# Algebraic models for multilinear dependence

Jason Morton

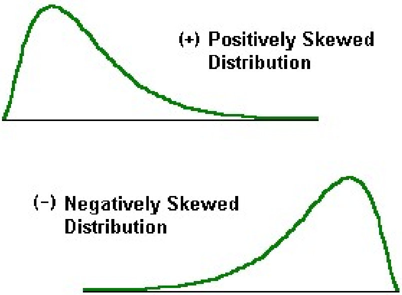
Stanford University

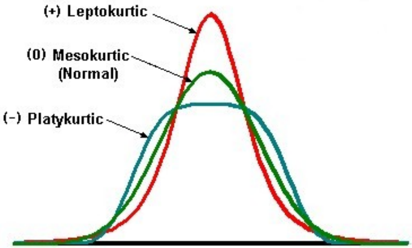
December 11, 2008
NIPS

Joint work with Lek-Heng Lim of Berkeley

# Univariate cumulants

Mean, variance, skewness and kurtosis describe the shape of a univariate distribution.



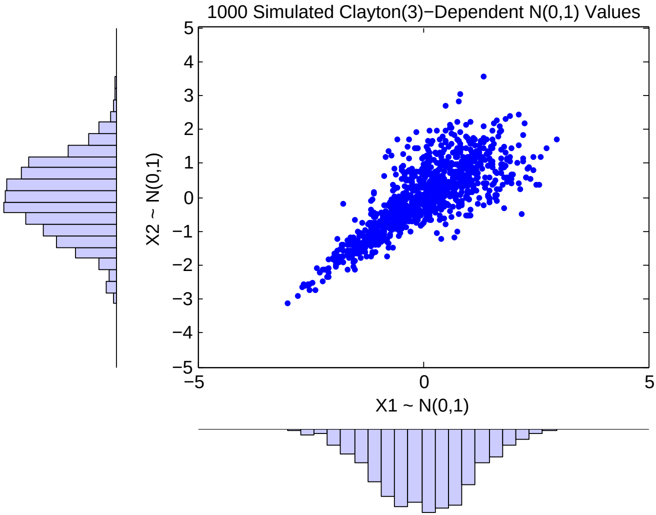
# Covariance matrices

The covariance matrix partly describes the dependence structure of a multivariate distribution.

- PCA
- Gaussian graphical models
- Optimization—bilinear form computes variance

But if the variables are not multivariate Gaussian, not the whole story.

# Even if marginals normal, dependence might not be



# Covariance matrix analogs: multivariate cumulants

- The cumulant tensors are the multivariate analog of skewness and kurtosis.
- They describe higher order dependence among random variables.

1. Definitions: tensors and cumulants
2. Properties of cumulant tensors
3. Insights and models from algebraic geometry
4. Algorithms from Riemannian geometry
5. Potential applications

# Symmetric tensors and actions

A tensor in coordinates is a multi-way array with a multilinear action.

- Tensor $[\![a_{ijk}]\!] \in \mathbb{R}^{r \times r \times r}$ is symmetric if it is invariant under all permutations of indices

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

Comes with an action:

- Symmetric multilinear matrix multiplication. If $Q$ is an $n \times r$ matrix, $T$ an $r \times r \times r$ tensor, make an $n \times n \times n$ tensor $K = (Q, Q, Q) \cdot T$ or just $Q \cdot T$ where

$$K_{\alpha\beta\gamma} = \sum_{i,j,k=1}^{r,r,r} q_{\alpha i} q_{\beta j} q_{\gamma k} t_{ijk}.$$

- If $T$ is $r \times r$ and $Q$ is $n \times r$, we have $Q \cdot T = QTQ^{\top}$; for $d > 2$ multiply on $3, 4, \ldots$ "sides" of the multi-way array.

# Moments and Cumulants are symmetric tensors

Vector-valued random variable $\mathbf{x} = (X_1, \ldots, X_n)$.
Three natural $d$-way tensors are:

- The $d$th non-central moment $s_{i_1,\ldots,s_d}$ of $\mathbf{x}$:

$$S_d(\mathbf{x}) = \left[\mathbb{E}(x_{i_1} x_{i_2} \cdots x_{id})\right]_{i_1,\ldots,i_d=1}^{n}.$$

- The $d$th central moment $S_d(\mathbf{x} - \mathbb{E}[\mathbf{x}])$, and
- The $d$th cumulant $\kappa_{i_1 \ldots i_d}$ of $\mathbf{x}$:

$$K_d(\mathbf{x}) = \left[\sum_{A_1 \sqcup \cdots \sqcup A_q = \{i_1,\ldots,i_d\}} (-1)^{q-1}(q-1)! s_{A_1} \ldots s_{A_q}\right]_{i_1,\ldots,i_d=1}^{n}.$$

# Measuring useful properties.

For univariate $x$, the cumulants $K_d(x)$ for $d = 1, 2, 3, 4$ are

- expectation $\kappa_i = \mathbb{E}[x]$,
- variance $\kappa_{ii} = \sigma^2$,
- skewness $\kappa_{iii}/\kappa_{ii}^{3/2}$, and
- kurtosis $\kappa_{iiii}/\kappa_{ii}^2$.

The tensor versions are the multivariate generalizations

$$\kappa_{ijk}$$

they provide a natural measure of non-Gaussianity.

# Alternative Definitions of Cumulants

- In terms of log characteristic function,

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

- In terms of Edgeworth series,

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

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index, $\mathbf{t}^{\boldsymbol{\alpha}} = t_1^{\alpha_1} \cdots t_d^{\alpha_d}$, and $\alpha! = \alpha_1! \cdots \alpha_d!$.

# Properties of cumulants: Multilinearity

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

$$K_d(A\mathbf{x}) = A \cdot K_d(\mathbf{x}),$$

  where $\cdot$ is the multilinear action.
- This makes factor models work: $\mathbf{y} = A\mathbf{x}$ implies $K_d^Y = A \cdot K_d^X$;
- For example, $K_2^Y = AK_2^X A^\top$ .
- Independent Components Analysis finds an $A$ to approximately diagonalize $K_d^X$.

# Properties of cumulants: Independence

Independence:

- $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are mutually independent of variables $\mathbf{y}_1, \ldots, \mathbf{y}_k$, we have
  $K_d(\mathbf{x}_1 + \mathbf{y}_1, \ldots, \mathbf{x}_k + \mathbf{y}_k) = K_d(\mathbf{x}_1, \ldots, \mathbf{x}_k) + K_d(\mathbf{y}_1, \ldots, \mathbf{y}_k)$.
- $K_{i_1,\ldots,i_n}(\mathbf{x}) = 0$ whenever there is a partition of $\{i_1, \ldots, i_n\}$ into two nonempty sets $I$ and $J$ such that $\mathbf{x}_I$ and $\mathbf{x}_J$ are independent.
- Why we want to diagonalize in independent component analysis
- Exploitable in other sparse cumulant techniques

# Properties of cumulants: Vanishing and Extending

- Gaussian: If $\mathbf{x}$ is multivariate normal, then $K_d(\mathbf{x}) = 0$ for all $d \geq 3$.
  - Why you might not have heard of them: for Gaussians, the covariance matrix does tell the whole story.
- Support: There are no distributions with a bound $n$ so that

$$K_d(\mathbf{x}) \begin{cases} \neq 0 & 3 \leq d \leq n, \\ = 0 & d > n. \end{cases}$$

  - Parametrization is trickier when $K_2$ doesn't tell the whole story.

# Making cumulants useful, tractable and estimable

Cumulant tensors are a useful generalization, but too big. They have $\binom{\#vars+d-1}{d}$ quantities, too many to

- learn with a reasonable amount of data,
- store, and
- optimize.

Needed: small, implicit models analogous to PCA

PCA: eigenvalue decomposition of a positive semidefinite real symmetric matrix. We need a tensor analog.

But, it isn't as easy as it looks. . .

# Tensor decomposition

Three possible generalizations are the same in the matrix case but not in the tensor case. For a $n \times n \times n$ tensor $T$,

| Name | minimum $r$ such that |
|---|---|
| Tensor rank | $T = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i$<br>not closed |
| Border rank | $T = \lim_{\epsilon \to 0}(S_\epsilon)$, $Trank(S_\epsilon) = r$<br>closed but hard to represent;<br>defining equations unknown. |
| Multilinear rank | $T = (A, B, C) \cdot K, K \in \mathbb{R}^{r \times r \times r}, A, B, C \in \mathbb{R}^{n \times r}$,<br>closed and understood. |

# Multilinear rank factor model

Let $\mathbf{y} = Y_1, \ldots, Y_n$ be a random vector. Write the $d$th order cumulant $K_d(\mathbf{y})$ as a best $r$-multilinear rank approximation in terms of the cumulant $K_d(\mathbf{x})$ of a smaller set of $r$ factors $\mathbf{x}$:

$$K_d^Y \approx Q \cdot K_d^X.$$

where

- $Q$ is orthonormal , and $Q^\top$ projects to the factors
- The column space of $Q$ defines the $s$-dim subspace which best explains the $d$th order dependence.
- In place of eigenvalues, we have the core tensor $K_d^X$, the cumulant of the factors.

Have model, need loss and algorithm.

# Principal cumulant components analysis

- Want factors/principal components that account for variation in all cumulants simultaneously

$$\min_{Q\in \mathrm{O}(n,r),\, \mathcal{C}_d\in \mathrm{S}^d(\mathbb{R}^r)} \sum_{d=1}^{\infty} \alpha_d \|\hat{K}_d(\mathbf{y}) - Q\cdot \mathcal{C}_d\|^2,$$

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

$$\mathrm{O}(n,r) \times \prod_{d=1}^{\infty} \mathrm{S}^d(\mathbb{R}^r).$$

- Reduces to optimization over a single Grassmannian $\mathrm{Gr}(n,r)$ of dimension $r(n-r)$,

$$\max_{Q\in \mathrm{Gr}(n,r)} \sum_{d=1}^{\infty} \alpha_d \|Q^\top \cdot \hat{\mathcal{K}}_d(\mathbf{y})\|^2.$$

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

# Geometric insights

- Secants of Veronese in $\mathsf{S}^d(\mathbb{R}^n)$ and rank subsets— difficult to study.
- Symmetric subspace variety in $\mathsf{S}^d(\mathbb{R}^n)$ — closed, easy to study.
- Stiefel manifold $\mathsf{O}(n, r)$ is set of $n \times r$ real matrices with orthonormal columns.
- Grassman manifold $\mathsf{Gr}(n, r)$ is set of equivalence classes of $\mathsf{O}(n, r)$ under left multiplication by $\mathsf{O}(n)$.
- Parametrization of $\mathsf{S}^d(\mathbb{R}^n)$ via

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

# 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 $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ 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

[Savas-Lim]

- 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 $X \in \mathrm{Gr}(n, r)$

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

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

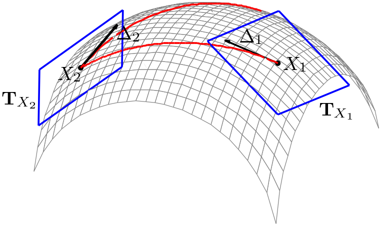
$$H : \Delta \in \mathbf{T}_X \to H\Delta \in \mathbf{T}_X.$$

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

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

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

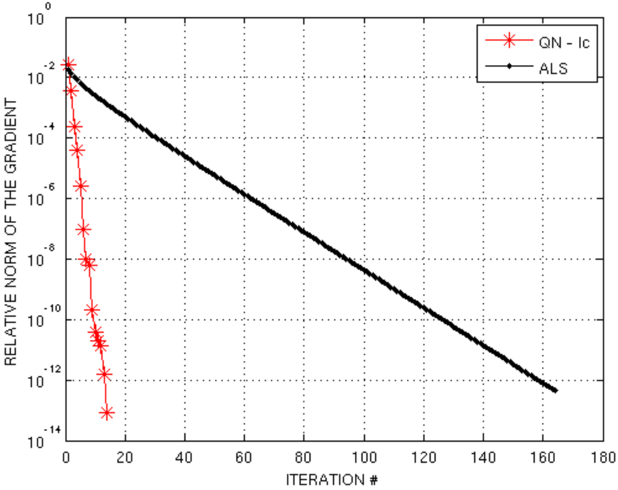
# L-BFGS on Grassmannian




- BFGS update must be adjusted: on the Grassmannian, the vectors are defined on different points belonging to different tangent spaces.
- Parallel transport along a geodesic to new position.
- Limited memory version.

# Convergence

- Compares favorably with Alternating Least Squares.



# Mean-variance portfolio optimization

Markowitz mean-variance portfolio optimization defines risk to be variance.

$$\min w^\top K_2(\mathbf{x}) w \qquad s.t. \qquad w^\top \mathbb{E}[\mathbf{x}] > \underline{r}$$

Evidence indicates that investors optimizing variance with respect to the covariance matrix accept unwanted skewness and kurtosis risk.

- Extreme example: selling out-of-the-money puts looks safe and uncorrelated
- Many hedge funds essentially do this

# Muti-moment portfolio optimization

So, take skewness and kurtosis into account in the objective.

- Need to use skewness $K_3$ and kurtosis $K_4$ tensors.
- Use low multilinear rank model to regularize and make optimization computable with many assets (linear vs. cubic)

With mean-zero returns in a #assets$= m \times n =$#periods matrix $A$,

- Choose an $s$, need $m \times s$ orthonormal projector $Q$
- Approximate cumulant $nK_d = A^\top \cdot \Delta_{d,n} \approx Q \cdot C$
- Multilinear forms $w^\top \cdot K_d \approx w^\top Q \cdot \frac{1}{n} C$ give variance, skewness and kurtosis

# Analogously to Eigenfaces,

Cumulants give features supplementing the PCA varimax subspace.

- In eigenfaces, we have a centered #pixels$= n \times m =$#images matrix $A$, $m \ll n$.
- The eigenvectors of the covariance matrix $K_2^P$ of the *pixels* are the eigenfaces.
- For efficiency, we compute the covariance matrix $K_2^{Images}$ of the *images* instead. The SVD gives both implicitly.

$$USV^\top = svd(A^\top)$$

$$mK_2^{Images} = A^\top A = U\Lambda U^\top$$

$$nK_2^{Pixels} = AA^\top = V\Lambda V^\top$$

Orthonormal columns of $V$, eigenvectors of $K_2^P$, are the eigenfaces.

# we can compute Skewfaces,

Centered #pixels= $n \times m$ =#images matrix $A$.

- Let $K_3^P$ be the (huge) third cumulant tensor of the pixels.
- Analogously, we want to compute it implicitly
- We just need the projector $\Pi$ onto the subspace of skewfaces that best explain $K_3^P$.

Let $A^\top = USV^\top$ with dims $(m^2, m^2, m \times n)$.

$$nS_3^I = A^\top \cdot \Delta_n = U \cdot S \cdot V^\top \cdot \Delta_n$$

$$mK_3^P = A \cdot \Delta_m = V \cdot (S \cdot U^\top \cdot \Delta_m)$$

Pick a small multilinear rank $s$. If $(S \cdot U^\top \cdot \Delta_m) \approx Q \cdot C_3$ for some $m \times s$ matrix $Q$ and NON-diagonal core tensor $C_3$,

$$mK_3^P \approx V \cdot Q \cdot C_3 = VQ \cdot C_3$$

and $\Pi = VQ$ is our orthonormal-column projection matrix onto the 'skewmax' subspace.

# and combine Eigen-, Skew-, and Kurto-faces.

Combine the information from multiple cumulants:

- Do the same for procedure for the kurtosis tensor (a little more complicated).
- Say we keep the first $r$ principal components (columns of $V$), $s$ 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 covariance matrix $K_2^P$,
- next $s$ vectors, with $W_{1:r}$, best explain the big skewness tensor $K_3^P$ of the pixels, and
- last $t$ vectors, with $W_{1:r+s}$, best explain pixel kurtosis $K_4^P$.

End

jason@math.stanford.edu