Analysis of cumulants

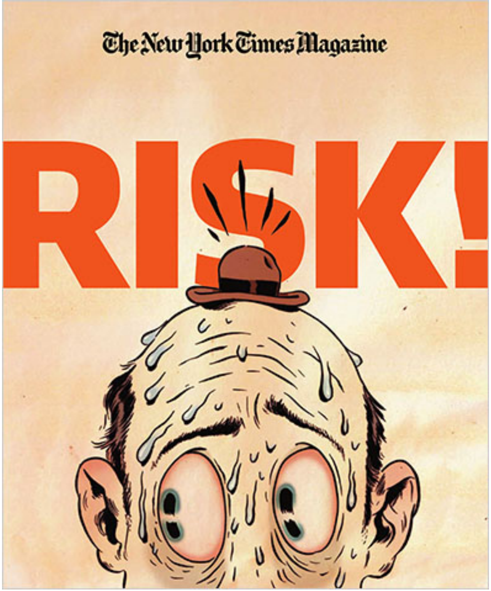
Lek-Heng Lim

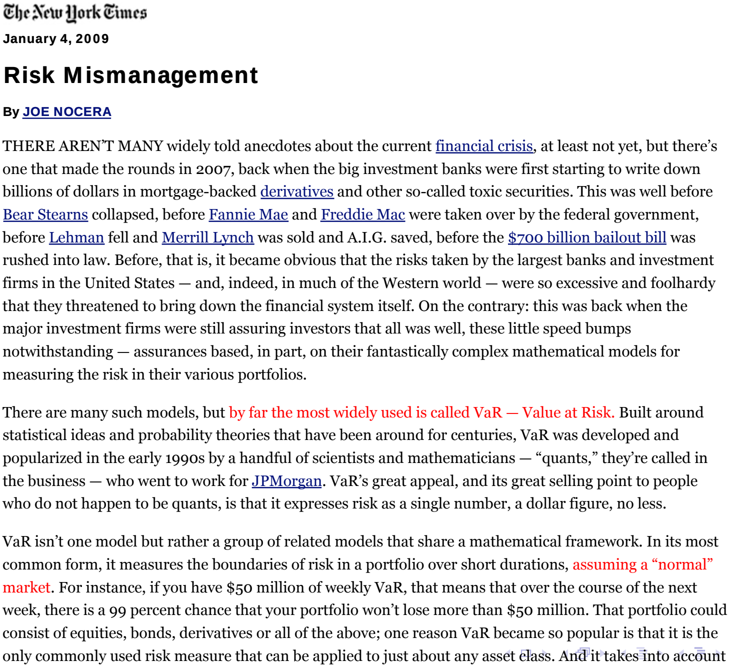
University of California, Berkeley

March 6, 2009

Joint work with Jason Morton

Cover Story: January 4, 2009





The New York Times

January 4, 2009

Risk Mismanagement

By JOE NOCERA

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 measuring the risk in their various portfolios.

There are many such models, but by far the most widely used is called VaR — Value at Risk. Built around statistical ideas and probability theories that have been around for centuries, VaR was developed and popularized in the early 1990s by a handful of scientists and mathematicians — "quants," they're called in the business — who went to work for JPMorgan. VaR's great appeal, and its great selling point to people who do not happen to be quants, is that it expresses risk as a single number, a dollar figure, no less.

VaR isn't one model but rather a group of related models that share a mathematical framework. In its most common form, it measures the boundaries of risk in a portfolio over short durations, assuming a "normal" market. For instance, if you have $50 million of weekly VaR, that means that over the course of the next week, there is a 99 percent chance that your portfolio won't lose more than $50 million. That portfolio could consist of equities, bonds, derivatives or all of the above; one reason VaR became so popular is that it is the only commonly used risk measure that can be applied to just about any asset class. And it takes into account

Why not Gaussian

- Log characteristic function

$$\log \mathrm{E}(\exp(i\langle \mathbf{t}, \mathbf{x}\rangle)) = \sum_{|\alpha|=1}^{\infty} i^{|\alpha|} \kappa_{\alpha}(\mathbf{x}) \frac{\mathbf{t}^{\alpha}}{\alpha!}.$$

- Gaussian assumption:

$$\infty = 2.$$

- If $\mathbf{x}$ is multivariate Gaussian, then

$$\log \mathrm{E}(\exp(i\langle \mathbf{t}, \mathbf{x}\rangle)) = i\langle \mathrm{E}(\mathbf{x}), \mathbf{t}\rangle + \frac{1}{2}\mathbf{t}^{\top} \operatorname{Cov}(\mathbf{x})\mathbf{t}.$$

- $\mathcal{K}_1(\mathbf{x})$ mean, $\mathcal{K}_2(\mathbf{x})$ (co)variance, $\mathcal{K}_3(\mathbf{x})$ (co)skewness, $\mathcal{K}_4(\mathbf{x})$ (co)kurtosis,. . . .
- Paul Wilmott: "The relationship between two assets can never be captured by a single scalar quantity."

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.

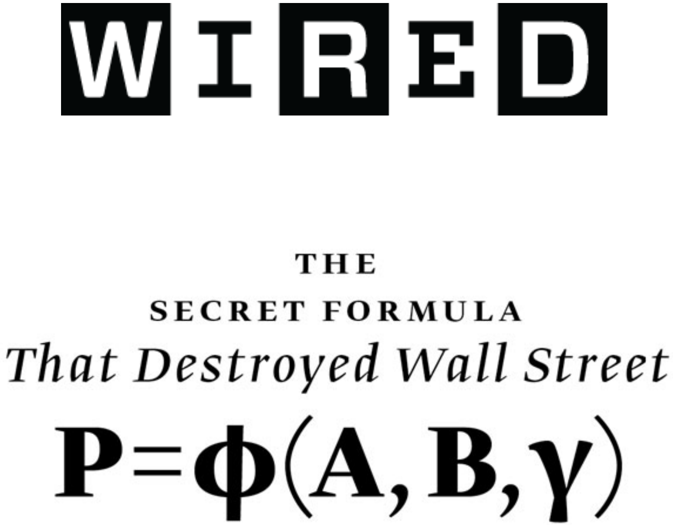
“This is a Stan O’Neal trade,” he said, referring to the former chief executive of Merrill Lynch. He clicked to a slide that showed a trade that made slow, steady profits — and then quickly spiraled downward for a giant, brutal loss.

“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.”

Eventually, though, you do start to get the point. Taleb says that Wall Street risk models, no matter how

Cover Story: March 1, 2009

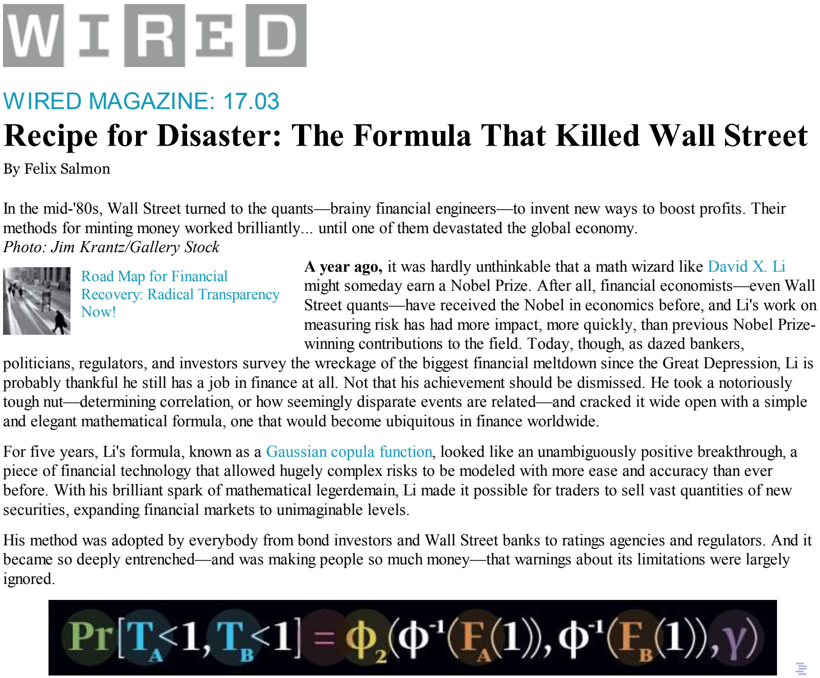


WIRED

THE SECRET FORMULA

That Destroyed Wall Street

$$\mathbf{P}=\boldsymbol{\phi}(\mathbf{A},\mathbf{B},\boldsymbol{\gamma})$$



WIRED

WIRED MAGAZINE: 17.03

Recipe for Disaster: The Formula That Killed Wall Street

By Felix Salmon

In the mid-'80s, Wall Street turned to the quants—brainy financial engineers—to invent new ways to boost profits. Their methods for minting money worked brilliantly... until one of them devastated the global economy.
Photo: Jim Krantz/Gallery Stock

Road Map for Financial Recovery: Radical Transparency Now!

A year ago, it was hardly unthinkable that a math wizard like David X. Li might someday earn a Nobel Prize. After all, financial economists—even Wall Street quants—have received the Nobel in economics before, and Li's work on measuring risk has had more impact, more quickly, than previous Nobel Prize-winning contributions to the field. Today, though, as dazed bankers, politicians, regulators, and investors survey the wreckage of the biggest financial meltdown since the Great Depression, Li is probably thankful he still has a job in finance at all. Not that his achievement should be dismissed. He took a notoriously tough nut—determining correlation, or how seemingly disparate events are related—and cracked it wide open with a simple and elegant mathematical formula, one that would become ubiquitous in finance worldwide.

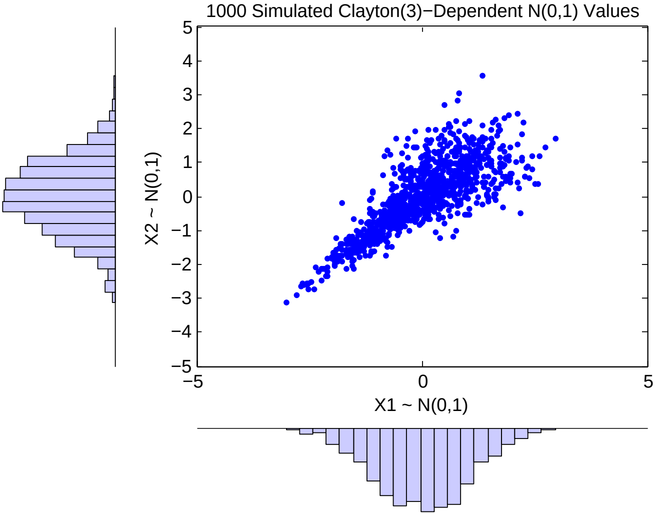
For five years, Li's formula, known as a Gaussian copula function, looked like an unambiguously positive breakthrough, a piece of financial technology that allowed hugely complex risks to be modeled with more ease and accuracy than ever before. With his brilliant spark of mathematical legerdemain, Li made it possible for traders to sell vast quantities of new securities, expanding financial markets to unimaginable levels.

His method was adopted by everybody from bond investors and Wall Street banks to ratings agencies and regulators. And it became so deeply entrenched—and was making people so much money—that warnings about its limitations were largely ignored.

$$\Pr[T_A<1, T_B<1] = \phi_2(\phi^{-1}(F_A(1)), \phi^{-1}(F_B(1)), \gamma)$$

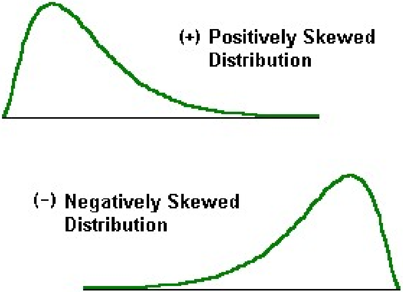
Why not copulas

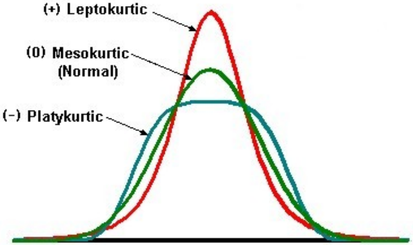
- Nassim Taleb: “Anything that relies on correlation is charlatanism.”
- Even if marginals normal, dependence might not be.



Cumulants

- **Univariate distribution:** First four cumulants are
  - mean $\mathcal{K}_1(x) = \mathrm{E}(x) = \mu$,
  - variance $\mathcal{K}_2(x) = \mathrm{Var}(x) = \sigma^2$,
  - skewness $\mathcal{K}_3(x) = \sigma^3 \, \mathrm{Skew}(x)$,
  - kurtosis $\mathcal{K}_4(x) = \sigma^4 \, \mathrm{Kurt}(x)$.


(+) Positively Skewed Distribution
(−) Negatively Skewed Distribution



- **Multivariate distribution:** Covariance matrix *partly* describes the dependence structure — enough for Gaussian. Cumulants describe higher order dependence among random variables.

Cumulants

- For multivariate $\mathbf{x}$, $\mathcal{K}_d(\mathbf{x}) = [\![\kappa_{j_1\cdots j_d}(\mathbf{x})]\!]$ are symmetric tensors of order d.
- In terms of Edgeworth expansion,

$$\log \mathrm{E}(\exp(i\langle \mathbf{t}, \mathbf{x}\rangle) = \sum_{|\alpha|=1}^{\infty} i^{|\alpha|}\kappa_\alpha(\mathbf{x})\frac{\mathbf{t}^\alpha}{\alpha!}, \quad \log \mathrm{E}(\exp(\langle \mathbf{t}, \mathbf{x}\rangle) = \sum_{|\alpha|=1}^{\infty} \kappa_\alpha(\mathbf{x})\frac{\mathbf{t}^\alpha}{\alpha!},$$

  $\alpha = (j_1, \dots, j_n)$ is a multi-index, $\mathbf{t}^\alpha = t_1^{j_1}\cdots t_n^{j_n}$, $\alpha! = j_1!\cdots j_n!$.
- Provide a natural measure of non-Gaussianity: If $\mathbf{x}$ Gaussian,

$$\mathcal{K}_d(\mathbf{x}) = 0 \quad \text{for all } d \geq 3.$$

- Gaussian assumption equivalent to quadratic approximation.
- **Non-Gaussian data:** Not enough to look at just mean and covariance.

Examples of cumulants

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

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

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

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

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

Discrete: $\mathbf{x} \sim \textsc{Multinomial}(n, \mathbf{q})$,

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

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

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

$$\mathcal{A} = \llbracket a_{ijk} \rrbracket_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$$

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.

Probably the source

Woldemar Voigt, *Die fundamentalen physikalischen Eigenschaften der Krystalle in elementarer Darstellung*, Verlag Von Veit, Leipzig, 1898.

> *"An abstract entity represented by an array of components that are functions of co-ordinates such that, under a transformation of co-ordinates, the new components are related to the transformation and to the original components in a* **definite way***."*

Definite way: multilinear matrix multiplication

- Correspond to change-of-bases transformations for tensors.
- 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: $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$, $X \in \mathbb{R}^{p \times l}$, $Y \in \mathbb{R}^{q \times m}$, $Z \in \mathbb{R}^{r \times n}$,

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

- ‘Right’ (covariant) multiplication: $(X, Y, Z) \cdot \mathcal{A} := \mathcal{A} \cdot (X^{\top}, Y^{\top}, Z^{\top})$.

Tensors inevitable in multivariate problems

- Expand multivariate $f(x_1, \dots, x_n)$ in power series

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

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

- a_0 scalar, $\mathbf{a}_1$ vector, A_2 matrix, $\mathcal{A}_d$ tensor of order d.
- **Lesson:** Important to look beyond the quadratic term.
- **Objective:** Want to better understand tensor-valued quantities.

Examples

- Mathematics
  - **Derivatives of univariate functions:** $f : \mathbb{R} \to \mathbb{R}$ smooth, $f'(x), f''(x), \dots, f^{(k)}(x) \in \mathbb{R}$.
  - **Derivatives of multivariate functions:** $f : \mathbb{R}^n \to \mathbb{R}$ smooth, $\operatorname{grad} f(\mathbf{x}) \in \mathbb{R}^n, \operatorname{Hess} f(\mathbf{x}) \in \mathbb{R}^{n\times n}, \dots, D^{(k)} f(\mathbf{x}) \in \mathbb{R}^{n\times\cdots\times n}$.
- Statistics
  - **Cumulants of random variables:** $\mathcal{K}_d(x) \in \mathbb{R}$.
  - **Cumulants of random vectors:** $\mathcal{K}_d(\mathbf{x}) = [\![\kappa_{j_1\cdots j_d}(\mathbf{x})]\!] \in \mathbb{R}^{n\times\cdots\times n}$.
- Physics
  - **Hooke's law in 1D:** x extension, F force, k spring constant,

$$F = -kx.$$

  - **Hooke's law in 3D:** $\mathbf{x} = (x_1, x_2, x_3)^\top$, elasticity tensor $\mathcal{C} \in \mathbb{R}^{3\times3\times3\times3}$, stress $\Sigma \in \mathbb{R}^{3\times3}$, strain $\Gamma \in \mathbb{R}^{3\times3}$

$$\sigma_{ij} = \sum\nolimits_{k,l=1}^{3} c_{ijkl}\gamma_{kl}.$$

Tensors in physics

- **Hooke's law again:** At a point $\mathbf{x} = (x_1, x_2, x_3)^\top$ in a linear anisotropic solid,

$$\sigma_{ij} = \sum\nolimits_{k,l=1}^{3} c_{ijkl}\gamma_{kl} - \sum\nolimits_{k=1}^{3} b_{ijk}e_k - ta_{ij}$$

  where elasticity tensor $\mathcal{C} \in \mathbb{R}^{3\times3\times3\times3}$, piezoelectric tensor $\mathcal{B} \in \mathbb{R}^{3\times3\times3}$, thermal tensor $A \in \mathbb{R}^{3\times3}$, stress $\Sigma \in \mathbb{R}^{3\times3}$, strain $\Gamma \in \mathbb{R}^{3\times3}$, electric field $\mathbf{e} \in \mathbb{R}^3$, temperature change $t \in \mathbb{R}$.
- **Invariant under change-of-coordinates:** If $\mathbf{y} = Q\mathbf{x}$, then

$$\bar{\sigma}_{ij} = \sum\nolimits_{k,l=1}^{3} \bar{c}_{ijkl}\bar{\gamma}_{kl} - \sum\nolimits_{k=1}^{3} \bar{b}_{ijk}\bar{e}_k - t\bar{a}_{ij}$$

  where

$$\overline{\mathcal{C}} = (Q, Q, Q, Q)\cdot\mathcal{C}, \quad \overline{\mathcal{B}} = (Q, Q, Q)\cdot\mathcal{B}, \quad \overline{A} = (Q, Q)\cdot A,$$
$$\overline{\Sigma} = (Q, Q)\cdot\Sigma, \quad \overline{\Gamma} = (Q, Q)\cdot\Gamma, \quad \bar{\mathbf{e}} = Q\mathbf{e}.$$

Tensors in computer science

- For $A = [a_{ij}], B = [b_{jk}] \in \mathbb{R}^{n\times n}$,

$$AB = \sum\nolimits_{i,j,k=1}^{n} a_{ik} b_{kj} E_{ij} = \sum\nolimits_{i,j,k=1}^{n} \varphi_{ik}(A)\varphi_{kj}(B) E_{ij}$$

  where $E_{ij} = \mathbf{e}_i \mathbf{e}_j^\top \in \mathbb{R}^{n\times n}$. Let

$$\mathcal{T}_n = \sum\nolimits_{i,j,k=1}^{n} \varphi_{ik} \otimes \varphi_{kj} \otimes E_{ij}.$$

- $\mathcal{T}_n$ is a tensor of order 3.
- $O(n^{2+\varepsilon})$ algorithm for multiplying two $n \times n$ matrices gives $O(n^{2+\varepsilon})$ algorithm for solving system of n linear equations [Strassen; 1969].
- **Conjecture.** $\operatorname{rank}_\otimes(T_n) = O(n^{2+\varepsilon})$.

Tensors in statistics

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

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

Humans cannot understand ‘raw’ tensors

Humans cannot make sense out of more than $O(n)$ numbers. For most people, $5 \leq n \leq 9$ [Miller; 1956].

- VaR: single number
  - Readily understandable.
  - Not sufficiently informative and discriminative.
- Covariance matrix: $O(n^2)$ numbers
  - Hard to make sense of without further processing.
  - For symmetric matrices, may perform eigenvalue decomposition.
  - Basis for PCA, MDS, ISOMAP, LLE, Laplacian Eigenmap, etc.
  - Used in clustering, classification, dimension reduction, feature identification, learning, prediction, visualization, etc.
- Cumulant of order d: $O(n^d)$ numbers
  - How to make sense of these?
  - Want analogue of ‘eigenvalue decomposition’ for symmetric tensors.
  - Principal Cumulant Component Analysis: finding components that simultaneously account for variation in cumulants of all orders.

Analyzing matrices

- Orbits of group action on $\mathbb{R}^{m\times n}$, $A \mapsto XAY^{-1}$
  - $\mathrm{GL}(m) \times \mathrm{GL}(n)$: $A = L_1 D L_2^\top$
    - L_1, L_2 unit lower triangular, D diagonal
  - $\mathrm{O}(m) \times \mathrm{O}(n)$: $A = U\Sigma V^\top$
    - U, V orthogonal, Σ diagonal
- Orbits of group action on $\mathbb{C}^{n\times n}$, $A \mapsto XAX^{-1}$
  - $\mathrm{GL}(n)$: $A = SJS^{-1}$
    - S nonsingular, J Jordan form
  - $\mathrm{O}(n)$: $A = QRQ^\top$
    - Q orthogonal, R upper triangular
- Orbits of group action on $\mathrm{S}^2(\mathbb{R}^n)$, $A \mapsto XAX^\top$
  - $\mathrm{O}(n)$: $A = Q\Lambda Q^\top$
    - Q orthogonal, Λ diagonal

What about tensors?

- Orbits of $\mathrm{GL}(m) \times \mathrm{GL}(n)$ action on matrix pencils $\mathbb{R}^{2\times m\times n}$.
  $(A_1, A_2) = (SK_1T^{-1}, SK_2T^{-1})$ where (K_1, K_2) Kronecker form:

$$\left(\begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}, \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & \ddots & \\ & & \ddots & 0 \\ & & & 1 \end{bmatrix} \right) \in \mathbb{R}^{2\times(p+1)\times p},$$

$$\left(\begin{bmatrix} 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \end{bmatrix} \right) \in \mathbb{R}^{2\times q\times(q+1)},$$

$$\left(\begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}, \begin{bmatrix} 0 & & & -a_0 \\ 1 & \ddots & & \vdots \\ & \ddots & 0 & -a_{r-2} \\ & & 1 & -a_{r-1} \end{bmatrix} \right) \in \mathbb{R}^{2\times r\times r}.$$

- No nice orbit classification for 3-tensors $\mathbb{R}^{l\times m\times n}$ when $l > 2$.
- Even harder for k-tensors $\mathbb{R}^{d_1\times\cdots\times d_k}$.

Another view of EVD and SVD

- Linear combination of rank-1 matrices.
- **Symmetric eigenvalue decomposition** of $A \in \mathsf{S}^2(\mathbb{R}^n)$,

$$A = V \Lambda V^\top = \sum_{i=1}^{\operatorname{rank}(A)} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i.$$

- **Singular value decomposition** of $A \in \mathbb{R}^{m \times n}$,

$$A = U \Sigma V^\top = \sum_{i=1}^{\operatorname{rank}(A)} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i.$$

- Nonlinear approximation: best r-term approximation by a dictionary of atoms.

Secant varieties

- For a nondegenerate variety $X \subseteq \mathbb{R}^n$, write

$$s_r(X) = \text{union of } s\text{-secants to } X \text{ for } s \leq r,$$

r-**secant quasiprojective variety** of X.

- *r*-**secant variety**,

$$\sigma_r(X) = \text{Zariski closure of } s_r(X).$$

- For purpose of analyzing tensors, X often one of the following:

$$\operatorname{Seg}(\mathbb{R}^{d_1}, \dots, \mathbb{R}^{d_p}) = \{\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_p \in \mathbb{R}^{d_1 \times \cdots \times d_p} \mid \mathbf{v}_i \in \mathbb{R}^{d_i}\},$$
$$\operatorname{Ver}_p(\mathbb{R}^n) = \{\mathbf{v} \otimes \cdots \otimes \mathbf{v} \in \mathsf{S}^p(\mathbb{R}^n) \mid \mathbf{v} \in \mathbb{R}^n\}.$$

- Respectively 'rank-1 tensors' and 'rank-1 symmetric tensors':

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

- Serious difficulty: $s_r(X) \neq \sigma_r(X)$ for $r > 1$, cf. [de Silva, L.; 08].

Other forms

- **Approximation theory:** Decomposing function into linear combination of separable functions,

$$f(x,y,z) = \sum\nolimits_{i=1}^{r} \lambda_i \varphi_i(x) \psi_i(y) \theta_i(z).$$

  Application: separation of variables for PDEs.

- **Operator theory:** Decomposing operator into linear combination of Kronecker products,

$$\Delta_3 = \Delta_1 \otimes I \otimes I + I \otimes \Delta_1 \otimes I + I \otimes I \otimes \Delta_1.$$

  Application: numerical operator calculus.

Other forms

- **Commutative algebra:** Decomposing homogeneous polynomial into linear combination of powers of linear forms,

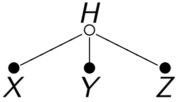
$$p_d(x, y, z) = \sum\nolimits_{i=1}^{r} \lambda_i (a_i x + b_i y + c_i z)^d.$$

  Application: independent components analysis.

- **Probability theory:** Decomposing probability density into conditional densities of random variables satisfying naïve Bayes:

$$\Pr(x, y, z) = \sum\nolimits_{h} \Pr(h) \Pr(x \mid h) \Pr(y \mid h) \Pr(z \mid h).$$

  Application: probabilistic latent semantic indexing.



Analyzing tensors

- $A \in \mathbb{R}^{m \times n}$.
  - **Singular value decomposition:**

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

    where $\operatorname{rank}(A) = r$, U, V orthonormal columns, $\Sigma = \operatorname{diag}[\sigma_1, \dots, \sigma_r]$.
- $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$. Can either keep diagonality of Σ or orthogonality of U and V but not both.
  - **Linear combination:**

    $$\mathcal{A} = (X, Y, Z) \cdot \Sigma = \sum\nolimits_{i=1}^{r} \sigma_i \mathbf{x}_i \otimes \mathbf{y}_i \otimes \mathbf{z}_i$$

    where $\operatorname{rank}_\otimes(A) = r$, X, Y, Z matrices, $\Sigma = \operatorname{diag}_{r \times r \times r}[\sigma_1, \dots, \sigma_r]$; r may exceed n.
  - **Multilinear combination:**

    $$\mathcal{A} = (U, V, W) \cdot \mathcal{C} = \sum\nolimits_{i,j,k=1}^{r_1, r_2, r_3} c_{ijk} \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{w}_k$$

    where $\operatorname{rank}_\boxplus(A) = (r_1, r_2, r_3)$, U, V, W orthonormal columns, $\mathcal{C} = [\![c_{ijk}]\!] \in \mathbb{R}^{r_1 \times r_2 \times r_3}$; $r_1, r_2, r_3 \le n$.

Tensor ranks (Hitchcock, 1927)

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

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

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

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

- In general, $r_1(\mathcal{A}) \neq r_2(\mathcal{A}) \neq r_3(\mathcal{A}) \neq \operatorname{rank}_{\otimes}(\mathcal{A})$.

Symmetric tensors as hypermatrices

- 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

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

Tensor ranks (Hitchcock, 1927)

- **Multilinear rank.** $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$. Then

$$r_1(\mathcal{A}) = r_2(\mathcal{A}) = r_3(\mathcal{A}).$$

- **Outer product rank.** $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$.

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

DARPA mathematical challenge eight

One of the twenty three mathematical challenges announced at DARPA Tech 2007.

Problem

Beyond convex optimization: *can linear algebra be replaced by algebraic geometry in a systematic way?*

- **Algebraic geometry in a slogan:** polynomials are to algebraic geometry what matrices are to linear algebra.
- Polynomial $f \in \mathbb{R}[x_1, \dots, x_n]$ of degree d can be expressed as

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

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

Symmetric tensors as polynomials

- $[\![a_{j_1\cdots j_p}]\!] \in \mathsf{S}^p(\mathbb{R}^n)$ associated with unique homogeneous polynomial $F \in \mathbb{C}[x_1, \dots, x_n]_p$ via

$$F(\mathbf{x}) = \sum\nolimits_{j_1,\dots,j_p=1}^{n} \binom{p}{d_1, \dots, d_n} a_{j_1\cdots j_p} x_1^{d_1} \cdots x_n^{d_n},$$

  $d_j =$ number of times index j appears in $j_1, \dots, j_p$,

$$d_1 + \cdots + d_n = p.$$

- $\mathsf{S}^p(\mathbb{C}^n) \cong \mathbb{C}[x_1, \dots, x_n]_p$.
- Waring problem for polynomials: what's the smallest r so that

$$p_d(x, y, z) = \sum\nolimits_{i=1}^{r} \lambda_i (a_i x + b_i y + c_i z)^d.$$

- Alexander-Hirschowitz theorem: what's the generic r?
- Equivalent formulation on the next slide.

One plausible EVD and SVD for hypermatrices

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

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

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

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

  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, $\sigma_i \in \mathbb{R}$.

Geometry of symmetric outer product decomposition

- Embedding

$$\nu_{n,p} : \mathbb{R}^n \to \mathsf{S}^p(\mathbb{R}^n) \cong \mathbb{R}[x_1, \dots, x_n]_p.$$

- Image $\nu_{n,p}(\mathbb{R}^n)$ is (real affine) **Veronese variety**, set of rank-1 symmetric tensors

$$\operatorname{Ver}_p(\mathbb{R}^n) = \{\mathbf{v}^{\otimes p} \in \mathsf{S}^p(\mathbb{R}^n) \mid \mathbf{v} \in \mathbb{R}^n\}.$$

- As polynomials,

$$\operatorname{Ver}_p(\mathbb{R}^n) = \{L(\mathbf{x})^p \in \mathbb{R}[x_1, \dots, x_n]_p \mid L(\mathbf{x}) = \alpha_1 x_1 + \dots + \alpha_n x_n\}.$$

- $\mathcal{A} \in \mathsf{S}^p(\mathbb{R}^n)$ has rank 2 iff it sits on a bisecant line through two points of $\operatorname{Ver}_p(\mathbb{R}^n)$, rank 3 iff it sits on a trisecant plane through three points of $\operatorname{Ver}_p(\mathbb{R}^n)$, etc.

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 $\operatorname{rank}_{\mathsf{S}}(\mathcal{A}_n) \leq 2$, $\operatorname{rank}_{\mathsf{S}}(\mathcal{A}) \geq p$, and

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

- See [Comon, Golub, L, Mourrain; 08] for details. Exact decomposition when $r < n$, algorithm of [Comon, Mourrain, Tsigaridas; 09]

Inherent difficulty

- The best r-term approximation problem for tensors has no solution in general (except for the nonnegative case).
- Eugene Lawler: "The Mystical Power of Twoness."
  - 2-SAT is easy, 3-SAT is hard;
  - 2-dimensional matching is easy, 3-dimensional matching is hard;
  - 2-body problem is easy, 3-body problem is hard;
  - 2-dimensional Ising model is easy, 3-dimensional Ising model is hard.
- Applies to tensors too:
  - 2-tensor rank is easy, 3-tensor rank is hard;
  - 2-tensor spectral norm is easy, 3-tensor spectral norm is hard;
  - 2-tensor approximation is easy, 3-tensor approximation is hard;
  - 2-tensor eigenvalue problem is easy, 3-tensor eigenvalue problem is hard.

Another plausible EVD and SVD for hypermatrices

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

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

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

- **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 $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.

Eigenvalue decompositions for symmetric tensors

Let $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$.

- **Symmetric outer product decomposition**

$$\mathcal{A} = \sum\nolimits_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i = (V, V, V) \cdot \Lambda.$$

  where $\operatorname{rank}_{\mathsf{S}}(A) = r$, $V \in \mathbb{R}^{n \times r}$, $\Lambda = \operatorname{diag}[\lambda_1, \dots, \lambda_r] \in \mathsf{S}^3(\mathbb{R}^n)$.

- In general, r can exceed n.
- **Symmetric multilinear decomposition**

$$\mathcal{A} = (U, U, U) \cdot \mathcal{C} = \sum\nolimits_{i,j,k=1}^{s} c_{ijk} \mathbf{u}_i \otimes \mathbf{u}_j \otimes \mathbf{u}_k$$

  where $\operatorname{rank}_{\boxplus}(A) = (s, s, s)$, $U \in \mathsf{O}(n, s)$, $\mathcal{C} = [\![c_{ijk}]\!] \in \mathsf{S}^3(\mathbb{R}^s)$.

- $s \leq n$.

Geometry of symmetric multilinear decomposition

- **Symmetric subspace variety**,

$$\begin{aligned}\mathrm{Sub}_r^p(\mathbb{R}^n) &= \{\mathcal{A} \in \mathsf{S}^p(\mathbb{R}^n) \mid \exists V \leq \mathbb{R}^n \text{ such that } \mathcal{A} \in \mathsf{S}^p(V)\} \\ &= \{\mathcal{A} \in \mathsf{S}^p(\mathbb{R}^n) \mid \operatorname{rank}_{\boxplus}(\mathcal{A}) \leq (r, r, r)\}.\end{aligned}$$

- Unsymmetric version,

$$\begin{aligned}&\mathrm{Sub}_{p,q,r}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n) \\ &\quad = \{\mathcal{A} \in \mathbb{R}^{l\times m\times n} \mid \exists U, V, W \text{ such that } \mathcal{A} \in U \otimes V \otimes W\} \\ &\quad\quad = \{\mathcal{A} \in \mathbb{R}^{l\times m\times n} \mid \operatorname{rank}_{\boxplus}(\mathcal{A}) \leq (p, q, r)\}.\end{aligned}$$

- Symmetric subspace variety in $\mathsf{S}^p(\mathbb{R}^n)$ — closed, irreducible, easier to study.
- Quasiprojective secant variety of Veronese in $\mathsf{S}^p(\mathbb{R}^n)$ — not closed, not irreducible, difficult to study.
- Reference: Forthcoming book [Landsberg, Morton; 09]

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 $Y = [\mathbf{y}_1, \dots, \mathbf{y}_n]$, $S = [\mathbf{s}_1, \dots, \mathbf{s}_n]$, $E = [\varepsilon_1, \dots, \varepsilon_n]$, 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

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

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

What if

- **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

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

- We have assumed $A = Q \in \mathrm{O}(n, r)$ since otherwise $A = QR$ and

$$\mathcal{K}_p(A\mathbf{s}) = (Q, \dots, Q) \cdot [(R, \dots, R) \cdot \mathcal{K}_p(\mathbf{s})].$$

- Recover orthonormal basis of subspace spanned by A.
- $\mathcal{C}_p \approx (R, \dots, R) \cdot \hat{\mathcal{K}}_p(\mathbf{s})$ not necessarily diagonal.

PCCA optimization

- Appears intractable: optimization over infinite-dimensional manifold
$$\mathrm{O}(n, r) \times \prod_{p=1}^{\infty} \mathsf{S}^p(\mathbb{R}^r).$$
- Surprising relaxation: optimization over a single Grassmannian $\mathrm{Gr}(n, r)$ of dimension $r(n-r)$,
$$\max_{Q \in \mathrm{Gr}(n,r)} \sum_{p=1}^{\infty} \alpha_p \|\hat{\mathcal{K}}_p(\mathbf{y}) \cdot (Q, \dots, Q)\|_F^2.$$
- In practice $\infty = 3$ or 4.

Grassmannian parameterization

- 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 via
$$\mathrm{Gr}(n,r) \times \mathsf{S}^p(\mathbb{R}^r) \to \mathsf{S}^p(\mathbb{R}^n).$$
- Image is $\mathrm{Sub}_r^p(\mathbb{R}^n)$.
- More generally
$$\mathrm{Gr}(n,r) \times \prod_{p=1}^{\infty} \mathsf{S}^p(\mathbb{R}^r) \to \prod_{p=1}^{\infty} \mathsf{S}^p(\mathbb{R}^n).$$
- Image is $\prod_{p=1}^{\infty} \mathrm{Sub}_r^p(\mathbb{R}^n)$.

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
$$\max_{X \in \mathrm{Gr}(n,r)} \|\mathcal{A} \cdot (X, \dots, X)\|_F\,.$$
- Efficient algorithm exists: Limited memory BFGS on Grassmannian [Savas, L; '09]

Thanks

- Details and data experiments: forthcoming paper on “Principal Cumulant Component Analysis” by Morton and L.