PRINCIPAL CUMULANT COMPONENT ANALYSIS

JASON MORTON AND LEK-HENG LIM

Abstract. Multivariate Gaussian data is completely characterized by its mean and covariance, yet modern non-Gaussian data makes higher-order statistics such as cumulants inevitable. For univariate data, the third and fourth scalar-valued cumulants are relatively well-studied as skewness and kurtosis. For multivariate data, these cumulants are tensor-valued, higher-order analogs of the covariance matrix capturing higher-order dependence in the data. In addition to their relative obscurity, there are few effective methods for analyzing these cumulant tensors. We propose a technique along the lines of Principal Component Analysis and Independent Component Analysis to analyze multivariate, non-Gaussian data motivated by the multilinear algebraic properties of cumulants. Our method relies on finding principal cumulant components that account for most of the variation in *all* higher-order cumulants, just as PCA obtains varimax components. An efficient algorithm based on limited-memory quasi-Newton maximization over a Grassmannian, using only standard matrix operations, may be used to find the principal cumulant components. Numerical experiments include forecasting higher portfolio moments and image dimension reduction.

1. Introduction

Data arising from modern-day applications like computational biology, computer vision, and finance are rarely well described by a multivariate Gaussian distribution; we need to examine not just the mean and covariance matrix, but higher order structure in the data. A timely example of the consequence of ignoring non-Gaussianity is the financial crisis, which has been attributed in part to the over-reliance on measures of risk appropriate primarily for Gaussians. Emphasizing variance allows risks to creep in through the higher moments. For a single continuous variable, mean, variance, skewness and kurtosis are commonly used to describe the distribution's shape. A negatively skewed distribution has a longer left tail and more mass on the right as compared to the Gaussian; a leptokurtotic distribution has fatter tails and a more concentrated central peak (Figure 1).

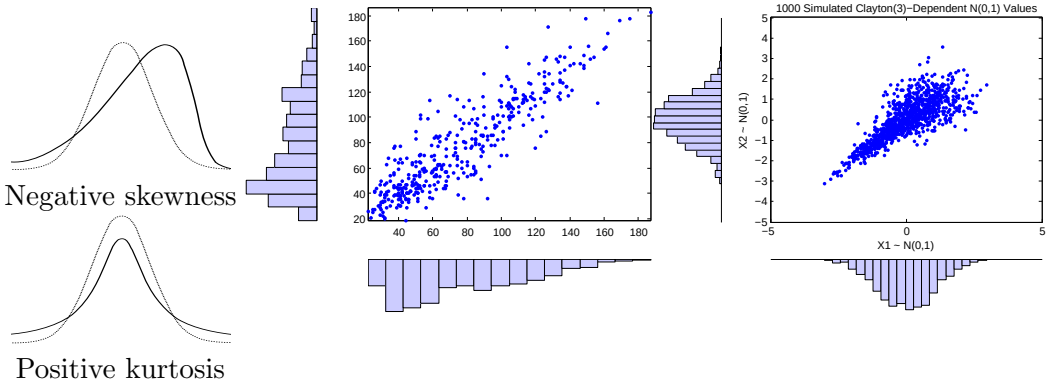


Figure 1. Left: Skewness and kurtosis describe univariate distribution shape for unimodal distributions. Middle: joint and marginals for two pixels in the ORL face database. Right: Gaussian marginals but not dependence.

For continuous multivariate data, covariance matrices partially describe the dependence structure. When the variables are multivariate Gaussian, this description is complete. Rank restrictions on the covariance matrix yields Principal Component Analysis (PCA) . The covariance matrix plays a critical role in optimization in finance and other areas involving optimization of risky payoffs, since it is the bilinear form which computes the variance of a linear combination of variables.

For multivariate, non-Gaussian data, the covariance matrix is an incomplete description of the dependence structure. Note that even with Gaussian marginals, dependence may be non-Gaussian (Figure 1). *Cumulant tensors* are the multivariate generalization of univariate skewness and kurtosis and the higher-order generalization of covariance matrices. For a $\mathbb{R}^p$-valued random variable, the tensor corresponding to skewness is a symmetric $p \times p \times p$ array, while the kurtosis tensor is symmetric $p \times p \times p \times p$. Analogously to the covariance matrix, these are *multilinear* forms that compute the skewness and kurtosis of a linear combination of variables. Unlike the covariance case and ICA [3, 6], these tensors *cannot in general be diagonalized*, so we must make other choices in modeling them.

Introduced as half-invariants [21], cumulants fell into disuse as they require large amounts of data to estimate and there were few methods to analyze them. We propose a simple analytic method with a computational and decompositional [20] flavor, overcoming several of the practical and theoretical barriers to their wider use. PCA finds orthogonal components that best explain the variation in the second-order cumulant (the covariance matrix). This is fine if the data is Gaussian since all higher-order cumulants vanish, but PCA is blind to the non-Gaussian structure in real-world data. In the proposed Principal Cumulant Component Analysis (PCCA), we find orthogonal components that best explain the variation in *all* the cumulants simultaneously. Eigenvalues are replaced by a small *core tensor* C that captures irreducible non-Gaussian dependence among the components.

PCCA is a principled way to incorporate higher-order statistics into PCA via (1) the use of multivariate cumulants, which are precisely the statistical objects generalizing covariance matrices to higher-order information; and (2) the analysis of cumulants via a multilinear model suggested by how cumulants transform.

While there are superficial similarities, PCCA differs statistically, mathematically, and algorithmically from all existing methods that attempt to take higher-order information into account. PCCA applies broadly to any scenario where it is reasonable to assume a linear generative model $\mathbf{y} = A\mathbf{x} + \mathbf{n}$ and so differs from ICA [3, 6], which requires a strong statistical independence assumption for $\mathbf{x}$. It differs from tensor decomposition methods such as 'Tucker model'-based methods [22, 24] and 'CANDECOMP/PARAFAC'-based methods [5, 11] or their nonnegative-constrained variants NMF/NTF [18]. These 'tensor' methods evolved from heuristic-based psychometrics models that lack strong mathematical or statistical justifications. PCCA is a method that truly uses *tensors* as opposed to just manipulating an array of numbers: Cumulants are honest tensors as they transform as a tensor should under a change of basis, and there is only one natural action, defined below in (1). Furthermore our algorithmic approach improves on the difficulties of d-mode SVD or HOSVD and alternating least squares [7]. These tensor decompositions lack rigorous justification in that they do not optimize well-defined objective functions (and so cannot have likelihood interpretations), converge to stationary points, nor represent projections to a well-defined space of models. More details on these issues can be found in Sections 2.3 and 2.5, and the journal version of this report.

A common argument against cumulants is their sensitivity to outliers. But this is precisely a problem with covariance-based methods: in a setting of risky payoffs they are too insensitive to outliers; a 50% drop in the stock market is an outlier, more likely with negative skew and positive kurtosis, but not one we should ignore. Moreover the issue is one of degree not kind: covariance K_2 is also more outlier sensitive than mean K_1, in the same way. PCCA uses outlier information with low-rank regularization. For simplicity, we have used unbiased estimates for cumulants in this

paper; but we note that other estimators/methods robust against outliers exist (e.g. peak shaving and shrinkage) and can provide more robust *inputs* to PCCA.

Our use of cumulants in PCCA differs substantially from the manners they are typically used in statistics. In statistics, the use of cumulants often rely on analytic solutions [15] and symbolic computation [1], which scale poorly with large data sets. On the other hand, our approach exploits robust numerical algorithms and takes advantage of the abundance of floating-point computation. This allows us to analyze $300\times300\times300$ skewness cumulants or $50\times50\times50\times50$ kurtosis cumulants with relative ease on a laptop computer.

In Section 2.1, we give the necessary definitions of tensor and cumulants; in Section 2.2 the properties that make them suitable for modeling dependence, and in Section 2.3 a geometric view on the difficulties of extending the notion of eigenvalue decomposition to symmetric tensors. In Section 2.4 we describe the multilinear factor model and loss that define PCCA. In Section 2.5 we show how to estimate the model using a manifold BFGS algorithm, and in Section 3 explore applications to dimension reduction and multi-moment portfolio optimization.

2. The Cumulant Factor Model

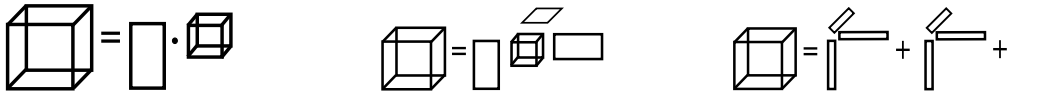
Suppose we are given a p-vector of random variables $\mathbf{y}$ that are not multivariate Gaussian or derived from a linear mixture of independent sources, but follow a law $\mathbf{y} = A\mathbf{x}+\mathbf{n}$ for some random r-vector $\mathbf{x}$, $p\times r$ matrix A, $r \ll p$, and independent Gaussian noise $\mathbf{n}$. For example, $\mathbf{y}$ could be the pixels in an image or the returns on a collection of assets. We show how to estimate these factors, describing their non-Gaussian dependence structure in terms of cumulants and how it propagates to the observed variables.

2.1. Tensors and cumulants.

A tensor in coordinates is a multi-way array with a multilinear action. A tensor $C = (c_{ijk}) \in \mathbb{R}^{p\times p\times p}$ is *symmetric* if it is invariant under all permutations of indices, $c_{ijk} = c_{ikj} = c_{jik} = c_{jki} = c_{kij} = c_{kji}$. We denote the set of dth order symmetric tensors with r dimensions in each mode by $\mathsf{S}^d(\mathbb{R}^r)$. The multilinear action is a symmetric multilinear matrix multiplication as follows. If Q is an $p\times r$ matrix, and C an $r\times r\times r$ tensor, define the $p\times p\times p$ tensor $K = (Q,Q,Q)\cdot C$ or just $K = Q\cdot C$ as

$$\kappa_{\ell mn} = \sum\nolimits_{i,j,k=1,1,1}^{r,r,r} q_{\ell i}q_{mj}q_{nk}c_{ijk}, \tag{1}$$

and similarly for d-way tensors; see Figure 2(a). If $d = 2$, so C is a $r\times r$ and Q is a $p\times r$ matrix, we have $Q\cdot C = QCQ^\top$. For $d > 2$, we multiply on $3, 4, \ldots$ "sides" of the multi-way array. Note that the operation is associative in the the sense of $Q_1\cdot(Q_2\cdot C) = (Q_1Q_2)\cdot C$.



(**a**) $K = Q\cdot C$ (**b**) $K = (X,Y,Z)\cdot C$ (**c**) $T = \mathbf{u}_1\otimes\mathbf{v}_1\otimes\mathbf{w}_1+\cdots$

FIGURE 2. **(a)** Symmetric multilinear action. Multilinear **(b)** and rank **(c)** tensor decompositions.

Multivariate moments and cumulants are symmetric tensors. For a vector-valued random variable $\mathbf{x} = (X_1,\ldots,X_n)$, three natural d-way tensors are:

- The dth non-central moment $s_{i_1,\ldots,i_d}$ of $\mathbf{x}$: $S_d(\mathbf{x}) = \big[\mathbb{E}(X_{i_1}X_{i_2}\cdots X_{id})\big]_{i_1,\ldots,i_d=1}^{p}$.
- The dth central moment $M_d(\mathbf{x}) = S_d(\mathbf{x}-\mathbb{E}[\mathbf{x}])$, and
- The dth cumulant $\kappa_{i_1\ldots i_d}$ of $\mathbf{x}$:

$$K_d(\mathbf{x}) = \Big[\sum\nolimits_P (-1)^{q-1}(q-1)!\,s_{P_1}\ldots s_{P_q}\Big]_{i_1,\ldots,i_d=1}^{p},$$

where the sum is over all partitions $P = P_1 \sqcup \cdots \sqcup P_q = \{i_1, \dots, i_d\}$ of the index set. The cumulant definition is just Möbius inversion, as $s_{i_1,\dots,i_d} = \sum_P \kappa_{P_1} \dots \kappa_{P_q}$ summing over partitions P as above. The dth cumulant is the central moment "surprise," the difference between the dth central moment and what we would expect it to be given the lower order moments; for example, $\kappa_{ijk\ell} = m_{ijk\ell} - (m_{ij}m_{k\ell} + m_{ik}m_{j\ell} + m_{i\ell}m_{jk})$. For Gaussians, the surprise is zero. For any distribution, cumulants and central moments are equal for $d \leq 3$. We use a simple bias-corrected version of this definition, known as k-statistics [10], to estimate cumulants from data; for example, $k_{ijk} = \frac{N}{(N-1)(N-2)} \sum_{t=1}^{N} Y_{ti}Y_{tj}Y_{tk}$ for an $N \times p$ data matrix Y. Future work will explore using various recent improvements to these estimators, such as peak-shaving and shrinkage (e.g. [25]).

For univariate x, the cumulants $K_d(x)$ for $d = 1, 2, 3, 4$ yield expectation $\kappa_i = \mathbb{E}[x]$ and variance $\kappa_{ii} = \sigma^2$, while skewness $= \kappa_{iii}/\kappa_{ii}^{3/2}$, and kurtosis $= \kappa_{iiii}/\kappa_{ii}^2$. Of course here the index $i \in \{1\}$, this being the univariate case. The tensor versions are the multivariate generalizations κ_{ijk}. They provide a natural measure of non-Gaussianity; for example, showing they vanish in the limit is one way to prove central limit theorems. We can also express them in terms of the log characteristic function, $\kappa_{\alpha_1\cdots\alpha_d}(\mathbf{x}) = (-i)^d \frac{\partial^d}{\partial t_{\alpha_1}\cdots\partial t_{\alpha_d}} \log \mathbb{E}[\exp(i\langle \mathbf{t}, \mathbf{x}\rangle)]\Big|_{\mathbf{t}=\mathbf{0}}$, or the Edgeworth series, $\log \mathbb{E}[\exp(i\langle \mathbf{t}, \mathbf{x}\rangle)] = \sum_{\alpha=0}^{\infty} i^{|\boldsymbol{\alpha}|}\kappa_{\boldsymbol{\alpha}}(\mathbf{x})\frac{\mathbf{t}^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!}$ where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$ is a multi-index, $\mathbf{t}^{\boldsymbol{\alpha}} = t_1^{\alpha_1} \cdots t_d^{\alpha_d}$, and $\boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_d!$. Note that for $z = \mathbf{h}^\top \mathbf{x}$ a linear combination with coefficients $\mathbf{h}$, $\mathbf{h}^\top \cdot K_3(\mathbf{x})/(\mathbf{h}^\top \cdot K_2(\mathbf{x}))^{3/2}$ is the skewness of z and $\mathbf{h}^\top \cdot K_4(\mathbf{x})/(\mathbf{h}^\top \cdot K_2(\mathbf{x}))^2$ is the kurtosis.

2.2. **Properties of cumulant tensors.** Cumulants have several important properties that make them useful and justify their slight additional complexity relative to moments [10, 15]. The first (also true of moments) is *multilinearity*. If $\mathbf{x}$ is a $\mathbb{R}^n$-valued random variable and $A \in \mathbb{R}^{m\times n}$

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

where $\cdot$ is the multilinear action (1). This generalizes the familiar covariance case ($d = 2$), where $K_2(A\mathbf{x}) = AK_2(\mathbf{x})A^\top$. This multilinearity is arguably the most important principle underlying factor models like PCA — finding a linear transformation of $\mathbf{x}$ that yields information about the data via matrix reduction of the covariance matrix $K_2(A\mathbf{x})$. This same principle, when applied to higher-order cumulants, motivates our PCCA model. Note that this multilinearity also means that the cumulant is a tensor, i.e. that it transforms correctly under change of basis (or a linear map). Independent Component Analysis (ICA) [6] finds an A to approximately diagonalize $K_d(\mathbf{x})$, thus recovering the mixing matrix.

The second is *independence*. If $\mathbf{x}_1, \dots, \mathbf{x}_p$ are mutually independent of variables $\mathbf{y}_1, \dots, \mathbf{y}_p$, we have $K_d(\mathbf{x}_1 + \mathbf{y}_1, \dots, \mathbf{x}_p + \mathbf{y}_p) = K_d(\mathbf{x}_1, \dots, \mathbf{x}_p) + K_d(\mathbf{y}_1, \dots, \mathbf{y}_p)$. Moreover $K_{i_1\cdots i_d}(\mathbf{x}) = 0$ whenever there is a partition of $\{i_1, \dots, i_d\}$ into two nonempty sets I and J such that $\mathbf{x}_I$ and $\mathbf{x}_J$ are independent. Thus diagonal third and fourth order cumulant tensors are a necessary condition for independent variables, which is why ICA imposes it on the factor cumulant. This property can also be exploited to derive other sparse cumulant techniques. Sparsity constraints (such as requiring diagonal cumulants) break rotational symmetry, and can thus be used for blind source separation.

The third property is *vanishing and extending*. If $\mathbf{x}$ is multivariate normal, then $K_d(\mathbf{x}) = 0$ for all $d \geq 3$. multivariate Gaussianity is a widespread assumption, this is one reason cumulant tensors are not more popular. The Gaussian is special: unfortunately [14], there are no distributions with a bound n so that $K_d(\mathbf{x}) \neq 0$ when $3 \leq d \leq n$, but $K_d(\mathbf{x}) = 0$ for $d > n$. This means that parameterization is trickier when working with nonzero cumulants above K_2; fixing only the first four cumulant tensors does not completely specify a distribution, and we must make additional assumptions such as a loss.

2.3. **Difficulties in extending the eigenvalue decomposition to tensors.** Cumulant tensors are a useful common generalization of skewness, kurtosis, and the covariance matrix, but they can be very big. The dth cumulant tensor of a p-valued random vector has $\binom{p+d-1}{d}$ quantities. This number quickly becomes too large to learn without a great deal of data, to optimize with, and even to store in memory. Thus we require small, implicit models analogous to PCA. PCA is just the eigenvalue decomposition of a positive semidefinite real symmetric matrix. Hence we need a tensor analog to this decomposition. However, there is some subtlety in extending the notion of eigenvalue decomposition to tensors.

In fact, three possible generalizations are the same in the matrix case but not in the tensor case. Let T be an $p \times p \times p$ tensor. The *tensor rank* is the minimum r such that T can be written as a sum of r rank one tensors, $T = \sum_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i$; see Figure 2(c). The set of such tensors, for $d > 2$, is not closed. Nevertheless approximation schemes have been devised, e.g. [5, 11], for an ill-posed objective function. The *border rank* is the minimum r such that T can be written as a *limit* of a sum of r rank one tensors $T = \lim_{\epsilon \to 0}(T_\epsilon)$, tensor rank$(T_\epsilon) = r$. Geometrically, these spaces correspond to secant varieties of Segre (in the non-symmetric case) or Veronese (in the symmetric case) varieties. They are closed sets but are hard to represent. As the tensor rank may jump arbitrarily for fixed border rank [8], we cannot know in advance how many parameters we need to store such an object. Its defining equations are unknown in general, and are a subject of active study in algebraic geometry; thus they are likely too complex to make an inviting model. The final concept is *multilinear rank*: the least r such that we can write $K = (X, Y, Z) \cdot C, C \in \mathbb{R}^{r \times r \times r}, X, Y, Z \in \mathbb{R}^{n \times r}$ (Figure 2(b)). The resulting *subspace variety* is also closed, but its equations are almost trivial and it is well understood, making it a good candidate for a model.

2.4. **Principal Cumulant Component Analysis.** Thus we can define a *multilinear rank factor model* as follows. Let $\mathbf{y} = (Y_1, \ldots, Y_n)$ be a random vector. Write the dth 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 $\mathbf{x}$:

$$K_d(\mathbf{y}) \approx Q \cdot K_d(\mathbf{x}). \tag{2}$$

where Q is orthonormal, and $Q^\top$ projects an observation $\mathbf{y}$ to its factor loadings $\mathbf{x}$. The column space of Q defines the r-dim subspace which best explains the dth order dependence (best in the sense of the approximation $\approx$, defined below). In place of eigenvalues, we have the core tensor $K_d(\mathbf{x})$, the *cumulant of the factors*. Critically, $K_d(\mathbf{x})$ is not necessarily diagonal, but captures irreducible higher-order non-Gaussian dependence among the factors $\mathbf{x}$.

The motivation for (2) may be explained as follows. For a linear generative model $\mathbf{y} = A\mathbf{x} + \mathbf{n}$ with independent noise where $A \in \mathbb{R}^{p \times r}$, the properties of cumulants discussed earlier imply that $K_d(\mathbf{y}) \in \mathsf{S}^d(\mathbb{R}^p)$, the cumulant of the observed factors, satisfies $K_d(\mathbf{y}) = A \cdot K_d(\mathbf{x}) + K_d(\mathbf{n})$. We will assume that the cumulant of the noise $K_d(\mathbf{n})$ is small (if $\mathbf{n}$ is Gaussian, the cumulant estimates $\hat{K}_d(\mathbf{n}) \to 0$ for $d > 2$ as sample size increases) and seek an approximation $K_d(\mathbf{y}) \approx A \cdot K_d(\mathbf{x})$. Since we may always perform QR-factorization to get $A = QR$ with $Q \in \mathrm{O}(p, r)$ and $R \in \mathbb{R}^{r \times r}$, we obtain $K_d(\mathbf{y}) \approx Q \cdot R \cdot K_d(\mathbf{x})$. Let $C = R \cdot K_d(\mathbf{x}) \in \mathsf{S}^d(\mathbb{R}^r)$. We then minimize an appropriate loss over all Q and C. A crucial point to note is that we are not attempting to recover A but the subspace spanned by the columns of A — as represented by an orthonormal basis Q for the column space of A. The columns of Q are the *principal cumulant components* that we seek. Indeed any orthonormal basis for the subspace would work and since the equivalence class of all orthonormal basis for A is precisely a point on a Grassmannian, the ultimate optimization problem that we solve is over such a space.

Given this model, it remains to define a loss (what we mean by $\approx$ in (2)) and provide an algorithm to compute this approximation. We choose least squares loss, $K_d(\mathbf{y}) \approx Q \cdot K_d(\mathbf{x})$ if $\|K_d(\mathbf{y}) - Q \cdot K_d(\mathbf{x})\|_2$ is small, though a multilinear operator norm would also make sense. We

denote the estimated cumulant by $\hat{K}_d$. There are at least two ways to combine the information appearing at different orders d.

First (PCCA1), we could ask for factors or principal components that account for variation in each cumulants *separately*, for $d = 2, 3, 4, \dots$:

$$\min_{Q_d \in \mathrm{O}(p,r),\, C_d \in \mathsf{S}^d(\mathbb{R}^r)} \|\hat{K}_d(\mathbf{y}) - Q_d \cdot C_d\|^2, \tag{3}$$

where Q_d is an orthonormal $n \times r$ matrix, and $C_d = R \cdot \hat{K}_d(\mathbf{x}) \approx R \cdot K_d(\mathbf{x})$ is the factor cumulant in each degree d. The projectors Q_d may then be used to extract separate sets of features or combined later. We pursue this approach in Section 3.2.

Second (PCCA2), we could ask for factors or principal components that account for variation in all cumulants *simultaneously* with a single mixing matrix Q

$$\min_{Q \in \mathrm{O}(p,r),\, C_d \in \mathsf{S}^d(\mathbb{R}^r)} \sum\nolimits_{d=2}^{\infty} \alpha_d \|\hat{K}_d(\mathbf{y}) - Q \cdot C_d\|^2. \tag{4}$$

Here the α_d weight the relative importance of the cumulants by degree. Again, $C_d \approx \hat{K}_d(\mathbf{x})$ is not necessarily diagonal.

Problem (4), of which (3) is a special case, appears intractable, an optimization over an infinite-dimensional manifold $\mathrm{O}(p,r) \times \prod_{d=2}^{\infty} \mathsf{S}^d(\mathbb{R}^r)$. However, as shown in Theorem 2.1, it reduces to optimization over a single Grassmannian $\mathrm{Gr}(p,r)$.

$$\max_{[Q] \in \mathrm{Gr}(p,r)} \sum\nolimits_{d=2}^{\infty} \alpha_d \|Q^\top \cdot \hat{K}_d(\mathbf{y})\|^2. \tag{5}$$

As a set, the Grassmannian is the set of r-dimensional subspaces of a p-dimensional space and has dimension $r(p-r)$. Note that two matrices represent the same point in $\mathrm{Gr}(p,r)$ if they have the same column space. PCA chooses the subspace (point on the Grassmannian) which best explains covariance; PCCA chooses the point which best explains covariance and the higher order cumulants. Note that one consequence of including higher order cumulants in our subspace selection criteria is that the covariance matrix obtained will not generally be diagonal. In practice we use declining weights $\alpha_d = O(\frac{1}{d!})$, which connect our method to the information-theoretic perspective [3, 6] on ICA. Let $\boldsymbol{\varphi}$ be the pdf of a multivariate Gaussian, and $p_{\mathbf{n}}$ of the noise term $\mathbf{n} := \mathbf{y} - Q\mathbf{x}$, then given that the noise is not too far from normal, the Kullback-Leibler divergence is approximately [12]

$$D_{\mathrm{KL}}(p_{\mathbf{n}}, \boldsymbol{\varphi}) = \int p_{\mathbf{n}}(\boldsymbol{\xi}) \log \frac{p_{\mathbf{n}}(\boldsymbol{\xi})}{p_{\boldsymbol{\varphi}}(\boldsymbol{\xi})}\, d\boldsymbol{\xi} \approx \frac{1}{3!}\|K_3(\mathbf{n})\|^2 + \frac{1}{4!}\|K_4(\mathbf{n})\|^2,$$

so we are looking for a mixing matrix Q to make the error as Gaussian as possible; finding it means we have successfully captured the higher-order dependence structure with our factor model. We truncate the summation to $D = 3$ or 4 as cumulants of higher order become increasingly inaccurate to estimate in practice. Of course, if accuracy of data can be guaranteed, D may be increased as appropriate.

Theorem 2.1. *Let $K_d \in \mathsf{S}^d(\mathbb{R}^p)$, $d = 2, \dots, D$. The point $[Q] \in \mathrm{Gr}(p,r)$ is a solution to the problem*

$$\max_{[Q] \in \mathrm{Gr}(p,r)} \sum\nolimits_{d=2}^{D} \alpha_d \|Q^\top \cdot K_d\|^2 \tag{6}$$

if and only if for all $Q \in [Q]$, $Q, (C_d := Q^\top \cdot K_d)_{d=2}^D$ is a solution to the problem

$$\min_{Q \in \mathrm{O}(p,r), C_d \in \mathsf{S}^d(\mathbb{R}^r)} \sum\nolimits_{d=2}^{D} \alpha_d \|K_d - Q \cdot C_d\|^2 \tag{7}$$

Proof. The objective (7) can be written $\sum_{d=2}^{D} \alpha_d (\|K_d\|^2 - 2\langle K_d, Q \cdot C_d\rangle + \|Q \cdot C_d\|^2)$. The first term can be ignored, and the second and third terms sum to $\sum_d \alpha_d \|Q^\top \cdot K_d\|^2$ as can be seen by the following. The normal equation in each degree implies each $C_d = Q^\top \cdot K_d$ at any solution $Q^*, (C_d^*)_{d=2}^D$, and $Q^\top$ is Q adjoint, so $\langle K_d, Q \cdot C_d\rangle = \langle K_d, QQ^\top \cdot K_d\rangle = \langle Q^\top K_d, Q^\top \cdot K_d\rangle = \|Q^\top \cdot K_d\|$

for all d. Since Q is an L^2 isometry $\mathbb{R}^r \to \mathbb{R}^p$, $\|Q \cdot C_d\|^2 = \|QQ^\top \cdot K_d\|^2 = \|Q^\top \cdot K_d\|^2$. Thus (7) is minimized if and only if $\sum_d \alpha_d \|Q^\top \cdot K_d\|^2$ is maximized. Moreover all $Q \in [Q]$ give the same value of the objective in each. □

2.5. **Grassmannian BFGS algorithm for tensor approximation.** In the case of $\mathrm{Gr}(p, r)$, [2] showed that many operations in Riemannian geometry can be reduced to numerical linear algebra operations, enabling the translation of gradient-based optimization algorithms to objectives defined on the Grassmann manifold. In this setting, all data, such as Hessian approximations, must be parallel transported between steps. This framework led to tensor approximation algorithms [9, 17]; we use the quasi-Newton BFGS algorithm of [17] to optimize (3) and (4). Such gradient-based optimization methods are theoretically and practically superior to the commonly used alternating least squares and HOSVD tensor approximation algorithms, which are not guaranteed to converge to a critical point of the objective and do not respect the problem's symmetry constraints (see Figure 3). Furthermore, the BFGS approximate Hessian on a Grassmannian is known to share the same optimality property as its Euclidean counterpart [17, Theorem 6.6]. Due to space limitations, implementation details will appear in the journal version.

3. Applications and Experiments

3.1. **Muti-moment portfolio optimization.** Markowitz mean-variance optimal portfolio theory defines risk to be variance. The investor selects a portfolio which minimizes variance subject to achieving a minimum return $\underline{r}$. For random vector $\mathbf{y}$ of returns on p assets, the optimal portfolio holdings vector $\mathbf{h} \in \mathbb{R}^p$ is the point on the solution locus of

$$\min \mathbf{h}^\top K_2(\mathbf{y})\mathbf{h} \qquad s.t. \qquad \mathbf{h}^\top \mathbb{E}[\mathbf{y}] > \underline{r} \tag{8}$$

tangent to the line crossing the return axis at the risk-free rate. Evidence [13] indicates that investors optimizing variance with respect to the covariance matrix accept unwanted negative skewness and excess kurtosis risk; this can be easily seen in hedge fund index returns (Figure 3). An extreme is that selling out-of-the-money puts looks safe and uncorrelated to the market; many hedge funds pursue strategies which are essentially equivalent.

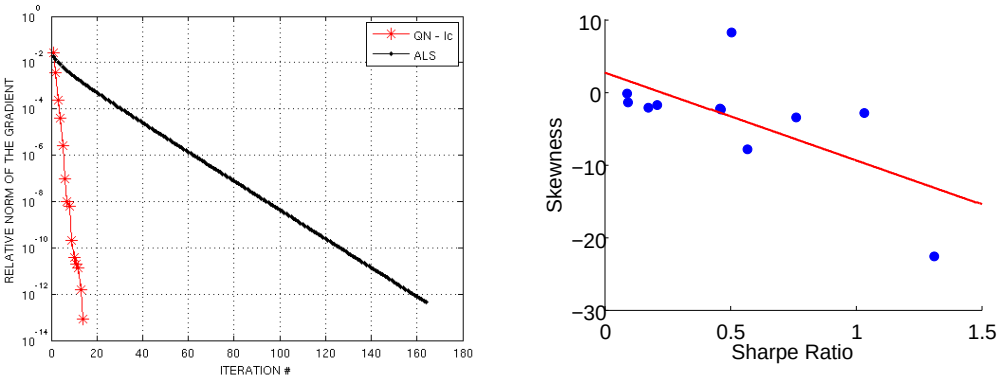


Figure 3. Left: Grassmannian BFGS compares favorably with Alternating Least Squares in estimating a PCCA multilinear factor model for the pixel skewness tensors in an image dimension reduction task. Right: Sharpe ratio (a mean-variance performance measure, excess return over standard deviation $\frac{\mu - \mu_f}{\sigma}$) vs. skewness in Hedge Fund Research Indices daily returns.

One way to avoid this is to take skewness and kurtosis into account in the objective function. Just as we need an estimate $\hat{K}_2(\mathbf{x})$ to optimize (8), we will need estimated skewness $\hat{K}_3$ and

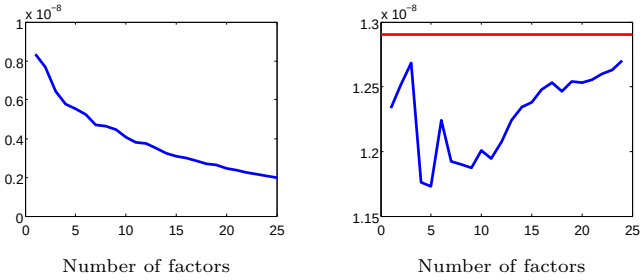


FIGURE 4. Error × number of skewness factors for in-sample estimated skewness (left) and generalization error; red line is error with full $\hat{K}_3(\mathbf{y})$ cumulant. Error estimates based on 10^4 randomly-generated 50-stock portfolios trained on 365 consecutive trading days, tested on the following 389.

kurtosis $\hat{K}_4$ tensors to optimize with respect to skewness and kurtosis. The low multilinear rank PCCA model helps in two ways. First, it regularizes, reducing the variance of our estimates of the entries of $K_d(\mathbf{x})$ at the cost of bias. See Figure 4 for an example. The benefits of this tradeoff in the covariance case are well established [19]. Secondly, it makes optimization feasible with many assets, as we may optimize first with respect to the factors (asset allocation) or easily compute the derivative of skewness, kurtosis, etc. with respect to each security weight. Consider the problem faced by an investor whose preferences are a linear combination of mean, variance, skewness, and kurtosis:

$$\min \sum_{d=2}^{4} \alpha_d \mathbf{h}^\top \cdot K_d(\mathbf{y}) \qquad s.t. \qquad \mathbf{h}^\top \mathbb{E}[\mathbf{y}] > \underline{r}. \tag{9}$$

With a multilinear rank-r mixing matrix model $\mathbf{y} = Q\mathbf{x}$ from PCCA, we can approximate the univariate portfolio cumulant $\phi_d(\mathbf{h}) := \mathbf{h}^\top \cdot K_d(\mathbf{y}) \approx \mathbf{h}^\top Q \cdot \hat{K}_d(\mathbf{x})$, so that with $\mathbf{w}^\top := \mathbf{h}^\top Q$ the factor loadings, $\frac{\partial \phi_d}{\partial h_i} = \sum_{\alpha=(1,\dots,1)}^{(r,\dots,r)} \kappa_\alpha \sum_{j=1}^{d} q_{\alpha_j i} \prod_{k\neq j} w_{\alpha_k}$. And in particular, for example, $\partial\text{skewness}/\partial h_i = \phi_2^{-2}(\phi_2 \partial_i \phi_3 - \phi_3 \partial_i \phi_2)$ gives us the estimated impact on portfolio skewness of an increase in weight h_i.

3.2. **Dimension Reduction: Skewfaces.** We consider a skew analog to eigenfaces [23], using the first of the PCCA techniques (PCCA1) to find an optimal subspace for each degree separately, discarding the core tensors, and combining features at the end. Thus we obtain "skewmax" features supplementing the PCA varimax subspace. In eigenfaces we begin with a centered #pixels= $p \times N$ =#images matrix M, with $N \ll p$. We used the ORL Database of Faces [16]. The eigenvectors of the covariance matrix $\hat{K}_2^{\text{Pixels}}$ of the *pixels* are the eigenfaces. For efficiency, we compute the covariance matrix $\hat{K}_2^{\text{Images}}$ of the *images* instead. The SVD gives both implicitly; if $USV^\top$ is the SVD of $M^\top$, with $U, S \in \mathbb{R}^{N\times N}$ and $V^\top \in \mathbb{R}^{N\times p}$, then $\hat{K}_2^{\text{Pixels}} = \frac{1}{N} MM^\top = \frac{1}{N} V\Lambda V^\top$. The orthonormal columns of V, eigenvectors of $\hat{K}_2^{\text{Pixels}}$, are the eigenfaces.

For the skew tensor version, let $\hat{K}_3^{\text{Pixels}}$ be the $10,304 \times 10,304 \times 10,304$ third cumulant tensor of the pixels. Analogously, we want to compute it implicitly, and we are only interested in the projector Π onto the subspace of skewfaces that best explains $\hat{K}_3^{\text{Pixels}}$. Let $M^\top = USV^\top$ be the SVD. Then with $\hat{K}_3$ denoting the operator computing the k-statistic for the third cumulant tensor, multilinearity implies

$$\hat{K}_3^{\text{Pixels}} = \hat{K}_3(VSU^\top) = V \cdot \hat{K}_3(SU^\top)$$

Pick a small multilinear rank r. If $\hat{K}_3(SU^\top) \approx Q \cdot C_3$ for some $N \times r$ matrix Q and non-diagonal $r \times r \times r$ core tensor C_3,

$$\hat{K}_3^{\text{Pixels}} \approx V \cdot Q \cdot C_3 = VQ \cdot C_3$$

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

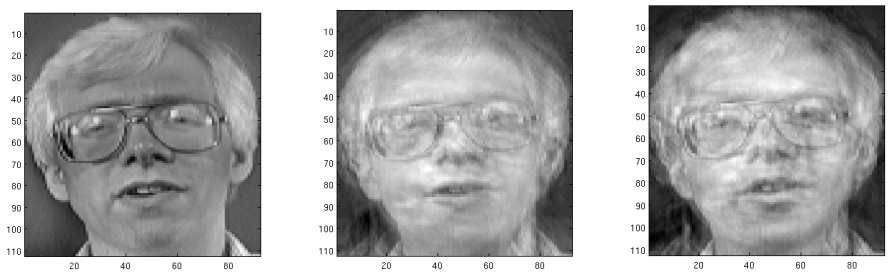


FIGURE 5. Original vs. reconstruction with 30 eigenvectors or 20 eigenvectors with 10 skewvectors.

We have combined the $d = 2$ and 3 tensors by orthogonalizing the skew factors with respect to the eigenfactors using a QR decomposition. Thus the first 20 vectors best explain the covariance matrix $\hat{K}_2^{\text{Pixels}}$, and the next 10 vectors, *together with* the first 20, best explain the big skewness tensor $\hat{K}_3^{\text{Pixels}}$ of the pixels. Typical reconstruction results are shown in Figure 5.

REFERENCES

[1] Andrews, D.F. and Stafford, J.E. (2000). *Symbolic computation for statistical inference.* Oxford University Press, New York, NY.

[2] Edelman, A., Arias, T., and Smith, S. (1999) The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, **20**(2), 303–353.

[3] Bell, A.J. and Sejnowski, T.J. (1996) Edges are the 'independent components'of natural scenes. In *Advances in Neural Information Processing Systems* 9, pp.831–837. Cambridge, MA: MIT Press.

[4] Bader, B.W. and Kolda, T.G. (2007) MATLAB Tensor Toolbox Version 2.2, `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox`.

[5] Carroll, J.D. and Chang, J.J. (1970) Analysis of individual differences in multidimensional scaling via n-way generalization of Eckart-Young decomposition. *Psychometrika*, **35**(3), 283–319.

[6] Comon, P. (1994). Independent component analysis: a new concept? *Signal Process.*, **36**(3), 287–314.

[7] De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, **21**(4), 1253–1278.

[8] De Silva, V. and Lim, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem, *SIAM J. Matrix Anal. Appl.*, **30**(3), 1084–1127.

[9] Eldén, L. and Savas, B. (2007). A Newton-Grassmann method for computing the best multilinear rank-(r_1, r_2, r_3) approximation of a tensor. *SIAM J. Matrix Anal. Appl.*, to appear.

[10] Fisher, R.A. (1929). Moments and product moments of sampling distributions. *Proc. London Math. Soc.*, **30**, 199–238.

[11] Harshman, R.A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics*, **16**, 1–84.

[12] Jones, M.C. and Sibson, R. (1987). What is projection pursuit? *J. R. Statist. Soc. A*, **150**(1):1–36.

[13] Kaiser, D.G., Schweizer, D., and Wu, L. (2008). Strategic hedge fund portfolio construction that incorporates higher moments, preprint, `http://ssrn.com/abstract=1080509`.

[14] Marcinkiewicz, J. (1938). Sur une propriete de la loi de Gauss. *Math. Z.*, **44**, 612–618.

[15] McCullagh, P. (1987) *Tensor methods in statistics*, Chapman and Hall.

[16] Samaria, F. and Harter, A. (1994). Parameterisation of a Stochastic Model for Human Face Identification. *IEEE Workshop on Applications of Computer Vision*, Sarasota (Florida). Database of Faces courtesy of AT&T Laboratories Cambridge, `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`.

[17] Savas, B. and Lim, L.-H. (2009), Quasi-Newton methods on Grassmannians and multilinear approximations of tensors and symmetric tensors. *preprint.*

[18] Shashua, A. and Hazan, T. (2005), Non-negative tensor factorization with applications to statistics and computer vision. *Proc. ICML.*

[19] Stefek, D. (2002) The Barra integrated model, Barra Research Insights.

[20] Stewart, G.W. (2000). The decompositional approach to matrix computation. *Comput. Sci. Eng.*, **2**(1), 50–59.

[21] Thiele, T.N. (1889). *The General Theory of Observations*.
[22] Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**(3), 279–311.
[23] Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. *Proc. CVPR*, 586–591.
[24] Vasilescu, M. and Terzopoulos, D. (2002). Multilinear analysis of image ensembles: TensorFaces. *Proc. ECCV*, Copenhagen, Denmark, 447–460.
[25] Welling, M. (2005). Robust higher order statistics. *Tenth International Workshop on Artificial Intelligence and Statistics*, Barbados, 405–412.

Department of Mathematics, Stanford University, Stanford, CA 94305-2125
E-mail address: jason@math.stanford.edu

Department of Mathematics, University of California, Berkeley, CA 94720-3840
E-mail address: lekheng@math.berkeley.edu