

Laplace approximation for measurement error correction in generalized linear (mixed) models

Michela Battauz

Abstract The Laplace approximation is applied to measurement error models when the likelihood function involves high dimensional integrals. The cases considered are generalized linear models with multiple covariates measured with error and generalized linear mixed models with measurement error in the covariates. The asymptotic order of the approximation and the asymptotic properties of the Laplace-based estimator for these models are shown to depend on the measurement error variance. Simulation studies are presented to illustrate the performance of the method.

Key words: Generalized linear models; Generalized linear mixed models; High dimensional integrals; Laplace approximation; Measurement errors; Random effects; Variance components.

1 Introduction

Likelihood methods for measurement error correction often involve integrals which have no closed-form solution. In these cases, numerical methods such as Gaussian quadrature are employed. Nevertheless, when high dimensional integrals are involved, such methods become computationally demanding or even unfeasible. Two cases of such type are considered in this paper.

The first one concerns generalized linear measurement error models (GLMeMs) where the dimension of the integral depends on the number of error-prone variables. Schafer (1987) proposes the use of Gauss-Hermite quadrature in maximum likelihood analysis to approximate integrals. As pointed out by Higdon and Schafer (2001), the inclusion of more than one explanatory variable measured with error is considerably more difficult because multidimensional quadrature is quite cumbersome.

Michela Battauz

University of Udine, Dept. of Statistics, via Treppo 18 33100 Udine, e-mail: battauz@dss.uniud.it

The second case regards generalized linear mixed measurement error models (GLMMeMs). Wang *et al.* (1998) analyze the bias in parameter estimation when the measurement error is not properly taken into account and propose the SIMEX method for measurement error correction. Wang *et al.* (1999) showed that a naive application of regression calibration is not suitable for generalized linear mixed models with measurement error in the covariates. When a likelihood-based approach is used in order to adjust for measurement error in mixed models, both the integration of the random effects and the true variables are required.

This paper proposes the use of the Laplace approximation for dealing with high dimensional integrals in measurement error models when a likelihood-based approach is followed. The Laplace approximation is as follows (Barndorff-Nielsen and Cox, 1989, Chap. 6)

$$\int_{\mathbb{R}^q} e^{-r(x)} dx \simeq (2\pi)^{(q/2)} |r''(\hat{x})|^{-1/2} e^{-r(\hat{x})} \quad (1)$$

where x is a q -dimensional parameter vector and $r(x)$ is a smooth function on $\mathbb{R}^q$ with a unique minimum at $\hat{x}$. Generally, $r(x)$ is of order $O(n)$ and the typical case is $r(x) = nr_1(x)$, where $r_1(x)$ is a function presenting the same properties of $r(x)$. For fixed $q < \infty$, the order of accuracy of the Laplace approximation is then $O(1/n)$. In the case considered here an unusual type of Laplace approximation results as the order of the function $r(x)$ does not depend on a single parameter n , but also on the measurement error variances. The accuracy of the approximation requires then a specific investigation.

The article is organized as follows. The models and some notation are introduced in Section 2. The likelihood functions are presented in Section 3, while the asymptotic order of the approximation and the asymptotic properties of the estimator are presented in Section 4. Section 5 regards some simulation studies, and Section 6 contains some concluding remarks.

2 Model specifications

Consider the GLMeM. Let Y_i be the response variable on unit $i = 1, \dots, m$. Let X_i denote a vector of true covariates with dimension p_1 and Z_i denote a vector of covariates measured without error with dimension p_2 . It is assumed that observations from different units are independent. Suppose that the responses come from a generalized linear model with density

$$f_{Y_i|X_i,Z_i}(y_i|x_i,z_i) = \exp \{ [y_i\gamma_i - b(\gamma_i)] / \tau^2 - c(y_i, \tau) \}, \quad (2)$$

mean $e(Y_i) = \mu_i$ and variance $\text{var}(Y_i) = \tau^2 v(\mu_i)$. The mean is related to the linear predictor through the link function g

$$g(\mu_i) = \beta_0 + X_i^\top \beta_x + Z_i^\top \beta_z,$$

where β_0 is the intercept, β_x is a p_1 -vector of coefficients of X_i and β_z is a p_2 -vector of coefficients of Z_i . In this paper the additive measurement error structure is assumed for the measurement error model, hence

$$W_i = X_i + U_i,$$

where W_i is a p_1 -vector of measurements, U_i is a p_1 -vector of measurement errors independent of X_i that follow a normal distribution with zero mean and variance $\Sigma_u = \Sigma_u(\psi_u)$, and ψ_u is a vector of parameters. In general, measurement error problems require the availability of extra information on the measurement error distribution (Carroll *et al.*, 2006, §2.3). In this paper the measurement error variance Σ_u is assumed to be known. Conditional on X_i and Z_i , the distribution of Y_i is assumed independent of W_i , which means that measurement error is nondifferential (Carroll *et al.*, 2006, §2.5). Furthermore, given the true variables X_i , the measurements W_i are assumed independent of any other variable. The structural model for the true variables is

$$X_i = \alpha Z_i + e_i, \quad (3)$$

where α is a $p_1 \times p_2$ -matrix of coefficients and e_i is a p_1 -vector of residuals that follow a normal distribution with zero mean and variance matrix $\Sigma_x = \Sigma_x(\psi_x)$, and ψ_x is a vector of parameters.

In case of GLMMs, let Y_{ij} denote the response variable of unit j in cluster i , where $j = 1, \dots, n_i$ and $i = 1, \dots, m$. The vector of p_1 variables measured with error is denoted X_{ij} and the vector of p_2 variables measured without error is denoted Z_{ij} . Conditional on the random effects, the responses come from a generalized linear model with predictors

$$\eta_{ij} = \beta_0 + X_{ij}^\top \beta_x + Z_{ij}^\top \beta_z + S_{ij}^\top b_i,$$

where β_x and β_z are vector of coefficients with dimensions p_1 and p_2 , S_{ij} is a subset of Z_{ij} with dimension q , b_i is a vector of random effects with dimension q that follows a normal distribution with zero mean and variance $D = D(\psi_b)$, and ψ_b is a vector of parameters. The random effects are independent of any other variable. The conditional means of the response variable are denoted by μ_{ij} , satisfying $g(\mu_{ij}) = \eta_{ij}$ for some link function g , and the conditional variance is $\tau^2 v(\mu_{ij})$. The density of Y_{ij} is

$$f_{Y_{ij}|b_i, X_{ij}, Z_{ij}}(y_{ij}|b_i, x_{ij}, z_{ij}) = \exp \{ [y_{ij}\gamma_i - b(\gamma_i)] / \tau^2 - c(y_{ij}, \tau) \}. \quad (4)$$

Note that the situation considered here concerns variables measured with error associated with fixed coefficients. The nondifferential measurement error assumption is taken. The specification of the model is completed by defining the measurement error model, that is

$$W_{ij} = X_{ij} + U_{ij},$$

where W_{ij} is a p_1 -vector of measurements, U_{ij} is a p_1 -vector of normally distributed measurement errors with zero mean and variance $\Sigma_u = \Sigma_u(\psi_u)$. The measurement

errors are assumed independent of the true variables and independent across units. Measurement error variances are assumed known. Furthermore, conditional on X_i , the measurements are assumed independent of other variables. Two cases are considered for the model for X_{ij} : the homogeneous and the heterogeneous case (Wang *et al.*, 1998). In the homogeneous case the true variables are marginally independent, that is

$$X_{ij} = \alpha^\top Z_{ij} + e_{ij},$$

where α is a p_2 -vector of coefficients and e_{ij} is a p_1 -vector of independent $N(0, \Sigma_x)$. In the heterogeneous case the model for the true variables presents random coefficients. Here we consider the case of a random intercept model

$$X_{ij} = \alpha^\top Z_{ij} + a_i + e_{ij},$$

where a_i are independent $N(0, \sigma_a^2)$, independent of other variables, and e_{ij} is a p_1 -vector of independent $N(0, \Sigma_x)$. In both cases, the vector of parameters entering Σ_x is ψ_x .

3 Likelihood functions

The likelihood function for GLMeMs can be expressed as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^m \int_{\mathbb{R}^{p_1}} f(Y_i|X_i, Z_i) f(W_i|X_i) f(X_i|Z_i) dX_i \\ &= \prod_{i=1}^m \int_{\mathbb{R}^{p_1}} \exp\{-r(X_i)\} dX_i, \end{aligned} \quad (5)$$

where $r(X_i) = -\log f(Y_i|X_i, Z_i) - \log f(W_i|X_i) - \log f(X_i|Z_i)$ and $\theta = (\beta_x^\top, \beta_z^\top, \tau^2, \psi_u^\top, \text{vec}(\alpha)^\top, \phi_x^\top)^\top$ is the vector of all the parameters. Let $\hat{X}_i$ denote the minimizer of $r(X_i)$ satisfying the equation $r'(X_i) = 0$. The Laplace approximation gives the following expression for the log-likelihood function $\tilde{\ell}(\theta) = -\sum_{i=1}^m r(\hat{X}_i) - \frac{1}{2} \sum_i \log |r''(\hat{X}_i)|$, where $\tilde{\ell}(\theta)$ denotes the approximated log-likelihood function and

$$r''(\hat{X}_i) = \frac{1}{\tau^2} \omega_i \beta_x \beta_x^\top + \Sigma_u^{-1} + \Sigma_x^{-1}, \quad (6)$$

where $\omega_i = [v(\mu_i) g_\mu^2(\mu_i)]^{-1}$ and $g_\mu(\mu_i)$ is the first derivative of $g(\mu_i)$ with respect to μ_i .

For simplicity, in mixed models we consider a single variable measured with error, that is $p_1 = 1$. Define $Y_i = (Y_{i1}, \dots, Y_{ini})^\top$, $X_i = (X_{i1}, \dots, X_{ini})^\top$, $Z_i = (Z_{i1}, \dots, Z_{ini})^\top$ and S_i and W_i similarly. In the homogeneous case, the vector of all the parameters is $\theta = (\beta_x^\top, \beta_z^\top, \tau^2, \psi_u^\top, \alpha^\top, \phi_x^\top, \psi_b)^\top$ and the likelihood function results

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^m \int_{\mathbf{R}^q} \int_{\mathbf{R}^{n_i}} \left(\prod_{j=1}^{n_i} f(Y_{ij}|X_{ij}, Z_{ij}, b_i) f(W_{ij}|X_{ij}) f(X_{ij}|Z_{ij}) \right) f(b_i) dX_i db_i \\
&= \prod_{i=1}^m \int_{\mathbf{R}^q} \int_{\mathbf{R}^{n_i}} \exp\{-r(b_i, X_i)\} dX_i db_i,
\end{aligned} \tag{7}$$

where $r(b_i, X_i) = -\sum_j (\log f(Y_{ij}|X_{ij}, Z_{ij}, b_i) + \log f(W_{ij}|X_{ij}) + \log f(X_{ij}|Z_{ij})) - \log f(b_i)$. Let $(\hat{b}_i, \hat{X}_i)^\top$ denote the minimizer of $r(b_i, X_i)$ satisfying the equation $r'(b_i, X_i) = 0$. The log-likelihood function is then approximated as $\tilde{\ell}(\theta) = -\sum_{i=1}^m \sum_{j=1}^{n_i} r(\hat{b}_i, \hat{X}_i) - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^{n_i} \log |r''(\hat{b}_i, \hat{X}_i)|$, where $r''(b_i, X_i)$ has various components, that are

$$\begin{aligned}
\frac{\partial^2 r(b_i, X_i)}{\partial b_i \partial b_i^\top} &= \frac{1}{\tau^2} \sum_{j=1}^{n_i} \omega_{ij} S_{ij} S_{ij}^\top + D^{-1} = O(n_i), \\
\frac{\partial^2 r(b_i, X_i)}{\partial X_{ij}^2} &= \frac{1}{\tau^2} \omega_{ij} \beta_x^2 + \frac{1}{\sigma_u^2} + \frac{1}{\sigma_x^2} = O(\sigma_u^{-2}), \\
\frac{\partial^2 r(b_i, X_i)}{\partial X_{ij} \partial X_{i'j}} &= 0, \quad \frac{\partial^2 r(b_i, X_i)}{\partial b_i \partial X_{ij}} = \frac{1}{\tau^2} S_{ij} \omega_{ij} \beta_x = O(1),
\end{aligned} \tag{8}$$

where $\omega_{ij} = [\nu(\mu_{ij}) g_\mu^2(\mu_{ij})]^{-1}$, $g_\mu(\mu_{ij})$ is the first derivative of $g(\mu_{ij})$ with respect to μ_{ij} , σ_u^2 represents the measurement error variance and σ_x^2 denotes the variance of X_{ij} . The asymptotic order of the second derivatives have been specified because they will be used in Section 4.

In the heterogeneous case, $\theta = (\beta_x^\top, \beta_z^\top, \tau^2, \psi_u^\top, \alpha^\top, \phi_x^\top, \psi_b, \sigma_a^2)^\top$ and the likelihood function is

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^m \int_{\mathbf{R}^q} \int_{\mathbf{R}} \int_{\mathbf{R}^{n_i}} \left(\prod_{j=1}^{n_i} f(Y_{ij}|X_{ij}, Z_{ij}, b_i) f(W_{ij}|X_{ij}) f(X_{ij}|Z_{ij}, a_i) \right) f(a_i) f(b_i) dX_i da_i db_i \\
&= \prod_{i=1}^m \int_{\mathbf{R}^q} \int_{\mathbf{R}} \int_{\mathbf{R}^{n_i}} \exp\{-r(a_i, b_i, X_i)\} dX_i da_i db_i,
\end{aligned} \tag{9}$$

where $r(a_i, b_i, X_i) = -\sum_j (\log f(Y_{ij}|X_{ij}, Z_{ij}, b_i) + \log f(W_{ij}|X_{ij}) + \log f(X_{ij}|Z_{ij}, a_i)) - \log f(a_i) - \log f(b_i)$. The Laplace approximation gives then $\tilde{\ell}(\theta) = -\sum_{i=1}^m \sum_{j=1}^{n_i} r(\hat{a}_i, \hat{b}_i, \hat{X}_i) - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^{n_i} \log |r''(\hat{a}_i, \hat{b}_i, \hat{X}_i)|$, where $(\hat{a}_i, \hat{b}_i, \hat{X}_i)^\top$ denote the minimizer of $r(a_i, b_i, X_i)$. The components of $r''(a_i, b_i, X_i)$ are those of (8) and

$$\begin{aligned}
\frac{\partial^2 r(a_i, b_i, X_i)}{\partial a_i^2} &= \frac{n_i}{\sigma_x^2} + \frac{1}{\sigma_a^2} = O(n_i), \\
\frac{\partial^2 r(a_i, b_i, X_i)}{\partial a_i \partial b_i} &= 0, \quad \frac{\partial^2 r(a_i, b_i, X_i)}{\partial a_i \partial X_{ij}} = \frac{1}{\sigma_x^2} = O(1),
\end{aligned} \tag{10}$$

where ω_{ij} is defined as in the homogeneous case.

4 Asymptotic properties

This section concerns the asymptotic order of the approximation and the asymptotic properties of the Laplace-based estimator.

The proofs are not reported here because of limited space, but they can be obtained along the lines of Vonesh (1996), who derived the rate of convergence of the estimator based on Laplace approximation for nonlinear mixed effects models.

In GLMeMs, the order of accuracy of the Laplace approximation results to depend on the magnitude of the largest measurement error variance. In particular the log-likelihood function (5) is approximated as

$$\ell(\theta) = \tilde{\ell}(\theta) + mO(\max_s \sigma_{us}^4),$$

where σ_{us}^2 is the measurement error variance of the s th variable. Let $\hat{\theta}$ be the Laplace-based maximum likelihood estimate satisfying $\tilde{\ell}_r(\hat{\theta}) = 0$. Under suitable regularity conditions on $\ell(\theta)$ and provided that $(\hat{\theta} - \theta) = o_p(1)$, it is possible to show that

$$(\hat{\theta} - \theta) = O_p \left[\max(m^{-1/2}, \max_s \sigma_{us}^4) \right].$$

Consider now the GLMMeMs. The order of accuracy of the Laplace approximation depends on both the measurement error variance and the number of units within clusters. Specifically, the log-likelihood functions (7) and (9) are approximated as

$$\ell(\theta) = \tilde{\ell}(\theta) + mO(\max(\sigma_u^4, \min(n_i)^{-1})).$$

Under suitable regularity conditions on $\ell(\theta)$ and provided that $(\hat{\theta} - \theta) = o_p(1)$, the rate of convergence of the Laplace-based maximum likelihood estimator is

$$(\hat{\theta} - \theta) = O_p(\max(m^{-1/2}, \sigma_u^4, \min(n_i)^{-1})).$$

5 Simulation studies

In order to evaluate the behavior of the estimator for GLMeMs when $m < \infty$ and $\sigma_{us} > 0$ for all s , and the behavior of the estimator for GLMMeMs when $m < \infty$, $n_i < \infty$ and $\sigma_u > 0$, some simulation studies have been conducted. The standard deviation of the true variables is set $\sigma_x = 1$ and measurement error standard deviations are taken constant across different variables and equal to 0.655, 0.5 and 0.333, so that the reliabilities result around 0.7, 0.8 and 0.9. Various estimation methods have been considered for each simulated dataset. The true method uses the true covariates and represents a benchmark. The naive method uses the observed covariates and ignores the presence of measurement error. Finally, the Laplace-based correction proposed in this paper is referred to as the adjusted method.

Table 1 Simulation of Logistic measurement error model

method	σ_u	Bias	RMSE
true	0.655	0.067	0.117
naive	0.655	-0.265	0.124
adjusted	0.655	0.057	0.180
true	0.500	0.053	0.139
naive	0.500	-0.169	0.109
adjusted	0.500	0.064	0.201
true	0.333	0.058	0.126
naive	0.333	-0.059	0.100
adjusted	0.333	0.062	0.159

For the GLMeM, a logistic regression model have been considered. Two variables measured with error have been generated. The true variables have zero mean and correlation equal to 0.5, while measurement errors are taken uncorrelated. The elements of β_x are both set equal to 1, while the intercept β_0 is set equal equal to 0. No covariates measured without error have been included. The number of observations m is taken equal to 100. There are 1000 simulations for each parameter setting. Table 1 reports the results. Since the two variables are perfectly symmetric, the coefficient of the second one is omitted.

Table 2 Simulation of logistic model in the homogeneous case

method	σ_u	n	β_x		β_z		ψ	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
true	0.655	5	0.017	0.224	0.011	0.164	-0.017	0.329
naive	0.655	5	-0.829	0.840	-0.162	0.212	-0.181	0.297
adjusted	0.655	5	-0.080	0.409	-0.014	0.192	-0.138	0.454
true	0.655	10	0.012	0.145	0.001	0.105	-0.006	0.202
naive	0.655	10	-0.831	0.836	-0.169	0.190	-0.161	0.221
adjusted	0.655	10	-0.090	0.251	-0.018	0.122	-0.081	0.253
true	0.500	5	0.008	0.212	0.012	0.152	-0.030	0.322
naive	0.500	5	-0.593	0.612	-0.112	0.176	-0.143	0.293
adjusted	0.500	5	-0.054	0.274	-0.004	0.167	-0.118	0.358
true	0.500	10	0.016	0.145	0.006	0.105	-0.012	0.205
naive	0.500	10	-0.587	0.596	-0.119	0.152	-0.128	0.210
adjusted	0.500	10	-0.032	0.190	-0.005	0.118	-0.066	0.226
true	0.333	5	0.017	0.217	0.015	0.161	-0.025	0.338
naive	0.333	5	-0.305	0.355	-0.050	0.158	-0.088	0.308
adjusted	0.333	5	0.000	0.243	0.012	0.167	-0.066	0.351
true	0.333	10	0.008	0.145	0.000	0.105	-0.013	0.200
naive	0.333	10	-0.311	0.335	-0.065	0.118	-0.075	0.195
adjusted	0.333	10	-0.004	0.164	-0.001	0.110	-0.035	0.210

A logistic regression model with random intercepts in the homogeneous case is been considered for GLMMeMs. One variable measured with error has been taken. The number of clusters m is taken equal to 100, and the number of units within clusters is constant across groups and it is $n_i = n = 5$ and 10. One variable measured

without error has been generated from a standard normal distribution. The coefficients are $\beta_0 = 0$, $\beta_x = 2$ and $\beta_z = 1$. The variance of the random intercepts, ψ , has been taken equal to 0.5. Results are based on 1000 datasets for each set of parameter values and are displayed in Table 2. The true and the naive method use the Laplace approximation for the integration of the random effects.

6 Discussion

The use of the Laplace approximation in measurement error models gives encouraging results. The method results to be applicable to problems involving high-dimensional integrals where other solutions become impractical or computationally intensive. Furthermore, it could be a convenient choice even when other methods are available in terms of computational time. A drawback of a likelihood-based approach is that it requires the specification of the distribution of the latent variables and the measurement error. Such an assumption could be restrictive and lead to inconsistent estimates and misleading inference if the distribution assumption is violated. Li and Hsiao (2004) consider the estimation of generalized linear models with covariate measurement error without making distributional assumptions on the latent variables or the measurement errors.

References

- Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman & Hall Ltd.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.
- Higdon, R. and Schafer, D. W. (2001). Maximum likelihood computations for regression with measurement error. *Computational Statistics & Data Analysis*, **35**(3), 283–299.
- Li, T. and Hsiao, C. (2004). Robust estimation of generalized linear models with measurement errors. *Journal of Econometrics*, **118**(1-2), 51–65.
- Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, **74**, 385–391.
- Vonesh, E. F. (1996). A note on the use of Laplace’s approximation for nonlinear mixed-effects models. *Biometrika*, **83**, 447–452.
- Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, **93**, 249–261.
- Wang, N., Lin, X., and Gutierrez, R. G. (1999). A bias correction regression calibration approach in generalized linear mixed measurement error models. *Communications in Statistics: Theory and Methods*, **28**, 217–232.