Statistica Sinica **22** (2012), 1639-1660
doi:http://dx.doi.org/10.5705/ss.2010.282

OUTLIER DETECTION AND TRIMMED ESTIMATION FOR GENERAL FUNCTIONAL DATA

Daniel Gervini

University of Wisconsin–Milwaukee

Abstract: This article introduces trimmed estimators for the mean and covariance function of general functional data. The estimators are based on a new measure of "outlyingness" or data depth that is well defined on any metric space, although we focus on Euclidean spaces. We compute the breakdown point of the estimators and show that the optimal breakdown point is attainable for the appropriate choice of tuning parameters. The small-sample behavior of the estimators is studied by simulation, and we show that they have better outlier-resistance properties than alternative estimators. This is confirmed by two data applications that also show that the outlyingness measure can be used as a graphical outlier-detection tool in functional spaces where visual screening of the data is difficult.

Key words and phrases: Breakdown point, data depth, robust statistics, stochastic processes.

1. Introduction

Many statistical applications involve data that does not fit into the classical univariate or multivariate frameworks, for example, growth curves, spectral curves, and time-dependent gene expression profiles. These are samples of functions, rather than numbers or vectors. The statistical analysis of function-valued data has received a lot of attention in recent years (see e.g., Ramsay and Silverman (2002); Serfling (2006), and references therein). However, most of the work on Functional Data Analysis has focused on univariate curves; in many applications, the sample functions are not univariate.

Consider, for example, excitation-emission matrices (EEMs), which are common in Chemometrics. When certain fluorescent substances are exposed to light of wavelength s, they emit light at wavelength t. The resulting light intensity X is then a bivariate function. Mortensen and Bro (2006) analyzed a collection of 338 such surfaces; the logarithms of four of them are shown in Figure 1. A movie showing the 338 surfaces in quick succession is available on the author's website; it is clear in this movie that there are some atypical surfaces in the data set. For example, Figure 1 shows that even after taking logarithms, the surface on the lower right corner is out of line compared to the other three. These

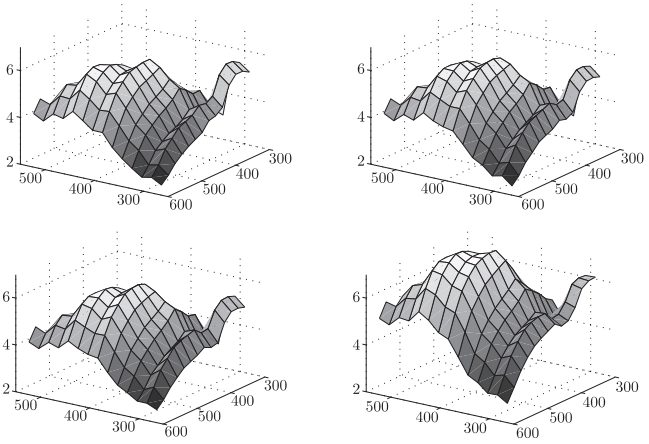

Figure 1. Excitation-Emission Matrices. Four samples of log-EEMs.

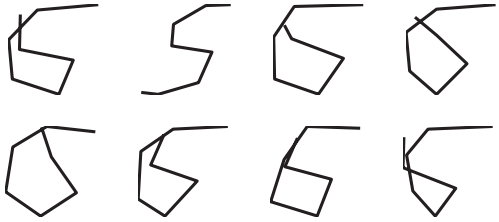

Figure 2. Handwritten Digits Example. Eight samples of the number "five".

atypical objects are what we refer to as "outliers" ; that is, objects that depart from the main modes of variability of the majority of the data. Note that since each functional object typically consists of many measurements taken at different time points (or wavelength points, in this case), a few of those measurements could be outlying without the whole surface being atypical. That kind of isolated measurement error is not of interest here; it has been addressed in the robust smoothing literature (e.g., Shi and Li (1995), Jiang and Mack (2001).)

As a second example, consider a digit recognition problem. Eight handwritten "fives" , from a total of 1,055 samples, are shown in Figure 2. Ignoring the distance between the pen tip and the writing pad, the planar trajectory of the pen tip is a curve over time. In Figure 2 we see that some of the handwritten digits look more like "sixes" than "fives" . The reason for this is going to be explained in Section 5, but it is clear at this point that the sample of "fives" is not homogeneous; it contains either isolated outliers or systematic clusters that are hard to guess a priori.

These examples show two things that are the main motivation for this paper: (i) functional data belonging to complicated spaces are encountered in practice,

and (ii) outliers may be present in a sample but, due to the complexity of the data, visual screening of the data set may be impractical or impossible. The problem of robust estimation in functional spaces has been addressed by some authors, including Locantore et al. (1999), Fraiman and Muniz (2001), Cuevas, Febrero, and Fraiman (2007), Gervini (2008), and López-Pintado and Romo (2009), but all of these papers deal with univariate curves. Some of these methods can be extended to more complex spaces in a more or less straightforward way, but some of them cannot. For example, the methods of Fraiman and Muniz (2001) and López-Pintado and Romo (2009) are based on data-depth notions that require an ordering of the response variables that cannot be extended to vector-valued functions like the handwritten digits in an obvious way. On the other hand, the projection-based methods discussed in Cuevas, Febrero, and Fraiman (2007) and the spatial median and the spherical principal components of Locantore et al. (1999) and Gervini (2008) can be extended to any Euclidean space, though Gervini (2008) found that the breakdown point of the spherical principal components is very low. A third goal here is to develope principal component estimators that are more robust than the spherical principal components, but not as computationally demanding as the projection-based methods of Cuevas, Febrero, and Fraiman (2007).

The estimators we introduce are based on a measure of "outlyingness" that can be defined on any metric space, but we restrict ourselves to Euclidean spaces, where principal components can also be defined. These estimators are easy to compute and turn out to have very good robustness properties. We prove in Section 3 that they can attain the optimal 50% breakdown point. In our simulation study (Section 4) they outperform most of the alternative estimators cited above. The paper also studies other theoretical properties in Section 3, and analyzes in more detail the two applications mentioned above (Section 5). Proofs of the theoretical results and an additional data application can be found in a technical supplement available on the author's webpage.

2. Trimmed Estimators Based on Interdistances

2.1. A measure of "outlyingness"

Let $\{X_1, \ldots, X_n\}$ be a sample in a linear space $\mathcal{H}$ endowed with an inner product $\langle \cdot, \cdot \rangle$, for instance, $\mathcal{L}^2(\mathbb{R})$ with its canonical inner product $\langle f, g \rangle = \int fg$. The inner product induces the norm $\|f\| = \langle f, f \rangle^{1/2}$ in $\mathcal{H}$, and this norm induces the distance function $d(f, g) = \|f - g\|$, so any Euclidean space is a metric space. Let us consider the set of interdistances $\{d(X_i, X_j)\}$. An observation X_i can be seen as an outlier if it's far from *most* other observations (not necessarily *all* of them, because outliers sometimes form clusters). Given $\alpha \in [0, 1]$, we define the α-radius r_i as the distance between X_i and the $\lceil \alpha n \rceil$-th closest observation, where

$\lceil x \rceil$ denotes the integer closest to x from above. This is the radius of the smallest ball centered at X_i that covers $100\alpha\%$ of the observations. Intuitively, r_i is small where the data is dense and large where the data is sparse (see Proposition 4 in Section 3). Therefore, the rank of r_i in the set $\{r_1, \ldots, r_n\}$ is a measure of the "outlyingness" of X_i: the more isolated X_i is, the larger is r_i compared to the other radii.

In principle the coverage parameter α can be any number between 0 and 1, but note that if there is a tight cluster of n^* outliers and $\lceil \alpha n \rceil < n^*$, then the radii of the outliers is small, perhaps even smaller than the radii of the "good" observations, which renders them useless for our purposes. Therefore α must be large enough that at least one good observation is captured by r_i whenever X_i is an outlier. Since n^* can be as large as $n/2$, in general only $\alpha \geq .50$ guarantees this. On the other hand, taking $\alpha > .50$ may cause the opposite problem: that an outlying observation is always captured by r_i when X_i is not an outlier, making the radii of the "good" observations too large (the formalization of these heuristics constitute the proof of Proposition 2 in Section 3). For these reasons we take $\alpha = 0.50$ for estimation purposes. However, for outlier-screening purposes it is instructive to see boxplots and histograms of the radii for values of α less than .50, outliers tend to emerge clearly and consistently as α increases.

At this point some comments about the actual computation of the interdistances are in order. First, note that all the interdistances can be computed from the set of inner products $\{\langle X_i, X_j \rangle\}$, since $d^2(X_i, X_j) = \langle X_i, X_i \rangle + \langle X_j, X_j \rangle - 2\langle X_i, X_j \rangle$. It is easy to compute the pairwise inner products when the sample objects have been pre-smoothed, or even if they have not been pre-smoothed but they were sampled on a regular common grid without much random error. In that case, a basic numerical integration method such as the trapezoidal rule gives accurate results (see Gervini (2008, Thm. 1)). If the X_is were sampled on sparse and irregular grids, perhaps with a different grid for each individual, then it is not possible to estimate all pairwise inner products and this method cannot be applied (the other methods mentioned in the introduction cannot be applied either, since they are based on pre-smoothed data).

2.2. Trimmed estimators

In addition to being useful outlying-screening tools, the radii can be used to construct robust estimators of the mean, the covariance function, and the principal components of the process under consideration. For a stochastic process X in $\mathcal{H}$ with $E(\|X\|^2) < \infty$, the mean operator $\mathfrak{M}$ and the covariance operator $\mathfrak{C}$ are defined as $\mathfrak{M} : \mathcal{H} \to \mathbb{R}$ given by $\mathfrak{M}f = E(\langle f, X \rangle)$, and $\mathfrak{C} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is given by $\mathfrak{C}(f, g) = \mathrm{cov}(\langle f, X \rangle, \langle g, X \rangle)$ (these quantities are well defined because $\langle f, X \rangle$ and $\langle g, X \rangle$ are real-valued random variables with finite variances for any f and

g in $\mathcal{H}$). By the Riesz Representation Theorem there exists a unique $\mu \in \mathcal{H}$ such that $\mathfrak{M}f = \langle f, \mu \rangle$, which we call $E(X)$ (this is one way to define the expectation of a stochastic process in a Euclidean space.)

In a Euclidean space it is also possible to define principal directions of variability, or principal components. The first principal component of X is the $\phi_1 \in \mathcal{H}$ that maximizes $\mathrm{var}(\langle f, X \rangle)$ among $f \in \mathcal{H}$ with $\|f\| = 1$; the second principal component is the $\phi_2 \in \mathcal{H}$ that maximizes $\mathrm{var}(\langle f, X \rangle)$ among $f \in \mathcal{H}$ with $\|f\| = 1$ and $\langle f, \phi_1 \rangle = 0$; and so on. It can be shown (Gohberg, Goldberg, and Kaashoek (2003, Chap. IV)) that the principal components are eigenfunctions of the covariance operator and they are countable; that is, $\mathfrak{C}(\phi_k, \cdot) = \lambda_k \langle \phi_k, \cdot \rangle$ with $\lambda_k \in \mathbb{R}$ and $\lambda_k \geq 0$.

The classical estimators of these quantities (the sample mean, covariance, and principal components) are not resistant to outliers. As a more robust alternative we propose trimmed estimators based on the radii. Specifically, given a trimming proportion $\beta \in [0, .50]$ we take $w(X_i) = \mathbb{I}\{r_i < r_{(\lceil (1-\beta)n \rceil)}\}$ and

$$\hat{\mu} = \frac{1}{\sum_{i=1}^n w(X_i)} \sum_{i=1}^n w(X_i) X_i, \tag{2.1}$$

$$\widehat{\mathfrak{C}}(f, g) = \frac{1}{\sum_{i=1}^n w(X_i)} \sum_{i=1}^n w(X_i) \langle X_i - \hat{\mu}, f \rangle \langle X_i - \hat{\mu}, g \rangle. \tag{2.2}$$

These are "hard-trimmed" estimators, where a 0-1 weight function is used. More generally, we can define weights of the form $w(X_i) = g(\mathrm{rank}(r_i)/n)$, where $g : [0, 1] \to \mathbb{R}^+$ is a bounded, non-negative and non-increasing function such that $g(t) > 0$ for $t < 1 - \beta$ and $g(t) = 0$ for $t \geq 1 - \beta$. "Soft-trimming" weights are obtained with a smooth function g such as

$$g(r) = \begin{cases} 1, & 0 \leq r \leq a, \\ (r - b)\left[\frac{1}{(a-b)} + \frac{(r-a)\{2r-(a+b)\}}{(b-a)^3}\right], & a \leq r \leq b, \\ 0, & r \geq b, \end{cases} \tag{2.3}$$

where $a = 1 - \beta_1$ for some $\beta_1 > \beta$, and $b = 1 - \beta$. This function downweights the largest $100\beta_1\%$ radii, and cuts off the largest $100\beta\%$ radii completely; we can take, for example, $\beta_1 = 0.50$ and $\beta = 0.20$.

Trimmed estimators based on various measures of data depth have been proposed in other contexts, in particular in multivariate analysis (Fraiman and Meloche (1999), Liu, Parelius, and Singh (1999), Serfling (2006), Zuo and Serfling (2000), Zuo, Cui, and He (2004)). The behavior of these estimators varies according to the specific data-depth measure that is being used but, as a general rule, their outlier resistance increases as β increases and their efficiency decreases

as β increases (see e.g., Stigler (1973); Van der Vaart (1998, Chap. 22); Maronna, Martin, and Yohai (2006, Chap. 2)). Since there is a trade-off between robustness and efficiency, we recommend choosing β in a data-driven way: a histogram of the radii usually gives a good idea of the proportion of outliers in the sample, and this value could be used as β. A more objective alternative, suggested by a referee, is to fit a mixture of two Gamma distributions to the sample of radii and take as β the proportion of observations in the smaller group. If instead of these data-driven choices of β the user prefers to use a fixed β, our simulations showed that "soft-trimming" weights like (2.3) are preferrable to "hard-trimming" weights (see Section 4).

Just as the radii (and therefore the weights $w(X_i)$) depend on the data only through the inner products $\{\langle X_i, X_j\rangle\}$, the principal components of (2.2) can be computed entirely from the inner products. As in Gervini (2008) and Jolliffe (2002, Chap. 3.5): if $\tilde{w}_i = w(X_i)/\sum_{i=1}^n w(X_i)$, then $\hat{\phi}_k = \sum_{i=1}^n (c_{ki}/l_k^{1/2})\tilde{w}_i^{1/2}(X_i - \hat{\mu})$ and $\hat{\lambda}_k = l_k$, where $\mathbf{c}_k$ is the kth unit-norm eigenvector of the matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ with elements $G_{ij} = \langle \tilde{w}_i^{1/2}(X_i - \hat{\mu}), \tilde{w}_j^{1/2}(X_j - \hat{\mu})\rangle$, and l_k is the kth eigenvalue (the G_{ij}s can be expressed entirely in terms of the $\langle X_i, X_j\rangle$s and the $\tilde{w}_i$s, after some algebra). The applicability of these estimators is then limited only by the possibility of computing all pairwise inner products. This is generally not possible if the data objects were sparsely and irregularly sampled, and alternative estimation methods must be sought. For instance, the reduced-rank t-model estimators of Gervini (2008), originally developed for sparsely sampled univariate curves, can be extended to more general functional spaces, but this is clearly outside the scope of this paper.

3. Properties of the Estimators

3.1. Finite-sample properties

Location and scatter estimators must satisfy certain equivariance properties to be proper measures of "location" and "scatter" . A location estimator must be translation equivariant: if $\hat{\mu}$ is the estimator based on the sample $\{X_1, \ldots, X_n\}$, then the estimator based on the sample $\{X_1 + c, \ldots, X_n + c\}$, with $c \in \mathcal{H}$, must be $\hat{\mu} + c$. Other desirable properties are scale and rotation equivariance: if $\hat{\mu}$ is the estimator based on the sample $\{X_1, \ldots, X_n\}$, then the estimator based on the sample $\{a\mathfrak{U}X_1, \ldots, a\mathfrak{U}X_n\}$, with $\mathfrak{U}$ a unitary operator and $a \in \mathbb{R}$, must be $a\mathfrak{U}\hat{\mu}$ (a unitary operator is $\mathfrak{U} : \mathcal{H} \to \mathcal{H}$ such that $\|\mathfrak{U}f\| = \|f\|$ for every $f \in \mathcal{H}$). A scatter estimator, on the other hand, must be translation invariant and rotation and scale equivariant in the following sense: if $\widehat{\mathfrak{C}}(\cdot,\cdot)$ is the covariance estimator based on the sample $\{X_1, \ldots, X_n\}$, then the covariance estimator based on the sample $\{a\mathfrak{U}X_1, \ldots, a\mathfrak{U}X_n\}$ must be $a^2\widehat{\mathfrak{C}}(\mathfrak{U}^*\cdot, \mathfrak{U}^*\cdot)$, where $\mathfrak{U}^*$ is the adjoint of $\mathfrak{U}$.

The rotation equivariance of $\widehat{\mathfrak{C}}$ automatically implies rotation equivariance of the principal component estimators obtained from $\widehat{\mathfrak{C}}$.

Our trimmed estimators satisfy these properties, as shown in Proposition 1. This is a consequence of the translation and rotation invariance of the radii, and therefore of the weights $w(X_i)$ (which, in addition, are scale invariant). Note that translation, scale and rotation invariance are properties that any "outlyingness" measure should satisfy: if an observation is considered an outlier for a given dataset, the same observation should be considered an outlier if the dataset is simply translated, rotated or re-scaled.

Proposition 1. Let $X_1, \ldots, X_n$ be a sample in $\mathcal{H}$, $a \neq 0$ a scalar, $b \in \mathcal{H}$, and $\mathfrak{U}$ a unitary operator. Let $\tilde{X}_i = a\mathfrak{U}X_i + b$ with $\{\tilde{d}_{ij}\}$ and $\{\tilde{r}_i\}$ the corresponding interdistances and radii, $\widehat{\tilde{\mu}}$, $\widehat{\tilde{\mathfrak{C}}}$, $\{\widehat{\tilde{\lambda}}_k\}$, and $\{\widehat{\tilde{\phi}}_k\}$ the corresponding estimators.

1. $\tilde{d}_{ij} = |a| d_{ij}$ for all i and j, and $\tilde{r}_i = |a|\, r_i$ for all i, so $\operatorname{rank}(\tilde{r}_i) = \operatorname{rank}(r_i)$ and $w(\tilde{X}_i) = w(X_i)$ for all i.
2. $\widehat{\tilde{\mu}} = a\mathfrak{U}\hat{\mu} + b$.
3. $\widehat{\tilde{\mathfrak{C}}}(f, g) = a^2\widehat{\mathfrak{C}}(\mathfrak{U}^* f, \mathfrak{U}^* g)$ for all f and g, so $\widehat{\tilde{\lambda}}_k = a^2\hat{\lambda}_k$ and $\widehat{\tilde{\phi}}_k = \mathfrak{U}\hat{\phi}_k$ for all k.

The robustness of an estimator is usually measured by the breakdown point (Donoho and Huber (1983)). The finite-sample breakdown point is the largest proportion of outliers that an estimator can tolerate. Here, given a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, let $\tilde{\mathcal{X}}_k$ be a contaminated sample obtained from $\mathcal{X}$ by changing k points arbitrarily; then the finite-sample breakdown point of $\hat{\mu}$ is $\varepsilon_n^*(\hat{\mu}) := k^*/n$, where k^* is the smallest k for which there is a sequence of contaminated samples $\{\tilde{\mathcal{X}}_k^{(m)}\}_{m\geq 1}$ such that $\|\hat{\mu}^{(m)}\| \underset{m\to\infty}{\longrightarrow} \infty$. The finite-sample breakdown point of $\widehat{\mathfrak{C}}$ is defined analogously. The asymptotic breakdown point is the limit of $\varepsilon_n^*(\hat{\mu})$ as n goes to infinity, if the limit exists. The highest asymptotic breakdown point attainable by an equivariant estimator is 0.50 (Lopuhaä and Rousseeuw (1991)).

Proposition 2. Suppose $w(X_i) = g(\operatorname{rank}(r_i)/n)$, with g satisfying the conditions given in Section 2. If $\alpha \leq 0.50$, $\lceil \alpha n \rceil \geq 3$, and $\beta \leq 0.50$, then $\varepsilon_n^*(\hat{\mu}) = \varepsilon_n^*(\widehat{\mathfrak{C}}) = \min(\lceil \alpha n \rceil, \lfloor \beta n \rfloor + 2)/n$, which tends to $\min(\alpha, \beta)$ when n goes to infinity.

This proposition shows that the asymptotic breakdown point of the trimmed estimators is $\min(\alpha, \beta)$. Then, if $\alpha = 0.50$, the breakdown point is just the trimming proportion β, and the optimal breakdown point can be attained with $\beta = 0.50$. In practice, though, such estimators are very inefficient when the actual proportion of outliers is much less than 50%, as we show by simulation in Section 4; a better alternative is to use "soft" trimming, as explained in Section 2.

3.2. Population versions and properties

The estimators (2.1) and (2.2) can be generalized to any probability measure P on $\mathcal{H}$, of which (2.1) and (2.2) can be seen as particular cases obtained for $P = P_n$, the empirical measure on the sample $\{X_1, \ldots, X_n\}$. One reason this generalization is useful is that it allows us to study the consistency of the estimators: since $P_n \to P$ when the X_is are i.i.d. with distribution P, under certain conditions (Fernholz (1983); Van der Vaart (1998, Chap. 20)) $\hat{\mu}$ and $\widehat{\mathfrak{C}}$ converge in probability to their respective population versions μ_P and $\mathfrak{C}_P$.

The derivation of μ_P and $\mathfrak{C}_P$ is as follows. Let X be a stochastic process with distribution P. Define $F_P(t; v) = P\{\|X - v\| \leq t\}$ for each $v \in \mathcal{H}$. The radius of the smallest ball centered at v with probability α is $r_P(v) = F_P^{-1}(\alpha; v)$, where $F_P^{-1}(\alpha; v) := \min\{t : F_P(t; v) \geq \alpha\}$ is the usual quantile function. Then $r_P(X)$ is the α-radius around X and, if $G_P(t) := P\{r_P(X) \leq t\}$, the weight function $w_P(v)$ has the form $w_P(v) = g[G_P\{r_P(v)\}]$, with g as in Section 2. Then

$$\mu_P = \frac{E_P\{w_P(X)X\}}{E_P\{w_P(X)\}},$$
$$\mathfrak{C}_P(f, g) = \frac{E_P\{w_P(X)\langle X - \mu_P, f\rangle\langle X - \mu_P, g\rangle\}}{E_P\{w_P(X)\}}.$$

The eigenvalues and eigenfunctions of $\mathfrak{C}_P$ are denoted by $\lambda_{k,P}$ and $\phi_{k,P}$, respectively.

A first result is that μ_P and $\mathfrak{C}_P$ are well-defined for *any* probability distribution P on $\mathcal{H}$, even if $\|X\|$ does not have finite moments of any order.

Proposition 3. For any $\alpha > 0$ there is a constant $K_{\alpha,P} \geq 0$ such that $\|v\| \leq r_P(v) + K_{\alpha,P}$ for all $v \in \mathcal{H}$. Therefore, if $\beta > 0$ then $E_P\{w_P(X)\|X\|^k\} < \infty$ for any $k \geq 0$.

The next proposition shows that $r_P(v)$ is really a measure of outlyingness, in the sense that $r_P(v)$ is larger in regions of $\mathcal{H}$ where P is less concentrated.

Proposition 4. If v and w are two points in $\mathcal{H}$ such that $P(B_\delta(v)) \geq P(B_\delta(w))$ for all $\delta > 0$ (where $B_\delta(v)$ denotes the ball with center v and radius δ), then $r_P(v) \leq r_P(w)$.

The equivariance of $\hat{\mu}$ and $\widehat{\mathfrak{C}}$ carries over to μ_P and $\mathfrak{C}_P$ (the proof is given in the technical supplement). A consequence of the translation equivariance of μ_P is the following.

Proposition 5. If $X - \mu_0$ and $\mu_0 - X$ are identically distributed, then $\mu_P = \mu_0$.

To study the population versions of the trimmed principal components, suppose X admits, with probability 1, the decomposition

$$X = \mu_0 + \sum_{k\in\mathcal{I}} \lambda_{0k}^{1/2} Z_k \phi_{0k}, \tag{3.1}$$

where $\mu_0 \in \mathcal{H}$, the Z_ks are real random variables, $\{\phi_{0k}\} \subset \mathcal{H}$ is an orthonormal system, and $\{\lambda_{0k}\}$ is a strictly positive non-increasing sequence with $\sum_{k\in\mathcal{I}} \lambda_{0k} < \infty$; the set of indices $\mathcal{I}$ is countable but may be finite or infinite. This decomposition holds, for instance, if $E(\|X\|^2) < \infty$, and is known as the Karhunen–Loève decomposition (Ash and Gardner (1975, Chap. 1.4)). In that case $E(X) = \mu_0$, the ϕ_{0k}'s and the λ_{0k}'s are the eigenfunctions and eigenvalues of the covariance operator, and $Z_k = \langle X - \mu_0, \phi_{0k}\rangle / \lambda_{0k}^{1/2}$ are uncorrelated with $E(Z_k) = 0$ and $\mathrm{var}(Z_k) = 1$.

Expansion (3.1) also holds in some situations where $E(\|X\|^2) = \infty$, and gives a meaningful notion of "heavy-tailed distributions" for functional spaces. For instance, if the Z_k's in (3.1) are independent, Kolmogorov's Three Series Theorem (Gikhman and Skorokhod (2004, p.384)) implies that $\sum_{k\in\mathcal{I}} \lambda_{0k}^{1/2} Z_k \phi_{0k}$ converges almost surely in $\mathcal{H}$ if and only if $\sum_{k\in\mathcal{I}} P(\lambda_{0k} Z_{0k}^2 > c) < \infty$ for every $c > 0$. The latter is satisfied whenever the λ_{0k}'s go to zero fast enough, even if the Z_k's do not have finite moments of any order. For example, if the Z_k's have a Cauchy distribution,

$$\sum_{k\in\mathcal{I}} P(\lambda_{0k} Z_k^2 > c) \le \sum_{k\in\mathcal{I}} \frac{2}{\pi} \left(\frac{\lambda_{0k}}{c}\right)^{1/2},$$

and the right-hand side is finite for any $c > 0$ as long as $\sum_{k\in\mathcal{I}} \lambda_{0k}^{1/2} < \infty$.

Model (3.1) can be used to characterize the two types of outliers that may be present in a functional sample. One type is observations that satisfy (3.1) but with extreme values of the Z_k's; these are called *intrinsic* outliers since they belong to the space generated by the ϕ_{0k}'s. Another type that does not follow (3.1) at all, is an *extrinsic* outlier that falls outside the subspace of $\mathcal{H}$ where the "good" data lives. To exemplify the difference, suppose that a sample of curves shows a prominent feature, such as a peak, and the leading principal component ϕ_{01} explains variability around this peak (a usual situation). An intrinsic outlier would be a curve with an unusual peak (either too flat or too sharp compared to the other curves), whereas an extrinsic outlier would be an observation with a peak at a location where the rest of the data shows no such feature. Our estimators can handle both types of outliers, since the interdistances make no distinction between the two types (although extrinsic outliers are easier to spot). The outliers considered in the simulations (Section 4) are intrinsic, while those in the examples (Section 5) are mostly extrinsic.

Note that under (3.1) the interdistances satisfy $d_{ij}^2 = \sum_{k\in\mathcal{I}} \lambda_{0k}(Z_{ki} - Z_{kj})^2$, so the distribution of the d_{ij}s (and therefore of the radii) depends entirely on the Z_k's and the λ_{0k}'s, not on μ_0 or the ϕ_{0k}'s. This implies the following.

Proposition 6. If (3.1) holds with independent and symmetrically distributed Z_k's, then

$$\mathfrak{C}_P(f,g) = \sum_{k\in\mathcal{I}} \tilde{\lambda}_{0k}\langle\phi_{0k}, f\rangle\langle\phi_{0k}, g\rangle \tag{3.2}$$

with

$$\tilde{\lambda}_{0k} = \frac{E_P\{w_P(X)|\langle X-\mu_0, \phi_{0k}\rangle|^2\}}{E_P\{w_P(X)\}} = \lambda_{0k}\,\frac{E_P\{w_P(X)Z_k^2\}}{E_P\{w_P(X)\}}. \tag{3.3}$$

The sequence $\{\tilde{\lambda}_{0k}\}$ is strictly positive but not necessarily decreasing, and it only depends on the distribution of $\{\lambda_{0k}^{1/2}Z_k\}$. In addition, if the Z_k's are identically distributed, then $\lambda_{0j} = \lambda_{0k}$ implies $\tilde{\lambda}_{0j} = \tilde{\lambda}_{0k}$, so that the multiplicity of the eigenvalues is preserved.

This result implies that the set of principal components of $\mathfrak{C}_P$, $\{\phi_{k,P}\}$, coincides with the set $\{\phi_{0k}\}$, but it cannot be said in general that $\phi_{k,P} = \phi_{0k}$ for each k, because the sequence $\{\tilde{\lambda}_{0k}\}$ is not necessarily decreasing. The reason is that although $\lambda_{0k}Z_k^2$ is stochastically greater than $\lambda_{0j}Z_j^2$ when $\lambda_{0k} > \lambda_{0j}$ and the Z_ks are identically distributed, this does not imply that $w_P(X)\lambda_{0k}Z_k^2$ is stochastically greater than $w_P(X)\lambda_{0j}Z_j^2$ in general. However, (3.3) does imply that $\tilde{\lambda}_{0k} > 0$ if and only if $\lambda_{0k} > 0$, so the dimension of the model is preserved.

4. Simulations

We ran a Monte Carlo study to assess the comparative performance of the following estimators.

- The sample mean and sample principal components.
- The spatial median and spherical principal components (Locantore et al. (1999), Gervini (2008)). The spatial median is defined as the $\hat{\mu}$ that minimizes $\sum_{i=1}^n \|X_i - \mu\|$, and the spherical principal components are defined as the principal components of the normalized sample $\{(X_i - \hat{\mu}) / \|X_i - \hat{\mu}\|\}$, the eigenfunctions of the covariance operator

$$\widehat{\mathfrak{C}}(f,g) := \frac{1}{n}\sum_{i=1}^n \langle\frac{X_i - \hat{\mu}}{\|X_i - \hat{\mu}\|}, f\rangle\langle\frac{X_i - \hat{\mu}}{\|X_i - \hat{\mu}\|}, g\rangle.$$

- Trimmed estimators based on the deviations $\|X_i - \hat{\mu}\|$, where $\hat{\mu}$ is the spatial median, with 20% and 50% trimming. The observations with the largest 20% or 50% deviations were eliminated and the mean and principal components of the remaining data is computed.

- Trimmed estimators based on h-depth (Cuevas, Febrero, and Fraiman (2007)), with 20% and 50% trimming. The h-depth of a datum z is defined as

$$\hat{f}_h(z) = \frac{1}{n}\sum_{i=1}^{n} K_h(\|z - X_i\|),$$

  where $K_h(t) = h^{-1}K(t/h)$ for some kernel function K. Following Cuevas, Febrero, and Fraiman (2007), we take K as the Gaussian density and h as the 20th percentile of the set of L_2-interdistances (there is no clear rationale for this choice but we used their tuning parameters in order to make our simulation results comparable). Note that a small, not a large, value of $\hat{f}_h(X_i)$ indicates that X_i is an outlier, so we trim those observations with small value of $\hat{f}_h$. In their extensive simulations, these estimators outperformed the estimators of Fraiman and Muniz (2001) and some projection-based estimators, so we did not include the latter in our simulations.
- Trimmed estimators based on band depth (López-Pintado and Romo (2009)), with 20% and 50% trimming. The band depth is computed as follows. Given real-valued functions $f_1, \ldots, f_k$ defined on some interval $I \subseteq \mathbb{R}$, with $k \geq 2$, the k-band spanned by these functions is

$$B(f_1, \ldots, f_k) := \{(t, y) : t \in I, y \in [\min_{1\leq i\leq k} f_i(t), \max_{1\leq i\leq k} f_i(t)]\}.$$

  For a given curve z and a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, let $BD_k(z; \mathcal{X})$ be the average number of sample k-bands that contain the graph of z:

$$BD_k(z; \mathcal{X}) = \binom{n}{k}^{-1} \sum_{1\leq i_1<\cdots<i_k\leq n} \mathbb{I}\{G(z) \subseteq B(X_{i_1}, \ldots, X_{i_k})\},$$

  where $G(z) := \{(t, z(t)) : t \in I\}$. The J-depth of the curve z is $D_J(z; \mathcal{X}) = \sum_{k=2}^{J} BD_k(z; \mathcal{X})$. As recommended by López-Pintado and Romo (2009), we use $J = 3$. Once again, outliers are indicated by small values of $D_J(z; \mathcal{X})$, so we trim the 100β% observations with smallest values of D_J.
- Our trimmed estimators with hard and soft rejection weights. For hard-rejection weights, 20% and 50% fixed trimming was considered, as well as the adaptive β estimated with Gamma mixtures; for the soft-rejection weight (2.3), the parameters $\beta_1 = 0.50$ and $\beta = 0.20$ were used. In all cases, the radii were computed with $\alpha = 0.50$ (simulations with $\alpha = 0.20$ were also run but not reported here, because the estimator's performance was uniformly worse than that for $\alpha = 0.50$).

The data were generated following (3.1) with $\mu_0 = 0$ and $\phi_{0k}(t) = \sqrt{2}\sin(\pi k t)$, for $t \in [0, 1]$. The Z_k's followed different distributions for each scenario, and

two sequences of eigenvalues were considered: a slow-decaying sequence $\lambda_{0k} = 1/\{k(k+1)\}$ (Model 1), and a fast-decaying sequence $\lambda_{0k} = 1/2^k$ (Model 2). Model 2 is practically a finite-dimensional model, since the first five terms account for 97% of the variability; Model 1, on the other hand, needs 31 terms to account for the same proportion of the variability. For data generation we truncated Model 1 at the 1,000th term and Model 2 at the 10th term, to represent 99.9% of the total variability. The sample size was $n = 50$ in all cases, and the curves were discretized at an equally spaced grid of 100 points. Each sampling situation was replicated 2,000 times. The mean absolute errors, reported in Tables 1 and 2, are accurate to two significant places (we do not report Monte Carlo standard errors for reasons of space and readability).

The Z_k's were (i) non-contaminated Normal data, (ii) outlier-contaminated Normal data, and *(iii)* non-Normal data (specifically, data with heavier tails than Normal). For case (i) we generated i.i.d. Z_k's as $N(0,1)$. For case (ii) we considered two scenarios: to study the robustness of the location estimators, we generated outliers by adding $3\phi_{01}(t)$ to $n\varepsilon$ sample curves (which creates a bias in $\hat{\mu}$); to study the robustness of the estimators of ϕ_1, we generated outliers by adding $3\phi_{02}(t)$ to $n\varepsilon/2$ sample curves and subtracting the same quantity from other $n\varepsilon/2$ sample curves (thus inflating the variability in the direction of ϕ_{02}, creating a bias in $\hat{\phi}_1$ without affecting $\hat{\mu}$). Four values of ε were considered: 0.10, 0.20, 0.30, and 0.40. For case *(iii)* we generated i.i.d. Z_k's with Student's t_ν distribution, with degrees of freedom ν at 1, 2, and 3.

Table 1 reports the mean absolute error $E(\|\hat{\mu} - \mu_0\|)$ for each estimator and each sampling distribution. Since there are 12 estimators and 8 sampling distributions it is hard to make conclusions at a glance. To facilitate comparisons, we ranked the estimators in increasing order of error for each sampling distribution, and computed the average rank for each estimator; this average rank is given in the last column of Table 1. We see that the comparative performance of the estimators was similar under both models. The soft-trimmed estimators showed the best overall performance, since they had the smallest average ranks; at the other extreme, the band-depth 50%-trimmed estimators showed the worst overall performance. Looking in more detail, we see that hard-trimmed estimators with 20% trimming performed poorly for contaminated Normal distributions with $\varepsilon > 0.20$ and for the Cauchy distribution. Among hard-trimmed estimators with 50% trimming, our estimator and the h-depth-based estimator were comparable, the former being better for contaminated Normals with $\varepsilon \geq 0.20$ (and significantly better for $\varepsilon = 0.40$) and the latter being slightly better in the other situations. The soft-trimmed estimator showed an intermediate behavior between the 20% and 50% hard-trimmed estimators; even at the most extreme cases of the 40% contaminated Normal and the Cauchy distribution, its estimation error

Table 1. Simulation Results. Mean absolute errors of location estimators.

		Normal	Contaminated Normal				Student			Ranking
Model	Estimator		10%	20%	30%	40%	t_1	t_2	t_3	
1	Mean	0.134	0.318	0.607	0.906	1.206	74.44	0.406	0.225	9.8
	Median	0.140	0.187	0.320	0.539	0.886	1.00	0.239	0.190	5.6
	Deviation (20%)	0.158	0.165	0.172	0.343	0.739	1.51	0.260	0.204	5.4
	Deviation (50%)	0.181	0.209	0.262	0.325	0.498	1.05	0.275	0.229	7.2
	h-depth (20%)	0.159	0.160	0.171	0.404	0.867	1.54	0.257	0.205	5.8
	h-depth (50%)	0.185	0.190	0.201	0.226	0.373	1.06	0.274	0.229	6.2
	Band depth (20%)	0.150	0.152	0.169	0.370	0.725	34.66	0.451	0.252	6.9
	Band depth (50%)	0.189	0.214	0.268	0.347	0.454	23.63	0.518	0.318	9.8
	Hard trimmed (20%)	0.165	0.164	0.166	0.288	0.634	1.45	0.256	0.208	4.5
	Hard trimmed (50%)	0.197	0.198	0.201	0.210	0.283	1.06	0.281	0.238	6.6
	Hard trimmed (adap.)	0.169	0.157	0.180	0.220	0.410	54.04	0.360	0.223	6.0
	Soft trimmed (20%)	0.175	0.175	0.177	0.198	0.396	1.10	0.253	0.211	4.4
2	Mean	0.132	0.322	0.606	0.906	1.20	9.423	0.384	0.223	10.1
	Median	0.141	0.192	0.317	0.530	0.871	0.348	0.208	0.180	4.6
	Deviation (20%)	0.161	0.164	0.175	0.352	0.728	0.506	0.224	0.194	4.6
	Deviation (50%)	0.183	0.207	0.258	0.318	0.488	0.369	0.235	0.210	6.6
	h-depth (20%)	0.162	0.162	0.174	0.434	0.908	0.547	0.228	0.194	5.2
	h-depth (50%)	0.185	0.193	0.204	0.220	0.427	0.369	0.235	0.215	5.9
	Band depth (20%)	0.149	0.243	0.438	0.695	0.985	1.112	0.281	0.222	8.9
	Band depth (50%)	0.173	0.416	0.630	0.877	1.18	1.166	0.363	0.278	10.9
	Hard trimmed (20%)	0.166	0.166	0.167	0.292	0.640	0.478	0.226	0.195	4.5
	Hard trimmed (50%)	0.195	0.197	0.203	0.208	0.304	0.377	0.240	0.221	6.3
	Hard trimmed (adap.)	0.170	0.158	0.182	0.229	0.429	10.325	0.336	0.216	6.4
	Soft trimmed (20%)	0.175	0.177	0.179	0.199	0.413	0.377	0.221	0.200	4.1

was not much larger than that of the 50% hard-trimmed estimator. The adaptive estimator also showed an intermediate behavior between the 20% and 50% hard-trimmed estimators, except for the Cauchy distribution, for which it does not even seem to be well defined (the same can be said for the estimators based on band depth); this is not entirely surprising, since the Cauchy distribution produces a single heavy-tailed distribution of radii rather than a mixture. All things considered, the soft-trimmed estimator seems to offer the best trade-off between robustness and efficiency.

For the principal component estimators, the mean absolute errors $E(\|\hat{\phi}_1 - \phi_{01}\|)$ are reported in Table 2. Breakdown of $\hat{\phi}_1$ occurs when $\hat{\phi}_1$ is orthogonal to ϕ_{01}, in which case $\|\hat{\phi}_1 - \phi_{01}\| = \sqrt{2}$, so the errors are always bounded. The best-ranked estimators were now the spherical principal components, which was rather unexpected; looking in more detail, we see that this was mostly because of their low errors for t distributions. Their performance for contaminated Normal distributions was not good, showing very large errors for contamination proportions as small as 20%. Our hard-trimmed estimators and the h-depth-based estimators showed comparable performances, although once again the soft-trimmed estimator offered a better trade-off between robustness and efficiency: although it breaks down for the 40%-contaminated Normal, it had much lower estimation errors than the 50%-hard-trimmed estimators for lower levels of contamination and for t distributions (even for the Cauchy). The adaptive estimator did not break down for the 40%-contaminated Normal, but it did for the Cauchy distribution.

The similar behavior of the estimators based on the radii and those based on the h-depth is not accidental, because both are based on metric notions of data depth. The α-radii do have certain advantages over the h-depth: the parameter α that defines the radii is an interpretable quantity, while the parameter h that defines the h-depth is an arbitrary bandwidth with an unknown effect on the estimator's properties. Also, the breakdown point of the estimators based on α-radii is known, while the breakdown point of the estimators based on h-depth is unknown.

In contrast with these metric notions of data depth, the band depth of López-Pintado and Romo (2009) is not based on distances but on the number of "bands" that cover each sample function. Therefore the trimmed estimators based on band depth behave very differently (in fact, much worse) than those based on α-radii or h-depth. Two additional disadvantages of band-depth trimming are that the determination of all the "bands" that cover a given curve is a combinatorial problem, unfeasible for large sample sizes, and that generalizing the concept of "band" to Euclidean spaces beyond univariate curves is not obvious.

Table 2. Simulation Results. Mean absolute errors of first principal component estimators.

		Normal	Contaminated Normal				Student			Ranking
Model	Estimator		10%	20%	30%	40%	t_1	t_2	t_3	
1	Sample p.c.	0.168	1.27	1.36	1.37	1.38	1.25	0.508	0.263	8.5
	Spherical p.c.	0.204	0.297	0.879	1.24	1.34	0.577	0.268	0.232	4.8
	Deviation (20%)	0.322	0.280	0.203	1.25	1.36	0.998	0.393	0.350	6.0
	Deviation (50%)	0.644	0.615	0.543	0.492	0.390	0.912	0.598	0.604	8.5
	h-depth (20%)	0.300	0.271	0.210	1.27	1.36	0.972	0.373	0.327	5.3
	h-depth (50%)	0.526	0.511	0.492	0.481	0.414	0.898	0.545	0.536	7.5
	Band depth (20%)	0.192	0.199	0.222	1.28	1.35	1.26	0.541	0.304	6.1
	Band depth (50%)	0.256	0.326	0.455	0.552	0.738	1.24	0.604	0.396	7.5
	Hard trimmed (20%)	0.310	0.288	0.218	1.12	1.36	0.979	0.400	0.347	5.9
	Hard trimmed (50%)	0.467	0.478	0.473	0.479	0.441	0.902	0.560	0.532	7.3
	Hard trimmed (adap.)	0.331	0.224	0.247	0.271	0.344	1.20	0.520	0.361	5.1
	Soft trimmed (20%)	0.347	0.335	0.282	0.268	1.25	0.778	0.418	0.377	5.6
2	Sample p.c.	0.224	1.29	1.36	1.37	1.38	0.966	0.583	0.396	9.0
	Spherical p.c.	0.281	0.493	1.06	1.27	1.34	0.479	0.362	0.327	4.9
	Deviation (20%)	0.439	0.387	0.287	1.25	1.36	0.688	0.473	0.448	5.9
	Deviation (50%)	0.697	0.673	0.636	0.575	0.490	0.698	0.652	0.657	8.5
	h-depth (20%)	0.392	0.366	0.310	1.29	1.36	0.659	0.449	0.416	4.9
	h-depth (50%)	0.569	0.577	0.572	0.574	0.564	0.675	0.613	0.592	7.3
	Band depth (20%)	0.311	0.485	1.05	1.32	1.36	0.733	0.471	0.409	7.1
	Band depth (50%)	0.491	0.947	1.02	1.06	1.21	0.594	0.462	0.461	6.5
	Hard trimmed (20%)	0.400	0.380	0.303	1.15	1.35	0.693	0.478	0.448	5.4
	Hard trimmed (50%)	0.524	0.545	0.541	0.566	0.633	0.700	0.621	0.598	7.8
	Hard trimmed (adap.)	0.432	0.307	0.363	0.415	0.570	0.893	0.548	0.480	5.6
	Soft trimmed (20%)	0.421	0.424	0.377	0.410	1.26	0.599	0.484	0.470	5.1

5. Examples

5.1. Excitation–emission matrices

As explained in Mortensen and Bro (2006), enzyme cultivation processes often require quick on-line adjustments that demand fast and reliable quality control tools. Samples of the cultivation broth are typically taken at regular time intervals, and enzyme activity is measured. The traditional off-line chemical analyses determine enzyme activity directly and accurately, but it may take hours or days to get the results back from the laboratory. An alternative is to employ multi-channel fluorescence sensors that produce immediate results in the form of excitation-emission matrices (EEMs), although enzyme activity can be determined only indirectly from the EEMs (via principal component regression or partial least squares).

Mortensen and Bro (2006) analyzed a dataset of 338 EEMs, available at `http://www.models.life.ku.dk/research/data/`. A movie showing these 338 EEMs in quick succession is available on the author's website. A few atypical EEMs can be spotted at the end of the movie. Taking logarithms of the EEMs ameliorates the effect of the outliers to some extent, but not completely, as Figure 1 shows (a movie showing the log-EEMs is also available on the author's website).

In principle, an EEM is a two-dimensional array consisting of light intensity measured at certain excitation and emission wavelengths. Mortensen and Bro (2006) used 15 excitation filters ranging from 270 to 550 nm, and 15 emission filters ranging from 310 to 590 nm; all filters had a maximum half-width of 20 nm. Since emission wavelength must be longer than excitation wavelength, the EEMs are actually triangular arrays: of the 15×15 possible excitation/emission combinations, only 120 yield actual measurements. This problem can be approached as a classical multivariate problem of dimension $p = 120$ and sample size $n = 338$, and some of the robust methods reviewed by Filzmoser et al. (2009) for the "large p, small n" problem could be applied. However, since light intensity is a continuous function of the excitation and emission wavelengths, it is more appropriate and statistically efficient to approach this problem as a functional-data problem; the 120 measurements are just an arbitrary discretization of the continuous surfaces $X(s,t)$ that live in $\mathcal{L}^2(\mathbb{R}^2)$. This is a Euclidean space with inner product $\langle f, g\rangle = \iint f(s,t)g(s,t)\, ds\, dt$, the mean of X is the bivariate function $\mu(s,t) = E\{X(s,t)\}$ and the covariance operator of X can be represented as

$$\mathfrak{C}(f,g) = \int\int\int\int \rho(s_1,t_1,s_2,t_2)\; f(s_1,t_1)\; g(s_2,t_2)\; ds_1\; dt_1\; ds_2\; dt_2,$$

where $\rho(s_1,t_1,s_2,t_2) = \operatorname{cov}\{X(s_1,t_1), X(s_2,t_2)\}$.

We carried out a principal component analysis on the log-EEMs. A histogram of the radii (Figure 3) shows that 15 observations are clear outliers,

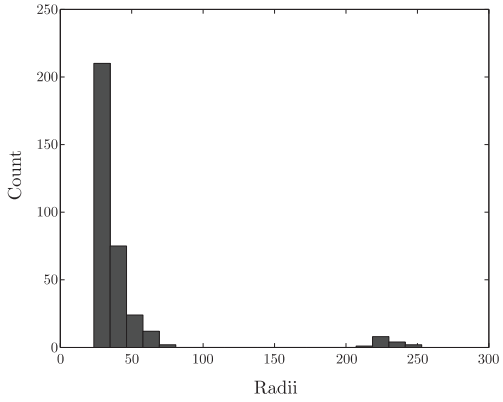

Figure 3. Excitation-Emission Matrices. Histogram of the radii with $\alpha = 0.50$.

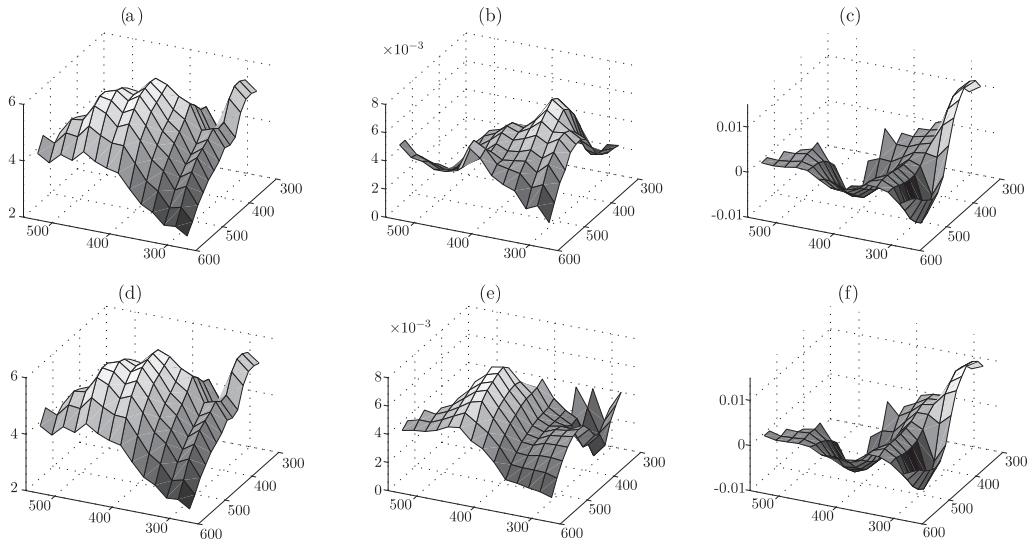

Figure 4. Excitation–Emission Matrices. (a) Trimmed mean, (b) first trimmed principal component, (c) second trimmed principal component, (d) sample mean, (e) first sample principal component, and (f) second sample principal component. Trimmed estimators were computed with 5% trimming.

so we computed the 5% trimmed mean and principal components (Figure 4). Among the 20 leading components, the first (Figure 4(b)) accounts for 59% of the variability, and the second (Figure 4(c)) accounts for 20% of the variability. We also computed the sample mean and principal components; among the 20 leading components, the first (Figure 4(e))accounts for 88% of the variability, and the second (Figure 4(f)) for only 5%.

While the sample mean (Figure 5(d)) is not very different from the trimmed mean (Figure 5(a)), the first principal component is seriously affected by the outliers. The first sample principal component only explains how the outliers vary from the "good" observations; it may be useful for outlier detection, but it's not associated with any genuine source of variability. The first trimmed component, in contrast, is genuinely the main direction of variability of the "clean" data.

The second trimmed component and the second sample principal component are very similar, but the latter underestimates the relative importance of the component, assigning it only 5% of the total variability. This is unfortunate because the second component is the one primarily associated with enzyme activity. Mortensen and Bro (2006) provided an enzyme activity measure for calibration, and the correlation coefficient (after eliminating the 15 outliers) between enzyme activity and the second trimmed component was .69. This association could be overlooked if the user based his analysis on the non-robust sample principal components and decided that the second component was negligible.

5.2. Handwritten digits

The planar trajectory of a pen tip is a curve $\mathbf{X}(t) = (x(t), y(t))$ in $\mathbb{R}^2$, where t is time. Then the analysis of handwritten digits can be approached as a functional data problem in the Euclidean space $\left(\mathcal{L}^2(\mathbb{R})\right)^2$ endowed with the inner product $\langle \mathbf{f}, \mathbf{g} \rangle = \int \mathbf{f}(t)^T \mathbf{g}(t) dt$. The mean trajectory is $\mu(t) = E\{\mathbf{X}(t)\}$ and the covariance operator can be represented as

$$\mathfrak{C}(\mathbf{f}, \mathbf{g}) = \iint \mathbf{f}(s)^T \mathbf{R}(s,t) \mathbf{g}(t) \; ds \; dt,$$

with $\mathbf{R}(s,t) = E[\{\mathbf{X}(s) - \mu(s)\}\{\mathbf{X}(t) - \mu(t)\}^T]$. In this section we analyze a set of 1,055 handwritten samples of the digit "five" , available at the Machine Learning Repository of the University of California at Irvine, `http://archive.ics.uci.edu/ml/`. The data was rotated and scaled so that x and y range between 0 and 100, and t between 0 and 1. Eight sample digits are shown in Figure 2.

A plot of the sample mean (Figure 5(a)) does not resemble a "five" or any other recognizable digit. To understand why this happens, we computed the radii for different values of α and noticed that their distribution becomes increasingly bimodal as α increases. The histogram for $\alpha = 0.50$ is shown in Figure 6. There are two neatly distinguishable groups: 627 observations with $r_i < 60$, and 428 observations with $r_i > 60$. The large number of observations in the second group suggests that the sample may be made up of two systematic clusters, rather than a single homogeneous group and a few isolated outliers.

This is confirmed by a plot of the 41% trimmed mean (Figure 5(b)), together with the mean of the observations that were cut off (Figure 5(c)). It turns out

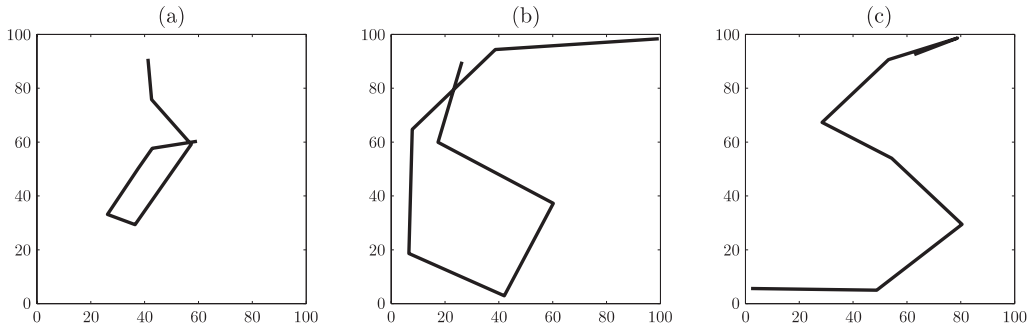

Figure 5. Handwritten Digits. (a) Sample mean, (b) 41% trimmed mean, and (c) mean of the trimmed observations.

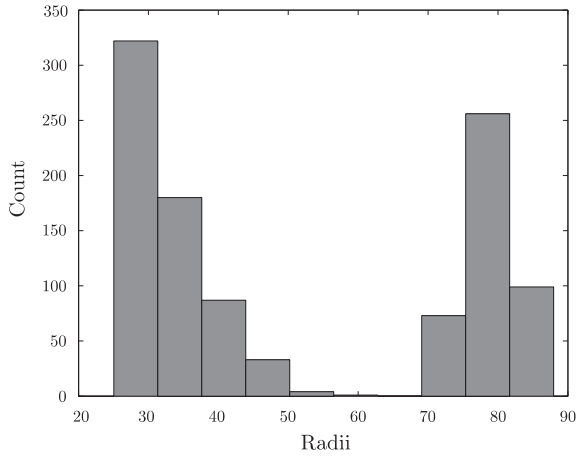

Figure 6. Handwritten Digits. Histogram of the radii with $\alpha = 0.50$.

that there are two ways to draw the number "five" . The most common is in two strokes, beginning at the upper left corner and moving downwards, then raising the pen to draw the top dash (our planar representation of the trajectory does not capture this vertical movement explicitly). The other way, less common, is to draw the number "five" in a single stroke, like the letter "S" . Figure 5(b) corresponds to the first class and Figure 5(c) corresponds to the second.

As in the EEMs example, the sample principal components do not provide much useful information except for discrimination. The trimmed principal components, on the other hand, do provide useful information about the directions of variability in the bigger cluster. The easiest way to interpret the principal components is to plot their effects on the mean (Figure 7). This figure shows the trimmed mean and the first two trimmed principal components (Figure 7(a,b)), as well as the mean and the first two principal components of the observations

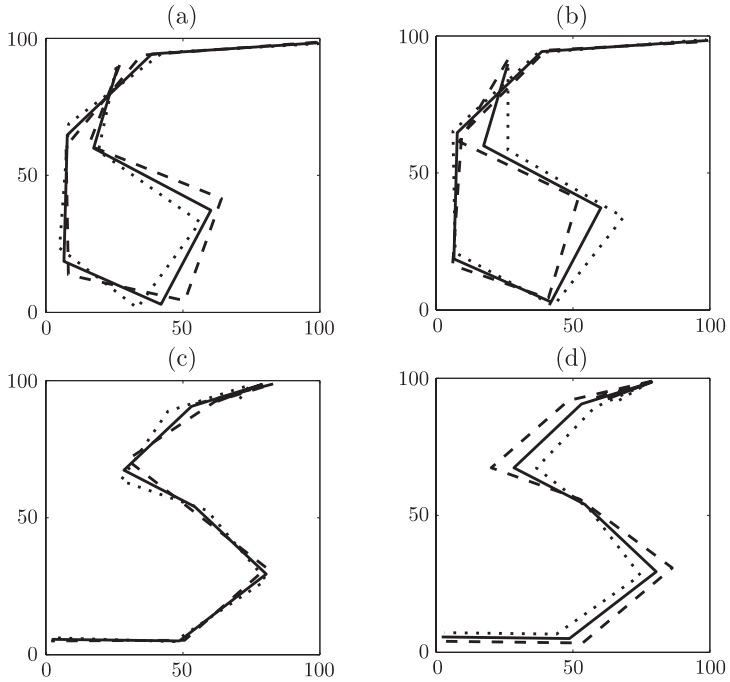

Figure 7. Handwritten Digits Example. Effect of the principal components on the mean (—— is the mean; − − − is the mean plus 5 times the principal component; · · · is the mean minus 5 times the principal component). (a) Trimmed mean and first trimmed component; (b) trimmed mean and second trimmed component; (c) mean and first component of the trimmed observations; (d) mean and second component of the trimmed observations.

that were cut off (Figure 7(c,d)). The first trimmed principal component (Figure 7(a)) explains 56% of the variability and is associated with variation in the inclination of the "belly" of the digit. The second trimmed principal component (Figure 7(b)) explains 14% of the variability and is mostly associated with variation in the inclination of the vertical dash. Regarding the components of the second type of "fives" , the first principal component (Figure 7(c)) accounts for 44% of the variability and is associated with variation in the "roundness" of the "five" : negative scores correspond to rounded "S-shaped" digits, while positive scores correspond to more angular "Z-shaped" digits. The second principal component (Figure 7(d)) accounts for 20% of the variability and explains variability in the width of the digit.

Acknowledgements

This research was supported by the National Science Foundation, grant DMS 0604396.

References

Ash, R. B. and Gardner, M. F. (1975). *Topics in Stochastic Processes*. Academic Press, New York.

Cuevas, A., Febrero, M. and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Comput. Statist.* **22**, 481-496.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont, CA.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Springer, New York.

Filzmoser, P., Serneels, S., Maronna, R. and Van Espen, P.J. (2009). Robust multivariate methods in chemometrics. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Vol. III), 681-722. Elsevier, Amsterdam.

Fraiman, R. and Meloche, J. (1999). Multivariate L-estimation. *Test* **8**, 255-317.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test* **10**, 419-40.

Gervini, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika* **95**, 587-600.

Gikhman, I. I. and Skorokhod, A. V. (2004). *The Theory of Stochastic Processes I*. Springer, New York.

Gohberg, I., Goldberg, S. and Kaashoek, M. A. (2003). *Basic Classes of Linear Operators*. Birkhäuser Verlag, Basel.

Jiang, J. C. and Mack, Y. P. (2001). Robust local polynomial regression for dependent data. *Statist. Sinica* **11**, 705-722.

Jolliffe, I. T. (2002). *Principal Component Analysis*. (2nd edition). Springer, New York.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27**, 783-858.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L. (1999). Robust principal component analysis for functional data (with discussion). *Test* **8**, 1-73.

López-Pintado, S., and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* **104**, 718-734.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19**, 229-248.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Mortensen, P. P. and Bro, R. (2006). Real-time monitoring and chemical profiling of a cultivation process. *Chemometrics and Intelligent Laboratory Systems* **84**, 106-113.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed.). Springer, New York.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications* (DIMACS, Vol. 72), 1-16.

Shi, P. D. and Li, G. Y. (1995). Global convergence rates of B-spline M-estimators in nonparametric regression. *Statist. Sinica* **5**, 303-318.

Stigler, S. M. (1973). The asymptotic distribution of the trimmed mean. *Ann. Statist.* **1**, 472-477.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Zuo, Y., and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28**, 461-482.

Zuo, Y., Cui, H., and He, X. (2004). On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *Ann. Statist.* **32**, 167-188.

Department of Mathematical Sciences, University of Wisconsin–Milwaukee, P.O. Box 413, Milwaukee, WI 53201, USA.

E-mail: gervini@uwm.edu

(Received November 2010; accepted October 2011)