Imputing missing distributions by LQD transformation and RKHS-based functional regression

Zhicheng Chen*

ABSTRACT

Data loss is a big problem in many online monitoring systems due to various reasons. Copula-based approaches are effective imputation methods for missing data imputation; however, such methods are highly dependent on a reliable distribution of missing data. This article proposed a functional regression approach for missing probability density function (PDF) imputation. PDFs are first transformed to a Hilbert space by the log quantile density (LQD) transformation. The transformed results of the response PDFs are approximated by the truncated Karhunen–Loève representation. Corresponding representation in the Hilbert space of a missing PDF is estimated by a vector-on-function regression model in reproducing kernel Hilbert space (RKHS), then mapping back to the density space by the inverse LQD transformation to obtain an imputation for the missing PDF. To address errors caused by the numerical integration in the inverse LQD transformation, original PDFs are aided by a PDF of uniform distribution. The effect of the added uniform distribution in the imputed result of a missing PDF can be separated by the warping function-based PDF estimation technique.

Keywords: missing data, probability density function, log quantile density transformation, functional regression, reproducing kernel Hilbert space.

1. Introduction

For a variety of reasons, data missing is a very common phenomenon in many online monitoring systems. When the amount of missing data is huge or certain applications require the full data, imputing missing values is very meaningful. Monitoring data collected by different sensors are usually correlated, therefore, harnessing correlations to impute missing data is a promising direction. Copula-based imputation methods are effective imputation methods for missing random values by harnessing probability distributions and correlations [1-6]. However, copula-based methods are highly dependent on reliable distributions of missing data. To reduce errors caused by imputation models in a copula-based approach, imputing missing distributions intelligently by reliable distribution learning technique is more preferable. Inspired by the newly developed log quantile density (LQD) transformation [7] and RKHS-based functional regression [8], this article proposed a new regression-based approach for missing distribution imputation.

2. Technical Background

The log quantile density (LQD) transformation is proposed by Petersen and Müller in 2016 [7]. Given a PDF $f(x)$ with support on $[0,1]$, the LQD transformation is mapping the PDF to

* PhD student; Email: 13B933002@hit.edu.cn; research area: structural health monitoring.

a Hilbert space by the following functional transformation

$$\psi(t) = \log(q(t)) = -\log\left\{ f\left(Q(t)\right)\right\}, \quad t \in [0,1] \tag{1}$$

where, $Q = F^{-1}$ is the quantile function of the PDF $f(x)$, i.e., the inverse function of the corresponding cumulative distribution function $F(x) = \int_{-\infty}^{x} f(\tau)d\tau$, $q(t)$ is the quantile density function, i.e., $q(t) = Q'(t) = \frac{d}{dt}F^{-1}(t) = \left[f\left(Q(t)\right)\right]^{-1}$. The transformed result $\psi(t)$ is an ordinary function and do not need to satisfy corresponding constraints of a PDF (i.e., nonnegative and integrating to one), thus general functional regression methods can be applied to the transformed results of distributions.

The transformed result $\psi(t)$ can be mapped back to the original density space by the following inverse LQD transformation

$$f(x) = \theta_\psi \exp\left\{-\psi\left(F(x)\right)\right\}, \quad F^{-1}(t) = \theta_\psi^{-1}\int_0^t e^{\psi(s)}ds \tag{2}$$

where $\theta_\psi = \int_0^1 e^{\psi(s)}ds$.

For more detailed discussion, readers are referred to [7].

3. Problem Formulation

In this section, we formalize the problem to be addressed in this article. Specifically, suppose we have obtained n pairs of correlated PDFs, i.e., $\{g_i, f_i\}_{i=1}^{n}$. Given a new PDF g_0, suppose the corresponding PDF f_0 is missing (due to no valid observed samples for density estimation), the task in this article is to develop an imputation method to obtain a substituted PDF to replace f_0 by harnessing the information of available PDFs, i.e., $\{g_i, f_i\}_{i=1}^{n} \cup g_0$.

All investigated PDFs in this study are assumed to be one-dimensional continuous PDFs with strictly positive support on $[0,1]$, distributions with general finite supports can be easily tackled by the scale transformation introduced in [9].

4. LQD Transformation and FPCA

Functional regression methods have good potential in dealing with the problem of missing curve imputation; however, general functional regression methods cannot be directly used in missing PDF imputation, because PDFs are special functions with the constraints of nonnegative and integrating to one. The newly developed log quantile density (LQD) transformation [7] provides a particularly promising approach to transform PDFs to ordinary functions by an invertible map; however, the numerical integration in the inverse LQD transformation (see Eq. (2)) may introduce significant errors for a PDF taken values approximately equal to zero in the interval $[0,\delta](\delta > 0)$. To illustrate this, consider a PDF $f(x)$ with finite support on $[0,\ 1]$, suppose $f(x)=\varepsilon$ (ε is a very small number that approximately equals to zero) when $x \in [0,\delta]$, from Eq. (1), it can be seen $f\left(Q(t)\right)$ is also near zero within the start interval $t \in [0,\delta']$, thus the corresponding transformed result by the log function tends to infinity, i.e., $\psi(t) \to \infty$ when $t \in [0, \delta']$, therefore, the integral in Eq. (2) calculated by numerical integration methods may introduce significant errors. To address this problem, the original PDF is pre-processed by adding a PDF of the uniform distribution on $[0,\ 1]$, i.e.,

$$f^*(x) = \alpha f(x) + (1-\alpha)u(x), \quad 0.5 \le \alpha < 0.9 \tag{3}$$

where, α is the combination coefficient, $u(x)$ is the PDF of the uniform distribution on $[0,\ 1]$, i.e.,

$$u(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & otherwise \end{cases} \tag{4}$$

The original PDF $f(x)$ can be approximately recovered from $f^*(x)$ by the newly developed warping function-based density estimation technique [9] aided by a known auxiliary PDF $h(x)$ with support on $[0,1]$, i.e.,

$$\hat{\gamma} = \underset{\gamma \in \Gamma}{\arg\min} \left\{ \int_0^1 \left| \alpha h(\gamma(\tau))\dot{\gamma}(\tau) + (1-\alpha)u(\tau) - f^*(\tau) \right| d\tau \right\} \tag{5a}$$

$$f(x) \approx h(\hat{\gamma}(x))\dot{\hat{\gamma}}(x) \tag{5b}$$

where, γ is the warping function used to transform the PDF $h(x)$ to get close to the target PDF $f(x)$, Γ is the set of all valid warping functions for one-dimensional continuous distributions with support on $[0,1]$, i.e.,

$$\Gamma = \left\{ \gamma \,\middle|\, \gamma \textit{ is invertible}, \gamma(0) = 0, \gamma(1) = 1, \gamma \textit{ is smooth}, \ \gamma^{-1} \textit{ is smooth} \right\} \tag{6}$$

For detailed discussion of warping function-based density estimator, readers are referred to [9].

After the aforementioned pretreatment, all PDFs are transformed to a Hilbert space by the LQD transformation (see Eq. (1)), i.e.,

$$\psi_i^{f^*}(t) = -\log\left\{ f_i^*\left(Q_{f,i}^*(t)\right) \right\}, \ i = 1,2,\cdots,n \tag{7a}$$

$$\psi_i^{g^*}(t) = -\log\left\{ g_i^*\left(Q_{g,i}^*(t)\right) \right\}, \ i = 0,1,\cdots,n \tag{7b}$$

where, $f_i^* = \alpha f_i + (1-\alpha)u$, $g_i^* = \alpha g_i + (1-\alpha)u$, $Q_{f,i}^*$ and $Q_{g,i}^*$ are quantile functions of f_i^* and g_i^*, respectively.

The functional principal component analysis (FPCA) is applied to reduce the dimensionality of $\left\{ \psi_i^{f^*}(t) \right\}_{i=1}^n$ (reasons for such treatment will detailed in the end of section 5). For detailed discussion of the FPCA technique, readers are referred to [7, 10]. In the FPCA framework, $\psi_i^{f^*}(t)$ can be approximated by the truncated Karhunen–Loève representation, i.e.,

$$\psi_i^{f^*}(t) \approx \mu_{\psi^{f^*}}(t) + \sum_{j=1}^m \xi_i^j \phi_j(t), \quad i = 1,2,\cdots,n \tag{8}$$

where, $\left\{ \phi_j(t) \right\}_{j=1}^m$ are eigenfunctions, $\mu_{\psi^{f^*}}$ is the estimated mean function of $\left\{ \psi_i^{f^*}(t) \right\}_{i=1}^n$, i.e.,

$\mu_{\psi^{f^*}} = \frac{1}{n}\sum_{i=1}^{n}\psi_i^{f^*}(t)$, $\{\xi_i^j\}_{j=1}^m$ are PFC scores given by $\xi_i^j = \int_0^1 \left(\psi_i^{f^*}(\tau) - \mu_{\psi^{f^*}}(\tau)\right)\phi_j(\tau)d\tau$, $j = 1, 2, \cdots, m$.

The feature of the function $\psi_i^{f^*}(t)$ can be characterized by the feature vector $\boldsymbol{\xi}_i = \left[\xi_i^1 \ \xi_i^2 \cdots \xi_i^m\right]$. By such treatment, the information of available PDFs $\{g_i, f_i\}_{i=1}^n \cup g_0$ are transformed to

$$\begin{Bmatrix} \boldsymbol{\xi}_1 \\ \psi_1^{g^*}(t) \end{Bmatrix}, \begin{Bmatrix} \boldsymbol{\xi}_2 \\ \psi_2^{g^*}(t) \end{Bmatrix}, \cdots, \begin{Bmatrix} \boldsymbol{\xi}_n \\ \psi_n^{g^*}(t) \end{Bmatrix}, \begin{Bmatrix} \text{missing} \\ \psi_0^{g^*}(t) \end{Bmatrix} \tag{9}$$

5. Vector-on-Function Regression Model

The remaining task is to develop a vector-on-function regression model for the structured data set in Eq. (9), which takes the form

$$\boldsymbol{\xi}_i = F_{reg}\left(\psi_i^{g^*}(t)\right) + \varepsilon_i, \ i = 0, 1, 2, \cdots, n \tag{10}$$

where, F_{reg} is the function-to-vector map that need to be estimated, ε_i is the error term.

Nonparametric regression in reproducing kernel Hilbert space (RKHS) provides a general nonlinear regression framework for various types of data (e.g., real numbers, vectors, functions, etc.). The RKHS-based vector-on-function regression model can be developed in a similar way as the general RKHS-based function-on-function regression model proposed by Lian [8]. In the RKHS framework, the regression function F_{reg} is solved by the following penalized minimization problem

$$\min_{F_{reg}\in \mathrm{H}} \sum_{i=1}^{n} \left\| \boldsymbol{\xi}_i - F_{reg}\left(\psi_i^{g^*}\right) \right\|_2^2 + \lambda \left\| F_{reg} \right\|_{\mathrm{H}} \tag{11}$$

where, $\|\cdot\|_2$ is the two-norm of vectors, $\|\cdot\|_{\mathrm{H}}$ is the norm defined in the reproducing kernel Hilbert space (for detailed discussion, readers are referred to [8]), $\lambda > 0$ is the smoothing parameter. According to the representer theorem, the solution of the above minimization problem takes the form

$$\hat{F}_{reg}\left(\psi_i^{g^*}\right) = \sum_{j=1}^{n} K\left(\psi_i^{g^*}, \psi_j^{g^*}\right)\boldsymbol{\beta}_j \tag{12}$$

where, $\{\boldsymbol{\beta}_j\}_{j=1}^n$ are vector coefficients with the same dimension as $\{\boldsymbol{\xi}_j\}_{j=1}^n$, $K(\cdot,\cdot)$ is the functional kernel, commonly used kernel is the Gaussian kernel

$$K\left(\psi_i^{g^*}, \psi_j^{g^*}\right) = \exp\left\{ -\frac{\int \left|\psi_i^{g^*}(\tau) - \psi_j^{g^*}(\tau)\right|^2 d\tau}{2\sigma^2} \right\} \tag{13}$$

From Eq. (12), it can be seen, undetermined coefficients of the solution to the minimization problem in Eq. (11) are $\{\boldsymbol{\beta}_j\}_{j=1}^n$, where $\boldsymbol{\beta}_j$ is a row vector takes the form $\boldsymbol{\beta}_j = \left[\beta_j^1 \ \beta_j^2 \cdots \beta_j^m\right] \in R^{1\times m}$, where m is the dimension of the response row vector $\boldsymbol{\xi}_i \in R^{1\times m}$.

$\left\{\boldsymbol{\beta}_j\right\}_{j=1}^n$ can be represented by a matrix

$$\mathbf{B}=\left[\boldsymbol{\beta}_1^{\mathrm{T}}\ \ \boldsymbol{\beta}_2^{\mathrm{T}}\cdots\boldsymbol{\beta}_n^{\mathrm{T}}\right]^{\mathrm{T}}\in R^{n\times m} \tag{14}$$

Then the equivalent form of the minimization problem in Eq. (11) is

$$\min_{\mathbf{B}}\left\{\operatorname{tr}\left((\mathbf{Y}-\mathbf{AB})(\mathbf{Y}-\mathbf{AB})^{\mathrm{T}}\right)+\lambda\operatorname{tr}\left(\mathbf{ABB}^{\mathrm{T}}\right)\right\} \tag{15}$$

where $\mathbf{Y}=\left[\boldsymbol{\xi}_1^{\mathrm{T}}\ \ \boldsymbol{\xi}_2^{\mathrm{T}}\cdots\boldsymbol{\xi}_n^{\mathrm{T}}\right]^{\mathrm{T}}\in R^{n\times m}$ and $\mathbf{A}=\left\{K\left(\psi_i^{g^*},\psi_j^{g^*}\right)\right\}\in R^{n\times n}$. The analytical solution of $\mathbf{B}$ in minimization problem in Eq. (15) is

$$\operatorname{vec}(\mathbf{B})=\left[\left(\mathbf{I}_{m\times m}\otimes\mathbf{A}\right)+\lambda\mathbf{I}_{mn\times mn}\right]^{-1}vec(\mathbf{Y}) \tag{16}$$

It can be seen the size of the matrix $\left(\mathbf{I}_{m\times m}\otimes\mathbf{A}\right)$ is $mn\times mn$, where m is the dimension of the feature vector of $\psi_i^{f^*}(t)$, i.e., $\boldsymbol{\xi}_i=\left[\xi_i^1\ \ \xi_i^2\cdots\xi_i^m\right]$ (see Eq. (8)), n is number of training functional samples, i.e., $\left\{\psi_i^{g^*}(t)\right\}_{i=1}^n$. The treatment of dimension reduction for $\left\{\psi_i^{f^*}(t)\right\}_{i=1}^n$ by the FPCA technique (see Eq. (8)) is very meaningful. If $\psi_i^{f^*}(t)$ is represented by corresponding values on regular grid $\{t_1,t_2,\cdots,t_T\}$ on $[0,1]$ as that used in the function-on-function regression model proposed by Lian [8], in general, dense grids are need to characterize a complex continuous function in consideration of the integral calculation in the inverse LQD transformation; such a approach will cost a huge amount of memory for formulating the matrix $\left(\mathbf{I}_{m\times m}\otimes\mathbf{A}\right)$; if the number of training functional samples are considerable large, an approach without dimension reduction may even lead to an insufficient memory errors and application failures. Additionally, predicting a response in lower dimension can help to improve accuracy. Therefore, in a practical engineering application, combining the FPCA-based dimension reduction technique with the vector-on-function regression model is more preferable than directly use the general RKHS-based function-on-function regression model in the problem of missing distribution imputation.

6. Missing Distribution Imputation

With the aforementioned vector-on-function regression model, the missing feature vector $\boldsymbol{\xi}_0$ can be estimated by

$$\hat{\boldsymbol{\xi}}_0=\hat{F}_{reg}\left(\psi_0^{g^*}\right)=\sum_{j=1}^n K\left(\psi_0^{g^*},\psi_j^{g^*}\right)\boldsymbol{\beta}_j \tag{17}$$

Then the missing representation of the PDF $f_0^*(x)$ in the Hilbert space can be estimated by

$$\hat{\psi}_0^{f^*}(t)\approx\mu_{\psi^{f^*}}(t)+\sum_{j=1}^m\xi_0^j\phi_j(t) \tag{18}$$

The PDF $f_0^*(x)$ can be subsequently estimated by applying the inverse LQD transformation to $\hat{\psi}_0^{f^*}(t)$. The target distribution $f_0(x)$ can finally be imputed after eliminating the effect of the added uniform distribution by the warping function-based PDF estimation technique, see Eq. (5), where the auxiliary PDF $h(x)$ can be set to be the PDF $g_0(x)$.

7. Conclusions

A new approach based on LQD transformation, FPCA and functional regression technique is proposed for missing distribution imputation. The integration in the inverse LQD transformation may introduce significant errors for some distributions, a pretreatment by adding a PDF of uniform distribution is proposed to address this problem, the newly developed warping function-based density estimation technique is proposed to recover the original PDF from a PDF mixed by the uniform distribution. The dimension reduction of response functions is very meaningful in memory saving and accuracy improvement. The proposed distribution imputation approach has good potential in providing more reliable distribution models for copula-based missing time series imputation.

8. Acknowledgments

This work is financially supported by the National Basic Research Program of China (Grant No. 2013CB036305), the National Natural Science Foundation of China (Grant No. 51638007), the Ministry of Science and Technology of China (Grant No. 2015DFG82080) and the China Scholarship Council (CSC).

References

[1] Käärik E. Imputation algorithm using copulas. *Metodoloskizvezki* 2006; 3(1): 109.

[2] Bárdossy A and Pegram G. Infilling missing precipitation records–A comparison of a new copula-based method with other techniques. *Journal of hydrology* 2014; 519: 1162-1170.

[3] Afrianti YS, Indratno SW and Pasaribu US. Imputation algorithm based on copula for missing value in time series data. In: *proceedings of International Conference on Technology, Informatics, Management, Engineering & Environment*, Bandung, Indonesia, 19-21, August 2014, pp. 252-257.

[4] Lascio FMLD, Giannerini S and Reale A. Exploring copulas for the imputation of complex dependent data. *Statistical Methods & Applications* 2015; 24(1):159-175.

[5] Lascio FMLD, Giannerini S and Reale A. Imputation of complex dependent data by conditional copulas: analytic versus semiparametric approach. In: *proceedings of the 21st International Conference on Computational Statistics*, Geneva, 19-22, August 2014, pp. 491-497.

[6] Chen Z. Modeling the Strain Monitoring Data in Structural Health Monitoring Using Copulas. *viXra preprint* viXra:1704.0277; 2017.

[7] Petersen A and Müller HG. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics* 2016; 44(1): 183-218.

[8] Lian H. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics* 2007; 35(4): 597-606.

[9] Dasgupta S, Pati D and Srivastava A. A Geometric Framework For Density Modeling. *arXiv preprint* arXiv:1701.05656; 2017.

[10] Kneip A and Utikal KJ. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* 2001; 96(454): 519-542.