Chemometrics and Intelligent Laboratory Systems 156 (2016) 21–35

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Tutorial article

Heavy tailed calibration model with Berkson measurement errors for replicated data

Betsabé Blas [a,*], Heleno Bolfarine [b], Victor H. Lachos [c]

[a] Departamento de Estatística, Universidade Federal de Pernambuco, Recife, PE, Brazil
[b] Departamento de Estatística, Universidade de São Paulo, São Paulo, SP, Brazil
[c] Departamento de Estatística, Universidade Estadual de Campinas, SP, Brazil

ARTICLE INFO

Article history:
Received 8 January 2016
Received in revised form 26 April 2016
Accepted 28 April 2016
Available online 13 May 2016

Keywords:
MC-EM algorithm
Scale mixtures of normal distributions
Controlled variable
Calibration model
Local influence

ABSTRACT

This work considers the so called controlled calibration model in which the independent variable is a controlled variable (Berkson type) and assumes that the measurement errors follow a scale mixtures of normal (SMN) distribution. The SMN family of distributions is an attractive class of symmetric distributions including the normal, Student-t, slash and contaminated normal distributions as special cases, providing a robust alternative to estimation in controlled calibration models in the absence of normality. An EM-type algorithm is developed, which is used to develop the local influence approach to assess the robustness aspects of the parameter estimates under four perturbation schemes. Results obtained from a real dataset in the area of chemistry are reported.

1. Introduction

In chemical analysis, a chemist wants to establish a calibration line in order to measure the amount of some chemical element in samples. Calibration models are intended to link a quantity of interest X (e.g. the concentration of a chemical compound) to a value Y obtained from a measurement device. In this context, a major concern is to build calibration models that are able to provide precise predictions for X from measured responses Y. There are two stages related to this calibration process: in the first stage, a calibration curve is established for the relation between the dependent variable Y and the independent variable X. In this stage, for each pre-fixed amount X, the measurement Y is made with a quick and inexpensive method. The pre-fixed amount X has been determined by an extremely accurate standard method that is slow and expensive. Afterwards, at the second stage, the measurement Y_0 corresponding to an unobserved value X_0 of X is observed.

When the concentration of the standard solution is pre-fixed by the chemist and a process is carried out attempting to attain it, errors are generated even though the standard method is extremely accurate. Hence, in this case the so-called controlled variable X arises (also known as Berkson-type variable, [1]), which is defined by the pre-fixed concentration value of the standard solution. Assuming additive error, it can be expressed by the equation $x = X + \delta$, where x is the unobserved variable, which represents the unknown real concentration, and δ is the related measurement error variable.

Motivated by chemical applications with replicated measurement of the variable Y, and as a generalization of the controlled calibration model proposed in [2], one can write the controlled calibration model with replicated measurement on the response variable (CCM) as follows:

$$Y_{ij} = \alpha + \beta x_i + \epsilon_{ij}, \quad j = 1, \cdots, m_i \text{ and } i = 1, \cdots, n, \tag{1}$$

$$x_i = X_i + \delta_i, \quad i = 1, \cdots, n, \tag{2}$$

$$Y_{0i} = \alpha + \beta X_0 + \epsilon_{0i}, \quad i = n+1, \cdots, n+r, \tag{3}$$

where the variables $X_1, X_2, \cdots, X_n$ are taken as pre-fixed values by the analyst. Typically, one assumes that the error variables are independent and normally distributed (iid), i.e., ϵ_{ij}, $\epsilon_{0l} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$ and $\delta_i \overset{\text{iid}}{\sim} N(0, \sigma_\delta^2)$. If $N_k(\mu, \sum)$ denotes the k-variate normal distribution with mean μ and covariate matrix $\sum$, then the first stage error model is given by:

$$\boldsymbol{\phi}_i = \begin{pmatrix} \epsilon_i \\ \delta_i \end{pmatrix} \overset{\text{iid}}{\sim} N_{m_i+1}(\mathbf{0}, \boldsymbol{\Psi}_i), \tag{4}$$

* Corresponding author.
E-mail address: betsabe@de.ufpe.br (B. Blas).

http://dx.doi.org/10.1016/j.chemolab.2016.04.014
0169-7439/

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^\top$ and $m_i + 1$ is the number of components of $\boldsymbol{\phi}_i$, which can vary from unit to unit in some applications, and the covariance matrix $\boldsymbol{\Psi}_i = \begin{pmatrix} \sigma_\epsilon^2 \mathbf{I}_{m_i} & 0 \\ 0 & \sigma_\delta^2 \end{pmatrix}$ is an $(m_i+1)\times(m_i+1)$ matrix of known form indexed by a set of unknown parameters σ_ϵ^2 and σ_δ^2. The second stage error model is given by:

$$\boldsymbol{\epsilon}_0 \sim N_r\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_r\right), \tag{5}$$

where $\boldsymbol{\epsilon}_0 = (\epsilon_{0n+1}, \ldots, \epsilon_{0n+r})^\top$ and $\mathbf{I}_r$ represents the $r \times r$ identity matrix.

In this case, the model parameters are $\alpha, \beta, X_0, \sigma_\epsilon^2$ and σ_δ^2 and the main interest is to estimate the quantity X_0. Moreover, the Berkson-type variable X_i is a fixed and known value, and $\boldsymbol{\epsilon}_i$ and x_i are unknown random quantities, $i = 1, \ldots, n$.

Despite its interesting usability the distribution of the measurement errors as well as the unobserved covariates are assumed to be Gaussian, but unfortunately, this normal assumption is too restrictive and suffers from lack of robustness, which may have important effects on inferences. Hence, a study of the properties of the controlled calibration model under nonstandard assumptions, such as normality, is very pertinent. Our approach is to replace the normal distribution by the scale mixtures of normal distributions [3], which is the most important subclass of the elliptically symmetric distributions. In this work we consider a classical approach for CCM assuming scale mixtures of normal (SMN) distributions. This extension results in a flexible class of models for robust estimation in CCM that contains as proper elements, the normal (N), Student-t (T), slash (SL) and the contaminated normal (CN) distributions. All these distributions present heavier tails than the normal one, and thus can be used for robust inference in many types of models.

The assessment of robustness aspects of the parameter estimates in statistical models has been an important concern of various researchers in recent decades. The deletion method, which consists in studying the impact on the parameter estimates after dropping individual observations, is probably the most employed technique to detect influential observations (see [4]). Nevertheless, research on the influence of small perturbations in the model/data on the parameter estimates has received increasing attention in recent years. This can be achieved by performing local influence analysis, a general statistical technique used to assess the stability of the estimation outputs with respect to the model inputs. Following the pioneering work of Cook [5], this area of research has received considerable attention in the recent statistical literature. However, as the observed log-likelihood function of the CCM with SMN distributions (hereafter, SMN-CCM) involves some integrals, direct application of Cook's approach [5] for SMN-CCM is very difficult, because these measures involve the first and second partial derivatives of this function. Zhu and Lee [6] developed an approach to perform local influence analysis for general statistical models with missing data by working with a Q-displacement function closely related to the conditional expectation of the complete-data log-likelihood at the E-step of the EM algorithm. Zhu et al. [7] developed a rigorous differential-geometrical framework for a perturbation model, named the perturbation manifold. This approach shows that the metric tensor of the perturbation manifold provides important information about selecting an appropriate perturbation for a specific model and, it also defines new influence measures for smooth objective functions. The theory of Zhu et al. [7] was developed on the basis of the observed data likelihood, therefore it is not suitable for application to complex latent variable models that involve observed-data likelihood with intractable integrals. An application of this approach can be found in [8], where this method is applied in functional comparative models with replicated data.

The paper is organized as follows. Section 2 presents the robust SMN-CCM and discusses the maximum likelihood (ML) estimation via the EM-algorithm. Additionally, the expected information matrix is derived analytically. In Section 3, we give a brief introduction to the local influence approach for models with incomplete-data and develop the method required for the SMN-CCM. The advantage of the proposed method is illustrated using a real chemical dataset in Section 4. Section 5 presents our concluding remarks and some mathematical expressions and figures are given in the Appendix.

2. Model formulation and estimation

2.1. The SMN class of distributions

In this section we define the SMN-CCM. Before this, let us recall that a SMN distribution is defined as the distribution of the p-dimensional random vector

$$\mathbf{Y} = \boldsymbol{\mu} + \kappa^{1/2}(U)\mathbf{Z}, \tag{6}$$

where $\boldsymbol{\mu}$ is a location vector, $\mathbf{Z}$ is a normal random vector with mean vector $\mathbf{0}$, variance–covariance matrix $\boldsymbol{\Sigma}$, $\kappa(.)$ is a weight function and U is a mixing positive random variable with cumulative distribution function (cdf) $H(u;\boldsymbol{\nu})$ and probability density function (pdf) $h(u;\boldsymbol{\nu})$, independent of $\mathbf{Z}$, where $\boldsymbol{\nu}$ is a scalar or parameter vector indexing the distribution of U. Given U, $\mathbf{Y}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\kappa(u)\boldsymbol{\Sigma}$, i.e., $\mathbf{Y}|U = u \sim N_p(\boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma})$. Hence, the pdf of $\mathbf{Y}$ is given by:

$$f(\mathbf{y}) = \int_0^\infty \phi_p(\mathbf{y}|\boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}) dH(u), \tag{7}$$

where $\phi_p(.|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the pdf of the p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariate matrix $\boldsymbol{\Sigma}$. A particular case of this distribution is the normal distribution, for which H is degenerate, with $\kappa(u) = 1$, $u > 0$. From here on, we denote $\text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; H)$ as the SMN distribution with the pdf Eq. (7). We notice that when $k(u) = u^{-1}$ in Eq. (6), the distribution of $\mathbf{Y}$ reduces to the normal/independent (NI) family discussed, for instance, in [9]. The multivariate T distribution with ν degrees of freedom, can be derived from the mixture model (7), by taking $\kappa(u) = 1/u$, where U is distributed as $Gamma(\nu/2, \nu/2)$, with $u > 0, \nu > 0$. The SL distribution arises when $\kappa(u) = 1/u$ and the distribution of U is $Beta(\nu, 1)$, $0 < u < 1$ and $\nu > 0$. The CN distribution is given when $\kappa(u) = 1/u$ and U is a discrete random variable taking one of two states.

2.2. The proposed model

Eqs. (1)–(2) from the CCM can be represented as:

$$Y_{ij} = \alpha + \beta X_i + \xi_{ij}, \tag{8}$$

where $\xi_{ij} = \epsilon_{ij} + \beta\delta_i$.

The SMN-CCM can be formulated as a generalization of the model defined in Eqs. (1)–(3), which can be obtained by considering Eq. (8) along with Eq. (3) , and the following assumptions:

$$\xi_{ij} \sim SMN_1\left(0, \beta^2\sigma_\delta^2 + \sigma_\epsilon^2; H\right), \quad j = 1, \ldots, m_i, i = 1, \ldots, n,$$
$$\epsilon_{0i} \sim SMN_1\left(0, \sigma_\epsilon^2; H\right), \quad i = n+1, \cdots, n+r,$$
$$cov\ (\delta_i, \epsilon_{ij}) = 0 \text{ for all } i, j, \quad cov\ (\delta_i, \epsilon_{0j}) = 0 \text{ for all } i, j,$$
$$cov\ (\epsilon_{ik}, \epsilon_{il}) = 0, k \neq l \text{ and } \quad cov\ (\epsilon_{0k}, \epsilon_{0l}) = 0, k \neq l.$$