SFB 649 Discussion Paper 2014-001

Principal Component Analysis in an Asymmetric Norm

Ngoc Mai Tran*
Maria Osipenko**
Wolfgang Karl Härdle**

* University of California at Berkeley, USA
** Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Principal Component Analysis in an Asymmetric Norm *

Ngoc Mai Tran[1], Maria Osipenko[2], and Wolfgang Karl Härdle[3]

[1]Department of Statistics, University of California at Berkeley, USA.
[2]Collaborative Research Center 649: *Economic Risk*,
Humboldt-Universität zu Berlin, Berlin, Germany.
[3]C.A.S.E.- Center for Applied Statistics & Economics,
Humboldt-Universität zu Berlin, Berlin, Germany.
Lee Kong Chian School of Business, Singapore Management University,
Singapore.

Abstract

Principal component analysis (PCA) is a widely used dimension reduction tool in the analysis of many kind of high-dimensional data. It is used in signal processing, mechanical ingeneering, psychometrics, and other fields under different names. It still bears the same mathematical idea: the decomposition of variation of a high dimensional object into uncorrelated factors or components. However, in many of the above applications, one is interested in capturing the tail variables of the data rather than variation around the mean. Such applications include weather related event curves, expected shortfalls, and speeding analysis among others. These are all high dimensional tail objects which one would like to study in a PCA fashion. The tail character though requires to do the dimension reduction in an asymmetric norm rather than the classical L_2-type orthogonal projection. We develop an analogue of PCA in an asymmetric norm. These norms cover both quantiles and expectiles, another tail event measure. The difficulty is that there is no natural basis, no 'principal components', to the k-dimensional subspace found. We propose two definitions of principal components and provide algorithms based on iterative least squares. We prove upper bounds on their convergence times, and compare their performances in a simulation study. We apply the algorithms to a Chinese weather dataset with a view to weather derivative pricing.

Keywords: principal components; asymmetric norm; dimension reduction; quantile; expectile.

JEL Classification: C38, C61, C63.

*This research was supported by Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

1 Introduction

When data come as curves without known functional form, the statistician faces immediately the need for dimension reduction. The conventional and widely used tool for such high dimensional curve data is principal component analysis (PCA). The basic principle of this technique is to treat the curves as random variations around a mean curve, and then orthogonalize the covariance operator into eigenfunctions and corresponding (random) loadings. The focus of this principle is on studying the variation around a mean curve. Loadings on (interpretable) eigenfunctions would then represent specific variations around the average. PCA or more generally functional PCA (FPCA) has been successfully applied in many fields such as gene expression measurements, weather, natural hazard, and environment studies, demographics, etc, see Jolliffe (2004), Crambes et al. (2009), and Chen and Müller (2012). One of the first applications is in Ramsay and Silverman (2005). They considered temperature curves recorded daily over a year at multiple stations in an area. The premise is that there are only a few principal components influencing the average temperature, and that the temperature curve from each station is well-approximated on average by a specific linear combinations of these factors. PCA approximates the mean of the data by a nested sequence of optimal subspaces of small dimensions. Thus the optimal subspace of dimension k comes with a natural basis, consisting of uncorrelated random curves (vectors), the principal components, playing the role of the factors aforementioned. Due to the nested structure of the optimal subspaces, one can compute the first few components using a greedy algorithm. The first principal component can be computed efficiently using iterative partial least squares.

In many of the above applications, one is not only interested in the variation around an average curve, but rather in features of the data that are expressible as scale (variance) or tail related functional data. In pricing of financial products where volatility is relevant, for example, the variation of the scale of risk factors is at the core of fair pricing. If one would like to construct weather derivatives or forecasts for the above FPCA example on temperature curves, one needs not only to know the variation across stations, but also the changing scale of the temperature curves, Campbell and Diebold (2005), Benth and Benth (2012), and Härdle and López Cabrera (2012). In climatological science, one is interested in the extremes of certain natural phenomena like drought or rainfall. A tail indicator like a quantile of a conditional distribution when indexed by an explanatory variable also constitutes a curve. Therefore, such a quantile curve collection may also be treated in an FPCA context. Yet another tail-describing curve is the expectile function. Like the quantile curve, it can be represented via a solution with respect to an asymmetric norm. Expectiles have as well numerous application areas, especially in the calculation of risk measures of a financial asset or a portfolio. Taylor (2008) shows how a widely accepted risk measure such as expected shortfall can be assessed via expectiles. Kuan et al. (2009) apply this tail measure in an autoregressive risk management context.

In this paper, we develop an analogue of PCA for quantiles and expectiles. The later, proposed by Newey and Powell (1987), is an analogue of the mean for quantiles. The quantile to level τ of a distribution with cdf F, assuming F is invertible, is defined as $q_\tau = F^{-1}(\tau)$. It is also the solution to the following optimization problem:

$$q_\tau = \arg\min_{q\in\mathbb{R}} \mathsf{E}\|X - q\|_{\tau,1}$$

where X is a random variable in $\mathbb{R}$ with distribution F, and $\|x\|_{\tau,\alpha}^{\alpha}$ is the asymmetric norm:

$$\|x\|_{\tau,\alpha}^{\alpha} = |I(x \leq 0) - \tau||x|^{\alpha}, \quad \alpha = 1. \tag{1}$$

Given $X_i \sim F, i = 1, \ldots, n$, one may formulate the estimation of the unknown quantile in a location model:

$$X_i = q_\tau + \varepsilon_i, \tag{2}$$

with the τ-quantile of the cdf of ε being zero. A natural estimate of q_τ in (2) is:

$$\hat{q}_\tau = \arg\min_{q \in \mathbb{R}} \sum_{i=1}^{n} \|X_i - q\|_{\tau,1}. \tag{3}$$

The estimator as written in (3) can be defined for $\mathbb{R}^p$-valued vectors, if the asymmetric norm is taken by applying (1) coordinatewise and then summing over the coordinates. Given this extension it can be used to analyse curves data when discretized on a regular grid as mensioned in Kneip and Utikal (2001).

Formulation (3) yields a statistical interpretation. In fact, if the noise ε_i in (2) follows a so-called asymmetric Laplace distribution $ALD(\tau)$, which has cdf proportional to the functional $\exp(-\| \cdot \|_{\tau,1})$, then (3) can be interpreted as a quasi maximum likelihood estimation of equation (2). Putting $\alpha = 2$ in (1) yields, via (3), a quasi likelihood interpretation based on an asymmetric normal distribution. Both cases $\alpha = 1$ and $\alpha = 2$ for $\tau \neq 0.5$ are indicators for a certain tail index. This paper aims to shed some light on how to create suitable subspace decompositions for such collections of tail index curves.

As noted in Guo et al. (2013), the first step in this problem corresponds to doing low-rank matrix approximation with weighted L_1 and L_2 norm, respectively, where the weights are sign-sensitive (see Section 1). Based on a proposal of Schnabel (2011) an iterative weighted least squares algorithm for expectiles is employed where the weights are updated in each iteration. This algorithm is guaranteed to converge, although not necessarily to the global minimum as we shall show below. Thus one can at least find a locally optimal k-dimensional subspace that best approximates a given quantile or expectile curve (vector). The difficulty is that the weight matrix is not of rank one, hence there is no natural basis, no 'principal components', to the k-dimensional subspace found. While this is a known problem in weighted low-rank matrix approximation, see Srebro and Jaakkola (2003), this problem has not been addressed before.

Furthermore, the definition of an optimal τ-expectile subspace employed in Guo et al. (2013) is not invariant under linear transformations of the data. That is, if one changes the basis of the data, the optimal τ-expectile subspace in the new basis is not necessarily a linear transform of that expressed in the old basis. This means one has to fix a basis for the data before computing the optimal τ-expectile subspace. This restricts the usefulness of this method to applications where there is a natural basis, such as in the Chinese weather dataset, where yearly temperature is expressed as a vector of 365 daily temperatures. Here one would be interested in capturing extreme daily temperature as opposed to extreme temperature expressed in a Fourier basis. However, in many other applications, invariance under change of basis is an important feature of PCA.

The contributions of our paper is two fold. After defining the basic concepts in the next section 2, we, first, work with the formulation in Guo et al. (2013) in section 3 and propose two natural bases, hence two definitions of principal components for the optimal subspace found. Second, in section 4 we propose an alternative definition of principal components for quantiles and expectiles, closely related to the definition of principal directions for quantiles of Fraiman and Pateiro-López (2012). This definition satisfies many nice properties, such as invariance under translations and linear transformations of the data. In particular, it returns the usual PCA basis under elliptically symmetric distributions. We then provide algorithms to compute the three versions of principal components aforementioned, based on iterative weighted least squares in section 5. We prove upper bounds on their convergence times in section 5.2 and compare their performances in a simulation study in section 6. In section 7 of our paper, we show an application to a Chinese weather dataset with a view to pricing weather derivatives. The last section summarizes our findings.

2 Quantiles and expectiles

2.1 Definitions

We now set up notations and recall the definitions of quantile and expectile. In the next two sections we specify the main optimization problems in $\mathbb{R}^p$.

For $y \in \mathbb{R}^p$, define $y_+ \overset{\text{def}}{=} \max(0, y)$, $y_- \overset{\text{def}}{=} \max(0, -y)$ coordinatewise. For $\tau \in (0,1)$, let $\|\cdot\|_1$ denote the L_1-norm in $\mathbb{R}^p$, that is, $\|y\|_1 = \sum_{j=1}^p |y_j|$. Define the asymmetric L_1-norm in $\mathbb{R}^p$ via

$$\|y\|_{\tau,1} = \tau\|y_+\|_1 + (1-\tau)\|y_-\|_1 = \sum_{j=1}^{p} |y_j| \cdot \{\tau I(y_j \geq 0) + (1-\tau) I(y_j < 0)\}.$$

Similarly, let $\|\cdot\|_2$ denote the L_2-norm in $\mathbb{R}^p$, $\|y\|_2^2 = \sum_{j=1}^p y_j^2$. Define the asymmetric L_2-norm in $\mathbb{R}^p$ via

$$\|y\|_{\tau,2}^2 = \tau\|y_+\|_2^2 + (1-\tau)\|y_-\|_2^2.$$

When $\tau = 1/2$, we recover a constant multiple of the L_1 and L_2-norms. These belong to the class of asymmetric norms with sign-sensitive weights, and have appeared in approximation theory, Cobzaş (2013). Some properties we use in this paper are the fact that these norms are convex, and their unit balls restricted to a given orthant in $\mathbb{R}^p$ are weighted simplices for the $\|\cdot\|_{\tau,1}$ norm, and axis-aligned ellipsoids for the $\|\cdot\|_{\tau,2}$ norm. In other words, they coincide with the unit balls of axis-aligned weighted L_1 and L_2 norms.

Let $Y \in \mathbb{R}^p$ be a random variable with cdf F. The τ-quantile $q_\tau(Y) \in \mathbb{R}^p$ of F_Y is the solution to the following optimization problem

$$q_\tau(Y) = \operatorname*{argmin}_{q \in \mathbb{R}^p} \mathsf{E}\|Y - q\|_{\tau,1}.$$

Similarly, the τ-expectile $e_\tau(Y) \in \mathbb{R}^p$ of F_Y is the solution to

$$e_\tau(Y) = \operatorname*{argmin}_{e \in \mathbb{R}^p} \mathsf{E}\|Y - e\|_{\tau,2}^2.$$

Since the asymmetric L_1 and L_2 norms are convex, the solution exists and is unique, assuming that $\mathsf{E}(Y)$ is finite. This definition guarantees that the τ-quantile $q_\tau(Y)$ is unique even when the cdf F is not invertible.

2.2 Properties

We collect some mathematical properties of quantiles and expectiles here. These will be useful for proving theorems in future sections. For convenience we shall suppress the dependence on Y where possible. We shall state the next proposition for the one-dimensional case, that is, $Y \in \mathbb{R}$. Analogous results in higher dimensions hold coordinatewise.

Proposition 2.1 (Properties of expectile Newey and Powell (1987))**.** *Let $Y \in \mathbb{R}$ be a random variable. Let F be its cdf, G be its first partial moment, defined as*

$$G(x) = \int_{-\infty}^{x} u \, dF(u).$$

Assume that $G(x) < \infty$ for all $x \in \mathbb{R}$

- *For $\tau \in (0,1)$, $e_\tau(Y + t) = e_\tau(Y) + t$ for $t \in \mathbb{R}$.*
- *For $\tau \in (0,1)$,*
$$e_\tau(sY) = \begin{cases} se_\tau(Y) & \text{for} \quad s \in \mathbb{R}, s > 0 \\ -se_{1-\tau}(Y) & \text{for} \quad s \in \mathbb{R}, s < 0 \end{cases}$$
- *$e_\tau = e_\tau(Y)$ is the τ-quantile of the distribution function T, i.e. $\tau = T(e_\tau)$ where*
$$T(x) = \frac{G(x) - xF(x)}{2\{G(x) - xF(x)\} + \{x - \int_{-\infty}^{\infty} u \, dF(u)\}}. \tag{4}$$

Corollary 2.1. *Suppose $Y \in \mathbb{R}^p$ has a symmetric distribution about 0, and it belongs to the location-scale family. Then for $\tau \in (0,1)$, $e_\tau(Y) = -e_{1-\tau}(Y)$, and*

$$e_\tau(sY + t) = se_\tau(Y) + t$$

for all $s \in \mathbb{R}, t \in \mathbb{R}^p$. In particular, if $Y \sim \mathrm{N}(\mu, \sigma^2)$, $Z \sim \mathrm{N}(0,1)$, then

$$e_\tau(Y) = |\sigma| e_\tau(Z) + \mu.$$

Suppose the cdf F is differentiable. Then $q_\tau(Y) = F^{-1}(\tau)$. Let $F_n^{-1} : (0,1) \to \mathbb{R}$ and $F^{-1} : (0,1) \to \mathbb{R}$ denote the empirical and population quantile function, respectively. A classical result of empirical process theory in van der Vaart and Wellner (1996, §2) states that

$$\sqrt{n}(F_n^{-1} - F^{-1})(t) \xrightarrow{\mathcal{L}} \frac{-W^0}{F'(F^{-1})}(t), \ t \in (0,1)$$

where W^0 denotes the Brownian bridge on $[0,1]$, and the convergence takes place over the Skorokhod space $D([0,1])$. Now, the last point of Proposition 2.1 states that the expectile is indeed the quantile of a function T. Thus, one may suspect that the expectile process also satisfies a similar statement. Indeed, we now make this concrete.

Theorem 2.1. *Let F be a differentiable cdf which defines a distribution with mean zero, variance σ^2. Let $e : (0,1) \to \mathbb{R}, \tau \mapsto e_\tau$ be the expectile function. Let F_n, e_n be the empirical versions. Then for any $0 < \delta < 1$,*

$$\sqrt{n}(e_n - e) \xrightarrow{\mathcal{L}} \mathcal{E},$$

where the convergence takes place over $D([\delta, 1-\delta])$, $\mathcal{E}$ is a stochastic process on $[\delta, 1-\delta]$, whose marginals are normally distributed with mean 0 *and variance*

$$\mathrm{Var}\{\mathcal{E}(\tau)\} = \frac{\mathsf{E}\{\tau(Y-e_\tau)_+ + (1-\tau)(e_\tau - Y)_+\}^2}{\left[\tau\{1-F(e_\tau)\} + (1-\tau)F(e_\tau)\right]^2} \tag{5}$$

for $\tau \in [\delta, 1-\delta]$.

For example, if $\tau = 1/2$, then $e_{1/2,n}$ and $e_{1/2}$ are just the empirical and population mean, and $\mathrm{Var}\{\mathcal{E}(1/2)\} = \sigma^2$. Thus we recover the classical central limit theorem. We first give an overview of the proof. We shall prove convergence for the inverse process of $e_{\tau,n}$ and e_τ, which is $T_n(e_\tau)$ and $T(e_\tau)$ as defined in Proposition 2.1. Then we invoke the result of Doss and Gill (1992) to show that the expectile process itself must also converge to a stochastic process $\mathcal{E}_\tau$. Finally, to derive the marginal distribution of $\mathcal{E}_\tau$ with e_τ being the solution of a *convex* optimization problem. Thus its asymptotic properties, in particular, its limit in distribution, can be derived involving a theorem of Hjort and Pollard (2011). Applying the result of Hjort and Pollard can only give finite dimensional convergence of the process $\sqrt{n}(e_{\tau,n} - e_\tau)$. On the other hand, it is possible to derive Theorem 2.1 using the result of Doss and Gill alone, however, the computation for the second moment of $\mathcal{E}_\tau$ is quite messy. Thus we choose to only derive the second moment properties of the process $\mathcal{E}_\tau$.

Proof. Note that the inverse process of $e_{\tau,n}$ and e_τ are $T_n : \mathbb{R} \to [0,1], e_\tau \mapsto T_n(e_\tau)$ and $T : \mathbb{R} \to [0,1], e_\tau \mapsto T(e_\tau)$ as defined by (4) in Proposition 2.1. To be clear,

$$T_n(x) = \frac{G_n(x) - xF_n(x)}{2\{G_n(x) - xF_n(x)\} + (x - \mu_n)},$$

where $G_n(x) = \int_{-\infty}^{x} u\, dF_n(u)$ is the empirical version of G, and $\mu_n = \int_{-\infty}^{\infty} u dF_n(u)$ is the empirical mean. By Newey and Powell (1987), the functions T_n, T are both distribution functions, and thus they are non-decreasing cadlag functions. We claim that the stochastic processes $\sqrt{n}(T_n - T)$ converges to some stochastic process in the Skorokhod space $D([-\infty,\infty])$. Indeed, note that the processes $\sqrt{n}(G_n - G)$ and $\sqrt{n}(F_n - F)$ both converge to some process on $D([-\infty,\infty])$. Similarly, assuming $\mu = 0$, $\sqrt{n}\mu_n$ converges to the normal distribution with mean 0, variance σ^2. The numerator of the fraction $\sqrt{n}\{T_n(x) - T(x)\}$ is

$$\sqrt{n}\{G_n(x) - G(x)\}x - \sqrt{n}\{F_n(x) - F(x)\}x^2 - \sqrt{n}\mu_n\{G(x) - xF(x)\}.$$

Since F has finite second moment, $|G(x) - xF(x)|$ is uniformly bounded for large x. Thus the above expression converges in distribution uniformly in x. Now, the denominator of the fraction $T_n(x) - T(x)$ is

$$[2\{G(x) - xF(x)\} + x][2\{G_n(x) - xF_n(x)\} + (x - \mu_n)],$$

which converges a.s. for all x to $[2\{G(x) - xF(x)\} + x]^2$, which is bounded away from 0. Thus the process $\sqrt{n}(T_n - T)$ converges in $D([-\infty,\infty])$.
By Doss and Gill (1992), this implies that the inverse processes $T_n^{-1} = e_n$, $T^{-1} = e$ must satisfy

$$\sqrt{n}(e_n - e) \xrightarrow{\mathcal{L}} \mathcal{E}$$

where the convergence takes place over $D([\delta, 1-\delta])$, $\mathcal{E}$ is a stochastic process on $[\delta, 1-\delta]$. Finally, to derive the marginal distribution of $\mathcal{E}_\tau$ with e_τ being the solution of a *convex* optimization problem with differentiable objective function. Thus the asymptotic properties of the empirical estimator $e_{\tau,n}$, and in particular, its limit in distribution, can be derived using the theorem of Hjort and Pollard (2011, Theorem 2). Explicitly, in their notation, fix $\tau \in [\delta, 1-\delta]$, and let $g_\tau(y,t) = \|y-t\|^2_{\tau,2} = \tau(y-t)^2_+ + (1-\tau)(t-y)^2_+$ be our objective function. Differentiate with respect to t, we find

$$g'_\tau(y,t) = 2\{-\tau(y-t)_+ + (1-\tau)(t-y)_+\}, \quad g''_\tau(y,t) = 2\{\tau I(y \geq t) + (1-\tau)I(y < t)\}.$$

Define $K = \mathsf{E}g'_\tau(Y,t)^2 = 4\mathsf{E}\{\tau(Y-e_\tau)_+ + (1-\tau)(e_\tau - Y)_+\}^2$, and

$$J = \mathsf{E}\{g''_\tau(Y,e_\tau)\} = \mathsf{E}2\{\tau I(y \geq t) + (1-\tau)I(y < t)\} = 2[\tau\{1-F(e_\tau)\} + (1-\tau)F(e_\tau)].$$

Now, since g is a convex differentiable function, as $t \to e_\tau$,

$$\mathsf{E}\{g_\tau(Y,t) - g_\tau(Y,e_\tau)\} = \frac{1}{2}\mathsf{E}\{g''_\tau(Y,e_\tau)\}(e_\tau - t)^2 + \mathcal{O}(|t|^2).$$

Therefore, by Hjort and Pollard (2011, Theorem 2), $\sqrt{n}\{e_{\tau,n} - e_\tau\}$ converges to a normal distribution with mean 0 and variance $J^{-1}KJ^{-1}$, which in our case simplifies to (5). □

3 Principal components as error minimizers

There are multiple, equivalent ways to define standard PCA, which generalize to different definitions of principal components for quantiles and expectiles. We focus on two formulations: minimizing the residual sum of squares, and maximizing the variance capture.

3.1 Review of PCA

Suppose we observe n vectors $Y_1, \ldots, Y_n \in \mathbb{R}^p$ with edf F_n. Write Y for the $n \times p$ data matrix. PCA solves for the k-dimensional affine subspace that best approximates $Y_1, \ldots, Y_n$ in L_2-norm. In matrix terms, we are looking for the constant $m^* \in \mathbb{R}^p$ and the matrix E^*_k, the rank-k matrix that best approximates $Y - \mathbf{1}(m^*)^\top$ in the Frobenius norm. That is,

$$(m^*_k, E^*_k) = \underset{m \in \mathbb{R}^p, E \in \mathbb{R}^{n\times p}: rank(E) = k}{\operatorname{argmin}} \|Y - \mathbf{1}m^\top - E\|^2_{1/2,2}. \tag{6}$$

As written, m is not well-defined: if (m, E) is a solution, then $(m + c, E - \mathbf{1}c^\top)$ is another equivalent solution for any c in the column space of E. Geometrically, this means we can express the affine subspace $m + E$ with respect to any chosen point m. It is intuitive to choose m to be the best constant in this affine subspace that approximates Y. By a least squares argument, the solution is $m^*_k = \mathsf{E}(Y)$. That is, it is independent of k and coincides with the best constant approximation to Y. Thus, it is sufficient to assume $\mathsf{E}(Y) = m \equiv 0$, and consider the optimization problem in (6) without the constant term.

Suppose Y is full rank and the eigenvalues of its covariance matrix are all distinct. Again by least squares argument, for $1 \leq k < p$, the column space of E^*_k is contained in the column space of E^*_{k+1}, and $E^*_{k+1} - E^*_k$ is the optimal rank-one approximation of

$Y - E_k^*$. This has two implications. Firstly, there exists a natural basis for E_k^*. Indeed, there exists a unique ordered sequence of orthonormal vectors $v_1, v_2, \ldots, v_p \in \mathbb{R}^p$ such that $E_1^* = U_1V_1^\top$,$E_2^* = U_2V_2^\top$, and so on, where the columns of V_k are the first k v_i's. The v_i's are called the *principal components*, or *factors*. For fixed k, V_k is the *component*, or *factor matrix*, and U_k is the *loading*.

Secondly, a greedy algorithm reduces computing the components $v_1, v_2, \ldots$ to computing the first component, in other words, solving (6) for $k = 1$. Every rank-one matrix $E \in \mathbb{R}^{n\times p}$ has a unique decomposition $E = UV^\top$ for $U \in \mathbb{R}^{n\times 1}$, $V \in \mathbb{R}^{p\times 1}$ with $V^\top V = 1$. Thus, solving (6) is equivalent to an unconstrained minimization problem over the pair of matrices (U, V) with objective

$$J(U, V) = \|Y - UV^\top\|^2_{1/2,2} = \sum_{i,j}(Y_{ij} - \sum_l U_{il}V_{jl})^2.$$

For fixed U, J is a quadratic in the entries of V, and vice versa. Since all local minima of J are global, see Srebro and Jaakkola (2003), J can be efficiently minimized using an iterative least squares algorithm, leading to an efficient method for performing PCA for small k in large datasets.

3.2 Analogues for expectiles

We now generalize the above definition of PCA to handle expectiles. The quantiles case follows similarly, and algorithms for L_1 matrix factorization can also be adapted to this case. Recall that we are looking for the best k-dimensional affine subspace which minimizes the asymmetric L_2-norm. The analogue of (6) is the following low-rank matrix approximation problem

$$(m_k^*, E_k^*) = \underset{m\in\mathbb{R}^p, E\in\mathbb{R}^{n\times p}: rank(E)=k}{\text{argmin}} \|Y - \mathbf{1}m^\top - E\|^2_{\tau,2}. \tag{7}$$

Again, we may define m to be the best constant approximation to Y on the affine subspace determined by (m, E). For a fixed affine subspace, such a constant is unique, and is the coordinatewise τ-expectile of the residuals $Y - E$. However, the expectile is not additive for $\tau \neq 1/2$. Thus in general, the column space of E_k^* is not a subspace of the column space E_{k+1}^*, the constant m_k^* depends on k, and is not equal to the τ-expectile $e_\tau(Y)$.

Let us fix k and consider the problem of computing m_k^* and E_k^*. Write a rank-k matrix E as $E = UV^\top$, where $U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{p\times k}$. Adjoin U with an all-1 column to form $\tilde{U}$, and adjoin m to the corresponding column of V to form $\tilde{V}$. Thus $\mathbf{1}m^\top + E = \tilde{U}\tilde{V}^\top$. Equation (7) is an unconstrained minimization problem over the pair of (adjoined) matrices $(\tilde{U}, \tilde{V})$ with minimization objective

$$J(\tilde{U}, \tilde{V}, W) = \|Y - \tilde{U}\tilde{V}^\top\|^2_{\tau,2} = \sum_{i,j} w_{ij}(Y_{ij} - m_j - \sum_l U_{il}V_{jl})^2.$$

where the weights w_{ij} are sign-dependent: $w_{ij} = \tau$ if $Y_{ij} - m_j - \sum_l U_{il}V_{lk} > 0$, $w_{ij} = 1 - \tau$ otherwise.

This objective function is not jointly convex in $\tilde{U}$ and $\tilde{V}$. However, for fixed $\tilde{U}$, in each coordinate ij, it is the asymmetric L_2-norm of a linear combination in the entries of $\tilde{V}$, and hence convex. Similarly, J is convex in $\tilde{U}$ for fixed $\tilde{V}$. Therefore, an iterative weighted least squares solution with weight update at each step is guaranteed to converge to a critical point of J (cf. Proposition 5.1). This algorithm (cf Algorithm 1) is called asymmetric weighted least squares (LAWS), see Newey and Powell (1987) and Schnabel (2011). While there are local minima, we find that the algorithm often finds the global minimum quite quickly, supporting similar observations in the literature for fixed weight matrix $[w_{ij}]$, as in Srebro and Jaakkola (2003).

For $k > 1$, the decomposition $E = UV^\top$ is not unique: for any $k \times k$ matrix R, the matrix $(UR, V(R^\top)^{-1})$ is another equivalent factorization. To specify a unique solution we need a choice for V. This is one of the unaddressed issues in Guo et al. (2013), and certainly a key difficulty. While there are algorithms to solve for (m_k^*, E_k^*) for fixed k, there is no natural basis for E_k^* which reveals information on E_j^* for $j < k$. Hence, we do not have a direct analogue for principal components for τ-expectiles.

To furnish a principal components basis for E_k^* based on LAWS, we propose two algorithms: TopDown and BottomUp. These are two definitions, described as algorithms, which output is a nested sequence of subspaces, each approximating E_j^* for $j = 1, \dots, k$. They lead to two different definitions of principal components.

Definition 3.1. Given data $Y \in \mathbb{R}^{n\times p}$ and an integer $k \geq 1$, the first k *TopDown principal components* are the outputs of the TopDown algorithm with input (Y, k). The first k *BottomUp principal components* are the outputs of the BottomUp algorithm with input (Y, k).

In TopDown, one first finds E_k^*. Then for $j = 1, 2, \dots, k-1$, one finds E_j, the best j-dimensional subspace approximation to $Y - m_k^*$, subjected to $E_{j-1} \subset E_j \subset E_k^*$. This defines a nested sequence of subspace $E_1 \subset E_2 \subset \dots \subset E_{k-1} \subset E_k^*$, and hence a basis for E_k^*, such that E_j is an approximation of the best j-dimensional subspace approximation to $Y - m_k^*$ contained in E_k^*. We solve (7) since (m_k^*, E_k^*) is the true minimizer in dimension k, and thus we knew the optimal constant term.

In BottomUp, one first finds E_1^*. Then for $j = 2, \dots, k$, one finds (m_j, E_j), the optimal j-dimensional affine subspace approximation to Y, subjected to $E_{j-1} \subset E_j$. In each step we re-estimate the constant term. Again, we obtain a nested sequence of subspaces $E_1^* \subset E_2 \subset \dots \subset E_k$, and constant terms $m_1, \dots, m_k$, where (m_j, E_j) is an approximation to the best affine j-dimensional subspace approximation to Y.

When $\tau = 1/2$, that is, when doing usual PCA, both algorithms correctly recover the principal components. For $\tau \neq 1/2$, they can produce different output. Interestingly, both in simulations and in practice, their outputs are not significantly different (see Sections 6 and 7). See Section 5 for a formal description of the TopDown and BottomUp algorithms and computational bounds on their convergence times.

3.3 Statistical properties

Even for $\tau = 1/2$, the objective function $J(U, V)$ is not simultaneously convex in both U and V, but it is a convex function when either one of the two arguments is kept fixed. By the same argument, one can show that the same property holds for $J(U, V, W)$. That is, if U is kept fixed, then $J(U, V, W)$ (which is now a function of V only, as W is a function of U and V) is convex in V. Similarly, if V is kept fixed, then $J(U, V, W)$ is a convex function in U. Applying the result of Hjort and Pollard (2011), we see that in each iteration, $V_n^{(t+1)}$ differs from $V^{(t+1)}$ by a term of order $\mathcal{O}(n^{-1/2})$. Thus, if the total number of iterations is small, one can prove consistency of the iterative least squares algorithm. We are not able to obtain a theoretical bound on the total number of iterations. In practice this does indeed seem to be small.

4 Principal components as maximizers of captured variance

4.1 Review of PCA

Again, suppose we observe n vectors $Y_1, \ldots, Y_n \in \mathbb{R}^p$. The first principal component ϕ^* is the unit vector in $\mathbb{R}^p$ which maximizes the variance of the data projected onto the subspace spanned by ϕ^*. That is,

$$\phi^* = \underset{\phi \in \mathbb{R}^p, \phi^\top \phi = 1}{\operatorname{argmax}} \operatorname{Var}(\phi\phi^\top Y_i : 1 \leq i \leq n) = \underset{\phi \in \mathbb{R}^p, \phi^\top \phi = 1}{\operatorname{argmax}} n^{-1} \sum_{i=1}^{n} (\phi^\top Y_i - \overline{\phi^\top Y})^2, \tag{8}$$

where $\overline{\phi^\top Y} = n^{-1} \sum_{i=1}^n \phi^\top Y_i = \phi^\top \bar{Y}$ is the mean of the projected data, or equivalently, the projection of the mean $\bar{Y}$ onto the subspace spanned by ϕ. Given that the first principal component is ϕ_1^*, the second principal component ϕ_2^* is the unit vector in $\mathbb{R}^p$ which maximizes the variance of the residual $Y_i - (\phi_1^*)^\top \bar{Y} - \phi_1^*(\phi_1^*)^\top Y_i$, and so on. In this formulation, the data does not have to be pre-centered. The sum $(\phi_1^*)^\top \bar{Y} + (\phi_2^*)^\top \bar{Y} + \ldots + (\phi_k^*)^\top \bar{Y}$ is the overall mean $\bar{Y}$ projected onto the subspace spanned by the first k principal components. For the benefit of comparison to Theorem 4.1, let us reformulate PCA as an optimization problem. Define

$$C = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^\top. \tag{9}$$

Then ϕ^* is the solution to the following optimization problem.

$$\begin{aligned} &\text{maximize } \phi^\top C \phi \\ &\text{subject to } \phi^\top \phi = 1. \end{aligned}$$

The principal component is not necessarily unique: if the covariance matrix is the identity, for example, then any unit vector ϕ would solve (8), and thus there is no unique principal component. In the discussions that follows, we implicitly assume that the principal component ϕ^* is unique. In other words, C has a unique largest eigenvalue.

4.2 An analogue for expectiles

Let $Y \in \mathbb{R}$ be a random variable with cdf F. We define its τ-*variance* to be

$$\operatorname{Var}_\tau(Y) = \mathsf{E}\|Y - e_\tau\|_{\tau,2}^2 = \min_{e \in \mathbb{R}} \mathsf{E}\|Y - e\|_{\tau,2}^2$$

where $e_\tau = e_\tau(Y)$ is the τ-expectile of Y. When $\tau = 1/2$, this reduces to the usual definition of variance. The following are immediate from Proposition 2.1

Proposition 4.1 (Properties of τ-variance)**.** *Let $Y \in \mathbb{R}$ be a random variable. For $\tau \in (0,1)$, the following statements hold.*

- $\mathrm{Var}_\tau(Y+c) = \mathrm{Var}_\tau(Y)$ *for* $c \in \mathbb{R}$
- $\mathrm{Var}_\tau(sY) = s^2\mathrm{Var}_\tau(Y)$ *for* $s \in \mathbb{R}, s > 0$.
- $\mathrm{Var}_\tau(-Y) = \mathrm{Var}_{1-\tau}(Y)$

Proof. The first two follow directly from corresponding properties for e_τ. We shall prove that last assertion. Recall that $e_\tau(-Y) = -e_{1-\tau}(Y)$. Thus

$$\begin{aligned}\mathrm{Var}_\tau(-Y) &= \mathsf{E}\| - Y - e_\tau(-Y)\|^2_{\tau,2} = \mathsf{E}\| - \{Y - e_{1-\tau}(Y)\}\|^2_{\tau,2} = \mathsf{E}\|Y - e_{1-\tau}(Y)\|^2_{1-\tau,2} \\ &= \mathrm{Var}_{1-\tau}(Y).\end{aligned}$$

□

If $\phi \in \mathbb{R}^p$ is a unit vector, that is, $\phi^\top\phi = 1$, then we define

$$\mathrm{Var}_\tau(\phi\phi^\top Y_i : 1 \le i \le n) = \mathrm{Var}_\tau(\phi^\top Y_i : 1 \le i \le n).$$

That is, the τ-variance of n vectors which are multiples of ϕ is just the τ-variance of the coefficients, which is a sequence of *real numbers*. Thus, the direct generalization of (8) would be

$$\begin{aligned}\phi^*_\tau &= \operatorname*{argmax}_{\phi\in\mathbb{R}^p,\phi^\top\phi=1} \mathrm{Var}_\tau(\phi\phi^\top Y_i : 1 \le i \le n) = \operatorname*{argmax}_{\phi\in\mathbb{R}^p,\phi^\top\phi=1} \mathrm{Var}_\tau(\phi^\top Y_i : 1 \le i \le n) && (10) \\ &= \operatorname*{argmax}_{\phi\in\mathbb{R}^p,\phi^\top\phi=1} n^{-1}\sum_{i=1}^n (\phi^\top Y_i - \mu_\tau)^2 w_i && (11)\end{aligned}$$

where $\mu_\tau \in \mathbb{R}$ is the τ-expectile of the sequence of n real numbers $\phi^\top Y_1, \dots \phi^\top Y_n$, and

$$w_i = \tau \text{ if } \sum_{j=1}^p Y_{ij}\phi_j > \mu_\tau, \text{ and } w_i = 1-\tau \text{ otherwise.} \tag{12}$$

Definition 4.1. Suppose we observe $Y_1, \dots, Y_n \in \mathbb{R}^p$. The first *principal expectile component* (PEC) ϕ^*_τ is the unit vector in $\mathbb{R}^p$ that maximizes the τ-variance of the data projected on the subspace spanned by ϕ^*_τ. That is, ϕ^*_τ solves (11).

'The' principal expectile component is not necessarily unique. In classical PCA, the first principal component is only unique if and only if the covariance matrix has a unique maximal eigenvalue. Even then, under this assumption, the principal component is only unique up to sign. That is, if ϕ is the principal component, then $-\phi$ is also a principal component. Principal expectile component, on the other hand, are sign-sensitive in general, unless if the distribution of Y is symmetric, or if $\tau = 1/2$. We make this observation concrete below, which is a Corollary of Proposition 4.1.

Corollary 4.1. *For $\tau \in (0,1)$, random variable $Y \in \mathbb{R}^p$, suppose ϕ^*_τ is a first τ-PEC of Y. Then*

$$-\phi^*_\tau = \phi^*_{1-\tau},$$

that is, $-\phi^*_\tau$ is also a first $(1-\tau)$-PEC of Y. Furthermore, if the distribution of Y is symmetric about 0, that is, $Y \overset{\mathcal{L}}{=} -Y$, then $-\phi^*_\tau$ is also a first τ-PEC of Y.

Proof. By Proposition 4.1, $\mathrm{Var}_\tau(\phi_\tau^{*\top} Y) = \mathrm{Var}_{1-\tau}\{(-\phi_\tau^{*\top})Y\}$. Thus if ϕ_τ^* solves (10) for τ, then $(-\phi_\tau)^*$ solves (10) for $1-\tau$. If the distribution of Y is symmetric about 0, then

$$\mathrm{Var}_\tau(\phi_\tau^{*\top} Y) = \mathrm{Var}_{1-\tau}\{\phi_\tau^{*\top}(-Y)\} = \mathrm{Var}_\tau(\phi_\tau^{*\top} Y).$$

In this case $-\phi_\tau^* = \phi_{1-\tau}^*$ is another τ-PEC of Y. □

Like in classical PCA, the other components are defined based on the residuals, and thus by definition, they are orthogonal to the previously found components. Therefore one obtains a nested sequence of subspace which captures the tail variations of the data.

By replacing the $\|\cdot\|_{\tau,2}^2$ norm with the $\|\cdot\|_{\tau,1}$ norm, one can define the analogue of principal component for quantiles. The analogue of τ-variance is the τ-deviation

$$\mathrm{Dev}_\tau(Y) = \mathsf{E}\|Y - q_\tau(Y)\|_{\tau,1} = \min_{q\in\mathbb{R}^p} \mathsf{E}\|Y - q\|_{\tau,1}.$$

The τ-deviation is linear rather than quadratic with respect to constants, that is, $\mathrm{Dev}_\tau(cY) = c\mathrm{Dev}_\tau(Y)$ for $c > 0$, we consider vectors in the L_1 unit ball rather than the L_2 unit ball. Define the τ-deviance of n vectors which are multiples of a vector $\psi \in \mathbb{R}^p$ to be the τ-deviance of the coefficients. That is,

$$\mathrm{Dev}_\tau(\psi\psi^\top Y_i : 1 \leq i \leq n) = \mathrm{Dev}_\tau(\psi^\top Y_i : 1 \leq i \leq n)$$

This leads to the optimization problem

$$\psi_\tau^* = \underset{\psi\in\mathbb{R}^p:\sum_j|\psi_j|=1}{\mathrm{argmax}} \mathrm{Dev}_\tau(\psi\psi^\top Y_i : 1 \leq i \leq n).$$

Definition 4.2. The first *principal quantile component* ψ_τ^* is the L_1-unit vector in $\mathbb{R}^p$ that maximizes the τ-deviation captured by the data projected on the subspace spanned by ψ_τ^*.

Generalizing principal components to quantiles via its interpretation as variance maximizer is not new. Fraiman and Pateiro-López (2012) define the first principal quantile direction ψ to be the one that maximizes the L_2 norm of the τ-quantile of the centered data, projected in the direction ψ. That is, ψ is the solution of

$$\max_{\psi\in\mathbb{R}^p:\psi^\top\psi=1} \|\psi^\top q_\tau(Y - \mathsf{E}Y)\|_{1/2,2}.$$

Their definition works for random variables in arbitrary Hilbert spaces. Kong and Mizera (2012) proposed the same definition but without centering Y at $\mathsf{E}Y$. These authors used the principal directions computed to study quantile level sets of distributions in small dimensions. Compared to these work, our definition is very natural, can be extended to Hilbert spaces, and in the case of expectile, satisfies many 'nice' properties, some of which are shared by the principal directions of Fraiman and Pateiro-López (2012). For example, the PEC coincides with the classical PC when the distribution of Y is elliptically symmetric.

Proposition 4.2. *[Properties of principal expectile component] Let $Y \in \mathbb{R}^p$ be a random variable, $\phi_\tau^*(Y)$ its unique first principal expectile component.*

1. *For any constant $c \in \mathbb{R}^p$, $\phi^*_\tau(Y + c) = \phi^*_\tau(Y)$. In words, the PEC is invariant under translations of the data.*

2. *If $B \in \mathbb{R}^{p \times p}$ is an orthogonal matrix, then $\phi^*_\tau(BY) = B\phi^*_\tau(Y)$. In words, the PEC respects change of basis.*

3. *If the distribution of Y is elliptically symmetric about some point $c \in \mathbb{R}^p$, that is, there exists an invertible $p \times p$ real matrix A such that $BA^{-1}(Y - c) \overset{\mathcal{L}}{=} A^{-1}(Y - c)$ for all orthogonal matrix B, then $\phi^*_\tau(Y) = \phi^*_{1/2}(Y)$. In this case, the PEC coincides with the classical PC regardless of τ.*

4. *If the distribution of Y is spherically symmetric about some point $c \in \mathbb{R}^p$, that is, $B(Y - c) \overset{\mathcal{L}}{=} Y - c$ for all orthogonal matrix B, then all directions are principal.*

Proof. By the first part of Proposition 4.1:

$$\begin{aligned}\mathrm{Var}_\tau\{\phi^\top(Y_i + c) : i = 1, \ldots, n\} &= \mathrm{Var}_\tau(\phi^\top Y_i + \phi^\top c : i = 1, \ldots, n)\\ &= \mathrm{Var}_\tau(\phi^\top Y_i : i = 1, \ldots, n).\end{aligned}$$

This proves the first statement. For the second, note that

$$\mathrm{Var}_\tau(\phi^\top BY_i : i = 1, \ldots, n) = \mathrm{Var}_\tau\{(B^\top\phi)^\top Y_i : i = 1, \ldots, n\}.$$

Thus if ϕ^*_τ is the first τ-PEC of Y, then $(B^\top)^{-1}\phi^*_\tau$ is the first τ-PEC of BY. But B is orthogonal, that is, $(B^\top)^{-1} = B$. hence $B\phi^*_\tau$ is the τ-PEC of BY. This proves the second statement. For the third statement, by statement 1, we can assume $c \equiv 0$. Thus $Y = AZ$ where $BZ \overset{\mathcal{L}}{=} Z$ for all orthogonal matrices B. Write A in its singular value decomposition $A = UDV$, where D is a diagonal matrix with positive values $D_{ii} = d_i$ for $i = 1, \ldots p$, and U and V are $p \times p$ orthogonal matrices. Choosing $B = V^{-1}$ gives

$$\phi^*_\tau(Y) = \phi^*_\tau(UDZ) = U\phi^*_\tau(DZ).$$

Now, by Proposition 4.1, since $d_j \geq 0$ for all j,

$$\mathrm{Var}_\tau(\phi^\top DZ) = \mathrm{Var}_\tau(\sum_{j=1}^p d_j Z_j \phi_j) = \sum_j \phi_j^2 d_j^2 \mathrm{Var}_\tau(Z_j).$$

Since $\sum_j \phi_j^2 = 1$, $\mathrm{Var}_\tau(\phi^\top DZ)$ lies in the convex hull of the p numbers $d_j^2\mathrm{Var}_\tau(Z_j)$ for $j = 1, \ldots p$. Therefore, it is maximized by setting ϕ to be the unit vector along the axis j with maximal $d_j^2\mathrm{Var}_\tau(Z_j)$. But $Z \overset{\mathcal{L}}{=} BZ$ for all orthogonal matrices B, thus $Z_j \overset{\mathcal{L}}{=} Z_k$, hence $\mathrm{Var}_\tau(Z_j) = \mathrm{Var}_\tau(Z_k)$ for all indices $j, k = 1, \ldots, p$. Thus $\mathrm{Var}_\tau(\phi^\top DZ)$ is maximized when ϕ is the unit vector along the axis j with maximal d_j. This is precisely the axis with maximal singular value of A, and hence is also the direction of the (classical) principal component of DZ. This proves the claim. The last statement follows immediately from the third statement. □

To compute the principal expectile component ϕ^*_τ, one needs to optimize the right-hand side of (11) over all unit vectors ϕ. Although this is a differentiable function in ϕ, optimizing it is a difficult problem, since μ_τ also depends on ϕ, and does not have a closed form solution. However, in certain situations, for given weights w_i, not only μ_τ but also ϕ^*_τ *have* closed form solutions.

Theorem 4.1. *Consider (11). Suppose we are given the true weights w_i, which are either τ or $1-\tau$. Let $\tau_+ = \{i \in \{1,\dots,n\} : w_i = \tau\}$ denote the set of observations Y_i with 'positive' labels, and $\tau_- = \{i \in \{1,\dots,n\} : w_i = 1-\tau\}$ denote its complement. Let n_+ and n_- be the sizes of the respective sets. Define an estimator $\hat{e}_\tau \in \mathbb{R}^p$ of the τ-expectile via*

$$\hat{e}_\tau = \frac{\tau \sum_{i\in\tau_+} Y_i + (1-\tau)\sum_{i\in\tau_-} Y_i}{\tau n_+ + (1-\tau)n_-}. \tag{13}$$

Define

$$C_\tau = \frac{\tau}{n}\left\{\sum_{i\in\tau_+}(Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top\right\} + \frac{1-\tau}{n}\left\{\sum_{i\in\tau_-}(Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top\right\}. \tag{14}$$

Then ϕ^*_τ is the solution to the following optimization problem.

$$\begin{aligned} &\text{maximize } \phi^\top C_\tau \phi \\ &\text{subject to } \phi^\top Y_i > \phi^\top \hat{e}_\tau \Leftrightarrow i \in \tau_+ \\ &\qquad\qquad \phi^\top\phi = 1. \end{aligned} \tag{15}$$

In particular, the PEC is the constrained classical PC of a weighted version of the covariance matrix of the data, centered at a constant possibly different from the mean.

Proof. Since the weights are the true weights coming from the true principal expectile component ϕ^*_τ, clearly ϕ^*_τ satisfies the constraint in (15). Now suppose ϕ is another vector in this constraint set. Then $\phi^\top \hat{e}_\tau$ is exactly μ_τ, the τ-expectile of the sequence of n real numbers $\phi^\top Y_1, \dots, \phi^\top Y_n$. Therefore, the quantity we need to maximize in (11) reads

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n (\phi^\top Y_i - \mu_\tau)^2 w_i &= \frac{\tau}{n}\sum_{i\in\tau_+}(\phi^\top Y_i - \phi^\top \hat{e}_\tau)^2 + \frac{1-\tau}{n}\sum_{i\in\tau_-}(\phi^\top Y_i - \phi^\top \hat{e}_\tau)^2 \\ &= \frac{\tau}{n}\sum_{i\in\tau_+}\phi^\top(Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top\phi + \frac{1-\tau}{n}\sum_{i\in\tau_-}\phi^\top(Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top\phi \\ &= \phi^\top C_\tau \phi. \end{aligned}$$

Thus the optimization problem above is indeed an equivalent formulation of (11), which was used to define ϕ^*_τ. Finally, the last observation follows by comparing the above with the optimization formulation for PCA, see the paragraph after (9). Indeed, when $\tau = 1/2$, $\hat{e}_{1/2} = \bar{Y}$, $C_{1/2} = C$, and we recover the classical PCA. □

Since $\hat{e}_\tau$ is a linear function in the Y_i, (15) defines a system of linear constraints in the entries of Y_i and ϕ^*_τ. Thus for each fixed sign sets (τ_+, τ_-), there exist (not necessarily unique) local optima $\phi^*_\tau(\tau_+, \tau_-)$. There are 2^n possible sign sets, one of which corresponds to the global optima ϕ^*_τ that we need. It is clear that finding the global optimum ϕ^*_τ by enumerating all possible sign sets is intractable. However, in many situations, the constraint in (15) is inactive. That is, the largest eigenvector of C_τ satisfies (15) for free. In such situations, we call ϕ^*_τ a *stable solution.* Just like classical PCA, stable solutions are unique for matrices C_τ with unique principal eigenvalue. More importantly, we have an efficient algorithm for finding stable solutions, if they exist.

Definition 4.3. For some given sets of weights $w = (w_i)$, define $e_\tau(w)$ via (13), $C_\tau(w)$ via (14). Let $\phi_\tau(w)$ be the largest eigenvector of $C_\tau(w)$. If $\phi_\tau(w)$ satisfies (15), we say that $\phi_\tau(w)$ is a locally *stable solution* with weight w.

To find locally stable solutions, one can solve (8) using iterative reweighted least squares: first initialize the w_i's, compute estimators $\mu_\tau(w)$ and $\phi_\tau(w)$ ignoring the constraint (15), update the weights via (12), and iterates. At each step of this algorithm, one finds the principal component of a weighted covariance matrix with some approximate weight. Since there are only finitely many possible weight sets, the algorithm is guaranteed to converge to a locally stable solution if it exists. In particular, if the true solution to (8) is stable, then for appropriate initial weights, the algorithm will find this value. We call this algorithm PrincipalExpectile. We give a formal description of this algorithm in Section 5.

4.3 Statistical properties

We now prove consistency of local maximizers of (8). The main theorem in this section is the following.

Theorem 4.2. *Fix $\tau > 0$. Let Y_n be the empirical version of Y, a random variable in $\mathbb{R}^p$ with finite second moment, distribution function F. Suppose $\phi^* = \phi^*_\tau$ is a unique global solution to (8) corresponding to Y. Then for sufficiently large n, for any sequence of global solutions ϕ^*_n corresponding to Y_n, we have*

$$\phi^*_n \overset{F-a.s.}{\longrightarrow} \phi^*$$

in $\mathbb{R}^p$ as $n \to \infty$.

For the proof, we first need the following lemma.

Lemma 4.1. *Let Y_n be the empirical version of Y, a random variable in $\mathbb{R}^p$ with finite second moment and distribution function F. Then uniformly over all $\phi \in \mathbb{R}^p$ with $\phi^\top \phi = 1$, and uniformly over all $\tau \in (0,1)$,*

$$\mathrm{Var}_\tau(Y_n^\top \phi) \overset{F-a.s.}{\longrightarrow} \mathrm{Var}_\tau(Y^\top \phi).$$

Proof. Since Y_n is the empirical version of Y and the set of all unit vectors $\phi \in \mathbb{R}^p, \phi^\top \phi = 1$ is compact, by the Cramer-Wold theorem, $Y_n^\top \phi \overset{\mathcal{L}}{\to} Y^\top \phi$ uniformly over all such unit vectors $\phi \in \mathbb{R}^p$. It then follows that e_τ and Var_τ, which are completely determined by the distribution function, also converge $F - a.s.$ uniformly over all ϕ. □

Proof of Theorem 4.2. Let $\mathbb{S}^{p-1}$ denote the unit sphere in $\mathbb{R}^p$. Equip $\mathbb{R}^p$ with the Euclidean norm $\|\cdot\|$. Define the map $V_Y : \mathbb{S}^{p-1} \to \mathbb{R}$, $V_Y(\phi) = \mathrm{Var}_\tau(Y^\top \phi)$. Fix $\epsilon > 0$. We shall prove that there exists a $\delta > 0$ such that the global minimum of V_{Y_n} is necessarily within δ-distance of ϕ^*.

Since V_Y is continuous, $\mathbb{S}^{p-1}$ is compact, and ϕ^* is unique, there exists a sufficiently small $\delta > 0$ such that

$$|V_Y(\phi) - V_Y(\phi^*)| < \epsilon \Rightarrow \|\phi - \phi^*\| < \delta$$

for $\phi \in \mathbb{S}^{p-1}$. In particular, if $\|\phi - \phi^*\| > \delta$, then

$$V_Y(\phi^*) + \epsilon < V_Y(\phi).$$

By Lemma 4.1, $V_{Y_n} \to V_Y$ as $n \to \infty$ uniformly over $\mathbb{S}^{p-1}$. In particular, there exists a large N such that for all $n > N$,

$$|V_{Y_n}(\phi) - V_Y(\phi)| < \epsilon/6$$

for all $\phi \in \mathbb{S}^{p-1}$. Thus for $\phi \in \mathbb{S}^{p-1}$ such that $\|\phi - \phi^*\| > \delta$,

$$V_{Y_n}(\phi) - V_Y(\phi^*) > \epsilon - \epsilon/6 = 5\epsilon/6.$$

Meanwhile, since V_Y is continuous, one can choose $\epsilon' = \epsilon/6$, and thus obtain δ' such that

$$|V_Y(\phi) - V_Y(\phi^*)| < \epsilon/6 \Leftarrow \|\phi - \phi^*\| < \delta'.$$

Then, for ϕ such that $\|\phi - \phi^*\| < \delta'$,

$$V_{Y_n}(\phi) - V_Y(\phi^*) \leq |V_{Y_n}(\phi) - V_Y(\phi)| + |V_Y(\phi) - V_Y(\phi^*)| < \epsilon/6 + \epsilon/6 = \epsilon/3.$$

So far we have shown that if $\|\phi - \phi^*\| > \delta$, then $V_{Y_n}(\phi)$ is at least $5\epsilon/6$ bigger than $V_Y(\phi^*)$. Meanwhile, if $\|\phi - \phi^*\| < \delta'$, then $V_{Y_n}(\phi)$ is at most $\epsilon/3$ bigger than $V_Y(\phi^*)$. Thus the global minimum ϕ_n^* of V_{Y_n} necessarily satisfy $\|\phi_n^* - \phi^*\| < \delta$. This completes the proof. □

5 Algorithms

5.1 TopDown and BottomUp

We now describe how iterative weighted least squares can be adapted to implement TopDown and BottomUp. We start with a description of the asymmetric weighted least squares (LAWS) algorithm of Newey and Powell Newey and Powell (1987). The basic algorithm outputs a subspace without the affine term, and needs to be adapted. See Guo et al. (2013) for a variation with smoothing penalty and spline basis.

Proposition 5.1. *The LAWS algorithm is well-defined, and is a gradient descent algorithm. Thus it converges to a critical point of the optimization problem* (6).

Proof. First, we note that the steps in the algorithm are well-defined. For fixed W and V, $J(U, V, W)$ is a quadratic in the entries of U. Thus the global minimum on line 8 has an explicit solution, see Srebro and Jaakkola (2003); Guo et al. (2013). A similar statement applies to line 9.

As noted in Section 3.2, $J(U, V, W)$ is not jointly convex in U and V, but as a function in U for fixed V, it is a convex, continuously differentiable, piecewise quadratic function. The statement holds for $J(U, V, W)$ as a function in V for fixed U. Hence lines 8 and 9 is one step in a Newton-Raphson algorithm on $J(U, V, W)$ for fixed V. Similarly, lines 10 and 11 is one step in a Newton-Raphson algorithm on $J(U, V, W)$ for fixed U. Thus the algorithm is a coordinatewise gradient descent on a coordinatewise convex function, hence converges. □

Algorithm 1 Asymmetric weighted least squares (LAWS)

```
1: Input: data $Y \in \mathbb{R}^{n\times p}$, positive integer $k < p$
2: Output: $\hat{E}^*_k$, an estimator of $E^*_k$, expressed in product form $\hat{E}^*_k = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n\times k}, \hat{V} \in \mathbb{R}^{p\times k}.\hat{U}, \hat{V}$ are unique up to multiplication by an invertible matrix.
3: procedure LAWS($Y, k$)
4:     Set $V^{(0)}$ to be some rank-$k$ $p \times k$ matrix.
5:     Set $W^{(0)} \in \mathbb{R}^{n\times p}$ to be 1/2 everywhere.
6:     Set $t = 0$.
7:     repeat
8:         Update $U$: Set $U^{(t+1)} = \text{argmin}_{U\in\mathbb{R}^{n\times k}} J(U, V^{(t)}, W^{(t)})$.
9:         Update $W$: Set $W_{ij}^{(t+1)} = \tau$ if $Y_{ij} - \sum_l U_{il}^{(t+1)} V_{lk}^{(t)} > 0$, $W_{ij}^{(t+1)} = 1 - \tau$ otherwise.
10:        Update $V$: Set $V^{(t+1)} = \text{argmin}_{V\in\mathbb{R}^{k\times p}} J(U^{(t+1)}, V, W^{(t+1)})$.
11:        Update $W$: Set $W_{ij}^{(t+1)} = \tau$ if $Y_{ij} - \sum_l U_{il}^{(t+1)} V_{lk}^{(t+1)} > 0$, $W_{ij}^{(t+1)} = 1 - \tau$ otherwise.
12:        Set t = t + 1
13:    until $U^{(t+1)} = U^{(t)}, V^{(t+1)} = V^{(t)}, W^{(t+1)} = W^{(t)}$.
14: return $\hat{E}_k = U^{(t)}(V^{(t)})^\top$.
15: end procedure
```

If some columns of U or V are pre-specified, one can run LAWS and not update these columns in lines 8 and 10. Thus one can use LAWS to find the optimal affine subspace by writing $\mathbf{1}m^\top + E = \tilde{U}\tilde{V}$ with the first column of $\tilde{U}$ constrained to be $\mathbf{1}$. Similarly, we can use this technique to solve the constrained optimization problems:

- *Find a rank-k approximation E_k whose span contains a given subspace of dimension $r < k$*
- Solution: Constrain the first r columns of $V^{(0)}$ to be a basis of the given subspace.
- *Find a rank-k approximation whose span lies within a given subspace of dimension $r > k$.*
- Solution: Let $B \in \mathbb{R}^{n\times r}$ be a basis of the given subspace. Then the optimization problem becomes
$$\min_{U\in\mathbb{R}^{r\times k}, V\in\mathbb{R}^{p\times k}} \|Y - BUV^\top\|^2_{\tau,2}.$$
  One can then apply the LAWS algorithm with variables U and V.
- *Find a rank-k approximation whose span contains a given subspace of dimension $r < k$, and is contained in a given subspace of dimension $R > k$.*
- Solution: Combine the previous two solutions.

With these tools, we now define the two algorithms, TopDown and BottomUp.

The TopDown algorithm requires the weights w_{ij} and the loadings on previous principal components to be re-evaluated when finding the next principal component. A variant of the algorithm would be to keep the weights w_{ij}. In this case, the algorithm is still well-defined. However, it will produce a different basis matrix $\hat{U}$, since the estimators are no longer optimal in the $\|\cdot\|^2_{\tau,2}$ norm.

Algorithm 2 TopDown

```
Input: data $Y \in \mathbb{R}^{n \times p}$, positive integer $k < p$
Output: $\hat{E}_k^*$, an estimator of $E_k^*$, expressed in product form $\hat{E}_k^* = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n \times k}, \hat{V} \in \mathbb{R}^{p \times k}$ are unique.
procedure TopDown($Y, k$)
    Use LAWS(Y,k) to find $\hat{m}_k^*, \hat{E}_k^*$. Write $\hat{E}_k^* = UV^\top$ for some orthonormal basis $U$.
    Use LAWS to find $\hat{U}_1$, the vector which spans the optimal subspace of dimension 1 contained in $U$.
    Use LAWS to find $\hat{U}_2$, where $(\hat{U}_1, \hat{U}_2)$ spans the optimal subspace of dimension 1 contained in $U$ and contains the span of $\hat{U}_1$
    Repeat the above step until obtains $\hat{U}$.
    Obtain $\hat{V}$ through the constraint $\hat{E}_k^* = \hat{U}\hat{V}^\top$.
return $\hat{m}_k^*, \hat{E}_k^*, \hat{U}, \hat{V}^\top$.
end procedure
```

Algorithm 3 BottomUp

```
Input: data $Y \in \mathbb{R}^{n \times p}$, positive integer $k < p$
Output: $\hat{E}_k^*$, an estimator of $E_k^*$, expressed in product form $\hat{E}_k^* = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n \times k}, \hat{V} \in \mathbb{R}^{p \times k}$ are unique.
procedure BottomUp($Y, k$)
    Use LAWS to find $\hat{E}_1^*$. Let $\hat{U}_1$ be the basis vector.
    Use LAWS to find $\hat{U}_2$ such that $(\hat{U}_1, \hat{U}_2)$ is the best two-dimensional approximation to $Y$, subjected to containing $\hat{U}_1$.
    Repeat the above step until obtains $\hat{U}$. We obtain $\hat{V}$ and $\hat{E}_k^*$ in the last iteration.
return $\hat{E}_k^*, \hat{U}, \hat{V}^\top$.
end procedure
```

5.2 Performance bounds of TopDown and BottomUp

We now show that the dependence on k only grows polylog in n. Thus both TopDown and BottomUp are fairly efficient algorithms even for large k.

Theorem 5.1. *For fixed V of dimension k, LAWS requires at most $\mathcal{O}\{\log(p)^k\}$ iterations, $\mathcal{O}\{npk^2 \log(p)^k\}$ flops to estimate U.*

In other words, if V has converged, LAWS needs at most $\mathcal{O}\{npk^2 \log(p)^k\}$ flops to estimate U. The role of U and V are interchangeable if we transpose Y. Thus if U has converged, LAWS needs at most $\mathcal{O}\{npk^2 \log(n)^k\}$ to estimate V. We do not have a bound for the number of iterations needed until convergence. In practice this seem to be of order log of n and p. For the proof of Theorem 5.1 we need the following two lemmas.

Lemma 5.1. *If $Y_1, \ldots, Y_n \in \mathbb{R}$ are n real numbers, then LAWS finds their τ-expectile e_τ in $\mathcal{O}\{\log(n)\}$ iterations.*

Proof. Given the weights $w_1, \ldots, w_n$, that is, given which Y_i's are above and below e_τ, the τ-expectile e_τ is a linear function in the Y_i as we saw in (13). As shown in Proposition 5.1, LAWS is equivalent to a Newton-Raphson algorithm on a piecewise quadratic function.

Since the points Y_i's are ordered, it takes $\mathcal{O}\{\log(n)\}$ to learn their true weights. Thus the algorithm converges in $\mathcal{O}\{\log(n)\}$ iterations. □

Lemma 5.2. *An affine line in $\mathbb{R}^p$ can intersect at most $2p$ orthants.*

Proof. Recall that an *orthant* of $\mathbb{R}^p$ is a subset of $\mathbb{R}^p$ where the sign of each coordinate is constrained to be either nonnegative or nonpositive. There are 2^p orthants in $\mathbb{R}^p$. Let $f(\lambda) = Y + \lambda v$ be our affine line, $\lambda \in \mathbb{R}, Y, v \in \mathbb{R}^p$. Let $\text{sgn} : \mathbb{R}^p \to \{\pm 1\}^p$ denote the sign function. Now, $\text{sgn}\{f(0)\} = \text{sgn}(Y), \text{sgn}\{f(\infty)\} = \text{sgn}(v)$, and $\text{sgn}\{f(\lambda)\}$ is a monotone increasing function in λ. As $\lambda \to \infty$, $\text{sgn}\{f(\lambda)\}$ goes from $\text{sgn}(Y)$ to $\text{sgn}(v)$ one bit flip at a time. Thus there are at most p flips, that is, the half-line $f(\lambda)$ for $\lambda \in [0, \infty)$ intersects at most p orthants. By a similar argument, the half-line $f(\lambda)$ for $\lambda \in (-\infty, 0)$ intersects at most p other orthants. This concludes the proof. □

Corollary 5.1. *An affine subspace of dimension k in $\mathbb{R}^p$ can intersect at most $\mathcal{O}(p^k)$ orthants.*

Proof. Fix any basis, say $\psi_1, \ldots, \psi_k$. By Lemma 5.2, ψ_1 can intersect at most $2p$ orthants. For each orthant of ψ_1, varying along ψ_2 can yield at most another $2p$ orthants. The proof follows by induction. (This is a rather liberal bound, but it is of the correct order for k small relative to p). □

Proof of Theorem 5.1. By Corollary 5.1, it is sufficient to consider the case $k = 1$. Fix V of dimension 1. Since U, V are column matrices, we write them in lower case letters u, v. Solving for each u_i is a separate problem, thus we have n separate optimization problem, and it is sufficient to prove the claim for each i for $i = 1, \ldots, n$.
Fix an i. As u_i varies, $Y_i - m_i - u_i v$ defines a line in $\mathbb{R}^p$. The weight vector $(w_{i1}, \ldots, w_{ip})$ only depend on which coordinates are the orthant of $\mathbb{R}^p$ in which $Y_i - m_i - u_i v$ is in. The later is equivalently to determining the weight of the p points $\frac{Y_i - m_i}{v_i}$. By Lemma 5.1, it takes $\mathcal{O}\{\log(p)\}$ for LAWS to determine the weights correctly. Thus LAWS takes at most $\mathcal{O}\{\log(p)\}$ iterations to converge, since each iteration involves estimating w, then v. Each iteration solves a weighted least squares, thus take $\mathcal{O}(npk^2)$. Hence for fixed v, LAWS can estimate u after at most $\mathcal{O}\{npk^2 \log(p)\}$ flops for $k = 1$. This concludes the proof for fixed v. By considering the transposed matrix Y, we see that the role of u and v are interchangeable. The conclusion follows similarly for fixed u. □

5.3 PrincipalExpectile

In this section we describe the PrincipalExpectile algorithm. This algorithm is used to compute the principal expectile component defined in Section 4. We shall describe the case $k = 1$, that is, the algorithm for computing the first principal expectile component only. To obtain higher order components, one iterates the algorithm over the residuals $Y_i - \hat{\phi}_1(\hat{\phi}_1^\top Y_i + \hat{\mu}_1)$, where $\hat{\mu}_1$ is the τ-expectile of the loadings $\hat{\phi}_1^\top Y_i$.

For n observations $Y_1, \ldots, Y_n$, there are at most 2^n possible labels for the Y_i's, and hence the algorithm has in total 2^n possible values for the w_i's. Thus either Algorithm 4 converges to a point which satisfies the properties of the optimal solution that Theorem 4.1 prescribes, or that it iterates infinitely over a cycle of finitely many possible values

Algorithm 4 PrincipalExpectile

```
Input: data $Y \in \mathbb{R}^{n \times p}$.
Output: a vector $\hat{\phi}$, an estimator of the first principal expectile component of $Y$.
procedure PrincipalExpectile($Y$)
    Initialize the weights $w_i^{(0)}$
    Set $t = 0$.
    repeat
        Let $\tau_+^{(t)}$ be the set of indices $i$ such that $w_i^{(t)} = \tau$, and $\tau_-^{(t)}$ be the complement.
        Compute $e_\tau^{(t)}$ as in equation (13) with sets $\tau_+^{(t)}, \tau_-^{(t)}$.
        Compute $C_\tau^{(t)}$ as in equation (14) with sets $\tau_+^{(t)}, \tau_-^{(t)}$.
        Set $\phi^{(t)}$ to be the largest eigenvector of $C_\tau^t (C_\tau^t)^\top$
        Set $\mu_\tau^{(t)}$ to be the $\tau$-expectile of $(\phi^{(t)})^\top Y_i$
        Update $w_i$: set $w_i^{(t+1)} = \tau$ if $(\phi^{(t)})^\top Y_i > \mu_\tau^{(t)}$, and set $w_i^{(t+1)} = 1 - \tau$ otherwise.
        Set t = t + 1
    until $w_i^t = w_i^{(t+1)}$ for all $i$.
return $\hat{\phi} = \phi^{(t)}$.
end procedure
```

of the w_i's. In particular, the true solution is a fixed point, and thus fixed points always exist. In practice, we find that the algorithm converges very quickly, and can get stuck in a finite cycle of values. In this case, one can jump to a different starting point and restart the algorithm. Choosing a good starting value is important in ensuring convergence. Since the τ-variance is a continuous function in τ, we find that in most cases, one can choose a good starting point by performing a sequence of such computations for a sequence of τ starting with $\tau = 1/2$, and set the initial weight to be that induced by the previous run of the algorithm for a slightly smaller (or larger) τ.

6 Simulation

To study the finite sample properties of the proposed algorithms we do a simulation study. We follow the simulation setup of Guo et al. (2013), that is, we simulate the data $Y_{ij}, i = 1, \dots, n, \; j = 1, \dots, p$ as

$$Y_{ij} = \mu(t_j) + f_1(t_j)\alpha_{1i} + f_2(t_j)\alpha_{2i} + \varepsilon_{ij}, \tag{16}$$

where t_j's are equidistant on [0,1], $\mu(t) = 1 + t + \exp\{-(t - 0.6)^2/0.05\}$ is the mean function, $f_1(t) = \sqrt{2}\sin(2\pi t)$ and $f_2(t) = \sqrt{2}\cos(2\pi t)$ are principal component curves, and ε_{ij} is a random noise.
We consider different settings 1 and 2 each with five error scenarios:

1. $\alpha_{1i} \sim \mathrm{N}(0, 36)$ and $\alpha_{2i} \sim \mathrm{N}(0, 9)$ are simulated independently across i with (1) iid errors drawn from $\mathrm{N}(0, \sigma^2)$, (2) iid errors from $t(5)$, (3) independent heteroscedastic errors from $\mathrm{N}\{0, \mu(t_j)\sigma^2\}$, (4) errors from $\log\mathrm{N}(0, \sigma^2)$ and (5) errors from a sum of two uniforms $U(0, \sigma^2)$ where σ^2=0.5.

2. $\alpha_{1i} \sim \mathrm{N}(0, 16)$ and $\alpha_{2i} \sim \mathrm{N}(0, 9)$ are simulated independently across i with (1) iid errors drawn from $\mathrm{N}(0, \sigma^2)$, (2) iid errors from $t(5)$, (3) independent heteroscedastic errors from $\mathrm{N}\{0, \mu(t_j)\sigma^2\}$, (4) errors from $\log\mathrm{N}(0, \sigma^2)$ and (5) errors from a sum of two uniforms $U(0, \sigma^2)$ where σ^2=1.

Note that the settings imply different ratios of coefficient-to-coefficient-to-noise variations. In the setting 1 scenario (1) we have a ratio 36:9:0.5, whereas in the setting 2 scenario (1) we have 16:9:1. Apart from standard Gaussian errors, we also consider "fat tailed" errors, heteroscedastic and skewed errors. We study the performance of the algorithms for three sample sizes: (i) small n=20, p=100; (ii) medium n=50, p=150; (iii) large n=100, p=200.

For every single case we repeat the simulation 500 times and record the mean computing times, the mean of the average mean squared error (MSE), its standard deviation, and convergence ratio for each algorithm. We label the run of the algorithm as unconverged whenever after 30 iterations and 50 restarts from a random starting point the algorithms fail to converge.

Convergence statistics are reported in Table 1 and computational time records are in Table 2. The results on the MSEs for both simulation settings are presented in Tables 3 and 4 respectively. For the ease of notation we write BUP for BottomUp, TD for TopDown and PEC for PrincipalExpectile.

The results for the settings 1 and 2 differ in the magnitude of the average MSE but there is no substantial qualitative difference in relative performance of the algorithms. BUP performs the worst of the three algorithms in terms of its MSE in all scenarios. TD and PEC are comparable in terms of their MSEs. PEC shows robustness against skewness and fat tails in the error distribution since it produces the lowest MSEs in scenarios (2) and (4). Yet TD tends to slightly outperform PEC in medium and large samples by errors close to iid normal or normal heteroscedastic; by small sample sizes PEC outperforms TD in all scenarios but (5).

Figures 1 and 2 illustrate the difference in the quality of component estimation for the 95% expectile when coefficient-to-coefficient-to-noise variation ratio changes (setting 1 versus setting 2 respectively). The results are shown for the error scenario (1) and small sample size. We observe that as the ratio changes from 36:9:0.5 (setting 1, Figure 1) to 16:9:1 (setting 2, Figure 2) the variability of the estimators of both component functions increases. The overall mean of the estimators remains very close to the true component functions.

sample		small			medium			large	
τ/rate	BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
0.900	0.02	0.00	0.24	0.01	0.00	0.23	0.00	0.00	0.20
0.950	0.18	0.03	0.22	0.05	0.00	0.26	0.06	0.00	0.21
0.975	0.43	0.22	0.21	0.23	0.04	0.25	0.17	0.00	0.24

Table 1: Nonconvergence rates of the algorithms by 500 simulation runs

PEC is the fastest algorithm as shown in Table 2. For large sample and high expectile level it is more than three times faster than TD and more than five times faster than BUP. Although the fastest from considered algorithms the PEC is considerably slower than the classical PCA routins with the average computational times 0.002 seconds for small, 0.005 seconds for medium, and 0.023 seconds for large sample (computed by function `prcomp` in package `stats` of statistical script language R).

sample		small			medium			large	
τ/sec	BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
0.900	1.24	0.70	0.57	2.91	1.59	1.39	7.53	4.02	2.71
0.950	1.64	1.13	0.55	4.01	2.68	1.57	10.53	6.88	3.03
0.975	2.36	2.05	0.56	5.56	4.59	1.56	14.62	10.96	3.54

Table 2: Average time in seconds for convergence of the algorithms by 500 simulations

The major draw back of PEC is the relative low convergence rate: for all sample sizes only around 80% of algorithm runs were convergent. In 20% cases the algorithm keeps iterating between two sets of weights which possibly indicates an adverse sample geometry, i.e. that two eigevalues of the scaled covariance matrix are too close to each other. TD, on the contrary, converges almost always in medium and large sample sizes.

We conclude that whenever the error distribution is fat-tailed or skewed, or by small samples PEC is likely produce more reliable results in terms of its MSE, whereas by errors close to normal and moderate or large samples TD is likely to produce smaller MSEs.

7 Application to Chinese Weather Data

Weather derivatives (WDs) are financial instruments written on weather indices as underlyings and are designed to trade with weather related risks. Temperature derivatives are WDs written on a temperature index such as the average temperature recorded at a prespecified weather station. As for financial derivatives risk factors of temperature are at the core of temperature derivative pricing. In this section we study the risk factors of temperature using daily average temperature data of 159 weather stations in China reported by Chinese Meteorological Administration for the years 1957 to 2009. We refer to this dataset as the Chinese weather dataset.

To conduct the analysis of the temperature risk factors which are relevant for pricing temperature derivatives we follow the well established methodology of Benth et al. (2007). That is, let T_{it} denote the average temperature at station i, $i = 1, 2, \ldots, n$ in time t. We consider each station i separately and using the whole time series of the average temperatures from 1957 to 2009 we fit the following model:

$$\begin{aligned} T_{it} &= X_{it} + \Lambda_{it} \\ \Lambda_{it} &= a_i + b_i t + c_i \sin(2\pi t/365) + d_i \cos(2\pi t/365) + g_i \sin(\pi t/365) + h_i \cos(\pi t/365) \quad (17) \\ X_{it} &= \sum_{j=1}^{10} \beta_{ij} X_{i,t-j} + \varepsilon_{it} \end{aligned}$$

We fit the model (17) to the temperature data of 159 stations and obtain the estimated residuals $\hat{\varepsilon}_{it}$.

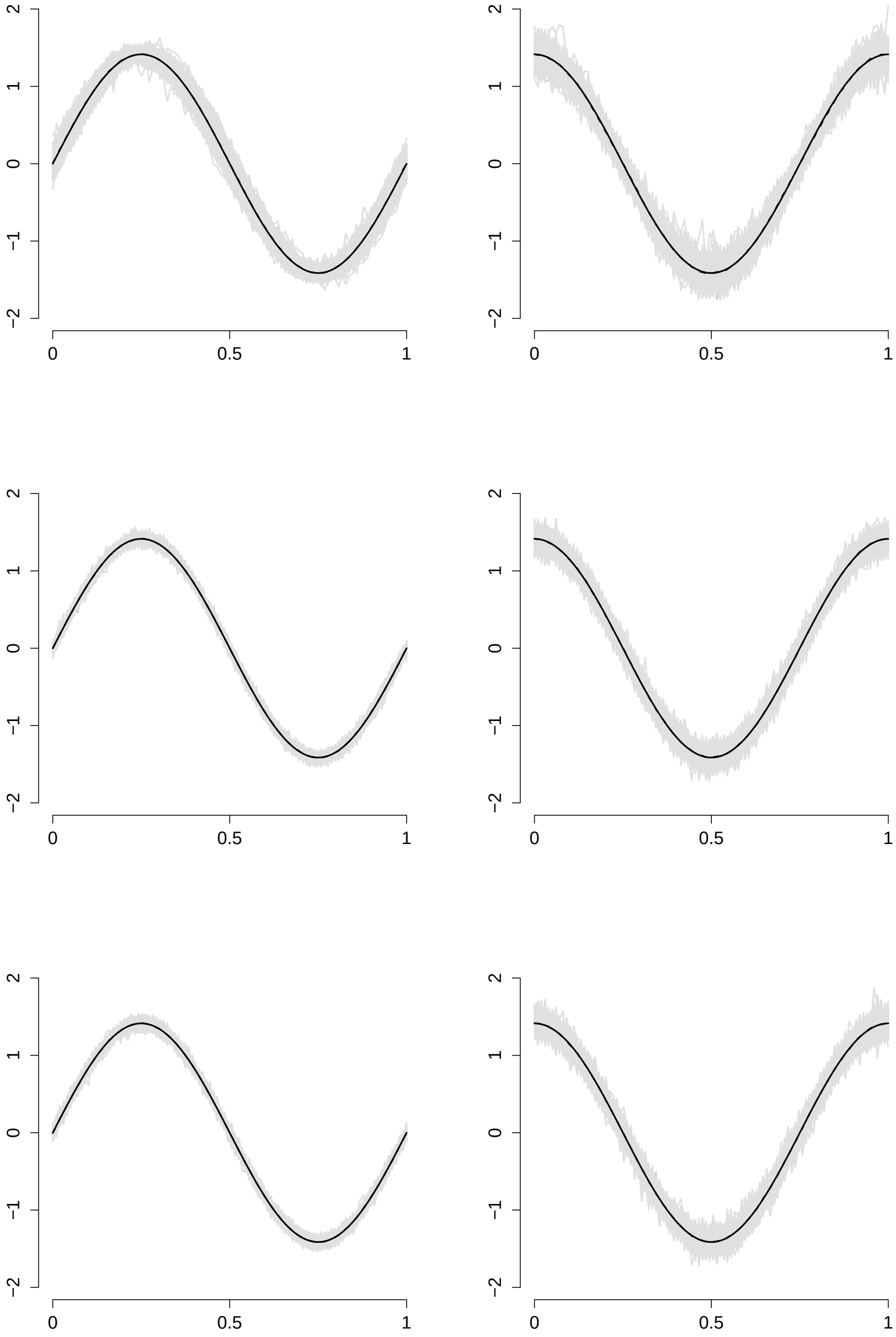

Figure 1: Estimated component functions (solid gray) by 500 simulation runs for simulation setting 1 scenario 1 small sample size and 95% expectile. The rows from the top to the bottom show respectively results produced by BUP, TD and PEC. Left panel corresponds to the first component function, right panel - to the second. The true functions are shown as solid black curves. The overall mean across simulation runs is shown as dashed black curve. The later can not be distinguished from the true curve.

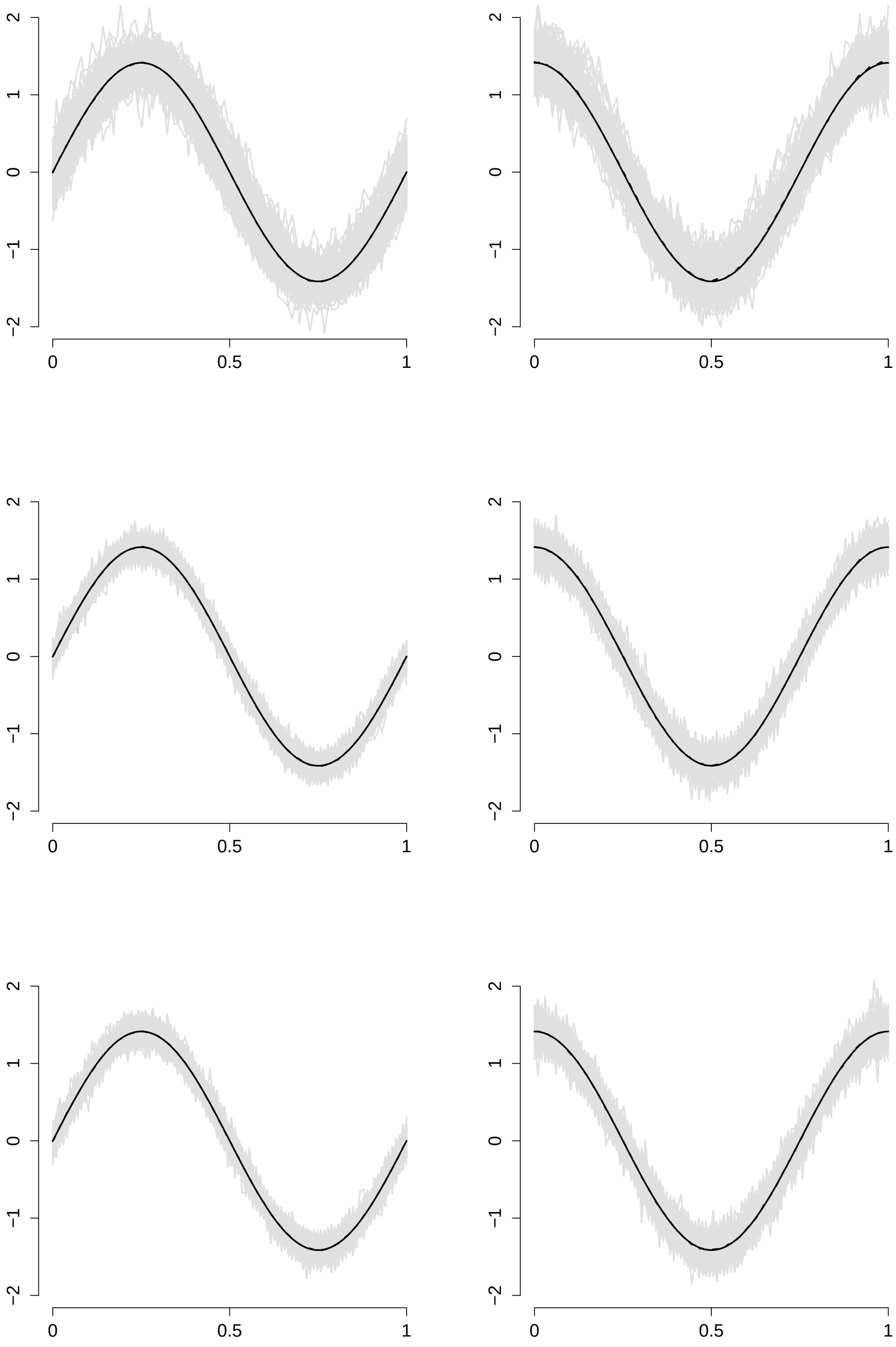

Figure 2: Estimated component functions (gray) by 500 simulation runs for simulation setting 2 scenario 1 small sample size and 95% expectile. The rows from the top to the bottom show respectively results produced by BUP, TD and PEC. Left panel corresponds to the first component function, right panel - to the second. The true functions are shown as solid black curves. The overall mean across simulation runs is shown as dashed black curve. The later can not be distinguished from the true curve.

scenario	τ	n=20, p=100			n=50, p=150			n=100, p=200		
		BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
(1)	0.900	0.2762	0.1216	0.1123	0.1339	0.0538	0.0632	0.0698	0.0297	0.0459
		(0.1997)	(0.0097)	(0.0111)	(0.1099)	(0.0033)	(0.0029)	(0.0552)	(0.0015)	(0.0014)
	0.950	0.3619	0.1568	0.1334	0.2323	0.0705	0.0727	0.1312	0.0394	0.051
		(0.2199)	(0.0123)	(0.0181)	(0.2076)	(0.0045)	(0.0044)	(0.1415)	(0.0020)	(0.0019)
	0.975	0.5064	0.2053	0.1601	0.3583	0.0944	0.0874	0.2157	0.0536	0.0594
		(0.2977)	(0.0154)	(0.0276)	(0.2989)	(0.0060)	(0.0075)	(0.2314)	(0.0027)	(0.0035)
(2)	0.900	0.7092	0.5421	0.3147	0.3382	0.2714	0.1494	0.1866	0.1548	0.0932
		(0.2382)	(0.1096)	(0.0685)	(0.1223)	(0.0727)	(0.0117)	(0.0522)	(0.0217)	(0.0050)
	0.950	1.105	0.7847	0.3854	0.5789	0.4440	0.1819	0.3316	0.2680	0.1101
		(0.4453)	(0.1646)	(0.0988)	(0.2664)	(0.1675)	(0.0192)	(0.1144)	(0.0575)	(0.0075)
	0.975	1.6066	1.1158	0.4709	0.9956	0.7033	0.2309	0.5780	0.4641	0.1358
		(0.7968)	(0.2106)	(0.1413)	(0.6936)	(0.2629)	(0.0341)	(0.2227)	(0.1175)	(0.0132)
(3)	0.900	0.4146	0.2300	0.2215	0.1829	0.1019	0.1270	0.0962	0.0562	0.0942
		(0.2413)	(0.0195)	(0.0236)	(0.1070)	(0.0065)	(0.0066)	(0.0510)	(0.0029)	(0.0032)
	0.950	0.6261	0.2966	0.2792	0.3538	0.1335	0.1622	0.1603	0.0746	0.1208
		(0.6313)	(0.0246)	(0.0369)	(1.1684)	(0.0088)	(0.0097)	(0.1135)	(0.0039)	(0.0045)
	0.975	0.8051	0.3885	0.3516	0.4879	0.1789	0.2109	0.2665	0.1016	0.1568
		(0.4516)	(0.0312)	(0.0527)	(0.3736)	(0.0118)	(0.0167)	(0.2234)	(0.0052)	(0.0077)
(4)	0.900	0.9162	0.8041	0.2226	0.4854	0.4510	0.1077	0.2876	0.2763	0.0697
		(0.2432)	(0.1532)	(0.0588)	(0.1093)	(0.0597)	(0.0089)	(0.0498)	(0.0247)	(0.0042)
	0.950	1.4972	1.2869	0.2725	0.9127	0.8092	0.1296	0.5585	0.5280	0.0812
		(0.4494)	(0.2337)	(0.0713)	(0.4895)	(0.1187)	(0.0142)	(0.1595)	(0.0554)	(0.0069)
	0.975	2.3371	1.9727	0.3331	1.5522	1.3387	0.1629	1.2223	0.9421	0.0995
		(1.0034)	(0.2835)	(0.0979)	(0.7483)	(0.1999)	(0.0248)	(1.4707)	(0.1110)	(0.0117)
(5)	0.900	0.0343	0.0091	0.0368	0.0298	0.0038	0.0315	0.0244	0.0021	0.0296
		(0.0224)	(0.0007)	(0.0013)	(0.0261)	(0.0002)	(0.0004)	(0.0238)	(0.0001)	(0.0002)
	0.950	0.1225	0.0110	0.0409	0.0351	0.0044	0.0345	0.0285	0.0023	0.0322
		(1.1145)	(0.0008)	(0.0020)	(0.0398)	(0.0003)	(0.0007)	0.0254	(0.0004)	(0.0004)
	0.975	0.0776	0.0135	0.0474	0.0455	0.0052	0.0397	0.0360	0.0027	0.0366
		(0.3266)	(0.0011)	(0.0034)	(0.0658)	(0.0003)	(0.0012)	(0.0309)	(0.0001)	(0.0006)

Table 3: average MSE and its standard deviation in brackets by 500 simulation runs for the first setting 1.

scenario	τ	n=20, p=100			n=50, p=150			n=100, p=200		
		BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
(1)	0.900	0.4484	0.2436	0.1988	0.2053	0.1077	0.1002	0.1109	0.0595	0.0660
		(0.2671)	(0.0195)	(0.0238)	(0.1273)	(0.0066)	(0.0058)	(0.0924)	(0.0030)	(0.0027)
	0.950	0.7021	0.314	0.2418	0.3681	0.1411	0.119	0.2075	0.0788	0.0761
		(0.4611)	(0.0246)	(0.0386)	(0.3066)	(0.0090)	(0.0091)	(0.2346)	(0.0039)	(0.0039)
	0.975	0.9218	0.4116	0.2945	0.5957	0.1890	0.1483	0.3364	0.1074	0.0925
		(0.5578)	(0.0312)	(0.0546)	(0.4751)	(0.0121)	(0.0152)	(0.3565)	(0.0053)	(0.0067)
(2)	0.900	0.7424	0.5427	0.3186	0.3560	0.2716	0.1502	0.2047	0.1549	0.0935
		(0.2933)	(0.1099)	(0.0762)	(0.1695)	(0.0728)	(0.0123)	(0.1886)	(0.0218)	(0.0050)
	0.950	1.1483	0.7855	0.3920	0.6656	0.4437	0.1832	0.3805	0.2684	0.1103
		(0.5078)	(0.1643)	(0.1096)	(0.6719)	(0.1658)	(0.0185)	(0.3563)	(0.0581)	(0.0075)
	0.975	1.7083	1.1095	0.4805	1.1714	0.7048	0.2342	0.6974	0.4648	0.1368
		(0.8614)	(0.1744)	(0.1493)	(0.9716)	(0.2652)	(0.0323)	(0.5981)	(0.1192)	(0.0126)
(3)	0.900	0.6616	0.4613	0.4093	0.2993	0.2041	0.2200	0.1684	0.1126	0.1540
		(0.2625)	(0.0392)	(0.0486)	(0.1163)	(0.0131)	(0.0134)	(0.1880)	(0.0058)	(0.0066)
	0.950	1.0027	0.5948	0.5229	0.4979	0.2675	0.2875	0.3031	0.1494	0.2042
		(0.5055)	(0.0495)	(0.0802)	(0.3671)	(0.0177)	(0.0215)	(0.4360)	(0.0077)	(0.0090)
	0.975	1.465	0.7811	0.6719	0.8605	0.3587	0.3831	0.5173	0.2036	0.2724
		(0.8018)	(0.0627)	(0.1154)	(0.8004)	(0.0237)	(0.0338)	(0.6708)	(0.0103)	(0.0156)
(4)	0.900	5.4073	5.2042	1.0318	3.3226	3.2871	0.4075	2.0358	2.0686	0.2295
		(2.1503)	(1.9812)	(0.9534)	(1.1548)	(1.0106)	(0.1258)	(0.6044)	(0.5259)	(0.1632)
	0.950	8.7171	8.0696	1.4256	6.5227	6.2094	0.5143	4.5541	4.4481	0.2939
		(2.8223)	(2.3418)	(1.4550)	(1.9576)	(1.5846)	(0.1540)	(1.4193)	(1.0287)	(0.3150)
	0.975	13.419	11.635	2.0054	11.202	9.8804	0.7372	8.9280	8.3663	0.3889
		(5.1223)	(1.6721)	(2.2733)	(4.0968)	(1.8550)	(0.5037)	(2.4679)	(2.7240)	(0.3161)
(5)	0.900	0.1135	0.0365	0.0572	0.0923	0.0153	0.0394	0.0561	0.0083	0.0333
		(0.0755)	(0.0027)	(0.0041)	(0.0878)	(0.0009)	(0.0011)	(0.0628)	(0.0004)	(0.0005)
	0.950	0.1430	0.0440	0.0651	0.1197	0.0177	0.0434	0.0896	0.0093	0.0356
		(0.1214)	(0.0034)	(0.0060)	(0.1033)	(0.0010)	(0.0018)	(0.0938)	(0.0005)	(0.0008)
	0.975	0.2489	0.0540	0.0769	0.1538	0.0209	0.0499	0.1145	0.0107	0.0396
		(0.6091)	(0.0042)	(0.0099)	(0.1272)	(0.0013)	(0.0031)	(0.1042)	(0.0006)	(0.0013)

Table 4: average MSE and its standard deviation in brackets by 500 simulation runs for the simulation setting 2.

It is crucial to study these risk factors $\hat{\varepsilon}_{it}$ for the pricing of WDs since the later relies heavily on the distributional properties of ε_{it}, frequently ε_{it} are assumed to be Gaussian, Benth et al. (2007) and Alaton et al. (2002). The findings of Campbell and Diebold (2005) reveal the importance of modeling conditional variances beside conditional means to capture the distributional features of temperature. We go beyond this and look at the scale factors in the tails of the ε_{it}'s.

To eliminate possible year-specific level and scale effects in ε_{it} for different years, we average and demean the $\hat{\varepsilon}_{it}$ day-wise (the 29th February was droped from the data) over all years, and presmooth them using 24 Fourier series.

We run the algorithms to estimate a collection of 159 expectile curves for the weather stations at each of the levels 5%, 50% and 95% with respect to days of a year from 1 to 365. We estimate first two principal component functions. As we show in Table 5 they already explain large portion of the sample variation.

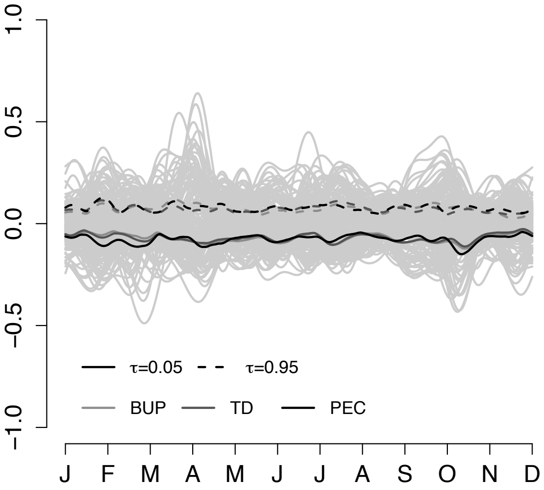

Figure 3: Averaged and smoothed residuals of temperature on 159 stations (gray) and the estimated constants by the algorithms. The horizontal axis features the months from January to December.

The estimation results of the three proposed algorithms are rather similar. On Figures 3 and 4 we present the estimated constant terms and the estimated two principal component functions for $\tau = 0.05$ and $\tau = 0.95$. The figures reveal that i. the estimators for the constant term produced by different algorithms are rather close to each other; ii. the estimators of the principal component functions do not show major difference between the algorithms eighter; iii. BUP and TD principal components are particularly close to each other.

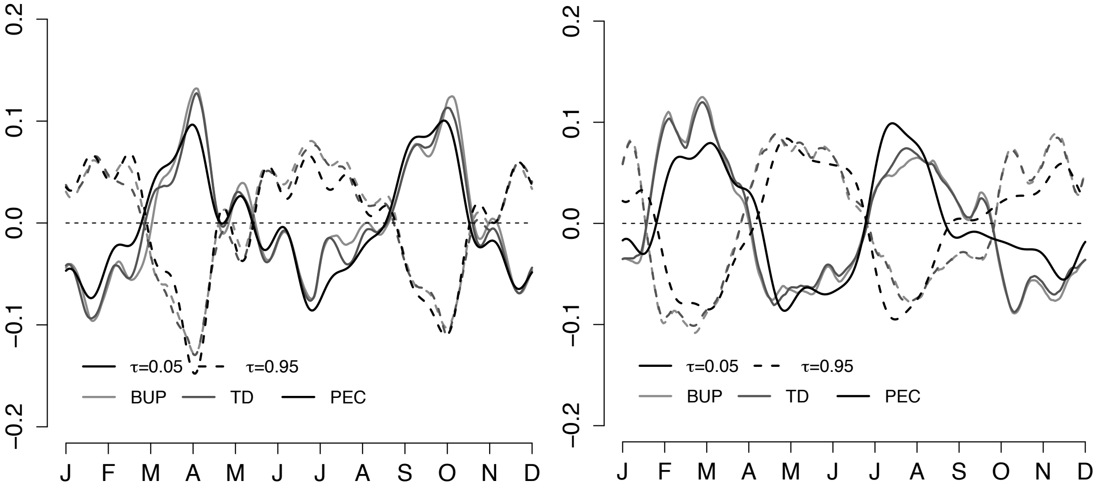

Figure 4: Left: the estimated first component function for the residuals of temperature. Right: the estimated second component function. The horizontal axis of all graphs features the months from January to December.

The obtained first and second components indicate changes in the temperature distribution from lighter to heavier tails and the other way around within a typical year. A positive score on the first component would mean lighter than average tails of the temperature distribution in spring and fall, and heavier than average tails in winter and summer. Similar, a positive score on the second component would indicate lighter than average tails of the temperature distribution in February, March, April, July, August, and September, and heavier than average tails during the rest of the year.

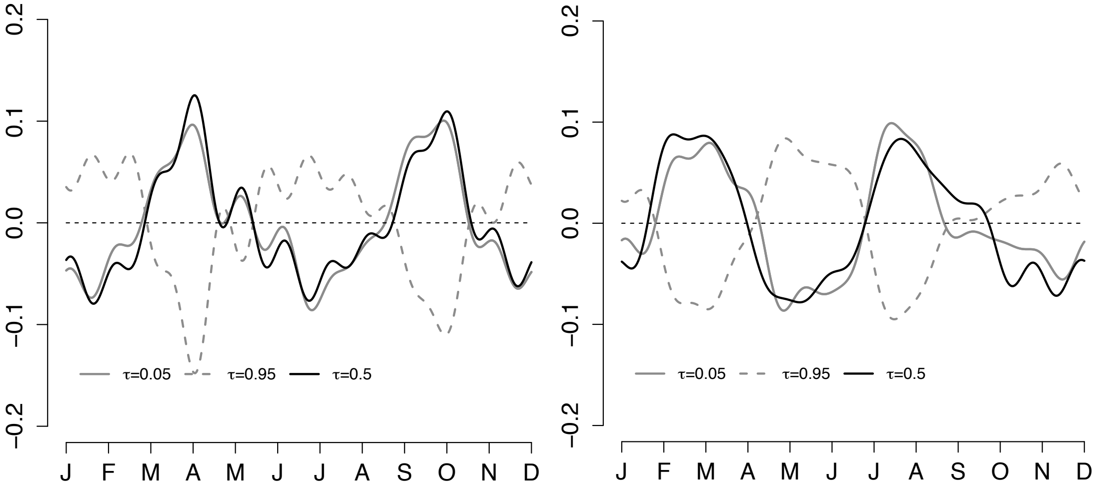

Figure 5: Left panel: the estimated first PEC for $\tau = \{0.05, 0.5, 0.95\}$. Right panel: the estimated second PEC for $\tau = \{0.05, 0.5, 0.95\}$.The horizontal axis of the graphs features the months from January to December abbreviated with the first letter.

		BUP	TD	PEC
$\tau =$	0.05	0.89	0.89	0.86
$\tau =$	0.50	0.82	0.82	0.82
$\tau =$	0.95	0.89	0.90	0.87

Table 5: Proportion of the explained variance by $K = 2$ for different τ-levels and each of the algorithms in the temperature residuals curves

In Figure 5 we show the principal component functions for $\tau = \{0.05, 0.5, 0.95\}$ obtained by PrincipalExpectile. We observe that the estimated principal component functions vary with τ and exhibit differences to the classical PCA where $\tau = 0.5$. By applying Proposition 4.2(3) to PEC, we conclude that the distribution of the considered temperature residuals is rather not an elliptically symmetric one. Thus, the normality assumption for pricing WDs on temperature as needed in the technology presented by Benth et al. (2007) might be violated for this data.

8 Summary

We proposed two definitions of principal components in an asymmetric norm and provided consistent algorithms based on iterative least squares. We derived the upper bounds on their convergence times as well as other useful properties of the resulting principal components in an asymmetric norm.

The algorithms TopDown and BottomUp minimize the projection error in a τ-asymmetric norm, and PrincipalExpectile algorithm maximizes the τ-variance of the low-dimensional projection. The later algorithm was shown to share 'nice' properties of PCA as invariance under translations and changes of basis, moreover, it coincides with classical PCA for elliptically symmetric distributions.

Using simulations we compared finite sample performance of the proposed algorithms. All algorithms appear to produce similar results. Overall performance of PrincipalExpectile and TopDown was very satisfactory in terms of the MSE, PrincipalExpectile showed robustness to 'fat-tails' and skewness of the data distribution.

We applied the algorithms to a Chinese weather dataset with a view to weather derivative pricing. Using a commonly accepted model for temperature of Benth et al. (2007), we estimated the first two principal component functions of the temperature residuals as functions of days of a year. The resulting component functions indicate relative changes in the tails of the temperature distribution from light to heavier and vice versa. Our further results question the validity of the normality assumption on the temperature residuals which is frequently used for pricing temperature based derivatives.

The proposed algorithms appear to be a good way to study extremes of multivariate data. They are easy to compute, relatively fast and their results are easy to interpret.

References

Alaton, P., B. Djehiche, and D. Stillberger (2002): "On Modelling and Pricing Weather Derivatives," *Applied Mathematical Finance*, 1, 1–20.

Benth, F., J. Benth, and S. Koekebakker (2007): "Putting a price on temperature," *Scandinavian Journal of Statistics*, 34, 746–767.

Benth, J. and F. Benth (2012): "A critical view on temperature modelling for application in weather derivatives markets," *Energy Economics*, 34, 592–602.

Campbell, S. and F. Diebold (2005): "Weather forecasting for weather derivatives," *Journal of the American Statistical Association*, 100, 6–16.

Chen, K. and H.-G. Müller (2012): "Conditional Quantile analysis when covariates are functions, with application to growth data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 874, 67–89.

Cobzaş, Ş. (2013): *Functional analysis in asymmetric normed spaces*, Springer.

Crambes, C., A. Kneip, and S. P. (2009): "Smooth splines estimators for functional linear regression," *Annals of Statistics*, 37, 35–72.

Doss, H. and R. Gill (1992): "An Elementary Approach to Weak Convergence for Quantile Processes, With Applications to Censored Survival Data," *Journal of the American Statistical Association*, 87, pp. 869–877.

Fraiman, R. and B. Pateiro-López (2012): "Quantiles for finite and infinite dimensional data," *Journal of Multivariate Analysis*, 108, 1–14.

Guo, M., L. Zhou, W. Härdle, and J. Huang (2013): "Functional Data Analysis for Generalized Quantile Regression," *Statistics and Computing*, doi: 10.1007/s11222-013-9425-1, 1–14.

Härdle, W. and B. López Cabrera (2012): "The Implied Market Price of Weather Risk," *Applied Mathematical Finance*, 19, 59–95.

Hjort, N. and D. Pollard (2011): "Asymptotics for minimisers of convex processes," *arXiv preprint arXiv:1107.3806.*

Jolliffe, I. (2004): *Principal component analysis*, Springer.

Kneip, A. and K. Utikal (2001): "Inference for Density Families Using Functional Principal Component Analysis," *Journal of the American Statistical Association*, 96, 519–532.

Kong, L. and I. Mizera (2012): "Quantile tomography: using quantiles with multivariate data," *Statistica Sinica*, 22, 1589–1610.

Kuan, C.-M., J.-H. Yeh, and Y.-C. Hsu (2009): "Assessing value at risk with CARE, the Conditional Autoregressive Expectile models," *Journal of Econometrics*, 150, 261–270.

Newey, W. and J. Powell (1987): "Asymmetric least squares estimation and testing," *Econometrica*, 819–847.

Ramsay, J. and B. Silverman (2005): *Functional data analysis*, Springer, New York.

Schnabel, S. (2011): "Expectile smoothing: new perspectives on asymmetric least squares. An application to life expectancy," Ph.D. thesis, Utrecht University.

Srebro, N. and T. Jaakkola (2003): "Weighted low-rank approximations," in *Machine Learning International Workshop*, vol. 20, 720.

Taylor, J. (2008): "Estimating Value at Risk and Expected Shortfall Using Expectiles," *Journal of Financial Econometrics*, 6, 231–252.

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, Springer.

SFB 649 Discussion Paper Series 2014

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Principal Component Analysis in an Asymmetric Norm" by Ngoc Mai Tran, Maria Osipenko and Wolfgang Karl Härdle, January 2014.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".