DENSITY-BASED INVERSE CALIBRATION WITH FUNCTIONAL PREDICTORS

Noslen Hernàndez*, Rolando J. Biscay**,
Nathalie Villa-Vialaneix***, Isneri Talavera*

* CENATAV, Havana, Cuba
** CIMFAV, Universidad de Valparaìso, Chile
*** SAMM, Université Paris 1, France

Abstract

A standard problem in chemometrics is calibration, which aims at predicting a scalar random variable Y from a spectrum X. However, if the main problem is to predict Y from X, the physical data generation mechanism is rather that the spectrum X (e.g., an absorbance spectrum) is explained by Y, which is often a chemical variable (e.g., concentration of a substance). Using this physical model $X = r(Y) + \epsilon$, we propose a nonparametric approach to solve statistical calibration with functional data and to predict Y from X. This approach is based on the conditional probability density of X given Y, $f(X|Y)$: the proposed predictor takes the form a weighted average of the observed values of Y, where the weights are derived from an nonparametric estimate of $f(X|Y)$. The estimation of $f(X|Y)$ is performed with standard nonparametric estimation methods: in the present paper, the proposed estimator is explicitely given in the realistic case where the error ϵ is supposed to fit a Gaussian distribution: r is first estimated with a Nadaraya-Watson kernel estimate and the explicit form of the $f(X|Y)$ in the Gaussian case is used to the estimate $\hat{f}(X|Y)$.

The method is computationally simple and easy to implement and does not require any specific assumptions on the conditional density of Y given X, unlike most approach in functional regression. The consistency of the approach can also be proved. Its efficiency is illustrated on simulated datasets and compared to other approaches designed to solve regression problems with functional predictors.

Keywords: calibration; functional regression; inverse regression; Gaussian process

1 Introduction and notations

Statistical calibration plays a crucial role in many areas of technology such as pharmacology, neuroscience and chemometrics [Osborne, 1991, Martens and Naes, 1989, Brown, 1993, Massart et al., 1997, Lavine and Workman, 2002, Walters and Rizzuto, 1988]: an observable random variable X is related to a variable of interest Y according to a statistical model specified by a conditional probability density $f(X/Y)$. A sample $\mathcal{D}$ of independent observations $(x_1, y_1), ..., (x_n, y_n)$ of (X, Y) is available (training

sample). The problem is to make statistical inferences about Y on the basis of the given statistical model, the data $\mathcal{D}$ and X. In particular, in spectroscopy, this framework is useful to model the case where some chemical variable Y (e.g., concentration of a substance) has to be predicted from a digitized function X (e.g., an absorbance spectrum). The conditional density $f(X/Y)$ thus represents the physical data generation mechanism in which the output spectrum X is determined by the input chemical concentration Y, plus some random perturbation mainly due to the measurement procedure.

Hereafter, we restrict ourselves to cases where the variable of interest Y takes real values and where the predictor X lies in a functional space, e.g., L_2, which have already been studied in [Cuevas et al., 2002, Hernández et al., 2012] in the case where Y is not supposed to be a random variable (fixed design).

2 Presentation of the method

The aforementionned problem is usually addressed through the estimation of the regression function $\gamma(x) = E(Y/X = x)$. In this paper, a new functional calibration method to estimate $\gamma(X)$ is introduced, which relies on assuming the following regression model:

$$X = r(Y) + \epsilon, \tag{1}$$

where ϵ is a random process (perturbation or noise), independent of Y, which is supposed to fit a Gaussian distribution, and r is a function from $\mathbb{R}$ into $\mathcal{X}$. Under this Gaussian distribution assumption, the conditional distribution $P(\cdot/y)$ is also a Gaussian distribution and is fully determined by its corresponding mean function $r(\cdot) = E(X/Y = \cdot)$, and its covariance operator Γ (not depending on y), which is a symmetric and positive Hilbert-Schmidt operator on the space $\mathcal{X}$. Thus, there exists an eigenvalue decomposition of Γ, $(\varphi_j, \lambda_j)_{j\geq 1}$ such that $(\lambda_j)_j$ is a decreasing sequence of positive real numbers, $(\varphi_j)_j$ are orthonormal functions on $\mathcal{X}$ and $\Gamma = \sum_j \lambda_j \varphi_j \otimes \varphi_j$ where $\varphi_j \otimes \varphi_j : h \in \mathcal{X} \to \langle \varphi_j, h\rangle \varphi_j$.

Suppose that the following usual regularity condition holds [Grenander, 1981, p. 271]: for each $y \in \mathbb{R}$, $\sum_{j=1}^{\infty} \frac{r_j^2(y)}{\lambda_j} < \infty$, where $r_j(y) = \langle r(y), \varphi_j\rangle$ for all $j \geq 1$. Then, the density $f(\cdot/y)$ of $P(\cdot/y)$ with respect to P_0 has the explicit form: $f(x/y) = \exp\left\{\sum_{j=1}^{\infty} \frac{r_j(y)}{\lambda_j}\left(x_j - \frac{r_j(y)}{2}\right)\right\}$, where $x_j = \langle x, \varphi_j\rangle$ for all $j \geq 1$.

Under these assumptions, and the one that the distribution of Y has a density $f_Y(y)$ (with respect to the Lebesgue measure on $\mathbb{R}$), the regression function can be written as $\gamma(x) = \frac{\int_{\mathbb{R}} f(x/y) f_Y(y) y dy}{f_X(x)}$, where $f_X(x) = \int_{\mathbb{R}} f(x/y) f_Y(y) dy$. which suggests the following (plug-in) estimate of $\gamma(x)$:

$$\hat{\gamma}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{f}(x/y_i) y_i}{\hat{f}_X(x)}, \tag{2}$$

where $\hat{f}(x/y)$ is an estimate of the density $f(x/y)$ of $P(\cdot/y)$ with respect to the measure P_0 and $\hat{f}_X(x)$ is defined by $\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x/y_i)$ and used to estimate the density $f_X(x)$ of X.

Finally, an estimate $\hat{f}(x/y)$ of $f(x/y)$ can be obtained through the following steps:

1. For each $t \in [0,1]$, compute an estimate $\hat{r}(\cdot)(t)$ of the function $r : y \mapsto r(y)(t)$ with a smoothing kernel method:

$$\hat{r}(y) = \frac{\sum_{i=1}^{n} K\left(\frac{y_i - y}{h}\right) x_i}{\sum_{i=1}^{n} K\left(\frac{y_i - y}{h}\right)} = \frac{\hat{m}(y)}{\hat{f}_Y(y)}, \tag{3}$$

where h is the bandwidth parameter, K an order k kernel, $\hat{m}(y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{y_i - y}{h}\right) x_i$ and $\hat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{y_i - y}{h}\right)$.

2. Obtain estimates $(\widehat{\varphi}_j, \hat{\lambda}_j)_j$ of the eigenfunctions and eigenvalues $(\varphi_j, \lambda_j)_j$ of the covariance Γ on the basis of the empirical covariance $\widehat{\Gamma}$ of the residuals $\hat{e}_i = x_i - \hat{r}(y_i)$, that is, $\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i \otimes \hat{e}_i$.

3. Estimate $f(x/y)$ by

$$\hat{f}(x/y) = \exp\left\{\sum_{j=1}^{p} \frac{\hat{r}_j(y)}{\hat{\lambda}_j}\left(\hat{x}_j - \frac{\hat{r}_j(y)}{2}\right)\right\}, \tag{4}$$

where $\hat{r}_j(y) = \langle \hat{r}(y), \widehat{\varphi}_j \rangle$, $\hat{x}_j = \langle x, \widehat{\varphi}_j \rangle$ for all $j \geq 1$ and $p = p(n)$ is an integer, smaller than n and such that $p(n) \to +\infty$.

Under technical assumptions, it can be proved that, for all $x \in \mathcal{X}$ such that $f_X(x) > 0$, we have:

$$\lim_{n \to +\infty} \hat{\gamma}(x) =^P \gamma(x).$$

3 Simulation

In this section, the feasibility and the performances of the nonparametric functional regression method described in Section 2 is discussed through a simulation study. A dataset was simulated in which values for the real random variable Y were drawn from a uniform distribution in the interval $[0, 10]$ and then, X was generated by using the following model:

$$X = \sin(Y)v_1 + \log(Y+1)v_5 + \epsilon$$

where $(v_i)_{i \geq 1}$ is the trigonometric basis of $\mathcal{X} = \mathcal{L}^2([0,1])$ (i.e., $v_{2k-1} = \sqrt{2}\cos(2\pi k t)$, and $v_{2k} = \sqrt{2}\sin(2\pi k t)$) and ϵ os a Gaussian process independent of Y with zero mean and covariance operator $\Gamma_e = \sum_{j \geq 1} \frac{1}{j} v_j \otimes v_j$. Training and a test samples were simulated with respective sizes $n_L = 300$ and $n_T = 200$.

Figure 1 compares the true $F(y)(t)$ to its estimated values for various values of y (top) and for various values of t (bottom). The results are very satisfactory given the fact that the data have a high level of noise (which clearly appears in the bottom of this figure). Figure 2 shows the results of the steps 2-3 of the estimation scheme: the estimated eigendecomposition of r is compared to the true one, and the predicted value for Y are compared to the true ones, both on training and test sets. The estimation of the eigendecomposition is also very satisfactory despite the high level of noise, and the comparison between training and test sets shows that the method does not overfit the data.

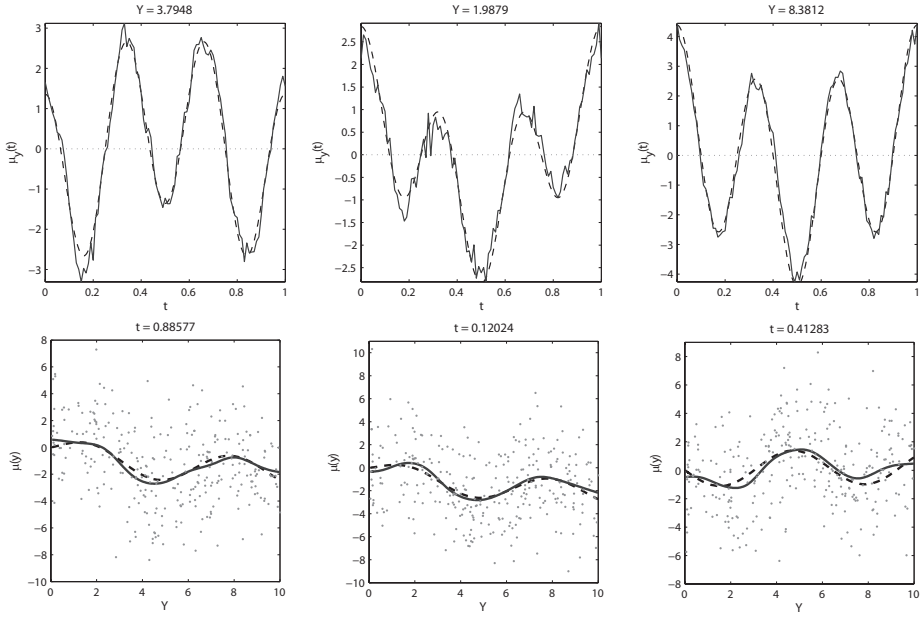

Figure 1: Model **M2**. Top: True values (discontinuous lines) and estimates (continuous lines) of $F(y)$ for various values of y. Bottom: true values and estimates of $F(\cdot)(t)$ for various values of t (bottom). The dots (bottom) are the simulated data $(x_i(t))_i$ in the training set.

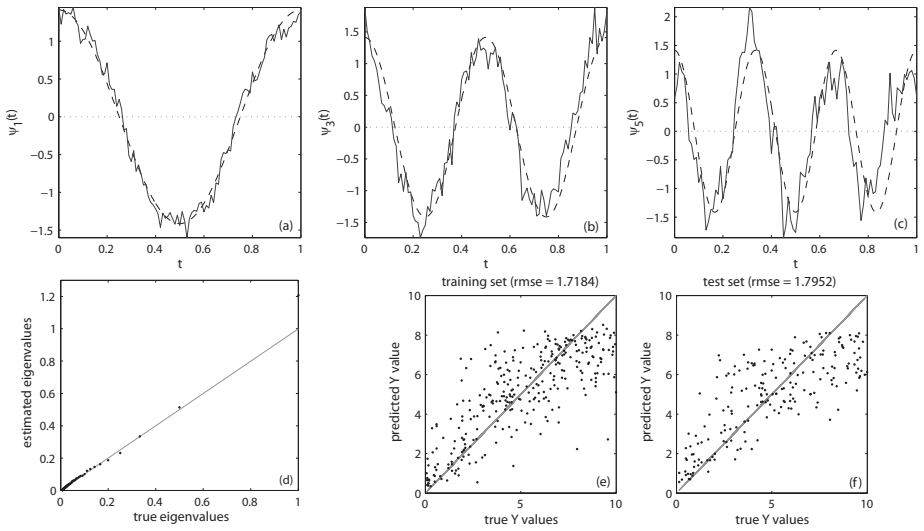

Figure 2: Model **M2**: (a-c): True (dashed line) and estimated eigenfunctions (continuous line); (d): estimated vs. true eigenvalues and (e-f): predicted values for Y vs. the true ones for training and test sets.

References

[Brown, 1993] Brown, P. (1993). *Measurement, Regression and Calibration*. Oxford.

[Cuevas et al., 2002] Cuevas, A., Febrero, M., and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *The Canadian Journal of Statistics*, 30:285–300.

[Grenander, 1981] Grenander, U. (1981). *Abstract Inference*. Berlin.

[Hernández et al., 2012] Hernández, N., Biscay, R., and Talavera, T. (2012). A non-Bayesian predictive approach for statistical calibration. *Journal of Statistical Computation and Simulation*, 82(4):529–545.

[Lavine and Workman, 2002] Lavine, B. and Workman, J. (2002). Fundamental reviews: chemometrics. *Analytical Chemistry*, 74:2763–2770.

[Martens and Naes, 1989] Martens, H. and Naes, T. (1989). *Multivariate Calibration*. Chichester.

[Massart et al., 1997] Massart, D., Vandeginste, B., Buydens, L., Jong, S., Lewi, P., and Smeyers-Verbeke, J. (1997). *Handbook of Chemometrics and Qualimetrics: Part B*. The Netherlands.

[Osborne, 1991] Osborne, C. (1991). Statistical calibration: a review. *International Statistical Review*, 59:309–336.

[Walters and Rizzuto, 1988] Walters, F. and Rizzuto, G. (1988). The calibration problem in statistics and its application to chemistry. *Analytical Letters*, 21:2069–2076.