UNIVERSITÀ DEGLI STUDI FIRENZE
DISIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Alternative estimating procedures for multiple membership logit models with mixed effects: indirect inference and data cloning

Anna Gottard, Giorgio Calzolari

DISIA WORKING PAPER
2014/07

Alternative estimating procedures for multiple membership logit models with mixed effects: indirect inference and data cloning

Anna Gottard and Giorgio Calzolari

Department of Statistics, Informatics, Applications - University of Florence

E-mail: gottard@disia.unifi.it, calzolari@disia.unifi.it

July 17, 2014

Abstract

Multiple-membership logit models with random effects are logit models for clustered binary data, where each statistical unit can belong to more than one group. For these models, the likelihood function is analytically intractable. We propose two different approaches for parameter estimation: data cloning and indirect inference. Data cloning computes maximum likelihood estimates, through the posterior distribution of an adequate Bayesian model fitted on cloned data. We implement a data cloning algorithm specific for the case of multiple-membership models. Indirect inference is a non-likelihood-based method which uses an auxiliary model to select sensible estimates. We propose an auxiliary model having the same dimension of parameter space as the target model, which is particularly convenient to reach good estimates very fast. A Monte Carlo experiment compares the two approaches on a set of simulated data. We report also Bayesian posterior mean and INLA hybrid data cloning estimates for comparison. Simulations

show a negligible loss of efficiency for the indirect inference estimator, compensated by a relevant computational gain. The approaches are then illustrated with a real example on matched paired data.

Keywords: Binary data, Bradley Terry models, intractable likelihood, integrated nested Laplace approximation, non-hierarchical random effects models.

1 Introduction

Many modern statistical applications involve inference for probabilistic models in the presence of unobserved relevant factors, acting on data with a clustered structure. This structure is usually hierarchical, but more complex situations can sometimes arise. For non-hieraarchical data, an interesting class of models consists of multiple membership models, introduced by Hill and Goldstein (1998), (see also Rasbash and Browne, 2001a; Browne et al., 2001). These models extend multilevel models for hierarchical, nested data, allowing a statistical unit (in the lower level of the groups hierarchy) to belong to more than one cluster (level-two unit). Examples of application can be found, for instance, in Fielding (2002) for evaluating the performance of the cost-effectiveness of advanced level teaching groups or in Roberts and Walwyn (2012) for a randomized trial on adolescent depression with cognitive behavioural therapy when patients have more than one therapist. See also Tranmer and Browne (2013).

Multiple membership clustering is formally defined as a map g from a finite set N of units to a finite set G of clusters, with $|G| = J$. Each element i in N is mapped to a finite subset of G, say G_i. Whenever each G_i is a singleton, then the map defines an ordinary hierarchical clustering, where clusters form a partition of N. On the contrary, in case of multiple membership clustering, the population of interest is assumed to be characterized by not disjoint sub-populations. Multiple membership models assign random effects for each element of the mapped grouping, supposing that, conditionally on a latent variable, units in the population are *iid*. In case of a binary response variables, Multiple

membership logit (MML) models may be express as

$$P(Y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{u}_i) = \frac{\exp\left\{\beta_0 + \sum_{l=1}^{p} \beta_l x_{li} + \sum_{j \in G_i} w_{ij} u_j\right\}}{1 + \exp\left\{\beta_0 + \sum_{l=1}^{p} \beta_l x_{li} + \sum_{j \in G_i} w_{ij} u_j\right\}} \tag{1}$$

where Y_i ($i = 1, \ldots n$, $n = |N|$) is the binary response of interest, $\boldsymbol{x}_i$ is the p-dimensional vector of observed explanatory variables, $\boldsymbol{u}_i$ is the set of unobserved random effects affecting i, being independent and identically distributed as $N(0, \tau^2)$, assumed to be independent of the set of explanatory variables. Each unobserved random effect has a weight w_{ij}, $|w_{ij}| < \infty$, specific for each unit i, chosen a priori depending on (careful) subject matter considerations. A typical example is in medical studies, when a patient in a hospital is assisted by more than one nurse. Each nurse has an effect on the patient's progress, which is taken into account by introducing several weighted random effects, the weights taking into account the time that each nurse spent with each patient. Weights usually sum to one, but alternatives are possible. Notice that, for each unit i, the adding variance due to the linear combination of the unobserved components is

$$\mathbb{V}\left[\sum_{j \in G_i} w_{ji} u_j\right] = \tau^2 \sum_{j} w_{ji}^2$$

which is less than τ^2, the adding variance in an ordinary hierarchical random intercept model. Different choices on weights can imply this variance to be greater than τ^2, as in the case of some Bradley and Terry models (Bradley and Terry, 1952) with random effects, that will be presented in Section 5, where for each unit the variance due to the random effects is $2\,\tau^2$.

The marginal likelihood function is

$$\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) = \int \prod_{i=1}^{n} \frac{\exp\left\{y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{u})\right\}}{1 + \exp\left\{\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{u}\right\}}\, \phi_J(\boldsymbol{u}; \tau^2)\, d\boldsymbol{u} \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2) \in \Theta \subseteq \mathbb{R}^{p+2}$ are the parameters of the model, $\boldsymbol{w}_i$ is a raw vector of length J collecting the random effect weights for unit i, with a zero in place j whenever $j \notin G_i$. Finally, $\phi(\cdot)_J$ is the J-variate density for the random effects, typically Normal

with zero mean and covariance matrix $\tau^2\mathbb{I}$. A common solution for estimating $\boldsymbol{\theta}$ is maximum likelihood, where estimates are obtained via maximizing (2) (or its logarithm)

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}).$$

The basic problem in MML models is that the likelihood function has no analytical expression and a multi-dimensional integral has to be numerically computed. In other words, MML models have an intractable likelihood function, which can not be easily evaluated, as the dimension of the integral is J. Ignoring the multiple membership clustering might bring to distort results, as shown by Chung and Beretvas (2011). Composite likelihoods such as quasi-likelihood or partial-likelihood have been shown to provide seriously biased and inconsistent estimators in the case of binary responses (Rodriguez and Goldman, 1995) leading to underestimation of the random effect variance. Consequently, a different solution has to be found. Several procedures has been proposed for estimation: Rasbash and Browne (2001b) compares the IGLS algorithm and an estimation procedure based on Monte Carlo Markov Chain (MCMC), showing MCMC procedure to be numerically stabler, although slower. At the moment, the most preferable procedure for MML models estimation is based on Bayesian paradigm and MCMC methods and such procedure has been implemented in the MLwiN package Browne (2012). However, some researchers could prefer to avoid Bayesian inference as unable of making an explicit choice of a priori distributions or for preferring frequentist inference. Moreover, it has been shown (Karl et al., 2012) that different priors can result in different estimates of model parameters, at least in case of continuous response variables.

To solve this inferential issue in a non-Bayesian framework, we are here proposing two different approaches: data cloning and indirect inference.

Data cloning (DC) (Lele et al., 2007, 2010) is a novel approach, developed in the context of hierarchical mixed models, to compute maximum likelihood (ML) estimates along with their asymptotic standard error. This approach is convenient whenever the posterior distributions of an adequate Bayesian model parameters can be computed analytically or by using a Monte Carlo Markov Chains (MCMC) methodology. DC procedure has been

then extended to nonlinear state–space models (Nadeem and Lele, 2012), generalized linear mixed models with two components of dispersions (Torabi, 2012) or spatial correlation (Baghishani and Mohammadzadeh, 2011) and to time series (Torabi and Shokoohi, 2012). Ponciano et al. (2012) used data cloning to assess parameter identifiability in phylogenetic models. In the following, we show how to successfully implement DC for solving the inferential problem for MML model parameters. Moreover, we show that both the ML and the DC estimator of $\boldsymbol{\theta}$ are consistent, under some regularity conditions.

The principal limit of the DC procedure is in the use of MCMC methods, so that the reaching of an adequate accuracy of the estimates requires a very large number of simulations, having in general a high computational cost. In alternative to data cloning, we propose a further method, indirect inference, that can be a much faster solution, still approximately inheriting the properties of the standard maximum likelihood estimates. Indirect inference is a class of estimators, including the generalized and simulated methods of moments as special cases. It was developed in the econometric framework in the early nineties (e.g. Gouriéroux et al., 1993; Gallant and Tauchen, 1996), and it has been put back in popularity because of its connection with approximate Bayesian computation techniques (Beaumont et al., 2002). Indirect inference has been applied in a variety of fields, such as, for example, financial models (see, among others, Gouriéroux and Monfort, 1996; Calzolari et al., 1998; Billio and Monfort, 2003; Sentana et al., 2008), regression models with measurement error (Kuk, 1995), hierarchical multilevel binary models (Mealli and Rampichini, 1999), robust indirect estimators and tests (Genton and Ronchetti, 2003; Czellar and Ronchetti, 2010) and α–stable stochastic volatility models (Lombardi and Calzolari, 2009). Interesting surveys on indirect inference are provided by Heggland and Frigessi (2004) and Jiang and Turnbull (2004).

In Section 2 we propose a data cloning estimator for MML models, which is proved to converge to the maximum likelihood estimator as the number of clones increases. In Section 3 we propose an indirect estimator for the same class of models. In particular, we show how both estimates and standard errors can be easily and fast derived adopting

an adequate auxiliary model. The comparison of the two classes of estimators is reported in Section 4. As an example we present in Section 5 an application on Bradley and Terry models with random effects. In Section 6 we present a brief summary and discussion on the proposal.

2 Data Cloning

Data Cloning (DC) is a recently developed procedure to compute maximum likelihood estimates and the inverse of the Fisher information matrix, by utilizing as an instrument the Bayesian paradigm and MCMC procedures. The idea has been anticipated by several approaches such as, for example, the *prior feedback* in Robert (1993) and *State Augmentation for Marginal Estimation* (Doucet et al., 2002). Other related works are Kuk (2003) and Jacquier et al. (2007).

DC exploits the well known result of Walker (1969) proving that, under suitable regularity conditions, the mean of the posterior distribution of a parameter θ tends to the maximum likelihood estimator as the sample size increases. Such convergence is proved to be independent of the prior distribution specification. DC consists in adopting a Bayesian model to a set of data cloned, that is replicated, several times till the posterior distribution becomes nearly degenerate with its mean converging to the ML estimate.

Let $\mathbf{y} = (y_1, \dots y_n)'$ be the observed vector of the binary response variable on a sample of size n. Let $\mathscr{L}(\boldsymbol{\theta}; \mathbf{y})$ be the likelihood function for a given set of data, $\pi(\boldsymbol{\theta})$ an arbitrary prior distribution for model parameters, and $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ the associate posterior distribution. To obtain maximum likelihood estimates, DC uses the (joint) pseudo-posterior distribution

$$\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)}) = \frac{\mathscr{L}_{(h)}(\boldsymbol{\theta}; \mathbf{y}_{(h)})\pi(\boldsymbol{\theta})}{C(\mathbf{y}_{(h)}, h)}, \tag{3}$$

where $\mathscr{L}_{(h)}(\boldsymbol{\theta}; \mathbf{y}_{(h)})$ represents the likelihood function on the data replicated h times, with $\mathbf{y}_{(h)} = (y_1, \dots y_n, \dots y_{n+1}, \dots\dots y_{hn})'$ and $C(\mathbf{y}, h) = \int \mathscr{L}_{(h)}(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ is the normalizing constant of the pseudo-posterior distribution.

Under regularity conditions (see Lele et al., 2010, Appendix A.1), the pseudo-posterior distribution degenerates towards ML estimates when h tends to infinity, invariantly to the choice of $\pi(\boldsymbol{\theta})$, as shown in Lele et al. (2010) Theorem A.2. In fact, in such a case, the pseudo-posterior distribution with h increasing tend to distribute normally, with mean equal to the ML estimates $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ of the model parameters and variance $1/h$ times the inverse of the Fisher information matrix:

$$\mathbb{E}\left[\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)})\right] \xrightarrow{h\to\infty} \widehat{\boldsymbol{\theta}}_{\text{MLE}}$$

$$\mathbb{V}\left[\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)})\right] \xrightarrow{h\to\infty} \frac{1}{h}\mathbb{V}\left[\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right]$$

The DC estimator is therefore defined as

$$\widehat{\boldsymbol{\theta}}_{\text{DC}} = \mathbb{E}\left[\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)})\right] \tag{4}$$

with h sufficiently large, with standard error

$$s.e.(\widehat{\boldsymbol{\theta}}_{\text{DC}}) = \sqrt{\frac{1}{h}\mathbb{V}\left[\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right]}. \tag{5}$$

Notice that the DC estimator is exactly the ML estimator when h reaches infinity. In practice, it is an approximation, which is good enough for an adequate h. Figure 1 shows the behaviour of the pseudo-posterior distribution as h increases for a parameter of a simple logit model with $n = 1\,000$. Notice that the distribution tends to collapse over the ML estimate, $\hat{\beta}$, and not at the *true* value of the parameter, that in this case was settled at 1. Factually, the use of cloned data does not improve the finite-sample properties of ML estimators, only helping to relieve their computation. To determine an adequate level for the number of clones h, one has to check whether the pseudo-posterior distribution has nearly degenerated. Lele et al. (2010) suggest to check if the largest eigenvalue of the posterior variance matrix is close to zero. Whenever a parameter is not identifiable, it has been showed (Lele et al., 2010, Theorem A.2) that this large eigenvalue does not converge to zero as h increases. Whenever the likelihood is flat for some parameters, for instance because these parameters are not identifiable or there is not enough information

in the data to identify them, then the largest eigenvalue of the posterior variance matrix does not decrease to zero when h increases. Consequently, DC procedure can be also utilized for checking parameters identifiability.

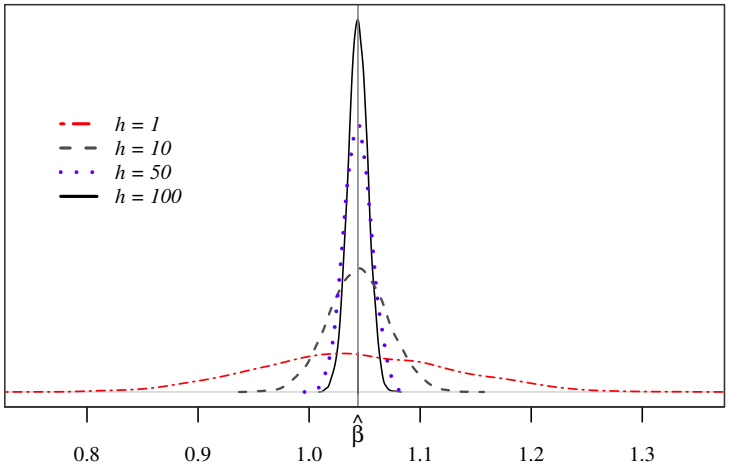

Figure 1: Example of pseudo-posterior distribution behaviour at increasing values of h

The DC estimators are invariant to the assumptions on the prior distribution, which can be chosen for computation convenience. A further proof of DC properties has been also given by Baghishani and Mohammadzadeh (2011).

Even if DC has been developed for hierarchical models, it can be adapted for the general class of non-hierarchical MML models. Indeed, the required assumptions (Lele et al., 2010, Appendix A.1) are fulfilled also by MML models, as they parallel logit mixed models a part for the assumption of a hierarchical structure of random effects, which is actually not used in Lele et al. (2010)'s Theorem A.2 proof. As a matter of fact, the absence of a hierarchy affects the dimensionality of the integral in (2), but not the form and the properties of the likelihood function. As shown in the proof of theorem 1 in the Appendix, a MML model can effectively be viewed as a General Generalized Linear Mixed Model (see, e.g. Jiang et al., 2013, for definition) with a particular multivariate latent variable. In the case of MML models, the pseudo-posterior distribution can be

specified as

$$\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)}, \mathbf{x}_{(h)}) = \frac{\left(\int \prod_{i=1}^{n_{(h)}} \frac{\exp\{y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{u})\}}{1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{u}\}} \, \phi_{J_{(h)}}(\boldsymbol{u}; \tau^2) \, d\boldsymbol{u} \right) \pi(\boldsymbol{\theta})}{C(\mathbf{y}_{(h)}, h)} \tag{6}$$

where $n_{(h)} = h \cdot n$, $J_{(h)} = h \cdot J$. The expression between parentheses is the likelihood function for h clones of the original data. For h sufficiently large, $\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)}, \mathbf{x}_{(h)})$ converges to a multivariate Normal distribution, with mean equal to the ML estimator $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ for the original data. Moreover, as the parameter space is continuous, the variance covariance matrix of these variates is h times the inverse of the observed Fisher Information matrix and can therefore be utilized to calculate asymptotic standard errors, to be used to obtain asymptotic confidence intervals.

The MCMC algorithms, such as Metropolis Haskings or Gibb sampling, allow to generate random variates from pseudo-posterior distribution (6), without computing the integrals either in the likelihood and in the denominator. Alternatively, the pseudo-posterior distribution can be approximated via integrated nested Laplace approximation. This alternative approximation is called *hybrid data cloning* and has been proposed by Baghishani et al. (2012).

The importance of ML estimator in MML models as defined in (1) and (2) and its approximation due to DC is assessed by the following two theorems, which derive by Jiang et al. (2013).

Theorem 1 *Assume* $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2) \in \Theta$*, with* Θ *being a convex subspace of* $\mathbb{R}^{p+2}$ *and* $\tau^2 > 0$*, of a MML model as defined in* (1) *and* (2)*. Let* M *be the largest subset of* N*,* $m = |M|$*, such that* $G_a \cap G_{a'} = \emptyset$ *and* $Y_a \perp\!\!\!\perp Y_{a'}$*, for all* $a, a' \in M$*. If* $m \to \infty$ *as* $n \to \infty$*, then, the ML estimator of* $\boldsymbol{\theta}$ *is consistent.*

Theorem 2 *Let* $\boldsymbol{\theta}$ *be the set of parameters of a MML model as defined in* (1)*. Assume* $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2) \in \Theta$*, with* Θ *being a convex subspace of* $\mathbb{R}^{p+2}$ *and* $\tau^2 > 0$*, of a MML model as defined in* (1) *and* (2)*. Let* $\widehat{\boldsymbol{\theta}}_{\text{DC}} = \mathbb{E}\left[\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)})\right]$ *be the DC estimator of* $\boldsymbol{\theta}$*, with*

$\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}_{(h)})$ *given in* (6). *If assumptions of Teorem 1 are fulfilled as* $n \to \infty$, *then* $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$ *is a consistent estimator of* $\boldsymbol{\theta}$ *with* $h \to \infty$.

Remark. Because $m \leq J$, the assumption of Theorem 1 that $m \to \infty$ as $n \to \infty$ can be satisfied only when $J \to \infty$ as $n \to \infty$ and the $n \times J$ matrix $\boldsymbol{W}$ having the raws the weights vector $\boldsymbol{w}_i$, $i = 1, \ldots n$, is sparse.

3 Indirect inference

Indirect inference was introduced by Smith (1993), Gouriéroux et al. (1993), and Gallant and Tauchen (1996). It is a simulation–based estimation procedure for a model, say $\mathcal{M}(\boldsymbol{\theta})$, with complex or intractable likelihood. It utilizes an *auxiliary* model $\widetilde{\mathcal{M}}(\boldsymbol{\eta})$ for estimating the parameters $\boldsymbol{\theta}$ of the model of interest $\mathcal{M}(\boldsymbol{\theta})$, granted that one is able to draw random samples from it, given a set of proposal values, $\boldsymbol{\theta}^*$. This is the case of the MML model as defined in (1). The auxiliary model parameters are estimated both on the observed data and on simulated samples. The estimates for $\mathcal{M}(\boldsymbol{\theta})$ are derived with a calibration procedure that compares these two sets of auxiliary model's estimates. Alternatively, a similar calibration procedure could be applied to the score of the likelihood (pseudo-likelihood) of the auxiliary model with simulated data. In the following we present the algorithm for finding the indirect estimates and standard errors for the MML model in (1).

The auxiliary model we propose for an MML model is perhaps the simplest possible: a linear model

$$x_i = g(\boldsymbol{x}_i, \boldsymbol{\eta}, e_i) = \gamma_0 + \sum_{l=1}^{p} \gamma_l x_{li} + \epsilon_i \tag{7}$$

where $\boldsymbol{\eta} = (\boldsymbol{\gamma}, \sigma^2) \in \Omega \subseteq \mathbb{R}^r$ is the parameter vector, $\epsilon_i = \sigma e_i$ are random terms independently distributed as Normal with zero mean and unknown variance σ^2. Notice that this proposal for $\widetilde{\mathcal{M}}(\boldsymbol{\eta})$ has $r = p + 2$ parameters, exactly the same of the MML model in (1). It is a case of exact identification, with a one–to–one correspondence between the parameters $(\boldsymbol{\beta}, \tau^2)$ versus $(\boldsymbol{\gamma}, \sigma^2)$. This choice is particularly convenient, as

will be seen in (11). The linear model provides a very rough approximation of the model of interest. Nevertheless, previous studies on hierarchical multilevel models revealed a good performance of this choice when used as the auxiliary model in indirect estimation (Calzolari et al., 2001).

The simple estimation of the auxiliary model parameters, based on the observed variables and expressed in the matricial form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, leads to biased (inconsistent) estimates $\widehat{\boldsymbol{\eta}} = (\hat{\gamma}, \hat{\sigma}^2)$, called *naive estimates*. Here, coefficients estimates are obtained straightforwardly as

$$\widehat{\boldsymbol{\gamma}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{8}$$

with $\boldsymbol{X}$ is the $n \times (p+1)$ matrix formed by the columns vectors $(\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ and $\boldsymbol{Y}$ is the n vector of the binary response variable. The computation of the variance parameter, $\hat{\sigma}^2$, requires some additional effort. Call n_j the number of statistical units i such that $w_{ji} \neq 0$. Therefore, $n_j = \sum_{i=1}^{n} \mathbb{1}_{j \in G_i}$. Then, an average residual among units in group j is here defined as

$$\overline{\epsilon}_j = \frac{1}{n_j} \sum_{i=1}^{n} \widehat{\epsilon}_i \, \mathbb{W}_{j \in G_i} \tag{9}$$

with $\mathbb{W}_{j \in G_i} = 1/w_{ij}$ when $w_{ij} \neq 0$ (that is $j \in G_i$) and 0 otherwise. Then, the naive estimator for the variance of the latent component is defined as

$$\widehat{\sigma}^2 = \frac{1}{J} \sum_{j=1}^{J} \overline{\epsilon}_j^2. \tag{10}$$

We denote $\widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$, the naive estimates that can be obtained using a simulated sample of size n^* of the response variable, $\widetilde{Y}_i(\boldsymbol{\theta}^*)$, $i = 1, 2, \ldots, n^*$, conditional on the observed explanatory variables $\boldsymbol{x}_i$ and the simulated values $\widetilde{\boldsymbol{U}}_j(\boldsymbol{\theta}^*)$, $j = 1, 2, \ldots, J$, at a given $\boldsymbol{\theta}^*$. The vector of J random effects is drawn only once and afterwards held fixed throughout the indirect estimation procedure. Consequently, a vector of random values $\widetilde{\boldsymbol{U}}_j$ is drawn once from a standard Normal distribution and $\widetilde{\boldsymbol{U}}_j(\boldsymbol{\theta}^*) = \tau^* \widetilde{\boldsymbol{U}}_j$.

If $\widehat{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$ are not too far in some sense, we can assume that the values $\boldsymbol{\theta}^*$ are good estimates of the parameter of interest. To detect the indirect estimator close enough to the *true* one, Gouriéroux et al. (1993) proposed

$$\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}^*} \left[\widehat{\boldsymbol{\eta}} - \widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)\right]' \Omega^{-1} \left[\widehat{\boldsymbol{\eta}} - \widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)\right] \tag{11}$$

where Ω is a positive definite weighting matrix. In case of exact identification, as presented here, estimates are unaffected by the choice of the matrix of weights Ω, as the minimization of the quadratic form (11) is obtained when $\widehat{\boldsymbol{\eta}} = \widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$. The tentative values for the *true* model parameters $\boldsymbol{\theta}^*$ are chosen iteratively until this equality is fulfilled. The iterative procedure for choosing tentative values of $\boldsymbol{\theta}^*$, called calibration, is here based on solving the implicit system of r equations $\widehat{\boldsymbol{\eta}} = \widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$ in the r unknowns $\boldsymbol{\theta}^*$, being in a case of exact identification. The solution of this implicit system of equations yielding to the indirect estimator cannot be written in closed form, as an analytic solution does not exist. Consequently, the problem has to be solved numerically, for instance using Newton Raphson. As in Calzolari et al. (1999) (page 18), we adopt the following updating equation

$$\boldsymbol{\theta}^*_{(k)} = \boldsymbol{\theta}^*_{(k-1)} + \delta \boldsymbol{A}^{-1}_{(k-1)} \left(\widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*_{(k-1)}) - \widehat{\boldsymbol{\eta}}\right) \tag{12}$$

where $\boldsymbol{\theta}^*_{(k)}$ is the value of the calibrated parameters after k iterations, $\boldsymbol{A}_{(k-1)}$ is a matrix that determines the direction of the k^{th} step, and δ is a real number between 0 and 1, determining the step size in the given direction. In particular, we take $\boldsymbol{A}$ equal to the Jacobian matrix of derivatives of the auxiliary parameters with respect to the parameters of interest. Derivatives are computed numerically, by finite differences method.

For some other types of models, when the $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ parameters are essentially the same, even if plugged into slightly different models, one can perform iterations using the identity matrix. This solution, called *Jacobi solution method* or *fine tuning* has been adopted in An and Liu (2000) and Mealli and Rampichini (1999). In the alternative, similar, approach proposed by Gallant and Tauchen (1996), calibration is aimed at minimizing a quadratic form based on the score of the likelihood, or pseudo-likelihood, of the auxiliary model. In this alternative approach, the score for the coefficients, for instance, is $\boldsymbol{X}'\widehat{\boldsymbol{\epsilon}}/n$, with $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}$.

Regarding the size of the simulated sample, one may adopt n^* equal to the observed

sample size n. Larger n^*, such as for instance $n^* = H \cdot n$, with a certain integer $H > 1$, can be adopted to improve estimates precision. Alternatively, as we shall do in our application and Monte Carlo experiments, H independent samples of length n can be produced, and $\widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$ computed as average of the H estimates. In most available applications of indirect inference, the value of H has usually been chosen between 10 and 100 (e.g. Gouriéroux et al., 1993; Sentana et al., 2008). The main reason for choosing H greater than 1 concerns the variance of the indirect estimator, whose formula includes a scalar factor equal to $(1 + 1/H)$: the larger is H, the smaller is the estimator variance. For the particular case of the MML models, we would suggest a much larger value of H, between 500 and 1000. This choice is not due to a further reduction of the estimator variance, but rather to obtain a smoother function $\widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$, numerically closer to a continuous function. The numerical derivatives can be then computed reliably with finite differences. In fact, the MML models are not only nonlinear, but also with discontinuous outcome. A discussion on the computational benefits due to a smoothing technique can be found in Calzolari and Di Iorio (2006).

The asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\theta}}$ follows straghtforwardly as for the generalized method of moments (Hansen, 1982)

$$V(\widehat{\boldsymbol{\theta}}) = \left[\frac{\partial \boldsymbol{\eta}'}{\partial \boldsymbol{\theta}} V(\widehat{\boldsymbol{\eta}})^{-1} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}'} \right]^{-1} . \tag{13}$$

Dealing with a just-identified case, the Jacobian $\partial \boldsymbol{\eta}' / \partial \boldsymbol{\theta}$ is a square matrix. The equation (13) can thus be also derived directly using the δ-method (e.g. Rao, 1973, p. 388). It is natural to adopt, as an estimate of the Jacobian matrix, the same matrix used in the calibration procedure (12), upon convergence. The variance-covariance matrix of the auxiliary parameters $V(\widehat{\boldsymbol{\eta}})$ can be then estimated as the sample variance-covariance matrix of the H vectors of length n, independently simulated, whose average has been computed to obtain $\widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$. This way of computing the estimate of $V(\widehat{\boldsymbol{\eta}})$ has the advantage of simplicity, with accuracy guaranteed by the large chosen value of H. In principle, a closed form evaluation should also be possible for the γ parameters, being the estimation of the auxiliary parameters based on OLS. However, no closed form can be easily found

for the parameter σ^2, whose estimates are computed as in equations (9) and (10).

4 A Monte Carlo simulation study

This section numerically investigates and compares the finite sample size performance of the two proposed estimating procedures. Moreover, for a better understanding, DC and indirect inference estimates are compared with estimates based on the mean of the posterior MCMC distributions of a Bayesian model, hence called *Bayesian*, and the hybrid data cloning estimates, *INLA Data Cloning*.

The comparative experiment was designed as follows: 500 samples of several dimensions has been generated by a MML model

$$\text{logit}\, P(Y_i = 1 \mid x_i, \boldsymbol{u}_i) = \beta_0 + \beta_1 x_i + \sum_{j \in G_i} w_{ij} u_j$$

with $\beta_0 = -2$, $\beta_1 = 1$, $|G_i| = 2$ for each $i = 1, \ldots n$, and $u_j \sim N(0, 1)$, that is $\tau^2 = 1$. The probability of a couple of groups to include a same unit was settled at 0.5. Each w_{i1} was randomly generated by a Uniform distribution, $U \sim U(0.3, 0.7)$, while $w_{i2} = 1 - w_{i1}$. For the DC, hybrid DC and Bayesian estimator, we set prior distributions $\beta_l \sim N(0, 10)$, $l = 0, 1$, moderately vague. The prior distribution for the variance τ^2 was quite informative, $U(0, 0.1)$. We opted for this prior selection to put in evidence the DC invariance property with respect to prior distribution choice.

We implemented the simulation study using the `dclone` package Sólymos (2010) in R (R Development Core Team, 2012) and JAGS Plummer (Plummer) for the DC estimates. Indirect inference estimates were computed by an *ad hoc* program in FORTRAN77. Bayesian estimates have been computed using JAGS Plummer (Plummer), while the hybrid data cloning estimates have been obtained with the `INLA` package in R (see for e.g., Martins et al., 2013, or the web site `http://www.r-inla.org`).

Table 1 reports the Monte Carlo mean of the estimates, their standard deviations and the Monte Carlo Mean Squared Errors (MSEs) over 500 Monte Carlo samples. Results

Table 1: Means, standard deviations (sd) and Mean Squared Errors (MSE) of parameter estimates on 500 Monte Carlo samples.

	Indirect inference			Data cloning			INLA Data Cloning			Bayesian Mean		
	Mean	sd	MSE	Mean	sd	MSE	Mean	sd	MSE	Mean	sd	MSE
					$J = 50$		$n = 651$					
β_0	-1.984	0.255	0.065	-1.999	0.249	0.062	-1.994	0.248	0.061	-1.951	0.244	0.062
β_1	0.999	0.085	0.007	1.006	0.079	0.006	1.003	0.079	0.006	0.985	0.079	0.007
τ^2	0.862	0.530	0.300	0.957	0.445	0.200	0.906	0.429	0.193	0.712	0.522	0.355
					$J = 100$		$n = 2\,481$					
β_0	-2.014	0.149	0.022	-2.006	0.143	0.020	-2.003	0.145	0.021	-2.005	0.142	0.020
β_1	1.001	0.045	0.002	1.001	0.044	0.002	1.001	0.044	0.002	1.002	0.044	0.002
τ^2	0.964	0.243	0.060	0.993	0.241	0.058	0.971	0.232	0.055	1.028	0.251	0.064
					$J = 200$		$n = 10\,038$					
β_0	-1.998	0.090	0.008	-2.000	0.089	0.008	-2.000	0.089	0.008	-1.999	0.089	0.008
β_1	1.000	0.021	0.000	1.002	0.020	0.000	1.001	0.020	0.000	1.001	0.020	0.000
τ^2	0.978	0.147	0.022	0.998	0.133	0.018	0.991	0.132	0.017	1.002	0.134	0.018

are uniformly good, with the ML estimator via DC, as expected, producing excellent results. Some interesting observations can be made. For small samples ($J = 50$), the best performance is by far due to the DC and Hybrid DC estimators. Regarding the estimates of τ^2, indirect inference estimator performs better than the Bayesian mean. However, this could be due to the challenging prior selection. DC and hybrid DC estimators provide roughly the same result. This result suggests that replacing INLA to MCMC algorithm speed up the estimating procedure without introducing any relevant approximation error.

Consistency of the DC estimator is confirmed by the simulations. For $J = 200$, the average estimates are very close to the true values and the mean squared error decreases as the sample size increases. Indirect inference performance is in line with the other procedures for the $\boldsymbol{\beta}$ parameters while its MSE is only lightly higher than the others.

Figure 2 illustrates the Monte Carlo sampling distributions for the two estimators of interest. For each parameter, the two distributions are quite close. The best performance concerns the parameter β_1, which is often the main parameter of interest. Concerning τ^2, the indirect inference estimator has usually a mode in a lower point that the due one, suggesting a frequent underestimate, even if of negligible amount. As sample size increases, it seems that the DC estimator reaches normality faster than the indirect inference estimator.

Table 2 summarizes the Monte Carlo experiments for the computation of estimators standard errors, both for DC and indirect inference. For the DC estimators, standard errors are computed as in (5), while equation (13) has ben used for the indirect inference estimators. On average, DC standard errors are generally lower than indirect inference ones, with the gap reducing as J and n increase. For $J = 50$, however, indirect inference standard errors show high variability, particularly for τ^2 standard error. This is due to sporadic cases, such as for example outliers, in which the algorithm fails to find a reasonable solution.

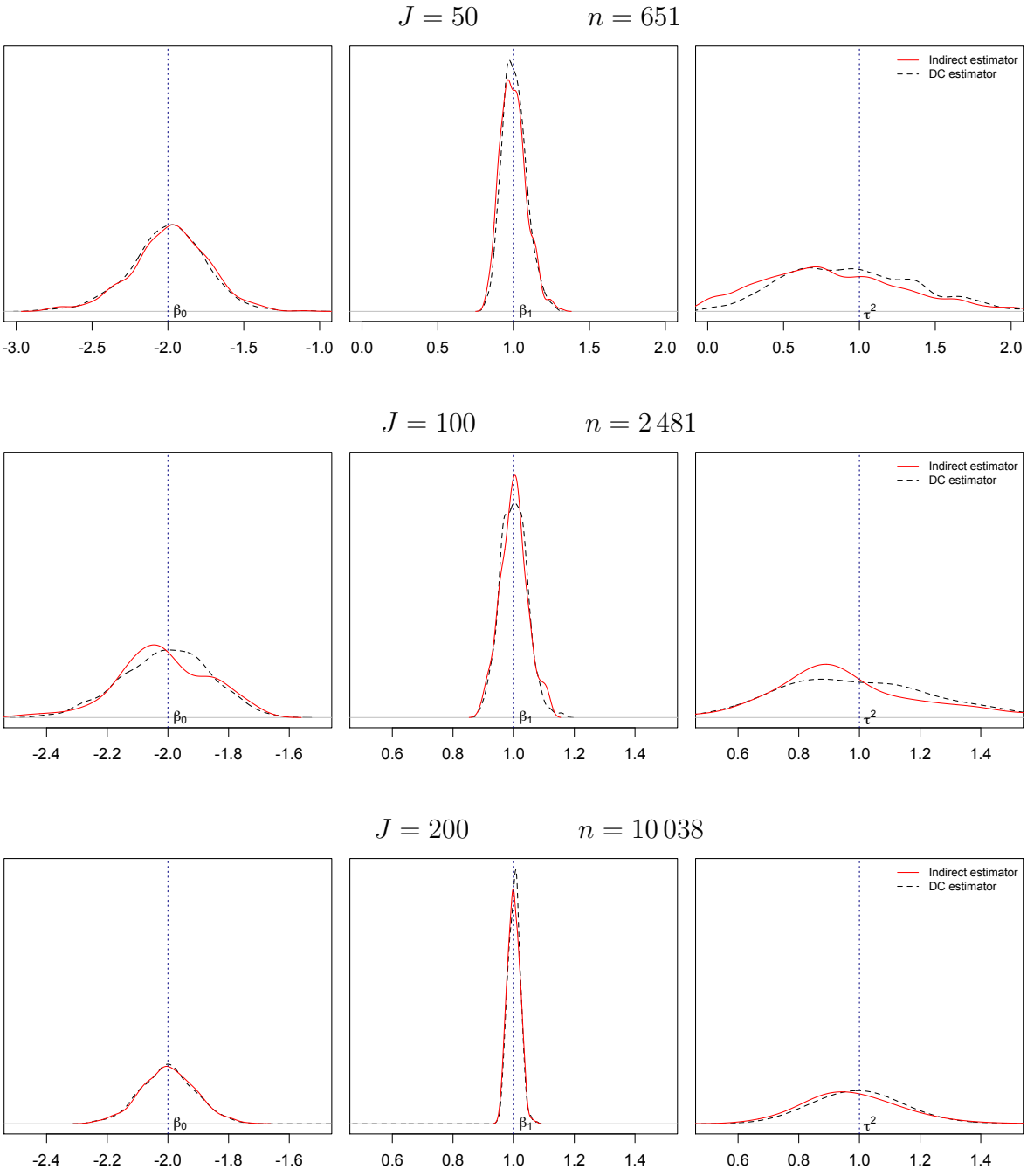

Figure 2: Monte Carlo sample distribution of parameter estimates.

Table 2: Means, standard deviations on standard error estimates (*s.e.*), and summaries on asymptotic confidence intervals (CI) on 500 Monte Carlo samples.

	Data cloning			Indirect inference		
	β_0	β_1	τ^2	β_0	β_1	τ^2
			$J = 50$	$n = 651$		
Mean *s.e.*	0.2442	0.0812	0.4413	0.3062	0.1208	0.8322
s.e. standard deviation	0.0261	0.0061	0.1204	0.8568	0.5128	4.0525
Average CI length	0.9572	0.3184	1.7297	1.2003	0.4735	3.2621
CI Monte Carlo coverage	0.9420	0.9600	0.9020	0.9420	0.9580	0.9040
			$J = 100$	$n = 2\,481$		
Mean *s.e.*	0.1441	0.0412	0.2317	0.1448	0.0436	0.2730
s.e. standard deviation	0.0101	0.0017	0.0395	0.0125	0.0025	0.0535
Average CI length	0.5647	0.1616	0.9081	0.5675	0.1709	1.0703
CI Monte Carlo coverage	0.9560	0.9360	0.9320	0.9500	0.9220	0.9540
			$J = 200$	$n = 10\,038$		
Mean *s.e.*	0.0870	0.0202	0.1309	0.0877	0.0213	0.1494
s.e. standard deviation	0.0047	0.0005	0.0141	0.0054	0.0008	0.0186
Average CI length	0.3409	0.0791	0.5130	0.3437	0.0837	0.5856
CI Monte Carlo coverage	0.9400	0.9680	0.9420	0.9400	0.9680	0.9360

5 A real example: a random effects Bradley & Terry model for next top model comparison

An interesting illustration of MML models consists of Bradley & Terry models with random effects. In this section, as an illustrative example, we apply DC and indirect inference estimating procedure to a Bradley & Terry random effects model on a well know data set, on Germany's next top models, year 2007. Data are available in package `psychotree` in R (R Core Team, 2014), and have been analyzed by Strobl et al. (2011) using a Bradley-Terry Model via recursive partitioning.

In this study of the Department of Psychology of University of Tübingen, a sample of 192 raters/judges, aged between 15 and 77 years, has been asked to judge the attractiveness of the 6 finalists of the second edition of Germany's next top models casting show (Barbara Meier, Anni Wendler, Hana Nitsche, Fiona Erdmann, Mandy, Graff, and Anja Platzer, in decreasing order according to the final ranking in the show). The choice was based on paired comparison: contestants photos were showed to each judge two-by-two, no ties admitted. As judge-specific explanatory variables, gender, age, and three questions about their interest/knowledge of the TV show were recorded.

The Bradley-Terry (BT) model (Bradley and Terry, 1952) is a model aiming at scoring a set of items on the basis of paired comparisons. Even if the field of application of BT models is quite wide, it is customary to adopt sport terminology: each element of the comparison is called *player*, each comparison *contest*, the score of each element is called *ability*. Considering each contest as the statistical unit, the BT model can be viewed as a generalized linear model in which the response variable is binary, assuming value 1 if the first element of the pair wins the comparison and 0 otherwise. We are going to briefly sketch these models to show as they are a particular case of MML models, remanding to (Firth, 2005; Turner and Firth, 2012; Cattelan et al., 2012) for a detailed description and updated review.

Denoting with a_i and a_j the positive–valued parameters representing the abilities of

players i and j respectively, the BT model assumes that the odds of i beating j are a_i/a_j. Then, the probability p_{ij} that player i beats player j, can be written as

$$\text{logit}(p_{ij}) = \lambda_i - \lambda_j$$

where $\lambda_l = \log a_l$, $l = i, j = 1, \ldots, I$, with I equals to 6 for Germany's next top models data. Assuming independence among contests, the abilities and log–abilities can be estimated by maximum likelihood by standard softwares, after imposing the opportune identifiability constraints, for instance $\lambda_1 = 0$. Extensions of the ordinary BT model have been considered to include specific characteristics for the players, the contests or the judges, or to admit the case of draw. In Germany's next top models data, the ability of a player can be supposed to depend on a set of contest-specific explanatory variables, for example the characteristic of the judge of the contest. Consequently, the corresponding BT model would assume that $\text{logit}(p_{ij}^k) = \lambda_i^k - \lambda_j^k$ with $\lambda_l^k = \sum_{r=1}^R \beta_{lr} x_{kr}$, where $k = 1, \ldots K$ is the judge indicator, $K = 192$ and $R = 4$ is the number of explanatory variables. Then,

$$\text{logit}(p_{ij}^k) = \sum_{r=1}^{R} (\beta_{ir} - \beta_{jr}) x_{kr}.$$

Notice that the total number of parameters, after the identifiability constraints, is $R(I - 1)$. Moreover, the abilities are decomposed into the sum of judge-specific abilities, and deterministically depend on the set of covariates included.

In Germany's next top models data as comparisons are carried out by several judges, each match (comparison between two specific models) is repeated more than once. Judgements made by the same judge are likely to be dependent. This aspect could suggest a lack of independence between contests, that can be addressed by including a specific random component, as

$$\lambda_l^k = \sum_{r=1}^{R} \beta_{lr} x_{kr} + u_{lk}$$

with $u_{lk} \sim N(0, \tau^2)$, that assumes a specific, unobserved value whenever the judge k scores player l. The resulting BT random effect model is a particular case of MML

model

$$\text{logit}(p_{ij}^k) = \sum_{r=1}^{R} (\beta_{ir} - \beta_{jr}) x_{kr} + u_{ik} - u_{jk}$$

in which the random effect weights sum to 0, being always 1 for the first component and -1 for the second. As in Germany's next top models data we are interested more in ranking that in explaining the effect of judge-specific covariates, we here adopt the following model

$$\text{logit}(p_{ij}^k) = \lambda_i^k - \lambda_j^k = \lambda_i - \lambda_j + u_{ik} - u_{jk}$$

that parsimoniously provides a top model scoring and takes into account for dependence between contests.

Table 3: Estimates of log-ability parameters for the top models (Barbara is the reference model) and standard errors via Data Cloning and indirect Inference

	Log-ability						
	Barbara	Anni	Hana	Fiona	Mandy	Anja	τ^2
Data cloning	0	-0.0208	1.1318	0.6116	-0.8610	-0.6247	6.0726
(s.e.)	-	(0.3110)	(0.3049)	(0.3047)	(0.3150)	(0.3096)	(0.7665)
Indirect inference	0	-0.0293	1.0927	0.6381	-0.8939	-0.6292	5.924
(s.e.)	-	(0.2963)	(0.2910)	(0.2924)	(0.3027)	(0.2915)	(0.8274)

Table 3 shows the estimates for model log-abilities, whose confidence intervals are depicted in Figure 3 and the variance of the unobserved component, via DC and indirect inference. In both cases Hana turned to be first, significantly better of both Barbara, the actual winner, and the second classified, Anni. These results confirm those obtained, using a different model and with different aim, by Strobl et al. (2011).

Estimates resulted from the two procedures are quite similar. We set $h = 80$ cloning for DC, corresponding to a largest eigenvalue of the posterior variance matrix equal to

0.0076. For Indirect Inference estimates, we set $H = 1000$ and starting values equal to the ordinary BT model, without random effects. Despite such similar results, the computational time necessary to reach the estimates was incredibly different: 15.62 seconds for Indirect Inference versus 13 210.55 seconds (over three hours and half) for Data Cloning (running on an Intel® i3 2120 pc, with 4 GB of Ram).

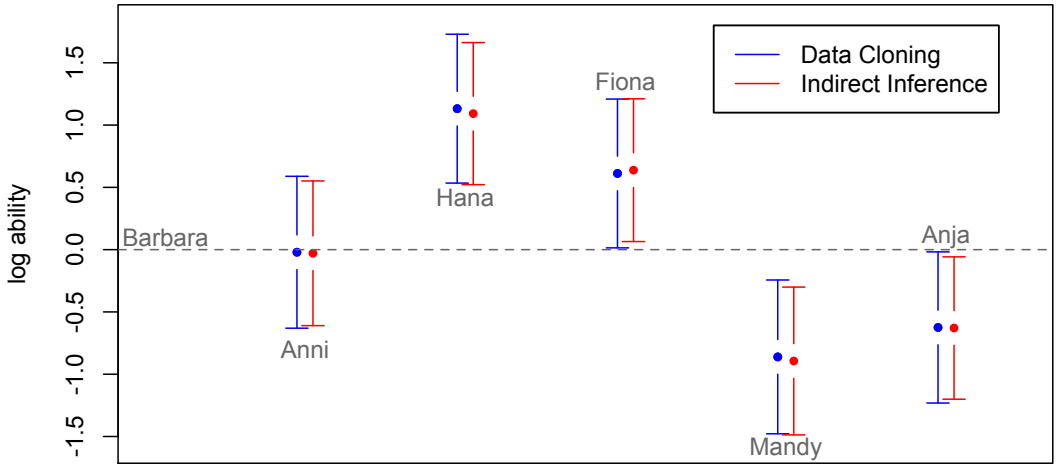

Figure 3: Asymptotic confidence intervals for models log-abilities

6 Final remarks

MML models are an interesting class of models for binary response variables, in which statistical units are supposed to belong to more than one group, in a non-hierarchical way. Computational difficulties make non Bayesian inference for these models particularly cumbersome. In particular, maximum likelihood inference entails the necessity of solving a high dimensional integral.

This paper provides two alternatives for estimating MML model parameters. The first procedure, Data Cloning, results in approximate maximum likelihood estimates, which

are shown to be consistent and asymptotically Normal. The second procedure, Indirect Inference, is based on an auxiliary model. Our strategy for this choice has been driven by accuracy, simplicity and computational speed of the algorithm.

The numerical study confirms the expected performance of the maximum likelihood estimates and highlights the good behaviour of Indirect Inference estimator. The basic conclusion is that the indirect estimator is slightly biased, but by far faster than both the data cloning and MCMC estimators.

The two procedures have been successfully applied to estimate a random effect Bradley and Terry model to analyse 2007 Germany's next top models data. Notice that both the estimation algorithms proposed can be easily extended to non-Gaussian and cross-classified random effects.

A Appendix

Proof of theorem 1 The theorem can be proved by verifying that Theorem 5 in Jiang et al. (2013) holds. MML models belong to the class of Generalized Linear Mixed Models as defined by Jiang et al. (2013): conditionally to a vector $\boldsymbol{u}$ of random effects, $Y_1, \dots Y_n$ are conditionally independent and have distribution belonging to the exponential family. The natural parameter is associated the conditional mean, with $\mathbb{E}[Y_i \mid \boldsymbol{u}] = g(\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u})$. In MML models, g is the inverse logit link and $\boldsymbol{z}_i$ is $\boldsymbol{w}_i'$ as defined in (2). Finally, $\boldsymbol{u} \sim N(\boldsymbol{0}, \tau^2\mathbb{I})$. Consequently, Theorem 5 Jiang et al. (2013) applies as far as the subset argument can be used satisfying theorem assumptions.

Consider the subset $\boldsymbol{y}_M = (y_a)$, $a \in M$, with probability distribution, under $\boldsymbol{\theta}$,

$$p_{\boldsymbol{\theta}}(y_a) = \mathbb{E}\left[\frac{\exp\left\{y_a(\mu+\xi)\right\}}{1+\exp(\mu+\xi)}\right]$$

where $\mu = \sum_{l=1}^{p} \beta_l x_{la}$ and $\xi_a = \sum_{j \in G_a} w_{aj} u_j$, $(\sum_{j \in G_a} w_{aj}^2)^{-1/2}\xi_a \sim N(0, \tau^2)$. Assuming

$m \to \infty$ as $n \to \infty$, the subset $\boldsymbol{y}_M$ satisfies Jiang et al. (2013)'s assumption (A1) and the so called Jiang's subset argument can be applied. Moreover, assumption (B2) holds as $\tau^2 > 0$, while assumption (B3) is fulfilled as $m^{-1}\log(n) \to 0$ paralleling the proof given in Section 4 Jiang et al. (2013). The remaining assumptions, (C1) and (C2) are fulfilled according to a proof paralleling that given in Section 6 in the Supplementary Material of Jiang et al. (2013). □

Proof of theorem 2 The proof derives by the consistency of the ML estimator of $\boldsymbol{\theta}$ and by the dominated convergence theorem. See Section 8 of the Supplementary Material of Jiang et al. (2013) for detailed proof. □

References

An, M. and M. Liu (2000). Using indirect inference to solve the initial-conditions problem. *Review of Economics and Statistics 82*(4), 656–667.

Baghishani, H. and M. Mohammadzadeh (2011). A data cloning algorithm for computing maximum likelihood estimates in spatial generalized linear mixed models. *Computational Statistics & Data Analysis 55*(4), 1748–1759.

Baghishani, H., H. Rue, and M. Mohammadzadeh (2012). On a hybrid data cloning method and its application in generalized linear mixed models. *Statistics and Computing 22*(2), 597–613.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate bayesian computation in population genetics. *Genetics 162*(4), 2025–2035.

Billio, M. and A. Monfort (2003). Kernel-based indirect inference. *Journal of Financial Econometrics 1*(3), 297–326.

Bradley, R. and M. Terry (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika 39*(3-4), 324.

Browne, W. (2012). Mcmc estimation in mlwin, v2.26. Centre for Multilevel Modelling, University of Bristol.

Browne, W., H. Goldstein, and J. Rasbash (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling 1*(2), 103.

Calzolari, G. and F. Di Iorio (2006). Discontinuities in indirect estimation: an application to ear models. *Computational Statistics and Data Analysis 50*, 2124–2136.

Calzolari, G., F. Di Iorio, and G. Fiorentini (1998). Control variates for variance reduction in indirect inference: interest rate models in continuous time. *The Econometrics Journal 1*(1), 100–112.

Calzolari, G., F. Di Iorio, and G. Fiorentini (1999). Indirect estimation of just-identified models with control variates. University of Firenze, Quaderni del Dipartimento di Statistica "G. Parenti".

Calzolari, G., F. Mealli, and C. Rampichini (2001). Alternative simulation-based estimators of logit models with random effects. University of Firenze, Quaderni del Dipartimento di Statistica "G. Parenti".

Cattelan, M. et al. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science 27*(3), 412–433.

Chung, H. and S. Beretvas (2011). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology 65*(2), 185–200.

Czellar, V. and E. Ronchetti (2010). Accurate and robust tests for indirect inference. *Biometrika 97*(3), 621–630.

Doucet, A., S. Godsill, and C. Robert (2002). Marginal maximum a posteriori estimation using markov chain monte carlo. *Statistics and Computing 12*(1), 77–84.

Fielding, A. (2002). Teaching groups as foci for evaluating performance in cost-effectiveness of gce advanced level provision: Some practical methodological innovations1. *School effectiveness and school improvement 13*(2), 225–246.

Firth, D. (2005). Bradley-terry models in r. *Journal of Statistical software 12*(1), 1–12.

Gallant, A. and G. Tauchen (1996). Which moments to match? *Econometric Theory 12*(04), 657–681.

Genton, M. and E. Ronchetti (2003). Robust indirect inference. *Journal of the American Statistical Association 98*(461), 67–76.

Gouriéroux, C. and A. Monfort (1996). *Simulation-based econometric methods*. Oxford University Press, USA.

Gouriéroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 8*(S1), S85–S118.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Heggland, K. and A. Frigessi (2004). Estimating functions in indirect inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 447–462.

Hill, P. and H. Goldstein (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral statistics 23*(2), 117–128.

Jacquier, E., M. Johannes, and N. Polson (2007). Mcmc maximum likelihood for latent state models. *Journal of Econometrics 137*(2), 615–640.

Jiang, J. et al. (2013). The subset argument and consistency of mle in glmm: Answer to an open problem and beyond. *The Annals of Statistics 41*(1), 177–195.

Jiang, W. and B. Turnbull (2004). The indirect method: inference based on intermediate statisticsa synthesis and examples. *Statistical Science 19*(2), 239–263.

Karl, A., Y. Yang, and S. Lohr (2012). Efficient maximum likelihood estimation of multiple membership linear mixed models, with an application to educational value-added assessments. *Computational Statistics & Data Analysis*.

Kuk, A. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B (Methodological) 57*, 395–407.

Kuk, A. (2003). Automatic choice of driving values in monte carlo likelihood approximation via posterior simulations. *Statistics and Computing 13*(2), 101–109.

Lele, S., B. Dennis, and F. Lutscher (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. *Ecology Letters 10*(7), 551–563.

Lele, S., K. Nadeem, and B. Schmuland (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association 105*(492), 1617–1625.

Lombardi, M. and G. Calzolari (2009). Indirect estimation of α-stable stochastic volatility models. *Computational Statistics & Data Analysis 53*(6), 2298–2308.

Martins, T. G., D. Simpson, F. Lindgren, and H. Rue (2013). Bayesian computing with inla: new features. *Computational Statistics & Data Analysis 67*, 68–83.

Mealli, F. and C. Rampichini (1999). Estimating binary multilevel models through indirect inference. *Computational Statistics & Data Analysis 29*(3), 313–324.

Nadeem, K. and S. Lele (2012). Likelihood based population viability analysis in the presence of observation error. *Oikos 121*(10), 1656–1664.

Plummer, M. Jags version 3.3.0 manual. *URL: http://http://sourceforge.net/projects/mcmc-jags/files/Manuals/3.x/jags_user_manual.pdf, year=2009*.

Ponciano, J., J. Burleigh, E. Braun, and M. Taper (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Systematic Biology*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York, Wiley, 2nd edition.

Rasbash, J. and W. Browne (2001a). Modelling non-hierarchical structures. *Multilevel modelling of health statistics*, 93–105.

Rasbash, J. and W. Browne (2001b). Non-hierarchical multilevel models. *Multilevel Modelling of Health Statistics*.

Robert, C. (1993). Prior feedback: A bayesian approach to maximum likelihood estimation. *Comput. Statist 8*, 279–294.

Roberts, C. and R. Walwyn (2012). Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Statistics in Medicine*. In press, available in early view.

Rodriguez, G. and N. Goldman (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 158*(1), 73–89.

Sentana, E., G. Calzolari, and G. Fiorentini (2008). Indirect estimation of large conditionally heteroskedastic factor models, with an application to the dow 30 stocks. *Journal of Econometrics 146*(1), 10–25.

Smith, A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics 8*(S1), S63–S84.

Sólymos, P. (2010). dclone: Data cloning in R. *The R Journal 2*(2), 29–37.

Strobl, C., F. Wickelmaier, and A. Zeileis (2011). Accounting for individual differences in bradley-terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics 36*(2), 135–153.

Torabi, M. (2012). Likelihood inference in generalized linear mixed models with two components of dispersion using data cloning. *Computational Statistics & Data Analysis 56*, 4259–4265.

Torabi, M. and F. Shokoohi (2012). Likelihood inference in small area estimation by combining time-series and cross-sectional data. *Journal of Multivariate Analysis*.

Tranmer, M., S. D. and W. Browne (2013). Multiple membership models for social network and group dependencies. *Journal of the Royal Statistical Society. Series A (to appear)*.

Turner, H. and D. Firth (2012). Bradley-terry models in r: The bradleyterry2 package. *Journal of Statistical software 48*(9), 1–12.

Walker, A. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 80–88.