Estimation for the Structured Bradley-Terry Model

Yuxi Jiang William Thomas James Thornton

April 26, 2019

1 Introduction

In this report we explore the Penalised Quasi-Likelihood (PQL) approach to Maximum Likelihood estimation for the structured Bradley-Terry model. In particular, we will investigate the estimator bias introduced by the Laplace method of integral approximation in the case of sparse contest data.

The structured Bradley-Terry model can be viewed as a special case of a Generalised Linear Mixed Model (GLMM). GLMMs are well studied and there are numerous fitting algorithms in the literature, including fully Bayesian procedures such as Markov Chain Monte Carlo (MCMC), Sequential Monte Carlo (Fan et al., 2008) or Variational Bayes (Hall et al., 2011). One could also appeal to Expectation Maximisation (EM) or Variational EM. We limit our scope to two fitting procedures: MCMC, as detailed by Zeger and Karim (1991), and Sequential Reduction, introduced by Ogden (2015).

We will give a brief introduction on the structured Bradley-Terry model in Section 2, followed by an explanation of the Penalised Quasi-Likelihood approach and an empirical study on its performance under sparse data structure in Section 3. The two fitting procedures that we present in this paper are detailed in Section 4 (MCMC) and Section 5 (Sequential Reduction) along with simulation studies that compares the performance of each of the methods to the PQL approach. Finally, conclusions and future research directions are available in Section 6.

2 The Bradley-Terry Model

In a pairwise contest between multiple 'players', the Bradley-Terry model assumes that the odds that player i defeats player j is given by α_i/α_j, where α_i and α_j are positive-valued parameters representing the 'ability' of players i and j respectively. That is:

$$\frac{\mathbb{P}(i\ beats\ j)}{\mathbb{P}(j\ beats\ i)} = \frac{\alpha_i}{\alpha_j} \quad \text{and} \quad \mathbb{P}(i\ beats\ j) = \frac{\alpha_i}{\alpha_j + \alpha_i}.$$

Alternatively, we may specify the Bradley-Terry model in a logit-linear form where $log(\alpha_i) = \lambda_i$:

$$logit(\mathbb{P}(i\ beats\ j)) = \lambda_i - \lambda_j$$

$$\mathbb{P}(i\ beats\ j) = \frac{1}{1 + e^{-(\lambda_i - \lambda_j)}} \tag{1}$$

If, for each player, we have available a set of explanatory variables, one may structure the Bradley-Terry model further by linking the explanatory variables to the ability parameters. For example, given explanatory variables $x_{i1}, ..., x_{ip}$ for player i, we can model the λ_i as:

$$\lambda_i = \sum_{p=1}^{P} \beta_p x_{i,p} + b_i,$$

where b_i denotes a set of random effects accounting for the variability between players, which are modelled as independent $\mathcal{N}(0, D)$ random variables. We can then model the difference in the abilities of players i and j by:

$$\eta_{i,j} = \lambda_i - \lambda_j = \sum_{p=1}^{P} \beta_p x_{i,p} - \sum_{p=1}^{P} \beta_r x_{j,p} + b_i - b_j \tag{2}$$

Hence,

$$\mathbb{P}(i \text{ } beats \text{ } j) = \frac{1}{1+e^{-\eta_{i,j}}}. \tag{3}$$

This model can therefore be viewed as a generalised linear mixed model (GLMM), where the response variables y_i are taken to be conditionally independent given the random effects b_i. In particular, this is a logistic regression for the probability of winning, or equivalently, a Binomial GLMM for the win/lose counts. Let $W_{i,j}$ denote the number of observed wins for player i when paired against player j and let $n_{i,j} = W_{i,j} + W_{j,i}$ denote the number of matches/fights between i and j. The likelihood of the coefficients $\beta = \beta_{[1:P]}$ and the variance D, given observed counts $W = \{W_{i,j}\}_{i\in[1:n],j\in[1:n]}$, is denoted by L, where the random effects $\boldsymbol{b} = b_{1:n}$ have been marginalised:

$$\begin{aligned} L(\beta, D) &= \int \prod_{i=1}^{n} \prod_{j:j>i} f(W_{i,j}|\eta_{i,j})\varphi_D(b_i)\mathbf{db} \\ &= \int \prod_{i=1}^{n} \prod_{j:j>i} \binom{n_{i,j}}{W_{i,j}} \left(\frac{1}{1+e^{-\eta_{i,j}}}\right)^{W_{i,j}} \left(\frac{1}{1+e^{\eta_{i,j}}}\right)^{W_{j,i}} \frac{1}{|D|^{-\frac{1}{2}}} e^{-\frac{1}{2}b_i D^{-1} b_i}\mathbf{db} \end{aligned} \tag{4}$$

Here, φ denotes the Normal density of the random effects (up to constant of proportionality), with variance D, while f denotes the Binomial distribution with corresponding probability based on (3) and count based on $n_{i,j}$.

It is difficult to compute the maximum likelihood estimates (MLE) for β and D due to the high-dimensional integral with respect to the random effects $\boldsymbol{b} = b_{1:n}$. Note that the integral cannot be decomposed further given the dependence of $\eta_{i,j}$ on b_i and b_j, as shown by (2).

3 Penalised Quasi-Likelihood

A common way of navigating this difficult integral is to use Laplace's method of integral approximation in tandem with the Penalised Quasi-Likelihood (PQL). This allows us to obtain parameter estimates without having to compute a high dimensional integral.

Let $\mathbf{y}$ denote the vector of observed responses and assume that the y_i are conditionally independent given a vector of random effects $\mathbf{b}$. A typical GLMM is of the form:

$$\mathbb{E}[\mathbf{y}|\mathbf{b}] = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}),$$

where $\mathbf{X}$ is a covariate matrix, $\mathbf{Z}$ is a model matrix of random effects and $\boldsymbol{\beta}$ is a vector of fixed coefficients. We also write $\mathbb{E}(y_i|\mathbf{b}) = \mu_i^b$ and $Var(y_i|\mathbf{b}) = \phi a_i v(\mu_i^b)$, where $v(.)$ is a specified variance function, a_i is a known constant and ϕ is a dispersion parameter. Taking $\mathbf{D} = \mathbf{D}(\theta)$ to be a covariance matrix depending on a vector of unknown variance components θ, we assume that $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D})$.

The integrated quasi-likelihood function is given by:

$$e^{ql(\boldsymbol{\beta},\theta)} \propto |\mathbf{D}|^{-1/2} \int \exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i, \mu_i^b) - \frac{1}{2}\mathbf{b}^T\mathbf{D}^{-1}\mathbf{b}\right\} \mathrm{d}\mathbf{b} \tag{5}$$

$$= |\mathbf{D}|^{-1/2} \int e^{f'(\mathbf{b})} \mathrm{d}\mathbf{b}, \tag{6}$$

where $d_i(y, \mu) = -2\int_y^\mu \frac{y-u}{a_i v(u)} du$ is the deviance measure of fit, f' is a function of $\mathbf{b}$ corresponding to the exponent within (5) and the true log-quasi-likelihood is denoted by $ql(\boldsymbol{\beta}, \theta)$.

Applying Laplace's method for integral approximation, in addition to some simplifications detailed in the appendix, one can re-write equation (5) as:

$$ql(\boldsymbol{\beta}, \theta) \approx -\frac{1}{2}\log|\mathbf{I} + \mathbf{Z}^T\mathbf{WZD}| - \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i, \mu_i^{\tilde{b}}) - \frac{1}{2}\tilde{\mathbf{b}}^T\mathbf{D}^{-1}\tilde{\mathbf{b}}. \tag{7}$$

We choose $\tilde{\mathbf{b}}$ to maximise the sum of the last two terms and select $\boldsymbol{\beta}$ to maximise the second term. Assuming that the first term in (7) is stable, that is, that the GLM iterated weights $\mathbf{W}$ vary slowly as a function of the mean, it does not have a large influence on the maximum. If we ignore this first term, we can then use Fisher's scoring algorithm with the score equations

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i^b)x_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} = 0 \quad \text{and} \quad \sum_{i=1}^{n} \frac{(y_i - \mu_i^b)z_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} = \mathbf{D}^{-1}\mathbf{b},$$

to fit a GLMM and obtain estimates of the fixed effect parameters, as well as the random effects.

3.1 Shortcomings of PQL

PQL does have its limitations however. As detailed above, the underlying approximation behind PQL is the Laplace method which assumes that f' in equation (6): has a global maximum $\tilde{\mathbf{b}}$, has negative second derivative at $\tilde{\mathbf{b}}$ and is twice continuously differentiable around $\tilde{\mathbf{b}}$. The intuition behind the Laplace approximation is that based on f' having the above properties, the contribution to the integral of the exponential of f' will primarily be around $\tilde{\mathbf{b}}$, and decreasing exponentially away from $\tilde{\mathbf{b}}$. This approximation is sometimes referred to as a 'Gaussian approximation' (Murphy, 2012).

In the context of the contests described in section 2, the observed data (y_i in equation (5)) will be from a Binomially distribution. In the case when the competition is dense and the players have multiple bilateral matches, the corresponding Binomial distribution will be 'close' to the Normal distribution by a central limit theorem argument and Stein's method, see Chen et al. (2010). Therefore, in the dense case, the Laplace method will give a close approximation. However, where the competitions are sparse, f' may be multi-modal and contribution to the integral in (6) may not be centred around any $\tilde{\mathbf{b}}$, hence the approximation in (7) will have larger error. We will see this effect in practice later in section 3.1.1.

3.1.1 Empirical Study

For n players, we simulate 'abilities' characterised by $\boldsymbol{\lambda}$ through covariates $\mathbf{x}$ and coefficients $\boldsymbol{\beta}$. We simulate $S = 500$ contests and $T = 1000$ 'matches' per contest. Across all the contests, we initiated the following static data, where β was chosen randomly:
Parameters: $\beta = (3.31, 4.42, 4.15)$, $D = 1$
Covariates: for $i \in [1 : n]$:

$$x = (x_{i,1}, x_{i,2}, x_{i,3})$$
$$x_{i,1} \sim \mathcal{N}(0,1) \quad x_{i,2} \sim \mathcal{N}(2,1) \quad x_{i,3} \sim \mathcal{N}(1,1)$$

For each contest $s \in [1 : S]$ and for $i \in [1 : n]$:

$$\lambda_i = \beta^T x + b_i$$
$$b_i \sim \mathcal{N}(0, D)$$

For $t \in [1 : T]$:

- Sample player 1 uniformly from $[1 : n]$
- Sample player 2 uniformly from $[1 : n] \setminus \{\text{player } 1\}$
- Select winner based on probability (1)

In Figure 1, we can see how the bias varies based on the sparsity of observations when we use the PQL method on simulated data. We vary only the number of players, noting that fewer players leads to each pair of players having a greater number of matches on average. Intuitively, this increased 'density' of results (or reduced 'sparsity') means that there are more pair-wise data-points which improves the accuracy of estimates. More theoretically, Laplace's method will provide a better approximation due to the observed Binomial data approaching normality, as detailed above.

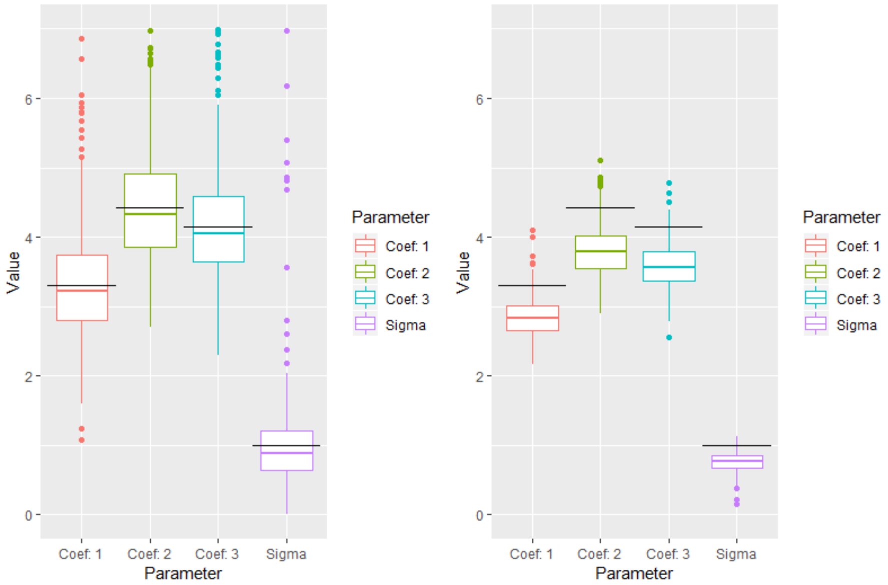

Figure 1: Box plots showing S simulated lots of $P = 3$ covariate coefficients estimates, β and random effect variance estimates, D, found from the PQL method on the Bradley Terry model. Horizontal black lines indicate the 'true' values from the data generating process. Left: $n = 10$, right: $n = 100$

4 MCMC for Bayesian Bradley Terry with Flat Priors

An alternative to PQL is a Markov Chain Monte Carlo (MCMC) approach under a Bayesian framework. A simple Bayesian approach would be to set Gaussian priors on the linear coefficients, β, and an inverse Chi-squared or Gamma distribution on the unknown variance, D. Prior information on the parameters may also help reduce the data requirement witnessed using PQL, however this will then raise the problem of hyper-parameter choice. To avoid this and target MLEs, flat priors are explored. Under non-informative priors, a MCMC approach to fit the Bradley Terry model can be used to compute parameter MLEs and avoids the bias induced by the Laplace approximation on sparse data. For full details on MCMC, including Metropolis-Hastings and Gibbs sampling see Roberts et al. (2004).

The dependency structure of the structured Bradley Terry model, detailed in Section 2, can be viewed as a directed graph as in Figure 2. In order to specify a Bayesian formulation based on the structured Bradley Terry model, one must set prior distributions on unknown parameters D, β_p for $i, j \in [1 : n]$, $p \in [1 : P]$. Zeger and Karim (1991) specify non-informative priors and detail a MCMC scheme to fit these parameters based on Gibbs Sampling and Rejection sampling nested within each Gibbs step.

The Gibbs sampling scheme suggested by Zeger and Karim (1991) is rather complex and targets the posterior distribution with flat priors detailed below, as follows:

1. Sample $\beta|\boldsymbol{b}, W$ using a non-informative prior:
   - Fit a Generalised Linear Model (GLM) where $\boldsymbol{b}$ are treated as non-random offsets in the linear predictor and recover estimated parameters $\hat{\beta} = \hat{\beta}_{1:P}$ and the covariance matrix of parameters, V.
   - Sample proposal $\beta^* \sim \mathcal{N}(\hat{\beta}, V)$.
   - Apply rejection sampling on the sampled β^* using the Binomial component of the likelihood function in equation (4), $\prod_{i=1}^{n} \prod_{j:j>i} f(W_{i,j}|\eta_{i,j})$, to induce a flat prior.
2. Sample variance $D|\mathbf{b}$ using the flat inverse Wishart prior (the inverse Chi-squared prior for the univariate case):
   - Sample from posterior Chi-squared distribution with parameters $\nu = n$, $s' = \sum b_i^2$.

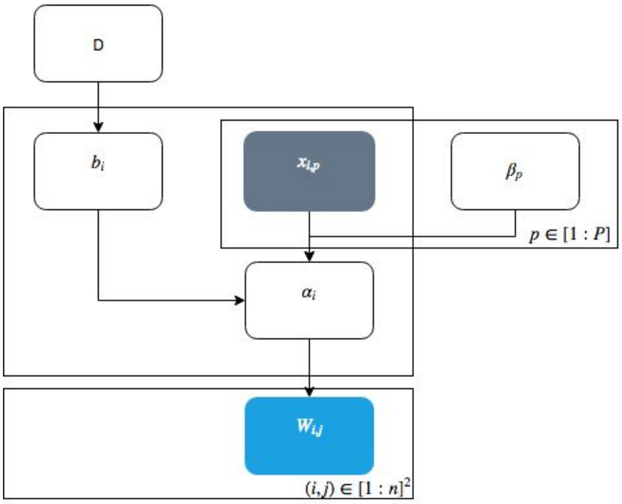

Figure 2: Plate diagram showing the dependency on observed win 'counts': $\{W_{ij}\}_{i,j}$ based on covariates $\{x_{i,j}\}_{i,j}$ and the unknowns: β_p, α_i, b_i and D

.

3. Sample errors $b_i|W, D, \beta$ for $i \in [1 : n]$ using rejection sampling for unnormalised target density, f_b, where $f_b(\mathbf{b}) = \prod_{i=1}^{n} \prod_{j:j>i} f(W_{i,j}|\eta_{i,j})\varphi_D(b_i)$:
    - Sample proposal b_i^* from $\mathcal{N}(0, 1)$, with density denoted $\varphi(\cdot)$.
    - Use numerical methods to find mode of f_b denoted: b_i^*.
    - Apply rejection sampling using normalising constant $f_b(b_i^*)/\varphi(b_i^*)$.

Figure 3 shows how the MCMC approach leads to sampled parameters that are closer on average to the underlying 'true' values. Although not shown explicitly in Figure 3, the average and mode of the sampled states are not visibly indistinguishable from the medians shown. The MCMC approach uses flat-priors, hence the posterior mode identified should coincide approximately with the MLE. The use of flat priors also means that the visible reduction in bias is due to the algorithm rather than from prior information.

5 Sequential Reduction Method

The sequential reduction method proposed by Ogden (2015) is another approach to approximate likelihoods in models by exploiting the dependence structure in the posterior distribution of the random effects.

For each random effect i, define the active set A_i to contain the indices of the non-zero elements of the i-th column of the random effect design matrix, $\mathbf{Z}$. The conditional independence structure can be illustrated graphically by constructing a graph $\mathcal{G}$ with

- A vertex for each random effect.
- An edge between vertices i and j if and only if $A_i \cap A_j \neq \phi$.

$\mathcal{G}$ is the posterior dependence graph for the random effects. Given the values of all the other random effects, if there is no edge between i and j, then b_i and b_j are conditionally independent in the posterior distribution of the random effects.

By writing the likelihood function as an integral over the random effects b_i, define:

$$g(b_1, \ldots, b_m|\beta, D, W) = \prod_{i=1}^{n} \prod_{j:j>i} f(W_{i,j}|\eta_{i,j}) \prod_{\ell=1}^{m} \varphi(b_\ell)$$

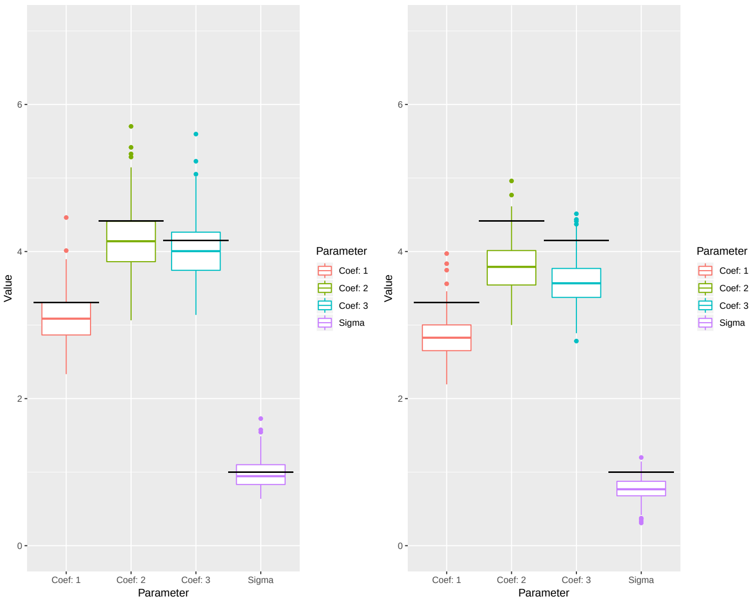

Figure 3: Same experiment as in section 3.1.1, with $n = 100$. Left: Boxplot of realised states in a single Markov Chain (of length 500) for Bayesian formulation detailed above. Right: $S = 500$ simulated PQL outputs on previously detailed inputs with $n = 100$. Horizontal black lines indicate the 'true' values from the data generating process.

to be the integrand, where g is proportional to the posterior density of the random effects given observations and parameters β and D. The likelihood we want to approximate becomes the normalising constant of the posterior distribution. By the Hammersley-Clifford theorem (Besag, 1974), since $g(\mathbf{b}|\beta, D, W) > 0$ for all $\mathbf{b}$, $g(\cdot|\beta, D, W)$ factorises over the maximal cliques of $\mathcal{G}$ as:

$$g(\mathbf{b}|\beta, D, W) = \prod_{C \in M(\mathcal{G})} g_C(\mathbf{b}_C)$$

for some functions $g_C(\cdot)$, where $M(\mathcal{G})$ denotes the set of all maximal cliques of $\mathcal{G}$. We can exploit this clique factorisation to approximate the desired likelihood through an iterative approach; where we first integrate the factorised form of $g(\mathbf{b}|\beta, D, W)$ with respect to b_1 to obtain a non-normalised posterior density of $\{b_2, \ldots, b_m\}$, then write the new posterior density as the product of maximal cliques of the new marginal posterior dependence graph and perform integration again.

Let $\mathcal{G}_i$ be the posterior dependence graph for $\{b_i, \ldots, b_m\}$, $M_i = M(\mathcal{G}_i)$ be the set of maximal cliques of $\mathcal{G}_i$, and N_i be the set containing the neighbours of vertex i. The general form of a sequential reduction method for approximating the model likelihood is given as follows:

1. Factorise $g(\mathbf{b}|\beta, D, W)$ over the maximal cliques M_1 of $\mathcal{G}_1$ as

$$\begin{aligned} g(b_1, \ldots, b_m|\beta, D, W) &= \tilde{g}(b_1, \ldots, b_m|\beta, D, W) \\ &= \prod_{C \in M_1} g_C^1(\mathbf{b}_C). \end{aligned}$$

2. For $i \in \{1, \ldots, m-1\}$, integrate with respect to b_i by integrating over the maximal cliques that contains vertex i

$$\int \tilde{g}(b_i, \ldots, b_m|\beta, D, W) db_i = \int \prod_{C \in M_i : C \subseteq N_i} g_C^i(b_C) db_i \prod_{\tilde{C} \in M_i : \tilde{C} \not\subseteq N_i} g_{\tilde{C}}^i(b_{\tilde{C}}).$$

   Let $g_{N_i}^1(b_i, \mathbf{b}_{N_i \setminus i}) = \prod_{C \in M_i : C \subseteq N_i} g_C^i(b_C)$, obtain

$$g_{N_i \setminus i}(\mathbf{b}_{N_i \setminus i}) = \int g_{N_i}(b_i, \mathbf{b}_{N_i \setminus i}) db_i$$

from integration (using a quadrature rule) and store an approximate representation $\tilde{g}_{N_i \setminus i}(\cdot)$ of the function $g_{N_i \setminus i}(\cdot)$. For all of the maximal cliques that are not subset of N_i, their functions remain unchanged, i.e. $g^{i+1}_{\tilde{C}}(\mathbf{b}_{\tilde{C}}) = g^{i}_{\tilde{C}}(\mathbf{b}_{\tilde{C}})$, and we have

$$\tilde{g}(b_{i+1}, \ldots, b_m | \beta, D, W) = \tilde{g}_{N_i \setminus i}(\mathbf{b}_{N_i \setminus i}) \prod_{\tilde{C} \in M_i : \tilde{C} \not\subseteq N_i} g^{i}_{\tilde{C}}(b_{\tilde{C}}).$$

3. Integrate $\tilde{g}(b_m|\beta, D, W)$ with respect to b_n to obtain the approximation to the model likelihood.

During the implementation of the sequential reduction method, three choices need to be made. The first choice is the factorisation of the maximal cliques M_1. The factorisation used in model implementation orders the cliques lexicographically, and a detailed explanation of this can be found in Ogden (2015). The second choice is the order in which we integrate out the random effects. Although the random effects can be integrated out in any order, it is more efficient to perform the integration in the order that minimises the cost of approximation. The desired ordering is such that the largest clique obtained by joining together all neighbours of the removed vertex is as small as possible at each stage. Over all possible orderings, the smallest possible value of the largest clique obtained is called the treewidth of the graph. However, the algorithms available to calculate the treewidth can be costly. The approach taken by Ogden (2015) is a constructive algorithm to find an upper bound for the treewidth for a reasonably good ordering that may not be optimal.

The final choice concerns the approximate representation $\tilde{g}_{N_i \setminus i}(\cdot)$, which we store for the purposes of evaluating the likelihood. In practice, this storage consists of a set of points at which $g_{N_i \setminus i}(\cdot)$ is to be evaluated, as well as a method of interpolation between those points. The approximate representation is made up of several stages, as detailed by Ogden (2015), but mainly seeks to minimize the size of the absolute error in the interpolation. The more points at which we evaluate the approximate representation, the more accurate the interpolation will be. However, evaluation at a large number of points can be computationally intensive, so it is important to place a suitable restriction on the number of evaluation points.

20 sets of data are simulated for each of the likelihood estimation methods, sequential reduction and PQL. All of the data are simulated according to the construction in Section 3.1.1, only the parameter β is different, with values $\beta = (0.451, 3.25, 2.71)$. A Bradley-Terry model is fitted to each dataset using the likelihood approximation method specified, the parameter estimates are collected and plotted in the box-plots in Figure 4. For the sequential reduction method, approximating the likelihood at each point takes on average 0.71 seconds, which is notably longer than the time taken for PQL. Sequential reduction shows a significant improvement to PQL in the accuracy of parameter estimation, however, we also noticed that the variances of estimators by sequential reduction are larger.

While sequential reduction shines in those situations where PQL struggles, particularly when faced with sparse, binary data (Ogden, 2015), this does not come without its own difficulties. In some scenarios, sequential reduction can become computationally intensive, meaning that other approximations to the likelihood should be considered. Unfortunately, this issue can arise even in situations where other methods, such as PQL and Laplace approximations, perform poorly (Ogden, 2015).

6 Conclusion

It is clear that there is significant bias in the PQL method for sparse contest data thus making the approach inappropriate for such data. With significant added complexity and computation the MCMC approach given by Zeger and Karim (1991) does appear to generate estimators with reduced bias, but still non-zero bias. It may be possible that informative priors, other Bayesian MCMC, Sequential Monte Carlo or Expectation Maximisation (EM), such as that advocated by Caron and Doucet (2012), may perform better. Indeed, it would be an interesting extension to incorporate covariates into Caron and Doucet (2012)'s model formulation.

A promising alternate to the Bayesian/MCMC route was that of Sequential Reduction, explored in Section 5. Visually this appeared to produce the least biased results of the methods attempted,

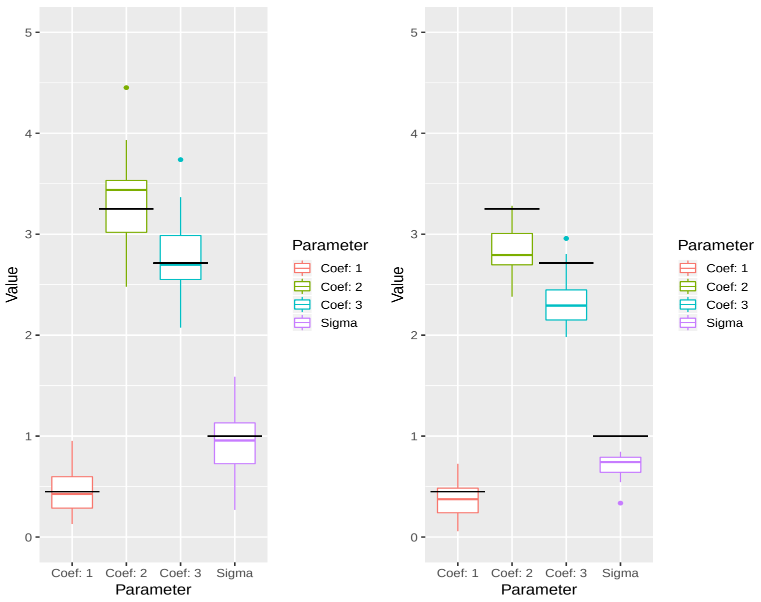

Figure 4: Similar experiment as in section 3.1.1, with $n = 100$, $S = 20$. Horizontal black lines indicate the 'true' values from the data generating process. Box-plots showing parameter estimation for simulated data with likelihood approximated by sequential reduction (left) and PQL (right).

however the implementation was computationally expensive, was unable to run for a large number of covariates and was time-consuming to run on even our toy dataset.

Further work is certainly required to identify a satisfactory approach.

References

Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

Francois Caron and Arnaud Doucet. Efficient Bayesian Inference for Generalized Bradley-Terry Models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.

Yanan Fan, David S Leslie, MP Wand, et al. Generalised Linear Mixed Model Analysis via Sequential Monte Carlo Sampling. *Electronic Journal of Statistics*, 2:916–938, 2008.

Peter Hall, Tung Pham, Matt P Wand, Shen SJ Wang, et al. Asymptotic Normality and Valid Inference for Gaussian Variational Approximation. *The Annals of Statistics*, 39(5):2502–2532, 2011.

K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. MIT Press, 2012. ISBN 9780262018029. URL `https://books.google.co.in/books?id=NZP6AQAAQBAJ`.

Helen E Ogden. A Sequential Reduction Method for Inference in Generalized Linear Mixed Models. *Electronic Journal of Statistics*, 9(1):135–152, 2015.

Gareth O Roberts, Jeffrey S Rosenthal, et al. General State Space Markov Chains and MCMC Algorithms. *Probability Surveys*, 1:20–71, 2004.

Scott L Zeger and M Rezaul Karim. Generalized Linear Models with Random Effects; A Gibbs Sampling Approach. *Journal of the American statistical association*, 86(413):79–86, 1991.

A Laplace Approximation Detail

The integrated quasi-likelihood function for GLMMs is given by:

$$e^{ql(\boldsymbol{\beta},\theta)} \propto |\mathbf{D}|^{-1/2} \int \exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i,\mu_i^b) - \frac{1}{2}\mathbf{b}^T\mathbf{D}^{-1}\mathbf{b}\right\} d\mathbf{b}$$

where $d_i(y,\mu) = -2\int_y^\mu \frac{y-u}{a_i v(u)}du$ is the deviance measure of fit and the true log-quasi-likelihood is denoted by $ql(\boldsymbol{\beta},\theta)$. This can be rewritten in the form:

$$e^{ql(\boldsymbol{\beta},\theta)} \propto c|\mathbf{D}|^{-1/2} \int \mathrm{e}^{-\kappa(\mathbf{b})} d\mathbf{b}.$$

Applying Laplace's method of integral approximation (essentially a Taylor series expansion around a minimum) yields the following:

$$ql(\boldsymbol{\beta},\theta) \approx -\frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\log|\kappa''(\tilde{\mathbf{b}})| - \kappa(\tilde{\mathbf{b}}),$$

where $\tilde{\mathbf{b}}$ is chosen to minimise $\kappa(\mathbf{b})$. That is, $\tilde{\mathbf{b}}$ solves the equation:

$$\kappa'(\mathbf{b}) = -\sum_{i=1}^{n} \frac{(y_i - \mu_i^b)z_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} + \mathbf{D}^{-1}\mathbf{b} = 0.$$

Differentiating, we obtain:

$$\kappa''(\mathbf{b}) = \sum_{i=1}^{n} \frac{z_i z_i^T}{\phi a_i v(\mu_i^b)[g'(\mu_i^b)]^2} + \mathbf{D}^{-1} + \mathbf{R} \approx \mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{D}^{-1},$$

where $\mathbf{R}$ is a remainder term and $\mathbf{W}$ is an $n\times n$ matrix with diagonal entries $w_i = \{\phi a_i v(\mu_i^b)[g'(\mu_i^b)]^2\}^{-1}$. These entries are the generalised linear model iterated weights. With these expressions, we can now write the log-quasi-likelihood as:

$$ql(\boldsymbol{\beta}) \approx -\frac{1}{2}\log|\mathbf{I} + \mathbf{Z}^T\mathbf{W}\mathbf{Z}\mathbf{D}| - \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i,\mu_i^{\tilde{b}}) - \frac{1}{2}\tilde{\mathbf{b}}^T\mathbf{D}^{-1}\tilde{\mathbf{b}}.$$