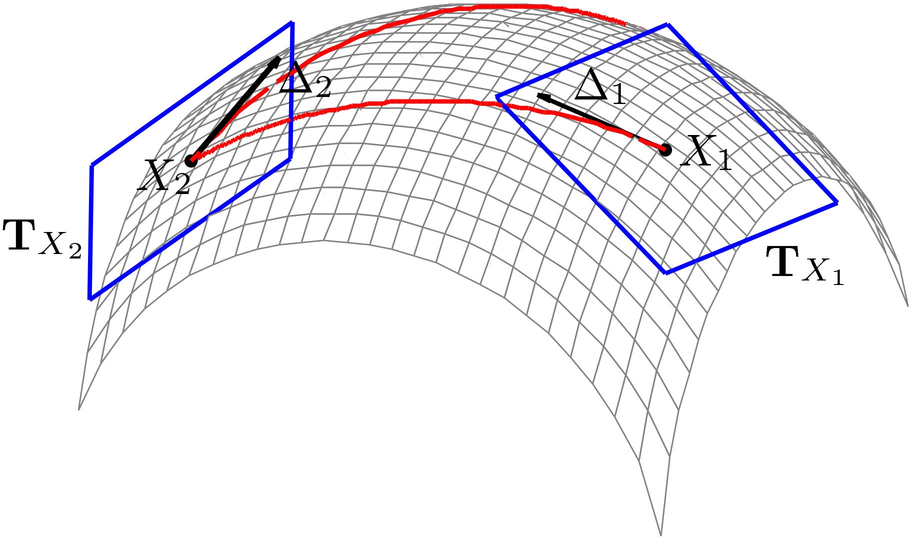
3. At $X \in \mathrm{Gr}(n, r)$, solve

$$H\Delta = -\nabla\Phi$$

   for search direction Δ.
4. Update iterate X: Move along geodesic from X in the direction given by Δ.

- [Arias, Edelman, Smith; 1999], [Eldén, Savas; 2008], [Savas, L.; 2008].

Picture



BFGS on Grassmannian

The BFGS update

$$H_{k+1} = H_k - \frac{H_k \mathbf{s}_k \mathbf{s}_k^\top H_k}{\mathbf{s}_k^\top H_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{y}_k}$$

where

$$\begin{aligned} \mathbf{s}_k &= \mathbf{x}_{k+1} - \mathbf{x}_k = t_k \mathbf{p}_k, \\ \mathbf{y}_k &= \nabla f_{k+1} - \nabla f_k. \end{aligned}$$

On Grassmannian the vectors are defined on different points belonging to different tangent spaces.

Different ways of parallel transporting vectors

$X \in \mathrm{Gr}(n, r)$, $\Delta_1, \Delta_2 \in \mathbf{T}_X$ and $X(t)$ geodesic path along Δ_1

- Parallel transport using global coordinates

$$\Delta_2(t) = T_{\Delta_1}(t)\Delta_2$$

  we have also

$$\Delta_1 = X_\perp D_1 \quad \text{and} \quad \Delta_2 = X_\perp D_2$$

  where $X_\perp$ basis for $\mathbf{T}_X$. Let $X(t)_\perp$ be basis for $\mathbf{T}_{X(t)}$.

- Parallel transport using local coordinates

$$\Delta_2(t) = X(t)_\perp D_2.$$

Parallel transport in local coordinates

All transported tangent vectors have the same coordinate representation in the basis $X(t)_\perp$ at all points on the path $X(t)$.

Plus: No need to transport the gradient or the Hessian.

Minus: Need to compute $X(t)_\perp$.

In global coordinate we compute

- $\mathbf{T}_{k+1} \ni \mathbf{s}_k = t_k T_{\Delta_k}(t_k)\mathbf{p}_k$
- $\mathbf{T}_{k+1} \ni \mathbf{y}_k = \nabla f_{k+1} - T_{\Delta_k}(t_k)\nabla f_k$
- $T_{\Delta_k}(t_k) H_k T_{\Delta_k}^{-1}(t_k) : \mathbf{T}_{k+1} \longrightarrow \mathbf{T}_{k+1}$

$$H_{k+1} = H_k - \frac{H_k \mathbf{s}_k \mathbf{s}_k^\top H_k}{\mathbf{s}_k^\top H_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{y}_k}$$

BFGS

Compact representation of BFGS in Euclidean space:

$$H_k = H_0 + \begin{bmatrix} S_k & H_0 Y_k \end{bmatrix} \begin{bmatrix} R_k^{-\top}(D_k + Y_k^\top H_0 Y_k) R_k^{-1} & -R_k^{-\top} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^\top \\ Y_k^\top H_0 \end{bmatrix}$$

where

$$\begin{aligned}
S_k &= [\mathbf{s}_0, \dots, \mathbf{s}_{k-1}], \\
Y_k &= [\mathbf{y}_0, \dots, \mathbf{y}_{k-1}], \\
D_k &= \operatorname{diag}\left[\mathbf{s}_0^\top \mathbf{y}_0, \dots, \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}\right], \\
R_k &= \begin{bmatrix} \mathbf{s}_0^\top \mathbf{y}_0 & \mathbf{s}_0^\top \mathbf{y}_1 & \cdots & \mathbf{s}_0^\top \mathbf{y}_{k-1} \\ 0 & \mathbf{s}_1^\top \mathbf{y}_1 & \cdots & \mathbf{s}_1^\top \mathbf{y}_{k-1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1} \end{bmatrix}.
\end{aligned}$$

L-BFGS

Limited memory BFGS [Byrd et al; 1994]. Replace H_0 by $\gamma_k I$ and keep the m most resent $\mathbf{s}_j$ and $\mathbf{y}_j$,

$$H_k = \gamma_k I + \begin{bmatrix} S_k & \gamma_k Y_k \end{bmatrix} \begin{bmatrix} R_k^{-\top}(D_k + \gamma_k Y_k^\top Y_k) R_k^{-1} & -R_k^{-\top} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^\top \\ \gamma_k Y_k^\top \end{bmatrix}$$

where

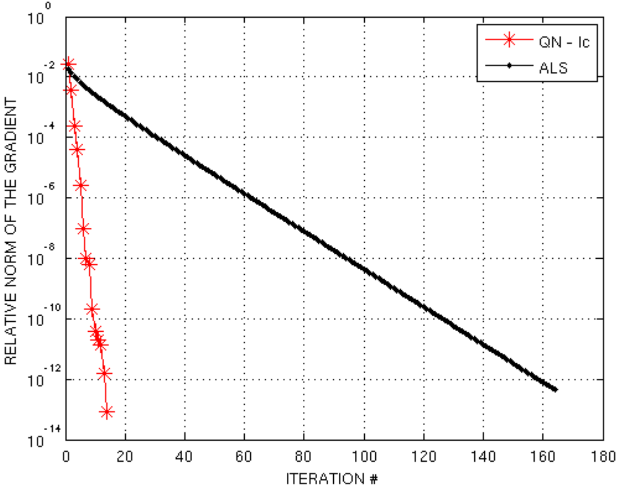
$$\begin{aligned} S_k &= [\mathbf{s}_{k-m}, \dots, \mathbf{s}_{k-1}], \\ Y_k &= [\mathbf{y}_{k-m}, \dots, \mathbf{y}_{k-1}], \\ D_k &= \operatorname{diag}\left[\mathbf{s}_{k-m}^\top \mathbf{y}_{k-m}, \dots, \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}\right], \\ R_k &= \begin{bmatrix} \mathbf{s}_{k-m}^\top \mathbf{y}_{k-m} & \mathbf{s}_{k-m}^\top \mathbf{y}_{k-m+1} & \cdots & \mathbf{s}_{k-m}^\top \mathbf{y}_{k-1} \\ 0 & \mathbf{s}_{k-m+1}^\top \mathbf{y}_{k-m+1} & \cdots & \mathbf{s}_{k-m+1}^\top \mathbf{y}_{k-1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1} \end{bmatrix}. \end{aligned}$$

L-BFGS on the Grassmannian

- In each iteration, parallel transport vectors in S_k and Y_k to $\mathbf{T}_k$, ie. perform
$$\bar{S}_k = TS_k, \qquad \bar{Y}_k = TY_k$$
where T is the transport matrix.
- No need to modify R_k or D_k
$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle T\mathbf{u}, T\mathbf{v} \rangle$$
where $\mathbf{u}, \mathbf{v} \in \mathbf{T}_k$ and $T\mathbf{u}, T\mathbf{v} \in \mathbf{T}_{k+1}$.
- H_k nonsingular, Hessian is singular. No problem $\mathbf{T}_k$ at $\mathbf{x}_k$ is invariant subspace of H_k, ie. if $\mathbf{v} \in \mathbf{T}_k$ then $H_k\mathbf{v} \in \mathbf{T}_k$.
- [Savas, L.; 2008]

Convergence

- Compares favorably with Alternating Least Squares.



Higher order eigenfaces

Principal cumulant subspaces supplement varimax subspace from PCA. Take face recognition for example, **eigenfaces** ($p = 2$) becomes **skewfaces** ($p = 3$) and **kurtofaces** ($p = 4$).

- Eigenfaces: given image $\times$ pixel matrix $A \in \mathbb{R}^{m \times n}$ with centered columns where $m \ll n$.
- Eigenvectors of pixel $\times$ pixel covariance matrix $\mathcal{K}_2^{\text{pixel}} \in \mathsf{S}^2(\mathbb{R}^n)$ are the eigenfaces.
- For efficiency, compute image $\times$ image covariance matrix $\mathcal{K}_2^{\text{image}} \in \mathsf{S}^2(\mathbb{R}^m)$ instead.
- SVD $A = U\Sigma V^\top$ gives both implicitly,

$$\mathcal{K}_2^{\text{image}} = \tfrac{1}{n}(A^\top, A^\top) \cdot \mathcal{I}_2 = \tfrac{1}{n}A^\top A = \tfrac{1}{n}V\Lambda V^\top,$$
$$\mathcal{K}_2^{\text{pixel}} = \tfrac{1}{n}(A, A) \cdot \mathcal{I}_2 = \tfrac{1}{m}AA^\top = \tfrac{1}{m}U\Lambda U^\top.$$

- Orthonormal columns of U, eigenvectors of $n\mathcal{K}_2^{\text{pixel}}$, are the eigenfaces.

Computing image and pixel skewness

- Want to implicitly compute $\mathcal{K}_3^{\text{pixel}} \in \mathsf{S}^3(\mathbb{R}^n)$, third cumulant tensor of the pixels (huge).
- Just need projector Π onto the subspace of skewfaces that best explain $\mathcal{K}_3^{\text{pixel}}$.
- Let $A = U\Sigma V^\top$, $U \in \mathrm{O}(n, m)$, $\Sigma \in \mathbb{R}^{m\times m}$, $V \in \mathrm{O}(m)$.

$$\begin{aligned}
\mathcal{K}_3^{\text{pixel}} &= \tfrac{1}{m}(A, A, A) \cdot \mathcal{I}_m \\
&= \tfrac{1}{m}(U, U, U) \cdot (\Sigma, \Sigma, \Sigma) \cdot (V^\top, V^\top, V^\top) \cdot \mathcal{I}_m \\
\mathcal{K}_3^{\text{image}} &= \tfrac{1}{n}(A^\top, A^\top, A^\top) \cdot \mathcal{I}_n \\
&= \tfrac{1}{n}(V, V, V) \cdot (\Sigma, \Sigma, \Sigma) \cdot (U^\top, U^\top, U^\top) \cdot \mathcal{I}_n
\end{aligned}$$

- $\mathcal{I}_n = [\![\delta_{ijk}]\!] \in \mathsf{S}^3(\mathbb{R}^n)$ is the 'Kronecker delta tensor', i.e. $\delta_{ijk} = 1$ iff $i = j = k$ and $\delta_{ijk} = 0$ otherwise.

Computing skewmax projection

- Define $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^m)$ by

$$\mathcal{A} = (\Sigma, \Sigma, \Sigma) \cdot (V^\top, V^\top, V^\top) \cdot \mathcal{I}_m$$

- Want $Q \in \mathsf{O}(m, s)$ and core tensor $\mathcal{C} \in \mathsf{S}^3(\mathbb{R}^s)$ not necessarily diagonal, so that $\mathcal{A} \approx (Q, Q, Q) \cdot \mathcal{C}$ and thus

$$\mathcal{K}_3^{\text{pixel}} \approx \tfrac{1}{m}(U, U, U) \cdot (Q, Q, Q) \cdot \mathcal{C} = \tfrac{1}{m}(UQ, UQ, UQ) \cdot \mathcal{C}.$$

- Solve

$$\min_{Q \in \mathsf{O}(m,s),\, \mathcal{C} \in \mathsf{S}^3(\mathbb{R}^s)} \|\mathcal{A} - (Q, Q, Q) \cdot \mathcal{C}\|_F$$

- $\Pi = UQ \in \mathsf{O}(n, s)$ is our orthonormal-column projection matrix onto the 'skewmax' subspace.
- Caveat: Q only determined up to $\mathsf{O}(s)$-equivalence. Not a problem if we are just interested in the associated subspace or its projector.

Combining eigen-, skew-, and kurto-faces

Combine information from multiple cumulants:

- Same procedure for the kurtosis tensor (a little more complicated).
- Say we keep the first r eigenfaces (columns of U), s skewfaces, and t kurtofaces. Their span is our optimal subspace.
- These three subspaces may overlap; orthogonalize the resulting $r + s + t$ column vectors to get a final projector.

This gives an orthonormal projector basis W for the column space of A; its

- first r vectors best explain the pixel covariance $\mathcal{K}_2^{\text{pixel}} \in \mathsf{S}^2(\mathbb{R}^n)$,
- next s vectors, with $W_{1:r}$, best explain the pixel skewness $\mathcal{K}_3^{\text{pixel}} \in \mathsf{S}^3(\mathbb{R}^n)$,
- last t vectors, with $W_{1:r+s}$, best explain pixel kurtosis $\mathcal{K}_4^{\text{pixel}} \in \mathsf{S}^4(\mathbb{R}^n)$.

Advertisement and acknowledgement

Jason Morton, "Algebraic models for multilinear dependence," SAMSI Workshop on *Algebraic Statistical Models*, Research Triangle Park, NC, January 15–17, 2009.

Thanks:

- J.M. Landsberg
- Berkant Savas