GOODNESS–OF–FIT TEST FOR SINGLE FUNCTIONAL INDEX MODEL

Lax Chan [1] and Aldo Goia[1]

[1] Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italy(e-mail: lax.chan@uniupo.it, aldo.goia@uniupo.it)

ABSTRACT: Motivated by a problem that commonly arise in the food industry, a methodology based on the Single Functional Index Model (SFIM) is proposed and a test procedure to specify the link function between the real response and the functional covariate is described and applied.

KEYWORDS: Functional regression, Goodness–of–fit test, Single Index Model

1 Introduction

In the food industry, to obtain the composition of a given substance in terms of protein, fat, moisture, etc. is an important task. Since a full–scale chemical analysis is often costly and time consuming, it is preferred to estimate that composition by using spectrometric curves which can be obtained easily as the absorption of a reflected light for various wavelengths. In that situation a regression model with a scalar response (the percentage of the component) and a functional covariate (the spectrometric curve, or a transformation of it) can be profitably used. Consider for instance the prediction of the fat proportion by using the near-infrared spectra of 39 milk specimens obtained by SCiO device recorded between 740 and 1070 nm in Figure 1 and originally considered in Riu *et al.*, 2020. This dataset has been used Di Brisco *et al.*, 2023, where some functional parametric and nonparametric regression models have been applied and compared.

One can note that, if full nonparametric approaches are exploratory but suffer of dimensionality problems, parametric models are easily interpreted but not flexible. A useful alternative in this research field can come from the semi-parametric regression approaches that combine flexibility and interpretability. In particular the class of Single Functional Index Model (SFIM) defines a relationship between the functional predictor X and the real–valued random variable Y through an unknown real link function g that acts on a projection of the functional predictor along an unknown direction θ, subject to an identifiability condition: $Y = g(\langle X, \theta \rangle) + \mathcal{E}$, where $\langle X, \theta \rangle = \int X(t)\,\theta(t)\,dt$, $\|\theta\|^2 = 1$ and

$\theta(t) > 0$ for a fixed t. A methodology which combines a spline approximation of the functional coefficient θ and the one-dimensional Nadaraya-Watson approach to estimate the link function g are proposed in Ferraty *et al.*, 2013. The main advantage in using SFIM is the possibility to work in the one dimensional analogue of an infinite dimensional problem, through the projective strategy, and hence to visualise an estimate of g from the observed data and hence suggests the nature of the relationship of X and Y. This allows to postulate a target link function g_0 and test its compatibility with the observed data at a significance level.

The new test procedure in the SFIM context based on the conditional moment test approach has been defined and analyzed in Chan *et al.*, 2023. This work aims to summarize the main features of such a test and apply it to the spectrometric example. In particular, after illustrating the basic principle of the test in Section 2, the application to the real data is discussed in Section 3.

2 The test principle

Consider the SFIM and define $\mathcal{G}_0 = \{g_0^{\beta} : \mathbb{R} \to \mathbb{R}, \beta \in \mathbb{R}^{d+1}\}$, where g_0^{β} is a known function depending on the parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_d) \in \mathbb{R}^{d+1}$, $d \geq 1$ integer. Consider then the following hypothesis:

$$H_0 : g \in \mathcal{G}_0 \qquad \text{vs.} \qquad H_1 : g \in \mathcal{G}_1$$

where $\mathcal{G}_1$ is a set of real functions g_1^{β} such that $\mathcal{G}_1 \cap \mathcal{G}_0 = \emptyset$.

Define $\mathcal{E} = Y - g_0^{\beta}(\langle X, \theta \rangle)$ and $\mathbb{E}[\mathcal{E}|X] = g(\langle X, \theta \rangle) - g_0^{\beta}(\langle X, \theta \rangle)$. The quantity $Q = \mathbb{E}[\mathcal{E}\mathbb{E}[\mathcal{E}|X] w(X)]$, where $w(X) > 0$ is a weight function, is null under H_0 and strictly positive under H_1.

To implement the test procedure, an empirical version of Q has to be derived from a sample (X_i, Y_i), $i = 1, \ldots, n$ drawn from (X, Y). Assuming the projection random variable $\langle X, \theta \rangle$ admits a positive probability density function f_θ, then a possible choice for the weight function is $w = f_\theta$. By taking a Nadaraya–Watson type nonparametric kernel estimate of $\mathbb{E}[\mathcal{E}|X]$ at the point X_i and a cross–validated kernel estimate of f_θ, the empirical version of Q is:

$$Q_n\left(\widehat{\theta}\right) = \frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \widehat{\mathcal{E}}_i \widehat{\mathcal{E}}_j K_{ij}^{\widehat{\theta}},$$

where $\widehat{\theta}$ is an estimate of θ and $\widehat{\mathcal{E}}_i = Y - g_0^{\widehat{\beta}}\left(\left\langle X_i, \widehat{\theta} \right\rangle\right)$, where $\widehat{\beta}$ is an estimate

for β. The standardised test statistic is $T_n = n\sqrt{h}Q_n\left(\widehat{\theta}\right)/\nu_n\left(\widehat{\theta}\right)$ where

$$\nu_n^2\left(\widehat{\theta}\right) = \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathcal{E}_i^2\mathcal{E}_j^2\left(K_{ij}^{\widehat{\theta}}\right)^2.$$

To compute the p-value and to derive the critical region of the test at the significance level α, the derivation of the asymptotic null distribution for T_n is required. Under appropriate assumptions, one can prove that $T_n \sim \mathcal{N}(0,1)$, as n diverges. Then one rejects the null hypothesis if $T_n \geq z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of the standard normal distribution. For further details, interested readers are invited to consult Chan *et al.*, 2023.

3 Application to spectrometric data

Consider the SFIM involving the original spectra as covariate and the quantitiy of fat as response. Some attempts with first and second derivatives of the spectrometric curves have been performed but with a deterioration in the quality of the prediction (and this is coherent with the models in Di Brisco *et al.*, 2023). In Figure 1 the estimates $\widehat{\theta}$ and $\widehat{g}$ of the direction θ and link function g are plotted. Observing the shape of the former, it seems that the relevant part of the spectrum in predicting the fat content is between about 950 and 1070 nm, whereas the latter suggests that a linear specification for the model seems not reasonable. For what concerns the prediction ability of that model, one used the RMSE, that is $\sum_i (y_i - \widehat{y}_i)^2 / \sum_i y_i^2$, and the MAPE, that is $\sum_i |y_i - \widehat{y}_i| / y_i$; the first index equals 0.015 and the second one 0.096.

At this stage it is possible to carry out the specification test; in particular the following polynomial and logistic null models are considered:

$$H_0^p : g_0(u) = \beta_0 + \sum_{j=1}^{p}\beta_j u \qquad H_0^{\log} : g_0(u) = e^{\beta_0+\beta_1 u}/(1+e^{\beta_0+\beta_1 u})$$

where $u = \left\langle x, \widehat{\theta}\right\rangle$ and $p = 1,2,3$ (corresponding to linear, quadratic and cubic link espectively). Since all the real parameters β_j are unknown, they are estimated by an OLS approach under the null hypothesis. The p-values calculated by using the asymptotic null distribution are: 0 for H_0^1, 0.035 for H_0^2, 0.207 for H_0^3 and 0 for $H_0^{\log}$. One can conclude that the linear, quadratic as well as logistic assumptions on the link function are not compatible with the empirical evidence, whereas a cubic link could be a good choice to model

the relationship. Therefore, a model to predict the content of fat Y in milk specimens starting from the spectrometric curve X can be specified as follows: $Y = 0.014 - 0.69 \cdot \langle X, \widehat{\theta} \rangle + 10.6 \cdot \langle X, \widehat{\theta} \rangle^2 - 33.2 \cdot \langle X, \widehat{\theta} \rangle^3 + \mathcal{E}$.

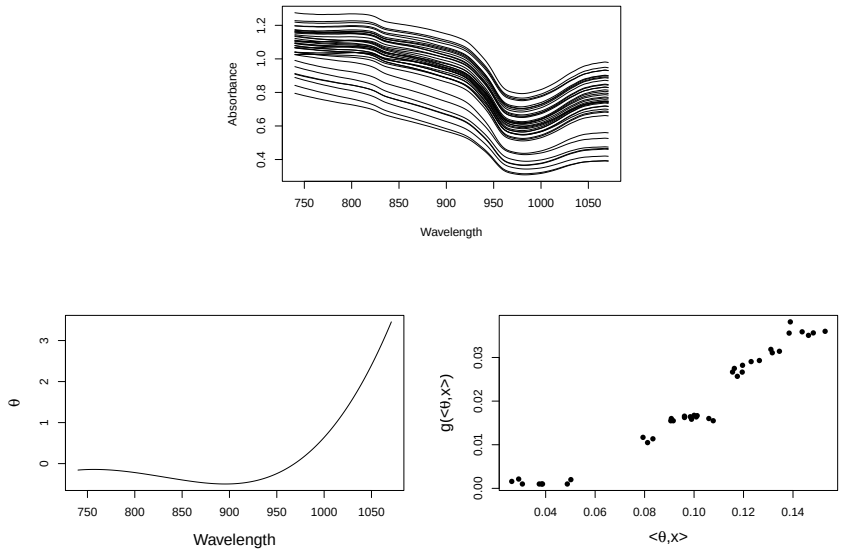

Figure 1. *Milk spectra recorded using SCiO device (top), Estimated direction* θ *(bottom left) and estimated link function (bottom right) for the SFIM.*

References

CHAN, L., DELSOL, L. & GOIA, A. 2023. A Link Function Specification Test in the Single Functional Index Model. *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-023-00545-7

DI BRISCO, A.M., BONGIORNO, E.G., GOIA, A. & MIGLIORATI, S. 2023. Bayesian flexible beta regression model with functional covariate. *Computational Statistics* **38**, 623-–645 .

FERRATY, F., GOIA, A., SALINELLI, E. & VIEU, P. 2013. Functional Projection Pursuit Regression, *Test* **22** , 293–320.

RIU, J., GORLA, G., CHAKIF, D., BOUQUÉ, R. & GIUSSANI, B. 2020. Rapid analysis of milk using low-cost pocket-size NIR spectrometers and multivariate analysis. *Foods*, **9** (8), 1090.