ON THE CONSEQUENCES OF MISSPECIFING ASSUMPTIONS CONCERNING RESIDUALS DISTRIBUTION IN A REPEATED MEASURES AND NONLINEAR MIXED MODELLING CONTEXT

Rachid el Halimi and Jordi Ocaña

Departament d'Estadística

Universitat de Barcelona

Abstract

In this paper we describe the results of a simulation study performed to elucidate the robustness of the Lindstrom and Bates (1990) approximation method under non-normality of the residuals, under different situations. Concerning the fixed effects, the observed coverage probabilities and the true bias and mean square error values, show that some aspects of this inferential approach are not completely reliable. When the true distribution of the residuals is asymmetrical, the true coverage is markedly lower than the nominal one. The best results are obtained for the skew normal distribution, and not for the normal distribution. On the other hand, the results are partially reversed concerning the random effects. Soybean genotypes data are used to illustrate the methods and to motivate the simulation scenarios.

1. Motivation and Introduction

The nonlinear mixed effects model is used to represent data in pharmacokinetics (Davidian and Giltinan, 1995), breast cancer dynamics (El Halimi et al., 2003), growth curves in Soybean genotypes data (Pinheiro and Bates, 2000) and other areas, where the within-individual model is a function of individual-scientific, scientifically meaningful parameters. It is well known that maximum likelihood estimation for nonlinear mixed effects models leads to a cumbersome integration problem, because random parameters appear inside the nonlinear expectation

function. To avoid this problem, several approximation have been proposed. The parametric approach to non-linear mixed-effects modeling using the LB-method (Lindstrom and Bates, 1990) is, essentially, based on the standard assumption of normality of the errors and random effects. But these assumptions may not always be realistic or, in any case, difficult to verify as they are not directly observed. In this paper we investigate the impact of the non-normality conditions on estimating fixed and random components parameters, via a Monte-Carlo simulation study by considering the Soybean genotypes model reported in Davidian and Giltinan (1995) and analyzed by Pinheiro and Bates (2000). Typical profiles are displayed in Figure 1, where the response of leaf weight are plotted by subject.

The goal of the study was to compare the growth patterns of two soybean genotypes, a commercial variety, Forrest (F) and an experimental strain, Plan Introduction #416937 (P). Data were collected during three years, from 1988 to 1990. At the beginning of the growing season in each year, 16 plots were planted with seeds; 8 plots with each genotype. Each plot was sampled eight to ten times at approximately weekly intervals. At each sampling time, six plants were randomly selected from each plot, leaves from this plant were weighted, and the average leaf weight per plant (in g) was calculated for each plot. Different plots in different sites were used in different years. The logistic model derived from Pinheiro and Bates (2000) is an appropriate characterization of leaf weight response, where the parameter may vary across subjects.

$$
\begin{aligned}
& y_{ij} = \frac{\delta_{1i}}{1+\exp\left[(\delta_{2i} - t_{ij})/\delta_{3i}\right]} + e_{ij} \\
& \begin{cases} \delta_{1i} = \delta_1 + \eta_{1i} \\ \delta_{2i} = \delta_2 + \eta_{2i} \\ \delta_{3i} = \delta_3 + \eta_{3i} \end{cases} .
\end{aligned} \tag{1.1}
$$

where y_{ij} represents the average leaf weight/plant in subject i, $i = 1, \cdots, 48$, at time t_{ij}. The random effects $\eta_i = (\eta_{1i},\ \eta_{2i},\ \eta_3)'$ are (0, D) and the e_{ij} are (0, σ^2) and are independent of the η_i. The association of the fixed effects δ with the random effects vector is represented by the

linear function above, where the subject-specific parameters δ_i are independent across i. But, as some of the analyses presented in El Halimi et al. (2003) and the profile of qq-norm of residuals (under homogeneity assumption) displayed in Figure 2 suggest, this assumption may not always be realistic. These violations of the assumptions of the model pose questions on the validity of the inferences made during the modeling process. As a first approach to answering these questions, we performed a simulation study emulating the conditions of the soybean genotypes studies described above.

2. Simulation study on the distributional assumptions

We carried out several simulation studies in which data were generated according to the soybean genotypes model given in equation (1.1), with known "population" or "true" parameter values. For fixed effects, these values were taken as $\delta = (19.26, 55, 8.4)'$. For random effects, the covariance matrix was

$$D = \begin{bmatrix} 25 & 2.50 & 4.00 \\ & 8.00 & 2.32 \\ & & 2.00 \end{bmatrix}.$$

The random effects were generated according to an expression equivalent to $\eta_i = Lh$, where h stands for a standardised version of the vector of random effects (common for all i) , generated from a normal distribution with zero mean and unit variance, and L stands for the lower triangular matrix resulting from the Cholesky decomposition of a covariance matrix D. These values were chosen near to the estimated values given by the splus 2000 implementation of LB-method (*nlme* function) and according to the maximum likelihood variant of the estimation procedure. The residuals or errors were generated in similar way, first as i.i.d standardised values and subsequently converted to values with standard deviation σ (in this particular cases σ=1) and according to the following marginal distributions:

N- Normal distribution, which represent the case where the usual assumption of normality on the errors is valid.

SN- Skew normal distribution with location parameter (a typical or central value that best describes the data, its effect is to translate the graph, relative to the standard normal distribution) $\zeta = -0.05$, scale parameter (its effect is to stretch out the graph) $\psi = 0.2$ and shape parameter (its effect is to allow the distribution to take a variety of shapes, depending on its value) $\varphi = 0.02$, as described in Azzalini (1986).

NPM- Non-parametric gaussian kernel estimated from experimental data with the optimal bandwidth implemented in S-plus for the gaussian kernel by Venables (1997) (*width.SJ* function).

E- Exponential distribution.

G- Gamma distribution.

The last two distributions (G and E) were used to represent a situation where the true distribution of errors is not symmetrically distributed and has heavier tails than expected from a normal distribution.

For each possible distribution of the residuals, series of 1000 simulated data sets were generated and processed to fit a nonlinear mixed model like (1.1), via the LB-method, maintaining the same starting values for ML estimation procedure. The resulting set of parameter estimates for each series was used to compute the summaries used to evaluate the performance of the inferential methods, like true coverage of the confidence intervals or the true estimator biases.

3. Results

Figures 3 to 5 contain box-plots for the "mixed" parameter estimates (fixed and random components) from all simulation runs. For each box-plot, the character in the abscises axis represents the simulated distribution residuals, while the simulated distribution of random effects was maintained normal. The horizontal continuous straight line represents the true value of the corresponding parameter. Thus, in this way, the box-plots give a graphical idea of the bias, standard deviation and dispersion of the "mixed" effects estimators, under different possible distributions.

Table 1 contains simulation summary statistics for the fixed-effects parameters estimates. In this table, if $\hat{\delta}_k$ stands for a fixed effect or parameter estimate at the k-th simulation replication and δ_T for the true value of the parameter (value used to generate the data), the summary measures for the estimators of the fixed effects are defined as:

MEAN: denotes the average of each series of 1000 estimates, $\frac{1}{1000}\sum_{k=1}^{1000}\hat{\delta}_k$,

BIAS: corresponds to the simulation estimate of bias, that is, MEAN – δ_T, the bias of a statistic indicates, on average, how much the estimator will over- or underestimate the "true" parameter value.

$\mathrm{C.I}_{\mathrm{Bias}}$: denotes approximate confidence intervals for bias, $\mathrm{BIAS} \pm \mathrm{Z}_\alpha \sqrt{\frac{S^2_{\hat{\delta}}}{1000}}$, where Z_α is the normal critical value at a 95% confidence level and $S^2_{\hat{\delta}} = \sum_{k=1}^{1000}\left(\hat{\delta}_k - \mathrm{MEAN}\right)^2 / 999$.

MSE: is the mean squared error, $\sum_{k=1}^{1000}\left(\hat{\delta}_k - \delta_T\right)^2 / 1000$, is a measure of its accuracy that takes into account both, bias and standard error.

COVERAGE-PROBABILITY: denotes the observed coverage of t-based 95% confidence intervals computed using the model-based standard errors and t-distribution critical values based on 362 degrees of freedom (the intervals provided by the *nlme* function),

PROB. LOW (PROB. UPP): denotes the proportions of δ_T lower (upper) than the lower (upper) bound of the confidence intervals (that is, they correspond to non-coverage probabilities).

Finally, Average width: denotes the arithmetic average of the 1000 observed lengths of the asymptotic intervals described above.

For the δ_1 parameter, from Figure 3 and Table 1 we infer that virtually there are not differences between the results for the five distributions under consideration, with respect to coverage, bias and MSE. For this last measure, the results are similar with values ranging between 0.54 and 0.67. None of the confidence interval coverages attain the nominal 95% value, but there is acceptable robustness, with coverages ranging from 94% under the sN residuals distribution to

91.5% under the exponential distribution. The NN case has a 92.1% coverage. The last column of Table 1 also shows very homogeneous results with respect to the width of the intervals. Additionally, the intervals are appreciably equitailed.

For the δ_2 parameter, the results shown in Table 1 demonstrate a negative bias indicating that, on average, the true parameter value is underestimated. The bias and MSE values are nearly identical for this parameter, except for the sN distribution, in which these measures are considerably smaller. The confidence intervals perform poorly in all cases, except for the sN residuals distribution. The coverage ranges from 79.9% for the exponential distribution (and 84.6% for the gamma distribution) to 93.5% for the sN distribution, with a poor 83.2% value for normal residuals. In correspondence with the coverages, the widths of the intervals show a similar tendency, with the best (shorter) value obtained under the sN distribution. The intervals have appreciably unequal tails, except for the sN case.

For the δ_3 parameter, we find that under model (1.1) and using the LB-method, the true parameter value is also systematically underestimated. We remark that the sN residuals distribution gives also the best results, not only from the point of view of the MSE values (0.06), but also from the point of view of coverage probability (90.7%) and equitaildness, but lower than the nominal 95% level. In all cases, the coverage probabilities are poor, ranging between 78.8% (exponential) to 90.7% for (sN), with 80% in the normal case, indicating a low robustness of the LB-method for this parameter. Again, all intervals have similar mean lengths and are not equitailed, except for the sN distribution, which is also associated to the best values.

These simulations confirm that the LB-method is not very efficient, even under normal conditions. But, in contrast with other simulation studies, like El Halimi et al. (2004) which is inspired in breast cancer data, a surprising result is that the best results for fixed effects are obtained when the errors are skew normal (sN). On the other hand, this result is not maintained for random components (D and σ). Concretely, the results for these model components are

displayed in Figures 4 to 5. Estimation under the nonlinear Soybean genotypes model tells a vastly different story. Here, all simulation conditions underestimate the random components of *D,* except the estimates under the sN distribution, which provide reasonable unbiased estimates, but perform considerably worse in terms of coverage probabilities, especially for the D_{33} parameter (24.5% of coverage probability). In almost all cases the coverage probabilities are lower than their nominal value for D_{12}, D_{13} and D_{33} and exceed their nominal value otherwise. Figure 5 displays the results for the parameter σ. Its estimation becomes less precise for the NPM and sN distributions, while the different measures of dispersion are in close agreement for E and G, and the magnitude of the variability is different otherwise. Regarding the coverage probabilities, results ranging from 99.3% for N to 0% for sN are observed.

4. Conclusions and discussion

Under the model (1.1) and according to the observed coverages for confidence intervals with a 95% nominal coverage, the results indicate that the LB-method is not robust when the normality assumptions of errors are violated, and even under normal conditions. In fact, the ML procedures seem not adequate even under normal conditions and for a large number of observations per subject and a large number of subjects.

The same results are also observed concerning the precision (length) of the intervals, and the bias and MSE of the point estimates. The performance of all these inferential methods is highly variable and dependent on the concrete simulated distributions and the concrete model parameters, in a complex and difficult to forecast way.

Acknowledgements

The research was supported by Instituto de Salud Carlos III FIS, grant 00/1130; and by Generalitat de Catalunya, grant 2001/SGR/00067.

References

Azzalini, A., 1986. Further results on a class of distributions which includes the normal ones. *Statistica*, 46, 199-208.

Davidian, M.; Giltinan, D.M. *Nonlinear Models for Repeated Measurement Data*; Chapman & Hall: London, **1995**.

El Halimi, R., Ocaña, J, Ruiz De Villa, M.C., Escrich, E., Solanas, M. Modeling Tumor Growth Data Using a Nonlinear Mixed-Effects Model. InterStat **2003**, http://jscs.stat.vt.edu/InterStat/ARTICLES/2003/abstracts/0309002.html-ssi

El Halimi, R., Ocaña, J., Ruiz de Villa, M.C. A simulation study on the robustness of parametric inference in a nonlinear mixed modelling context. www.mathpreprints.com, **2004**.

Lindstrom, M.J.; Bates, D.M. Nonlinear Mixed Effects Models for Repeated Measures Data. Biometrics **1990**, *46*, 673-687.

Pinheiro, J.C.; Bates, D.M. *Mixed-Effects Models in S and S-Plus*; Springer-Verlag: Berlin, **2000**.

Venables, et al. Modern Applied Statistics with S-plus, New York: Springer- Verlag, **1997**.

Table 1: summary results for fixed effect parameters estimates for 1000 simulations

DISTRIBUTIONS	RANDOM EFFECTS	ERRORS	MEAN	Bias±C.I$_{Bias}$	MSE	COVERAGE- PROBABILITY (ASYMP. INTERVALS)			Average width
						PROB LOW	PROB.	PROB. UPP	
	Fixed effect parameter δ_1: Maximum likelihood (ML)								
	N	**N**	18.99208	-0.2679243±0.04776559	0.665096	1%	92.1%	6.9%	2.876276
	N	**sN**	19.15177	-0.1082278±0.04559323	0.552286	1.9%	94%	4.1%	2.798492
	N	NPM	19.04983	-0.2101692±0.04359019	0.5382893	1.2%	93.8%	5%	2.87612
	N	**E**	18.95573	-0.3042746±0.04281623	0.6537758	0.9%	91.5%	7.6%	2.897845
	N	**G**	18.96339	-0.2966082±0.04650596	0.6504091	0.5%	91.6%	7.9%	2.899373
	Fixed effect parameter δ_2: Maximum likelihood (ML)								
	N	**N**	54.52249	-0.4775082±0.03030467	0.4668351	0.1%	83.2%	16.7%	1.95098
	N	**sN**	54.96844	-0.03156185±0.02606712	0.1776974	2.7%	93.5%	3.8%	1.591223
	N	NPM	54.56497	-0.4350315±0.03086829	0.4370394	0.1%	83.7%	16.2.9%	1.927866
	N	**E**	54.50822	-0.491783±0.03020465	0.5211324	0.3%	79.9%	19.8%	1.974908
	N	**G**	54.53612	-0.4638765±0.0309442	0.4641885	0.2%	84.6%	15.2%	1.97972
	Fixed effect parameter δ_3: Maximum likelihood (ML)								
	N	**N**	8.100563	-0.2994368±0.01805612	0.1744442	0%	80%	20%	1.117007
	N	**sN**	8.290249	-0.1097514±0.0133505	0.05839522	0.6%	90.7%	8.7%	0.8053575
	N	NPM	8.166408	-0.2335917±0.0178202	0.1371466	0%	84.5%	15.5%	1.124151
	N	**E**	8.079644	-0.3203559±0.01745259	0.1958706	0.1%	78.8%	21.1%	1.140723
	N	**G**	8.100581	-0.2994189±0.01805035	0.1743792	0.1%	80.3%	19.6%	1.137076
	Soybean genotypes model: $y_{ij} = \frac{\delta_{1i}}{1+\exp\left[(\delta_{2i} - t_{ij})/\delta_{3i}\right]} + e_{ij}$								

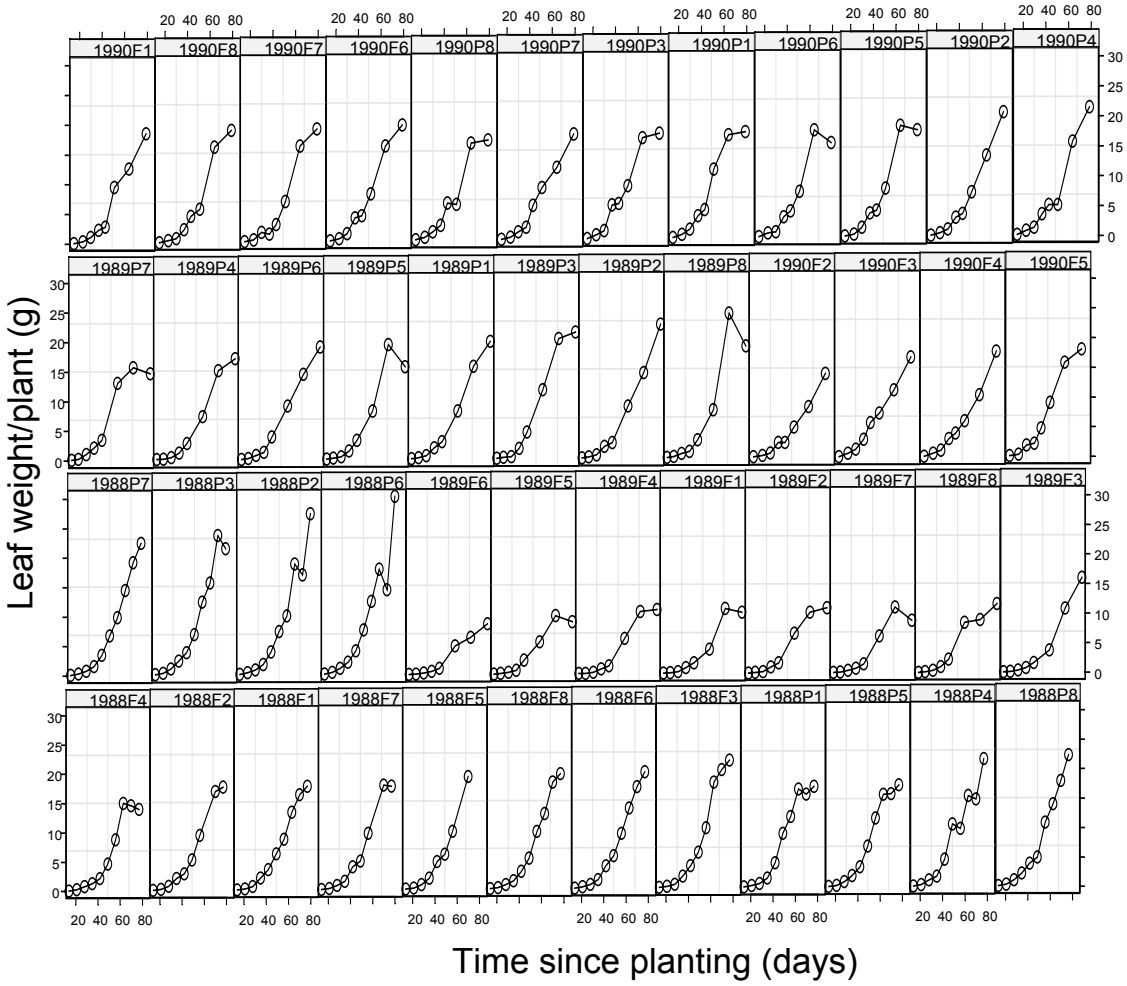

Figure1: Growth curves in Soybean genotypes data

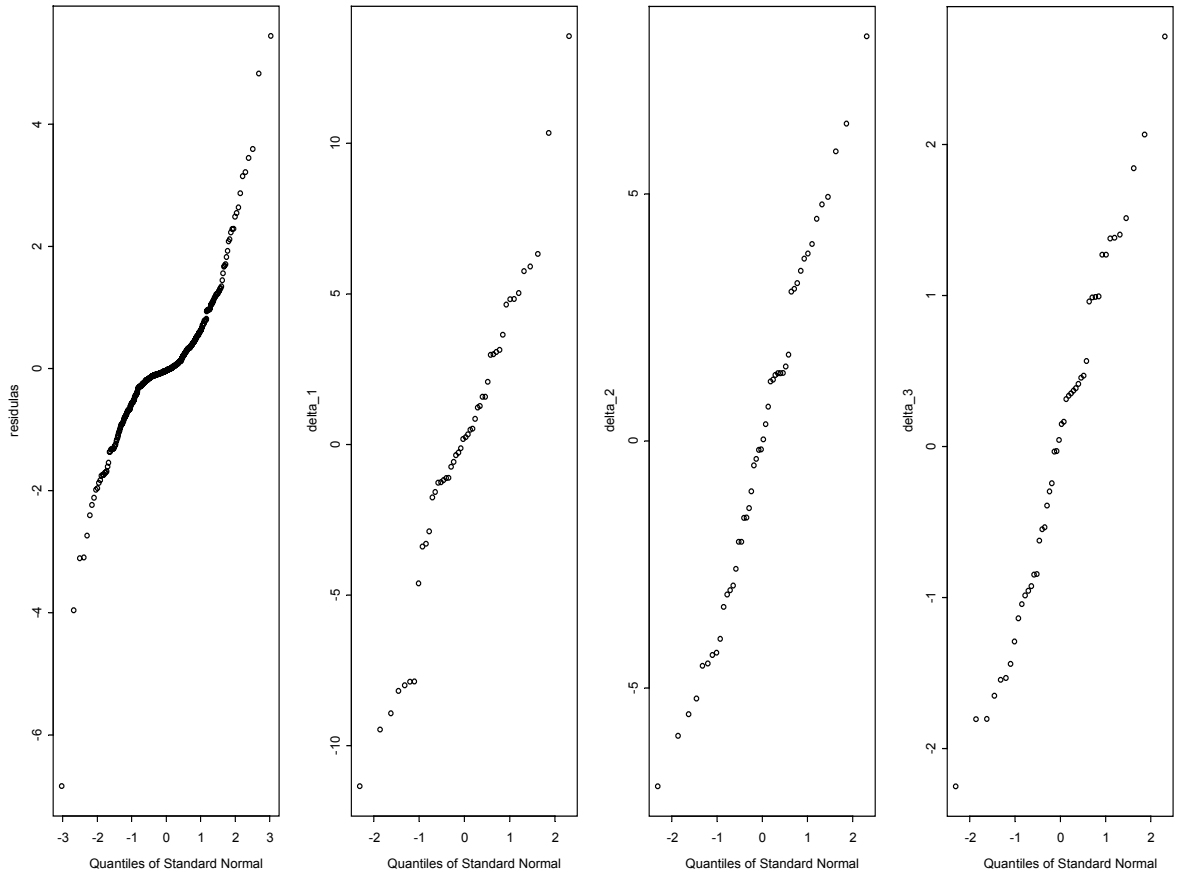

Figure 2: qq-norm of residuals and random effects for Soybean data

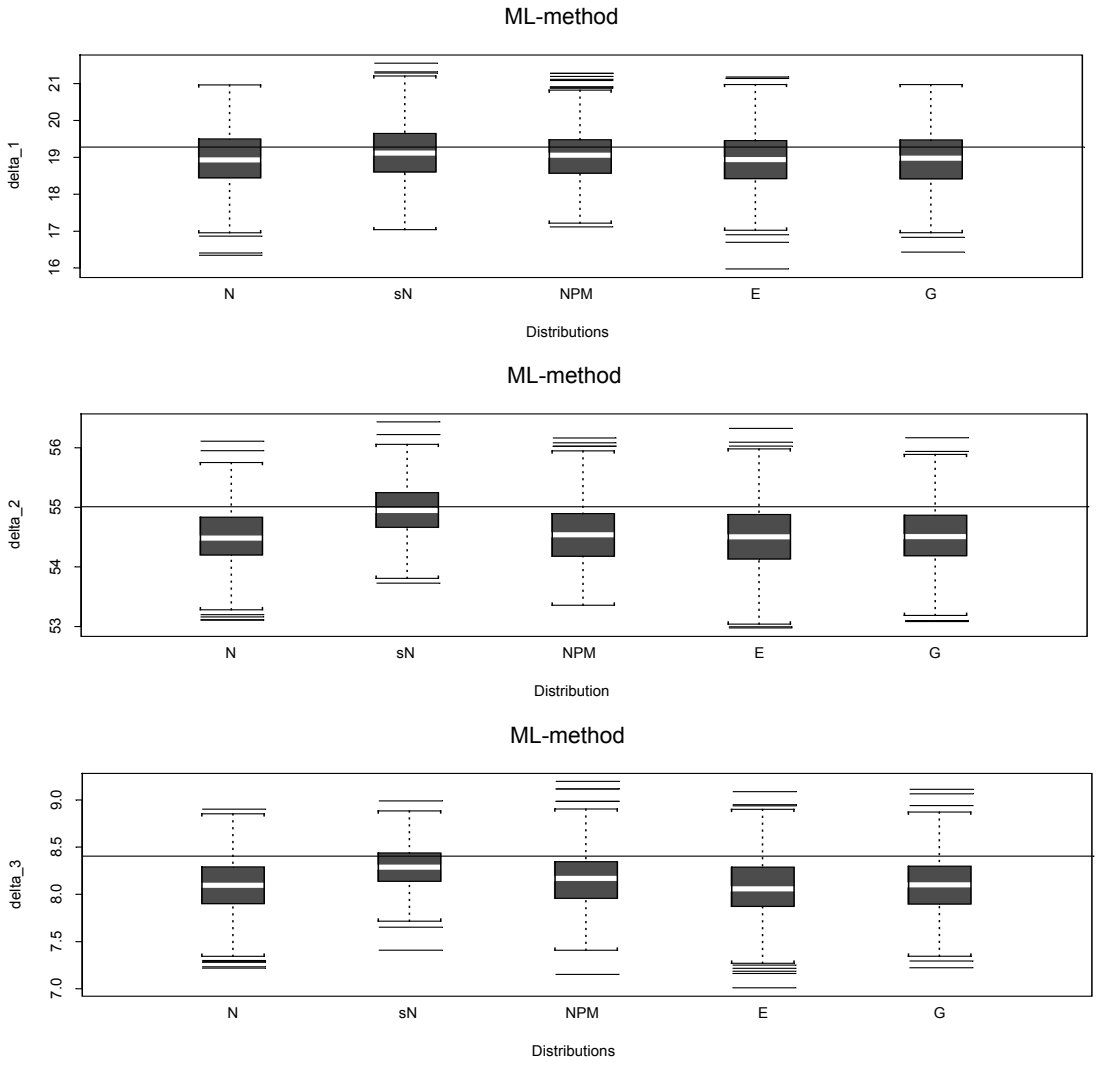

Figure 3: Box- plot of fixed effects of simulation results for soybean genotypes model

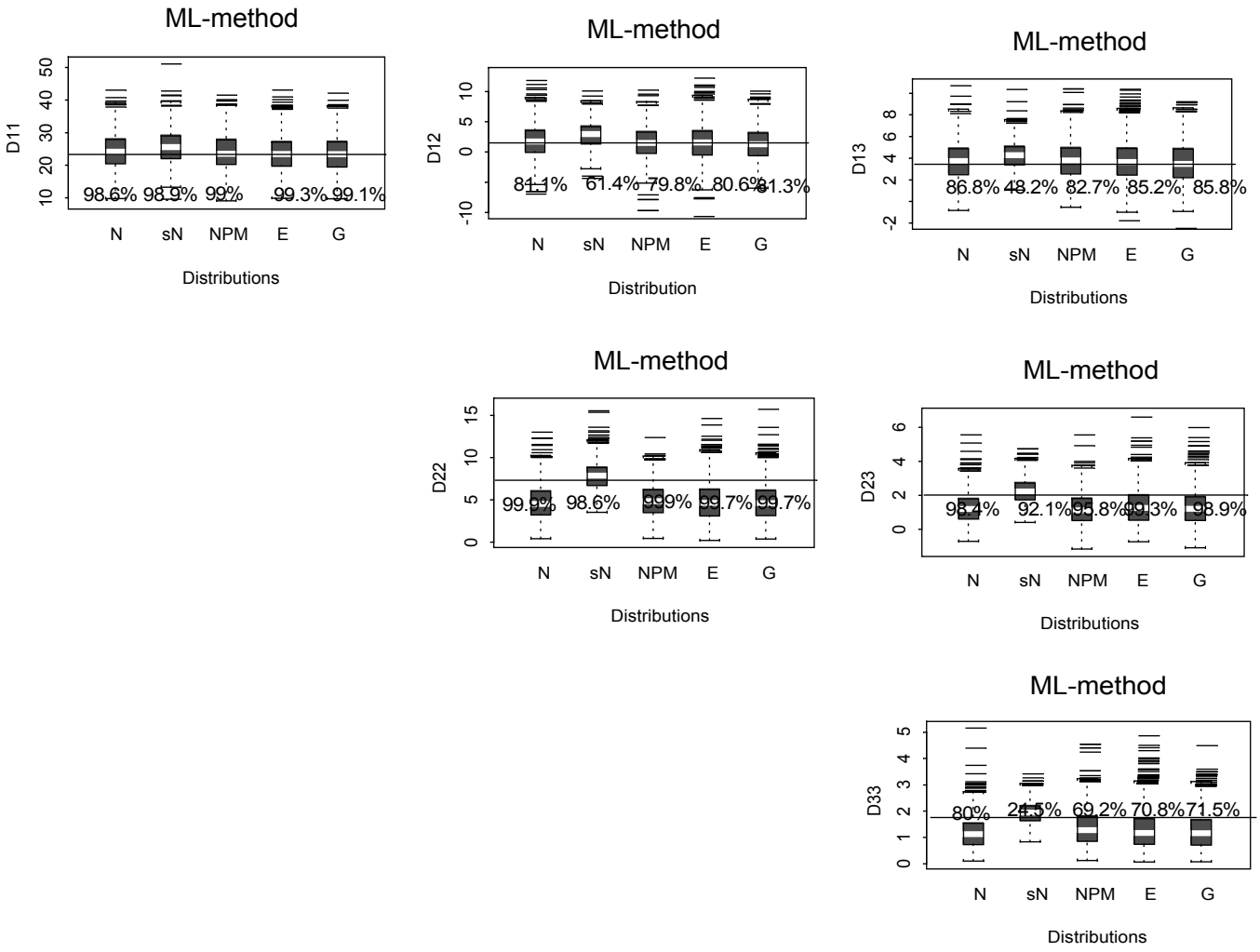

Figure 4: Box-plots for variance-covariance parameters of Soybean genotypes model

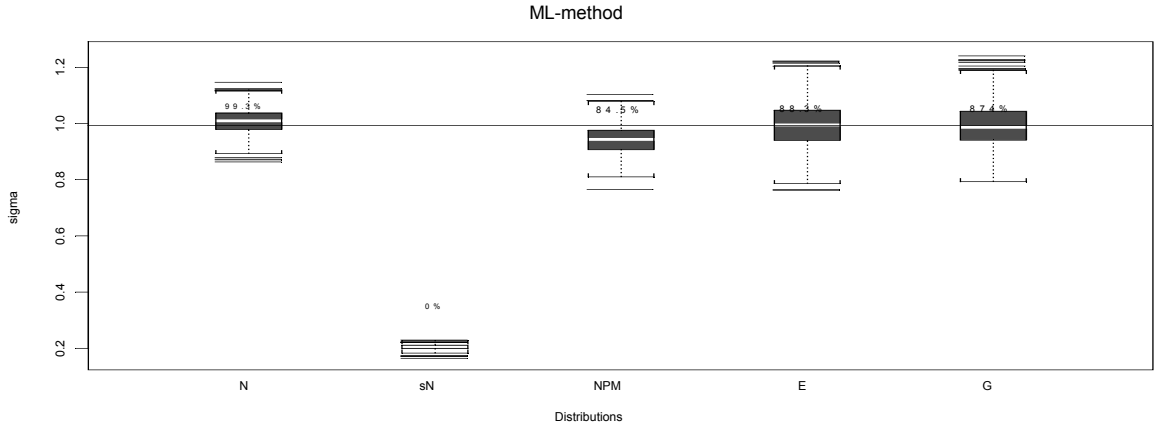

Figure 5: Box-plots for sigma parameter of genotype model