A definition of depth for functional observations

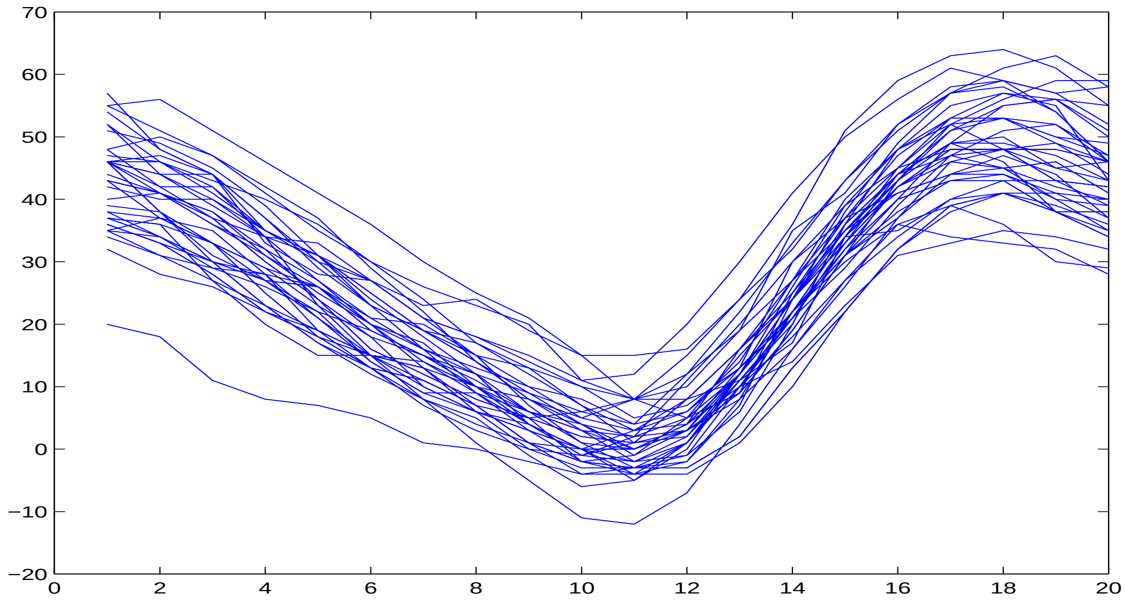
Sara López-Pintado and Juan Romo

Departamento de Estadística y Econometría

Universidad Carlos III de Madrid

Madrid, Spain

1. MOTIVATION AND BACKGROUND



- Question: which one is the deepest function ?

The observations

$$x_1(t), x_2(t), ..., x_n(t)$$

are n functions defined on an interval I.

Why considering functional data?

1. In many areas of knowledge the process generating the data provides us in a natural way with a set of functions.

2. Many problems are better approached if the observations are treated as continuous functions.

3. Each curve from the sample can be observed at different points and the separation of these points can be irregular.

4. Technological advance with the development of progressively more precise and sophisticated equipment makes possible the acquisition of a large number of data, usually called high frequency data, that allow us to express the data as functions.

- Goal: to introduce a definition of depth for functional data. This concept will be used to measure the centrality of a curve with respect to a set of curves. E.g.: to define the deepest function.

- The functional depth provides a center-outward ordering of a sample of curves. Order statistics will be defined. ($L-$statistics).

- The idea of deepest point of a set of data allows to classify a new observation by using the distance to a class deepest point.

The notion of depth has been extensively studied in the multivariate context.

Some definitions of data depth are:

1. The Mahalanobis depth (Mahalanobis, 1936).

2. The half-space depth (Hodges, 1955, Tukey, 1975).

3. The Oja depth (Oja, 1983).

4. The simplicial depth (Liu, 1990).

5. The majority depth (Singh, 1991).

6. The projection depth (Zuo, 2003).

Liu (1990), and Zuo and Serfling (2000) introduce general conditions to define a notion of statistical depth.

Key properties a concept of depth should verify:

- Affine invariance
- Maximality at center
- Monotonicity relative to deepest point
- Vanishing at infinity

Fraiman and Muniz (2001) defined a concept of depth for functional data.

Let $X_1(t), ..., X_n(t)$ be i.i.d. stochastic processes defined on [0,1]. Let F_t be the univariate marginal distribution of $X_1(t)$. Let D_n be any concept of depth in $\mathbb{R}$. Consider for every $t \in [0, 1]$

$$D_n(X_i(t)) = Z_i(t),$$

(univariate depth of $X_i(t)$ at t with respect to $X_1(t), ..., X_n(t)$).

Defining

$$I_i = \int_0^1 Z_i(t)dt, \qquad 1 \leq i \leq n,$$

the set of functions $X_1(t), ..., X_n(t)$ can be ordered according to the value of I_i.

2. A NEW CONCEPT OF DEPTH FOR FUNCTIONAL DATA

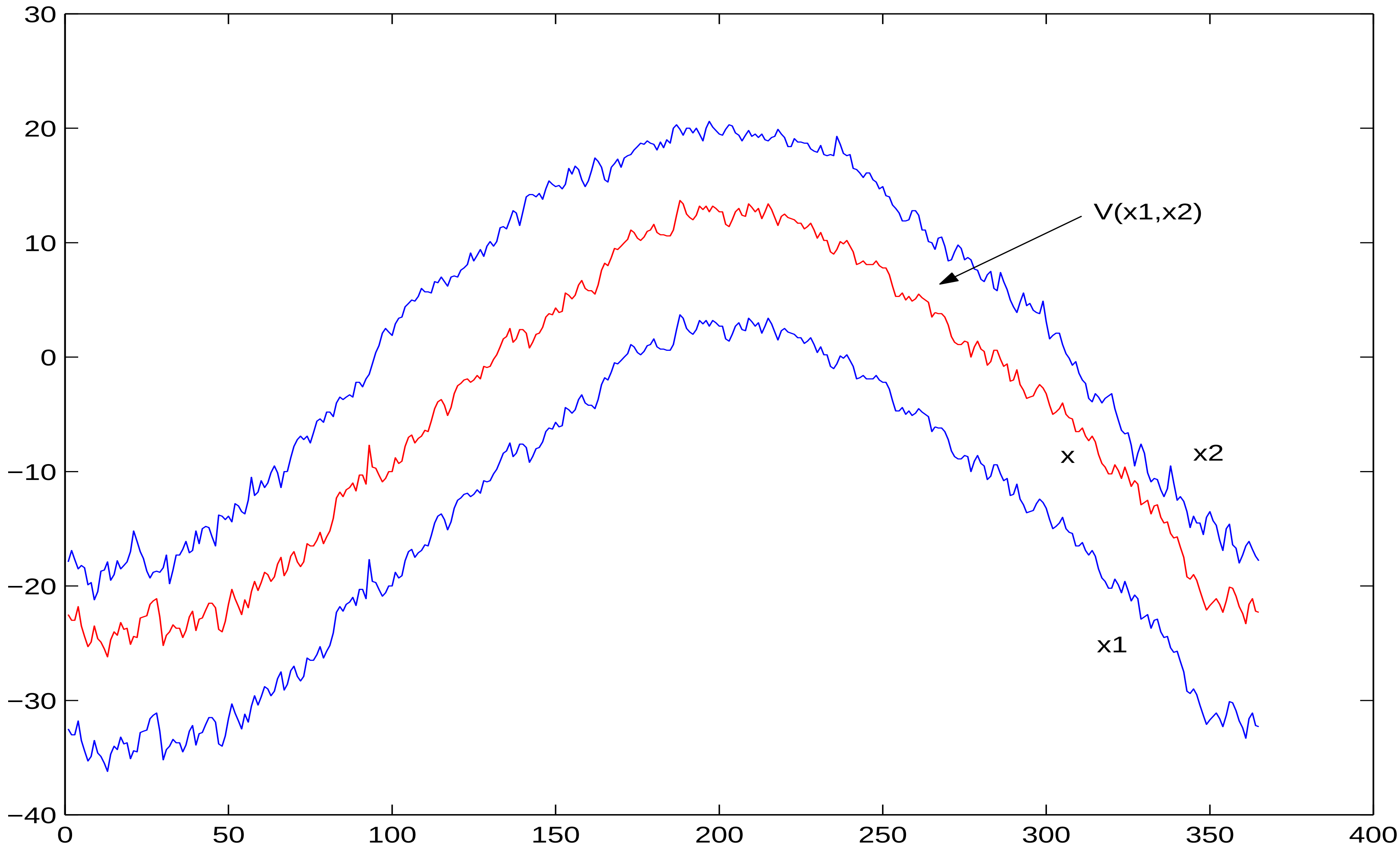
Let $x_1(t), ..., x_n(t)$ be a sample of functions. Define

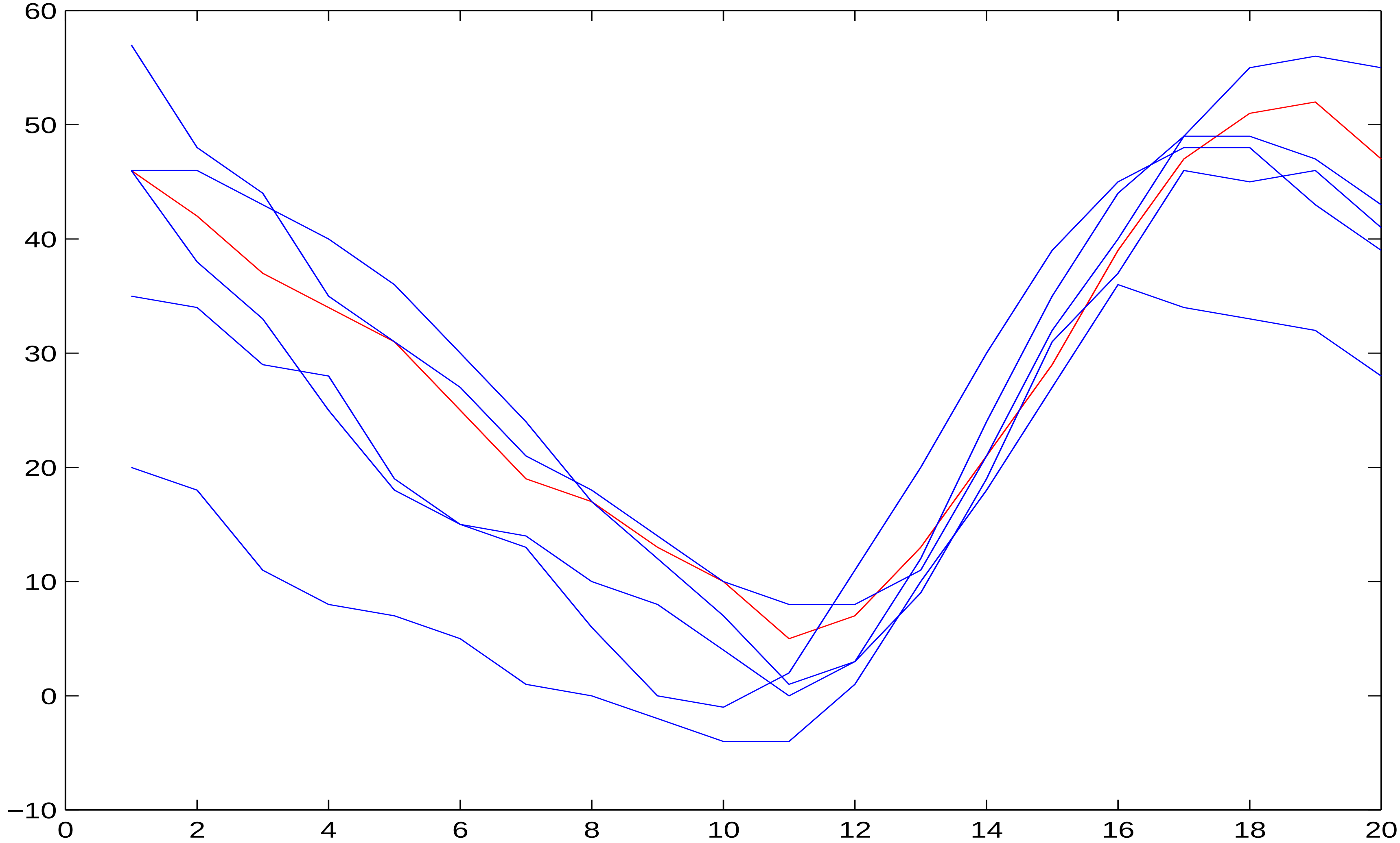
$$V(x_{i_1}, ..., x_{i_k}) = \left\{ x \colon \min_{r=1,\ldots,k} \{x_{i_r}(t)\} \leq x(t) \leq \max_{r=1,\ldots,k} \{x_{i_r}(t)\}\,,\ t \in [0,1] \right\}$$

(Functions whose graphs belong to the area delimited by the graphs of $x_{i_1}, x_{i_2}, ..., x_{i_k}$).

Equivalently,

$$V = \left\{ x(t) = \alpha_t \min_{r=1,\ldots,k} \{x_{i_r}(t)\} + (1-\alpha_t) \max_{r=1,\ldots,k} \{x_{i_r}(t)\}\,,\ t \in [0,1],\ \alpha_t \in [0,1] \right\}$$


V(x1,x2)
x
x2
x1
30
20
10
0
−10
−20
−30
−40
0
50
100
150
200
250
300
350
400


60
50
40
30
20
10
0
−10
0
2
4
6
8
10
12
14
16
18
20

The J-depth for x is:

$$S_{n,J}(x) = \sum_{j=2}^{J} S_n^{j)}(x),$$

where

$$S_n^{j)}(x) = \frac{\sum_{1 \leq i_1 < i_2 < \ldots i_j \leq n} I(x \in V(x_{i_1}, x_{i_2}, \ldots, x_{i_j}))}{\binom{n}{j}}$$

are proportions of bands containing x; this gives a center-outward ordering of the sample of curves.

A deepest function $\widehat{\mu}_{n,J}$ will satisfy:

$$\widehat{\mu}_{n,J} = \underset{x \in \{x_1, \ldots, x_n\}}{\arg\max} \; S_{n,J}(x)$$

The population version is

$$S_J(x) = \sum_{j=2}^{J} S^{j)}(x) = \sum_{j=2}^{J} P(x \in V(x_1, x_2, ..., x_j)),$$

and a population deepest function is a function μ_J maximizing $S_J(\cdot)$.

Example: Trimmed mean for functional data

The functional version of the $\alpha-$trimmed mean will be the average of the $n - [n\alpha]$ deepest observations:

$$\widehat{m}_{n,J}^{\alpha} = \frac{\sum_{i=1}^{n} I_{[\beta,+\infty]}(S_{n,J}(x_i))x_i}{\sum_{i=1}^{n} I_{[\beta,+\infty]}(S_{n,J}(x_i))}, \ \beta > 0,$$

where $\frac{1}{n}\left(\sum_{i=1}^{n} I_{[\beta,+\infty]}(S_{n,J}(x_i))\right) \simeq 1 - \alpha.$

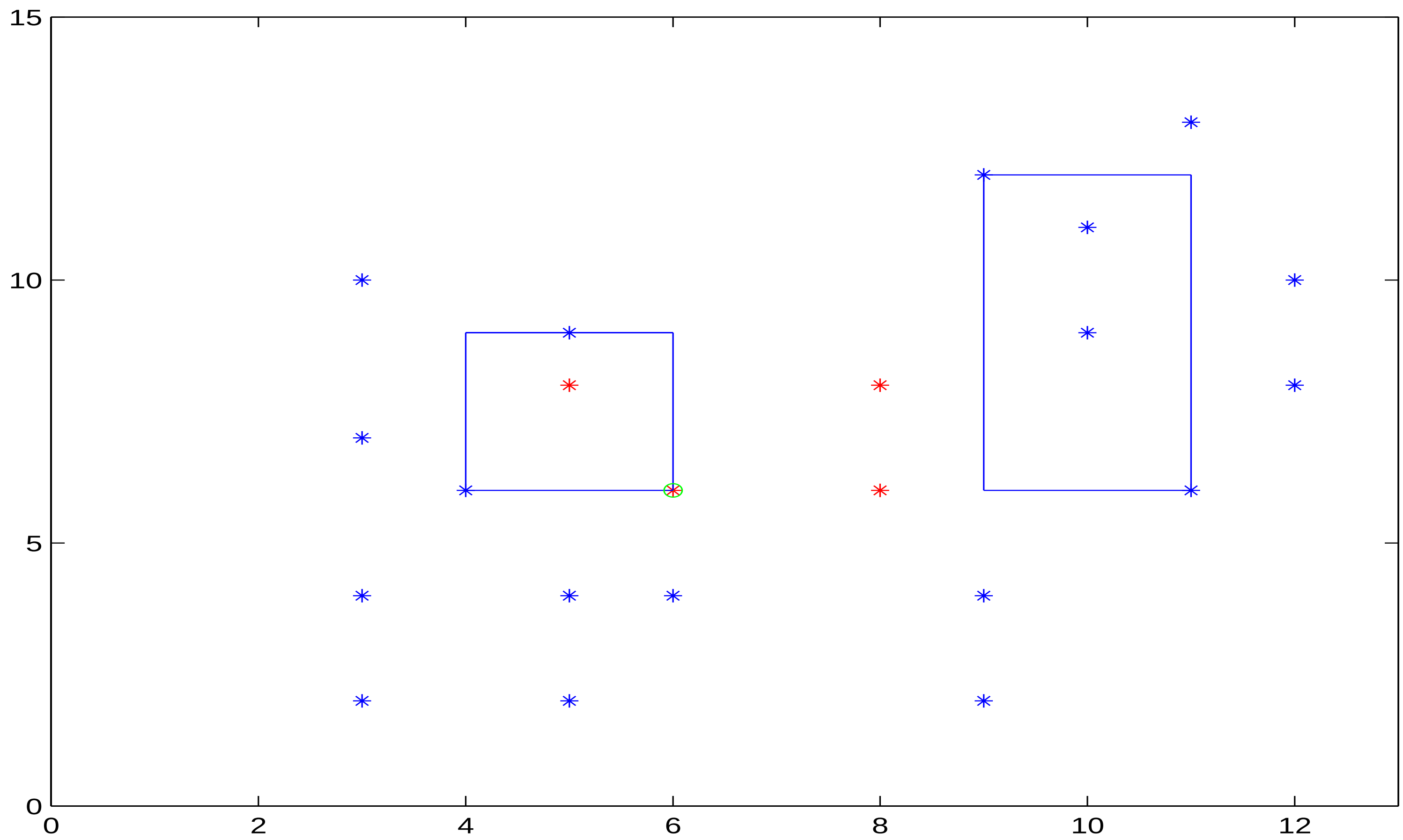
3. FINITE-DIMENSIONAL VERSION

Let F be a probability distribution in $\mathbb{R}^d$; $d \geq 1$. Let $\{y_1, ..., y_n\}$ be a random sample from F.

A multivariate observation can be seen as a function defined on $\{1, 2, ..., d\}$: $y(l)$ is the $l - th$ component of the vector y.

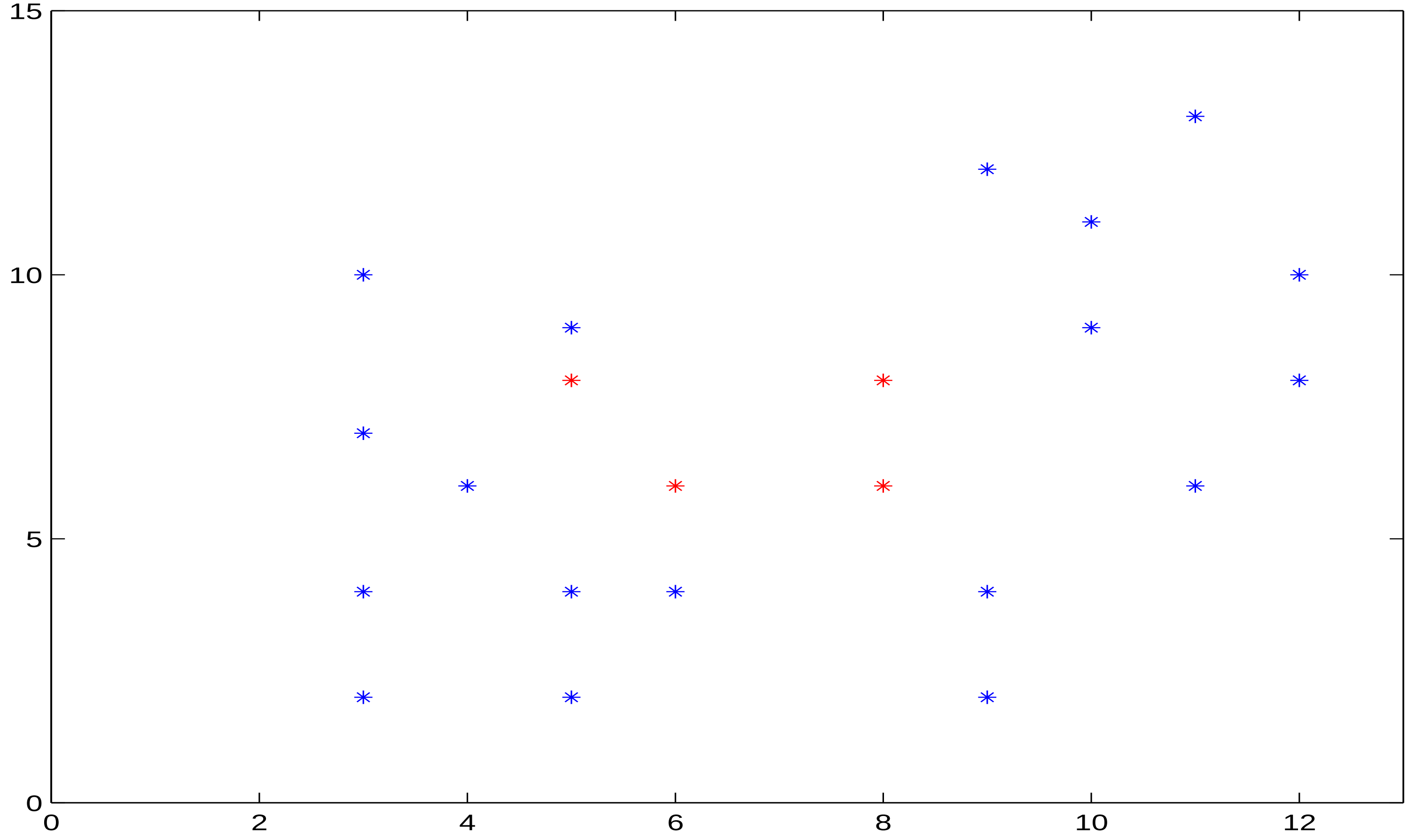
For $d = 2$, the finite dimensional band $V(y_1, y_2, ..., y_j)$ is the interval in the plane determined by the following four vertices

$$\left(\min_{k=1,\ldots,j}\{y_k(1)\}, \min_{k=1,\ldots,j}\{y_k(2)\}\right), \quad \left(\min_{k=1,\ldots,j}\{y_k(1)\}, \max_{k=1,\ldots,j}\{y_k(2)\}\right)$$
$$\left(\max_{k=1,\ldots,j}\{y_k(1)\}, \max_{k=1,\ldots,j}\{y_k(2)\}\right), \quad \left(\max_{k=1,\ldots,j}\{y_k(1)\}, \min_{k=1,\ldots,j}\{y_k(2)\}\right)$$



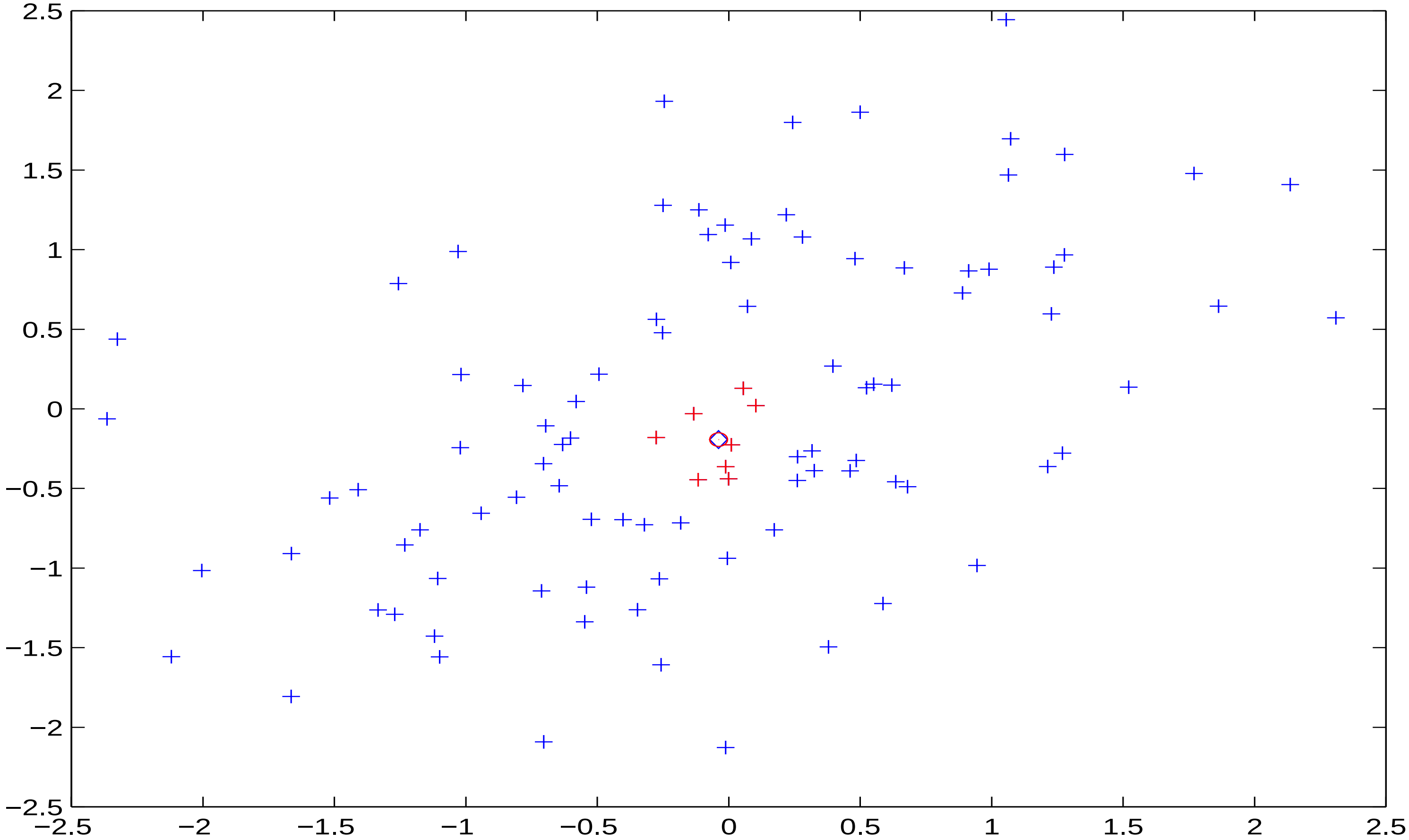
$S_n^{j)}(y)$ is the proportion of intervals $V(y_{i_1}, y_{i_2}, ..., y_{i_j})$ determined by j sample points containing y.

Example: deepest points for $S_{n,2}(\cdot)$

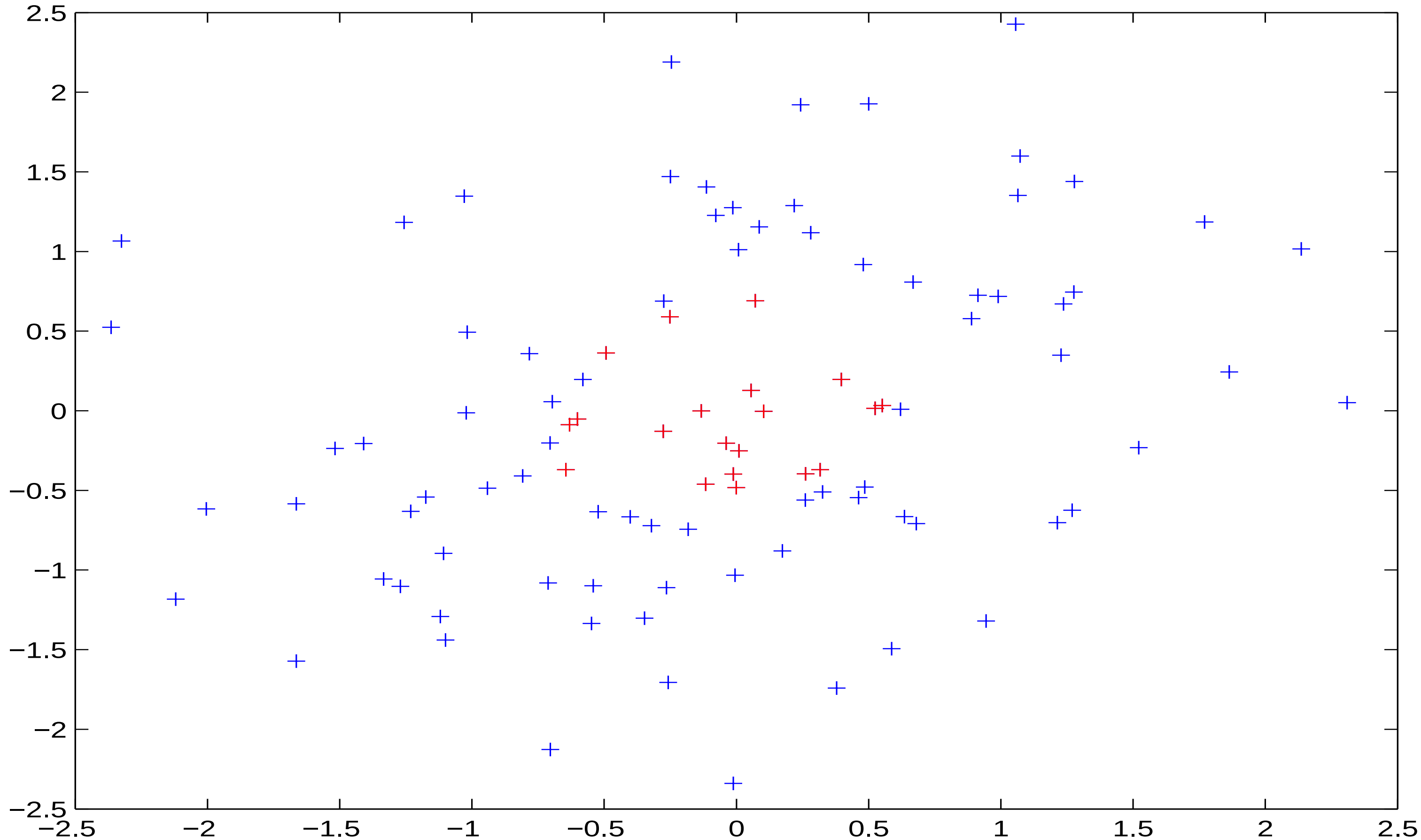


Remark: They essentially coincide with Liu's simplicial deepest points.

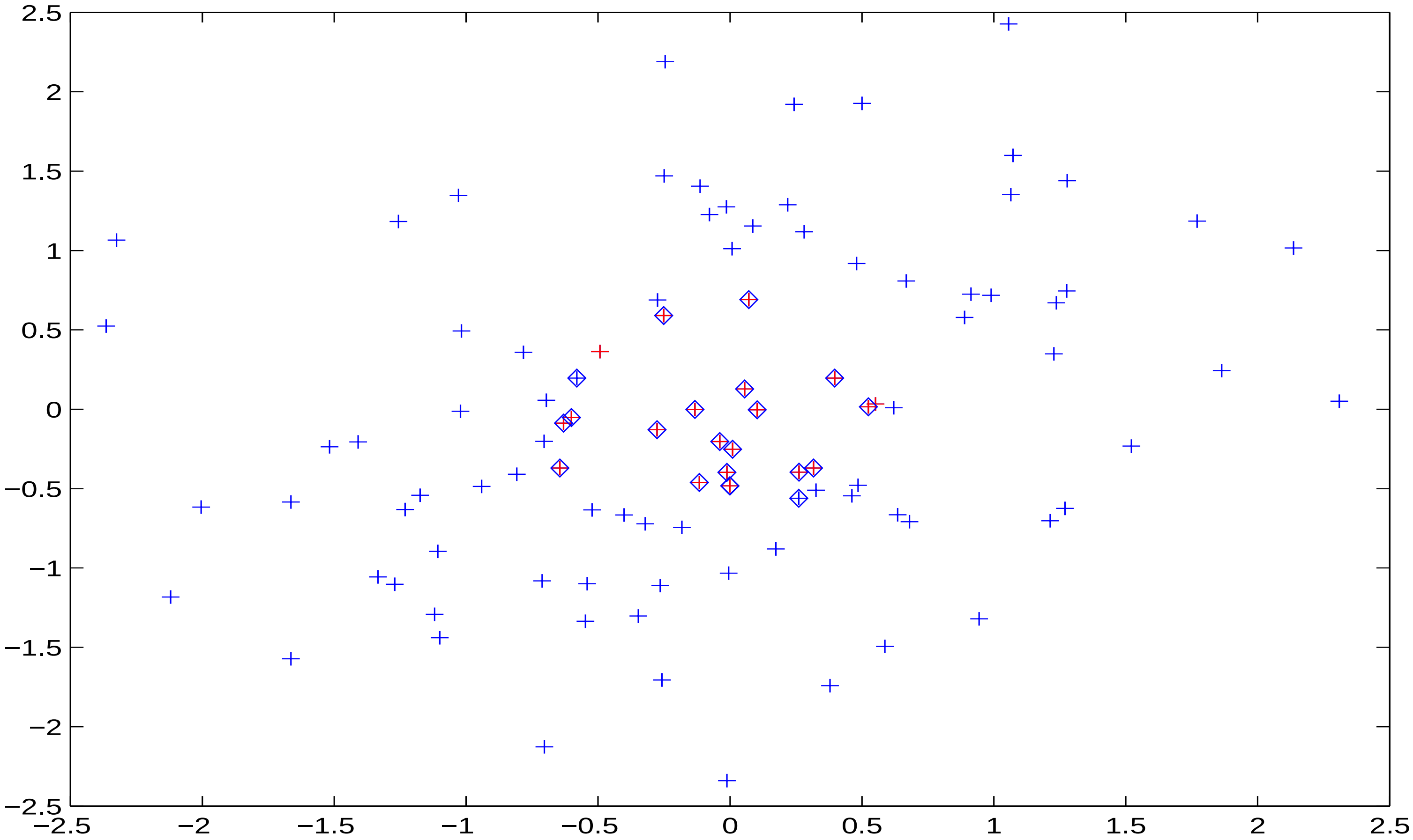
Another example: deepest points for $S_{n,3}(\cdot)$



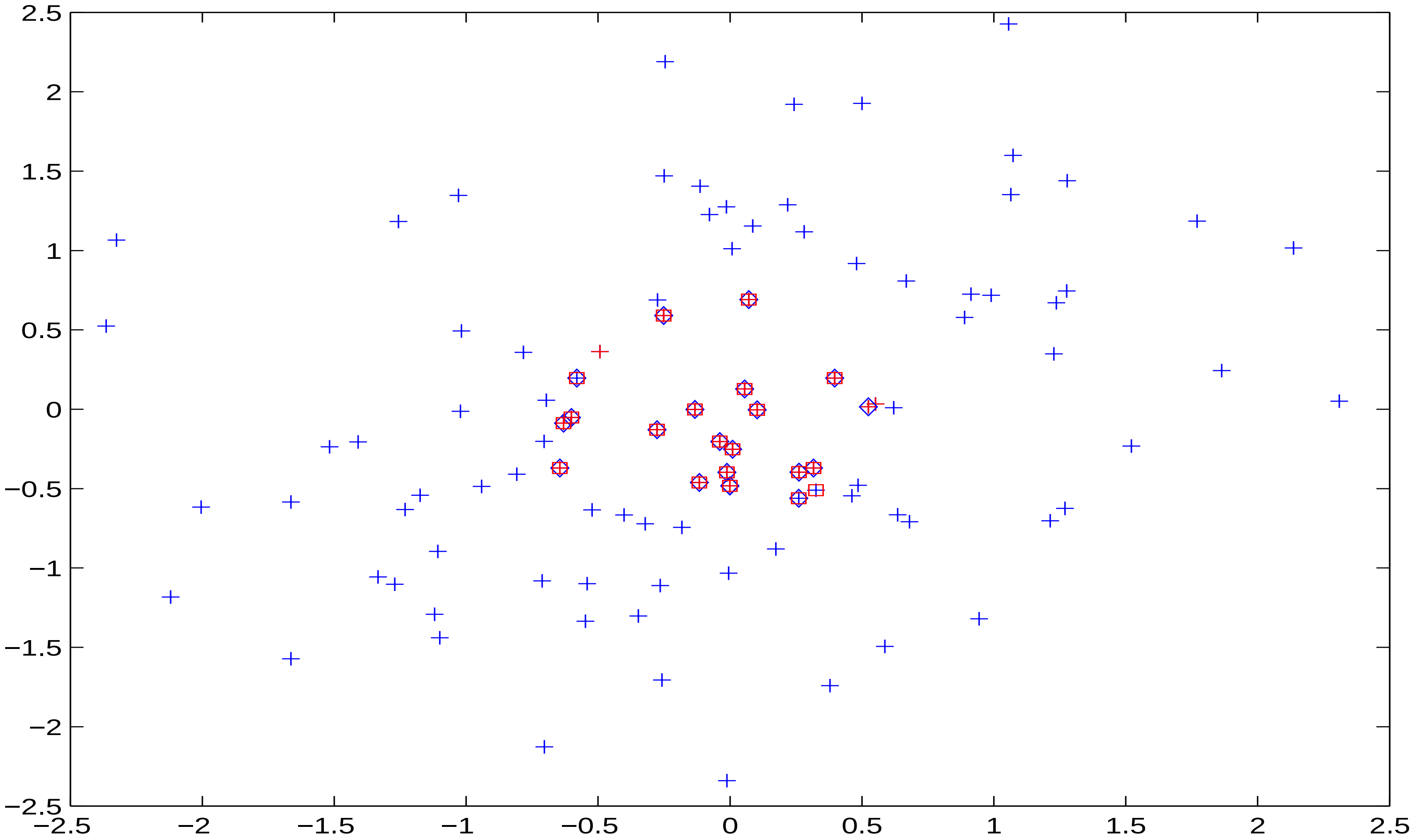
How does the choice of J affect the depth?



$J = 2$



$J = 3$



$J = 4$

4. SOME PROPERTIES

Finite-dimensional data:

1. The deepest point in $\mathbb{R}$ (with S_J) coincides with the usual univariate median; moreover, the order induced by S_J is independent of J.

2. $S_J(\cdot)$ is invariant under transformations of type $T(y) = A * y + b$, where A is a diagonal and invertible $d \times d$ matrix and $b \in \mathbb{R}^d$:

$$S_{J,T}(Ty) = S_J(y)$$

3. If F is absolutely continuous and symmetric then $S_J(\alpha y)$ is a monotone nonincreasing function in $\alpha \geq 0$ for all $y \in \mathbb{R}^d$.

4. $S_J(\cdot)$ vanishes at infinity:

$$\sup_{\|y\|_\infty \geqslant M} S_J(y) \to 0 \quad \text{if } M \to \infty$$

5. If the marginal distributions of F are absolutely continuous then $S_J(\cdot)$ is continuous.

6. $S_{n,J}(\cdot)$ is strongly consistent:

$$S_{n,J}(y) \overset{a.s.}{\to} S_J(y) \ \text{ when } n \to \infty, \quad y \in \mathbb{R}^d$$

7. $S_{n,J}(y)$ is uniformly consistent:

$$\sup_{y \in R^d} |S_{n,J}(y) - S_J(y)| \longrightarrow 0 \quad \textit{a.s.} \quad \text{as} \quad n \longrightarrow \infty$$

8. If $S_J(\cdot)$ is uniquely maximized at μ, and μ_n is a sequence of random variables satisfying $S_{n,J}(\mu_n) = \sup_{x \in R^d} S_{n,J}(x)$, then

$$\mu_n \longrightarrow \mu \quad \textit{a.s.} \quad \text{as} \quad n \longrightarrow \infty$$

Functional data:

1. Let $Q \cap [0,1] = \{q_1, q_2, ..., q_n, ...\}$ and $x_n = (x(q_1), ..., x(q_n))$. Then:

$$S_J(x_n) \to S_J(x), \qquad \text{when} \quad n \to \infty$$

2. $S_J(\cdot)$ is invariant under transformations of type $T(x) = a(t) * x(t) + b(t)$:

$$S_{J,T}(Tx) = S_J(x)$$

3.

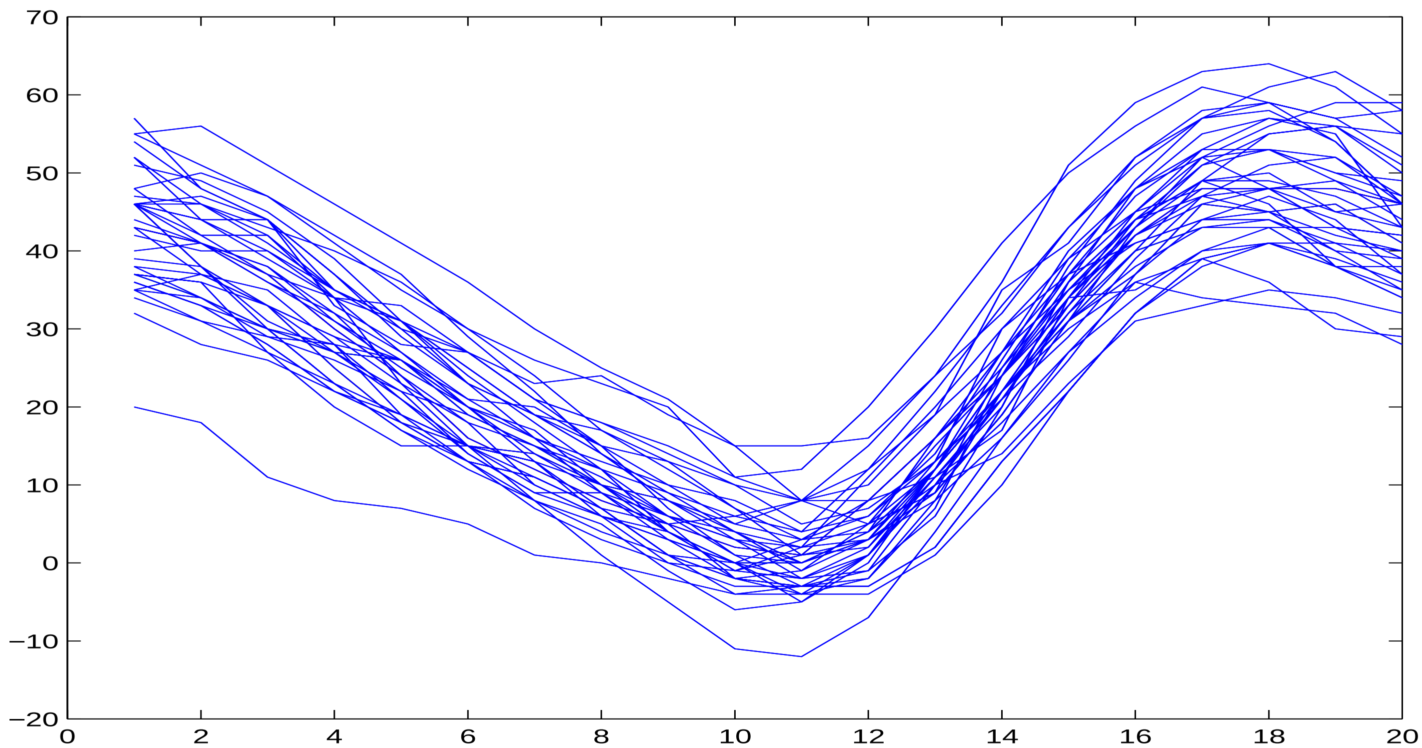
$$\sup_{\|x\|_\infty \geqslant M} S_J(x) \to 0 \quad \text{if } M \to \infty$$

4. $S_J(\cdot)$ is continuous.

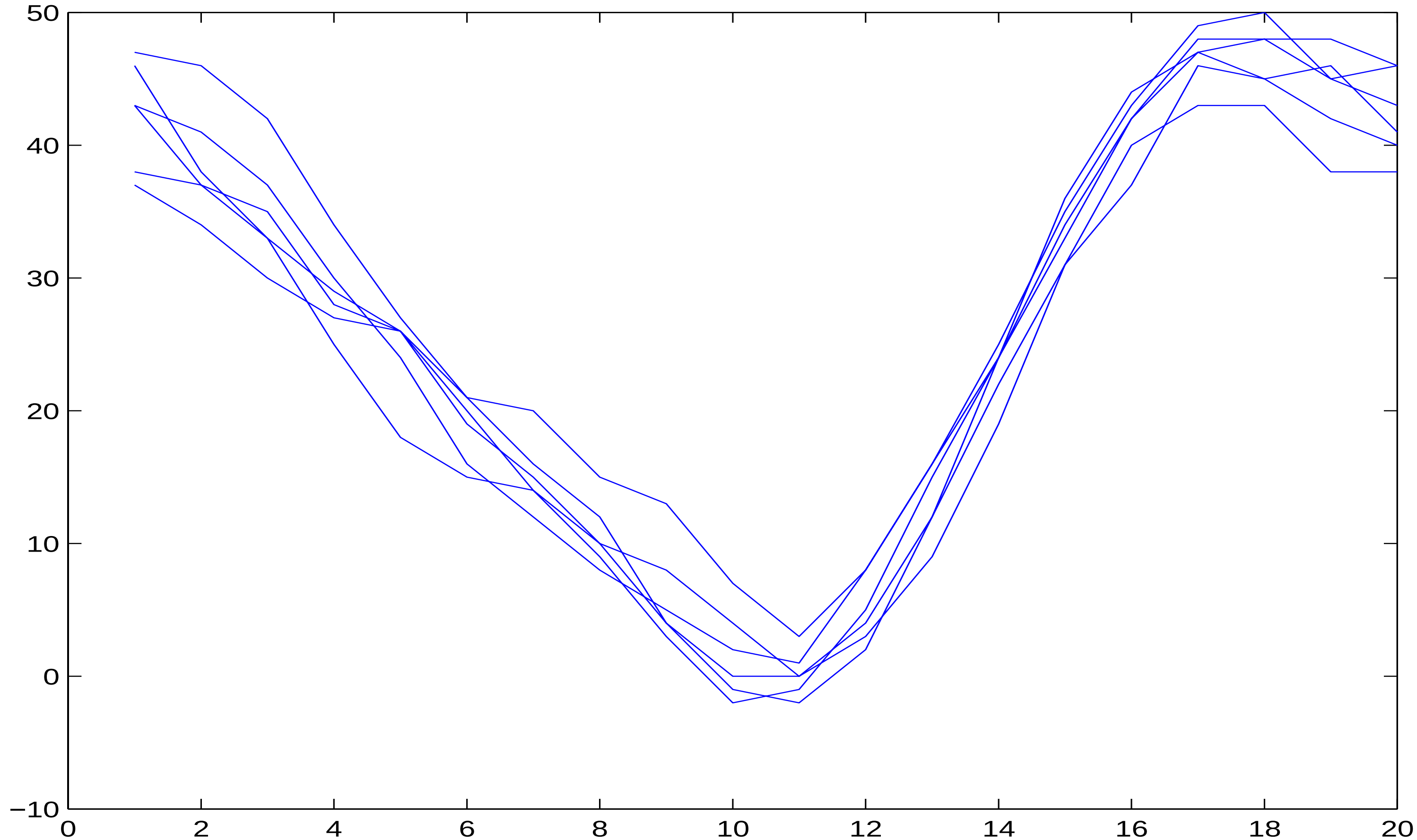
5. $S_{n,J}(x)$ is a consistent estimator: for any x,

$$S_{n,J}(x) \xrightarrow{a.s.} S_J(x) \quad \text{when } n \to \infty$$

5. APPLICATIONS

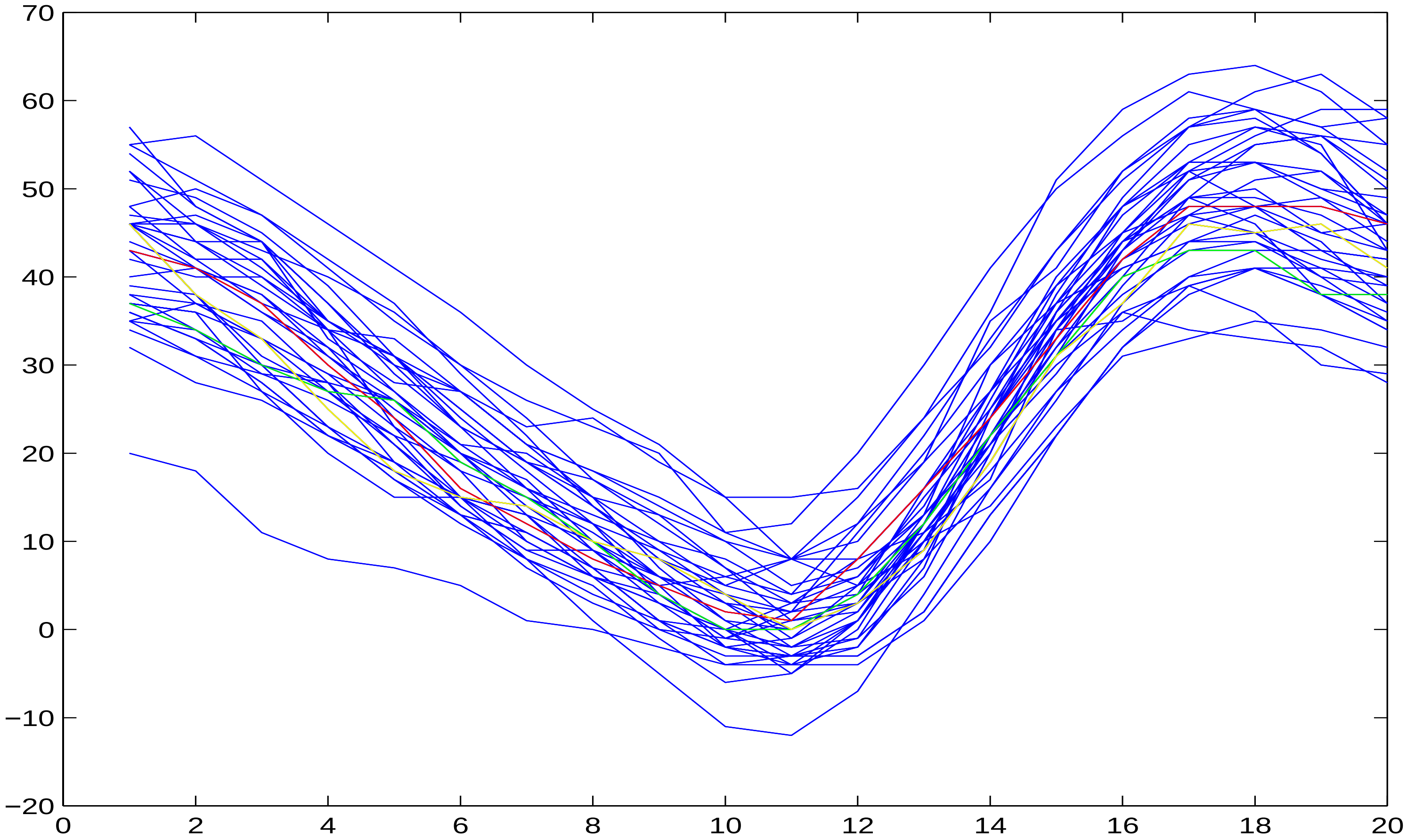


- Angles in the sagittal plane formed by the hip as 39 children go through a gait cycle. (Ramsay and Silverman, 1997)

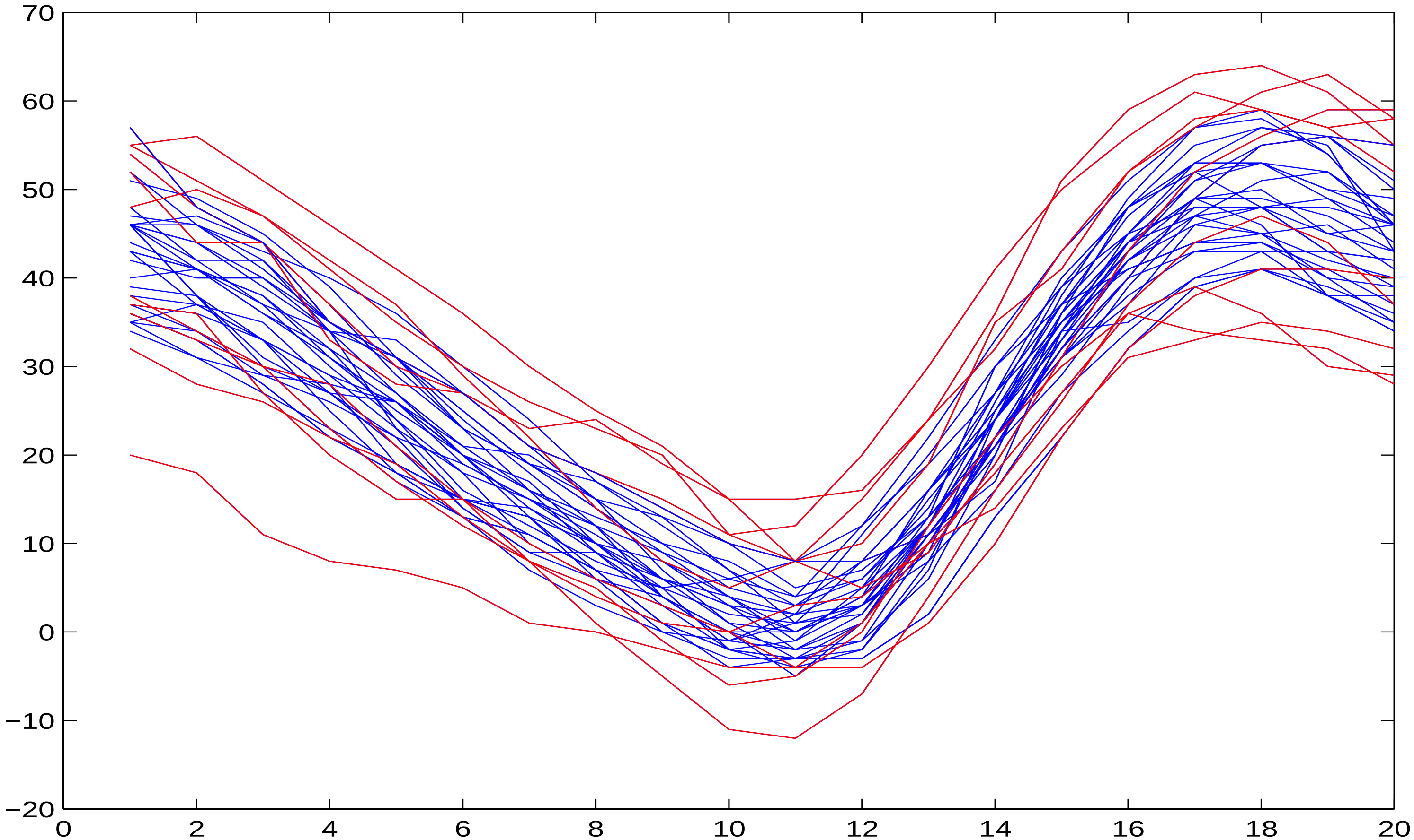


- Six deepest curves ($J = 5$).

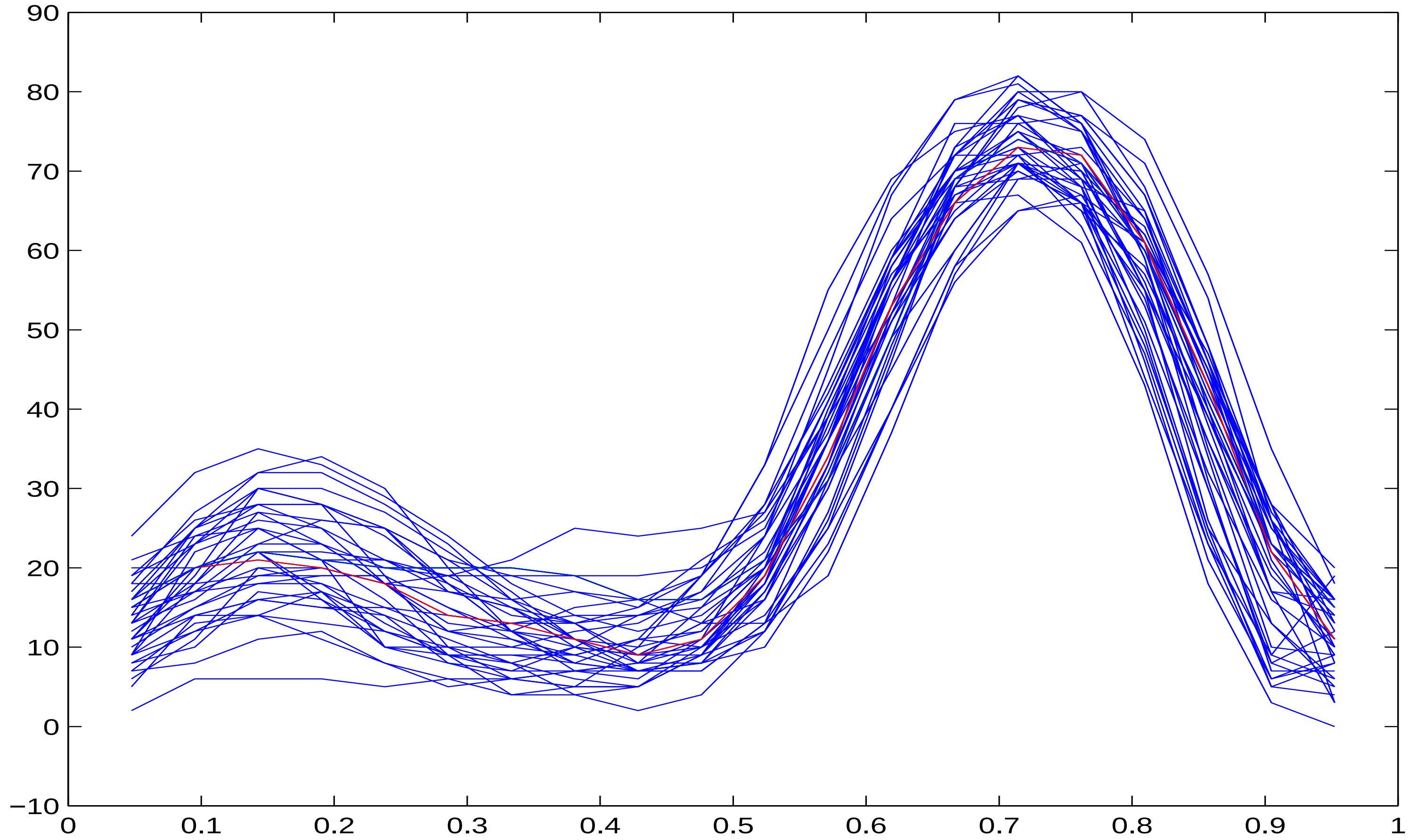
- The index $S_{n,J}$ when J increases gives the same centered-outward order.



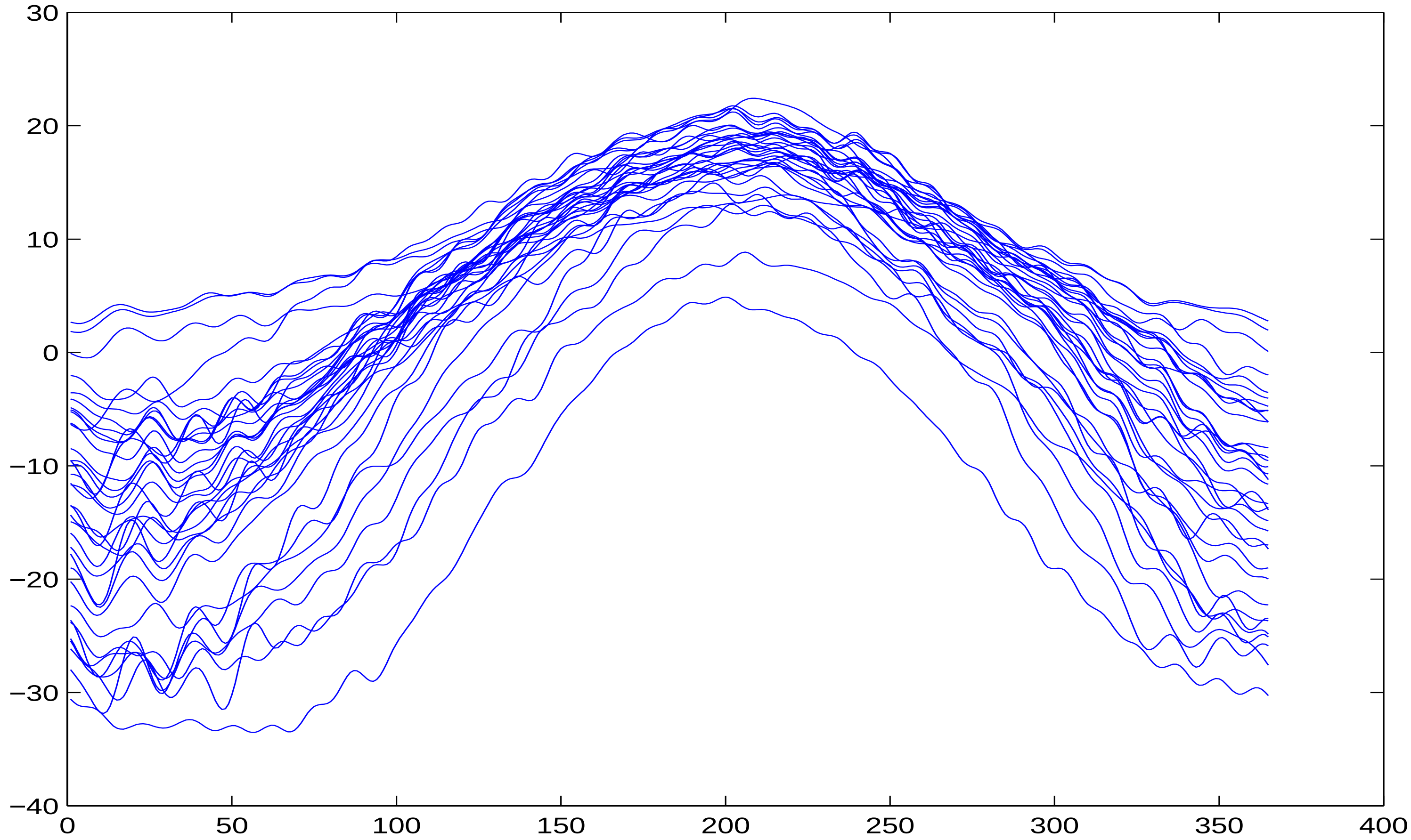
- Three deepest curves represented with colours red, green and yellow. The red curve is the median function.



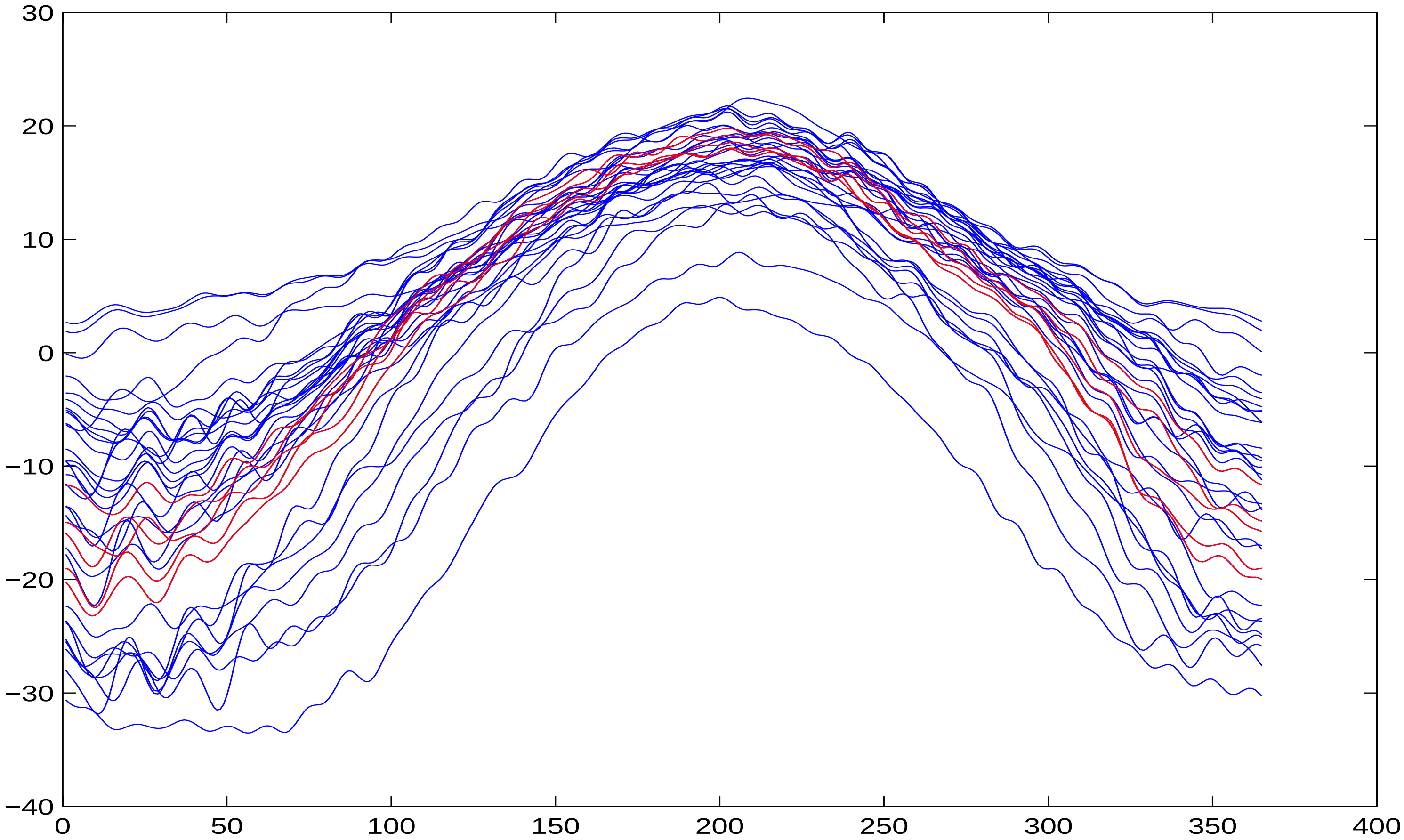
- The ten less deepest curves are in red.



- Angles in the sagittal plane formed by the knee as 39 children go through a gait cycle. The curve in red is the deepest one.



- Daily temperature in different weather stations in Canada during one year. The raw data were smoothed considering a Fourier basis with 65 elements in the basis.



- Five deepest curves with, $S_{n,J}$, $J = 3$.

6. CONCLUSIONS

- A new definition of depth for functional observations is introduced.

- This concept of depth can be particularized to the finite-dimensional case and is an alternative definition of depth for multivariate data.

- It verifies essentially the properties established by Liu (1990) and Zuo and Serfling (2000).

- It is convenient for high-dimensional data because regardless of the dimension of the data, low values of J can be considered.

REFERENCES

Arcones, M.A. and Giné, E. (1993). *Limit theorems for U-processes.*. Ann. Probab. **21**, 1494-1542.

Arcones, M.A., Chen, Z. and Giné, E. (1994). *Estimators related to U-processes with applications to multivariate medians: asymptotic normality.*. Ann. Probab. **21**, 1494-1542.

Fraiman, R. and Muniz, G. (2001). *Trimmed mean for functional data*. Test. **10**, 419-440.

Liu, R. (1990). *On a notion of data depth based on random simplices*. Ann. Statist. **18,** 405-414.

Liu, R. (1995). *Control charts for multivariate processes*. J. Amer. Statist. Assoc. **90,** 1380-1388.

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

Mahalanobis, P. C. (1936). *On the generalized distance in statistics*. Proc. Nat. Acad. Sci. India **12,** 49-55.

Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate Analysis*. Academic Press, New York.

Oja, H. (1983). *Descriptive statistics for multivariate distributions*. Statist. Probab. Lett. **1**, 327-332.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer Verlag, New York.

Ramsay, J.O. and B.W. Silverman (1997). *Functional Data Analysis*. Springer Verlag, New York.

Rousseeuw, P.J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Small, C.G. (1990). *A survey of multidimensional medians*. International Statistitical Review **58,** 263-277.

Tukey, J. (1975). *Mathematics and picturing data*. In Proceedings of the 1975 International Congress of Mathematics **2**, 523-531.

Yeh, A. and Singh, K. (1997). *Balanced confidence sets based on the Tukey depth*. J. Roy. Statist. Soc. Ser. B **3,** 639-652.

Zuo, Y. and Serfling, R.J. (2000). *General Notions of Statistical Depth Function*, Ann. Statist. **28**, 461-482.