
Article:

Jochmans, Koen and Otsu, Taisuke (2019) Likelihood corrections for two-way models. Annals of Economics and Statistics, 134 (134). pp. 227-242. ISSN 1968-3863

10.15609/annaeconstat2009.134.0227

LIKELIHOOD CORRECTIONS FOR TWO-WAY MODELS

By Koen Jochmans* and Taisuke Otsu†

University of Cambridge and London School of Economics

Initial draft: January 5, 2016. This version: February 19, 2018.

The use of two-way fixed-effect models is widespread. The presence of incidental parameter bias, however, invalidates statistical inference based on the likelihood. In this paper we consider modifications to the (profile) likelihood that yield asymptotically unbiased estimators as well as likelihood-ratio and score tests with correct size. The modifications are widely applicable and easy to implement. Our examples illustrate that the modifications can lead to dramatic improvements relative to the maximum likelihood method both in terms of point estimation and inference.

1. Introduction. Two-way fixed-effect models arise in many areas of applied economics. Many models for panel data, in addition to the usual individual-specific effects, routinely include time dummies to account for aggregate time effects (see, e.g., Hahn and Moon 2006). Models for linked data sets, too, typically feature different fixed effects for each type of agent. Two well-known examples are models for employer-employee data (Abowd, Kramarz and Margolis 1999) and gravity models for trade data on import and export behaviors of firms or countries (Anderson and van Wincoop 2003; Helpman, Melitz and Rubinstein 2008).

It is known since the work of Neyman and Scott [1948] that models with fixed effects pose a serious theoretical challenge. The incidental parameter problem has received substantial attention for one-way fixed-effect models; see Arellano and Honoré [2001], Arellano and Hahn [2007], and Arellano and Bonhomme [2011] for overviews with different emphases and for references. A main conclusion of this literature is that bias correction is needed to justify inference based on maximum likelihood. This recommendation is based on

*Supported by the European Research Council through grant n° 715787.

†Supported by the European Research Council through grant n° SNP 615882.

Keywords and phrases: asymptotic bias, bias correction, fixed effects, information bias, modified profile likelihood, panel data, MCMC, penalization, rectangular-array asymptotics.

the rectangular-array asymptotics (Li, Lindsay and Waterman 2003; Sartori 2003) where both the number of strata and the number of observations per stratum grow large. Such an asymptotic approximation is suitable for data sets where none of the dimensions is negligibly small compared to the other, which are in increased supply.

In spite of their popularity in applied work, the pursuit of estimators of two-way models that enjoy sound theoretical properties has taken off only recently. Charbonneau [2017] and Jochmans [2017a;b] have invoked sufficiency arguments for binary-choice and multiplicative-error models. Such an approach is attractive as it yields estimating equations that are free of fixed effects but its applicability is inherently limited in scope. Taking the rectangular-array perspective, Fernández-Val and Weidner [2016] and Chen, Fernández-Val and Weidner [2014] have characterized the leading bias terms in the maximum likelihood estimator of quite general two-way models with additive and interactive fixed effects, respectively. These results enable bias correction of the maximum likelihood estimator by subtracting from it a plug-in estimator of the bias. Such an approach is a natural extension of the ones taken in Hahn and Newey [2004] and Dhaene and Jochmans [2015] for one-way models.

In this paper we present likelihood corrections for two-way models that lead to asymptotically valid inference under rectangular-array asymptotics. Inference by modified likelihoods has a long history in statistics; see, e.g., Barndorff-Nielsen [1983], Cox and Reid [1987], DiCiccio et al. [1996], and Severini [1998a;b]. In the econometric literature on one-way panels, their use has been advocated by Arellano and Hahn [2006; 2007]. While the resulting point estimator enjoys similar theoretical properties as the bias-corrected estimators of Fernández-Val and Weidner [2016], modifying the likelihood has several implications that may lead researchers to prefer it over correcting the bias in the maximum-likelihood estimator.

First, the correction term has a simple generic form, depending only on the score and Hessian matrix for the nuisance parameters. As such we do not need to know the precise functional form of the bias, which is model specific, and implementation does not depend on whether the nuisance parameters are scalars or vectors. Second, correcting the likelihood leads to estimators that are invariant with respect to interest-preserving reparametrizations.

Third, correcting the likelihood function not only leads to point estimators with reduced bias, but also directly improves the likelihood-ratio and score statistics. Finally, our modified likelihoods can be combined with Markov chain Monte Carlo techniques to obtain point estimators and confidence regions with attractive frequentist properties by simulation. This avoids numerical optimization and estimation of the asymptotic variance, where calculations of higher-order derivatives of the profile likelihood are required.

The rest of the paper is organized as follows. Section 2 introduces the problem at hand and derive the leading bias term in the profile likelihood. In Section 3 we use these findings to set up the likelihood corrections that lead to superior inference methods. In Section 4 we discuss examples with some numerical evidences.

2. Models with two-way fixed effects. Consider an $n \times m$ sample of independent observations $\{z_{ij} : i = 1, \ldots, n, j = 1, \ldots, m\}$. Suppose that the density of z_{ij} (relative to some dominating measure) is specified to be

$$f(z_{ij}; \theta, \alpha_i, \gamma_j).$$

The function f is known up to the finite-dimensional parameters θ and the fixed effects α_i and γ_j, which may be vectors. The goal is to perform inference on θ. The vectors $\alpha = (\alpha_1', \ldots, \alpha_n')'$ and $\gamma = (\gamma_1', \ldots, \gamma_m')'$ are nuisance parameters.

2.1. *Profile log-likelihood.* Let $\lambda = (\alpha', \gamma')'$. The log-likelihood function for all parameters is

$$\ell(\theta, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{m} \log f(z_{ij}; \theta, \alpha_i, \gamma_j).$$

The maximum likelihood estimator of θ is given by $\hat{\theta} = \arg\max_\theta \hat{\ell}(\theta)$, where $\hat{\ell}(\theta)$ is the profile log-likelihood,

$$\hat{\ell}(\theta) = \ell(\theta, \hat{\lambda}(\theta)),$$

and $\hat{\lambda}(\theta)$ is the maximum likelihood estimator of the nuisance parameters for a given θ, i.e.,

$$\hat{\lambda}(\theta) = \arg\max_{\lambda} \ell(\theta, \lambda).$$

In many cases this optimization needs to be performed under a normalization constraint on the fixed effects. For example, if the density depends on (α_i, γ_j) only through $\alpha_i+\gamma_j$, then we cannot hope to learn the mean of each effect, so we would impose, for example, $\sum_i \alpha_i = \sum_j \gamma_j$. We leave this normalization implicit for most of the paper.

It is well-known that inference based on the profile likelihood performs poorly when the dimension of the nuisance parameters is large relative to the sample size. In general, profiling out the nuisance parameters α and γ introduces bias in the profile score function, which are of order $O(n)$ and $O(m)$, respectively. The sources of these bias terms are estimation errors in $\hat{\alpha}(\theta)$ and $\hat{\gamma}(\theta)$, respectively. Under the asymptotics where m remains fixed while $n \to \infty$, the dimension of α grows with the sample size, and this leads to the incidental parameter problem as studied in the seminal work of Neyman and Scott [1948]. Under the asymptotics where both $n, m \to \infty$, the dimensions of both α and γ grow with the sample size. In this case the behavior of $\sqrt{nm}(\hat{\theta} - \theta)$ depends on the relative magnitude of n and m. Moreover, its bias is of order $O(n/m) + O(m/n)$, which diverges unless n and m grow at the same rate, and this motivates the rectangular-array asymptotics, i.e., an asymptotic embedding in which $n/m \to \rho^2$ for some $\rho \in (0, \infty)$.

Under rectangular-array asymptotics, the maximum-likelihood estimator is asymptotically biased. This implies that confidence intervals based on the asymptotic distribution are incorrectly centered, and that likelihood-ratio and score tests suffer from size distortion. In what follows we consider modifications to the profile log-likelihood that yields correct inference under the rectangular-array asymptotics.

2.2. *Information bias.* The profile log-likelihood can be seen as a plug-in version of the (infeasible) target log-likelihood

$$\ell(\theta) = \ell(\theta, \lambda(\theta)),$$

where

$$\lambda(\theta) = \arg\max_{\lambda} \mathbb{E}\left(\ell(\theta, \lambda)\right).$$

Replacing $\lambda(\theta) = (\alpha(\theta)', \gamma(\theta)')'$ with the estimator $\hat{\lambda}(\theta) = (\hat{\alpha}(\theta)', \hat{\gamma}(\theta)')'$ introduces bias. To see this, let

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

and define the covariance matrix

$$\Omega(\theta) = \mathbb{E}\left(V(\theta)V(\theta)'\right).$$

Under certain regularity conditions (see, e.g., Fernández-Val and Weidner 2016 and Chen, Fernández-Val and Weidner 2014) we have

$$\hat{\lambda}(\theta) - \lambda(\theta) = \Sigma(\theta)^{-1}V(\theta) + O_p(n^{-1} \vee m^{-1}).$$

Together with an expansion of $\hat{\ell}(\theta) = \ell(\theta, \hat{\lambda}(\theta))$ around $\hat{\lambda}(\theta) = \lambda(\theta)$, the difference between the profile log-likelihood and its target takes the form of

$$\hat{\ell}(\theta) - \ell(\theta) = \frac{1}{2}V(\theta)'\Sigma(\theta)^{-1}V(\theta) + O_p(n^{-1/2} \vee m^{-1/2}).$$

Therefore,

$$\beta(\theta) = \mathbb{E}\left(\hat{\ell}(\theta) - \ell(\theta)\right) = \frac{1}{2}\,\mathrm{trace}\left(\Sigma(\theta)^{-1}\Omega(\theta)\right) + O(n^{-1/2} \vee m^{-1/2}).$$

Here, the leading bias term arises from the estimation noise in the fixed effects. Typically, it will be of order

$$\beta(\theta) = O(n) + O(m),$$

where the first term arises from the estimation noise in $\hat{\alpha}(\theta)$ and the second term stems from imprecision in $\hat{\gamma}(\theta)$. Under the usual regularity conditions that allow for integration and differentiation to be interchanged, the above bias in the profile log-likelihood function implies that the bias in the score equation takes the form $\beta'(\theta)$, which leads to the asymptotic bias in the maximum likelihood estimator.

3. Modified log-likelihood. A plug-in estimator of the bias term $\beta(\theta)$ based on the maximum likelihood estimator is

$$\check{\beta}(\theta) = \frac{1}{2}\,\mathrm{trace}\left(\hat{\Sigma}(\theta)^{-1}\hat{\Omega}(\theta)\right)$$

where the matrices $\hat{\Sigma}(\theta)$ and $\hat{\Omega}(\theta)$ are sample counterparts to $\Sigma(\theta)$ and $\Omega(\theta)$, respectively, obtained by using the plug-in estimator $\hat{\lambda}(\theta)$. Subtracting this estimator of the bias term from the profile log-likelihood yields the modified profile log-likelihood function

$$\check{\ell}(\theta) = \hat{\ell}(\theta) - \check{\beta}(\theta),$$

which yields a superior approximation to the target likelihood $\ell(\theta)$ as $n, m \to \infty$ with $n/m \to \rho^2$.

3.1. *Asymptotically-unbiased estimation.* Now we consider the maximum modified likelihood estimator

$$\check{\theta} = \arg\max_{\theta} \check{\ell}(\theta).$$

Let

$$I_\theta = -\,\mathbb{E}\left(\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\right),$$

be the Fisher information. Under standard regularity conditions we obtain

$$\check{\theta} - \theta \overset{a}{\sim} N(0, I_\theta^{-1}), \tag{3.1}$$

as $n, m \to \infty$ so that $n/m \to \rho^2$. This conclusion is to be contrasted with the corresponding result for the maximum likelihood estimator, which reads

$$\hat{\theta} - \theta \overset{a}{\sim} N(I_\theta^{-1}\beta'(\theta), I_\theta^{-1}),$$

as $n, m \to \infty$ so that $n/m \to \rho^2$.

The distributional result in (3.1) permits valid inference based on the Wald principle. However, given the lack of invariance of the Wald statistic to formulation of the null hypothesis, we may equally consider the likelihood-ratio statistic. For testing $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$, for example, the modified likelihood-ratio statistic is

$$-2\left(\check{\ell}(\theta_0) - \check{\ell}(\check{\theta})\right).$$

By virtue of the correction term $\check{\beta}(\theta)$, under the null, this statistic will be well-approximated by a χ^2 random variable. This modified likelihood-ratio statistic has usual benefits. It is invariant to how the null is formulated and

does not require a plug-in estimator of the information matrix. Likewise, the correction term implies

$$\frac{\partial \check{\ell}(\theta)}{\partial \theta} = \frac{\partial \hat{\ell}(\theta)}{\partial \theta} - \beta'(\theta),$$

which constitutes an improved approximation to the infeasible score $\partial \ell(\theta)/\partial \theta$. Hence, letting $\dot{\theta}$ denote the constrained maximizer of $\check{\ell}(\theta)$ under the null, and writing $\dot{I}_\theta$ for an estimator of the information under the null, the score statistic

$$\left(\frac{\partial \check{\ell}(\theta)}{\partial \theta}\bigg|_{\theta=\dot{\theta}} \right)' \dot{I}_\theta^{-1} \left(\frac{\partial \check{\ell}(\theta)}{\partial \theta}\bigg|_{\theta=\dot{\theta}} \right),$$

leads to size-correct inference in large samples.

3.2. *Local correction term.* The estimator $\check{\beta}(\theta)$ of the bias term does not use the likelihood structure. As such, it is equally applicable to quasi-likelihood or more general M-estimation problems. In the likelihood setting, under correct specification, we can use the fact that the information equality holds at the true parameter value to construct the alternative correction term

$$\tilde{\beta}(\theta) = -\frac{1}{2} \log \det \hat{\Sigma}(\theta) + \frac{1}{2} \log \det \hat{\Omega}(\theta),$$

and corresponding modified log-likelihood

$$\tilde{\ell}(\theta) = \hat{\ell}(\theta) - \tilde{\beta}(\theta).$$

The derivation of $\tilde{\beta}(\theta)$ from $\check{\beta}(\theta)$ follows in a similar fashion as discussed in Arellano and Hahn [2006]. The function $\tilde{\beta}(\theta)$ can be understood as an extension of DiCiccio et al. [1996] to two-way models. Following Pace and Salvan [2006], it can also be seen as a generalization of the approximate conditional log-likelihood developed by Cox and Reid [1987] which, in our context, would be

$$\hat{\ell}(\theta) + \frac{1}{2} \log \det \hat{\Sigma}(\theta),$$

to situations where θ and λ need not be information orthogonal.

3.3. *Estimation and inference via MCMC.* Numerical optimization of $\check{\ell}(\theta)$ (and $\tilde{\ell}(\theta)$) and estimation of the information I_θ may prove to be quite cumbersome in complicated models. Fortunately, we may resort to the use

of conventional Markov chain Monte Carlo methods and draw from the 'posterior'

$$\check{p}(\theta) \propto e^{\check{\ell}(\theta)}.$$

By the argument of Chernozhukov and Hong [2003], draws $\{\theta_*\}$ from the above posterior will be approximately

$$\theta_* \overset{a}{\sim} N(\overline{\theta}, I_\theta^{-1}),$$

where the mean parameter $\overline{\theta}$ is a consistent estimator of θ. Thus, aside from the mean, median and mode of the posterior being point estimators with bias-reduced properties, the variance of the posterior draws is a valid point estimator for the information. Furthermore, valid (frequentist) confidence sets can be constructed directly from the posterior. Therefore, if desired, both numerical optimization of the modified likelihood and direct estimation of the information can be avoided.

4. Examples. We now set up the modified log-likelihood function for some specific problems and provide simulation evidences.

4.1. *Linear model.* Our first example is a simple extension of the classic Neyman and Scott [1948] problem, where outcomes are generated as

$$z_{ij} \sim N(\alpha_i + \gamma_j, \theta).$$

The likelihood is

$$\ell(\theta, \lambda) = -\frac{nm}{2}\log\theta - \frac{\sum_{i=1}^n \sum_{j=1}^m (z_{ij} - \alpha_i - \gamma_j)^2}{2\theta}.$$

This model is overparametrized because adding a constant to all α_i and subtracting the same constant from all γ_j leaves the likelihood unchanged. We thus normalize the fixed effects by setting $\alpha_1 = 0$. So, the dimension of the nuisance parameters is $n + m - 1$.

A calculation shows that

$$\hat{\theta} = \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m ((z_{ij} - \overline{z}) - (\overline{z}_i - \overline{z}) - (\overline{z}_j - \overline{z}))^2$$

for $\overline{z}_i = m^{-1}\sum_{j=1}^m z_{ij}$, $\overline{z}_j = n^{-1}\sum_{i=1}^n z_{ij}$, and $\overline{z} = (nm)^{-1}\sum_{i=1}^n\sum_{j=1}^m z_{ij}$. In large samples,

$$\hat{\theta} - \theta \overset{a}{\sim} N\left(-\frac{\theta}{n} - \frac{\theta}{m} + \frac{\theta}{nm}, \frac{2\theta^2}{nm}\right). \tag{4.1}$$

Thus, here, the maximum likelihood estimator underestimates the variance on average.

Note that the log-likelihood is symmetric in the nuisance parameters. Moreover,

$$\frac{\partial \log f(z_{ij};\theta,\alpha_i,\gamma_j)}{\partial\alpha_i} = \frac{\partial \log f(z_{ij};\theta,\alpha_i,\gamma_j)}{\partial\gamma_j} = \frac{z_{ij}-\alpha_i-\gamma_j}{\theta} = \frac{\varepsilon_{ij}(\alpha_i,\gamma_j)}{\theta},$$

say. The plug-in estimator of $\varepsilon_{ij}(\alpha_i,\gamma_j)$ based on the maximum-likelihood estimator is

$$\hat{\varepsilon}_{ij} = (z_{ij}-\overline{z}) - (\overline{z}_i-\overline{z}) - (\overline{z}_j-\overline{z}),$$

which is independent of θ. Thus, if we partition the $(n+m-1)\times(n+m-1)$ covariance matrix of the score vector as

$$\hat{\Omega}(\theta) = \frac{1}{\theta^2}\begin{pmatrix} \hat{\Omega}_{\alpha\alpha} & \hat{\Omega}_{\alpha\gamma} \\ \hat{\Omega}_{\gamma\alpha} & \hat{\Omega}_{\gamma\gamma} \end{pmatrix},$$

then we have

$$(\hat{\Omega}_{\alpha\alpha})_{i,i'} = \begin{cases} \sum_j \hat{\varepsilon}^2_{(i+1)j} & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}, \qquad (\hat{\Omega}_{\gamma\gamma})_{j,j'} = \begin{cases} \sum_i \hat{\varepsilon}^2_{ij} & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases},$$

and

$$(\hat{\Omega}_{\alpha\gamma})_{i,j} = (\hat{\Omega}_{\alpha\gamma})_{j,i} = \hat{\varepsilon}^2_{(i+1)j},$$

where i ranges over $1,\ldots,n-1$ and j over $1,\ldots,m$. Also we have

$$\frac{\partial^2 \log f(z_{ij};\theta,\alpha_i,\gamma_j)}{\partial\alpha_i^2} = \frac{\partial^2 \log f(z_{ij};\theta,\alpha_i,\gamma_j)}{\partial\gamma_j^2} = \frac{\partial^2 \log f(z_{ij};\theta,\alpha_i,\gamma_j)}{\partial\alpha_i\partial\gamma_j} = -\frac{1}{\theta}.$$

It follows that the information matrix for the nuisance parameters does not depend on λ. Its plug-in estimator is

$$\hat{\Sigma}(\theta) = \frac{1}{\theta}\begin{pmatrix} m\,I_{n-1} & \iota_{n-1}\,\iota_m' \\ \iota_m\,\iota_{n-1}' & n\,I_m \end{pmatrix},$$

where I_n is the $n \times n$ identity matrix and ι_n denotes an n-vector of ones. By standard formulae for partitioned matrix inversion, it holds

$$\hat{\Sigma}(\theta)^{-1} = \theta \begin{pmatrix} m^{-1} I_{n-1} & 0 \\ 0 & n^{-1} I_m \end{pmatrix} + \frac{\theta}{m} \begin{pmatrix} \iota_{n-1}\iota'_{n-1} & -\iota_{n-1}\iota'_m \\ -\iota_m \iota'_{n-1} & \frac{n-1}{n}\iota_m \iota'_m \end{pmatrix}.$$

A small calculation then yields

$$\check{\beta}(\theta) = \frac{1}{2}\text{trace}(\hat{\Sigma}(\theta)^{-1}\hat{\Omega}(\theta)) = \frac{1}{2\theta}\frac{\sum_{i=1}^n \sum_{j=1}^m \hat{\varepsilon}_{ij}^2}{m} + \frac{1}{2\theta}\frac{\sum_{i=1}^n \sum_{j=1}^m \hat{\varepsilon}_{ij}^2}{n},$$

which is of order $O(n) + O(m)$. The modified log-likelihood has the simple form

$$\check{\ell}(\theta) = -\frac{nm}{2}\log\theta - \frac{nm+n+m}{nm}\frac{\sum_{i=1}^n \sum_{j=1}^m ((z_{ij}-\overline{z})-(\overline{z}_i-\overline{z})-(\overline{z}_j-\overline{z}))^2}{2\theta}.$$

The intuition of the modification in this example follows from the usual degrees-of-freedom argument. Moreover,

$$\check{\theta} = \frac{nm+n+m}{nm}\hat{\theta} = \hat{\theta} + \frac{\hat{\theta}}{n} + \frac{\hat{\theta}}{m},$$

which, together with (4.1), shows that the modified log-likelihood removes the leading bias from $\hat{\theta}$. In this example, the estimator obtained coincides with the bias-corrected maximum likelihood estimator.

Alternatively, a calculation shows that the local correction term that uses the likelihood setting, up to a constant, equals

$$\tilde{\beta}(\theta) = -\frac{n+m-1}{2}\log\theta,$$

Hence, an alternative modified log-likelihood here is

$$\tilde{\ell}(\theta) = -\frac{(n-1)(m-1)}{2}\log\theta - \frac{\sum_{i=1}^n \sum_{j=1}^m ((z_{ij}-\overline{z})-(\overline{z}_i-\overline{z})-(\overline{z}_j-\overline{z}))^2}{2\theta}.$$

Its maximizer is

$$\tilde{\theta} = \frac{nm}{(n-1)(m-1)}\hat{\theta} = \frac{1}{(n-1)(m-1)}\sum_{i=1}^n\sum_{j=1}^m ((z_{ij}-\overline{z})-(\overline{z}_i-\overline{z})-(\overline{z}_j-\overline{z}))^2,$$

which is exactly unbiased.

To further illustrate we present simulation results for the Neyman-Scott problem in Table 1. We fix $\theta = 1$ and present results for $n = m = 10$ and

$n = m = 20$, which suffice to make our point for this model. All results are obtained over $10,000$ Monte Carlo replications and are invariant to the distributions of the α_i's and γ_j's.

Table 1 provides the bias and standard deviation (obtained over the Monte Carlo replications) of the maximum-likelihood estimator $\hat{\theta}$ and of the modified-likelihood estimators $\check{\theta}$ and $\tilde{\theta}$. The table also contains the same statistics for the mean of the respective posteriors computed via MCMC, $\check{\theta}_*$ and $\tilde{\theta}_*$. Additionally we report (the average of) the standard error for each estimator, as well as the ratio of the standard error to the standard deviation. For maximum likelihood, the standard error is estimated by the plug-in estimator $\sqrt{2}\hat{\theta}/\sqrt{nm}$. The standard errors for $\check{\theta}$ and $\tilde{\theta}$ are obtained similarly. For $\check{\theta}_*$ and $\tilde{\theta}_*$, the standard errors are obtained as the standard deviation of the respective Markov chains. Finally, the table also reports the empirical size of two-sided tests for the null hypothesis that $\theta = 1$ with theoretical size equal to $\tau = .01, .05, .10$. We consider the Wald statistic for all estimators, the likelihood-ratio statistic, and (Bayesian) credible intervals based on the posterior quantiles.

The results show that the bias in the maximum likelihood estimator is of the same order as its standard deviation. Consequently, both the Wald and likelihood-ratio statistic are heavily size distorted. This is so for all the significance levels and for all the sample sizes considered. The bias is clearly seen to be $O(n^{-1})+O(m^{-1})$. All the modified estimators have much less bias. Moreover, the numerical results confirm our calculation that $\tilde{\theta}$ is unbiased. Further, the bias is consistently small relative to the standard deviation. As a result, the performance of all test statistics improves dramatically relative to maximum likelihood.

4.2. *Factor model.* Our second illustration is a stripped-down version of the model in Bai [2009]. Here,

$$z_{ij} \sim N(\alpha_i \gamma_j, \theta).$$

This differs from the classic Neyman and Scott [1948] example in that, now, the fixed effects enter in a multiplicative manner as opposed to additive. This is a non-trivial complication. The model can be interpreted as a factor model with heterogeneous factor loadings.

TABLE 1
Simulation results for the Neyman-Scott problem

	$n = m = 10$				
	$\hat{\theta}$	$\check{\theta}$	$\check{\theta}_*$	$\tilde{\theta}$	$\tilde{\theta}_*$
bias	-0.189	-0.027	0.017	0.001	0.068
std. dev.	0.128	0.153	0.160	0.158	0.167
std. err.	0.115	0.138	0.148	0.142	0.174
ratio	0.897	0.897	0.925	0.897	1.039
Wald					
0.01	0.281	0.049	0.030	0.037	0.009
0.05	0.431	0.107	0.072	0.086	0.041
0.10	0.524	0.166	0.128	0.143	0.086
LR					
0.01	0.148	0.025	—	0.010	—
0.05	0.322	0.083	—	0.052	—
0.10	0.440	0.152	—	0.104	—
Bayes					
0.01	—	—	0.028	—	0.017
0.05	—	—	0.084	—	0.063
0.10	—	—	0.142	—	0.115
	$n = m = 20$				
	$\hat{\theta}$	$\check{\theta}$	$\check{\theta}_*$	$\tilde{\theta}$	$\tilde{\theta}_*$
bias	-0.098	-0.008	0.003	0.000	0.014
std. dev.	0.067	0.074	0.075	0.075	0.076
std. err.	0.064	0.070	0.071	0.071	0.075
ratio	0.947	0.947	0.942	0.947	0.990
Wald					
0.01	0.202	0.021	0.021	0.018	0.013
0.05	0.372	0.072	0.071	0.066	0.055
0.10	0.473	0.126	0.125	0.119	0.103
LR					
0.01	0.137	0.015	—	0.010	—
0.05	0.312	0.065	—	0.050	—
0.10	0.428	0.121	—	0.099	—
Bayes					
0.01	—	—	0.027	—	0.022
0.05	—	—	0.076	—	0.063
0.10	—	—	0.132	—	0.113

The likelihood function is

$$\ell(\theta, \lambda) = -\frac{nm}{2} \log \theta - \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (z_{ij} - \alpha_i \gamma_j)^2}{2\theta}.$$

The scale of the effects is not identified. One possible normalization is to set $\sum_i \alpha_i^2 = \sum_j \gamma_j^2$.

The score vector for has entries

$$\frac{\partial \ell(\theta, \lambda)}{\partial \alpha_i} = \sum_{j=1}^{m} \frac{(z_{ij} - \alpha_i \gamma_j)\, \gamma_j}{\theta}, \qquad i = 1, \ldots, n,$$

$$\frac{\partial \ell(\theta, \lambda)}{\partial \gamma_j} = \sum_{i=1}^{n} \frac{(z_{ij} - \alpha_i \gamma_j)\, \alpha_i}{\theta}, \qquad j = 1, \ldots, m.$$

The estimator $\hat{\lambda}(\theta)$ does not depend on θ and can be found by iterating on the first-order conditions for α and γ. Given the estimators $\hat{\alpha}_i, \hat{\gamma}_j$ we find the estimator of θ to be

$$\hat{\theta} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} (z_{ij} - \hat{\alpha}_i \hat{\gamma}_j)^2.$$

The plug-in estimator of the $(n + m) \times (n + m)$ covariance matrix of the score for the incidental parameters is

$$\hat{\Omega}(\theta) = \frac{1}{\theta^2} \begin{pmatrix} \hat{\Omega}_{\alpha\alpha} & \hat{\Omega}_{\alpha\gamma} \\ \hat{\Omega}_{\gamma\alpha} & \hat{\Omega}_{\gamma\gamma} \end{pmatrix},$$

for $n \times n$ and $m \times m$ diagonal matrices $\hat{\Omega}_{\alpha\alpha}(\theta)$ and $\hat{\Omega}_{\gamma\gamma}(\theta)$ whose entries are

$$(\hat{\Omega}_{\alpha\alpha})_{i,i'} = \begin{cases} \sum_{j=1}^{m} (z_{ij} - \hat{\alpha}_i \hat{\gamma}_j)^2 \hat{\gamma}_j^2 & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}$$

and

$$(\hat{\Omega}_{\gamma\gamma})_{j,j'} = \begin{cases} \sum_{i=1}^{n} (z_{ij} - \hat{\alpha}_i \hat{\gamma}_j)^2 \hat{\alpha}_i^2 & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases},$$

respectively, and $n \times m$ and $m \times n$ submatrices $\hat{\Omega}_{\alpha\gamma}$ and $\hat{\Omega}_{\gamma\alpha}$ whose entries are

$$(\hat{\Omega}_{\alpha\gamma})_{i,j} = (\hat{\Omega}_{\gamma\alpha})_{j,i} = (z_{ij} - \hat{\alpha}_i \hat{\gamma}_j)^2\, \hat{\alpha}_i \hat{\gamma}_j.$$

The Hessian matrix is now estimated by

$$\hat{\Sigma}(\theta) = \frac{1}{\theta} \begin{pmatrix} \hat{\Sigma}_{\alpha\alpha} & \hat{\Sigma}_{\alpha\gamma} \\ \hat{\Sigma}_{\gamma\alpha} & \hat{\Sigma}_{\gamma\gamma} \end{pmatrix},$$

where, with $\hat{s} = \sum_{i=1}^{n} \hat{\alpha}_i^2 = \sum_{j=1}^{m} \hat{\gamma}_j^2$, we have

$$\Sigma_{\alpha\alpha} = \hat{s}\, I_n, \qquad \Sigma_{\gamma\gamma} = \hat{s}\, I_m, \qquad (\Sigma_{\alpha\gamma})_{i,j} = \hat{\alpha}_i\hat{\gamma}_j - (z_{ij} - \hat{\alpha}_i\hat{\gamma}_j) = (\Sigma_{\gamma\alpha})_{j,i}.$$

Combining these expressions lead to the bias estimator $\check{\beta}(\theta)$, which we omit here for brevity. Note that here, again, the local correction term is very simple and equals

$$\tilde{\beta}(\theta) = -\frac{n+m}{2}\log\theta,$$

up to a constant.

Table 2, which has the same layout as Table 1, provides numerical results for the factor model. The conclusions are essentially the same as those drawn in the previous subsection. Inference based on maximum likelihood performs poorly. The modified likelihoods provide estimators with negligible bias and test statistics with good size properties.

4.3. *Binary-choice model.* Our third example is a regression model for a binary outcome y_{ij}. Here, $z_{ij} = (y_{ij}, x_{ij}')'$ and we condition on x_{ij}; so, $f(z_{ij};\theta,\alpha_i,\gamma_j) = f(y_{ij}|x_{ij};\theta,\alpha_i,\gamma_j)$ is the probability mass function of a Bernoulli random variable. A logistic version has

$$P(y_{ij}=1|x_{ij},\alpha_i,\gamma_j) = \frac{1}{1+e^{-(\alpha_i+\gamma_j+x_{ij}'\theta)}} = \mu_{ij}(\theta;\alpha_i,\gamma_j) \text{ (say)}.$$

The mean of the fixed effects is again not identified, and so we normalize $\alpha_1 = 0$.

Let

$$\varepsilon_{ij}(\theta,\alpha_i,\gamma_j) = y_{ij} - \mu_{ij}(\theta;\alpha_i,\gamma_j),$$

and write its maximum likelihood estimator (which is not available in closed form) as

$$\hat{\varepsilon}_{ij}(\theta) = \varepsilon_{ij}(\theta,\hat{\alpha}_i(\theta),\hat{\gamma}_j(\theta)).$$

Then

$$\frac{\partial\ell(\theta,\lambda)}{\partial\alpha_{i-1}} = \sum_j \varepsilon_{ij}(\theta,\alpha_i,\gamma_j), \qquad \frac{\partial\ell(\theta,\lambda)}{\partial\gamma_j} = \sum_i \varepsilon_{ij}(\theta,\alpha_i,\gamma_j),$$

where i ranges over $2,\ldots,n$ and j ranges over $1,\ldots,m$. The components of the matrix

$$\hat{\Omega}(\theta) = \begin{pmatrix} \hat{\Omega}_{\alpha\alpha}(\theta) & \hat{\Omega}_{\alpha\gamma}(\theta) \\ \hat{\Omega}_{\gamma\alpha}(\theta) & \hat{\Omega}_{\gamma\gamma}(\theta) \end{pmatrix},$$

TABLE 2
Simulation results for the Bai problem

	$n = m = 10$				
	$\hat{\theta}$	$\check{\theta}$	$\check{\theta}_*$	$\tilde{\theta}$	$\tilde{\theta}_*$
bias	-0.187	-0.020	0.033	0.017	0.086
std. dev.	0.127	0.157	0.165	0.159	0.169
std. err.	0.115	0.139	0.150	0.144	0.178
ratio	0.907	0.885	0.913	0.907	1.056
Wald					
0.01	0.272	0.046	0.024	0.026	0.007
0.05	0.422	0.105	0.071	0.077	0.036
0.10	0.517	0.162	0.130	0.133	0.086
LR					
0.01	0.139	0.024	—	0.011	—
0.05	0.311	0.087	—	0.048	—
0.10	0.430	0.149	—	0.100	—
Bayes					
0.01	—	—	0.030	—	0.019
0.05	—	—	0.086	—	0.068
0.10	—	—	0.145	—	0.120
	$n = m = 20$				
	$\hat{\theta}$	$\check{\theta}$	$\check{\theta}_*$	$\tilde{\theta}$	$\tilde{\theta}_*$
bias	-0.098	-0.006	0.007	0.003	0.017
std. dev.	0.067	0.074	0.075	0.074	0.076
std. err.	0.064	0.070	0.071	0.071	0.076
ratio	0.959	0.953	0.945	0.959	1.004
Wald					
0.01	0.199	0.023	0.022	0.019	0.012
0.05	0.361	0.071	0.068	0.064	0.054
0.10	0.471	0.120	0.122	0.114	0.102
LR					
0.01	0.131	0.015	—	0.010	—
0.05	0.303	0.067	—	0.053	—
0.10	0.426	0.119	—	0.100	—
Bayes					
0.01	—	—	0.029	—	0.022
0.05	—	—	0.075	—	0.065
0.10	—	—	0.128	—	0.114

are of the form

$$(\hat{\Omega}_{\alpha\alpha}(\theta))_{i,i'} = \begin{cases} \sum_{j=1}^{m} \hat{\varepsilon}^2_{(i+1)j}(\theta) & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases},$$
$$(\hat{\Omega}_{\gamma\gamma}(\theta))_{j,j'} = \begin{cases} \sum_{i=1}^{n} \hat{\varepsilon}^2_{ij}(\theta) & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases},$$

and

$$(\hat{\Omega}_{\alpha\gamma}(\theta))_{i,j} = (\hat{\Omega}_{\gamma\alpha}(\theta))_{j,i} = \hat{\varepsilon}^2_{(i+1)j}(\theta).$$

To state the plug-in estimator of the information matrix, let

$$\sigma_{ij}(\theta, \alpha_i, \gamma_i) = \mu_{ij}(\theta, \alpha_i, \gamma_i)\big(1 - \mu_{ij}(\theta, \alpha_i, \gamma_i)\big),$$

which is the logistic density function at observation z_{ij} for given parameter values, and let $\hat{\sigma}_{ij}(\theta) = \sigma_{ij}(\theta, \hat{\alpha}_i(\theta), \hat{\gamma}_i(\theta))$. Then

$$\hat{\Sigma}(\theta) = \begin{pmatrix} \hat{\Sigma}_{\alpha\alpha}(\theta) & \hat{\Sigma}_{\alpha\gamma}(\theta) \\ \hat{\Sigma}_{\gamma\alpha}(\theta) & \hat{\Sigma}_{\gamma\gamma}(\theta) \end{pmatrix},$$

where

$$[\hat{\Sigma}_{\alpha\alpha}(\theta)]_{i,i'} = \begin{cases} \sum_{j=1}^{m} \hat{\sigma}_{(i+1)j}(\theta) & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}$$

and

$$[\hat{\Sigma}_{\gamma\gamma}(\theta)]_{j,j'} = \begin{cases} \sum_{i=1}^{n} \hat{\sigma}_{ij}(\theta) & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases}$$

are $(n-1) \times (n-1)$ and $m \times m$ diagonal matrices of order m and n, respectively, and the $(n-1) \times m$ submatrices $\hat{\Sigma}_{\alpha\gamma}(\theta)$ and $\hat{\Sigma}_{\gamma\alpha}(\theta)$ have entries

$$(\hat{\Sigma}_{\alpha\gamma}(\theta))_{i,j} = (\hat{\Sigma}_{\gamma\alpha}(\theta))_{j,i} = \hat{\sigma}_{(i+1)j}(\theta),$$

each of which is of order one.

Simulation results for designs where $\theta = 1$, x_{ij} is univariate logistic, and all fixed effects are set to zero are reported in Table 3. We provide results, based on $1,000$ replications, for samples of size $n = m = 20$ and $n = m = 40$. For the estimators $\hat{\theta}$, $\breve{\theta}$, and $\tilde{\theta}$, the standard error is estimated as the inverse of the empirical information. Experimentation with the outer product of the score vector gave very similar results. In contrast, inference based on the quasi-Bayesian estimators does not require an expression for the asymptotic variance. Here, again, the bias in $\hat{\theta}$ is clearly visible and the associated

TABLE 3
Simulation results for the Holland and Leinhardt problem

	$n = m = 20$				
	$\hat{\theta}$	$\check{\theta}$	$\check{\theta}_*$	$\tilde{\theta}$	$\tilde{\theta}_*$
bias	0.151	0.031	0.014	0.046	0.029
std. dev.	0.142	0.123	0.118	0.126	0.120
std. err.	0.129	0.118	0.107	0.119	0.111
ratio	0.903	0.953	0.906	0.945	0.920
Wald					
0.01	0.057	0.009	0.023	0.012	0.013
0.05	0.188	0.060	0.080	0.062	0.080
0.10	0.300	0.113	0.134	0.123	0.132
LR					
0.01	0.085	0.020	—	0.025	—
0.05	0.228	0.077	—	0.079	—
0.10	0.333	0.138	—	0.154	—
Bayes					
0.01	—	—	0.037	—	0.036
0.05	—	—	0.085	—	0.091
0.10	—	—	0.137	—	0.137
	$n = m = 40$				
	$\hat{\theta}$	$\check{\theta}$	$\check{\theta}_*$	$\tilde{\theta}$	$\tilde{\theta}_*$
bias	0.065	0.006	0.003	0.010	0.007
std. dev.	0.062	0.058	0.058	0.058	0.059
std. err.	0.058	0.056	0.055	0.056	0.055
ratio	0.938	0.969	0.957	0.967	0.938
Wald					
0.01	0.071	0.013	0.022	0.014	0.032
0.05	0.192	0.052	0.078	0.053	0.084
0.10	0.282	0.109	0.132	0.108	0.149
LR					
0.01	0.080	0.016	—	0.015	—
0.05	0.210	0.061	—	0.061	—
0.10	0.299	0.120	—	0.114	—
Bayes					
0.01	—	—	0.056	—	0.061
0.05	—	—	0.095	—	0.100
0.10	—	—	0.150	—	0.152

test statistics substantially overreject. Basing inference on the modified likelihood largely removes the bias and takes care of the overrejection problem in the test statistics.

The argument here does not depend on the logistic distribution. Other link functions, such as a probit or a log-log are equally admissible. Indeed, more generally, generic nonlinear regression models are amenable to our approach, and there is no reason to presume that the performance of the correction would not be as good as it is found to perform here.

References.

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica* **67** 251–333.

Anderson, J. E. and van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review* **93** 170–192.

Arellano, M. and Bonhomme, S. (2011). Nonlinear panel data analysis. *Annual Review of Economics* **3** 395–424.

Arellano, M. and Hahn, J. (2006). A Likelihood-Based Approximate Solution to the Incidental Parameter Problem in Dynamic Nonlinear Models with Multiple Effects. Unpublished manuscript.

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.

Arellano, M. and Honoré, B. E. (2001). Panel Data Models: Some Recent Developments. In *Handbook of Econometrics*, (J. J. Heckman and E. Leamer, eds.) **5** 53 3229–3329. Elsevier.

Bai, J. (2009). Panel data models with interactive effects. *Econometrica* **77** 1229–1279.

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

Charbonneau, K. B. (2017). Multiple fixed effects in binary response panel data models. *Econometrics Journal* **20** S1–S13.

Chen, M., Fernández-Val, I. and Weidner, M. (2014). Nonlinear panel models with interactive effects. Mimeo.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics* **115** 293–346.

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

Dhaene, G. and Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies* **82** 991–1030.

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.

Fernández-Val, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics* **196** 291–312.

Hahn, J. and Moon, H. R. (2006). Reducing bias of MLE in a dynamic panel model. *Econometric Theory* **22** 499–512.

Hahn, J. and Newey, W. K. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* **72** 1295–1319.

Helpman, E., Melitz, M. and Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *Quarterly Journal of Economics* **123** 441–487.

Jochmans, K. (2017a). Semiparametric analysis of network formation. Forthcoming in *Journal of Business and Economic Statistics*.

Jochmans, K. (2017b). Two-way models for gravity. *Review of Economics and Statistics* **99** 478–485.

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.

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

Pace, L. and Salvan, A. (2006). Adjustments of profile likelihood from a new perspective. *Journal of Statistical Planning and Inference* **136** 3554–3564.

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

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

Severini, T. A. (1998b). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85** 507–522.

Faculty of Economics
University of Cambridge
Sidgwick Avenue
Cambridge CB3 9DD
United Kingdom
E-mail: kj345@cam.ac.uk

Department of Economics
London School of Economics
Houghton Street
London WC2A 2AE
United Kingdom
E-mail: t.otsu@lse.ac.uk