MOX-Report No. 15/2016

A taxonomy of outlier detection methods for robust classification in multivariate functional data

Ieva, F.; Paganoni, A.M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

A taxonomy of outlier detection methods for robust classification in multivariate functional data

Francesca Ieva* and Anna Maria Paganoni$^\sharp$

April 6, 2016

$\star$ Dipartimento di Matematica “F. Enriques”
Università degli Studi di Milano
via Saldini, 50, 20133 Milano, Italy
$\sharp$ MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy

francesca.ieva@unimi.it, anna.paganoni@polimi.it

Keywords: Multivariate functional data, Outlier detection, Depth measures.

Abstract

We propose a new method for robust classification of multivariate functional data. We exploit the joint use of two different depth measures, generalizing the outliergram to the multivariate functional framework, aiming at detecting and discarding both shape and magnitude outliers in order to robustify the reference samples of data, composed by G different known groups. We asses by means of a simulation study method's performance in comparison with different outlier detection methods. Finally we consider a real dataset: we classify a data minimizing a suitable distance from the center of reference groups. We compare performance of supervised classification on test sets training the algorithm on original dataset and on the robustified one, respectively.

1 Introduction

Nowadays biomedical and healthcare studies produce more and more frequently data like signals, images and vital parameters (or a combination of these). This drives statistical research towards the identification of suitable models and inferential techniques for handling the complexity of such data. We focus our study on multivariate functional data, i.e., data where each observation is a set of possibly correlated functions of time observed at discrete points. These functions can be viewed as trajectories of stochastic processes defined on a given infinite dimensional functional space. Among others, well-knonw examples of such kind of data are the bivariate gait data set containing the simultaneous variation of the hip and knee angles of children (see Ramsay and Silverman (2005)), daily temperatures measured at 3, 9 and 12 cm below ground during 21 days (see Berrendero et al. (2011)), or 8-lead electrocardiagram (ECG) data studied in Ieva et al. (2013).

In such a context, the problem of outlier detection is a crucial point for a number of reasons. In fact, outliers are often considered as an error or noise, instead, they may carry important information on the phenomenon under study. Moreover, if not properly identified, they may lead to model misspecification, biased parameter estimation and incorrect results, especially in Functional Data Analysis, where the number of available statistical units is lower than the number of parameters. It is therefore important to identify them prior to modelling and data analysis. Febrero-Bande et al. (2008) identify two reasons for the presence of outliers in functional data. First, gross errors can be caused by errors in measurements and recording or typing mistakes, which should be identified and corrected if possible. Second, outliers can be correctly observed data curves that do not follow the same pattern as that of the majority of the curves. Moreover, in spite of the multivariate context, in the functional one there is no general definition of outliers. In fact, their nature is at least threefold: data may be *amplitude outlier* (the direct generalization of multivariate outliers), *shape outlier* (i.e., outliers with reapect to the phase), or *covariance outlier*, i.e., generated by a model that is different from the model of the central bulk of data in term of the variance-covariance operator, as detailed in Tarabelloni and Ieva (2016). In any case, their presence often depends on assumptions regarding the hidden structure of data and the applied detection method. Yet, some definitions are general enough to cope with various types of data and methods. In Hawkins (1980) the author defines an outlier as “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism”.

In general, there are two ways to treat outliers in a data sample: i) to apply outlier-detection tools and remove outlying observations from the dataset before starting the analysis and the inference; ii) to robustify the estimators adopted for carrying out the inference. Any combination of the two are also allowed. The first option requires to point out and tune suitable methods for the effective identification of outliers. Its aim is the *robustification* of the sample. In other words, it focuses on mantaining only the most representative observations, purging the sample of outliers. The second method, on the other hand, directly targets the robustification of estimators, in order not to let them to be affected by the presence of outlying observations. This is more concerned with the field of *robust statistics*. Some examples of robust estimators and practices may be found in Gervini (2008), Kraus and Panaretos (2012), Tarabelloni and Ieva (2016), among others.

In this paper we adopt the first approach and address the problem of robustifying a dataset of curves in order to improve the performances in classification and prediction of the cluster a new statistical unit belongs to. In particular we consider an application to the study of electocardiographic curves (ECGs). The dataset under study gathers more than two thousand signals with an associated clinical diagnosis. In particular, for this study we deal with three different groups (that we will call reference samples): healthy poeple, patients affected by Right and Left Bundle Brunch Block (a particular kind of Acute Myocardial Infarction) respectively. Our goal is to carry out a semi-automatic diagnosis of a new signal allocating it to one of the three groups on the basis of the curve morphology. Since this operation may strongly be biased by the presence of outliers, we aim at proving that the prediction performance may be improved if we robustify the reference samples. In fact, the presence of atypical observations may affect the statistical analysis even more in the context of supervised clustering, making the recognition of similarity patterns among curves more difficult. Therefore, we propose a new method (namely *MOUT - Multivariate OUTliergram*), consisting in the joint use of two different depth measures, to detect and discard both shape and magnitude outliers in order to robustify the reference samples and improve the assignment procedure of a new statistical unit, as explained in the following.

The new tool we propose is not the only one existing for outlier detection in multivariate functional case. Therefore, another main aim of this paper is to compare its performance with those provided by the main competitors, i.e., the Multivariate Functional Boxplot (MFB) proposed in Ieva and Paganoni (2013b), the Central Stability Plot (CSP) proposed in Hubert et al. (2015) and the Time OUtlier (TOU), consisting in the application of the Half-Space Depth of Claeskens et al. (2014) at each time point, declaring outliers those data labelled as outlier in one at least one time instance. All the methods are presented and discussed in details in Section 3. In so doing, we would like to provide a taxonomy of outlier detection methods for multivariate functional data.

The paper is organized as follows: In Section 2 the proposal of the new outlier detection method for multivariate functional data is given, starting from the definition of the indexes that are needed for its construction; in Section 3 the performances of the different multivariate functional outlier detection methods are tested and compared using some simulated datasets of multivariate functional data. Without loss of generality, we considered bivariate functional data samples and contaminated them whit different percentage of outliers. Finally, in Section 4 a real case study on Electrocardiographic (ECG) signals is considered, in order to show the impact of robustification in a real setting. Then result are discussed, together with potential further developments in Section 5.

All the analyses are carried out using the software `R` (R Core Team (2013)). The codes are available from the authors upon request, and a library for the computation of depth indexes in (multivariate) functional framework is available upon request at the BitBucket Repository (https://bitbucket.org/ntarabelloni/roahd).

2 Outlier Detection Tools in Multivariate Functional Data

As mentioned in the Introduction, there are mainly two approaches for dealing with outliers in multivariate functional case. Here we face the problem of outlier detection in multivariate functional data following a diagnostic approach, i.e., iteratively discarding outliers from the sample until no more outliers are detected. This is aimed at robustifying the samples as a preliminar step before performing an assignment of a new statistical units to one of the groups that are present in our data.

To this aim we will propose a new method for outlier detection in multivariate functional setting, based on the outliergram introduced in Arribas-Gil and Romo (2014), namely the Multivariate OUTliergram (MOUT). This generalization is introduced and relative results are provd in Subsection (2.1).

As the original version of the outliergram, MOUT performs at best in detecting shape outliers, i.e., data that are outliers in terms of phase more than in amplitude. These are often more complex to identify, since the ways a data may be outlier in terms of shape are much more with respect to the ways a data may be outlier in terms of amplitude (essentially traslational way).

Since, as we said before, no general definition of outlier exists, the same is for the methods adopted for identifying them, and a number of issuses arise (see Tarabelloni and Ieva (2016) for deeper discussion on this). This is the reason why we decided to rely on the features provided by the outliergram, showin that they are more effective than the other proposals, at least in the majority o cases where shape outliers are present. We then propose in Section (3) a comparison of the performances o out method with three competitors based on some results and graphical tools introduced in Sun and Genton (2012), Ieva and Paganoni (2013b) and Hubert et al. (2015).

2.1 The multivariate outliergram

For introducing the MOUT, let us start recalling the definition of Modified Band Depth (MBD) for univariate functional data introduced in Lopez-Pintado and Romo (2009) and Lopez-Pintado and Romo (2011). Given a stochastic process X taking values on the space $\mathcal{C}(I)$ of real continuous functions on the compact interval I, the empirical version of the band depth of order $J \geq 2$ for a function $f \in \mathcal{C}(I)$ is

$$\mathrm{BD}_X^J(f) = \sum_{j=1}^{J} \binom{N}{j}^{-1} \sum_{i_1 < i_2 < \ldots, < i_j} \mathbb{I}\left\{ G(f) \in B(f_{i_1}, \ldots, f_{i_j}), \quad \forall t \in I \right\}, \qquad (1)$$

where the subset of the plane $G(f) = \{(t, f(t)) : t \in I\}$ is the graph of the function f. $B(f_1, f_2, \ldots, f_j)$ is the band in $\mathbb{R}^2$ delimited by $f_1, f_2, \ldots, f_j$, realizations of independent copies of the stochastic process X, is defined as

$$B(f_1, f_2, ..., f_j) = \{(t, y(t)) : t \in I, \min_{r=1,...,j} f_r(t) \leq y(t) \leq \max_{r=1,...,j} f_r(t)\},\ j = 2, ..., J.$$

To overcome the problem of heavy ties due to the presence of the indicator function, Lopez-Pintado and Romo (2009) proposed the Modified Band Depth (MBD), where the time interval that f spends in the random band is weighted over I. The empirical version of the MBD is

$$\mathrm{MBD}_X^J(f) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} \tilde{\lambda}\{E(f; f_{i_1}, ..., f_{i_j})\}, \tag{2}$$

where $E(f) := E(f; f_{i_1}, ..., f_{i_j})$=$\{t \in I, \min_{r=i_1,...,i_j} f_r(t) \leq f(t) \leq \max_{r=i_1,...,i_j} f_r(t)\}$ and $\tilde{\lambda}(g) = \lambda(E(g))/\lambda(I)$ with λ the Lebesgue measure on I. In Lopez-Pintado and Romo (2009), authors state that while the choice of J clearly increases the magnitude of depth, it does not affect the induced ordering and therefore the ranks. This was supported by a simulation study in Tarabelloni et al. (2015). So given a set of curves $(f, f_1, ...f_n)$ the MBD of f, that we will denote by $MBD^J_{\{f_1,...,f_n\}}(f)$, measures the proportion of time interval I where the graph of f belongs to the envelopes of the j-tuples $(g_{i_1}, ..., g_{i_j})$, $j = 2, \ldots, J$.

On the other hand, we recall also the definition and Modified Epigraph Index (MEI) for univariate functional data introduced introduced in Lopez-Pintado and Romo (2011). Given a stochastic process X taking values on the space $\mathcal{C}(I)$ the empirical version of the epigraph index of a function $f \in \mathcal{C}(I)$ is

$$EI_X(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{f_i(t) \geq f(t), \quad \forall t \in I\right\}, \tag{3}$$

where $f_1, f_2, ..., f_n$ are realizations of independent copies of the stochastic process X. As before, to overcome the problem of heavy ties is more suitable to use the MEI, whose empirical version is

$$MEI_X(f) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\lambda}(\{t \in I, f_i(t) \geq f(t)\}). \tag{4}$$

So given a set of curves $(f, f_1, ...f_n)$ the MEI of f, that we will denote by $MEI^J_{\{f_1,...,f_n\}}(f)$, accounts for the mean proportion of time interval I where f lies below the curves of the sample.

Two are the different notions of depth measures for multivariate functional data proposed in literature: see Claeskens et al. (2014) and Ieva and Paganoni (2013b). Here we propose a generalization also of the MEI according to the one of MBD introduced in Ieva and Paganoni (2013b). Let $\mathbf{X}$ be a stochastic process taking values in the space $C(I; \mathbb{R}^h)$ of continuous functions $\mathbf{f} = (f_1, ..., f_h)$:

$I \to \mathbb{R}^h$, where I is a compact interval of $\mathbb{R}$. We have a dataset F_n constituted of $n \in \mathbb{N}$ sample observations of this process, which we indicate by $\mathbf{f}_1, \dots, \mathbf{f}_n$, $\mathbf{f}_j = (f_{j1}, ..., f_{jh})$. The MBD of $\mathbf{f}$ with respect to F_n becomes then

$$MBD^J_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}) = \sum_{k=1}^{h} p_k MBD^J_{\{f_{1k},\dots,f_{nk}\}}(f_k); \tag{5}$$

with $p_k > 0 \ \forall \ k = 1, ..., h, \quad \sum_{k=1}^{h} p_k = 1$. Analogously the MEI of $\mathbf{f}$ with respect to F_n is

$$MEI_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}) = \sum_{k=1}^{h} p_k MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k), \tag{6}$$

with $p_k > 0 \ \forall \ k = 1, ..., h, \quad \sum_{k=1}^{h} p_k = 1$. In (5) and (6) the curves that form the envelops are the components of the curves in F_n.

The idea of the paper is to jointly use the two different indexes (5) and (6) to detect and discard both shape and amplitude outliers in order to robustify a dataset. To this aim we generalize to multivariate functional data the result proved in Theorem 2.2 of Arribas-Gil and Romo (2014).

Theorem

Let F_n and $\mathbf{f}$ be in the space $\mathcal{C}(I; \mathbb{R}^h)$ of the continuous vector functions. Then

$$MBD^J_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}) \leq a_0 + a_1 MEI_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}) + a_2 n^2 (MEI_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}))^2, \tag{7}$$

where $a_0 = a_2 = -2/(n(n-1))$ and $a_1 = 2(n+1)/(n-1)$.

Proof

Using Theorem 2.2 of Arribas-Gil and Romo (2014)

$$MBD^J_{\{f_{1k},\dots,f_{nk}\}}(f_k) \leq a_0 + a_1 MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k) + a_2 n^2 (MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k))^2,$$

$\forall k = 1, ..., h$. By Jensen inequality

$$\sum_{k=1}^{h} p_k (MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k))^2 \geq \left(\sum_{k=1}^{h} p_k MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k) \right)^2 = MEI_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f})^2.$$

So, since $a_2 < 0$ and a_0, a_1 and a_2 are independent of k

$$\begin{aligned} MBD \quad & {}^J_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}) = \sum_{k=1}^{h} p_k MBD^J_{\{f_{1k},\dots,f_{nk}\}}(f_k) \\ \leq \quad & a_0 + a_1 \sum_{k=1}^{h} p_k MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k) + a_2 \left(\sum_{k=1}^{h} p_k MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k) \right)^2 \\ = \quad & a_0 + a_1 MEI_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}) + a_2 n^2 (MEI_{\{\mathbf{f_1},\dots,\mathbf{f_1}\}}(\mathbf{f}))^2. \end{aligned}$$

The inequality (7) allows us to generalize the outliergram proposed in Arribas-Gil and Romo (2014) to the case of multivariate functional data and propose the multivariate outliergram as a tool to detect outliers. Consider a a scatterplot of multivariate MBD vs multivariate MEI of data: the points far from the quadratic boundary (7) correspond to shape outliers as well as data with very low values of MBD are potential magnitude outliers. We can robustify a dataset discarding data by studying the distribution of the distances $d_i = MBD_i - (a_0 + a_1 MEI_i + a_2 n^2 (MEI_i)^2$, $i = 1, ..., n$. The data with $d_i \geq Q_{d3} + 1.5 IQR_d$, where Q_{d3} and IQR_D are the third quartile and inter-quartile range of $d_1, ..., d_n$, are classified as outliers.

As pointed out in Arribas-Gil and Romo (2014) this procedure might fail with curves with extreme multivariate MEI values (near 0 or 1) and consequently very low multivariate MBD independently of their possible atypical shape behaviour. So also in the multivariate functional framework we iteratively shift toward the center of the sample these extreme curves one by one. If for a curve the new distance between the updated multivariate MEI and relative quadratic boundary becomes outlier, we discard also this curve.

Last but not least, the choice of the MBD from Ieva and Paganoni (2013b) implies a choice for the weights p_ks averaging the contribution of each component of the multivariate unctional data. This is usually problem-driven, and in general no gold standards have still been given so far. If there is no a priori knowledge about the dependence structure between components they could be chosen uniformly (i.e. $p_k = 1/h, \forall\ k = 1, ..., h$). In Tarabelloni et al. (2015) a different choice have been proposed, taking into account the distance between the estimated variance-covariance operators of the two groups identified by the binary outcome which was the focus of the study.

In this paper we propose a choice for the weights p_k that should take in account the variability of each component of the multivariate functional process that generates data. Let us focus on the stochastic process generating data $\mathbf{X}$ with law $P_{\mathbf{X}}$. Since the time interval is compact the space $C(I; \mathbb{R}^h)$ is embedded in $L^2(I; \mathbb{R}^h)$ the space of square integrable functions $\mathbf{f} = (f_1, ..., f_h) : I \to \mathbb{R}^h$. So we assume that $\mathbf{X}$ take values on $L^2(I; \mathbb{R}^h)$. Let $\mu_l(t) = \mathbb{E}[X_l(t)]$, for each $t \in I$, denote the mean function of the $l-$component $X_l(t)$, for $1 \leq l \leq h$, then

$$\boldsymbol{\mu}_{\mathbf{X}}(t) := (\mu_1(t), \ldots, \mu_h(t))^T = \mathbb{E}[\mathbf{X}(t)]$$

is the mean function of $\mathbf{X}$. The covariance operator $\mathcal{V}_{\mathbf{X}}$ of $\mathbf{X}$ is a linear compact integral operator from $L^2(I; \mathbb{R}^h)$ to $L^2(I; \mathbb{R}^h)$ acting on a function $\mathbf{g}$ as follows:

$$(\mathcal{V}_{\mathbf{X}} \mathbf{g})(s) = \int_I V_{\mathbf{X}}(s,t) \mathbf{g}(t) dt, \tag{8}$$

The kernel $V_{\mathbf{X}}(s,t)$ is defined by

$$V_{\mathbf{X}}(s,t) = \mathbb{E}\left[(\mathbf{X}(s) - \boldsymbol{\mu}_{\mathbf{X}}(s)) \otimes (\mathbf{X}(t) - \boldsymbol{\mu}_{\mathbf{X}}(t)) \right], \quad s,t \in I \tag{9}$$

where $\otimes$ is an outer product in $\mathbb{R}^h$. For s, t fixed, $V_{\mathbf{X}}(s,t)$ is a $h \times h$ matrix, whose elements will be denoted as $V_{\mathbf{X}}^{kq}(s,t)$, for $k, q = 1, ..., h$. Let denote $V_{X_k}(s,t)$ the diagonal element $V_{\mathbf{X}}^{kk}(s,t)$. We propose to set the weight of each component proportional to the inverse of spectral norm of the variance-covariance operator of the corresponding component:

$$q_k = 1/\lambda_k^{(1)}, \quad \text{and} \quad p_k = \frac{q_k}{\sum q_k}. \tag{10}$$

where $\lambda_k^{(1)}$ is the maximum eigenvalue of the $V_{X_k}(s,t)$ operator.
In conclusion, given a sample of multivariate functional data using the choice of weights in (10), we can compute both the MBD, as in (5), and the MEI, as in (6), of each curve with respect to the sample. Then we can construct the MOUT and detect the potential outliers. We want to show that discarding the outliers pointed out using the MOUT, i.e., robustifying the data sample, improves the performances of the classification procedures.

3 Simulation studies

This section is devoted to compare the performance of the proposed procedure (MOUT) in detecting outliers with other multivariate functional outlier detection methods through a simulation study. As mentioned before, we compare four different techniques:

- **MOUT - Multivariate OUTliergram**, the multivariate functional outlier detection method proposed in Section 2.

- **MFB - Multivariate Functional Boxoplot**, the graphical tool proposed in Ieva and Paganoni (2013b). In this case, according to the order induced by the MBD in (5) and with the choice of p_ks proposed in (10), the region which contains the 50% of most central curves of the sample is constructed, then inflated by a factor $F = 1.5$ to build the fences of the functional boxplot. Given the envelope of the functions entirely contained inside the inflated region, the data crossing these fences even for one time instance are considered outliers.

- **CSP - Centrality-Stability Plot**, the method built according to the procedure described in Equation (14) of Hubert et al. (2015). For each data in the plot we compute the vertical distance between the point and the related theoretical bound. The points whose the related distances are outliers in the univariate distribution of distances of each component are considered outliers as a whole.

- **TOU - Time OUtliers** for halfspace depth. In this case, for each data we compute the halfspace depth as in Claeskens et al. (2014) for each time point. The data for which we have at least one time point outlier are considered outliers as multivariate functional data.

The MOUT, thanks to its definition and to the joint use of two different depth indexes, should perform optimally in detection of outliers in shape, whereas the MFB has been constructed mainly for detection in magnitude. Nevertheless, since the truth of the last sentence depends also on the choice of the inflation factor to construct the fences, it might work not optimally if the usual value of F is not properly tuned (see Tarabelloni and Ieva (2016)). On the other hand, both CSP and TOU are based on multivariate depths integrated over time interval; for this reason, good performances in outlier detection are expected, but also a possible large rate of false positive cases.

3.1 Models tested in the Simulation Study

We now test the four methods mentioned above on a set of simulated multivariate functional data contaminated with different percentage of outliers. We generated, without loss of generality, 200 bivariate curves contaminated by three different types of outliers (amplitude, shape and covariance outliers, as detailed in the following), with different percentage of outliers ($\nu = \{0.05, 0.1, 0.15, 0.2\}$ respectively). For each case, we applied the four methods for multivariate functional outlier detection described above and computed the proportion of correctly and falsely identified outliers. Results are reported in Table 1.

As **reference model** we choose a bivariate gaussian process

$$(X, Y), X(t) = \mu_X(t) + \mathbb{Z}_X(t), Y(t) = \mu_Y(t) + \mathbb{Z}_Y(t)$$

with means

$$\begin{aligned} \mu_X(t) &= \sin(2\pi t), \qquad t \in I = [0, 1], & (11) \\ \mu_Y(t) &= \sin(4\pi t), \qquad t \in I = [0, 1] & (12) \end{aligned}$$

and exponential Matérn covariance functions

$$\begin{aligned} \text{Cov}(Z_X(s), Z_X(t) = C_X(s, t) &= \alpha_X \exp\left(-\beta_X |s - t|\right), & (13) \\ \text{Cov}(Z_Y(s), Z_Y(t) = C_Y(s, t) &= \alpha_Y \exp\left(-\beta_Y |s - t|\right), \quad s, t \in I, & (14) \end{aligned}$$

being $\text{Cor}(Z_X(t), Z_Y(t)) = \rho$. In the following, where not explicitly declared, we will set $\alpha_X = 0.5, \alpha_Y = 0.7, \beta_X = \beta_Y = 0.4$ and the correlation between the two components is $\rho = 0.7$.

The procedures are then tested in context where contaminations are present in terms of magnitude outliers (Model 1), i.e., curves that lie far from the range of the majority bulk of data, shape outliers (Model 2), i.e. curves that present a different pattern with respect to the rest of the data, and covariance outliers

(Model 3), i.e., curves generated by a model that is different from the model of the majority of data just in term of the variance and covariance operator. For each case, the four different percentage of contaminations mentioned before ($\nu = \{0.05, 0.1, 0.15, 0.2\}$ respectively) are considered. These choices for the reference models produce as output multivariate functional data that show features typical of real functional data. The parameters of Matérn covariance tune the presence of significant exponential-like time covariance functions and a strong correlation between the two components.

Model 1 (magnitude outliers)

The first case of contamination is the one with magnitude outliers only. These are related to amplitude variability, and are the analogue of the outlyingness concept in the multivariate context. They are in general curves that lie far from the range of the majority bulk of data and are often easy to detect.

The data can be written as

$$X_i(t) = (5/2 + w_i^X)\mu_X(t) + Z_i^X(t), \tag{15}$$

$$Y_i(t) = (5/2 + w_i^Y)\mu_Y(t) + Z_i^Y(t), \tag{16}$$

where w_i^X and w_i^Y are an exponential sample, i.e. $w_i^X \sim \mathcal{E}(2)$, $w_i^Y \sim \mathcal{E}(2)$ μ_X, μ_Y are the means of the reference model and $(Z_i^X(t), Z_i^Y(t))$ is a realization from a centered stochastic bivariate gaussian process with the same exponential covariance as the reference model. The random multiplicative terms in the pointwise mean functions drive the presence of magnitude outliers, see Figure 1.

Model 2 (shape outliers)

The second case of contamination is the one with only shape outliers. These are related to phase variability and does not have a counterpart in classic multivariate statistics. In general they can be hidden in the middle of the sample of curves and the detection of these outliers could be difficult.

The contaminated data can be written as

$$X_i(t) = w_i^X \tilde{\mu}_X(t) + Z_i^X(t), \tag{17}$$

$$Y_i(t) = w_i^Y \tilde{\mu}_Y(t) + Z_i^Y(t), \tag{18}$$

where w_i^X and w_i^Y are an exponential sample, i.e., $w_i^X \sim \mathcal{E}(2)$, $w_i^Y \sim \mathcal{E}(2)$, $\tilde{\mu_X}(t) = \sin(2\pi(t - 0.5)), \tilde{\mu_Y}(t) = \sin(4\pi(t - 0.25))$ are the means of the contaminated model and $(Z_i^X(t), Z_i^Y(t))$ is a realization from a centered stochastic bivariate gaussian process with the same exponential covariance as the reference model. The shift in time dependence of the pointwise mean functions drive the presence of shape outliers, see Figure 2.

Model 3 (covariance outliers)

The third case of contamination is the one with the so called covariance outliers. These data are generated changing, with respect to the reference model, the parameters of the variance covariance structure.

The contaminated data can be written as

$$X_i(t) = w_i^X \mu_X(t) + Z_i^X(t), \quad (19)$$
$$Y_i(t) = w_i^Y \mu_Y(t) + Z_i^Y(t), \quad (20)$$

where w_i^X and w_i^Y are an exponential sample, i.e. $w_i^X \sim \mathcal{E}(2)$, $w_i^Y \sim \mathcal{E}(2)$, μ_X, μ_Y are the means of the reference model and $(Z_i^X(t), Z_i^Y(t))$ is a realization from a centered stochastic bivariate gaussian process with exponential Matérn covariance functions, as in the reference model, but such that $\alpha_X = 1.5, \alpha_Y = 1.7, \beta_X = \beta_Y = 1$ and the correlation between the tho components is $\rho = 0$. Even if the covariance operator has the same Matérn structure of the reference model in (13), the different choice for the parameters α, β and ρ drives the presence of covariance outliers, see Figure 3.

In Figures 1, 2 and 3 we show the curves generated with the three different models in a single simulation run and with a percentage of contaminated data equal to $\nu = 0.1$.

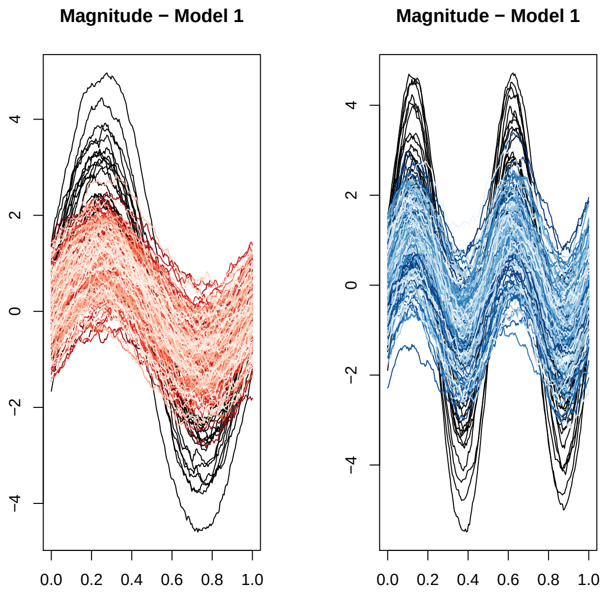

Figure 1: Model 1 (magnitude outliers), ν=0.1

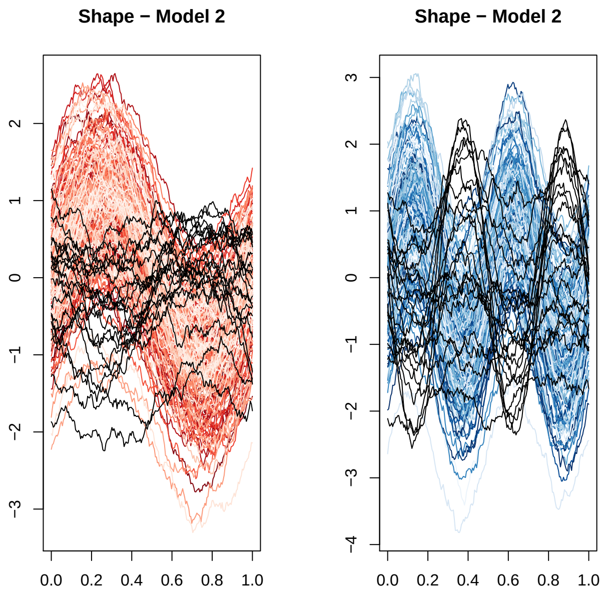

Figure 2: Model 2 (shape outliers), ν=0.1

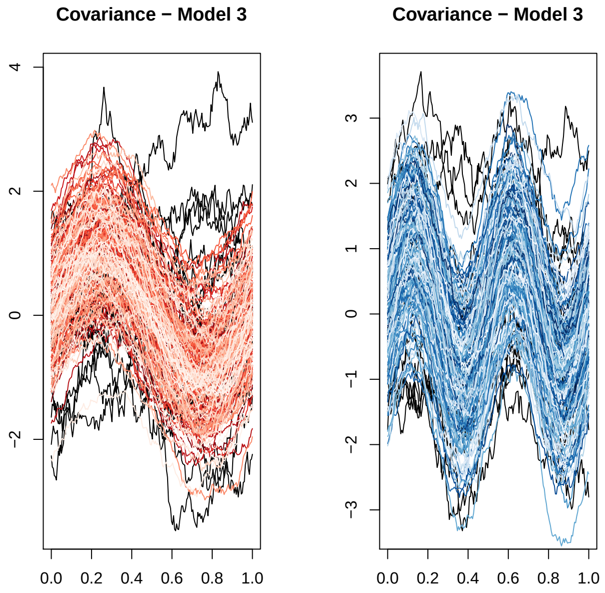

Figure 3: Model 3 (covariance outliers), ν=0.1

	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
$\nu = 0.05$						
MOUT	0.9890 (0.0314)	0.0293 (0.0144)	0.9030 (0.1087)	0.0304 (0.0145)	0.9000 (0.0000)	0.0316 (0.0000)
MFB	0.1920 (0.1292)	0.0001 (0.0007)	0.0680 (0.0764)	0.0001 (0.0005)	0.0980 (0.0943)	0.0001 (0.0005)
CSP	1.0000 (0.0000)	0.0386 (0.0151)	0.9830 (0.0451)	0.0365 (0.0116)	0.9610 (0.0650)	0.0394 (0.0137)
TOU	1.0000 (0.0000)	0.0261 (0.0293)	0.9230 (0.0897)	0.02670 (0.0272)	0.9180 (0.0869)	0.0235 (0.0184)
$\nu = 0.1$						
MOUT	0.9685 (0.0406)	0.0172 (0.0111)	0.7745 (0.1038)	0.0179 (0.0119)	0.6920 (0.1189)	0.0216 (0.0109)
MFB	0.1310 (0.0918)	0.0001 (0.0006)	0.0430 (0.0503)	0.0001 (0.0006)	0.0650 (0.0601)	0.0001 (0.0008)
CSP	0.8880 (0.1604)	0.0198 (0.0115)	0.6520 (0.2122)	0.0208 (0.0126)	0.9060 (0.0641)	0.0244 (0.0118)
TOU	0.9650 (0.0626)	0.0242 (0.0142)	0.6445 (0.1410)	0.0274 (0.0291)	0.8935 (0.0642)	0.0203 (0.0210)
$\nu = 0.15$						
MOUT	0.8410 (0.0823)	0.0065 (0.0063)	0.5637 (0.1264)	0.0087 (0.0078)	0.5767 (0.1148)	0.0113 (0.0084)
MFB	0.1140 (0.0815)	0.0001 (0.0006)	0.0260 (0.0353)	0.0000 (0.0000)	0.0530 (0.0416)	0.0000 (0.0000)
CSP	0.1287 (0.0881)	0.0264 (0.0159)	0.2017 (0.0920)	0.0261 (0.0148)	0.8110 (0.0865)	0.0117 (0.0076)
TOU	0.2647 (0.0950)	0.0236 (0.0142)	0.2510 (0.0859)	0.0218 (0.0129)	0.8580 (0.0652)	0.0214 (0.0327)
$\nu = 0.2$						
MOUT	0.4322 (0.1623)	0.0011 (0.0024)	0.3150 (0.1082)	0.0029 (0.0037)	0.4512 (0.1045)	0.0041 (0.0056)
MFB	0.0628 (0.0550)	0.0000 (0.0000)	0.0102 (0.0171)	0.0000 (0.0000)	0.0370 (0.0303)	0.0000 (0.0000)
CSP	0.0305 (0.0281)	0.0491 (0.0201)	0.0930 (0.0496)	0.0346 (0.0200)	0.6755 (0.0928)	0.0046 (0.0046)
TOU	0.1535 (0.0589)	0.0221 (0.0139)	0.1703 (0.0653)	0.0208 (0.0119)	0.8098 (0.0808)	0.0117 (0.0084)

Table 1: Mean (standard deviation) over 100 simulation runs of the proportion of true positive p_c and false positive p_f in the three different Models

Table 1 reports the obtained results, in term of the proportion of true positive p_c (i.e., the number of correctly identified outliers over the number of outliers in the dataset) and the proportion of false positive p_f (i.e., the number of wrongly identified outliers over the number of non-outlying curves in the dataset).
Let us observe that the proposed multivariate outliergram (MOUT) performs very well in particular in shape outliers detection, as expected, also for high values of contamination. The MFB has a very low efficiency in detecting also magnitude outliers, and it guarantees no false positive cases. The two methods based on integrated multivariate depths (CSP and TOU) show very remarkable performance in all the three methods; however the rate of false positive il in general higher than the parallel index of MOUT, and the percentage of correctly identified outliers decreases in particular for magnitude and shape outliers as long as the percentage of contaminated data ν increases. Supported by the results of the simulation study we decide to go deeper in studying MOUT properties and to propose it as a preprocessing instrument to robustify a sample of curves in order to improve classification procedures. In the following section we will present the use of MOUT in a classification problem of real data: electrocardiographic signals of healthy and pathological subjects.

4 Application to ECG signals

In this section we apply the method MOUT proposed in Section 2 for robustifying a dataset of vital signals (ECGs) arising from a real case study where the basic statistical unit (the patient) is characterized by a 8-variate function (the ECG), which describes his/her heart dynamics on the eight leads I, II, V1, V2, V3, V4, V5 and V6. ECG traces are collected in the PROMETEO database. PROMETEO project (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero) has been started in 2008 with the aim of spreading the intensive use of ECGs as pre-hospital diagnostic tool. The project was also a way of constructing a new database of ECGs with features never recorded before in any other data collection on heart diseases. In fact, both physiological signals (i.e., signals from patients not really having heart diseases) and pathological ones are present. See Indino (2015) for a detailed description of the dataset and of the techniques to denoise, smooth and register data.

For each statistical unit in the dataset we have a list of technical information, useful for signal processing and analysis. More precisely, they include waves repolarisation and depolarisation times, landmarks indicating onset and offset times of the main ECG's subintervals and an automatic diagnosis, established by the commercial Mortara-Rangoni VERITAS$^{\text{TM}}$ algorithm. We used these automatic diagnoses to label the ECG traces we analysed. Moreover, for each patient a reference beat signal lasting 1.2 seconds and measured on a grid of 1200 points (msec) is provided. It is built from the heartbeat rhythm of the patient measured over 10 seconds (10000 sampled time points). We then analyze 8 curves (one for each lead of the ECG) for each patient, representing the patient's "Median" beat for that lead. This representative heartbeat is a trace of a single cardiac cycle (heartbeat), i.e., of a P wave, a QRS complex, a T wave, and a U wave, which are normally visible in 50% to 75% of ECGs (see Ieva et al. (2013) for further details).

Among the diseases that are present in the PROMETEO database, we focus our analysis on $G = 3$ groups of patients: Normal (i.e., healthy people), those affected by Left Bundle Branch Blocks (LBBBs) and Right Bundle Branch Blocks (RBBBs), a particular kind of Myocardial Infarction, which are among the most common and easily detectable through the analysis of the ECG signal.

So, the dataset we analyse consists of the ECG signals of $n = 2102$ subjects, among which 1602 are Normals, 224 are affected by LBBB and 276 by RBBB. Figure 4 shows denoised and registered lead I data we consider for our analysis (see Ieva et al. (2013) for further details on wavelet denoising and landmarks registration adopted for preprocessing data).
The analysis is carried out separately for each of the G groups present in the original dataset. We randomly splitted each group in a training (85%) and a test set (15%): for each element of the test set we compute the L^2 distance with respect to the mean of each training set group and we classify it according to a minimization criterium.

In each group of data (Normals, LBBBs and RBBBs) separately we carried out the robustification procedure detailed above, discarding the outliers pointed out both by the multivariate functional boxplot and by the multivariate outliergram. The robustified dataset is then composed by $n = 2020$ subjects, among

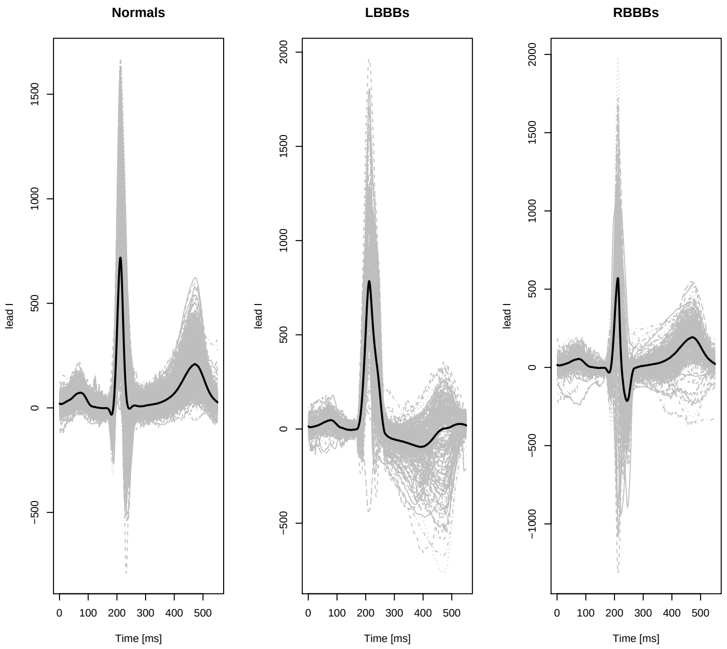

Figure 4: Denoised and registered data (lead I) for the three groups of patients with superimposed the mean functions (black solid lines).

which 1564 are Normals, 205 are affected by LBBB and 251 by RBBB.

We then tested the efficacy of this procedure classifying the curves of the test sets according to the following procedure: we compute the L^2 distances between each curve of this test set and the mean of the three groups and we assign a curve to the group that minimizes this distance. We repeated this analysis $k = 40$ times to avoid bias selection.

The mean (± sd) correct classification rate, over the $k = 40$ runs of the procedure is $0.9108(\pm 0.0163)$ if we consider the robustified training datasets, while it is $0.9014(\pm 0.0148)$ in the case of the original (i.e., non robustified training datasets). There is statistical evidence (p-value of Wilcoxon comparison test equal to 0.0075) to conclude that the distribution of correct classification rates over the 40 repetitions after robustification is stochastically greater than the one computed before robustification. Similar results are obtained using the median, instead of the mean, in the minimization of the L^2 distances.

5 Conclusions and further developments

In this paper a new method, namely MOUT - Multivariate OUTliergram, for robustifying a dataset of multivariate functional data is proposed. It is aimed at identifying and discarding multivariate functional outliers from a dataset in order to enforce the performances of a classifier in terms of prediction for a new statistical unit. In fact, the presence of outliers in high dimensional context like Functional Data Analysis may may lead to model misspecification, biased parameter estimation and incorrect results.

Since in the (multivariate) functional one there is no general definition of outliers, in this work we considered many different scenarios for testing the performances of the new method. We settled suitable simulation studies considering amplitude, shape and covariance outliers with different percentage of outliers, and for each of them compared the new technique proposed with a list of competitors from the literature. From the discussion of the results, it emerged that MOUT performs very well in identifying shape outliers, but also provides satisfying performance in all the other contexts.

We finally applied the MOUT to a real case study on ECG curves, aimed at performing a semi-automatic diagnosis for Acute Myocardial Infarction. After the robustificaton carried out through the MOUT, the performance of the classifier improved significantly.

In summary, this paper provides a taxonomy of the main methods for outlier detection in multivariate functional context, testing their performances in many different scenarios. Moreover, it discusses a crucial topic that is becoming more and more central in high dimensional contexts like FDA, that is how to treat outliers once you are able to define and identify them. The computational tools proposed in the paper are available in the free language `R`, suitably organized in the package `roahd`, available at `https://bitbucket.org/ntarabelloni/roahd`.

References

A. Arribas-Gil and J. Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 2014.

K. Bache and M. Lichman. *UCI Machine Learning Repository*, 2013. URL http://archive.ics.uci.edu/ml.

J. Berrendero, A. Justel, and M. Svarc. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9):2619 – 2634, 2011. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2011.03.011.

G. Claeskens, M. Hubert, L. Slaets, and K. Vakili. Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109(505):411–423, 2014.

M. Dyrby, S. Engelsen, L. Nørgaard, M. Bruhn, and L. Lundsberg-Nielsen. Chemometric quantitation of the active substance in a pharmaceutical tablet using nearinfrared (nir) transmittance and nir ft-raman spectra. *Applied Spectroscopy*, 56(5):579585, 2002.

M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Outlier detection in functional data by depth measures, with application to identify abnormal no$_x$ levels. *Environmetrics*, 19(4):331–345, 2008.

D. Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587600, 2008.

D. M. Hawkins. *Identification of outliers*. Springer, 1980.

V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, 2004.

M. Hubert, P. Rousseeuw, and P. Segaert. Multivariate functional outlier detection. *Statistical Methods and Applications*, 24(2):177–202, 2015.

F. Ieva and A. Paganoni. Risk prediction for myocardial infarction via generalized functional regression models. *Statistical Methods in Medical Research*, In press, 2013a.

F. Ieva and A. M. Paganoni. Depth measures for multivariate functional data. *Communication in Statistics - Theory and Methods*, 42(7):1265 – 1276, 2013b.

F. Ieva, A. Paganoni, D. Pigoli, and V. Vitelli. Multivariate functional clustering for the morphological analysis of ecg curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):401 – 418, 2013.

F. Indino. Analisi statistica di dati ad alta dimensionalit : un'applicazione ai segnali elettrocardiografici. Master's thesis, Politecnico di Milano, 2015. URL https://www.politesi.polimi.it/handle/10589/107284.

J.Li and R. Liu. New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*, 19:686 – 696, 2004.

D. Kraus and V. M. Panaretos. Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99(4):813832, 2012.

F. Larsen, F. van den Berg, and S. Engelsen. An exploratory chemometric study of nmr spectra of table wines. *Journal of Chemometrics*, 20(5):198–208, 2006.

R. Liu and K. Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252 – 260, 1993.

S. Lopez-Pintado and J. Romo. Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 51(10):4957–4968, 2007.

S. Lopez-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718 – 734, 2009.

S. Lopez-Pintado and J. Romo. A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55:1679 – 1695, 2011.

S. Lopez-Pintado, Y. Sun, and M. Genton. Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8:321–338, 2014.

D. Pigoli, J. Aston, I. Dryden, and P. Secchi. Distances and inference for covariance functions. *Biometrika*, 101(2):409 – 422, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, New York, second edition, 2005.

P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, third edition, 2003.

Y. Sun and M. Genton. Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics*, 23(1):5364, 2012.

N. Tarabelloni. *Tools for computational statistics coded in C++*, 2013. URL https://github.com/ntarabelloni/HPCS.

N. Tarabelloni and F. Ieva. On data robustification in functional data analysis. MOX Report 03/2016, Department of Mathematics - Politecnico di Milano, 2016. URL https://www.mate.polimi.it/biblioteca/add/qmox/03-2016.pdf.

N. Tarabelloni, F. Ieva, R. Biasi, and A. Paganoni. Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiographic signals. *The International Journal of Biostatistics*, (To appear), 2015.

J. Tuckey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver*, volume 2, pages 523 – 531, 1975.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

14/2016 Bonomi, D.; Manzoni, A.; Quarteroni, A.
A matrix discrete empirical interpolation method for the efficient model reduction of parametrized nonlinear PDEs: application to nonlinear elasticity problems

13/2016 Guerciotti, B; Vergara, C; Ippolito, S; Quarteroni, A; Antona, C; Scrofani, R.
Computational study of the risk of restenosis in coronary bypasses

12/2016 Bartezzaghi, A.; Dedè, L.; Quarteroni, A.
Isogeometric Analysis of Geometric Partial Differential Equations

11/2016 Zhu, S.; Dedè, L.; Quarteroni, A.
Isogeometric Analysis and proper orthogonal decomposition for the acoustic wave equation

10/2016 Flemisch, B.; Fumagalli, A.; Scotti, A.
A review of the XFEM-based approximation of flow in fractured porous media

08/2016 Dassi, F.; Perotto, S.; Si, H.; Streckenbach, T.
A priori anisotropic mesh adaptation driven by a higher dimensional embedding

09/2016 Rizzo, C.B.; de Barros, F.P.J.; Perotto, S.; Oldani, L.; Guadagnini, A.
Relative impact of advective and dispersive processes on the efficiency of POD-based model reduction for solute transport in porous media

07/2016 Pacciarini, P.; Gervasio, P.; Quarteroni, A.
Spectral Based Discontinuous Galerkin Reduced Basis Element Method for Parametrized Stokes Problems

05/2016 Alfio Quarteroni, A.; Lassila, T.; Rossi, S.; Ruiz-Baier, R.
Integrated Heart - Coupling multiscale and multiphysics models for the simulation of the cardiac function

06/2016 Micheletti, S.; Perotto, S.; Signorini, M.
Anisotropic mesh adaptation for the generalized Ambrosio-Tortorelli functional with application to brittle fracture