Journal of Multivariate Analysis 101 (2010) 512–524

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Semiparametric Bayesian measurement error modeling

María P. Casanova [a], Pilar Iglesias [b], Heleno Bolfarine [c], Victor H. Salinas [d,∗], Alexis Peña [b]

[a] *Departamento de Estadística, Facultad de Ciencias Físicas y Matemáticas, Universidad of Concepción, Chile*
[b] *Departamento de Estadística, Facultad de Matemática, Pontificia Universidad Católica de Chile, Chile*
[c] *Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil*
[d] *Departamento de Matemática y Ciencia de la Computación, Facultad de Ciencia, Universidad de Santiago de Chile, Chile*

ARTICLE INFO

Article history:
Received 8 May 2007
Available online 27 November 2009

AMS subject classifications:
62A15
62J05

Keywords:
Classical measurement error model
Hierarchical elliptical model
Posterior distribution
Dirichlet process
Gibbs sampling

ABSTRACT

This work presents a Bayesian semiparametric approach for dealing with regression models where the covariate is measured with error. Given that (1) the error normality assumption is very restrictive, and (2) assuming a specific elliptical distribution for errors (Student-t for example), may be somewhat presumptuous; there is need for more flexible methods, in terms of assuming only symmetry of errors (admitting unknown kurtosis). In this sense, the main advantage of this extended Bayesian approach is the possibility of considering generalizations of the elliptical family of models by using Dirichlet process priors in dependent and independent situations. Conditional posterior distributions are implemented, allowing the use of Markov Chain Monte Carlo (MCMC), to generate the posterior distributions. An interesting result shown is that the Dirichlet process prior is not updated in the case of the dependent elliptical model. Furthermore, an analysis of a real data set is reported to illustrate the usefulness of our approach, in dealing with outliers. Finally, semiparametric proposed models and parametric normal model are compared, graphically with the posterior distribution density of the coefficients.

1. Introduction

The simple linear regression measurement error model (MEM) is defined by the equations

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{1}$$

$$z_i = x_i + u_i, \tag{2}$$

$i = 1, \ldots, n$, indicating that the true value x_i is unknown and that its estimate z_i is available, generated by the random mechanism (2), which is called additive MEM. Situations where x is measured with error abound in the literature. Fuller [1] and Carroll et al. [2,3] report several situations where such problems occur. Bayesian parametric methodologies for the MEM model are considered in Dellaportas and Stephens [4], Mallik and Gelfand [5], Richardson and Gilks [6] and Bolfarine and Arellano-Valle [7], among others. A nonparametric approach to the dichotomous situation using a logistic model is considered in Muller and Roeder [8]. One simple situation occurs when the interest is focused on studying the relationship between the production of a certain cereal and the amount of nitrogen in the soil. This relation can only be measured by laboratory analysis, and certainly involves measurement error. Berkson [9] models the true unobserved value as depending on the observed value plus an error term (see [4]);

$$x_i = z_i + u_i.$$

∗ Corresponding author.
E-mail addresses: macasanova@udec.cl (M.P. Casanova), pliz@mat.puc.cl (P. Iglesias), hbolfar@ime.usp.br (H. Bolfarine), v_h_salinas@yahoo.com (V.H. Salinas), alexis.pena@sas.com (A. Peña).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2009.11.004

In contrast, our work models the observed value z_i as the sum of true value x_i and the error term. Therefore, this article investigates the classical measurement error model (and not the Berkson model), hereinafter called simply measurement error model.

Typically, the normality assumption is considered for the error vector, namely,

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} \sim N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \phi_e & 0 \\ 0 & \phi_u \end{pmatrix}\right], \quad i = 1, \dots, n. \tag{3}$$

Here, the normality assumption for the error vector in (3) is replaced by the ellipticity assumption. As is well known, the elliptical models let us consider two different cases. The first case, called the *independent model*, considers independent observations, which are typically harder to deal with but generate robust models, contrasting with outlying observations. The second case, the *dependent model*, presents robustness within the family that is being considered. A detailed discussion on properties and differences between both models is presented in Bolfarine and Arellano-Valle [7]. In the independent case considered first, we have that

$$\begin{aligned} e_i &\overset{iid}{\sim} El_1(0, \phi_e^{-1}, h_e), \\ u_i &\overset{iid}{\sim} El_1(0, \phi_u^{-1}, h_u), \end{aligned} \tag{4}$$

where

$$\begin{aligned} h_e(u) &= \int_0^\infty \frac{1}{\sqrt{2\pi w}} \mathrm{e}^{-\frac{u}{2w}} \mathrm{d}G_e(w), \\ h_u(u) &= \int_0^\infty \frac{1}{\sqrt{2\pi w}} \mathrm{e}^{-\frac{u}{2w}} \mathrm{d}G_u(w), \end{aligned} \tag{5}$$

with $G_e(\cdot)$ and $G_u(\cdot)$ mixing distributions satisfying $G_e(0) = G_u(0) = 0$. Also, $El_1(\mu, \phi, h)$ denotes an elliptical random variable with location μ, scale ϕ^{-1} and generator function $h(\cdot)$. In this case, the elliptical model is a compound normal model. See Fang et al. [10] and Arellano-Valle et al. [11] for important reviews on elliptical distributions. As considered in Bolfarine and Arellano-Valle [7], the independent elliptical model is non-differential.

The dependent elliptical MEM works with the joint distributions of $\mathbf{e} = (e_1, \dots, e_n)^t$ and of $\mathbf{u} = (u_1, \dots, u_n)^t$, which follows by considering that

$$\begin{aligned} \mathbf{e} &\sim El_n(\mathbf{0}, \phi_e^{-1} I_n, h_e), \\ \mathbf{u} &\sim El_n(\mathbf{0}, \phi_u^{-1} I_n, h_u), \end{aligned}$$

where I_n denotes the n-dimensional identity matrix. The generator functions h_e and h_u satisfy (5) for mixing distributions $G_e(\cdot)$ and $G_u(\cdot)$ such that $G_e(0) = G_u(0) = 0$. Note that the usual MEM normal model proceeds by letting G_e and G_u degenerate into a single point, in which case the components of $\mathbf{e}$ are independent. Similar results hold for the vector $\mathbf{u}$. However, for models different from the normal (Student-t model, for example), elements of $\mathbf{e}$ are uncorrelated but dependent. As considered in Bolfarine and Arellano-Valle [7], the dependent elliptical model is a weak differential measurement error model.

Zellner [12] shows the goodness in the analysis of measurement error model, considering the term of error with dependent elliptical distribution. The case addressed by this author, is the Student-t model. For the elliptical case analyzed by us, as mentioned in our article, the Dependent Model presents fewer difficulties in the analysis because the semiparametric case leads to the Parametric one.

The dependent assumption implies that the joint distribution of the terms of error is Multivariate Elliptical, a broader distribution, and therefore more flexible, than the joint distribution of independent elliptical variables (Independent Elliptical model). This author presents a few illustrations that validate the importance and usefulness of this model. In general their applicability is related with the logic associate with the concept of dependent error, for the particular faced problem.

This concept is linked to the dependence of the observations, which results in such errors. Some cases are, for example, grouped individuals (consumers, enterprisers), the runs of a single experiment (measurements with a single instrument, market returns in a particular moment), etc.

In this paper, we extend both the elliptical models described above by considering that the mixing distributions generating the elliptical family follows discrete mixtures as well as a Dirichlet process prior. We believe that the family of models considered are more general than the family of models used in Muller and Roeder [8] to describe the distribution of the unknown (true) covariate x. An alternative robust classical approach (termed corrected score) in the sense that it is not required to specify a distribution for x, appears in Nakamura [13]. However, the corrected score approach does not seem to be for some of the models (Student-t, for example) studied in this paper.

The paper is organized as follows. Section 2 presents the semiparametric approach for the independent elliptical MEM. Discrete and Dirichlet process priors are considered for the mixing parameter. Posterior implementation is carried out using the MCMC implementation considered in [14,15]. Section 3 investigates the dependent elliptical situation, where it is shown that the situation of the Dirichlet process prior reduces to the usual parametric mixture model with gamma distributions as the mixing distribution. Section 4 presents the computational implementation. Section 4.1 presents a real data application which illustrates the usefulness of the presented approach. Section 4.2 compares graphically semiparametric and parametric methods, based on simulated data. Finally, Section 5 presents some extensions and areas of future research.