Fast estimation of the median covariation matrix with application to online robust principal components analysis

Hervé Cardot1 · Antoine Godichon-Baggioni1

Received: 6 April 2016 / Accepted: 28 November 2016 / Published online: 8 December 2016 © Sociedad de Estadística e Investigación Operativa 2016

Abstract The geometric median covariation matrix is a robust multivariate indicator of dispersion which can be extended to infinite dimensional spaces. We define estimators, based on recursive algorithms, that can be simply updated at each new observation and are able to deal rapidly with large samples of high-dimensional data without being obliged to store all the data in memory. Asymptotic convergence properties of the recursive algorithms are studied under weak conditions in general separable Hilbert spaces. The computation of the principal components can also be performed online and this approach can be useful for online outlier detection. A simulation study clearly shows that this robust indicator is a competitive alternative to minimum covariance determinant when the dimension of the data is small and robust principal components analysis based on projection pursuit and spherical projections for high-dimension data. An illustration on a large sample and high-dimensional dataset consisting of individual TV audiences measured at a minute scale over a period of 24 h confirms the interest of considering the robust principal components analysis based on the median covariation matrix. All studied algorithms are available in the R package Gmedian on CRAN.

Keywords Functional data · Geometric median · $L_1$-median · Recursive robust estimation · Stochastic gradient

Electronic supplementary material The online version of this article (doi:10.1007/s11749-016-0519-x) contains supplementary material, which is available to authorized users.

Hervé Cardot
herve.cardot@u-bourgogne.fr
Antoine Godichon-Baggioni
antoine.godichon@u-bourgogne.fr

1 Institut de Mathématiques de Bourgogne, Université de Bourgogne Franche-Comté, 9, Rue Alain Savary, 21078 Dijon, France
Mathematics Subject Classification 62G05 · 62L20

1 Introduction

Principal components analysis is one of the most useful statistical tools to extract information by reducing the dimension when one has to analyze large samples of multivariate or functional data (see e.g. Jolliffe 2002; Ramsay and Silverman 2005). When both the dimension and the sample size are large, outlying observations may be difficult to detect automatically. Principal components, which are derived from the spectral analysis of the covariance matrix, can be very sensitive to outliers (see Devlin et al. 1981) and many robust procedures for principal components analysis have been considered in the literature (see Hubert et al. 2008; Huber and Ronchetti 2009; Maronna 2006).

The most popular approaches are probably the minimum covariance determinant estimator (see Rousseeuw and Driessen 1999) and the robust projection pursuit (see Croux and Ruiz-Gazen 2005; Croux et al. 2007). Robust PCA based on projection pursuit has been extended to deal with functional data in Hyndman and Ullah (2007) and Bali et al. (2011). Adopting another point of view, robust modifications of the covariance matrix, based on projection of the data onto the unit sphere, have been proposed in Locantore et al. (1999) (see also Gervini 2008; Taskinen et al. 2012).

We consider in this work another robust way of measuring association between variables that can be extended directly to variables taking values in infinite dimension spaces. For a random vector $X$ with center $m$, the covariance matrix, if it exists, is defined as the expectation $E[(X - m)(X - m)^T]$. Our robust approach is based on the notion of median in matrix spaces and we consider the median of $(X - m)(X - m)^T$, that we call median covariation matrix (MCM), instead of the mean. There are different ways of extending the notion of median in a multidimensional setting (see Small 1990) and we choose to consider the geometric median (also called spatial or L1-median) because it is defined as the solution of a convex optimization problem (see Cardot et al. 2013). As shown in Kraus and Panaretos (2012) the MCM has the same eigenspaces as the usual covariance matrix when the distribution of the data is symmetric and the second-order moment is finite so that considering the MCM to compute principal components can be of interest. Different algorithms can be considered to get effective estimators of the MCM. When the dimension of the data is not too high and the sample size is not too large, Weiszfeld’s algorithm (see Weiszfeld 1937; Vardi and Zhang 2000) can be directly used to estimate effectively both the geometric median and the median covariation matrix. When both the dimension and the sample size are large this static algorithm which requires to store all the data may be inappropriate and ineffective.

The aim of this work is twofold. At first, we show how the stochastic algorithms developed by Cardot et al. (2013) for the geometric median in Hilbert spaces can be adapted to estimate recursively and simultaneously the geometric median as well as the median covariation matrix without being obliged to store all the data. By this recursive property, these algorithms, which are very fast, can naturally deal with data that are observed sequentially and provide a natural update of the estimators at each
new observation. Another advantage compared to classical approaches is that such recursive algorithms will not require to store all the data. Second, this work aims at highlighting the interest of considering the median covariation matrix to perform principal components analysis of large samples of high-dimensional contaminated data.

The paper is organized as follows: the median covariation matrix as well as some of its properties is given in Sect. 2. The recursive estimators of the MCM are defined and some practical issues are discussed in Sect. 3. A simple modification of the algorithms is proposed to ensure that the MCM estimator is positive semi-definite. We explain how the eigenelements of the estimator of the MCM can be updated online without being obliged to perform a new spectral decomposition at each new observation. We also describe how this approach can be useful to treat densely discretized functional data by dealing with their vectors of coordinates in basis functions (such as splines, wavelets or Fourier series). In Sect. 4, almost sure and quadratic mean consistency results are given in the general framework of separable Hilbert spaces. In Sect. 5, a comparison with some classic robust PCA techniques is made on simulated data. The interest of considering the MCM is also highlighted on the analysis of individual TV audiences, a large sample of high-dimensional data which, because of its dimension, cannot be analyzed in a reasonable time with classical robust PCA approaches. Perspectives for future research are discussed in Sect. 6. The proofs as well as a description of Weiszfeld’s algorithm in our context are gathered in a supplementary file.

2 Definitions and framework

2.1 The geometric median

Let \( H \) be a separable Hilbert space (for example \( H = \mathbb{R}^d \) or \( H = L^2(I) \), for some closed interval \( I \subset \mathbb{R} \)). We denote by \( \langle \ldots \rangle \) its inner product and by \( \| \cdot \| \) the associated norm. We consider a random variable \( X \) that takes values in \( H \). The geometric median (also called \( L^1 \)-median or spatial median, see Small 1990; Gervini 2008) \( m \in H \) of \( X \) is defined by

\[
m := \arg \min_{u \in H} \mathbb{E} \left[ \| X - u \| - \| X \| \right].
\]

The point \( m \) is thus such that the mean distance from \( m \) to all the points of the distribution of \( X \) is minimum. It is uniquely defined under broad assumptions on the distribution of \( X \) (see Kemperman 1987) which can be expressed as follows:

**Assumption 1** There exist two linearly independent unit vectors \( (u_1, u_2) \in H^2 \), such that

\[
\text{Var}(\langle u, X \rangle) > 0, \quad \text{for} \quad u \in \{u_1, u_2\}.
\]

If the distribution of \( X - m \) is symmetric around zero and if \( X \) admits a first moment that is finite, then the geometric median is equal to the expectation of \( X \), \( m = \mathbb{E} [X] \).
Note, however, that the general definition (1) does not require to assume that the first-order moment of $\| X \|$ is finite since $|E[\|X - u\| - \|X\|]| \leq \|u\|$. As noted in Lopuhaä and Rousseeuw (1991) and Gervini (2008), the geometric median has a 0.5 breakdown point.

2.2 The (geometric) median covariation matrix (MCM)

We now consider the special vector space, denoted by $S(H)$, of $d \times d$ matrices when $H = \mathbb{R}^d$, or for general separable Hilbert spaces $H$, the vector space of linear operators mapping $H \to H$. Denoting by $\{e_j, j \in J\}$ an orthonormal basis in $H$, the vector space $S(H)$ equipped with the following inner product

$$\langle A, B \rangle_F = \sum_{j \in J} \langle Ae_j, Be_j \rangle$$

is a separable Hilbert space. Remark that in $S(\mathbb{R}^d)$, we have

$$\langle A, B \rangle_F = \text{tr} \left( A^T B \right),$$

where $A^T$ is the transpose matrix of $A$. The induced norm is the well-known Frobenius norm (also called Hilbert–Schmidt norm) and is denoted by $\| \cdot \|_F$.

When $X$ has finite second-order moments, with expectation $E[X] = \mu$, the covariance matrix of $X$, $E[(X - \mu)(X - \mu)^T]$ can be defined as the minimum argument, over all the elements belonging to $S(H)$, of the functional $G_{\mu,2} : S(H) \to \mathbb{R}$, with

$$G_{\mu,2}(\Gamma) = E \left[ \| (X - \mu)(X - \mu)^T - \Gamma \|_F^2 - \| (X - \mu)(X - \mu)^T \|_F^2 \right].$$

Note that in general Hilbert spaces with inner product $\langle ., . \rangle$, operator $(X - \mu)(X - \mu)^T$ should be understood as the operator $u \in H \mapsto \langle u, (X - \mu)(X - \mu) \rangle$. The MCM is obtained by removing the squares in function $G_{\mu,2}$ and by replacing the expectation $\mu$ by the geometric median $m$ to get a more robust indicator of covariation that can be seen as the median of the random variable $Y := (X - m)(X - m)^T$, in the space $S(H)$ equipped with the norm $\| \cdot \|_F$. The (geometric) Median Covariation Matrix of $X$ is thus defined as follows:

$$\Gamma_m := \arg \min_{V \in S(H)} E \left[ \| (X - m)(X - m)^T - V \|_F - \| (X - m)(X - m)^T \|_F \right].$$

The second term at the right-hand side of (4) prevents from having to introduce hypotheses on the existence of the moments of $X$. The MCM is unique provided that the support of $Y$ is not concentrated on a line and Assumption 1 can be rephrased as follows in $S(H)$:
**Assumption 2** There exist two linearly independent unit vectors \((V_1, V_2) \in S(H)^2\), such that

\[
\text{Var}((V, Y)_F) > 0, \quad \text{for } V \in \{V_1, V_2\}.
\]

We can remark that Assumptions 1 and 2 are strongly connected. Indeed, if Assumption 1 holds, then \(\text{Var}((u, X)) > 0\) for \(u \in \{u_1, u_2\}\). Consider the rank one matrices \(V_1 = u_1 u_1^T\) and \(V_2 = u_2 u_2^T\), we have \((V_1, Y)_F = (u_1, X - m)^2\) which has a strictly positive variance when the distribution of \(X\) has no atom. More generally, \(\text{Var}((V_1, Y)_F) > 0\) unless there is a scalar \(a > 0\) such that \(P[\langle u_1, X - m \rangle = a] = P[\langle u_1, X - m \rangle = -a] = \frac{1}{2}\) (assuming also that \(P[X - m = 0] = 0\)).

For \(h \in H\), we define the functional \(G_h : S(H) \rightarrow \mathbb{R}\) by

\[
G_h(V) := \mathbb{E}\left[\left\| (X - h)(X - h)^T - V \right\|^2_F - \left\| (X - h)(X - h)^T - V \right\|^2_F \right].
\]

(5)

Let us now introduce a new technical assumption, which ensures that the functional \(G_h\) is twice Fréchet-differentiable.

**Assumption 3** There is a constant \(C\) such that for all \(h \in H\) and all \(V \in S(H)\)

\[
(a): \quad \mathbb{E}\left[\left\| (X - h)(X - h)^T - V \right\|^{-1}_F \right] \leq C.
\]

\[
(b): \quad \mathbb{E}\left[\left\| (X - h)(X - h)^T - V \right\|^2_F \right] \leq C.
\]

This assumption implicitly forces the distribution of \((X - h)(X - h)^T\) to have no atoms. It is more “likely” to be satisfied when the dimension \(d\) of the data is large (see Chaudhuri 1992; Cardot et al. 2013 for a discussion) and it could be weakened as in Cardot et al. (2013) by allowing points, necessarily different from the MCM \(\Gamma_m\), to have strictly positive masses.

**2.3 Some interesting properties of the MCM**

Let us denote by \(\nabla G_h(V)\) the gradient of functional \(G_h(V)\). It is well defined if Assumption 3(a) holds and we have

\[
\nabla G_h(V) = -\mathbb{E}\left[\frac{(X - h)(X - h)^T - V}{\left\| (X - h)(X - h)^T - V \right\|}_F \right].
\]

(6)

The MCM \(\Gamma_m\) is the unique zero of \(\nabla G_m\), i.e. \(\nabla G_m(\Gamma_m) = 0\), so that it can be implicitly written as follows:

\[
\Gamma_m = \frac{1}{\mathbb{E}\left[\left\| (X - m)(X - m)^T - \Gamma_m \right\|_F \right]} \mathbb{E}\left[\frac{(X - m)(X - m)^T}{\left\| (X - m)(X - m)^T - \Gamma_m \right\|_F \right].
\]

(7)
It clearly appears from (7) that $\Gamma_m$ is a bounded, symmetric and semi-definite positive operator in $S(H)$.

As stated in Proposition 2 of Kraus and Panaretos (2012), the MCM has a remarkable stability property which motivates its uses for performing robust principal components analysis when the distribution of $X$ is symmetric, with finite second moment, i.e $\mathbb{E}\left[\|X\|^2\right] < \infty$. In that case, the covariance operator of $X$, $\Sigma = \mathbb{E}\left[(X - m)(X - m)^T\right]$, which is well defined, and $\Gamma_m$ share the same eigenvectors: if $e_j$ is an eigenvector of $\Sigma$ with corresponding eigenvalue $\lambda_j$, then $\Gamma_m e_j = \tilde{\lambda}_j e_j$, for some non-negative value $\tilde{\lambda}_j$. This important result means that for Gaussian and more generally symmetric distribution (with finite second-order moments), the covariance operator and the median covariation operator have the same eigenspaces. Note that it is also conjectured in Kraus and Panaretos (2012) that the order of the eigenfunctions is also the same.

For a probability measure $F_0$ defined on $H$, we denote by $m(F_0)$ (resp. $\Gamma_m(F_0)$) the corresponding geometric median (resp. median covariation matrix). Consider, for $\epsilon \in [0, 1)$, the $\epsilon$-contamination neighborhood of the probability measure $F_0$,

$$\mathcal{P}_\epsilon(F_0) = \{ F \mid F = (1 - \epsilon)F_0 + \epsilon G, \ G \ \text{a probability measure on} \ H \}.$$

Being spatial medians in the separable Hilbert spaces $H$ and $S(H)$, the geometric median $m(F_0)$ and the MCM, $\Gamma_m(F_0)$, are also robust indicators of central location in $H$ and $S(H)$ with a 0.5 asymptotic breakdown point (see Theorem 4 in Gervini 2008 for a proof in a general Hilbert space framework). This means that, for all $\epsilon \in (0, 1/2)$, there exist $a$ and $b$ such that for all $F \in \mathcal{P}_\epsilon(F_0)$, $\|m(F)\| \leq a$ and $\|\Gamma_m(F)\|_F \leq b$.

Note that in our high-dimensional context the smallest eigenvalue of the covariance matrix can be equal to zero or very close to zero. Moreover, if the dimension of $H$ is not finite and the covariance operator exists, then 0 is necessarily an accumulation point of the eigenvalues (because their sum is finite). Consequently, breakdown point criterions that have been defined in the literature to deal with covariance matrices and that also take account of the perturbation of the smallest eigenvalues (see e.g. Lopuhaä and Rousseeuw 1991) are not adapted to our high-dimensional framework.

### 3 The recursive algorithms

#### 3.1 Averaged stochastic gradient algorithms

We suppose now that we have i.i.d. copies $X_1, \ldots, X_n, \ldots$ of random variables with the same law as $X$.

For simplicity, we temporarily suppose that the median $m$ of $X$ is known. We consider a sequence of (learning) weights $\gamma_n = c_\gamma/n^\alpha$, with $c_\gamma > 0$ and $1/2 < \alpha < 1$ (the choice of values for these tuning parameters is discussed in Sect. 3.4) and we define the recursive estimation procedure as follows:
Fast estimation of the median covariation... 

\[
W_{n+1} = W_n + \gamma_n \frac{(X_{n+1} - m)(X_{n+1} - m)^T - W_n}{\| (X_{n+1} - m)(X_{n+1} - m)^T - W_n \|_F}
\]

\[
\overline{W}_{n+1} = \overline{W}_n - \frac{1}{n+1} (\overline{W}_n - W_{n+1})
\]

(8)

This algorithm can be seen as a particular case of the averaged stochastic gradient algorithm studied in Cardot et al. (2013) in the general context of Hilbert spaces. Indeed, the first recursive algorithm (8) is a stochastic gradient algorithm since the direction given at the right-hand side of (8) is an approximation to the gradient of functional \( G_m \),

\[
\mathbb{E} \left[ \frac{(X_{n+1} - m)(X_{n+1} - m)^T - W_n}{\| (X_{n+1} - m)(X_{n+1} - m)^T - W_n \|_F} \bigg| \mathcal{F}_n \right] = \nabla G_m(W_n),
\]

where \( \mathcal{F}_n = \sigma(X_1, \ldots, X_n) \) is the \( \sigma \)-algebra generated by \( X_1, \ldots, X_n \). The final estimator \( \overline{W}_n \) is obtained by averaging the past values of the first algorithm. The averaging step (see Polyak and Juditsky 1992), i.e. the computation of the arithmetical mean of the past values of a slowly convergent estimator, permits to obtain a new and efficient estimator converging at a parametric rate, with the same asymptotic variance as the empirical risk minimizer.

In most of the cases the value of \( m \) is unknown so that it is also required to estimate the median. To build an estimator of the MCM, we suggest to estimate simultaneously \( m \) and \( \Gamma_m \) by considering two averaged stochastic gradient algorithms that are running simultaneously. For \( n \geq 1 \),

\[
m_{n+1} = m_n + \gamma_n^{(m)} \frac{X_{n+1} - m_n}{\| X_{n+1} - m_n \|}
\]

(9)

\[
\overline{m}_{n+1} = \overline{m}_n - \frac{1}{n+1} (\overline{m}_n - m_{n+1})
\]

(10)

\[
V_{n+1} = V_n + \gamma_n \frac{(X_{n+1} - \overline{m}_n)(X_{n+1} - \overline{m}_n)^T - V_n}{\| (X_{n+1} - \overline{m}_n)(X_{n+1} - \overline{m}_n)^T - V_n \|_F}
\]

(11)

\[
\overline{V}_{n+1} = \overline{V}_n - \frac{1}{n+1} (\overline{V}_n - V_{n+1})
\]

(12)

where the averaged recursive estimator \( \overline{m}_{n+1} \) (resp. \( \overline{V}_{n+1} \)) of the median \( m \) (resp. of the median covariation matrix \( \Gamma_m \)) are controlled by sequences of descent steps \( \gamma_n^{(m)} \) and \( \gamma_n \) which are chosen to satisfy

\[
\gamma_n := c_{\gamma} n^{-\alpha}, \quad \gamma_n^{(m)} := c_{m} n^{-\alpha}, \quad \text{with} \quad c_{m} > 0, \ c_{\gamma} > 0, \ \text{and} \ \alpha \in \left( \frac{1}{2}, 1 \right).
\]

(13)

The practical choice of the values of the tuning parameters \( c_{m}, c_{\gamma} \) and \( \alpha \) is discussed in Sect. 3.4. Note that by construction, even if \( V_n \) is positive semi-definite, \( V_{n+1} \) may not be a positive semi-definite matrix when the learning steps do not satisfy the constraint...
Projecting $V_{n+1}$ onto the closed convex cone of positive semi-definite matrices would require to compute the eigenvalues of $V_{n+1}$ which is time consuming in high dimension even if $V_{n+1}$ is a rank one perturbation to $V_n$ (see Gu and Eisenstat 1994). We consider the following simple approximation to this projection which consists in replacing in (11) the descent step $\gamma_n$ by a thresholded one,

$$\gamma_{n, \text{pos}} = \min \left( \gamma_n, \left\| (X_{n+1} - \bar{m}_n)(X_{n+1} - \bar{m}_n)^T - V_n \right\|_F \right)$$

which ensures that $V_{n+1}$ remains positive semi-definite when $V_n$ is positive semi-definite. The use of this non linear modification of the descent steps as well as an initialization of the recursive algorithm (11) with a positive semi-definite matrix (for example $V_0 = 0$) ensures that for all $n \geq 1$, $V_n$ and $\bar{V}_n$ are positive semi-definite matrices. Remark that this choice of learning rates do not have any influence on the asymptotic behavior of the algorithm since, with Assumption 3(b) and Markov’s inequality, we have $\mathbb{P} \left[ \gamma_n \neq \gamma_{n, \text{pos}} \right] \leq C \gamma_n^2$. For more details, see Remark B.5 in the Supplementary file.

### 3.2 Dealing with functional data

In many applications, the data are functions of a continuous argument (e.g., time or frequencies) observed on a dense grid of points $0 \leq t_1 < \cdots < t_p \leq 1$. The observations are $X_i = (X_i(t_1), \ldots, X_i(t_p)) \in \mathbb{R}^p$, $i = 1, \ldots, n$, with $p$ potentially very large. Reducing the dimension by approximating the $p$ dimensional discretized observations by their projections onto $d$ basis functions, with $d$ much smaller than $p$, allows important gains in storage of the data, computation time and eventually statistical accuracy.

If the underlying random function $X_i(t)$, $t \in [0, 1]$ is smooth, it can be advantageous (see, e.g. Ramsay and Silverman 2005) to consider an approximation $\tilde{X}_i$ of $X_i$ in a low $d$-dimensional functional space, with $d < p$, span by $d$ smooth basis functions, $B_1, \ldots, B_d$, chosen by the statistician (such as splines or Fourier basis),

$$X_i(t) \approx \tilde{X}_i(t) = \sum_{j=1}^d \beta_{ij} B_j(t), \quad t \in [0, 1].$$

Let us denote by $\beta_i = (\beta_{i1}, \ldots, \beta_{id}) \in \mathbb{R}^d$ the vector of coordinates of $\tilde{X}_i$. As explained in Ramsay and Silverman (2005), linear smoothers are generally employed and in that case, there exists a $d \times p$ matrix $A$ such that $\beta_i = A X_i$, $i = 1, 2, \ldots$

---

1 In this subsection, vectors and matrices will be denoted with bold symbols and letters to make a clear distinction with functions and operators.
Introducing the $d \times d$ Gram matrix $M$ whose elements are $[M]_{jk} = \langle B_j, B_k \rangle$, for $j, k = 1, \ldots, d$, we have that $\| \tilde{X}_i \|_H = \| \beta_i \|_M$ with $\| \beta_i \|^2_M = \beta_i^T M \beta_i$. Consequently, looking for the geometric median of $\tilde{X}_i$, $i = 1, \ldots$ in the Hilbert space $H$ is equivalent to looking for the geometric median of $\beta_i$, $i = 1, \ldots$ in the Euclidean space $\mathbb{R}^d$ equipped with the metric $M$. To use classical softwares, one can consider orthonormal basis so that the classical Euclidean norm can be used and a vector of basis orthonormal functions $\phi_1, \ldots, \phi_d$ is obtained as follows $(\phi_1(t), \ldots, \phi_d(t)) = M^{-1/2}(B_1(t), \ldots, B_d(t))$. The vector of coordinates of $\tilde{X}_i$ in the basis $\phi_1, \ldots, \phi_d$ is simply $\mathbf{M}^{1/2} \beta_i$.

Similarly, the Frobenius norm can be computed in a matrix way. Let us consider the $d^2$ basis functions $\phi_{\ell k}$, for $\ell, k \in \{1, \ldots, d\}$, defined as follows, in the space $S(H)$. For every $\phi \in H$ we define $\phi_{\ell k} \phi = \langle \phi_{\ell k}, \phi \rangle \phi_k$. It is an orthonormal family in $S(H)$, since $\langle \phi_{\ell k}, \phi_{\ell' k'} \rangle_F = \langle \phi_k, \phi_k' \rangle \sum_{j=1}^d \langle \phi_j, \phi_{\ell} \rangle \langle \phi_j, \phi_{\ell'} \rangle = 1$ if $k' = k$ and $\ell' = \ell$ and 0 otherwise. Consequently the Frobenius norm of an element $\Delta \in S(H)$ with coordinates $\delta_{j k}$ is simply $\| \Delta \|_F = \sqrt{\text{tr}(\Delta^T \Delta)}$ where is the $d \times d$ matrix with generic elements $\delta_{\ell k} = \langle \Delta, \phi_{\ell k} \rangle_F = \langle \Delta \phi_{\ell}, \phi_k \rangle$, $\ell, k = 1, \ldots, d$. For example, $\| \tilde{X}_i \tilde{X}_i^T \|_F = \text{tr}(\mathbf{M}^{1/2} \beta_i \beta_i^T \mathbf{M}^{1/2})$.

Consequently, one can perform simply the iterative procedure described in (9)–(12) in the space of coordinates with classical norms in $\mathbb{R}^d$ and $S(\mathbb{R}^d)$ and then go back to the functional spaces thanks to the orthonormal basis $\phi_1, \ldots, \phi_d$ for the median function and $\{\phi_{\ell k}, k, \ell = 1, \ldots, d\}$ for the median covariation matrix.

### 3.3 Online estimation of the principal components

It is also possible to approximate recursively the $q$ eigenvectors (unique up to sign) of $\Gamma_m$ associated with the $q$ largest eigenvalues without being obliged to perform a spectral decomposition of $\bar{V}_{n+1}$ at each new observation. Many recursive strategies can be employed (see Cardot and Degras 2015 for a review on various recursive estimation procedures of the eigenelements of a covariance matrix). Because of its simplicity and its accuracy, we suggest using the following recursive algorithm:

$$u_{j,n+1} = u_{j,n} + \frac{1}{n+1} \left( \bar{V}_{n+1} \frac{u_{j,n}}{\| u_{j,n} \|} - u_{j,n} \right), \quad j = 1, \ldots, q,$$

combined with an orthogonalization by deflation of $u_{1,n+1}, \ldots, u_{q,n+1}$. This recursive algorithm is based on ideas developed by Weng et al. (2003) that are related to the power method for extracting eigenvectors. If we assume that the $q$ first eigenvalues $\lambda_1 > \cdots > \lambda_q$ are distinct, the estimated eigenvectors $u_{1,n+1}, \ldots, u_{q,n+1}$, which are uniquely determined up to sign change, tend to $\lambda_1 u_1, \ldots, \lambda_q u_q$.

Once the eigenvectors are computed, it is possible to compute the principal components as well as indices of outlyingness for each new observation (see Hubert et al. 2008 for a review of outliers detection with multivariate approaches).
3.4 Complexity, memory and choice of the learning steps

The recursive algorithms (11) and (12) require each $O(d^2)$ elementary operations at each update. With the additional online estimation given in (16) of the $q$ eigenvectors associated with the $q$ largest eigenvalues, $O(qd^2)$ additional operations are required. The orthogonalization procedure only requires $O(q^2d)$ elementary operations.

Note that the use of classical Newton–Raphson algorithms for estimating the MCM (see Fritz et al. 2012) cannot be envisaged for high-dimensional data since the computation or the approximation of the Hessian matrix would require $O(d^4)$ elementary operations. The well-known and fast Weiszfeld’s algorithm requires $O(nd^2)$ elementary operations for each sample with size $n$. However, the estimation cannot be updated automatically if the data arrive sequentially. Another drawback compared to the recursive algorithms studied in this paper is that all the data must be stored in memory, which is of order $O(nd^2)$ elements whereas the recursive techniques developed here only require an amount of memory of order $O(d^2)$.

The performances of the recursive algorithms depend on the sequence of learning steps $\gamma_n^{(m)}$ and $\gamma_n$ which are generally parametrized as in (13). The value of parameter $\alpha$, which should be in $(1/2, 1)$ is often chosen to be $\alpha = 2/3$ or $\alpha = 3/4$. If the values of $c_m$ and $c_\gamma$ are too small, the algorithms may get stuck close to their initial values and that is not desired. On the other hand, large values of $c_m$ and $c_\gamma$ allow the algorithms to have large oscillations and consequently a better exploration of the parameter space. Previous empirical studies (see Cardot et al. 2010, 2013) have shown that, thanks to the averaging step, estimator $\overline{m}_n$ performs well and is not too sensitive to the choice of $c_m$, provided that the value of $c_m$ is not too small. An intuitive explanation could be that here the recursive process is in some sense “self-normalized” since the deviations at each iteration in (8) have unit norm and finding some universal values for $c_m$ seems to be possible. Usual values for $c_m$ and $c_\gamma$ are in the interval (Bosq 2000; Kemperman 1987) and in all the computations made in this work we have considered $c_m = c_\gamma = 2$ and $\alpha = 3/4$ (see also Sect. 5).

4 Asymptotic properties

We state in this Section some convergence properties of the recursive estimators. All the proofs are in a supplementary files.

As noted before, when $m$ is known, $\overline{W}_n$ can be seen as an averaged stochastic gradient estimator of the geometric median of $Y = (X - m)(X - m)^T$ in the particular Hilbert space $S(H)$ equipped with the Frobenius norm. The asymptotic weak convergence of such an estimator has been studied in Cardot et al. (2013) in the general framework of Hilbert spaces, while non asymptotic bounds for the deviation of $\overline{W}_n$ around $I_m$ can be derived with the results given in Cardot et al. (2016) and finally $L^p$ rates of convergence can be deduced from Godichon-Baggioni (2016).

The more realistic case in which $m$ is unknown is more complicated because $\overline{V}_n$ depends on $\overline{m}_n$ which is also estimated recursively with the same data. We first state the strong consistency of the estimators $\overline{V}_n$ and $\overline{V}_n$.

**Theorem 1** If Assumptions 1–3(b) hold and the learning sequences $(\gamma_n, \gamma_n^{(m)})$ satisfy (13), we have
Fast estimation of the median covariation...

\[ \lim_{n \to \infty} \| V_n - \Gamma_m \|_F = 0 \text{ a.s.}, \]

and

\[ \lim_{n \to \infty} \| V_n - \Gamma_m \|_F = 0 \text{ a.s.}. \]

The obtention of the rate convergence of the averaged recursive algorithm relies on a fine control of the asymptotic behavior of the Robbins–Monro algorithms, as stated in the following theorem:

**Theorem 2** If Assumptions 1–3(b) hold and the learning sequences \((\gamma_n), (\gamma_n^{(m)})\) satisfy (13), there is a positive constant \(C'\), and for all \(\beta \in (\alpha, 2\alpha)\), there is a positive constant \(C_\beta\) such that for all \(n \geq 1\),

\[ E \left[ \left\| V_n - \Gamma_m \right\|^2_F \right] \leq \frac{C'}{n^{\alpha}}, \]

\[ E \left[ \left\| V_{n+1} - \Gamma_m \right\|^4_F \right] \leq \frac{C_\beta}{n^{\beta}}. \]

The obtention of an upper bound for the rate of convergence at the order four of the Robbins–Monro algorithm is crucial in the proofs. Furthermore, the following result ensures that the exhibited rate in quadratic mean is the optimal one:

**Theorem 3** If Assumptions 1–3(b) hold and the learning sequences \((\gamma_n), (\gamma_n^{(m)})\) satisfy (13), there is a positive constant \(c'\) such that for all \(n \geq 1\),

\[ E \left[ \left\| V_n - \Gamma_m \right\|^2_F \right] \geq \frac{c'}{n^{\alpha}}. \]

Finally, the following theorem is the most important theoretical result of this work. It shows that, in spite of the fact that it only considers the observed data one by one, the averaged recursive estimation procedure gives an estimator which has a classical parametric \(1/n\) rate of convergence in the Hilbert–Schmidt norm.

**Theorem 4** If Assumptions 1–3(b) hold and the learning sequences \((\gamma_n), (\gamma_n^{(m)})\) satisfy (13), there is a positive constant \(K'\) such that for all \(n \geq 1\),

\[ E \left[ \left\| V_n - \Gamma_m \right\|^2_F \right] \leq \frac{K'}{n}. \]

Assuming the eigenvalues of \(\Gamma_m\) are of multiplicity one, it can be deduced from Theorem 4 and Lemma 4.3 in Bosq (2000) the convergence in quadratic mean of the eigenvectors of \(V_n\) towards the corresponding (up to sign) eigenvector of \(\Gamma_m\).

Although the only difference between the sequence of estimators \((W_n)\) and \((V_n)\) is that the median \(m\) is replaced by an estimation \(\overline{m}_n\), the obtention of Theorem 4 is far from being direct because of the non linearity of the problem. The proofs, which are based on new induction steps compared to Cardot et al. (2013), allow to get better convergence rates in quadratic mean (the log \(n\) factor appearing in their Proposition 3.2...
is now replaced by a constant $K'$) even if this new framework is much more complicated because two averaged non-linear algorithms are running simultaneously. One can also note that the techniques generally employed to deal with two time scale Robbins Monro algorithms (see Mokkadem and Pelletier 2006 for the multivariate case) require assumptions on the rest of the Taylor expansion and the finite dimension of the data that are too restrictive in our framework (see the proofs in the supplementary file).

5 An illustration on simulated and real data

A small comparison with other classical robust PCA techniques is performed in this section considering data in relatively high dimension but samples with moderate sizes. This permits to compare our approach with classical robust PCA techniques, which are generally not designed to deal with large samples of high-dimensional data. In our comparison, we have employed the following well-known robust techniques: robust projection pursuit (see Croux and Ruiz-Gazen 2005; Croux et al. 2007) and denoted by RPP in the following, minimum covariance determinant (see Rousseeuw and Driessen 1999) denoted by MCD, and spherical PCA (see Locantore et al. 1999), denoted by sphPCA. The computations were made in the R language (R Development Core Team 2010), with the help of packages pcaPP and rrcov. For reproducible research, our codes for computing the MCM have been posted on CRAN in the Gmedian package. We will denote by MCM(R) the recursive estimator $\hat{V}_n$ defined in (12) and by MCM(R+) its positive semi-definite modification whose learning weights are defined in (14). These algorithms depend on tuning parameters $c_m$ and $c_\gamma$ and $\alpha$ that have not been particularly optimized in this simulation study. We have noted that the simulation results were very stable and did not depend much on the value of $c_m$ and $c_\gamma$ for $c_m, c_\gamma \in (Bali et al. 2011; Kemperman 1987)$. The values that we have chosen are the default values of the Gmedian package, $\alpha = 3/4$ and $c_m = c_\gamma = 2$.

If the size of the data $n \times d$ is not too large, an effective way for estimating $\Gamma_m$ is to employ Weiszfeld’s algorithm (see Weiszfeld 1937; Vard and Zhang 2000 as well the Supplementary file for a description of the algorithm in our particular situation). The estimate obtained thanks to Weiszfeld’s algorithm is denoted by MCM(W) in the following. Note that other optimization algorithms which may be preferred in small dimension (see Fritz et al. 2012) have not been considered here since they would require the computation of the Hessian matrix, whose size is $d^4$, and this would lead to much slower algorithms. Note that, except for our recursive estimators MCM(R) and MCM(R+), all the robust methods considered here require to store all the data in memory and do not admit a natural updating scheme when the data arrive sequentially so that they should be completely ran again at each new observation.

5.1 Simulation protocol

Independent realizations of a random variable $X_{\text{obs}} \in \mathbb{R}^d$ are drawn, where

$$X_{\text{obs}} = (1 - O(\delta))X + O(\delta)\epsilon,$$

(17)
is a mixture of two distributions and \( X, O \) and \( \epsilon \) are independent random variables. The random vector \( X \) has a centered Gaussian distribution in \( \mathbb{R}^d \) with covariance matrix \([\Sigma]\ell,j = \min(\ell, j)/d\) and can be thought as a discretized version of a Brownian sample path in \([0, 1]\). The multivariate contamination comes from \( \epsilon \), with different rates of contamination controlled by the Bernoulli variable \( O(\delta) \), independent from \( X \) and \( \epsilon \), with \( P(O(\delta) = 1) = \delta \) and \( P(O(\delta) = 0) = 1 - \delta \). Three different scenarios are considered for the distribution of the contamination \( \epsilon \):

- The elements of vector \( \epsilon \) are \( d \) independent realizations of a Student \( t \) distribution with one degree of freedom. This means that the first moment of \( X_{\text{obs}} \) is not defined when \( \delta > 0 \).
- The elements of vector \( \epsilon \) are \( d \) independent realizations of a Student \( t \) distribution with two degrees of freedom. This means that the second moment of \( X_{\text{obs}} \) is not defined when \( \delta > 0 \).
- The vector \( \epsilon \) is distributed as a “reverse time” Brownian motion. It has a Gaussian centered distribution, with covariance matrix \([\Sigma_\epsilon]\ell,j = 2\min(d - \ell, d - j)/d\). The covariance matrix of \( X_{\text{obs}} \) is \((1 - \delta)\Sigma + \delta\Sigma_\epsilon\).

The estimation error of the eigenspaces associated with the largest eigenvalues is evaluated by considering the squared Frobenius norm between the associated orthogonal projectors. Denoting by \( P_q \) the orthogonal projector onto the space generated by the \( q \) eigenvectors of the covariance matrix \( \Sigma \) associated with the \( q \) largest eigenvalues and by \( \widehat{P}_q \) an estimation, we consider the following loss criterion,

\[
R(\widehat{P}_q, P_q) = \text{tr}\left[ (\widehat{P}_q - P_q)^T (\widehat{P}_q - P_q) \right] = 2q - 2\text{tr}[\widehat{P}_q P_q].
\]

(18)

Smaller values of \( R(\widehat{P}_q, P_q) \) indicate better estimations. Note that we always have \( R(\widehat{P}_q, P_q) \leq 2q \) and \( R(\widehat{P}_q, P_q) = 2q \) means that the eigenspaces generated by the true and the estimated eigenvectors are orthogonal.

5.2 Comparison with classical robust PCA techniques

We first compare the performances of the estimators of the MCM based on the Weiszfeld’s algorithm, MCM(W), and the recursive algorithms, MCM(R) and MCM(R+), with the classical robust PCA techniques described before. We also make a comparison with an “oracle” estimate of the covariance matrix based on the non contaminated \( X \) data (see (17)).

We generated samples of \( X_{\text{obs}} \) with size \( n = 500 \) and dimension \( d \in \{50, 200\} \), over 500 replications. Different levels of contamination are considered: \( \delta \in \{0, 0.02, 0.05, 0.10, 0.20\} \). For both dimensions \( d = 50 \) and \( d = 200 \), the first eigenvalue of the covariance matrix of \( X \) represents about 81% of the total variance, and the second one about 9%.

We first examine the performances of the different approaches when there is no contamination. The median errors of estimation of the eigenspace generated by the first two eigenvectors \( (q = 2) \), according to criterion (18), are given in Table 1 for non contaminated data \( (\delta = 0) \). As expected, the “Oracle”, which is the classical PCA
Table 1  Median estimation errors, according to criterion $R(\hat{P}_q, P_q)$ with a dimension $q = 2$, for non contaminated samples of size $n = 200$, over 500 Monte Carlo experiments.

| Method         | $d = 50$ | $d = 200$ |
|----------------|---------|---------|
|                | $t_1$ $df$ | $t_2$ $df$ | inv. B. | $t_1$ $df$ | $t_2$ $df$ | inv. B. |
| PCA            | 0.0156  | 0.0199  |
| MCD            | 0.0208  | 0.0211  |
| MCM(W)         | 0.0243  | 0.0287  |
| MCM(R)         | 0.0246  | 0.0275  |
| SphPCA         | 0.0200  | 0.0209  |
| RPP            | 0.0220  | 0.0246  |

Table 2  Median estimation errors, according to criterion $R(\hat{P}_q, P_q)$ with a dimension $q = 2$, for datasets with a sample size $n = 200$, over 500 Monte Carlo experiments.

| $\delta$ | Method         | $d = 50$ | $d = 200$ |
|----------|----------------|---------|---------|
|          |                | $t_1$ $df$ | $t_2$ $df$ | inv. B. | $t_1$ $df$ | $t_2$ $df$ | inv. B. |
| 2%       | PCA            | 3.13    | 1.04    | 0.698  | 3.95    | 1.87    | 0.731  |
|          | RPP            | 0.086   | 0.097   | 0.090  | 0.085   | 0.094   | 0.084  |
|          | MCD            | 0.022   | 0.021   | 0.021  | –       | –       | –      |
|          | Sph. PCA       | 0.028   | 0.029   | 0.027  | 0.027   | 0.030   | 0.028  |
|          | MCM (Weiszfeld)| 0.021   | 0.021   | 0.021  | 0.021   | 0.022   | 0.022  |
|          | MCM (R+)       | 0.022   | 0.022   | 0.024  | 0.023   | 0.023   | 0.025  |
|          | MCM (R)        | 0.026   | 0.025   | 0.027  | 0.026   | 0.027   | 0.028  |
| 5%       | PCA            | 3.82    | 1.91    | 0.862  | 3.96    | 1.98    | 0.910  |
|          | RPP            | 0.090   | 0.103   | 0.093  | 0.089   | 0.098   | 0.087  |
|          | MCD            | 0.022   | 0.023   | 0.021  | –       | –       | –      |
|          | Sph. PCA       | 0.029   | 0.031   | 0.033  | 0.029   | 0.031   | 0.034  |
|          | MCM (Weiszfeld)| 0.023   | 0.023   | 0.028  | 0.022   | 0.023   | 0.030  |
|          | MCM (R+)       | 0.025   | 0.024   | 0.035  | 0.024   | 0.024   | 0.039  |
|          | MCM (R)        | 0.029   | 0.027   | 0.037  | 0.028   | 0.028   | 0.040  |
| 10%      | PCA            | 3.83    | 1.96    | 1.03   | 3.96    | 1.99    | 1.10   |
|          | RPP            | 0.107   | 0.108   | 0.099  | 0.088   | 0.101   | 0.097  |
|          | MCD            | 0.023   | 0.022   | 0.023  | –       | –       | –      |
|          | Sph. PCA       | 0.033   | 0.033   | 0.054  | 0.031   | 0.033   | 0.057  |
|          | MCM (Weiszfeld)| 0.025   | 0.026   | 0.059  | 0.023   | 0.024   | 0.056  |
|          | MCM (R+)       | 0.030   | 0.027   | 0.089  | 0.027   | 0.027   | 0.086  |
|          | MCM (R)        | 0.035   | 0.032   | 0.088  | 0.032   | 0.031   | 0.086  |
| 20%      | PCA            | 3.84    | 2.02    | 1.19   | 3.96    | 2.01    | 1.25   |
|          | RPP            | 0.110   | 0.135   | 0.138  | 0.091   | 0.122   | 0.137  |
|          | MCD            | 0.025   | 0.026   | 0.026  | –       | –       | –      |
|          | Sph. PCA       | 0.037   | 0.038   | 0.140  | 0.034   | 0.037   | 0.150  |
|          | MCM (Weiszfeld)| 0.030   | 0.030   | 0.174  | 0.026   | 0.028   | 0.181  |
|          | MCM (R+)       | 0.044   | 0.036   | 0.255  | 0.038   | 0.032   | 0.256  |
|          | MCM (R)        | 0.050   | 0.041   | 0.251  | 0.042   | 0.037   | 0.256  |

in this situation, provides the best estimations of the eigenspaces. When the dimension is not large ($d = 50$), the MCD and the median covariation matrix, estimated by Weiszfeld’s algorithm or by the modified MCM(R+) recursive estimator, behave

 Springer
well and similarly. When the dimension gets larger \((d = 200)\), the MCD cannot be used anymore and the algorithms MCM(W) and MCM(R+) provide most effective estimations of the eigenspaces. In both situations, the robust projection pursuit is the less efficient approach.

When the data are contaminated, the median errors of estimation of the eigenspace generated by the first two eigenvectors \((q = 2)\), according to criterion (18), are given in Table 2. In Fig. 1, the distribution of the estimation error \(R(\hat{P}_q, P_q)\) is drawn for the different approaches.

We can make the following remarks: at first note that even when the level of contamination is small (2 and 5%), the performances of classical PCA are strongly affected by the presence of outlying values in such (large) dimensions. When \(d = 50\), the MCD algorithm and the MCM estimation provide the best estimations of the original two-dimensional eigenspace, whereas when \(d\) gets larger (\(d = n = 200\)), the MCD estimator cannot be used anymore (by construction) and the MCM estimators, obtained with Weiszfeld’s and the semi-positive definite recursive algorithm, remain the most accurate. We can also remark that the recursive MCM algorithms, which are designed to deal with very large samples, perform well even for such moderate sample sizes (see also Fig. 1). The modification of the descent step suggested in (14), which corresponds to estimator MCM(R+), permits to improve the accuracy the initial MCM estimator, specially when the noise level is not small. The performances of the spherical PCA are slightly less accurate whereas the median error of the robust projection pursuit is always the largest among the robust estimators. When the contamination is highly structured temporally and the level of contamination is not small (contamination by a reverse time Brownian motion, with \(\delta = 0.20\)), the behavior of
5.3 Online estimation of the principal components

We now consider an experiment in high dimension, $d = 1000$, and evaluate the ability of the recursive algorithm defined in (16) to estimate recursively the eigenvectors of $\Gamma_m$ associated to the largest eigenvalues. Note that due to the high dimension of the data and limited computation time, we only make comparison of the recursive robust techniques with the classical PCA. For this we generate growing samples and compute for each sample size the approximation error of the different (fast) strategies to the true eigenspace generated by the $q$ eigenvectors associated with the $q$ largest eigenvalues of $\Gamma_m$.

We have drawn in Fig. 2, the evolution of the mean (over 100 replications) approximation error $R(\hat{P}_q, P_q)$, for a dimension $q = 3$, as a function of the sample size for samples contaminated by a 2 degrees of freedom Student $t$ distribution with a rate $\delta = 0.1$. An important fact is that the recursive algorithm which approximates recursively the eigenelements behaves very well and we can see nearly no difference between the spectral decomposition of $\overline{\nabla}_n$ (denoted by MCM in Fig. 2) and the estimates produced with the sequential algorithm (16) for sample sizes larger than a few
hundreds. We can also note that the error made by the classical PCA is always very high and does not decrease with the sample size.

### 5.4 Robust PCA of TV audience

The last example is a high dimension and large sample case. Individual TV audiences are measured, by the French company Médiamétrie, every minute for a panel of $n = 5422$ people over a period of 24 h, $d = 1440$ (see Cardot et al. 2012 for a more detailed presentation of the data). With a classical PCA, the first eigenspace represents 24.4% of the total variability, whereas the second one reproduces 13.5% of the total variance, the third one 9.64% and the fourth one 6.79%. Thus, more than 54% of the variability of the data can be captured in a four-dimensional space. Taking account of the large dimension of the data, these values indicate a high temporal correlation.

Because of the large dimension of the data, the Weiszfeld’s algorithm as well as the other robust PCA techniques cannot be used anymore in a reasonable time with a personal computer. The MCM has been computed thanks to the recursive algorithm given in (12) in approximately 3 min on a laptop in the R language (without any specific C routine).

As seen in Fig. 3, the first two eigenvectors obtained by a classical PCA and the robust PCA based on the MCM are rather different. This is confirmed by the relatively large distance between the two corresponding eigenspaces, $R(\hat{p}_{2}^{PCA}, \hat{p}_{2}^{MCM}) = 0.56$. The first robust eigenvector puts the stress on the time period comprised between 1000 and 1200 min, whereas the first non robust eigenvector focuses, with a smaller intensity, on a larger period of time comprised between 600 and 1200 min. The second robust
eigenvector differentiates between people watching TV during the period between 890 and 1050 min (negative value of the second principal component) and people watching TV between minutes 1090 and 1220 (positive value of the second principal component). Rather surprisingly, we have found that the third and fourth eigenvectors of the non robust and robust covariance matrices look quite similar (Fig. 4).

6 Concluding remarks

The simulation study and the illustration on real data indicate that performing robust principal components analysis via the median covariation matrix, which can bring new information compared to classical PCA, is a relevant alternative to more classical robust principal components analysis techniques. The use of recursive algorithms permits to perform robust PCA on very large datasets, in which outlying observations may be hard to detect. Another interest of the use of such sequential algorithms is that estimation of the median covariation matrix as well as the principal components can be performed online with automatic update at each new observation and without being obliged to store all the data in memory.

A deeper study of the asymptotic behavior of the recursive algorithms would certainly deserve further investigations. Proving the asymptotic normality and obtaining the limiting variance of the sequence of estimators $\overline{V}_n$ when $m$ is unknown would be of great interest. This is a challenging issue that is beyond the scope of the paper and would require to study the joint weak convergence of the two simultaneous recursive averaged estimators of $m$ and $\Gamma_m$. The use of the MCM could be interesting to robustify the estimation in many different statistical models, particularly with functional data.
For example, it could be employed as an alternative to robust functional projection pursuit in robust functional time series prediction or for robust estimation in functional linear regression, with the introduction of the median cross-covariation matrix.

Acknowledgements We thank the two anonymous referees for their comments and suggestions that helped us to improve the presentation of the paper. We thank the company Médiamétrie for allowing us to illustrate our methodologies with their data. We also thank Dr. Peggy Cénac for a careful reading of the proofs.

References

Bali J-L, Boente G, Tyler D-E, Wang J-L (2011) Robust functional principal components: a projection-pursuit approach. Ann Stat 39:2852–2882
Bosq D (2000) Linear processes in function spaces, vol 149. Lecture notes in statistics, theory and applications, Springer, New York,
Cardot H, Cénac P, Chaouch M (2010) Stochastic approximation to the multivariate and the functional median. In: Lechevallier Y, Saporta G (eds) Compstat 2010. Springer, New York, pp 421–428
Cardot, H, Cénac P, Godichon-Baggioni A (2016) Online estimation of the geometric median in Hilbert spaces: non asymptotic confidence balls. Ann Stat arXiv:1501.06930
Cardot H, Cénac P, Monnez J-M (2012) A fast and recursive algorithm for clustering large datasets with k-medians. Comput Stat Data Anal 56:1434–1449
Cardot H, Cénac P, Zitt P-A (2013) Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. Bernoulli 19:18–43
Cardot H, Degras D (2015) Online principal components analysis: which algorithm to choose? Tech Rep arXiv:1511.03688
Chaudhuri P (1992) Multivariate location estimation using extension of $R$-estimates through $U$-statistics type approach. Ann Stat 20(2):897–916
Croux C, Filzmoser P, Oliveira M (2007) Algorithms for projection-pursuit robust principal component analysis. Chemometr Intell Lab Syst 87:218–225
Croux C, Ruiz-Gazen A (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited. J Multivar Anal 95:206–226
Devlin S, Gnanadesikan R, Kettenring J (1981) Robust estimation of dispersion matrices and principal components. J Am Stat Assoc 76:354–362
Fritz H, Filzmoser P, Croux C (2012) A comparison of algorithms for the multivariate $L_1$-median. Comput Stat 27:393–410
Gervini D (2008) Robust functional estimation using the median and spherical principal components. Biometrika 95(3):587–600
Godichon-Baggioni A (2016) Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms; $L^p$ and almost sure rates of convergence. J Multivar Anal 146:209–222
Gu M, Eisenstat S (1994) A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem. SIAM J Matrix Anal Appl 15:1266–1276
Huber P, Ronchetti E (2009) Robust statistics. Wiley, Amsterdam
Hubert M, Rousseeuw P, Van Aelst S (2008) High-breakdown robust multivariate methods. Stat Sci 13:92–119
Hyndman R, Ullah S (2007) Robust forecasting of mortality and fertility rates: a functional data approach. Comput Stat Data Anal 51:4942–4956
Jolliffe I (2002) Principal components analysis, 2nd edn. Springer, New York
Kemperman JHB (1987) The median of a finite measure on a Banach space. In: Statistical data analysis based on the $L_1$-norm and related methods (Neuchâtel, 1987). North-Holland, Amsterdam, pp 217–230
Kraus D, Panaretos VM (2012) Dispersion operators and resistant second-order functional data analysis. Biometrika 99:813–832
Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K (1999) Robust principal components for functional data. Test 8:1–73
Lopuhaä HP, Rousseeuw PJ (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann Stat 19(1):229–248
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics. Wiley series in probability and statistics, theory and methods. Wiley, Chichester
Mokkadem A, Pelletier M (2006) Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. Ann Appl Probab 16(3):1671–1702
Polyak B, Juditsky A (1992) Acceleration of stochastic approximation. SIAM J Control Optim 30:838–855
R Development Core Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna ISBN 3-900051-07-0
Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York
Rousseeuw P, van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
Small CG (1990) A survey of multidimensional medians. Int Stat Rev/Revue Int Stat 58(3):263–277
Taskinen S, Koch I, Oja H (2012) Robustifying principal components analysis with spatial sign vectors. Stat Prob Lett 82:765–774
Vardi Y, Zhang C-H (2000) The multivariate $L_1$-median and associated data depth. Proc Natl Acad Sci USA 97(4):1423–1426
Weiszfeld E (1937) On the point for which the sum of the distances to n given points is minimum. Tohoku Math J 43:355–386
Weng J, Zhang Y, Hwang W-S (2003) Candid covariance-free incremental principal component analysis. IEEE Trans Pattern Anal Mach Intell 25:1034–1040