UiO : Department of Mathematics
University of Oslo

Likelihood corrections for the incidental parameter problem in Poisson distributed panel data

Jonas Øren

Master's Thesis, Spring 2021

This master's thesis is submitted under the master's programme *Stochastic Modelling, Statistics and Risk Analysis*, with programme option *Statistics*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Acknowledgements

First I would like to thank my supervisor Riccardo De Bin for introducing me to this exciting topic and for providing guidance and feedback on my progress and on my writing. With your help, writing this thesis has been both enjoyable and rewarding.

I would also like to thank my fellow students in study hall B801. Without the many lunch breaks and chats the two years leading up to this thesis would not have been the same.

Lastly I would like to thank Anna, Oliver and Ingrid for proof reading the thesis before delivery, and I would like to give a special thanks to Anna for keeping me fed and motivated in times of need.

Innhold

Figurer

Tabeller

KAPITTEL 1

Introduction

In statistical analysis of data we are interested in understanding or estimating certain aspects of the probability distribution generating the data. The aspects we are interested in does not always relate to all the parameters of the probability distribution. In this case the parameters that are not of primary interest is termed «nuisance parameters» and need to be dealt with in order to gain insight on the parameters of interest.

We will in this thesis focus on a special case of this problem where the nuisance parameters are especially problematic. The problem we are interested in is when the number of nuisance parameters increases with the data, such that increasing the number of observations does not improve estimates of the nuisance parameters. The problem was introduced by the statistician Jerzy Neyman and his student Elizabeth Scott in their article from 1948 (Neyman og Scott, 1948). Neyman and Scott called it the «incidental parameter problem».

Briefly explained the incidental parameter problem occurs when parameters that are not of interest increase in number with the data at a rate large enough to disturb the maximum likelihood estimates of the parameters that are of interest, in the sense that the maximum likelihood estimators for the parameters of interest will no longer be consistent.

The impact of the paper was not as large as perhaps it should have been. In his survey of the status of the problem since the release of the Neyman and Scott paper in 1948, Tony Lancaster (Lancaster, 2000) noted on how little attention it had received in economics despite its prevalence in many economic applications, including a problem that is of primary focus for this thesis, namely fixed effects in panel data.

Panel data can be used to model many phenomena of interest in economics as given in the following examples given by Karyne B. Charbonneau in her paper «Multiple Fixed Effects in Nonlinear Panel Data Models» (Charbonneau, 2012). Abowd, Kramarz og Margolis (1999) studied wage determinants using matched firm-employee data with fixed effects for both firms and workers in an influential paper. In a similar fashion papers by Aaronson, Barrow og Sander (2007) and by Rivkin, Hanushek og Kain (2005) studied acedemic achievement using matched data between students and teachers. One can also, as is the main motivation for the paper by Charbonneau (2012), apply the fixed effects gravity equations model to estimate factors influencing international trade, such as distance between countries, historical connections and diplomatic relations. In this case the Poisson distribution is commonly used to model quantity of goods traded.

There are several proposed solutions to the incidental parameter problem. In this thesis we will compare an approximate conditional likelihood derived by Charbonneau (2012) and four modifications to the profile likelihood presented by Pace og Salvan (2006), when applied to a model of Poisson distributed panel data with two fixed effect. We will study their behaviour on simulated data, and in a simulation study compare the accuracy of estimators based on these likelihood functions.

The thesis is organized as follows. In Chapter 2 we present the incidental parameter problem with an overview of the literature and several proposed solutions, including the conditional likelihood and profile likelihood modifications suggested by Pace og Salvan (2006). We also describe the model for Poisson distributed panel data with two fixed effect, and the approximate conditional likelihood derived by Charbonneau (2012). In Chapter 3 we derive the profile likelihood corrections presented by Pace og Salvan (2006) for the Poisson panel data model, and in Chapter 4 we study the behaviour of the profile likelihood, four modifications of the profile likelihood and the approximate conditional likelihood when applied to the model presented in Chapter 2. In a simulation study we compare the mean squared errors of their respective estimators. The code used for the simulations is presented in Appendix C.

KAPITTEL 2

Background

2.1 The incidental parameter problem

In their paper, Neyman og Scott (1948) describe parameters as either «structural» or «incidental». Structural parameters appear in the probability distribution of every random variable, while incidental parameters appear in the distribution of a finite number of random variables.

We want good estimates of the structural parameters, but removal of the incidental parameters by way of maximum likelihood estimation is marred by what Neyman and Scott termed «inconsistent observations». Each observation contains information about the parameters of its probability distribution, but in the incidental parameter problem information about the incidental parameters does not increase with the data.

Let $\theta \in \Theta$ denote the structural parameters, and let $\lambda \in \Lambda$ denote the incidental (or nuisance[1]) parameters of the probability distribution of a random variable Y distributed according to the density $f(y\,;\theta,\lambda)$. Let $\mathcal{L}(\theta,\lambda\,;y)$ denote the likelihood, defined as the density of Y viewed as a function of θ and λ with y given. The classical way to do maximum likelihood estimation of θ is done in two steps. First we find $\hat{\lambda}_\theta$ as

$$\hat{\lambda}_\theta = \max_{\lambda}\{\mathcal{L}(\theta,\lambda\,;y) : \lambda \in \Lambda\}$$

the values of λ that maximizes the likelihood given θ. Then we find the estimate of θ as

$$\hat{\theta} = \max_{\theta}\{\mathcal{L}(\theta,\hat{\lambda}_\theta\,;y) : \theta \in \Theta\}$$

the value that maximizes the likelihood evaluated in $\hat{\lambda}_\theta$ the maximizing value of λ given θ. The function $\mathcal{L}(\theta,\hat{\lambda}_\theta\,;y)$ is called the profile likelihood.

In the incidental parameter problem setting the bias from the first stage estimation of the incidental parameters λ carries over to the second stage estimations of the structural parameters θ.

Neyman og Scott (1948) defined the incidental parameter problem by the following proposition.

Proposition 2.1.1. *Maximum-likelihood estimates of the structural parameters relating to a partially consistent series of observations need not be consistent.*

[1] The term «nuisance» parameters generally refer to parameters that are not of primary interest, while «incidental» parameters are nuisance parameters that increase in number with the sample size (Lancaster, 2000).

To illustrate the proposition consider the following example.

Example 2.1.2. Let $\{x_{ij}\}$ be independent and normally distributed such that the density of x_{ij} is given by

$$f(x_{ij}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_{ij} - \alpha_i)^2/2\sigma^2\}, \quad i = 1, \ldots, s,\ j = 1, \ldots, n_i,\ s \to \infty.$$

Because the $\{\alpha_i\}$ only appears in the distribution of n_i random variables they are considered incidental, while σ^2 is considered structural because it appears in the distribution of every random variable.

To demonstrate the proposition take for simplicity the samples related to α_i to be constant, $n_i = n$, for all i, the maximum-likelihood estimate of α_i is given by $\hat{\alpha}_i = \overline{x}_i$. From a well known result about the sum of squared differences from the mean we have that the maximum-likelihood estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)}{sn} \sim \frac{\sigma^2 \chi_1^2(s[n-1])}{sn}$$

where χ_1^2 denotes a random variable following a chi-square distribution with one degree of freedom. Therefore $\hat{\sigma}^2$ has expectation $\sigma^2(n-1)/n$ for every s. Since $\hat{\sigma}^2$ is biased for every s it is an inconsistent estimator for σ^2.

The important insight is that the n_i are fixed, which means we do not get more information on the incidental parameters when s increases and the bias will be constant.

2.2 The profile likelihood

The maximum of the likelihood for a given value of the parameter of interest, in our notation the function $\mathcal{L}(\theta, \hat{\lambda}_\theta(\theta)\,;\,y)$, is usually termed the profile likelihood. Often $\ell(\theta, \lambda\,;\,y) = \log(\mathcal{L}(\theta, \lambda\,;\,y)$ is used for inference as the log-function is a monotone increasing function and thus has the same maximizing value, while often also simplifying calculations. We will also refer to $\ell_P(\theta) = \log(\mathcal{L}(\theta, \hat{\lambda}_\theta(\theta)\,;\,y))$ as the profile log-likelihood.

Although $\ell_P(\theta)$ is not a genuine log-likelihood for θ, in that it is no longer a probability distribution viewed as a function of its parameters and with observed values of the stochastic variable taken as constant, it has many desirable properties. As summarized by Pace og Salvan (2006) it is invariant under interest respecting reparameterizations, it is maximised by the maximum likelihood estimate and, under mild regularity conditions, the corresponding log-likelihood ratio statistic has the usual χ^2 with q degrees of freedom as its asymptotic null distribution.

However, as pointed out by Berger et al. (1999) and Pace og Salvan (2006), it does not take the sampling variability of $\hat{\lambda}_\theta$ properly into account. One effect of this is that the score computed from the profile likelihood typically has score and information bias of order[2] $O(1)$ (McCullagh og Tibshirani, 1990). When we say the profile likelihood has information and score bias of order $O(1)$ we refer to the fact that a proper likelihood $\ell(\theta, \lambda\,;\,y)$ has

$$\mathrm{E}\left(\frac{\partial \ell(\theta, \lambda\,;\,y)}{\partial \theta}\right) = 0$$

[2] See Appendix A on page 33 for a definition of $O(f(n))$.

and

$$\mathrm{E}\left(\frac{\partial^2 \ell(\theta, \lambda\,;y)}{\partial\theta\partial\theta^T}\right) + \mathrm{E}\left\{\left(\frac{\partial \ell(\theta, \lambda\,;y)}{\partial\theta}\right)\left(\frac{\partial \ell(\theta, \lambda\,;y)}{\partial\theta}\right)^T\right\} = 0.$$

while for the profile likelihood

$$\mathrm{E}\left(\frac{\partial \ell_P(\theta)}{\partial\theta}\right) = O(1)$$

and

$$\mathrm{E}\left(\frac{\partial^2 \ell_P(\theta)}{\partial\theta\partial\theta^T}\right) + \mathrm{E}\left\{\left(\frac{\partial \ell_P(\theta)}{\partial\theta}\right)\left(\frac{\partial \ell_P(\theta)}{\partial\theta}\right)^T\right\} = O(1).$$

As a practical consequence, the usual χ^2 and normal approximations for the null distributions of the profile likelihood ratio statistic and of its signed version for a scalar θ may be poor, leading to systematically misleading inferences (Pace og Salvan, 2006).

The likelihood in the maximum point may not be representative of the entire likelihood, or even locally around this point. By only using the maximizing point for λ the profile likelihood ignores the uncertainty inherent in estimation of $\hat{\lambda}$.

When we in treat a function as a likelihood, as we do with the profile likelihood, we often term this function a pseudo likelihood, and the log of this function a pseudo log-likelihood. Thus we are looking for pseudo likelihoods that will improve on the profile likelihood when we have the incidental parameter problem.

2.3 Proposed solutions to the incidental parameter problem

The question of how to solve the incidental parameter problem depends first and foremost on which school we wish to apply. For a Bayesian statistician the treatment of incidental parameters is clear – integrate them from the likelihood with respect to a prior distribution conditioned on all remaining known or unknown parameters. The issue is then how to choose this prior (Lancaster, 2000).

We will in this thesis only study the frequentist approach, except perhaps when considering the integrated likelihoods approach, which draws much inspiration from the Bayesian school. In the frequentist setting there are several proposed solutions. In the following, we will present some of these.

2.3.1 Conditional likelihood

One proposed method for solving the incidental parameter problem, is the so-called conditional likelihood, as described by Lancaster (2000) in the following way.

If we can find a statistic S such that the likelihood of Y, $\mathcal{L}(y \mid \theta, \lambda)$, factors into a part containing only the incidental parameters λ, and a part containing the structural parameter θ and that is independent of the incidental parameters, in the following way

$$\mathcal{L}(y \mid \theta, \lambda) = \mathcal{L}_1(S \mid \lambda)\,\mathcal{L}_2(y \mid S, \theta) \tag{2.1}$$

we can make inference on θ based only on the second part $\mathcal{L}_2(y \mid S, \theta)$. If the parameter space for θ does not depend on that for λ (variation independence) and standard regularity conditions are satisfied this will provide consistent inference for λ.

If the likelihood does not factor in the original parametrization we may be able to find a reparameterization from θ, λ to θ, λ^* such that the likelihood does factor. Then the same arguments apply and consistent inference can be based on $\mathcal{L}_2$.

When Equation (2.1) on the previous page applies, possibly after a reparameterization of the incidental parameters, α, λ are termed *likelihood orthogonal*, because we then have that

$$\frac{\partial^2 \log \mathcal{L}}{\partial \alpha \partial \lambda} = 0. \tag{2.2}$$

Other factorizations

If λ, θ can not be made likelihood orthogonal but the likelihood factors as

$$\mathcal{L}(y \mid \lambda, \theta) = \mathcal{L}_1(S \mid \lambda, \theta)\, \mathcal{L}_2(y \mid S, \theta) \tag{2.3}$$

or

$$\mathcal{L}(y \mid \lambda, \theta) = \mathcal{L}_1(S \mid \theta)\, \mathcal{L}_2(y \mid S, \lambda, \theta) \tag{2.4}$$

inference may be made from $\mathcal{L}_2$ in (2.3) or $\mathcal{L}1$ in (2.4), which is free of the incidental parameter (Lancaster, 2000). The terms $\mathcal{L}_1$ in (2.3) and $\mathcal{L}_2$ in (2.4) depend on θ, so when they are ignored, and inference on θ is based on the remaining terms there is some loss of information. But supporters of this approach suggest that one loses only information about θ that is inextricably tied with the unknown parameter λ (J. O. Berger mfl., 1999).

2.3.2 Likelihood corrections

We now present the main approach for this thesis, which concerns implementing corrections to the profile likelihood. These corrections are summarized in the paper "*Adjustments of the profile likelihood from a new perspective*" by Pace og Salvan (2006). Let $\xi_0 = (\theta_0, \lambda_0)$ denote the true parameter values. Pace og Salvan (2006) defines the least favorable target log-likelihood as

$$\ell_T(\theta) = \ell(\xi_\theta)$$

where $\xi_\theta = (\theta, \lambda_\theta)$ and λ_θ is the maximizer of $\mathrm{E}_0\{\ell(\theta, \lambda)\}$ with respect to λ for fixed θ. They seek to create a pseudo log-likelihood $\ell_{PS}(\theta)$ that is an unbiased estimator of $\mathrm{E}_0(\ell_T(\theta))$, where $\mathrm{E}_0(\cdot) = \mathrm{E}(\cdot \mid \xi_0)$ denotes the expectation under ξ_0, the true value of the parameters.

Before we present Pace and Salvan's arguments for this pseudo log-likelihood we define some notation. Let

$$j(\xi) = -\frac{\partial^2 \ell(\xi)}{\partial \xi \partial \xi^T} \tag{2.5}$$

denote the observed information matrix evaluated in ξ, and let

$$j_{\theta\theta} = -\frac{\partial^2 \ell(\xi)}{\partial\theta\partial\theta^T} \tag{2.6}$$

denote a block of the observed information matrix. In a similar fashion, let $j_{\theta\lambda}$, $j_{\lambda\theta}$ and $j_{\lambda\lambda}$ denote the remaining blocks of the observed information matrix. Let also

$$i(\xi) = \mathrm{E}(j(\xi))$$

denote the expected information, and in a similar fashion to the observed information let $i_{\theta\theta}$, $i_{\theta\lambda}$, $i_{\lambda\theta}$ and $i_{\lambda\lambda}$ denote the blocks of the expected information matrix.

Motivations for the corrected likelihoods

When a suitable reduced marginal or conditional model exists whose densities depend only on θ, inference about θ may be based on the corresponding log-likelihood. The modified profile likelihood $\ell_M(\theta)$ created by Barndorff-Nielsen (Barndorff-Nielsen, 1980 and Barndorff-Nielsen, 1983) is an approximation to integrated and conditional likelihoods. Assume that the minimal sufficient statistic for the model is a one-to-one function of $(\hat{\theta}, \hat{\lambda}, a)$, where a is an ancillary statistic, either exactly or approximately, so that $\ell(\theta, \lambda\,; y) = \ell(\theta, \lambda\,; \hat{\theta}, \hat{\lambda}, a)$. Then,

$$\ell_M(\theta) = \ell_P(\theta) - \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\xi}_\theta)| - \log\left|\frac{\partial\hat{\lambda}_\theta}{\partial\hat{\lambda}}\right|, \tag{2.7}$$

with

$$\left|\frac{\partial\hat{\lambda}_\theta}{\partial\hat{\lambda}}\right| = \frac{|\ell_{\lambda;\hat{\lambda}}(\hat{\xi}_\theta)|}{|j_{\lambda\lambda}(\hat{\xi}_\theta)|},$$

where

$$\ell_{\lambda;\hat{\lambda}}(\hat{\xi}_\theta) = \frac{\partial^2 \ell(\theta, \lambda\,; \hat{\theta}, \hat{\lambda}, a)}{\partial\lambda\partial\hat{\lambda}^T}$$

are the sample space derivatives. This modified profile likelihood has score bias of order $O(n^{-1})$ (Pace og Salvan, 2006). And also has information bias of order $O(n^{-1})$ (DiCiccio mfl., 1996). Calculation of sample space derivatives is straightforward only in special classes of models, such as exponential family models. When θ and λ are orthogonal we have[3] $\log|\partial\hat{\lambda}_\theta/\partial\hat{\lambda}| = O_p(n^{-1})$ in the moderate-deviation neighbourhoods, i.e., for $\theta - \hat{\theta} = O_p(n^{-1/2})$. Which means that the pseudo log-likelihood

$$\ell_A(\theta) = \ell_P(\theta) - \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\xi}_\theta)|, \tag{2.8}$$

proposed by Cox og Reid (1987), is an approximation of $\ell_M(\theta)$ with error of order $O_p(n^{-1})$ in the moderate-deviation neighbourhoods (Pace og Salvan, 2006). Severini (1998a) proposed the modified profile likelihood $\ell_{AE}^{III}(\theta)$, which we will consider in Equation (2.12), as an approximation to $\ell_M(\theta)$.

[3] We use the notation $O_p(g(n))$ for stochastic boundedness. For details see Appendix A on page 33.

Simulation results by Diciccio og Martin (1993), Diciccio og Stern (1994) and Sartori mfl. (1999), show that inference based on the modified profile likelihood is quite accurate, even in the presence of many nuisance parameters, and when a marginal or conditional target likelihood does not exist (Pace og Salvan, 2006). Further, Sartori et al. (2003) studies the distribution of a directed likelihood[4]

$$\operatorname{sgn}(\hat{\theta}_0 - \theta)[2\{\ell_0(\hat{\theta}) - \ell_0(\theta)\}]^{1/2}$$

for θ with respect to a likelihood ℓ_0, with a maximizing value $\hat{\theta}_0$ and compares one calculated from an adjusted profile likelihood having score bias of order $O(n^{-1})$ with one calculated from the profile likelihood. They show that it is closer to the distribution of a directed likelihood calculated from a genuine likelihood.

Connection with the least favorable target likelihood

Pace og Salvan (2006) give the following arguments for seeking a pseudo log-likelihood $\ell_{PS}(\theta)$ that is an unbiased estimator of $\mathrm{E}_0(\ell_T(\theta))$. If for every $\xi_0 \in \Xi$ we have that $\mathrm{E}_0(\ell_{PS}) = \mathrm{E}_0(\ell_T(\theta))$ then $\ell_{PS}(\theta)$ has some desirable properties, also found in a genuine likelihood. In particular it satisfies

$$\mathrm{E}_0(\ell_{PS}(\theta_0)) > \mathrm{E}_0(\ell_{PS}(\theta)),$$

for $\theta \neq \theta_0$, and thus $\partial\ell_{PS}(\theta)/\partial\theta$ is an unbiased estimating function for θ_0. Moreover, at the true θ_0, the expected curvature of $\ell_{PS}(\theta)$ gives the correct information in that minus the expected Hessian at $\theta = \theta_0$ will coincide with

$$i_{\theta\theta\cdot\lambda}(\xi_0) = i_{\theta\theta}(\xi_0) - i_{\theta\lambda}(\xi_0) i_{\lambda\lambda}(\xi_0)^{-1} i_{\lambda\theta}(\xi_0),$$

the partial expected information for θ. Also, under regularity conditions, $\hat{\lambda}_\theta$ is a consistent estimator for λ_θ (Huber mfl., 1967).

In addition Pace og Salvan (2006) mentions the following property. Let $\tilde{\lambda}_\theta$ be a function of θ such that $\tilde{\lambda}_{\theta_0} = \lambda_0$ and let $\ell_R(\theta) = \ell(\theta, \tilde{\lambda}_\theta)$ be a generic log-likelihood for θ obtained through model restriction. Then

$$\mathrm{E}_0[\ell_R(\theta_0) - \ell_R(\theta)] \geq \mathrm{E}_0[\ell_T(\theta_0) - \ell_T(\theta)],$$

which means that for any given $\theta \neq \theta_0$, the curve ξ_θ minimises the Kullback–Leibler divergence between $f(y\,; \theta, \tilde{\lambda}_\theta)$ and $f(y\,; \xi_0)$ among all possible curves $(\theta, \tilde{\lambda}_\theta)$ with $\tilde{\lambda}_{\theta_0} = \lambda_0$.

In constructing $\ell_{PS}(\theta)$ Pace og Salvan (2006) view the profile likelihood, $\ell_P(\theta)$, as an estimate of $\mathrm{E}_0(\ell_T(\theta))$ with bias of order $O(1)$ and suggest using an adjustment term $a(\theta)$ for correcting the profile likelihood, resulting in an adjusted likelihood of the form

$$\ell_{AE}(\theta) = \ell_P(\theta) - a(\theta). \tag{2.9}$$

Where the adjustment term $a(\theta)$ estimates the bias

$$b(\theta\,; \xi_0) = \mathrm{E}_0(\ell_P(\theta) - \ell_T(\theta)).$$

[4] Also known as a signed likelihood ratio.

Pace og Salvan (2006) show that the following three adjusted likelihoods are asymptotically equivalent versions of $\ell_{AE}(\theta)$:

$$\ell^{I}_{AE}(\theta) = \ell_P(\theta) - \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\xi}_\theta)| - \frac{1}{2}\log|V_{\hat{\xi}}(\hat{\lambda}_\theta)| \tag{2.10}$$

$$\ell^{II}_{AE}(\theta) = \ell_P(\theta) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\xi}_\theta)| - \frac{1}{2}\log|\upsilon_{\lambda\lambda}(\hat{\xi}_\theta, \hat{\xi}_\theta\,;\hat{\xi})| \tag{2.11}$$

$$\ell^{III}_{AE}(\theta) = \ell_P(\theta) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\xi}_\theta)| - \log|\upsilon_{\lambda\lambda}(\hat{\xi}_\theta, \hat{\xi}\,;\hat{\xi})| \tag{2.12}$$

with $V_\theta(\cdot) = \mathrm{Var}(\cdot \mid \theta)$, and

$$\upsilon_{\lambda\lambda}(\xi_1, \xi_2\,;\xi_0) = \mathrm{E}_{\xi_0}(\ell_\lambda(\xi_1)\ell_\lambda(\xi_2)^T). \tag{2.13}$$

Many available adjustments to the profile log-likelihood may be seen as connected to these versions of ℓ_{AE} (Pace og Salvan, 2006).

Bootstrap estimation of the bias corrected profile likelihood

As suggested by Pace og Salvan (2006) we can also use bootstrapping to create the simulation adjusted estimative log-likelihood $\ell_{SA}(\theta)$ as an estimate of the bias corrected profile likelihood $\ell_{AE}(\theta)$ from (2.9) in the following way.

Algorithm 1: Bootstrap estimation of modified profile likelihood

```
Estimate $\hat{\theta}$ and $\hat{\lambda}$, the unconstrained maximum likelihood estimates of $\theta$
  and $\lambda$;
for $r = 1, \ldots, R$ do
    Generate bootstrap sample $y_r$ from $f(y\,;\hat{\theta}, \hat{\lambda})$;
    Calculate $\hat{\lambda}^*_\theta(r)$, the constrained maximum likelihood estimate of $\lambda$
      given $\theta$;
end
Estimate $\ell_{SA}(\theta)$ by
    $\ell_{SA}(\theta) = \frac{1}{R}\sum_{r=1}^{R} \ell(\theta, \hat{\lambda}^*_\theta(r)\,;y)$
```

Although this estimator may be computationally intensive, it has the advantage of being fairly simple to implement. Pace og Salvan (2006) also suggest this estimator on the basis that it does not require an explicit nuisance parameterisation, as it only involves constrained maximisation, and that it is invariant under interest respecting reparameterizations.

Through expansions one can show that $\ell_{SA}(\theta)$ is asymptotically of the form $\ell_{AE}(\theta)$ in (2.9), estimating $a(\theta)$ in a similar fashion as $\ell^{I}_{AE}(\theta)$ in (2.10) (Pace og Salvan, 2006).

2.3.3 Integrated Likelihoods

Another solution to the incidental parameter problem is to simply eliminate the nuisance parameter by integration of the likelihood (with respect to the

Lebesgue measure) with a weight function $\pi(\theta \mid \lambda)$. The resulting *integrated likelihood*

$$\overline{L}(\theta) = \int_{\Lambda} \mathcal{L}(\theta, \lambda \mid y)\pi(\theta \mid \lambda) d\lambda. \tag{2.14}$$

According to J. O. Berger mfl. (1999) likelihood methods which operate solely on the likelihood $\mathcal{L}(\theta)$ can also be used with an integrated likelihood. For example using the mode, $\hat{\theta}$, of $\mathcal{L}(\theta)$ as the estimate of θ and using (when θ is a p dimentional vector)

$$C = \left\{\theta : -2\log\left(\mathcal{L}(\theta)/\,\mathcal{L}(\hat{\theta})\right) \leq \chi^2_p(1-a)\right\}$$

as an approximate $100(1-a)\%$ confidence set for θ, where $\chi^2_p(1-a)$ is he $(1-a)$th quantile of the chi squared distribution with p degrees of freedom (Sweeting, 1995).

Integrated likelihoods has been studied for the incidental parameter problem in the frequentist context by De Bin mfl. (2015), but as we did not implement integrated likelihoods in this thesis we will not go into further detail here. The interested reader can find more in Appendix B.

KAPITTEL 3

Likelihood corrections

For the rest of this thesis we will focus on a model of Poisson distributed variables in two way panel data with fixed effects.

3.1 Poisson model with two fixed effects

We want to model trade data where y_{ij} models trade flow in number of goods exported from country i and imported by country j. This means that there are no observations y_{ii}, as no country trades with itself. We have a panel of n individuals with $n \times (n-1)$ observations where each observation is independently Poisson distributed

$$y_{ij} \sim \text{Pois}(\lambda_{ij}) \quad i \neq j$$

with

$$\lambda_{ij} = \exp(x_{ij}\beta + \mu_i + \alpha_j).$$

where μ_i and α_j are individual fixed effects, x_{ij} is a vector of explanatory variables and β a vector of k elements $(\beta_1\ \beta_2\ \cdots\ \beta_k)^T$. The pdf of y_{ij} is

$$f(x\,;\beta) = \frac{\lambda_{ij}^{y_{ij}} e^{-\lambda_{ij}}}{y_{ij}!}, \quad i \neq j.$$

The model is motivated by the gravity equations that are frequently used to model international trade. Charbonneau (2012), Helpman mfl. (2008), and Silva og Tenreyro (2006) use such nonlinear models with fixed effects for importing and exporting countries. This model is also relevant for other areas, such as labor economics, where a wage equation might contain both worker and firm fixed effects, or industrial organization, where knowledge diffusion equations using patent data can include citing and cited country fixed effects (Charbonneau, 2012).

When modelling trade with the gravity equations, the Poisson model with two fixed effects is frequently used (Charbonneau, 2012). Hausman mfl. (1984) used a conditional maximum likelihood approach to develop what is now called the fixed effect Poisson estimator for the Poisson model with one fixed effect. But Lancaster (2002) shows that there really is not an incidental parameter problem in the Poisson model with one fixed effect. The maximum likelihood estimator, the maximum likelihood estimator conditioned on the sufficient statistic and a Bayes posterior[1] all yield as $n \to \infty$ the same consistent answer Lancaster

[1] after integrating the reparameterized nuisance parameter with respect to any proper prior exhibiting independence between the nuisance parameters and parameters of interest

(2000). But for the Poisson model with two fixed effects the incidental parameter problem remains as shown by Charbonneau (2012).

3.2 Conditional likelihood for a Poisson model with fixed effects

We now consider the conditional likelihood for the proposed model by first giving a quick summary of a conditional likelihood for a model with one fixed effect, leading to the derivation of a conditional likelihood for the model with two fixed effects.

Conditional likelihood for a Poisson model with one fixed effect

Charbonneau (2012) presents the conditional likelihood for a Poisson model with one fixed effect. Consider first the case with two observations for each individual. Let

$$y_{it} \sim \text{Pois}(\exp(x_{it}\beta + \alpha_i)).$$

The distribution of y_{i1} given that $y_{i1} + y_{i2} = K$ is given by

$$y_{i1} \mid (y_{i1} + y_{i2} = K) \sim \text{Binom}\left(K, \frac{\exp(x_{i1}\beta)}{\exp(x_{i1}\beta) + \exp(x_{i2}\beta)}\right).$$

This does not involve the fixed effects, and will therefore give consistent estimates of β.

The fixed effect Poisson estimator of Hausman mfl. (1984) extends this logic to several observations for each individual in the following way. Suppose we have observations of n individuals, y_{ij}, observed over T periods. Let

$$y_{ij} \sim \text{Pois}(\lambda_{it} = \exp(x_{it}\beta + \alpha_i + \alpha_0)),$$

where α_i represents individual fixed effects and α_0 is the overall intercept. We then have

$$\text{P}(y_{it} \mid x_{it}, \alpha_i) = \frac{e^{-\lambda_{it}}\lambda_{it}^{y_{it}}}{y_{it}!}. \tag{3.1}$$

The incidental parameter problem prevents us from consistently estimating the parameters in (3.1) by maximum likelihood. To solve this problem, Hausman mfl. (1984) follow Andersen (1970) and Andersen (1972) and condition on the sum $\sum_t y_{it}$ (cited by Charbonneau, 2012). This is a sufficient statistic for α_i, and it is well known that the distribution of y_{it} conditional on $\sum_t y_{it}$ is a multinomial distribution such that

$$\begin{aligned}\text{P}(y_{i1}, y_{i2}, \cdots, y_{iT} \mid \sum y_{it}) &= \frac{\text{P}(y_{i1}, y_{i2}, \cdots, y_{i,T-1}, \sum_{t=1}^{T} y_{it} - \sum_{t=1}^{T-1})}{\text{P}(\sum y_{it})} \\ &= \frac{\frac{e^{-\Sigma_t \lambda_{it}} \prod_t \lambda_{it}^{y_{it}}}{\prod_t (y_{it}!)}}{\frac{e^{-\Sigma_t \lambda_{it}} \left(\sum_t \lambda_{it}\right)^{\Sigma_t y_{it}}}{\left(\sum y_{it}\right)!}}\end{aligned}$$

$$= \frac{\left(\sum_t y_{it}\right)!}{\prod_t (y_{it}!)} \prod_t \left[\frac{\lambda_{it}}{\sum_t \lambda_{it}}\right]^{y_{it}}.$$

The term on the right can be simplified to

$$\frac{e^{x_{it}\beta+\mu_i}}{\sum_t e^{x_{it}\beta+\mu_i}} = \frac{e^{x_{it}\beta}}{\sum_t e^{x_{it}\beta}}$$

which does not depend on the fixed effects. Thus we can use it to produce a likelihood function to consistently estimate the parameter β.

Conditional likelihood for a Poisson model with two fixed effect

In order to eliminate the nuisance parameters in the model Charbonneau (2012) conditions on the vector of sums of the columns $\{\sum_i y_{ij}\}$, denoted c, and the vector of sums of the rows $\{\sum_j y_{ij}\}$, denoted r, which are the sufficient statistics for the fixed effects.

Let Y denote the vector of observations $\{y_{ij}\}$, and let μ, α denote the vectors of fixed effects. Let $x = \{x_{ij}\}$ denote the covariates, and β the parameter or vector of parameters of interest. Define also $\mathcal{Q}$ to be the set of all possible distributions of y_{ij} such that the sum of the rows is given by r and the sums of columns is given by c. Charbonneau (2012) shows that the resulting conditional probability distribution of the data is

$$\begin{aligned} \mathrm{P}(Y \mid r, c, \alpha, \mu, x, \beta) &= \frac{\mathrm{P}(Y \mid \alpha, \mu, x, \beta)}{\mathrm{P}(r, c \mid \alpha, \mu, x, \beta)} \\ &= \frac{\mathrm{P}(Y \mid \alpha, \mu, x, \beta)}{\sum_{Y' \in \mathcal{Q}} \mathrm{P}(Y' \mid \alpha, \mu, x, \beta)} \\ &= \frac{\dfrac{e^{-\Sigma_i \Sigma_j \lambda_{ij}} \prod_{i,j=1}^{n} \lambda_{ij}^{y_{ij}}}{\prod_{i,j=1}^{n} (y_{ij}!)}}{\sum_{Y' \in \mathcal{Q}} \dfrac{e^{-\Sigma_i \Sigma_j \lambda_{ij}} \prod_{i,j=1}^{n} \lambda_{ij}^{y'_{ij}}}{\prod_{i,j=1}^{n} (y'_{ij}!)}}. \end{aligned}$$

This is the probability of Y over the sum of probabilities of all possible Y' that have the same sum of rows and columns. Implementing this sum is not computationally feasible, as there will be too many possible Y''s in realistic applications, such as those reflecting trade between countries. In stead Charbonneau (2012) presents the following estimator. Compare a matrix Y to one other alternative matrix Y' with the same sum of rows and columns, which results in a likelihood function that does not depend on the fixed effects in the following way

$$\begin{aligned} &\frac{\mathrm{P}(Y \mid \alpha, \mu, x, \beta)}{\mathrm{P}(Y \mid \alpha, \mu, x, \beta) + \mathrm{P}(Y' \mid \alpha, \mu, x, \beta)} \\ &\qquad = \left(1 + \frac{\prod_{i,j=1}^{n} (y_{ij}!)}{\prod_{i,j=1}^{n} (y'_{ij}!)} \left(e^{\sum_i \sum_j x_{ij}\beta(y'_{ij} - y_{ij})}\right)\right)^{-1}. \end{aligned}$$

To implement this for estimation, Charbonneau (2012) suggests selecting a random l and k for each observation ij to compose a small 2×2 matrix. Then generate a second matrix, Y' that has the same sum of columns and rows. For each pair ij, the procedure can be repeated T times, using the estimate that minimizes

$$\ell_{CL}(\beta) = \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n} \log\left(1+\frac{y_{ij}!y_{ik}!y_{lj}!y_{lk}!}{y'_{ij}!y'_{ik}!y'_{lj}!y'_{lk}!}\big(\exp\big\{\textstyle\sum_{a,b\in\{i,j,l,k\}} \beta x_{ab}(y'_{ab}-y_{ab})\big\}\big)\right) \tag{3.2}$$

with

$$\begin{aligned}&\exp\big\{\textstyle\sum_{a,b\in\{i,j,l,k\}} \beta x_{ab}(y'_{ab}-y_{ab})\big\}\\ &= \exp\Big\{\beta\big[x_{ij}(y'_{ij}-y_{ij}) + x_{ik}(y'_{ik}-y_{ik}) + x_{lj}(y'_{lj}-y_{lj}) + x_{lk}(y'_{lk}-y_{lk})\big]\Big\}\end{aligned}$$

3.3 Corrected likelihoods

We now implement the likelihood corrections suggested by Pace og Salvan (2006) for a model with Poisson panel data with two fixed effects, as described in section 3.1.

The log-likelihood for this model is given by

$$\ell(\beta,\mu,\alpha) = \ell(\beta,\mu,\alpha\,;x,y) = \sum_{ij} y_{ij}\log(\lambda_{ij}) - \lambda_{ij} - \log(y_{ij}!). \tag{3.3}$$

Observe that the score for a single β_h, μ_k and α_l is

$$\frac{\partial\ell(\beta,\mu,\alpha)}{\partial(\beta_h,\mu_k,\alpha_l)} = \begin{bmatrix}\sum_{ij} y_{ij}x_{ijh} - x_{ijh}\lambda_{ij}\\ \sum_j y_{kj} - \lambda_{kj}\\ \sum_i y_{il} - \lambda_{il}\end{bmatrix}. \tag{3.4}$$

To describe the information matrix, observe first that

$$-\frac{\partial^2\ell(\beta,\mu,\alpha)}{\partial\mu_i\partial\mu_{i'}^T} = \begin{cases}\sum_j \lambda_{ij}, & \text{for } i=i'.\\ 0, & \text{for } i\neq i'.\end{cases} \tag{3.5}$$

$$-\frac{\partial^2\ell(\beta,\mu,\alpha)}{\partial\alpha_i\partial\alpha_{i'}^T} = \begin{cases}\sum_i \lambda_{ij}, & \text{for } i=i'.\\ 0, & \text{for } i\neq i'.\end{cases} \tag{3.6}$$

$$-\frac{\partial^2\ell(\beta,\mu,\alpha)}{\partial\mu_i\partial\alpha_j^T} = -\frac{\partial^2\ell(\beta,\mu,\alpha)}{\partial\alpha_j\partial\mu_i^T} = \lambda_{ij}, \quad \text{for } i\neq j. \tag{3.7}$$

Recall that there is no observation y_{ii}, and thus there is no observation containing both μ_i and α_i.

Note that there is no stochastic element y_{ij} in the observed information for the nuisance parameters, μ and α, and therefore the observed and the expected information matrix for the nuisance parameters will be the same. From (3.5), (3.6) and (3.7) we see that

$$j_{\xi\xi}(\beta,\, \mu,\, \alpha) = -\,\mathrm{E}\left(\frac{\partial^2 \ell(\beta, \mu, \alpha)}{\partial(\mu,\ \alpha)\partial(\mu,\ \alpha)^T}\right) = \begin{bmatrix} i_{\alpha\alpha} & i_{\alpha\beta} \\ i_{\beta\alpha} & i_{\beta\beta} \end{bmatrix}$$

Where

$$[i_{\alpha\alpha}]_{i,i'} = \begin{cases} \sum_j \lambda_{ij}, & \text{if } i = i' \\ 0, & \text{otherwise.} \end{cases}$$

$$[i_{\beta\beta}]_{i,i'} = \begin{cases} \sum_j \lambda_{ji}, & \text{if } i = i' \\ 0, & \text{otherwise.} \end{cases}$$

$$[i_{\alpha\beta}]_{i,i'} = \lambda_{i(i'+1)}.$$
$$[i_{\beta\alpha}]_{i,i'} = \lambda_{(i+1)i'}.$$
$$[i_{\alpha\beta}]_{i,i'} = 0, \quad \text{for } i = i'.$$

3.4 Adjusted likelihoods

Let $\xi = (\mu,\, \alpha)$ and let $\hat{\xi}_\beta = (\hat{\mu}_\beta,\, \hat{\alpha}_\beta)$ denote the constrained maximum likelihood estimates of μ and α given β and let $\hat{\xi}_{\hat{\beta}} = (\hat{\mu}_{\hat{\beta}},\, \hat{\alpha}_{\hat{\beta}})$ denote their unconstrained maximum likelihood estimates. Let also $\theta = (\beta, \mu, \alpha)$, $\hat{\theta}_\beta = (\beta, \hat{\mu}, \hat{\alpha})$ and $\hat{\theta}_{\hat{\beta}} = (\hat{\beta}, \hat{\mu}, \hat{\alpha})$.

From Equation (2.13)

$$v_{\theta,\theta}(\theta_1, \theta_2\,; \theta_0) = \mathrm{E}_{\theta_0}\left[\ell_\theta(\theta_1) \cdot \ell_\theta(\theta_2)^T\right].$$

Pace and Salvans first correction

We want to implement

$$\ell^I_{AE}(\beta) = \ell_P(\beta) - \frac{1}{2}\log|j_{\xi\xi}(\hat{\theta}_\beta)| - \frac{1}{2}\log|\,\mathrm{V}_{\hat{\xi}}(\hat{\xi}_\beta)|.$$

From a well known property of maximum likelihood estimators the variance of the estimator is asymptotically the expected information for the variables to be estimated (Casella og R. L. Berger, 2002, s. 472).

$$\mathrm{V}_{\hat{\xi}}(\hat{\xi}_\beta) \xrightarrow{\mathrm{P}} \mathrm{E}_{\hat{\xi}}\left[\frac{\partial^2 \ell(\beta_0, \mu_0, \alpha_0\,; y)}{\partial(\mu, \alpha)\partial(\mu, \alpha)^T}\right] = \mathrm{E}_{\hat{\xi}}(j_{\xi\xi}(\theta_0)) = j_{\xi\xi}(\theta_0)$$

with $(\beta_0, \mu_0, \alpha_0)$ the true values of (β, μ, α), and the last equality results from the fact that, as shown above, for this model the observed and expected information is the same. As the maximum likelihood estimator is a consistent estimator we have that (Casella og R. L. Berger, 2002, s. 472)

$$j_{\xi\xi}(\hat{\theta}_\beta) \xrightarrow{\mathrm{P}} j_{\xi\xi}(\theta_0)$$

which means we can use $j_{\xi\xi}(\hat{\xi}_\beta)$ to estimate $\mathrm{V}_{\hat{\xi}}(\hat{\xi}_\beta))$. Thus our estimator for $\ell^I_{AE}(\beta)$ is

$$\ell^I_{AE}(\beta) = \ell_P(\beta) - \log|j_{\xi\xi}(\hat{\theta}_\beta)|. \tag{3.8}$$

Note that this is very close to the modified likelihood ℓ_A in Equation (2.8) that was proposed by Cox og Reid (1987).

Pace and Salvans second correction

We now implement the second estimator in Equation (2.11). In the notation for the current model

$$\ell^{II}_{AE}(\beta) = \ell_P(\beta) + \frac{1}{2}\log|j_{\xi\xi}(\hat{\xi}_\beta)| - \frac{1}{2}\log|v_{\theta\theta}(\hat{\theta}_\beta, \hat{\theta}_\beta\,; \hat{\theta})|$$

We need to find the expression for

$$v_{\theta,\theta}(\hat{\theta}_\beta, \hat{\theta}_\beta\,; \hat{\theta}) = \mathrm{E}_{\hat{\beta},\,\hat{\xi}}\left[\ell_\xi(\hat{\theta}_\beta)\cdot\ell_\xi(\hat{\theta}_\beta)^T\right].$$

For ease of readability, let $\mathrm{E}_\beta(y) = \mathrm{E}(y \mid \beta, \hat{\xi})$ denote the expectation evaluated in $(\beta, \hat{\xi})$, and similarly let $\mathrm{Cov}_\beta(y) = \mathrm{Cov}_{\beta,\hat{\xi}}(y)$. Let also $\hat{\lambda}_{ij(\beta)} = \mathrm{E}_\beta(y_{ij}) = \exp\{\beta x_{ij} + \hat{\mu}_i + \hat{\alpha}_i\}$. The matrix $v_{\xi,\xi}(\hat{\theta}_\beta, \hat{\theta}_\beta\,; \hat{\theta})$ will be of the form

$$v_{\xi,\xi}(\hat{\theta}_\beta, \hat{\theta}_\beta\,; \hat{\theta}) = \begin{bmatrix} v_{\mu\mu} & v_{\mu\alpha} \\ v_{\alpha\mu} & v_{\alpha\alpha} \end{bmatrix}.$$

Using the score equations in Equation (3.4) on page 14 we find element (i, i') in matrix $v_{\mu\mu}$ as

$$\begin{aligned}
\left[v_{\mu\mu}\right]_{i,i'} &= \mathrm{E}_{\hat{\beta}}\Big(\sum_j (y_{ij} - \hat{\lambda}_{ij(\beta)}) \cdot \sum_j (y_{i'j} - \hat{\lambda}_{i'j(\beta)})\Big) \\
&= \sum_{jk} \mathrm{E}_{\hat{\beta}}(y_{ij}y_{i'k}) - 2\sum_{jk} \mathrm{E}_{\hat{\beta}}(y_{ij})\hat{\lambda}_{i'k(\beta)} + \sum_{jk} \hat{\lambda}_{ij(\beta)}\hat{\lambda}_{i'k(\beta)} \\
&= \sum_{jk} \big[\mathrm{Cov}_{\hat{\beta}}(y_{ij}, y_{i'k}) + \mathrm{E}_{\hat{\beta}}(y_{ij})\,\mathrm{E}_{\hat{\beta}}(y_{i'k})\big] \\
&\quad - 2\sum_{jk} \mathrm{E}_{\hat{\beta}}(y_{ij})\hat{\lambda}_{i'k(\beta)} + \sum_{jk} \hat{\lambda}_{ij(\beta)}\hat{\lambda}_{i'k(\beta)} \\
&= \sum_{jk} \mathrm{Cov}_{\hat{\beta}}(y_{ij}, y_{i'k}) + \sum_{jk} \big[\mathrm{E}_{\hat{\beta}}(y_{ij}) - \hat{\lambda}_{ij(\beta)}\big]\big[\mathrm{E}_{\hat{\beta}}(y_{i'k}) - \hat{\lambda}_{i'k(\beta)}\big] \\
&= \sum_{jk} \mathrm{Cov}_{\hat{\beta}}(y_{ij}, y_{i'k}) + \sum_{jk} \big[\hat{\lambda}_{ij(\hat{\beta})} - \hat{\lambda}_{ij(\beta)}\big]\big[\hat{\lambda}_{i'k(\hat{\beta})} - \hat{\lambda}_{i'k(\beta)}\big] \\
&= \sum_{jk} \mathrm{Cov}_{\hat{\beta}}(y_{ij}, y_{i'k}) + \sum_{j} \big[\hat{\lambda}_{ij(\hat{\beta})} - \hat{\lambda}_{ij(\beta)}\big] \sum_{j} \big[\hat{\lambda}_{i'j(\hat{\beta})} - \hat{\lambda}_{i'j(\beta)}\big].
\end{aligned}$$

In the third equality we use the formula $\mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2)$. From the independence of the observations $\{y_{ij}\}$ and using the fact that, for the Poisson distribution, $\mathrm{Var}(y_{ij}) = \lambda_{ij}$

$$\begin{aligned}
\sum_{jk} \mathrm{Cov}_{\hat{\beta}}(y_{ij}, y_{i'k}) &= \begin{cases} \sum_j \mathrm{Var}_{\hat{\beta}}(y_{ij}), & \text{for } i = i' \\ 0, & \text{for } i \neq i'. \end{cases} \\
&= \begin{cases} \sum_j \hat{\lambda}_{\hat{\beta}}, & \text{for } i = i' \\ 0, & \text{for } i \neq i'. \end{cases}
\end{aligned}$$

Similarly

$$\left[v_{\alpha\alpha}\right]_{i,i'} = \mathrm{E}_{\hat{\beta}}\Big(\sum_j (y_{ji} - \hat{\lambda}_{ji(\beta)}) \cdot \sum_j (y_{ji'} - \hat{\lambda}_{ji'(\beta)})\Big)$$

$$= \sum_{jk} \mathrm{Cov}_{\hat\beta}(y_{ji}, y_{ki'}) + \sum_j \big[\hat\lambda_{ji(\hat\beta)} - \hat\lambda_{ji(\beta)}\big] \sum_j \big[\hat\lambda_{ji'(\hat\beta)} - \hat\lambda_{ji'(\beta)}\big].$$

with

$$\begin{aligned}
\sum_{jk} \mathrm{Cov}_{\hat\beta}(y_{ji}, y_{ki'}) &= \begin{cases} \sum_j \mathrm{Var}_{\hat\beta}(y_{ji}), & \text{for } i = i' \\ 0, & \text{for } i \neq i'. \end{cases} \\
&= \begin{cases} \sum_j \hat\lambda_{\hat\beta}, & \text{for } i = i' \\ 0, & \text{for } i \neq i'. \end{cases}
\end{aligned}$$

and

$$\begin{aligned}
\big[v_{\mu\alpha}\big]_{i,i'} &= \big[v_{\alpha\mu}\big]_{i',i} \\
&= \mathrm{E}_{\hat\beta}\Big(\sum_j (y_{ij} - \hat\lambda_{ij(\beta)}) \cdot \sum_j (y_{ji'} - \hat\lambda_{ji'(\beta)})\Big) \\
&= \sum_{jk} \mathrm{Cov}_{\hat\beta}(y_{ij}, y_{ki'}) + \sum_j \big[\hat\lambda_{ij(\hat\beta)} - \hat\lambda_{ij(\beta)}\big] \sum_j \big[\hat\lambda_{ji'(\hat\beta)} - \hat\lambda_{ji'(\beta)}\big], \\
&= \mathrm{Var}_{\hat\beta}(y_{ii'}) + \sum_j \big[\hat\lambda_{ij(\hat\beta)} - \hat\lambda_{ij(\beta)}\big] \sum_j \big[\hat\lambda_{ji'(\hat\beta)} - \hat\lambda_{ji'(\beta)}\big], \\
&= \hat\lambda_{ii'\hat\beta} + \sum_j \big[\hat\lambda_{ij(\hat\beta)} - \hat\lambda_{ij(\beta)}\big] \sum_j \big[\hat\lambda_{ji'(\hat\beta)} - \hat\lambda_{ji'(\beta)}\big], \qquad \text{for } i \neq i', \\
\big[v_{\mu\alpha}\big]_{i,i'} &= \big[v_{\alpha\mu}\big]_{i',i} = 0, \qquad \text{for } i = i'.
\end{aligned}$$

For ease of implementation we will define the matrix

$$M = \begin{bmatrix} M_{\mu\mu} & M_{\mu\alpha} \\ M_{\alpha\mu} & M_{\alpha\alpha} \end{bmatrix}$$

with elements

$$\big[M_{\mu\mu}\big]_{i,i'} = \sum_j \big[\hat\lambda_{ij(\hat\beta)} - \hat\lambda_{ij(\beta)}\big] \sum_j \big[\hat\lambda_{i'j(\hat\beta)} - \hat\lambda_{i'j(\beta)}\big].$$

$$\big[M_{\alpha\alpha}\big]_{i,i'} = \sum_j \big[\hat\lambda_{ji(\hat\beta)} - \hat\lambda_{ji(\beta)}\big] \sum_j \big[\hat\lambda_{ji'(\hat\beta)} - \hat\lambda_{ji'(\beta)}\big].$$

$$\big[M_{\mu\alpha}\big]_{i,i'} = \sum_j \big[\hat\lambda_{ij(\hat\beta)} - \hat\lambda_{ij(\beta)}\big] \sum_j \big[\hat\lambda_{ji'(\hat\beta)} - \hat\lambda_{ji'(\beta)}\big].$$

and

$$\big[M_{\alpha\mu}\big]_{i',i} = \big[M_{\mu\alpha}\big]_{i,i'}$$

We then have

$$v_{\xi,\xi}(\hat\theta_\beta, \hat\theta_\beta\,; \hat\theta) = j_{\xi\xi}(\hat\theta) + M$$

and

$$\ell^{II}_{AE}(\beta) = \ell_P(\beta) + \frac{1}{2}\log|j_{\xi\xi}(\hat\xi_\beta)| - \frac{1}{2}\log|j_{\xi\xi}(\hat\theta) + M| \tag{3.9}$$

Pace and Salvans third correction

To implement the third corrected likelihood

$$\ell_{AE}^{III}(\beta) = \ell_P(\beta) + \frac{1}{2}\log|j_{\xi\xi}(\hat{\theta}_\beta)| - \log|v_{\theta\theta}(\hat{\theta}_\beta, \hat{\theta}\,;\hat{\theta})|$$

we first observe that the covariance matrix $v_{\theta,\theta}(\hat{\theta}_\beta, \hat{\theta}\,;\hat{\theta})$ is on the form

$$v_{\theta,\theta}(\hat{\theta}_\beta, \hat{\theta}\,;\hat{\theta}) = \mathrm{E}_{\hat{\theta}}\left[\ell_\xi(\hat{\theta}_\beta)\cdot\ell_\xi(\hat{\theta})^T\right] = \begin{bmatrix} v^*_{\mu\mu} & v^*_{\mu\alpha} \\ v^*_{\alpha\mu} & v^*_{\alpha\alpha} \end{bmatrix}$$

with elements in the upper left block given by

$$\begin{aligned}
\left[v^*_{\mu\mu}\right]_{i,i'} &= \mathrm{E}_{\hat{\theta}}\Big(\sum_j (y_{ij} - \hat{\lambda}_{ij(\beta)})\cdot\sum_j \big(y_{i'j} - \lambda_{i'j(\hat{\beta})}]\big)\Big) \\
&= \mathrm{E}_{\hat{\theta}}\Big(\sum_j (y_{ij} - \hat{\lambda}_{ij(\beta)})\cdot\sum_j \big(y_{i'j} - \mathrm{E}_{\hat{\theta}}[y_{i'j}]\big)\Big) \\
&= \mathrm{E}_{\hat{\theta}}\Big(\sum_{jk} y_{ij}y_{i'k} - \sum_{jk} y_{ij}\,\mathrm{E}_{\hat{\theta}}(y_{i'j}) \\
&\qquad - \sum_{jk}\hat{\lambda}_{ij(\beta)}y_{i'k} + \sum_{jk}\hat{\lambda}_{ij(\beta)}\,\mathrm{E}_{\hat{\theta}}(y_{i'k})\Big) \\
&= \sum_{jk}\mathrm{E}_{\hat{\theta}}(y_{ij}y_{i'k}) - \sum_{jk}\mathrm{E}_{\hat{\theta}}(y_{ij})\,\mathrm{E}_{\hat{\theta}}(y_{i'j}) \\
&\qquad - \sum_{jk}\hat{\lambda}_{ik(\beta)}\,\mathrm{E}_{\hat{\theta}}(y_{i'j}) + \sum_{jk}\hat{\lambda}_{ij(\beta)}\,\mathrm{E}_{\hat{\theta}}(y_{i'k}) \\
&= \sum_{jk}\left[\mathrm{E}_{\hat{\theta}}(y_{ij}y_{i'k}) - \mathrm{E}_{\hat{\theta}}(y_{ij})\,\mathrm{E}_{\hat{\theta}}(y_{i'j})\right] \\
&\qquad - \sum_{jk}\hat{\lambda}_{ik(\beta)}\,\mathrm{E}_{\hat{\theta}}(y_{i'j}) + \sum_{jk}\hat{\lambda}_{ij(\beta)}\,\mathrm{E}_{\hat{\theta}}(y_{i'k}) \\
&= \sum_{jk}\mathrm{Cov}_{\hat{\theta}}(y_{ij}, y_{i'k}) + 0 \\
&= \begin{cases} \sum_j \lambda_{ij(\hat{\theta})}, & \text{for } i = i', \\ 0, & \text{for } i \neq i'. \end{cases}
\end{aligned}$$

Similarly

$$\begin{aligned}
\left[v^*_{\alpha\alpha}\right]_{i,i'} &= \mathrm{E}_{\hat{\theta}}\Big(\sum_j (y_{ji} - \hat{\lambda}_{ji(\beta)})\cdot\sum_j \big(y_{ji'} - \lambda_{ji'(\hat{\beta})}]\big)\Big) \\
&= \begin{cases} \sum_j \lambda_{ji(\hat{\theta})}, & \text{for } i = i', \\ 0, & \text{for } i \neq i'. \end{cases}
\end{aligned}$$

$$\left[v^*_{\mu\alpha}\right]_{i,i'} = \left[v^*_{\alpha\mu}\right]_{i',i} = \mathrm{E}_{\hat{\theta}}\Big(\sum_j (y_{ij} - \hat{\lambda}_{ij(\beta)})\cdot\sum_j \big(y_{ji'} - \lambda_{ji'(\hat{\beta})}]\big)\Big)$$

$$= \begin{cases} 0, & \text{for } i = i', \\ \lambda_{ii'(\hat{\theta})}, & \text{for } i \neq i'. \end{cases}$$

Thus for this model

$$v_{\theta,\theta}(\hat{\theta}_\beta, \hat{\theta}\,; \hat{\theta}) = j_{\xi\xi}(\hat{\theta})$$

and

$$\ell_{AE}^{III}(\beta) = \ell_P(\beta) + \frac{1}{2} \log|j_{\xi\xi}(\hat{\theta}_\beta)| - \log|j_{\xi\xi}(\hat{\theta})| \tag{3.10}$$

KAPITTEL 4

Results

In order to compare the usefulness of the described pseudo-likelihoods we will compare their performance on simulated data. With simulated data we can study the performance of the different pseudo-likelihoods on data that closely resembles data from real world applications. It also allows for comparison of performance on different sample sizes. To ease implementations, without loss of generality for the results, we will include observations for y_{ij} with $i = j$, even though these observations are not present in trade data applications.

We only consider the case where we have one parameter of interest β. This makes the implementation easier and the computational burden smaller, as optimization over multiple parameters is both difficult and computationally intensive. With a simple implementation it is easier to separate aspects of the pseudo-likelihood of interest from computational quirks of optimization.

We will implement and compare the profile log-likelihood $\ell_P(\beta)$, described in section 2.2, the approximate conditional likelihood $\ell_{CL}(\beta)$ from Equation 3.2, in this section denoted as ”the conditional likelihood”, the likelihood corrections $\ell^I_{AE}(\beta)$, $\ell^{II}_{AE}(\beta)$ and $\ell^{III}_{AE}(\beta)$ from Equations (3.8), (3.9) and (3.10), and the bootstrap estimated corrected likelihood $\ell_{BS}(\beta)$ from section 2.3.2. The modified likelihood $\ell_A(\beta)$ in Equation (2.8) consistently gave the same estimates as $\ell^I_{AE}(\beta)$ in preliminary studies and therefore is not presented here.

We will generate samples of simulated data and for each sample estimate β using the mentioned pseudolikelihoods. In a preliminary study, we inspect the estimates for 10 samples and study plots of the pseudo log-likelihoods for a single sample. In a simulation study we calculate the root of the mean of squared errors (RMSE) of estimates for 250 samples. For an estimator $\tilde{\beta}$ with estimates $\{\tilde{\beta}_i\}$ RMSE is given by

$$\text{RMSE} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\beta}_i - \beta_0)^2.$$

To estimate the maximizing β for each pseudo log-likelihood we compute the value of the pseudolikelihoods over a grid of evenly spaced β values, ensuring that the true β_0 is included, and use the maximizing value from this grid as an estimate. The grid is chosen to be evenly spaced in the interval $\big[(1-30)\cdot\beta_0,\ (1+30)\cdot\beta_0\big]$, to ensure the candidate β solutions represent a wide interval relative to β_0. This way we are not ”cheating” too much, by only presenting candidates close to the true parameter value. The simulations are all done in R (R Core Team, 2021) and computational estimation of maximum likelihood estimates is done with the `optim` package using the *BFGS*, algorithm as this algorithm seems to have a good reputation for multivariate optimization and seems to give reasonably good results for our data.

4.1 Simulating data

To simulate the data we draw

$$y_{ij} \sim \text{Pois}(\lambda_{ij}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n \tag{4.1}$$

from the Poisson distribution with parameter

$$\lambda_{ij} = \exp(x_{ij}\beta + \mu_i + \alpha_j).$$

We will set $x_{ij} = 1$ for all i, j and $\mu_i = \alpha_j = 1$, for all i, j. Thus we are assuming all incidental parameters have an equal impact on the observations.

4.2 Settings

Selection of the precision parameter T for $\ell_{BS}(\beta)$ in Equation (3.2) was done heuristically. With larger T more alternative sample points are used for comparison in the estimator. The theoretical conditional likelihood compares with all possible alternative sample points, which is computationally infeasible. Setting $T = 1000$ was not very intensive, and increasing to $T = 3000$ did not produce noticeably different results. We therefore chose $T = 1000$ for the entire simulation study.

Selection of R, the number of bootstrap samples in the bootstrap estimator $\ell_{BS}(\beta)$ from section 2.3.2, was also done heuristically. The results were very good with $R = 70$, and did not improve with larger R. We therefore use $R = 70$ for the entire simulation study.

The model is in its current state overparameterized for parameter estimation. Since the nuisance parameter enter the model through $\mu_i + \alpha_j$ for y_{ij} adding a constant to all μ and subtracting the same constant from all α leaves the likelihood unchanged. Therefore we need to standardize, for example by setting $\mu_1 = a$. In our simulations we let $a = 1$, thus assuming that we are guessing the correct value.

4.3 Preliminary study

We first study the case of a small sample of 10 × 10 observations. Figures 4.1, 4.2 and 4.3 show series of estimation errors $\hat{\theta} - \theta_0$ for 10 different samples, with respectively $\beta_0 = 0.2$, $\beta_0 = 1$ and $\beta_0 = 10$. Figure 4.4 shows log-transformed log-likelihood of a single sample. The transformation $-\log(-\ell(\beta_i))$ was used on the plotted log-likelihoods in order to accentuate teaks that were relatively small compared to the fluctuations in the likelihoods, especially for β values far away from the true β_0.

As expected the profile likelihood estimator does not perform very well. The estimates are very unstable, and often far away from the true value.

The three likelihood corrections $\ell^{I}_{AE}(\beta)$, $\ell^{II}_{AE}(\beta)$ and $\ell^{III}_{AE}(\beta)$ do not seem to improve upon the profile likelihood for any of the three values of β_0.

The third corrected likelihood $\ell^{III}_{AE}(\beta)$ had problems producing estimates. The problem seems to lie with the matrix $j_{\xi\xi}(\hat{\theta})$, which only relies on $\hat{\theta}$, the unconstrained maximum likelihood estimate for the parameters. The estimated determinant for this matrix is close to zero, and sometimes negative, which is a problem as we expect this matrix to be positive definite. Since we take the log of this determinant, a negative determinant will result in `NA` values polluting the computations.

The approximate conditional likelihood $\ell_{CL}(\beta)$ performs differently for the three values for β_0. For $\beta_0 = 0.2$ and $\beta_0 = 1$ the estimates from $\ell_{CL}(\beta)$ closely resembles those based on the profile likelihood, but does not seem to consistently improve on it, while for $\beta_0 = 10$ the approximate conditional likelihood estimator is completely off, choosing values on the border of the candidate β values. This is surprising, considering that larger β_0 has increased effect on the observations relative to the nuisance parameters, and thus it should be easier for the estimators to identify.

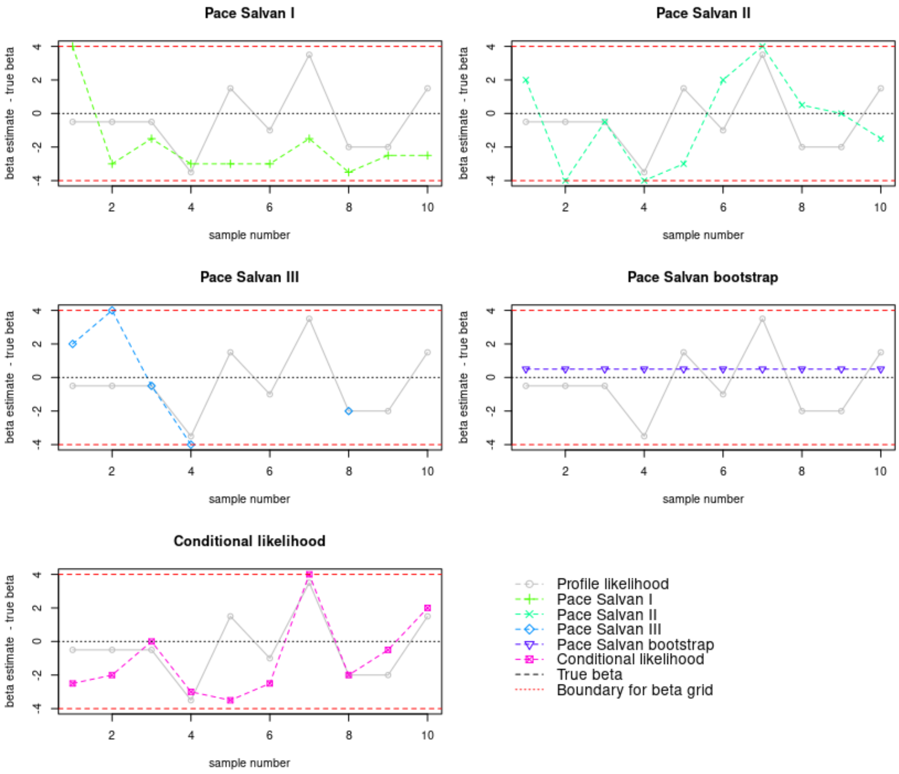
The bootstrap estimator $\ell_{BS}(\theta)$ on the other hand provides very good estimates for all of the samples and for all three values of β_0. Although for $\beta_0 = 0.2$ the estimator overestimates β_0 slightly in each sample. It seems that the influence of β_0 on the sampling probabilities is too small for the estimator to detect properly. For $\beta_0 = 1$ and $\beta_0 = 10$ the estimates are spot on for each sample, keeping in mind that the candidate values are a rough grid, and the exact true value will probably not be detected by the estimator when the researcher selecting the proposed β values does not know the true value, or when another maximization algorithm is used.

If we inspect the plotted log-likelihoods for a sample with $\beta_0 = 1$ in 4.4 we see that the profile likelihood and its corrections are very flat, except for some large dips for values of β much larger than β_0. The likelihood corrections $\ell^{I}_{AE}(\beta)$, $\ell^{II}_{AE}(\beta)$ and $\ell^{III}_{AE}(\beta)$, do not compute values for all of the proposed β values, and there does not seem to be a pattern to where it is not able to compute a value. In the simulations the estimated information matrices $j_{\xi\xi}(\hat{\theta})$ and $j_{\xi\xi}(\hat{\theta}_\beta)$ either have elements too large to be represented by the computer, resulting in `Inf` datatypes which pollute the estimates, or their computationally estimated determinant is negative which as mentioned is not good. With larger datasets this problem increased, and for 100 × 100 observations neither of the three likelihood corrections was able to compute values for any supplied β.

The flatness of the likelihoods, and the missing values for the corrected likelihoods, explains the high variability and low precision of their resulting estimates. The approximate conditional likelihood $\ell_{CL}(\beta)$ is highly erratic across the proposed β values, with many peaks that will confuse an optimization algorithm such as the Nelder Mead algorithm, or other methods that tends to get stuck on local maxima. The fact that the corrected likelihoods are not able to consistently compute values make these useless for optimization methods other than perhaps for grid search. The plot of $\ell_{BS}(\theta)$ on the other hand has a clear peak at the true value and has a smooth trajectory. It is therefore likely to produce good estimates for such maximization algorithms.

We also briefly study how the pseudo log-likelihoods perform on a larger dataset of 20 × 20 observations. A plot of estimation errors is shown in Figure

4.5, and plotted log transformed log-likelihoods are shown in Figure 4.6. The only notable difference here is that $\ell_{CL}(\beta)$ seems to perform slightly better than the profile likelihood estimator. The other pseudo log-likelihoods perform the same as they did in the smaller sample case.



Figur 4.1: Estimate errors $\hat{\beta} - \beta_0$ with $\beta_0 = 0.2$ for 10 samples with 10×10 observations.

4.4 Simulation study

We simulated 250 samples from the model with $\beta_0 = 1$ and for each sample we computed the pseudo log-likelihoods for a grid of β values and choose the maximizing β as an estimate of β_0. The grid of candidate estimates was chosen in the same manner as as in the preliminary study, resulting in a grid of 21 evenly spaced candidate values in the interval $[-29, 31]$.

Table 4.1 shows the square root of the mean of squared errors (RMSE) of the estimates. Table 4.1 also shows the number of samples where the pseudo log-likelihood fails to produce an estimate (NA). The results are in line with the preliminary study on smaller samples with $\beta_0 = 1$. The estimates from profile likelihood are on average far away from the true value, with a RMSE of 14.35, which is again as expected as the profile likelihood suffers from the incidental parameter problem. The estimates from the three likelihood corrections $\ell^{I}_{AE}(\beta)$, $\ell^{II}_{AE}(\beta)$ and $\ell^{III}_{AE}(\beta)$ has a larger RMSE than the profile likelihood, as we expected based on their performance on the preliminary study, and the plots of their corresponding pseudo log-likelihoods. The approximate conditional likelihood $\ell_{CL}(\beta)$ produces the worst estimates with a RMSE of 18.9, while the estimates from $\ell_{BS}(\theta)$ has a RSME of 0, which means β_0 is selected for each sample.

	RMSE	num. NA
Profile likelihood	14.35	0
Pace & Salvan I	16.14	0
Pace & Salvan II	15.56	3
Pace & Salvan III	14.95	133
Pace & Salvan Bootstrap	0.00	0
Approx. conditional likelihood	18.85	0

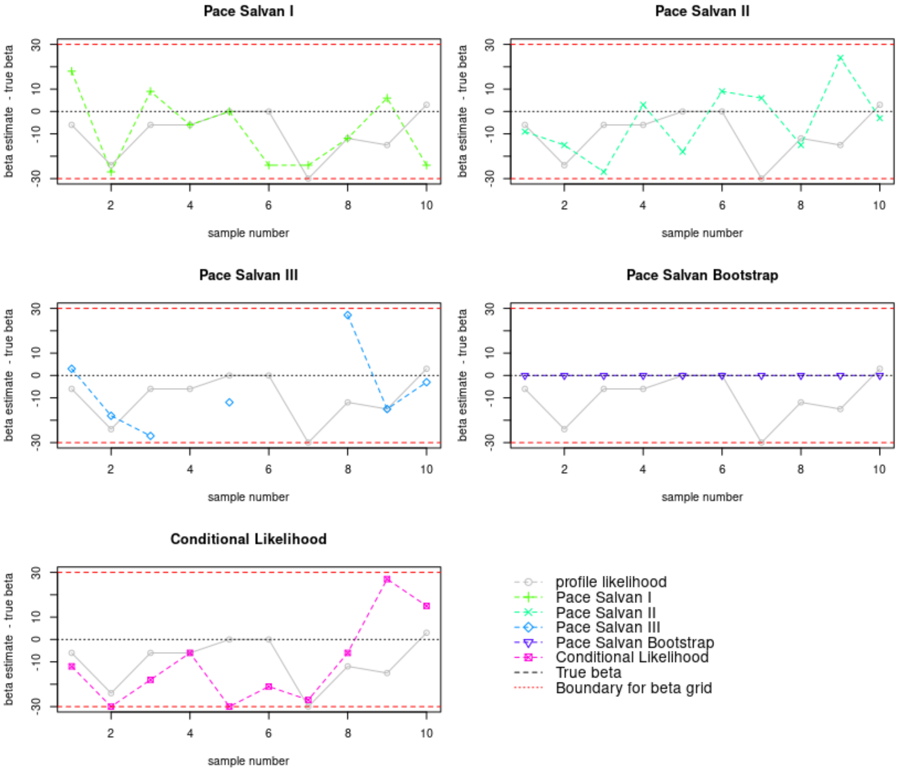
Tabell 4.1: Root mean square error for estimates of β based on 250 samples, with $\beta_0 = 1$.

4.5 Discussion

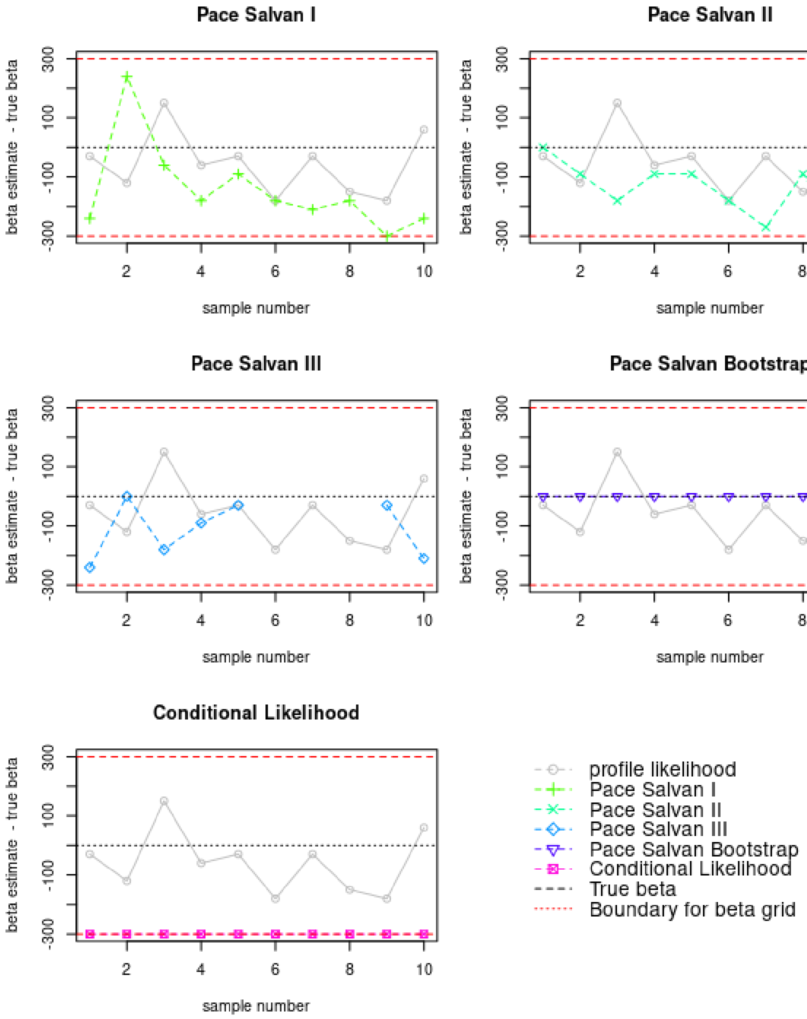
The bootstrap estimated corrected likelihood seems to completely solve the nuisance parameter problem for this model selecting the true parameter for each sample in the simulation study with $\beta_0 = 1$. In the preliminary study with $\beta_0 = 0.2$ the estimates from $\ell_{BS}(\theta)$ were biased upwards, indicating that when the parameter of interest is small relative to the incidental parameters the induced bias on estimates of the parameter of interest is more pronounced. Although the estimates are remarkably good, we must keep in mind that in our simulations we chose a grid of 21 candidate β that contains the true parameter β_0, with the two neighbouring candidates a distance of $\beta_0/21$ away. We do not know where in $[(1-1/21)\beta_0, (1+1/21)\beta_0]$ the bootstrap likelihood $\ell_{BS}(\theta)$ will have it's peak, but it is unlikely that the maximizing point will be exactly at β_0 for samples with 10×10 observations. A drawback of the bootstrap estimated corrected likelihood is that it is computationally intensive. In more realistic applications one will have data with more than 100 individuals, resulting in 100×100 observations requiring constrained maximum likelihood estimates of 200 incidental parameters for each bootstrap iteration.

The three corrected likelihoods $\ell^{I}_{AE}(\beta)$, $\ell^{II}_{AE}(\beta)$ and $\ell^{III}_{AE}(\beta)$ did not perform very well. There are several possible causes for their instability and faulty computations. They all rely on numerically estimated determinants and on multivariate optimization algorithms for determining maximum likelihood estimates, which may not give consistent results. There may also be rounding errors disturbing the computations. With larger datasets, for 100×100 observations or more neither of the three likelihood corrections was able to compute values for any supplied β, which means they can not be used for most applications.

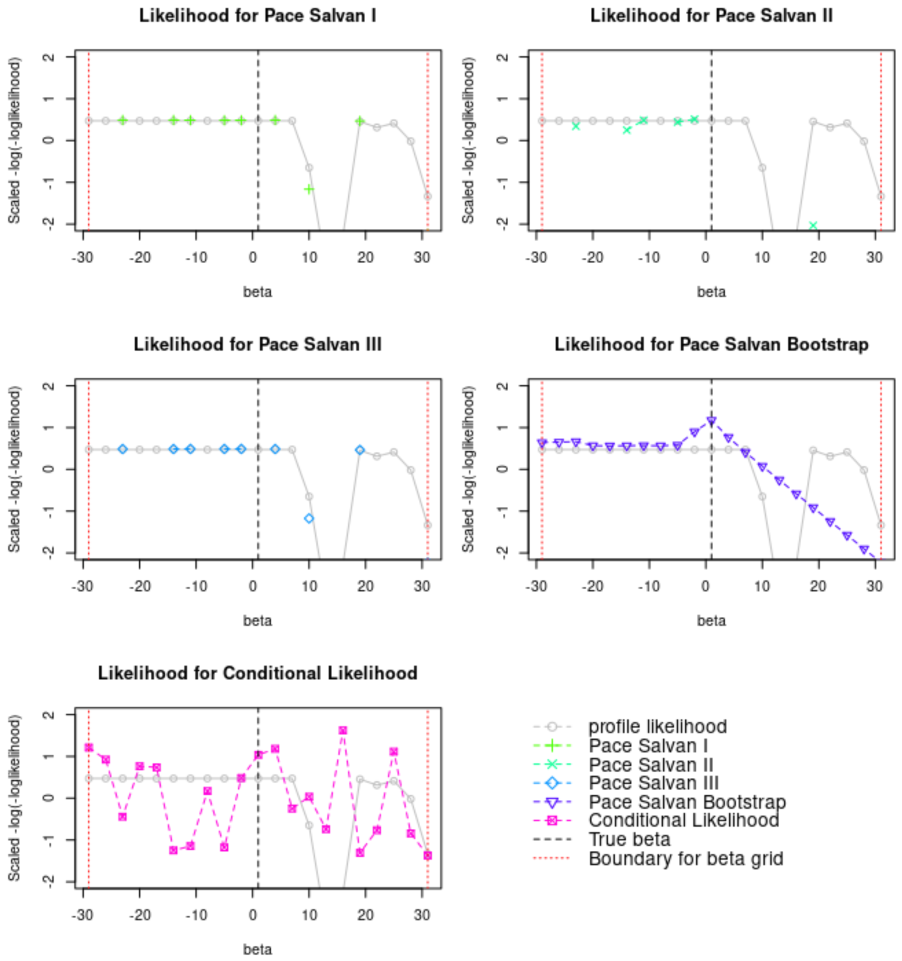
The approximate conditional likelihood $\ell_{CL}(\beta)$ had the worst performance in this simulation study. It may be that choosing an even larger precision parameter T would have eventually improved the estimates, but due to time constraints we were not able to test the method with a larger value for T.



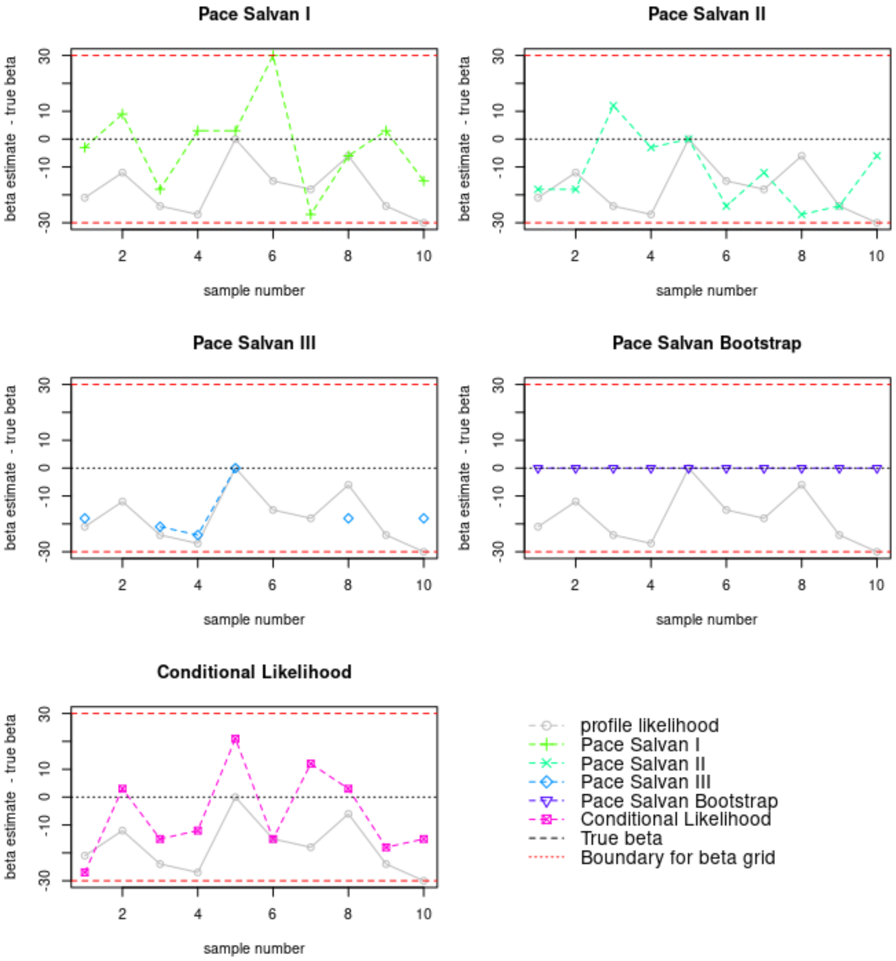
Figur 4.2: Estimate errors $\hat{\beta} - \beta_0$ with $\beta_0 = 1$ for 10 samples with 10×10 observations.



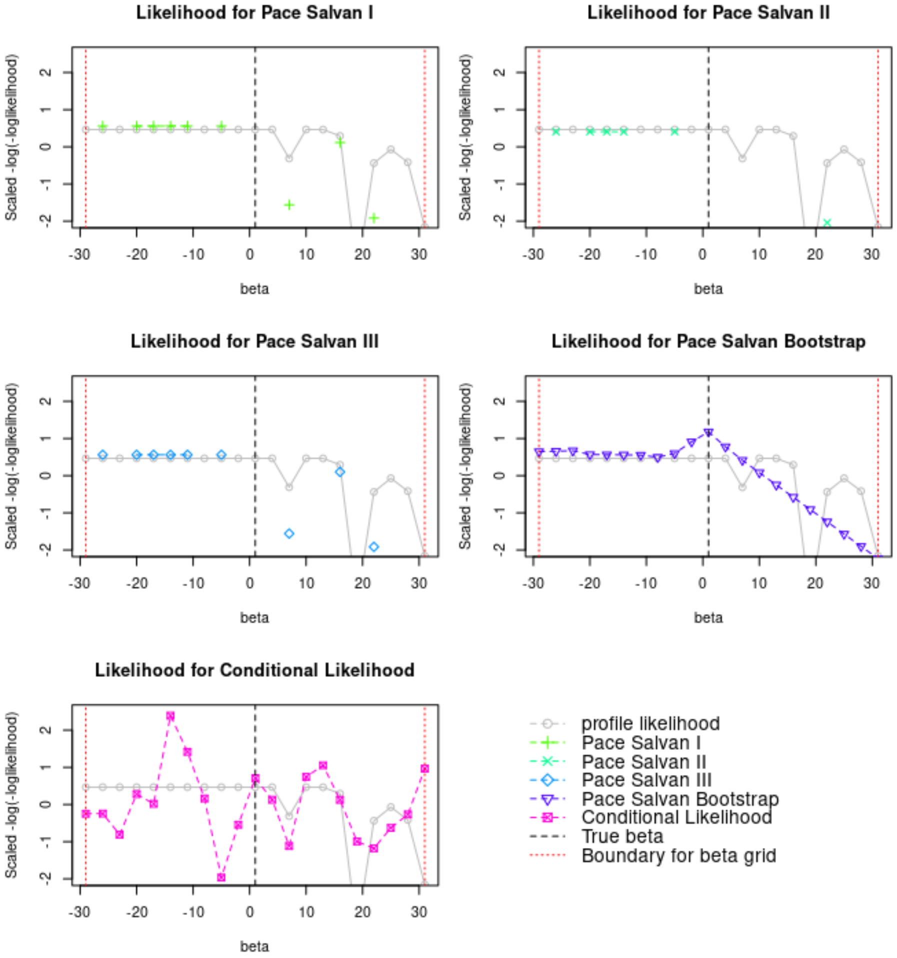
Figur 4.3: Estimate errors $\hat{\beta} - \beta_0$ with $\beta_0 = 10$ for 10 samples with 10×10 observations.



Figur 4.4: log transformed likelihoods with $\beta_0 = 1$ for one sample with 10×10 observations.



Figur 4.5: Estimate errors $\hat{\beta} - \beta_0$ with $\beta_0 = 1$ for 10 samples with 20×20 observations.



Figur 4.6: log transformed likelihoods with $\beta_0 = 1$ for one sample with 20×20 observations.

KAPITTEL 5

Conclusion

We have studied the incidental problem first introduced by Neyman og Scott (1948) in a model with Poisson distributed panel data with two fixed effects. We have considered several proposed solutions to the incidental parameter problem. The conditional likelihood derives the distribution of the data conditioned on a sufficient parameter, Charbonneau (2012) derived an approximate conditional likelihood for the model considered in this thesis. Corrected likelihoods seek to adjust the bias in the expected information of the profile likelihood or, from the perspective of Pace og Salvan (2006), to estimate a least favorable target likelihood.

We have derived the corrected profile likelihoods suggested by Pace og Salvan (2006) for the model considered. We have then compared these proposed solutions and the profile likelihood in a simulation study, by first inspecting the behaviour of the likelihoods in a preliminary study, and then by using the chosen pseudo likelihoods to estimate the parameter of interest on multiple samples and comparing their mean squared error.

The bootstrap estimate of corrected profile likelihood ℓ_{BS} by Pace og Salvan (2006) consistently and accurately estimated the parameter of interest in the simulated data as long as the parameter of interest was sufficiently large compared to the incidental parameters. When the parameter of interest has less influence on the sampling probabilities ℓ_{BS} had some bias, but still gave consistent estimates that vastly improved on the profile likelihood.

The corrected profile likelihoods that were based on estimated information matrices did not improve the bias of the profile likelihood and also suffered from computational issues, often struggling to produce values. With larger datasets closer to the size of real data we were unable to produce values for these likelihoods, which means they can not be used in most applications.

Our implementation of the approximate conditional likelihood ℓ_{CL} did not improve on the profile likelihood. The mean squared error of the estimates from the data in the simulation study were larger than the estimates from the profile likelihood, and a plot of the likelihood for one sample revealed an erratic likelihood with many local maxima, and not a smooth curve with a clear peak, as we would expect. Since ℓ_{CL} is an approximation of a conditional likelihood it is possible that we simply did not allocate enough time and computations to make the approximation sufficient. Still, with the same amount of computational power or time available, the bootstrap estimated corrected likelihood ℓ_{BS} gave better estimations and was also easier to implement.

For further work we would suggest studying the behaviour of the approximate conditional likelihood with more computations allocated to improve the approximation, in order to see if the method will then produce a more useful pseudo likelihood. It would be interesting to compare estimates from ℓ_{BS} with other current estimates on real world data. For example data on trade between countries. The gravity equations for trade between countries frequently use the Poisson distribution to model panel data with fixed effects, with the parameters of interest being country specific traits and relations between the countries.

We also briefly touched upon the integrated likelihoods approach to eliminating the incidental parameter problem, with details given in Appendix B. Integrated likelihoods have been shown to produce very accurate predictions and confidence intervals in problems with many nuisance parameters, and it would be interesting to see how an implementation of this approach compares to the pseudo likelihoods studied in this thesis.

Appendices

TILLEGG A

Big-Oh notation

Big-Oh notation is used to indicate the order by which a function $f(n)$ will grow as n increases.

A.1 Big-Oh notation

The statement

$$f(n) = O(g(n))$$

is taken to mean that there exists a positive real number M and a real number N such that

$$|f(n)| \leq Mg(n) \quad \forall\, n \geq N.$$

A.2 Big-Oh notation for stochastic boundedness

We let $O_p(g(n))$ denote big-Oh notation for stochastic boundedness, meaning if we write

$$f_n = O_p(g(n))$$

we are stating that for any $\epsilon > 0$ there exists finite numbers $M > 0$ and $N > 0$ such that

$$P(|f_n/g(n)| > M) < \epsilon, \quad \forall\, n > N.$$

TILLEGG B

Integrated likelihoods

Inspired by Bayesian statistics we will refer to the weight function as a prior, although it does not need to meet the requirements for a probability density function.

Advantages

As mentioned the profile likelihood ignores the uncertainty inherent in the estimation of the nuisance parameters λ. Integrated likelihoods addresses this problem by averaging over the possible values of λ (J. O. Berger mfl., 1999).

According to Severini (2010) it can be shown that integrated likelihood functions have the same type of optimality properties as the likelihood function in models without a nuisance parameter (Wald, 1950).

Use of integrated likelihoods in likelihood ratio statistics

Severini (2010) has studied the properties of likelihood ratio statistics based on integrated likelihoods. The statistic

$$\overline{R} = \operatorname{sgn}(\overline{\theta} - \theta)[2\{\overline{\ell}(\overline{\theta}) - \overline{\ell}(\theta)\}]^{1/2},$$

with $\overline{\ell}(\theta) = \log(\overline{L})$ and $\overline{\theta}$ the maximizer of $\overline{\ell}(\theta)$, is asymptotically normally distributed to the second order, and under certain conditions close to the standard normal distribution. If the information available for θ is large we can satisfy these conditions by using what Severini terms the zero-score-expectation parametrization, $\phi \equiv \phi(\theta, \lambda; \hat{\theta})$, of the nuisance parameters and choosing a prior $\pi(\phi \mid \theta)$ that, in general terms, does not depend strongly on θ (see Appendix B on the following page).

Severini (2010) recommends integrated likelihood ratios for their well behaved nature. They have been shown to produce finite confidence sets where the profile likelihood will (unreasonably) produce infinite sets (Ghosh mfl., 2006) and to produce a single confidence interval from a unimodal integrated likelihood, where the profile likelihood is bimodal with two resulting confidence intervals (Malley mfl., 2003).

Selection of nuisance parameterization and prior

Severini (2007) shows in his paper that in order for an integrated likelihood to be a useful pseudolikelihood function we should choose a parameterization of the nuisance parameter γ that is unrelated to θ and choose a prior $\pi(\gamma \mid \theta)$ that does not depend on θ.

By unrelated Severini means that $\hat{\gamma}_\theta$, the maximum likelihood estimator of γ for fixed θ, is approximately constant as a function of θ. The definition of unrelated, and the arguments leading to these criteria will be presented after we introduce Severini's suggested method for finding such a parameterization.

The zero-score-expectation parameter

Severini (2007) defines the zero-score-expectation parameter, $\phi \equiv \phi(\theta, \lambda\,; \hat{\theta})$, as the data dependent parameter that solves the implicit equation

$$\mathrm{E}\{\ell_\lambda(\theta,\lambda) \mid \hat{\theta}, \phi\} \equiv \mathrm{E}\{\ell_\lambda(\theta,\lambda) \mid \theta_0, \lambda_0\}\Big|_{(\theta_0,\lambda_0)=(\hat{\theta},\phi)}$$

This parameter depends on the data, and in a Bayesian setting this would be a problem as a parameter of a probability distribution can not be data-dependent, but it does not cause problems in a likelihood function where the data are considered fixed.

Severini (2007) shows that ϕ is strongly unrelated to θ, and that in models where $\mathrm{E}\{\ell(\theta,\lambda) \mid \hat{\theta}, \hat{\lambda}\} = \ell(\theta,\lambda)$, such as for full-rank exponential family models with loglikelihood of the form $\ell(\theta,\lambda) = c(\theta,\lambda)^T x - d(\theta,\lambda)$, the stronger property $\hat{\phi}_\theta = \hat{\phi}$ holds for all θ.

Let $\mathcal{L}^*(\theta,\phi)$ denote the likelihood function in terms of (θ,ϕ). Then the integrated likelihood for θ with respect to a density $\pi(\phi)$ for ϕ is given by

$$\overline{L}(\theta) = \int \mathcal{L}^*(\theta,\phi)\pi(\phi)d\phi.$$

By showing that this integrated likelihood is an approximation to the modified likelihood proposed by Barndorff-Nielsen (1983), Severini shows that this integrated likelihood will be score unbiased and information unbiased to order $O(n^{-1})$.

Definition of unrelated

To define relatedness, we first need a definition of deviations, which in broad measures describes how far a parameter is from its maximum likelihood value. Severini defines moderate deviations of θ as $\theta = \hat{\theta} + O(n^{-1/2})$ and large deviations of θ as θ as $\theta = \hat{\theta} + O(1)$.

Severini then defines relatedness in the following way. A parameter γ is weakly unrelated to θ if $\hat{\gamma}_\theta = \hat{\gamma} + O(n^{-1})$ for moderate deviations of θ. An implication of this is that if γ and θ are orthogonal parameters, then γ is *weakly unrelated* to θ (Severini, 2007 suggests Cox og Barndorff-Nielsen, 1994 and Pace og Salvan, 1997 for examples).

Severini also defines γ to be *strongly unrelated* to θ if $\hat{\gamma}_\theta = \hat{\gamma} + O(n^{-1/2})$ for large deviations of of θ. As noted by Severini, an orthogonal nuisance parameter is not in general strongly unrelated to θ.

Selection of a prior for the nuisance parametes

In the following, we list Severini's arguments for why, in order to construct an integrated likelihood function that is useful for non-Bayesian inference, we should construct a nuisance parameter ϕ that is strongly unrelated to θ and then choose a prior density for ϕ that does not depend on θ.

First, if there exists a nuisance parameter γ such that the likelihood factors as

$$L_1(\theta)L_2(\gamma)$$

then $L_1(\theta)$ can be used as a likelihood for θ. We would therefore like an integrated likelihood to correspond to L_1 in this case. If $\pi(\lambda \mid \theta)$ is such that θ and γ are independent then

$$\int_\Lambda L_1(\theta)L_2(\gamma)\pi(\lambda \mid \theta)d\lambda = L_1(\theta)\int_\Lambda L_2(\gamma)\pi(\lambda \mid \theta)d\lambda \propto L_1(\theta)$$

In other words the integrated likelihood will correspond to L_1 provided that $\pi(\lambda \mid \theta)$ is such that θ and γ are independent. Since γ does not depend on θ, as seen from the factorization of the likelihood, this property suggests that, for non-Bayesian inference about θ, $\pi(\theta \mid \lambda)$ should be chosen so that unrelated parameters are independent.

As mentioned, two important frequentist properties of a genuine likelihood are score unbiasedness and information unbiasedness. In general, $\mathrm{E}(\bar{\ell}_\theta(\theta) \mid \lambda)$ and $\mathrm{E}(\bar{\ell}_{\theta\theta}(\theta) + \bar{\ell}_\theta(\theta)\bar{\ell}_\theta(\theta)^T \mid \lambda)$ are both $O(1)$ as $n \to \infty$ (Severini, 1998b). But if the model is parameterized by a nuisance parameter γ that is weakly unrelated to θ and $\pi(\gamma \mid \theta)$ does not depend on θ, then $\mathrm{E}(\ell_\theta(\theta) \mid \lambda) = O(n^{-1})$ (Severini, 1998b).

If γ is strongly unrelated to θ and $\pi(\gamma \mid \theta)$ does not depend on θ, then $\mathrm{E}(\bar{\ell}_{\theta\theta}(\theta) + \bar{\ell}_\theta(\theta)\bar{\ell}_\theta(\theta)^T \mid \lambda)$ is also $O(n^{-1})$ (Severini, 1998b).

Again, this analysis suggests that $\pi(\gamma \mid \theta)$ should be chosen so that, if a nuisance parameter γ is strongly unrelated to θ, then θ and γ are independent under $\pi(\gamma \mid \theta)$.

Lastly Severini argues that we will not want the integrated likelihood to be sensitive to the choice of prior, as the choice of prior is somewhat arbitrary. For instance, in case where $\mathcal{L}(\theta, \lambda) = L_1(\theta)L_2(\gamma)$ any prior density $\pi(\lambda \mid \theta)$ under which θ and γ are independent yields the same integrated likelihood. But what about the case when the likelihood does not factor neatly? Severini shows that if the model is parameterized in terms of a nuisance parameter γ that is weakly unrelated to θ and $\pi(\gamma \mid \theta)$ does not depend on θ, then, for θ in the moderate deviation range, $\bar{L}(\theta)$ does not depend on the form of $\pi(\gamma)$, if terms of order n^{-1} are ignored, and if γ is strongly unrelated to θ then in addition, for θ in the large deviation range, $\bar{L}(\theta)$ does not depend on the form of $\pi(\gamma)$ if terms of order $n^{-1/2}$ are ignored. Which again means that if we want an integrated likelihood function to not depend heavily on incidental features of the prior, the prior should be chosen so that parameters that are strongly unrelated are independent.

TILLEGG C

Code

- The functions used to simulate the data and find maximum likelihood estimates is presented in Appendix C.1.
- The functions used to compute the approximate conditional likelihood in Equation (3.2) is presented in Appendix C.2 on page 39.
- The functions used to compute the Bootstrap estimated corrected likelihood from section 2.3.2 is presented in Appendix C.3 on page 40.
- The functions used to compute the profile likelihood and the corrected likelihoods $\ell^{I}_{AE}(\beta)$, $\ell^{II}_{AE}(\beta)$ and $\ell^{III}_{AE}(\beta)$ from Equations (3.8), (3.9) and (3.10) is presented in Appendix C.4 on page 41.

C.1 Sampling data, and maximum likelihood estimation

```
sample_x_and_parameters <- function(beta, m, n) {
  # x <- array(runif(m*n, min = 0, max = 1), c(m,n))
  # TODO temp
  x <- array(1, c(m,n))
  # mu <- runif(m, min = 0, max = a)
  # alpha <- runif(n, min = 0, max = a)

  mu <- rep(1, m)
  alpha <- rep(1, n)

  return( list(x=x, mu=mu, alpha=alpha) )
}

sample_y <- function(beta, x, m, n, mu, alpha) {
  theta <- outer(mu, alpha, FUN = "+")
  eta <- x*beta + theta
  y <- matrix(rpois(n = n*m, lambda = exp(eta)),
              nrow = m)
  return(y)
}

negative_loglikelihood_beta <- function(params, beta, m, n, x, y, mu_1)
```

```r
{
  mu <- c(mu_1, params[1:(m-1)])
  alpha <- params[m:(n+m-1)]
  eta <- outer(mu, alpha,
               FUN = function(mu, alpha)
               {
                 (x*beta + mu + alpha)
               }
  )
  lambda <- exp(eta)
  return(
    - sum(y*eta) + sum(lambda)
  )
}

negative_loglikelihood_ur <- function(params, m, n, x, y, mu_1)
  #' Negative loglikelihood for unrestricted maximum likelihood
  #' estimation of beta mu and alpha
{
  mu <- c(mu_1, params[1:(m-1)])
  alpha <- params[m:(n+m-1)]
  beta <- params[n+m]
  eta <- outer(mu, alpha,
                  FUN = function(mu, alpha)
                  {
                    (x*beta + alpha + mu)
                  }
  )
  return(
    - sum(y*eta) + sum(exp(eta))
  )
}

# TODO: rename
theta_mle_given_beta <- function(beta, x, y, m, n, mu_1) {
  #' Restricted maximum likelihood estimation given beta

  optim_out <- optim(fn = negative_loglikelihood_beta, par = rep(0.5, n+m-1)
                     , beta = beta
                     , m = m
                     , n = n
                     , x = x
                     , y = y
                     , mu_1 = mu_1
                     , method = "BFGS"
  )
  optim_estimate <- optim_out$par
  mu_hat <- optim_estimate[1:(m-1)]
  alpha_hat <- optim_estimate[m:(n+m-1)]

  return(list(
    optim_estimate = optim_estimate,
    mu_hat = mu_hat,
    alpha_hat = alpha_hat
  )
  )
}

theta_mle_ur <- function(x, y, m, n, mu_1) {
  #' Unrestricted maximum likelihood estimates of all parameters

  optim_out <- optim(fn = negative_loglikelihood_ur, par = rep(0.5, n+m)
                     , method = "BFGS"
                     , m = m
                     , n = n
                     , x = x
                     , y = y
                     , mu_1 = mu_1
  )
  optim_estimate <- optim_out$par
  mu_hat <- optim_estimate[1:(m-1)]
  alpha_hat <- optim_estimate[m:(n+m-1)]
  beta_hat <- optim_estimate[m+n]

  return(list(
    optim_estimate = optim_estimate,
    mu_hat = mu_hat, # Does not include mu_1 = 0
    alpha_hat = alpha_hat,
    beta_hat = beta_hat
  )
  )
}
```

C.2 Approximate conditional likelihood

```
negative_conditional_loglikelihood <- function(beta, x, y, m, n, nt) {
  #' Computes an estimate of the negative of the conditional log-likelihood
  #' from Charboonneu (2012) "Multiple Fixed Effects in Nonlinear Panel Data Models".
  #'
  #' nt = number of alternative y's to compare to. Increases computations by O(n*m)
  #'
  cond_loglik <- 0
  for (t in 1:nt) {
    for (i in 1:m) {
      for (j in 1:n) {
        # Choose a submatrix from y and x
        l <- sample(1:m, 1)
        k <- sample(1:n, 1)
        y_sub <- array(c(y[i,j], y[i,k], y[l,j], y[l,k]), c(2,2))
        x_sub <- array(c(x[i,j], x[i,k], x[l,j], x[l,k]), c(2,2))

        # Construct alternative y matrix, with same row- and column sums
        y_alt <- array(rep(NA, 4), c(2,2))
        mar_sums <- c(colSums(y_sub), rowSums(y_sub))

        # First value must allow for row and column sums to be the same as in y_ijlk
        y_alt[1] <- sample(
          max(0, (y_sub[1,1]-y_sub[2,2])):min(mar_sums[1], mar_sums[3]),
          1)
        y_alt[2,1] <- mar_sums[1] - y_alt[1,1]
        y_alt[1,2] <- mar_sums[3] - y_alt[1,1]
        y_alt[2,2] <- mar_sums[2] - y_alt[1,2]

        cond_log_ijt <- log(1 + prod(factorial(y_sub)/factorial(y_alt))
                                      * exp(beta*sum(x_sub*(y_alt-y_sub))))

        # Add term if not NA
        if (!is.na(cond_log_ijt))
          cond_loglik <- cond_loglik + cond_log_ijt
      }
    }
  }
  # Return negative log-likelihood
  return(cond_loglik)
}
negative_conditional_loglikelihood <- Vectorize(negative_conditional_loglikelihood
                                                , vectorize.args = "beta")
```

C.3 Bootstrap estimated corrected likelihood

```
bootstrap_profile_loglikelihood <- function(beta, x, y, m, n,
                                            beta_ur, mu_ur, alpha_ur, R, mu_1) {
  #'
  #' Returns bootstrapped corrected profile likelihood given beta
  #' R (int): number of bootstrap iterations
  #'

  Rinv = 1/R
  bs_corrected_loglik <- 0
  theta <- outer(c(mu_1, mu_ur), alpha_ur, FUN = "+")
  lda <- exp(x*beta + theta)
  for (r in 1:R) {
    y_bs <- matrix(rpois(n = n*m, lambda = lda), nrow = m)



    optim_estimate <- theta_mle_given_beta(beta=beta, x=x, y=y_bs, m=m, n=n, mu_1)
    mu_hat_beta    <- optim_estimate$mu_hat
    alpha_hat_beta <- optim_estimate$alpha_hat

    eta_beta = outer(c(mu_1, mu_hat_beta),
                     alpha_hat_beta,
                     FUN = function(mu, alpha)
                     {
                       (x*beta + mu + alpha)
                     }
    )
    # Add current bootstrap estimate to the mean estimator
    bs_corrected_loglik <- bs_corrected_loglik +
      Rinv*sum( y*eta_beta - exp(eta_beta) - lfactorial(y) )
  }

  return(bs_corrected_loglik)
}
bootstrap_profile_loglikelihood <- Vectorize(bootstrap_profile_loglikelihood,
                                             vectorize.args = "beta")
```

C.4 Corrected likelihoods

```
corrected_profile_loglikelihood <- function(beta, x, y, m, n,
                                            beta_ur, mu_ur, alpha_ur, mu_1) {
  #' Returns corrected profile likelihood given beta
  #'

  optim_estimate <- theta_mle_given_beta(beta=beta, x=x, y=y, m=m, n=n, mu_1=mu_1)
  # Estimates given beta
  mu_hat_beta = optim_estimate$mu_hat
  alpha_hat_beta = optim_estimate$alpha_hat
  num_fisher_info = optim_estimate$num_fisher_info

  # Full linear operatar (contains mu_1 = 0)
  eta_beta_full = outer(c(mu_1, mu_hat_beta),
                        alpha_hat_beta,
                        FUN = function(mu, alpha)
                        {
                          (x*beta + mu + alpha)
                        }
  )
  lambda_beta_full <- exp(eta_beta_full)

  # Remove the mu_1 row, as we are not estimating mu_1
  lambda_beta <- lambda_beta_full[2:m, ]

  observed_info_aa <- diag(rowSums(lambda_beta))
  observed_info_gg <- diag(colSums(lambda_beta))
  observed_info_ag <- lambda_beta
  j_xx_beta <- rbind(cbind(   observed_info_aa,  observed_info_ag ),
               cbind( t(observed_info_ag), observed_info_gg ))

det_j_xx_beta <- det(j_xx_beta)


# Unrestricted ml estimate of covariance matrix
lambda_ur <- outer(c(mu_1, mu_ur),
                   alpha_ur,
                   FUN = function(mu, alpha)
                   {
                     exp(x*beta_ur + mu + alpha)
                   }
)

lambda_ur <- lambda_ur[2:m,]

Cov_aa <- diag(rowSums(lambda_ur))
Cov_gg <- diag(colSums(lambda_ur))
Cov_ag <- lambda_ur
j_xx_beta <- rbind(cbind(   Cov_aa,  Cov_ag ),
               cbind( t(Cov_ag), Cov_gg ))

det_jxx_beta <- det(j_xx_beta)

# Profile likelihood
profile_loglik <- sum( y*eta_beta_full - lambda_beta_full - lfactorial(y)  )


# Pace & Salvan
Pace_Salvan_I <- profile_loglik - log(det_j_xx_beta)
Pace_Salvan_II  <- NA
Pace_Salvan_III <- NA

# Pace & Salvan - correction II
  # We will implement:
    # M_11[i,i'] = sum_j( lambda_diff_ij ) * sum_j( lambda_diff_i'j )
    # M_22[i,i'] = sum_j( lambda_diff_ji ) * sum_j( lambda_diff_ji' )
    # M_12[i,i'] = sum_j( lambda_diff_ij ) * sum_j( lambda_diff_ji' )
    # M_21[i,i'] = sum_j( lambda_diff_i'j ) * sum_j( lambda_diff_ji )
lambda_diff <- lambda_ur - lambda_beta
diffsum_ij <- rowSums(lambda_diff)  # dd[i] = sum_j( lambda_diff_ij )
diffsum_ji <- colSums(lambda_diff)  # dd[i] = sum_j( lambda_diff_ji )
M_11 <- outer(diffsum_ij, diffsum_ij, "*")
M_22 <- outer(diffsum_ji, diffsum_ji, "*")
M_12 <- outer(diffsum_ij, diffsum_ji, "*")
M_21 <- t(M_12)

M <- rbind(cbind( M_11, M_12 ),
           cbind( M_21, M_22 ))
j_xx_M_beta <- j_xx_beta + M
```

```
  det_j_xx_M_beta <- det(j_xx_M_beta)
  Pace_Salvan_II <- profile_loglik + 0.5*log(det_j_xx_beta) - 0.5*log(det_j_xx_M_beta)

  Pace_Salvan_III <- profile_loglik + 0.5*log(det_j_xx_beta) - log(det_jxx_beta)

  cox_approximate_mod_lik <- profile_loglik - 0.5*log(det_j_xx_beta)

  likelihood_esimates <-     c( profile_loglik,
                                cox_approximate_mod_lik,
                                Pace_Salvan_I,
                                Pace_Salvan_II,
                                Pace_Salvan_III
  )
  # Set infinity estemates to NA
  likelihood_esimates[is.infinite(likelihood_esimates)] <- NA

  return(
    likelihood_esimates
  )
}
corrected_profile_loglikelihood <- Vectorize(corrected_profile_loglikelihood,
                                             vectorize.args = "beta")
```

Bibliografi

Abowd, J. M., Kramarz, F. & Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, årg. 67nr. 2, 251–333.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society: Series B (Methodological)*, årg. 32nr. 2, 283–301.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society: Series B (Methodological)*, årg. 34nr. 1, 42–54.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. Biometrika, 67, 293-310. *Mathematical Reviews (MathSciNet): MR581727 Zentralblatt MATH*, årg. 434.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, årg. 70nr. 2, 343–365.

Berger, J. O., Liseo, B., Wolpert, R. L. Med flere. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical science*, årg. 14nr. 1, 1–28.

Casella, G. & Berger, R. L. (2002). *Statistical inference*. Thomson Learning.

Charbonneau, K. (2012). Multiple fixed effects in nonlinear panel data models: theory and evidence. *Princeton University*.

Cox, D. R. & Barndorff-Nielsen, O. E. (1994). *Inference and asymptotics* (Bd. 52). CRC Press.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, årg. 49nr. 1, 1–18.

De Bin, R., Sartori, N. [Nicola], Severini, T. A. Med flere. (2015). Integrated likelihoods in models with stratum nuisance parameters. *Electronic Journal of Statistics*, årg. 9nr. 1, 1474–1491.

Diciccio, T. J. & Martin, M. A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, årg. 55nr. 1, 305–316.

Diciccio, T. J. & Stern, S. E. (1994). Constructing approximately standard normal pivots from signed roots of adjusted likelihood ratio statistics. *Scandinavian journal of statistics*, 447–460.

DiCiccio, T. J., Stern, S. E., Martin, M. A. & Young, G. A. (1996). Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society: Series B (Methodological)*, årg. 58nr. 1, 189–203.

Ghosh, M., Datta, G. S., Kim, D. & Sweeting, T. J. (2006). Likelihood-based inference for the ratios of regression coefficients in linear models. *Annals of the Institute of Statistical Mathematics*, årg. 58nr. 3, 457–473.

Hausman, J. A., Hall, B. H. & Griliches, Z. (1984). *Econometric models for count data with an application to the patents-R&D relationship* (tekn. rapp.). national bureau of economic research.

Helpman, E., Melitz, M. & Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The quarterly journal of economics*, årg. 123nr. 2, 441–487.

Huber, P. J. Med flere. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, I *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. University of California Press.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of econometrics*, årg. 95nr. 2, 391–413. https://doi.org/https://doi.org/10.1016/S0304-4076(99)00044-5

Lancaster, T. (2002). Orthogonal Parameters and Panel Data. *The Review of Economic Studies*, årg. 69nr. 3, 647–666. http://www.jstor.org/stable/1556713

Malley, J. D., Redner, R. A., Severini, T. A., Badner, J. A., Pajevic, S. & Bailey-Wilson, J. E. (2003). Estimation of linkage and association from allele transmission data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, årg. 45nr. 3, 349–366.

McCullagh, P. & Tibshirani, R. (1990). A Simple Method for the Adjustment of Profile Likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, årg. 52nr. 2, 325–344. http://www.jstor.org/stable/2345439

Neyman, J. & Scott, E. L. (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, årg. 16nr. 1, 1–32. http://www.jstor.org/stable/1914288

Pace, L. & Salvan, A. (1997). *Principles of statistical inference: from a Neo-Fisherian perspective* (Bd. 4). World scientific.

Pace, L. & Salvan, A. (2006). Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference*, årg. 136nr. 10, 3554–3564.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rivkin, S. G., Hanushek, E. A. & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, årg. 73nr. 2, 417–458.

Sartori, N., Bellio, R., Salvan, A. & Pace, L. (1999). Miscellanea. The directed modified profile likelihood in models with many nuisance parameters. *Biometrika*, årg. 86nr. 3, 735–742.

Sartori, N. [Nicola], Salvan, A. & Pace, L. (2003). A note on directed adjusted profile likelihoods. *Journal of statistical planning and inference*, årg. 110nr. 1-2, 1–9.

Severini, T. A. (1998a). An approximation to the modified profile likelihood function. *Biometrika*, årg. 85nr. 2, 403–411.

Severini, T. A. (1998b). Likelihood functions for the elimination of nuisance parameters. *Biometrika*, årg. 85, 507–522.

Severini, T. A. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika*, årg. 94nr. 3, 529–542.

Severini, T. A. (2010). Likelihood ratio statistics based on an integrated likelihood. *Biometrika*, årg. 97nr. 2, 481–496.

Silva, J. S. & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, årg. 88nr. 4, 641–658.

Sweeting, T. J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika*, årg. 82nr. 1, 1–23.

Wald, A. (1950). Statistical decision functions.

Aaronson, D., Barrow, L. & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of labor Economics*, årg. 25nr. 1, 95–135.