Scandinavian Journal of Statistics Theory and Applications

doi: 10.1111/j.1467-9469.2012.00818.x


# Outcome Prediction for Heart Failure Telemonitoring Via Generalized Linear Models with Functional Covariates

STEFANO BARALDO, FRANCESCA IEVA and ANNA MARIA PAGANONI

*MOX – Modeling and Scientific Computing, Department of Mathematics 'F. Brioschi', Politecnico di Milano*

VALERIA VITELLI

*Chair on Systems Science and the Energetic Challenge, European Foundation for New Energy – Électricité de France, École Centrale Paris and Supélec (École Supérieure d'Électricité)*

**ABSTRACT. An effective methodology for dealing with data extracted from clinical surveys on heart failure linked to the Public Health Database is proposed. A model for recurrent events is used for modelling the occurrence of hospital readmissions in time, thus deriving a suitable way to compute individual cumulative hazard functions. Estimated cumulative hazard trajectories are then treated as functional data, and they are used as covariates along with clinical survey data within the framework of generalized linear models with functional covariates.**

*Key words:* functional data analysis, generalized linear models, Public Health Database, recurrent events processes

## 1. Introduction

Heart failure is a degenerative disease known worldwide as one of the most frequent causes of hospitalization among the eldest in the population. Since the frequency of crises undergone by a given patient increases along time, a growing employment of health care resources in terms of money, structures and personnel is needed. The necessity of a cost-effective solution for the care of this and other chronic pathologies has led to the experimentation of telemedicine as a possibly convenient strategy (Capomolla *et al.*, 2004; Scalvini *et al.*, 2004; Giordano *et al.*, 2008).

The basic idea of telemonitoring is to keep the patient at home and to instruct her/him about the use of monitoring instruments, which send registered information (ECG, body weight, heart frequency, etc.) to the health institution by a network connection. The physician in charge evaluates received data to properly manage the home care programme, for example by modifying drug doses and by scheduling visits.

Telemonitoring databases contain information mostly regarding the telemonitoring period itself, such as duration of the period, number of ECGs transmitted to the hospital, clinical parameters at starting and ending times, the NYHA severity class of the patient's pathology, some features regarding the last hospitalization and others. The telemonitoring outcome, that is, the conclusion of the planned period (usually 180 days, for what concerns the programmes considered in this work) without interruption by adverse events, should be related to the patients' clinical history to get better insight into the effectiveness and applicability of this strategy. To this aim, in this study we consider Hospital Dimission Forms (*Schede di Dimissione Ospedaliera*, or SDO for short) extracted from the Italian administrative Public Health Database (PHD), which gathers detailed information about hospitalization periods. The use of information about hospitalizations to study telemonitoring outcome is an innova-

tive approach, since no standard methodology exists to exploit these kinds of data. Moreover, data conveyed by administrative databases are used for clinical investigation for the first time in Italy and in the field of telemonitoring.

Since heart failure is a pathology that alternates phases of stability to sudden worsenings of the patient's condition, it is not possible to assume a stationary pattern for critical events. When dealing with time dependent observations of localized events, a natural modelling approach, yet new in the field of telemonitoring, is to consider each patient's hospitalizations as points of a non-stationary, doubly stochastic counting process. The model we consider is a Cox-type one, a specification of the general class of models introduced in Peña & Hollander (2004) and applied in González *et al.* (2005) to the study of intervention effects after cancer relapse. This class of models allows one to take into account many aspects that influence hospitalization risk and to compute the realized trajectories of the cumulative hazard process underlying the hospitalizations counting process; these longitudinal data reduce complex characteristics of the patient's clinical history to a single curve that represents each patient's instantaneous risk of hospitalization. Cumulative hazard processes are then studied in the light of functional data analysis techniques (see Ramsay & Silverman (2005) for a general presentation of the subject) to identify their main features, and used to construct a generalized linear model with functional covariates for predicting telemonitoring outcome.

The methodology proposed in this work can be divided in two distinct phases, based on the sources of data involved. First, hospital admission historical data from the PHD are used to fit a counting process model, since it is reasonable to assume that hospitalizations reflect the aggravation of the patient's condition. Hence, data gathered in the PHD are involved in the construction of hazard functions, longitudinal data that reflect the evolution of rehospitalization risk before the application of the treatment (in our case, telemonitoring). The second phase is motivated by the main objective of this work, which is to predict the outcome of a treatment (in our case, the regular conclusion of a telemonitoring period) based on available data. In this part of the analysis the data collected in the clinical survey are enriched with information gained from the preceding clinical history of the patient, represented by the longitudinal data estimated in the first phase. The two parts of the analysis could be performed also as stand-alone, the first analyzing the risk of hospital readmission for a given time period and class of hospitalization causes (e.g. cardiovascular), the second being a supervised classification method independent from the first. Nonetheless, a problem driven interaction between the two sources of information can bring improvements in the prediction of the outcome of interest.

The paper is structured as follows. Section 2 describes the theoretical and methodological framework. For what concerns the extraction of pre-telemonitoring longitudinal information, the model for recurrent events is first introduced. Then the smoothing of cumulative hazard functions obtained by realized trajectories of the recurrent event processes is detailed, and the dimensional reduction via functional principal components is described. Finally, generalized linear models with functional covariates are presented, as they will be used for the prediction of the telemonitoring outcome. Section 3 presents in detail the motivating application, practical issues and results. Finally, section 4 contains some concluding remarks and discussion of future works.

## 2. Theoretical framework

### *2.1. Model for recurrent events*

First, we introduce the counting process model that describes the patient-specific progression of hospital readmission risk in time.

Let $(\mathcal{F}_t)_{t\in I}$ be a filtration associated to the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $I=[0,\tau]$. We define the counting process $(N(t))_{t\in I}$ adapted to $(\mathcal{F}_t)_{t\in I}$ as follows:

$$N(t)=\sum_{j=0}^{\infty} I\{S_j \le \min(t,\tau)\}, \tag{1}$$

where $S_j$ represents the calendar time of the $j$th occurrence of the observed event and $\tau$ represents a random censoring time for the process.

Under the standard assumption that $N$ is a submartingale such that the class of random variables $N(T)$, with $T$ an arbitrary stopping time, is uniformly integrable (i.e. $N$ is a *class D* submartingale), the Doob–Meyer decomposition theorem states that there exists a unique predictable, non-decreasing, *cadlag* (right-continuous with left limits) and integrable compensator (or *cumulative hazard*) process $(\Lambda(t))_{t\in I}$ such that

$$M=N-\Lambda \tag{2}$$

is a zero-mean, uniformly integrable martingale [see, for example, Andersen *et al.* (1993)]. Hence the distribution of event times is completely characterized by the knowledge of process $\Lambda$, on which modelling efforts can then be focused. We will assume that

$$\Lambda(t)=\int_0^t C(s)\lambda(s)\,\mathrm{d}s, \tag{3}$$

where $C(s)=I\{s\le\tau\}$ is the *at-risk process*, and $(\lambda(s))_{s\in I}$ is called *hazard function*, or *intensity process*.

A wide variety of models for the intensity process can be found in counting processes literature, ranging from simple Poisson processes to multiplicative hazard models (Cox, 1972), additive models, frailty and dynamic models (see for instance Andersen *et al.* (1993) and Aalen *et al.* (2006) for presentations and discussions on various possibilities). Our choice for the target problem is the following Cox-type model: for $i=1,\ldots,n$, the $i$th subject has covariate vector $\mathbf{X}_i(t)=(X_{i1}(t),\ldots,X_{iq}(t))^T$ (eventually time dependent) and the intensity is

$$\lambda(t\,|\,\mathbf{X}_i)=\lambda_0(t)\alpha^{N_i(t^-)}\exp[\boldsymbol{\beta}^T\mathbf{X}_i(t)], \tag{4}$$

where $\lambda_0(t)$ is an unknown baseline hazard function, $\alpha$ is a real parameter and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_q)^T$ a $q$-dimensional vector of real coefficients.

We choose to account for unobserved heterogeneity by using the dynamic component $\alpha^{N_i(t^-)}$ instead of a frailty variable, that is, a multiplicative random effect. Dynamic and frailty modelling can be seen as two related methods for describing subject heterogeneity, but the former is more general and flexible (see Aalen *et al.* (2006) for a discussion on these two approaches); however, results obtained in the two cases are compared in section 3.2. The dependence of intensity on process state is modelled by the term $\alpha^{N_i(t^-)}$ because of its clear interpretation: values of $\alpha$ higher than 1 indicate that a new event implies a worsening of the patient's condition, increasing future rehospitalization risk, vice versa for $\alpha$ values lower than 1. We also assume the baseline intensity $\lambda_0$ to be a function of the total time $t$, but a wide range of choices is valid within the same framework; see for example the concept of *effective age* introduced in Peña & Hollander (2004).

Adding a censoring variable to account for different observation intervals, the model for cumulative hazard can be written as follows, for patients $i=1,\ldots,n$:

$$\Lambda_i(t\,|\,\mathbf{X}_i)=\int_0^t C_i(s)\lambda_0(s)\alpha^{N_i(s^-)}\exp[\boldsymbol{\beta}^T\mathbf{X}_i(s)]\,\mathrm{d}s, \tag{5}$$

where $C_i(s)=I\{s\le\tau_i\}$, i.e. subjects have different, mutually independent censoring times $\tau_i$.

Independent censorship as defined in Kalbfleisch & Prentice (1980) can be assumed for the considered problem, as will be clear from the application described in section 3.

### 2.2. Cumulative hazard smoothing and reconstruction

Semiparametric estimation of cumulative hazard, as proposed in Peña *et al.* (2007), produces a step function estimate $\hat{\Lambda}_0$ of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$ that has the following expression: defining $t_j$ as the $j$th observed jump time of the aggregated process $N(t) = \sum_{i=1}^n N_i(t)$ and $\tau = \max_{i=1,\dots,n} \tau_i$, then

$$\hat{\Lambda}_0(t) = \sum_{t_j \le t} \frac{1}{\sum_{i=1}^n C_i(t_j)\hat{\alpha}^{N_i(t_j^-)} \exp[\hat{\boldsymbol{\beta}}^T \mathbf{X}_i(t_j)]}, \quad t \in (0, \tau],$$

where $\hat{\alpha}$ and $\hat{\beta}$ are maximum likelihood estimates of $\alpha$ and $\boldsymbol{\beta}$.

Assuming the true underlying $\Lambda_0$ function to be absolutely continuous, we deal with the issue of smoothing its estimate $\hat{\Lambda}_0$ before moving to the reconstruction of cumulative hazard process realizations for each patient. The function $\Lambda_0(t)$ has two features that we want the smoothing procedure to preserve: increasing monotonicity and $\Lambda_0(0) = 0$. A fast method for smoothing functional data while enforcing desired constraints has been proposed in He & Ng (1999) and consists in a minimum absolute deviation estimate of coefficients for a B-spline basis expansion: these are computed by minimizing the distance induced by $\|\cdot\|_1$ of the smoothed estimate evaluations $\{\tilde{\Lambda}_0(t_j)\}_{j=1,2,\dots}$ from $\{\hat{\Lambda}_0(t_j)\}_{j=1,2,\dots}$, thus obtaining the desired regularized curve $\tilde{\Lambda}_0$.

We then reconstruct the realizations of processes $\Lambda_i(t)$ for $i = 1, \dots, n$ under the chosen model, since we intend to use them as patient-specific functional data. We can rewrite the model for intensity in a form that allows to plug in $\tilde{\Lambda}_0$ instead of an estimate of $\lambda_0$. Letting $(t_1^{(i)}, \dots, t_{N_i(\tau_i)}^{(i)})$ be the jump times for patient $i$ and setting $t_0^{(i)} = 0$ and $t_{N_i(\tau_i)+1}^{(i)} = \tau_i$, we have

$$\begin{aligned}\Lambda_i(t) &= \int_0^t \lambda_0(s)\alpha^{N_i(s^-)} \exp[\boldsymbol{\beta}^T \mathbf{X}_i(s)]\,\mathrm{d}s \\ &= \sum_{k=0}^{N_i(t^-)} \mathrm{e}^{k\log\alpha} \int_{t_k^{(i)}}^{t_{k+1}^{(i)} \wedge t} \lambda_0(s) \exp[\boldsymbol{\beta}^T \mathbf{X}_i(s)]\,\mathrm{d}s, \quad i = 1, \dots, n, \end{aligned} \tag{6}$$

for $t \in [0, \tau_i)$. If the covariate vector $\mathbf{X}_i^T$ is differentiable on each interval $[t_k^{(i)}, t_{k+1}^{(i)})$ for $i = 1, \dots, n$ and $k = 1, \dots, N_i(\tau_i)$, integrating by parts it is possible to express each $\Lambda_i$ as a function of $\Lambda_0$, so that an estimate of $\lambda_0$ will not be required.

As a qualitative validation of the fitted model, we can compare the averages of counting and cumulative hazard processes: taking conditional expectations in (2) we have $\mathbb{E}[\Lambda_i(t) \,|\, \mathbf{X}_i(t)] = \mathbb{E}[N_i(t) \,|\, \mathbf{X}_i(t)]$ for $i = 1, \dots, n$. The comparison is not straightforward when curves have different censoring times, in particular if faster growing curves have higher probability of earlier censoring (this is common for risk curves, as frailer patients die earlier). In this situation the naive pointwise sample mean $\mu_n(t) = \sum_{i=1}^n \Lambda_i(t) C_i(t)/n(t)$, with $n(t) = \sum_{i=1}^n C_i(t)$, is not monotone and underestimates expected values for large times. In this case, a mean process computed by cumulation of mean increments $\tilde{\mu}_n(x_k) = \sum_{j=1}^k \sum_{i=1}^n C_i(x_j)[\Lambda_i(x_j) - \Lambda_i(x_{j-1})]/n(x_j)$, with $\{x_0, x_1, x_2, \dots\}$ any overset of $\{0, (\tau_i)_{i=1,\dots,n}\}$, is more accurate, since it enforces monotonicity by definition when all sample curves are monotone. As pointed out in Crowell (1992), this estimator is unbiased and consistent, and in the case of positively correlated increments it is likely to have lower variance with respect to the pointwise one.

### *2.3. Functional principal component analysis*

From this point, the reconstructed realizations $\tilde{\Lambda}_i(t)$ are considered as functional data, and will be later included as covariates in a generalized linear model for telemonitoring outcome prediction. Since these data are high-dimensional, a common strategy is to perform a suitable dimensional reduction, choosing only relevant components of a proper basis expansion.

Consider the functional ANOVA decomposition of data, as suggested in Müller & Stadtmüller (2005),

$$\tilde{\Lambda}_i(t) = \mu(t) + D_i(t) + \varepsilon_i(t), \quad i = 1, \ldots, n \tag{7}$$

where $\mu(t)$ is the population mean function, $D_i(t)$ is the residual for subject $i$ and $\varepsilon_i(t)$ a noise term. One of the possibilities for representing $\tilde{\Lambda}_i(t)$ is to use functional principal component analysis (FPCA), that is, the Karhunen–Loève decomposition (see for example Ferraty & Vieu (2006) for some theoretical results and Ramsay & Silverman (2005) for practical details). At this point we assume that functional data are known on a common domain $T$, thus allowing to estimate a common Karhunen–Loève basis. Once eigenfunctions $\{\psi_k\}_{k\in\mathbb{N}}$ and eigenvalues $\{v_k\}_{k\in\mathbb{N}}$ of the covariance operator for $\tilde{\Lambda}$ have been found, we express the functional ANOVA decomposition (7) in the following form

$$\tilde{\Lambda}_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \quad i = 1, \ldots, n,$$

where $\xi_{ik} = \int_T D_i(s)\psi_k(s)\,\mathrm{d}s$ is the $k$th score for subject $i$. Eigenfunction-eigenvalue couples $\{(\psi_k, v_k)\}_{k\in\mathbb{N}}$ completely explain modes of variation of data, in the sense that eigenfunctions ordered with respect to the associated eigenvalues represent orthonormal directions of decreasing residual variability. It is thus possible to represent data using just the most relevant modes of variation, represented by the first $K$ elements of $\{\psi_k\}_{k\in\mathbb{N}}$. A natural criterion for choosing the number $K$ of components to be included in the truncated representation is the following: fix a threshold $c$ for the proportion of variance they should globally describe, which is based on eigenvalues, and choose the minimum $K$ such that:

$$\frac{\sum_{k=1}^{K} v_k}{\sum_{k=1}^{m} v_k} \geq c,$$

where $m$ is the number of abscissa values on which functional data are known. We then use the approximation

$$\tilde{\Lambda}_i^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \quad i = 1, \ldots, n.$$

For the sake of simplicity, from now on we will write $\tilde{\Lambda}_i(t)$ even when its truncated basis expansion $\tilde{\Lambda}_i^K(t)$ is used.

### *2.4. Generalized linear models with functional covariates*

While the methodological elements presented up to this point regarded the extraction of longitudinal pre-telemonitoring information, the present section involves the analysis of data from clinical surveys, eventually enriched with information obtained from estimated cumulative hazard trajectories.

Let us consider a logistic regression model, where the response variable is Bernoulli distributed, $Y_i \sim B(p_i)$ for $i \in 1, \ldots, n$ and $\theta_i = \log(p_i/(1-p_i))$, while $\theta_i$ is a linear function of covariates related to subject $i$. This set of covariates is composed by data from the clinical

survey, that is, information recorded at the beginning of the telemonitoring period; we augment this dataset adding the functions $D_i(t)$, $i=1,\ldots,n$ introduced in section 2.3, so that $\theta_i$, $i=1,\ldots,n$, assumes the (truncated) form

$$\theta_i = \int_T D_i(t)\delta(t)\,\mathrm{d}t + \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\gamma} \approx \int_T \delta(t)\sum_{k=1}^{K}\zeta_{ik}\psi_k(t)\,\mathrm{d}t + \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\gamma},$$

where $\delta : T \mapsto \mathbb{R}$ is a functional parameter, $\boldsymbol{\gamma} \in \mathbb{R}^q$ is a vector of parameters to be estimated and $\mathbf{w}_i \in \mathbb{R}^q$, for $i=1,\ldots,n$, is a vector of covariates available from the clinical survey. If also $\delta(\cdot)$ is represented using the principal components basis, that is, $\delta(t)=\sum_{j=1}^{K}\delta_j\psi_j(t)$, for the orthonormality of $\{\psi_k\}_{k\in\mathbb{N}}$ we obtain

$$\theta_i = \sum_{k=1}^{K}\zeta_{ik}\delta_k + \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\gamma}, \quad i=1,\ldots,n.$$

In the end, the model is reduced to a classical logistic regression, in which the unknowns are represented by the parameter vector $(\delta_1,\ldots,\delta_K,\gamma_1,\ldots,\gamma_q)$. The same approach can be extended without further difficulties to the more general context of generalized linear models with different responses and link functions.

## 3. Application to telemonitoring data analysis and results

In Lombardia region an experimentation of heart failure telemonitoring started in 2003, involving 34 health care institutions (see CEFRIEL (2010) for an overview of programme and protocols). Four studies (*Criteria*, *Piano Urbano*, *Nuove Reti Sanitarie* and *Telemaco*) were devoted to collect, under prior informed consent, information about telemonitoring periods, then gathered in a comprehensive database. Each record of this database refers to a single telemonitoring period, and contains anagraphical data of the involved patient, number of transmitted electrocardiograms, diagnosis and disease etiology at the last hospital admission and other relevant clinical quantities.

The enrolment protocol adopted during the period 2004–8 includes adult citizens of Lombardia with a NYHA class (describing the severity of heart disease) of III or IV who experienced at least one hospitalization for heart failure during the 6 months before the beginning of telemonitoring. The telemonitoring period is planned for a 180-day duration, with possible re-enrolment under particular conditions. The period may be interrupted, by protocol, if a hospitalization lasting more than 8 days occurs or because of surgical intervention; we considered these two cases and decease as 'negative' conclusions of the treatment, while other types of 'external' events that forced interruption, such as a change of dwelling or the decision by the patient herself/himself to stop the therapy, have been considered as drop-outs.

Since data regarding telemonitoring periods have a limited scope, we requested an interrogation of the regional administrative PHD, to obtain hospital discharge data (SDOs) stored during the five years of interest. Each one of these records contains extensive information about a single hospitalization, such as date, duration, drugs received, DRG (Diagnosis Related Group, a classification code for patients discharged from hospital, based on the type of resources used during the stay) and other data of clinical interest. This information is used to estimate pre-telemonitoring hazard functions, so the subset of records regarding just pre-telemonitoring hospital admissions for each patient has been extracted from the full database. Each one of the available subjects is identified by a unique code, derived from an anonymizing procedure applied to her/his identity number, and used to retrieve from the SDO database her/his hospitalization history in the period 2004–8. The linkage and matching of

information between these two databases resulted in the constitution of an initial sample of 1081 patients. A reduction of this dataset was operated in the second part of the analysis, to include only subjects whose telemonitoring period started at least in 2006. In this way, a 2-year time window before telemonitoring is available for all of them, ensuring better background for predictive tasks.

The risk of hospital readmission is obviously zero during each hospital stay, which typically lasts some days (mean duration = $17.18 \pm 28.03$ days). We deal with this issue by removing the hospital stay period from the process time count and by counting just once consecutive hospitalization periods that were eventually registered as multiple in the database because of a patient's transfer to a different structure or for any other reason. The time variable is expressed in days passed from 1st January 2004.

The following analyses have been carried out using the statistical software R (R Development Core Team, 2009). For hazard estimation packages `gcmrec` (González *et al.*, 2009) and `frailtypack` (González *et al.*, 2010) have been used, while package cobs (Ng & Maechler, 2009) has been used for constrained smoothing.

### *3.1. Hazard estimation and dimensional reduction*

The first step of the analysis consists in the estimation of model (5) for cumulative hazard functions, using the procedure explained in section 2.2.

The beginning of telemonitoring is introduced as a censoring time $\tau_i$, $i = 1, \ldots, n$, for the hospitalization counting processes, assuming that this event does not influence preceding hospitalizations; this assumption seems reasonable, on the basis of the enrolment protocol.

Following medical advice, subject age is included as covariate $X_i(s)$ in (5), providing the following model for patients $i = 1, \ldots, n$

$$\Lambda_i(t \mid \mathbf{X}_i) = \int_0^t C_i(s)\lambda_0(s)\alpha^{N_i(s^-)} \exp[\beta X_i(s)]\,\mathrm{d}s,$$

with $C_i(s) = I\{s \leq \tau_i\}$. Including age at $t = 0$ as a time-independent covariate would yield similar results in the following, but we include it as a continuous variable, because hospital admissions occur for each patients at different distances from the beginning of observation time. Moreover, this choice is coherent with the assumption that the influence of a covariate on hazard does not depend on the particular time instant when we start observing the subject.

Estimated baseline cumulative hazard $\hat{\Lambda}_0(t)$ is represented in Fig. 1 (dashed line), while parameter estimates are shown in Table 1. We notice that parameter $\alpha$, describing the effect of a new event on the risk of future hospitalizations, is significantly higher than 1, according to a one-sided hypothesis test; this means that a new event represents an increase in rehospitalization risk. Parameter $\beta$, related to the age covariate, is surprisingly negative, indicating that the risk of rehospitalization is slightly lower for older patients; this could be explained by the fact that in the old population considered (the mean age is $67.82 \pm 11.19$) subjects who survived up to a higher age are the less frail ones.

Since the non-parametric estimator used for $\Lambda_0(t)$ produces a step function, we perform a smoothing of this estimate with the method proposed in section 2.2; for the B-Spline basis, we choose polynomial order 2 and 20 equally spaced knots. A comparison between the non-parametric estimate and the B-spline smoothed estimate is shown in Fig. 1. In this picture we also notice that the cumulative baseline hazard function $\hat{\Lambda}_0(t)$ has a convex behaviour, describing a gradual increase of instantaneous risk due to the disease, common to the whole population.

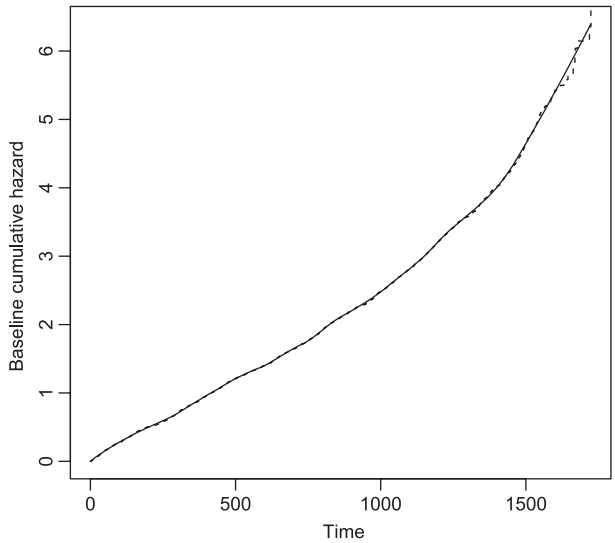


*Fig. 1.* Results of baseline cumulative hazard function estimation: $\hat{\Lambda}_0$ (dashed line) and its smoothed version $\tilde{\Lambda}_0$ (solid line).

Table 1. *Results of hazard parameters estimation*

| | Estimate | Std. dev. | *p*-value |
|---|---|---|---|
| $\alpha$ | 1.21 | 0.00887 | $<2\cdot10^{-16}$* |
| $\beta$ | −0.00336 | 0.00172 | 0.051 |

*p*-value * refers to a test with null hypothesis $\alpha \leq 1$. Both tests are carried out with a normal approximation for maximum likelihood estimators.

Once $\tilde{\Lambda}_0$ has been computed, we can reconstruct individual cumulative hazard processes, letting $\mathbf{X}_i(t) = X_i(t)$, representing the age of patient $i$. Plugging $(X_i(t), N_i(t), \tau_i, \hat{\alpha}, \hat{\beta})$ for $i = 1, \ldots, n$ in (6) gives the reconstruction of cumulative hazard processes for all the considered patients, shown in Fig. 2A. To verify that the condition $\mathbb{E}[\Lambda_i(t) \,|\, X_i(t)] = \mathbb{E}[N_i(t) \,|\, X_i(t)]$ holds, it is possible to visualize the mean functions of point processes and of cumulative hazard processes, computed using the cumulative mean increment estimator $\tilde{\mu}_n$ suggested in section 2.2. To address the problem of computing this conditional expectation, we can split the sample in classes $A_{c_1}, A_{c_2}, \ldots$ of similar initial age and average on these classes to approximate $\mathbb{E}[\Lambda_i(t) \,|\, X_i(t)]$ and $\mathbb{E}[N_i(t) \,|\, X_i(t)]$. For example, the martingale residual trajectories and their average for subjects belonging to the age class $A_{60} = \{i : a_i \in (55, 65]\}$ are shown in Fig. 2B; we can see that residuals $\hat{M}_i(t) = N_i(t) - \tilde{\Lambda}_i(t)$, $i \in A_{60}$, seem to have the expected behaviour.

Figure 3 shows a comparison between average curves computed using pointwise estimator $\mu_n$ and estimator $\tilde{\mu}_n$, respectively; in the left panel we notice that average curves estimated with $\mu_n$ are non-monotone and heavily biased because of right censoring, while average curves estimated with $\tilde{\mu}_n$, depicted in the right panel, suffer from censoring only at the very right end of the domain.

A dimensional reduction of the functional dataset of cumulative hazard functions is operated via principal component analysis, so that FPC scores can be used as indicators of pre-telemonitoring risk status and be added to other variables retrieved from the telemonitoring

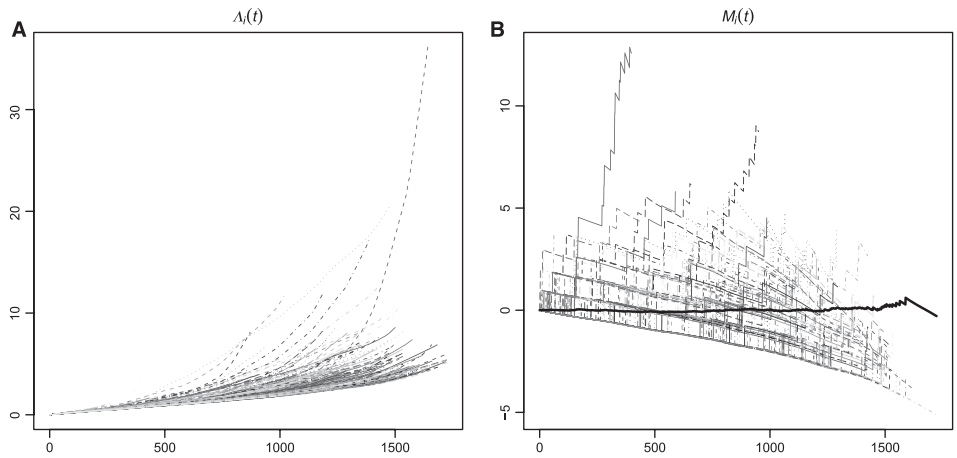


*Fig. 2.* Estimated trajectories and martingale residuals. (A) Reconstructed realizations of cumulative hazard processes. (B) Trajectories of residuals $\hat{M}_i(t) = N_i(t) - \tilde{\Lambda}_i(t)$, $i \in A_{60}$, and their average (thick line).

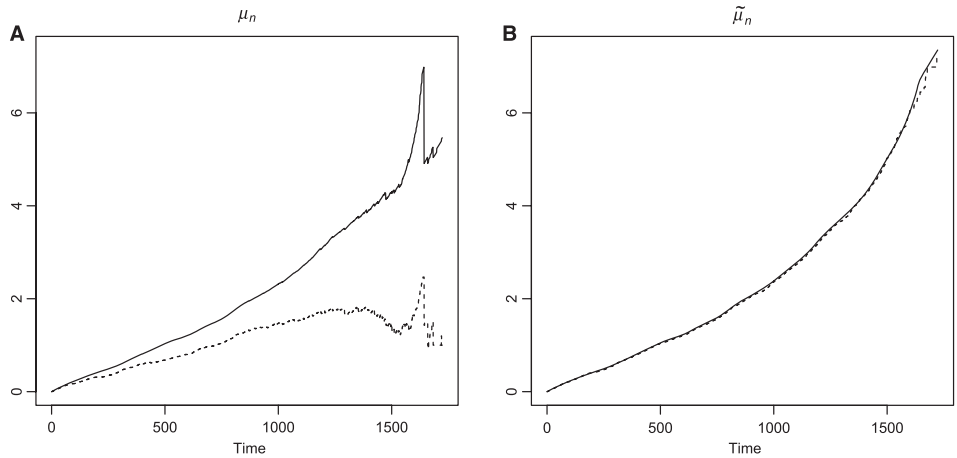


*Fig. 3.* Average curves of counting process data (dashed lines) and of reconstructed cumulative hazard functions (solid lines) for age class $A_{60}$. (A) Mean curves obtained with the pointwise estimator $\mu_n$. (B) Mean curves obtained with estimator $\tilde{\mu}_n$.

survey in the subsequent logistic regression model. To avoid the problem of censoring, as previously mentioned, we choose patients for which at least 2 years of clinical history before telemonitoring are available in our records, and restrict the time window for our analyses to exactly the 2 years preceding telemonitoring. Doing so, we obtain a dataset of $n = 747$ curves, evaluated on a grid of length $m = 730$ (hazard functions were computed on a uniform time grid with daily spacing).

Before proceeding to principal component analysis, curves are centred by subtracting their mean function $\mu_n(t)$ (which coincides with estimator $\tilde{\mu}_n(t)$ on the chosen subset of data); moreover, the noise term $\varepsilon_i(t)$ is discarded, since curves have already been estimated with smoothness.

We shall now select the components to be considered in the subsequent analysis. A simple and effective criterion consists in choosing the first $K$ components, such that their associated

Table 2. *First $K=2$ eigenvalues obtained with FPCA*

| | $v_1$ | $v_2$ |
|---|---|---|
| Value | 777.04 | 45.64 |
| % variance | 94.08 | 5.53 |
| Cum. % variance | 94.08 | 99.60 |

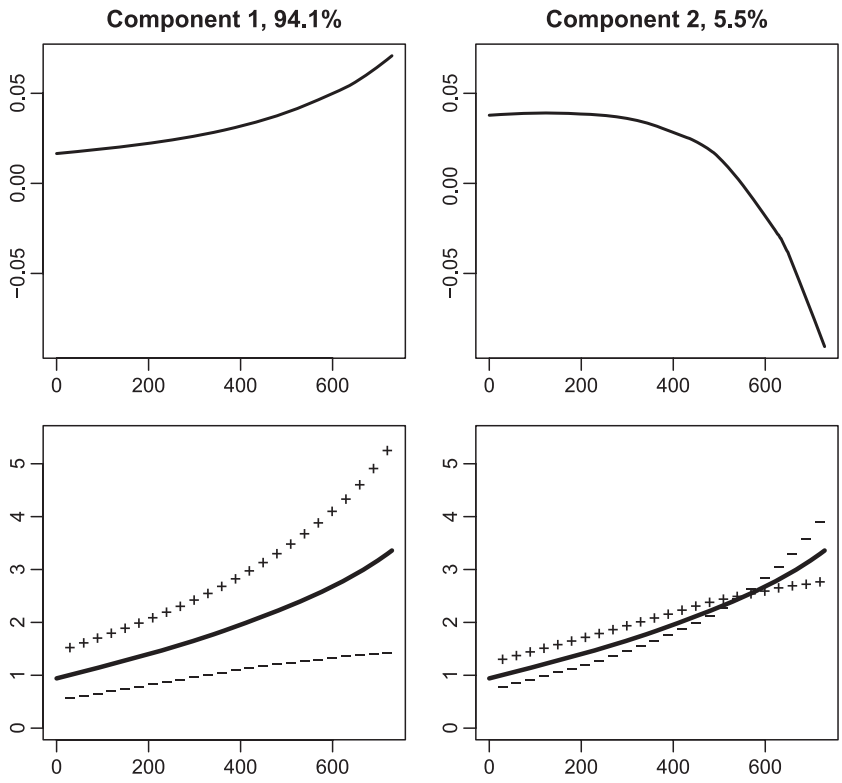


*Fig. 4.* In the upper panels, first $K=2$ eigenfunctions obtained with FPCA; in the lower panels, representation of $\mu_n(t)$ (solid line) and $\mu_n(t) \pm v_k \phi_k(t)$ ('+' or '−', respectively), $k=1,2$.

eigenvalues explain a proportion of variance $c>95$ per cent. This criterion leads to the choice of the first $K=2$ components, as detailed in Table 2.

Figure 4 shows the two relevant functional principal components in the top panels and $\mu_n(t) \pm v_k \phi_k(t)$, $k=1,2$ in the bottom panels. The first component is monotone increasing, and is highly dominant in the description of data curves, while the second one is decreasing, characterizing curves that do not grow very fast also on a long time period.

### *3.2. Generalized linear model estimation*

The final goal of this work is the prediction of telemonitoring outcome, defined as a binary variable with value 1 if telemonitoring has regular conclusion and 0 if the period is terminated by an adverse event, that is, hospitalization, surgical intervention or decease. The scores of principal components 1 and 2 of individual hazard functions are considered as covariates that summarize the features of the patients' pre-telemonitoring clinical history, and they

are used to predict telemonitoring outcome together with a subset of the variables present in the telemonitoring database. In particular, clinicians suggested to take into account diagnosis and etiology of the last hospitalization before telemonitoring and sex. Variable diagnosis has 3 levels (*congestive*, *left* or *unspecified* heart failure), while etiology has 5 levels (*hypertensive*, *hyschaemic*, *primary*, *valve*, *other*). Following the notation introduced in section 2.4, the probability of normal outcome $p_i$ takes the following form

$$p_i = \frac{\exp\left(\sum_{k=1}^{2} \zeta_{ik}\delta_k + \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)}{1+\exp\left(\sum_{k=1}^{2} \zeta_{ik}\delta_k + \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)}, \quad \text{for } i=1,\ldots,n,$$

where $p_i = \mathbb{E}(Y_i \mid \mathbf{w}_i, \boldsymbol{\zeta}_{i1}, \boldsymbol{\zeta}_{i2})$ and $\mathbf{w}_i$ is composed by dummy variables representing cathegorical covariates sex, diagnosis and etiology.

The model output is reported in Table 3. Scores 1 and 2 are both significant, and their signs are coherent with a possible interpretation: principal component 1 is an increasing function, so a larger score, which represents a steeper cumulative hazard process, implies a lower probability of regular conclusion; component 2, instead, is decreasing, and its estimated coefficient has opposite sign, indicating that patients who have lower cumulative hazard for long times have a higher probability of normal conclusion of the telemonitoring period. Also, we can notice a slight dependence on etiology; in particular, valvular etiology seems to increase the probability of early conclusion of telemonitoring caused by an adverse event. Instead, there is no significant difference in the probability of adverse events among either men and women, nor among subjects with different types of diagnoses.

We shall now compare the approach proposed in this work with a frailty model for hazard functions, in which estimated frailties are used as predictors for the logistic regression. The shared gamma frailty model is formulated as

$$\lambda(t \mid \mathbf{X}_i, Z_i) = Z_i \lambda_0(t) \alpha^{N_i(t^-)} \exp\left[\boldsymbol{\beta}^T \mathbf{X}_i(t)\right],$$

where $Z_1,\ldots,Z_n$ are i.i.d. from a Gamma$(1/\theta, 1/\theta)$ distribution; notice that $\theta$ represents the frailty variance, so that in the limit $\theta \to 0$ the variable is constant and equal to 1. Since we are interested in individual hazard processes, a full likelihood approach, such as the one proposed in Rondeau *et al.* (2003), is not viable, because random effects are integrated out and only population parameters are estimated. Anyway, this approach is useful to assess the significance of random effects, while a computationally heavier EM algorithm can be used to estimate individual $z_i$s. We used the R package `frailtypack` to check the significance of the parameters in the frailty model, while the estimates of the $z_i$s have been obtained with

Table 3. *Estimates, standard errors and p-values for the parameters of the logistic regression with FPCA scores*

| Parameter | Estimate | Std. error | *p*-value |
|---|---|---|---|
| $\gamma_0$ (Intercept) | 14.8108 | 437.0832 | 0.9730 |
| $\gamma_1$ (Sex) | 0.1557 | 0.2167 | 0.4726 |
| $\gamma_2$ (Etiology – Hypertensive) | 0.0669 | 0.4469 | 0.8810 |
| $\gamma_3$ (Etiology – Hyschaemic) | −0.0187 | 0.2506 | 0.9405 |
| $\gamma_4$ (Etiology – Primary) | −0.0819 | 0.3199 | 0.7981 |
| $\gamma_5$ (Etiology – Valve) | −0.8867 | 0.4673 | 0.0790 |
| $\gamma_6$ (Etiology – Other) | −0.6599 | 0.3593 | 0.1248 |
| $\gamma_7$ (Diagnosis – Congestive) | −13.8204 | 437.0832 | 0.9748 |
| $\gamma_8$ (Diagnosis – Left) | −13.1587 | 437.0832 | 0.9760 |
| $\gamma_9$ (Diagnosis – Unspecified) | −13.6343 | 437.0833 | 0.9751 |
| $\delta_1$ (FPCA score 1) | −0.0144 | 0.0039 | 0.0003 |
| $\delta_2$ (FPCA score 2) | 0.0567 | 0.020490 | 0.0056 |

Table 4. *Estimates, standard errors and p-values for the parameters of the logistic regression with estimated frailties*

| Parameter | Estimate | Std. error | $p$-value |
|---|---|---|---|
| $\gamma_0$ (Intercept) | 15.2024 | 436.2230 | 0.9722 |
| $\gamma_1$ (Sex) | 0.1513 | 0.2160 | 0.4833 |
| $\gamma_2$ (Etiology – Hypertensive) | 0.15180 | 0.4460 | 0.7336 |
| $\gamma_3$ (Etiology – Hyschaemic) | 0.1492 | 0.2503 | 0.5511 |
| $\gamma_4$ (Etiology – Primary) | −0.0452 | 0.3084 | 0.8834 |
| $\gamma_5$ (Etiology – Valve) | −0.9230 | 0.4657 | 0.0474 |
| $\gamma_6$ (Etiology – Other) | −0.4346 | 0.3624 | 0.2305 |
| $\gamma_7$ (Diagnosis – Congestive) | −13.7154 | 436.2230 | 0.9749 |
| $\gamma_8$ (Diagnosis – Left) | −13.0917 | 436.2230 | 0.9761 |
| $\gamma_9$ (Diagnosis – Unspecified) | −13.688 | 436.2230 | 0.9750 |
| $Z$ (Frailty) | −0.6865 | 0.1769 | 0.0001 |

Table 5. *Comparison between the logistic regression model with FPCA scores* ($\theta=0$) *and with estimated frailties* ($\theta=0.68$): *AIC, Brier score and cross-validation error (MSE)*

| Parameter | AIC | Brier score | CV error |
|---|---|---|---|
| $\theta=0$ | 688.6547 | 0.1614153 | 0.1675431 |
| $\theta=0.68$ | 691.8656 | 0.1630643 | 0.1688114 |

package `gcmrec`, using the EM algorithm presented in Peña *et al.* (2007). When the multiplicative frailty component is considered, $\theta \simeq 0.68$ with a significant difference from 0 (a Wald test gives a $p$-value near 0), while age and process state $N_i(t)$ in this case are not significant ($p$-values $>10\%$). On this basis, we fit the model without covariates using the EM algorithm implemented in package `gcmrec`, and use the estimated frailty realizations $z_1, \ldots, z_n$ as predictors in a logistic regression model; the output of this model is presented in Table 4.

Table 5 shows a comparison between the logistic regression fits with FPC scores and estimated frailty realizations. Notice that the model with FPC scores has lower AIC with respect to the model with frailties, while the lower Brier score denotes a better fit to data. Moreover, a lower leave-one-out cross-validation error, computed using a mean square error cost function, denotes a higher predictive power for the first model.

The fact that by using FPC scores related to dominant principal components we potentially introduce many covariates for the logistic regression, while the frailty approach produces only one predictor, should be considered as an additional feature of the FPC approach. Estimated frailty realizations represent just an amplification factor for population hazard, while the functional principal component analysis is capable of catching arbitrarily complex behaviours of intensity in time. In this case, the second score describes a late stabilization of the patient's health, a phenomenon that is not captured by the frailty variable. The estimation process for the shared Gamma frailty model with respect to the estimation process for the first model, despite the FPCA step, is also computationally heavier.

## 4. Concluding remarks

In this work, a novel approach to the analysis of telemonitoring data has been proposed, aimed at getting deeper insight on the patient's health status using data from clinical registries and PHD. The presented methodology, involving database integration, counting process modelling of critical events and generalized linear models, can be applied to the study of many different pathologies, thanks to its capability of dealing with complex data.

The counting process model is a natural way of representing the occurrence of hospitalizations in time, and enables us to include a large piece of information contained in the PHD to describe the clinical history of a patient. The model used is very general and allows to describe complex dynamics in an easily interpretable form.

The obtained trajectories are thus studied as functional covariates in the framework of generalized linear models with functional covariates, which offers a powerful tool to analyze dependencies and to perform classification and prediction in a wide range of applications. Using the FPCA it is possible to perform dimensional reduction of functional data, allowing to use well established methods for GLM estimation and inference in a multivariate setting and borrowing strength from both techniques.

The obtained results could have an impact on the planning of this care strategy and provide support to the decision of allocating a patient to telemonitoring, basing on her/his probability of concluding it regularly. Further development of this framework in cooperation with medical staff will include the selection and use of various different time dependent and independent variables to study telemonitoring effectiveness on quality of life, mortality and costs, and could lead to the definition of a useful tool for health care assessment and treatment planning.

## 5. Acknowledgements

This work is part of the experimental home telemonitoring programme for chronic heart failure 'Nuove Reti Sanitarie – PTS', promoted by Regione Lombardia. The authors wish to thank the staffs from Regione Lombardia, CEFRIEL and the Department of Management, Economics and Industrial Engineering of Politecnico di Milano and Dr. Amerigo Giordano.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

The R scripts **GLMHF_Supp.R** and **GLMHF_funcs_Supp.R** for executing the whole procedure described in the article, as well as artificial SDO and telemonitoring R binary datasets **PHD_data_Supp.RData** and **survey_data_Supp.RData** for testing purposes, are available as supporting material.

## References

Aalen, O. O., Borgan, Ø. & Gjessing, H. K. (2006). *Survival and event history analysis: a process point of view*. Springer-Verlag, New York.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

Capomolla, S., Pinna, G., La Rovere, M. T., Maestri, R., Ceresa, M., Ferrari, M., Febo, O., Caporotondi, A., Guazzotti, G., Lenta, F., Baldin, S., Mortara, A. & Cobelli, F. (2004). Heart failure case disease management program: a pilot study of home telemonitoring versus usual care. *Eur. Heart J. Suppl.* **6 (Supplement F)** F91–F98.

CEFRIEL (2010). http://ftp.cefriel.it/nrs/.

Cox, D. R. (1972). Regression models and life tables. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **34**, 187–220.

Crowell, J. I. (1992). Nonparametric estimation of a process mean from censored data. *Stat. Probab. Lett.* **15**, 253–257.

Ferraty, F. & Vieu, P. (2006). *Nonparametric functional data analysis*. Springer-Verlag, New York.

Giordano, A., Scalvini, S., Zanelli, E., Corrà, U., Longobardi, G. L., Ricci, V. A., Baiardi, P. & Glisenti, F. (2008). Multicenter randomised trial on home-based telemanagement to prevent hospital readmission of patients with chronic heart failure. *Int. J. of Cardiol.* **131**, 192–199.

González, J. R., Peña, E. A. & Slate, E. H. (2005). Modelling intervention effects after cancer relapses. *Stat. Med.* **24**, 3959–3975.

González, J. R., Rondeau, V. & Mazroui, Y. (2010). *frailtypack: Frailty models using maximum penalized likelihood estimation*. Avaliable on http://cran.r-project.org/web/packages/frailtypack/index.html.

González, J. R., Slate, E. H. & Peña, E. A. (2009). *gcmrec: General class of models for recurrent event data*. Available on http://cran.r-project.org/web/packages/gcmrec/index.html.

He, X. & Ng, P. T. (1999). Cobs: Qualitatively constrained smoothing via linear programming. *Comput. Statist.* **14**, 315–337.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure data*. John Wiley & Sons, New York.

Müller, H.-G. & Stadtmüller, U. (2005). Generalized Functional Linear Models. *Ann. Stat.* **33**, 774–805.

Ng, P. T. & Maechler, M. (2009). *cobs: Cobs – constrained B-splines (sparse matrix based)*. Available on http://cran.r-project.org/web/packages/cobs/index.html.

Peña, E. A. & Hollander, M. (2004). Models for recurrent events in reliability and survival analysis. In *Mathematical reliability: An expository perspective*, chap. 6. (eds R. Soyer, T. Mazzuchi & N. Singpurwalla) 493–514. Kluwer Academic Publishers, Dordrecht.

Peña, E. A., Slate, E. H. & González, J. R. (2007). Semiparametric inference for a general class of models for recurrent events. *J. Stat. Plan. Infer.* **137**, 1727–1747.

R Development Core Team (2009). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Available on http://cran.r-project.org/.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional data analysis*. Springer Science+Business Media, New York.

Rondeau, V., Commenges, D. & Joly, P. (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data. Anal.* **9**, 139–153.

Scalvini, S., Zanelli, E., Volterrani, M., Martinelli, G., Baratti, D., Buscaya, O., Baiardi, P., Glisenti, F. & Giordano, A. (2004). A pilot study of nurse-led, home-based telecardiology for patients with chronic heart failure. *J. Telemed. Telecare* **10**, 113–117.

*Received December 2010, in final form September 2012*

Stefano Baraldo, Politecnico di Milano, Dipartimento di Matematica 'F. Brioschi', piazza Leonardo Da Vinci 32, 20133 Milano, Italy.
E-mail: stefano1.baraldo@mail.polimi.it