Numerical Multilinear Algebra III

Lek-Heng Lim

University of California, Berkeley

January 5–7, 2009

(Contains joint work with Jason Morton of Stanford and Berkant Savas of UT Austin

Current affairs

Question

What lesson in multilinear algebra did we learn from the current global economic/financial crisis?

Current affairs

Question

What lesson in multilinear algebra did we learn from the current global economic/financial crisis?

- **One answer:** it's important to look beyond the quadratic term.

Current affairs

Question

What lesson in multilinear algebra did we learn from the current global economic/financial crisis?

- **One answer:** it's important to look beyond the quadratic term.
- **Taylor approximation:** multivariate $f(x_1, \dots, x_n)$ approximated as

$$f(\mathbf{x}) \approx a_0 + \mathbf{a}_1^\top \mathbf{x} + \mathbf{x}^\top A_2 \mathbf{x} + \mathcal{A}_3(\mathbf{x}, \mathbf{x}, \mathbf{x}) + \cdots + \mathcal{A}_d(\mathbf{x}, \dots, \mathbf{x}) + \cdots .$$

$$a_0 \in \mathbb{R}, \mathbf{a}_1 \in \mathbb{R}^n, A_2 \in \mathbb{R}^{n \times n}, \mathcal{A}_3 \in \mathbb{R}^{n \times n \times n}, \dots .$$

Current affairs

Question

What lesson in multilinear algebra did we learn from the current global economic/financial crisis?

- **One answer:** it's important to look beyond the quadratic term.
- **Taylor approximation:** multivariate $f(x_1, \dots, x_n)$ approximated as

$$f(\mathbf{x}) \approx a_0 + \mathbf{a}_1^\top \mathbf{x} + \mathbf{x}^\top A_2 \mathbf{x} + \mathcal{A}_3(\mathbf{x}, \mathbf{x}, \mathbf{x}) + \cdots + \mathcal{A}_d(\mathbf{x}, \dots, \mathbf{x}) + \cdots .$$

  $a_0 \in \mathbb{R}, \mathbf{a}_1 \in \mathbb{R}^n, A_2 \in \mathbb{R}^{n \times n}, \mathcal{A}_3 \in \mathbb{R}^{n \times n \times n}, \dots$.
- Numerical linear algebra: $d = 2$.

Current affairs

Question

What lesson in multilinear algebra did we learn from the current global economic/financial crisis?

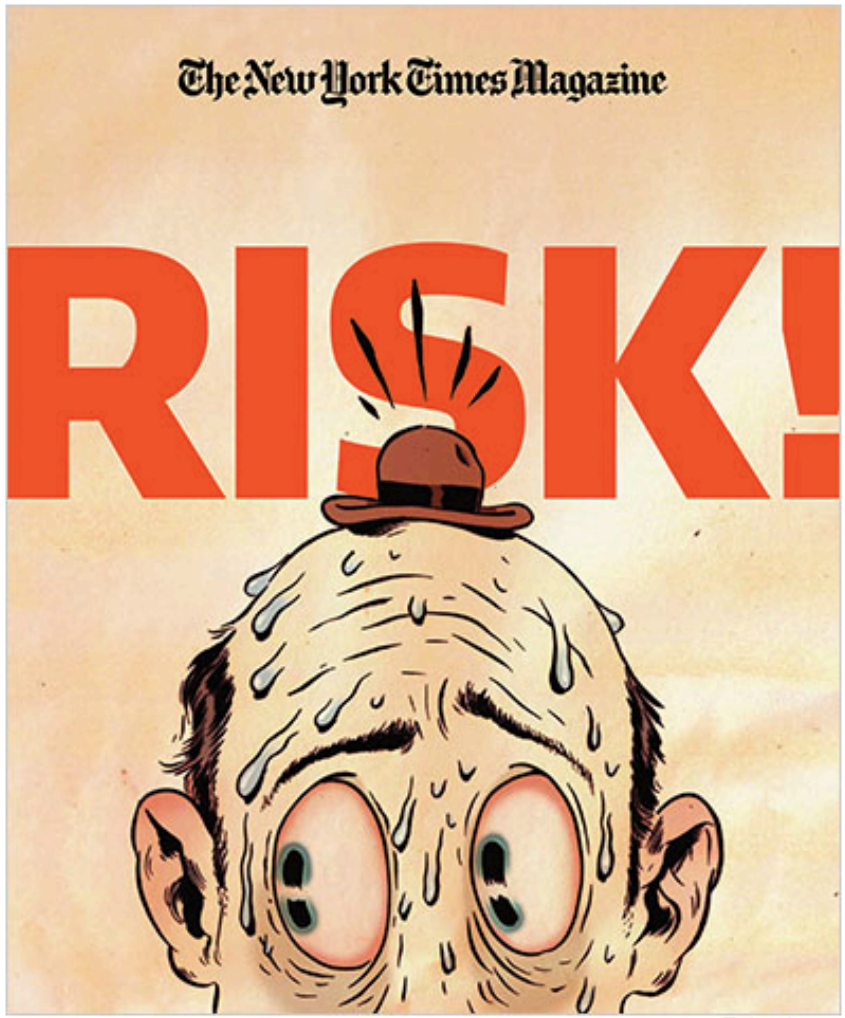
- **One answer:** it's important to look beyond the quadratic term.
- **Taylor approximation:** multivariate $f(x_1, \dots, x_n)$ approximated as

$$f(\mathbf{x}) \approx a_0 + \mathbf{a}_1^\top \mathbf{x} + \mathbf{x}^\top A_2 \mathbf{x} + \mathcal{A}_3(\mathbf{x}, \mathbf{x}, \mathbf{x}) + \cdots + \mathcal{A}_d(\mathbf{x}, \dots, \mathbf{x}) + \cdots .$$

  $a_0 \in \mathbb{R}, \mathbf{a}_1 \in \mathbb{R}^n, A_2 \in \mathbb{R}^{n \times n}, \mathcal{A}_3 \in \mathbb{R}^{n \times n \times n}, \dots$.
- Numerical linear algebra: $d = 2$.
- Numerical multilinear algebra: $d > 2$.

The New York Times Magazine

1.4.2009



Zohar Lazar

Risk Mismanagement

By JOE NOCERA

Were the measures used to evaluate Wall Street trades flawed? Or was the mistake ignoring them?

· Times Topics: Credit Crisis

The Way We Live Now

The Senator Track

By LISA BELKIN

Why Caroline Kennedy's "experience" counts.

· Times Topics: Caroline Kennedy

· Post a Comment

· More at The Motherlode Blog

THE MEDIUM

We Interrupt This Program

By VIRGINIA HEFFERNAN

Hulu, the streaming-video service, offers a new (old?) paradigm for watching TV and movies online.

· Post a Comment

QUESTIONS FOR JOAN RIVERS

Cutup

Interview by DEBORAH SOLOMON

The comedian talks about plastic surgery as a business decision, Barack Obama's ears and getting a little work done in a recession.

· Times Topics: Joan Rivers

· Past Questions For ... Columns

ON LANGUAGE

Bleeping Expletives

By WILLIAM SAFIRE

Bonfire of the profanities.

January 4, 2009

Risk Mismanagement

By JOE NOCERA

'The story that I have to tell is marked all the way through by a persistent tension between those who assert that the best decisions are based on quantification and numbers, determined by the patterns of the past, and those who base their decisions on more subjective degrees of belief about the uncertain future. This is a controversy that has never been resolved.'

— FROM THE INTRODUCTION TO "AGAINST THE GODS: THE REMARKABLE STORY OF RISK," BY PETER L. BERNSTEIN

THERE AREN'T MANY widely told anecdotes about the current financial crisis, at least not yet, but there's one that made the rounds in 2007, back when the big investment banks were first starting to write down billions of dollars in mortgage-backed derivatives and other so-called toxic securities. This was well before Bear Stearns collapsed, before Fannie Mae and Freddie Mac were taken over by the federal government, before Lehman fell and Merrill Lynch was sold and A.I.G. saved, before the $700 billion bailout bill was rushed into law. Before, that is, it became obvious that the risks taken by the largest banks and investment firms in the United States — and, indeed, in much of the Western world — were so excessive and foolhardy that they threatened to bring down the financial system itself. On the contrary: this was back when the major investment firms were still assuring investors that all was well, these little speed bumps notwithstanding — assurances based, in part, on their fantastically complex mathematical models for

Risk managers use VaR to quantify their firm's risk positions to their board. In the late 1990s, as the use of derivatives was exploding, the Securities and Exchange Commission ruled that firms had to include a quantitative disclosure of market risks in their financial statements for the convenience of investors, and VaR became the main tool for doing so. Around the same time, an important international rule-making body, the Basel Committee on Banking Supervision, went even further to validate VaR by saying that firms and banks could rely on their own internal VaR calculations to set their capital requirements. So long as their VaR was reasonably low, the amount of money they had to set aside to cover risks that might go bad could also be low.

Given the calamity that has since occurred, there has been a great deal of talk, even in quant circles, that this widespread institutional reliance on VaR was a terrible mistake. At the very least, the risks that VaR measured did not include the biggest risk of all: the possibility of a financial meltdown. "Risk modeling didn't help as much as it should have," says Aaron Brown, a former risk manager at Morgan Stanley who now works at AQR, a big quant-oriented hedge fund. A risk consultant named Marc Groz says, "VaR is a very limited tool." David Einhorn, who founded Greenlight Capital, a prominent hedge fund, wrote not long ago that VaR was "relatively useless as a risk-management tool and potentially catastrophic when its use creates a false sense of security among senior managers and watchdogs. This is like an air bag that works all the time, except when you have a car accident." Nassim Nicholas Taleb, the best-selling author of "The Black Swan," has crusaded against VaR for more than a decade. He calls it, flatly, "a fraud."

How then do we account for that story that made the rounds in the summer of 2007? It concerns Goldman Sachs, the one Wall Street firm that was not, at that time, taking a hit for billions of dollars of suddenly devalued mortgage-backed securities. Reporters wanted to understand how Goldman had somehow sidestepped the disaster that had befallen everyone else. What they discovered was that in December 2006, Goldman's various indicators, including VaR and other risk models, began suggesting that something was wrong. Not hugely wrong, mind you, but wrong enough to warrant a closer look.

"We look at the P.& L. of our businesses every day," said Goldman Sachs' chief financial officer, David

It's not every day that an options trader becomes famous by writing a book, but that's what Taleb did, first with "Fooled by Randomness," which was published in 2001 and became an immediate cult classic on Wall Street, and more recently with "The Black Swan: The Impact of the Highly Improbable," which came out in 2007 and landed on a number of best-seller lists. He also went from being primarily an options trader to what he always really wanted to be: a public intellectual. When I made the mistake of asking him one day whether he was an adjunct professor, he quickly corrected me. "I'm the Distinguished Professor of Risk Engineering at N.Y.U.," he responded. "It's the highest title they give in that department." Humility is not among his virtues. On his Web site he has a link that reads, "Quotes from 'The Black Swan' that the imbeciles did not want to hear."

"How many of you took statistics at Columbia?" he asked as he began his lecture. Most of the hands in the room shot up. "You wasted your money," he sniffed. Behind him was a slide of Mickey Mouse that he had put up on the screen, he said, because it represented "Mickey Mouse probabilities." That pretty much sums up his view of business-school statistics and probability courses.

Taleb's ideas can be difficult to follow, in part because he uses the language of academic statisticians; words like "Gaussian," "kurtosis" and "variance" roll off his tongue. But it's also because he speaks in a kind of brusque shorthand, acting as if any fool should be able to follow his train of thought, which he can't be bothered to fully explain.

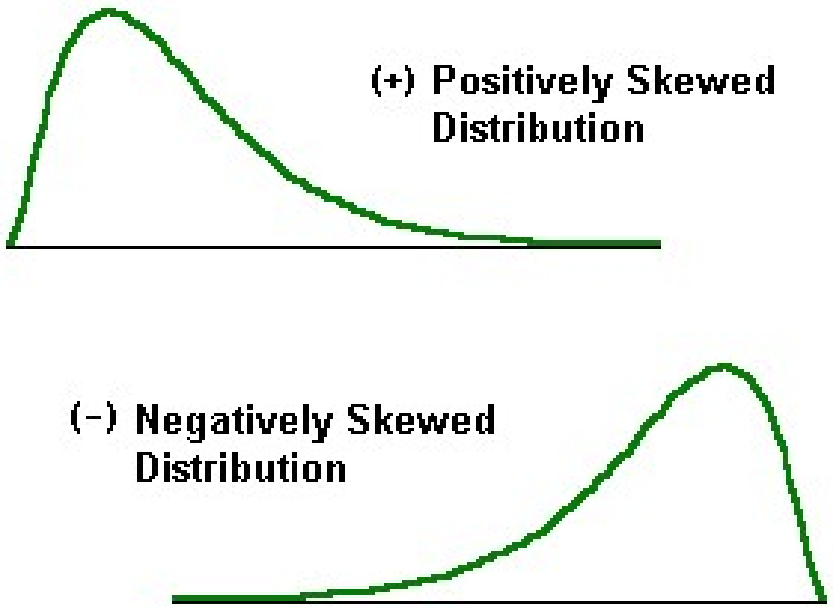
"This is a Stan O'Neal trade," he said, referring to the former chief executive of Merrill Lynch. He clicked to a slide that showed a trade that made slow, steady profits — and then quickly spiraled downward for a giant, brutal loss.

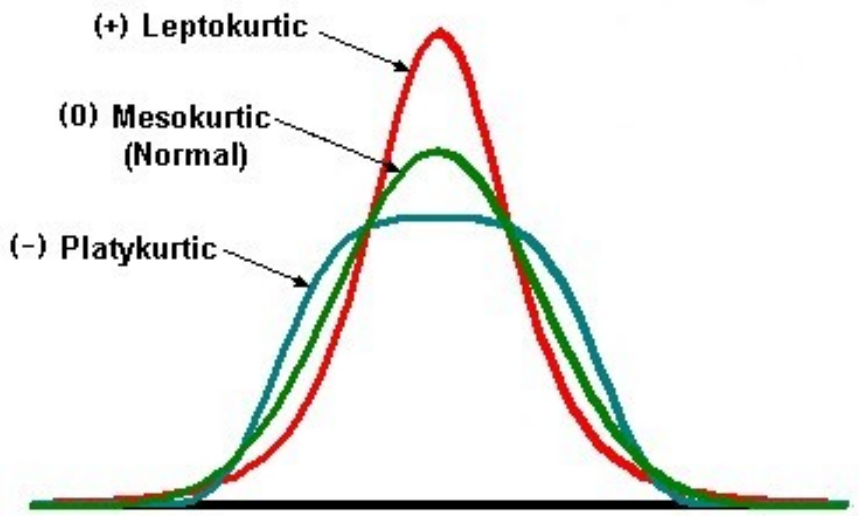
"Why do people measure risks against events that took place in 1987?" he asked, referring to Black Monday, the October day when the U.S. market lost more than 20 percent of its value and has been used ever since as the worst-case scenario in many risk models. "Why is that a benchmark? I call it future-blindness.

"If you have a pilot flying a plane who doesn't understand there can be storms, what is going to happen?" he asked. "He is not going to have a magnificent flight. Any small error is going to crash a plane. This is why the crisis that happened was predictable."

Univariate cumulants

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





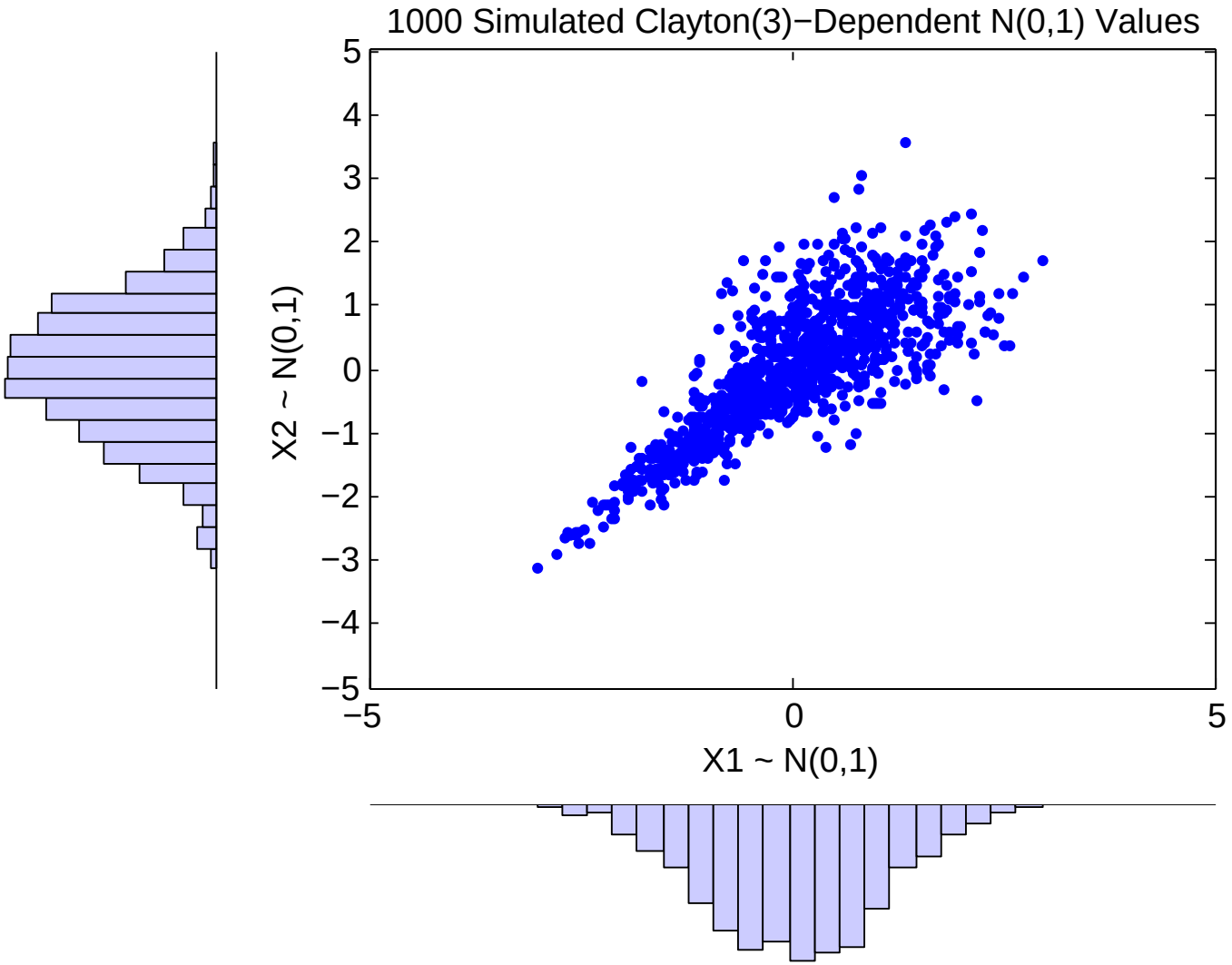
Covariance matrices

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.

Recap: 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.

Recap: multilinear matrix multiplication

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

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

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

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

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

Recap: 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.
- $\mathrm{S}^p(\mathbb{R}^n)$ denotes set of all order-p symmetric tensors.
- Symmetric multilinear matrix multiplication $\mathcal{C} = (X, X, X) \cdot \mathcal{A}$ where

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

Examples of symmetric tensors

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

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

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

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

Cumulants

- In terms of log characteristic and cumulant generating functions,

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

- In terms of Edgeworth expansion,

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

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

Measures useful properties

- For univariate x, the cumulants $\mathcal{K}_d(x)$ for $d = 1, 2, 3, 4$ are
  - expectation $\kappa_i = \mathrm{E}(x)$,
  - 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 of κ_{ijk}.
- They provide a natural measure of non-Gaussianity.

Properties of cumulants

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

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

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

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

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

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

Support: There are no distributions where

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

Examples of cumulants

Univariate: $\mathcal{K}_p(x)$ for $p = 1, 2, 3, 4$ are mean, variance, skewness, kurtosis (unnormalized)

Discrete: $x \sim \text{Poi}(\lambda)$, $\mathcal{K}_p(x) = \lambda$ for all p.

Continuous: $x \sim \text{U}([0, 1])$, $\mathcal{K}_p(x) = B_p/p$ where $B_p = p$th Bernoulli number.

Nonexistent: $x \sim \text{t}(3)$, $\mathcal{K}_p(x)$ does not exist for all $p \geq 3$.

Multivariate: $\mathcal{K}_1(\mathbf{x}) = \text{E}(\mathbf{x})$ and $\mathcal{K}_2(\mathbf{x}) = \text{Cov}(\mathbf{x})$.

Discrete: $\mathbf{x} \sim \text{Mult}(n, \mathbf{q})$,

$$\kappa_{j_1 \cdots j_p}(\mathbf{x}) = n \frac{\partial^p}{\partial t_{j_1} \cdots \partial t_{j_p}} \log(q_1 e^{t_1 x_1} + \cdots + q_k e^{t_k x_k}) \Big|_{t_1, \ldots, t_k = 0}.$$

Continuous: $\mathbf{x} \sim \text{N}(\mu, \Sigma)$, $\mathcal{K}_p(\mathbf{x}) = 0$ for all $p \geq 3$.

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

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

In terms of matrix multiplication

Data often presented as $Y \in \mathbb{R}^{m \times n}$, e.g. gene $\times$ microarray, text $\times$ document, person $\times$ image, user $\times$ movie, webpage $\times$ webpage etc.

> *"And so we now have eigengenes, eigenarrays and eigenexpression in the world of transcriptomics, eigenproteins and eigenprofiles in proteomics, eigenpathways in metabolomics, and eigenSNPs in genetics. There are also eigenimages and eigenfaces in image analysis, and eigenpatterns in seismology. In fact, if you put any word you like after eigen- into a Google query box and hit return, I guarantee a result. Yes, even eigenresult and eigenGoogle!"*
>
> — *Terry Speed, IMS Bulletin, April 2008*

- Mean centered, otherwise $\mathbf{y} = \mathbf{x} - \mathrm{E}(\mathbf{x})$.
- $\hat{\mathcal{K}}_1(\mathbf{y}) = \mathbf{0}$.
- $\hat{\mathcal{K}}_2(\mathbf{y}) = \frac{1}{n-1} YY^\top = \frac{1}{n-1}(Y, Y) \cdot I_{n \times n}$.
- $\hat{\mathcal{K}}_3(\mathbf{y}) = \frac{n}{(n-1)(n-2)}(Y, Y, Y) \cdot \mathcal{I}_{n \times n \times n}$.
- $\mathcal{I}_{n \times n \times n} = [\![\delta_{ijk}]\!] \in \mathrm{S}^3(\mathbb{R}^n)$ is the 'Kronecker delta tensor', i.e. $\delta_{ijk} = 1$ if $i = j = k$ and $\delta_{ijk} = 0$ otherwise.

Factor analysis

- Linear generative model

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

  noise $\varepsilon \in \mathbb{R}^m$, 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, $\mathbf{s}$, ε, but need to recover $\mathbf{s}$ and sometimes A from multiple observations of $\mathbf{y}$.
- Time series of observations, get matrices $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$, $S = [\mathbf{s}_1, \ldots, \mathbf{s}_n]$, $E = [\varepsilon_1, \ldots, \varepsilon_n]$, 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: **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: **s** statistically independent entries, ε Gaussian

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

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

What if

- **s** not Gaussian, e.g. power-law distributed data in social networks.
- **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

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

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

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

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

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

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

Recap: tensor ranks

- **Matrix rank.** $A \in \mathbb{R}^{m \times n}$.

$$\begin{aligned} \operatorname{rank}(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet 1}, \dots, A_{\bullet n}\}) && \text{(column rank)} \\ &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{1\bullet}, \dots, A_{m\bullet}\}) && \text{(row rank)} \\ &= \min\{r \mid A = \textstyle\sum_{i=1}^{r} \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}\} && \text{(outer product rank).} \end{aligned}$$

- **Multilinear rank.** $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$. $\operatorname{rank}_{\boxplus}(A) = (r_1(A), r_2(A), r_3(A))$,

$$\begin{aligned} r_1(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{1\bullet\bullet}, \dots, A_{l\bullet\bullet}\}) \\ r_2(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet 1\bullet}, \dots, A_{\bullet m\bullet}\}) \\ r_3(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet\bullet 1}, \dots, A_{\bullet\bullet n}\}) \end{aligned}$$

- **Outer product rank.** $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$.

$$\operatorname{rank}_{\otimes}(\mathcal{A}) = \min\{r \mid \mathcal{A} = \textstyle\sum_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i\}$$

where $\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} := [\![u_i v_j w_k]\!]_{i,j,k=1}^{l,m,n}$.

Recap: matrix EVD and SVD

- Rank revealing decompositions.
- **Symmetric eigenvalue decomposition** of $A \in \mathrm{S}^2(\mathbb{R}^n)$,

$$A = V\Lambda V^\top = \sum_{i=1}^r \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i,$$

  where $\operatorname{rank}(A) = r$, $V \in \mathrm{O}(n)$ eigenvectors, Λ eigenvalues.
- **Singular value decomposition** of $A \in \mathbb{R}^{m\times n}$,

$$A = U\Sigma V^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i$$

  where $\operatorname{rank}(\mathcal{A}) = r$, $U \in \mathrm{O}(m)$ left singular vectors, $V \in \mathrm{O}(n)$ right singular vectors, Σ singular values.
- Ditto for **nonnegative matrix decomposition**.

Recap: one plausible EVD and SVD for hypermatrices

- Rank revealing decompositions associated with the outer product rank.
- **Symmetric outer product decomposition** of $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$,

$$\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i$$

  where $\operatorname{rank}_{\mathrm{S}}(A) = r$, $\mathbf{v}_i$ unit vector, $\lambda_i \in \mathbb{R}$.
- **Outer product decomposition** of $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$,

$$\mathcal{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i$$

  where $\operatorname{rank}_{\otimes}(\mathcal{A}) = r$, $\mathbf{u}_i \in \mathbb{R}^l, \mathbf{v}_i \in \mathbb{R}^m, \mathbf{w}_i \in \mathbb{R}^n$ unit vectors, $\sigma_i \in \mathbb{R}$.
- Ditto for **nonnegative outer product decomposition**.

Recap: another plausible EVD and SVD for hypermatrices

- Rank revealing decompositions associated with the multilinear rank.
- **Singular value decomposition** of $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$,

$$\mathcal{A} = (U, V, W) \cdot \mathcal{C}$$

where $\operatorname{rank}_{\boxplus}(A) = (r_1, r_2, r_3)$, $U \in \mathbb{R}^{l \times r_1}$, $V \in \mathbb{R}^{m \times r_2}$, $W \in \mathbb{R}^{n \times r_3}$ have orthonormal columns and $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.

- **Symmetric eigenvalue decomposition** of $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$,

$$\mathcal{A} = (U, U, U) \cdot \mathcal{C}$$

where $\operatorname{rank}_{\boxplus}(A) = (r, r, r)$, $U \in \mathbb{R}^{n \times r}$ has orthonormal columns and $\mathcal{C} \in \mathrm{S}^3(\mathbb{R}^r)$.

- Ditto for **nonnegative multilinear decomposition**.

Outer product approximation is ill-behaved

- Approximation of a homogeneous polynomial by a sum of powers of linear forms (e.g. Independent Components Analysis).
- Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ be linearly independent. Define for $n \in \mathbb{N}$,

$$A_n := n\left[\mathbf{x} + \frac{1}{n}\mathbf{y}\right]^{\otimes p} - n\mathbf{x}^{\otimes p}$$

- Define

$$\mathcal{A} := \mathbf{x} \otimes \mathbf{y} \otimes \cdots \otimes \mathbf{y} + \mathbf{y} \otimes \mathbf{x} \otimes \cdots \otimes \mathbf{y} + \cdots + \mathbf{y} \otimes \mathbf{y} \otimes \cdots \otimes \mathbf{x}.$$

- Then $\operatorname{rank}_S(\mathcal{A}_n) \leq 2$, $\operatorname{rank}_S(\mathcal{A}) \geq p$, and

$$\lim_{n\to\infty} \mathcal{A}_n = \mathcal{A}.$$

- See [Comon, Golub, L, Mourrain; 08] for details.

Happens to operators too

- Approximation of an operator by a sum of Kronecker product of lower-dimensional operators (e.g. Numerical Operator Calculus).
- For linearly independent operators $P_i, Q_i : V_i \to W_i$, $i = 1, 2, 3$, let $\mathcal{D} : V_1 \otimes V_2 \otimes V_3 \to W_1 \otimes W_2 \otimes W_3$ be

$$\mathcal{D} := P_1 \otimes Q_2 \otimes Q_3 + Q_1 \otimes Q_2 \otimes P_3 + Q_1 \otimes Q_2 \otimes P_3.$$

- If finite-dimensional, then '$\otimes$' may be taken to be Kronecker product of matrices.
- For $n \in \mathbb{N}$,

$$\mathcal{D}_n := n\left[P_1 + \frac{1}{n}Q_1\right] \otimes \left[P_2 + \frac{1}{n}Q_2\right] \otimes \left[P_3 + \frac{1}{n}Q_3\right] - nP_1 \otimes P_2 \otimes P_3.$$

- Then

$$\lim_{n\to\infty} \mathcal{D}_n = \mathcal{D}.$$

Some geometric notions

- Secants of Veronese in $\mathrm{S}^p(\mathbb{R}^n)$ — not closed, not irreducible, difficult to study.
- Symmetric subspace variety in $\mathrm{S}^p(\mathbb{R}^n)$ — 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 $\mathrm{S}^p(\mathbb{R}^n)$ via

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

- More generally

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

From Stieffel to Grassmann

- Given $\mathcal{A} \in \mathrm{S}^p(\mathbb{R}^n)$, some $r \ll n$, want

$$\min_{X \in \mathrm{O}(n,r),\, \mathcal{C} \in \mathrm{S}^p(\mathbb{R}^r)} \|\mathcal{A} - (X, \dots, X) \cdot \mathcal{C}\|_F,$$

- 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 $\mathcal{C}_* \in \mathrm{S}^p(\mathbb{R}^r)$ 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 $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ 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

- 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

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

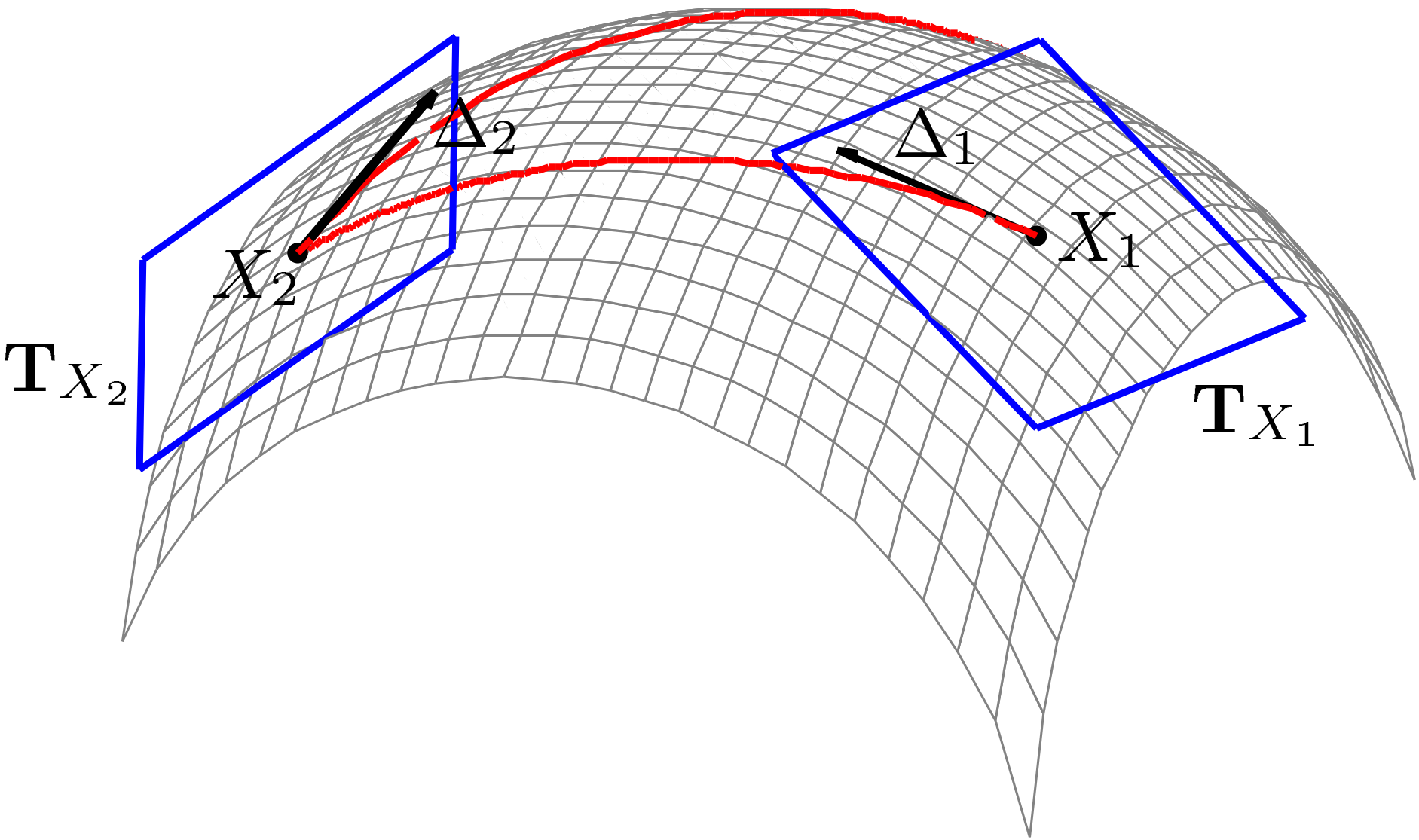
3. At $X \in \mathrm{Gr}(n, r)$, solve

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

   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

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

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

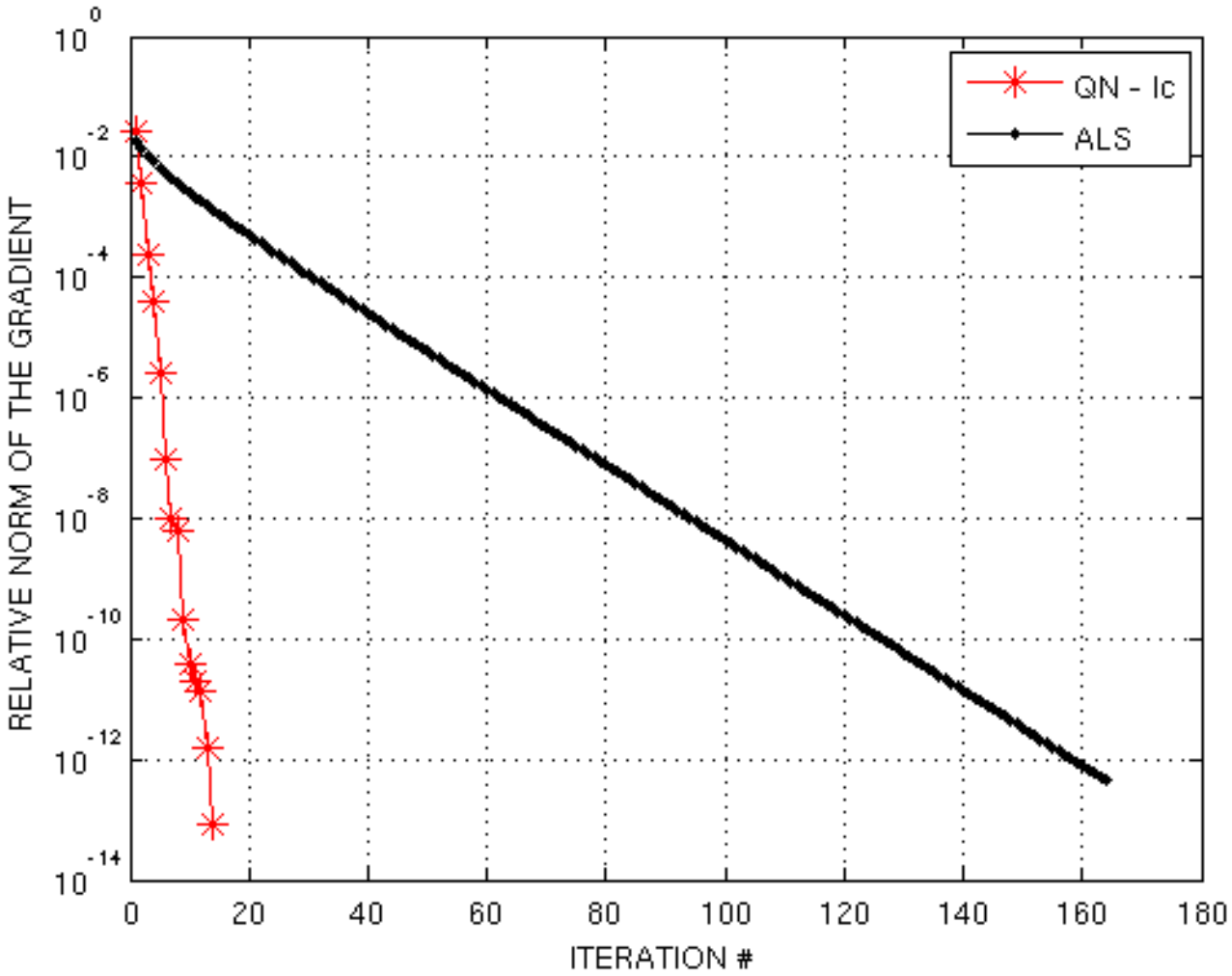
$$\begin{aligned} S_k &= [\mathbf{s}_{k-m}, \dots, \mathbf{s}_{k-1}], \\ Y_k &= [\mathbf{y}_{k-m}, \dots, \mathbf{y}_{k-1}], \\ D_k &= \operatorname{diag}\left[\mathbf{s}_{k-m}^\top \mathbf{y}_{k-m}, \dots, \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}\right], \\ R_k &= \begin{bmatrix} \mathbf{s}_{k-m}^\top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

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,

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

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

Computing image and pixel skewness

- Want to implicitly compute $\mathcal{K}_3^{\text{pixel}} \in \mathrm{S}^3(\mathbb{R}^n)$, 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 $A = U\Sigma V^\top$, $U \in \mathrm{O}(n, m)$, $\Sigma \in \mathbb{R}^{m\times m}$, $V \in \mathrm{O}(m)$.

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

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

Computing skewmax projection

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

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

- Want $Q \in \mathrm{O}(m, s)$ and core tensor $\mathcal{C} \in \mathrm{S}^3(\mathbb{R}^s)$ 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

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

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

Combining eigen-, skew-, and kurtofaces

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

Acknowledgement

Joint work with:

- Jason Morton, Stanford University
- Berkant Savas, University of Texas at Austin