MOX-Report No. 09/2018

# Profile Monitoring of Probability Density Functions via Simplicial Functional PCA with application to Image Data

Menafoglio, A.; Grasso, M.; Secchi, P.; Colosimo, B.M.

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

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

# Profile Monitoring of Probability Density Functions via Simplicial Functional PCA with application to Image Data

A. Menafoglio[1]*, M. Grasso[2], P. Secchi[1]. B.M. Colosimo[2]*

[1]MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
[2]Dipartimento di Meccanica, Politecnico di Milano
*biancamaria.colosimo@polimi.it

**Abstract**

The advance of sensor and information technologies is leading to data-rich industrial environments, where big amounts of data are potentially available. In this scenario, image data play a relevant role, as they can easily describe many phenomena of interest. This study focuses on images where several and similar features of interest are randomly distributed and characterized by no spatially correlated structure. Examples are pores in parts obtained via casting or additive manufacturing, voids in metal foams and light-weight components, grains in metallographic analysis, etc. The proposed approach consists of summarizing the random occurrences of the observed features via its (empirical) probability density function (PDF). In particular, a novel approach for PDF monitoring is proposed. It is based on simplicial functional principal component analysis (SFPCA), which is performed by applying an isometric isomorphism between the space of density functions, i.e., the Bayes space $B^2$, and the space of square integrable functions $L^2$. A simulation study shows the enhanced monitoring performances provided by the SFPCA-based profile monitoring against other competitors proposed in the literature. Eventually, a real case study dealing with the quality control of foamed materials production is discussed, to highlight a practical use of the proposed methodology.

**Keywords**: statistical process control, image-based process monitoring, functional data analysis, constrained curves, Bayes space

## 1 Introduction

In the recent years, we are experiencing a quick evolution towards digitalized factories where continuously evolving sensor and information technologies are shaping data-rich industrial environments. On the one hand, the use of novel in-line sensing solutions

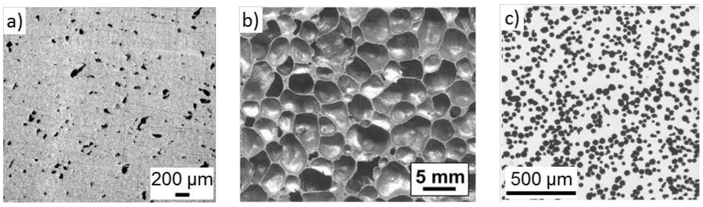

**Figure 1:** Examples of image data for quality monitoring applications: a) pores in products obtained via metal additive manufacturing , b) cross-section of a metal foam , c) micrographs of steel powder particles for metal additive manufacturing

(e.g., machine vision systems, non-contact in-line metrology, etc.), allows one to link the quality and stability of processes to high-frequency streams of images, surface elevation maps, 3D data clouds, etc. On the other hand, emerging production technologies (e.g., additive manufacturing, Gibson et al., 2010), pave the way to products characterized by more and more complicated shapes and lightweight structures.

In this framework, Wells et al., (2013) and Wang and Tsung (2005) suggested an approach for statistical process control (SPC) of high-dimensional data via profile monitoring. An extended literature has been devoted to the suite of profile monitoring techniques so far: the interested reader may refer to Woodall et al., 2004 and Noorossana et al., 2012 for an overview. The main idea of Wang and Tsung (2005), later extended by Wells et al., (2013), consists of translating huge sample size data or high density point clouds into linear profiles through the use of Q-Q plots, and then applying traditional profile monitoring to check the stability of Q-Q plot parameters. This paper applies a similar rationale for monitoring the occurrence of random shapes in image data, as pores in casting/additively-manufactured components (Fig. 1 a), voids in lightweight metal foams (Fig. 1 b), or powder grains for additive manufacturing (Fig. 1 c). In particular, we attempt to generalize the seminal idea proposed by Wang and Tsung (2005) and Wells et al. (2013) by monitoring the functional shape of probability density functions (PDFs) rather than Q-Q plot parameters.

With reference to the examples shown, we assume that one synthetic descriptor is primarily of interest (usually the area of pores/voids/grains) and no spatial correlation structure is observable (i.e., the features are randomly distributed within the image area). As a matter of fact, the proposed approach deals with univariate PDFs, although extensions to the multivariate case can be easily envisaged. In case of spatially correlated features, the proposed method can be applied to residuals obtained after removing the correlation structure (that should be separately monitored with an additional control chart).

PDFs represent a special case of functional data (Ramsay, 2005) that, in principle, can be modelled and monitored via functional principal component analysis (FPCA) (Colosimo and Pacella, 2007; 2010), provided that the constrained nature of PDFs is appropriately taken into account. As a matter of fact, when the functional data of interest is a PDF, $f(x)$, two constraints have to be satisfied, namely (i) $f(x) > 0$ and

(ii) $\int f(x) = 1$. By performing PCA on PDFs, even if the loadings should form an empirical basis for the original data, densities approximated on the basis of the retained functional principal components (FPCs) may violate the above constraints (Delicado (2011), Hron et al. (2016)). This may have a detrimental effect both on process monitoring performances and on dimensionality reduction capabilities. A number of authors (e.g., Egozcue et al. (2006), Van den Boogaart et al., 2010, Delicado (2011), Van den Boogaart et al. (2014), Menafoglio et al. (2014, 2016a, 2016b), Hron et al. (2016)) pointed out that PDFs can be interpreted as functional compositional data, i.e., functional observations carrying only relative information, which are usually collected in the form of constrained data integrating to a constant. Traditional FDA techniques operate in the space of square-integrable real measurable functions $L^2$, whereas compositional data entails the use of a different space, known as Bayes space, $B^2$ (Egozcue et al. (2006), Van den Boogaart et al., 2010, Egozcue et al. (2013), Van den Boogaart et al. (2014)), that generalizes to the functional setting the well-known Aitchison geometry for compositional data (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001; Egozcue and Pawlowsky-Glahn, 2006; Egozcue, 2009; Pawlowsky-Glahn and Buccianti, 2011). Therefore, the theory of Bayes spaces can be used to extend the applicative domain of FDA techniques to probability density curves.

This study presents a novel approach for FPCA-based profile monitoring of empirical PDFs by exploiting their inner constrained nature. The proposed methodology relies on a variant of the FPCA known as simplicial functional principal component analysis (SFPCA), where the term "simplicial" refers to the use of an infinite dimensional simplex. The method is based on applying an isometric isomorphism between the space of density functions $B^2$ and $L^2$. Such an isomorphism allows one to resort to traditional FDA tools by preserving the capability of properly dealing with density curve constraints.

A motivating real case study dealing with the quality control of foamed materials production is presented and discussed. Metal foams are special cases of porous metals with a cellular structure characterized by interesting combinations of physical and mechanical properties (Banhart, 2001). A simulation analysis is discussed to demonstrate the enhanced monitoring performances provided by the SFPCA-based profile monitoring against the traditional FPCA-based monitoring and other benchmarks, including the Q-Q plot-based scheme and simple Shewhart's control charts.

The remaining part of this work is organized as follow. Section 2 introduces the real case study; Section 3 describes the SFPCA technique and how it is implemented into the proposed profile monitoring framework; Section 4 presents the simulation study and the comparison against competitor methods; Section 5 presents the results achieved on real data; Section 6 eventually concludes the paper.

## 2 A motivating real case study

The use of image data for part quality inspection is becoming more widespread in industry (Qiu, 2005; Yan et al., 2015; Megahed et al., 2011). Random porous and cellular materials represent a category of complicated structures where novel SPC methods are needed to cope with the challenging nature of quality inspection data (Kim et al., 2014; Zhuravleva et al., 2013; Campoli et al., 2013; Banhart, 2001). In this framework, practitioners need effective methods to describe, in a synthetic way, the quality signatures enclosed in cross-section images. Indeed, the quality of the part is related to descriptors (e.g., size or shape of random features) that can be roughly summarized by their first few statistical moments. However, a better description can be ascribed to the whole PDF of the feature descriptors. As a matter of fact, the shape of the PDF can be used as a quality signature to determine both the quality of the part and the stability of the process.

This Section presents a real case study regarding the characterization of porous materials known as "metal foams" via image analysis. Metal foams are a special case of highly porous materials with a cellular structure characterized by interesting combinations of physical and mechanical properties, i.e., high stiffness at low specific weight or high gas permeability at high thermal conductivity (Banhart, 2001; Strano, 2001; Villa et al., 2011). In this framework, the improvement of the production process is aimed at achieving a better reproducibility and predictability of the morphological and structural homogeneity of cellular structures.

Two aluminium foam samples produced via the powder compact melting technique (Banhart, 2001) are considered in our study. The samples, called sample A and sample B, respectively, were analysed via optical image analysis, which consists of cutting the sample, polishing the slices and capturing high contrast pictures where cell membranes and the interior of the cells appear in different brightness (Banhart, 2001). Despite being a destructive technique, it is quite useful during the material development and process setup phase to determine the pore size distribution and/or to perform a shape analysis of the cells. The two samples, of diameter $D = 35$ mm, were produced under the same process conditions, but different methods for slice polishing on the preparation plane were adopted. Thus, a process monitoring tool is expected to detect a shift in the pore-size distribution whose assignable cause is a change of the polishing treatment. Both the samples were cut in such a way that the distance between consecutive sections was higher than the larger expected pore size, in order to minimize the between-section correlation. Fig. 2 shows the binary images generated after basic image pre-processing steps for a sub-sample of 10 sections from Sample A (27 sections in total) and a sub-sample of 10 sections from Sample B (30 sections in total). The images of all the sections are shown in the Supplementary Material.

In both the samples, the distribution of the cell sizes exhibit a certain randomness, typical of the powder compact melting processes. The pore-size descriptor used in this study is the "area ratio", which is defined as:

$$A_r(i,j) = \frac{A_i(j)}{A_{tot}(j)}, \quad i = 1, ..., N_j;\ j = 1, 2, ... \tag{1}$$

where $A_i(j)$ is the area of the $i$-th pore in the $j$-th section, $N_j$ is the number of pores in the $j$-th section and $A_{tot}(j)$ is the total area of the $j$-th section. In order to determine if the $A_r(i,j)$ exhibits some statistically significant spatial autocorrelation, the

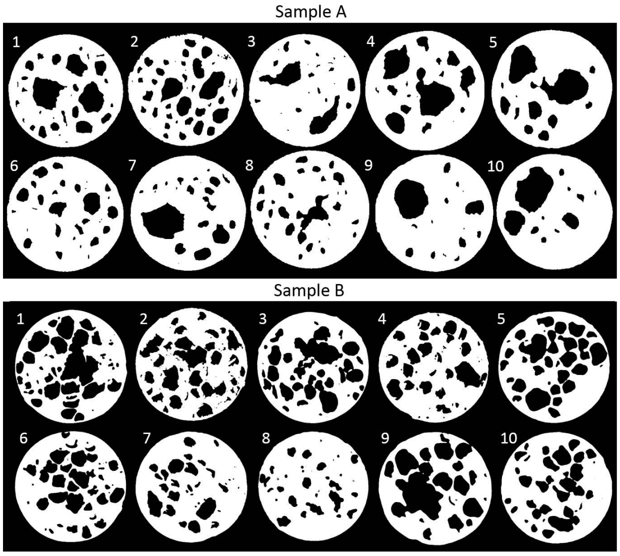

**Figure 2:** An example of 10 sections from Sample A (top) and Sample B (bottom).

global Moran's test (Bivand and Piras, 2015) was applied to the metal foam data in both the samples, revealing no significant spatial autocorrelation for the analysed sections. Due to the lack of spatial autocorrelation, the probability density curves of the $A_r(i, j)$ descriptor can be used as synthetic process signatures. This allows transforming the information content enclosed by original cross-section images into 1D curves that can be monitored via profile monitoring methods.

# 3 Proposed methodology

The proposed profile monitoring approach envisages a training phase (a.k.a. Phase I in SPC literature) and a monitoring phase (Phase II). The training phase aims at (i) characterizing the natural (in-control) variability of the process described in terms of PDFs of the univariate statistical descriptor of interest, and (ii) estimating the control limits for process monitoring. During the training phase, three major steps are needed: (i) PDF estimation, (ii) SPFCA and selection of the number $K$ of SFPCs to retain and (iii) multivariate control chart design for the scores associated with the retained SFPCs and the SFPCA reconstruction error along the directions orthogonal to the first $K$ SFPCs. During the monitoring phase, the PDF of each new sample is estimated and it is projected onto the space spanned by the $K$ retained SFPCs. If the control statistics violate the previously designed control limits, an alarm is signalled, otherwise the new density curve is deemed representative of an in-control state. The different steps of the proposed approach are schematically depicted in Fig. 3 and described in more detail in the following subsection. For a brief review of the FPCA methodology the reader is referred to Colosimo and Pacella (2007; 2010).

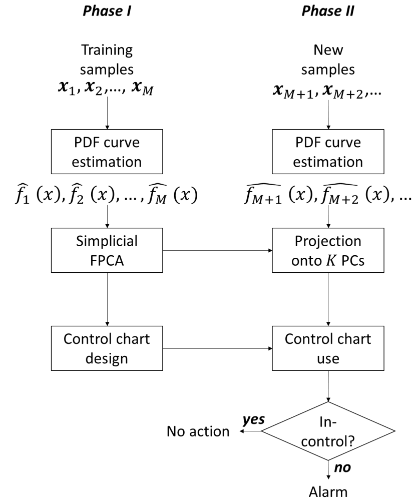

**Figure 3:** Scheme of the proposed approach

## 3.1 Probability density curve estimation

Having acquired the training sample of descriptors extracted from the $j$-th sample, denoted by $\mathbf{x}_j = (x_{1j}, ..., x_{N_j j})$ with $j = 1, 2, ...$, we first aim to compute a smooth estimate of the underlying distribution, described equivalently by the PDF $f_j(x)$ or by the cumulative distribution function (CDF) $F_j$. We will always assume $F_j$ to be continuous, and supported on a compact domain $[0, 1]$, for $= 1, 2, ...$ . Note that the case of a general compact support $[a, b]$ is obtained through the variable transformation $x = \frac{(t-a)}{(b-a)}$, with $t \in [a, b]$. Given a sample of i.i.d. observations $x_{1j}, ..., x_{N_j j}$ from $F_j$, a (discontinuous) non-parametric estimator for $F_j$ is given by the Empirical Cumulative Distribution Function (ECDF), denoted by $F_{N_j}$ and defined as

$$F_{N_j}(x) = \frac{1}{N_j} \sum_{i=1}^{N_j} I_{x_{ij} < x}. \tag{2}$$

Although $F_{N_j}$ is strongly consistent for $F_j$, it does not provide a smooth estimate of the underlying CDF, since it has jump discontinuities (of amplitude $1/N_j$) in correspondence of the observations. The problem of smoothly estimating the underlying CDF $F_{N_j}$ can be overcome by smoothing the ECDF through the use of, e.g., Bernstein Polynomials. The use of Bernstein Polynomials to this purpose is well documented in the literature (e.g., Vitale (1975), Petrone (1999), Leblanc (2010) and references therein). We here follow the approach of Babu et al. (2002), who proposed the estimator

$$\widehat{F}_j(x; N_j, B_j) = \sum_{k=0}^{B_j} F_{N_j}(k/B_j) b_{k,B_j}(x), \tag{3}$$

where $b_{k,B_j}(x) = B_j k x^k (1-x)^{B_j-k}$, $k = 0, ..., B_j$. Estimator (3) is still strongly consistent for $F_j$ (Babu et al., 2002), but it is also continuous and allows computing the associated (smooth) PDF $\widehat{f}_j$ as

$$\widehat{f}_j(x; N_j, B_j) = B_j \sum_{k=0}^{B_j-1} \left( F_{N_j}((k+1)/B_j) - F_{N_j}(k/B_j) \right) b_{k,B_j-1}(x). \tag{4}$$

As opposed to kernel smoothing estimators (e.g., Rosenblatt 1956; Parzen 1962; Silverman 1986), the estimator is well suited for distributions with compact support, as those we analyse. Note that estimator (3) only depends on $F_{N_j}$ and on the number $B_j$ of Bernstein polynomials. The higher $B_j$, the better the fitting of the ECDF, at the expense of a more fluctuating estimate (i.e., bias-variance trade-off). Based on simulations, Babu et al. (2002) found acceptable the use of $B_j = \frac{N_j}{\log(N_j)}$. We here set $B_j = N_j$ to avoid over-smoothing of the resulting PDFs, which may turn in losing interesting features of the samples and power to discriminate between in-control and out-of-control conditions. More refined methods to estimate the optimal value of $B_j$ have been recently proposed by Dutta (2016), but are not considered here for the sake of limiting the computational cost of the procedure.

### 3.2 SFPCA of probability density curves

Given the smooth density functions $\widehat{f}_1, ..., \widehat{f}_M$ estimated for the indicators of each of the $M$ Phase I realizations, we now aim to explore the variability of the dataset and consistently reduce its dimensionality, while properly accounting for the data constraints (i.e., positivity and integral constraint). We here consider the densities $\widehat{f}_1, ..., \widehat{f}_M$ as elements of the Bayes space $B^2$, that is the space of (equivalence classes) of positive functions integrating to a constant, with square-integrable logarithm, i.e.,

$$B^2 = \left\{ f : f > 0, \int_0^1 f(t)dt = c, \log(f) \in L^2 \right\}. \tag{5}$$

The space $B^2$ can be equipped with the operations of perturbation $\oplus$ (that plays the role of the sum) and powering $\odot$ (i.e., the product by a constant)

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_0^1 f(\tau)g(\tau)d\tau}; \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_0^1 f(\tau)^\alpha d\tau}, \tag{6}$$

and with the inner product

$$\langle f, g \rangle = \frac{1}{2} \int_0^1 \int_0^1 \log \frac{f(t)}{f(s)} \log \frac{g(t)}{g(s)} dt ds. \tag{7}$$

This space was designed by Egozcue et al. (2006) and Van den Boogaart et al. (2014) precisely to represent the salient features of density functions when interpreted in the light of compositional data analysis. For instance, the information conveyed by compositional data is well-known to be relative (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001), that is, the relevant information is provided by ratios between the parts (i.e., the point evaluation of the functions) rather that by the absolute value of the functions themselves. This is precisely expressed by the inner product in (7), that generalizes to the functional case the Aitchison inner product for multivariate compositional data (Pawlowsky-Glahn and Egozcue, 2001). In addition, the operations between densities defined in (6) guarantee that taking a linear combinations of densities results in a density function, unlike the usual geometry of $L^2$. Besides this, these operations have meaningful interpretation in mathematical statistics, e.g., perturbation can be interpreted as a Bayesian update of information (Egozcue et al., 2013).

Egozcue *et al.* (2006) proved that the space $B^2$ equipped with the operations in (6) and the inner product in (7) is a separable Hilbert space. This allows to generalize most methods in FDA – that are typically developed for data in $L^2$ – to the Bayes space setting. Amongst these, we here focus on Simplicial Functional Principal Component Analysis (SFPCA, Hron *et al.*, 2016), that aims to reduce the dimensionality of a dataset of density functions. Given $\widehat{f}_1, ..., \widehat{f}_M$, we aim to find the directions in $B^2$, denoted by $\zeta_1, \ldots, \zeta_{M-1}$, along which the dataset displays the maximum variability. Formally, $\zeta_1$ maximizes

$$\sum_{j=1}^{M} \langle \widehat{f}_j \ominus \overline{f},\ \zeta \rangle^2 \quad \text{subject to} \quad \|\zeta\| = 1, \tag{8}$$

where $\overline{f} = \frac{1}{M} \odot \oplus_{j=1}^{M} \widehat{f}_j$ is the sample mean, and for $j > 1$, $\zeta_j$ maximizes

$$\sum_{j=1}^{M} \langle \widehat{f}_j \ominus \overline{f},\ \zeta \rangle^2 \quad \text{subject to} \quad \|\zeta\| = 1,\ \langle \zeta, \zeta_i \rangle = 0, i < j. \tag{9}$$

Note that $\hat{f}_j \ominus \overline{f}$ is the observation in the $j$-th sample, $\hat{f}_j$, centered with respect to the sample mean $\overline{f}$, whereas $\langle \hat{f}_j \ominus \overline{f},\ \zeta \rangle$ represents the projection of the $j$-th centered observation along the direction in $B^2$ identified by $\zeta$. Hence, the objective functional in (8) and (9) is the sample variance of the projections along the generic direction $\zeta$, that has to be maximized to find the principal directions. Hron *et al.* (2016) proved that minimization of (8)-(9) can be performed by solving an equivalent FPCA problem in $L^2$, on a transformed dataset. More precisely, one can map the dataset of density curves $\hat{f}_1, \ldots,\ \hat{f}_M$ from $B^2$ in $L^2$ by using the centered log-ratio (clr) transformation, defined, for $f \in B^2$, as

$$\left[clr\left(f\right)\right](t) = \log\left(f\left(t\right)\right) \ - \int_0^1 \log\left(f\left(\tau\right)\right) d\tau, \quad t \in [0,1]\,. \tag{10}$$

From the mathematical viewpoint, transformation (10) is an isometric isomorphism between $B^2$ and $L^2$, i.e., it is a map allowing to represent elements in $B^2$ as elements of $L^2$, preserving their distances. In practice, having transformed the density curves $\hat{f}_1, \ldots, \hat{f}_M$ in elements of $L^2$, $y_1, \ldots, y_M$, one can then apply to the latter dataset the usual FPCA, obtaining (a) the FPCs $\xi_i$, $i = 1, 2, \ldots, M-1$, that are linked to the SFPCs by the relation $\xi_i = clr(\zeta_i)$, $i = 1, 2, \ldots, M-1$; and (b) the scores $z_{ji}$ of the $j$-th curve $y_j$ along the $i$-th FPC, i.e., $z_{ji} = \int_0^1 y\left(t\right) \xi_i\left(t\right) dt$. The latter scores are equivalent to the scores of the $j$-th density along the $i$-th SFPC $\zeta_i$, i.e., $z_{ji} = \langle \hat{f}_j \ominus \overline{f},\ \zeta_i \rangle$. The scores $z_{ji}$ can be treated by using the usual Euclidean geometry, as they are coordinates with respect to an orthonormal basis of $B^2$. For additional details on FPCA, we refer the reader to Ramsay and Silverman (2001).

For the purpose of reducing the dimensionality of the dataset based on SFPCA, one can employ very similar techniques as in FPCA. For instance, one can compute the variability displayed by the scores along a given principal direction $\zeta_i$ in $B^2$ through the associated eigenvalue $\rho_i$, $i = 1, \ldots, M-1$. Note that the latter is equivalently found as the eigenvalue associated with the $i$-th FPC in $L^2$, $\xi_i$. One can then retain the minimum number, $K$, of SFPCs allowing to express a given amount of the total variability (e.g., 95% or 98%), quantified as $\sum_{i=1}^{K} \rho_i / \sum_{i=1}^{\infty} \rho_i$.

### 3.3 Profile monitoring

The profile monitoring procedure based on the SFPCA is equivalent to the one based on the FPCA described by Colosimo and Pacella (2007). It requires the computation of two statistics: one is the Hotelling's $T^2$ statistic, used to detect shifts along the directions of the first $K$ SFPCs:

$$T_j^2\left(K\right) = \sum_{i=1}^{K} \frac{z_{ji}^2}{\rho_i}, \quad j = 1, 2, \ldots, \tag{11}$$

where $\rho_i$ is the $i$-th eigenvalue and $z_{ji}$ is the $j$-th score associated with the $i$-th SFPC. The second is the sum of prediction error statistic, used to detect shifts along directions orthogonal to the ones associated with the first $K$ SFPCs:

$$SPE_j\left(K\right) = \langle \hat{f}_j^* \ominus \hat{f}_j,\ \hat{f}_j^* \ominus \hat{f}_j \rangle, \quad j = 1, 2, \ldots, \tag{12}$$

where $\hat{f}_j^* = \oplus_{i=1}^{K} z_{ji} \odot \zeta_i$ is the reconstruction of the $j$-th density curve after retaining the first $K$ SFPCs. Two control charts can be designed to monitor the $T_j^2\left(K\right)$ and $SPE_j\left(K\right)$ statistics with probability limits, analogously to the FPCA-based SPC scheme (Colosimo and Pacella, 2007; 2010).

## 4 Simulation study

A simulation study is presented to demonstrate the effectiveness of the SFPCA for profile monitoring of PDFs. The framework of the study is inspired by pore-size probability distribution monitoring problems. To this aim, synthetic porous structures were generated by adapting the algorithm presented in Tschopp *et al.* (2008). The output of the algorithm at each execution run is a 2D binary image where connected components (i.e., the pores) with fixed aspect ratio equal to 1 (i.e., circular pores) but random radius are spread within a square region of fixed size without overlapping (see Tschopp *et al.*, 2008 for details). The pore sizes were controlled by defining the PDF of the pore radius, $f(r)$. A random sample of $N$ radius values, $r_1, \ldots, r_N$, was drawn from $f(r)$ and the corresponding pores were created. This synthetic structure generation process was repeated $J$ times to simulate $J$ random sections of a porous structure. The dataset used for monitoring purposes consists of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J$, where $\boldsymbol{x}_j = \left[log(A_{1j}), \ldots, log(A_{Nj})\right]^T$ and $A_{ij}$ is area of the $i$-th pore of the $j$-th section based on pixel counting.

Three different scenarios for the generation of both in-control and out-of-control pore size distributions were considered. In Scenario 1, the in-control distribution of pore radii is lognormal and the out-of-control condition consists of a shift from a unimodal to a bimodal distribution, where the additional term affects the right tail of the pore-size distribution, corresponding to a density increase of larger pores. In Scenario 2, the in-control distribution of pore radii is chi-squared and the out-of-control condition consists of a shift towards a bimodal distribution affecting the upper tail. In Scenario 3, the in-control distribution is bimodal and the out-of-control condition consists of a variance increase of one of the two components. The three scenarios were selected because they provide realistic pore structures and challenging out-of-control deviations that may be difficult to detect with traditional monitoring approaches. The details about the simulation of different scenarios are provided hereafter.

**Scenario 1** In-control (IC) distribution:

$$r_{ij} \sim log\mathcal{N}\left(1, 0.5\right), \quad i = 1, \ldots, N,\ r_{ij} \in [0,\ 50]\,, \quad j = 1, \ldots, M.$$

Out-of-control (OOC) distribution:

$$r_{ij} \sim \frac{1}{1 + w_1\left(sev\right)}\, log\mathcal{N}\left(1, 0.5\right) + \frac{w_1\left(sev\right)}{1 + w_1\left(sev\right)}\, \mathcal{N}\left(20, 5\right), \\ i = 1, \ldots, N,\ r_{ij} \in [0,\ 50]\,, \quad j = 1, \ldots, J - M,$$

where $M$ is the number of in-control sections, $sev = 1, \ldots, 4$, is the severity level index, and $J - M$ is the number of out-of-control sections at each severity level. $w_1\left(sev\right)$ is a severity-dependant weight such that: $w_1\left(sev\right) \in \{0.001,\ 0.01,\ 0.02,\ 0.03\}$.

**Scenario 2** In-control distribution:

$$r_{ij} \sim \chi^2\left(3\right), \quad i = 1, \ldots, N,\ r_{ij} \in [0\ 50]\,, \quad j = 1, \ldots, M.$$

Out-of-control distribution:

$$r_{ij} \sim \frac{1}{1 + w_1\left(sev\right)}\, \chi^2\left(3\right) + \frac{w_1\left(sev\right)}{1 + w_1\left(sev\right)}\, \mathcal{N}\left(20, 5\right), \\ i = 1, \ldots, N,\ r_{ij} \in [0\ 50]\,, \quad j = 1, \ldots, J - M,$$

where $w_1\left(sev\right) \in \{0.005,\ 0.015,\ 0.025,\ 0.035\}$.

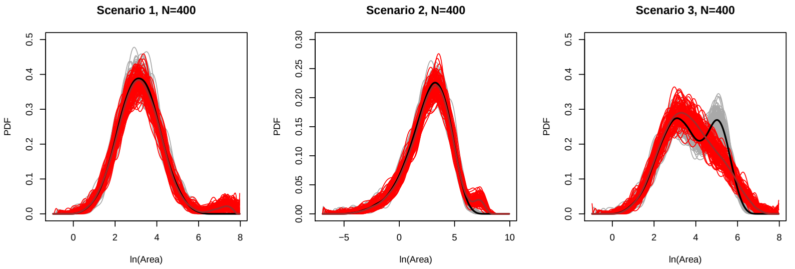

**Figure 4:** Comparison between 100 in-control density curves (grey lines) and 100 out-of-control density curves at the highest severity level (red lines) for different scenarios at $N = 400$; cross-sectional average curves are depicted with solid thick lines.

**Scenario 3** In-control distribution:

$$r_{ij} \sim 0.7\, log\mathcal{N}\,(1, 0.5) + 0.3\, log\mathcal{N}\,(2, 0.25)\,,\ i = 1, ..., N, r_{ij} \in [0, 50],\ j = 1, \ldots, M;$$

Out-of-control distribution:

$$r_{ij} \sim 0.7\, log\mathcal{N}\,(1, 0.5) + 0.3\, log\mathcal{N}\,\left(2, \sigma^2(\mathrm{sev})\right),\ i = 1, ..., N, r_{ij} \in [0, 50],\ j = 1, \ldots, J - M;$$

where $\sigma^2_{sev} \in \{0.30,\ 0.045,\ 0.40,\ 0.45\}$.

For each scenario, two cases were considered, one with $N = 400$ and one with $N = 200$, in order to investigate the effect of the number of pores on the density-based profile monitoring performances. As a way of illustration, Fig. 4 shows a superimposition of 100 in-control PDFs and 100 out-of-control PDFs in the three scenarios at $N = 400$ (the out-of-control curves corresponds to the highest severity level, $sev = 4$). Further examples of in-control and out-of-control porous structures and corresponding PDFs are shown in the Supplementary Material.

Four different control charts were applied and compared to determine if the generated porous sections were in-control or not. The first control chart is a Shewhart's control chart $\overline{X} - S$ (Montgomery, 2008) with probability limits applied to the sample mean of the $\boldsymbol{x}_j$ descriptor values. This is representative of a traditional approach for quality monitoring via pore area computation. The second is the Q-Q plot-based approach proposed by Wang and Tsung (2005) and extended by Wells *et al.*, (2013). Three variables are monitored in this case, i.e., the intercept and the slope of the fitted Q-Q plot, and its mean square error (MSE). For a fair comparison, the first two variables were monitored via a Hotelling's $T^2$ chart and the MSE via a univariate chart for individual observations. The third is a profile monitoring approach based on FPCA (Colosimo and Pacella, 2007), where the FPCA methodology is applied to the $\hat{f}_j$ density curves. This is representative of a traditional profile monitoring scheme applied without keeping into account the constrained nature of the density curves. The third is the proposed approach based on SFPCA, where the SFPCA is applied to the density curves $\hat{f}_j$, $j = 1, \ldots, J$. Regarding both the FPCA-based and SFPCA-based methods, the number of retained FPCs was selected such that a given percentage of data

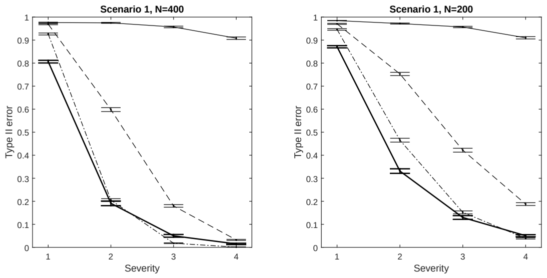

**Figure 5:** Type II error, $\beta$, and 95% confidence intervals in Scenario 1 ($N = 400$, left panel and $N = 200$, right panel); thick solid line: SPFCA-based approach; solid line: FPCA-based approach; dash-dot line: Q-Q plot approach; dashed line: Shewharts approach

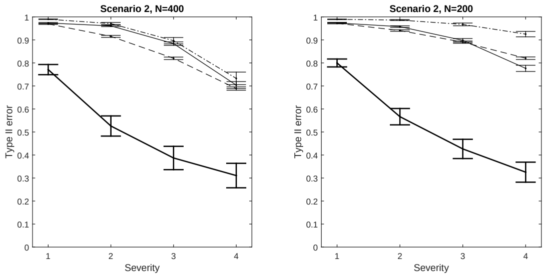

**Figure 6:** Type II error, $\beta$, and 95% confidence intervals in Scenario 2 ($N = 400$, left panel and $N = 200$, right panel); thick solid line: SPFCA-based approach; solid line: FPCA-based approach; dash-dot line: Q-Q plot approach; dashed line: Shewharts approach

variability was explained. Due to the smooth nature of the curves, few FPCs are expected to explain a large percentage of the overall variability. Because of this, a high threshold at 98% was used. For each scenario, 100 simulation runs were performed at each severity level. Each run consisted of $M = 1100$ in-control density curves and 250 out-of-control density curves. The $M$ curves were split into two datasets. A dataset of size $M_1 = 100$ was used as Phase I dataset for control chart and principal component parameter estimation. A dataset of size $M_2 = 1000$ was used to estimate the empirical control limits corresponding to a Type I error $\alpha = 0.01$.

Fig. 5, 6 and 7 show that the SFPCA-based approach outperforms both the traditional Shewhart's control chart and the FPCA-based method in all the simulated scenarios. In Scenario 1 and 2, the out-of-control condition locally affects the PDF in a portion of the domain (the upper tail) with small natural variability. The FPCA-based $T^2$ statistic is not able to detect the shift and, due to its local nature, the FPCA-based SPE statistic is poorly affected as well. The SFPCA-based approach, instead, is able to detect the shift thanks to a better characterization and reconstruction of the natural variability of the PDF curves. Moreover, it is worth to notice that, in Scenario 1, the FPCA is even outperformed by the Shewhart's control chart. From previous studies (Kim et

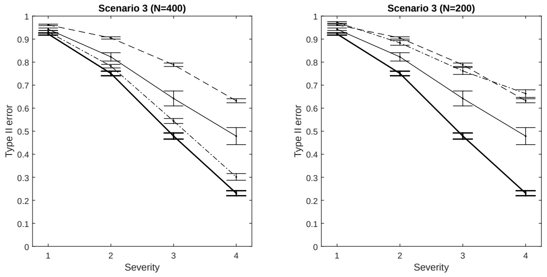

**Figure 7:** Type II error, $\beta$, and 95% confidence intervals in Scenario 3 ($N = 400$, left panel and $N = 200$, right panel); thick solid line: SPFCA-based approach; solid line: FPCA-based approach; dash-dot line: Q-Q plot approach; dashed line: Shewharts approach.

al., 2014; Woodall et al., 2004; Colosimo and Pacella, 2010), it is known that profile monitoring schemes, particularly FPCA-based ones, are more effective than traditional control charts. However, in the presence of PDF profiles, the conventional FPCA is not only unable to preserve the constrained nature of the curves themselves, but it may be also less influenced by small local shape variations than traditional methods. Because of this, the conventional FPCA is not the best candidate for the development of profile monitoring schemes applied to probability density curves, and the SFPCA should be preferred instead. In Scenario 3, the shape variation of the bimodal distribution poorly affects the sample mean and variance of the monitored descriptor, which reduces the effectiveness of the Shewhart's control chart. Again, the SFPCA-based approach is the one that better captures the occurred shape modifications and yields to the highest detection power. However, the larger is the portion of the PDF affected by the out-of-control shift, the lower is the gap between the SFPCA- and FPCA-based approach performances. Regarding the Q-Q plot-based method, it provides very good performances in Scenario 1, where the log-area descriptor exhibits an in-control normal distribution. It also outperforms the FPCA-based one in Scenario 3 when $N = 400$, despite of the non-normal distribution. Nevertheless, its performances worsen as the number of pores in the section decreases and it is outperformed by our proposed approach in Scenario 2 and 3. As a matter of fact, the Q-Q plot-based method is thought for in-control distributions that are normal or transformable to normal and its reliability decreases as the actual PDF deviates from this assumption. Thus, the SPFCA-based approach can be regarded as a generalization of profile monitoring methods to PDFs regardless of the in-control distribution. Table 1 reports the number of retained FPCs to explain about 98% of the overall variability of the density curves included into the Phase I dataset. Table 1 shows that the SFPCA requires only 2 SFPCs to explain such a percentage of variability, whereas the FPCA requires about 10 – 14 FPCs. This makes the SFPCA not only more effective than the FPCA for profile monitoring of density curves, but also more efficient in terms of dimensionality reduction.

**Table 1:** Average number of retained FPCs (sample standard deviation in brackets), at $N = 400$ and $N = 200$.

| Scenario | N. of PCs ($N = 400$), $K$ | | N. of PCs ($N = 200$), $K$ | |
|---|---|---|---|---|
| | SFPCA | FPCA | SFPCA | FPCA |
| 1 | 2.02 (0.1407) | 13.5 (0.5025) | 2 (0) | 9.98 (0.1482) |
| 2 | 2.05 (0.219) | 14.96 (0.4245) | 2.4 (0.4926) | 10.08 (0.3082) |
| 3 | 2 (0) | 14.4 (0.4924) | 2 (0) | 10.7 (0.4606) |

# 5 Real case study results

The metal foam case study introduced in Section 2 is here adopted to test the proposed approach on real data. In order to compare the SFPCA-based against previously mentioned competitor techniques, the Sample A data was used as Phase I dataset (i.e., representative of in-control patterns) whereas the Sample B data was used as Phase II (i.e., representative of out-of-control patterns). Fig. 8 shows some results achieved by applying the SFPCA to Sample A. Fig. 8 shows the Pareto plot (or scree plot) (a), the variability of the scores depending on the number of SFPCs (b), the loadings corresponding to the first 4 SFPCs estimated via SFPCA (c), their counterpart in $L^2$ (d) and additional plots to enhance interpretability (e-h). The latter displays the mean density $\overline{f}$ perturbed by plus/minus each of the first four SFPCs powered by the standard deviation along the corresponding direction, i.e., $\overline{f} \oplus (\pm\sqrt{\rho_i}) \odot \zeta_i$, $i = 1, \dots, 4$. This kind of plots shows typical behaviour of density curves associated with high/low scores along the considered SFPC. For instance, Fig 8 (e) evidences that densities with high scores along the first SFPC tend to be more concentrated on low values of pore size and vice versa. The second SFPC (Fig 8 (f)) mainly characterises the left tail of the distribution: high scores are presented by densities concentrated on very low small sized pores, low scores by densities with lighter left tails. The third SFPC, Fig 8 (g), is associated with the modality of the densities (unimodal for high scores, slightly bimodal for low scores), whereas no clear interpretation is obtained for the fourth SFPC.

By setting a threshold on the percentage of explained variance at 98%, the FPCA requires retaining $K = 4$ FPCs, whereas the SFPCA requires retaining $K = 3$ SFPCs. Fig. 9 shows a comparison of the smoothed PDFs for Sample A (grey curves in background) and Sample B (red curves in foreground). For both samples, the density curves exhibit a quite large shape variability. However, in Sample B there is an inflation of the left tail corresponding to a higher fraction of small pores than in Sample A. Moreover, in Sample B, some curves, especially the ones associated with the 23-rd and 25-th sections, exhibit the largest peak at quite low values of the descriptor, due to a predominance of small and medium size pores. These differences are caused by the different polishing and section preparation treatments applied on the two samples.

Table 2 compares the Type II error for all the competing techniques considered in

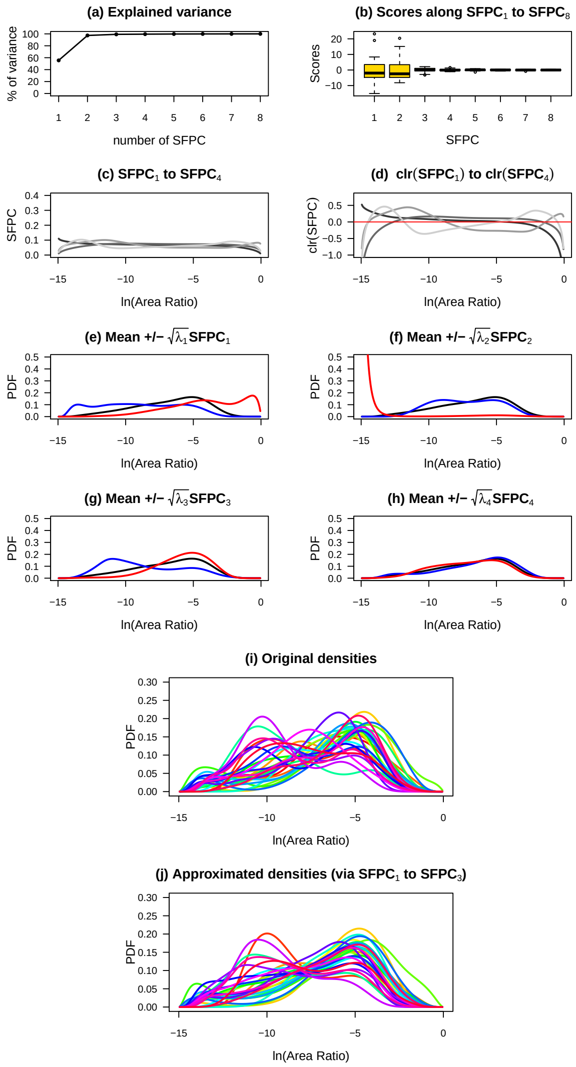

**Figure 8:** Results of the SFPCA applied to Sample A: (a) Pareto plot for the SFPCA; (b) score variability as a function of the number of PCs; (c) loading in $B^2$ associated with the first four SFPCs (i.e., $\zeta_1, \dots, \zeta_4$); (d) loading in $L^2$ associated with the first four SFPCs (i.e., $\xi_1, \dots, \xi_4$); (e) to (h) plot of $\overline{f} \oplus (\pm\sqrt{\rho_i}) \odot \zeta_i$, $i = 1, \dots, 4$.

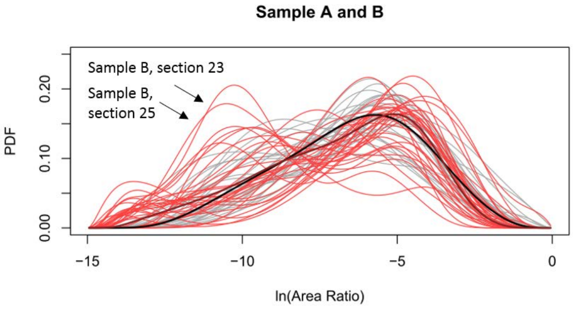

**Figure 9:** Superimposition of Sample A density curves (grey lines in background) and Sample B density curves (red lines in foreground); solid thick lines corresponds to cross-sectional average curves.

**Table 2:** Type II errors for the metal foam case study

| Method | Type II error |
|---|---|
| SFPCA | 0.1 |
| FPCA | 0.1333 |
| Q-Q plot | 0.5333 |
| Shewhart | 0.4 |

Section 3, with $\alpha = 0.01$. Table 2 shows that the SFPCA-based and FPCA-based approaches perform better than the Q-Q plot-based approach and the traditional Shewart's control chart. Indeed, the departure from normality shown in Fig. 9 makes these two latter competitors poorly effective in detecting the distributional change between the two samples. Similarly to the simulation study, the SFPCA-based approach should be preferred to the FPCA-based one. Fig. 10 and Fig. 11 compare the SFPCA-based and FPCA-based control charts. In the SFPCA-based approach, both the $T^2$ and $SPE$ control charts clearly signal a sustained shift in terms of PDF shape change between Sample A and Sample B. Only three sections of Sample B (section 4, 9 and 30) are classified as in-control observations, whereas all the remaining sections are signalled by at least one of the two control charts. Fig. 10 also shows that the two largest peaks of both the control statistics in Phase II correspond to the 23-rd and 25-th section of Sample B, whose density curves were indicated in Fig. 9. In the FPCA-based approach, instead, the $T^2$ signals only five data points, without showing any evident sustained shift. The difference between the two samples is almost entirely captured by the $SPE$ control chart alone, which means that the major differences occur along principal directions that explain a small portion of the Phase I data variability. In addition, the FPCA-based method is less effective in signalling the 23-rd and 25-th section of Sample B, where

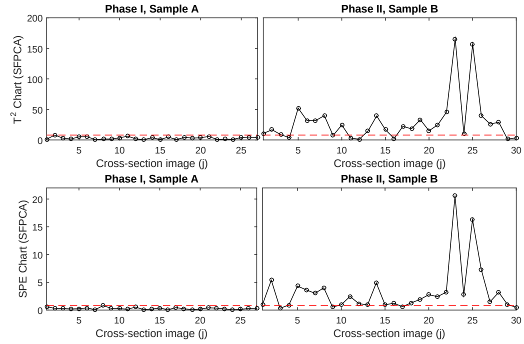

**Figure 10:** SFPCA-based control charts for the metal foam case study.

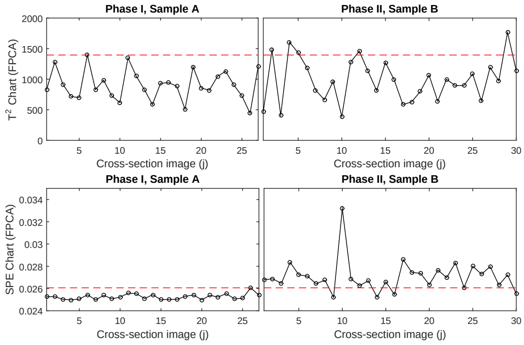

**Figure 11:** FPCA-based control charts for the metal foam case study.

the visual analysis of curves revealed the larger departure from the Phase I pattern. This is caused by a different definition of the loadings affected by a non-appropriate curve projection. Because of this, the SFPCA is a more effective approach to capture the differences of PDF shapes.

## 6 Conclusion

The development of novel SPC tools must face the need to synthesize the information enclosed by the original data and, by transforming it to a format that is easier to handle and more convenient from a computational viewpoint. The underlying idea of our study consists of transforming a high-dimensional data sample (e.g., an image) into a profile by analysing the PDF of a descriptor of interest, whose values are computed for a large number of random features. Although probability density curves represent a special case of functional data, most profile monitoring methods are not appropriate for constrained functions. Therefore, we proposed a profile monitoring approach that relies on a simplicial variant of the FPCA, which allows preserving the constrained nature of the data thanks to the Bayes space geometry. The simulation study showed that the SFPCA-based approach outperformed the FPCA-based profile monitoring in all the simulated scenarios. The results highlighted that, in the presence of probability density curves, the well-known FPCA is not only unable to preserve the constrained nature of the curves, but it also yields a reduced out-of-control detection power, especially in the presence of local shifts in the tails of the distribution. Because of this, the SFPCA should be preferred to FPCA for the development of profile monitoring schemes applied to probability density curves. The comparison study also showed that the SFPCA requires a smaller number of principal components to explain the same amount of variability explained by the FPCA. This makes the SFPCA not only more effective than the FPCA for profile monitoring of PDFs, but also more efficient in terms of dimensionality reduction. Moreover, the proposed approach can be considered a generalization of the Q-Q plot-based method to any in-control distribution.

The real case study in metal foam production confirmed the higher effectiveness of the SFPCA-based approach against the competitors to capture the shape modifications of PDFs. Generally speaking, the use of PDFs in process monitoring problems represents a compromise between data synthesis and goodness of process signature characterization, but it yields a loss of information about possible spatial or temporal dependencies. In principle, the proposed approach represents a suitable choice when spatial/temporal auto-correlation modelling is not necessary or not feasible. But it can also be used in parallel to other monitoring methods when spatial/temporal models are available, to couple a distributional data analysis to other modelling paradigms.

Future studies will be aimed at extending the proposed methodologies to problems where more than one single statistical descriptor is of interest. In those cases, rather than monitoring the marginal PDFs, the signature of the process can be related to the multidimensional shape of the joint PDF of monitored descriptors. The SFPCA method can be extended to the multivariate case, but additional research efforts are needed to formalize the method and to determine its performances.

# References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. In: Monographs on Statistics and Applied Probability, Chapman and Hall Ltd., London, UK, p. 416. (Reprinted 2003 with additional material by The Blackburn Press).

Banhart, J. (2001). Manufacture, characterisation and application of cellular metals and metal foams. *Progress in materials science*,46(6), 559–632.

Bivand, R., Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, 63(19), 1–36.

Campoli, G., Borleffs, M. S., Yavari, S. A., Wauthle, R., Weinans, H., & Zadpoor, A. A. (2013). Mechanical properties of open-cell metallic biomaterials manufactured using additive manufacturing. *Materials & Design*, 49, 957–965.

Colosimo, B.M., Pacella, M. (2007). On the Use of Principal Component Analysis to Identify Systematic Patterns in Roundness Profiles. *Quality and Reliability Engineering International*, 23(6), 707–725.

Colosimo, B.M., Pacella, M. (2010). A Comparison Study of Control Charts for Statistical Monitoring of Functional Data. *International Journal of Production Research*, 48(6), 1575–1601.

Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55(1), 401–420.

Egozcue, J.J., Díaz-Barrero, J.L., Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica (English Series)* 22(4), 1175–1182.

Egozcue, J.J., Pawlowsky-Glahn, V., (2006). *Simplicial geometry for compositional data*. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (Eds.), Compositional Data Analysis in the Geosciences: From Theory to Practice. In: Special Publications, vol. 264. Geological Society, London, pp. 145–160.

Egozcue, J.J., Pawlowsky-Glahn, V., Tolosana-Delgado, R., Ortego, M.I., van den Boogaart, K.G. (2013). *Bayes spaces: use of improper distributions and exponential families*. *Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat.*, 107, 475–486.

Gibson, I., Rosen, D. W., Stucker, B. (2010). *Additive manufacturing technologies*. New York: Springer.

Hron, K., Menafoglio, A., Templ, M., Hruzova, K., Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 94, 330–350.

Kim, T. B., Yue, S., Zhang, Z., Jones, E., Jones, J. R., & Lee, P. D. (2014). Additive manufactured porous titanium structures: Through-process quantification of pore and strut networks. *Journal of Materials Processing Technology*, 214(11), 2706–2715.

Megahed, F. M., Woodall, W. H., & Camelio, J. A. (2011). A review and perspective on control charting with image data. *Journal of Quality Technology*, 43(2), 83–98.

Menafoglio, A., Guadagnini, A., Secchi, P. (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*. 28(7), 1835–1851.

Menafoglio, A., Secchi, P., Guadagnini, A., (2016a). A Class-Kriging predictor for Functional Compositions with Application to Particle-Size Curves in Heterogeneous Aquifers. *Mathematical Geosciences*, 48(4), 463–485.

Menafoglio, A., Guadagnini, A., Secchi, P., (2016b). Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach. *Water Resources Research*, 52, 5708–5726.

Montgomery D. C. (2008). *Introduction to Statistical Quality Control*. John Wiley & Sons, 6th Ed.

Noorossana R., Saghaei A., Amiri A. (2012). *Statistical Analysis of Profile Monitoring*. John Wiley & Sons.

Pawlowsky-Glahn, V., Buccianti, A. (Eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd., Chichester, UK, p. 378.

Pawlowsky-Glahn, V., Egozcue, J.J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*. 15(5), 384–398.

Qiu, P. (2005). *Image processing and jump regression analysis*.(Vol. 599). John Wiley & Sons.

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

Sharratt, B. M. (2015). Non-Destructive Techniques and Technologies for Qualification of Additive Manufactured Parts and Processes. A literature Review. Contract Report DRDC-RDDC-2015-C035, Victoria, BC.

Strano, M. (2011). A New FEM Approach for Simulation of Metal Foam Filled Tubes. *Journal of Manufacturing Science and Engineering*, 133, 061003, 1–11.

Tschopp, M. A., Wilks, G. B., & Spowart, J. E. (2008). Multi-scale characterization of orthotropic microstructures. *Modelling and Simulation in Materials Science and Engineering*, 16(6), 065009.

Van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *Statist. Oper. Res. Trans*, 34(2), 201–222.

Van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Aust. N. Z. J. Stat.*, 56(2), 171–194.

Villa, A., Strano, M., Mussi, V., (2011). Optimization of Design and Manufacturing Process of Metal Foam Filled AntiIntrusion Bars. in: *AIP Conference Proceedings* 1353. 1656–1661.

Wang, K., & Tsung, F. (2005). Using profile monitoring techniques for a data-rich environment with huge sample size. *Quality and Reliability Engineering International* 21(7). 677–688.

Wells, L. J., Megahed, F. M., Niziolek, C. B., Camelio, J. A., & Woodall, W. H. (2013). Statistical process monitoring approach for high-density point clouds. *Journal of Intelligent Manufacturing* 24(6). 1267–1279.

Yan, H., Paynabar, K., & Shi, J. (2015). Image-based process monitoring using low-rank tensor decomposition. *IEEE Transactions on Automation Science and Engineering*, 12(1), 216–227.

Woodall, W.H., Spitzner, D.J., Montgomery, D.C., Gupta, S. (2004). Using Control Charts to Monitor Process and Product Quality Profiles. *Journal of Quality Technology* 36(3). 309–320.

Zhuravleva, K., Bönisch, M., Prashanth, K. G., Hempel, U., Helth, A., Gemming, T., Calin, M., Scudino, S., Schultz, L., Gebert, A. (2013). Production of porous $\beta$-Type Ti–40Nb alloy for biomedical applications: comparison of selective laser melting and hot pressing. *Materials* 6(12). 5700–5712.

# MOX Technical Reports, last issues

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

**08/2018** Bonaventura, L.; Casella, F.; Delpopolo, L.; Ranade, A.;
*A self adjusting multirate algorithm based on the TR-BDF2 method*

**06/2018** Antonietti, P.F.; Mazzieri, I.
*High-order Discontinuous Galerkin methods for the elastodynamics equation on polygonal and polyhedral meshes*

**07/2018** Ieva, F.; Bitonti, D.
*Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data*

**05/2018** Pagani, S.; Manzoni, A.; Quarteroni, A.
*Numerical approximation of parametrized problems in cardiac electrophysiology by a local reduced basis method*

**03/2018** Antonietti, P. F.; Houston, P.; Pennesi, G.
*Fast numerical integration on polytopic meshes with applications to discontinuous Galerkin finite element methods*

**04/2018** Ekin, T.; Ieva, F.; Ruggeri, F.; Soyer, R.
*Statistical Medical Fraud Assessment: Exposition to an Emerging Field*

**02/2018** Canuto, C.; Nochetto, R. H.; Stevenson, R.; Verani, M.
*A saturation property for the spectral-Galerkin approximation of a Dirichlet problem in a square*

**01/2018** Berrone, S.; Bonito, A.; Stevenson, R.; Verani, M.
*An optimal adaptive Fictitious Domain Method*

**67/2017** Esterhazy, S.; Schneider, F.; Mazzieri, I; Bokelmann, G.
*Insights into the modeling of seismic waves for the detection of underground cavities*

**68/2017** Paolucci, R.; Infantino, M.; Mazzieri, I.; Özcebe, A.G.; Smerzini, C.; Stupazzini, M.
*3D physics-based numerical simulations: advantages and current limitations of a new frontier to earthquake ground motion prediction. The Istanbul case study.*