Functional PCA of measures for investigating the influence of bioturbation on sediment structure

Claude Manté, Georges Stora

Functional PCA of measures for investigating the influence of bioturbation on sediment structure

Claude Mante, *Aix-Marseille Universite*, claude.mante@univ-amu.fr
Georges Stora, *Aix-Marseille Universite*, georges.stora@univ-amu.fr

Abstract. After describing the main characteristics of grain-size curves, we recall previous results about Principal Components Analysis of absolutely continuous measures, in connection with grain-size curves analysis. This method simultaneously takes into account a chosen reference probability (r.p.) μ (associated with a Radon-Nikodym derivation operator), and the imposed sampling mesh T_p. The point is that it amounts to usual PCA in some metrics $M^{-}(T_p;\mu)$; consequently, analyzing a set of grain-size curves in reference to different r.p.s amounts to carry out PCA with different metrics. Three complementary r.p.s were chosen to analyze a set of 552 grain-size curves issued from an experiment designed for investigating the influence of a *Polychaetes, Nereis diversicolor* , on the sediment structure. The obtained results show that this worm is actually able to alter the sediment. Furthermore, it is shown that its influence depends on its density in the sedimentary column, but not on its position.

Keywords. Functional Data Analysis, Radon-Nikodym derivative, Metrics in PCA, Homogeneity test, ANOVA, Sedimentology

1 Introduction

The interpretation of sedimentological field data is frequently based on the analysis of cumulative curves produced by sieving sediments, or by other devices devoted to particle-size analysis. Such grain-size curves possess four main characteristics:

1. the curve F_ν associated with each sediment ν is supported by some fixed common interval $[a,b] \subset [0,+\infty[$

2. F_ν is similar to a distribution function: $F_\nu(x) = \nu([a,x[)$ is the relative weight (or number, *etc.*) of particles of ν smaller than x

3. the grain-size x is associated with a scale: classically, x is expressed according either metric units (MU) or the so-called ϕ-units ($\phi := \ln(MU)$) [2]

4. each curve is sampled according to some common p-mesh T_p depending on the device used.

Now, why should we care about "bioturbation"? This term refers to the movement of particles generated by the displacement of organisms in the sedimentary column. Recently, using grain-size analysis, Ciutat *et al* [3] have shown that tubicifid worms are able to alter the structure of the sedimentary column. We will study the ability of *Nereis diversicolor* (another worm, living in the first 20 cm of the sediment, in brackish water ecosystems) to do a similar job. This *Polychaete* is also aggregative (one can generally observe densities > 800 organisms per m^2), and builds many galleries down to 20 cm in the sediment.

2 What is special about functional PCA (FPCA) of measures?

Notice first that a probability is not a function. It is generally represented by its distribution function (depending on a fixed scale), or by its Radon-Nikodym density, relative to some reference probability (r.p.). In this section, we recall previous works about FPCA of grain-size distributions (or measures), taking into account the influence of the chosen r.p. on the similarity between grain-size curves, in connection with possible scale changes.

A functional setting for PCA of grain-size curves [13, 14]

Let μ be a r.p. equivalent to Lebesgue's mesure λ on $[a,b]$, ν a sediment, $F_\nu(x) := \nu([a,x[)$ the associated curve, and $f_\nu := \frac{d\nu}{d\mu} \in L^2_\mu$. It has been proven [13] that:

1. the integral operator $\widetilde{\Im} : f_\nu \mapsto F_\nu \in \mathcal{L}\left(L^2_\mu, L^2\right)$ has a bounded inverse iff $F_\nu \in H_\mu$, the reproducing kernel Hilbert space (r.k.H.s.) of kernel $K^\mu(x,y) := \mu([a, \inf(x,y)])$

2. the restricted operator $\Im \in L\left(L^2_\mu, H_\mu\right)$ is unitary, and thus $\|F_{\nu_1} - F_{\nu_2}\|^2_{H_\mu} = \int \left(\frac{d\nu_1}{d\mu} - \frac{d\nu_2}{d\mu}\right)^2 d\mu$

3. to any p-mesh $T_p \subset]a,b[$ is associated a unique r.k.H.s. $H^p_\mu \subsetneq H_\mu$.

Remark: *thanks to (2), we don't need to compute derivatives: the metrics does the job.*
The role of scale changes is clarified by the following isometry theorem.

Theorem 2.1. *Let η and ν be in H_μ, and S be a scale change (homeomorphism). Then we have:*

$$\|\eta - \nu\|_{H_\mu} = \|S_*\nu - S_*\eta\|_{H_{S_*\mu}}$$

where $S_*\mu$ denotes the probability induced by S.

As a consequence, once a reference probability space (r.p.s.) $\{[a,b], \mathcal{B}([a,b]), \mu\}$ has been fixed ($\mathcal{B}$ denotes the borelian σ-field), the distance between two measures is independent of the scale used, if the same transformation is applied to the r.p. In other words, if S is increasing

$$\{[a,b], \mathcal{B}([a,b]), \mu\} \approx \{[S(a),S(b)], \mathcal{B}([S(a),S(b)]), S_*\mu\}$$

and both these spaces belong to a common equivalence class of r.p.s .

Corollary 2.2. *(Standardization) If the d.f. F_μ associated with μ is strictly increasing, we have:*

$$\|\eta - \nu\|_{H_\mu} = \|F_{\mu*}\nu - F_{\mu*}\eta\|_{L^2[0,1]}.$$

The symbolical case

In [14], we only considered symbolical (as opposed to empirical) reference probabilities, in connection with scale changes. In such cases, the reference probability is symbolically expressed - for instance, its density relative to Lebesgue's measure is a given function $f(x)$.
We defined in that paper three different r.k.H.s. designed for FPCA of grain-size curves:

1. H_{MU}, corresponding to the metric units scale, associated with the r.p.s.

$$\{[a,b], \mathcal{B}([a,b]), \mathcal{U}\} \approx \{[\ln(a), \ln(b)], \mathcal{B}([\ln(a), \ln(b)]), \ln_* \mathcal{U}\}$$

   where $\mathcal{U}$ denotes the uniform probability

2. H_ϕ, naturally associated with ϕ-units, and with the r.p.s.

$$\{[\ln(a), \ln(b)], \mathcal{B}([\ln(a), \ln(b)]), \mathcal{U}\}$$

3. $H_{\tau_{cr}}$, associated with sediment transport theory (τ_{cr} is a critical shear stress function, closely linked with erosion).

Remarks: *working in H_{MU} amounts using an exponential r.p. with ϕ-units, because $\frac{d\ln_* \mathcal{U}}{d\lambda} = \frac{\exp(\bullet)}{(b-a)}$. Reciprocally, since $\frac{d\exp_* \mathcal{U}}{d\lambda} = \frac{1}{(\bullet)(\ln(b)-\ln(a))}$, working in H_ϕ amounts using as a reference the truncated Pareto of parameters $(1, a, b)$ in the system MU; according to Devoto and Martinez [5], it is actually a relevant distribution for ground rocks distributions.*

In our case, FPCA in H_μ does not only depends on μ, but also on the common mesh T_p. More precisely, FPCA in H_μ^p amounts to usual PCA in some metrics $M^-(T_p; \mu)$ [14], and it was shown [13] that this metrics is particularly simple and well-conditioned (μ-optimal) when T_p consists of fractiles of μ.

Definition 2.3. *$T_p = \{t_1, \cdots, t_p\}$ is μ-optimal if:*

$$\forall\, 1 \le k \le p,\ F_\mu(t_k) = \frac{k}{p}.$$

Thus, FPCA in H_μ^p merely amounts to the spectral decomposition of the matrix $V_p \circ M^-(T_p; \mu)$, where V_p is the usual covariance matrix computed from the grain-size curves.
The situation would be different if there were no common mesh, or if we faced a family $\mathcal{F} \subset H_\mu$ of distributions of individual measurements, *i.e.* of empirical distribution functions (e.d.f.s). In this cases, the sample $\mathcal{F} := \{F_{\nu_i} : 1 \le i \le N\}$ consists of e.d.f.s, and $F_{\bar{\nu}}$ will denote their average. FPCA of $\mathcal{F}$ takes place in the whole space H_μ, and the empirical covariance operator is $\mathcal{V}/N$, where the operator $\mathcal{V}$ is given by [9]:

$$\mathcal{V}\xi = \sum_{i=1}^{N} \langle F_{\nu_i} - F_{\bar{\nu}}, \xi \rangle_{H_\mu} (F_{\nu_i} - F_{\bar{\nu}}).$$

Because the reproducing kernel $K^\mu(x,y)$ is continuous, H_μ is separable ([17], p.126) and possesses countable Hilbert bases. Consequently, $\mathcal{V}$ can be identified with a semi-infinite matrix. In order that its eigendecomposition is computationally tractable, it is classical to restrict oneself to a subspace of rank $p \leq N$, generated for instance by vectors of some orthonormal basis (to build). A more elegant solution was proposed in [9]: it consisted in calculating the $N \times N$ matrix $\mathcal{M}$ of scalar products between centered densities, of general term $\mathcal{M}_{ij} := \left\langle F_{\nu_i} - F_{\bar{\nu}}, F_{\nu_j} - F_{\bar{\nu}} \right\rangle_{H_\mu}$. Thanks to the r.k.H.s. properties, these integrals can be approximated from the interpolated e.d.f.s (see the appendix A.1 of [13]). Afterwards, the best Euclidean representation of $\mathcal{F}$ can be obtained from the eigendecomposition of the matrix of general term $\mathcal{M}_{ij} - \mathcal{M}_{\bullet j} - \mathcal{M}_{i\bullet} + \mathcal{M}_{\bullet\bullet}$ ([12], see also [9] p. 522); this analysis is formally equivalent to the spectral decomposition of $\mathcal{V}$. Thus, if N is not too big, FPCA of $\mathcal{F}$ could be carried on by using this alternative method. Other connections between the method proposed in [9] and ours were reported in [14].

The mesh used

It is noteworthy that here, the sizes measured were in geometrical progression. This kind of mesh is universally adopted by geologists, both for practical and theoretical reasons. From the practical side [2], "the grade scale must be designed to accommodate comparatively large class (size) intervals from the coarsest detrital materials and extremely small intervals for the smallest particles. A geometrical scale is suitable for this purpose." From the theoretical side, Kolmogorov proved in 1940 [10] that under suitable conditions, the frequency distribution of the size (**any** size criterion: length, volume, *etc.*) of particles under grinding tends to be lognormal. One can easily determine μ-optimal meshes in the symbolical case, but notice that the actual mesh T_p imposed by experimental condition has no reason for being optimal. This important point is tackled in the next section.

The empirical case

Suppose now that the r.p. corresponds to a given "source" sediment ν_0, and that the data consists in a sample $\{\nu_1, \cdots, \nu_N\}$ of sediments stemming from different alterations of ν_0. It is quite natural to consider the densities $\left\{\frac{d\nu_1}{d\nu_0}, \cdots, \frac{d\nu_N}{d\nu_0}\right\}$ for describing in H_{ν_0} the sedimentary evolution; this is indeed the core of the McLaren & Bowles theory of Sediment Transport [18].

In this paper, ν_0 will be a typical control sediment named Ta_3, some kind of initial unperturbed state. But is T_p well-suited for FPCA in H_{ν_0}? When it is the case (*i.e.* T_p is ν_0-optimal), FPCA in $H^p_{\nu_0}$ amounts to classical PCA of the associated histograms. But generally, T_p has not been designed for that purpose. Consequently, in the empirical case, we have to face the "fractiles problem": given a r.p. ν_0 and an imposed mesh T_p, find a ν_0-optimal (or quasi-optimal) sub-mesh $T_K \subset T_p$ such that: $\forall\, 1 \leq i \leq K$, $\widehat{F_{\nu_0}}\left(T^i_K\right) \approx \frac{i}{K}$, where $\widehat{F_{\nu_0}}$ is an approximation of F_{ν_0}. In [15], we proposed a solution to this problem, based on the Bernstein operator. This operator has nice shape-preserving properties but unfortunately, its convergence is sluggish [4]. Consequently, we superseded it by a more rapidly convergent sequence of approximants, proposed by Sevy [21] and based on iterates of this operator.

The reader can see on Figure 1 how a quasi-optimal sub-mesh, well-suited for the particular control sediment $\nu_0 = Ta_3$, could be extracted from T_p. Notice that the approximation $\widehat{F_{\nu_0}}$ based on iterated operators is much better than the classical Bernstein approximation.

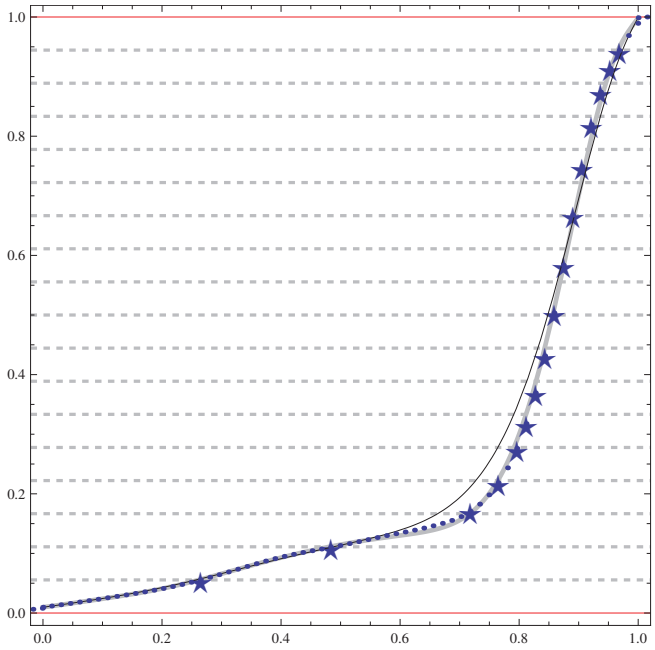

Figure 1: The approximate d.f.s associated with some reference sediment $\nu_0 = Ta_3$ (dotted curve). Thin black curve: usual Bernstein approximation; in gray: $\widehat{F_{\nu_0}}$. Stars correspond to the obtained quasi-optimal sub-mesh.

3 An experiment with Nereis diversicolor

In order to investigate the impact of *Nereis diversicolor* (*N.d.*) on sedimentary structures, an experiment was carried out. Some quantity of natural sediment was sterilized (frozen at -20°C), mixed and split up into 29 tanks of 14 cm depth. Five tanks were kept as control ones (*i.e.* without any worm), while the other ones were colonized by a fixed number $norg \in \{1, 2, 4, 8\}$ of living *N.d.*; three tanks were associated with each value of $norg$ (replicates). At the end of the experiment, each tank was cut into 20 slices, and one (control tanks) or two (colonized tanks) samples were took into each slice. In the case of colonized tanks, in each slice, a sample was taken inside a gallery (IG sediments) and another one outside galleries ($NCOG$ sediments. Notice that this was done visually; thus, one cannot be absolutely sure that $NCOG$ sediments were not altered by *N.d.*; that is why they were named $NCOG$ (Not Control OG). On the contrary, control sediments (COG) are necessarily outside galleries.

This experiment resulted in a set of 552 grain-size curves (the 28 missing sediments could not be sampled); these curves were sampled according to a mesh of size $p = 92$. Since distances between sediments depend on the chosen reference probability, we performed three complementary analyses, and kept in each case the three first principal components (always corresponding to more than 90% of total variance):

1. PCA in H_ϕ (symbolical, focuses on coarse particles; 94% of total variance)

2. PCA in H_{MU} (symbolical and "equitable", because of the uniform weight in the MU system; 91% of total variance)

3. PCA in H_{Ta3} (empirical, focuses on the control unperturbed sediment $Ta3$; 92% of total variance).

Notice that in the case of PCA in H_{Ta3}, all the distributions, originally sampled according a 92-knots mesh, were sub-sampled according to the 17-knots sub-mesh adapted for Ta_3, the reference sediment (see Figure 1).

Now, does *N.d.* causes noticeable alterations to sediments grain-size structure? We will break this vague question into three simpler ones:

1. can differences between Control (COG) and Not Control Outside Gallery ($NCOG$) sediments be statistically detected?

2. if it is not the case, are Inside Gallery (IG) and Outside Gallery ($OG := COG \cup NCOG$) sediments statistically different?

3. if $IG \neq OG$, is there an influence on IG sediments of:

   - depth (20 *slices*)
   - the number of organisms, $norg \in \{1, 2, 4, 8\}$?

Homogeneity of the OG sediments group in factor space

We first suppose that the three first components of the control group obey a Gaussian distribution:

$$\left(X_1^{COG}, X_2^{COG}, X_3^{COG}\right)' \sim N\left(\mu_{COG}, \Sigma_{COG}\right).$$

It is possible to test whether each NCOG sediment $X := (X_1, X_2, X_3)'$ belongs to the COG group by using the distribution of the associated Mahalanobis distance. The test statistic is [11]:

$$\left(X - \hat{\mu}_{COG}\right)' \hat{\Sigma}_{COG}^{-1} \left(X - \hat{\mu}_{COG}\right) \sim \frac{3\left(n^2 - 1\right)}{n\left(n - 3\right)} F_{3,n-3}. \tag{1}$$

The obtained results are summed up in Table 1; we can see that for all the analyses, more than 95% of the $NCOG$ sediments fell into the 95% Gaussian confidence region (associated with formula (1)) of the control group. Thus one can reasonably infer that $NCOG \approx COG$. On the contrary, since about 50% of the IG sediments fell in these regions, we conclude that $IG \neq COG$.

	In $\mathcal{E}(MU)$	Out $\mathcal{E}(MU)$	In $\mathcal{E}(\phi)$	Out $\mathcal{E}(\phi)$	In $\mathcal{E}(Ta3)$	Out $\mathcal{E}(Ta3)$
NCOG	**0.955**	0.044	**0.959**	0.04	**0.988**	0.012
IG	0.422	**0.577**	0.495	**0.505**	0.51	**0.49**

Table 1: Comparison of the groups $NCOG$ and IG with the control group COG, based on 95% Gaussian confidence regions of the control data in factor spaces.

Are OG and IC sediments identical?

Let us now compare the IG sediments with the whole OG group. We can see on Table 2 that in every analysis, about 60% of the IG sediments fall outside the 95% OG Mahalanobis confidence regions in R^3.

This is illustrated from a slightly different viewpoint on Figure 2, in the special case of PCA in H_{MU}. To obtain this figure, we determined the 95% confidence ellipsoid of the OG group in R^3 (i.e. the ellipsoid holding 95% of the OG points). Its parameters are (c, ρ, Ω), where c is the center, ρ is the vector of moments of inertia and $\Omega = (\varpi_1, \varpi_2, \varpi_3)$ is the matrix of normed principal direction. On Figure 2, all the sediments were projected on the plane generated by the principal directions of the ellipsoid, ϖ_1 and ϖ_2. It is noteworthy that for all the analyses (PCA in H_ϕ, H_{MU} and H_{Ta3}), about 60% of the IG sediments (the "altered sediments", clearly different from the OG ones) fell outside the corresponding confidence ellipsoids again.

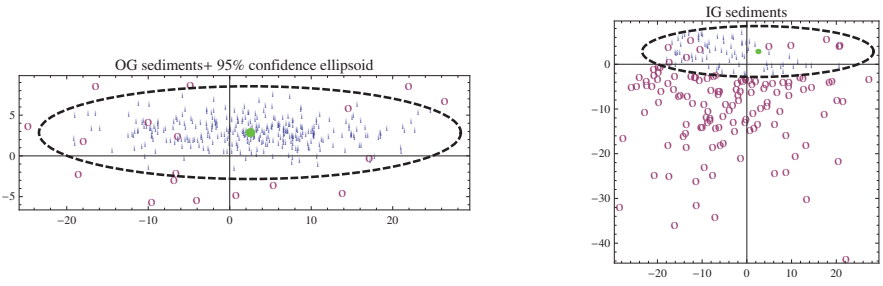

Figure 2: Representation of OG (left panel) and IG (right panel) sediments, projected on the principal confidence ellipse. Because of possible perspective errors, we labeled by a i points inside the ellipsoid, and by a o points outside.

	In $\mathcal{E}(MU)$	Out $\mathcal{E}(MU)$	In $\mathcal{E}(\phi)$	Out $\mathcal{E}(\phi)$	In $\mathcal{E}(Ta3)$	Out $\mathcal{E}(Ta3)$
OG	**0.948**	0.052	**0.948**	0.052	**0.948**	0.052
IG	0.383	**0.616**	0.393	**0.607**	0.432	**0.568**

Table 2: Comparison of the groups OG and IG, based on 95% Gaussian confidence regions for the group OG in factor spaces.

Since these analyses focus on complementary aspects of grain-size distributions, it is natural to consider the three sets of altered sediments: $AsMU$, $As\phi$ and $AsTa3$. It is noteworthy that $AsTa3 \subsetneq As\phi \cup AsMU$, and that $As\phi \cup AsMU$ corresponded to about 65% of the IG sediments, while about 83% of the altered sediments were common to these three sets.

In conclusion, one will agree that $IG \neq OG$, and that *N.d.* actually altered important features of the sedimentary structure.

Testing slice and norg effects

For that purpose, we merely used ANOVA, separately for each component. We found that generally, the factor *slice* had no significant influence on the three first components of all the analyses. On the contrary, as the reader can see on Tables 3, 4 and 5, the factor *norg* always significantly contributed to at least two main components of each FPCA.
We also investigated the pairwise equality of means by using classical tests: Tukey-Kramer, Duncan and Newman-Keuls. Since they gave consistent issues, we only show those from the former, which is especially recommended [22].

	variance (%)	F ratio	p-value	distinguishable pairs (Tukey-Kramer)
PC1	68	34.13	$< 10^{-9}$	{1,2},{1,4},{1,8},{2,8},{4,8}
PC2	18	5.87	0.00073	{1,8},{4,8}
PC3	8	13.8	$< 10^{-6}$	{1,2},{1,4},{1,8}

Table 3: Main results of the ANOVA in H_ϕ.

In all cases, multiple comparisons didn't evidenced differences between *norg*=2 and *norg*=4; consequently, these classes were merged in a unique class denoted $2 \cup 4$. One can deduce from Table 3 that the best plane of PCA in H_ϕ for illustrating the influence of *norg* on the curves structure is generated by the first and the third components. The sediments are projected on this plane on Figure 3.

	variance (%)	F ratio	p-value	distinguishable pairs (Tukey-Kramer)
PC1	52	1.65	0.18	None
PC2	22.5	26.3	$< 10^{-9}$	{1,2},{1,4},{1,8},{2,8},{4,8}
PC3	17	12.76	$< 10^{-6}$	{1,2},{1,4},{1,8},{2,8},{4,8}

Table 4: Main results of the ANOVA in H_{MU}.

	variance (%)	F ratio	p-value	distinguishable pairs (Tukey-Kramer)
PC1	50	14.47	$< 10^{-6}$	{1,2},{1,4},{1,8},{2,8}
PC2	27.4	21.45	$< 10^{-9}$	{1,2},{1,4},{1,8},{2,8},{4,8}
PC3	15.6	0.65	0.582	None

Table 5: Main results of the ANOVA in H_{Ta3}.

In conclusion, the number of *N.d.*s present in the sedimentary column had an influence on the sediment structure, irrespective of the depth. Surprisingly, the alteration of the sediment decreased with $norg$, as one can see on Figure 3. A single worm (small disks) was more perturbing than 2 or 4 worms (medium size disks), while 8 worms (big disks) seem quite inefficient. This could be explained by mixing: each *N.d.* probably destroys the job of its fellow creatures...

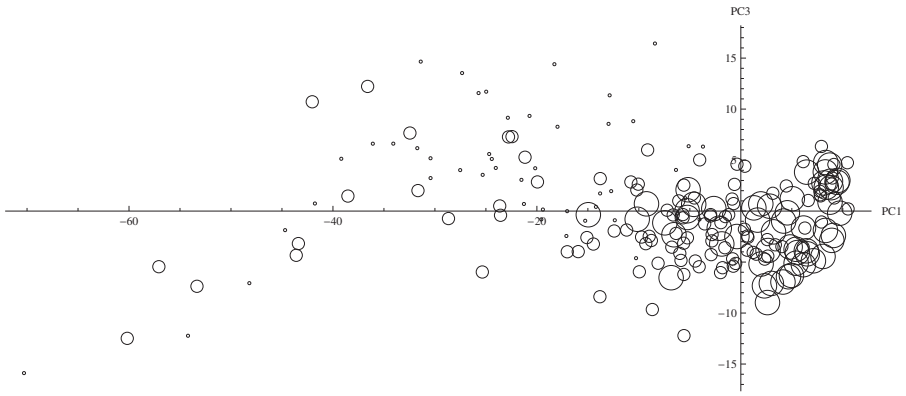

Figure 3: Representation of the IG sediments, for PCA in H_ϕ. The diameter of each disk depends on $norg \in \{1, 2 \cup 4, 8\}$.

4 Discussion

We will first tackle an important point raised by one of the referees and, second, discuss relationships between our work and multivariate exploratory methods prevailing in the Geosciences community.

FPCA or FLDA?

In Section 3, we examined only the three first components of each analysis for testing differences between the classes OG, $NCOG$ and IG. But would the assertion $NCOG \approx COG$ remain acceptable in a larger space, *e.g.* of dimension $d = 10$?
We proceeded that way because the three first components corresponded to a high level of variance ($> 90\%$) and because keeping too many components may be hazardous, due to possible degeneracies of $\sum_{COG}$ and $\sum_{OG}$ causing numerical problems in Formula (1). But we agree that sometimes classes are not "visible" in the first principal planes of an exploratory analysis. For instance, in [8], we obtained satisfactory results with 7 principal components, but none of them was able to separate the considered classes. In such cases, it is necessary to supersede FPCA by Functional Linear Discriminant Analysis (FLDA) whose goal is to separate classes, irrespective of their quality of representation. As a consequence, the components issued from FLDA may not correspond to important structural aspects of the data. Furthermore, such axes may signify nothing: it is well-known ([20], Ch. 10) that in the functional case, one must beware of spurious correlations due to the great number of explicative variables. Consequently, most authors recommend to include a regularization step in FLDA. The most sophisticated regularization method is penalization ([20], Ch. 12), while the simplest one consists in smoothing

(or filtering) the data by keeping only the first d components of a preliminary FPCA [7, 8]. Thus, in both cases (FPCA or FLDA), the point is to determine the right dimension d before practicing either method. This can be done by incorporating sphericity tests (see for instance [19]) in the procedures.

Connections with compositional data analysis

The proposed method is widely applicable for analyzing families of measures, but, of course, other methods are available, associated with other geometrical settings for FPCA [9, 20]. This was discussed in [14], and we will rather focus on connections with other methods, introduced in Geosciences by Aitchison [1]. Egozcue *et al.* [6] proposed a nice generalization of Aitchison's geometry [1] to probability density function (relative to Lebesgue's measure), but did not mention neither the possibility of another reference measure, nor the possibility of scale changes. Tolosana-Delgado *et al.* [23] introduced a reference distribution $N(z)$, but with a goal different from ours: it is essentially used for building an orthonormal basis for the Hilbert space $A^2(N)$ of densities associated with N and Aitchison's geometry. But the main problem with compositional methods is the possible presence of zeros (or very small values) in the data (see for instance [16]). This results in a heavy constraint upon the function of $A^2(N)$: their logarithm must be square-integrable with respect to N. On the contrary, our method can be widely used, even with signed measures [14], and we have no problem with zeros, at least in the symbolical case.

In the empirical case, the situation is different: firstly, the common interval $[a, b]$ should be included in the support of the "source" sediment ν_0, but this is not enough to avoid numerical difficulties: it is possible that there exists some ν_0-negligible interval $\varnothing \neq [\alpha, \beta] \subsetneq [a, b]$, such that $\nu_0([\alpha, \beta]) \approx 0$. Solving the "fractile problem" gives a practical solution to this problem, because if T_K is (quasi-)optimal, there is an index i such that $[\alpha, \beta] \subsetneq [T_K^{i-1}, T_K^i[\cup[T_K^i, T_K^{i+1}[$, with $\nu_0\left([T_K^{i-1}, T_K^i[\right) \approx \nu_0\left([T_K^i, T_K^{i+1}[\right) \approx \frac{1}{K}$. Thus, PCA in $H_{\nu_0}^K$ is not problematic. This construction should be also useful for applying compositional methods to grain-size (or similar) curves. Indeed, the most widely used compositional method (Logcontrast PCA [1]) consists in performing PCA on vectors of the K-simplex, after transformation by the centered logratio function $clr(s) := \ln\left(\frac{s}{g(s)}\right)$, where $g(s) := \left(\prod_{i=1}^{K} s_i\right)^{1/K}$ is the geometrical mean. Denoting $I_K := (1, \cdots, 1)/K$ the neutral perturbation [1] on the K-simplex, one can observe that $clr(I_K) = 0$. Thus, using an (quasi-)optimal mesh put the reference distribution near the neutral perturbation, while its image under clr put it close to the origin of the cloud of sub-sampled distributions, which is quite desirable.

Bibliography

[1] Aitchison, J. (2003), *The statistical analysis of compositional data*, The Blackburn Press, revised ed.

[2] Buller, A. T. and McManus, J. (1972), *Simple metric sedimentary statistics used to recognize different environments*, Sedimentology, **18**, 1-21.

[3] Ciutat, C., Weber, O., Gerino, M. and Boudou, A. (2006), *Stratigraphic effects of tubicifid in freshwater sediments: a kinetic study based on X-ray images and grain-size analysis*, Acta Oecologica, **30**, 228-237.

[4] Davis, P. J. (1963), *Interpolation and approximation*, Blaisdell, New York.

[5] Devoto, D. and Martinez, S. (1998), *Truncated Pareto law and oresize distribution of ground rocks*, Mathematical Geology, **30**, 661-673.

[6] Egozcue, J.J., Diaz-Barrero, J.L. and Pawlowsky-Glahn, V. (2006), *Hilbert space of probability density functions based on Aitchison geometry*, Acta Mathematica Sinica, English series, **22**, **4**, 1175-1182.

[7] Ferraty, F. and Vieu, P. (2003), *Curves discrimination: a non-parametrice functional approach*, Computational Satistics & Data Analysis, **44**, 161-173.

[8] Khelil, A., Mante, C. and David, P. (1997), *Localisation et discrimination de signaux acoustiques de bulles d'air par des techniques statistiques*, Traitement du Signal, **14**, **2**, 151-159.

[9] Kneip, A. and Utikal K. J. (2001),*Inference for density families using Functional Principal Component Analysis (with discussion)*, J. Amer. Soc. **96**, 519-542.

[10] Kolmogorov, A. N. (1992), *On the logarithmic normal distribution of particle sizes under grinding*, In: Shiryayev, A. N. (Ed.), Selected works of A. N. Kolmogorov, vol. 2, Kluwer Academic Publishers, London, 281-284.

[11] Krzanowski, W.J. (2005), *Principles of Multivariate Analysis*, Oxford University Press, revised ed.

[12] Lavit, C. and Escoufier, Y. (1994), The ACT (STATIS method), Computational Statistics & Data Analysis, **18**, 97-119.

[13] Mante, C. (1999), *The use of regularization methods in computing Radon- Nikodym derivatives. Application to grain-size distributions*, SIAM Journal on Scientific Computing, **21**, **2**, 455-472.

[14] Mante, C., Yao, A.F. and Degiovanni,C. (2007), *Principal Components Analysis of measures, with special emphasis on grain-size curves*, Computational Statistics & Data Analysis, **51**, 4969-4983.

[15] Mante, C. (2012), *Application of iterated Bernstein operators to distribution function and density approximation*, Applied Mathematics and Computation, doi:10.1016/j.amc.2012.02.073.

[16] Martin-Fernandez, J.A., Barcelo-Vidal, C. and Pawlowski-Glahn, V. (2003), *Dealing with zeros and missing values in compositional data sets using nonparametric imputation*, Mathematical Geology, **35**, **3**, 253-278.

[17] Mate, L. (1989), *Hilbert space methods in Science and Engineering*, Adam Hilger, Bristol.

[18] McLaren, P., Hill, S.H. and Bowles, D. (2007), *Deriving transport pathways in a sediment trend analysis (STA)*, Sedimentary Geology, **202**, 489-498.

[19] Perez-Neto, P. R., Jackson, D. A. and Somers K. M. (2005), *How many principal components? Stopping rules for determining the number of non-trivial axes revisited*, Computational Statistics & Data Analysis, **49**, 974-997.

[20] Ramsay, J. O. and Silverman B. W. (1997), *Functional Data Analysis*, Springer Series in Statistics, Berlin.

[21] Sevy, J.C. (1995), *Lagrange and least-square polynomials as limits of linear combinations of iterates of Bernstein and Durrmeyer polynomials*, Journal of Approximation Theory, **80**, 267-271.

[22] Stoline, M. R. (1981), *The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in One-way ANOVA designs*, The american Statistician, **35**, **3**, 134-140.

[23] Tolosana-Delgado, R., van den Boogaart, K.G., Mikes, T. and von Eynatten, H. (2008), *Statistical treatment of grain-size curves and empirical distributions: densities as compositions?*. In: Daunis-i-Estadella, J. and Martin-Fernandez, J.A. (Eds.), Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, May 27-30, University of Girona, Girona (Spain).