Robust functional estimation using the median and spherical principal components

BY DANIEL GERVINI

Department of Mathematical Sciences, University of Wisconsin, P.O. Box 413, Milwaukee, Wisconsin 53201, U.S.A.

gervini@uwm.edu

SUMMARY

We present robust estimators for the mean and the principal components of a stochastic process in $L^2(\mathbb{R})$. Robustness and asymptotic properties of the estimators are studied theoretically, by simulation and by example. It is shown that the proposed estimators are generally more robust to outliers than the commonly used sample mean and principal components, although their properties depend on the spacings of the eigenvalues of the covariance function.

Some key words: Breakdown point; Influence function; Nonparametric regression; Outlier detection; Stochastic process.

1. INTRODUCTION

The behaviour of a stochastic process in $L^2(\mathbb{R})$ is largely determined by its mean and covariance function. The estimators of these quantities most commonly used are the sample mean and the sample covariance function. A general overview of the problem and a extensive list of references can be found in Ramsay & Silverman (2005). More recent work, dealing with estimation of the principal components of the covariance function, includes Gervini (2006), Hall & Hosseini-Nasab (2006), Hall et al. (2006) and Yao & Lee (2006).

A serious drawback of these estimators is their extreme sensitivity to atypical curves. Outliers do occur in functional samples, as Fig. shows. These curves are trajectories of the lower lip of an individual repeatedly pronouncing the word

'bob' (Malfait & Ramsay, 2003). Three of these curves seem to be out of line, with the second peak occurring at a later time than for the rest of the curves. To determine whether or not these curves are within the normal range of variability, it is necessary to estimate accurately the covariance function or its principal components. However, we shall see in § 8 that these atypical curves distort the common principal component estimators and, as a result, they are not identified as outliers. On the other hand, the robust estimators we propose do detect these curves as outliers. In general, visual detection of outliers is not as easy as in Fig. , so it is important to develop robust estimators that can handle outliers in an automatic way.

Not much work has been done in the area of robust functional data analysis. Of course, it is always possible to reduce the functional problem to a multivariate one by evaluating the curves on a common output grid or by using the coefficients of a basis expansion, as in Locantore et al. (1999). However, that discretization of the problem has a number of disadvantages. For example, it requires two data-smoothing steps, first to evaluate the sample curves on a finite grid and then to reconstruct the functional estimators from this grid. The theoretical properties of these procedures are uncertain, and they clearly incur avoidable smoothing bias. A fully functional approach to the problem is preferable. In this context we can mention the trimmed means proposed by Fraiman & Muniz (2001) and the depth-based estimators of Cuevas et al. (2007) and López-Pintado & Romo (2007). In more abstract settings, the median was studied by Kemperman (1987).

In this article we present a fully functional approach to robust estimation of the mean and principal components. The estimators we propose are, essentially, functional versions of the multivariate median and spherical principal components (Locantore et al., 1999). Although some of the results in this article are straightforward extensions of those in Locantore et al. (1999), Boente & Fraiman (1999) and Marden (1999), many others are original. For reasons of space we omit proofs and technical details; they can be found in a technical supplement posted on the author's website.

2. THE FUNCTIONAL MEDIAN

Let P be the probability function associated with a stochastic process X on $L^2(T)$, where T is a closed interval in $\mathbb{R}$. On $L^2(T)$ we define the usual inner product $\langle f, g\rangle = \int fg$ and norm $\|f\| = \langle f, f\rangle^{1/2}$. The functional median of X, which we denote by $M(P)$, is defined as the minimizer of

$$F_P(y) = \int (\|x - y\| - \|x\|)P(dx)$$

over all $y \in L^2(T)$. Note that $F_P(y)$ is well defined even if $\int \|x\|P(dx) = \infty$. Since F_P is convex and $\lim_{y\to\infty} F_P(y) = +\infty$ (Kemperman, 1987, Lemma 2.3), there is at least one finite minimizer of F_P, and the minimizer is unique unless P is concentrated on a one-dimensional set of the form $\{s\phi_0 : s \in \mathbb{R}\}$ for some $\phi_0 \in L^2(T)$ (Kemperman, 1987, Theorem 2.17). In that case the set of minimizers of $F_P(y)$ is of the form $\{s\phi_0 : s \in [a, b]\}$ and one can choose $M(P) = \{(a + b)/2\}\phi_0$. In this way, the functional median is well defined and translation-invariant for any stochastic process X in $L^2(T)$. Note that if X is symmetric about μ, in the sense that $X - \mu$ and $\mu - X$ have the same distribution, then $M(P) = \mu$ because $F_P(\mu + y) = F_P(\mu - y)$ for all $y \in L^2(T)$. For a random sample $X_1, \ldots, X_n$ the median is $\tilde{\mu}_n = M(P_n)$, where P_n is the empirical measure. In other words,

$$\tilde{\mu}_n = \arg \min_{y\in L^2(T)} \sum_{i=1}^{n} \|X_i - y\| . \tag{1}$$

Estimating equations for $M(P)$ can be obtained using Gateaux differentials (Luenberger, 1969, Ch. 7), as shown in the Appendix. It turns out that $M(P)$ must satisfy

$$\int \frac{\{x - M(P)\}}{\|x - M(P)\|} P(dx) = 0, \tag{2}$$

whenever $P[\{M(P)\}] = 0$. For the sample median $\tilde{\mu}_n$ the estimating equation

would be

$$\sum_{i=1}^{n} \frac{X_i - \tilde{\mu}_n}{\|X_i - \tilde{\mu}_n\|} = 0, \tag{3}$$

provided that $\tilde{\mu}_n \neq X_i$ for all i.

An important consequence of (3) is that

$$\tilde{\mu}_n = \sum_{i=1}^{n} w_i X_i, \text{ with } w_i \geq 0 \text{ and } \sum_{i=1}^{n} w_i = 1. \tag{4}$$

To be explicit, $w_i = \|X_i - \tilde{\mu}_n\|^{-1} / \sum_{i=1}^{n} \|X_i - \tilde{\mu}_n\|^{-1}$ when (3) holds, but, even if $\tilde{\mu}_n = X_i$ for some i, the representation (4) is still valid. Therefore, the sample median can be found by minimizing

$$\sum_{i=1}^{n} \|X_i - \sum_{j=1}^{n} w_j X_j\| = \sum_{i=1}^{n} \{(e_i - w)^{\mathrm{T}} G (e_i - w)\}^{1/2} \tag{5}$$

with respect to $w = (w_1, \ldots, w_n)^{\mathrm{T}}$, subject to the restrictions $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$, where e_i is the ith canonical vector in $\mathbb{R}^n$ and G is the inner-product matrix with elements $G_{ij} = \langle X_i, X_j \rangle$.

In this way, we have transformed the functional minimization problem (1) into a convex n-dimensional minimization problem (5). Although n can be quite large, the minimizer can be found very easily with the algorithm given in the Appendix. This algorithm uses the sample mean as starting point and converges to the median by adjusting the weights w_i on each step. Note that only the inner-product matrix G is used, not the actual values of the X_i's. This is an important practical advantage, since the inner products $\langle X_i, X_j \rangle$ can usually be estimated from the raw data, without smoothing.

For instance, suppose the raw data consist of vectors $x_1, \ldots, x_n$ with $x_{ij} = X_i(t_j) + \varepsilon_{ij}$, $t_1 < \cdots < t_m$ is the common input grid and $\{\varepsilon_{ij}\}$ are independent

random errors, with $E(\varepsilon_{ij}) = 0$ and $\mathrm{var}(\varepsilon_{ij}) = \sigma_i^2$. Then G_{ik} can be estimated by

$$\hat{G}_{ik} = \sum_{j=1}^{m-1} \Big(\frac{x_{ij}x_{kj} + x_{i,j+1}x_{k,j+1}}{2} \Big) (t_{j+1} - t_j) - \hat{\sigma}_i^2 \, \delta_{ik}, \tag{6}$$

where δ_{ik} is Kronecker's delta and $\hat{\sigma}_i^2$ is a consistent nonparametric variance estimator, such as the one proposed by Gasser et al. (1986). Note that $\hat{G}_{ik}$ is basically a bias-corrected version of the trapezoidal rule. In practice, the correction term can be omitted if the noise level is small. The following theorem shows that $\hat{G}_{ik}$ is consistent, conditionally on $X_1, \ldots, X_n$, when the number of observations per curve goes to infinity.

THEOREM 1 . *Suppose that $X_1, \ldots, X_n$ are continuous, $E(\varepsilon_{ij}^4) < \infty$, $\max_{1 \le j \le m-1} (t_{j+1} - t_j) \to 0$ as $m \to \infty$, and $\hat{\sigma}_i^2 \to \sigma_i^2$ in probability conditionally on X_i as $m \to \infty$, for all i. Then $\hat{G}_{ik} \to G_{ik}$ in probability conditionally on X_i and X_k as $m \to \infty$, for all i, k.*

3. SPHERICAL PRINCIPAL COMPONENTS

If a stochastic process X in $L^2(T)$ has a continuous covariance function $\rho(s,t) = \mathrm{cov}\{X(s), X(t)\}$, there exist a complete orthonormal system $\{\phi_k\}$ in $L_2(T)$ and a sequence of real numbers $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ such that $\rho(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t)$, with the series converging absolutely and uniformly on $T \times T$ (Ash & Gardner, 1975). Since ϕ_k is an eigenfunction of ρ with eigenvalue λ_k, it is usually estimated by the kth eigenfunction of the sample covariance function. However, these estimators are very sensitive to atypical curves.

As alternative estimators, we propose the eigenfunctions of the weighted covariance function

$$\tilde{\rho}_n(s,t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{X_i(s) - \tilde{\mu}_n(s)\}\{X_i(t) - \tilde{\mu}_n(t)\}}{\|X_i - \tilde{\mu}_n\|^2}.$$

Note that $\tilde{\rho}_n$ is just the sample covariance function of the centred curves projected on the unit sphere, so we call the eigenfunctions $\tilde{\phi}_{n,k}$ of $\tilde{\rho}_n$ 'spherical principal components', as in Locantore et al. (1999). Once again, it is possible to compute the spherical principal components using only the inner-product matrix G; details are given in the Appendix.

4. CONSISTENCY OF THE ESTIMATORS

The next theorem shows that the median is consistent, in the sense of weak convergence, under very general conditions. Although this is generally weaker than consistency in L^2 norm, for finite-dimensional subspaces of $L^2(T)$ like the one generated by model (8) below, both forms of convergence are equivalent.

THEOREM 2 . *Let $X_1, \ldots, X_n$ be a random sample of a probability function P_0 that is not concentrated on a one-dimensional set of $L^2(T)$, so that $M(P_0)$ is unique. Then $\langle \tilde{\mu}_n, f \rangle \to \langle M(P_0), f \rangle$ almost surely for any $f \in L^2(T)$.*

Theorem 2 covers all cases of practical interest except for the shape-invariant model $X(t) = (1 + \sigma Z)\mu(t)$, where Z is a random variable symmetric about 0. However, this is essentially a univariate problem, since $\tilde{\mu}_n = \{1+\sigma \mathrm{med}(Z_i)\}\mu(t)$ and $M(P_0) = \{1+\sigma \mathrm{med}(Z)\}\mu(t)$, where $\mathrm{med}(Z_i)$ and $\mathrm{med}(Z)$ are the univariate sample and population medians, respectively. The properties of these univariate estimators are well known.

Regarding the spherical principal components, it follows from Theorem 2 and the Strong Law of Large Numbers that $\tilde{\rho}_n$ converges almost surely, in the sense of weak convergence in $L^2(T \times T)$, to

$$\tilde{\rho}(s,t) = E\left[\frac{\{X(s) - \mu(s)\}\{X(t) - \mu(t)\}}{\|X - \mu\|^2}\right].$$

Therefore, if $\tilde{\phi}_k$, the kth eigenfunction of $\tilde{\rho}$, has multiplicity one, the sample spherical component $\tilde{\phi}_{n,k}$ is strongly consistent for the right choice of sign, or equivalently $|\langle \tilde{\phi}_{n,k}, \tilde{\phi}_k \rangle| \to 1$ almost surely (Dauxois et al., 1982, Proposition 4). The

question is whether or not $\tilde{\phi}_k = \phi_k$. To answer this, we need stronger model assumptions on X.

It is known that, if $\mu(t) = E\{X(t)\}$ and $\rho(s,t)$ are continuous, then $X(t)$ admits the expansion

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} Z_k \lambda_k^{1/2} \phi_k(t),\ t \in T, \tag{7}$$

where $\{Z_k\}$ are uncorrelated random variables with zero mean and unit variance. The right-hand side of (7), known as the Karhunen–Loève decomposition, converges in square mean and uniformly on T (Ash & Gardner, 1975, p. 38).

In most cases, the sequence of eigenvalues decreases so fast that the tail components can be ignored for practical purposes. Without much loss of generality, therefore, we will assume a finite p-component model,

$$\begin{aligned} X(t) &= \mu(t) + \sum_{k=1}^{p} Z_k \lambda_k^{1/2} \phi_k(t),\ t \in T, \\ &\lambda_1 \geq \cdots \geq \lambda_p > 0, \\ &Z = (Z_1, \ldots, Z_p)^{\mathrm{T}} \text{ has symmetric and exchangeable marginals.} \end{aligned} \tag{8}$$

Note that we do not assume finite moments of Z of any order. The assumption that Z has symmetric and exchangeable marginals is satisfied by all spherical distributions (Bilodeau & Brenner, 1999), including the multivariate Normal and multivariate t distributions, and also by some non-spherical distributions such as the uniform on the hypercube and independent-marginal t distributions. If Z does have finite second moments, marginal symmetry implies that $E(Z_i Z_j) = 0$ for $i \neq j$, so that $\rho(s,t) = E(Z_1^2) \sum_{k=1}^{p} \lambda_k \phi_k(s) \phi_k(t)$. Consequently, all models (8) with finite second moments have the same principal components, up to the usual sign ambiguity, and the same eigenvalue ratios $\lambda_2/\lambda_1, \ldots, \lambda_p/\lambda_1$.

THEOREM 3 . *Let* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ *and* $\Phi(t) = (\phi_1(t), \ldots, \phi_p(t))^{\mathrm{T}}$. *Under model* (8), *we have* $\tilde{\rho}(s,t) = \Phi(s)^{\mathrm{T}} \Lambda^{1/2} \Omega \Lambda^{1/2} \Phi(t)$ *with* $\Omega = E\left\{ZZ^{\mathrm{T}}/(Z^{\mathrm{T}}\Lambda Z)\right\}$

diagonal. Then the eigenvalues of $\tilde{\rho}$ *are* $\tilde{\lambda}_k := \lambda_k \Omega_{kk}$, $k = 1, \ldots, p$, *with respective eigenfunctions* $\phi_1, \ldots, \phi_p$. *Moreover,* $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_p$, *and* $\tilde{\lambda}_k > \tilde{\lambda}_{k+1}$ *if* $\lambda_k > \lambda_{k+1}$ *and* Z_k *has a nondegenerate distribution.*

Note that, if Z has a spherical distribution, $\Omega = E\{UU^{\mathrm{T}}/(U^{\mathrm{T}}\Lambda U)\}$ with $U \in \mathbb{R}^p$ uniformly distributed on the unit sphere, so that Ω depends on $\{\lambda_k\}$ and p but not on the specific distribution of Z. In other words, Ω is the same for all spherical models.

The main consequence of Theorem 3 is that $\tilde{\rho}$ has exactly the same eigenfunctions as ρ, in the same order, and, if Z is not degenerate, with the same multiplicity. However, the eigenvalues of $\tilde{\rho}$ are not the same as those of ρ. To estimate the λ_k's consistently we suggest using a robust variance estimator of the component scores $\{\langle X_i - \tilde{\mu}_n, \tilde{\phi}_{n,k} \rangle\}$, such as the median squares, calibrated for the specific distribution of Z; the calibration is not necessary if only the eigenvalue ratios need to be estimated.

5. Robustness and asymptotic normality of the median

Given $\varepsilon \in [0, 1)$, consider the ε-contamination neighbourhood of a probability measure P_0,

$$\mathcal{N}_\varepsilon(P_0) = \left\{(1-\varepsilon)P_0 + \varepsilon Q : \ Q \text{ a probability measure on } L^2(T)\right\}.$$

The maximum bias of the median over $\mathcal{N}_\varepsilon(P_0)$ is defined as

$$B_M(\varepsilon) = \sup\{\|M(P) - \mu\| : P \in \mathcal{N}_\varepsilon(P_0)\}.$$

Note that $B_M(\varepsilon)$ need not be finite for $\varepsilon > 0$, and in fact it will be infinite for ε large enough. We therefore define the breakdown point as $\varepsilon^*_M = \inf\{\varepsilon : B_M(\varepsilon) = \infty\}$.

We will show two important properties of $B_M(\varepsilon)$, namely, that it goes to zero when the contamination proportion goes to zero, and that $\varepsilon^*_M = 0.50$. A

translation-equivariant estimator of location cannot have a breakdown point greater than 0.50 (Maronna et al., 2006, p. 76), so the median attains the optimal upper bound. Part (i) of the next theorem establishes the so-called 'exact fit property', which means that, if a single point has probability greater than 0.50, then that point has to be the median.

THEOREM 4 . (i) *If* $P(\{z\}) > 1/2$ *for some* $z \in L^2(T)$, *then* $M(P) = z$ *and the median is unique, and, if* $P(\{z\}) = 1/2$, *the median may not be unique but* z *is one of the possible medians.*
(ii) $\varepsilon^*_M = 0.50$.
(iii) *If* μ *is the unique median of* P_0, $\lim_{\varepsilon \to 0} B_M(\varepsilon) = 0$.

Another measure of robustness is the influence function, which measures the sensitivity of an estimator to clustered outliers. The influence function is defined as

$$\text{IF}_M(z) = \lim_{\varepsilon \to 0} \varepsilon^{-1}\{M(P_{\varepsilon,z}) - M(P_0)\},$$

provided the limit exists, where $P_{\varepsilon,z} = (1-\varepsilon)P_0 + \varepsilon\delta_z$ and δ_z is the point-mass probability at $z \in L^2(T)$. The gross-error sensitivity is defined as

$$\gamma^*_M = \sup\{\|\text{IF}_M(z)\| : z \in L^2(T)\}.$$

The next theorem shows that $\gamma^*_M < \infty$, meaning that clustered outliers have only a limited effect on the median regardless of their location.

THEOREM 5 . *If we assume* (8) *as the central model, and if* Z *has nondegenerate marginals, no atoms in a neighbourhood of* 0 *and* $E(\|Z\|^{-1}) < \infty$, *then*

$$\begin{aligned}
\text{IF}_M(z) &= \frac{1}{c}\frac{z-\mu}{\|z-\mu\|} + \sum_{k=1}^{p} \frac{\lambda_k \xi_k}{c(c-\lambda_k\xi_k)} \langle \phi_k, \frac{z-\mu}{\|z-\mu\|} \rangle \phi_k, \text{ for } z \neq \mu, \\
\text{IF}_M(\mu) &= 0, \\
\gamma^*_M &= \frac{1}{\min_{1\le k\le p}(c-\lambda_k\xi_k)},
\end{aligned}$$

where $c = E\{(Z^{\mathrm{T}}\Lambda Z)^{-1/2}\}$ *and* $\xi_k = E\{Z_k^2(Z^{\mathrm{T}}\Lambda Z)^{-3/2}\}$.

The assumption $E(\|Z\|^{-1}) < \infty$ in Theorem 5 is not very constraining; it is satisfied if $\|Z\|$ has a bounded density. If Z has a spherical distribution, $c = E(\|Z\|^{-1})E\{(U^{\mathrm{T}}\Lambda U)^{-1/2}\}$ and $\xi_k = E(\|Z\|^{-1})E\{U_k^2(U^{\mathrm{T}}\Lambda U)^{-3/2}\}$, where $U \in \mathbb{R}^p$ has the uniform distribution on the unit sphere. Then the ratios $\{\xi_k/c\}$ are the same for any Z with spherical distribution, and the influence function can be factorized as $\mathrm{IF}_M(z) = E\{\|Z\|^{-1}\}G(z)$, with $G(z)$ depending only on $\{\lambda_k\}$ and $\{\phi_k\}$ but not on the specific distribution of Z. The k for which $\min_{1\le k\le p}(c - \lambda_k\xi_k)$ is attained also depends on the sequence $\{\lambda_k\}$ alone.

The influence function also determines the asymptotic covariance of $n^{1/2}(\tilde{\mu}_n - \mu)$ when $\tilde{\mu}_n$ is asymptotically Normal, as shown next.

THEOREM 6 . *Under the same conditions of Theorem* 5, $n^{1/2}(\tilde{\mu}_n - \mu)$ *converges in distribution to a Gaussian random element of* $L^2(T)$ *with mean zero and covariance function* $\rho_M(s,t) = E_{P_0}\{\mathrm{IF}_M(X)(s)\mathrm{IF}_M(X)(t)\}$. *To be explicit,*

$$\rho_M(s,t) = \sum_{k=1}^{p} \frac{\tilde{\lambda}_k}{(c - \lambda_k\xi_k)^2}\,\phi_k(s)\phi_k(t),$$

with $\tilde{\lambda}_k$ *as in Theorem* 3.

Using Theorem 6 we can compute the asymptotic relative efficiency of the median with respect to the sample mean. Since $n^{1/2}(\bar{X} - \mu)$ is asymptotically Gaussian with mean zero and covariance function $\rho(s,t)$, we can define $\mathrm{ARE}(\tilde{\mu}_n, \bar{X})$ as the worst possible asymptotic relative efficiency of a linear function of the estimators,

$$\begin{aligned}
\mathrm{ARE}(\tilde{\mu}_n, \bar{X}) &= \inf_{f\neq 0} \frac{\int\int \rho(s,t)f(s)f(t)\,ds\,dt}{\int\int \rho_M(s,t)f(s)f(t)\,ds\,dt} \\
&= \min_{1\le k\le p} \frac{\lambda_k E(Z_1^2)(c - \lambda_k\xi_k)^2}{\tilde{\lambda}_k} \\
&= E(Z_1^2)c^2 \min_{1\le k\le p} \frac{(1 - \lambda_k\xi_k/c)^2}{\Omega_{kk}}.
\end{aligned}$$

For every Z with elliptical distribution, the k at which the minimum is attained depends only on the λ_k's, although the actual value of $\text{ARE}(\tilde{\mu}_n, \bar{X})$ does depend on the specific distribution of Z. Table 1 shows a few values of $\text{ARE}(\tilde{\mu}_n, \bar{X})$ when Z has a multivariate t distribution with ν degrees of freedom and the eigenvalues are $\lambda_k = 1/k^r$, $k = 1, \ldots, p$. For each p we observe that, with ν fixed, the relative efficiency of the median decreases as r increases; and, with r fixed, the relative efficiency decreases as ν increases. In other words, the median is more efficient for heavy-tailed distributions that are not too elliptical. Comparisons across p are less clear-cut: for $r = 1, 2$, the median is more efficient for $p = 5$ than for $p = 2$, but for $r = 4$ the opposite is true, so we cannot draw general conclusions.

6. Robustness of the spherical principal components

Let $\mathcal{N}_\varepsilon(P_0)$ be, as before, the ε-contamination neighbourhood of P_0 and let $\Phi_k(P)$ be the kth spherical principal component of a random process with distribution P. The maximum bias of Φ_k over $\mathcal{N}_\varepsilon(P_0)$ can be defined as

$$B_{\Phi_k}(\varepsilon) = \sup\{\arccos|\langle \Phi_k(P), \phi_k\rangle| : P \in \mathcal{N}_\varepsilon(P_0)\}.$$

The bias is always bounded because $\|\Phi_k(P)\| = 1$ by definition, but we can say that breakdown occurs if $\Phi_k(P)$ is orthogonal to ϕ_k. Then we define the breakdown point of Φ_k as $\varepsilon^*_{\Phi_k} = \inf\{\varepsilon : B_{\Phi_k}(\varepsilon) = \pi/2\}$. The next theorem gives useful, although somewhat discouraging, upper bounds for $\varepsilon^*_{\Phi_k}$.

Theorem 7 . *If we assume* (8) *as the central model, if* $\lambda_1 > \cdots > \lambda_p > 0$ *and* Z *has nondegenerate marginals, then*

$$\begin{aligned}
\varepsilon^*_{\Phi_k} &\leq \min\left\{\frac{\tilde{\lambda}_{k-1} - \tilde{\lambda}_k}{1 + \tilde{\lambda}_{k-1} - \tilde{\lambda}_k}, \frac{\tilde{\lambda}_k - \tilde{\lambda}_{k+1}}{1 + \tilde{\lambda}_k - \tilde{\lambda}_{k+1}}\right\},\ k = 2, \ldots, p-1,\\
\varepsilon^*_{\Phi_1} &\leq \frac{\tilde{\lambda}_1 - \tilde{\lambda}_2}{1 + \tilde{\lambda}_1 - \tilde{\lambda}_2},\\
\varepsilon^*_{\Phi_p} &\leq \min\left\{\frac{\tilde{\lambda}_{p-1} - \tilde{\lambda}_p}{1 + \tilde{\lambda}_{p-1} - \tilde{\lambda}_p}, \frac{\tilde{\lambda}_p}{1 + \tilde{\lambda}_p}\right\}.
\end{aligned}$$

The proof of Theorem 7 is instructive. It shows that symmetric point-mass contaminations of the form $P_{\varepsilon,z} = (1-\varepsilon)P_0 + \varepsilon(0.5\delta_{\mu+z} + 0.5\delta_{\mu-z})$, with either $z = \phi_k$ or $z = \phi_{k+1}$, may cause breakdown of Φ_k. These contaminations cause overestimation of the variance either in the direction of ϕ_k, making $\Phi_k(P_{\varepsilon,z}) = \pm\phi_{k-1}$, or in the direction of ϕ_{k+1}, making $\Phi_k(P_{\varepsilon,z}) = \pm\phi_{k+1}$. The upper bounds given in Theorem 7 depend on the spacings of the eigenvalues, which is a rather negative feature of the estimators, since $\tilde{\lambda}_j - \tilde{\lambda}_{j+1} \to 0$ when $\lambda_j - \lambda_{j+1} \to 0$ and then $\varepsilon^*_{\Phi_k}$ can be arbitrarily close to zero if λ_{k-1}, λ_k and λ_{k+1} are too close to one another.

The next Theorem gives the influence function of Φ_k.

THEOREM 8 . *Under the same assumptions as in Theorem* 7,

$$\begin{aligned}
\text{IF}_{\Phi_k}(z) &= \frac{\zeta_k(z)}{\tilde{\lambda}_k}\left\{\frac{z-\mu}{\|z-\mu\|} - \zeta_k(z)\phi_k + \sum_{\substack{j=1\\ j\neq k}}^{p} \frac{\tilde{\lambda}_j\zeta_j(z)}{(\tilde{\lambda}_k - \tilde{\lambda}_j)}\phi_j\right\}, \textit{ if } z \neq \mu, \\
\text{IF}_{\Phi_k}(\mu) &= 0, \textit{ if } z = \mu, \\
\gamma^*_{\Phi_k} &= \frac{1}{2c_k},
\end{aligned}$$

where $\zeta_j(z) = \langle z-\mu, \phi_j\rangle / \|z-\mu\|$ *and* $c_k = \min[\tilde{\lambda}_k, \{\min|\tilde{\lambda}_k - \tilde{\lambda}_j| : j = 1,\ldots,p,\ j \neq k\}]$.

Again, we see that the influence function of Φ_k depends strongly on the eigenvalue spacings. If $\lambda_{k-1} - \lambda_k \to 0$ or $\lambda_k - \lambda_{k+1} \to 0$ then $\gamma^*_{\Phi_k} \to \infty$, which implies that the bias of Φ_k may be very large even for small values of ε. Nevertheless, $\gamma^*_{\Phi_k} < \infty$ for any given sequence of distinct eigenvalues, so the spherical principal components are always robust in this sense.

7. SIMULATIONS

We ran some simulations to study the finite-sample behaviour of the estimators. Samples were generated from model (8) with $\mu = 0$, $p = 5$ and $\phi_k(t) =$

$\sqrt{2}\sin(\pi kt)$, for $t \in [0,1]$. The actual observations $x_1, \ldots, x_n$ were sampled on a grid of $m = 30$ equispaced points in $[0,1]$. As eigenvalues, we considered three different sequences: $\lambda_k^{(1)} = 1/k$, $\lambda_k^{(2)} = 1/k^2$ and $\lambda_k^{(3)} = 1/k^4$, for $k = 1, \ldots, 5$.

Our first goal was to study the estimation error of the median under heavy-tailed models, so we generated observations from model (8) with Z following multivariate t distributions with ν degrees of freedom, for $\nu = 1, \ldots, 15$. The sample size was $n = 100$, and 2000 samples were generated for each model. The estimation error for each sample was measured by the average squared error, $\text{ASE}(\hat{\mu}) = \sum_{j=1}^{m}\{\hat{\mu}(t_j) - \mu(t_j)\}^2/m$, and the mean $\text{ASE}(\hat{\mu})$ over the 2000 simulated samples was computed, which is a Monte Carlo estimate of the mean average squared error $\text{MASE}(\hat{\mu}) = E\{\text{ASE}(\hat{\mu})\}$. Figure plots the relative MASE of the median with respect to the sample mean. As expected, the median is a much better estimator for heavy-tailed distributions. However, even at worst, the MASE of the median is only 20% larger than that of the mean.

We also studied the robustness of the median under point-mass contaminations of a Normal model. We generated $n = 100$ curves from model (8) with $Z \sim N(0, I)$, and replaced εn of them by $K\lambda_j^{1/2}\phi_j(t)$. We considered several choices for the parameters: $K = 2, 4, 8, 16$, $j = 1, 3$ and $\varepsilon = 0.05, 0.10, \ldots, 0.45$. Each sampling situation was replicated 200 times. As error measure we used the mean root average squared error, $E\{\text{ASE}^{1/2}(\hat{\mu})\}$, which is the finite-sample equivalent of the bias function $\|M(P) - \mu\|$ used in § 5. It would take up too much space to report the simulation results in full detail, but we can mention that, for each ε and K, the estimation error is larger for contaminations in the ϕ_1-direction than in the ϕ_3-direction, as expected from Theorem 5, since $\min_{1\le k\le p}(c - \lambda_k\xi_k)$ is attained at $k = 1$ for the three sequences of eigenvalues under consideration. Moreover, for each ε the estimation error increases with K, but for contaminations in the ϕ_1-direction there is little difference between $K = 8$ and $K = 16$, so we only show the results for $K = 16$; see Fig. (a). The most favourable situation for the median is now the most 'elliptical' case, $\lambda^{(3)}$, and the worst is $\lambda^{(1)}$. This is not unexpected, since the gross error sensitivities are 1.79 for $\lambda^{(1)}$, 1.52 for $\lambda^{(2)}$ and

1.33 for $\lambda^{(3)}$.

Finally, we studied the robustness of the spherical components under point-mass contaminated Normal models. Again, we considered many types of contamination but report only the worst-case scenario, where εn observations are replaced by $K\lambda_2^{1/2}\phi_2(t)$ with $K = 16$, for $\varepsilon = 0.05, 0.10, \ldots, 0.45$. The sample size was again $n = 200$ and each sampling situation was replicated 200 times. The estimation error was now measured by the absolute angle, $\text{AA}(\hat{\phi}_k) = \arccos\{|\sum_{j=1}^{m} \hat{\phi}_k(t_j)\phi_k(t_j)/(m-1)|\}$, and $\text{MAA}(\hat{\phi}_k) = E\{\text{AA}(\hat{\phi}_k)\}$. The simulated $\text{MAA}(\hat{\phi}_1)$ curves are shown in Fig. (b); we do not show the corresponding curves for the sample principal components because they are practically constant at $\pi/2$, the breakdown value. The worst case for estimation of ϕ_1 corresponds to the most 'spherical' sequence $\lambda^{(1)}$, because the first two eigenvalues are too close to one another. To be specific, $\tilde{\lambda}^{(1)} = (0.35, 0.23, 0.17, 0.14, 0.11)$, $\tilde{\lambda}^{(2)} = (0.51, 0.22, 0.13, 0.08, 0.05)$ and $\tilde{\lambda}^{(3)} = (0.75, 0.17, 0.05, 0.02, 0.01)$, so from Theorems 7 and 8 we have $\varepsilon^*_{\Phi_1} \leq 0.11$ and $\gamma^*_{\Phi_1} = 3.91$ for $\lambda^{(1)}$, $\varepsilon^*_{\Phi_1} \leq 0.22$ and $\gamma^*_{\Phi_1} = 1.72$ for $\lambda^{(2)}$, and $\varepsilon^*_{\Phi_1} \leq 0.37$ and $\gamma^*_{\Phi_1} = 0.86$ for $\lambda^{(3)}$, which is consistent with the $\text{MAA}(\hat{\phi}_1)$ curves shown in Fig. (b).

In conclusion, we can say that, although the behaviour of the median and the spherical components is strongly dependent on the model eigenvalues, these estimators do represent an improvement over the sample mean and principal components from the point of view of robustness, and this improvement may be quite substantial for some models.

8. EXAMPLE: LIP MOVEMENT IN HUMAN SPEECH

Understanding the relationship between lip movement and time of activation of different face muscles is a central problem in the physiological study of human speech. In the experiment reported by Malfait & Ramsay (2003), a subject was instructed to say the word 'bob' 32 times. We did not have access to the raw data, but the smoothed data are publicly available at http://www.stats.ox.ac.uk/˜silverma/fdacasebook/lipemg.html. The trajectories of the centre of the lower lip are shown

in Figs and (a). All of these curves show a similar pattern: a peak corresponding to the first /b/, a plateau corresponding to the /o/, and a second peak corresponding to the last /b/. In addition to amplitude variability, these curves show substantial time variability, especially for the second /b/. Note that three of the curves, highlighted in Fig. (a), exhibit large first peaks and considerably delayed second peaks.

The first spherical and sample principal components are shown in Fig. (b). To understand better the type of variability explained by these components, we plotted the mean and the median plus/minus a constant times the corresponding components in Figs (c) and (d). Clearly, the three atypical curves, which represent only 10% of the data, have a large influence on the sample principal component. It is not easy to interpret this component because of the mixed effect of amplitude and time variability. Nevertheless, it is clear from Fig. (d) that a positive component score will be associated with curves that show a large first peak and a delayed second peak, which is precisely the pattern shown by the three outlying curves.

A residual plot is helpful to detect outliers. Given estimators of the mean, $\hat{\mu}$, and the principal components, $\{\hat{\phi}_k\}$, we define the component scores $s_{ik} = \langle x_i - \hat{\mu}, \hat{\phi}_k \rangle$ and the residual trajectories $r_i(t) = x_i(t) - \hat{\mu}(t) - \sum_{k=1}^{q} s_{ik}\hat{\phi}_k(t)$. For this particular example we use five components, since they account for practically all of the systematic variability. Figure (a) shows the residual curves using the robust estimators, and it is clear that three curves have large negative residuals around $t = 0.5$, which is the point where the second peak occurs for most of the sample curves. To identify these curves, we plotted $\|r_i\|^2$ versus i in Fig. (c); the three outliers are observations 24, 25 and 27. In contrast, the plots corresponding to the sample mean and principal components, see Figs (b) and (d), do not show any unusual observations, because of the masking effect of the outliers.

ACKNOWLEDGEMENT

This research was supported by a grant from the U. S. National Science Foundation.

APPENDIX
Technical details

Derivation of (2). Given $y \in L^2(T)$, the real-valued function $G(s) = F_P\{M(P)+sy\}$, $s \in \mathbb{R}$, has a minimum at $s = 0$ by definition of $M(P)$. Therefore $G'(0) = 0$, provided that G is differentiable at 0. It is easy to check that if $P[\{M(P)\}] = 0$ then G is differentiable at 0 and

$$G'(0) = -\int \frac{\langle x - M(P), y\rangle}{\|x - M(P)\|} P(dx) = \langle -\int \frac{\{x - M(P)\}}{\|x - M(P)\|} P(dx), y\rangle.$$

Since $G'(0) = 0$ for all $y \in L^2(T)$, (2) follows.

Computation of the median and the spherical components. To compute the median we propose the following reweighting algorithm, similar to Gower (1974) and Vardi & Zhang (2000).

Step 1. *Initialization.* Set $w_i^{(0)} = 1/n$ and $d_i^{(0)} = \{\left(e_i - w^{(0)}\right)^{\mathrm{T}} G \left(e_i - w^{(0)}\right)\}^{\frac{1}{2}}$.

Step 2. *Iteration.* Repeat the following stages for $k \geq 1$ until convergence.

(i) If all $d_i^{(k-1)} > 0$, set $w_i^{(k)} = (d_i^{(k-1)})^{-1} / \sum_{i=1}^n (d_i^{(k-1)})^{-1}$. Otherwise, let $\mathcal{I} = \{i : d_i^{(k-1)} = 0\}$, and set $w_i^{(k)} = 1/\mathrm{card}(\mathcal{I})$ for $i \in \mathcal{I}$ and $w_i^{(k)} = 0$ for $i \notin \mathcal{I}$;

(ii) Set $d_i^{(k)} = \{\left(e_i - w^{(k)}\right)^{\mathrm{T}} G \left(e_i - w^{(k)}\right)\}^{1/2}$.

For the spherical components, we use an idea similar to one in Kneip (1999). Let $U_i = (X_i - \tilde{\mu}_n)/\|X_i - \tilde{\mu}_n\|$ and let $\tilde{G}$ be the $n \times n$ matrix with elements $\tilde{G}_{ij} = \langle U_i, U_j\rangle$. Note that

$$\tilde{G}_{ij} = \frac{(e_i - w_0)^{\mathrm{T}} G (e_j - w_0)}{\{(e_i - w_0)^{\mathrm{T}} G (e_i - w_0)\}^{1/2} \{(e_j - w_0)^{\mathrm{T}} G (e_j - w_0)\}^{1/2}},$$

where w_0 is the vector of weights in the representation (4) of $\tilde{\mu}_n$, and $\tilde{G}$ is symmetric and nonnegative definite, since $w^{\mathrm{T}} \tilde{G} w = \|\sum_{i=1}^n w_i U_i\|^2$. Let l_k be the

kth eigenvalue of $\tilde{G}$ with corresponding eigenvector v_k, $\tilde{\lambda}_{n,k} = l_k/n$ and let $w_k = v_k/\sqrt{l_k}$, provided that $l_k > 0$. Then $\tilde{\lambda}_{n,k}$ is the kth eigenvalue of $\tilde{\rho}_n$, with corresponding eigenfunction

$$\tilde{\phi}_{n,k} = \sum_{i=1}^{n} w_{ki} \frac{(X_i - \tilde{\mu}_n)}{\|X_i - \tilde{\mu}_n\|}. \tag{A1}$$

To see (A1), note that $\tilde{\lambda}_{n,k}$ and $\tilde{\phi}_{n,k}$ satisfy, by definition, $\langle \tilde{\rho}_n(\cdot, t), \tilde{\phi}_{n,k} \rangle = \tilde{\lambda}_{n,k} \tilde{\phi}_{n,k}(t)$ for every $t \in T$, with $\|\tilde{\phi}_{n,k}\| = 1$ and $\tilde{\lambda}_{n,k} = \max \iint \tilde{\rho}_n(s,t)\phi(s)\phi(t)\, ds\, dt$ among those $\phi \in L^2(T)$ with $\|\phi\| = 1$ and $\langle \phi, \tilde{\phi}_j \rangle = 0$, $j = 1, \ldots, k-1$. The equation $\langle \tilde{\rho}_n(\cdot, t), \tilde{\phi}_{n,k} \rangle = \tilde{\lambda}_{n,k} \tilde{\phi}_{n,k}(t)$ can be written as

$$\frac{1}{n} \sum_{i=1}^{n} U_i(t) \langle U_i, \tilde{\phi}_{n,k} \rangle = \tilde{\lambda}_{n,k} \tilde{\phi}_{n,k}(t),$$

so that $\tilde{\phi}_{n,k} = \sum_{i=1}^{n} w_{ki} U_i$ with $w_{ki} = \langle U_i, \tilde{\phi}_{n,k} \rangle / (n\tilde{\lambda}_{n,k})$. Therefore, without loss of generality we can restrict the maximization problem to those $\phi \in L^2(T)$ of the form $\phi = \sum_{i=1}^{n} w_i U_i$. In that case, $\|\phi\|^2 = w^{\mathrm{T}} \tilde{G} w$ and $\int \int \tilde{\rho}_n(s,t)\phi(s)\phi(t)\, ds\, dt = n^{-1} w^{\mathrm{T}} \tilde{G}^2 w$. Thus $\tilde{\lambda}_{n,k} = n^{-1} \max w^{\mathrm{T}} \tilde{G}^2 w$ subject to the restrictions $w^{\mathrm{T}} \tilde{G} w = 1$ and $w^{\mathrm{T}} \tilde{G} w_j = 0$, for $j = 1, \ldots, k-1$. If we let $v = \tilde{G}^{1/2} w$, it follows that $\tilde{\lambda}_{n,k} = l_k/n$ and $w_k = \tilde{G}^{-1/2} v_k$, where l_k is the kth eigenvalue of $\tilde{G}$ and v_k is the associated eigenvector. Note that $\tilde{G}^{-1/2} v_k = v_k/\sqrt{l_k}$, so that w_k is as claimed.

REFERENCES

ASH, R. B. & GARDNER, M. F. (1975). *Topics in Stochastic Processes*. New York: Academic Press.

BILODEAU, M. & BRENNER, D. (1999). *Theory of Multivariate Statistics*. New York: Springer-Verlag.

BOENTE, G. & FRAIMAN, R. (1999). Comment on a paper by Locantore et al. *Test* **8**, 28–35.

CUEVAS, A., FEBRERO, M. & FRAIMAN, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Comp. Statist.* **22**, 481–96.

DAUXOIS, J., POUSSE, A. & ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Mult. Anal.* **12**, 136–54.

FRAIMAN, R. & MUNIZ, G. (2001). Trimmed means for functional data. *Test* **10**, 419–40.

GASSER, T., SROKA, L. & JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–33.

GERVINI, D. (2006). Free-knot spline smoothing for functional data. *J. R. Statist. Soc.* B **68**, 671–87.

GOWER, J. C. (1974). The mediancentre. *Appl. Statist.* **23**, 466–70.

HALL, P. & HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Statist. Soc.* B **68**, 109–26.

HALL, P., MULLER, H.-G. & WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–517.

KEMPERMAN, J. H. B. (1987). The median of a finite measure on a Banach space. In *Statistical Analysis Based on the L_1-norm and Related Methods* (*Neuchâtel, 1987*), Ed. Yadolah Dodge, pp. 217–30. Amsterdam: North Holland.

KNEIP, A. (1999). Comment on a paper by Locantore et al. *Test* **8**, 50–4.

LOCANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. & COHEN, K. L. (1999). Robust principal components for functional data (with Discussion). *Test* **8**, 1–73.

LOPEZ-PINTADO, S. & ROMO, J. (2007). Depth-based inference for functional data. *Comp. Statist. Data Anal.* **51**, 4957–68.

LUENBERGER, D. G. (1969). *Optimization by Vector Space Methods*. New York: John Wiley.

MALFAIT, N. & RAMSAY, J. O. (2003). The historical functional linear model. *Can. J. Statist*. **31**, 115–28.

MARDEN, J. (1999). Some robust estimates of principal components. *Statist. Prob. Lett.* **43**, 349–59.

MARONNA, R. A., MARTIN, R. D. & YOHAI, V. J. (2006). *Robust Statistics: Theory and Methods*. New York: John Wiley.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis,* 2nd ed. New York: Springer-Verlag.

VARDI, Y. & ZHANG, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Nat. Acad. Sci. U.S.A.* **97**, 1423–6.

YAO, F. & LEE, T. C. M. (2006). Penalized spline models for functional principal component analysis. *J. R. Statist. Soc.* B **68**, 3–25.

	$p=2$			$p=5$		
$\nu \diagdown r$	1	2	4	1	2	4
5	1.18	1.14	1.03	1.33	1.24	0.91
10	0.93	0.90	0.81	1.05	0.98	0.72
15	0.87	0.85	0.76	0.99	0.92	0.68
20	0.84	0.82	0.73	0.96	0.89	0.65

Table 1: Values of ARE of the median for t distributions with ν degrees of freedom and eigenvalue sequences $\lambda_k = 1/k^r$, $k = 1, \ldots, p$.

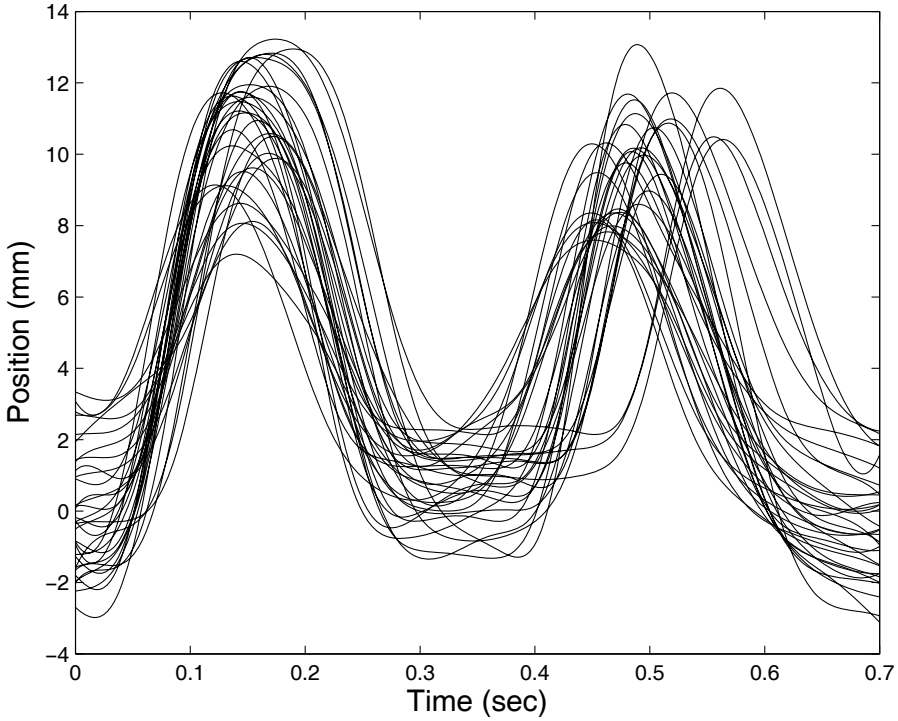

Figure 1: Lip-movement data. Smoothed lower-lip trajectories of an individual pronouncing ‘bob’ 32 times.

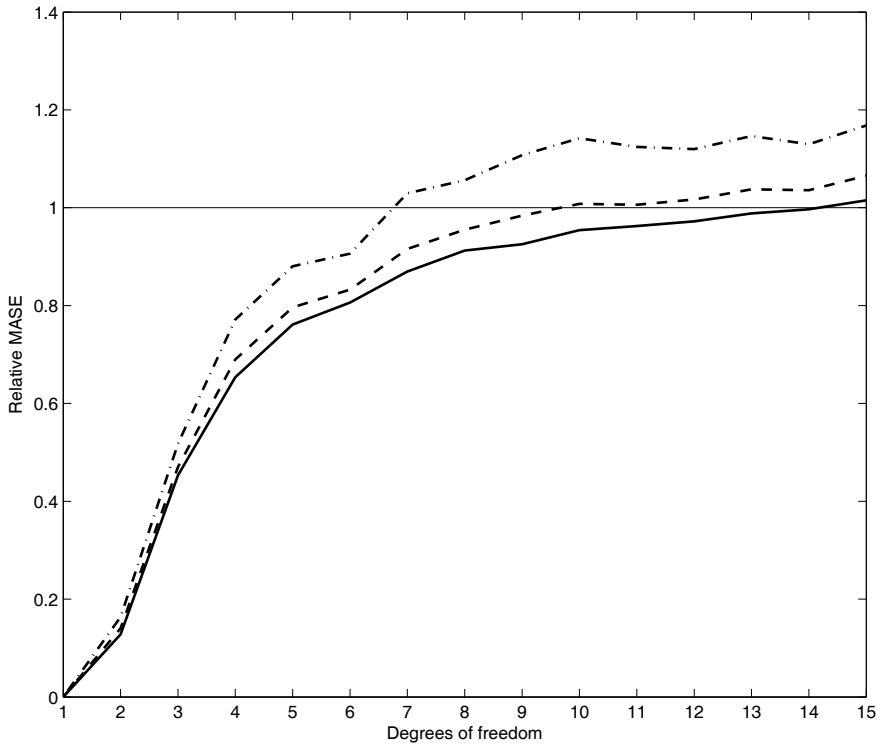

Figure 2: Simulated data. Relative mean average squared error of the spatial median with respect to the sample mean under multivariate t models, for eigenvalue sequences $\lambda^{(1)}$, solid line, $\lambda^{(2)}$, dashed line, and $\lambda^{(3)}$, dash-dotted line.

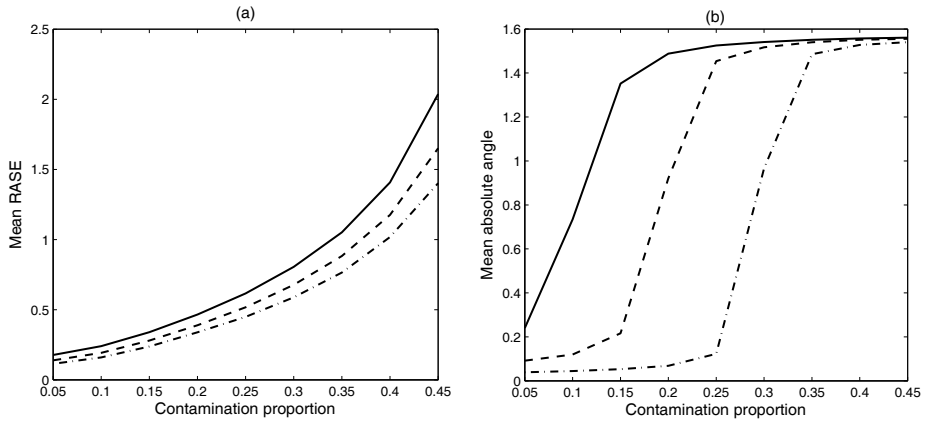

Figure 3: Simulated data from point-mass contaminated Normal models. (a) mean root average squared error of the spatial median, (b) mean absolute angle of the first spherical principal component, for eigenvalue sequences $\lambda^{(1)}$, solid line, $\lambda^{(2)}$, dashed line, and $\lambda^{(3)}$, dash-dotted line.

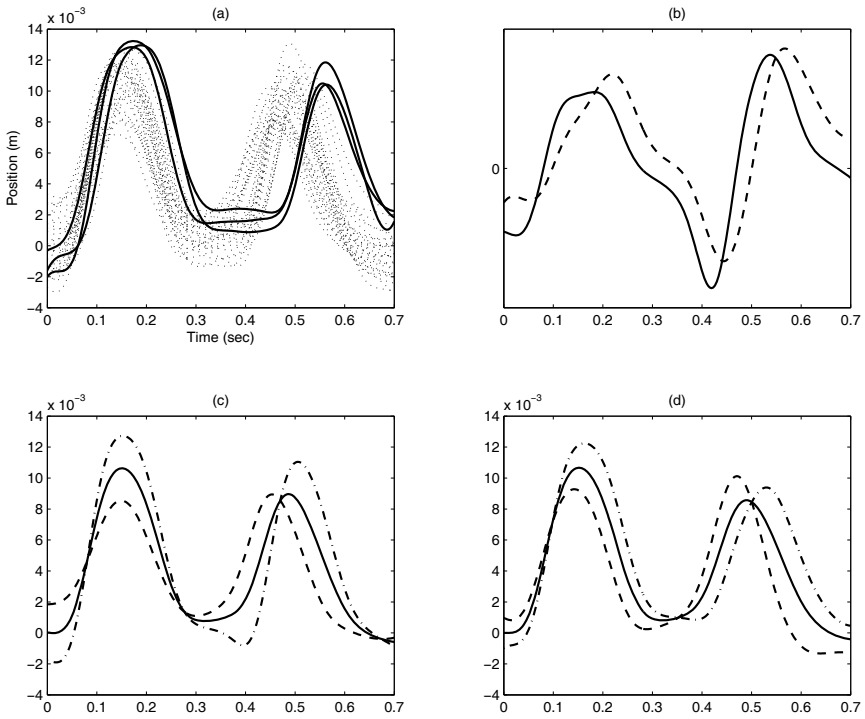

Figure 4: Lip movement data. (a) Smooth sample curves, with three atypical curves highlighted in solid lines. (b) First sample principal component, dashed line, and first spherical component, solid line. (c) Median, solid line, plus and minus a constant times the first spherical principal component, dash-dot and dashed lines respectively. (d) Sample mean, solid line, plus and minus a constant times the first sample principal component, dash-dotted and dashed lines respectively.

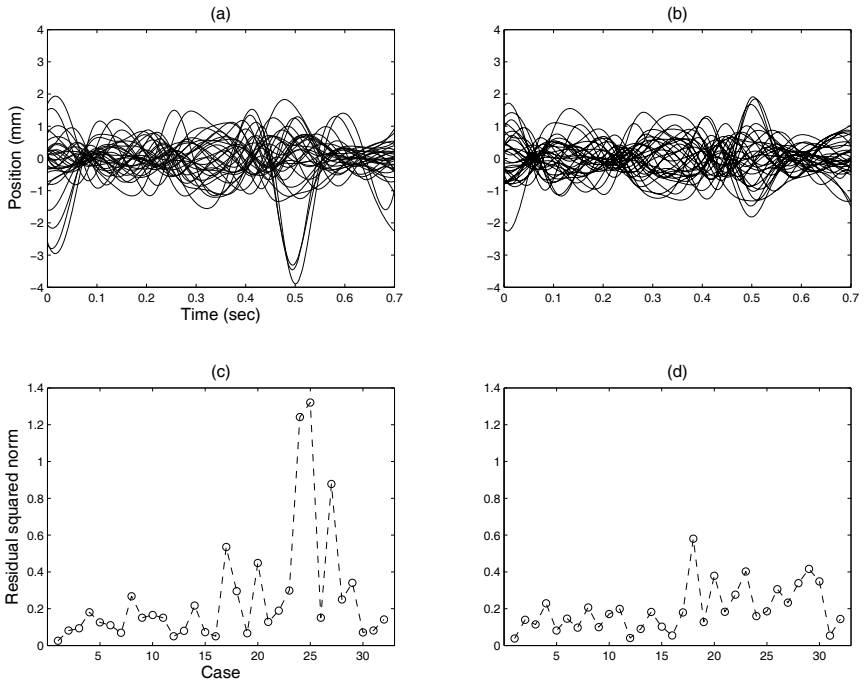

Figure 5: Lip movement data. Residual curves based on robust estimators are shown in (a) and based on the classical estimators are shown in (b). Squared norms of the residuals for each sample curve are shown in (c) for the robust estimators and in (d) for the classical estimators.