Simulation based bias correction methods for complex problems

DUPUIS LOZERON, Elise

Abstract
Nowadays, the increase in data size and model complexity has led to increasingly difficult estimation problems. The numerical aspects of the estimation procedure can indeed be very challenging. To solve these estimation problems, approximate methods such as pseudo-likelihood functions or approximated estimating equations can be used as these methods are typically easier to implement numerically although they can lead to inconsistent and/or biased estimators. In this thesis, we propose a unified framework to compare four existing bias reduction estimators, two of them are based on indirect inference and two are based of bootstrap. We derive the asymptotic and finite sample properties of these bias correction methods. We demonstrate the equivalence between one version of the indirect inference and the iterative bootstrap which both correct sample biases up to the order $n^{-3}$. Therefore, our results provide different tools to correct the asymptotic as well as finite sample biases of estimators and give insight as to which method should be applied according to the problem at hand. We then apply these bias reduction [...]
SIMULATION BASED BIAS CORRECTION METHODS FOR COMPLEX PROBLEMS

by

Elise Dupuis Lozeron

A thesis submitted to the
Geneva School of Economics and Management,
University of Geneva, Switzerland,
in fulfillment of the requirements for the degree of
PhD in Statistics

Members of the thesis committee:
Prof. Maria-Pia Victoria-Feser, Co-Adviser, University of Geneva
Prof. Stéphane Guerrier, Co-Adviser, University of Illinois
Prof. Eva Cantoni, Chair, University of Geneva
Prof. Yanyuan Ma, University of South-Carolina

Thesis No. 17
November 2015
Remerciements

En premier lieu, j’aimerais remercier mes deux co-directeurs de thèse, le prof. Maria-Pia Victoria-Feser et le prof. Stéphane Guerrier. Leur aide précieuse et leurs encouragements tout au long de cette thèse m’ont permis de mener à bien et de finaliser ce travail de doctorat. Leurs conseils et leur bienveillance ont représentés une aide inestimable.

Je voudrais aussi remercier les autres membres de mon jury, le Prof. Eva Cantoni et le Prof. Yanyuan Ma, pour leur lecture attentive et leurs questions et commentaires pertinents.

J’aimerais ensuite remercier tous mes collègues de l’Université de Genève. Tout d’abord, les membres, anciens et actuels, du groupe de Méthodologie et d’Analyse de Données de la Faculté de Psychologie, à savoir le Prof. Olivier Renaud pour les discussions statistiques très enrichissantes que j’ai eu avec lui, le Dr. Katia Iglesias pour son amitié et sa grand générosité, ainsi que le Dr. Guillaume Fürst et Victorin Luisier pour les très bons moments passés ensemble. Je remercie également mes collègues et amis de la Faculté d’Économie et de Management: Roberto Molinari pour tout son soutien, Samuel Orso, Rose Hiquet, le Dr. Marco-Olivier Boldi et le Dr. Kustrim Reka, avec lesquels j’ai partagés de très grands moments et de nombreux fous rires, ainsi que Marco Avella, Daniel Flores, le Dr. Stéphane Rothen et le Dr. Dominique Couturier.

Je tiens également à remercier mes collègues et amis des Hôpitaux Universitaires de Genève et du Centre Hospitalier Universitaire Vaudois, et en particulier le Dr. Dan Adler pour son soutien indéfectible, le Prof. Pierre-Olivier Bridexau qui m’a accueilli dans le service de pneumologie des HUG, Sandrine Rudaz pour son amitié précieuse et le Dr. Angèle Gayet-Ageron.

Mes parents, ma soeur Camille et ma famille m’ont beaucoup aidé par le soutien qu’ils m’ont apporté depuis toujours. Quant à ma fille Lucie, sa présence près de moi, dans ma vie quotidienne, est un grand bonheur. Je la remercie en particulier pour sa patience envers mes indisponibilités à la fin de cette thèse. Je voudrais aussi remercier Jessica Lozeron pour sa grande gentillesse. Enfin, je remercie ma très chère amie Catrin Schulte-Hillen pour son soutien sans faille, et son immense disponibilité.
Abstract

Nowadays, the increase in data size and model complexity has led to increasingly difficult estimation problems. The numerical aspects of the estimation procedure can indeed be very challenging. To solve these estimation problems, approximate methods such as pseudo-likelihood functions or approximated estimating equations can be used as these methods are typically easier to implement numerically although they can lead to inconsistent and/or biased estimators.

In this thesis, we propose a unified framework to compare four existing bias reduction estimators, two of them are based on indirect inference and two are based on bootstrap. We derive the asymptotic and finite sample properties of these bias correction methods. We demonstrate the equivalence between one version of the indirect inference and the iterative bootstrap which both correct sample biases up to the order $n^{-3}$. Therefore, our results provide different tools to correct the asymptotic as well as finite sample biases of estimators and give insight as to which method should be applied according to the problem at hand.

We then apply these bias reduction techniques to robust estimation of income distributions. We used a very simple starting estimator which is known to be robust but not consistent and correct its bias with indirect inference. This is a very general way to construct robust estimators for complex models. A second illustration is provided by the estimation of Generalized Linear Latent Variable Models. We were able to compute unbiased estimates for these very complex models that have a large number of parameters without employing numerical integration techniques. As a by-product, bias reduction techniques allow to compute a goodness-of-fit test statistic for latent variable models.
RÉSUMÉ

De nos jours, l’augmentation de la taille des données récoltées et de la complexité des modèles utilisés pour les analyser donnent lieu à des problèmes d’estimation de plus en plus difficiles à résoudre. Pour ce faire, des techniques numériques de plus en plus sophistiquées sont mises en œuvre. Une solution consiste également à utiliser des méthodes d’approximation numériques comme les méthodes de pseudo-vraisemblance et les équations d’estimation approximées. Cependant, ces méthodes d’approximation, bien qu’elles facilitent les aspects numériques, peuvent résulter en des estimateurs biaisés et/ou non convergents.

Dans cette thèse, nous comparons quatre méthodes de réduction du biais, deux d’entre elles sont basées sur l’inference indirecte et deux sur le bootstrap. Les propriétés asymptotiques et en échantillon fini des estimateurs ainsi obtenus sont ensuite dérivées. Nous démontrons l’équivalence entre une version de l’inference indirecte et la version itérative du bootstrap qui toutes deux permettent d’obtenir des estimateurs dont le biais est d’ordre $n^{-3}$. Les résultats obtenus dans cette thèse mettent en évidence que ces méthodes ne corrigent pas toutes le même type de biais et permettent ainsi de savoir quelle méthode utiliser suivant le problème auquel l’on est confronté.

Nous avons ensuite appliqué ces techniques de réduction du biais à la modélisation robuste des distributions de revenus. Dans un premier temps un estimateur biaisé mais robuste et simple à calculer est obtenu et, dans un deuxième temps, l’inference indirecte est utilisée pour corriger son biais. Cette procédure est une façon simple d’obtenir un estimateur robuste et de plus est applicable pour de nombreux modèles. Nous nous sommes ensuite intéressés à l’estimation des modèles linéaires généralisés à variables latentes. En utilisant les méthodes étudiées dans cette thèse nous avons pu estimer les paramètres de ces modèles complexes sans utiliser de méthode d’intégration numérique. De plus, ces méthodes permettent de calculer un test d’adéquation pour les modèles linéaires généralisés à variables latentes.
CONTENTS

Remerciements i

Abstract iii

Résumé v

List of Abbreviations 1

1 Introduction 3
  1.1 Indirect Inference 5
  1.2 Bias correction methods via Bootstrap 6

2 Statistical Properties of Simulation Based methods 9
  2.1 Mathematical setup 9
  2.2 Statistical Properties of Simulation Based Methods for Bias Reduction 13
  2.3 A simple example 19
  2.4 Asymptotic normality 21

3 Robust Estimation of Income Distribution 23
  3.1 Income distribution and inequality index 23
  3.2 Robust estimation for income distribution 24
  3.3 Simulation study 26
  3.4 Case study 33

4 Generalized Linear Latent Variable Models 35
  4.1 Latent variable models 35
  4.2 GLLVM 36
  4.3 Starting estimator and iterative bootstrap algorithm 37
  4.4 Simulation study 39
4.5 Goodness-of-fit test ............................................. 44
   4.5.1 Goodness-of-fit test statistic .......................... 44
   4.5.2 Simulation Study ........................................... 45

5 Conclusions ..................................................... 49
To my family.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>Approximate Bayesian Computation</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>FA</td>
<td>Factor Analysis</td>
</tr>
<tr>
<td>GLLVM</td>
<td>Generalized Linear Latent Variable Model</td>
</tr>
<tr>
<td>GLM</td>
<td>Generalized Linear Model</td>
</tr>
<tr>
<td>GLMM</td>
<td>Generalized Linear Mixed Model</td>
</tr>
<tr>
<td>GMM</td>
<td>Generalized Method of Moments</td>
</tr>
<tr>
<td>GOF</td>
<td>Goodness-Of-Fit</td>
</tr>
<tr>
<td>IF</td>
<td>Influence Function</td>
</tr>
<tr>
<td>ML(E)</td>
<td>Maximum Likelihood (Estimator)</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>OBRRE</td>
<td>Optimal B-Robust Estimators</td>
</tr>
<tr>
<td>PQL</td>
<td>Penalized Quasi-Likelihood</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Recent developments in Statistics have led to increasingly complex models. These models often have a likelihood that has no closed form which results in difficult optimization problems for parameter estimation. This difficulty is also encountered in the Bayesian framework where methods such as the Approximate Bayesian Computation (ABC) (Beaumont et al., 2002) try to bypass the intractability of the likelihood to obtain an approximate sample from the posterior distribution.

Examples of these complex models can be found in many domains such as finance where stochastic models representing different continuous processes are proposed for observations that however only represent the joint effect of these processes in discrete time. These models are extremely important and popular in Finance and Actuarial Science where they are often used, among others, for asset pricing (Campbell et al., 1997) and for claims reserves (Wüthrich and Merz, 2008). Another widely used class of models that has an intractable likelihood are the Generalized Liner Mixed Model (GLMM) (McCulloch et al., 2008). In these models, the addition of random-effects aims at dealing with repeated measurements (e.g. longitudinal designs, clustered data) in classic Generalized Linear Model (GLM) settings but entails a considerable increase in complexity for the estimation procedure. In fact, the random effects must be integrated out to obtain the marginal likelihood which can lead to non-closed form objective functions. The same challenges are encountered in the Generalized Linear Latent Variable Models (GLLVM) (Bartholomew et al., 2011). In this case, the realizations of the latent (unobserved) variables are treated as missing values and integrated out from the likelihood function. In the domain of robust statistics, the use of weights in the estimating equations may demand the computation of correction terms to achieve Fisher consistency. Indeed, if the data are supposed to be generated from a model having a non-symmetric density for example, the expectation of a weighted score function will be non-zero. This correction can render the computation very difficult or almost infeasible because the estimating process requires to solve integrals that are intractable in most cases. The difficulty caused by the computation of consistency correction terms is also encountered in the robust estimation of many other models.

To overcome the mentioned numerical challenges, many strategies have been used and usually differ from model to model. For example, when the likelihood has no closed form due to multidimensional integrals, several approximations or techniques of numerical integration have been used. In the framework of GLMM, one widely used approximation is the Penalized Quasi-Likelihood (PQL) (Breslow and Clayton, 1993) which, from an algorithmic point of view, consists in computing a working vector of responses (also called pseudo-data) and iterative weights to fit a linear mixed model to obtain estimates for the
fixed effect parameters and predictions for the random effects similarly to the way that
the GLM is fitted via the iterative reweighted least squares algorithm. Based on these
steps, estimates for the variance of the random effects are computed through restricted
maximum likelihood. PQL has been widely implemented in software and works reason-
ably well except for cases where the data is of binary nature (Rodriguez and Goldman,
2001, Goldstein and Rasbash, 1996). Nowadays, more sophisticated approximations are
used for GLMM given the increase in computational power. For example, Laplace approx-
imations and adaptive Gauss-Hermite quadrature are becoming a standard in common
software both for GLMM (Bates et al., 2015) and GLLVM (Huber et al., 2004). The idea
behind the Laplace approximation is to use a second-order Taylor expansion to simplify
the integral and to find an approximate closed-form expression. Adaptive quadratures
increase the accuracy of standard Gauss-Hermite quadrature by centering the quadra-
ture locations under the peak of the integrand. Laplace approximations can be seen as
an adaptive quadrature with only one quadrature node. Pinheiro and Bates (1995) sug-
gest to see Gaussian quadrature as a deterministic version of Monte Carlo integration
whereas adaptive quadrature is the deterministic version of importance sampling. In the
mixed models setting, the grid of quadrature locations are centered around the conditional
modes of the random effects. These approximations are known to provide nearly unbiased
estimators (Pinheiro and Chao, 2006). However they are quite difficult to implement, es-
pecially the optimization of such an approximated likelihood can be quite cumbersome
from a numerical point of view. Moreover, an increase in the number of random effects
tells an increase in the dimension of the integrals which, particularly when the random
effects are crossed, results in computationally intensive problems.

Composite likelihood (Lindsay, 1988) is another approximation method that has been
used to deal with intractable likelihoods. For example, Bello and Varin (2005) approxi-
mated the likelihood function of Generalized Linear Mixed Models (GLMM) with binary
responses and crossed random effects using a pairwise likelihood function formed by the
product of the bivariate probabilities for all possible pairs of observations sharing at least
one common random term. This considerably reduces the dimension of the integrals in
the pairwise likelihood compared to the “true” likelihood which can be then evaluated by
simple quadrature techniques. If the elements are properly chosen, the composite likeli-
hood estimator is consistent and normally distributed (Varin, 2008; Varin et al., 2011).
This kind of approximation has also been used in GLLVM where, for example, Katsikat-
sou et al. (2012) implemented pairwise likelihood estimation in factor analysis models
with ordinal data.

However, all the above strategies that are applied to overcome computational difficul-
ties in estimating complex models are mostly tailor-made to specific (classes of) models
and their application outside the models they are used for is, in general, not straightfor-
ward, if not impossible. Another important class of strategies used to deal with estimation
of complex models are simulation-based methods. Monte Carlo expectation maximization
(Wei and Tanner, 1990) and Markov chain Monte Carlo (Robert and Casella, 1999) are
widely used elements of this class. Their main drawbacks are the computational time
required and the assessment of convergence which requires careful evaluation. In this
thesis, we study the properties of several simulation-based methods which can be em-
ployed to correct sample and asymptotic biases. The resulting properties allow one to
build a framework defining simulation-based estimators that can be implemented very
generally for complex models. The estimators considered in this paper are two-step es-
1.1. Indirect Inference

Indirect estimators (Gouriéroux et al., 1993, Smith, 1993, Gallant and Tauchen, 1996) are a general class of estimators for a model \( F_\theta \) based on auxiliary statistics which can for example be moments or parameters of an auxiliary model. This estimator can be written as the minimum of a quadratic form involving the auxiliary statistic implied by a value of \( \theta \) which can be a known function or can be computed via simulations.

Gouriéroux et al. (1993) presented indirect inference as a simulation-based method aimed at estimating and making inference on the parameters of a complex model whose likelihood is intractable but from which it is easy to simulate data. Instead of using the likelihood of the model, the idea is to use an incorrect (auxiliary) criterion which does not provide a consistent estimator for the parameter of interest. The auxiliary model should have at least the same number of parameters as the true model. Estimates for the auxiliary model are computed on the observations supposedly generated from the correct model and on simulated values drawn from the correct model \( F_\theta \) for a chosen value of \( \theta \). The aim is to find \( \theta \), via minimization of a quadratic form, such that the values of auxiliary estimator found on the observed data and on the simulated data are close enough.

The auxiliary criterion can be an approximation of the likelihood of the correct model or the likelihood of an approximate model. However, it can also be another function of the data such as specific moments. Gallant and Tauchen (1996) proposed to use the scores of the auxiliary model as a criterion function and to search the value of \( \theta \) that generates data for which these scores, evaluated at the auxiliary estimates, are close to 0. Indeed for the observed data the score vector is equal to 0, as implied by the first order conditions. Both approaches can be viewed as extensions of the generalized method of moments (GMM) where the parameters or the scores of the auxiliary model are used to define the GMM criterion function. This also generalizes the method of simulated moments (McFadden, 1989, Pakes and Pollard, 1989) where the moments of a GMM have no analytic form and are instead computed from simulated data. Therefore, indirect inference defines a very broad class of estimators using a general minimization problem based on an incorrect model.

When the observed sample size increases, the auxiliary estimates \( \hat{\pi} \) converge to a pseudo-true value \( \pi(\theta) \) and so the auxiliary estimate defines a mapping from the parameter space of \( \theta \) to the parameter space of the auxiliary model. The aim of indirect inference is to inverse this mapping and to approximate the point \( \theta(\hat{\pi}) \). Therefore the indirect estimator is defined as:

\[
\hat{\theta} = \arg\min_{\theta \in \Theta} ||\hat{\pi} - \pi(\theta)||^2_{\Phi},
\]

where \( \Phi \) is a matrix chosen in a suitable way (see Chapter 2.2) and \( ||X||^2_{\Phi} \) is equal to
\( X^T \Phi X \). If \( \Phi = I \) this corresponds to the squared \( L^2 \) norm. The function \( \pi(\cdot) \) is often called the binding function and is assumed to be one-to-one. Usually this function is unknown or, to put it in another way, has no analytic form and therefore cannot be inverted to find an estimate for \( \theta \). For this reason, it is estimated through simulations allowing to invert the mapping from \( \theta \) to the auxiliary estimator. From an algorithmic point of view, it is essential to keep the same random seed to generate the simulated data to achieve convergence of the procedure. This ensures that the changes in simulated values, and hence in the estimated function \( \pi(\theta) \), are due to the changes in \( \theta \), otherwise the mapping cannot be inverted. Gouriéroux et al. (1993) showed that when the auxiliary estimator is asymptotically normal, the indirect estimator can be consistent and asymptotically normal under some technical conditions. Thus, indirect inference can be seen in some cases as a method for asymptotic bias correction. Gouriéroux et al. (2000) also studied the small sample properties of indirect inference procedures. Using Edgeworth expansions of the indirect estimator they showed that, for an infinite number of simulations, indirect inference performs a second order bias correction.

Indirect inference has been applied in many different areas of statistics including continuous time financial models (Gourieroux and Monfort, 1997, Billio and Monfort, 2003, Sentana et al., 2008), random effects models (Mealli and Rampichini, 1999, Jiang et al., 1999), robustness (Genton and Ronchetti, 2003, Czellar and Ronchetti, 2010), latent variables models (Moustaki and Victoria-Feser, 2006). Recently, indirect inference has also received increased attention in Bayesian statistics, more specifically in ABC (Drovandi et al., 2011). ABC was developed for models for which the likelihood is intractable impeding the estimation of the posterior distribution through Monte Carlo methods. Data are simulated from proposed values of the parameter of interest and if they are close enough to the observed original data, the proposed values have a non-zero probability of being in the approximate posterior distribution. A common way to measure the distance between observed and simulated data is to compare summary statistics for which an important issue is the choice of these statistics. Drovandi et al. (2011) proposed to use parameters of an auxiliary model as summary statistics. Indeed, in the indirect inference framework, the auxiliary model is often supposed to be an incorrect criterion that nevertheless captures important features of the observed data (Smith, 2008).

### 1.2 Bias correction methods via Bootstrap

As noted before, indirect inference can be seen in some cases as an asymptotic bias correction method. What is less known is that it also has finite sample bias reduction properties. Small sample bias correction methods have been regularly proposed in the literature since common estimators such as the Maximum Likelihood Estimator (MLE) are consistent but often biased in finite samples. Jackknife is one of the first methods that was proposed (Tukey, 1958, Quenouille, 1956) to perform correction without having an explicit expression for the bias. Schucany et al. (1971) introduced a way of deriving higher order bias correction by combining two biased estimators which includes the jackknife as a special case. Bootstrap (Efron and Tibshirani, 1993) is another popular method to correct finite sample bias. It is a simple and automatic way to perform bias correction using resampling, without relying on an explicit form of the bias function or trying to compute its expansion.
However, MacKinnon and Smith (1998) have argued that performing bias correction in an automated way without exploring the shape of the bias function can lead to an increase of the variance or, even worse, of the mean squared error of the estimator. They proposed different procedures according to whether the bias is a constant, a linear or a non-linear function of the true parameter. From a similar perspective, one can cite the empirical bias correction procedure of Pfeffermann and Correa (2012) which selects the best bias function via cross-validation among several candidate functions. They obtained better results in terms of bias correction and mean squared error than the bootstrap and the jackknife estimators but their procedure seems slightly complicated for it to be regularly used for complex models.

In this work, we focus on bias correction techniques using the bootstrap as a way to derive estimators for complex models from simpler but inconsistent estimators. This is a non-standard use of these techniques, unlike indirect inference. Kuk (1995) proposed an iterative version of the bootstrap method to precisely correct an inconsistent but easy-to-compute starting estimator and applied it to GLMM estimation. However, theoretical results on the order of correction that the iterative bootstrap delivers were not given and are instead provided in this work. Mealli and Rampichini (1999) compared Kuk’s iterative procedure to indirect inference in a simulation study with binary GLMM using PQL as an inconsistent starting estimator. Their results showed that both methods are comparable and lead to consistent estimators for which this work gives a more formal proof of their empirical results.

More specifically, we consider four existing simulation methods, two of them being based on the indirect inference method and two on the bootstrap. In Chapter 2 we formally set up the bias-correction frameworks for two versions of indirect inference along with the bootstrap and the iterative bootstrap. The two versions of indirect inference basically differ in the way the simulation is made by either simulating $H$ samples of size $n$ or one sample of size $nH$ while the conventional bootstrap corresponds to the first step of the iterative bootstrap. We then study the finite sample and asymptotic properties of the four resulting estimators to find out that indirect inference with $H$ samples of size $n$ is equivalent to the iterative bootstrap, both having more accurate finite sample properties in terms of bias and variance over the other two. Basically, a higher order bias reduction is achieved when using indirect inference with $H$ samples of size $n$ over the one sample of size $nH$, and the same is true for the iterative bootstrap over conventional bootstrap. The relatively small price to pay is an increase (of a quantifiable order) in the sample variance of the simulation-based estimator compared to the sample variance of the direct consistent estimator if the latter can be computed without numerical approximations. As a by-product, some of the simulation-based estimators can have a reduced sample bias of an order up to $n^{-3}$. As an illustration of the different order of correction made by these four simulation methods, in section 2.3 of Chapter 2 we provide a very simple example. We propose simulation-based estimators that are both easy to compute and unbiased in different settings. The first one, described in Chapter 3, highlights the features of these methods for robust estimation through an application to income distribution estimation while the second one is provided in Chapter 4 and concerns the estimation of GLLVM. Part of the material presented here has been already submitted for publication in Guerrier et al. (2015a).
In this Chapter, we will define the four simulation-based estimators and how they are obtained. Two of them are based on the indirect inference method and two on the bootstrap. We also give the conditions under which they are consistent estimators and study their bias reduction properties when the sample size is finite. To this end, we compute their expectation and variance.

### 2.1 Mathematical setup

Generally, we define \( \hat{\pi} = \hat{\pi}(\theta, n) \) as an estimator of \( \theta_0 \in \Theta \subseteq \mathbb{R}^p \) based on a sample of size \( n \) generated from \( F_{\theta_0} \), with \( \Theta \) compact. Let \( \hat{\pi} \overset{p}{\to} \pi_0 \), with \( \pi_0 \equiv \pi(\theta_0) \). Typically \( \pi \neq \theta \), hence \( \hat{\pi} \) is not a consistent estimator of \( \theta_0 \) but is easy to compute.

We also define \( \hat{\theta} \) as a consistent estimator of \( \theta \) based on \( \hat{\pi} \). Assume \( E[\hat{\pi}] \) exists, where \( E[\cdot] \) denotes the expectation under \( F_{\theta_0} \). We write the bias of \( \hat{\pi} \) as:

\[
E[\hat{\pi}] - \theta = D\theta + c(n),
\]

where \( D \) is a \( p \times p \) matrix of fixed coefficients that does not depend on \( n \) or \( \theta \), \( c(n) \) is a fixed and bounded quantity that depends on \( n \) and satisfies \( \lim_{n \to \infty} c(n) = 0 \).

We can then write

\[
\hat{\pi}(\theta, n) = \pi(\theta, n) + v(\theta, n),
\]

(2.1)

where \( \pi(\theta, n) = E[\hat{\pi}(\theta, n)] = \theta + D\theta(\theta, n) + c(n) \) and \( v(\theta, n) \equiv \hat{\pi} - E[\hat{\pi}] \) is a mean zero random vector.

We now describe four bias correction strategies. The first two are indirect estimators in the sense of Gouriéroux et al., 1993. The third Strategy is the first-order bootstrap correction of Efron and Tibshirani, 1993, and the last Strategy is an iterative version of the third Strategy proposed in Kuk, 1995.

**Strategy 2.1:** We define the indirect estimator based on \( H \) samples of size \( n \) as

\[
\hat{\theta} = \text{argmin}_{\theta \in \Theta} \| \hat{\pi} - \pi(\theta, n) \|^2_{\Phi},
\]

(2.2)

where \( \Phi \) is a symmetric positive definite matrix of the same dimension as \( \hat{\pi} \). If \( \Phi \) is estimated by \( \hat{\Phi} \) then \( \hat{\Phi} \overset{p}{\to} \Phi \), \( \Phi \) being positive definite. \( \Phi \) does not necessarily depend on \( \theta \).
Here, 
\[
\hat{\pi}(\theta, n) = \frac{1}{H} \sum_{h=1}^{H} \hat{\pi}_h(\theta, n),
\] (2.3)
where \(\hat{\pi}_h(\theta, n)\) denotes the value of \(\hat{\pi}\) obtained on the \(h\)th simulated sample of size \(n\) under \(F_\theta\).

**Strategy 2.2:** We define the indirect estimator based on one sample of size \(n_H\) as
\[
\hat{\theta} = \arg\min_{\theta \in \Theta} ||\hat{\pi} - \hat{\pi}(\theta, n_H)||^2_n,
\] (2.4)
where \(\hat{\pi}(\theta, n_H)\) denotes the value of \(\hat{\pi}\) obtained on a simulated sample of size \(n_H\) under \(F_\theta\).

Strategy 2.1 and Strategy 2.2 differ in the simulation of either \(H\) samples of size \(n\) or one sample of size \(n_H\) and so they differ in the way they estimate \(\pi(\theta)\). We will show that Strategy 2.1 achieves a higher order bias reduction than Strategy 2.2.

**Strategy 2.3:** The bootstrap bias corrected estimator is defined as:
\[
\hat{\theta}_B = \hat{\pi} + (\hat{\pi} - \hat{\pi}(\pi, n)),
\] (2.5)
where \(\hat{\pi}(\pi, n)\) is calculated as in (2.3) but with \(\theta\) replaced by \(\pi\). Here the bias \(E[\hat{\pi}] - \theta\) is assessed through sampling \(H\) samples of size \(n\) from \(F_\pi\).

**Strategy 2.4:** At iteration \(k\), the iterative bootstrap bias corrected estimator \(\hat{\theta}_B^{(k)}\) is defined as:
\[
\hat{\theta}_B^{(k)} = \hat{\pi} + \left(\hat{\theta}_B^{(k-1)} - \pi\left(\hat{\theta}_B^{(k-1)}, n\right)\right),
\] (2.6)
with 
\[
\pi\left(\hat{\theta}_B^{(k-1)}, n\right) = \frac{1}{H} \sum_{h=1}^{H} \hat{\pi}_h\left(\hat{\theta}_B^{(k-1)}, n\right).
\]
We define \(\hat{\theta}_B\) as the limit of \(\hat{\theta}_B^{(k)}\) once the procedure converges.

Before presenting the statistical properties of the four bias correction methods, we first describe the assumptions.

**Assumption A.1:** The true parameter value \(\theta_0\) satisfies \(||\theta_0||_1 < \infty\).

**Assumption A.2:** At any sample size \(n\), \(\pi(\theta, n)\) as a function of \(\theta\) is one-to-one on \(\Theta\).

**Assumption A.3:** Let \(n^{-\alpha}V_{\theta,n}\) be the variance-covariance matrix of \(v(\theta, n)\) given in (2.1). Then \(\alpha > 0\), and \(|V_{\theta,n}^{(i,j)}| < \infty\), for all \(i, j = 1, \ldots, p\), for any \(\theta \in \Theta\) and at any sample size \(n\).

**Assumption A.4:** If \(\Phi\) is estimated by \(\hat{\Phi}\) then \(\hat{\Phi} \xrightarrow{p} \Phi\), with \(\Phi\) being positive definite.

**Assumption A.5:** \(\pi(\theta)\) is stochastically equicontinuous and \(\hat{\pi}(\theta_0)\) is equicontinuous.

Assumption A.1 simply requires all components of the true parameter \(\theta\) to be finite. Assumption A.2 ensures identifiability. Assumption A.3 requires that the variance of \(v(\theta, n)\) goes to zero as \(n\) increases. For example, in many estimation procedures, \(\alpha = 1\).

It should be noted that, by definition and from (2.1), \(n^{-\alpha}V_{\theta,n}\) is the variance-covariance
2.1. Mathematical setup

Matrix of $\hat{\pi}$. Assumption A.4 requires that the estimated $\hat{\Phi}$ converges to $\Phi$, which is positive definite, when $n \to \infty$.

Assumptions A.1, A.2, A.3, A.4 are very mild and they directly lead to Corollary 2.1 and 2.2. Assumption A.5 is required to prove the uniform convergence in probability of $\hat{\pi}$ which is needed for Corollary 2.2.

**Corollary 2.1**: As functions of $\theta$, both $\hat{\pi}(\theta, n)$ and $\pi(\theta, nH)$ are one-to-one on $\Theta$ when $\max(n, H) \to \infty$.

**Proof**: Using (2.3) as well as Assumption A.3, we have that

$$\lim_{H \to \infty} \hat{\pi}(\theta, n) = \lim_{H \to \infty} \frac{1}{H} \sum_{h=1}^{H} \hat{\pi}_h(\theta, n) = \pi(\theta, n) + \lim_{H \to \infty} \frac{1}{H} \sum_{h=1}^{H} v_h(\theta, n) = \pi(\theta, n)$$

$$\lim_{H \to \infty} \hat{\pi}(\theta, nH) = \lim_{H \to \infty} \pi(\theta, nH) + \lim_{H \to \infty} v(\theta, nH) = \pi(\theta)$$

$$\lim_{n \to \infty} \hat{\pi}(\theta, n) = \lim_{n \to \infty} \frac{1}{H} \sum_{h=1}^{H} \hat{\pi}_h(\theta, n) = \pi(\theta) + \frac{1}{H} \sum_{n=1}^{H} \lim_{n \to \infty} v_h(\theta, n) = \pi(\theta)$$

$$\lim_{n \to \infty} \hat{\pi}(\theta, nH) = \lim_{n \to \infty} \pi(\theta, nH) + \lim_{n \to \infty} v(\theta, nH) = \pi(\theta)$$

where $\pi(\theta) = \lim_{n \to \infty} \pi(\theta, n)$. The proof for the case where $(n, H) \to \infty$ is trivial. By Assumption A.2 we have that $\hat{\pi}(\theta)$ and $\pi(\theta)$ are both one-to-one on $\Theta$ if $\max(n, H) \to \infty$.

**Remark A**: The equivalent of Corollary 2.1 is not needed in the Gouriéroux et al. (1993) framework as they used the argument that, for a fixed $H$, $\pi(\theta, n)$ goes to $\pi(\theta)$ which by assumption is one-to-one. In our case, we let $H \to \infty$ and keep $n$ fixed.

**Corollary 2.2**: Under Assumptions A.1 to A.5, and for a fixed $H$, $\hat{\theta}$ defined in (2.2), $\tilde{\theta}$ defined in (2.4) and $\tilde{\theta}_B$ defined in (2.6) are consistent estimators of $\theta$.

**Proof**: For $\hat{\theta}$ and $\tilde{\theta}$ the proof is a direct application of Theorem 1.1 of Harris and Mátyás, 1999.

Assuming that $\hat{\pi} \xrightarrow{P} \pi_0$, that the parameter set $\Theta$ is compact and under Assumption A.5 we have that $\hat{\pi} \xrightarrow{P} \pi_0$ uniformly. Let $g(\theta, n) = \mathbb{E}[\hat{\pi} - \pi(\theta, n)]$ then we have that $g(\theta_0, n) = 0$. Moreover, we have by Assumption A.2 that $g(\theta, n) = 0$ if and only if $\theta = \theta_0$ since $\pi(\theta, n)$ is one-to-one on $\Theta$.

Let $g^*(\theta, n) = \mathbb{E}[\hat{\pi} - \pi(\theta, n)]$ then by Corollary 2.1 we have that, for a fixed $H \in \mathbb{N}$ $g^*(\theta, n) = 0$ if and only if $\theta = \theta_0$. Under Assumption A.4, and because $\hat{\pi}(\theta)$ converges uniformly to $\pi(\theta_0)$, we can verify that the conditions of Theorem 1.1 of Harris and Mátyás, 1999 are satisfied. Therefore $\tilde{\theta}$ defined in (2.2) is consistent.

Using the same argument, we have that $\hat{\theta}$ defined in (2.4) is also a consistent estimator of $\theta_0$.

For $\tilde{\theta}_B$, we have that at convergence of the iterative bootstrap procedure

$$\tilde{\theta}_B^{(k-1)} = \tilde{\theta}_B^{(k)} \equiv \tilde{\theta}_B.$$

Using (2.6) we can write

$$\tilde{\theta}_B = \pi(\theta, n) + (\tilde{\theta}_B - \tilde{\pi}(\tilde{\theta}_B, n)),$$
hence
\[ \hat{\pi}(\theta, n) = \pi(\hat{\theta}_B, n) + \frac{1}{H} \sum_{h=1}^{H} v_h(\hat{\theta}_B, n). \]

Using assumption A.3 in equation (2.1) we have
\[ \lim_{n \to \infty} \hat{\pi}(\theta, n) = \lim_{n \to \infty} \pi(\theta, n) + \lim_{n \to \infty} v(\theta, n) = \pi(\theta). \]

Since \( \lim_{n \to \infty} v_h(\hat{\theta}_B, n) = 0, \forall h \), then \( \lim_{n \to \infty} \pi(\hat{\theta}_B, n) = \pi(\theta) \) and by Assumption A.2, \( \hat{\theta}_B \xrightarrow{p} \theta \).

Next, we impose some additional assumptions on the form of the vector \( b(\theta, n) \) that contributes to the leading term of the bias vector.

**Assumption A.6:**
\[ b(\theta, n) = [b_j(\theta_j, n)]_{j=1,...,p} \] where \( b_j(\theta_j, n) = \sum_{i=0}^{q_j} a_{ij} \left( \frac{\theta_j}{n} \right)^i < \infty. \)

Assumption A.6 requires the \( j \)th component of the vector \( b(\theta, n) \) to be a polynomial in \( \theta_j/n \). Polynomial approximations are often used to assess bias (see e.g. Efron and Tibshirani, 1993). However, for a vector parameter \( \theta \), the bias of its \( j \)th component may also depend on other elements of \( \theta \) when \( D \) is not diagonal. Therefore, we allow a more flexible bias form described by \( Db(\theta, n) + c(n) \) in (2.1) so that the bias of estimating \( \theta_j \) is a linear combination of such polynomials.

The next assumptions impose some regularity conditions on the matrix \( D \), the coefficients \( a_{ij} \) and the term \( c(n) \).

**Assumption A.7:** Let \( \hat{\theta}^o \equiv \lim_{H \to \infty} \hat{\theta} \). The coefficients \( a_{ij} \) satisfy that for \( i, j = 1, ..., p \):
\[
\begin{align*}
\text{var} \left[ \sum_{i=2}^{q_j} \frac{a_{ij}}{n^{i-2}} (\hat{\theta}_j^o)^i \right] &= d_j < \infty, \\
\text{var} \left[ \sum_{i=2}^{q_j} \frac{a_{ij}}{n^{i-2}} (\hat{\pi}_j^o)^i \right] &= \tilde{d}_j < \infty, \\
\text{cov} \left[ \sum_{i=1}^{q_j} \frac{a_{ij}}{n^{i-1}} (\hat{\theta}_j^o)^i; \sum_{i=1}^{q_k} \frac{a_{ik}}{n^{i-1}} (\hat{\theta}_k^o)^i \right] &= c_{jk} < \infty, \\
\text{cov} \left[ \sum_{i=1}^{q_j} \frac{a_{ij}}{n^{i-1}} (\hat{\pi}_j^o)^i; \sum_{i=1}^{q_k} \frac{a_{ik}}{n^{i-1}} (\hat{\pi}_k^o)^i \right] &= f_{jk} < \infty.
\end{align*}
\]

**Assumption A.8:** \( I + \frac{1}{n} DA_1 \) is invertible, where \( A_1 = \text{diag}(a_{1j})_{j=1,...,p} \).

**Assumption A.9:** \( c(n) = O\left(n^{-\delta}\right) \), where \( 0 < \delta < \infty \).

Assumption A.9 increases the flexibility of the auxiliary estimator by allowing its bias to have an additional vanishing quantity in addition to a fixed function of the sample size and the parameter to be estimated.
2.2 Statistical Properties of Simulation Based Methods for Bias Reduction

We now study the convergence rate of the four bias reduction methods presented in Section 2.1. In Theorem 2.1, we show that indirect inference provides consistent estimators with a fast convergence rate when the number of simulated samples $H$ is large. Moreover, the variance of the indirect estimator is only slightly inflated compared to the variance of the initial estimator, with an increment of order $O(n^{-2})$. This increment tends to be smaller than the variance of the initial estimator. The conditions as well as the results in Theorem 2.1 are different from those in Gouriéroux et al., 2000, where it is shown that indirect inference automatically operates a second order bias correction.

**Theorem 2.1:** Under Assumptions A.1 to A.8, the corrected estimator $\hat{\theta}$ defined in (2.2) satisfies

$$
E \left[ \lim_{H \to \infty} \hat{\theta} \right] = \theta_0 + O \left( n^{-\min(3,\alpha+2)} \right),
$$

$$
\text{var} \left( \lim_{H \to \infty} \hat{\theta} \right) = n^{-\alpha} V_{\theta,n} + O \left( n^{-\min(2,1+\alpha)} \right).
$$

**Proof:** $\hat{\theta}$ is implicitly defined as the solution in $\theta$ of

$$
\frac{\partial}{\partial \theta} \| \hat{\pi} - \bar{\pi}(\theta, n) \|^2_{\Phi} = 0.
$$

Since $\Phi$ is positive definite and $\bar{\pi}(\theta, n)$ is one-to-one by Corollary 2.1 for $H \to \infty$, we have that $\hat{\theta}$ is actually implicitly defined (when taking the limit in $H$) as

$$
\hat{\pi} - \lim_{H \to \infty} \bar{\pi}(\hat{\theta}, n) = 0.
$$

Using (2.1) and (2.3), we have

$$
\hat{\pi} = \theta_0 + Db(\theta_0, n) + c(n) = \lim_{H \to \infty} \bar{\pi}(\hat{\theta}, n) = \hat{\theta} + Db(\hat{\theta}, n) + c(n).
$$

Using the definition $\hat{\theta}^o \equiv \lim_{H \to \infty} \hat{\theta}$ and rearranging the terms, we get

$$
\hat{\theta}^o = \theta_0 + D \left[ b(\theta_0, n) - b(\hat{\theta}^o, n) \right] + v(\theta_0, n). \tag{2.7}
$$

We now consider the variance of $\hat{\theta}^o$. Using (2.7), we get

$$
\text{var} \left( \hat{\theta}^o \right) = \text{var} \left( v(\theta_0, n) \right) + D \text{var} \left( b(\hat{\theta}^o, n) \right) D^T - \frac{2}{n} D \left[ r_{j,k} \right]_{j,k=1,...,p},
$$

$$
\overset{\Delta}{=} \text{var} \left( v(\theta_0, n) \right) + \frac{1}{n^2} DCD^T - \frac{2}{n} D \left[ r_{j,k} \right]_{j,k=1,...,p}, \tag{2.8}
$$

with $C = [c_{jk}]_{j,k=1,...,p}$ and $r_{j,k} = \text{cov} \left[ \sum_{i=1}^{q_j} \frac{a_{ij}}{n^{0.5}} \left( \hat{\theta}^o_j \right)^i ; v_k(\theta_0, n) \right]$. Using the Cauchy-Schwarz inequality, we have

$$
r_{j,k}^2 \leq \text{var} \left[ \sum_{i=1}^{q_j} \frac{a_{ij}}{n^{0.5}} \left( \hat{\theta}^o_j \right)^i \right] \text{var} \left[ v_k(\theta_0, n) \right]. \tag{2.9}
$$
By Assumption A.3, we have that \( \text{var} [v_k(\theta_0, n)] = O(n^{-\alpha}) \). We can write
\[
\text{var} \left[ \sum_{i=1}^{n} \frac{a_{ij}}{n^{i-1}} \left( \hat{\theta}_j^i \right) \right] = a_{2,j} \text{var} [\hat{\theta}_j] + \text{var} \left[ \sum_{i=2}^{q_i} \frac{a_{ij}}{n^{i-1}} \left( \hat{\theta}_j^i \right) \right] + 2 \text{cov} \left( a_{1,j} \hat{\theta}_j; \sum_{i=2}^{q_i} \frac{a_{ij}}{n^{i-1}} \left( \hat{\theta}_j^i \right) \right).
\]

Now we study each term of the equation above:
\[
\text{var} \left[ \hat{\theta}_j^o \right] = \text{var} \hat{\theta}_j^o, \\
= (2.7) \text{var} \left[ \theta_0 + D \left( b(\theta_{0,j}, n) - b(\hat{\theta}_j, n) \right) + v(\theta_0, n) \right]_j, \\
= \text{D var} \left[ b(\hat{\theta}_j, n) \right] \text{D}^T + \text{var} [v(\theta_0, n)] - 2 \text{cov} \left( \text{Db}(\hat{\theta}_j, n); v(\theta_0, n) \right)_j.
\]
\( \text{D var} \left[ b(\hat{\theta}_j, n) \right] \text{D}^T = \frac{1}{n^2} \text{DCD}^T \) and is of order \( O(n^{-2}) \) by Assumption A.7 whereas \( \text{var} [v(\theta_0, n)] \) is of order \( O(n^{-\alpha}) \) by Assumption A.3. Applying the Cauchy-Schwarz inequality again, we have that \( \text{cov} \left( \text{Db}(\hat{\theta}_j, n); v(\theta_0, n) \right) \) is of order \( O(n^{-\min(2, \alpha)}) \). Therefore
\[
\text{var} \left[ \hat{\theta}_j^o \right] = O(n^{-\min(2, \alpha)}).
\]
The second term is equal to:
\[
\text{var} \left[ \frac{1}{n^2} \sum_{i=2}^{q_i} \frac{a_{ij}}{n^{i-2}} \left( \hat{\theta}_j^i \right) \right] = \frac{1}{n^2} \text{var} \left[ \sum_{i=2}^{q_i} \frac{a_{ij}}{n^{i-2}} \left( \hat{\theta}_j^i \right) \right] = \frac{1}{n^2} d_j,
\]
and is of order \( O(n^{-2}) \) as \( d_j \) is bounded by Assumption A.7. The last term is at least of order \( O(n^{-\min(2, \alpha)}) \) by the Cauchy-Schwartz inequality. We obtain that:
\[
\text{var} \left[ \sum_{i=1}^{q_i} \frac{a_{ij}}{n^{i-1}} \left( \hat{\theta}_j^i \right) \right] = O(n^{-\min(2, \alpha)}).
\]
Using this result in (2.9), it follows that \( r_{j,k}^2 \) is of order \( O(n^{-\min(2, \alpha + 2)}) \).
This implies that:
\[
\text{var} \left( \hat{\theta}_j^o \right) = \text{var} \left( v(\theta_0, n) \right) + \frac{1}{n^2} \text{DCD}^T + O \left( n^{-\min(1+\alpha, 2+\alpha/2)} \right),
\]
\[
\overset{\text{A.7}}{=} n^{-\alpha} V_{\theta, n} + O \left( n^{-\min(2, 1+\alpha)} \right),
\]
where \( C = [c_{jk}]_{j,k=1,...,p} \). This verifies the second part of the proof.

Taking expectations on both sides of (2.7), we get
\[
\text{E} \left[ \hat{\theta}_j^o \right] = \theta_0 + \text{D E} \left[ b(\theta_0, n) - b(\hat{\theta}_j, n) \right],
\]
\[
\overset{\text{A.6}}{=} \theta_0 + \text{D E} \left[ \sum_{i=0}^{q_i} a_{ij} \left( \frac{\theta_0}{n} \right)^i - \sum_{i=0}^{q_i} a_{ij} \left( \frac{\hat{\theta}_j}{n} \right)^i \right]_{j=1,...,p},
\]
\[
= \theta_0 + \text{D E} \left[ \frac{1}{n} A_1 (\theta_0 - \hat{\theta}_j^o) \right] + \text{D E} [u] = \theta_0 + \frac{1}{n} D A_1 (\theta_0 - \text{E} \left[ \hat{\theta}_j^o \right]) + \text{D E} [u],
\]
\[
\overset{\text{(2.11)}}{=}
\]
where $A_1 = \text{diag} \left( a_{ij} \right)_{j=1,...,p}$ and $u = \left[ \sum_{i=2}^{q_i} \frac{a_{ij}}{n^2} \left( (\theta_{0,j})^i - (\hat{\theta}_{(j)})^i \right) \right]_{j=1,...,p}$. Rearranging the terms, we have

$$
E \left[ \hat{\theta}^o \right] - \theta_0 = \frac{1}{n} DA_1 \left( \theta_0 - E \left[ \hat{\theta}^o \right] \right) + DE \left[ u \right],
$$

$$
I \left( E \left[ \hat{\theta}^o \right] - \theta_0 \right) + \frac{1}{n} DA_1 \left( E \left[ \hat{\theta}^o \right] - \theta_0 \right) = DE \left[ u \right],
$$

$$
\left( I + \frac{1}{n} DA_1 \right) \left( E \left[ \hat{\theta}^o \right] - \theta_0 \right) = DE \left[ u \right],
$$

(2.12)

We now consider the $j^{\text{th}}$ element of $E \left[ u \right]$. Letting $u_j^* = \sum_{i=3}^{q_i} \frac{a_{ij}}{n^2} \left( (\theta_{0,j})^i - (\hat{\theta}_{(j)})^i \right)$, we obtain

$$
E \left[ u_j \right] = E \left[ \frac{a_{2j}}{n^2} \left( \theta_{0,j}^2 - \hat{\theta}_{(j)}^2 \right) \right] + E \left[ u_j^* \right],
$$

$$
\overset{A.7}{=} \frac{a_{2j}}{n^2} \left( \theta_{0,j}^2 - E \left[ \hat{\theta}_{(j)}^2 \right] \right) + O \left( n^{-3} \right),
$$

$$
= \frac{a_{2j}}{n^2} \left( \theta_{0,j}^2 - E \left[ \hat{\theta}_{j}^2 \right] - \text{var} \left[ \hat{\theta}_{j} \right] \right) + O \left( n^{-3} \right),
$$

$$
= \frac{a_{2j}}{n^2} \left( \theta_{0,j} - E \left[ \hat{\theta}_{j} \right] \right) \left( \theta_{0,j} + E \left[ \hat{\theta}_{j} \right] \right) - \frac{a_{2j}}{n^2} \text{var} \left[ \hat{\theta}_{j} \right] + O \left( n^{-3} \right),
$$

$$
= \frac{a_{2j}}{n^2} \left( \theta_{0,j} - E \left[ \hat{\theta}_{j} \right] \right) \left( \theta_{0,j} + E \left[ \hat{\theta}_{j} \right] \right) + O \left( n^{-\min(3,\alpha+2)} \right).
$$

(2.13)

From equation (2.12) we can write

$$
\left( \theta_{0,j} - E \left[ \hat{\theta}_{j} \right] \right) = - \left( B E \left[ u_{(j)} \right] \right).
$$

Since matrix $B$ is bounded, the terms $\left( \theta_{0,j} - E \left[ \hat{\theta}_{j} \right] \right)$ and $E \left[ u_{(j)} \right]$ converge at the same rate. Using (2.13), we have that (at least) $E \left[ u_{(j)} \right] = O \left( n^{-2} \right)$. Moreover, since the term $\left( \theta_{0,j} + E \left[ \hat{\theta}_{j} \right] \right)$ is bounded by Assumptions A.1 and A.6, we obtain

$$
\frac{a_{2j}}{n^2} \left( \theta_{0,j} - E \left[ \hat{\theta}_{j} \right] \right) \left( \theta_{0,j} + E \left[ \hat{\theta}_{j} \right] \right) = O \left( n^{-4} \right).
$$

(2.14)

Now, substituting (2.14) in (2.13), we obtain

$$
E \left[ u_{(j)} \right] = O \left( n^{-4} \right) + O \left( n^{-\min(3,\alpha+2)} \right),
$$

and, finally, we have

$$
E \left[ \hat{\theta}^o \right] - \theta_0 = O \left( n^{-\min(3,\alpha+2)} \right),
$$

which concludes the proof.

In Theorem 2.2, we show that for the indirect inference procedure the asymptotic properties differ according to whether we use a single sample of size $nH$ or $H$ samples of size $n$. More specifically, the bias of the single-sample based estimator $\hat{\theta}$ is larger than that of the multi-sample based estimator $\hat{\theta}$ while its variance is smaller.
Theorem 2.2: Under Assumptions A.1 to A.6 and Assumption A.9, the corrected estimator \( \tilde{\theta} \) defined in (2.4) satisfies

\[
\mathbb{E} \left[ \lim_{H \to \infty} \tilde{\theta} \right] = \theta_0 + \mathcal{O} \left( n^{-\min(1,\delta)} \right),
\]

\[
\text{var} \left( \lim_{H \to \infty} \tilde{\theta} \right) = n^{-\alpha} \mathbb{V}_{\theta,n}.
\]

Proof: \( \tilde{\theta} \) is implicitly defined as the solution in \( \theta \) of

\[
\frac{\partial}{\partial \theta} \| \hat{\pi} - \tilde{\pi}(\theta, nH) \|^2_\Phi = 0.
\]

Since \( \Phi \) is positive definite and \( \tilde{\pi}(\theta, nH) \) is one-to-one by Corollary 2.1 for \( H \to \infty \) we have that \( \tilde{\theta} \) is actually implicitly defined as

\[
\hat{\pi} - \lim_{H \to \infty} \tilde{\pi}(\tilde{\theta}, nH) = 0.
\]

Using (2.1), we can write

\[
\hat{\pi} = \theta_0 + D b(\theta_0, n) + c(n) + v(\theta_0, n),
\]

\[
\tilde{\pi}(\tilde{\theta}, nH) = \tilde{\theta} + D b(\tilde{\theta}, nH) + c(nH) + v(\tilde{\theta}, nH).
\] (2.15)

When \( H \to \infty \), we have that \( c(nH) \to 0 \) and \( b(\tilde{\theta}, nH) \to A_0 \equiv (a_{0j})_{j=1,...,p} \). By definition, \( v(\tilde{\theta}, nH) \) has expectation zero and variance \( (nH)^{-\alpha} \mathbb{V}_{\theta,nH} \) which implies that \( \lim_{H \to \infty} v(\tilde{\theta}, nH) = 0 \) in probability. Thus, rearranging (2.15) when \( H \to \infty \) and writing \( \tilde{\theta}^o \equiv \lim_{H \to \infty} \tilde{\theta} \), we get

\[
\tilde{\theta}^o = \theta_0 + D [b(\theta_0, n) - A_0] + v(\theta_0, n) + c(n).
\] (2.16)

Since \( \theta_0 + D [b(\theta_0, n) - A_0] + c(n) \) is a non-random quantity, we have

\[
\text{var}(\tilde{\theta}^o) = n^{-\alpha} \mathbb{V}_{\theta_0,n},
\]

which verifies the second part of proposition 2.2. Taking the expectation of \( \tilde{\theta}^o \) in (2.16), with \( \mathbb{E}[v(\theta_0, n)] = 0 \), yields

\[
\mathbb{E} \left[ \tilde{\theta}^o \right] = \theta_0 + D \mathbb{E} [b(\theta_0, n) - A_0] + c(n),
\]

\[
A.6 \quad \theta_0 + D \mathbb{E} \left[ \sum_{i=1}^{q_i} a_{ij} \left( \frac{\theta_{0,j}}{n} \right)^j \right]_{j=1,...,p} + c(n),
\]

\[
A.9 \quad \theta_0 + \mathcal{O} \left( n^{-\min(1,\delta)} \right),
\]

which verifies the first part of Theorem 2.2.

In Theorem 2.3 and Corollary 2.3, we derive the properties of the one step bootstrap estimator \( \hat{\theta}_B \). We show that the asymptotic bias is of the same order as the indirect inference estimator using one sample of size \( nH \) when \( \delta \geq 1 \) and is of larger order when \( \delta < 1 \), while the variance is identical to the first leading order. Since in most practical situations \( \delta = 1 \), the one step iterative bootstrap estimator can be considered as an alternative to the one sample indirect inference estimator.
Theorem 2.3: Under Assumptions A.1 to A.7, the corrected estimator $\hat{\theta}_B$ defined in (2.5) satisfies

$$\mathbb{E} \left[ \lim_{H \to \infty} \hat{\theta}_B \right] = \theta_0 + O\left(n^{-1}\right),$$

$$\text{var} \left( \lim_{H \to \infty} \hat{\theta}_B \right) = n^{-\alpha} \mathbb{V}_{\theta,n} + O\left(n^{-\min(2,\alpha+1)}\right).$$

Proof:

Using (2.1) we can write that

$$\hat{\theta}_B = 2\hat{\pi} - \frac{1}{H} \sum_{h=1}^{H} \hat{\pi}_h(\hat{\pi},n),$$

$$= 2\hat{\pi} - \frac{1}{H} \sum_{h=1}^{H} \left( \hat{\pi} + Dh(\hat{\pi},n) + c(n) + v_h(\hat{\pi},n) \right),$$

$$= \hat{\pi} - Dh(\hat{\pi},n) - c(n) - \frac{1}{H} \sum_{h=1}^{H} v_h(\hat{\pi},n),$$

$$= \theta_0 + Dh(\theta_0,n) + v(\theta_0,n) - Dh(\hat{\pi},n) - \frac{1}{H} \sum_{h=1}^{H} v_h(\hat{\pi},n).$$

Since $\lim_{H \to \infty} \frac{1}{H} \sum_{h=1}^{H} v_h(\hat{\pi},n) = 0$ in probability, letting $\hat{\theta}_B^0 \equiv \lim_{H \to \infty} \hat{\theta}_B$, we have

$$\hat{\theta}_B^0 = \theta_0 + D [b(\theta_0,n) - b(\hat{\pi},n)] + v(\theta_0,n). \quad (2.17)$$

We define $m_{j,k} = \text{cov} \left[ \sum_{i=1}^{q_j} \frac{a_{ij}}{n+1} (\hat{\pi})^i \right]_{j,k} ; v_k(\theta_0,n)$. Using Assumption A.3, A.7 and the Cauchy-Schwarz inequality it can be shown, similarly to $r_{j,k}$, that $m_{j,k} = O(n^{-\min(\alpha,1+\alpha/2)})$. We have

$$\text{var}(\hat{\theta}_B^0) = \text{cov} \left[ Db(\hat{\pi},n) + v(\theta_0,n) ; Db(\hat{\pi},n) + v(\theta_0,n) \right],$$

$$= \text{var} \left( v(\theta_0,n) \right) + D \text{var} \left( b(\hat{\pi},n) \right) D^T - \frac{2}{n} D [m_{j,k}]_{j,k=1,...,p},$$

$$= n^{-\alpha} \mathbb{V}_{\theta_0,n} + \frac{1}{n^2} D F D^T + O\left(n^{-\min(1+\alpha,2+\alpha/2)}\right)$$

$$\equiv n^{-\alpha} \mathbb{V}_{\theta_0,n} + O\left(n^{-\min(2,1+\alpha)}\right), \quad (2.18)$$

where $F = [f_{jk}]_{j,k=1,...,p}$. Taking the expectation on both sides of (2.17), we get

$$\mathbb{E} \left[ \hat{\theta}_B^0 \right] = \theta_0 + D \mathbb{E} \left[ b(\theta_0,n) - b(\hat{\pi},n) \right],$$

$$\overset{A.6}{=} \theta_0 + D \mathbb{E} \left[ \sum_{i=0}^{q_i} a_{ij} \left( \frac{\theta_{0,j}}{n} \right)^i - \sum_{i=0}^{q_j} a_{ij} \left( \frac{\hat{\pi}_j}{n} \right)^i \right]_{j=1,...,p},$$

$$= \theta_0 + D \mathbb{E} \left[ \frac{1}{n} \text{diag}(A_1) (\theta_0 - \hat{\pi}) \right] - D \mathbb{E} \left[ w \right],$$

$$= \theta_0 + \frac{1}{n} D \text{diag}(A_1) (\theta_0 - \mathbb{E}\hat{\pi}) - D \mathbb{E} \left[ w \right], \quad (2.19)$$

where

$$w = \left[ \sum_{i=2}^{q_j} a_{ij} \left( \frac{\theta_{0,j}}{n^i} - (\hat{\pi}_j)^i \right) \right]_{j=1,...,p}. $$
If we consider the $j^{th}$ element of $E[w]$, we have
\[
E[w_j] = E \left[ \frac{a_{2j}}{n^2} \left( \theta_{0,j}^2 - \hat{\theta}_j^2 \right) \right] + E[w_j] = \frac{a_{2j}}{n^2} \left( \theta_{0,j}^2 - \hat{\theta}_j^2 \right) + O \left( n^{-3} \right),
\]
where $w_j^* = \sum_{i=3}^{q_j} \frac{a_{ij}}{n} \left( (\theta_{0,j})^i - (\hat{\theta}_j)^i \right)$. Therefore we can conclude that $E[w]$ is at most $O(n^{-2})$. The second element of (2.19) can be developed as follows
\[
\frac{1}{n} D \text{diag}(A_1) (\theta_0 - E[\hat{\pi}]) = -\frac{1}{n} D \text{diag}(A_1) (D b (\theta_0, n) + c(n)) = -\frac{1}{n} D \text{diag}(A_1) D \left[ \sum_{i=0}^{q_j} a_{ij} \left( \frac{\theta_{0,j}}{n} \right)^i \right]_{j=1, \ldots, n}
\]
\[
= -\frac{1}{n} D \text{diag}(A_1) [c(n)] + \frac{1}{n} D \text{diag}(A_1) D A_0
\]
\[
= -\frac{1}{n} D \text{diag}(A_1) D \left[ \sum_{i=1}^{q_j} a_{ij} \left( \frac{\theta_{0,j}}{n} \right)^i \right]_{j=1, \ldots, n} - \frac{1}{n} D \text{diag}(A_1) [c(n)],
\]
which is therefore $O(n^{-1})$. Hence, we obtain
\[
E[\hat{\theta}_B] = \theta_0 + O(n^{-1}),
\]
which concludes the proof.

\[\square\]

**Corollary 2.3:** If $a_{0j}$ defined in Assumption A.6 satisfies $a_{0,j} = 0$, $j = 1, \ldots, p$, then under the same conditions as in Theorem 2.3 and Assumption A.9, we have
\[
E \left[ \lim_{H \to \infty} \hat{\theta}_B \right] = \theta_0 + O \left( n^{-\min(2,1+\delta)} \right),
\]
\[
\text{var} \left( \lim_{H \to \infty} \hat{\theta}_B \right) = n^{-\alpha} \text{V}_{\theta,n} + O(n^{-\min(2,\alpha+1)}).
\]

**Proof:** The proof is immediate from (2.20). Indeed, if we assume a different form of the bias function, that is without the constant terms (i.e. $a_{0j} = 0$) then $A_0 = 0$ and (2.20) would reduce to
\[
\frac{1}{n} D \text{diag}(A_1) (\theta_0 - E[\hat{\pi}]) = -\frac{1}{n} D \text{diag}(A_1) D \left[ \sum_{i=1}^{q_j} a_{ij} \left( \frac{\theta_{0,j}}{n} \right)^i \right]_{j=1, \ldots, n} - \frac{1}{n} D \text{diag}(A_1) [c(n)],
\]
which is $O(n^{-\min(2,1+\delta)})$. We hence obtain
\[
E[\hat{\theta}_B^o] = \theta_0 + O(n^{-\min(2,1+\delta)}).
Since constant terms do not influence the variance of the estimator, the variance of $\hat{\theta}_B^o$ is the same as that given in (2.18).

Theorem 2.4 derives the asymptotic properties of the iterative bootstrap bias correction estimator $\tilde{\theta}_B$. We show that it has the same properties as the one sample indirect inference estimator $\hat{\theta}$.

**Theorem 2.4:** Under Assumptions A.1 to A.8, the corrected estimator $\tilde{\theta}_B$ defined in (2.6) satisfies

$$
\mathbb{E} \left[ \lim_{H \to \infty} \tilde{\theta}_B \right] = \theta_0 + O \left( n^{-\min(3,\alpha+2)} \right),
$$

$$
\text{var} \left( \lim_{H \to \infty} \tilde{\theta}_B \right) = n^{-\alpha} \mathbf{v}_{\theta,n} + O \left( n^{-\min(2,\alpha+1)} \right).
$$

**Proof:**

At iteration $k$, $\tilde{\theta}_B^{(k)}$ is explicitly defined as:

$$
\tilde{\theta}_B^{(k)} = \pi + \left( \tilde{\theta}_B^{(k-1)} - \pi \left( \tilde{\theta}_B^{(k-1)}, n \right) \right),
$$

$$
= \pi + \tilde{\theta}_B^{(k-1)} - \frac{1}{H} \sum_{h=1}^{H} \pi_h \left( \tilde{\theta}_B^{(k-1)}, n \right),
$$

$$
\overset{(2.1)}{=} \theta_0 + D_b \left( \theta_0, n \right) + c(n) + v \left( \theta_0, n \right) + \tilde{\theta}_B^{(k-1)},
$$

$$
- \frac{1}{H} \sum_{h=1}^{H} \left( \tilde{\theta}_B^{(k-1)} - D_b \left( \tilde{\theta}_B^{(k-1)}, n \right) + c(n) + v_h \left( \tilde{\theta}_B^{(k-1)}, n \right) \right),
$$

$$
= \theta_0 + D_b \left( \theta_0, n \right) + v \left( \theta_0, n \right) - D_b \left( \tilde{\theta}_B^{(k-1)}, n \right) - \frac{1}{H} \sum_{h=1}^{H} v_h \left( \tilde{\theta}_B^{(k-1)}, n \right).
$$

When the iterated bootstrap reaches convergence, we have that $\tilde{\theta}_B^{(k-1)} = \tilde{\theta}_B^{(k)} = \tilde{\theta}_B$. Hence

$$
\tilde{\theta}_B = \theta_0 + D_b \left( \theta_0, n \right) + v \left( \theta_0, n \right) - D_b \left( \tilde{\theta}_B, n \right) - \frac{1}{H} \sum_{h=1}^{H} v_h \left( \tilde{\theta}_B, n \right).
$$

When $H \to \infty$, we get

$$
\tilde{\theta}_B^o = \theta_0 + b \left( \theta_0, n \right) - b \left( \tilde{\theta}_B^o, n \right) + v \left( \theta_0, n \right),
$$

which is the same as (2.7) and hence $\tilde{\theta}_B$ has the same properties as $\hat{\theta}$.

---

### 2.3 A simple example

We illustrate the performance of the various bias correction procedures discussed above in a simple simulation example. Let $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$; we wish to estimate $\sigma^2$ based on the sample $(x_i)_{i=1,\ldots,n}$ using

$$
\hat{\sigma}^2 = \frac{1}{n+5} \sum_{i=1}^{n} (x_i - \bar{x})^2 - \frac{1}{n} \hat{\sigma}^2
$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. 


Note that $\mathbb{E}[\hat{\pi}] = \sigma^2(n-1)/(n+5) - 1/n$, hence it is an asymptotically biased estimator of $\sigma^2$. In addition, it is easily verified that $\delta = 1$ in the expansion of $\mathbb{E}[\hat{\pi}]$ (see Assumption A.9) and $\alpha = 1$ in the variance of $\hat{\pi}$ (see Assumption A.3).

We now correct the bias of $\hat{\pi}$ using the methods discussed in Chapter 2.1 and compare their performance with the conventional unbiased estimator

$$
\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.
$$

Let $\hat{\sigma}^2$, $\hat{\sigma}^2_B$ and $\hat{\sigma}^2_B$ denote the estimator based on Strategies 2.1, 2.2, 2.3 and 2.4 respectively. Comparing the results from Theorem 2.2, 2.3 and the bias of $\hat{\pi}$, we expect that $\hat{\sigma}^2$ and $\hat{\sigma}^2_B$ will not improve much over the original estimator $\hat{\pi}$ since their biases have the same order $O(n^{-1})$. On the contrary, $\hat{\sigma}^2$ and $\hat{\sigma}^2_B$ will have a bias of order $O(n^{-3})$, hence will be much improved over the original estimator $\hat{\pi}$.

Figure 2.1 shows the empirical distribution of the different estimators obtained with 1000 Monte Carlo replications with $\sigma^2 = 10$, $\mu = 0$, $n = 10$ and $H = 1000$. As expected, $\hat{\sigma}^2$ and $\hat{\sigma}^2_B$ did not improve over $\hat{\pi}$ too much and still have visible bias, while $\hat{\sigma}^2$ and $\hat{\sigma}^2_B$ appear to be unbiased and have nearly identical performance as the unbiased estimator $\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$.

Table 2.1 shows that the bias correction obtained by Strategies 2.1, 2.3 and 2.4 does not lead to an increase of the Mean-Squared Error (MSE) because the empirical MSE of $\hat{\sigma}^2_B$, $\hat{\sigma}^2_B$ or $\hat{\sigma}^2$ is smaller than the one of $\hat{\pi}$. The bias correction obtained by Strategies 2.1, 2.3 and 2.4 does not entail an important increase of the variance as a side-effect.
2.4 Asymptotic normality

We will now show that the corrected estimators $\hat{\theta}$ and $\tilde{\theta}$ defined in (2.2) and (2.4) are asymptotically normal under certain conditions that we will now state.

**Assumption A.10:** $\lim_{n \to \infty} g(\theta, n) \equiv g(\theta) = E[\hat{\pi} - \pi(\theta)]$ is continuously differentiable with respect to $\theta$ on $\Theta$.

**Assumption A.11:** Let $P(\theta) = \partial \pi(\theta)/\partial \theta^T$, then for any sequence $\theta^*$ such that $\theta^* \overset{P}{\to} \theta_0$ we have that $P(\theta^*) - P \overset{P}{\to} 0$ where $P$ is a sequence of matrices that do not depend on $\theta$. Moreover, we assume that $P^T(\theta) \Phi P(\theta)$ is full rank.

**Assumption A.12:** We suppose that $g(\theta)$ satisfies a central limit theorem, so that $\sqrt{n} g(\hat{\theta}) \overset{L}{\to} \mathcal{N}(0, V_{\theta})$.

**Theorem 2.5:** When $\lim_{H \to \infty}$, under Assumptions A.1 to A.5, and Assumptions A.10 to A.12, the corrected estimators $\hat{\theta}$ and $\tilde{\theta}$ defined in (2.2) and (2.4) respectively are asymptotically normal with the following asymptotic distribution:

$$\sqrt{n} (\hat{\theta} - \theta_0) \overset{L}{\to} \mathcal{N}(0, \Sigma)$$

and

$$\sqrt{n} (\tilde{\theta} - \theta_0) \overset{L}{\to} \mathcal{N}(0, \Sigma)$$

with $\Sigma = QV_{\theta}Q^T$ and $Q = \left(P^T(\theta) \Phi P(\theta)\right)^{-1} P^T(\theta) \Phi$.

**Proof:** In proof of Corollary 2.1, we show that $\lim_{H \to \infty} \hat{\pi}(\theta, n) = \pi(\theta, n)$. So, when $\lim_{H \to \infty}$, the indirect estimator $\hat{\theta}$ defined in (2.2) as

$$\hat{\theta} = \arg \min_{\theta \in \Theta} ||\hat{\pi} - \pi(\theta, n)||^2_\Phi$$

(2.21)

is equivalent to the classical GMM estimator defined as

$$\hat{\theta}_{GMM} = \arg \min_{\theta \in \Theta} ||\hat{\pi} - \pi(\theta, n)||^2_\Phi$$

(2.22)

i.e. $\hat{\theta} \overset{P}{\to} \hat{\theta}_{GMM}$, by the continuous mapping theorem (see e.g. DasGupta (2008), Theorem 1.14). Then, when $H \to \infty$, $\hat{\theta}$ has the same asymptotic normal distribution as a GMM estimator (see also Theorem 1.2. of Harris and Mátyás (1999)).
The proof is the same for $\hat{\theta}$ as $\lim_{H \to \infty} \hat{\pi}(\theta, nH) = \pi(\theta) \quad \square$

**Remark B:** It is not necessary for $H \to \infty$ in Theorem 2.5. If $H$ is a constant, then $\hat{\theta}$ is asymptotically normal with variance equal to

$$
\lim_{n \to \infty} \text{var} \left( \sqrt{n} \hat{\theta} \right) = \Sigma'
$$

with $\Sigma' = P(\theta) V_{\theta} P(\theta)^T + \frac{1}{H} P(\theta) V_{\theta} P(\theta)^T$.

*The same is true for $\tilde{\theta}$.**
Chapter 3

Robust Estimation of Income Distribution

3.1 Income distribution and inequality index

The distribution of income among people has been widely studied in Economics. One aim of this study is to understand how the total income in a given society is distributed among its economic units and which economic and social factors have an influence on this distribution. Another objective is to determine how a suitable measure of inequality can be obtained from this distribution. Statistical models have been used, for example, to explain how the income is generated and which factors influence its generation. Moreover these models have been used to estimate the parameter of probability density functions to describe income distribution and to compute inequality indexes. In this chapter, we will focus on the estimation of the parameters of income distribution models in a robust manner.

The first work on modeling income distributions was done by Pareto (1896). His statistical study of personal income distributions among different economies, states and time periods delivers the Pareto distribution which well describes the decreasing linear relation observed between the logarithm of the number of economic units having an income greater than a certain value and the logarithm of this income value. This distribution has an important property that all income distribution models should fulfil. From this initial work, a large number of probability functions has been proposed to describe income distributions, for example the Gamma distribution (Amoroso, 1925), the Lognormal (Gibrat, 1931), the Singh-Maddala (Singh and Maddala, 1976) or the Dagum (Dagum, 1977).

Statistical aspects in the analysis of income distributions, for example, have to deal with the choice of the model that best suits the data (using for instance the Cox statistic (Cox, 1961) or the test developed by Atkinson (1970)) and the estimation of the parameters of the chosen model. In this thesis only the last aspect will be treated. Estimates for the parameters can also be used to compute measures of income inequality. Many different measures have been proposed to quantify income inequality. One of the first indices was proposed by Gini (1912). It is also probably the most popular one. This index takes values between 0 and 1, a value of 1 denoting a perfectly unequal society. Following Gini’s work, many different indices have been proposed (see e.g. Cowell, 1977 and references therein). Inequality indices can also be computed without using parametric models for income distribution. Instead one can use the empirical distribution of the data without choosing a model.

However, if one chooses to use parametric models it could be interesting to use robust
methods. Indeed, income data are most often obtained from surveys and it is well-known that the quality of such data is often not perfect because the sources of error are numerous. It is also a field subject to extreme observations. Robust methods can be even more interesting if the parameter estimates are used as inputs in more complex economic models. Victoria-Feser and Ronchetti (1994) have proposed robust estimation methods for income data. They also developed these robust techniques for grouped data (Victoria-Feser and Ronchetti, 1997), which is often the format of income data. Cowell and Victoria-Feser (1996) also showed that computing inequality measures directly from the sample without using a parametric model is subject to robustness issues. Therefore, in order to obtain robust inequality indices, they suggest to rely on robust estimation of the income distribution parameters and then plug-in these robust estimates in the computation of inequality indices.

3.2 Robust estimation for income distribution

Let \( \theta_0 \) denote the parameter vector of an income distribution \( F_{\theta_0} \) that we wish to estimate. In the robustness paradigm, one supposes that the data generating distribution is not exactly the postulated model \( F_{\theta_0} \) but rather a deviation model of the form

\[
(1 - \varepsilon)F_{\theta_0} + \varepsilon F_C,
\]

where \( F_C \) is an unspecified contamination distribution function and \( \varepsilon \in [0, 1] \) is the amount of deviation from the postulated model \( F_{\theta_0} \). To estimate the income distribution \( F_{\theta_0} \) and use it to study aspects of income inequality such as poverty, a key component is to understand how \( \varepsilon > 0 \) affects the estimation of \( \theta_0 \) and hence the conclusions of the study. If \( \varepsilon \) is large, which means that a large proportion of the data are generated from \( F_C \), an arbitrary contamination distribution other than \( F_{\theta_0} \), then we may conclude that the analysis of the model \( F_{\theta_0} \) will not yield useful results and we may want to use another model. However, if \( \varepsilon \) is relatively small, we may expect the analysis of \( F_{\theta_0} \) to be indicative. In other words, the analysis should have certain robustness properties under model deviation.

Victoria-Feser and Ronchetti (1994) used the Optimal B-Robust Estimators (OBRE) (Hampel et al., 1986) as robust estimators of income distribution, which are the most efficient estimators among the class of M-estimators with bounded Influence Function (IF). The IF is a measure of local robustness and describes the relative change in an estimator caused by a small amount of contamination at any point. It can also be viewed as a linearization of the asymptotic bias of the estimator caused by contamination in the data. Therefore, a robust estimator must have a bounded IF, which is the case of the OBRE estimators.

The OBRE estimator is the root in \( \theta \) of the estimating equation

\[
\sum_{i=1}^{n} \psi(x_i; \theta) = \sum_{i=1}^{n} A(\theta) [s(x_i; \theta) - a(\theta)] w(x_i; \theta, c) = 0,
\]

where \( s(x; \theta) \equiv \partial \log \{f(x; \theta)\}/\partial \theta \) is the score function with \( f(x; \theta) \) being the probability density function of \( F_{\theta} \), and \( w(x; \theta, c) \) being a weight function that bounds \( A(\theta) [s(x; \theta) - a(\theta)] \), such as the Huber or the Tukey bisquare weight function (see e.g. Hampel et al., 1986).
The constant $c$ adjusts the bound on the IF where the smaller the $c$, the more robust the estimator is, but also the less efficient it is compared to the MLE which corresponds to the case where $c = \infty$ for the Huber weight function. The $p \times p$ matrix $A(\theta)$ and the $p \times 1$ vector $a(\theta)$ are defined implicitly by

$$
\mathbb{E} \left[ \psi (x_i; \theta) \psi (x_i; \theta)^T \right] = \left( A(\theta)^T A(\theta) \right)^{-1},$

$$
\mathbb{E} [\psi (x_i; \theta)] = 0.
$$

(3.3)

The vector $a(\theta)$ is a consistency correction, whereas $A(\theta)$ is often chosen to maximize the estimation efficiency.

However, the OBRE is in general difficult to compute, mainly because the integrals in (3.3) are not analytically tractable for income distribution models. Due to these numerical difficulties, the OBRE is hardly applicable for “complex” densities.

As in Guerrier et al. (2015b), we propose to use a simpler robust starting estimator that is inconsistent but easy to compute and to correct for consistency by means of indirect inference. We propose a weighted MLE $\hat{\pi}$ which is the solution of

$$
\sum_{i=1}^{n} \psi(x_i; \pi) = \sum_{i=1}^{n} w(x_i; \pi, c) s(x_i; \pi) = 0,
$$

(3.4)

where $w(x_i; \pi, c)$ are weights such that $w(x_i; \pi, c) s(x_i; \pi)$ is bounded for all $i$. Genton and Ronchetti (2003) showed that if the starting estimator has a bounded IF then the estimator obtained from indirect inference has an IF bounded by the same constant. Comparing equation (3.2) and (3.4) it is clear that the weighted MLE corresponds to the OBRE with $A(\theta) = I$ and $a(\theta) = 0$. The last equality implies that the weighted MLE is not expected to be a consistent estimator of $\theta_0$ if the density of the model is asymmetric because, in this case, the expectation of the weighted score function is not equal to zero at the model. This is often the case for income distributions since income data are generally asymmetric.

To derive the statistical properties of the bias reduction techniques presented in Section 2.2 we made the assumption that $\hat{\pi} \xrightarrow{D} \pi_0$. The weighted MLE is an $M$-estimator. Huber (1967) uses the following assumptions to show the consistency of an $M$-estimator.

**Assumption A.13:** For each fixed $\pi \in \Theta$, a compact parameter set, $\psi(x; \pi)$ is measurable.

**Assumption A.14:** The expected value $\mathbb{E}[\psi(x; \pi)]$ exists for all $\pi \in \Theta$, and has a unique zero at $\pi = \pi_0$.

**Assumption A.15:** As the neighbourhood $U$ of $\pi$ shrinks to $\{\pi\}$

$$
\mathbb{E} \left[ \sup_{\pi \in U} |\psi(x, \pi_j) - \psi(x, \pi_j)| \right] \to 0.
$$

**Theorem 3.1:** If the Assumptions A.13 to A.15 hold, then $\hat{\pi}$ converges uniformly in probability to $\pi_0$.

**Proof:** See proof of Theorem 2 in Huber (1967).

Assumptions A.13 to A.15 are mild assumptions to show the consistency of $M$-estimators. Assumption A.14 refers to the existence and the uniqueness of the solution.
of the $\psi$ function and Assumption A.15 implies the continuity of the expectation of this function. These assumptions are difficult to verify. Stronger assumptions that are easier to verify can be used, as in Duncan (1987). For example, in the latter the assumption that $\psi(x; \pi)$ is continuous is made instead of Assumption A.15.

Using simulation based bias correction methods is then a simple way to obtain robust estimators for income distributions. Indeed it is in general quite easy to compute a weighted MLE as a robust starting estimator but, in most cases, it is not consistent and needs to be corrected for its asymptotic bias through simulations. The resulting estimator is thus consistent and robust.

\section{Simulation study}

In this section we will present the results of three simulation studies. The first two study the robust estimation of the parameters of a Lomax (Simulation 3.1) and a Dagum (Simulation 3.2) distribution respectively. The Lomax distribution is a two-parameter distribution with the following density function

$$f(x; q, b) = \frac{q}{b} \left(1 + \frac{x}{b}\right)^{-(q+1)}, \quad x > 0,$$

with $q$ being the shape parameter and $b$ the scale parameter, with $q > 0$ and $b > 0$. It is a special case of a Pareto type II distribution. The Dagum is a more flexible distribution than the Lomax as it has a third parameter. It is a generalized Beta 2 with one of the three shape parameters equal to one. Its density function is as follows

$$f(x; a, b, p) = \frac{a p x^{a p - 1}}{b^p [1 + (x/b)^{a p + 1}], \quad x > 0},$$

with $a > 0$, $b > 0$ and $p > 0$. These two simulation studies are designed to illustrate the bias correction performed by the two indirect estimators, $\hat{\theta}$ which is based on $H$ samples of size $n$, and $\tilde{\theta}$, which is based on one sample of size $nH$. As a starting estimator we used the weighted MLE ($\hat{\pi}$) presented in (3.4) where $s(x_i; \pi)$ is the score function, i.e. the derivative of the log-density with respect to each parameter. The score functions of the Lomax are:

$$s_1(x; q, b) = \frac{\partial}{\partial q} \log f(x; q, b) = \frac{x(q + 1)}{(b^2(x/b + 1))} - \frac{1}{b},$$

$$s_2(x; q, b) = \frac{\partial}{\partial b} \log f(x; q, b) = \frac{1}{q} - \log(\frac{1}{b}x + 1).$$
For the Dagum distribution, the score functions are given by

\[
\begin{align*}
    s_1(x; a, b, p) &= \frac{\partial}{\partial a} \log f(x; a, b, p) \\
    &= a \log \left( \frac{x}{b} \right) (1 + p) - a \log \left( \frac{x}{b} \right) (1 + p) + a p (\log(b) - \log(x)) - 1 , \\
    s_2(x; a, b, p) &= \frac{\partial}{\partial b} \log f(x; a, b, p) = a \frac{b}{b} - a \frac{(p + 1)}{b}, \\
    s_3(x; a, b, p) &= \frac{\partial}{\partial p} \log f(x; a, b, p) = a \log(x) - a \log(b) - \log\left( \left( \frac{x}{b} \right)^a + 1 \right) + \frac{1}{p}.
\end{align*}
\]

We use the Tukey bisquare weight function, which is defined as

\[
w(x_i; \pi, c) = \begin{cases} 
    \left( 1 - \left( \frac{||s(x_i; \pi)||}{c} \right)^2 \right)^2, & ||s(x_i; \pi)|| \leq c, \\
    0, & ||s(x_i; \pi)|| > c.
\end{cases}
\]

The constant \( c \) was chosen via Monte Carlo simulations to provide an efficiency of 80% compared to the MLE. We implement an algorithm to search \( c \) on a grid. For each point on that grid we compute the weighted MLE and the MLE obtained on data simulated under the true parameters. The efficiency is equal to the ratio of the determinant of the empirical covariance matrices of both estimators at the power one over the number of parameters (Serfling, 2009). Both indirect estimators are supposed to correct the asymptotic bias of \( \hat{\pi} \) which is known to be an inconsistent estimator for the parameters of an asymmetric distribution, as the Lomax and the Dagum distributions.

The third simulation study compares the MLE estimator, the weighted MLE and the two indirect estimators using the MLE as a starting estimator when the sample size is small. It is a well-known issue that the MLE is biased in small samples for the estimation of the parameters of income distributions (see e.g. Giles et al., 2013 for the Lomax distribution). Therefore the finite sample bias of the MLE should be corrected only using the indirect inference estimator obtained with \( H \) samples of size \( n \) as it has a bias of order \( O(n^{-2}) \) at least, whereas the indirect inference estimator obtained with one sample of size \( nH \) has a bias of order \( O(n^{-1}) \) at best.

Simulation 3.1. The values of the true parameters for the Lomax density function are respectively \( b = 5 \) and \( q = 2 \). We compute \( \hat{\pi}_{ML} \) as a benchmark estimator, the weighted MLE \( \hat{\pi}, \hat{\theta} \) (Strategy 2.1) and \( \tilde{\theta} \) (Strategy 2.2). For all simulations we use \( c = 13 \) in the weighted MLE estimating equations. We simulated data from two different settings:

- in the first one \( n = 1000 \) to compare the asymptotic bias correction properties of \( \hat{\theta} \) and \( \tilde{\theta} \),
- in the second one \( n = 1000 \) but with random contamination of 1% of the data which are multiplied by 1000 to test the robustness of \( \hat{\theta} \) and \( \tilde{\theta} \).

We always use \( H = 100 \) for \( \hat{\theta} \) and \( \tilde{\theta} \).
Figure 3.1: Empirical distribution of the different estimators (defined in Simulation 3.1) obtained with 100 samples of size \( n = 1000 \) from a Lomax distribution with parameters \( b = 5 \) (left panel) and \( q = 2 \) (right panel). Graphs (a) and (b) correspond to the setting without contamination whereas graphs (c) and (d) correspond to the setting where a randomly chosen 1\% of the data are multiplied by 1000.
3.3. Simulation study

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MSE</th>
<th>Bias</th>
<th>Parameter</th>
<th>MSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\pi}$</td>
<td>1.63 $\cdot 10^0$</td>
<td>$-8.58 \cdot 10^{-1}$</td>
<td>$\hat{\pi}$</td>
<td>1.79 $\cdot 10^{-1}$</td>
<td>$2.14 \cdot 10^0$</td>
</tr>
<tr>
<td>$\hat{\theta}$ (Strategy 2.2)</td>
<td>8.03 $\cdot 10^{-1}$</td>
<td>$-1.26 \cdot 10^0$</td>
<td>$\hat{\theta}$ (Strategy 2.2)</td>
<td>6.35 $\cdot 10^{-2}$</td>
<td>$1.74 \cdot 10^0$</td>
</tr>
<tr>
<td>$\hat{\theta}$ (Strategy 2.1)</td>
<td>7.34 $\cdot 10^{-1}$</td>
<td>$-1.30 \cdot 10^0$</td>
<td>$\hat{\theta}$ (Strategy 2.1)</td>
<td>5.63 $\cdot 10^{-2}$</td>
<td>$1.70 \cdot 10^0$</td>
</tr>
</tbody>
</table>

Table 3.1: Empirical MSE and bias of the different estimators (defined in Simulation 3.3) obtained with 100 samples of size $n = 1000$ from a Lomax distribution without contamination.

Figure 3.1 shows the empirical distribution of the different estimators obtained with 100 Monte Carlo replications. Part (a) and (b) of Figure 3.1 corresponds to the 1st setting and show that both strategies 2.1 and 2.2 correct the asymptotic bias of the weighted MLE. This was expected from theorem 2.1 and 2.2, as both estimators have a rate of convergence of $O(n^{-1})$ at least. The performance of both indirect estimators is similar to the asymptotically optimal estimator, i.e. the MLE. Their empirical variances are also similar to the one of the starting estimator. When the data is contaminated (part (c) and (d) of Figure 3.1) both indirect estimators correct the asymptotic bias of the weighted MLE and are still unbiased despite the contamination, thus leading to robust estimators. Not surprisingly, the MLE exhibits a clear bias as it is not robust.

Table 3.1 presents the empirical MSE for the weighted MLE and the estimators obtained through Strategy 2.1 and Strategy 2.2 using the weighted MLE as a starting estimator. The MSE of the corrected estimators $\hat{\theta}$ and $\tilde{\theta}$ are approximately half of the MSE of the weighted MLE. This suggests that the bias correction performed by Strategy 2.1 and Strategy 2.2 does not lead to an increase of the variance.

Simulation 3.2. The values of the true parameters for the Dagum density function are respectively $a = 2$, $b = 1$ and $p = 0.5$. We compute $\hat{\pi}_{ML}$ as a benchmark estimator, the weighted MLE $\hat{\pi}$, $\hat{\theta}$ and $\tilde{\theta}$. For all simulations we use $c = 9$ in the weighted MLE estimating equations. We simulated data from two different settings:

- in the first one $n = 1000$ to compare the asymptotic bias correction properties of $\hat{\theta}$ and $\tilde{\theta}$,
- in the second one $n = 1000$ but with random contamination of 1% of the data which are multiplied by 3000 to test the robustness of $\hat{\theta}$ and $\tilde{\theta}$.

We always use $H = 100$ for $\hat{\theta}$ and $\tilde{\theta}$.

Figure 3.2 shows the empirical distribution of the different estimators obtained with 100 Monte Carlo replications. As was already observed in the results of Simulation 3.1, indirect estimators provide robust and unbiased estimators using a simple starting estimator. Yet the robust estimation of the three parameters distribution like the Dagum is known to be very complex (Victoria-Feser, 1993). We emphasized that the robust estimation approach taken here is not specific to income distributions. It can be used for any model for which a score function can be specified. This opens the door to simple computation of robust estimators especially for complex models.

Simulation 3.3. We simulate $n = 50$ data from a Lomax with true parameters $b = 5$ and $q = 2$ to compare the small sample bias correction properties of $\hat{\theta}$ and $\tilde{\theta}$. We compute
Figure 3.2: Empirical distribution of the different estimators (defined in Simulation 3.2) obtained with 100 samples of size \( n = 1000 \) from a Dagum distribution with parameters \( a = 2 \) (graphs (a) and (d)), \( b = 1 \) (graphs (b) and (e)) and \( p = 0.5 \) (graphs (c) and (f)). Graphs on the top row correspond to the setting without contamination whereas graphs on the bottom row correspond to the setting where a randomly chosen 1% of the data are multiplied by 3000.
3.3. Simulation study

Figure 3.3: Empirical distribution of the different estimators (defined in Simulation 3.3) obtained with 100 samples of size $n = 50$ from a Lomax distribution with parameters $b = 5$ (graph (a)) and $q = 2$ (graph (b)).

$\hat{\pi}_{ML}$, the weighted MLE $\hat{\pi}$, $\hat{\theta}$ and $\tilde{\theta}$, the latter two with $\hat{\pi}_{ML}$ as a starting estimator. For all simulations we use $c = 13$ in the weighted MLE estimating equations and $H = 100$ in the indirect inference procedure.

Figure 3.3 shows the empirical distribution of the different estimators obtained with 100 Monte Carlo replications. The estimation of the scale parameter $b$ by the MLE is biased as the sample size is small. This small sample bias is corrected only by Strategy 2.1 and not by Strategy 2.2, as expected. Even if analytic second-order bias expressions have been proposed in the case of the Lomax distribution (Giles et al., 2013) it is worth noting that it is by far not the case for all income distribution whereas simulation based bias correction methods can always be applied.

The empirical MSE presented in Table 3.2 show that the bias correction performed by Strategy 2.1 decreases the bias and the MSE of the MLE due to the small sample size. This is particularly true for parameter $b$. Here again, the bias correction is not obtained at the cost of an increase of the variance.

In this simulation study, we did not compute the estimator obtained by iterative
bootstrap as it is asymptotically equivalent to the indirect inference estimator obtained by Strategy 2.1. Iterative bootstrap has to be preferred for models with many parameters, as optimization required by indirect inference is difficult to perform in such cases. When only asymptotic bias has to be corrected, as in Simulation 3.1 and 3.2, indirect inference using only one sample of size $nH$ (Strategy 2.2) is recommended as it takes less computational time.

<table>
<thead>
<tr>
<th>Parameter $b$</th>
<th>Parameter $q$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
</tr>
<tr>
<td>$\hat{\pi}_{ML}$</td>
<td>$2.41 \cdot 10^1$</td>
</tr>
<tr>
<td>$\hat{\theta}$ (Strategy 2.2)</td>
<td>$3.68 \cdot 10^1$</td>
</tr>
<tr>
<td>$\hat{\theta}$ (Strategy 2.1)</td>
<td>$1.74 \cdot 10^1$</td>
</tr>
</tbody>
</table>

Table 3.2: Empirical MSE and bias of the different estimators (defined in Simulation 3.3) obtained with 100 samples of size $n = 50$ from a Lomax distribution.
3.4 Case study

In order to illustrate the use of robust estimation for the study of income distributions and inequality indices, we apply our estimation method to income data issued from a longitudinal study of Swiss households, the Swiss Household Panel. We model the yearly households income (OECD equivalence scale) for two different years: 1999 and 2013. We remove missing data and rescale the data by a factor of $10^{-3}$. We fit a Dagum model, using MLE and weighted MLE corrected by Strategy 2.1. For the weighted MLE, we use a tuning constant that provides an efficiency of 80% compared to the MLE. For the indirect inference procedure, we chose $H = 100$.

Using the estimates of the parameters we also compute the Gini index, as an income inequality measure. For the Dagum model the Gini index is given by

$$I_G = \frac{\Gamma\left(p\right)\Gamma\left(2p + 1/a\right)}{\Gamma\left(2p\right)\Gamma\left(p + 1/a\right)} - 1.$$  (3.5)

Table 3.3: MLE and Robust estimates, with 95%CI, for the parameters of the Dagum distribution and associated Gini indices for the Swiss Household data.

<table>
<thead>
<tr>
<th>Year</th>
<th>Parameter</th>
<th>MLE</th>
<th>95% CI</th>
<th>Robust</th>
<th>95% CI</th>
<th>Empirical</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>$a$</td>
<td>3.25</td>
<td>[3.09;3.31]</td>
<td>3.21</td>
<td>[3.06;3.36]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>50.65</td>
<td>[48.02;51.55]</td>
<td>52.79</td>
<td>[50.62;54.98]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p$</td>
<td>0.78</td>
<td>[0.71;0.81]</td>
<td>0.71</td>
<td>[0.65;0.78]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gini Index</td>
<td>0.327</td>
<td>[0.318;0.330]</td>
<td>0.340</td>
<td>[0.331;0.348]</td>
<td>0.324</td>
<td>[0.314;0.327]</td>
</tr>
<tr>
<td>2013</td>
<td>$a$</td>
<td>3.73</td>
<td>[3.56;3.79]</td>
<td>3.65</td>
<td>[3.49;3.80]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>72.77</td>
<td>[69.48;73.93]</td>
<td>74.06</td>
<td>[71.35;76.78]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p$</td>
<td>0.75</td>
<td>[0.68;0.77]</td>
<td>0.72</td>
<td>[0.66;0.78]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gini Index</td>
<td>0.290</td>
<td>[0.282;0.293]</td>
<td>0.300</td>
<td>[0.293;0.307]</td>
<td>0.294</td>
<td>[0.282;0.299]</td>
</tr>
</tbody>
</table>

The estimates for the parameters of the Dagum distribution obtained by the MLE and the robust estimator as well as the Gini index computed from these estimates are presented in Table 3.3. The 95% Confidence Intervals (CI) are obtained by parametric bootstrap for the MLE and the robust estimator and by non-parametric bootstrap for the empirical estimation of the Gini index. The difference between the robust and the MLE estimates are quite small. Nevertheless, when looking at the 95% CI, there are differences between the MLE and the robust estimates for parameter $b$.

Figure 3.4 represents the two estimated distributions with the histogram of the empirical distribution for 1999 and 2013. Both estimated densities provide very good fits. The Dagum is a very flexible distribution and can accommodate very extreme observations. This can explain why there is little difference between the two estimation techniques despite quite extreme values. The same was been noticed by Victoria-Feser (1993) in the modeling of income data from the USA and the UK with the Dagum distribution.

The Gini indices are also quite similar, whether they are computed from MLE or robust estimates. The empirical version of the index is almost the same as the one obtained with MLE. Their 95% CI show that they are not different. The indices obtained
Figure 3.4: Robust (blue dashed line) and MLE (red solid line) density estimation of the Dagum model on the Swiss Household income data for 1999 and 2013.

The analysis of these data shows that robust and MLE estimation are different, even with an income distribution which is known to be flexible and so can accommodate quite extreme observations. The empirical version of the inequality index is quite similar to the one obtained by MLE and so robust estimation has a real added value.
Chapter 4

Generalized Linear Latent Variable Models

4.1 Latent variable models

Latent variables models have now become a very popular statistical techniques in various area of research such as psychology, social sciences or economics. They are powerful models to describe a set of multivariate data using one or more underlying variables and for exploring the interrelationships among observed variables. They are also widely used for the construction of measurement scales. Factor analysis (FA) (Jöreskog, 1967) belongs to the family of latent variables models. In such a model, manifest (or observed) variables, which are supposed to be multivariate normal, are linked to latent variables (or factors). MLE is performed from the sample covariance matrix of the manifest variables.

However, in many applied fields, the measures are very seldom taken on a normal scale. For example, survey data may contain variables which are measured on a binary or a categorical scale. In these cases, one approach that has been developed assumes that the manifest variables are indirect observations of normal underlying variables (Jöreskog, 1990, B. Muthén, 1978). One can apply the same ML estimation as for normal manifest variables using polychoric, tetrachoric or polyserial correlations but this is known to provide inconsistent estimators Huber et al. (2004).

Another approach, that was developed for example by Bartholomew (1984) and Moustaki and Knott (2000), proposes a more general framework inspired from GLM (McCullagh and Nelder, 1989) for latent variables models in the case of normal and non-normal observations called GLLVM. The manifest variables are assumed to follow, conditionally on the latent variables, a distribution that belongs to the exponential family, for example the multinomial distribution for ordinal data. One difficulty of this approach is that it requires multiple integration in the score equations of the MLE. Indeed, as the latent variables are not observed they must be integrated out from the likelihood function. When the manifest variables are not normal, the integral in the score equations have no closed form. Thus the maximization of the likelihood require numerical integration.

Various approximations have been proposed for the estimation of GLLVM. Moustaki and Knott (2000) used simple Gaussian quadrature as numerical approximation method but this turns out to be inaccurate when the number of latent variables is larger than two. Adaptive Gaussian quadrature (Rabe-Hesketh et al., 2002) substantially improve the estimation accuracy but is computationally intensive. These approximations are used for
example in the softwares Mplus and Stata (GLLAM package). Huber et al. (2004) use the Laplace approximation. Both Laplace approximation and adaptive Gaussian quadrature lead to unbiased estimators. However these methods have an important drawback from a practical point of view: they require a lot of tuning in order to make the optimization of the approximated likelihood successful.

In order to avoid part of this numerical problems, we propose instead to first compute a starting estimator which is not consistent but easy to compute and then correct for its bias via Strategy 2.4.

### 4.2 GLLVM

Latent variable models are used to find a set of factors (or latent variables) denoted by $z_k$, $k = 1, \ldots, q$, that conveys most of the information contained in a set of response variables denoted by $x^{(j)}$, $j = 1, \ldots, p$, with $p$ (much) larger than $q$. A crucial assumption is that all the dependence structure of the manifest variables is explained by the factors. This is known as the assumption of conditional or local independence.

In GLLVM the conditional expectation of the manifest variables is linked to the latent variables through a link function in the following manner.

$$
\nu_j(\mathbb{E}(x^{(j)}|z)) = \lambda^{(j)T}z.
$$

Here $\lambda^{(j)} = (\lambda_0^{(j)}, \ldots, \lambda_q^{(j)})^T = (\lambda_0^{(j)}, \lambda_{(2)}^{(j)})^T$, where $\lambda_{(2)}^{(j)}$ are the loadings, $z = (1, z_1, \ldots, z_q)^T = (1, z_{(2)})^T$ and $\nu_j$ are the link functions. The latent variables can be seen as covariates in the linear predictor $\lambda^{(j)T}z$.

The conditional distribution of the manifest variables given the latent ones is supposed to belong to the exponential family, i.e.

$$
g_j(x^{(j)}|z) = \exp \left( \frac{x^{(j)}u_j(\lambda^{(j)T}z) - b(u_j(\lambda^{(j)T}z))}{\phi_j} + c_j(x^{(j)}, \phi_j) \right),
$$

where $\phi_j$ is the scale parameter and $u_j(\lambda^{(j)T}z)$ is the so-called canonical parameter. The form of the functions $u_j(\cdot)$, $b(\cdot)$ and $c_j(\cdot)$ depends on the specific distribution $g_j$, $u_j(\cdot)$ being the identity function when $\nu_j$ is the canonical link.

Due to the assumption of conditional independence, the joint conditional distribution of the manifest variables is

$$
\prod_{j=1}^{p} g_j(x^{(j)}|z).
$$

The density of the latent variables, denoted by $h(z_{(2)})$ is assumed to be multivariate standard normal. This implies that the latent variables are independant, but this assumption can be relaxed. The joint distribution of the manifest and latent variables is then

$$
\prod_{j=1}^{p} g_j(x^{(j)}|z)h(z_{(2)}).
$$
4.3 Starting estimator and iterative bootstrap algorithm

As the latent variables are not observed, their realizations are treated as missing and are integrated out. This gives the following marginal density for the manifest variables

\[ f_{\lambda, \phi}(x) = \int \left\{ \prod_{j=1}^{p} g_j(x^{(j)} | z) \right\} h(z_{(2)}) dz_{(2)}. \]  

(4.5)

Given a sample of \( n \) observations \( x_1, \ldots, x_n \) where \( x_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \), \( i = 1, \ldots, n \), the log-likelihood of the loadings \( \lambda \) and the scale parameters \( \phi \) is

\[
l(\lambda, \phi | x) = \sum_{i=1}^{n} \log f_{\lambda, \phi}(x_i),
\]

\[
= \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \prod_{j=1}^{p} \exp \left\{ \frac{x_i^{(j)} u_j(\lambda^{(j)} T z) - b_j(u_j(\lambda^{(j)} T z))}{\phi_j} \right. 
+ c_j(x^{(j)}, \phi_j) \left. \right\} h(z_{(2)}) dz_{(2)}.
\]

(4.6)

MLE for \( \lambda \) and \( \phi \) are the zeros of the derivatives of (4.6) with respect to \( \lambda \) and \( \phi \) and necessitate approximations of the integrals.

### 4.3 Starting estimator and iterative bootstrap algorithm

To motivate the choice of our starting estimator we will first expose the estimator proposed by Huber et al. (2004) which is based on the application of the Laplace method to GLLVM, and explain how it could be roughly approximated. The marginal density given in (4.5) can also be expressed as

\[ f_{\lambda, \phi}(x) = \int \exp (pQ(\lambda, \phi, z, x)) dz_{(2)}, \]  

(4.7)

where

\[
Q(\lambda, \phi, z, x) = \frac{1}{p} \left[ \sum_{j=1}^{p} \left( \frac{x_i^{(j)} u_j(\lambda^{(j)} T z) - b_j(u_j(\lambda^{(j)} T z))}{\phi_j} + c_j(x^{(j)}, \phi_j) \right) \right. 
- \left. \frac{z_i^{(2)} z_{(2)}}{2} - \frac{q}{2} \log(2\pi) \right].
\]

(4.8)

The idea behind Laplace approximation is to approximate \( Q(\lambda, \phi, z, x) \) by a second-order Taylor expansion around the value of \( z \) that maximizes \( Q(\cdot) \). By applying it to (4.7) we get

\[ f_{\lambda, \phi}(x) = \left(\frac{2\pi}{p}\right)^{q/2} \det [-U(\hat{z})]^{-1/2} \exp (pQ(\lambda, \phi, \hat{z}, x)) \left(1 + O((p^{-1}))\right), \]  

(4.9)

with

\[ U(\hat{z}) = \frac{\partial^2 Q(\lambda, \phi, z, x)}{\partial z^T \partial z} \bigg|_{z=\hat{z}} = -\frac{1}{p} \Gamma(\lambda, \phi, \hat{z}), \]
Chapter 4. Generalized Linear Latent Variable Models

\( \hat{z} \) is the maximum of \( Q(\lambda, \phi, z, x) \). From (4.8) and (4.9) the approximate log-likelihood function is then

\[
\tilde{l}(\lambda, \phi|x_i) = \sum_{i=1}^{n} \left( - \frac{1}{2} \log(\det[\Gamma(\lambda, \phi, \hat{z}_i)]) + \sum_{j=1}^{p} \left( x_{i}^{(j)} u_j(\lambda^{(j)T} \hat{z}_i) - b(u_j(\lambda^{(j)T} \hat{z}_i)) \right) + c_j(x_{i}^{(j)}, \phi_j) \right) - \frac{\hat{z}_i^{T}(2) \hat{z}_i(2)}{2} - \frac{q}{2} \log(2\pi) \quad (4.10)
\]

The score equations are obtained by differentiating (4.10) with respect to \( \lambda \) and \( \phi \). As these equations also depend on \( \hat{z}_i \) the algorithm used to obtain the Laplace approximated MLE for \( \lambda \) and \( \phi \) must alternatively maximizes (4.8) with respect to \( z \) and (4.10) with respect to \( \lambda \) and \( \phi \).

Ignoring the first term in the the approximate log-likelihood in (4.10) yields to a function whose maximum in \( \lambda \) and \( \phi \) corresponds to the PQL estimator for these parameters in the context of GLMM (Breslow and Clayton, 1993). Indeed, maximizing (4.10) without the first term for \( \lambda \) and \( \phi \) provides the same estimators that would result from the maximization of a “likelihood” where the latent scores \( z_i \) would be considered as parameters and which could be written as

\[
l^{\star}(\lambda, \phi, z_i|x_i) = \sum_{i=1}^{n} \left( \sum_{j=1}^{p} \left( x_{i}^{(j)} u_j(\lambda^{(j)T} z_i) - b(u_j(\lambda^{(j)T} z_i)) \right) + c_j(x_{i}^{(j)}, \phi_j) \right) - \frac{z_i^{T}(2) z_i(2)}{2} - \frac{q}{2} \log(2\pi), \quad (4.11)
\]

Equation (4.11) also correspond to what would be the \( h \)-likelihood proposed by Lee and Nelder (1996) for GLMM in the context of latent variable models. Neither PQL nor \( h \)-likelihood is a classical likelihood as the latent scores (or the random effects in the GLMM) are treated as parameters. However, the \( \hat{z}_i \) obtained by the maximization of the \( h \)-likelihood are the same than that obtained by the maximization of \( Q(\cdot) \) in Huber et al. (2004) framework. They are also the Empirical Bayes modal, because they are the value of \( z_i \) that maximizes the posterior distribution of the latent scores given the data and the estimated parameter \( \lambda \) and \( \phi \).

If the \( z_i \) were considered as fixed in (4.11), the maximum of \( l^{\star} \) for \( \lambda \) and \( \phi \) would be the same as the MLE of a GLM, in which the value for the covariates would be the known values of \( z_i \). Therefore, as a starting estimator, we proposed to find estimates for the loadings and the scale parameter by fitting a GLM on the data with predictions of the latent scores as covariates. These predictions for the \( z_i \) are found in a preliminary step where we perform a FA on the data as if they were normal and compute the Bartlett estimator of the latent scores (Bartlett, 1950). A similar idea has been used in Sardy and Victoria-Feser (2012). This initial estimator is thus obviously not consistent but easily computed and implemented.

We then correct its bias by the iterative bootstrap defined in Strategy 2.4. The number of parameters in GLLVM is generally high, which precludes the use of indirect estimators.
as they necessitate an optimization over the vector of parameters. Iterative bootstrap is asymptotically equivalent to indirect inference with $H$ samples of size $n$ and thus corrects both asymptotic and finite sample bias.

The algorithm used to implement the iterative bootstrap in the case of GLLVM is given in Algorithm 1.

**Algorithm 1** Iterative bootstrap for GLLVM estimation.

**Require:**
- Data set $x = x_1, \ldots, x_n$

1: procedure **Starting estimator** $\hat{\pi}$ on $x$
2: Do a FA on $x$
3: Compute Bartlett scores $\hat{z}_{FA}$ from loadings and uniqueness obtained by FA
4: Do a GLM using $x$ as responses and $\hat{z}_{FA}$ as covariates
5: Store $\hat{\pi} = [\hat{\lambda}_{GLM}, \hat{\phi}_{GLM}]$
6: Set $i \leftarrow 1$
7: Set $\tilde{\theta}_B^{(i)} \leftarrow \hat{\pi}$
8: Generate $H n \times 1$ vectors $z^{(k)} \sim N(0, 1)$, $k = 1, \ldots, q$
9: Set the seed
10: while $\text{crit} > \text{eps}$ do
11: for $h : 1 \rightarrow H$ do
12: Generate $n \times 1$ vectors $x^{(j)} \sim g_j$, $j = 1, \ldots, p$, using $z^*$ and $\tilde{\theta}_B^{(i)}$
13: Compute $\hat{\pi}_h(\tilde{\theta}_B^{(i)}, n) = \frac{1}{H} \sum_{h=1}^{H} \hat{\pi}_h(\tilde{\theta}_B^{(i)}, n)$
14: $\tilde{\theta}_B^{(i+1)} \leftarrow \tilde{\pi} + \left(\tilde{\theta}_B^{(i)} - \bar{\pi}(\tilde{\theta}_B^{(i)}, n)\right)$
15: $\text{crit} \leftarrow ||\tilde{\theta}_B^{(i+1)} - \tilde{\theta}_B^{(i)}||_2$
16: $i \leftarrow i + 1$
17: return $\tilde{\theta}_B \leftarrow \tilde{\theta}_B^{(i+1)}$

4.4 Simulation study

In this section we will present the results from a simulation study performed with ordinal manifest variables. In this case $g_j(x^{(j)}|z)$ is the multinomial distribution. There is no scale parameter $\phi_j$ for the multinomial. Specifically, for $j$, let $m_j$ be the number of categories and $\gamma_s^{(j)}$ be the probability that the manifest variable is smaller or equal to $s$, where $s = 1, \ldots, m_j - 1$. We set a proportional odds model

$$\gamma_s^{(j)} = \frac{e^{\lambda_0^{(j)}}}{1 + e^{\lambda_0^{(j)}}},$$

where $\lambda_0 = (\lambda_0^{(j)}, \lambda_1^{(j)})^T = (\lambda_0^{(j)}, \lambda_1^{(j)}, \ldots, \lambda_{m_j}^{(j)})^T$, $\lambda_0^{(j)}$ being the threshold for the $s$th category with 

$$-\infty < \lambda_0^{(j)} \leq \lambda_0^{(j)} \leq \ldots \lambda_0^{(j)} < \infty.$$ 

To compute the starting estimator, we have to perform an ordinal GLM to obtain $\hat{\lambda}_{GLM}$. For computational reasons, we implement a modified version of the MLE of an ordinal
GLM. First, the starting estimates for the thresholds $\lambda_{0s}^{(j)}, s = 1, \ldots, m^{(j)} - 1$ are computed using the empirical cumulative log-odds, i.e. the one obtained at $z_{(2)} = 0$ as

$$\hat{\lambda}_{0s}^{Emp} = \log \left( \frac{\hat{\gamma}_{0s}^{Emp}}{1 - \hat{\gamma}_{0s}^{Emp}} \right),$$

(4.13)

with $\hat{\gamma}_{0s}^{Emp} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_{i}^{(j)} \leq s)$. Then, $\hat{\lambda}_{GLM}$ are found as the MLE estimator of an ordinal GLM. The log-likelihood for one manifest variable $x^{(j)}$ is written as

$$\sum_{i=1}^{n} \left( \sum_{s=1}^{m^{(j)}-1} x_{i,s}^{(j)} \log \frac{\gamma_{i,s}^{(j)}}{\gamma_{i,s+1}^{(j)} - \gamma_{i,s}^{(j)}} - x_{i,s+1}^{(j)} \log \frac{\gamma_{i,s+1}^{(j)}}{\gamma_{i,s+1}^{(j)} - \gamma_{i,s}^{(j)}} \right),$$

where $x_{i,s}^{(j)} = \mathbb{I}(x_{i}^{(j)} \leq s)$ and $\gamma_{i,s}^{(j)}$ are obtained with $\hat{\lambda}_{0s}^{Emp}$ and $\hat{z}_{i,FA}$.

In this simulation study, the estimator obtained by iterative bootstrap is compared in terms of bias to the MLE obtained via Mplus (L. Muthén and B. Muthén, 1998-2011) which uses adaptive Gauss quadrature to approximate the integrals in (4.6).

**Simulation 4.1.** We generate 120 samples of size $n = 300$ from a model with

- three independent latent variables, $q = 3$
- fifteen manifest variables, $p = 15$
- five categories by manifest variables, $m^{(j)} = 5$

The true parameter values are given in Table 4.1. The thresholds are not constrained to have the same values across manifest variables, even if we chose the same true value for the threshold of the same level across all manifests. To ensure unique solution in $\lambda$ we have to fix at least $q(q-1)/2$ loadings at 0 (Huber et al., 2004). Otherwise, a rotation of $\lambda$ by an orthogonal matrix will also be a solution for the same model. In this example we set 15 loadings to 0. This yields $(m^{(j)} - 1 + q) \times p - 15 = 90$ parameters to estimate.

We compute the starting estimator $\hat{\pi}$, the iterative bootstrap bias corrected estimator $\tilde{\theta} \ (\text{Strategy 2.4})$ and the approximated MLE obtained via Mplus $\hat{\theta}_{ML}$ as a benchmark estimator.

Figure 4.1 shows the empirical distribution of the different estimators (centered at the true value) for four loadings (the $\lambda_{0s}^{(j)}$’s). For other loading estimates, the results are very similar. As expected, the starting estimator ($\hat{\pi}$) is biased, but applying the iterative bootstrap provides a nearly unbiased estimator ($\tilde{\theta}_{B}$), which is comparable to the one obtained using adaptive quadrature to approximate the likelihood function ($\hat{\theta}_{ML}$), with a slight loss of efficiency. The same can be said when we inspect the threshold estimates (the $\lambda_{0s}^{(j)}$’s), see Figure 4.2.
Figure 4.1: Empirical distribution of the estimates, centered at the true value, of four loadings obtained by the starting inconsistent estimator ($\hat{\pi}$), the estimator based on iterative bootstrap ($\tilde{\theta}_B$) and the approximated MLE obtained using Mplus ($\hat{\theta}_{MLE}$).
Figure 4.2: Empirical distribution of the bias of the estimates, centered at the true value, of four thresholds obtained by the starting inconsistent estimator (\(\hat{\pi}\)), the estimator based on iterative bootstrap (\(\tilde{\theta}_B\)) and the MLE obtained using Mplus (\(\hat{\theta}_{ML}\)).
Table 4.1: True parameter values used in Simulation 4.4.

<table>
<thead>
<tr>
<th>(\lambda_{01})</th>
<th>(\lambda_{02})</th>
<th>(\lambda_{03})</th>
<th>(\lambda_{04})</th>
<th>(\lambda_1)</th>
<th>(\lambda_2)</th>
<th>(\lambda_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.8</td>
<td>0</td>
<td>-1.0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.0</td>
<td>0</td>
<td>2.0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>-2.0</td>
<td>2.3</td>
<td>1.0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.2</td>
<td>0</td>
<td>-1.0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>-1.6</td>
<td>-1.6</td>
<td>0.9</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.0</td>
<td>0</td>
<td>-0.7</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>0</td>
<td>-0.7</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>-1.2</td>
<td>2.7</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>0</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.3</td>
<td>-1.0</td>
<td>-0.9</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>-0.8</td>
<td>0.8</td>
<td>2.0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>0</td>
<td>-1.8</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
<td>1.6</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>2.0</td>
<td>0</td>
<td>-1.8</td>
</tr>
</tbody>
</table>
4.5 Goodness-of-fit test

4.5.1 Goodness-of-fit test statistic

Goodness-Of-Fit (GOF) measures are an important tool to assess the quality of latent variable models. In such models, it is not totally satisfactory to compare models with criteria such as the Akaike Information Criterion (AIC). Indeed, it is quite difficult to justify which models are to be used in the comparison, as not only the number of loadings different from 0 is unknown, i.e. the fact that a particular manifest variable is linked or not to a particular latent variable, but also the number of latent variables. To assess if the postulated theoretical model offers an acceptable fit to the data, as for instance in the case of the construction of measurement scales, one really need a GOF measure.

Several GOF indexes are available for latent variable models when manifest variables are normally distributed. When they are not multivariate normal, GOF measures have been proposed based on the distance between the covariance matrix constrained by the model and the unconstrained covariance matrix, estimated with polychoric correlations for example. However in the non-normal case their null asymptotic distribution is no more valid (see e.g. Marsh et al., 2004).

Conne et al. (2010) propose a GOF test for GLLVM based on the comparison of a distance among the latent scores predicted by the model estimated with the Laplace approximation and the corresponding distance among the observed manifest variables. They use distances developed by Kaufman and Rousseeuw (1990) for cluster analysis, which take into account the nature of the manifest variables. The p-value of the test is computed by a modified parametric bootstrap. Both simulations under the null hypothesis and under the alternative show the good statistical properties of this test. Nevertheless, both the estimation of the model and the computation of the p-value are computationally intensive, which precludes the applicability of this test.

Here we propose a GOF measure based on the distance between \( \hat{\pi} \) and \( \bar{\pi}(\tilde{\theta}_B,n) \) in the Mahalanobis sense. The idea is that, at the solution, both quantities should be close if the model is correct. However, this should not happen when the model is wrong, in which case \( \hat{\pi} \) is implied by another model than the one used to compute \( \bar{\pi}(\tilde{\theta}_B,n) \). The GOF test statistic that we propose is equal to

\[
S = ||\hat{\pi} - \bar{\pi}(\tilde{\theta}_B,n)||^2_{\Phi}. \tag{4.14}
\]

\( S \) can be seen as the value of the objective function of estimator \( \hat{\theta} \), evaluated at \( \theta = \tilde{\theta}_B \). As \( \tilde{\theta}_B \) obtained by Strategy 2.4 and \( \hat{\theta} \) obtained by Strategy 2.1 are equivalent when \( H \) is large, \( S \) should be close to 0 if \( \tilde{\theta}_B \) correctly estimates the value of \( \theta \) that generates the data. \( \Phi \) is equal to the identity matrix as it is not used in Strategy 2.4.

The same kind of idea has been used by Hansen (1982) for testing over-identifying restrictions in GMM models and by Gouriéroux et al. (1993) as a global specification test for indirect estimators. In these cases, the dimension of \( \pi \) is larger than that of \( \theta \). The asymptotic distribution of these tests statistics is a chi-square with \((\text{dim } \pi - \text{dim } \theta)\) degrees of freedom. In our situation \( \pi \) and \( \theta \) have the same dimension so the asymptotic distribution would be a chi-square with 0 degree of freedom. For this reason, we cannot used it to compute the p-value of the GOF test in our framework. A standardization...
constant would be needed to avoid that the distribution of the test tends to a Dirac as
$S \frac{D}{\rightarrow} 0$. It can be conjectured that $n^\alpha S \frac{D}{\rightarrow} F \text{ as } n \rightarrow \infty$, where $F$ is a continuous
distribution, but this is left for further research together with the asymptotic properties
of this GOF test. Therefore, we compute the p-value of this test by parametric bootstrap
according to Algorithm 2.

**Algorithm 2 P-value for GOF $S$.**

Require:
- Data set $\mathbf{x} = (x_1, \ldots, x_n)$

1: **procedure** Computation of $S$
2: Compute $\hat{\theta}_B$ using a model $M_1$ that supposedly generate $\mathbf{x}$
3: Compute $\pi(\hat{\theta}_B, n) = \frac{1}{H} \sum_{h=1}^{H} \hat{\pi}_h(\hat{\theta}_B, n)$
4: Compute $S = \|\hat{\pi} - \pi(\hat{\theta}_B, n)\|_2^2$
5: Store $S$
6: **procedure** Computation of the p-value
7: for $b : 1 \rightarrow B$ do
8: Generate one sample $\mathbf{x}^*$ under $M_1$ with $\hat{\theta}_B$
9: Compute $\hat{\theta}_B^b$
10: Compute $S^b$
11: p-value $= \frac{1}{B} \sum_{b=1}^{B} 1_{S > S^b}$

**4.5.2 Simulation Study**

In Simulation 4.2 we study the distribution of the p-value obtained by parametric boot-
strap for $S$ under the null hypothesis, whereas in Simulation 4.3 we study its distribution
under one alternative.

**Simulation 4.2.** We generate 1000 samples of size $n = 300$ from a model with 2 latent
variables, 10 ordinal manifest variables, each having 5 levels. The true values for the
thresholds and the loadings of Model $M_0$ are given in Table 4.2. For each sample we
compute $\hat{\theta}_B$ by iterative bootstrap, $S$ and its p-value according to the Algorithm 2. To
compute one p-value by the parametric bootstrap described in the Algorithm 2 we used
$B = 100$ bootstrap samples. Figure 4.3 shows the resulting empirical distribution of the
p-value, which looks very close to a uniform distribution. The observed empirical level
for an $\alpha$-level of 5% is equal to 0.052.

**Simulation 4.3.** We generate 1000 samples of size $n = 300$ from a model with 3 latent
variables, 10 ordinal manifest variables, each having 5 levels. The true values for the
thresholds and the loadings of Model $M_1$ are given in Table 4.3. For each sample we
compute $\hat{\theta}_B$ by iterative bootstrap, $S$ and its p-value according to the Algorithm 2. The
entire procedure is identical to the one used in Simulation 4.2 except that the starting
1000 samples are generate from a different model than $M_0$, i.e. $M_1$, whereas $M_0$ is
the model used in the estimation of $\hat{\theta}_B$ and the simulation of the samples during the
parametric bootstrap for the computation of the p-value. Figure 4.4 shows the resulting
empirical distribution of the p-value, which clearly deviates from a uniform distribution
with a shift toward 0.

Figure 4.3 shows that the p-value obtained by parametric bootstrap for the GOF
Figure 4.3: Empirical distribution of the p-value for the GOF test $S$ under $H_0$.

Figure 4.4: Empirical distribution of the p-value for the GOF test $S$ under $H_1$. 
4.5. Goodness-of-fit test

<table>
<thead>
<tr>
<th>$\lambda_{01}$</th>
<th>$\lambda_{02}$</th>
<th>$\lambda_{03}$</th>
<th>$\lambda_{04}$</th>
<th>$\lambda_{1}$</th>
<th>$\lambda_{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.8</td>
<td>1.3</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>0</td>
<td>-0.8</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.2</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>-1.6</td>
<td>2.0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>-1.2</td>
<td>0</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>0</td>
<td>2.3</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.4</td>
<td>1.4</td>
<td>2.5</td>
<td>1.2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.2: True parameter values used in Simulation 4.2 (Model $\mathcal{M}_0$).

<table>
<thead>
<tr>
<th>$\lambda_{01}$</th>
<th>$\lambda_{02}$</th>
<th>$\lambda_{03}$</th>
<th>$\lambda_{04}$</th>
<th>$\lambda_{1}$</th>
<th>$\lambda_{2}$</th>
<th>$\lambda_{3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.8</td>
<td>1.3</td>
<td>0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>-2.0</td>
<td>-0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
<td>0</td>
<td>1.5</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.2</td>
<td>-1.7</td>
<td>1.0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>-1.0</td>
<td>2.0</td>
<td>0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>-2.0</td>
<td>0</td>
<td>-1.2</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>-1.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>0</td>
<td>0.5</td>
<td>-2.0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>0</td>
<td>1.4</td>
<td>0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.8</td>
<td>-1.9</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 4.3: True parameter values used in Simulation 4.3 (Model $\mathcal{M}_1$)

Statistic $S$ under $H_0$ follows a uniform distribution whereas under $H_1$ its distribution is clearly shifted toward 0 as shown by Figure 4.4. Therefore the parametric bootstrap proposed in Algorithm 2 seems to be a reliable way to obtain a p-value for $S$, as we do not know its asymptotic distribution. Further research is needed to study the empirical behaviour of this test statistic in terms of power for instance, but this seems to be a promising way to derive GOF test for GLLVM.
In chapter 2 we formally define the four simulation-based bias correction methods. Strategy 2.1 and Strategy 2.2 are based on indirect inference, the first one using $H$ samples of size $n$ in the estimation of the binding function $\pi(\theta)$ and the second one using only one sample of size $nH$. Strategy 2.3 and Strategy 2.4 are based on bootstrap, the first one being the conventional bootstrap bias correction and the second one being the iterated version of the first one. We show that the estimators obtained by Strategy 2.1, Strategy 2.2 and Strategy 2.4 are consistent for a finite $H$. The estimator obtained by Strategy 2.3 is consistent if we assume that the polynomial form of the bias has no constant term. We then study the finite sample properties of the different bias corrected estimators. We show that Strategy 2.1 leads to an estimator which has a bias of order $O(n^{-3})$, whereas the bias of the one obtained by Strategy 2.2 is of order $O(n^{-1})$. This higher order bias correction is achieved with a relatively small increase of variance. We show that in most practical situations, the one step bootstrap estimator is equivalent to the indirect estimator using one sample of size $nH$ in terms of bias and variance. Finally, we prove that the iterative bootstrap estimator is equivalent to the indirect estimator using $H$ samples of size $n$.

In chapter 3 we consider robust estimation of parameters of income distribution models. We propose to use the weighted MLE as a robust starting estimator, which is a very simple estimator which is known to be inconsistent, and then correct for its bias by indirect inference. In different simulation studies we show that both indirect estimators allow to correct for asymptotic bias thus leading to a consistent and robust estimator. Simulation study 3.3 confirms that only the indirect estimator using $H$ samples of size $n$ corrects for small sample bias. To correct for asymptotic bias, the indirect estimator using one sample of size $nH$ can be preferred as it is generally faster from a computational point of view. It is worth noting that the robust approach used here is not specific to income distributions. It can be used for any model for which a score function can be specified. This opens the door to simple computation of robust estimators especially for complex models.

In chapter 4 we propose an estimation method for GLLVM that avoids numerical integration and complex optimization of the approximated likelihood. Instead, we use a simple and easy-to-compute starting estimator that is not consistent and, in a second step, we correct for its bias by iterative bootstrap. The large number of parameters of these models precludes the use of indirect inference. A simulation study shows that our estimator is nearly unbiased and is comparable, with a slight loss of efficiency, to the one proposed by Mplus based on adaptive quadrature. Moreover, we propose a GOF test statistic using the idea that the iterative bootstrap is asymptotically equivalent to
the indirect estimator with $H$ samples of size $n$. Therefore, if we plugged the estimates obtained by iterative bootstrap in the objective function of the indirect estimator, the value of this function should be very close to 0 if the model is correct. The p-value is obtained by parametric bootstrap. Simulations show that it has a uniform distribution under $H_0$ and a distribution shifted toward 0 under $H_1$. 
## List of Tables

2.1 Empirical MSE and bias of the different estimators defined in Section 2.1 obtained with 1000 Monte Carlo replications with $\sigma^2 = 10$, $\mu = 0$, $n = 10$ and $H = 1000$. ................................................. 21

3.1 Empirical MSE and bias of the different estimators (defined in Simulation 3.3) obtained with 100 samples of size $n = 1000$ from a Lomax distribution without contamination. ................................................. 29

3.2 Empirical MSE and bias of the different estimators (defined in Simulation 3.3) obtained with 100 samples of size $n = 50$ from a Lomax distribution. . 32

3.3 MLE and Robust estimates, with 95%CI, for the parameters of the Dagum distribution and associated Gini indices for the Swiss Household data. . . 33

4.1 True parameter values used in Simulation 4.4. ............................... 43

4.2 True parameter values used in Simulation 4.2 (Model $M_0$). ............................... 47

4.3 True parameter values used in Simulation 4.3.(Model $M_1$) ............................... 47
List of Figures

2.1 Empirical distribution of the different estimators defined in Section 2.1 obtained with 1000 Monte Carlo replications with $\sigma^2 = 10$, $\mu = 0$, $n = 10$ and $H = 1000$. $\hat{\sigma}^2$, $\tilde{\sigma}^2$, $\hat{\sigma}_B^2$ and $\tilde{\sigma}_B^2$ are the estimators based on Strategies 2.1, 2.2, 2.3 and 2.4, respectively. The last boxplot corresponds to the unbiased estimator $\sum_{i=1}^{n}(x_i - \bar{x})^2/(n - 1)$. .................. 20

3.1 Empirical distribution of the different estimators (defined in Simulation 3.1) obtained with 100 samples of size $n = 1000$ from a Lomax distribution with parameters $b = 5$ (left panel) and $q = 2$ (right panel). Graphs (a) and (b) correspond to the setting without contamination whereas graphs (c) and (d) correspond to the setting where a randomly chosen 1% of the data are multiplied by 1000. .......................... 28

3.2 Empirical distribution of the different estimators (defined in Simulation 3.2) obtained with 100 samples of size $n = 1000$ from a Dagum distribution with parameters $a = 2$ (graphs (a) and (d)), $b = 1$ (graphs (b) and (e)) and $p = 0.5$ (graphs (c) and (f)). Graphs on the top row correspond to the setting without contamination whereas graphs on the bottom row correspond to the setting where a randomly chosen 1% of the data are multiplied by 3000. .......................... 30

3.3 Empirical distribution of the different estimators (defined in Simulation 3.3) obtained with 100 samples of size $n = 50$ from a Lomax distribution with parameters $b = 5$ (graph (a)) and $q = 2$ (graph (b)). .......................... 31

3.4 Robust (blue dashed line) and MLE (red solid line) density estimation of the Dagum model on the Swiss Household income data for 1999 and 2013. 34

4.1 Empirical distribution of the estimates, centered at the true value, of four loadings obtained by the starting inconsistent estimator $(\hat{\pi})$, the estimator based on iterative bootstrap $(\tilde{\theta}_B)$ and the approximated MLE obtained using Mplus $(\hat{\theta}_{ML})$. .......................... 41

4.2 Empirical distribution of the bias of the estimates, centered at the true value, of four thresholds obtained by the starting inconsistent estimator $(\hat{\pi})$, the estimator based on iterative bootstrap $(\tilde{\theta}_B)$ and the MLE obtained using Mplus $(\hat{\theta}_{ML})$. .......................... 42

4.3 Empirical distribution of the p-value for the GOF test $S$ under $H_0$. 46

4.4 Empirical distribution of the p-value for the GOF test $S$ under $H_1$. 46


