Calibration model with scale mixtures of skew-normal distributions

Betsabé Grimalda Blas Achic[1]
Marcos Antonio Alves Pereira[2]
[1]Federal University of Pernambuco, Brazil
[2]Federal University of Piauí, Brazil
betsabe@de.ufpe.br, imarcosweb@gmail.com

Abstract

This work presents a new statistical linear calibration model with replication by assuming that the error model follows the family of scale mixtures of skew-normal distributions, which is a class of asymmetric thick-tailed distributions that includes the skew-normal distribution. In the literature, most of calibration models assume that the errors are normally distributed, however, the normal distribution is extremely sensitive to atypical observations and asymmetry. The estimation of the model parameters are done numerically by the EM algorithm. The new approach is applied to a real data set from chemical analysis.

1. Introduction

The calibration models are usually composed of two stages. In the first stage, dependent variables are observed in function of pre-fixed independent variables. In the second stage, only the dependent variables are observed in function of an unknown quantity. The relationship between the both independent and dependent variable is established in the two stages, thus the parameters model are estimated. In chemical analysis, the purpose of the calibration model is usually to establish a quantitative relation over the two stages, for the first stage it is between several known concentrations and their corresponding signals, and on the second stage it is between an unknown concentration and their corresponding signals, and the main interest is estimate this unknown concentration (see Blas *et al.*,2007).
This paper discusses a new calibration model with replicated response variable by assuming that the error model follows a family of scale mixtures of skew-normal (SMSN) distributions, as introduced by Branco and Dey (2001). Calibration models in the literature largely assume that the errors are normally distributed, however, the normal distribution is not suitable in the presence of atypical or discordant observations and also to the asymmetry.
We can say that a random variable Y follows a SMSN distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ, and it can be denoted as $Y \sim SMSN(\mu, \sigma^2, \lambda)$. The probability density function (pdf) of Y is given by

$$f(y) = 2\int_0^\infty \phi(y|\mu, \sigma^2\kappa(u))\Phi\left(\lambda\frac{y-\mu}{\sigma}\right) dH(u;\tau),$$

where $y \in \mathbb{R}$, $\phi(.)$ denotes the density of univariate normal distribution with mean μ and variance $\sigma^2 > 0$ and $\Phi(.)$ is the distribution function of the standard univariate normal distribution. U is a random variable with distribution function $H(.,\tau)$ and density $h(.,\tau)$ and τ is a scalar or vector parameter indexing the distribution of U. In this work we consider $\kappa(u) = 1/u$, which leads to good mathematical properties.
The SMSN family is a flexible class of distributions for robust estimation since it contains asymmetric distributions and all the symmetric class of scale mixture normal(SMN) distributions. One particular case is the skew-normal (SN) distribution which is arrived when H is degenerated, with $u = 1$. The SMSN class also includes distributions such as the skew-t (ST), skew-slash (SSL) and the skew-potential exponential(SPE) distribution.

2. Linear Calibration Model with SMSN Distributions Error

The SMSN linear calibration model is given by

$$y_{ij} = \alpha + \beta x_i + \epsilon_{ij},\ i = 1,\dots,n,\ j = 1,\dots,r_i, \quad (1)$$
$$y_{0i} = \alpha + \beta x_0 + \epsilon_{0i},\ i = n+1,\dots,n+m, \quad (2)$$

where y_{ij} and y_{0i} are observed responses for the fixed value x_i and the unknown quantity x_0, respectively. α, β and x_0 are unknown parameters. ϵ_{ij} and ϵ_{0i} are independent and identically distributed (iid) SMSN with 0 location parameter, scale parameter σ^2 and skewness parameter λ. The EM algorithm for the proposed model parameters are presented in the following.

The model (1-2) can be written hierarchically as

$$\begin{aligned}
&Y_{ij}|T_{ij} = t_{ij}, U_{ij} = u_{ij} \overset{\text{iid}}{\sim} N\left(\alpha + \beta x_i + t_{ij}\frac{\sigma\lambda\kappa(u_{ij})}{\sqrt{s}}, \frac{\sigma^2\kappa(u_{ij})}{s}\right)\\
&U_{ij} \overset{\text{iid}}{\sim} H(u_{ij};\tau),\ \ T_{ij} \overset{\text{iid}}{\sim} NH(0,1),\ i = 1,\dots,n,\ j = 1,\dots,r_i,\\
&Y_{0i}|T_{0i} = t_{0i}, U_{0i} = u_{0i} \overset{\text{iid}}{\sim} N\left(\alpha + \beta x_0 + t_{0i}\frac{\sigma\lambda\kappa(u_{0i})}{\sqrt{s_0}}, \frac{\sigma^2\kappa(u_{0i})}{s_0}\right)\\
&U_{0i} \overset{\text{iid}}{\sim} H(u_{0i};\tau),\ \ T_{0i} \overset{\text{iid}}{\sim} HN(0,1),\ i = n+1,\dots,n+m,
\end{aligned}$$

where $HN(0,1)$ denotes the half-$N(0,1)$ distribution, $s = 1 + \lambda^2\kappa(u_{ij})$ and $s_0 = 1 + \lambda^2\kappa(u_{0i})$. The parameter τ from the mixing variable is fixed previously, as recomended by Lange K. L. *et al.* (1989). Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top,\dots,\boldsymbol{y}_R^\top)$, $\boldsymbol{u} = (u_1,\dots,u_R)^\top$, $\boldsymbol{t} = (t_1,\dots,t_R)^\top$, $\boldsymbol{y_0} = (y_{01},\dots,y_{0m})^\top$, $\boldsymbol{u_0} = (u_{01},\dots,u_{0m})^\top$, $\boldsymbol{t_0} = (t_{01},\dots,t_{0m})^\top$ and $R = \sum_{i=1}^n r_i$. Then, under the hierarchical model (1-2), it follows the complete log-likelihood function $\ell_c(\boldsymbol{\theta}|\boldsymbol{y_c})$ associated with $\boldsymbol{y_c} = (\boldsymbol{y}^\top, \boldsymbol{y_0}^\top, \boldsymbol{u}^\top, \boldsymbol{u_0}^\top, \boldsymbol{t}^\top, \boldsymbol{t_0}^\top)^\top$.
Let $\boldsymbol{\theta}^{(p)} = (\alpha^{(p)}, \beta^{(p)}, \sigma^{2(p)}, \lambda^{(p)}, x_0^{(p)})^\top$ be the estimates of $\boldsymbol{\theta}$ at the pth iteration. It follows, after some simple algebra, that the conditional expectation of the complete log-likelihood function has the form

$$\begin{aligned}
Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= \mathbb{E}\left[\ell_c(\boldsymbol{\theta}|\boldsymbol{y_c})|\boldsymbol{y}, \boldsymbol{y_0}, \widehat{\boldsymbol{\theta}}^{(p)}\right] = -R\log\sigma^{2(p)} - m\log\sigma^{2(p)}\\
&-\frac{1}{2\sigma^{2(p)}}\sum_{i=1}^n\sum_{j=1}^{r_i}(y_{ij} - \alpha^{(p)} - \beta^{(p)}x_i)^2(\widehat{\kappa_{ij}}^{(p)} + \lambda^{(p)^2})\\
&-\frac{1}{2\sigma^{2(p)}}\sum_{i=1}^n\sum_{j=1}^{r_i}\widehat{t_{ij}^2}^{(p)} + \frac{\lambda^{(p)}}{\sigma^{2(p)}}\sum_{i=1}^n\sum_{j=1}^{r_i}(y_{ij} - \alpha^{(p)} - \beta^{(p)}x_i)\widehat{t_{ij}}^{(p)}\\
&-\frac{1}{2\sigma^{2(p)}}\sum_{i=n+1}^{n+m}(y_{0i} - \alpha^{(p)} - \beta^{(p)}x_0{}^{(p)})^2(\widehat{\kappa_{0i}}^{(p)} + \lambda^{(p)^2})\\
&-\frac{1}{2\sigma^{2(p)}}\sum_{i=n+1}^{n+m}\widehat{t_{0i}^2}^{(p)} + \frac{\lambda^{(p)}}{\sigma^{2(p)}}\sum_{i=n+1}^{n+m}(y_{0i} - \alpha^{(p)} - \beta^{(p)}x_0{}^{(p)})\widehat{t_{0i}}^{(p)}.
\end{aligned}$$

with $\widehat{t_{ij}} = \mathbb{E}(T_{ij}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{ij})$, $\widehat{t_{ij}^2} = \mathbb{E}(T_{ij}^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{ij})$, $\widehat{t_{0i}} = \mathbb{E}(T_{0i}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{0i})$ and $\widehat{t_{0i}^2} = \mathbb{E}(T_{0i}^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{0i})$. Thus, we have the following EM algorithm steps:

E-step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(p)}$, compute $\widehat{t_{ij}}^{(p)}$, $\widehat{t_{ij}^2}^{(p)}$, $\widehat{t_{0i}}^{(p)}$ and $\widehat{t_{0i}^2}^{(p)}$, where $\widehat{t_{ij}} = \mathbb{E}(T_{ij}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{ij})$, $\widehat{t_{ij}^2} = \mathbb{E}(T_{ij}^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{ij})$, $\widehat{t_{0i}} = \mathbb{E}(T_{0i}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{0i})$ and $\widehat{t_{0i}^2} = \mathbb{E}(T_{0i}^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{0i})$.
M-step: Update $\hat{\boldsymbol{\theta}}^{(p+1)}$ by maximizing $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$, which leads to the following closed form expressions:

$$\begin{aligned}
\hat{\alpha}^{(p+1)} &= \left[\widehat{\boldsymbol{\kappa}}^{(p)^\top}\mathbf{1_R} + \widehat{\boldsymbol{\kappa_0}}^{(p)^\top}\mathbf{1_m} + (R+m)\lambda^{(p)^2}\right]^{-1}\\
&\left[\left(\boldsymbol{y}^\top - \beta^{(p)}\boldsymbol{x}^\top\right)\widehat{\boldsymbol{\kappa}}^{(p)} + \lambda^{(p)}\left(\lambda^{(p)}\boldsymbol{y}^\top - \widehat{\boldsymbol{t}}^{(p)^\top} - \lambda^{(p)}\beta^{(p)}\boldsymbol{x}^\top\right)\mathbf{1_R} + \boldsymbol{y_0}^\top\widehat{\boldsymbol{\kappa_0}}^{(p)}\right.\\
&\left.+\left(\lambda^{(p)^2}\boldsymbol{y_0}^\top - \lambda^{(p)}\widehat{\boldsymbol{t_0}}^{(p)^\top} - \beta^{(p)}x_0^{(p)}\widehat{\boldsymbol{\kappa_0}}^{(p)^\top}\right)\mathbf{1_m} - m\beta^{(p)}x_0^{(p)}\lambda^{(p)^2}\right]\\
\hat{\beta}^{(p+1)} &= \left[\boldsymbol{x}^\top\left(\boldsymbol{D}(\widehat{\boldsymbol{\kappa}}^{(p)}) + \lambda^{(p)^2}\mathbf{I_R}\right)\boldsymbol{x} + x_0^{(p)^2}\left(\widehat{\boldsymbol{\kappa_0}}^{(p)^\top}\mathbf{1_m} + \lambda^{(p)^2}m\right)\right]^{-1}\\
&\left\{\boldsymbol{x}^\top\left[\boldsymbol{D}(\widehat{\boldsymbol{\kappa}}^{(p)})\boldsymbol{y} - \alpha^{(p)}\widehat{\boldsymbol{\kappa}}^{(p)} + \lambda^{(p)^2}\left(\boldsymbol{y} - \alpha^{(p)}\mathbf{1_R}\right) - \lambda^{(p)}\widehat{\boldsymbol{t}}^{(p)}\right] + x_0^{(p)}\times\right.\\
&\left.\left[\boldsymbol{y_0}^\top\widehat{\boldsymbol{\kappa_0}}^{(p)} + \left(\lambda^{(p)^2}\boldsymbol{y_0}^\top - \alpha^{(p)}\widehat{\boldsymbol{\kappa_0}}^{(p)^\top} - \lambda^{(p)}\widehat{\boldsymbol{t_0}}^{(p)^\top}\right)\mathbf{1_m} - m\lambda^{(p)^2}\alpha^{(p)}\right]\right\}\\
\widehat{\sigma^2}^{(p+1)} &= [2(R+m)]^{-1}\left[\left(\boldsymbol{\eta}^{(p)^\top}\boldsymbol{D}(\widehat{\boldsymbol{\kappa}}^{(p)}) + \lambda^{(p)^2}\boldsymbol{\eta}^{(p)^\top} - 2\lambda^{(p)}\widehat{\boldsymbol{t}}^{(p)^\top}\right)\boldsymbol{\eta}^{(p)} + \widehat{\boldsymbol{t}}^{2(p)^\top}\mathbf{1_R}\right.\\
&\left.+\left(\boldsymbol{\eta_0}^{(p)^\top}\boldsymbol{D}(\widehat{\kappa}_0^{(p)}) + \lambda^{(p)^2}\boldsymbol{\eta_0}^{(p)^\top} - 2\lambda^{(p)}\widehat{\boldsymbol{t_0}}^{(p)^\top}\right)\boldsymbol{\eta_0}^{(p)} + \widehat{\boldsymbol{t_0}}^{2(p)^\top}\mathbf{1_m}\right]\\
\hat{\lambda}^{(p+1)} &= \left[\boldsymbol{\eta}^{(p)^\top}\boldsymbol{\eta}^{(p)} + \boldsymbol{\eta_0}^{(p)^\top}\boldsymbol{\eta_0}^{(p)}\right]^{-1}\left[\widehat{\boldsymbol{t}}^{(p)^\top}\boldsymbol{\eta}^{(p)} + \widehat{\boldsymbol{t}_0}^{(p)^\top}\boldsymbol{\eta_0}^{(p)}\right]\\
\hat{x}_0^{(p+1)} &= \left[\beta^{(p)}\left(\widehat{\boldsymbol{\kappa_0}}^{(p)^\top}\mathbf{1_m} + m\lambda^{(p)^2}\right)\right]^{-1}\\
&\left[\boldsymbol{y_0}^\top\widehat{\boldsymbol{\kappa_0}}^{(p)} - \left(\alpha^{(p)}\widehat{\boldsymbol{\kappa_0}}^{(p)^\top} + \lambda^{(p)}\widehat{\boldsymbol{t_0}}^{(p)^\top}\right)\mathbf{1_m} + \lambda^{(p)^2}\left(\boldsymbol{y_0}^\top\mathbf{1_m} - m\alpha^{(p)}\right)\right].
\end{aligned}$$

where $\boldsymbol{\eta}^{(p)} = \boldsymbol{y} - \alpha^{(p)}\mathbf{1_R} - \beta^{(p)}\boldsymbol{x}$, $\boldsymbol{\eta_0}^{(p)} = \boldsymbol{y_0} - \left(\alpha^{(p)} + \beta^{(p)}x_0^{(p)}\right)\mathbf{1_m}$, $\boldsymbol{D}(\boldsymbol{A}) = \mathrm{Diag}(a_1, a_2, \dots)$, $\mathbf{1_k}$ denotes an k-dimensional column vector of ones and $\mathbf{I_R}$ is an identity matrix of order R.
The Fisher-information matrix is used to calculate the covariance matrices associated to the maximum-likelihood estimates.

3. Application

We fit the SMNS calibration model to the real data set discussed by Neto *et al.* (2007) which is given in Table 1. Triplicate absorbance readings $\boldsymbol{y_i} = (y_{i1}, y_{i2}, y_{i3})$ were taken for each zinc standard concentration x_i.

Table 1: *Zinc concentration (mg/l) and triplicate absorbance readings.*

Concentration	Absorbance		
x_i	y_{i1}	y_{i2}	y_{i3}
0.0	0.696	0.696	0.706
0.5	7.632	7.688	7.603
1.0	**14.804**	**14.861**	**14.731**
2.0	28.895	29.156	29.322
3.0	43.993	43.574	44.699

To show the ability of our approach to deal with chemical data, we use the triplicate absorbance readings $\boldsymbol{y_0} = (14.804, 14.861, 14.731)$ from Table 1 to represent the data from the second stage of the calibration model, which is related to the true concentration 1.0, thus the response variables y_{ij}, $i = 1, 2, 4, 5$ and $j = 1,\dots,3$ belong to the first stage calibration model. We consider the SN, ST, SSL and SPE distributions from the SMSN class and as suggested by Lange *et al.* (1989) the log-likelihood was used for choosing among values of τ.

Table 2: *Parameter estimates for the* ST, SN, SSL *and* SPE *distributions.*

Distribution	Parameters				Criteria	
	α	β	x_0	$U(\widehat{x_0})$	AIC	BIC
ST	0.497	14.195	1.002	0.003	-80.35	-76.80
($\tau = 2$)	(0.019)	(0.011)	(0.002)			
SN	0.300	14.290	1.004	0.045	-7.99	-4.45
	(0.224)	(0.112)	(0.023)			
SSL	0.491	14.198	1.002	0.004	-70.47	-66.93
($\tau = 1.5$)	(0.002)	(0.007)	(0.002)			
SPE	0.312	14.285	1.001	0.043	-9.03	-5.59
($\tau = 0.8$)	(0.251)	(0.125)	(0.022)			

Table 2 presents the parameter estimates, the estimated asymptotic standard errors and the standard uncertainty $U(\widehat{x_0})$, which is the confidence interval amplitude divided by 2, from the proposed calibration model for the ST, SN, SSL and SPE distributions with appropriates values of τ. Table 2 also shows information criteria values of $\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2k$ and $\mathrm{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + k\log(R+m)$ as a means of ranking each fitted model. The aim is to find the model with the lowest value of the selected information criterion. We observe that among the four distributions the SPE distribution has the smaller vies and the ST distribution has the smaller standard error related to the parameter x_0, and according to the both criteria the ST distribution is more suitable than the SN, SSL and SPE distributions.

4. Conclusions

In this work, we propose the SMSN calibration model which has additional parameters that can be used for adjusting skewness and heavy-tailedness simultaneously and provide more robust procedures than the ones that use the SN (and normal) distribution, and that was observed in the application section.

Acknowledgment: The first author thanks **FACEPE** partial financial support. The second author thanks the scholarship support from **CAPES**.

5. Bibliography

Blas, B., Sandoval, M. C. an Yoshida, O. S. (2007). Homoscedastic controlled calibration model. *Journal of Chemometrics*, **21**, 145 – 155.

Branco,M.D. and Dey, D.K. (2001). A class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99 – 113.

Lange, K.L., Little, J.A. and Taylor, M.G. (1989). Robust Statistical modeling using the t distribution. *Journal of the American Statistical Association*, **84**, 881 – 896.

Neto, B. B., Scarminio, I. S. and Bruns, R. E. (2007). *Como fazer experimentos: pesquisa e desenvolvimento na ciência e na indústria*. São Paulo: Editora da Unicamp.