**Made available by Hasselt University Library in https://documentserver.uhasselt.be**

# Clustering multiply imputed multivariate high-dimensional longitudinal profiles

Peer-reviewed author version

BRUCKERS, Liesbeth; MOLENBERGHS, Geert & DENDALE, Paul (2017)
Clustering multiply imputed multivariate high-dimensional longitudinal profiles. In: BIOMETRICAL JOURNAL, 59(5), p. 998-1015.

DOI: 10.1002/bimj.201500027
Handle: http://hdl.handle.net/1942/24959

Biometrics XX, 1–23 DOI: 00.0000/j.0000-0000.0000.00000.x
Month 20XX

# Clustering Multiply Imputed Multivariate High-Dimensional Longitudinal Profiles

**Liesbeth Bruckers[1,*], Geert Molenberghs[2], Paul Dendale[3]**

[1]I-BioStat, Universiteit Hasselt, Agoralaan, Diepenbeek, Belgium

[2]I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek,Belgium

and I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium.

[3]Jessa Hospital, Heart Centre Hasselt, Hasselt, Belgium

and Universiteit Hasselt, Faculty of Medicine and Life Sciences, Diepenbeek, Belgium

*email:* liesbeth.bruckers@uhasselt.be

Summary: In this paper a procedure for clustering high dimensional multivariate data with missing observations is proposed. Functional data analysis often utilizes dimension reduction techniques such as principal component analysis. Dimension reduction techniques require complete data matrices. To overcome this problem, the data were completed by means of multiple imputation. Each imputed data set was subjected to a cluster procedure for multivariate functional data. Consensus clustering was subsequently applied to summarize the ensemble of partitions into the final cluster result. The uncertainty in cluster membership, due to missing data, was characterized by means of the agreement between the members of the ensemble and the fuzziness of the consensus clustering. The usefulness of the method was illustrated on the heart failure data.

Key words: Cluster analysis; Data reduction; Functional data analysis; Missing data.

## 1. Introduction

Repeated measures and multivariate outcomes are very common in social, behavioral, and educational sciences, as well as in clinical trials. A lot of methodological work has been done to extend cluster analysis to these complex data structures, in particular repeated measures.

When analyzing repeated measurements data, individual differences in evolution are generally captured by random effects, often via linear mixed models (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). Individual differences can also be described by latent trajectory classes (Land and Nagin, 1996; Nagin and Land, 1993, Nagin, 1999; Nagin and Tremblay, 2001) or by growth mixture models (Muthén and Shedden, 1999; Muthén and Muthén, 1998, 2007). When, for each patient, more than a single outcome is measured over time, a multivariate set of longitudinal profiles is obtained. Interest could be in finding subgroups of patients that are similar in their evolution over time for the various repeated sequences. Examples can be found in Nagin and Tremblay (2001) and Nagin (2005). Growth mixture modeling for a multivariate longitudinal data setting is often problematic. When the number of repeated outcomes is large, computational problems are likely in the estimation process due to the high dimension of the joint distribution of the random effects. Alternative approaches, such as a two-stage method (Putter et al, 2008), a latent variable model for repeated measures assuming an underlying quantity of main interest (Roy and Lin, 2000), and an algorithm using pseudo-likelihood and ideas based on $k$-means clustering (Bruckers et al, 2014) have been explored.

Nowadays, data complexity and dimensionality are enhanced by novel data collection techniques. These techniques permit observations to be densely sampled over a continuum, usually time. The data then reflect the influence of a (set of) smooth function(s) underlying and generating the observations. Often, the evolutions are not easily described by a mathematical formula. The dependencies between these so-called functional data curves can be analyzed by

methods from the functional data analyses framework. As usual, observed heterogeneity can be corrected for via explanatory variables. Unobserved sources of population heterogeneity can be investigated via cluster analyses, where the main objective is to classify patients into homogenous groups. However, clustering functional data in general requires first a reduction of the high dimension of the data.

Cluster analyses and data reduction techniques are hampered by missing values—an issue often intertwined with longitudinal data. In the regression context, a multiple imputation procedure (Rubin, 1987; Schafer, 1997; Carpenter and Kenward, 2013) can be applied to quantify the extra uncertainty in estimators of population parameters due to the missing values. Applying a cluster algorithm on the imputed data set results in multiple partitionings of the patients. It is however not so clear how uncertainty due to the imputation process needs to be reflected in the final result. Basagaña et al. (2013) present a framework for multiple imputation in cluster analysis. They suggest ways to report how the final number of clusters, the result of a variable selection procedure and the assignment of individuals to clusters is affected by the missing values. Their final decision on a patient's cluster membership is based on a majority vote.

We propose to approach the problem as a combinatorial optimization problem to summarize the cluster ensemble into a single consolidated clustering and at the same time measure the missing data influence in the cluster analyses. In this paper, we apply a model-based clustering technique to a multivariate functional data set after multiple imputation. The concept of functional data is briefly introduced in Section 3. The final data analysis brings together a number of statistical techniques that are briefly introduced: multivariate functional data and functional principal component analyses, as a data reduction technique, are described in Section 4. A summary overview of cluster methods for functional data is

given in Section 6. The ensemble method for clustering is the topic of Section 7. The various steps of the proposed procedure are graphically displayed in Figure 1.

[Figure 1 about here.]

Section 9 illustrates the methodology on telemonitoring data for chronic heart failure patients, introduced in Section 8. Daily measurements — of blood pressure, heart rate and body weight — are collected to better monitor a patient's instantaneous risk for heart failure.

## 2. Multiple Imputation

Data reduction techniques, like principal component analysis, require rectangular data structures. Records with missing values are discarded in the analyses. To circumnavigate this problem, multiple imputation was used to create a set of complete/rectangular data sets.

Multiple imputation is a popular tool for dealing with data when they are only partially observed (Rubin, 1987; Schafer, 1997; Carpenter and Kenward, 2013; Molenberghs and Kenward, 2007). The idea is to use the observed information to impute a sensible value for the missing ones. To reflect the uncertainty in this prediction, missing values are imputed multiple times. Multiple imputation is appealing because it results in complete data sets, that can be analyzed with standard statistical techniques. Two routes can be followed: multivariate or fully conditionally specified imputation (Schafer, 1997; Little and Rubin, 2002; Van Buuren et al, 1999; Raghunathan et al, 2001). Both approaches assume the missing data to be missing at random (MAR, Little and Rubin, 2002). Under the MAR assumption, the probability that an observation is missing, is driven only by the observed data, implying that no extra information is contained in the missing part of the data.

Standard imputation models applied to longitudinal data can lead to absurd results (Honaker and King, 2010). Imputations falling far from previous and subsequent observations, or imputations that are very implausible on the basis of common sense. Honaker and King

(2010) developed the software package AMELIA that facilitates imputation of (among others) smooth time-series patterns. AMELIA implements a so-called EMB algorithm. This algorithm combines the classical EM procedure with a bootstrap approach to take draws from the posterior.

## 3. Functional Data

Functional data analysis (FDA) can be seen as an extension of classical multivariate methods where data are not vectors but rather functions or curves. Functional data describe a process that changes smoothly and continuously over a domain. Often, this domain is time, resulting in repeated measurement data, but it can be anything, such as, for example space or energy. Data in many fields result from a process that is functional. Ramsay and Silverman (2005) provide many examples.

In functional data analysis, the existence of a smooth function $x$ is assumed. This function gives rise to data $y_j$, superimposed by measurement error $\varepsilon_j$, usually observed at discrete time points $t_j$, such that $y_j = x(t_j) + \varepsilon_j$. Although the curves are sampled for a finite set of time-points, the observations are supposed to belong to an infinite-dimensional space. The functional form of the data is often reconstructed from the discrete observations by assuming that the finite-dimensional space is spanned by a basis of functions. Consider a basis $\phi = \{\phi_1, \ldots, \phi_K\}$ and represent the functional data $x_i(t)$, for patient $i$, by a linear combination of the $K$ basis functions: $x_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t)$. The basis coefficients are estimated so that the constructed curve optimally fits the data for a certain degree of smoothing. The number of basis functions can be chosen in terms of a bias-variance trade-off (Ramsay and Silverman, 2005).

## 4. Principal Component Analysis of Functional Data

For high-dimensional multivariate data, a dimension reduction via principal component analysis (PCA) is usually performed prior to applying a statistical procedure in order to avoid the effects of the curse of dimensionality. The principal components (Hotelling, 1933), in the multivariate situation when data for $N$ subjects is obtained for $p$ variables, are defined as:

$$f_{im} = \sum_{j=1}^{p} \beta_{jm} x_{ij}, \quad i = 1, \ldots, N. \tag{1}$$

with $\beta_{jm}$ a set of orthogonal weights that maximize the variation in the $f_{im}$. The solutions to this maximization problem are given by the eigenvectors of the eigenequation $V\boldsymbol{\beta} = \lambda\boldsymbol{\beta}$, with $V$ the $p \times p$ sample variance-covariance matrix. A sequence of eigenvalue-eigenvector pairs $(\lambda_m, \boldsymbol{\beta_m})$ satisfies this eigenequation, with $\boldsymbol{\beta_m}$ orthogonal.

For functional data, a continuous index $s$ is taking over the role of the discrete index $j$ in (1). The principal component scores, for univariate functional data, are obtained as the inner product of two functions, the weight function and the data function (Ramsay and Silverman, 2005):

$$f_i = \int \beta(s) x_i(s) ds, \quad i = 1, \ldots, N.$$

A sequence of weight functions $\beta_m(s)$ is chosen such that they define the most important modes of variation in the curves, conditional on the weights to be orthonormal. So,

(1) $\frac{1}{N}(\int \beta_m x_i)^2$ is maximal,

(2) $\|\beta_m^2\| = \int (\beta_m)^2 = 1$,

(3) $\int \beta_m \beta_k = 0, \;\; k \neq m.$

Functional principal component analysis is also tantamount to solving an eigenequation. Define the sample variance-covariance function as $v(s,t) = \frac{1}{N}\sum_{i=1}^{N} x_i(s) x_i(t)$. Then $V$, in the functional version of PCA, is a variance operator and transforms a function $\beta$ as $V\beta =$

$\int v(.,t)\beta(t)dt$. The eigenequation can then be expressed as:

$$V\beta(s) = \int v(s,t)\beta(t)dt = \lambda\beta(s), \tag{2}$$

where $\beta$ are eigenfunctions now instead of vectors. Web Appendix A describes how the solutions to this continuous functional eigenanalysis problem (2) can be obtained.

When extending functional PCA to $M$-variate functional data, the weight functions become $M$-vector functions $\beta = (\beta^1, \ldots \beta^M)'$, with $\beta^l$ depicting the variation in the $l^{th}$ dimension (Berrendero et al, 2011; Ramsay and Silverman, 2005). The principal component scores are again linear combinations of the data:

$$f_i = \sum_{m=1}^{M} \int \boldsymbol{\beta^m} x_i^m,$$

where the weight functions $\boldsymbol{\beta}^m$ are solutions of an eigenequation system $\boldsymbol{V\beta} = \lambda\boldsymbol{\beta}$. $\boldsymbol{V}$ is the covariance operator as defined before, $v_{ii}(s,t)$ is the covariance operator for the $i^{th}$ functional data dimension and $v_{ij}(s,t)$ the cross-covariance operator between dimensions $i$ and $j$. The eigenequation translates to a system of equations:

$$\begin{cases} v_{11}\boldsymbol{\beta^1} + v_{12}\boldsymbol{\beta^2} + \ldots + v_{1m}\boldsymbol{\beta^m} = \lambda\boldsymbol{\beta^1}, \\ v_{21}\boldsymbol{\beta^1} + v_{22}\boldsymbol{\beta^2} + \ldots + v_{2m}\boldsymbol{\beta^m} = \lambda\boldsymbol{\beta^2}, \\ \qquad\qquad\qquad \vdots \\ v_{m1}\boldsymbol{\beta^1} + v_{m2}\boldsymbol{\beta^2} + \ldots + v_{mm}\boldsymbol{\beta^m} = \lambda\boldsymbol{\beta^m}. \end{cases}$$

In practice, a standard principal component analysis is carried out on a vector $Z_i$ concatenating all data functions of patient $i$.

## 5. Density for Functional Data

Model-based clustering identifies homogenous subgroups of patients using a mixture model for the density function of the data. Delaigle and Hall (2010) use the Karhunen-Loève expansion to introduce the notion of a probability density for functional data.

The basis, yielding a minimum value for the total mean squared error when decomposing

a stochastic process $\boldsymbol{X}(t)$ as an infinite linear combination, is the set of orthogonal eigenfunctions of the process itself:

$$\boldsymbol{X}(t) = \boldsymbol{\mu}(t) + \sum_{j=1}^{\infty} f_j \beta_j(t).$$

If $\boldsymbol{\mu}(t) = 0$, i.e., for a centered process, the composition is referred to as the Karhunen-Loève expansion (Karhunen, 1947; Loève, 1978). The basis coefficients are random variables, in contrast to the coefficients resulting from, for example, a polynomial basis. The random variables $f_j$ are uncorrelated, have zero mean and variance $\lambda_j$. We denote the distribution of $f_j$ by $\mathfrak{f}_j$. The variables $f_j$ follow a Gaussian distribution and are stochastically independent for a Gaussian process.

Let $p(\boldsymbol{x}|h) = P(\|\boldsymbol{X} - \boldsymbol{x}\| \leqslant h)$ for $h > 0$ and $\|\boldsymbol{X} - \boldsymbol{x}\|$ the $L_2$-distance between $\boldsymbol{X}$ and $\boldsymbol{x}$. Then, $p(\boldsymbol{x}|h)$ is the probability that $\boldsymbol{X}$ belongs to a ball of radius $h$ centered at $\boldsymbol{x}$. Delaigle and Hall (2010) show that this probability can be written as a product of the densities $\mathfrak{f}_j$, corresponding to the largest eigenvalues:

$$\log\ p(\boldsymbol{x}|h) = C_1(r, \theta) + \sum_{j=1}^{r} \log \mathfrak{f}_j(f_j) + O(r), \tag{3}$$

where $\mathfrak{f}_j(f_j) = \mathfrak{f}_j(f_j(\boldsymbol{x}))$ is the density of the $j$ principal component score evaluated for the $j$ component score for $\boldsymbol{x}$; $r = r(h)$ diverges to infinity as $h$ decreases to zero, and $C_1$ depends on $h$ and on the infinite eigenvalue sequence, $\theta$. Based on (3), a natural surrogate for the log density of functional data is provided by the average of log densities of the $r$ largest principal components. This log-density $l(\boldsymbol{x}|r) = r^{-1} \sum_{j=1}^{r} \log \mathfrak{f}_j\,(f_j)$ captures variation with $\boldsymbol{x}$ up to order $r$.

## 6. Clustering of Functional Data

An excellent review of approaches to clustering functional data is presented by Jacques and Preda (2013). They classify the approaches into four categories: raw-data clustering, two-stage procedures, model-based procedures and nonparametric techniques for clustering

functional data. We opt for a model-based clustering, using principal components. This procedure tackles the functional nature of the data, simultaneously performs a data reduction and cluster exercise, while at the same time allowing for complex covariance structures in the multivariate longitudinal profiles.

Jacques and Preda use the approximation of the probability density for functional random variables to fit a parametric mixture model to univariate functional data (Jacques and Preda, 2012) and to multivariate functional data (Jacques and Preda, 2013). We briefly summarize the different steps of their algorithm.

Assume the existence of a latent group indicator $Z = (Z^1, \cdots, Z^K)$ for $K$ clusters. For subject $i$, $Z_i^g = 1$ if its curves $\boldsymbol{x}_i$ belong to group $g$, 0 otherwise. Let $Z$ have a multinomial distribution with mixing proportions $\pi_1, \ldots, \pi_K$ ($\sum_{g=1}^K \pi^k = 1$). Under these assumptions, the unconditional approximated density of $\boldsymbol{X}$ is equal to

$$\mathfrak{f}_{\mathbf{X}}^{(q)}(\boldsymbol{x}; \theta) = \sum_{g=1}^{K} \pi_g \prod_{j=1}^{q_g} \mathfrak{f}_{j,g}(f_{j,g}(\boldsymbol{x}); \lambda_{j,g}).$$

When $\boldsymbol{X}$ is a Gaussian process, the $\mathfrak{f}_{j,g}$ are Gaussian. The parameters $\theta = \{(\pi_g, \lambda_{1,g}, \cdots, \lambda_{q_g,g})_{1 \leqslant g \leqslant K}\}$ and $q = (q_1, \cdots, q_K)$ are estimated by maximizing the pseudo completed log-likelihood via an iterative EM algorithm:

$$L^{(q)}(\theta; \{X_1, \cdots, X_n\}, \{Z_1, \cdots, Z_n\}) = \sum_{i=1}^{n} \sum_{g=1}^{K} Z_i^g \left( \log(\pi_g) + \sum_{j=1}^{q_g} \log(\mathfrak{f}_{j,g}(f_{i,j,g}(\boldsymbol{x}_i)) \right),$$

where $f_{i,j,g}$ is the $j^{th}$ principal component of curves $\boldsymbol{x}_i$ belonging to group $g$.

At iteration $h$, the E-step of the EM-algorithm evaluates the conditional expectation of the pseudo completed log-likelihood, with respect to unknown $Z_i^g$, given the observed data

and current parameter estimates:

$$\begin{aligned}
\Theta(\theta, \theta^{(h)}) &= E_{\theta^{(h)}}[L^{(q)}(\theta; \boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x}] \\
&= \sum_{i=1}^{n}\sum_{g=1}^{K} E_{\theta^{(h)}}[Z_i^g|\boldsymbol{X} = \boldsymbol{x}] \left(\log(\pi_g) + \sum_{j=1}^{q_g} \log(\mathfrak{f}_{j,g}(f_{i,j,g}(\boldsymbol{x_i}; \lambda_{j,g})))\right) \\
&\simeq \sum_{i=1}^{n}\sum_{g=1}^{K} \frac{\pi_g \prod_{j=1}^{q_g} \mathfrak{f}_{j,g}(f_{i,j,g}(\boldsymbol{x_i}); \lambda_{j,g})}{\sum_{g=1}^{K} \pi_k \prod_{j=1}^{q_g} \mathfrak{f}_{j,g}(f_{i,j,g}(\boldsymbol{x_i}); \lambda_{j,g})} \\
&\quad \times \left(\log(\pi_g) + \sum_{j=1}^{q_g} \log(\mathfrak{f}_{j,g}(f_{i,j,g}(\boldsymbol{x_i}; \lambda_{j,g})))\right)
\end{aligned}$$

where $\mathfrak{f}_{j,g}(f_{i,j,g}(\boldsymbol{x_i}); \lambda_{j,g})$ is the value of $\mathfrak{f}_{j,g}$ for $\boldsymbol{X}_i = \boldsymbol{x}_i$.

Before executing the M-step, Jacques and Preda (2013) update the group-specific principal components $f_{j,g}$. For this purpose, a weighted principal component analyses is fitted, with weights $E_{\theta^{(h)}}[Z_i^g|\boldsymbol{X} = \boldsymbol{x}]$. Furthermore, the class-specific dimensions $q_g$ are selected by means of the scree-test of Cattell (Cattell, 1966). After these intermediate steps, the M-step maximizes $\Theta(\theta, \theta^{(h)})$ with respect to $\theta$.

Jacques and Preda note that this procedure does not guarantee an increase in the pseudo likelihood between two iterations. The reason for this is that an approximation to the density of functional data is used. They advise to pre-run the algorithm a couple of times with different (random) starting values, using a small number of iterations. The best solution among these is then to be used as the starting point for the algorithm with a large number of iterations (Biernacki, 2004). This empirical strategy increases the chance of convergence to a local maxima.

## 7. Consensus Clustering

Cluster ensembles are collections of individual solutions to a given clustering problem (Strehl and Ghosh, 2002). Let $\mathfrak{X} = \{x_1, x_2, \ldots, x_n\}$ denote a set of objects/samples, where each $x_i$ is some $\alpha$-dimensional data vector. A partitioning of the $n$ objects into $k$ clusters can be represented as a set of $k$ sets of objects $(\mathcal{C}_l|l = 1, \ldots k)$ or as a label vector $\lambda \in \mathbb{N}^n$. The cluster algorithm (function) to obtain this label vector is called a clusterer $\Phi$. The label vector $\lambda$

containing the class identifiers is not unique. The class labels can be permuted arbitrarily without changing the underlying partition. The resulting partition can be a soft (fuzzy) or a hard (crisp) partition. Results obtained from applying different clusterers $\Phi$ on a dataset can be quite different but all equally plausible. The problem of combining multiple partitionings into a single clustering is referred to as cluster ensembles. It is assumed that the consensus cluster is less likely to be biased towards the models ($\Phi$) used in the separate analyses and more likely to reflect the underlying structure of the data. Day (1986) and Leclerc (1998) studied the consensus of hard partitions; fuzzy consensus clustering has been investigated by Gordon and Vichi (2001).

Intuitively, the final consensus is the partition of the $n$ objects that shares most information with the original clusterings. Consensus clustering synthesizes the information in the elements of a cluster ensemble into a single clustering, often by minimizing a criterion function measuring how (dis)similar consensus candidates are from the ensemble (the so-called optimization approach to consensus clustering). Since there is no relation between the labels assigned to object $i$ by a clusterer ($\Phi_1$) and another clusterer ($\Phi_2$) the cluster ensemble problem is more difficult than a classifier ensemble problem. This label correspondence issue is the main problem that has to be dealt with when clustering ensembles. The problem can be solved via the Hungarian method (Kuhn, 1955). An additional issue is that the number and shape of the input clusters may be different and that the optimal final number of clusters is often not known in advance.

To state the cluster ensemble as a problem of mapping a set of $r$ labelings, $\lambda^{(1,\cdots,r)}$, to a single consensus clustering, $\lambda$, a consensus function $\Gamma$, $\mathbb{N}^{n\times r} \rightarrow \mathbb{N}^n$ is needed: $\Gamma : \{\lambda^{(q)}|q \in \{1,\cdots,r\}\} \rightarrow \lambda$. An estimate $\widehat{\lambda}$ is often obtained by maximizing (minimizing) a criterion/objective function measuring how (dis)similar consensus candidates are from the ensemble. Measures for dissimilarity and similarity are key ingredient to clustering

(ensembles). Let $d$ be a suitable dissimilarity measure; most popular criterion functions are of the form

$$L(\lambda) = \sum w_b d(\lambda^b, \lambda)^p, \tag{4}$$

where $w_b$ is a weight given to element $\lambda^b$ of the ensemble, and $p \geqslant 1$. If $p = 1$ the consensus solution is called a median of the ensemble, while $p = 2$ gives least squares consensus partitions (Gordon, 1999). A variety of methods are available to minimize criteria of this form; fixed-point algorithms for soft Euclidean and Manhattan consensus partitions, greedy algorithms, SUMT algorithms, and exact solvers (Hornik, 2005). A multiplicity of (dis)similarity measures are described in the literature. Among the ones commonly used are the Euclidean and Manhattan dissimilarity of the memberships (Dimitriadou, Weingessel and Hornik 2002), the Rand index (Rand 1971, Gordon 1999), Normalized Mutual Information (Strehl and Ghosh 2002), the Katz-Powell index (Katz and Powell 1953), the Jaccard index, etc. The maximization in (4) ranges over all possible $k$-partitions (Sthrel and Ghosh, 2002). An exhaustive search over all possible clusterings with $k$ labels for the one with the maximum ANMI is in general not possible. Dimitriadon, Weingessel and Hornik (2002) have shown that optimal matching can be determined very efficiently when agreement is expressed as Euclidean partition dissimilarity. Web Appendix B illustrates the idea of concensus clustering for the normalized mutual information (NMI).

To evaluate the reliability of a partition of a data set, the fuzziness in the partitioning can be investigated. In fuzzy clustering, a data point does not completely belong to just one cluster but has a probability of belonging to each cluster. Points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center. The uncertainty of a fuzzy partition can be quantified via the the Partition Coefficient, $\sum_{i,j} \mu_{i,j}^2$, and the Partition Entropy, $\sum_{i,j} H(\mu_{i,j})$, where $H(u) = u\log(u) - (1-u)\log(1-u)$ (Bezdek 1981).

## 8. Telemonitoring in the Management of Heart Failure Study

The intent of the TElemonitoring in the MAnagement of Heart Failure study (TEMA-HF1) was to investigate whether intensive follow-up of chronic heart failure patients—through modern communication technology, facilitating collaboration between general practitioners and a heart failure clinic—could reduce mortality and re-hospitalization rate. Details regarding the design and results of TEMA-HF1 are reported elsewhere (Dendale, 2012).

Chronic heart failure (CHF) is characterized by recurrent hospitalizations due to fluid overload and/or worsening of renal function. To reduce morbidity, mortality and healthcare cost, regular adjustment of the treatment of CHF patients is needed. Telemonitoring proofs to be a valuable instrument to collect day-to-day measurements of important parameters, resulting at the end in an improved clinical outcome.

In the TEMA-HF1 study, 160 CHF patients, hospitalized in 7 Belgian hospitals, were included between April 2008 and June 2010. Patients were randomly assigned to receive usual care (UC) after discharge, or to be intensively followed for up to 6 months by telemonitoring (TM). To illustrate the methodology outlined in this article, only data from the TM group was used. For this group, the telemonitoring device daily transferred data on body weight, blood pressure (systolic and diastolic), and heart rate. Missing information on two consecutive days provoked an alert, patients were contacted to motivate them to make the measurements. At baseline, additional patient characteristics were collected: sex, age, heart rhythm, cardiac muscle fibre stretch as measured through NTprobBNP, a fitness indicator (according NYHA class indication) and the left ventricle ejection fraction (LVEF), which is a measure of heart performance. Four TM patients left the study prematurely for motivational reasons, 4 died during the course of the 6 month study, and 16 were hospitalized at least once for heart failure related reasons.

Although alerts were sent out when the longitudinal measurement were missing for two

consecutive days, quite some missingness is present in the data. Twenty-eight percent of the patients did receive an alert concerning missing information for the heart rate, 64% concerning the blood pressure measurements and 84% concerning body weight (Dendale, 2012).

The ability, of the 4 daily-measured biomarkers, to discriminate between patients needing re-hospitalization in the near future and patients not needing to be hospitalized, has been investigate by Njeri Njagi et al. (2013). They fitted a joint model for the time to re-hospitalization and the longitudinal biomarker. The model results in a dynamic prediction, i.e., a patient-specific probability for re-hospitalization. This probability is estimated based on info in the longitudinal biomarker (the level of the biomarker and changes in the biomarker), and can (theoretically) be updated daily with every new value of the biomarker being collected.

## 9. Results

Information about the extent of missingness in the heart failure data is presented in Tables 1 and 2.

[Table 1 about here.]

[Table 2 about here.]

Baseline characteristics are fairly complete. About one out of four patients does not have information for the six minute walking test (WALK). On average, 76% of the patients' daily measurements for the biomakers were recorded. Meaning that on average for 137 days out of 180, heart rate, diastolic and systolic blood pressure were communicated to the heart failure clinic by means of the telemonitoring device. The heart failure data has particular features. Heart rate and blood pressure are recorded by the same device and thus simultaneously missing or present. The periods lacking telemonitoring data, are, in general,

not too long (average duration is 6 days, median duration is 1 day). However, some patients are featured by longer periods of lacking data. About 5% of the periods, with missing info on consecutive days, lasts longer than 2 weeks. Fifteen patients (8%) dropped out and left the study prematurely (before day 170). The mean follow-up time is 163 days.

The EMB algorithm implemented in AMELIA (Honaker and King, 2009) was used to obtain ten complete data matrices. A natural logarithm transformation was applied to the longitudinal measurements of heart rate, blood pressure, and body weight in order to normalize the distributions. The imputation model included all patients' baseline characteristic. For the daily-measured biomakers a smooth model over time was imposed, with patient specific time trends. Specifically, a cubic spline model was specified. The EM algorithm can suffer from numerical instability when the number of parameters in the imputation model is high and/or when the degree of missingness is high. Therefore, a ridge prior of 10% was used. Multiple imputation leads to valid results when the imputation model is correctly specified and missingness is missing-at-random (MAR). MAR cannot be formally tested for. But the accuracy of the imputed values can be judged by over-imputing. Each observed value, in succession, is treated as if it was missing. After a large number of imputations, it can be investigated if the actual observed value falls within the range of imputed values. Based on this technique it can be concluded the imputation model is acceptable (graph not shown).

The model-based cluster algorithm for multivariate functional data, described in Section 6, was then carried out on each completed data set. Basically the method boils down to applying a parametric mixture model to the surrogate density of the functional data. Multivariate functional principal components analyses is a key building block for as much as the surrogate density function is determined by the PC scores. Since the units of the four biomarkers are different (kg, bpm, and mm Hg), the data were first normalized, $\boldsymbol{Y}(t) = R(t,t)^{-1}X(t))$ with $R(t,t) = \sqrt{(V(t,t))}$, whereupon the contribution of the 4 biomarkers, in defining the

principal components, is the same. The response profiles were first smoothed by means of a cubic spline basis with 69 basis functions. A patient's evolution in diastolic, and systolic blood pressure, heart rate and weight can be well summarized by the first three principal component scores. Sixty-nine percent of the variability in these biomarkers is explained by three principal components: 28% (range 27–29%) is attributable to the first principal component, 22% (range 21–25%) to the second principal component and finally the third component adds another 19% (range 18–20%). These are percentages averaged over the ten imputed data sets. Graphical displays of the normalized curves of the 4 responses and of the principal component scores can be found in Web Appendix C.

The model-based cluster algorithm was applied to the surrogate densities of each of the ten completed data sets separately. For each data set, the algorithm was initialized by running fifty random initializations, for 40 iterations. The random initialization resulting in the best solution (i.e., the highest pseudo likelihood value), is used as the starting point for a longer algorithm with 500 iterations. The threshold of the Cattell scree test was set to 0.05. An increase in the pseudo log likelihood value less than 1e-5 was specified as the stopping criteria. Code for R developed by Jacques and Preda (2013) was used.

For the obtained soft two-class solutions, information about the cluster sizes, the estimated orders for the surrogate density functions, and the fuzziness are given in Table 3. The Euclidean agreement between the 10 elements of the ensemble ranges from 0.67 (data set 4 and 10) to 0.94 (data set 3 and 6), with a mean Euclidean agreement of 0.80.

[Table 3 about here.]

The agreement among the ten imputed data sets is of particular interest. This measurement quantifies the uncertainty in partitioning the heart failure patients, induced by the presence of missing data. The two-class cluster solution for member 4 of the ensemble, results in a partition of (31,49) patients, for member 6 this is (15,65).

Subsequently, two-class consensus clustering was used to synthesize the information in the 10 partitions—resulting from the model-based clustering— into a single clustering. The Euclidean distance was used as dissimilarity measure, and the consensus solution was obtained by maximizing the objective function. A fixed-point algorithm, implemented in the R package CLUE (Hornik, 2005), was used. This algorithm results in a soft consensus partition.

The results are presented in Table 3. Partitioning of the 80 patients, based on their profiles for diastolic and systolic blood pressure, heart rate, and weight results in groups of sizes 63 and 17. The average agreement between the consensus clustering and the 10 members of the ensemble equals 0.78 (range 0.65–0.86). The fact that a patient is not necessarily assigned to the same cluster for each of the 10 imputed data sets introduces uncertainty in the consensus cluster assignment. This uncertainty is measurable via a patient's probability of belonging to the cluster. The normalized partition coefficient—measuring the uncertainty in a fuzzy partition — equals 0.36 for the resulting consensus clustering. The fuzziness for the consensus clustering is generally higher than the fuzziness of the 10 members of the ensemble. The fuzziness for the consensus result, reflects uncertainty in allocation as present in any cluster procedure, increased by uncertainty due to missing information in a patient's profile. The cluster allocation is clear cut for most patients. For the 65 patients assigned to cluster 1, the average probability of belonging to cluster 1 is 87%. For cluster 2 this probability equals 89%. No relation has been found between the proportion of missingness in a patient's pattern and its cluster membership. Twelve patients (19%) of cluster 1 were re-hospitalized at least once during the study, in cluster 2 four patients (24%) were re-hospitalized at least once. This difference is not statistically significant ($\chi^2$ =0.005, $p$-value = 0.94).

It is well documented (Hajnal and Loosveldt, 2000; Bradley and Fayyad (1998); Pena, Lozano and Larranaga, 1999) that cluster results are sensitive to the preferred algorithm

and the randomly selected starting values. Likewise for the proposed method, alternative options and settings could lead to different partitions of the heart failure data.

The final step in the outlined procedure (Section 7), i.e., the consensus clustering, involves a number of choices. The (dis)similarity measure, the objective function, and the optimization algorithm have to be decided. For the heart failure data, Web Appendix D describes the susceptibility of the method in terms of some of these choices. The choice of the distance measure and procedure to optimize the objective function was not very important. The choice of the scree-test threshold, or the number of principal components, to be used in the approximation of the surrogate density, on the other hand does influence the cluster results. For the first imputed data set, the number of principal components was forced to be equal for the two clusters, and changed from 1–10. It is seen that the optimization only converges when four principal components are used; and that the number of patients with unstable group allocation increases with the number of principal components diverging from 4. For the heart failure data, we conclude that the final cluster result is rather sensitive to the number of principal components selected by the scree-test.

## 10. Discussion

In this paper, a procedure for clustering high dimensional multivariate data with missing observations is proposed. Functional data analysis often utilizes dimension reduction techniques such as principal component analysis. Dimension reduction techniques require complete data matrices. To overcome this problem, the data were completed by means of multiple imputation. Each imputed data set was subjected to a cluster procedure for multivariate functional data. Consensus clustering was subsequently applied to summarize the ensemble of partitions into the final cluster result.

The uncertainty in cluster membership, due to missing data, was characterized by means of the agreement between the members of the ensemble and the fuzziness of the consensus

clustering. The usefulness of the method was illustrated on the heart failure data. However, a number of topics are still open for further investigation.

The functional representation of raw data in general involves some smoothing. In this work the data was smoothed by a cubic spline basis with 69 basis functions. But alternative smoothing methods—including other basis function, local weighting methods and roughness penalty approaches—could have been used. They all have in common that smoothing parameters (e.g., the number of basis functions, bandwidth of kernel function or penalty parameters) have to be optimally chosen.

The class-specific orders, used to describe the pseudo likelihood, are chosen through the threshold of the Cattle scree test. This is a heuristic method. Other heuristic and statistical procedures could be used to determine the number of components to be retained (Jackson, 1993).

Information criteria like AIC and BIC are generally used to determine the optimal number of clusters. These criteria can be obtained from the pseudo likelihood, but are not very useful. Only relative comparisons between a set of models attempting to fit a given dataset can be done with these. The amount of data used in the algorithm, depends on the class-specific orders resulting from the Cattle Scree test. Thus it is not guaranteed that the data used in different models is identical, which hampers the determination of the number of clusters.

Breaban and Luchian (2011) have defined a new information criterion, CritCF. This criterion takes into account the number of clusters and the number of variables for ranking partitions. This criterion could be valuable in addressing two issues at once, the issue of selecting the class-specific orders and the issue of determining the optimal number of clusters.

The proposed algorithm was applied on ten completed data sets, but the choice of the number of imputed data sets is still an open topic.

**Acknowledgements**

The authors gratefully acknowledge support from IAP research Network P7/06 of the Belgian government (Belgian Science Policy).

**Supplementary Materials**

Web Appendix A, referenced in Section 4, Appendix B, referenced in Section 7 and Appendices C and D, referenced in Section 9 are available with this paper at the Biometrics website on Wiley Online Library.

**References**

Basagaña, X., Barrera-Gömez, J., Benet, M., Antö, J. M., and Garcia-Aymerich, J. (2013). A framework for multiple imputation in cluster analysis. *American Journal of Epidemiology*, **177,** 718–25.

Berrendero, J. R., Justel, A., and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, **55,** 2619–2634.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers Norwell, MA, USA.

Biernacki, C. (2004). Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures. *Statistics and Computing*, **14,** 267–279.

Bradley, P. and Fayyad, U. (1998). Defining Initial Points for K-Means Clustering. *Proceedings of the 15th International Conference on Machine Learning (ICML98)*, Microsoft Research, May 1998.

Breaban, M. and Luchian, H. (2011). A unifying criterion for unsupervised clustering and feature selection. *Pattern Recognition*, **44,** 854-865.

Bruckers, L., Molenberghs, G. and Drinkenburg, W. (2014). A Cluster Algorithm for Multivariate Longitudinal Data. *Submitted*.

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and Its Application.* New York: John Wiley & Sons.

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research*, **1,** 245–276.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley & Sons, Inc.wi

Day, W. H. E. (1986). Foreword: comparison and Consensus of Classifications. *Journal of Classification*, **3,** 183–185.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics*, **38,** 1171–1193.

Dendale, P., De Keulenaer, G., Troisfontaines, P., Weytjens, C., Mullens, W., Elegeert, Y., Ector, B., Houbrechts, M., Willekens, K., and Hansen D. (2012). Effect of a Telemonitoring-facilitated Collaboration Between General Practitioner and Heart Failure Clinic on Mortality and Rehospitalization Rates in Severe Heart Failure The TEMA-HF 1 (Telemonitoring in the Management of Heart Failure) Study. *European Journal of Heart Failure*, **14,** 333–340.

Dimitriadou, E., Weingessel, A., and Hornik, K. (2002). A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, **16,** 901–912.

Gordon, A. D. (1999). *Classification* (second edition). Chapman & Hall/CRC, Boca Raton, Florida.

Gordon, A. D. and Vichi M. (2001). Fuzzy partition models for fitting a set of partitions. *Psychometrika*, **66,** 229–248.

Hajnal, I. and Loosveldt, G. (2000). *Data Analysis, Classification, and Related Methods Studies in Classification, Data Analysis, and Knowledge Organization.* Springer Berlin

Heidelberg.

Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software*, **14**.

Honaker, J., King, G., and Blackwell, M. (2009), *Amelia II: A Program for Missing Data.*

Honaker, J. and King, G. (2010). What to Do about Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science*, **54,** 561–581.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology*, **24,** 417–441 and 498–520.

Jackson, D.A. (1993). Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology*, **74,** 2201–2014.

Jacques, J. and Preda, C. (2013). Model-based clustering of multivariate functional data. *Computational Statistics and Data Analysis*, in press.

Jacques, J. and Preda, C. (2012). Model-based clustering of functional data. *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges*, 459–464.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice Hall.

Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.,*, **37,** 1–79.

Katz, L. and Powell, J. H. (1953). A proposed index of the conformity of one sociometric measurement to another. *Psychometrika*, **18,** 249–256.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, **2,** 83–97.

Land, K. C. and Nagin, D. S. (1996). Micromodels of Criminal Careers: A Synthesis of the Criminal Careers and Life Course Approaches via Semiparametric Mixed Poisson Regression Models, with Empirical Models. *Journal of Quantitative Criminology*, **12,** 163-191.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38,** 963–974.

Leclerc, B. (1998). *Consensus of classifications: the case of trees. In: Advances in Data Science and Classification, Studies in Classification, Data Analysis and Knowledge Organization*. Berlin, Springer-Verlag, pp. 81-90.

Loève, M. (1978). *Probability theory. Vol. II, 4th ed. Graduate Texts in Mathematics*. Springer-Verlag.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons.

Muthén, L. K. and Muthén, B. O. (1998-2007). *Mplus User's Guide. Fourth edition*. Los Angeles, CA: Muthén and Muthén.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM-algorithm. *Biometrics*, **55,** 463–469.

Nagin, D. S. and Land, K. C. (1993). Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, **31,** 327-362.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-bases approach. *Psychological Methods*, **4,** 139–157.

Nagin, D. S. and Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviours: a group-based method. *Psychological Methods*, **6**, 18–34.

Nagin, D. S. (2005). *Group-based Modeling of Development*. Cambridge, MA.: Harvard University Press.

Njeru Njagi, E., Rizopoulos, D., Molenberghs, G., Dendale, P., and Willekens, K. (2013). A

joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. *Statistical Modelling*, **13,** 179–198.

Pena J., Lozano J. and Larranaga P. (1999). An Empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Letters*, **20,** 1027–1040.

Putter, H., Vos, T., de Haes, H., and van Houwelingen, H. (2008). Joint analysis of multiple longitudinal outcomes: Application of a latent class model. *Statistics in Medicine*, **27,** 6228–6249.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, **27,** 85–95.

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

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66,** 846–850.

Roy, J. and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, **56,** 1047–1054.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman and Hall.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles. A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, **3** 583–617.

van Buuren S., Boshuizen H. C., and Knook D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, **18,** 681–694.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Figure 1 about here.]

*Received Month* 20XX. *Revised Month* 20XX. *Accepted Month* 20XX.

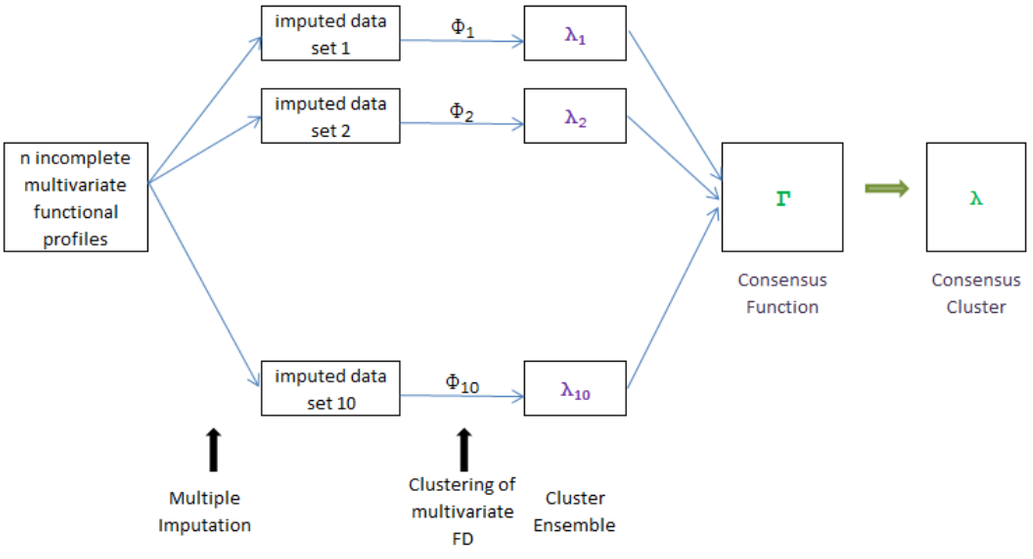


**Figure 1.** Steps of the proposed procedure.

**Table 1**
*Number of patients with missing information at baseline.*

| Characteristic | # of patients | Characteristic | # of patients |
|---|---|---|---|
| Age | 0 | LVEF | 2 |
| Gender | 0 | NTPROBNP | 4 |
| Diastolic Blood Pressure | 0 | REG-AF | 0 |
| Systolic Blood Pressure | 0 | NYHA | 0 |
| Heart Rate | 0 | WALK | 26 |
| Weight | 0 | | |

**Table 2**
*Percentage of days with missing information.*

| Biomarker | mean | median |
|---|---|---|
| Diastolic Blood Pressure | 24 | 14 |
| Systolic Blood Pressure | 24 | 14 |
| Heart Rate | 24 | 14 |
| Weight | 20 | 7 |

**Table 3**
*Number of patients assigned to clusters 1 and 2.*

| | Imputed Dataset | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | consensus |
| | # of patients | | | | | | | | | | |
| Cluster 1 | 63 | 62 | 63 | 49 | 62 | 65 | 53 | 62 | 65 | 62 | 63 |
| Cluster 2 | 17 | 18 | 17 | 31 | 18 | 15 | 27 | 18 | 15 | 18 | 17 |
| | # of principal components | | | | | | | | | | |
| Cluster 1 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 7 | 6 | 7 | - |
| Cluster 2 | 5 | 5 | 5 | 53 | 5 | 5 | 6 | 6 | 5 | 6 | - |
| Fuzziness | 0.24 | 0.22 | 0.25 | 0.37 | 0.28 | 0.23 | 0.53 | 0.27 | 0.22 | 0.22 | 0.36 |