Modified-likelihood estimation of the β-model

Koen Jochmans

SciencesPo
DEPARTMENT OF ECONOMICS

Discussion paper 2016-01

MODIFIED-LIKELIHOOD ESTIMATION OF THE *B*-MODEL

Koen Jochmans

Sciences Po Economics Discussion Papers

MODIFIED-LIKELIHOOD ESTIMATION OF THE β-MODEL

By Koen Jochmans

Sciences Po, Paris

February 9, 2016

We consider point estimation and inference based on modifications of the profile likelihood in models for dyadic interactions between n agents featuring agent-specific parameters. This setup covers the β-model of network formation and generalizations thereof. The maximum-likelihood estimator of such models has bias and standard deviation of $O(n^{-1})$ and so is asymptotically biased. Estimation based on modified likelihoods leads to estimators that are asymptotically unbiased and likelihood-ratio tests that exhibit correct size. We apply the modifications to versions of the β-model for network formation and of the Bradley-Terry model for paired comparisons.

1. Introduction. A growing literature has uncovered the importance of interactions between agents through social networks as drivers for economic outcomes. Examples include employment opportunities (Calvó-Armengol and Jackson 2004), risk sharing (Fafchamps and Gubert 2007; Jackson, Rodriguez-Barraquer and Tan 2012), and also educational achievements (Calvó-Armengol, Patacchini and Zenou 2009).

A leading approach to statistical modelling of dyadic interaction is through the inclusion of agent-specific parameters (see, e.g., Snijders 2011). One of the most popular applications of this paradigm is the β-model for network formation. In this model, agent fixed effects capture degree heterogeneity in link formation (Chatterjee, Diaconis and Sly 2011, Rinaldo, Petrovic and Fienberg 2013, and Yan and Xu 2013 are recent references). Graham [2015] augments the standard β-model with observable dyad characteristics. This allows to empirically distinguish degree heterogeneity from homophily (see Jackson 2008 and McPherson, Smith-Lovin and Cook 2001 for discussion on the importance of homophily in network formation). Clearly, estimation of such fixed-effect models is non-standard as the number of parameters grows

Keywords and phrases: asymptotic bias, β-model, Bradley-Terry model, fixed effects, modified profile likelihood, paired comparisons, matching, network formation, undirected random graph

with the sample size. Inference on the homophily parameter is plagued by asymptotic bias that needs to be corrected for. The bias problem comes from the presence of the agent-specific parameters in the model, and is similar to the well-known incidental-parameter problem (Neyman and Scott 1948) in models for panel data. The same problem appears in more general models for dyadic interactions between heterogenous agents, such as in the references given above.

The problem of inference in the presence of many nuisance parameters has a long history in statistics. In this paper we look at generic estimation problems for dyadic data and argue in favor of inference based on modified likelihood functions. In its most general form, the modified likelihood is a bias-corrected version of the profile likelihood, that is, of the likelihood after having profiled-out the nuisance parameters. The adjustment is both general and simple in form, involving only the score and Hessian of the likelihood with respect to the nuisance parameters. The adjustment term removes the leading bias from the profile likelihood and leads to asymptotically-unbiased inference and likelihood ratio statistics that are χ^2-distributed under the null. The form of the adjustment can be specialized by using the likelihood structure (as in DiCiccio et al. 1996), in which case the modified likelihood penalizes the profile likelihood for deviations from the information equality, arising due to the estimation noise in the fixed effects.[1]

We work out the modifications to the profile likelihood in a linear version of the β-model and in a linear version of the Bradley and Terry [1952] model for paired comparisons. These simple illustrations give insight in how the adjustments work. We next apply them to the β-model of Graham [2015], and evaluate our approach using his simulation designs. We find that both modifications dramatically improve on maximum likelihood in terms of bias and mean squared error as well as reliability of statistical inference, and that they can be more reliable than bias-correcting the maximum-likelihood estimator.

[1] It can be further simplified when an information-orthogonal reparametrization of β_i exist, as in Cox and Reid [1987] and Lancaster [2002]. However, as such reparametrizations do not exist in general (see, e.g., Severini 2000) we do not consider such modifications further here.

2. Fixed-effect models for dyadic data. We consider data on dyadic interactions between n agents. For each of $n(n-1)/2$ distinct agent pairs (i,j) with $i<j$ we observe the random variable z_{ij}. The data may be multivariate. For example, we may observe an outcome y_{ij} generated by pair (i,j) together with a vector of dyad characteristics x_{ij}, in which case we have $z_{ij}=(y_{ij},x_{ij}')'$.

The density of z_{ij} (relative to some dominating measure) takes the form

$$f(z_{ij};\vartheta,\beta_i,\beta_j),$$

where ϑ and $\beta_1,\ldots,\beta_n$ are unknown Euclidean parameters. Models of this form are relevant for the analysis of network formation (see, for example, Fafchamps and Gubert 2007 and Attanasio et al. 2009 for applications), for studying strategic behavior among agents, or for the construction of rankings (Bradley and Terry 1952).

Our goal will be to perform inference on ϑ treating the β_i as fixed effects. As is well known, the maximum-likelihood estimator of ϑ generally performs poorly when the number of nuisance parameters is large relative to the sample size (see, for example, Neyman and Scott 1948 and Li, Lindsay and Waterman 2003). We will consider modifications of the maximum-likelihood method that yield estimators with good statistical properties. Before doing so, we first discuss two leading types of models that fit into our framework.

2.1. *Models with complementarity.* In one important class of models the parameters β_i and β_j enter the density of the dyad (i,j) additively, that is,

$$f(z_{ij};\vartheta,\beta_i,\beta_j)=f(z_{ij};\vartheta,\beta_i+\beta_j).$$

Such models arise, for example, when agents i and j interact to generate output, and working together creates a surplus.

A leading example of such a model is the β-model of network formation (Chatterjee, Diaconis and Sly 2011; Graham 2015). Here, $z_{ij}=(y_{ij},x_{ij})'$ and the probability distribution of y_{ij} given x_{ij} is Bernoulli with parameter

$$\mathbb{P}(y_{ij}=1|x_{ij};\vartheta,\beta_i,\beta_j)=F(\beta_i+\beta_j+x_{ij}\vartheta)$$

for a cumulative distribution function F. The likelihood of i and j to engage in a relation with one another is increasing in $\beta_i+\beta_j$ and also in $x_{ij}\vartheta$.

The first factor represents degree heterogeneity among agents. Some agents are more open to interaction than others, and engaging in a relation is a reciprocal decision. The second factor can capture homophily, that is, the fact that agents with similar characteristics are more likely to interact. If x_{ij} is a measure of distance (or dissimilarity), for example, homophily would correspond to $\vartheta < 0$. Disentangling these two factors is a long-standing problem. See McPherson, Smith-Lovin and Cook [2001] for an overview on the large literature on homophily.

More generally, one may wish to analyze an output y_{ij} produced by the dyad (i, j) rather than the production decision. Such a problem can often be cast into a (nonlinear) regression model of the form

$$y_{ij} = t(\beta_i + \beta_j + x'_{ij}\vartheta, \varepsilon_{ij})$$

for a transformation function t and a latent disturbance ε_{ij}. The motivation of the index would be as before. One application of this type of regression model would be a gravity equation for bilateral trade flows featuring country fixed effects (see Head and Mayer 2014 for an overview of the large literature on gravity equations).

2.2. *Models with competition.* An alternative specification can feature agent-specific parameters to express substitutability or competition effects. An example is the model of Bradley and Terry [1952] for paired comparisons. Again letting $z_{ij} = (y_{ij}, x_{ij})'$, this model has

$$\mathbb{P}(y_{ij} = 1 | x_{ij}; \vartheta, \beta_i, \beta_j) = F(\beta_i - \beta_j + x_{ij}\vartheta).$$

In contrast to the β-model, here, this probability is increasing in β_i and decreasing in β_j. In the same vain, x_{ij} could be a the difference between various characteristics of the agents in the dyad. One application of such a framework is in modelling the outcome of contests (Simons and Yao 1999). In an industrial-organization context, the framework could be used to model strategic interactions such as market-entry decisions.

Generalizations of the Bradley and Terry [1952] model are derived in Hunter [2004], and the techniques discussed below will equally be applicable to such generalizations.

Contrary to in the models with complementarities above, here, the mean of the β_i is undetermined. Therefore, a normalization needs to be imposed

when estimating these models. One common normalization is $\sum_{i=1}^{n} \beta_i = 0$. An alternative normalization would be to set $\beta_i = 0$ for some chosen i. The choice of normalization is irrelevant for estimation and inference on ϑ.

3. Estimation and inference. The log-likelihood for our data is

$$\ell(\vartheta, \beta) = \sum_{i=1}^{n} \sum_{i<j} \log f(z_{ij}; \vartheta, \beta_i, \beta_j),$$

where we let $\beta = (\beta_1, \ldots, \beta_n)'$. For simplicity of exposition we ignore any normalization that may be needed on β to achieve identification. When a normalization of the form $c(\beta) = 0$ is needed, everything to follow goes through on replacing $\ell(\vartheta, \beta)$ by the constrained likelihood $\ell(\vartheta, \beta) - \lambda\, c(\beta)$, where λ denotes the Lagrange multiplier. We will give a detailed example below.

It is useful to recall that the maximum-likelihood estimator of ϑ can be expressed as

$$\hat{\vartheta} = \arg\max_{\vartheta} \hat{\ell}(\vartheta),$$

where $\hat{\ell}(\vartheta) = \ell(\vartheta, \hat{\beta}(\vartheta))$, with

$$\hat{\beta}(\vartheta) = \arg\max_{\beta} \ell(\vartheta, \beta),$$

is the profile likelihood.

Inference based on the profile likelihood performs poorly, even in large samples, because the dimension of β is n, which grows with the sample size $n(n-1)/2$. Quite generally, estimating the n parameters β_i along with ϑ will imply that

$$\mathbb{E}(\hat{\vartheta} - \vartheta) = O(n^{-1}).$$

As $\mathbb{E}((\hat{\vartheta} - \vartheta)^2) = O(n^{-2})$, bias and standard deviation are of the same order of magnitude, and the maximum-likelihood estimator is asymptotically biased.

3.1. *Modified profile likelihood.* Estimation and inference in the presence of nuisance parameters has a long history in statistics. The seminal work of Barndorff-Nielsen [1983] and Cox and Reid [1987] contains modifications to the profile likelihood that lead to superior inference. More recent work

includes DiCiccio et al. [1996] and Severini [1998]. Modified likelihoods have been found to solve the incidental-parameter problem in models for panel data under so-called rectangular-array asymptotics (as defined in Li, Lindsay and Waterman 2003). See, notably, Sartori [2003] and Arellano and Hahn [2007]. We argue that they can equally be used to yield asymptotically-valid inference in the current context.

In its simplest form, modified likelihoods can be understood as yielding a superior approximation to the target likelihood

$$\ell(\vartheta) = \ell(\vartheta, \beta(\vartheta)), \qquad \beta(\vartheta) = \arg\max_{\beta} \mathbb{E}(\ell(\vartheta, \beta)).$$

Moreover, the profile likelihood is the sample counterpart to this infeasible likelihood. Replacing $\beta(\vartheta)$ with $\hat{\beta}(\vartheta)$ introduces bias that leads to invalid inference.

Under regularity conditions we have that

$$\hat{\beta}(\theta) - \beta(\vartheta) = \Sigma(\vartheta)^{-1} V(\vartheta) + O_p(n^{-1}),$$

where we introduce

$$V(\vartheta) = \left.\frac{\partial \ell(\vartheta, \beta)}{\partial \beta}\right|_{\beta=\beta(\vartheta)}, \qquad \Sigma(\vartheta) = -\mathbb{E}\left(\left.\frac{\partial^2 \ell(\vartheta, \beta)}{\partial \beta \partial \beta'}\right)\right|_{\beta=\beta(\vartheta)}.$$

An expansion of the profile likelihood around $\beta(\vartheta)$ yields

$$\begin{aligned}\hat{\ell}(\vartheta) - \ell(\vartheta) = \quad & (\hat{\beta}(\theta) - \beta(\vartheta))' V(\vartheta) \\ & - \frac{1}{2}(\hat{\beta}(\theta) - \beta(\vartheta))' \Sigma(\vartheta)(\hat{\beta}(\theta) - \beta(\vartheta)) + O_p(n^{-1/2}).\end{aligned}$$

Combining the two expansions and taking expectations then shows that the bias of the profile likelihood is of the form

$$\mathbb{E}(\hat{\ell}(\vartheta) - \ell(\vartheta)) = \frac{1}{2}\text{trace}(\Sigma(\vartheta)^{-1}\Omega(\vartheta)) + O(n^{-1/2})$$

for

$$\Omega(\vartheta) = \mathbb{E}[V(\vartheta)\, V(\vartheta)'],$$

the variance of $V(\vartheta)$.

A modified likelihood then is

$$\dot{\ell}(\vartheta) = \hat{\ell}(\vartheta) - \frac{1}{2}\text{trace}(\hat{\Sigma}(\vartheta)^{-1}\hat{\Omega}(\vartheta)),$$

where we define the plug-in estimators

$$\hat{\Sigma}(\vartheta) = \hat{\Sigma}(\vartheta, \hat{\beta}(\vartheta)), \qquad \hat{\Omega}(\vartheta) = \hat{\Omega}(\vartheta, \hat{\beta}(\vartheta)),$$

for matrices

$$-(\hat{\Sigma}(\vartheta,\beta))_{i,j} = \begin{cases} \sum_{i<j} \frac{\partial^2 \log f(z_{ij};\vartheta,\beta_i,\beta_j)}{\partial \beta_i^2} + \sum_{i>j} \frac{\partial^2 \log f(z_{ji};\vartheta,\beta_j,\beta_i)}{\partial \beta_i^2} & \text{if } i = j \\ \frac{\partial^2 \log f(z_{ij};\vartheta,\beta_i,\beta_j)}{\partial \beta_i \partial \beta_j} & \text{if } i < j \\ \frac{\partial^2 \log f(z_{ji};\vartheta,\beta_j,\beta_i)}{\partial \beta_i \partial \beta_j} & \text{if } i > j \end{cases}$$

and

$$(\hat{\Omega}(\vartheta,\beta))_{i,j} = \begin{cases} \sum_{i<j} \left(\frac{\partial \log f(z_{ij};\vartheta,\beta_i,\beta_j)}{\partial \beta_i}\right)^2 + \sum_{i>j} \left(\frac{\partial \log f(z_{ji};\vartheta,\beta_j,\beta_i)}{\partial \beta_i}\right)^2 & \text{if } i = j \\ \left(\frac{\partial \log f(z_{ij};\vartheta,\beta_i,\beta_j)}{\partial \beta_i}\right)^2 & \text{if } i < j \\ \left(\frac{\partial \log f(z_{ji};\vartheta,\beta_j,\beta_i)}{\partial \beta_i}\right)^2 & \text{if } i > j \end{cases}$$

In large samples, this modification removes the leading bias from the profile likelihood. Consequently, in large samples, the likelihood-ratio statistic has correct size and

$$\dot{\vartheta} = \arg\max_{\vartheta} \dot{\ell}(\vartheta),$$

will have bias $o(n^{-1})$. Furthermore, under regularity conditions, we have the limit result

$$\dot{H}(\dot{\vartheta})^{1/2}(\dot{\vartheta} - \vartheta) \xrightarrow{d} \mathcal{N}(0, I_{\dim \vartheta})$$

as $n \to \infty$, where we let

$$\dot{H}(\vartheta) = -\frac{\partial^2 \dot{\ell}(\vartheta)}{\partial \vartheta \partial \vartheta'}$$

be the observed Fisher information for ϑ derived from $\dot{\ell}(\vartheta)$.

Following the arguments in Arellano and Hahn [2007] we can exploit the likelihood structure to get

$$\frac{1}{2}\text{trace}(\hat{\Sigma}(\vartheta)^{-1}\hat{\Omega}(\vartheta)) = -\frac{1}{2}\log(\det \hat{\Sigma}(\vartheta)) + \frac{1}{2}\log(\det \hat{\Omega}(\vartheta)) + O(n^{-1}),$$

which validates the alternative modified likelihood

$$\ddot{\ell}(\vartheta) = \hat{\ell}(\vartheta) + \frac{1}{2}\log(\det \hat{\Sigma}(\vartheta)) - \frac{1}{2}\log(\det \hat{\Omega}(\vartheta));$$

see DiCiccio et al. [1996]. Its maximizer, say $\ddot{\vartheta}$ satisfies the same asymptotic properties as $\dot{\vartheta}$.

3.2. *Illustration: A linear β-model.* Consider the following extension of the classic many normal means problem of Neyman and Scott [1948]. Data are generated as

$$z_{ij} \sim \mathcal{N}(\beta_i + \beta_j, \vartheta),$$

and are independent across dyads. The likelihood function for all parameters (ignoring constants) is

$$\ell(\vartheta, \beta) = -\frac{1}{2}\frac{n(n-1)}{2}\log\vartheta - \frac{1}{2}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij} - \beta_i - \beta_j)^2}{\vartheta}.$$

Its first two derivatives with respect to the β_i are

$$\frac{\partial\ell(\vartheta,\beta)}{\partial\beta_i} = \sum_{i<j}\frac{z_{ij} - \beta_i - \beta_j}{\vartheta} + \sum_{i>j}\frac{z_{ji} - \beta_j - \beta_i}{\vartheta}$$

and

$$\frac{\partial^2\ell(\vartheta,\beta)}{\partial\beta_i\partial\beta_j} = \begin{cases} -\frac{(n-1)}{\vartheta} & \text{if } i = j \\ -\frac{1}{\vartheta} & \text{if } i \neq j \end{cases}.$$

Let $\tilde{z}_i = (n-2)^{-1}\sum_{i<j} z_{ij} + (n-2)^{-1}\sum_{i>j} z_{ji}$ and $\overline{z} = (2(n-1)^{-1}\sum_{i=1}^{n}\tilde{z}_i$. Solving for the maximum-likelihood estimator of β_i gives $\hat{\beta}_i = \tilde{z}_i - \overline{z}$ for any ϑ. The profile likelihood is therefore

$$\hat{\ell}(\vartheta) = -\frac{n(n-1)}{2}\log\vartheta - \frac{1}{2}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij} - (\tilde{z}_i - \overline{z}) - (\tilde{z}_j - \overline{z}))^2}{\vartheta},$$

and its maximizer is

$$\hat{\vartheta} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i<j}(z_{ij} - (\tilde{z}_i - \overline{z}) - (\tilde{z}_j - \overline{z}))^2.$$

Some tedious but straightforward calculations yield

$$\hat{\vartheta} - \vartheta \sim \mathcal{N}\left(-\frac{2}{n-1}\vartheta,\ \frac{n-3}{n-1}\frac{2\,\vartheta^2}{n(n-1)/2}\right),$$

which confirms that the maximum-likelihood estimator of ϑ suffers from asymptotic bias.

To set up the modified likelihood, first note that

$$(\hat{\Sigma}(\vartheta))_{i,j} = \begin{cases} \frac{n-1}{\vartheta} & \text{if } i = j \\ \frac{1}{\vartheta} & \text{if } i \neq j \end{cases}, \quad (\hat{\Sigma}(\vartheta)^{-1})_{i,j} = \begin{cases} \frac{\vartheta}{2}\frac{2n-3}{(n-1)(n-2)} & \text{if } i = j \\ -\frac{\vartheta}{2}\frac{1}{(n-1)(n-2)} & \text{if } i \neq j \end{cases},$$

and that

$$(\hat{\Omega}(\vartheta))_{i,j} = \begin{cases} \sum_{i<j} \frac{(z_{ij}-(\tilde{z}_i-\overline{z})-(\tilde{z}_j-\overline{z}))^2}{\vartheta^2} + \sum_{i>j} \frac{(z_{ji}-(\tilde{z}_j-\overline{z})-(\tilde{z}_i-\overline{z}))^2}{\vartheta^2} & \text{if } i=j \\ \frac{(z_{ij}-(\tilde{z}_i-\overline{z})-(\tilde{z}_j-\overline{z}))^2}{\vartheta^2} & \text{if } i<j \\ \frac{(z_{ji}-(\tilde{z}_j-\overline{z})-(\tilde{z}_i-\overline{z}))^2}{\vartheta^2} & \text{if } i>j \end{cases}.$$

It is then easily seen that

$$\frac{1}{2}\text{trace}(\hat{\Sigma}(\vartheta)^{-1}\hat{\Omega}(\vartheta)) = \frac{1}{2}\,\frac{2}{n-1}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij}-(\tilde{z}_i-\overline{z})-(\tilde{z}_j-\overline{z}))^2}{\vartheta}.$$

From this we obtain

$$\dot{\ell}(\vartheta) = -\frac{n(n-1)}{2}\log\vartheta - \left(1+\frac{2}{n-1}\right)\frac{1}{2}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij}-(\tilde{z}_i-\overline{z})-(\tilde{z}_j-\overline{z}))^2}{\vartheta},$$

and its maximizer

$$\dot{\vartheta} = \frac{n+1}{n-1}\,\hat{\vartheta} = \hat{\vartheta} + \frac{2}{n-1}\hat{\vartheta}.$$

Clearly, this estimator removes the leading bias from the maximum-likelihood estimator. Moreover,

$$\dot{\vartheta}-\vartheta \sim \mathcal{N}\left(-\left(\frac{2}{n-1}\right)^2\vartheta,\ \frac{n(n(n-1)-5)}{(n-1)^3}\,\frac{2\,\vartheta^2}{n(n-1)/2}\right),$$

which shows that the remaining bias in the point estimator is small relative to its standard deviation.

As an alternative correction, we may exploit the likelihood structure to adjust the profile likelihood by the term

$$-\frac{1}{2}\log(\det\hat{\Sigma}(\vartheta)) + \frac{1}{2}\log(\det\hat{\Omega}(\vartheta)) = \frac{n}{2}\,\log\,\vartheta + c,$$

where c is a constant that does not depend on ϑ. This yields the modification

$$\ddot{\ell}(\vartheta) = -\frac{n(n-3)}{2}\log\vartheta - \frac{1}{2}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij}-(\tilde{z}_i-\overline{z})-(\tilde{z}_j-\overline{z}))^2}{\vartheta},$$

whose maximizer satisfies

$$\ddot{\vartheta}-\vartheta \sim \mathcal{N}\left(0, \frac{2\,\vartheta^2}{n(n-3)/2}\right).$$

This estimator is exactly unbiased.

To give an idea of the magnitude of the bias in this problem, Table 1 contains the bias and standard deviation of the estimators $\hat{\vartheta}$, $\dot{\vartheta}$, and $\ddot{\vartheta}$ for various sample sizes n and variance parameter fixed to $\vartheta = 1$. These results are invariant to the value of the β_i.

TABLE 1
Many normal means

n	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$
		bias			standard deviation	
5	-0.5000	-0.2500	0.0000	0.3162	0.4743	0.6325
10	-0.2222	-0.0494	0.0000	0.1859	0.2272	0.2390
15	-0.1429	-0.0204	0.0000	0.1278	0.1460	0.1491
20	-0.1053	-0.0111	0.0000	0.0970	0.1073	0.1085
25	-0.0833	-0.0069	0.0000	0.0782	0.0847	0.0853
50	-0.0408	-0.0017	0.0000	0.0396	0.0412	0.0413
75	-0.0270	-0.0007	0.0000	0.0265	0.0272	0.0272
100	-0.0202	-0.0004	0.0000	0.0199	0.0203	0.0203

3.3. *Illustration: A linear Bradley-Terry model.* As an alternative to the Neyman and Scott [1948] model with complementarities, now suppose that

$$z_{ij} \sim \mathcal{N}(\beta_i - \beta_j, \vartheta)$$

independently across dyads. This model is overparametrized as, clearly, the mean of the β_i is not identified. A common normalization in this type of model is $\sum_{i=1}^{n} \beta_i = 0$ (Simons and Yao 1999), and we will maintain it here. The constrained likelihood is

$$-\frac{1}{2}\frac{n(n-1)}{2}\log\vartheta - \frac{1}{2}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij} - \beta_i + \beta_j)^2}{\vartheta} + \lambda\sum_{i=1}^{n}\beta_i,$$

where λ is the Lagrange multiplier for our normalization constraint. The first-order condition for the constrained problem for β_i for a given ϑ equals

$$\frac{\sum_{i<j} z_{ij} - \sum_{i>j} z_{ji}}{\vartheta} - \frac{n}{\vartheta}\beta_i = 0.$$

This gives

$$\hat{\beta}_i = \frac{\sum_{i<j} z_{ij} - \sum_{i>j} z_{ji}}{n} = \tilde{z}_i \quad \text{(say)}$$

for all i and any ϑ. Observe that the sign of $\hat{\beta}_i$ is driven by the comparison of the magnitudes of $\sum_{i<j} z_{ij}$ and $\sum_{i>j} z_{ji}$. Also note that $\sum_{i=1}^{n} \hat{\beta}_i = 0$ holds. We therefore have

$$\hat{\ell}(\vartheta) = -\frac{1}{2}\frac{n(n-1)}{2}\log\vartheta - \frac{1}{2}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij} - \tilde{z}_i + \tilde{z}_j)^2}{\vartheta},$$

and with it, the maximum-likelihood estimator

$$\hat{\vartheta} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i<j}(z_{ij} - \tilde{z}_i + \tilde{z}_j)^2.$$

A calculation shows that $\mathbb{E}(\hat{\vartheta} - \vartheta) = -2n^{-1}$.

It is immediate that

$$\hat{\Sigma}(\vartheta) = \operatorname{diag}\left(\frac{n}{\vartheta}\right), \qquad \hat{\Sigma}(\vartheta)^{-1} = \operatorname{diag}\left(\frac{\vartheta}{n}\right),$$

and that

$$(\hat{\Omega}(\vartheta))_{i,j} = \begin{cases} \sum_{i<j}\frac{(z_{ij}-\tilde{z}_i+\tilde{z}_j)^2}{\vartheta^2} + \sum_{i>j}\frac{(z_{ji}-\tilde{z}_j+\tilde{z}_i)^2}{\vartheta^2} & \text{if } i = j \\ \frac{(z_{ij}-\tilde{z}_i+\tilde{z}_j)^2}{\vartheta^2} & \text{if } i < j \\ \frac{(z_{ji}-\tilde{z}_j+\tilde{z}_i)^2}{\vartheta^2} & \text{if } i > j \end{cases}.$$

Therefore,

$$\dot{\ell}(\vartheta) = -\frac{1}{2}\frac{n(n-1)}{2}\log\vartheta - \frac{1}{2}\left(1+\frac{2}{n}\right)\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij} - \tilde{z}_i + \tilde{z}_j)^2}{\vartheta},$$

$$\ddot{\ell}(\vartheta) = -\frac{1}{2}\frac{n(n-3)}{2}\log\vartheta - \frac{1}{2}\sum_{i=1}^{n}\sum_{i<j}\frac{(z_{ij} - \tilde{z}_i + \tilde{z}_j)^2}{\vartheta}.$$

The corresponding estimators are

$$\dot{\vartheta} = \left(1+\frac{2}{n}\right)\hat{\vartheta}, \qquad \ddot{\vartheta} = \frac{n-1}{n-3}\,\hat{\vartheta} = \left(1+\frac{2}{n-3}\right)\hat{\vartheta}.$$

Both remove the leading bias from the maximum-likelihood estimator, as

$$\mathbb{E}(\dot{\vartheta} - \vartheta) = -\frac{4}{n^2} = O(n^{-2}), \qquad \mathbb{E}(\ddot{\vartheta} - \vartheta) = \frac{2}{n(n-3)} = O(n^{-2}),$$

but, in this case, neither is exactly unbiased. The first estimator has bias that is strictly negative (for any finite n). The second estimator overcorrects and has strictly positive bias. The second-order bias is monotone in n. We have

$$\frac{4}{n^2} > \frac{2}{n(n-3)}$$

for all $n > 7$. As $n \to \infty$,

$$\sqrt{\frac{n(n-1)}{2}}(\dot{\vartheta} - \vartheta) \xrightarrow{d} \mathcal{N}(0, 2\vartheta^2),$$

and $\|\ddot{\vartheta} - \dot{\vartheta}\| = o_p(n^{-1})$; that is, the two modifications to the likelihood yield asymptotically-equivalent estimators.

4. Application to the β-model. The β-model of network formation (Chatterjee, Diaconis and Sly 2011) generates Bernoulli outcome variables with success probability

$$\mathbb{P}(y_{ij} = 1 | x_{ij}; \vartheta, \beta_i, \beta_j) = F(\beta_i + \beta_j + x_{ij}'\vartheta),$$

where $F(a) = (1 + e^{-a})^{-1}$ is the logistic cumulative distribution function.

4.1. *Modified profile likelihood.* The likelihood function, conditional on the regressors, is

$$\ell(\vartheta, \beta) = \sum_{i=1}^{n} \sum_{i<j} y_{ij} \log F_{ij}(\vartheta, \beta_i, \beta_j) + (1 - y_{ij}) \log (1 - F_{ij}(\vartheta, \beta_i, \beta_j)),$$

where we let $F_{ij}(\vartheta, \beta_i, \beta_j) = F(\beta_i + \beta_j + x_{ij}'\vartheta)$.

For a given value of ϑ, the score the incidental parameters has elements

$$\frac{\partial \ell(\vartheta, \beta)}{\partial \beta_i} = \sum_{i<j} y_{ij} - F_{ij}(\vartheta, \beta_i, \beta_j) + \sum_{i>j} y_{ji} - F_{ji}(\vartheta, \beta_j, \beta_i)$$

while the $n \times n$ Hessian matrix has (i, j)th-entry equal to

$$\frac{\partial^2 \ell(\vartheta, \beta)}{\partial \beta_i \partial \beta_j} = \begin{cases} -\sum_{i<j} f_{ij}(\vartheta, \beta_i, \beta_j) - \sum_{i>j} f_{ji}(\vartheta, \beta_j, \beta_i) & \text{if } i = j \\ -f_{ij}(\vartheta, \beta_i, \beta_j) & \text{if } i < j \\ -f_{ji}(\vartheta, \beta_j, \beta_i) & \text{if } i > j \end{cases},$$

for $f_{ij}(\vartheta, \beta_i, \beta_j) = F_{ij}(\vartheta, \beta_i, \beta_j)\,(1 - F_{ij}(\vartheta, \beta_i, \beta_j))$. The maximum-likelihood estimator of the β_i for a given value of ϑ is not available in closed form and needs to be computed numerically. Because the likelihood is globally concave, Newton's algorithm is well-suited for the task, and will typically find the solution in two or three iterations.

Introduce the shorthands

$$\hat{F}_{ij}(\vartheta) = F_{ij}(\vartheta, \hat{\beta}_i(\vartheta), \hat{\beta}_j(\vartheta)), \quad \hat{f}_{ij}(\vartheta) = f_{ij}(\vartheta, \hat{\beta}_i(\vartheta), \hat{\beta}_j(\vartheta)).$$

The profile likelihood is

$$\hat{\ell}(\vartheta) = \sum_{i=1}^{n} \sum_{i<j} y_{ij} \log \hat{F}_{ij}(\vartheta) + (1 - y_{ij}) \log (1 - \hat{F}_{ij}(\vartheta)),$$

and a modified likelihood is readily constructed by appropriately combining the matrices $\hat{\Sigma}(\vartheta)$ and $\hat{\Omega}(\vartheta)$.

4.2. *Simulation experiments.* We next present the results from a Monte Carlo experiment. The designs are borrowed from Graham [2015]. All designs are of the following form. Let $u_i \in \{-1, 1\}$ so that $\mathbb{P}(u_i = 1) = \frac{1}{2}$. We generate the dyad covariate as

$$x_{ij} = u_i\, u_j,$$

and the fixed effects as

$$\beta_i = \mu + \gamma_1 \frac{1 + u_i}{2} + \gamma_2 \frac{1 - u_i}{2} + u_i,$$

where $v_i \sim \text{Beta}(\lambda_1, \lambda_2)$. We set $\mu = -\lambda_1(\lambda_1 + \lambda_2)^{-1}$, so that $\mu + v_i$ has mean zero, and will consider several choices for the parameters (γ_1, γ_2) and (λ_1, λ_2). The parameter choices are summarized in Table 2. In the first four designs (A1–A4), the β_i are drawn independently of x_{ij} from symmetric Beta distributions. In the next four designs (B1–B4) the β_i are generated from skewed distributions that depend on u_i (and thus correlate with the regressor x_{ij}).

We simulate $10,000$ data sets for each design for $n \in \{25, 50, 75, 100\}$ and $\vartheta = 1$. Because the results across designs are qualitatively very similar, we present the full set of results only for Design A1 (Table 3). Tables 4 and 5

TABLE 2
Simulation designs for the β-model

Design	γ_1	γ_2	λ_1	λ_2
A1	0	0	1	1
A2	-0.25	-0.25	1	1
A3	-0.75	-0.75	1	1
A4	-1.25	-1.25	1	1
B1	0	0.50	0.25	0.75
B2	-0.50	0	0.25	0.75
B3	-1.00	-0.50	0.25	0.75
B4	-1.50	-1.00	0.25	0.75

provide and results for $n \in \{50, 100\}$ for all designs. Each table contains the mean and median bias of ϑ, $\dot{\vartheta}$, and $\ddot{\vartheta}$, along with their standard deviation and their interquartile range (both across the Monte Carlo replications). The tables also provide the empirical size of the likelihood ratio test for the null that $\vartheta = 1$ for theoretical size $\alpha \in \{.05, .10\}$. Because the results for $n = 100$ can be compared (up to Monte Carlo error) to the numerical results collected in Graham [2015, Table 2], Table 5 contains two additional columns in which we reproduce the results for his analytically bias-corrected maximum-likelihood estimator ($\tilde{\vartheta}$) and his 'tetrad logit' estimator ($\check{\vartheta}$). The latter is based on moment conditions that are free of β_i using a sufficiency argument. Bias correcting $\hat{\vartheta}$ does not salvage the likelihood ratio statistic, and the conditional likelihood function of the 'tetrad logit' estimator does not satisfy the information equality. Hence, the results on size for these two estimators are based on the Wald statistic.

Table 3 clearly shows that both the bias and standard deviation of $\hat{\vartheta}$ are $O(n^{-1})$. Consequently, the likelihood ratio test is size distorted even in large samples. Point estimation through the modified likelihoods gives estimators with small bias relative to their standard error. Even for $n = 25$, the bias is only about 20% of the bias in maximum likelihood estimator. In larger samples, the estimators are essentially unbiased. Interestingly, both $\dot{\vartheta}$ and $\ddot{\vartheta}$ are also less volatile than is $\hat{\vartheta}$. Thus, at least here, bias correction does not come at the cost of an increase in dispersion. Together with the substantial decrease in mean squared error, inference, too, improves dramatically. The likelihood ratio statistic for both $\dot{\ell}(\vartheta)$ and $\ddot{\ell}(\vartheta)$ have near theoretical size for all n.

TABLE 3
β-model. Design A1 for all n

n	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$
	mean bias			standard deviation		
25	0.1098	0.0204	0.0304	0.1897	0.1560	0.1572
50	0.0492	0.0045	0.0071	0.0717	0.0679	0.0681
75	0.0320	0.0020	0.0032	0.0467	0.0450	0.0451
100	0.0237	0.0011	0.0017	0.0341	0.0332	0.0332
	median bias			interquartile range		
25	0.1029	0.0154	0.0253	0.2069	0.1873	0.1889
50	0.0487	0.0042	0.0067	0.0961	0.0913	0.0914
75	0.0316	0.0017	0.0028	0.0630	0.0607	0.0608
100	0.0236	0.0010	0.0017	0.0464	0.0450	0.0451
	empirical size ($\alpha = .10$)			empirical size ($\alpha = .05$)		
25	0.1937	0.1134	0.1147	0.1142	0.0627	0.0637
50	0.1896	0.1128	0.1125	0.1178	0.0558	0.0555
75	0.1866	0.1092	0.1081	0.1142	0.0575	0.0569
100	0.1890	0.1042	0.1025	0.1103	0.0520	0.0513

Tables 4 and 5 show that all conclusions from Design A1 carry over to the other designs. Moreover, the introduction of correlation between regressors and heterogenous coefficients or skewing the distribution from which the latter are drawn does not prevent the modified likelihood to improve on maximum likelihood both in terms of point estimation and inference. A comparison of the two tables clearly shows that both the bias and standard deviation of $\hat{\vartheta}$ shrink by a factor of one half as n doubles, again illustrating that both are of order n^{-1}. The subsequent reduction in bias by considering $\dot{\vartheta}$ and $\ddot{\vartheta}$ and improvement in the likelihood ratio test are manifest for all designs.

Table 5 further shows that the modified-likelihood approach outperforms bias correction of the maximum-likelihood estimator in Designs A3 and B3 and, in particular, in Designs A4 and B4, where bias correction of maximum likelihood introduces additionial bias relative to $\hat{\vartheta}$. The additional bias also leads to a large deterioration of the empirical size of the Wald statistic associated with $\breve{\vartheta}$. The performance of the modified likelihood is comparable to Graham's 'tetrad logit' estimator $\breve{\vartheta}$ in terms of bias, and it tends to be somewhat more accurate in terms of the empirical size of the associated hypothesis tests.

TABLE 4
β-model. All designs for $n = 50$

Design	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$
		mean bias			standard deviation	
A1	0.0492	0.0045	0.0071	0.0717	0.0679	0.0681
A2	0.0499	0.0054	0.0079	0.0742	0.0704	0.0705
A3	0.0467	0.0033	0.0047	0.0933	0.0890	0.0891
A4	0.0497	0.0049	0.0024	0.1391	0.1335	0.1335
B1	0.0526	0.0073	0.0096	0.0768	0.0728	0.0729
B2	0.0490	0.0035	0.0059	0.0747	0.0707	0.0708
B3	0.0493	0.0046	0.0060	0.0936	0.0891	0.0891
B4	0.0500	0.0043	0.0005	0.1380	0.1320	0.1316
		median bias			interquartile range	
A1	0.0487	0.0042	0.0067	0.0961	0.0913	0.0914
A2	0.0482	0.0040	0.0064	0.0995	0.0943	0.0945
A3	0.0441	0.0008	0.0022	0.1247	0.1191	0.1191
A4	0.0412	-0.0032	-0.0059	0.1827	0.1748	0.1748
B1	0.0513	0.0061	0.0084	0.1034	0.0981	0.0982
B2	0.0479	0.0024	0.0049	0.0999	0.0948	0.0949
B3	0.0470	0.0024	0.0039	0.1252	0.1195	0.1196
B4	0.0438	-0.0018	-0.0052	0.1827	0.1740	0.1743
	empirical size ($\alpha = .10$)			empirical size ($\alpha = .05$)		
A1	0.1896	0.1128	0.1125	0.1178	0.0558	0.0555
A2	0.1857	0.1135	0.1118	0.1139	0.0602	0.0603
A3	0.1565	0.1098	0.1082	0.0878	0.0581	0.0563
A4	0.1287	0.1095	0.1083	0.0664	0.0594	0.0592
B1	0.1902	0.1141	0.1112	0.1146	0.0582	0.0579
B2	0.1801	0.1081	0.1049	0.1040	0.0574	0.0564
B3	0.1498	0.1052	0.1030	0.0830	0.0554	0.0538
B4	0.1236	0.1064	0.1067	0.0634	0.0543	0.0551

TABLE 5. *β-model. All designs for* $n = 100$

Design	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$	$\tilde{\vartheta}$	$\check{\vartheta}$	$\hat{\vartheta}$	$\dot{\vartheta}$	$\ddot{\vartheta}$	$\tilde{\vartheta}$	$\check{\vartheta}$
			mean bias					standard deviation		
A1	0.0237	0.0011	0.0017	—	—	0.0341	0.0332	0.0332	0.0340	0.0340
A2	0.0240	0.0015	0.0022	—	—	0.0352	0.0343	0.0343	0.0350	0.0350
A3	0.0215	0.0003	0.0006	—	—	0.0444	0.0434	0.0434	0.0460	0.0440
A4	0.0206	0.0001	-0.0004	—	—	0.0645	0.0633	0.0633	0.0790	0.0650
B1	0.0241	0.0013	0.0019	—	—	0.0356	0.0346	0.0346	0.0360	0.0370
B2	0.0246	0.0015	0.0021	—	—	0.0359	0.0349	0.0349	0.0340	0.0340
B3	0.0222	0.0004	0.0007	—	—	0.0450	0.0439	0.0439	0.0470	0.0460
B4	0.0231	0.0020	0.0015	—	—	0.0651	0.0639	0.0639	0.0810	0.0670
			median bias					interquartile range		
A1	0.0236	0.0010	0.0017	0.0010	-0.0010	0.0464	0.0450	0.0451	—	—
A2	0.0234	0.0010	0.0017	0.0020	0.0000	0.0471	0.0458	0.0458	—	—
A3	0.0199	-0.0013	-0.0010	0.0190	0.0010	0.0598	0.0584	0.0584	—	—
A4	0.0181	-0.0022	-0.0028	0.1060	0.0050	0.0870	0.0854	0.0855	—	—
B1	0.0240	0.0012	0.0019	0.0040	0.0010	0.0482	0.0469	0.0469	—	—
B2	0.0247	0.0017	0.0022	0.0030	0.0010	0.0484	0.0472	0.0472	—	—
B3	0.0213	-0.0007	-0.0003	0.0170	0.0000	0.0604	0.0591	0.0591	—	—
B4	0.0212	0.0002	-0.0003	0.0860	0.0020	0.0867	0.0852	0.0851	—	—
			empirical size ($\alpha = .10$)					empirical size ($\alpha = .05$)		
A1	0.1890	0.1042	0.1025	0.1020	0.0760	0.1103	0.0520	0.0513	0.0550	0.0360
A2	0.1846	0.1044	0.1026	0.0970	0.0810	0.1081	0.0533	0.0526	0.0570	0.0420
A3	0.1447	0.1051	0.1037	0.1260	0.0800	0.0805	0.0522	0.0515	0.0680	0.0380
A4	0.1149	0.1003	0.1002	0.4670	0.0810	0.0593	0.0505	0.0500	0.3520	0.0360
B1	0.1748	0.1002	0.0978	0.1050	0.0810	0.0999	0.0494	0.0488	0.0610	0.0340
B2	0.1846	0.1064	0.1050	0.0870	0.0610	0.1089	0.0533	0.0521	0.0350	0.0250
B3	0.1437	0.1050	0.1036	0.1400	0.0850	0.0812	0.0522	0.0516	0.0760	0.0380
B4	0.1174	0.1007	0.1003	0.3910	0.0900	0.0621	0.0509	0.0507	0.2960	0.0430

References.

Arellano, M. and Hahn, J. (2007). Understanding bias in nonlinear panel models: Some recent developments. In *Advances In Economics and Econometrics* (R. W. Blundell, W. K. Newey and T. Persson, eds.) **III**. Econometric Society. Cambridge University Press.

Attanasio, O., Barr, A., Cardenas, J. C., Genicot, G. and Meghir, C. (2009). Risk pooling, risk preferences, and social networks. *American Economic Journal: Applied Microeconomics* **4** 134–167.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39** 324–325.

Calvó-Armengol, A. and Jackson, M. O. (2004). The effects of social networks on employment and inequality. *American Economic Review* **94** 426–454.

Calvó-Armengol, A., Patacchini, E. and Zenou, Y. (2009). Peer effects and social networks in education. *Review of Economic Studies* **76** 1239–1267.

Chatterjee, S., Diaconis, P. and Sly, A. (2011). Random graphs with a given degree sequence. *Annals of Applied Probability* **21** 1400–1435.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B* **49** 1–39.

DiCiccio, T. J., Martin, M. A., Stern, S. E. and Young, A. (1996). Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society, Series B* **58** 189–203.

Fafchamps, M. and Gubert, F. (2007). Risk sharing and network formation. *American Economic Review* **97** 75–79.

Graham, B. S. (2015). An econometric model of link formation with degree heterogeneity. Mimeo.

Head, K. and Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of International Economics* (G. Gopinath, E. Helpman and K. Rogoff, eds.) **IV** 131–195. Elsevier.

Hunter, D. (2004). MM algorithms for generalized Bradley-Terry models. *Annals of Statistics* **32** 384–406.

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton University Press.

Jackson, M. O., Rodriguez-Barraquer, T. and Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *American Economic Review* **102** 1857–1897.

Lancaster, T. (2002). Orthogonal Parameters and Panel Data. *Review of Economic Studies* **69** 647–666.

Li, H., Lindsay, B. G. and Waterman, R. P. (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B* **65** 191–208.

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology* **27** 415–444.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

Rinaldo, A., Petrovic, S. and Fienberg, S. (2013). Maximum likelihood estimation in the β-model. *Annals of Statistics* **41** 1085–1110.

Sartori, N. (2003). Modified profile likelihood in models with stratum nuisance parameters. *Biometrika* **90** 533–549.

Severini, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika* **85** 403–411.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford University Press.

Simons, G. and Yao, Y. C. (1999). Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *Annals of Statistics* **27** 1041–1060.

Snijders, T. A. B. (2011). Statistical models for social networks. *Annual Review of Sociology* **37** 129–151.

Yan, T. and Xu, J. (2013). A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika* **100** 519–524.

Department of Economics
Sciences Po
28 rue des Saints Pères
75007 Paris
France
E-mail: koen.jochmans@sciencespo.fr