Contributions to simulation-based estimation methods

Orso, Samuel

This publication URL: https://archive-ouverte.unige.ch/unige:121536

Publication DOI: 10.13097/archive-ouverte/unige:121536

Contributions to simulation-based estimation methods

by

Samuel Orso

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-advisor, University of Geneva
Prof. Stéphane Guerrier, Co-advisor, Pennsylvania State University
Prof. Stefan Sperlich, Chair, University of Geneva
Prof. Yanyuan Ma, Pennsylvania State University

Thesis No. 66
January 2019

La Faculté d'économie et de management, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par-là, émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 29 janvier 2019

Le doyen,
Marcelo OLARREAGA

Impression d'après le manuscrit de l'auteur.

Acknowledgements

This thesis has greatly benefitted from the academic and moral support of many, and as hard as it seems, I would like to express my gratitude to all of them.

Foremost, I would like to thank my two supervisors Prof. Maria-Pia Victoria-Feser and Prof. Stéphane Guerrier. I would not be writting these lines if it was not for them. They made me discover statistics as an under-graduate student in business administration and have never let me down ever since. I am not sure I would have had their patience with someone who had such a poor background. Stéphane has been very influencial in my decisions to undertake studies in statistics and to pursue a PhD. He has always supported me and is very inspirational. Maria-Pia has never stopped me for trying my own way for research, better, she made it possible regardless of the circumstances. I admire her commitment and all the positive energies she puts into work.

I would like to thank Prof. Stefan Sperlich, the chair of my jury, for the interest he took in my work, for his constructive feedbacks and for his professionalism. Likewise, I would like to thank Prof. Yanyuan Ma who honoured me by accepting to be my external expert and whose instructive comments improved the content of this thesis.

I would also like to thank my incredible colleagues. I am forever indebted to Marco, *a.k.a.* Marc-small-oh, and Marc-Olivier, *a.k.a.* Marc-Big-Oh, for their valuable discussions from which this thesis has benefitted. I shared unforgettable moments with the "*dîner de cons*" team: Kustrim, Elise, Rose, Marc-Olivier, Mattia and Dany. I enjoyed sportive time with the "*badminton*" team: Rami, Walid, Mark, Haotian, Mattia and Dany. Throughout these years, I have always loved sharing ideas with my colleagues at the GSEM. Among them, I would like to acknowledge: Jean-Christophe, Jonathan, Ozgu, Roberto, Sebastian, Justine, Ingrid, Steffen, Mark, Pierre-Yves, Linda, Setareh, Laura, Guillaume, Cesare, Julien, Alban, Jiajun, Benjamin, Kasia, Mucyo, Yuming *et al.* My office mates Haotian and Gaëtan were very supportive all along my dissertation and I specially want to thank them.

I would like to thank my parents, Charly and Corine, my three sisters and their partners, Joëlle, Lucien, Myriam, Estelle and Sergio, for their love and for always being supportive. A special though goes to my two lovely little nieces Anouk and Justine. This thesis would have been impossible without the support of my in-laws Jordan, Margarita and Aleks. On a great deal of occasions, Margarita has volunteered to compensate my lack of availability to my family. This thesis was one of the last thing, if not the last, that I started before my son Isaac was born. He encouraged me a lot for example by saying: "*Mon papa fait des statistiques*" [1], or with the many drawings that covers half of my wall at the office. My daughter Tania was born later. Her arrival gave me the courage to attack the last marathon that represents the writing of a thesis. Her good sleeps were the encouragement I needed. Many of the ideas that came up to me during the thesis

[1] French for: "*My dad does statistics*"

happened during the bedtime. These little reflections of myself have greatly impacted this thesis. Finally, I would like to thank my soulmate and wife Linda for all these years of sharing and giving. She shared my ups and down. She never stopped believing in me. She gave me all the forces that leads me to say today: “I have *Ph.*inishe*D*”.

Abstract

The focus of this thesis is twofold. First, it delivers a new look at existing simulation-based methods for statistical inference in parametric problems. Emphasis is placed on finite sample theoretical properties and computational efficiency. In particular, a simple and computationally efficient method for inference is proposed. It is shown that exact inference may be claimed in theory in some situations even though sample size is kept fixed. Numerical examples demonstrate the wide applicability of this method. Second, a general class of flexible models for dependent random phenomena is studied. Emphasis is placed on problems of point estimations due to the presence of outliers or because of the underlying computational burden. To tackle these issues, a new multi-step robust and computationally efficient estimator is proposed. Asymptotic properties are studied along with illustrative examples.

Résumé

Cette thèse comporte deux parties. Premièrement, elle délivre un nouveau regard sur différentes méthodes par simulations développées pour faire de l'inférence statistique sur des problèmes paramétriques. Le focus est porté sur les propriétés théoriques en échantillon fini et les problèmes computationnelles. En particulier, une méthode simple et computationnellement éfficiente est proposée. Il est démontré qu'il est possible d'obtenir une inférence exacte dans certaines situations tout en gardant la taille d'échantillon fixe. Des examples numériques illustrent le vaste champ d'application de cette méthode. Deuxièmement, une classe générale de modèles pour des variables aléatoires dépendantes est étudié. Les sujets principaux sont les problèmes d'estimation ponctuelles dûs à la présence de données aberrantes ou à cause de problèmes computationnelles. Afin d'adresser ces questions, un nouvel estimateur robuste et computationnellement efficient est proposé. Ses propriétés asymptotiques sont étudiées. Des examples viennent illustrer ces résultats.

Contents

To Linda, Isaac and Tania

Introduction

The initial goal of this thesis was to develop an estimator both robust to outliers and computationally feasible when modeling multivariate data with copula models. While pursuing this objective, many discoveries have been made, enlarging thereby significantly the scope of study. This manuscript presents a substantial part of these findings in two separate chapters:

1. The proposal of a multi-purpose finite-sample inferential framework.

2. A framework for multi-step robust estimators tailored to copula models.

These two chapters may be read as stand-alone and separate research papers. However, they have many common grounds, which are duly noticed along the chapters. A common philosophy that underlies both chapters is the *computational feasibility* of the methods. The numerical aspect of the proposed methodologies is indeed a permanent motivation for both chapters. In a broad sense, what is understood by feasible is that, for a given method requesting numerical resources, a user should be able to make inference about unknown quantities of a parametric probabilist model within a reasonable time. Of course, this reasonability depends upon the computational power at the user's disposal, the software implementation and the limit in time fixed by, or imposed to, the user and is therefore intrisincally subjective. Nonetheless, computational feasibility is highlighted on many occasions, it is especially appreciated under the lights brought by alternative methods. A third and final chapter combining useful theoretical results for the two other chapters ends the thesis.

Chapter 1 has the widest scope of application. It proposes a new look at existing simulation-based methods by demonstrating in the most general setting where apparently separated methods are in fact equivalent, and where they are not. Capitalizing on this comparison, a computationally efficient method is proposed for which the finite sample and asymptotic properties are studied. In particular, it is shown that exact inference may be claimed when the sample size is fixed under strong but commonly encountered situations. Walkthrough and numerical examples demonstrate the applicability of the method.

Chapter 2 is a research paper specific to parametric multivariate dependence models. A general class of data generating process is studied leading to the proposition of an estimating procedure. The interest is guided by situations where estimating the dependence is a complex problem requiring multi-steps procedures. A special emphasis is on estimators that are robust to outliers. The asymptotic properties of the proposed method are examined as well as recommendation of practical implementation. The concept of robustness is theoretically formalized with the influence function. It permits to appreciate not only the proposed procedure, but also the robustness of other estimators. This chapter ends with simulation studies and an application to the mobility of income in time.

This thesis has also been frequently ponctuated by opportunities to work on a variety of research projects. These opportunities materialize in the following list of publications and advanced manuscripts:

- 📕 Guerrier, S., Orso, S., Victoria-Feser, M.-P. "*Inference for Index Functionals*". In: *Econometrics*, 6(2), 22, April 2018. https://doi.org/10.3390/econometrics6020022

- 📕 Branca, M., Orso, S., Molinari, R., Xu, H., Guerrier, S., Zhang, Y., Mili, N. "*Is Non-Metastatic Cutaneous Melanoma Predictable through Genomic Biomakers?*" In: *Melanoma Research*, 28(1):21–29, February 2018. https://doi.org/10.1097/CMR.0000000000000412

- 📕 Guerrier, S., Mili, N., Molinari, R., Orso, S., Avella-Medina, M. and Ma, Y. "*A Predictive Based Regression Algorithm for Gene Network Selection*". In: *Frontiers in Genetics*, 7:97, 2016. https://doi.org/10.3389/fgene.2016.00097

- 👥 Montet, X., Hofmeister, J., Burgmeister, S., Orso, S., Mili, N., Guerrier, S., Victoria-Feser, M.-P., Muller., H. "*Lung Nodule Classification by Data Mining and Artificial Intelligence*". Submitted to *Nature Medecine.*

- 📝 Guerrier, S., Karemera, M., Orso, S., Victoria-Feser, M.-P. "*On the Properties of Simulation-based Estimators in High Dimensions*". Working paper. https://arxiv.org/abs/1810.04443.

- 📝 Orso, S., Clausen, P., Guerrier, S., Skaloud, J. "*Estimation of Inertial Sensor Stochastic Characteristics under Varying Environmental Conditions*". Working paper.

The first paper deals with inference problem concerning inequality indices in welfare economics. In particular, it is demonstrated that in the situation where the data generating mechanism is a parametric model, only a subset of the parameters needs to be estimated in order to conduct inference on a targeted index. It relates to the present manuscript by the methodology employed. The three following papers are unrelated to this thesis. In the oldest, an heuristic algorithm is developed for selecting input variables based on their predictive capacity. The two other published or submitted papers highlight the particular interest that such methodology found in gene selection and radiomic problems. The two last research manuscripts are advanced working papers. On a theoretical ground, the document entitled "*On the Properties of Simulation-based Estimators in High Dimensions*" exposes the capicity of a simulation-based methods to correct for bias in a general paradigm of estimation. This study is closely related to the present thesis. The other working paper is unrelated to the thesis, it develops a method to include external variables in the modelisation of the stochastic signals issued from inertial sensors. Nowadays, this is considered as an important problem for navigation system.

List of notation

$\overset{d}{=}$:	equality in distribution
$\xrightarrow{p}$:	convergence in probability
$\rightsquigarrow$:	convergence in distribution
$o, \mathcal{O}$:	order symbols
$o_p, \mathcal{O}_p$:	stochastic order symbols
$\mathcal{B}(c, r), \mathcal{B}^c(c, r)$:	open, closed ball of center c and radius r
$\mathbb{E}$:	expectation
Cov:	covariance
argmin, argzero:	arguments of minima, roots
diag:	diagonal of a matrix
rank:	rank of a matrix
trace:	trace of a matrix

1

SwiZs: Switched Z-estimators

On peut tromper mille fois une personne, on peut tromper une fois mille personnes, mais on ne peut pas tromper mille fois mille personnes.

– Emile, *La Cité de la Peur*

1.1 Introduction

The algorithmic principle of the bootstrap method is quite simple: reiterate the mechanism that produces an estimator on pseudo-samples. But when it comes to estimators that are numerically complicated to obtain, the bootstrap is less attractive to use due to the numerical burden. If one estimator is hard to find, reiterating compounds this issue. Paraphrasing Emile in the French comedy *La Cité de la Peur*: we can implement the bootstrap when the estimator is simple to obtain or we can compute a numerically complex point estimator, but it is too computationally cumbersome to do both.

Although this limitation is purely practical and tends to be reduced by the ever increasing computational power at our disposal, everyone would agree that it is nonetheless attractive to have a method that frees the user from the computational burden, or at least provides an answer within a reasonable time. In this chapter, we explore a special case of the efficient method of moments ([GT96]) that encompasses both the computation of numerically complex estimators and of a "bootstrap distribution" at a reduced cost. The idea deviates from the algorithmic principle of the bootstrap: the proposed method no longer attempts at reproducing the sample mechanism that lead to an estimator, but instead, tries to find every estimators that may have produced the observed sample, or more often, some statistics on the sample.

The idea is not new though, several methods follow this pattern. The indirect inference method ([GMR93; Smi93]) similarly attempts at finding the point estimate that lead to statistics obtained from simulated samples as close as possible to the same statistics on the observed sample. Mostly used in econometric and financial contexts, indirect inference has been successfully applied to the estimation of stable distribution ([GRV11]), stochastic volatility models ([Mon98; LC09]), financial contingent claims ([PY09]), dynamic panel models ([GPY10]), dynamic stochastic equilibrium models ([DGR07]), continuous time models ([GT10]), diffusion processes ([BSZ98]); but it has also been used in queueing theory ([HF04]), robust estimation of generalized linear latent variable models ([MVF06]), robust income distribution ([Gue+18b]), high dimensional generalized linear model and penalized regression ([Gue+18a]). Often presented as the Bayesian counterpart of the indirect inference, the approximate Bayesian computation ([Tav+97; Pri+99]) aims at finding the values that match the statistics computed on simulated samples and the statistics on the observed sample, with a certain degree approximation. The method has however grown in a different context of applications. For example, it has been successfully employed in population genetics ([BZB02]), in ecology ([Bea10]), in evolutionary biology ([Cor+08; Wil+10]). Less popular, R.A. Fisher's fiducial inference (see for instance [Fis22; Fis30; Fis33; Fis35; Fis56]) and related methods such as the generalized fiducial inference ([Han09; Han13; Han+16]), D.A.S. Fraser's structural inference ([Fra68], see also [DSZ73]), Dempster-Shafer theory ([Sha76; Dem08]) and inferential models ([ML13; ML15; Mar15]) follow a similar pattern, the main idea being to find all possible values that permit to generate simulated sample as close as possible to the observed sample, but without specifying any prior distribution.

Regardless of the difference in philosophy of the aforementioned methods, they have in common that they are usually very demanding in computational resources when implemented for non-trivial applications. This is a major difference with the approach we endorse in this chapter. By letting the statistics be the solution of an estimating function of the same dimension as the quantity of interest, we demonstrate that it is possible to bypass the computation of the same statistics on simulated sample by directly estimating the quantity of interest within the estimating function, resulting thereby in a potential significant gain in computational time. In Section 1.3, we demonstrate in finite sample that under some weak conditions the estimators resulting from our approach is equivalent to the estimators one would have obtained using certain forms of indirect inference, approximate Bayesian computation or fiducial inference approaches, whereas it is different than parametric bootstrap estimators, except in the case of a location parameter. This section innovates on two aspects. First, it implicates that our approach can be employed in practice to solve problems that relate to indirect inference, approximated Bayesian compuation and fiducial inference in a computationally efficient manner. Second, it proves or disproves formally the link between the aforementioned methods, and this in the most general situation as the results remain true for any sample size.

Constructing tests or confidence regions that controls over the error rates in the long-run is probably one of the most important problem in statistics ever since at least Neyman-Pearson famous article [NP33]. Yet, the theoretical justification for most methods in statistics is asymptotic. The bootstrap for example, despite its simplicity and its widespread usage is an asymptotic method ([Hal92]); for the other methods, see for example [Fra+18] for approximate Bayesian computation, [GMR93] for indirect inference and [Han+16] for generalized fiducial inference. There are in general no claim about the exactness of the inferential procedures in finite sample (see [Mar15] for one of the exceptions). In Section 1.4, we study theoretically the frequentist error rates of confidence regions constructed on the distribution issued from our proposed approach. In particular, we demonstrate under some strong, but frequently encountered, conditions that the confidence regions have exact coverage probabilities in finite sample. Asymptotic justification is nonetheless provided in Section 1.5. In addition, we bear the comparison with the asymptotic properties of indirect inference method to conclude that, surprisingly, both approaches reach the same conclusion but under distinct conditions. Some leads are evoked, but we lack to elucidate the fundamental reason behind such discrepancy.

Although the proposed method is first and foremost computational, surprisingly in some situations explicit closed-form solutions may be found. We gather a non-exhaustive number of such examples, some important, in Section 1.6. The numerical study in Section 1.7 ends this chapter. We study via Monte Carlo simulations the coverage probabilities obtained from our approach and compare with others on a variety of problems. We conclude that in most situations, exact coverage probability computed within a reasonable computational time can be claimed with our method.

1.2 Setup

Let $\mathbb{N}$ ($\mathbb{N}^+$) be the sets of all positive integers including (excluding) 0. For any positive integer n, let $\mathbb{N}_n$ be the set whose elements are the integers $0, 1, 2, \dots, n$; similarly $\mathbb{N}_n^+ = \{1, 2, \dots, n\}$.

We consider a sequence of random variables $\{\mathbf{x}_i : i \in \mathbb{N}_n^+\}$, possibly multivariate, to follow an assumely known distribution $F_{\boldsymbol{\theta}}$, indexed by a vector of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathrm{I\!R}^p$. We suppose that it is easy to generate artificial samples $\mathbf{x}^*$ from $F_{\boldsymbol{\theta}}$. Specifically, we generate the random variable $\mathbf{x}$ with a known algorithm that associates $\boldsymbol{\theta}$ and a random variable $\mathbf{u}$. We denote the generating mechanism as follows:

$$\mathbf{x} = \mathbf{g}(\boldsymbol{\theta}, \mathbf{u}).$$

The random variable $\mathbf{u}$ follows a known model $F_{\mathbf{u}}$ that does not depend on $\boldsymbol{\theta}$. Using this notation, the observed sample is $\mathbf{x}_0 = \mathbf{g}(\boldsymbol{\theta}_0, \mathbf{u}_0)$ and the artificial sample is $\mathbf{x}^* = \mathbf{g}(\boldsymbol{\theta}, \mathbf{u}^*)$, where $\mathbf{u}_0$ and $\mathbf{u}^*$ are realizations of $\mathbf{u}$.

Example 1.1 (Normal). *Suppose $\mathbf{x} \sim \mathcal{N}(\theta, 1)$, then four examples of possible generating mechanism are:*

1. $\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) = \boldsymbol{\theta} + \mathbf{u}$ *where* $\mathbf{u} \sim \mathcal{N}(0, 1)$,
2. $\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) = \boldsymbol{\theta} + \sqrt{2}\,\mathrm{erf}^{-1}(2\mathbf{u} - 1)$ *where* $\mathbf{u} \sim \mathcal{U}(0, 1)$ *and* $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\,\mathrm{d}t$ *is the error function,*
3. $\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) = \boldsymbol{\theta} + \sqrt{-2\ln(\mathbf{u}_1)}\cos(2\pi\mathbf{u}_2)$ *where* $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)^T$, $\mathbf{u}_1 \sim \mathcal{U}(0, 1)$ *and* $\mathbf{u}_2 \sim \mathcal{U}(0, 1)$,
4. $\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) = \boldsymbol{\theta} + \mathbf{u}_2\sqrt{\frac{-2\ln(\mathbf{u}_3)}{\mathbf{u}_3}}$ *where* $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$, $\mathbf{u}_3 = \mathbf{u}_1 + \mathbf{u}_2$, $\mathbf{u}_1 \sim \mathcal{U}(0, 1)$, $\mathbf{u}_2 \sim \mathcal{U}(0, 1)$.

A possible counter-example is the following: $\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) = \mathbf{u} - \boldsymbol{\theta}$ *where* $\mathbf{u} \sim \mathcal{N}(2\boldsymbol{\theta}, 1)$*. Clearly* $\mathbf{x} = \mathbf{g}(\boldsymbol{\theta}, \mathbf{u})$*, but this* $\mathbf{g}$ *is not adequate because the distribution of* $\mathbf{u}$ *depends on* $\boldsymbol{\theta}$*.*

We now define the estimators we wish to study.

Definition 1.2 (SwiZs). *We consider the following sequence of estimators:*

$$\hat{\boldsymbol{\pi}}_n \in \boldsymbol{\Pi}_n = \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}}{\mathrm{argzero}}\, \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}\left(\mathbf{g}\left(\boldsymbol{\theta}_0, \mathbf{u}_{0i}\right), \boldsymbol{\pi}\right) = \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}}{\mathrm{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0, \mathbf{u}_0, \boldsymbol{\pi}\right),$$

$$\hat{\boldsymbol{\theta}}_n^{(s)} \in \boldsymbol{\Theta}_n^{(s)} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\mathrm{argzero}}\, \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}\left(\mathbf{g}\left(\boldsymbol{\theta}, \mathbf{u}_{si}^*\right), \hat{\boldsymbol{\pi}}_n\right) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\mathrm{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}_s^*, \hat{\boldsymbol{\pi}}_n\right),$$

where $\boldsymbol{\phi}$ is an estimating function and $s \in \mathbb{N}_S^+$. The estimators $\hat{\boldsymbol{\pi}}_n$ are referred as the auxiliary estimators. Any sequence of estimators $\{\hat{\boldsymbol{\theta}}_n^{(s)} : s \in \mathbb{N}_S^+\}$ is called Switched Z-estimators, or in short, SwiZs. The collection of the solutions is $\boldsymbol{\Theta}_n = \cup_{s \in \mathbb{N}_S^+} \boldsymbol{\Theta}_n^{(s)}$.

Remark 1.3. *The SwiZs in the Definition 1.2 may arguably be viewed as a special case of the Efficient Method of Moment (EMM) estimator proposed by [GT96]. Indeed, to have an EMM estimator the only modification to the Definition 1.2 is*

$$\hat{\boldsymbol{\theta}}^{(s)}_{EMM,n} \in \boldsymbol{\Theta}^{(s)}_{EMM,n} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argzero}}\, \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}^*_{sh}, \hat{\boldsymbol{\pi}}_n\right),$$

where $H \in \mathbb{N}^+$. Ergo, the SwiZs and EMM coincide whenever $H = 1$. Note that in general the EMM is defined with H large and $S = 1$.

1.3 Equivalent methods

As already remarked, the SwiZs does not appear to be a new estimator. The SwiZs in fact offers a new point of view to different existing methods as it federates several techniques under the same hat. In this Section, we show the equivalence or disequivalence of the SwiZs to other existing methods, for any sample size n, to conclude that the distribution obtained by the SwiZs is (approximatively) a Bayesian posterior, and thereby that it is valid for the purpose of inference.

The EMM and the indirect inference estimator of [Smi93; GMR93] are known to have the same asymptotic distribution when $\dim(\boldsymbol{\pi}) = \dim(\boldsymbol{\theta})$ (see Proposition 4.1 in [GM96]). In the next result, we demonstrate that the SwiZs and a certain form of indirect inference estimator are equivalent for any n.

Definition 1.4 (indirect inference estimators)**.** *Let $\hat{\boldsymbol{\pi}}_n$ and $\{\mathbf{u}_j : j \in \mathbb{N}\}$ be defined as in the Definition 1.2. We consider the following sequence of estimators, for $s \in \mathbb{N}^+_S$:*

$$\hat{\boldsymbol{\pi}}^{(s)}_{II,n}(\boldsymbol{\theta}) \in \boldsymbol{\Pi}^{(s)}_{II,n} = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\text{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}^*_s, \boldsymbol{\pi}\right), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

$$\hat{\boldsymbol{\theta}}^{(s)}_{II,n} \in \boldsymbol{\Theta}^{(s)}_{II,n} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argzero}}\, d\left(\hat{\boldsymbol{\pi}}_n, \hat{\boldsymbol{\pi}}^{(s)}_{II,n}(\boldsymbol{\theta})\right), \quad \hat{\boldsymbol{\pi}}_n \in \boldsymbol{\Pi}_n, \quad \hat{\boldsymbol{\pi}}^{(s)}_{II,n} \in \boldsymbol{\Pi}^{(s)}_n,$$

where d is a metric. We call $\{\hat{\boldsymbol{\theta}}^{(s)}_{II,n} : s \in \mathbb{N}^+_S\}$ the indirect inference estimators. The collections of solutions are denoted $\boldsymbol{\Pi}_{II,n} = \cup_{s\in\mathbb{N}^+_S}\boldsymbol{\Pi}^{(s)}_{II,n}$ and $\boldsymbol{\Theta}_{II,n} = \cup_{s\in\mathbb{N}^+_S}\boldsymbol{\Theta}^{(s)}_{II,n}$.

Remark 1.5. *In Definition 1.4, we are implicitly assuming that $\boldsymbol{\Theta}$ contains at least one of, possibly many zeros, of the distance between the auxiliary estimators on the sample and the pseudo-sample. Therefore, the theory is the same for any measure of distance that we denote generically by d.*

Remark 1.6. *The indirect inference estimators in Definition 1.4 is a special case of the more general form*

$$\hat{\boldsymbol{\theta}}^{(s)}_{II,B,m} \in \boldsymbol{\Theta}^{(s)}_{II,B,m} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argzero}}\, d\left(\hat{\boldsymbol{\pi}}_n, \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\pi}}^{(s)}_{II,b,m}(\boldsymbol{\theta})\right),$$

$B \in \mathbb{N}^+$, $m \geq n$. In Definition 1.4 we fixed $B = 1$ and $m = n$. [GMR93] considered two cases: first, B large, $m = n$ and $S = 1$, second, $B = 1$, m large and $S = 1$. For both cases, the ℓ_2-norm was used as the measure of distance (see the preceding remark).

Assumption 1.7 (uniqueness)**.** *For all $(\boldsymbol{\theta}, s) \in \boldsymbol{\Theta} \times \mathbb{N}_S$, $\text{argzero}_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\, \boldsymbol{\Phi}_n(\boldsymbol{\theta}, \mathbf{u}_s, \boldsymbol{\pi})$ has a unique solution*

Theorem 1.8 (Equivalence SwiZs/indirect inference)**.** *If Assumption 1.7 is satisfied, then the following holds for any $s \in \mathbb{N}_S^+$:*

$$\boldsymbol{\Theta}_n^{(s)} = \boldsymbol{\Theta}_{II,n}^{(s)}.$$

Theorem 1.8 is striking because it concludes that a certain form of EMM, the SwiZs, and indirect inference estimators (as in Definition 1.4) are actually the very same estimators, not only asymptotically, but for any sample size, and under a very mild condition. Indeed, Assumption 1.7 requires the roots of the estimating function to be well separated so there exists a unique solution. This requirement is unrestrictive and it is typically satisfied. One may even wonder what would be the purpose of an estimating function for which Assumption 1.7 would not hold. In this spirit, Assumption 1.7 may be qualified as the "minimum criterion" for choosing an estimating function.

Even if the optimizer is perfect, Theorem 1.8 does not imply that the exact same values are found using the SwiZs or the indirect inference estimators, but that they belong to the same set of solutions, and thereby that they share the same statistical properties. Hence, Theorem 1.8 offers us two different ways of computing the same estimators. Simple calculations however show that the SwiZs is computationally more attractive. Indeed, if we let k denotes the cost evaluation of $\boldsymbol{\Phi}_n$, l the numbers of evaluations of $\boldsymbol{\Phi}_n$ for obtaining an auxiliary estimator or the final estimator, then the SwiZs has a total cost of roughly $\mathcal{O}(2kl)$ whereas it is $\mathcal{O}(kl + kl^2)$ for the indirect inference estimator, so a reduction in order of $\mathcal{O}(kl^2)$. This computational efficiency of the SwiZs accounts for the fact that it is not necessary to compute $\hat{\boldsymbol{\pi}}_{\text{II},n}$, and thus avoids the numerical problem of the indirect inference estimator of having an optimization nested within an optimization. This discrepancy is also, quite surprisingly, reflected in the theory we develop in Section 1.4 for the finite sample properties and in Section 1.5 for the asymptotic properties.

At first glance, the SwiZs may appear similar to the parametric bootstrap (see the Definiton 1.9 below). If we strengthen our assumptions and think of the auxiliary estimator as an unbiased estimator of $\boldsymbol{\theta}$, it is natural to think of the SwiZs and the parametric bootstrap as being equivalent. In any cases, both methods use the exact same ingredients, so we may wonder whether actually they are the same. The next result demonstrates that in fact, they will be seldom equivalent.

Definition 1.9 (parametric bootstrap)**.** *Let $\hat{\boldsymbol{\pi}}_n$ and $\{\mathbf{u}_j : j \in \mathbb{N}\}$ be defined as in Definition 1.2. We consider the following sequence of estimators:*

$$\hat{\boldsymbol{\theta}}_{Boot,n}^{(s)} \in \boldsymbol{\Theta}_{Boot,n}^{(s)} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\pi}}_n, \mathbf{u}_s^*, \boldsymbol{\theta}\right), \quad s \in \mathbb{N}_S^+.$$

The collection of the solutions is $\boldsymbol{\Theta}_{Boot,n} = \cup_{s\in\mathbb{N}_S^+}\boldsymbol{\Theta}_{Boot,n}^{(s)}$.

Remark 1.10. *For the solutions $\boldsymbol{\Theta}_{Boot,n}^{(s)}$ in Definition 1.9 to be nonempty, the parametric bootstrap requires that $\boldsymbol{\Pi}_n \subset \boldsymbol{\Theta}$. The SwiZs has not such requirement.*

Assumption 1.11. *The zeros of the estimating functions are symmetric on $(\boldsymbol{\theta}, \boldsymbol{\pi})$, that is*

$$\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \mathbf{u}_s, \boldsymbol{\pi}) = \boldsymbol{\Phi}_n(\boldsymbol{\pi}, \mathbf{u}_s, \boldsymbol{\theta}) = \mathbf{0}.$$

Theorem 1.12 (equivalence SwiZs/parametric bootstrap)**.** *If and only if Assumption 1.11 is satisfied, then it holds that*

$$\boldsymbol{\Theta}_n^{(s)} = \boldsymbol{\Theta}_{Boot,n}^{(s)}.$$

Assumption 1.11 is very restrictive, so Theorem 1.12 suggests that in general the SwiZs and the parametric bootstrap are not equivalent. This may appear as a surprise as only the argument $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are interchanged in the estimating function. Then, if they are different, the question of which one should be preferred naturally arises. We do not attempt at answering this question, but we rather prefer to stimulate debates by giving motivations for using the SwiZs. Popularized by [Efr79], the bootstrap has been a long-standing technique for (frequentist) statistician, it is relatively straightforward to implement and has a well-established theory (see for instance [Hal92]). On the other hand, although the idea of the SwiZs has been arguably around for decades (see the comparison with the fiducial inference at the end of this section), we lack evidence of its widespread usage, at least not under the form presented here. When facing situations where $\hat{\boldsymbol{\pi}}_n$ is an unbiased estimator of $\boldsymbol{\theta}_0$, compared to the parametric bootstrap, the SwiZs is more demanding for the implementation and is generally less numerically efficient (see Section 1.7) suggesting that solving $\boldsymbol{\Phi}_n(\boldsymbol{\theta},\boldsymbol{\pi})$ in $\boldsymbol{\theta}$ is computationally more involved than in $\boldsymbol{\pi}$. However, in all the other situations where for example $\hat{\boldsymbol{\pi}}_n$ may be an (asymptotically) biased estimator of $\boldsymbol{\theta}_0$, a sample statistic or a consistent estimator of a different model, the parametric bootstrap cannot be invoked directly, at least not with the same form as in Definition 1.12. Indeed, the parametric bootstrap requires $\hat{\boldsymbol{\pi}}_n$ to be a consistent estimator of $\boldsymbol{\theta}_0$. Therefore, when considering complex model for which a consistent estimator is not readily available at a reasonable cost, the SwiZs may be computationally more attractive. The rest of this section aims at demonstrating that the distribution of the SwiZs is valid for the purpose of inference, whereas the following section theorizes the inferential properties of the SwiZs in finite sample for which Sections 1.6 and 1.7 gather evidences. But before, having emphasized their differences, we would like to share a rather common problem on which the parametric bootstrap and the SwiZs are equivalent.

The condition under which the SwiZs and the parametric bootstrap are equivalent (Assumption 1.11) is very strong and generally not met. There is one situation however where this condition holds, if the inferential problem is on the parameter of a location family as formalized in the next Proposition 1.13.

Proposition 1.13 (equivalence SwiZs/parametric bootstrap in location family problems)**.** *Suppose that x is a univariate random variable identically and independently distributed according to a location family, that is $x \overset{d}{=} \theta + y$, where $\theta \in \mathbb{R}$ is the location parameter. If the auxiliary parameter is estimated by the sample average and x is symmetric around 0, that is $x \overset{d}{=} -x$, then*

$$\boldsymbol{\Theta}_n^{(s)} = \boldsymbol{\Theta}_{Boot,n}^{(s)}.$$

The conditions which satisfies Proposition 1.13 are restrictive. Indeed, they are satisfied for location families for which the centered random variable is symmetric. Proposition 1.13 holds for example with a Gaussian, a Student, a Cauchy and a Laplace random variables (variance and degrees of freedom known), but not, for example, for a generalized extreme value, a skewed Laplace and a skewed t random variables (even with non-location parameters being fixed). The proof uses an average as the auxiliary estimator, but it should be easily extended to other estimator of location such as the trimmed mean. Proposition 1.13 is illustrated with a Cauchy random variable in Example 1.51 of Section 1.6.

Although the parametric bootstrap and the SwiZs will lead rarely to the same estimators, in spite of the similitude of their forms, the next result demonstrates that the distribution of the SwiZs corresponds in fact to (some sort of) a Bayesian posterior. Likewise the indirect inference, the approximate Bayesian computation (ABC) techniques were proposed to respond to complex problems. The two techniques are often presented to be respectively the frequentist and the Bayesian approaches to a same problem and have even been mixed sometimes (see [DPL15]). We now show under what conditions the SwiZs and the ABC are equivalent, but before, we need to give more precision on what type of ABC. Often dated back to [DG84], the ABC has evolved and covers now a broad-spectrum of techniques such as rejection sampling (see e.g. [Tav+97; Pri+99]), the Markov chain Monte Carlo (see e.g. [Mar+03; BCS07]), the sequential Monte Carlo sampling (see e.g. [SFT07; Bea+09; Ton+09]) among others (see [Mar+12] for a review). The equivalence between the SwiZs and the ABC is demonstrated with a rejection sampling presented in the next definition. However, the note of [Sis+10] suggests that this result may be extended to Markov chain Monte Carlo and sequential Monte Carlo sampling algorithms. We leave such rigorous demonstration for further research.

Definition 1.14 (Approximate Bayesian Computation (ABC) estimators)**.** *Let* $\hat{\boldsymbol{\pi}}_n$ *and* $\{\mathbf{u}_j : j \in \mathbb{N}\}$ *be defined as in Definition 1.2. Let* $\hat{\boldsymbol{\pi}}^{(s)}_{II,n}(\boldsymbol{\theta})$ *be defined as in Definition 1.4. We consider the following algorithm. For a given* $\varepsilon \geq 0$*, for a given infinite sequence* $\{\mathbf{u}_s : s \in \mathbb{N}^+_S\}$*, for a given infinite sequence of empty sets* $\{\boldsymbol{\Theta}^{(s)}_{ABC,n}(\varepsilon) : s \in \mathbb{N}^+_S\}$*, for a given prior distribution* $\mathcal{P}$ *of* $\boldsymbol{\theta}$*, repeat (indefinitely) the following steps:*

1. *Generate* $\boldsymbol{\theta}^\star \sim \mathcal{P}$.
2. *Compute* $\hat{\boldsymbol{\pi}}^{(s)}_{II,n}(\boldsymbol{\theta}^\star)$.
3. *If the following criterion is satisfied*

$$d\left(\hat{\boldsymbol{\pi}}_n, \hat{\boldsymbol{\pi}}^{(s)}_{II,n}(\boldsymbol{\theta}^\star)\right) \leq \varepsilon,$$

add $\boldsymbol{\theta}^\star$ *to the set* $\boldsymbol{\Theta}^{(s)}_{ABC,n}$*, i.e.* $\boldsymbol{\Theta}^{(s)}_{ABC,n}(\varepsilon) = \boldsymbol{\Theta}^{(s)}_{ABC,n}(\varepsilon) \cup \{\boldsymbol{\theta}^\star\}$.

For a given $s \in \mathbb{N}^+_S$*, we denote by* $\hat{\boldsymbol{\theta}}^{(s)}_{ABC,n}(\varepsilon)$ *an element of* $\boldsymbol{\Theta}^{(s)}_{ABC,n}(\varepsilon)$*. The collection of the solutions is denoted* $\boldsymbol{\Theta}_{ABC,n}(\varepsilon) = \cup_{s\in\mathbb{N}^+}\boldsymbol{\Theta}^{(s)}_{ABC,n}(\varepsilon)$.

Remark 1.15. *The ABC algorithm presented in Definition 1.14 is a specific version of the simple accept/reject algorithm proposed by [Tav+97; Pri+99], where the auxiliary estimators are the solution of an estimating function and the dimensions of* $\boldsymbol{\pi}$ *and* $\boldsymbol{\theta}$ *are the same.*

Definition 1.16 (posterior distribution)**.** *The distribution of the infinite sequence* $\{\hat{\boldsymbol{\theta}}^{(s)}_{ABC,n}(\varepsilon) : s \in \mathbb{N}^+_S\}$ *issued from Definition 1.14 is referred to as the* $(\varepsilon, \hat{\boldsymbol{\pi}}_n)$*-approximate posterior distribution. If* $\varepsilon = 0$*, we have the* $\hat{\boldsymbol{\pi}}_n$*-approximate posterior distribution. If* $\hat{\boldsymbol{\pi}}_n$ *is a sufficient statistic, we have the* ε*-approximate posterior distribution. If both* $\varepsilon = 0$ *and* $\hat{\boldsymbol{\pi}}_n$ *is sufficient, then we simply refer to the posterior distribution.*

Remark 1.17. *In Definition 1.16, we mention two sources of approximation to the posterior distribution,* ε *and* $\hat{\boldsymbol{\pi}}_n$*. There is actually a third source of approximation stemming from the number of simulations* S*, if indeed* $S < \infty$*. Since it is common to every methods presented, it is left implicit.*

Assumption 1.18 (existence of a prior)**.** *For every $s \in \mathbb{N}_S^+$ and for all n, there exists a prior distribution $\mathscr{P}$ such that*

$$\lim_{\varepsilon \downarrow 0} \Pr\left(d\left(\hat{\boldsymbol{\pi}}_n, \hat{\boldsymbol{\pi}}_{II,n}^{(s)}(\boldsymbol{\theta}^\star)\right) \le \varepsilon\right) = 1, \quad \boldsymbol{\theta}^\star \sim \mathscr{P}.$$

Theorem 1.19 (Equivalence SwiZs/ABC)**.** *If Assumptions 1.7 and 1.18 are satisfied, then the following holds:*

$$\boldsymbol{\Theta}_n^{(s)} = \lim_{\varepsilon \downarrow 0} \boldsymbol{\Theta}_{ABC,n}^{(s)}(\varepsilon).$$

From Theorem 1.19 and Definition 1.16, we have clearly established that the distribution obtained by the SwiZs is a $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution. Yet, the conclusion reached by Theorem 1.19 is surprising at two different levels: first, Theorem 1.19 implies the possibility of obtaining an $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution without specifying explicitly a prior distribution by using the SwiZs, second, whereas, for each $s \in \mathbb{N}_S^+$, it would in general require a very large number of sampled $\boldsymbol{\theta}^\star$ for the ABC to approach an $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution ($\varepsilon = 0$), it is obtainable by the SwiZs at a much reduced cost. Indeed, for a given $s \in \mathbb{N}_S^+$, it demands in general a considerable number of attempts to sample a $\boldsymbol{\theta}^\star$ that satisfies the matching criterion with an error of $\varepsilon \approx 0$, whereas it is replaced by one optimization for the SwiZs, so it may be more computationally efficient to use the SwiZs. Note also that in the situation where one has a prior knowledge on $\boldsymbol{\theta}$, the SwiZs may be modified, for example, by including an importance sampling weight, in the same fashion that the ABC would be modified when the prior distribution is improper (see e.g. [DMDJ06]). However, for some problems, the optimizations to obtain the SwiZs distribution may be numerically cumbersomes and the ABC may prove itself a facilitating alternative (for example [FP12] argued in this direction for some of their examples when comparing the indirect inference and the ABC).

Switching between the SwiZS and the ABC algorithms for estimating a posterior poses the fundamental and practical question of which prior distribution to use. Assumption 1.18 stating that a prior distribution exists is very reasonable and widely accepted (although a frequentist fundamentalist may argue differently), but the result of Theorem 1.19 brings at least three questions: which prior distribution satisfies both the SwiZs and the ABC at the same time, whether the prior distribution under which Theorem 1.19 holds is unique and whether there is an “optimal” prior in the numerical sense (that would produce $\boldsymbol{\theta}^\star$ satisfying “rapidly” the matching criteria as defined at the point 3 of Definition 1.14). We do not answer these questions because, firstly, the numerical problems we face in Section 1.7 are achievable quite efficiently by the SwiZs, secondly, they would deserve much more attention than what we are able to conduct in the present. Thus, we content ourselves by mentioning only briefly studies made on this direction. In order to approach this topic, we first need to present an ultimate technique.

The possibility of obtaining an (approximate) Bayesian posterior without specifying explicitly a prior distribution on the parameters of interest inescapably links the SwiZs to R.A. Fisher's controversial fiducial inference (see for instance [Fis22; Fis30; Fis33; Fis35; Fis56]). Here we keep the SwiZs neutral and do not aim at reanimating any debate. It is delicate to give an unequivocal definition of the fiducial inference as it has changed on many occasion over time (see [Zab92] for a comprehensive historical review) and we rather give the presentation with the generalized fiducial inference proposed by [Han09] (see also [Han13; Han+16]) which includes R.A. Fisher's fiducial inference. Other efforts to generalize R.A. Fisher's fiducial inference include Fraser's structural inference ([Fra68], see also [DSZ73]), the Dempster-Shafer theory ([Sha76; Dem08], see also [ZL11]) refined later with the concept of inferential models ([ML13; ML15]). As argued by [Han09], Fraser's structural inference may be viewed as a special case of the generalized fiducial inference where the generating function $\mathbf{g}$ has a specific structure. The concept of inferential models is similar to the generalized fiducial inference in appearance but they differ in their respective theory. The departure point of the inferential models is to conduct inference with the conditional distribution of the pivotal quantity $\mathbf{u}$ given $\mathbf{x}_0$ after the sample has been observed. It is argued that keeping $\mathbf{u} \sim F_{\mathbf{u}}$ after the sample has been observed makes the whole procedure subjective ([ML15]), but the idea is essentially a gain in efficiency of the estimators. Also this idea is sound (see Lemma 1.30 in the next section), we do not see how it can be applied for the practical examples we use in Section 1.7, and more fundamentally, we do not understand how such conditional distribution may be built without some form of prior (and arguably subjective) knowledge on $\mathbf{u}_0$. We therefore leave such consideration for further research and limit the equivalence to the generalized fiducial inference given in the next definition.

Definition 1.20 (Generalized fiducial inference)**.** *The generalized fiducial distribution is given by*

$$\hat{\boldsymbol{\theta}}^{(s)}_{GFD,n} \in \boldsymbol{\Theta}^{(s)}_{GFD,n} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argzero}}\, d\left(\mathbf{x}, \mathbf{g}\left(\boldsymbol{\theta}, \mathbf{u}^*_s\right)\right).$$

Remark 1.21. *The generalized fiducial distribution in Definition 1.20 is slightly more specific than usually defined in the literature. In Definition 1 in [Han+16], it is given by*

$$\lim_{\varepsilon\downarrow 0}\left[\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left\|\mathbf{x} - \mathbf{g}\left(\boldsymbol{\theta}, \mathbf{u}^*_s\right)\right\| \,\middle|\, \min_{\boldsymbol{\theta}} \left\|\mathbf{x} - \mathbf{g}\left(\boldsymbol{\theta}, \mathbf{u}^*_s\right)\right\| \leq \varepsilon\right],$$

for any norm. Here, in addition, we assume that $\boldsymbol{\Theta}$ contains at least one of, possibly many, zeros.

If we let the sample size equals the dimension of the parameter of interest, $n = p$, then it is obvious from their definitions that the generalized fiducial distribution and the indirect inference estimators are equivalent. We formalize this finding for the sake of the presentation.

Assumption 1.22. *The followings hold:*

i. $\hat{\boldsymbol{\pi}}_n = \mathbf{x}$;

ii. $\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}, \mathbf{u})$.

Proposition 1.23. *If Assumption 1.22 is satisfied, then the following holds:*

$$\boldsymbol{\Theta}^{(s)}_{II,n} = \boldsymbol{\Theta}^{(s)}_{GFD,n}.$$

Also the link between the indirect inference and the generalized fiducial inference seems self-evident, it was, at the best of our knowledge, never mentioned in the literature. It may be explained by the two different goals that each of these methods target, that may respectively be loosely summarized as finding a point-estimate of a complex problem and making Bayesian inference without using a prior distribution. Having established this equivalence, the connection with the SwiZs is direct from Theorem 1.8 and formalize in the next proposition.

Proposition 1.24. *If Assumptions 1.7 and 1.22 are satisfied, then the following holds:*

$$\mathbf{\Theta}_n^{(s)} = \mathbf{\Theta}_{GFD,n}^{(s)}.$$

In the light of Proposition 1.24, the SwiZs may appear equivalent to the generalized fiducial inference under a very restrictive condition. Indeed, the only possibility for Assumption 1.22 to hold is that the sample size must equal the dimension of the problem. But we would be willing to concede that this apparent rigidity is thiner as one may propose to use sufficient statistics with minimal reduction on the sample, thereby leaving n greater than p, and Proposition 1.23 would still hold. Such situation however is confined to problems dealing with exponential families as demonstrated by the Pitman-Koopman-Darmois theorem, so in general, when n is greater than p and the problem at hand is outside of the exponential family, the SwiZs and the generalized fiducial inference are not equivalent.

Although the link between the generalized fiducial inference and the indirect inference has remained silent, the connection with the former to the ABC has been much more emphased. Indeed, the algorithms proposed to solve the generalized fiducial inference problems are mostly borrowed from the ABC literature (see [HLL14]). Therefore, the discussion we conducted above on the numerical aspects of the SwiZs and the ABC still holds here, the SwiZs may be an efficient alternative to solve the generalized fiducial inference problem.

The generalized fiducial inference is also linked by [Han+16] to what may be called “non-informative” prior approaches (see [KW94] for a broad discussion of this concept). More specifically, it appears that some distribution resulting from the generalized fiducial inference corresponds to the posterior distribution obtained by [Fra+10] based on a data-dependent prior proportional to the likelihood function in the absence of information. This result enlarges the previous vision brought by [Lin58] that concluded that R.A. Fisher’s fiducial inference is “Bayes inconsistent” (in the sense that the Bayes’ theorem cannot be invoked) apart from problems on the Gaussian and the gamma distributions. [Lin58]’s results relied on a narrower definition of fiducial inference than brought by the generalized fiducial inference, so whether the generalized fiducial inference has become Bayes consistent for broader problems nor [Fra+10] approach with an uninformative prior is Bayes inconsistent remains an open question. But most importantly, the strong link between the generalized fiducial inference and this non-informative prior approach reveals the common goal towards which of these approaches tends, which might be stated as tackling the individual subjectivism in the Bayesian inference that has been one of the major subject of criticism ever since at least [Fis22].

Last but not least, we complete the loop by the following Corollary which is a consequence of Theorems 1.8, 1.12 and 1.19, and Propositions 1.23 and 1.24.

Corollary 1.25. *We have the followings:*

i. If Assumptions 1.7 and 1.18 are satisfied, then $\mathbf{\Theta}_{II,n}^{(s)} = \lim_{\varepsilon\downarrow 0} \mathbf{\Theta}_{ABC,n}^{(s)}(\varepsilon)$*;*

ii. If Assumptions 1.7, 1.18 and 1.11 are satisfied, then $\mathbf{\Theta}_{Boot,n}^{(s)} = \lim_{\varepsilon\downarrow 0} \mathbf{\Theta}_{ABC,n}^{(s)}(\varepsilon)$*;*

iii. If Assumptions 1.7 and 1.11 are satisfied, then $\mathbf{\Theta}_{II,n}^{(s)} = \lim_{\varepsilon\downarrow 0} \mathbf{\Theta}_{Boot,n}^{(s)}(\varepsilon)$*;*

iv. If Assumptions 1.7, 1.11 and 1.22 are satisfied, then $\mathbf{\Theta}_{Boot,n}^{(s)} = \lim_{\varepsilon\downarrow 0} \mathbf{\Theta}_{GFD,n}^{(s)}(\varepsilon)$*;*

v. If Assumptions 1.7, 1.18 and 1.22 are satisfied, then $\mathbf{\Theta}_{ABC,n}^{(s)} = \lim_{\varepsilon\downarrow 0} \mathbf{\Theta}_{GFD,n}^{(s)}(\varepsilon)$*.*

1.4 Exact frequentist inference in finite sample

Having demonstrated that the distribution of the SwiZs sequence, for a single experiment, is approximatively a Bayesian posterior, we now turn our interest to the long-run statistical properties of the SwiZs. Our point of view here is frequentist, that is we suppose that we have an indefinite number of independent trials with fixed sample size n and fixed $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$. For each experiment we compute an exact α-credible set, as given in the Definition 1.27 below, using the SwiZs independently: the knowledge acquired on an experiment is not used as a prior to compute the SwiZs on another experiment. The goal of this Section is to demonstrate under what conditions the SwiZs leads to exact frequentist inference when the sample size is fixed.

Definition 1.26 (sets of quantiles)**.** *Let $F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}$ be a $\hat{\boldsymbol{\pi}}_n$-approximate posterior cumulative distribution function. We define the following sets of quantiles:*

1. *Let $\underline{Q}_\alpha = \left\{\hat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}_n, \alpha \in (0,1) : F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}(\hat{\boldsymbol{\theta}}_n) \leq \alpha\right\}$ be the set of all $\hat{\boldsymbol{\theta}}_n$ for which $F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}$ is below the threshold α.*
2. *Let $\overline{Q}_\alpha = \left\{\hat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}_n, \alpha \in (0,1) : F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}(\hat{\boldsymbol{\theta}}_n) \geq 1-\alpha\right\}$ be the set of all $\hat{\boldsymbol{\theta}}_n$ for which $F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}$ is above the threshold $1-\alpha$.*

Definition 1.27 (credible set)**.** *Let $F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}$ be a $\hat{\boldsymbol{\pi}}_n$-approximate posterior cumulative distribution function. A set $C_{\hat{\boldsymbol{\pi}}_n}$ is said to be an α-credible set if*

$$\Pr\left(\hat{\boldsymbol{\theta}}_n \in C_{\hat{\boldsymbol{\pi}}_n} | \hat{\boldsymbol{\pi}}_n\right) \geq 1-\alpha, \quad \alpha \in (0,1), \tag{1.1}$$

where

$$C_{\hat{\boldsymbol{\pi}}_n} = \boldsymbol{\Theta}_n \setminus \left\{\underline{Q}_{\alpha_1} \cup \overline{Q}_{\alpha_2}\right\}, \quad \alpha_1 + \alpha_2 = \alpha.$$

If we replace "$\geq$" by the equal sign in (1.1), we say that the coverage probability of $C_{\hat{\boldsymbol{\pi}}_n}$ is exact.

Definition 1.27 is standard in the Bayesian literature (see e.g. [Rob07]). Note that an α-credbile set can have an exact coverage only if the random variable is absolutely continuous. Such credible set is referred to as an "exact α-credible set".

The next result gives a mean to verify the exactness of frequentist coverage of an exact α-credible set.

Proposition 1.28 (Exact frequentist coverage)**.** *If a $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution evaluated at $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_n$ is a realization from a standard uniform variate identically and independently distributed, $F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}_0) = u$, $u \sim \mathcal{U}(0,1)$, then every exact α-credible set built from the quantiles of $F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}$ leads to exact frequentist coverage probability in the sense that $\Pr\left(C_{\hat{\boldsymbol{\pi}}_n} \ni \boldsymbol{\theta}_0\right) = 1-\alpha$ (unconditionally).*

Proposition 1.28 states that if the cumulative distribution function (cdf), obtained from the SwiZs, variates (across independent trials!) uniformly around $\boldsymbol{\theta}_0$ (fixed!), so does any quantities computed from the percentiles of this cdf, leading to exact coverage in the long-run. The proof relies on Borel's strong law of large number. Although this result may be qualified of unorthodox by mixing both Bayesian posterior and frequentist properties, it arises very naturally. Replacing $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution by any conditional distribution on $\hat{\boldsymbol{\pi}}_n$ in Proposition 1.28 leads to the same result. This proposition is similar in form to the concept of confidence distribution formulated by [SH02] and later refined by [SXS+05; XSS11; XS13]. The confidence distribution is however a concept entirely frequentist and could not be directly exploited here. The general theoretical studies on the finite sample frequentist properties are quite rare in the literature, we should eventually mention the study of [Mar15], although the theory developped is around inferential models and different than our, the author uses the same criterion of uniformly distributed quantity to demonstrate the frequentist properties.

Remark 1.29. *In Proposition 1.28, we use a standard uniform variable as a mean to verify the frequentist properties. With the current statement of the proposition, other distributions with support in* $[0,1]$ *may be candidates to verify the exactness of the frequentist coverage. However, if we restrain the frequentist exactness to be* $\Pr(C_{\hat{\boldsymbol{\pi}}_n} \ni \boldsymbol{\theta}_0) = 1-\alpha$, $\Pr(\overline{Q}_{\alpha_2} \ni \boldsymbol{\theta}_0) = \alpha_2$ *and* $\Pr(\underline{Q}_{\alpha_1} \ni \boldsymbol{\theta}_0) = \alpha_1$*, for* $\alpha = \alpha_1+\alpha_2$*, then the uniform distribution would be the only candidate.*

In the light of Proposition 1.28, we now give the conditions under which the distribution of the sequence $\{\hat{\boldsymbol{\theta}}_n^{(s)} : s \in \mathbb{N}^+\}$, $F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}$, leads to exact frequentist coverage probabilities. We begin with a lemma which is essential in the construction of our argument.

Lemma 1.30. *If the mapping* $\boldsymbol{\pi} \mapsto \boldsymbol{\Phi}_n$ *has unique zero in* $\boldsymbol{\Pi}$ *and the mapping* $\boldsymbol{\theta} \mapsto \boldsymbol{\Phi}_n$ *has unique zero in* $\boldsymbol{\Theta}$*, then the following holds*

$$\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}_0, \hat{\boldsymbol{\pi}}_n\right).$$

The idea behind Lemma 1.30 is that if one knew the true pivotal quantity $\mathbf{u}_0$ that generated the data, then one could directly recover the true quantity of interest $\boldsymbol{\theta}_0$ from the sample. Of course, both $\mathbf{u}_0$ and $\boldsymbol{\theta}_0$ are unknown (otherwise statisticians would be an extinct species!), but here we are exploiting the idea that, for a sufficiently large number of simulations S, at some point we will generate $\mathbf{u}_s$ “close enough” to $\mathbf{u}_0$. This idea is reflected in the following assumption.

Assumption 1.31. *Let* $\boldsymbol{\Theta}_n \subseteq \boldsymbol{\Theta}$ *be the set of the solutions of the SwiZs in the Definition 1.2. We have the following:*

$$\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_n.$$

The following functions are essential for convenient data reduction.

Assumption 1.32 (data reduction)**.** *We have:*

i. There exists a Borel measurable surjection such that $\mathbf{b}(\mathbf{u})$ *has the same dimension as* $\mathbf{x}$*.*

ii. There exists a Borel measurable surjection such that $\mathbf{h}\circ\mathbf{b}(\mathbf{u})$ *has the same dimension as* $\boldsymbol{\theta}$*.*

Remark 1.33. *The function* $\mathbf{b}$ *allows to work with a random variable of the same dimension as the observed variable. Indeed we have*

$$\mathbf{x} \overset{d}{=} \mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) \overset{d}{=} \boldsymbol{g} \circ (\mathrm{id}_{\boldsymbol{\Theta}} \times \mathbf{b})(\boldsymbol{\theta}, \mathbf{u}) \overset{d}{=} \boldsymbol{g}(\boldsymbol{\theta}, \mathbf{v}),$$

where $\mathbf{v} = \mathbf{b}(\mathbf{u})$ *has the same dimension as* $\mathbf{x}$ *and* $\mathrm{id}_{\boldsymbol{\Theta}}$ *is the identity function on the set* $\boldsymbol{\Theta}$*. On the other hand, the function* $\mathbf{h}$ *allows us to deal with random variables of the same dimension as* $\boldsymbol{\theta}$*, and thus* $\boldsymbol{\pi}$*.*

Remark 1.34. *In Assumption 1.32, by saying the functions* $\mathbf{h}$ *and* $\mathbf{b}$ *are Borel measurable, we want to emphasis thereby that after applying these functions we still work with random variables, which is essential here.*

To fix ideas, we consider the following example:

Example 1.35 (Explicit form for $\mathbf{h}$ and $\mathbf{b}$)**.** *As in Example 1.1, suppose that* $\mathbf{x} = x_1, \cdots, x_n$ *is identically and independently distributed according to* $\mathcal{N}(\theta, \sigma^2)$*, where* σ^2 *is known, and consider the generating function* $\mathbf{g} \in \mathcal{G}$

$$\mathbf{g}(\theta, \mathbf{u}, \sigma^2) = \theta + \sigma\sqrt{-2\ln(u_1)}\cos(2\pi u_2),$$

where u_{1i}, u_{2i}*,* $i = 1, \cdots, n$*, are identically and independently distributed according to* $\mathcal{U}(0,1)$*. Letting* $\mathbf{v} \equiv \mathbf{b}(\mathbf{u}) = \sqrt{-2\ln(u_1)}\cos(2\pi u_2)$*, we clearly have that* $\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_n)$ *is a random variable of the same dimension as* $\mathbf{x}$*. Now, if we consider* $\mathbf{h}$ *as the function that averages its argument, we have* $w \equiv \mathbf{h} \circ \mathbf{b}(\mathbf{u}) = {}^1\!/_n \sum_{i=1}^n v_i$*, so by properties of Gaussian random variable we have that* w *has a Gaussian distribution with mean 0 and variance* ${}^1\!/_n$*. Since* w *is a scalar, it has the same dimensions as* θ*.*

Example 1.35 shows explicit forms for functions in Assumption 1.32. It is however not requested to have an explicit form as we will see. Indeed, under Assumption 1.32, we can construct the following estimating function:

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}^*, \boldsymbol{\pi}\right) = \boldsymbol{\varphi}_p\left(\boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\pi}\right),$$

where $\mathbf{w} = \mathbf{h} \circ \mathbf{b}(\mathbf{u}^*)$ is a p-dimensional random variable. The index p in the estimating function $\boldsymbol{\varphi}_p$ aims at emphasing that $\mathbf{w}$ has the same dimensions as $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, which is essential in our argument. Since the sample size n and dimension p are fixed here, it is disturbing. For some fixed $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ and $\boldsymbol{\pi}_1 \in \boldsymbol{\Pi}$, it clearly holds that:

$$\begin{aligned}
\hat{\boldsymbol{\pi}}_n &= \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}^*, \boldsymbol{\pi}\right) = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\operatorname{argzero}}\, \boldsymbol{\varphi}_p\left(\boldsymbol{\theta}_1, \mathbf{w}, \boldsymbol{\pi}\right),\\
\hat{\boldsymbol{\theta}}_n &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}^*, \boldsymbol{\pi}_1\right) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argzero}}\, \boldsymbol{\varphi}_p\left(\boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\pi}_1\right).
\end{aligned}$$

Assumption 1.36 (characterization of $\boldsymbol{\varphi}_p$)**.** *Let* $\boldsymbol{\Theta}_n \subseteq \boldsymbol{\Theta}$ *and* W_n *be open subsets of* $\mathbb{R}^p$*. Let* $\hat{\boldsymbol{\pi}}_n$ *be the unique solution of* $\boldsymbol{\Phi}_n(\boldsymbol{\theta}_0, \mathbf{u}_0, \boldsymbol{\pi})$*. Let* $\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w}) \equiv \boldsymbol{\varphi}_p(\boldsymbol{\theta}, \mathbf{w}, \hat{\boldsymbol{\pi}}_n)$ *be the map where* $\hat{\boldsymbol{\pi}}_n$ *is fixed. We have the followings:*

i. $\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n} \in C^1\left(\boldsymbol{\Theta}_n \times W_n, \mathbb{R}^p\right)$ *is once continuously differentiable on* $(\boldsymbol{\Theta}_n \times W_n) \setminus K_n$*, where* $K_n \subset \boldsymbol{\Theta}_n \times W_n$ *is at most countable,*

ii. $\det\left(D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w})\right) \neq 0$*,* $\det\left(D_{\mathbf{w}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w})\right) \neq 0$ *for every* $(\boldsymbol{\theta}, \mathbf{w}) \in (\boldsymbol{\Theta}_n \times W_n) \setminus K_n$*,*

iii. $\lim_{\|(\boldsymbol{\theta},\mathbf{w})\|\to\infty} \|\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w})\| = \infty$*.*

Assumption 1.37 (characterization of $\boldsymbol{\varphi}_p$ II)**.** *Let* $\boldsymbol{\Theta}_n \subseteq \boldsymbol{\Theta}$*,* W_n *and* $\boldsymbol{\Pi}_n \subseteq \boldsymbol{\Pi}$ *be open subsets of* $\mathbb{R}^p$*. Let* $\boldsymbol{\varphi}_{\boldsymbol{\theta}_1}(\mathbf{w}, \boldsymbol{\pi}) \equiv \boldsymbol{\varphi}_p(\boldsymbol{\theta}_1, \mathbf{w}, \boldsymbol{\pi})$ *be the map where* $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ *is fixed. Let* $\boldsymbol{\varphi}_{\mathbf{w}_1}(\boldsymbol{\theta}, \boldsymbol{\pi}) \equiv \boldsymbol{\varphi}_p(\boldsymbol{\theta}, \mathbf{w}_1, \boldsymbol{\pi})$ *be the map where* $\mathbf{w}_1 \in W_n$ *is fixed. We have the followings:*

i. $\boldsymbol{\varphi}_{\boldsymbol{\theta}_1} \in C^1\left(W_n \times \boldsymbol{\Pi}_n, \mathbb{R}^p\right)$ *is once continuously differentiable on* $(W_n \times \boldsymbol{\Pi}_n) \setminus K_{1n}$*, where* $K_{1n} \subset W_n \times \boldsymbol{\Pi}_n$ *is at most countable,* *ii.* $\boldsymbol{\varphi}_{\mathbf{w}_1} \in \mathcal{C}^1\left(\boldsymbol{\Theta}_n \times \boldsymbol{\Pi}_n, \mathrm{I\!R}^p\right)$ *is once continuously differentiable on* $(\boldsymbol{\Theta}_n \times \boldsymbol{\Pi}_n) \setminus K_{2n}$*, where* $K_{2n} \subset \boldsymbol{\Theta}_n \times \boldsymbol{\Pi}_n$ *is at most countable,*

iii. $\det\left(D_{\mathbf{w}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_1}(\mathbf{w}, \boldsymbol{\pi})\right) \neq 0$*,* $\det\left(D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_1}(\mathbf{w}, \boldsymbol{\pi})\right) \neq 0$ *for every* $(\mathbf{w}, \boldsymbol{\pi}) \in (W_n \times \boldsymbol{\Pi}_n) \setminus K_{1n}$*,*

iv. $\det\left(D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\mathbf{w}_1}(\boldsymbol{\theta}, \boldsymbol{\pi})\right) \neq 0$*,* $\det\left(D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\mathbf{w}_1}(\boldsymbol{\theta}, \boldsymbol{\pi})\right) \neq 0$ *for every* $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in (\boldsymbol{\Theta}_n \times \boldsymbol{\Pi}_n) \setminus K_{2n}$*,*

v. $\lim_{\|(\mathbf{w},\boldsymbol{\pi})\| \to \infty} \|\boldsymbol{\varphi}_{\boldsymbol{\theta}_1}(\mathbf{w}, \boldsymbol{\pi})\| = \infty$*,*

vi. $\lim_{\|(\boldsymbol{\theta},\boldsymbol{\pi})\| \to \infty} \|\boldsymbol{\varphi}_{\mathbf{w}_1}(\boldsymbol{\theta}, \boldsymbol{\pi})\| = \infty$*.*

Theorem 1.38. *If Assumptions 1.32 and 1.31 and one of Assumptions 1.36 or 1.37 are satisfied, then the followings hold:*

1. *There is a* $\mathcal{C}^1$*-diffeomorphism map* $\mathbf{a} : W_n \to \boldsymbol{\Theta}_n$ *such that the distribution function of* $\hat{\boldsymbol{\theta}}_n$ *given* $\hat{\boldsymbol{\pi}}_n$ *is*

$$\int_{\boldsymbol{\Theta}_n} f_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}\left(\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n\right) \mathrm{d}\boldsymbol{\theta} = \int_{W_n} f\left(\mathbf{a}(\mathbf{w})|\hat{\boldsymbol{\pi}}_n\right) |J(\mathbf{w}|\hat{\boldsymbol{\pi}}_n)| \,\mathrm{d}\mathbf{w},$$

where

$$J(\mathbf{w}|\hat{\boldsymbol{\pi}}_n) = \frac{\det\left(D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\mathbf{a}(\mathbf{w}), \mathbf{w})\right)}{\det\left(D_{\mathbf{w}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\mathbf{a}(\mathbf{w}), \mathbf{w})\right)}.$$

2. *For all* $\alpha \in (0, 1)$*, every exact* α*-credible set built from the percentiles of the distribution function have exact frequentist coverage probabilities.*

Theorem 1.38 is very powerful as it concludes that the SwiZs (Assumptions 1.32, 1.31 and 1.36) and the indirect inference estimators (Assumption 1.32, 1.31 and 1.37) have exact frequentist coverage probabilities in finite sample. Our argument is based on the possibility of changing variables from $\hat{\boldsymbol{\theta}}_n$ to $\mathbf{w}$, but also from $\mathbf{w}$ to $\hat{\boldsymbol{\theta}}_n$ (hence the diffeomorphism). This argument may appear tautological, but this is actually because we are able to make this change-of-variable in both directions that the conlcusion of Theorem 1.38 is possible (see the parametric bootstrap in Examples 1.52 and 1.54 for counter-examples). The result is very general because we do not suppose that we know explicitly the estimators $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\pi}}_n$, neither the random variable $\mathbf{w}$. Because of their unknown form, we employ a global implicit function theorem for our proof which permits to characterize the derivative of these estimators through their estimating function. One of the conclusion of the global implicit function theorem is the existence of a unique and global invertible function $\mathbf{a}$. It seems not possible to reach the conclusion of Theorem 1.38 with a local implicit function theorem (usually encountered in textbooks), but it may be of interest for further research as some conditions may accordingly be relaxed.

Although powerful, Theorem 1.38's conditions are restrictive or difficult to inspect, but not hard to believe as we now explain. First, the existence of the random variable $\mathbf{w}$ depends on the possibility to have data reduction as expressed in Assumption 1.32. We do not need to know explicitly $\mathbf{w}$ and $\mathbf{w}$ does not need to be unique, so essentially Assumption 1.32 holds for every problem for a which a maximum likelihood estimator exists (see e.g. [HMC05], Theorem 2 in Chapter 7); see also [FFS10; ML15] for the construction of $\mathbf{w}$ by conditioning. Yet, it remains unclear if this condition holds in the situations when the likelihood function does not exist. The indirect inference and ABC literatures are overflowing with examples where the likelihood is not tractable, but one should keep in mind that such situation does not exclude the existence of a maximum likelihood, it is simply impractical to obtain one. Second, Assumption 1.31 states that the true value $\boldsymbol{\theta}_0$ belongs to the set of solutions. This condition can typically only be verified in simulations when controlling all the parameters of the experiment, although it is not critical to believe such condition holds when making a very large number of simulations S. We interpret the inclusion of the set of solutions to $\boldsymbol{\Theta}$ as follows: once $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is fixed, it is not necessary to explore the whole set $\boldsymbol{\Theta}$ (that would require S to be extremly large), but an area sufficiently large of $\boldsymbol{\Theta}$ such that it includes $\boldsymbol{\theta}_0$. Third, Assumptions 1.36 and 1.37 are more technical and concerns the finite sample behavior of the estimating functions of, respectively, the SwiZs and the indirect inference estimators. Although we cannot conclude that Assumption 1.36 is weaker than Assumption 1.37, it seems easier to deal with the former.

Assumption 1.36 (*i*) requires the estimating function to be once continuously differentiable in $\boldsymbol{\theta}$ and $\mathbf{w}$ almost everywhere. The estimators $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\pi}}_n$ are not known in an explicit form, but they can be characterized by their derivatives using an implicit function theorem argument. Since $\boldsymbol{\theta}$ and $\mathbf{w}$ appears in the generating function $\boldsymbol{g}$, this assumption may typically be verified with the example at hand using a chain rule argument: the estimating function must be once continuously differentiable in the observations represented by $\boldsymbol{g}$, and $\boldsymbol{g}$ must be once continuously differentiable in both its arguments. Discrete random variables are automatically ruled out by this last requirement, but this should not appear as a surprise as exactness of the coverage cannot be claimed in general for discrete distribution (see e.g. [Cai05]). The smoothness requirement on the estimating function excludes for example estimators based on order statistics. In general, relying on non-smooth estimating function leads to less efficient estimators and less stable numerical solutions, but they may be an easier estimating function to choose in situations where it is not clear which one to select. Although, non-smooth estimating functions and discrete random variables are dismissed, the condition may nearly be satisfied when considering a n large enough. Assumption 1.37 (*i*, *ii*) requires in addition the estimating equation to be once continuously differentiable in $\boldsymbol{\pi}$.

Assumption 1.36 (*ii*), as well as Assumption 1.37 (*iii*, *iv*), essentially necessitate the estimating function to be "not too flat" globally. It is one of the weakest condition to have invertibility of the Jacobian matrices. Usually only one of the Jacobian has such requirement for an implicit function theorem, but since we are targeting a C^1-diffeomorphism, we strenghten the assumption on both Jacobians. Once verified the first derivative of the estimating function as explained in the preceding paragraph, the non-nullity of determinant may be appreciated, it typically depends on the model and the choosen estimating function. An example for which this condition is not globally satisfied is when considering robust estimators as the estimating function is constant on an uncountable set once exceeding some threeshold. This consideration gives raise to the question on whether this condition may be relaxed to hold only locally, condition which would be satisfied by the robust estimators, but Example 1.61 with the robust Lomax distribution in the Section 1.7 seems to indicate the opposite direction.

Assumption 1.36 (*iii*), as well as Assumption 1.37 (*v*, *vi*), is a necessary and sufficient condition to invoke Palais' global inversion theorem ([Pal59]) which is a key component of the global implicit function theorem of [Cri17] we use. It can be verified in two steps by, first, letting $\boldsymbol{g}$ diverges in the estimating function, and then letting $\boldsymbol{\theta}$ and $\mathbf{w}$ diverges in $\boldsymbol{g}$. Once again, robust estimators do not fulfill this requirement as their estimating functions do not diverge with $\boldsymbol{g}$ but rather stay constant.

Theorem 1.38 is derived under sufficient conditions. In its actual form, although very general, it excludes some specific estimating functions and non-absolutly continuous random variable. It is of both practical and theoretical interest to develop results for a wider-range of situations. Such considerations are left for further research.

We finish this section by considering a special, though maybe common, case where the auxiliary estimator is known in an explicit form. Suppose $\hat{\boldsymbol{\pi}}_n = \mathbf{h}(\mathbf{x}_0)$ where $\mathbf{h}$ is a known (surjective) function of the observations (see Assumption 1.32). We can define a (new) indirect inference estimator as follows:

$$\hat{\boldsymbol{\theta}}^{(s)}_{\mathrm{II},n} \in \boldsymbol{\Theta}^{(s)}_{\mathrm{II},n} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argzero}}\, d\left[\mathbf{h}(\mathbf{x}_0), \mathbf{g}(\boldsymbol{\theta}, \mathbf{w}_s)\right]. \tag{1.2}$$

Remark 1.39. *The estimator defined in Equation 1.2 is a special case of the indirect inference estimators as expressed in Definition 1.4, and thus of the SwiZs by Theorem 1.8, where the auxiliary estimators $\hat{\boldsymbol{\pi}}_n$ and $\hat{\boldsymbol{\pi}}_{II,n}$ are known in an explicit form.*

Assumption 1.40 (characterization of $\boldsymbol{g}$)**.** *Let $\boldsymbol{\Theta}_n \subseteq \boldsymbol{\Theta}$, W_n be subsets of $\mathrm{I\!R}^p$ and $K_n \subset \boldsymbol{\Theta}_n \times W_n$ be at most countable. The followings hold:*

i. $\boldsymbol{g} \in \mathcal{C}^1\left(\boldsymbol{\Theta}_n \times W_n, \mathrm{I\!R}^p\right)$ *is once continuously differentiable on* $\left(\boldsymbol{\Theta}_n \times W_n\right) \setminus K_n$,

ii. $\det(D_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{\theta}, \mathbf{w})) \neq 0$ *and* $\det(D_{\mathbf{w}}\boldsymbol{g}(\boldsymbol{\theta}, \mathbf{w})) \neq 0$ *for every* $(\boldsymbol{\theta}, \mathbf{w}) \in (\boldsymbol{\Theta}_n \times W_n) \setminus K_n$,

iii. $\lim_{\|(\boldsymbol{\theta},\mathbf{w})\|\to\infty} \|\boldsymbol{g}(\boldsymbol{\theta}, \mathbf{w})\| = \infty$.

Proposition 1.41. *If Assumptions 1.32, 1.31 and 1.40 are satisfied, then the conclusions (1) and (2) of Theorem 1.38 hold. In particular, the distribution function is:*

$$\int_{\boldsymbol{\Theta}_n} f_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}\left(\boldsymbol{\theta}|\mathbf{h}(\mathbf{x}_0)\right) \mathrm{d}\boldsymbol{\theta} = \int_{W_n} f\left(\mathbf{a}(\mathbf{w})|\mathbf{h}(\mathbf{x}_0)\right) |J(\mathbf{w}|\mathbf{h}(\mathbf{x}_0))| \,\mathrm{d}\mathbf{w},$$

where

$$J(\mathbf{w}|\mathbf{h}(\mathbf{x}_0)) = \frac{\det\left(D_{\boldsymbol{\theta}}\boldsymbol{g}(\mathbf{a}(\mathbf{w}), \mathbf{w})\right)}{\det\left(D_{\mathbf{w}}\boldsymbol{g}(\mathbf{a}(\mathbf{w}), \mathbf{w})\right)}.$$

The message of Proposition 1.41 is fascinating: once the auxiliary estimator is known in an explicit form, the conditions to reach the conclusion of Theorem 1.38 simplify accounting for the fact that the implicit function theorem is no longer necessary. The discussion we have after Theorem 1.38 still holds, but the verification process of the conditions is reduced to inspecting the generating function.

1.5 Asymptotic properties

When $n \to \infty$, different assumptions than in Section 1.4 may be considered to derive the distribution of the SwiZs. By Theorem 1.8, the SwiZs in Definition 1.2 and the indirect inference estimators in Definition 1.4 are equivalent for any n. Yet, due to their different forms, the conditions to derive their asymptotic properties differ, at least in appearance. We treat both the asymptotic properties of the SwiZs and the indirect inference estimators in an unified fashioned and highlight their differences. We do not attempt at giving the weakest conditions possible as our goal is primarly to demonstrate in what theoretical aspect the SwiZs is different from the indirect inference estimators. The asymptotic properties of the indirect inference estimators were already derived by several authors in the literature, and we refer to [GM96], Chapter 4, for the comparison.

The following conditions are sufficient to prove the consistency of any estimator $\hat{\boldsymbol{\theta}}_n^{(s)}$ in Defintions 1.2 and 1.4. When it is clear from the context, we simply drop the suffix and denote $\hat{\boldsymbol{\theta}}_n$ for any of these estimators.

Assumption 1.42. *The followings hold:*

i. The sets $\boldsymbol{\Theta}, \boldsymbol{\Pi}$ *are compact,*

ii. For every $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \boldsymbol{\Pi}$*,* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and* $\mathbf{u} \sim F_{\mathbf{u}}$*, there exists a random value* $A_n = \mathcal{O}_p(1)$ *such that, for a sufficiently large* n*,*

$$\left\|\boldsymbol{\Phi}_n(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi}_1) - \boldsymbol{\Phi}_n(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi}_2)\right\| \leq A_n \left\|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\right\|,$$

iii. For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*,* $\boldsymbol{\pi} \in \boldsymbol{\Pi}$*, the estimating function* $\boldsymbol{\Phi}_n\left(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi}\right)$ *converges pointwise to* $\boldsymbol{\Phi}(\boldsymbol{\theta},\boldsymbol{\pi})$*.*

iv. For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*,* $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \boldsymbol{\Pi}$*, we have*

$$\boldsymbol{\Phi}\left(\boldsymbol{\theta},\boldsymbol{\pi}_1\right) = \boldsymbol{\Phi}\left(\boldsymbol{\theta},\boldsymbol{\pi}_2\right),$$

if and only if $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$*.*

Assumption 1.43 (SwiZs)**.** *The followings hold:*

i. For every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$*,* $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ *and* $\mathbf{u} \sim F_{\mathbf{u}}$*, there exists a random value* $B_n = \mathcal{O}_p(1)$ *such that, for sufficiently large* n*,*

$$\left\|\boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}) - \boldsymbol{\Phi}_n(\boldsymbol{\theta}_2,\mathbf{u},\boldsymbol{\pi})\right\| \leq B_n \left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|,$$

ii. For every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$*,* $\boldsymbol{\pi} \in \boldsymbol{\Pi}$*, we have*

$$\boldsymbol{\Phi}\left(\boldsymbol{\theta}_1,\boldsymbol{\pi}\right) = \boldsymbol{\Phi}\left(\boldsymbol{\theta}_2,\boldsymbol{\pi}\right),$$

if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$*.*

Assumption 1.44 (IIE)**.** *The followings hold:*

i. For every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$*, there exists a random value* $C_n = \mathcal{O}_p(1)$ *such that, for sufficiently large* n*,*

$$\left\|\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta}_1) - \hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta}_2)\right\| \leq C_n \left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|;$$

ii. Let $\boldsymbol{\pi}(\boldsymbol{\theta})$ *denotes the mapping towards which* $\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta})$ *converges pointwise for every* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*. For every* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$*, we have*

$$\boldsymbol{\pi}(\boldsymbol{\theta}_1) = \boldsymbol{\pi}(\boldsymbol{\theta}_2),$$

if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$*.*

Theorem 1.45 (consistency)**.** *Let* $\{\hat{\boldsymbol{\pi}}_n\}$ *be a sequence of estimators of* $\{\boldsymbol{\Phi}_n(\boldsymbol{\pi})\}$*. For any fix* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, let* $\{\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta})\}$ *be the sequence of estimators of* $\{\boldsymbol{\Phi}_n(\boldsymbol{\theta},\boldsymbol{\pi})\}$*. Let* $\{\hat{\boldsymbol{\theta}}_n\}$ *be a sequence of estimators of* $\{\boldsymbol{\Phi}_n(\boldsymbol{\theta})\}$*. We have the following:*

1. *If Assumption 1.42 holds, then any sequence* $\{\hat{\boldsymbol{\pi}}_n\}$ *converges in probability to* $\boldsymbol{\pi}_0$ *and any sequence* $\{\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta})\}$ *converges in probability to* $\boldsymbol{\pi}(\boldsymbol{\theta})$*;*

2. *Moreover, if one of Assumptions 1.43 or 1.44 holds, then any sequence $\{\hat{\boldsymbol{\theta}}_n\}$ converges in probability to $\boldsymbol{\theta}_0$.*

Theorem 1.45 demonstrates the consistency of $\hat{\boldsymbol{\theta}}_n$ under two sets of conditions. Assumptions 1.42 and 1.44, or the conditions that are implied by these Assumptions, are regular in the literature of the indirect inference estimators (see [GM96], Chapter 4). More specifically, the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}$, usually referred to as the “binding” function (see e.g. [GMR93]) or the “bridge relationship” (see [JT04]), is central in the argument and is required to have a one-to-one relationship (Assumption 1.44 (*ii*)). Surprisingly, in Theorem 1.45, such requirement may be substitued by the bijectivity of the deterministic estimating function with respect to $\boldsymbol{\theta}$ (Assumption 1.43 (*ii*)). Whereas the bijectivity of $\boldsymbol{\pi}(\boldsymbol{\theta})$ can typically only be assumed (if $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}$ was known explicitly, then one would not need to use the indirect inference estimator unless of course one would be willing to lose statistical efficiency and numerical stability for no gain), there is more hope for Assumption 1.43 (*ii*) to be verifiable. Since both Assumptions 1.43 and 1.44 leads to the same conclusion, one would expect some strong connections between them. Since $\boldsymbol{\pi}(\boldsymbol{\theta})$ may be interpreted as the implicit solution of $\boldsymbol{\Phi}(\boldsymbol{\theta}, \boldsymbol{\pi}(\boldsymbol{\theta})) = \mathbf{0}$, it seems possible to link both Assumptions with the help of an implicit function theorem, but it typically requires further conditions on the derivatives of $\boldsymbol{\Phi}$ that are not necessary for obtaining the consistency results, and we thus leave such considerations for further research.

Proving the consistency of an estimator relies on two major conditions (see Lemma 3.1): the uniform convergence of the stochastic objective function and the bijectivity of the deterministic objective function (Assumption 1.42 (*iv*), Assumption 1.43 (*ii*), Assumption 1.44 (*ii*)). This second condition is referred to as the identifiability condition. It can sometimes be verified, or sometimes it is only assumed to hold, but it is typically appreciated in accordance with the chosen probabilistic model. Discrepancy among approaches mainly occurs on the demonstration of the uniform convergence. Here we rely on a stochastic version of the classical Arzelà-Ascoli theorem (see Lemma 3.3), see [Vaa98] for alternative approaches based on the theory of empirical processes. To satisfy this theorem, we require the parameter sets to be compact (Assumption 1.42 (*i*)), the stochastic objective function to converges pointwise (Assumption 1.42 (*iii*)) and the stochastic objective function to be Lipschitz (Assumption 1.42 (*ii*), Assumption 1.43 (*i*), Assumption 1.44 (*i*)). Note that the last requirement is in fact for the objective function to be stochastically equicontinuous, requirement verified by the Lipschitz condition (see Lemma 3.4), see also [PP94] for a broad discussion on this condition and alternatives. Some authors proposed to relax the compactness condition, see for example [Hub67], but this is generally not a sensitive issue in practice. The pointwise convergence of the stochastic objective function may be appreciated up to further details depending on the context. For identically and independently distributed observations, typically the weak law of large numbers may be employed, thus requiring the stochastic objective function to have the same finite expected value across the observations. Other law of large numbers results may be used for serially dependent processes (see the Chapter 7 of [Ham94]) and for non-identically distributed processes (see [And88]), each results having its own conditions to satisfy.

We now turn our interest to the asymptotic distribution of an estimator $\hat{\boldsymbol{\theta}}_n$. Likewise the consistency result, the following sufficient conditions, are separated to outline the difference between the SwiZs and the indirect inference estimators.

Assumption 1.46. *The followings hold:*

i. Let $\boldsymbol{\Theta}^\circ, \boldsymbol{\Pi}^\circ$, the interior sets of $\boldsymbol{\Theta}, \boldsymbol{\Pi}$, be open and convex subsets of $\mathrm{I\!R}^p$, *ii.* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^\circ$ *and* $\boldsymbol{\pi}_0 \in \boldsymbol{\Pi}^\circ$*,*

iii. $\boldsymbol{\Phi}_n \in \mathcal{C}^1\left(\boldsymbol{\Theta}^\circ \times \boldsymbol{\Pi}^\circ, \mathbb{R}^p \times \mathbb{R}^p\right)$ *when n is sufficiently large,*

iv. For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}^\circ, \boldsymbol{\pi} \in \boldsymbol{\Pi}^\circ$*,* $D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi}), D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi})$ *converge pointwise to* $D_{\boldsymbol{\pi}}\boldsymbol{\Phi}(\boldsymbol{\theta},\boldsymbol{\pi}) \equiv \mathbf{K}(\boldsymbol{\theta},\boldsymbol{\pi}), D_{\boldsymbol{\theta}}\boldsymbol{\Phi}(\boldsymbol{\theta},\boldsymbol{\pi}) \equiv \mathbf{J}(\boldsymbol{\theta},\boldsymbol{\pi})$*,*

v. $\mathbf{K} \equiv \mathbf{K}(\boldsymbol{\theta}_0,\boldsymbol{\pi}_0), \mathbf{J} \equiv \mathbf{J}(\boldsymbol{\theta}_0,\boldsymbol{\pi}_0)$ *are nonsingular,*

vi. $n^{1/2}\boldsymbol{\Phi}_n(\boldsymbol{\theta}_0,\mathbf{u},\boldsymbol{\pi}_0) \rightsquigarrow \mathcal{N}\left(\mathbf{0},\mathbf{Q}\right)$*,* $\|\mathbf{Q}\|_\infty < \infty$*.*

Assumption 1.47 (SwiZs II)**.** *For every* $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \boldsymbol{\Pi}^\circ$*,* $\boldsymbol{\theta} \in \boldsymbol{\Theta}^\circ$ *and* $\mathbf{u} \sim F_{\mathbf{u}}$*, there exists a random value* $E_n = \mathcal{O}_p(1)$ *such that, for sufficiently large n,*

$$\left\|D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi}_1) - D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi}_2)\right\| \leq E_n \left\|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\right\|.$$

Assumption 1.48 (IIE II)**.** *The followings hold:*

i. $\hat{\boldsymbol{\pi}}_{II,n} \in \mathcal{C}^1(\boldsymbol{\Theta}^\circ, \mathbb{R}^p)$ *for sufficiently large n;*

ii. For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}^\circ$*,* $D_{\boldsymbol{\theta}}\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta})$ *converges pointwise to* $D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta})$*.*

Theorem 1.49 (asymptotic normality)**.** *If the conditions of Theorem 1.45 are satisfied, we have the following additional results:*

1. *If Assumption 1.46 holds, then*

$$n^{1/2}\left(\hat{\boldsymbol{\pi}}_n - \boldsymbol{\pi}_0\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{K}^{-1}\mathbf{Q}\mathbf{K}^{-T}\right),$$

and

$$n^{1/2}\left(\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta}) - \boldsymbol{\pi}(\boldsymbol{\theta})\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{K}^{-1}\mathbf{Q}\mathbf{K}^{-T}\right);$$

2. *Moreover, if Assumption 1.47 or 1.48 holds, then*

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, 2\mathbf{J}^{-1}\mathbf{Q}\mathbf{J}^{-T}\right).$$

Theorem 1.49 gives the asymptotic distribution of both the auxiliary estimator and the estimator of interest. The conditions to derive the asymptotic distribution of the auxiliary estimator as expressed in Assumption 1.46 is regular for most estimators in the statistical literature. The proof of the first statement relies on the possibility to apply a delta method (see Lemma 3.8), which requires the estimating function to be once continuously differentiable (Assumption 1.46 (*i*), (*ii*) and (*iii*)). The case where this condition is not met is typically when $\boldsymbol{\theta}_0$ is a boundary point of $\boldsymbol{\Theta}$. Not devoid of interest, this case is atypical and deserve to be treated on its own, this situation is therefore excluded by Assumption 1.46 (*ii*). In contrast, relaxing the smoothness requirement on the estimating function has received a much larger attention in the literature (see [Hub67; NM94; Vaa98] among others). Here we content ourselves with the stronger smooth condition on the estimating function (Assumption 1.46 (*iii*)), maybe because it is largely admitted, but also maybe because the smoothness of the estimating function is already required when n is finite by Theorem 1.38 to demonstrate the exact coverage probabilities, a situation that encourages us to consider smooth estimating function in the practical examples. The conditions for the Jacobian matrices to exist (Assumption 1.46 (*iv*)) and to be invertible (Assumption 1.46 (*v*)) are regular ones. The last condition is that a central limit theorem is applicable on the estimation equation (Assumption 1.46 (*vi*)). This statement is very general and its validity depends upon the context. For identically and independently distributed observations, one typically needs to verify Lindeberg's conditions ([Lin22]), which essentially requires that the two first moments exist and are finite. The requirements are similar if the observations are non-identically observed (see e.g. [Bil12]). The conditions are also similar for stationary processes (see e.g. [Wu11], for a review). Note eventually that, also as minor as it might be, the delta method (which is essentially a mean value theorem) largely in use in the statistical literature has recently been shown to be wrongly used by many for vector-valued function ([Fen+13]), this flaw has been taken into account in the present (see Lemma 3.8 for more details).

The proof of the second statement of Theorem 1.49 on the asymptotic distribution of the estimator of interest is more specific to the indirect inference literature. Compared to the proof of the first statement, it requires in addition that, for n large enough, the binding function to be asymptotically differentiable with respect to $\boldsymbol{\theta}$ for the indirect inference estimator (Assumption 1.48) or the derivative of the estimating function with respect to $\boldsymbol{\theta}$ to be stochastically Lipschitz for the SwiZs (Assumption 1.47). For the same arguments we presented after the consistency Theorem 1.45, it may be more practical to verify Assumption 1.47 as the verification of Assumption 1.48 is impossible, at least directly, as the binding function is unknown. This is actually not entirely true as one may express the derivative of the binding function by invoking an implicit function theorem, the condition then may be verified on the resulting explicit derivative. The proof we use under Assumption 1.48 uses this mechanism, the derivative of the binding function is thus given by

$$D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta}) = -\mathbf{K}^{-1}\mathbf{J},$$

for every $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ (see the proof in Appendix for more details). It is only by using this implicit function theorem argument that the exact same explicit distribution for both the SwiZs and the indirect inference estimators may be obtained. The same idea may be used then to find the derivative of $\hat{\boldsymbol{\pi}}_{\mathrm{II},n}(\boldsymbol{\theta})$ and verify Assumption 1.48. Note eventually that [GM96] have an extra condition not required here (but that would as well be required) because they include a stochastic covariate with their indirect inference estimator.

Having demonstrated the asymptotic properties of one of the SwiZs estimators, $\hat{\boldsymbol{\theta}}_n^{(s)}$, $s \in \mathbb{N}_S^+$, we finish this section by giving the property of the average of the SwiZs sequence. The mean is an interesting estimator on its own and it is often considered as a point estimate in a Bayesian context.

Proposition 1.50. *Let $\bar{\boldsymbol{\theta}}_n$ be the average of $\{\hat{\boldsymbol{\theta}}_n^{(s)} : s \in \mathbb{N}_S^+\}$. If the conditions of Theorem 1.49 are satisfied, then it holds that*

$$n^{1/2}\left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \gamma\mathbf{J}^{-1}\mathbf{Q}\mathbf{J}^{-T}\right),$$

where the factor $\gamma = 1 + 1/S$.

The discussion of the proof and the condition to obtain Theorem 1.49 are also valid for Proposition 1.50. The only point that deserves further explanations is on the factor γ. This factor accounts for the numerical approximation of the $\hat{\boldsymbol{\pi}}_n$-approximate posterior when S is finite. It is not surprising though for someone familiar with the indirect inference literature. What may appear unclear is how this factor pass from 2 for one the SwiZs estimate in Theorem 1.49 to $\gamma < 2$ for the mean in Proposition 1.49. If the $\{\hat{\boldsymbol{\theta}}_n^{(s)} :$

$s \in \mathbb{N}_S^+\}$ are independent, then it is well-known from the properties of the convolution of independent Gaussian random variables that γ should equal 2. In fact, the pivotal quantities $\{\mathbf{u}_s : s \in \mathbb{N}_S^+\}$ are indeed independent, but each of the $\{\hat{\boldsymbol{\theta}}_n^{(s)} : s \in \mathbb{N}_S^+\}$ shares a "common factor", namely $\hat{\boldsymbol{\pi}}_n$, and thus this common variability may be reduced by increasing S. Note eventually that the average estimator in Proposition 1.50 has the same asymptotic distribution as the two indirect inference estimators considered by [GMR93] (given that the dimension of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ matches and that our implicit function theorem argument is used).

1.6 Examples

In this section, we illustrate the finite sample results of the Section 1.4 with some examples for which explicit solutions exist. Indeed, for all the examples, we are able to demonstrate analytically that the SwiZs' $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution follows a uniform distribution when evaluated at the true value $\boldsymbol{\theta}_0$, and thus concluding by Proposition 1.28 that any confidence regions built from the percentiles of this posterior have exact coverage probabilities in the long-run. In addition, and maybe more surprisingly, for most examples we are able to derive the explicit posterior distribution that the SwiZs targets. This message is formidable, one may not even need computations to characterize the distribution of $\hat{\boldsymbol{\theta}}_n$ given $\hat{\boldsymbol{\pi}}_n$, but as one may foresee, these favorable situations are limited in numbers. Lastly, we illustrate Proposition 1.13 on the equivalence between the SwiZs and the parametric bootstrap with a Cauchy random variable in Example 1.51 to conclude that they are indeed the same. Since the SwiZs and the parametric bootstrap are seldom equivalent (see the discussion after Theorem 1.12), we also demonstrate the nonequivalence of the two methods in the case of uniform random variable with unknown upper bound (Example 1.52) and a gamma random variable with unknown rate (Example 1.54). The considerations of this section are not only theoretical but also practical as we treat the linear regression (Example 1.56) and the geometric Brownian motion when observed irregularly (Example 1.59), two models widely use.

Example 1.51 (Cauchy with unknown location). *Let $x_i \sim$ Cauchy(θ, σ), $\sigma > 0$ known, $i = 1, \dots, n$, be identically and independently distributed. Consider the generating function $g(\theta, u) = \theta + u$ where $u \sim$ Cauchy$(0, \sigma)$ and the average as the (explicit) auxiliary estimator, $\hat{\pi}_n = \bar{x}$. We have*

$$\hat{\pi}_{II,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(\theta, u_i) = \theta + w,$$

where $w = \frac{1}{n}\sum_{i=1}^{n} u_i$. By the properties of the Cauchy distribution, we have that $w \sim$ Cauchy$(0, \sigma)$, that is the average of independent Cauchy variables has the same distribution of one of its components. Let $\hat{\theta}_n$ be the solution of $d(\hat{\pi}_n, \hat{\theta}_n + w) = 0$, hence we have the explicit solution $\hat{\theta}_n = \hat{\pi}_n - w$. Note that by symmetry of w around 0 we have $w \overset{d}{=} -w$, so $\hat{\theta}_n = \hat{\pi}_n + w$. We therefore have that

$$\begin{aligned}\Pr\left(\hat{\theta}_n \leq \theta_0 | \hat{\pi}_n\right) &= \Pr\left(\hat{\pi}_n + w \leq \theta_0 | \hat{\pi}_n\right) \\ &= \Pr\left(\theta_0 - w_0 + w \leq \theta_0 | \theta_0, w_0\right) \\ &= \Pr\left(w \leq w_0\right) \sim \mathcal{U}(0,1),\end{aligned}$$

and by Proposition 1.28 the coverage obtained on the percentiles of the distribution of $\hat{\theta}_n|\hat{\pi}_n$ are exact in the long-run (frequentist).

The distribution of $\hat{\theta}_n|\hat{\pi}_n$ can be known in an explicit form. From the solution of $\hat{\theta}_n$, we let $w = a(\theta) = \hat{\pi}_n + \theta$. Following Proposition 1.41, we have

$$f_{\hat{\theta}_n}(\theta|\hat{\pi}_n) = f_w(a(\theta)|\hat{\pi}_n)\left|\frac{\frac{\partial}{\partial\theta}g(\theta,w)}{\frac{\partial}{\partial w}g(\theta,w)}\right|.$$

Since $g(\theta,w) = \theta + w$, the scaling factor is 1 and $\hat{\theta}_n|\hat{\pi}_n \sim$ Cauchy$(\hat{\pi}_n,\sigma)$.

Eventually, we illustrate Theorem 1.12, more specifically Proposition 1.13, by showing that the parametric bootstrap is equivalent. The bootstrap estimators is $\hat{\theta}_{Boot,n} = \frac{1}{n}\sum_{i=1}^n g(\hat{\pi}_n,u_i) = \hat{\pi}_n + w$. It follows immediately that $\hat{\theta}_n = \hat{\theta}_{Boot,n}$ and both estimators are equivalently distributed.

Example 1.52 (uniform with unknown upper bound)**.** *Let $x_i \sim \mathcal{U}(0,\theta)$, $i = 1,\dots,n$, be identically and independently distributed. Consider the generating function $g(\theta,u) = u\theta$ where $u \sim \mathcal{U}(0,1)$ and the (explicit) auxiliary estimator $\max_i x_i$. Clearly, $\max_i x_i = \theta\max_i u_i$. Denote $w = \max_i u_i$ so the auxiliary estimator on the sample is $\hat{\pi}_n = w_0\theta_0$. Now define the estimator $\hat{\theta}_n$ to be the solution such that $d(\hat{\pi}_n,\hat{\theta}w) = 0$. An explicit solution exists and is given by $\hat{\theta}_n = \frac{\theta_0 w_0}{w}$. We therefore have that*

$$\Pr\left(\hat{\theta}_n \leq \theta_0|\hat{\pi}_n\right) = \Pr\left(\frac{\theta_0 w_0}{w} \leq \theta_0|\theta_0,w_0\right) = \Pr\left(w^{-1} \leq w_0^{-1}\right) \sim \mathcal{U}(0,1),$$

and by Proposition 1.28 the coverage obtained on the percentiles of the distribution of $\hat{\theta}_n$ are exact in the frequentist sense.

We can even go further by expliciting the distribution of $\hat{\theta}_n$ given $\hat{\pi}_n$. Let define the mapping $a(\theta) = \frac{\theta_0 w_0}{\theta}$. By the change-of-variable formula we obtain:

$$f_{\hat{\theta}_n}(\theta|\hat{\pi}_n) = f_w(a(\theta)|\hat{\pi}_n)\left|\frac{\partial}{\partial\theta}a(\theta)\right|.$$

The maximum of n standard uniform random variables has the density $f_w(w) = nw^{n-1}$. The derivative is given by $\partial a(\theta)/\partial\theta = -\theta_0 w_0/\theta^2$. Note that by Proposition 1.41 we equivalently have

$$\left.\frac{\frac{\partial}{\partial\theta}g(\theta,w)}{\frac{\partial}{\partial w}g(\theta,w)}\right|_{w=a(\theta)} = \left.\frac{w}{\theta}\right|_{w=\theta_0 w_0/\theta} = \frac{\theta_0 w_0}{\theta^2}.$$

Hence, we eventually obtain:

$$f_{\hat{\theta}_n}(\theta|\hat{\pi}_n) = \frac{n\hat{\pi}_n^n}{\theta^{n+1}}, \quad \hat{\pi}_n = \theta_0 w_0.$$

Note that $\hat{\pi}_n$ is a sufficient statistic. Therefore we have obtained that the posterior distribution of $\hat{\theta}_n$ given $\hat{\pi}_n$ is a Pareto distribution parametrized by $\hat{\pi}_n$, the minimum value of the support, and the sample size n, as the shape parameter.

In view of the preceding display, it is not difficult to develop a similar result for the parametric bootstrap (see the Definition 1.9). The bootstrap estimator solution is simple, it is given by $\hat{\theta}_{Boot,n} = \max_i u_i\hat{\pi}_n = \theta_0 w_0 w$. We thus obtain

$$\Pr\left(\hat{\theta}_{Boot,n} \leq \theta_0|\hat{\pi}_n\right) = \Pr\left(\theta_0 w_0 w \leq \theta_0|\theta_0,w_0\right) = \Pr\left(w \leq w_0^{-1}\right),$$

so it cannot be concluded that $F_{\hat{\theta}_{Boot,n}|\hat{\pi}_n}(\theta_0)$ follows a uniform distribution and we cannot invoke Proposition 1.28. Note that however we cannot exclude that the parametric bootstrap leads to exact coverage probability in virtue of Proposition 1.28 (see Remark 1.29). The parametric bootstrap is well-known to be inadequate in such problem. This fact may be made more explicit as we give now the distribution of the parametric bootstrap estimators. Let define the mapping $w = b(\tilde{\theta}) = \frac{\tilde{\theta}}{\theta_0 w_0}$. Note that $b(\theta_0) = 1/w_0 \neq w_0$. We obtain by the change-of-variable formula

$$f_{\hat{\theta}_{Boot,n}}\left(\tilde{\theta}|\hat{\pi}_n\right) = f_w\left(b(\tilde{\theta})|\hat{\pi}_n\right)\left|\frac{\partial}{\partial\tilde{\theta}}b(\tilde{\theta})\right| = \frac{n\tilde{\theta}^{n-1}}{\hat{\pi}_n^n}.$$

This distribution is known to be the power-function distribution, a special case of the Pearson Type I distribution (see [JKB94]). More interestingly, we have the following relationship between the parametric bootstrap and the SwiZs estimates:

$$\hat{\theta}_{Boot,n} \overset{d}{=} \frac{1}{\hat{\theta}_n}.$$

Ultimately, note that the support of the distribution of $\hat{\theta}_{Boot,n}$ is $(0, \hat{\pi}_n)$ whereas it is $(\hat{\pi}_n, +\infty)$ for the SwiZs, so both distributions never cross! Since $\hat{\pi}_n$ is systematically bias downward the true value θ_0, the coverage of the parametric bootstrap is always null. We illustrate this fact in the next figure.

Example 1.53 (exponential with unknown rate parameter)**.** *Let $x_i \sim \mathcal{E}(\theta)$, $i = 1, \ldots, n$, be identically and independently distributed. Consider the generating function $g(\theta, u) = \frac{u}{\theta}$, where $u \sim \Gamma(1, 1)$, and the inverse of the average as auxiliary estimator, denoted $\bar{x}^{-1}$. Clearly we have $\bar{x}^{-1} = \theta/w$, where $w = \sum_{i=1}^n u_i/n$, so $\hat{\pi}_n = \theta_0/w_0$. The solution of $d(\hat{\pi}_n, \theta/w) = 0$ in θ is denoted $\hat{\theta}_n$, it is given by $\hat{\theta}_n = \theta_0 w/w_0 = w\hat{\pi}_n$. We therefore have*

$$\Pr\left(\hat{\theta}_n \leq \theta_0|\hat{\pi}_n\right) = \Pr\left(w \leq w_0\right) \sim \mathcal{U}(0, 1).$$

It results from Proposition 1.28 that any intervals built from the percentiles of the distribution of $\hat{\theta}_n$ has exact frequentist coverage. The distsribution can be found in explicit form. We have by the additive property of the Gamma distribution that $w \sim \Gamma(n, 1/n)$ (shape-rate parametrization). It immediately results from the change-of-variable formula that

$$\hat{\theta}_n|\hat{\pi}_n \sim \Gamma\left(n, \sum_{i=1}^n x_i\right).$$

Note that $\hat{\pi}_n$ is a sufficient statistic so the obtained distribution is a posterior distribution.

This last example on an exponential variate can be (slightly) generalized to a gamma random variable as follows.

Example 1.54 (gamma with unknown rate parameter)**.** *Consider the exact same setup as in Example 1.53 with the exception that $x_i \sim \Gamma(\alpha, \theta)$ and $u \sim \Gamma(\alpha, 1)$, where $\alpha > 0$ is a known shape parameter. Following the same steps as in Example 1.53 we find the following posterior distribution:*

$$\hat{\theta}_n|\hat{\pi}_n \sim \Gamma\left(\alpha n, \sum_{i=1}^n x_i\right).$$

We also have that any intervals built from the percentiles of the posterior have exact frequentist coverage probabilities.

In view of this display and Example 1.53, we can derive the distribution of the parametric bootstrap. The estimator is obtained as follows:

$$\hat{\theta}_{Boot,n} = \frac{n}{\sum_{i=1}^n g(\hat{\pi}_n, u_i)} = \frac{\hat{\pi}_n}{w},$$

where $w \sim \Gamma(n\alpha, 1/n)$. It follows by the inverse of gamma variate and the change-of-variable formula that

$$\hat{\theta}_{Boot,n} \sim \Gamma^{-1}\left(n\alpha, \sum_{i=1}^n x_i\right),$$

so $\hat{\theta}_{Boot,n} \overset{d}{=} 1/\hat{\theta}_n$. Since $\hat{\pi}_n = \theta_0/w_0$, we can also conclude that the parametric bootstrap is not uniformly distributed:

$$\Pr\left(\hat{\theta}_{Boot,n} \leq \theta_0 | \hat{\pi}_n\right) = \Pr\left(\frac{\theta_0}{w_0 w} \leq \hat{\pi}_n | \theta_0, w_0\right) = \Pr\left(\frac{1}{w} \leq w_0\right).$$

The posterior distribution we obtained for the SwiZs in the last example coincides with the fiducial distribution [see Table 1 VM15], [see Example 21.2 KS61]. This correspondance is not surprising in view of the discussion held after Proposition 1.24. Indeed the gamma distribution is a member of the exponential family and we use a sufficient statistics as the auxiliary estimator, so the SwiZs and the generalized fiducial distribution are equivalent.

We now turn our attention to more general examples where $\boldsymbol{\theta}$ is not a scalar.

Example 1.55 (normal with unknown mean and unknown variance)**.** *Let $x_i \sim \mathcal{N}(\mu, \sigma^2)$ be identically and independently distributed and consider $g(\mu, \sigma^2, u) = \mu + \sigma u$ where $u \sim \mathcal{N}(0,1)$. Take the following auxiliary estimator, $\hat{\boldsymbol{\pi}}_n = (\bar{x}, ks^2)^T = \mathbf{h}(x)$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ and $k \in \mathbb{R}$ is any constant. Note for example that $k < 0$, so the auxiliary estimator of the variance may be negative. Indeed the SwiZs accepts situation for which $\boldsymbol{\Pi} \cap \boldsymbol{\Theta} = \emptyset$, it is clearly not the case of the parametric bootstrap for example (see Remark 1.10). We have that*

$$\mathbf{w} = \mathbf{h}(u) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n u_i \\ \sum_{i=1}^n \left(u_i - \frac{1}{n}\sum_{j=1}^n u_j\right)^2 \end{pmatrix}.$$

An explicit solution exists for $d(\hat{\boldsymbol{\pi}}_n, g(\mu, \sigma^2, \mathbf{w})) = 0$ in (μ, σ^2) and is given by

$$\hat{\boldsymbol{\theta}}_n = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \bar{x}_0 - \hat{\sigma} w_1 \\ \frac{s_0^2}{w_2} \end{pmatrix} = \mathbf{a}(\mathbf{w}).$$

Note that $\bar{x}_0 = \mu_0 + \sigma_0 w_{0,1}$ and $s_0^2 = \sigma_0^2 w_{0,2}$. We obtain the following

$$\begin{aligned}\Pr\left(\hat{\boldsymbol{\theta}}_n \leq \boldsymbol{\theta}_0\right) &= \Pr\left(\begin{pmatrix} \mu_0 + \sigma_0 w_{0,1} - \sigma_0 w_1 \sqrt{\frac{w_{0,2}}{w_2}} \\ \sigma_0^2 \frac{w_{0,2}}{w_2} \end{pmatrix} \leq \begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix}\right) \\ &= \Pr\left(\begin{pmatrix} \frac{w_1}{\sqrt{w_2}} \\ \frac{1}{w_2} \end{pmatrix} \leq \begin{pmatrix} \frac{w_{0,1}}{\sqrt{w_{0,2}}} \\ \frac{1}{w_{0,2}} \end{pmatrix}\right) \sim \mathcal{U}(0,1).\end{aligned}$$

Therefore, by Proposition 1.28, any region built from the percentiles of the posterior distribution of $\hat{\boldsymbol{\theta}}_n$ has exact frequentist coverage. This posterior distribution has a closed form.

Note that $w_1 \sim \mathcal{N}(0, 1/n)$. Once realized that $u_i - \frac{1}{n}\sum_{j=1}^n u_j \sim \mathcal{N}(0, (n-1)/n)$, it is not difficult to obtain that $w_2 \sim \Gamma(n/2, n/2(n-1))$, a gamma random variable (shape-rate parametrization). It is straightforward to remark that

$$\hat{\mu}|(\hat{\sigma}^2, \hat{\boldsymbol{\pi}}_n) \sim \mathcal{N}\left(\bar{x}_0, \frac{\hat{\sigma}^2}{n}\right), \quad \hat{\sigma}^2 \sim \Gamma^{-1}\left(\frac{n}{2}, \frac{s_0^2 n}{2(n-1)}\right),$$

where Γ^{-1} represents the inverse gamma distribution. The joint distribution is known in the Bayesian literature as the normal-inverse-gamma distribution (see [Koc07]). We thus have the following joint distribution

$$\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n \sim \mathcal{N}\text{-}\Gamma^{-1}\left(\bar{x}_0, n, \frac{n}{2}, \frac{s_0^2 n}{2(n-1)}\right).$$

The distribution of $\hat{\mu}$ unconditionnaly on $\hat{\sigma}^2$ is a non-standardized t-distribution with n degrees of freedom,

$$\hat{\mu}|\hat{\boldsymbol{\pi}}_n \sim t\left(\bar{x}_0, \frac{s_0^2 n}{n-1}, n\right).$$

The results on the normal distribution (Example 1.55) can be generalized to the linear regression.

Example 1.56 (linear regression)**.** *Consider the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and $\dim(\boldsymbol{\beta}) = p$. Suppose the matrix $\mathbf{X}^T\mathbf{X}$ is of full rank. A natural generating function is $\mathbf{g}(\boldsymbol{\beta}, \sigma^2, \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u}$ where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ (see Example 1.1 for other suggestions). Take the ordinary least squares as the auxiliary estimator so we have the following explicit form:*

$$\hat{\boldsymbol{\pi}}_n = \begin{pmatrix} \hat{\boldsymbol{\pi}}_1 \\ \hat{\pi}_2 \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}_0 \\ k\mathbf{y}_0^T\mathbf{P}\mathbf{y}_0 \end{pmatrix},$$

where $\mathbf{P} = \mathbf{I}_n - \mathbf{H}$ is the projection matrix, $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is the hat matrix, $\mathbf{y}_0$ denotes the observed responses and $k \in \mathbb{R}$ is any constant. Note that $\mathbf{P}$ and $\mathbf{H}$ are symmetric idempotent matrices and that $\mathbf{P}\mathbf{X} = \mathbf{0}$. An explicit solution exists for $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}^T\ \hat{\sigma}^2)^T$. To find it, we use the indirect inference estimator, which by Theorem 1.8 is the equivalent to the SwiZs estimator. Using $\mathbf{y} \overset{d}{=} \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u}$, we have

$$\hat{\boldsymbol{\pi}}_{II,n}(\boldsymbol{\theta}) = \begin{pmatrix} \hat{\boldsymbol{\pi}}_1(\boldsymbol{\theta}) \\ \hat{\pi}_2(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u}\right) \\ k\sigma^2\mathbf{u}^T\mathbf{P}\mathbf{u} \end{pmatrix}.$$

Since $\hat{\pi}_2(\boldsymbol{\theta})$ depends only on σ^2, solving $d(\hat{\pi}_2, \hat{\pi}_2(\boldsymbol{\theta})) = 0$ in σ^2 leads to

$$\hat{\sigma}^2 = \frac{\mathbf{y}_0^T\mathbf{P}\mathbf{y}_0}{\mathbf{u}^T\mathbf{P}\mathbf{u}}.$$

On the other hand, solving $d(\hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_1(\boldsymbol{\theta})) = \mathbf{0}$ in $\boldsymbol{\beta}$ leads to

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y}_0 + \hat{\sigma}\mathbf{u}\right).$$

Since $\mathbf{y}_0 = \mathbf{X}\boldsymbol{\beta}_0 + \sigma_0\mathbf{u}_0$, *we obtain the following:*

$$\begin{aligned}\Pr\left(\hat{\boldsymbol{\theta}}_n \leq \boldsymbol{\theta}_0\right) &= \Pr\left(\hat{\boldsymbol{\beta}} \leq \boldsymbol{\beta}_0, \hat{\sigma}^2 \leq \sigma_0^2\right) \\ &= \Pr\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta}_0 + \sigma_0\mathbf{u}_0 + \hat{\sigma}\mathbf{u}\right) \leq \boldsymbol{\beta}_0, \frac{\left(\mathbf{X}\boldsymbol{\beta}_0 + \sigma_0\mathbf{u}_0\right)^T\mathbf{P}\left(\mathbf{X}\boldsymbol{\beta}_0 + \sigma_0\mathbf{u}_0\right)}{\mathbf{u}^T\mathbf{P}\mathbf{u}} \leq \sigma_0^2\right) \\ &= \Pr\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\sigma_0\mathbf{u}_0 - \hat{\sigma}\mathbf{u}\right) \leq \mathbf{0}, \frac{\sigma_0^2\mathbf{u}_0^T\mathbf{P}\mathbf{u}_0}{\mathbf{u}^T\mathbf{P}\mathbf{u}} \leq \sigma_0^2\right) \\ &= \Pr\left(\frac{\mathbf{X}^T\mathbf{u}}{\sqrt{\mathbf{u}^T\mathbf{P}\mathbf{u}}} \leq \frac{\mathbf{X}^T\mathbf{u}_0}{\sqrt{\mathbf{u}_0^T\mathbf{P}\mathbf{u}_0}}, \frac{1}{\mathbf{u}^T\mathbf{P}\mathbf{u}} \leq \frac{1}{\mathbf{u}_0^T\mathbf{P}\mathbf{u}_0}\right) \sim \mathcal{U}(0,1).\end{aligned}$$

Note that at the third equality we use the fact that $\mathbf{u} \overset{d}{=} -\mathbf{u}$ *since* $\mathbf{u}$ *is symmetric around* $\mathbf{0}$. *The last development, together with Proposition 1.28, demonstrates that any region built on the percentiles of the distribution of* $\hat{\boldsymbol{\theta}}_n$ *leads to exact frequentist coverage probabilities. The distribution of* $\hat{\boldsymbol{\theta}}_n$ *can be obtained in an explicit form.*

Since $\mathbf{P}$ *is symmetric and idempotent, it is well known that* $\mathbf{u}^T\mathbf{P}\mathbf{u} \sim \chi^2_{n-p}$ *[see Theorem 5.1.1 MP92]. Hence we obtain that*

$$\hat{\boldsymbol{\beta}}|(\hat{\sigma}^2, \hat{\boldsymbol{\pi}}_n) \sim \mathcal{N}\left(\hat{\boldsymbol{\pi}}_1, \hat{\sigma}^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right), \quad \hat{\sigma}^2|\hat{\boldsymbol{\pi}}_n \sim \Gamma^{-1}\left(\frac{n-p}{2}, \frac{\mathbf{y}_0^T\mathbf{P}\mathbf{y}_0}{2}\right).$$

As shown in Example 1.55, it follows that the joint distribution of $\hat{\boldsymbol{\theta}}_n$ *conditionally on* $\hat{\boldsymbol{\pi}}_n$ *is a normal-inverse-gamma distribution*

$$\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n \sim \mathcal{N}\text{-}\Gamma^{-1}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}_0, \left(\mathbf{X}^T\mathbf{X}\right)^{-1}, \frac{n-p}{2}, \frac{\mathbf{y}_0^T\mathbf{P}\mathbf{y}_0}{2}\right),$$

and the distribution of $\hat{\boldsymbol{\beta}}$, *unconditionally on* $\hat{\sigma}^2$, *is a multivariate non-standardized* t *distribution with* $n-p$ *degrees of freedom*

$$\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\pi}}_n \sim t\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}_0, \frac{\mathbf{y}_0^T\mathbf{P}\mathbf{y}_0}{n-p}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}, n-p\right).$$

In this last example on the linear regression, we employed the OLS as the auxiliary estimator, which is known to be an unbiased estimator. In fact, it is not a necessity to have unbiased auxiliary estimator. The next example illustrate this point.

Example 1.57 (ridge regression)**.** *Consider the same setup as in Example 1.56,* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ *and* $\operatorname{rank}(\mathbf{X}^T\mathbf{X}) = p$. *Take the ridge estimator as the auxiliary estimator, so for the regression coefficients we have*

$$\hat{\boldsymbol{\pi}}_1^R = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y}_0,$$

for some constant $\lambda \in \mathrm{I\!R}$. *Consider the squared residuals as an estimator of the variance, so after few manipulations, we obtain*

$$\hat{\pi}_2^R = k\mathbf{y}_0^T\mathbf{P}_\lambda\mathbf{P}_\lambda\mathbf{y}_0,$$

where $\mathbf{P}_\lambda \equiv \mathbf{I}_n - \mathbf{H}_\lambda$, $\mathbf{H}_\lambda \equiv \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T$, $k \in \mathrm{I\!R}$ *is any constant. Note that* $\mathbf{P}_\lambda$ *is symmetric but not idempotent. As in Example 1.56, let's use the indirect inference estimator with* $\mathbf{y} \overset{d}{=} \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u}$. *We obtain*

$$\hat{\boldsymbol{\pi}}_{II,n}^R(\boldsymbol{\theta}) = \begin{pmatrix}\hat{\boldsymbol{\pi}}_1^R(\boldsymbol{\theta}) \\ \hat{\pi}_2^R(\boldsymbol{\theta})\end{pmatrix} = \begin{pmatrix}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u}\right) \\ k\left(\mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u}\right)^T\mathbf{P}_\lambda\mathbf{P}_\lambda\left(\mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u}\right)\end{pmatrix}.$$

Let $\tilde{\boldsymbol{\beta}}$ denotes the solution of $d(\hat{\boldsymbol{\pi}}_1^R, \hat{\boldsymbol{\pi}}_1^R(\boldsymbol{\theta})) = 0$ in $\boldsymbol{\beta}$. We have the explicit solution given by

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y}_0 - \tilde{\sigma}\mathbf{u}\right).$$

Using $\tilde{\boldsymbol{\beta}}$ in $\hat{\pi}_2^R(\boldsymbol{\theta})$ leads to

$$\hat{\pi}_2^R(\tilde{\boldsymbol{\theta}}) = k(\mathbf{H}\mathbf{y}_0 - \tilde{\sigma}\mathbf{P}\mathbf{u})^T\mathbf{P}_\lambda\mathbf{P}_\lambda\left(\mathbf{H}\mathbf{y}_0 - \tilde{\sigma}\mathbf{P}\mathbf{u}\right),$$

where $\mathbf{H} \equiv \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}$ and $\mathbf{P} \equiv \mathbf{I}_n - \mathbf{H}$. We have the followings: $\mathbf{H}\mathbf{H}_\lambda = \mathbf{H}_\lambda$, $\mathbf{P}\mathbf{P}_\lambda = \mathbf{P}$ and $\mathbf{P}\mathbf{H} = \mathbf{0}$. Finding $\tilde{\sigma}^2$ such that $d(\hat{\pi}_2^R, \hat{\pi}_2^R(\tilde{\boldsymbol{\theta}})) = 0$ gives

$$\tilde{\sigma}^2\mathbf{u}^T\mathbf{P}\mathbf{u} + \mathbf{y}_0^T\mathbf{H}\mathbf{P}_\lambda\mathbf{P}_\lambda\mathbf{H}\mathbf{y}_0 - \mathbf{y}_0^T\mathbf{P}_\lambda\mathbf{P}_\lambda\mathbf{y}_0 = 0,$$

which leads to the following solution:

$$\tilde{\sigma}^2 = \frac{\mathbf{y}_0^T\mathbf{P}\mathbf{y}_0}{\mathbf{u}^T\mathbf{P}\mathbf{u}}.$$

Therefore, $\tilde{\sigma}^2$ is the same as $\hat{\sigma}^2$ we found in Example 1.56, and we directly have that $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. As a consequence, the distribution of $\tilde{\boldsymbol{\theta}}$ is exactly the same as $\hat{\boldsymbol{\theta}}_n$ in Example 1.56 and the frequentist coverage probabilities are exact.

From Example 1.55 on the normal distribution, the derivation to closely related distribution is straightforward, as we see now with the log-normal distribution.

Example 1.58 (log-normal with unknown mean and unknown variance)**.** *Let $x_i \sim \log\text{-}\mathcal{N}(\mu, \sigma^2)$ be identically and independently distributed and consider $g(\mu, \sigma^2, u) = e^\mu e^{\sigma u}$ where $u \sim \mathcal{N}(0, 1)$. If we take the maximum likelihood estimator as the auxiliary estimator, we have*

$$\hat{\boldsymbol{\pi}}_n = \begin{pmatrix}\hat{\pi}_1\\ \hat{\pi}_2\end{pmatrix} = \begin{pmatrix}\frac{1}{n}\sum_{i=1}^n \ln(x_i)\\ \sum_{i=1}^n\left(\ln(x_i) - \frac{1}{n}\sum_{j=1}^n \ln(x_j)\right)^2\end{pmatrix}$$

The solution is the following

$$\hat{\boldsymbol{\theta}}_n = \begin{pmatrix}\hat{\mu}\\ \hat{\sigma}^2\end{pmatrix} = \begin{pmatrix}\hat{\pi}_1 - \hat{\sigma}w_1\\ \frac{\hat{\pi}_2}{w_2},\end{pmatrix}$$

where $w_1 = \frac{1}{n}\sum_{i=1}^n u_i$ and $w_2 = \sum_{i=1}^n\left(u_i - \frac{1}{n}\sum_{j=1}^n u_j\right)^2$. It is the same solution as Example 1.55, hence the posterior distribution of $\hat{\boldsymbol{\theta}}_n$ is normal-inverse-gamma and any α-credible region built on this posterior have exact frequentist coverage.

Having illustrated the theory for random variable that are identically and independently distributed, we now show a last example on time series data. Note that (variations of) this example is numerically studied in [GMR93].

Example 1.59 (irregularly observed geometric Brownian motion with unknown drift and unknown volatility)**.** *Consider the stochastic differential equation*

$$dy_t = \mu y_t dt + \sigma y_t dW_t,$$

where $\{W_t : t \geq 0\}$ is a Wiener process and $\boldsymbol{\theta} = (\mu\ \sigma^2)^T$ are the drift and volatility parameters. An explicit solution to Itô's integral exists and is given by

$$y_t = y_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right].$$

Suppose we observe the process at n *points in time:* $t_1 < t_2 < \ldots < t_n$, $\forall i \; t_i \in \mathbb{R}^+$. *Define the difference in time by* $\Delta_i = t_i - t_{i-1}$, *so we have* $n-1$ *time differences. Note that all the time differences are positive,* $\Delta_i > 0$, *and we allow the process to be irregularly observed,* $\Delta_i \neq \Delta_j, i \neq j$. *Instead of working directly with the process* $\{y_{t_i} : i \geq 1\}$, *it is more convenient to work with the following transformation of the process* $\{x_{t_i} = \ln(y_{t_i}/y_{t_{i-1}}) : i \geq 2\}$. *Indeed, we have*

$$x_{t_i} = \left(\mu - \frac{1}{2}\sigma^2\right)\Delta_i + \sigma\left(W_{t_i} - W_{t_{i-1}}\right).$$

By the properties of the Wiener process, we have $W_{t_i} - W_{t_{i-1}} \sim \mathcal{N}(0, \Delta_i)$ *and* $W_{t_i} - W_{t_{i-1}}$ *is independent from* $W_{t_j} - W_{t_{j-1}}$ *for* $i \neq j$. *Hence the vector* $\mathbf{x} = (x_{t_2} \; \ldots \; x_{t_n})^T$ *is independentely but non-identically distributed according to the joint normal distribution*

$$\mathbf{x} \sim \mathcal{N}\left(\left(\mu - \frac{1}{2}\sigma^2\right)\boldsymbol{\Delta}, \sigma^2\Sigma\right),$$

where $\boldsymbol{\Delta} = (\Delta_2 \; \ldots \; \Delta_n)^T$ *and* $\Sigma = \operatorname{diag}(\boldsymbol{\Delta})$. *Note that* $\boldsymbol{\Delta} = \Sigma\mathbf{1}_{n-1}$, *where* $\mathbf{1}_{n-1}$ *is a vector of* $n-1$ *ones, and* $\boldsymbol{\Delta}^T\mathbf{1}_{n-1} = \boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2}$ *since all the* Δ *are positives.*

We consider the following auxiliary estimators:

$$\hat{\boldsymbol{\pi}}_n = \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_0^T\mathbf{1}_{n-1} \\ \mathbf{x}_0^T\Sigma^{-1}\mathbf{x}_0 \end{pmatrix}.$$

Since $\mathbf{x} \overset{d}{=} (\mu - \sigma^2/2)\boldsymbol{\Delta} + \sigma\Sigma^{1/2}\mathbf{z}$, *where* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n-1})$, *we obtain the following indirect inference estimators (or equivalently SwiZs),*

$$\hat{\pi}_1(\boldsymbol{\theta}) = \left[\left(\mu - \frac{1}{2}\sigma^2\right)\boldsymbol{\Delta} + \sigma\Sigma^{1/2}\mathbf{z}\right]^T\mathbf{1}_{n-1} = \left(\mu - \frac{1}{2}\sigma^2\right)\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2} + \sigma\mathbf{z}^T\boldsymbol{\Delta}^{1/2},$$

and

$$\begin{aligned}\hat{\pi}_2(\boldsymbol{\theta}) &= \left[\left(\mu - \frac{1}{2}\sigma^2\right)\boldsymbol{\Delta} + \sigma\Sigma^{1/2}\mathbf{z}\right]^T\Sigma^{-1}\left[\left(\mu - \frac{1}{2}\sigma^2\right)\boldsymbol{\Delta} + \sigma\Sigma^{1/2}\mathbf{z}\right] \\ &= \left(\mu - \frac{1}{2}\sigma^2\right)^2\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2} + 2\sigma\left(\mu - \frac{1}{2}\sigma^2\right)\mathbf{z}^T\boldsymbol{\Delta}^{1/2} + \sigma^2\mathbf{z}^T\mathbf{z}.\end{aligned}$$

Solving $d(\hat{\pi}_1, \hat{\pi}_1(\hat{\boldsymbol{\theta}})) = 0$ *in* $\hat{\mu}$ *gives*

$$\hat{\mu} = \frac{1}{2}\hat{\sigma}^2 - \hat{\sigma}\mathbf{z}^T\boldsymbol{\Delta}^{1/2}\left(\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2}\right)^{-1} + \mathbf{x}_0^T\mathbf{1}_{n-1}\left(\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2}\right)^{-1}. \tag{1.3}$$

Now solving $d(\hat{\pi}_2, \hat{\pi}_2(\hat{\boldsymbol{\theta}})) = 0$ *in* $\hat{\sigma}^2$ *and substituing* $\hat{\mu}$ *by the above expression in* (1.3) *leads to*

$$\hat{\sigma}^2 = \frac{\mathbf{x}_0^T\mathbf{Q}\mathbf{x}_0}{\mathbf{z}^T\mathbf{P}\mathbf{z}},$$

where $\mathbf{P} = \mathbf{I}_{n-1} - \boldsymbol{\Delta}^{1/2}\left(\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2}\right)^{-1}\boldsymbol{\Delta}^{T/2}$ *is symmetric and idempotent, and* $\mathbf{Q} = \Sigma^{-1} - \mathbf{1}_{n-1}\left(\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2}\right)^{-1}\mathbf{1}_{n-1}^T$. *By the properties of the rank of a matrix, we have* $\operatorname{rank}(\mathbf{P}) = \operatorname{trace}(\mathbf{P}) = n-2$. *Note that by independence* $\mathbf{z}^T\Delta^{1/2} \overset{d}{=} z(\Delta^{T/2}\Delta^{1/2})$, *where* z *is a* *single standard normal random variable. Similarly to the example on the linear regression (Example 1.56), we obtain the explicit distributions*

$$\hat{\mu}|\left(\hat{\boldsymbol{\pi}}_n,\hat{\sigma}^2\right)\sim\mathcal{N}\left(\frac{1}{2}\hat{\sigma}^2+\mathbf{x}_0^T\mathbf{1}_{n-1}\left(\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2}\right)^{-1},\hat{\sigma}^2\left(\boldsymbol{\Delta}^{T/2}\boldsymbol{\Delta}^{1/2}\right)^{-1}\right),$$
$$\hat{\sigma}^2|\hat{\boldsymbol{\pi}}_n\sim\Gamma^{-1}\left(\frac{n-2}{2},\frac{\mathbf{x}_0^T\mathbf{Q}\mathbf{x}_0}{2}\right).$$

As with Example 1.56, this findings suggest that $\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n$ is jointly distributed according to a normal-inverse-gamma distribution. However, $\hat{\sigma}^2$ appears in the mean of $\hat{\mu}|(\hat{\boldsymbol{\pi}}_n,\hat{\sigma}^2)$ so such conclusion is not straightforward. We leave the derivation of the joint distribution and the distribution of $\hat{\mu}$ unconditionnal on $\hat{\sigma}^2$ for further research.

We now demonstrate that the $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution of $\hat{\boldsymbol{\theta}}_n$ leads to exact frequentist coverage probabilities. Once realized that $\Sigma^{-1}=\Sigma^{-1/2}\Sigma^{-1/2}$, $\Sigma^{1/2}\mathbf{1}_{n-1}=\boldsymbol{\Delta}^{1/2}$, and $\boldsymbol{\Delta}^T\Sigma^{-1}=\mathbf{1}_{n-1}$, it is not difficult to show that $\boldsymbol{\Delta}^T\mathbf{Q}\boldsymbol{\Delta}=0$, $\boldsymbol{\Delta}^T\mathbf{Q}\Sigma^{1/2}=0$ and $\Sigma^{1/2}\mathbf{Q}\Sigma^{1/2}=\mathbf{P}$. Since $\mathbf{x}_0=(\mu_0-\sigma_0^2/2)\boldsymbol{\Delta}+\sigma_0\Sigma^{1/2}\mathbf{z}_0$, we obtain

$$\hat{\sigma}^2=\sigma_0^2\frac{\mathbf{z}_0^T\mathbf{P}\mathbf{z}_0}{\mathbf{z}^T\mathbf{P}\mathbf{z}}=\sigma_0^2\frac{w_0}{w},$$
$$\hat{\mu}=\frac{\sigma_0^2}{2}\frac{w_0}{w}-\sigma_0\sqrt{\frac{w_0}{w}}z+\mu_0-\frac{1}{2}\sigma_0^2+\sigma_0z_0.$$

Therefore,

$$\Pr\left(\hat{\mu}\leq\mu_0,\ \hat{\sigma}^2\leq\sigma_0^2\right)=\Pr\left(\frac{\sigma_0^2}{2}\frac{w_0}{w}-\sigma_0\sqrt{\frac{w_0}{w}}z+-\frac{1}{2}\sigma_0^2+\sigma_0z_0\leq0,\ \frac{w_0}{w}\leq1\right)$$
$$=\Pr\left(\frac{k_0}{w}-\frac{z}{\sqrt{w}}\leq\frac{k_0}{w_0}-\frac{z_0}{\sqrt{w_0}},\ w^{-1}\leq w_0^{-1}\right)\sim\mathcal{U}(0,1),$$

where $k_0=\sigma_0\sqrt{w_0}/2$. Thus, any region on the joint distribution of $\hat{\boldsymbol{\theta}}_n$ leads to exact frequentist coverage by Proposition 1.28.

1.7 Simulation study

The main goal of this section is threefold. First, we illustrate the results of the Section 1.4 on the frequentist properties in finite sample of the SwiZs in the general case where no solutions are known in explicit forms, as opposed to the Section 1.6, and thus requiring numerical solutions. In order to achieve this point, we measure at different levels the empirical coverage probabilities of the intervals built from the percentiles of the $\hat{\boldsymbol{\pi}}_n$-approximate posterior obtained by the SwiZs. Note that for $\dim(\boldsymbol{\theta})>1$, we only considered marginal intervals to avoid a supplementary layer of numerical nuisance, the coverage probabilities are not concerned by this choice, only the length of the intervals. Second, we elaborate on the verification of the conditions of Theorem 1.38 with the examples at hand. As already motivated, the emphasis is on the estimating function. It seems easier to verify Assumption 1.36 than Assumption 1.37, since only one of them is necessary to satisfy Theorem 1.38, we concentrate our efforts on the former. We also brighten the study up to situations where Assumption 1.36 does not entirely hold or cannot be verified to measure its consequences empirically. Third, we give the general idea on how to implement the SwiZs. Indeed, anyone familiar with the numerical problem of solving a point estimator such as the maximum likelihood estimator has a very good idea on how to obtain the auxiliary estimator $\hat{\boldsymbol{\pi}}_n$. Solving the estimating function for the parameters of interest is very similar, it requires the exact same tools but has the inconvenient of needing further analytical derivations and implementations details. As already remarked, the parametric bootstrap does not possess such inconvenient. The counterpart is that the SwiZs may lead to exact coverage probabilities. The motto "no pain, no gain" is particularly relevant here. For this purpose, the parametric bootstrap is proposed as the point of comparison for all the examples of this section. We measure the computational time as experienced by the user in order to appreciate the numerical burden. In case both the SwiZs and the parametric bootstrap have very similar coverage probabilities, we also quantify the length of the intervals as a mean of comparison.

As a subsidiary goal of this section, we study the point estimates of the SwiZs. Indeed, the indirect inference is also a method for reducing the small sample bias of an initial (auxiliary) estimator, even in situations where it may be "unnatural" to call such method, as for example, when a maximum likelihood estimator may be easily obtained (see [Gue+18b]). Since the SwiZs is a special case of indirect inference, it would be interesting to gauge the ability of the SwiZs to correct the bias. We explore the properties of the mean and the median of the SwiZs. This choice is arbitrary but largely admitted.

There are common factors in the implementation of all the examples of this section so we start by mentioning them by category. For the design, we use $M = 10,000$ independent trials so we can appreciate the coverage probabilities up to the fourth digit. We evaluate numerically the $\hat{\boldsymbol{\pi}}_n$-approximate posterior distribution of the SwiZs and the parametric bootstrap distribution based on $S = 10,000$ replicates. We measure the coverage probabilities at $50\%, 75\%, 90\%, 95\%$ and 99% levels. Although sometimes we do not report all of them for more clarity of the presentation, they are however shown in Appendix for more transparency. For the implementation, we write our code in C++ with the help of the Armadillo ([SC16]), the Eigen ([GJ+10]) and the Boost ([Boo18]) libraries. Since the statistical community uses mainly R in academia ([R C17]), we were able to use the C++ implementation within R thanks to the Rcpp ([EB17]), the RcppArmadillo ([ES14]), the RcppEigen ([BE13]) and the BH ([EEK16]) packages. For the numerical optimization, instead of solving directly the root of the estimating function, we try to find the minimum of the squared ℓ_2-norm of the estimating function. This is a more practical solution and permits to use quasi-Newton routines (see the Chapter 11 of [NW06] for a broader discussion). Note that taking the norm of the estimating function is a trick also used when proving the consistency of $\hat{\boldsymbol{\pi}}_n$ and $\hat{\boldsymbol{\theta}}_n$. All the optimizations are conducted by the Limited memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton routine proposed by [Noc80; LN89], more specifically, we employ the C implementation made available by [ON10] and accessible for R user through the RcppNumerical ([Qiu+18]) package. The optimization are performed with the default values. The starting values for solving the auxiliary estimator are obtained using the differential evolution algorithm of [SP97] and made available to R user by the RcppDE ([RS16]) package. For finding $\hat{\boldsymbol{\theta}}_n$, we simply use $\hat{\boldsymbol{\pi}}_n$ as the starting values. There are constraints on the parameters of most models we considered in this section. We use variable transformation when necessary for the parameters to comply to these constraints. All the numerical evaluation are performed at the Baobab cluster of the University of Geneva on 16 parallelized threads. The variation in time due to the different type of nodes treating our demand in the cluster has not been taken into account in the reported time because, first, it appeared to be minor, and second, the SwiZs and the comparative methods are performed simultaneously so they would be equivalently concerned.

We select five different scenarii. First, we start with a toy example by considering a standard Student's t-distribution with unknown degrees of freedom (Example 1.60). Although the Student distribution is ubiquitous in statistics since at least Gosset's Biometrika paper ([Stu08]), there are no simple tractable way to construct an interval of uncertainty around the degrees of freedom. In addition, the degrees of freedom is a parameter that gauges the tail of the distribution and is not particularly easy to handle. The existence of the moments of this distribution depends upon the values that this parameter takes. We take a particular interest in small values of this parameter for which, for example the variance or the kurtosis are infinite.

Example 1.60 (standard t-distribution with unknown degrees of freedom). *Let $x_i \sim t(\theta)$, $i = 1, \cdots, n$, be identically and independently distributed with density*

$$f(x_i, \theta) = \frac{\left(1 + \frac{x_i^2}{\theta}\right)^{-\frac{\theta+1}{2}}}{\sqrt{\theta}\mathcal{B}\left(\frac{1}{2}, \frac{\theta}{2}\right)}, \tag{1.4}$$

where θ represents the degrees of freedom and $\mathcal{B}$ is the beta function. We consider the likelihood score function as the estimating function and we take the MLE as the auxiliary estimator. In this situation, Θ and Π are equivalent, and thus, there are no reasons to disqualify the parametric bootstrap. Substituing θ by π in the Equation 1.4, taking then the derivative with respect to π of the log-density leads to the following

$$\Phi_n(\theta, \mathbf{u}, \pi) = \psi\left(\frac{\pi+1}{2}\right) - \psi\left(\frac{\pi}{2}\right) - \frac{1}{n}\sum_{i=1}^n \ln\left(\frac{g(\theta, \mathbf{u}_i)^2 + 1}{\pi}\right) + \frac{1}{n}\sum_{i=1}^n \frac{g(\theta, \mathbf{u}_i)^2 - 1}{g(\theta, \mathbf{u}_i)^2 + \pi},$$

where ψ is the digamma function. We now verify Assumption 1.36 so Theorem 1.38 can be invoked. Suppose Assumption 1.32 holds so we can write the following scalar-valued function

$$\varphi_{\hat{\pi}_n}(\theta, w) = \frac{1}{2}\psi\left(\frac{\hat{\pi}_n+1}{2}\right) - \frac{1}{2}\psi\left(\frac{\hat{\pi}_n}{2}\right) - \frac{1}{2}\ln\left(\frac{g(\theta, w)^2 + 1}{\hat{\pi}_n}\right) + \frac{1}{2}\frac{g(\theta, w)^2 - 1}{g(\theta, w)^2 + \hat{\pi}_n},$$

where $\hat{\pi}_n$ is fixed. The first derivative with respect to θ is given by

$$\frac{\partial}{\partial\theta}\varphi_{\hat{\pi}_n}(\theta, w) = g(\theta, w)\frac{\partial}{\partial\theta}g(\theta, w)\left[\frac{\hat{\pi}_n - 1}{\left(g(\theta, w)^2 + \hat{\pi}_n\right)^2} - \frac{1}{g(\theta, w)^2 + 1}\right]. \tag{1.5}$$

Substituing $(\partial/\partial\theta)g$ by $(\partial/\partial w)g$ gives the first derivative with respect to w. The derivative exists everywhere so $K_n = \emptyset$. Therefore, if the generating function $g(\theta, w)$ is once continuously differentiable in both its arguments then Assumption 1.36 (i) is satisfied.

The determinant here is $|\frac{\partial}{\partial\theta}\varphi_{\hat{\pi}_n}(\theta, w)|$. It will be zero on a countable set of points: if $g(\theta, w) = 0$, if $(\partial/\partial\theta)g(\theta, w) = 0$ or if the rightest term of the Equation 1.5 is 0. Substituing $(\partial/\partial\theta)g$ by $(\partial/\partial w)g$ gives the same analysis. Hence, the determinant of the derivatives of the estimating function is almost everywhere non-null and Assumption 1.36 (ii) is satisfied.

Eventually, we clearly have that

$$\lim_{|g|\to\infty} |\varphi_{\hat{\pi}_n}(\theta, w)| = +\infty.$$

As a consequence, given that $\lim_{\|(\theta,w)\|\to\infty}|g(\theta,w)|=\infty$*, Assumption 1.36 (iii) is satisfied.*

In the light of these findings, the choice of generating function is crucial and there are many candidates [see e.g. Dev86]. The inverse cumulative distribution function is a natural choice, but a numerically complicated one in this case. Indeed, following the Boost C++ library [Boo18], it can be obtained by

$$g_1(\theta,u_1)=\operatorname{sign}\left(u_1-\frac{1}{2}\right)\left(\frac{\theta(1-z)}{z}\right)^{1/2},$$

where $u_1\sim\mathcal{U}(0,1)$ *and* z *is equal to the incomplete beta function inverse parametrized by* θ *and depending on* u_1*. An alternative choice, numerically and analytically simpler, is to consider Bailey's polar algorithm [Bai94], which is given by*

$$g_2(\theta,\mathbf{u}_2)=u_{2,1}\sqrt{\frac{\theta}{u_{2,2}}\left(u_{2,2}^{-2/\theta}-1\right)},$$

where $u_{2,2}\overset{d}{=}u_{2,1}^2+u_{2,3}^2$ *if* $u_{2,2}\leq 1$ *and* $u_{2,1},u_{2,3}\sim\mathcal{U}(-1,1)$*. Clearly* $g_2(\theta,\mathbf{u}_2)$ *is once continuously differentiable in each of its arguments and the limit is* $\lim_{(\theta,u_{2,1},u_{2,2})\to(\infty,1,1)}|g_2(\theta,u_{2,1},u_{2,3})|=\infty$*. Hence, even if* w *is unknown, these results strongly suggests that the conditions of Theorem 1.38 hold, and as a conclusion, any intervals built on the percentiles of the distribution of* $\hat{\theta}_n$ *given* $\hat{\pi}_n$ *have exact frequentist coverage.*

The coverage probabilities in the Table 1.1 below are computed for three different values of $\theta_0=\{1.5,3.5,6\}$ *and a sample size of* $n=50$*. When* $\theta_0=1.5$*, the variance of a Student's random variable is infinite and the skewness and kurtosis of the distribution are undefined. When* $\theta_0=3.5$*, the variance is finite and the kurtosis is infinite. When* $\theta_0=6$*, the first five moment exists.*

The SwiZs is accurate at all the confidence levels with a maximum discrepancy of 1.39% in absolute value. This is very reasonable considering the numerical task we perform. In comparison, the parametric bootstrap has a minimum discrepancy of 0.87% *for an average of* 4.44%*. The SwiZs is also more efficient, it dominates the parametric bootstrap with a median interval length systematically smaller. The parametric bootstrap is however about six times faster than the SwiZs to compute the intervals. The comparison is not totally fair in disfavor of the SwiZs as we were able here to use directly the log-likelihood for the parametric bootstrap, which is numerically simpler to evaluate than the estimating functions. We also bear the comparison with the bias-corrected and accelerated (BCa) resampling bootstrap of [ET94]. Performances of this bootstrap scheme are comparable to the parametric bootstrap. Finally, when considered in absolute value, 0.2 second do not seem to be a hard effort for obtaining interval which is nearly exact and shorter.*

Second, we consider a more practical case with the two-parameters Lomax distribution ([Lom54]) (Example 1.61), also known as the Pareto II distribution. This distribution has been used to characterise wealth and income distributions as well as business and actuarial losses (see [KK03] and the references therein). Because of this close relationship to the application, we also measure the coverage probabilities of the Gini index, the value-at-risk and the expected shortfall, quantities that may be of interest for the practitioner. The maximum likelihood estimator has been shown in [GFG13] to suffer from small sample bias when n is relatively small and the parameters are close to the boundary of the parameter space. We add their proposal for bias adjustment to the basket of comparative methods. To keep the comparison fair, we use a similar simulation scenario to the ones

		SwiZs			parametric bootstrap			BCa bootstrap		
θ_0	α	$\hat{c}$	$\bar{I}$	$\bar{s}$	$\hat{c}$	$\bar{I}$	$\bar{s}$	$\hat{c}$	$\bar{I}$	$\bar{s}$
1.5	50%	50.66%	0.5129	0.1622	49.13%	0.5794	0.0358	47.69%	0.4906	0.0333
	75%	75.39%	0.8839		73.27%	1.0504		71.64%	0.8607	
	90%	90.15%	1.2861		87.03%	1.6734		86.64%	1.2815	
	95%	94.68%	1.5540		91.42%	2.1935		91.82%	1.5800	
	99%	98.84%	2.1052		96.05%	3.8820		97.13%	2.2714	
3.5	50%	50.08%	1.7594	0.2010	47.65%	2.8832	0.0349	44.94%	1.8716	0.0322
	75%	74.62%	3.2780		70.36%	6.6243		68.80%	3.7372	
	90%	90.39%	5.2129		84.50%	20.665		84.36%	6.5202	
	95%	94.85%	6.8416		89.63%	240.11		90.62%	9.6584	
	99%	98.73%	10.788		95.11%	3104.1		95.60%	29.011	
6	50%	48.61%	4.2027	0.2093	46.54%	11.463	0.0342	44.29%	4.6886	0.0305
	75%	74.39%	8.3688		68.34%	245.75		69.99%	12.245	
	90%	89.56%	16.087		80.83%	2586.4		87.45%	41.335	
	95%	94.61%	26.250		85.06%	3376.8		93.05%	515.51	
	99%	98.90%	361.28		95.55%	4827.0		95.94%	2261.8	

Table 1.1: $\hat{c}$: estimated coverage probabilities, $\bar{I}$: median interval length, $\bar{s}$: average time in seconds to compute the intervals for one trial.

they proposed, which were also motivated by their closeness to situations encountered in practice. Situations where the Lomax distribution is employed has been shown to suffer from influential outliers ever since at least [VFR94b], we therefore consider, in a second time, the weighted maximum likelihood ([FS94]) as the auxiliary estimator to gain robustness. Interestingly, the weighted maximum likelihood estimator is generally not a consistent estimator (see [DM02; MVF06]) so the parametric bootstrap cannot be invoked directly, whereas, on the countrary, the SwiZs may be employed without any particular care.

Example 1.61 (two-parameters Lomax distribution). *Let $x_i \sim Lomax(\boldsymbol{\theta})$, $i = 1, \cdots, n$, $\boldsymbol{\theta} = (b, q)$, be identically and independently distributed with density*

$$f(x_i, \boldsymbol{\theta}) = \frac{q}{b}\left(1 + \frac{x_i}{b}\right)^{-q-1}, \quad x_i > 0, \tag{1.6}$$

where $b, q > 0$ are shape parameters. We consider the likelihood score function as the estimating function and we take the MLE as the auxiliary estimator. The parameter sets $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$ are equivalent with this setup, and thus, the parametric bootstrap may be employed. Substituing $\boldsymbol{\theta}$ by $\boldsymbol{\pi}$ in the Equation 1.6, taking then the derivative with respect to $\boldsymbol{\pi}$ of the log-density leads to the following

$$\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}) = \begin{pmatrix} \frac{1}{\pi_2} - \sum_{i=1}^n \log\left(1 + \frac{g(\boldsymbol{\theta}, \mathbf{u}_i)}{\pi_1}\right) \\ -\frac{1}{\pi_1} + \frac{(\pi_2+1)}{\pi_1} \sum_{i=1}^n \frac{g(\boldsymbol{\theta}, \mathbf{u}_i)}{\pi_1 + g(\boldsymbol{\theta}, \mathbf{u}_i)} \end{pmatrix}.$$

We now verify Assumption 1.36 so Theorem 1.38 can be invoked. Suppose Assumption 1.32 on the existence of a random variable with the same dimensions as $\boldsymbol{\theta}$ holds, *and let denote it by* $\mathbf{w} = (w_1 \; w_2)^T$. *Now assume that we can re-express the estimating function as follows*

$$\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w}) = \begin{pmatrix} \frac{1}{\hat{\pi}_2} - \log\left(1 + \frac{g(\boldsymbol{\theta}, w_1)}{\hat{\pi}_1}\right) \\ \frac{(\hat{\pi}_2 + 1) g(\boldsymbol{\theta}, w_2)}{\hat{\pi}_1^2 + \hat{\pi}_1 g(\boldsymbol{\theta}, w_2)} - \frac{1}{\hat{\pi}_1} \end{pmatrix},$$

where $\hat{\boldsymbol{\pi}}_n$ *is fixed. The Jacobian matrix with respect to* $\boldsymbol{\theta}$ *is given by*

$$D_{\boldsymbol{\theta}} \boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w}) = \begin{pmatrix} \kappa_1(\boldsymbol{\theta}) D_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, w_1) \\ \kappa_2(\boldsymbol{\theta}) D_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, w_2) \end{pmatrix},$$

where

$$\begin{aligned} \kappa_1(\boldsymbol{\theta}) &= \frac{-1}{\hat{\pi}_1 + g(\boldsymbol{\theta}, w_1)} \\ \kappa_2(\boldsymbol{\theta}) &= \frac{\hat{\pi}_1^2 (\hat{\pi}_2 + 1)}{\left(\hat{\pi}_1^2 + \hat{\pi}_1 g(\boldsymbol{\theta}, w_2)\right)^2}. \end{aligned}$$

Note that $\hat{\boldsymbol{\pi}}_n$ *and* $g(\boldsymbol{\theta}, \mathbf{w})$ *are strictly positive, so* $\kappa_1(\boldsymbol{\theta}) < 0$ *and* $\kappa_2(\boldsymbol{\theta}) > 0$. *Substituing* $D_{\boldsymbol{\theta}} g$ *by* $D_{\mathbf{w}} g$ *leads to the Jacobian matrix with respect to* $\mathbf{w}$, *given by*

$$D_{\mathbf{w}} \boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w}) = \begin{pmatrix} \kappa_1(\boldsymbol{\theta}) \frac{\partial}{\partial w_1} g(\boldsymbol{\theta}, w_1) & 0 \\ 0 & \kappa_2(\boldsymbol{\theta}) \frac{\partial}{\partial w_2} g(\boldsymbol{\theta}, w_2) \end{pmatrix}.$$

We see by inspection that the derivatives are defined everywhere and $\mathbf{K}_n = \{\emptyset\}$. *If* $D_{\boldsymbol{\theta}} g$ *and* $D_{\mathbf{w}} g$ *exist and are continuous, then Assumption 1.36 (i) is satisfied.*

The determinants are given by

$$\begin{aligned} \det\left(D_{\boldsymbol{\theta}} \boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w})\right) &= \kappa(\boldsymbol{\theta}, \mathbf{w}) \left[\frac{\partial}{\partial a} g(\boldsymbol{\theta}, w_1) \frac{\partial}{\partial b} g(\boldsymbol{\theta}, w_2) - \frac{\partial}{\partial a} g(\boldsymbol{\theta}, w_2) \frac{\partial}{\partial b} g(\boldsymbol{\theta}, w_1)\right] \\ \det\left(D_{\boldsymbol{\theta}} \boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w})\right) &= \kappa(\boldsymbol{\theta}, \mathbf{w}) \frac{\partial}{\partial w_1} g(\boldsymbol{\theta}, w_1) \frac{\partial}{\partial w_2} g(\boldsymbol{\theta}, w_2), \end{aligned}$$

where $\kappa(\boldsymbol{\theta}, \mathbf{w}) = \kappa_1(\boldsymbol{\theta}) \kappa_2(\boldsymbol{\theta})$ *and* $\kappa(\boldsymbol{\theta}, \mathbf{w}) < 0$. *The only scenarii where these determinants are zero are whether all the partial derivatives are zero, or if* $(\partial/\partial a) g(\boldsymbol{\theta}, w_1)(\partial/\partial b) g(\boldsymbol{\theta}, w_2) = (\partial/\partial a) g(\boldsymbol{\theta}, w_2)\ (\partial/\partial b) g(\boldsymbol{\theta}, w_1)$. *Since the Lomax random variables are absolutely continuous, it is impossible for the generating function to be flat on* $\boldsymbol{\theta}$ *and on* $\mathbf{w}$, *except maybe in extreme cases. Therefore, situations where the determinants are zero are countable, and Assumption 1.36 (ii) is satisfied.*

Suppose the generating function satisfies the following property:

$$\lim_{\|(\boldsymbol{\theta}, w_1)\| \to \infty} g(\boldsymbol{\theta}, w_1) = \infty.$$

Since the limit of the natural logarithm tends to infinity when its argument diverges, we clearly have that

$$\lim_{\|(\boldsymbol{\theta}, \mathbf{w})\| \to \infty} \|\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}, \mathbf{w})\| = +\infty,$$

and as a consequence, Assumption 1.36 (iii) is satisfied.

It remains to demonstrate that a generating function satisfies the above properties. A natural and computationally easy choice for the generating function is the inverse cdf, it is given by

$$g(\boldsymbol{\theta}, u) = b + bu^{-1/q}, \quad u \sim \mathcal{U}(0,1).$$

Clearly the generating function is once continuously differentiable in each (b, q, u)*. The only possibilities for the partial derivatives of* g *to be zero are whether* $q = \{+\infty\}$ *or* $u = \{0\}$*. The generating function tends to infinity when* b *diverges whereas it remains constant when* q *or* u *diverges. All these findings strongly suggest that Theorem 1.38 is applicable here, and as a conclusion that any intervals built on the percentiles of the SwiZs distribution lead to exact frequentist coverage probabilities.*

However, the situation is less optimistic with the weighted maximum likelihood. Indeed, the estimating function is typically modified as follows:

$$\widetilde{\boldsymbol{\Phi}}_n\left(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}\right) = \mathrm{w}(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}, k)\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}\right),$$

where $\mathrm{w}(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}, k)$ *is some weight function typically taking values in* $[0, 1]$ *that depends upon a tuning constant* k*. Usual weight functions are Huber's type ([Hub+64]) and Tukey's biweighted function ([BT74]); see [Ham+11] for a textbook on robust statistics. For an estimating function to be robust, the weight function either decreases to 0 or remains constant for large values of* x*. As a consquence, at least two out of the three hypothesis of Assumption 1.36 do not hold. Indeed, the determinants will be zero on an uncountable set and* $\lim_{\|(\boldsymbol{\theta},\mathbf{w})\|\to\infty} \widetilde{\boldsymbol{\Phi}}_n < \infty$*.*

For the simulations, we set $\boldsymbol{\theta}_0 = (2 \;\; 2.3)^T$ *and use* $n = \{35, 50, 100, 150, 250, 500\}$ *as sample sizes. As already mentioned, this setup is close to the ones proposed in [GFG13], and we thus add their proposal for correcting the bias of the maximum likelihood estimator to the basket of the compared methods. The bias-adjustment estimator is given by*

$$\hat{\boldsymbol{\theta}}^{(s)}_{BA,n} = \hat{\boldsymbol{\pi}}_n - \mathbf{B}(\hat{\boldsymbol{\pi}}_n)\mathbf{A}(\hat{\boldsymbol{\pi}}_n)\,\mathrm{vec}\left(\mathbf{B}(\hat{\boldsymbol{\pi}}_n)\right),$$

where

$$\mathbf{A}(\boldsymbol{\pi}) = n\begin{pmatrix} \frac{2\pi_2}{\pi_1^3(\pi_2+2)(\pi_2+3)} & \frac{-1}{\pi_1^2(\pi_2+1)(\pi_2+2)} & \frac{\pi_2}{\pi_1^2(\pi_2+2)^2} & \frac{-1}{\pi_1(\pi_2+1)^2} \\ \frac{-1}{\pi_1^2(\pi_2+1)(\pi_2+2)} & 0 & \frac{-1}{\pi_1(\pi_2+1)^2} & \frac{1}{\pi_2^3} \end{pmatrix},$$

and

$$\mathbf{B}^{-1}(\boldsymbol{\pi}) = n\begin{pmatrix} \frac{\pi_2}{\pi_1^2(\pi_2+2)} & \frac{-1}{\pi_1(\pi_2+1)} \\ \frac{-1}{\pi_1(\pi_2+2)} & \frac{1}{\pi_2^2} \end{pmatrix}.$$

All the detailed results of simulation are in Appendix 1.D.1. In Figure 1.1, we discover that the SwiZs has very accurate coverage probabilities at all levels and all sample sizes which seems in accordance with Theorem 1.38 and the subsequent verification analysis for this example. For sample sizes greater or equal to 250, the parametric bootstrap and the bias-adjustment proposal of [GFG13] meet the performance of the SwiZs at almost every levels. However, below a sample of 150, the performance of the bias-adjustment are catastrophic. This may only be explained by the following phenomenon: the maximum likelihood is adjusted too severely for small values of n*, and for a large proportion of the time the resulting bias-adjusted estimator is out of the parameter space* $\boldsymbol{\Theta}$*. We report in Table 1.2 our empirical findings. This phenomenon affects not only the coverage probabilities but also the variation of this estimator (Figure 1.3) and the length of the confidence intervals (Figure 1.2). Here we opted for discarding the inadmissible values (negative), thereby reducing artificially the variance and the length of the confidence intervals of the*

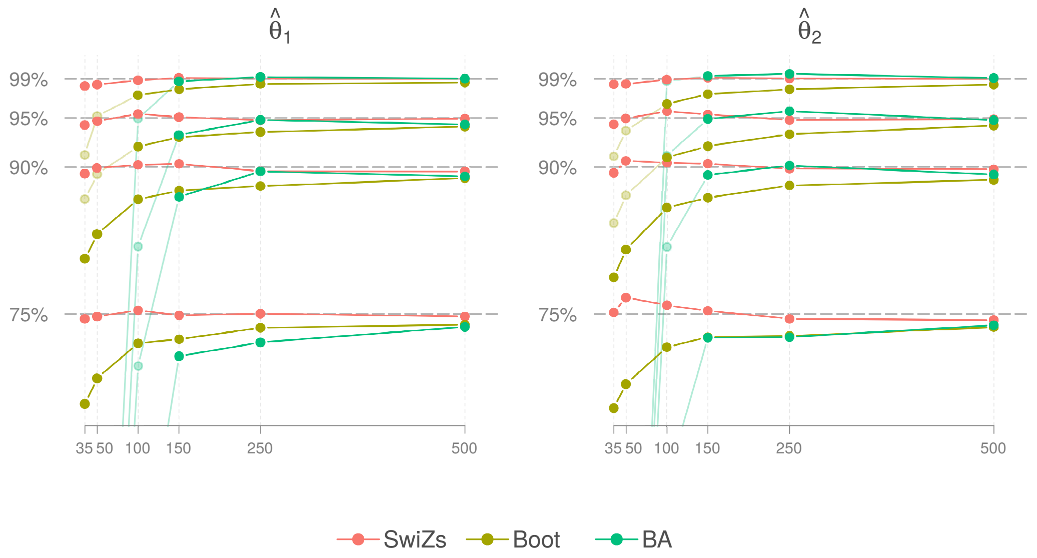

Figure 1.1: Coverage probabilities of the SwiZs, the parametric bootstrap (Boot) and the bias-adjustment (BA) proposal of [GFG13] for different sample sizes. *On the left panel* is the coverage for the first estimator, and the second is on the *right*. The gray horizontal dotted-lines indicate the perfect coverage probabilities. The closer to these lines is the better.

bias-adjustment. All the other methods considered do not suffer from the positivity constrain on $\boldsymbol{\theta}$ *and thus we do not attempt to tackle this limitation of the bias-adjustment method. The SwiZs has shorter uncertainty intervals than the parametric bootstrap,*

$n = 35$	$n = 50$	$n = 100$	$n = 150$
38.78%	21.94%	3.02%	0.40%

Table 1.2: Empirical proportion of times the bias-adjusted maximum likelihood estimator is jointly out of the parameter space $\boldsymbol{\Theta}$.

however it is more demanding in computational efforts (Figure 1.2). The computational comparison is not entirely fair in disfavor of the SwiZs as here we take advantage that the maximum likelihood estimator can be optimized directly on the log-likelihood, which is numerically easier to evaluate than the likelihood scores that constitues the estimating function. An unexpected good surprise emerges from Figure 1.3 where it seems that taking the median of the SwiZs leads to almost median unbiased point estimators. The same may be said when using the weighted maximum likelihood as the auxiliary estimator (Figure 1.5). However, using a robust estimator as the auxiliary parameter do not offer interesting coverage probabilities in small samples (Figure 1.4), which seems to indicate that Assumption 1.36 may not be easily relaxed. The parametric bootstrap unsurprisingly fails completely when considering an inconsistent estimator. Eventually, the empirical distributions in Figure 1.6 reminds us of the difficulty of estimating confidence regions.

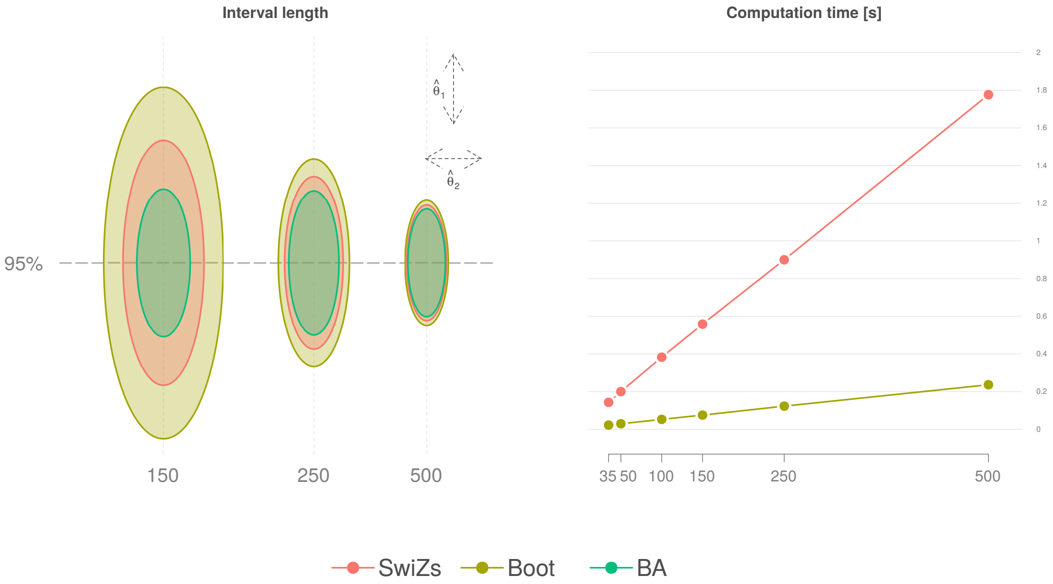

Figure 1.2: *On the left panel*: representation of the median interval lengths for a confidence level of 95% for the SwiZs, the parametric bootstrap (Boot) and the bias-adjustment (BA) proposal of [GFG13] for three different sample sizes. The ellipses are just a representation and do not reflect the real shapes of the confidence regions. All the ellipses are on the same scale. The centre of the ellipses is chosen for aesthetical reason and have no special meaning. The y-axis corresponds to the median interval length of the first parameter, the x-axis the one of the second parameter. The smaller the ellipse is, the better it is. *On the right panel*: the average computational time in seconds of the SwiZs and the Boot for the different sample sizes. Note that the computational time of the the BA (not on the figure) is quasi-identical to the Boot. The lower is the better.

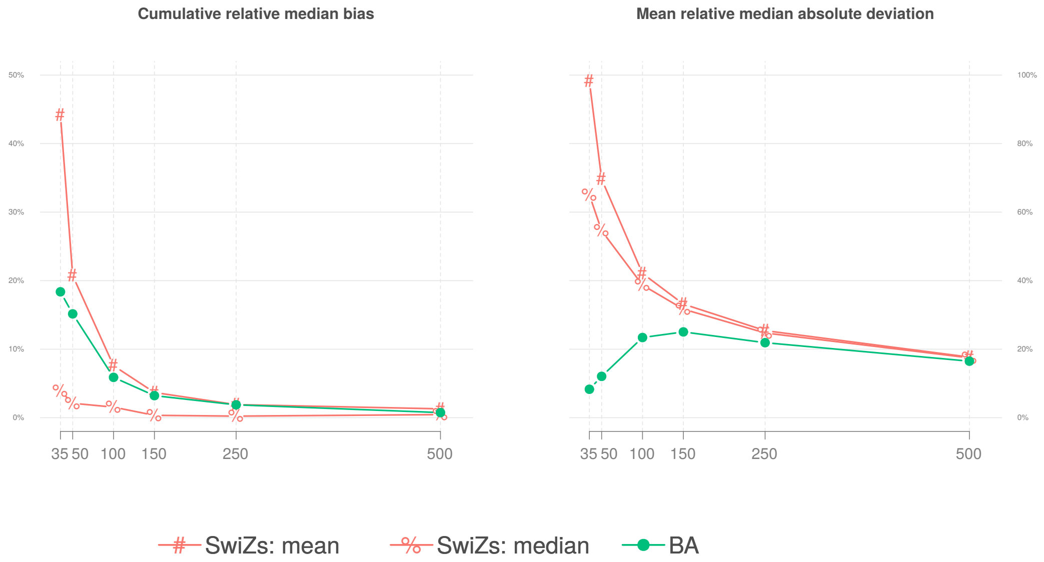

Figure 1.3: *On the left panel*: the sum of absolute value of the median bias for the two estimators divided by their respective true values for the mean of SwiZs distribution, the median of the SwiZs distribution and the bias-adjustment (BA) proposal of [GFG13] evaluated on the different sample sizes. *On the right panel*: likewise the *left panel*, but for a different measure: the average of the median absolute deviation for the two estimators divided by their respective true values. The lower is the better.

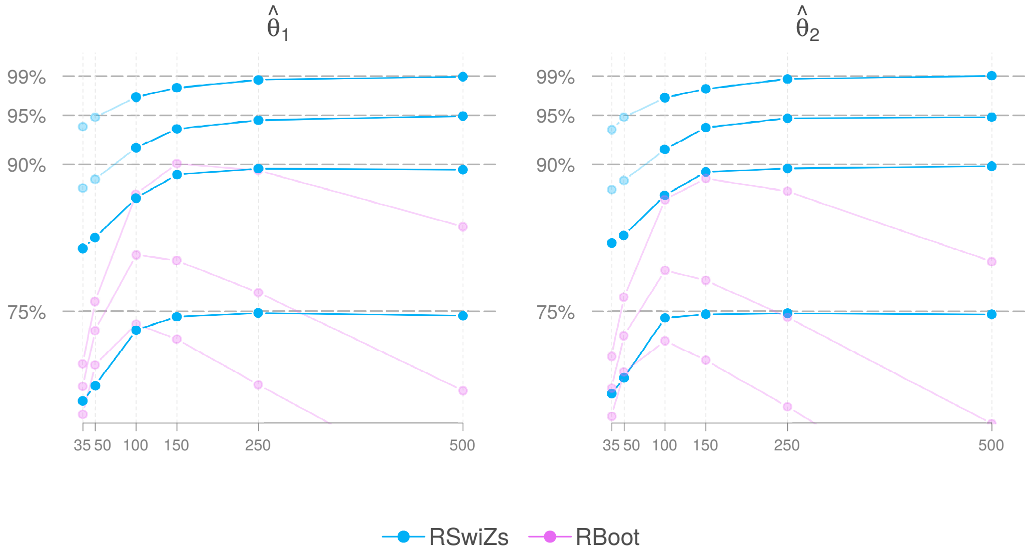

Figure 1.4: Coverage probabilities for different sample sizes of the SwiZs (RSwiZs) and the parametric bootstrap (RBoot) when taking the weighted maximum likelihood as auxiliary estimator. *On the left panel* is the coverage for the first estimator, and the second is on the *right*. The gray horizontal dotted-lines indicate the perfect coverage probabilities. The closer to these lines is the better.

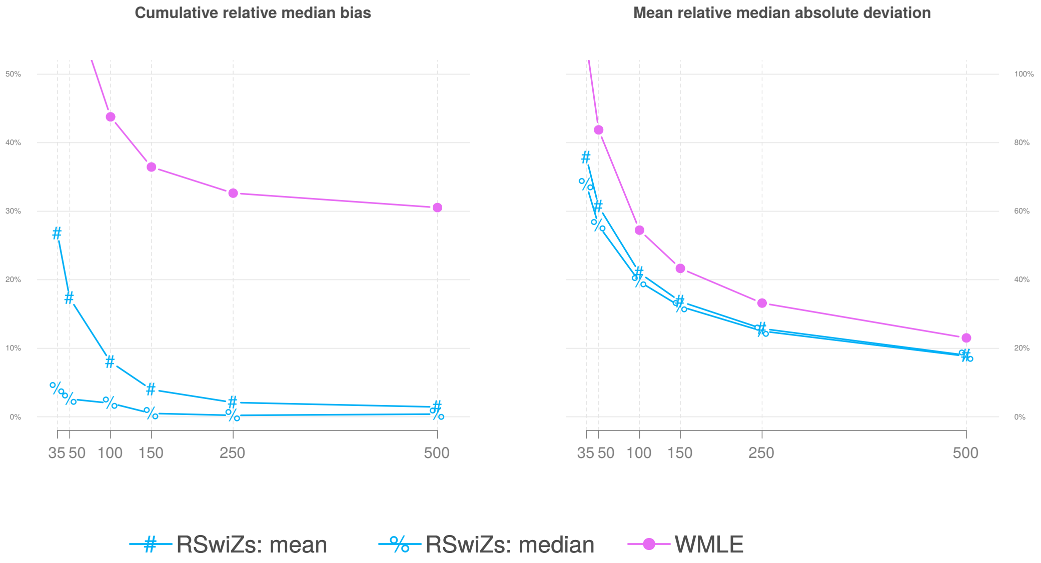

Figure 1.5: *On the left panel*: the sum of absolute value of the median bias for the two estimators divided by their respective true values for different sample sizes for the mean of SwiZs distribution (RSwiZs: mean), the median of the SwiZs distribution (RSwiZs: median) when considering the weighted maximum likelihood (WMLE) as the auxiliary estimator. *On the right panel*: likewise the *left panel*, but for a different measure: the average of the median absolute deviation for the two estimators divided by their respective true values. The lower is the better.

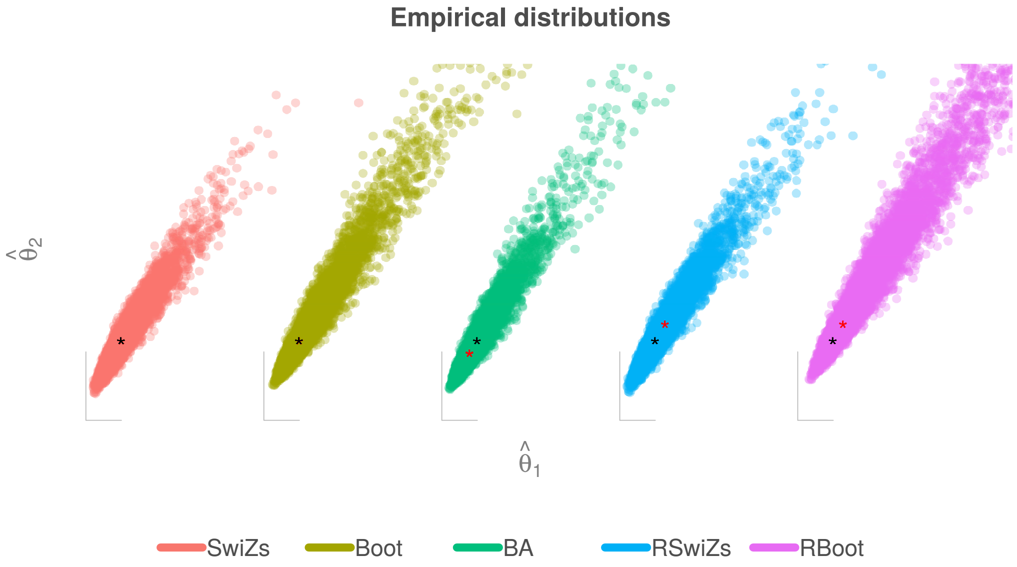

Figure 1.6: Empirical conditional distribution for a given $\hat{\boldsymbol{\pi}}_n$ and a sample size of $n = 100$ of the SwiZs, the parametric bootstrap (Boot), the bias-adjustment proposal of [GFG13] (BA) when considering the maximum likelihood as the auxiliary estimator and the SwiZs (RSwiZs) and the parametric bootstrap (RBoot) when considering the weighted maximum likelihood as the auxiliary estimator. The *black star* represents $\boldsymbol{\theta}_0 = [2\ 2.3]^T$ whereas the *red stars* indicate the values of $\hat{\boldsymbol{\pi}}_n$: the maximum likelihood estimator for SwiZs and Boot, the bias-adjustment for BA and the weighted maximum likelihood estimator for RSwiZs and RBoot. The "*try square*" at *bottom-left-corner* of each distribution has both sides of length 2 and has its corner exactly at the $(0, 0)$-coordinate.

Third, we investigate a linear mixed-model. These models are very common in statistics as they incorporate both parameters associated with an entire population and parameters associated with individual experimental units facilitating thereby the study of, for examples, longitudinal data, multilevel data and repeated measure data. Although being widespread, the inference on the parameters remain a formidable task. We study a rather simple model, namely the random intercept and random slope model when data is balanced.

Example 1.62 (random intercept and random slope linear mixed model)**.** *Consider the following balanced Gaussian mixed linear model expressed for the ith individual as*

$$\mathbf{y}_i = (\beta_0 + \alpha_i)\mathbf{1}_m + (\beta_1 + \gamma_i)\mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, n,$$

where $\boldsymbol{\epsilon}_i, \alpha_i$ and γ_i are identically and independently distributed according to centered Gaussian distributions with respective variances $\sigma^2_\epsilon \mathbf{I}_m, \sigma^2_\alpha$ and σ^2_γ, m being the number of replicates, the same for each individual, and $\mathbf{1}_m$ is a vector of m ones. The vector of parameters of interest is $\boldsymbol{\theta} = \left(\beta_0, \beta_1, \sigma^2_\epsilon, \sigma^2_\alpha, \sigma^2_\gamma\right)^T$. Let $\boldsymbol{\pi} = (\pi_0, \dots, \pi_4)^T$ be the corresponding vector of auxiliary parameters. We take the MLE as the auxiliary estimator and thus consider the likelihood score function as the estimating function. With this setup, the parameter spaces $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$ are equivalent, and the parametric bootstrap may be employed. Denote by $N = nm$ the total sample size. The negative log-likelihood may be expressed as

$$\boldsymbol{\ell}\left(\mathbf{y}, \boldsymbol{\theta}\right) = k + \frac{1}{2N}\sum_{i=1}^{n} \log\left(\det\left(\boldsymbol{\Omega}_i(\boldsymbol{\theta})\right)\right) + (\mathbf{y}_i - \beta_0\mathbf{1}_m - \beta_1\mathbf{x}_i)^T\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\theta})(\mathbf{y}_i - \beta_0\mathbf{1}_m - \beta_1\mathbf{x}_i),$$

for some constant k and where $\boldsymbol{\Omega}_i(\boldsymbol{\theta}) = \sigma^2_\epsilon\mathbf{I}_m + \sigma^2_\alpha\mathbf{1}_m\mathbf{1}_m^T + \sigma^2_\gamma\mathbf{x}_i\mathbf{x}_i^T$ is clearly a symmetric positive definite matrix. Taking the derivatives with respect to $\boldsymbol{\theta}$, then substituing $\boldsymbol{\theta}$ by $\boldsymbol{\pi}$ and $\mathbf{y}_i$ by $\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}_i)$ leads to

$$\boldsymbol{\Phi}_N\left(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}\right) = \begin{pmatrix} \frac{-1}{N}\sum_{i=1}^n \mathbf{z}^T(\boldsymbol{\theta}, \mathbf{u}_i, \boldsymbol{\pi})\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\pi})\mathbf{1}_m \\ \\ \frac{-1}{N}\sum_{i=1}^n \mathbf{z}^T(\boldsymbol{\theta}, \mathbf{u}_i, \boldsymbol{\pi})\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\pi})\mathbf{x}_i \\ \\ \frac{1}{2N}\sum_{i=1}^n \operatorname{trace}\left(\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\pi})\frac{\partial}{\partial\pi_j}\boldsymbol{\Omega}_i(\boldsymbol{\pi})\right) \\ \quad -\mathbf{z}^T(\boldsymbol{\theta}, \mathbf{u}_i, \boldsymbol{\pi})\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\pi})\frac{\partial}{\partial\pi_j}\boldsymbol{\Omega}_i(\boldsymbol{\pi}) \\ \quad \times\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\pi})\mathbf{z}(\boldsymbol{\theta}, \mathbf{u}_i, \boldsymbol{\pi}), \quad j = 2, 3, 4 \end{pmatrix},$$

where $\mathbf{z}(\boldsymbol{\theta}, \mathbf{u}_i, \boldsymbol{\pi}) = \mathbf{g}(\boldsymbol{\theta}, \mathbf{u}_i) - \pi_0\mathbf{1}_m - \pi_1\mathbf{x}_i$ (see also [Jia07] for more details on these derivations). The derivatives of $\boldsymbol{\Omega}_i(\boldsymbol{\pi})$ are easily obtained: $(\partial/\partial\pi_2)\boldsymbol{\Omega}_i(\boldsymbol{\pi}) = \mathbf{I}_m$, $(\partial/\partial\pi_3)\boldsymbol{\Omega}_i(\boldsymbol{\pi}) = \mathbf{1}_m\mathbf{1}_m^T$ and $(\partial/\partial\pi_4)\boldsymbol{\Omega}_i(\boldsymbol{\pi}) = \mathbf{x}_i\mathbf{x}_i^T$. Since they do not depend on parameters, let denotes $(\partial/\partial\pi_j)\boldsymbol{\Omega}_i(\boldsymbol{\pi}) \equiv \mathbf{D}_{ij}$.

We now motivate the possibility to employ Theorem 1.38 by verifying Assumption 1.36. First, we suppose that a random variable $\mathbf{w}$ of the same dimension as $\boldsymbol{\theta}$ exists. Then, we assume that the estimating function may be re-expressed as follows:

$$\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_N}\left(\boldsymbol{\theta}, \mathbf{w}\right) = \begin{pmatrix} \frac{-1}{N}\sum_{i=1}^n \mathbf{z}_i^T(\boldsymbol{\theta}, w_0, \hat{\boldsymbol{\pi}}_N)\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{1}_m \\ \\ \frac{-1}{N}\sum_{i=1}^n \mathbf{z}_i^T(\boldsymbol{\theta}, w_1, \hat{\boldsymbol{\pi}}_N)\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{x}_i \\ \\ \frac{1}{2N}\sum_{i=1}^n \operatorname{trace}\left(\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{D}_{ij}\right) \\ \quad -\mathbf{z}_i^T(\boldsymbol{\theta}, w_j, \hat{\boldsymbol{\pi}}_N)\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{D}_{ij} \\ \quad \times\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{z}_i(\boldsymbol{\theta}, w_j, \hat{\boldsymbol{\pi}}_N), \quad j = 2, 3, 4 \end{pmatrix},$$

where $\mathbf{z}_i(\boldsymbol{\theta}, w_j, \hat{\boldsymbol{\pi}}_N) = \mathbf{g}(\boldsymbol{\theta}, w_j) - \hat{\pi}_0\mathbf{1}_m - \hat{\pi}_1\mathbf{x}_i$, $j = 0, 1, 2, 3, 4$, and $\hat{\boldsymbol{\pi}}_N$ is fixed. The Jacobian matrix with respect to $\boldsymbol{\theta}$ is given by

$$D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_N}(\boldsymbol{\theta}, \mathbf{w}) = \begin{pmatrix} \frac{-1}{N}\sum_{i=1}^{n} D_{\boldsymbol{\theta}}\mathbf{g}^T(\boldsymbol{\theta}, w_0)\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{1}_m \\ \\ \frac{-1}{N}\sum_{i=1}^{n} D_{\boldsymbol{\theta}}\mathbf{g}^T(\boldsymbol{\theta}, w_1)\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{x}_i \\ \\ \frac{-1}{N}\sum_{i=1}^{n} D_{\boldsymbol{\theta}}\mathbf{g}^T(\boldsymbol{\theta}, w_j)\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{D}_{ij} \\ \quad\times\boldsymbol{\Omega}_i^{-1}(\hat{\boldsymbol{\pi}}_N)\mathbf{g}(\boldsymbol{\theta}, w_j), \quad j = 2, 3, 4 \end{pmatrix}.$$

Substituing $D_{\boldsymbol{\theta}}\mathbf{g}^T$ by $D_{\mathbf{w}}\mathbf{g}^T$ in the above delivers immediately the Jacobian matrix with respect to $\mathbf{w}$. Note that this second Jacobian is a diagonal matrix. Clearly, the differentiability and continuity of $\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_N}$ depends exclusively upon the differentiability and continuity of $\mathbf{g}$. Ergo, if $D_{\boldsymbol{\theta}}\mathbf{g}$ and $D_{\mathbf{w}}\mathbf{g}$ exist and are continuous, then Assumption 1.36 (i) holds.

These Jacobian matrices may have a null determinant under two circumstances: whether the generating function $\mathbf{g}$ is flat on $\boldsymbol{\theta}$ and/or $\mathbf{w}$, and/or whether they are linearly dependent. Since the Normal distribution is absolutely continuous, $\mathbf{g}$ may be flat only on extreme cases. The Jacobian $D_{\mathbf{w}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_N}$ is a diagonal matrix, so its determinant is null if and only if one of its diagonal element is null. Since both the design and $\hat{\boldsymbol{\pi}}_N$ are fixed, situations where $D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_N}$ is linearly dependent may occur if the vectors $(\partial/\partial\theta_j)\mathbf{g}(\boldsymbol{\theta}, \mathbf{w}) = k(\partial/\partial\theta_{j'})\mathbf{g}(\boldsymbol{\theta}, \mathbf{w}), j \neq j'$, for some constant $k \in \mathbb{R}$. But because $\mathbf{w}$ is random, this situation is unlikely to occur, and, depending on $\mathbf{g}$, Assumption 1.36 (ii) is plausible.

Eventually, it clearly holds that

$$\lim_{\|(\boldsymbol{\theta}, \mathbf{w})\| \to \infty} \|\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_N}(\boldsymbol{\theta}, \mathbf{w})\| = \infty$$

if $\|\mathbf{g}(\boldsymbol{\theta}, \mathbf{w})\| \to \infty$ as $\|(\boldsymbol{\theta}, \mathbf{w})\| \to \infty$, so Assumption 1.36 (iii) is satisfied given that $\mathbf{g}$ fulfills the requirement.

Once again, the plausibility of Assumption 1.36 is up to the choice of the generating function. A popular choice is the following:

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}_i) = \beta_0\mathbf{1}_m + \beta_1\mathbf{x}_i + \mathbf{C}_i(\boldsymbol{\theta})\mathbf{u}_i, \quad \mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m),$$

where $\mathbf{C}_i(\boldsymbol{\theta})$ is the lower triangular Cholesky factor such that $\mathbf{C}_i(\boldsymbol{\theta})\mathbf{C}_i^T(\boldsymbol{\theta}) = \boldsymbol{\Omega}_i(\boldsymbol{\theta})$. It is straightforward to remark that $\mathbf{g}$ is once continuously differentiable in β_0, β_1 and $\mathbf{u}_i$. For the variances components, the partial derivatives of the Cholesky factor is given by Theorem A.1 in [Sär13]:

$$\frac{\partial}{\partial\theta_j}\mathbf{C}_i(\boldsymbol{\theta}) = \mathbf{C}_i(\boldsymbol{\theta})L\left(\mathbf{C}_i^{-1}(\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\boldsymbol{\Omega}_i(\boldsymbol{\theta})\mathbf{C}_i^{-T}(\boldsymbol{\theta})\right), \quad j = 2, 3, 4,$$

where the function L returns the lower triangular and half of the diagonal elements of the inputed matrix, that is:

$$L_{ij}(\mathbf{A}) = \begin{cases} \mathbf{A}_{ij}, & i > j, \\ \frac{1}{2}\mathbf{A}_{ij}, & i = j, \\ 0, & i < j. \end{cases}$$

The partial derivatives of the covariance matrix are given by: $(\partial/\partial\sigma_\epsilon^2)\boldsymbol{\Omega}_i(\boldsymbol{\theta}) = \mathbf{I}_m$, $(\partial/\partial\sigma_\alpha^2)\boldsymbol{\Omega}_i(\boldsymbol{\theta}) = \mathbf{1}_m\mathbf{1}_m^T$ and $(\partial/\partial\sigma_\gamma^2)\boldsymbol{\Omega}_i(\boldsymbol{\theta}) = \mathbf{x}_i\mathbf{x}_i^T$. Hence, $\mathbf{C}_i(\boldsymbol{\theta})$ is once differentiable. For the continuity of the partial derivative of $\mathbf{C}_i(\boldsymbol{\theta})$, note that $\mathbf{C}_i(\boldsymbol{\theta})$ and $\mathbf{C}_i^{-1}(\boldsymbol{\theta})$ are once differentiable and thus continuous. Indeed, $(\partial/\partial\theta_j)\mathbf{C}_i^{-1}(\boldsymbol{\theta}) = -\mathbf{C}_i^{-1}(\boldsymbol{\theta})[(\partial/\partial\theta_j)\mathbf{C}_i(\boldsymbol{\theta})]\mathbf{C}_i^{-1}(\boldsymbol{\theta})$. Eventually,

$(\partial/\partial\theta_j)\boldsymbol{\Omega}_i(\boldsymbol{\theta})$ is constant in $\boldsymbol{\theta}$, and therefore continuous. Since matrix product preserves the continuity, the Cholesky factor is once continuously differentiable. The partial derivatives of $\mathbf{g}$ may be zero if the design is null or if the pivotal quantity is zero, two extreme situations unlikely encountered. It is straightforward to remark that the estimating function diverges as $\boldsymbol{\theta}$ and $\mathbf{u}_i$ tends to infinity. All these findings make usage of Theorem 1.38 highly plausible.

Let us turn our attention to simulations. We set $\boldsymbol{\theta}_0 = (1, 0.5, 0.5^2, 0.5^2, 0.2^2)^T$ and considered $n = m = \{5, 10, 20, 40\}$ such that $N = nm = \{25,\ 100,\ 400,\ 1,600\}$. The detailed results of simulations may be found in the tables of Appendix 1.D.2. In Figure 1.7, we can observe the outstanding performances of the SwiZs in terms of coverage probabilities, which supports our analysis and the possibility of using Theorem 1.38. The parametric bootstrap meets the performance of the SwiZs as the sample size increases, however, when the sample size is small, it is off the ideal level for the variance components. The length of the marginal intervals of uncertainty are comparable between the two methods, except for the smallest sample size considered where it is anyway harder to interpret the size of the interval of the parametric bootstrap since it is off the confidence level. We also bear the comparison with profile likelihood confidence intervals which are based on likelihood ratio test. The coverage probabilities are almost undistinguishable from the SwiZs whereas interval lengths for variance components are the shortest. We interpret such good performances as follows: first, as shown in Example 1.56 on linear regression, asymptotic and finite sample distributions coincides in theory, coincidance that may be still hold in the present case with balanced linear mixed model; second, larger intervals accounts for the fact that no simulations are needed. A good surprise appears in Figure 1.8 where the median of the SwiZs shows good performances in terms of relative median bias.

Fourth, we study inference in queueing theory models (see [Sho+18] for a monograph). In particular, we re-investigate the M/G/1 model studied by [HF04; BF10; FP12]. Although the underlying process is relatively simple, there is no known closed-form for the likelihood function and inference is not easy to conduct.

Example 1.63 (M/G/1-queueing model). *Consider the following stochastic process*

$$x_i = \begin{cases} v_i, & \text{if } \sigma_i^\varepsilon \leq \sigma_{i-1}^x, \\ v_i + \sigma_i^\varepsilon - \sigma_{i-1}^x, & \text{if } \sigma_i^\varepsilon > \sigma_{i-1}^x, \end{cases}$$

for $i = 1, \cdots, n$, where $\sigma_i^\varepsilon = \sum_{j=1}^i \varepsilon_j$, $\sigma_i^x = \sum_{j=1}^i x_j$, v_i is identically and independently distributed according to a uniform distribution $\mathcal{U}(\theta_1, \theta_2)$, $0 \leq \theta_1 < \theta_2 < \infty$ and ε_i is identically and independently distributed according to an exponential distribution $\mathcal{E}(\theta_3)$, $\theta_3 > 0$. In queueing theory, random variables have special meaning, for the ith customer: x_i represents interdeparture time, v_i is service time and ε_i corresponds to interarrival time. Only the interdeparture times x_i are observed, v_i and ε_i are latent. All past information influence the current observation and therefore this process is not Markovian. Finding an "appropriate" auxiliary estimator is challenging as we now discuss.

In this context, semi-automatic ABC approaches by [BF10] and [FP12] use several quantiles as summary statistics for the auxiliary estimator. This method cannot be employed here for the SwiZs because, first, the restriction that $\dim(\boldsymbol{\theta}) = \dim(\boldsymbol{\pi})$ would be violated, and second, the quantiles are non-differentiables with respect to $\mathbf{g}$ and consequently, as already discussed, Assumptions 1.36 and 1.37 would not hold. However, [HF04] present

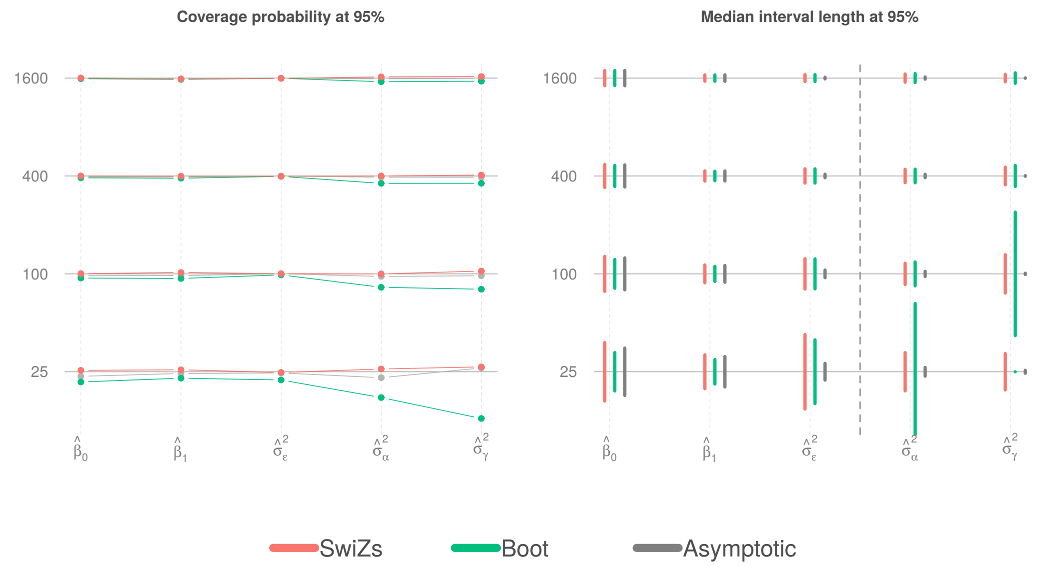

Figure 1.7: *On the left panel*: Representation of the coverage probabilities for different sample sizes of the SwiZs, the parametric bootstrap (Boot) and the confindence intervals based on the likelihood ratio test (Asymptotic) for the five estimators. The gray line represents the ideal level of 95% coverage probabilitiy. *On the right panel*: median length of the marginal intervals of uncertainty at a level of 95%. For graphical reason, the lengths corresponding to $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_\gamma$ *on the right* is downsized by a factor of 5 compared to the lengths corresponding to the other estimators.

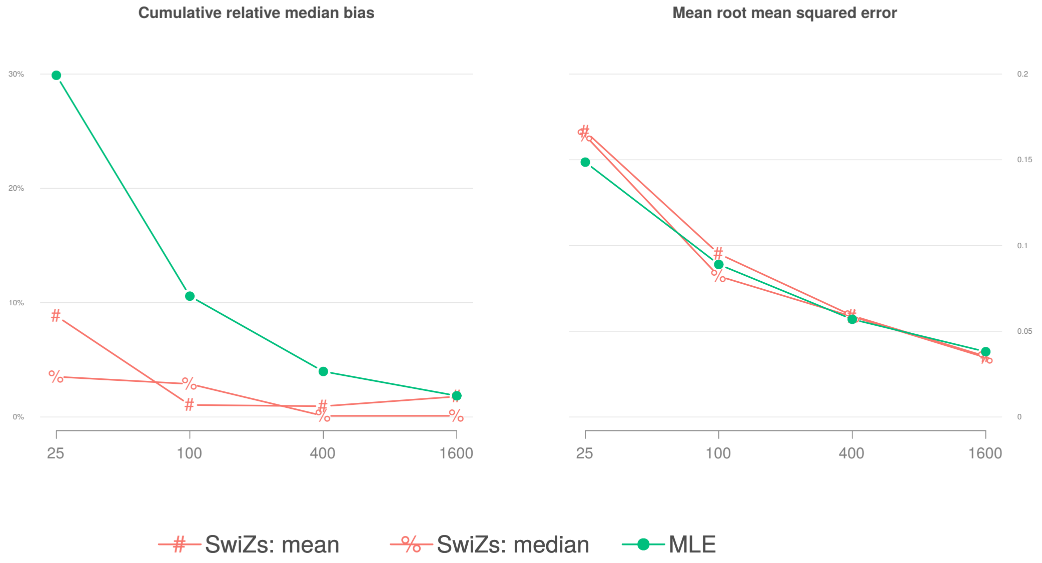

Figure 1.8: *On the left panel*: the sum of absolute value of the median bias for the five estimators divided by their respective true values for different sample sizes for the mean of SwiZs distribution (SwiZs: mean), the median of the SwiZs distribution (SwiZs: median) and the maximum likelihood estimator (MLE). *On the right panel*: likewise the *left panel*, but for a different measure: the average of root mean squared error for the five estimators. For both panels, the lower is the better.

different choices and motivate a particular auxiliary model with the following closed-form:

$$f(x_i,\boldsymbol{\pi}) = \begin{cases} 0, & \textit{if } x_i \leq \pi_1, \\ (\pi_2-\pi_1)^{-1}\left[1-\alpha\exp\left(-\pi_3^{-1}(x_i-\pi_1)\right)\right], & \textit{if } \pi_1 < x_i \leq \pi_2, \\ \frac{\alpha}{\pi_2-\pi_1}\left[\exp\left(-\pi_3^{-1}(x_i-\pi_2)\right)-\exp\left(-\pi_3^{-1}(x_i-\pi_1)\right)\right], & \textit{if } x_i > \pi_2, \end{cases}$$

where $-1 \leq \alpha \leq 1$ is some constant. Motivations for this auxiliary model are based on a graphical analysis of the sensitivity of $\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and the root mean squared errors performances of $\hat{\boldsymbol{\theta}}_n$ on simulations. Unfortunately, Assumption 1.36 is not satisfied with this choice. Indeed, by taking the likelihood scores of the auxiliary model as the estimating equation, one can realize that the score relative to π_2 is

$$\Phi_{n,2}(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi}) = \begin{cases} 0, & \textit{if } g(\boldsymbol{\theta},\mathbf{u}) < \pi_1, \\ \frac{1}{\pi_2-\pi_1}, & \textit{if } \pi_1 \leq g(\boldsymbol{\theta},\mathbf{u}) < \pi_2, \\ \frac{1}{\pi_2-\pi_1} - \frac{\pi_3^{-1}e^{\pi_2/\pi_3}}{e^{\pi_2/\pi_3}-e^{\pi_1/\pi_3}}, & \textit{if } g(\boldsymbol{\theta},\mathbf{u}) \geq \pi_2, \end{cases}$$

hence, it does not depend on $\boldsymbol{\theta}$! This result implies directly that all the partial derivatives with respect to $\boldsymbol{\theta}$ and $\mathbf{w}$ are null and $\det(\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}) = 0$ for all $(\boldsymbol{\theta},\mathbf{w}) \in (\boldsymbol{\Theta}_n \times W_n)$. Assumption 1.37 is also violated and Theorem 1.38 cannot be invoked. Worse, the behaviour of this score does not depend on n and the identifiability condition in Assumption 1.43 (ii) does not hold since $\Phi_2(\boldsymbol{\theta}_1,\boldsymbol{\pi}) = \Phi_2(\boldsymbol{\theta}_2,\boldsymbol{\pi})$ for all $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \in \boldsymbol{\Theta}$, so using this auxiliary model does not lead to a consistent estimator. It is however not clear whether Assumption 1.44, the alternative to Assumption 1.43, holds or not because the quantities to verify are unknown. Note however that in view of the equivalence theorem between the SwiZs *and the indirect inference estimator (Theorem 1.8), it would appear as a contradiction for Assumption 1.43 not to hold but Assumption 1.44 to be satisfied.*

[HF04] idea is to select an auxiliary model where $\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})$ is both sensitive to $\boldsymbol{\theta}$ and efficient for a given $\boldsymbol{\theta}$. Since they justify their choice on a graphical analysis with simulated samples, one may wonder whether the authors were unlucky or misleaded by the graphics on this particular example. In fact, although $\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})$ is unknown in an explicit form, its Jacobian may be derived explicitly by mean of an implicit function theorem, so for a given $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ we have:

$$D_{\boldsymbol{\theta}}\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}_1) = -\left[D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}, \boldsymbol{\pi}\right)\Big|_{\boldsymbol{\pi}=\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}_1)} \right]^{-1} D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}, \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}_1)\right).$$

The Jacobian $D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n$ is non zero. Yet, as already discussed, the second partial derivative of $\boldsymbol{\Phi}_n$ with respect to $\boldsymbol{\theta}$ is null. Because only the second row of $D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n$ has zero entries, there is no reason to believe that $D_{\boldsymbol{\theta}}\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})$ has zero entries. Consequently, the authors were not misleaded by the gaphics or unlucky, it is the criterion itself that is misleading.

We now face ourselves to the delicate task of choosing an auxiliary model which non-only respects the constraint $\dim(\boldsymbol{\theta}) = \dim(\boldsymbol{\theta})$, but also makes Assumption 1.36 plausible. In view of this particular M/G/1 stochastic process, using the convolution between a gamma with shape parameter n and unknown rate parameter and a uniform distributions may be a "natural" choice, yet, terms computationally complicated to evaluate readily appear. We propose instead of using Fréchet's three parameters extreme value distribution, whose density is given, for $i = 1, \dots, n$, by:

$$f(x_i, \boldsymbol{\pi}) = \frac{\pi_1}{\pi_2}\left(\frac{x_i - \pi_3}{\pi_2}\right)^{-1-\pi_1} \exp\left\{ -\left(\frac{x_i - \pi_3}{\pi_2}\right)^{-\pi_1} \right\}, \quad \textit{if } x_i > \pi_3,$$

where $\pi_1 > 0$ is a shape parameter, $\pi_2 > 0$ is a scale parameter and $\pi_3 \in \mathbb{R}$ is a parameter representing the location of the minimum. The relationship between π_3 and θ_1 as the minimum of the distribution seems natural and we thus further constrain here π_3 to be non-negative, so $\boldsymbol{\pi} > 0$. However, the existence of a potential link between $(\theta_2, \theta_3)^T$ and $(\pi_1, \pi_2)^T$ is not self-evident, but certainly that the shape (π_1) and scale (π_2) parameters offer enough flexibility to "encompass" the distribution of the M/G/1 stochastic process as illustrated in Figure 1.9. Note that the "closeness" between M/G/1 and Fréchet models is also dependent on the parametrization. It remains to advocate this choice in the light of Assumption 1.36. We take the maximum likelihood estimator of Fréchet's distribution as the auxiliary estimator and thus the likelihood score as the estimating function, which is given by:

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}\right) = \begin{pmatrix} \frac{-1}{\pi_1} + \frac{1}{n}\sum_{i=1}^n \log\left(\frac{\mathbf{g}(\boldsymbol{\theta},\mathbf{u}_i)-\pi_3}{\pi_2}\right)\left[1 - \left(\frac{\mathbf{g}(\boldsymbol{\theta},\mathbf{u}_i)-\pi_3}{\pi_2}\right)^{-\pi_1}\right] \\ -\frac{\pi_1}{\pi_2}\frac{1}{n}\sum_{i=1}^n \left[1 - \left(\frac{\mathbf{g}(\boldsymbol{\theta},\mathbf{u}_i)-\pi_3}{\pi_2}\right)^{-\pi_1}\right] \\ \frac{-1}{n}\sum_{i=1}^n \frac{1+\pi_1}{\mathbf{g}(\boldsymbol{\theta},\mathbf{u}_i)-\pi_3} + \frac{\pi_1}{\pi_2}\frac{1}{n}\sum_{i=1}^n \left(\frac{\mathbf{g}(\boldsymbol{\theta},\mathbf{u}_i)-\pi_3}{\pi_2}\right)^{-\pi_1-1} \end{pmatrix}.$$

Let us assume that a random variable $\mathbf{w}$ with the same dimension as $\boldsymbol{\theta}$ exists such that the estimating function may be expressed as follows:

$$\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}\left(\boldsymbol{\theta}, \mathbf{w}\right) = \begin{pmatrix} \frac{-1}{\hat{\pi}_1} + \log\left(z_1\right)\left[1 - {z_1}^{-\hat{\pi}_1}\right] \\ -\frac{\hat{\pi}_1}{\hat{\pi}_2}\left[1 - {z_2}^{-\hat{\pi}_1}\right] \\ -\frac{(1+\hat{\pi}_1){z_3}^{-1}}{\hat{\pi}_2} + \frac{\hat{\pi}_1}{\hat{\pi}_2}{z_3}^{-\hat{\pi}_1-1} \end{pmatrix},$$

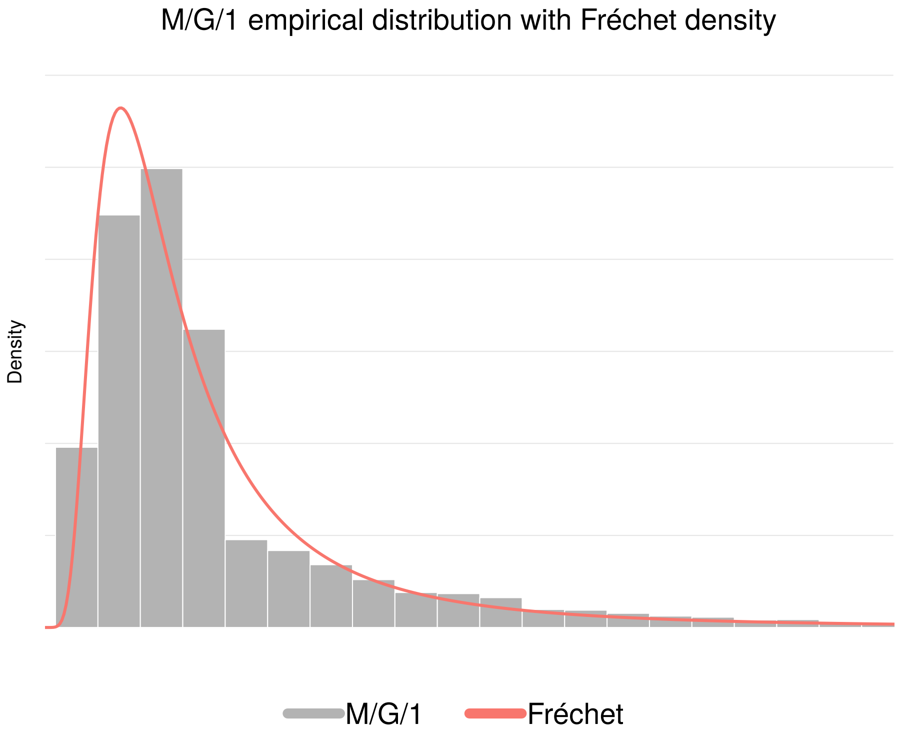

Figure 1.9: Histogram of a simulated M/G/1 stochastic process of size $n = 10^4$ on which the density (solid line) of Fréchet distribution has been added. The true parameter is $\boldsymbol{\theta}_0 = [0.3\ 0.9\ 1]^T$, the auxiliary estimator we obtain here is approximately $\hat{\boldsymbol{\pi}}_n = [0.02\ 0.60\ 2.05]^T$.

where $\hat{\boldsymbol{\pi}}_n$ is fixed and $z_i \equiv \frac{\mathbf{g}(\boldsymbol{\theta},w_i)-\hat{\pi}_3}{\hat{\pi}_2}$, $i=1,2,3$. The Jacobian matrix with respect to $\boldsymbol{\theta}$ is give by:

$$D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta},\mathbf{w}) = \begin{pmatrix} D_{\boldsymbol{\theta}^T} g(\boldsymbol{\theta},w_1)\left[\frac{z_1^{-1}}{\hat{\pi}_2}\left(1-z_1^{-\hat{\pi}_1}\right)+\frac{\hat{\pi}_1}{\hat{\pi}_2}\log\left(z_1\right)z_1^{-\hat{\pi}_1-1}\right] \\ D_{\boldsymbol{\theta}^T} g(\boldsymbol{\theta},w_2)\left[-\frac{\hat{\pi}_1^2}{\hat{\pi}_2^2}z_2^{-\hat{\pi}_1-1}\right] \\ D_{\boldsymbol{\theta}^T} g(\boldsymbol{\theta},w_3)\left[\frac{(\hat{\pi}_1-1)}{\hat{\pi}_2^2}z_3^{-2}-\frac{\hat{\pi}_1(\hat{\pi}_1+1)}{\hat{\pi}_2^2}z_3^{-\hat{\pi}_1-2}\right] \end{pmatrix}.$$

Substituing $D_{\boldsymbol{\theta}}\mathbf{g}^T$ by $D_{\mathbf{w}}\mathbf{g}^T$ in the above equation gives the Jacobian matrix with respect to $\mathbf{w}$, a matrix which is diagonal. It is straightforward to remark that the differentiability and continuity depends exclusively on the smoothness of $\mathbf{g}$. Thus, if $\mathbf{g}$ is once continuously differentiable in both $\boldsymbol{\theta}$ and $\mathbf{w}$, then Assumption 1.36 (i) holds.

Concerning the determinant of these Jacobian matrices, they may be null only on unlikely situations: first, if $\mathbf{g}$ equals $\hat{\pi}_3$ then z_i is zero for $i=1,2,3$, second, if $D_{\boldsymbol{\theta}}\mathbf{g}$ or $D_{\mathbf{w}}\mathbf{g}$ are zeros. The choice of $\mathbf{g}$ may be guided by this restriction so typically the determinants may be null, but only on a countable set, and Assumption 1.36 (ii) is verified. For Assumption 1.36 (iii), it is straightforward to remark that

$$\lim_{\|(\boldsymbol{\theta},\mathbf{w})\|\to\infty}\|\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta},\mathbf{w})\|,$$

as long as $\lim_{\|(\boldsymbol{\theta},\mathbf{w})\|\to\infty}\|\mathbf{g}(\boldsymbol{\theta},\mathbf{w})\|=\infty$, since $\log(z_1)$ would diverge. Depending on g, Assumption 1.36 (iii) is satisfied.

Therefore, the plausibility of Assumption 1.36 is up to the choice of the generating equation g. Here, the choice is quasi immediate as it is driven by the form of the process:

$$g(\boldsymbol{\theta},\mathbf{u}_i) = \begin{cases} v_i(\boldsymbol{\theta}), & \textit{if } \sigma_i^{\varepsilon}(\boldsymbol{\theta}) \leq \sigma_{i-1}^{g}(\boldsymbol{\theta}), \\ v_i(\boldsymbol{\theta})+\sigma_i^{\varepsilon}(\boldsymbol{\theta})-\sigma_{i-1}^{g}(\boldsymbol{\theta}), & \textit{if } \sigma_i^{\varepsilon}(\boldsymbol{\theta}) > \sigma_{i-1}^{g}(\boldsymbol{\theta}), \end{cases}$$

where $\mathbf{u}_i=(u_{1i},u_{2i})^T$, $u_{ji}\sim\mathcal{U}(0,1)$, $j=1,2$, u_{1i} and u_{2i} are independent, $v_i(\boldsymbol{\theta}) \overset{d}{=} \theta_1+(\theta_2-\theta_1)u_{1i}$, $\sigma_i^{\varepsilon}(\boldsymbol{\theta})=\sum_{j=1}^{i}\varepsilon_j(\boldsymbol{\theta})$, $\varepsilon_j(\boldsymbol{\theta})=-\theta_3^{-1}\log(u_{2j})$ and $\sigma_i^{g}=\sum_{j=1}^{i}g(\boldsymbol{\theta},\mathbf{u}_j)$. Let E_i corresponds to the event $\{\sigma_i^{\varepsilon}(\boldsymbol{\theta})\leq\sigma_{i-1}^{g}(\boldsymbol{\theta})\}$ and $\bar{E}_i$ be the contrary. The partial derivatives may be found recursively as follows:

$$\frac{\partial}{\partial\theta_1}g(\boldsymbol{\theta},\mathbf{u}_i) = \begin{cases} 1-u_{1i}, & \textit{if } i=1, \\ 1-u_{1i}, & \textit{if } i>1 \textit{ and } E_i, \\ 1-u_{1i}-\sum_{j=1}^{i-1}\frac{\partial}{\partial\theta_1}g(\boldsymbol{\theta},\mathbf{u}_j), & \textit{if } i>1 \textit{ and } \bar{E}_i. \end{cases}$$

$$\frac{\partial}{\partial\theta_2}g(\boldsymbol{\theta},\mathbf{u}_i) = \begin{cases} u_{1i}, & \textit{if } i=1, \\ u_{1i}, & \textit{if } i>1 \textit{ and } E_i, \\ u_{1i}-\sum_{j=1}^{i-1}\frac{\partial}{\partial\theta_2}g(\boldsymbol{\theta},\mathbf{u}_j), & \textit{if } i>1 \textit{ and } \bar{E}_i. \end{cases}$$

$$\frac{\partial}{\partial\theta_3}g(\boldsymbol{\theta},\mathbf{u}_i) = \begin{cases} 0, & \textit{if } i=1, \\ 0, & \textit{if } i>1 \textit{ and } E_i, \\ -\frac{1}{\theta_3^2}\sum_{j=1}^{i}\log(u_{2j})-\sum_{j=1}^{i-1}\frac{\partial}{\partial\theta_3}g(\boldsymbol{\theta},\mathbf{u}_j), & \textit{if } i>1 \textit{ and } \bar{E}_i. \end{cases}$$

$$\frac{\partial}{\partial u_1}g(\boldsymbol{\theta},\mathbf{u}_i) = \begin{cases} \theta_2-\theta_1, & \textit{if } i=1, \\ \theta_2-\theta_1, & \textit{if } i>1 \textit{ and } E_i, \\ \theta_2-\theta_1-\sum_{j=1}^{i-1}\frac{\partial}{\partial u_1}g(\boldsymbol{\theta},\mathbf{u}_j), & \textit{if } i>1 \textit{ and } \bar{E}_i. \end{cases}$$

$$\frac{\partial}{\partial u_2}g(\boldsymbol{\theta},\mathbf{u}_i) = \begin{cases} 0, & \textit{if } i=1, \\ 0, & \textit{if } i>1 \textit{ and } E_i, \\ -\theta_3^{-1}\sum_{j=1}^{i}\frac{1}{u_{2j}}-\sum_{j=1}^{i-1}\frac{\partial}{\partial u_2}g(\boldsymbol{\theta},\mathbf{u}_j), & \textit{if } i>1 \textit{ and } \bar{E}_i. \end{cases}$$

	SwiZs	indirect inference	parametric bootstrap
Average time [*seconds*]	0.97	134.18	197.15
Total time [*hours*]	2.7	372.5	547.4

Table 1.3: Average time in seconds to estimate a conditional distribution on $S = 10,000$ points and total time in hours for the $M = 10,000$ independent trials.

Clearly g is once continuously differentiable in both its arguments with non-zero derivatives. Eventually, we have that $v_i(\boldsymbol{\theta})$ goes to ∞ when $\theta_1 \to \infty$, $\theta_2 \to \infty$ and $u_{1i} \to 1$, whereas $\varepsilon_i(\boldsymbol{\theta})$ tends to zero whenever $\theta_3 \to \infty$ and $u_{2i} \to 1$. It is not clear whether $v_i(\boldsymbol{\theta}) + \sigma_i^{\varepsilon}(\boldsymbol{\theta}) - \sigma_i^{g}(\boldsymbol{\theta})$ diverges or converges to 0 when $\|(\boldsymbol{\theta}, \mathbf{u}_i)\| \to \infty$, but in any case $\|g(\boldsymbol{\theta}, \mathbf{u}_i)\|$ tends to ∞ since $v_i(\boldsymbol{\theta})$ diverges. As a consequence, Assumption 1.36 is highly plausible and thus Theorem 1.38 seems invokable.

For the simulation, we set $\boldsymbol{\theta}_0 = [0.3\ 0.9\ 1]^T$ and $n = 100$ as in [HF04]. We compare the SwiZs with indirect inference in Definition 1.4 and the parametric bootstrap using the indirect inference with $B = 1$ as the initial consistent estimator (see Definition 1.9). By Theorem 1.8, the SwiZs and the indirect inference are equivalent, but as argued, the price for obtaining the inidirect inference is higher so here we seek empirical evidence, and Table 1.3 speaks for itself, the difference is indeed monstrous. The parametric bootstrap is even worse in terms of computational time. It is maybe good to remind the reader that the comparison is fair: all three methods benefits from the same level of implementation and uses the very same technology. The complete results may be found in Appendix 1.D.3.

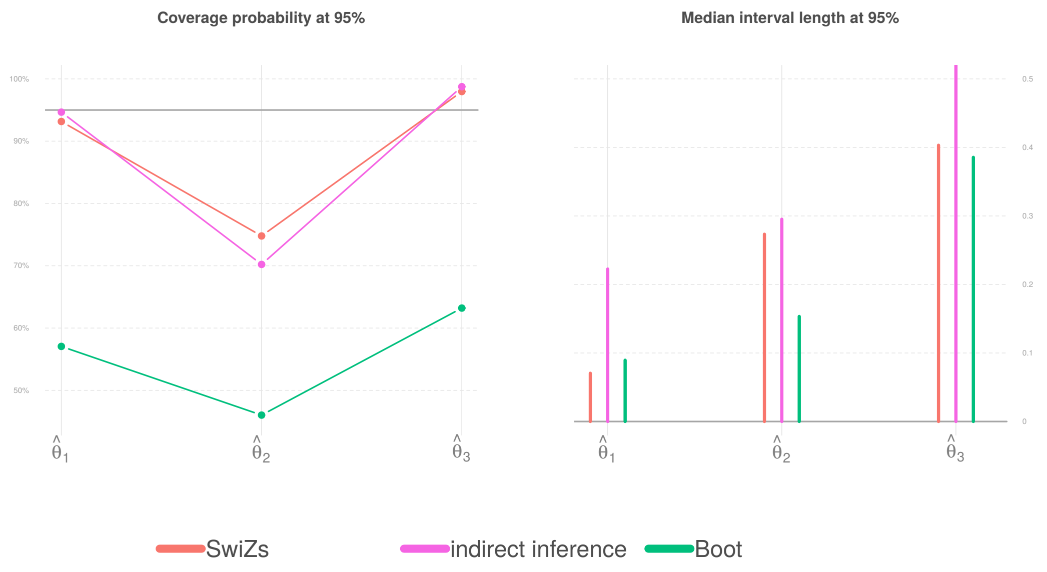

Figure 1.10: *On the left panel*: Representation of the 95% coverage probability (ideal is gray line) of the SwiZs, the indirect inference and the parametric bootstrap with indirect inference as initial estimator. The closer to the gray line is the better. *On the right panel*: Illustration of the median interval lengths at a target level of 95%. The shorter is the better.

In Figure 1.10 we can realize that the SwiZs do not offer an exact coverage in this case,

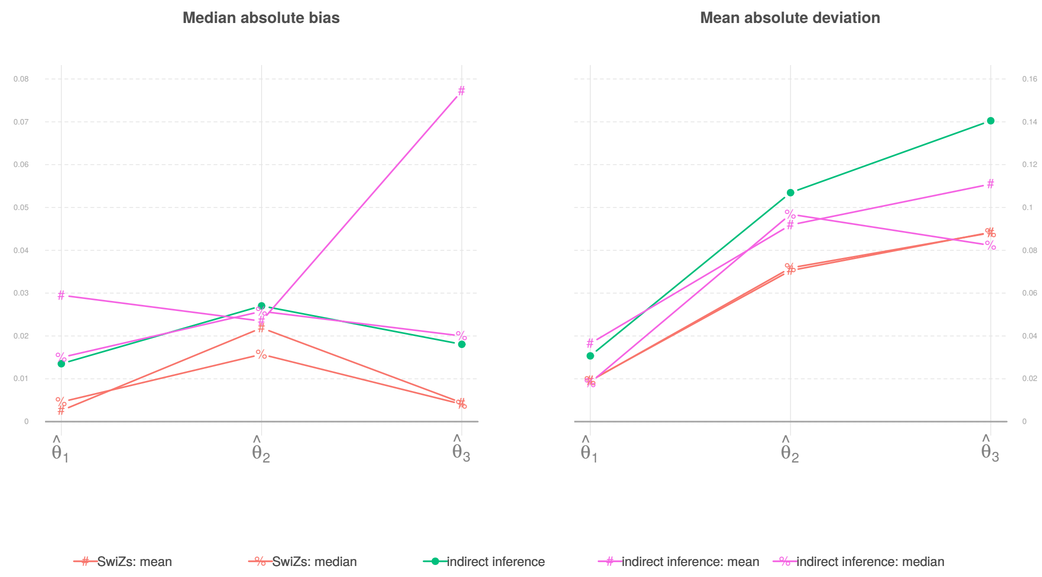

Figure 1.11: *On the left panel*: Median absolute bias of point estimators: mean and median on the SwiZs and indirect inference distributions plus the indirect inference with $B = 1$. *On the right panel*: same as *left panel* with a different measure: mean absolute deviation. For both panel, the lower is the better.

it is even far from ideal for $\hat{\theta}_2$. It is nonetheless better than the parametric bootstrap. Especially the coverage of $\hat{\theta}_1$ and $\hat{\theta}_3$ are close to the ideal level. Considering the context of this simulation: moderate sample size, no closed-form for the likelihood, the results are very encouraging. A good surprise appears from Figure 1.11 where the SwiZs demonstrates better performances of its point estimates (mean and median) compared to indirect inference approaches in termes of absolute median bias and mean absolute deviation.

It is however not clear which one, if not both, we should blame for failure of missing exact coverage probability between our analysis on the applicability of Theorem 1.38 to this case or the numerical optimization procedure. The previous examples seem to indicate for the latter. To this end, we re-run the same experiment only for the SwiZs (for pure operational reason) by changing the starting values to be the true parameter $\boldsymbol{\theta}_0$ to measure the implication. Indeed, starting values are a sensitive matter for quasi-Newton routine and since $\hat{\boldsymbol{\pi}}_n$ is not a consistent estimator of $\boldsymbol{\theta}_0$, using it as a starting value might have a persistent influence on the sequence $\{\hat{\boldsymbol{\theta}}_n^{(s)} : s \in \mathbb{N}_S^+\}$. Results are reported in Table in Appendix 1.D.3. The coverage probabilities of $\hat{\theta}_1$ and $\hat{\theta}_3$ becomes nearly perfect, which shows that indeed good starting values may reduce the numerical error in the coverage probabilities. However, coverage probability for $\hat{\theta}_2$ persistently shows result off the desired levels, which seems rather to indicate a problem related to the applicability of Theorem 1.38. Increasing the sample size to $n = 1,000$ (see Table 1.19) makes the coverage of all three parameters nearly perfect.

Fifth and last, we consider logistic regression. This is certainly one of the most widely used statistical model in practice. This case is challenging at least on two aspects. First, the random variable is discrete and the finite sample theory in Section 1.4 does not hold. Second, the generating function is non-differentiable with respect to $\boldsymbol{\theta}$, therefore gradient-based optimization routines cannot be employed. In what follows, we circumvent this inconvenient by smoothing the generating function. To this end, we start by introducing the continuous latent representation of the logistic regression.

Example 1.64. *Suppose we have the model*

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$ and ϵ_i, $i = 1, \cdots, n$, are identically and independently distributed according to a logistic distribution with mean 0 and unity variance. This distribution belongs to symmetric location-scale families. It is similar to the Gaussian distribution with heavier tails. The unknwon parameters $\boldsymbol{\theta}$ of this model could be easily estimated by the ordinary least squares:

$$\hat{\boldsymbol{\pi}}_n = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}.$$

The corresponding estimating function is:

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\pi}\right) = \mathbf{X}^T\mathbf{X}\boldsymbol{\pi} - \mathbf{X}^T\mathbf{g}\left(\boldsymbol{\theta}, \mathbf{u}\right).$$

A straightforward generating function is $\mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) = \mathbf{X}\boldsymbol{\theta} + \mathbf{u}$ where $u_i \sim \text{Logistic}(0, 1)$. Evaluating this function at $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}_n$ leads to

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}, \hat{\boldsymbol{\pi}}_n\right) = \mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{u}.$$

Solving the root of this function in $\boldsymbol{\theta}$ gives the following explicit solution:

$$\hat{\boldsymbol{\theta}}_n = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\boldsymbol{y} - \mathbf{u}\right). \tag{1.7}$$

Following Example 1.56 on linear regression, it is easy to show that inference based on the distribution of this estimator leads to exact frequentist coverage probabilities.

Let us turn our attention to logistic regression. In this case, $\boldsymbol{y}$ is not observed. Instead, we observe a binary random variable $\mathbf{y}$, whose elements are:

$$y_i = \begin{cases} 1, & \mathbf{X}_i\boldsymbol{\theta} + \epsilon_i \geq 0, \\ 0, & \mathbf{X}_i\boldsymbol{\theta} + \epsilon_i < 0, \end{cases}$$

where $\mathbf{X}_i$ is the ith row of $\mathbf{X}$. Saying it differently, this consideration implies that the generating function is modified to the following indicator function:

$$\mathbf{g}\left(\boldsymbol{\theta}, u_i\right) = \mathbf{1}\left\{\mathbf{X}_i\boldsymbol{\theta} + u_i \geq 0\right\}.$$

Clearly, this change implies that $\boldsymbol{\Phi}_n$ has a flat Jacobian matrix and Assumptions 1.36 and 1.37 do not hold*. Moreover, this problem becomes numerically more invloved, especially if we want to pursue with a gradient-based optimization routine. As mentionned, in practice we seek the solution of the following problem:*

$$\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left\|\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{g}(\boldsymbol{\theta}, \mathbf{u})\right\|_2^2 \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} f(\boldsymbol{\theta}). \tag{1.8}$$

Note that $\mathbf{X}^T\mathbf{y}$ is the sufficient statistic for a logistic regression (see Chapter 2 in [MN89]). The gradient of $f(\boldsymbol{\theta})$ is

$$-D_{\boldsymbol{\theta}}\mathbf{g}(\boldsymbol{\theta}, \mathbf{u})\mathbf{X}\left[\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{g}(\boldsymbol{\theta}, \mathbf{u})\right].$$

However, the Jacobian $D_{\boldsymbol{\theta}}\mathbf{g}(\boldsymbol{\theta},\mathbf{u})$ is 0 almost everywhere and alternatives are necessary for using gradient-based methods. A possibility is to smooth $\mathbf{g}(\boldsymbol{\theta},\mathbf{u})$ by using for example a sigmoid function:

$$\mathbf{g}(\boldsymbol{\theta},u_i)=\lim_{t\to0}\frac{1}{1+\exp\left(-{}^{(\mathbf{x}_i\boldsymbol{\theta}+u_i)}/_{t}\right)}.$$

The value of t tunes the approximation and the value of the gradient. However, from our experience, large values of t, say $t>0.1$, leads to poor results and small values, say $t<0.1$, leads to numerical instability. We thus prefer to use a different strategy by taking $-f(\boldsymbol{\theta})$ as the gardient. This strategy corresponds to the iterative bootstrap procedure ([Gue+18b]); see Section 2.4. In Figure 1.12, we illustrate the difference between these two approximations and the "ideal" distribution we would have obtained by observing the continuous underlying latent process. Clearly, the loss of information induced from the possibility of only observing a binary outcome results in an increase of variability. Nonetheless, the difference is not enormous. Both approximations leads to similar distributions in terms of shapes. We can notice a little difference in their modes. Since the iterative bootstrap approximation is numerically advantageous, we use it in the next study.

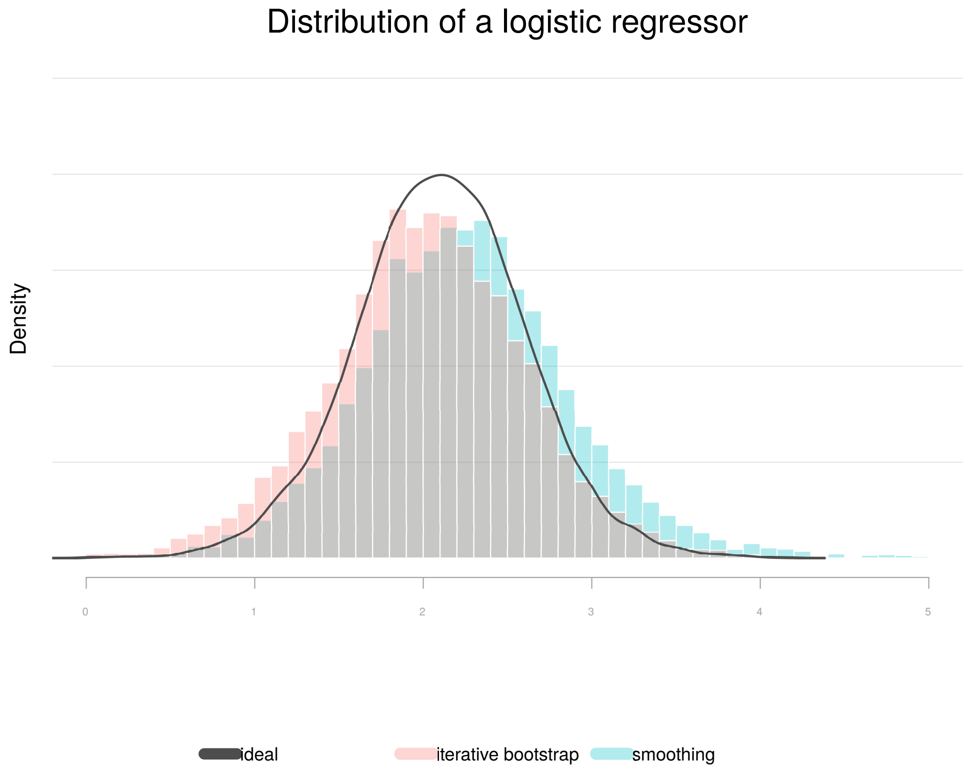

Figure 1.12: Simulated SwisZ distribution of a single logistic regression with coefficient $\boldsymbol{\theta}=2$ and sample size of 10. "Ideal" is (1.7). "Smoothing" approximates the gradient with a sigmoid function and $t=0.01$. "Iterative bootstrap" uses $-f(\boldsymbol{\theta})$ as the gradient.

For simulation, we setup $\boldsymbol{\theta}_0=(0,5,5,-7,-7,\underbrace{0,\dots,0}_{15})^T$ and sample size $n=200$. We compare coverage probabilities of 95% confidence intervals obtained by the SwiZs and by

	SwiZs	asymptotic
θ_1	0.9442	0.9187
θ_2	0.9398	0.8115
θ_3	0.9382	0.8121
θ_4	0.9432	0.7688
θ_5	0.9450	0.7737
θ_6	0.9397	0.9233
θ_7	0.9357	0.9170
θ_8	0.9398	0.9237
θ_9	0.9391	0.9218
θ_{10}	0.9400	0.9208
θ_{11}	0.9424	0.9208
θ_{12}	0.9375	0.9214
θ_{13}	0.9368	0.9204
θ_{14}	0.9389	0.9210
θ_{15}	0.9400	0.9207
θ_{16}	0.9400	0.9183
θ_{17}	0.9361	0.9183
θ_{18}	0.9449	0.9241
θ_{19}	0.9412	0.9218
θ_{20}	0.9427	0.9240

Table 1.4: 95% coverage probabilities of confidence intervals from the SwiZs and asymptotic theory.

asymptotic theory. We report results in Table 1.4. We can clearly see that the SwiZs have the most precise confidence intervals for all coefficients with coverage close to the target level of 95%.

References for Chapter 1

[And88] Donald WK Andrews. “Laws of large numbers for dependent non-identically distributed random variables”. In: *Econometric theory* 4.3 (1988), pp. 458–467.

[Bai94] Ralph W Bailey. “Polar generation of random variates with the t-distribution”. In: *Mathematics of Computation* 62.206 (1994), pp. 779–781.

[BCS07] Paola Bortot, Stuart G Coles, and Scott A Sisson. “Inference for stereological extremes”. In: *Journal of the American Statistical Association* 102.477 (2007), pp. 84–92.

[BE13] Douglas Bates and Dirk Eddelbuettel. “Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package”. In: *Journal of Statistical Software* 52.5 (2013), pp. 1–24. URL: http://www.jstatsoft.org/v52/i05/.

[Bea+09] Mark A Beaumont et al. “Adaptive approximate Bayesian computation”. In: *Biometrika* 96.4 (2009), pp. 983–990.

[Bea10] Mark A Beaumont. “Approximate Bayesian computation in evolution and ecology”. In: *Annual review of ecology, evolution, and systematics* 41 (2010), pp. 379–406.

[BF10] Michael GB Blum and Olivier François. “Non-linear regression models for Approximate Bayesian Computation”. In: *Statistics and Computing* 20.1 (2010), pp. 63–73.

[Bil12] Patrick Billingsley. *Probability and Measure*. Vol. 939. John Wiley & Sons, 2012.

[Boo18] Boost. *Boost C++ Libraries*. http://www.boost.org/. Last accessed 2018-06-03. 2018.

[BSZ98] Laurence Broze, Olivier Scaillet, and Jean-Michel Zakoian. “Quasi-indirect inference for diffusion processes”. In: *Econometric Theory* 14.2 (1998), pp. 161–186.

[BT74] Albert E Beaton and John W Tukey. “The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data”. In: *Technometrics* 16.2 (1974), pp. 147–185.

[BZB02] Mark A Beaumont, Wenyang Zhang, and David J Balding. “Approximate Bayesian computation in population genetics”. In: *Genetics* 162.4 (2002), pp. 2025–2035.

[Cai05] T Tony Cai. “One-sided confidence intervals in discrete distributions”. In: *Journal of Statistical planning and inference* 131.1 (2005), pp. 63–88.

[Cor+08] Jean-Marie Cornuet et al. “Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation”. In: *Bioinformatics* 24.23 (2008), pp. 2713–2719.

[Cri17] Mihai Cristea. “On global implicit function theorem”. In: *Journal of Mathematical Analysis and Applications* 456.2 (2017), pp. 1290–1302.

[Dem08] Arthur P Dempster. “The dempster-shafer calculus for statisticians.” In: *International Journal of approximate reasoning* 48.2 (2008), pp. 365–377.

[Dev86] Luc Devroye. *Non-uniform random variate generation*. Springer-Verlag, New York, 1986.

[DG84] Peter J Diggle and Richard J Gratton. "Monte Carlo methods of inference for implicit statistical models". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1984), pp. 193–227.

[DGR07] Ramdan Dridi, Alain Guay, and Eric Renault. "Indirect inference and calibration of dynamic stochastic general equilibrium models". In: *Journal of Econometrics* 136.2 (2007), pp. 397–430.

[DM02] Debbie J Dupuis and Stephan Morgenthaler. "Robust weighted likelihood estimators with an application to bivariate extreme value problems". In: *Canadian Journal of Statistics* 30.1 (2002), pp. 17–36.

[DMDJ06] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. "Sequential monte carlo samplers". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.3 (2006), pp. 411–436.

[DPL15] Christopher C Drovandi, Anthony N Pettitt, and Anthony Lee. "Bayesian indirect inference using a parametric auxiliary model". In: *Statistical Science* 30.1 (2015), pp. 72–95.

[DSZ73] A Philip Dawid, Mervyn Stone, and James V Zidek. "Marginalization paradoxes in Bayesian and structural inference". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1973), pp. 189–233.

[EB17] Dirk Eddelbuettel and James Joseph Balamuta. "Extending *R* with *C++*: A Brief Introduction to *Rcpp*". In: *PeerJ Preprints* 5 (2017), e3188v1. ISSN: 2167-9843. URL: https://doi.org/10.7287/peerj.preprints.3188v1.

[EEK16] Dirk Eddelbuettel, John W. Emerson, and Michael J. Kane. *BH: Boost C++ Header Files*. R package version 1.62.0-1. 2016. URL: https://CRAN.R-project.org/package=

[Efr79] B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26.

[ES14] Dirk Eddelbuettel and Conrad Sanderson. "RcppArmadillo: Accelerating R with high-performance C++ linear algebra". In: *Computational Statistics and Data Analysis* 71 (2014), pp. 1054–1063. URL: http://dx.doi.org/10.1016/j.csda.2013.02.

[ET94] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[Fen+13] Changyong Feng et al. "The Mean Value Theorem and Taylor's Expansion in Statistics". In: *The American Statistician* 67.4 (2013), pp. 245–248.

[FFS10] Ailana M Fraser, Donald AS Fraser, and Ana-Maria Staicu. "Second order ancillary: A differential view from continuity". In: *Bernoulli* (2010), pp. 1208–1223.

[Fis22] R.A. Fisher. "On the mathematical foundations of theoretical statistics". In: *Phil. Trans. R. Soc. Lond. A* 222.594-604 (1922), pp. 309–368.

[Fis30] R.A. Fisher. "Inverse probability". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 26. 4. Cambridge University Press. 1930, pp. 528–535.

[Fis33] R.A. Fisher. “The concepts of inverse probability and fiducial probability referring to unknown parameters”. In: *Proc. R. Soc. Lond. A* 139.838 (1933), pp. 343–348.

[Fis35] R.A. Fisher. “The fiducial argument in statistical inference”. In: *Annals of eugenics* 6.4 (1935), pp. 391–398.

[Fis56] R.A. Fisher. *Statistical methods and scientific inference.* Oxford, England: Hafner Publishing Co., 1956.

[FP12] Paul Fearnhead and Dennis Prangle. “Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 419–474.

[Fra+10] DAS Fraser et al. “Default priors for Bayesian and frequentist inference”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.5 (2010), pp. 631–654.

[Fra+18] David T Frazier et al. “Asymptotic properties of approximate Bayesian computation”. In: *Biometrika* 105.3 (2018), pp. 593–607.

[Fra68] D.A.S. Fraser. *The structure of inference.* Wiley, New York, 1968.

[FS94] C Field and B Smith. “Robust estimation: A weighted maximum likelihood approach”. In: *International Statistical Review/Revue Internationale de Statistique* (1994), pp. 405–424.

[GFG13] David E Giles, Hui Feng, and Ryan T Godwin. “On the bias of the maximum likelihood estimator for the two-parameter Lomax distribution”. In: *Communications in Statistics-Theory and Methods* 42.11 (2013), pp. 1934–1950.

[GJ+10] Gaël Guennebaud, Benoît Jacob, et al. *Eigen v3.* http://eigen.tuxfamily.org. 2010.

[GM96] Christian Gourieroux and Alain Monfort. *Simulation-based econometric methods.* Oxford university press, 1996.

[GMR93] Christian Gourieroux, Alain Monfort, and Eric Renault. “Indirect inference”. In: *Journal of applied econometrics* 8.S1 (1993).

[GPY10] Christian Gouriéroux, Peter CB Phillips, and Jun Yu. “Indirect inference for dynamic panel models”. In: *Journal of Econometrics* 157.1 (2010), pp. 68–77.

[GRV11] René Garcia, Eric Renault, and David Veredas. “Estimation of stable distributions by indirect inference”. In: *Journal of Econometrics* 161.2 (2011), pp. 325–337.

[GT10] A Ronald Gallant and George Tauchen. “Simulated score methods and indirect inference for continuous-time models”. In: *Handbook of financial econometrics* 1 (2010), pp. 427–477.

[GT96] A Ronald Gallant and George Tauchen. “Which moments to match?” In: *Econometric Theory* 12.4 (1996), pp. 657–681.

[Gue+18a] Stéphane Guerrier et al. “On the Properties of Simulation-based Estimators in High Dimensions”. In: *arXiv preprint arXiv:1810.04443* (2018).

[Gue+18b] Stéphane Guerrier et al. "Simulation-Based Bias Correction Methods for Complex Models". In: *Journal of the American Statistical Association* (2018), pp. 1–12.

[Hal92] Peter Hall. *The bootstrap and edgeworth expansion.* Springer-Verlag, New York, 1992.

[Ham+11] Frank R Hampel et al. *Robust statistics: the approach based on influence functions.* Vol. 196. John Wiley & Sons, 2011.

[Ham94] James Douglas Hamilton. *Time series analysis.* Vol. 2. Princeton university press Princeton, NJ, 1994.

[Han+16] Jan Hannig et al. "Generalized fiducial inference: A review and new results". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 1346–1361.

[Han09] Jan Hannig. "On generalized fiducial inference". In: *Statistica Sinica* (2009), pp. 491–544.

[Han13] Jan Hannig. "Generalized fiducial inference via discretization". In: *Statistica Sinica* (2013), pp. 489–514.

[HF04] Knut Heggland and Arnoldo Frigessi. "Estimating functions in indirect inference". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.2 (2004), pp. 447–462.

[HLL14] Jan Hannig, Randy CS Lai, and Thomas CM Lee. "Computational issues of generalized fiducial inference". In: *Computational Statistics & Data Analysis* 71 (2014), pp. 849–858.

[HMC05] Robert V Hogg, Joseph McKean, and Allen T Craig. *Introduction to mathematical statistics.* Pearson Education, 2005.

[Hub+64] Peter J Huber et al. "Robust estimation of a location parameter". In: *The annals of mathematical statistics* 35.1 (1964), pp. 73–101.

[Hub67] Peter J Huber. "The behavior of maximum likelihood estimates under nonstandard conditions". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.* Vol. 1. 1. University of California Press. 1967, pp. 221–233.

[Jia07] Jiming Jiang. *Linear and generalized linear mixed models and their applications.* Springer Science & Business Media, 2007.

[JKB94] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions.* 2nd. Vol. 1. John Wiley & Sons, Inc., 1994.

[JT04] Wenxin Jiang and Bruce Turnbull. "The indirect method: inference based on intermediate statistics—a synthesis and examples". In: *Statistical Science* 19.2 (2004), pp. 239–263.

[KK03] Christian Kleiber and Samuel Kotz. *Statistical size distributions in economics and actuarial sciences.* Vol. 470. John Wiley & Sons, 2003.

[Koc07] Karl-Rudolf Koch. *Introduction to Bayesian statistics.* Springer Science & Business Media, 2007.

[KS61] Maurice Kendall and Alan Stuart. *The advanced theory of statistics*. 3rd. Vol. 2nd: Inference and relationship. Charles Griffin & Company Limited, 1961.

[KW94] Robert E Kass and Larry Wasserman. "Formal rules for selecting prior distributions: A review and annotated bibliography". In: *Journal of the American Statistical Association* (1994).

[LC09] Marco J Lombardi and Giorgio Calzolari. "Indirect estimation of α-stable stochastic volatility models". In: *Computational Statistics & Data Analysis* 53.6 (2009), pp. 2298–2308.

[Lin22] Jarl Waldemar Lindeberg. "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung". In: *Mathematische Zeitschrift* 15.1 (1922), pp. 211–225.

[Lin58] Dennis V Lindley. "Fiducial distributions and Bayes' theorem". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1958), pp. 102–107.

[LN89] Dong C Liu and Jorge Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3 (1989), pp. 503–528.

[Lom54] KS Lomax. "Business failures: Another example of the analysis of failure data". In: *Journal of the American Statistical Association* 49.268 (1954), pp. 847–852.

[Mar+03] Paul Marjoram et al. "Markov chain Monte Carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15324–15328.

[Mar+12] Jean-Michel Marin et al. "Approximate Bayesian computational methods". In: *Statistics and Computing* 22.6 (2012), pp. 1167–1180.

[Mar15] Ryan Martin. "Plausibility functions and exact frequentist inference". In: *Journal of the American Statistical Association* 110.512 (2015), pp. 1552–1561.

[ML13] Ryan Martin and Chuanhai Liu. "Inferential models: A framework for prior-free posterior probabilistic inference". In: *Journal of the American Statistical Association* 108.501 (2013), pp. 301–313.

[ML15] Ryan Martin and Chuanhai Liu. "Conditional inferential models: combining information for prior-free probabilistic inference". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.1 (2015), pp. 195–217.

[MN89] Peter McCullagh and John A Nelder. *Generalized linear models*. Vol. 37. CRC press, 1989.

[Mon98] Chiara Monfardini. "Estimating stochastic volatility models through indirect inference". In: *The Econometrics Journal* 1.1 (1998), pp. 113–128.

[MP92] Arakaparampil M Mathai and Serge B Provost. *Quadratic forms in random variables: theory and applications*. Dekker, 1992.

[MVF06] Irini Moustaki and Maria-Pia Victoria-Feser. "Bounded-influence robust estimation in generalized linear latent variable models". In: *Journal of the American Statistical Association* 101.474 (2006), pp. 644–653.

[NM94] Whitney K Newey and Daniel McFadden. "Large sample estimation and hypothesis testing". In: *Handbook of econometrics* 4 (1994), pp. 2111–2245.

[Noc80] Jorge Nocedal. "Updating quasi-Newton matrices with limited storage". In: *Mathematics of computation* 35.151 (1980), pp. 773–782.

[NP33] Jerzy Neyman and Egon S Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Phil. Trans. R. Soc. Lond. A* 231.694-706 (1933), pp. 289–337.

[NW06] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. 2nd. Springer, 2006.

[ON10] Naoaki Okazaki and J Nocedal. *libLBFGS: a library of Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)*. 2010. URL: http://www.chokkan.org/softwar

[Pal59] Richard S Palais. "Natural operations on differential forms". In: *Transactions of the American Mathematical Society* 92.1 (1959), pp. 125–141.

[PP94] Benedikt M Pötscher and Ingmar R Prucha. "Generic uniform convergence and equicontinuity concepts for random functions: An exploration of the basic structure". In: *Journal of Econometrics* 60.1-2 (1994), pp. 23–63.

[Pri+99] Jonathan K Pritchard et al. "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." In: *Molecular biology and evolution* 16.12 (1999), pp. 1791–1798.

[PY09] Peter CB Phillips and Jun Yu. "Simulation-based estimation of contingent-claims prices". In: *The Review of Financial Studies* 22.9 (2009), pp. 3669–3705.

[Qiu+18] Yixuan Qiu et al. *RcppNumerical: 'Rcpp' Integration for Numerical Computing Libraries*. R package version 0.3-2. 2018. URL: https://CRAN.R-project.org/package=Rcp

[Rob07] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[RS16] Dirk Eddelbuettel extending DEoptim which itself is based on DE-Engine (by Rainer Storn). *RcppDE: Global Optimization by Differential Evolution in C++*. R package version 0.1.5. 2016. URL: https://CRAN.R-project.org/package=RcppDE.

[SC16] Conrad Sanderson and Ryan Curtin. "Armadillo: a template-based C++ library for linear algebra". In: *Journal of Open Source Software* (2016).

[SFT07] Scott A Sisson, Yanan Fan, and Mark M Tanaka. "Sequential monte carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 104.6 (2007), pp. 1760–1765.

[SH02] Tore Schweder and Nils Lid Hjort. "Confidence and likelihood". In: *Scandinavian Journal of Statistics* 29.2 (2002), pp. 309–332.

[Sha76] Glenn Shafer. *A mathematical theory of evidence*. Vol. 42. Princeton university press, 1976.

[Sho+18] John F Shortle et al. *Fundamentals of queueing theory*. Vol. 399. John Wiley & Sons, 2018.

[Sis+10] SA Sisson et al. “A note on target distribution ambiguity of likelihood-free samplers”. In: *arXiv preprint arXiv:1005.5201* (2010).

[Smi93] Anthony A Smith. “Estimating nonlinear time-series models using simulated vector autoregressions”. In: *Journal of Applied Econometrics* 8.S1 (1993).

[SP97] Rainer Storn and Kenneth Price. “Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces”. In: *Journal of global optimization* 11.4 (1997), pp. 341–359.

[Stu08] Student. “The probable error of a mean”. In: *Biometrika* (1908), pp. 1–25.

[SXS+05] Kesar Singh, Minge Xie, William E Strawderman, et al. “Combining information from independent sources through confidence distributions”. In: *The Annals of Statistics* 33.1 (2005), pp. 159–183.

[Sär13] Simo Särkkä. *Bayesian filtering and smoothing.* Vol. 3. Cambridge University Press, 2013.

[Tav+97] Simon Tavaré et al. “Inferring coalescence times from DNA sequence data”. In: *Genetics* 145.2 (1997), pp. 505–518.

[Ton+09] Tina Toni et al. “Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems”. In: *Journal of the Royal Society Interface* 6.31 (2009), pp. 187–202.

[Vaa98] Aad W Van der Vaart. *Asymptotic statistics.* Vol. 3. Cambridge university press, 1998.

[VFR94b] Maria-Pia Victoria-Feser and Elvezio Ronchetti. “Robust methods for personal-income distribution models”. In: *Canadian Journal of Statistics* 22.2 (1994), pp. 247–258.

[VM15] Piero Veronese and Eugenio Melilli. “Fiducial and confidence distributions for real exponential families”. In: *Scandinavian Journal of Statistics* 42.2 (2015), pp. 471–484.

[Wil+10] Richard D Wilkinson et al. “Dating primate divergences through an integrated analysis of palaeontological and molecular data”. In: *Systematic Biology* 60.1 (2010), pp. 16–31.

[Wu11] Wei Biao Wu. “Asymptotic theory for stationary processes”. In: *Statistics and its Interface* 4.2 (2011), pp. 207–226.

[XS13] Min-ge Xie and Kesar Singh. “Confidence distribution, the frequentist distribution estimator of a parameter: A review”. In: *International Statistical Review* 81.1 (2013), pp. 3–39.

[XSS11] Minge Xie, Kesar Singh, and William E Strawderman. “Confidence distributions and a unifying framework for meta-analysis”. In: *Journal of the American Statistical Association* 106.493 (2011), pp. 320–333.

[Zab92] Sandy L Zabell. “RA Fisher and fiducial argument”. In: *Statistical Science* 7.3 (1992), pp. 369–387.

[ZL11] Jianchun Zhang and Chuanhai Liu. “Dempster-Shafer inference with weak beliefs”. In: *Statistica Sinica* (2011), pp. 475–494.

[R C17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

Appendix

1.A Technical results

Lemma 1.A.1. *Let X and Y be open subsets of $\mathbb{R}^n$. If $\mathbf{f} : X \to Y$ is a C^1-diffeomorphism, then the Jacobian matrices of the maps $x \mapsto \mathbf{f}$ and $y \mapsto \mathbf{f}^{-1}$ are invertible, and the derivatives at the points $a \in X$ and $b \in Y$, are given by:*

$$D_x\mathbf{f}(a) = \left[D_y\mathbf{f}^{-1}|_{y=\mathbf{f}(a)}\right]^{-1}, \quad D_y\mathbf{f}(b) = \left[D_x\mathbf{f}|_{x=\mathbf{f}^{-1}(b)}\right]^{-1}.$$

Proof. By assumption, $\mathbf{f}$ is invertible, once continuously differentiable and $\mathbf{f}^{-1}$ is once continuously differentiable.

We have $\mathbf{f}^{-1} \circ \mathbf{f} = \mathrm{id}_X$, where id_X is the identity function on the set X. Fix $a \in X$. By the chain rule, the derivative at a is the following:

$$D_y\mathbf{f}^{-1}\left(\mathbf{f}(a)\right) D_x\mathbf{f}(a) = \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the identity matrix. Since $D_y\mathbf{f}^{-1}$ and $D_x\mathbf{f}$ are square matrices, we have:

$$\det\left(D_y\mathbf{f}^{-1}(\mathbf{f}(a))\right) \det\left(D_x\mathbf{f}(a)\right) = 1.$$

The determinants cannot be 0, there are either 1 or -1 for both matrices, ergo, the Jacobian are invertible and we can write

$$D_x\mathbf{f}(a) = \left[D_y\mathbf{f}^{-1}(\mathbf{f}(a))\right]^{-1}.$$

The proof for $\mathbf{f} \circ \mathbf{f}^{-1} = \mathrm{id}_Y$ follows by symmetry. □

Lemma 1.A.2. *Let $\mathbf{\Theta}$ and W be open subsets of $\mathbb{R}^p$. If there exists a C^1-diffeomorphic mapping $\mathbf{a} : W \to \mathbf{\Theta}$, that is, $\mathbf{w} \mapsto \mathbf{a}$ is continuously once differentialbe in $\mathbf{\Theta} \times W$ and the inverse map $\boldsymbol{\theta} \mapsto \mathbf{a}^{-1}$ is continuously once differentiable in $\mathbf{\Theta} \times W$, then the cumulative distribution function of $\{\hat{\boldsymbol{\theta}}_n^{(s)} : s \in \mathbb{N}\}$ is given by:*

$$\int_{\mathbf{\Theta}_n} f_{\hat{\boldsymbol{\theta}}_n}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\pi}}_n\right) \mathrm{d}\boldsymbol{\theta} = \int_W f_{\mathbf{w}}\left(\mathbf{a}(\mathbf{w})|\hat{\boldsymbol{\pi}}_n\right) \frac{1}{\left|\det\left(D_{\mathbf{w}}\mathbf{a}(\mathbf{w})\right)\right|} \mathrm{d}\mathbf{w},$$

provided that f is a nonnegative Borel function and $\Pr\left(\hat{\boldsymbol{\pi}}_n \neq \emptyset\right) = 1$.

Proof of Lemma 1.A.2. By assumption, $\mathbf{w} \mapsto \mathbf{a}$ is a C^1-diffeomorphism so by Lemma 1.A.1 the Jacobian of $\mathbf{a}$ and $\mathbf{a}^{-1}$ are invertible. All the conditions of the change-of-variable formula for multidimensional Lebesgue integral in [Bil12, Theorem 17.2, p.239] are satisfied, so we obtain

$$\int_{\mathbf{\Theta}_n} f_{\hat{\boldsymbol{\theta}}_n}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\pi}}_n\right) \mathrm{d}\boldsymbol{\theta} = \int_{\mathbf{a}^{-1}(\mathbf{\Theta}_n)} f_{\mathbf{w}}\left(\mathbf{a}^{-1}(\boldsymbol{\theta})|\hat{\boldsymbol{\pi}}_n\right) \det\left(D_{\boldsymbol{\theta}}\mathbf{a}^{-1}(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta}$$

By Lemma 1.A.1, we have that $D_{\boldsymbol{\theta}}\mathbf{a}^{-1} = \left[D_{\mathbf{w}}\mathbf{a}\right]^{-1}$. Taking the determinant ends the proof. □

1.B Finite sample

Proof of Theorem 1.8. We proceed by showing first that $\boldsymbol{\Theta}_{\mathrm{II},n}^{(s)} \subset \boldsymbol{\Theta}_n^{(s)}$, and second that $\boldsymbol{\Theta}_{\mathrm{II},n}^{(s)} \supset \boldsymbol{\Theta}_n^{(s)}$.

It follows from Assumption 1.7 that $\hat{\boldsymbol{\pi}}_n$ is the unique solution of $\operatorname{argzero}_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \boldsymbol{\Phi}_n(\boldsymbol{\theta}_0, \mathbf{u}_0, \boldsymbol{\pi})$, ergo $\boldsymbol{\Pi}_n$ in the Definition 1.2 is a singleton.

(1). Fix $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_{\mathrm{II},n}^{(s)}$. By Definition 1.4, it holds that

$$\hat{\boldsymbol{\pi}}_n = \hat{\boldsymbol{\pi}}_{\mathrm{II},n}^{(s)}\left(\boldsymbol{\theta}_1\right), \quad \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_{\mathrm{II},n}^{(s)}(\boldsymbol{\theta}_1)\right) = \mathbf{0},$$

where $\hat{\boldsymbol{\pi}}_{\mathrm{II},n}^{(s)}$ is the unique solution of $\operatorname{argzero}_{\boldsymbol{\theta}\in\boldsymbol{\Pi}} \boldsymbol{\Phi}_n(\boldsymbol{\theta}_1, \mathbf{u}_s, \boldsymbol{\pi})$. Ergo, it holds as well that

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) = \mathbf{0},$$

implying that $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_n^{(s)}$ by Definition 1.2. Thus $\boldsymbol{\Theta}_{\mathrm{II},n}^{(s)} \subset \boldsymbol{\Theta}_n^{(s)}$.

(2). Fix $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_n$. By Definition 1.2 we have

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_2, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) = \mathbf{0}.$$

By Definition 1.4, we also have

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_2, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_{\mathrm{II},n}^{(s)}(\boldsymbol{\theta}_2)\right) = \mathbf{0},$$

where $\hat{\boldsymbol{\pi}}_{\mathrm{II},n}^{(s)}(\boldsymbol{\theta}_2)$ is the unique solution of $\operatorname{argzero}_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \boldsymbol{\Phi}_n(\boldsymbol{\theta}_2, \mathbf{u}_s, \boldsymbol{\pi})$. It follows that $\hat{\boldsymbol{\pi}}_n = \hat{\boldsymbol{\pi}}_{\mathrm{II},n}^{(s)}\left(\boldsymbol{\theta}_2\right)$ uniquely, implying that $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_{\mathrm{II},n}^{(s)}$ by Definition 1.4. Thus $\boldsymbol{\Theta}_{\mathrm{II},n}^{(s)} \supset \boldsymbol{\Theta}_n^{(s)}$, which concludes the proof. □

Proof of Theorem 1.12. We proceed by showing first that (A) $\boldsymbol{\Theta}_n^{(s)} = \boldsymbol{\Theta}_{\mathrm{Boot},n}^{(s)}$ implies (B) $\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \mathbf{u}_s, \boldsymbol{\pi}) = \boldsymbol{\Phi}_n(\boldsymbol{\pi}, \mathbf{u}_s, \boldsymbol{\theta}) = \mathbf{0}$, then that (B) implies (A).

1. Suppose (A) holds. Fix $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_n^{(s)}$ and $\hat{\boldsymbol{\pi}}_n \in \boldsymbol{\Pi}_n$. We have by the Definition 1.2

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) = \mathbf{0}.$$

By (A), we also have that $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_{\mathrm{Boot},n}^{(s)}$ so by the Definition 1.9

$$\boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\pi}}_n, \mathbf{u}_s, \boldsymbol{\theta}_1\right) = \mathbf{0}.$$

Since both estimating equations equal zero, we have

$$\boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\pi}}_n, \mathbf{u}_s, \boldsymbol{\theta}_1\right) = \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) = \mathbf{0}.$$

Hence (A) implies (B).

2. Suppose now that (B) holds. Fix $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_n^{(s)}$ and $\hat{\boldsymbol{\pi}}_n \in \boldsymbol{\Pi}_n$ so $\boldsymbol{\Phi}_n(\boldsymbol{\theta}_1, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n) = \mathbf{0}$. By (B), we have

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) = \boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\pi}}_n, \mathbf{u}_s, \boldsymbol{\theta}_1\right) = \mathbf{0},$$

so $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_{\mathrm{Boot},n}^{(s)}$ and thus $\boldsymbol{\Theta}_n^{(s)} \subset \boldsymbol{\Theta}_{\mathrm{Boot},n}^{(s)}$. The same argument shows that $\boldsymbol{\Theta}_n^{(s)} \supset \boldsymbol{\Theta}_{\mathrm{Boot},n}^{(s)}$ which ends the proof. □

Proof of Proposition 1.13. Since $\hat{\pi}_n = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n x_i$, the sample average, we can write the following estimating equation

$$\hat{\pi}_n = \underset{\pi\in\Pi}{\text{argzero}}\left(\bar{\mathbf{x}} - \pi\right) = \underset{\pi\in\Pi}{\text{argzero}}\,\Phi_n\left(\theta_0, \mathbf{u}_0, \pi\right),$$

where $x \overset{d}{=} g(\theta_0, u_0)$. Since x follows a location family, we have that $x \overset{d}{=} \theta_0 + g(0, u_0) \overset{d}{=} \theta_0 + y$.

The SwiZs is defined as

$$\hat{\theta}_n^{(s)} = \underset{\theta\in\Theta}{\text{argzero}}\,\Phi_n\left(\theta, \mathbf{u}_s, \hat{\pi}_n\right).$$

On the other hand, the parametric bootstrap estimator is

$$\hat{\theta}_{\text{Boot},n}^{(s)} = \underset{\theta\in\Theta}{\text{argzero}}\,\Phi_n\left(\hat{\pi}_n, \mathbf{u}_s, \theta\right).$$

Eventually, we obtain that

$$\begin{aligned}
\Phi_n\left(\hat{\theta}_n^{(s)}, \mathbf{u}_s, \hat{\pi}_n\right) &= \hat{\theta}_n^{(s)} + \bar{\mathbf{y}} - \hat{\pi}_n = 0,\\
\Phi_n\left(\hat{\pi}_n, \mathbf{u}_s, \hat{\theta}_n^{(s)}\right) &= \hat{\pi}_n + \bar{\mathbf{y}} - \hat{\theta}_{\text{Boot},n}^{(s)}\\
&= -\hat{\pi}_n + \bar{\mathbf{y}} + \hat{\theta}_{\text{Boot},n}^{(s)}\\
&= \Phi_n\left(\hat{\theta}_{\text{Boot},n}^{(s)}, \mathbf{u}_s, \hat{\pi}_n\right) = 0,
\end{aligned}$$

where we use the fact that $\bar{\mathbf{y}} \overset{d}{=} -\bar{\mathbf{y}}$. Therefore, $\hat{\theta}_n^{(s)} = \hat{\theta}_{\text{Boot},n}^{(s)}$, or equivalently $\Phi_n\left(\theta, \mathbf{u}_s, \pi\right) = \Phi_n\left(\pi, \mathbf{u}_s, \theta\right) = 0$, which ends the proof. □

Proof of Theorem 1.19. Fix $\varepsilon = 0$. The Theorem 1.8 is satisfied so $\boldsymbol{\Theta}_n^{(s)} = \boldsymbol{\Theta}_{\text{II},n}^{(s)}$ for any s. It is sufficient then to prove $\boldsymbol{\Theta}_{\text{ABC,n}}^{(s)}(0) = \boldsymbol{\Theta}_{\text{II},n}^{(s)}$ for any $s \in \mathbb{N}_S^+$. We proceed by verifying that first $\boldsymbol{\Theta}_{\text{ABC},n}^{(s)}(0) \subset \boldsymbol{\Theta}_{\text{II},n}^{(s)}$, and second that $\boldsymbol{\Theta}_{\text{ABC},n}^{(s)}(0) \supset \boldsymbol{\Theta}_{\text{II},n}^{(s)}$.

(1). Fix $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_{\text{ABC},n}^{(s)}(0)$. By the Assumption 1.18, $\boldsymbol{\theta}_1$ is also a realization from the prior distribution $\mathscr{P}$. By Definition 1.14, we have

$$d\left(\hat{\boldsymbol{\pi}}_n, \hat{\boldsymbol{\pi}}_{\text{II},n}^{(s)}(\boldsymbol{\theta}_1)\right) = 0.$$

By Definition 1.4, $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_{\text{II},n}^{(s)}$, thus $\boldsymbol{\Theta}_{\text{ABC},n}^{(s)}(0) \subset \boldsymbol{\Theta}_{\text{II},n}^{(s)}$.

(2). Fix $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_{\text{II},n}^{(s)}$. By Definition 1.4, we have

$$d\left(\hat{\boldsymbol{\pi}}_n, \hat{\boldsymbol{\pi}}_{\text{II},n}^{(s)}(\boldsymbol{\theta}_2)\right) = 0.$$

By Assumption 1.18 and Definition 1.14, $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_{\text{ABC},n}^{(s)}(0)$, ergo $\boldsymbol{\Theta}_{\text{ABC},n}^{(s)}(0) \supset \boldsymbol{\Theta}_{\text{II},n}^{(s)}$, which ends the proof. □

Proof of Proposition 1.28. Fix $\alpha_1, \alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0,1)$. Since we consider an exact α-credible set $C_{\hat{\boldsymbol{\pi}}_n}$, we have

$$\begin{aligned}
1 - \alpha &= \Pr\left(\boldsymbol{\theta} \in C_{\hat{\boldsymbol{\pi}}_n} | \hat{\boldsymbol{\pi}}_n\right)\\
&= \Pr\left(\boldsymbol{\theta} \in \boldsymbol{\Theta}_n \setminus \{\underline{Q}_{\alpha_1} \cup \overline{Q}_{\alpha_2}\}\right)\\
&= \Pr\left(F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}) \in (\alpha_1, 1 - \alpha_2)\right).
\end{aligned}$$

Consider the event $E = \{u \in (\alpha_1, 1-\alpha_2)\}$ taking value one with probability p if u is inside the interval and 0 otherwise. Let $u = F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}_0)$ so at each trial there is one such event. Now consider indefinitely many trials, so we have $\{E_i : i \in \mathbb{N}^+\}$ where $\mathbb{E}(E_i) = \Pr(E_i = 1) = p_i$. Denote by N is the number of trials. The frequentist coverage probability is given by

$$\lim_{N\to\infty} \frac{\sum_{i=1}^N E_i}{N}.$$

By assumption, u is an independent standard uniform variable, so the events are independent and $p_i = 1-\alpha$ for all $i \geq 1$ and for every $\alpha \in (0,1)$. It follows that $\{E_i : i \in \mathbb{N}^+\}$ are identically and independently distributed Bernoulli random variables. The proof follows by Borel's strong law of large numbers (see [Wen91]). □

Proof of Lemma 1.30. Fix $\mathbf{u}_0$. Fix $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$. By definition we have

$$\hat{\boldsymbol{\pi}}_n = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}_0, \boldsymbol{\pi}\right).$$

By assumption, the following equation

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_1, \mathbf{u}_0, \hat{\boldsymbol{\pi}}_n\right) = \mathbf{0}$$

is uniquely defined. Now fix $\boldsymbol{\pi}_1 \in \boldsymbol{\Pi}$. By definition we have

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}, \mathbf{u}_0, \boldsymbol{\pi}_1\right),$$

and by assumption

$$\boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\theta}}_n, \mathbf{u}_0, \boldsymbol{\pi}_1\right) = \mathbf{0}$$

is uniquely defined. It follows that $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_n$ if and only if $\boldsymbol{\pi}_1 = \hat{\boldsymbol{\pi}}_n$. □

Proof of Theorem 1.38. We gives the demonstration under the Assumptions 1.36 and 1.37 separately.

1. We proceed by showing that we have a C^1-diffeomorphism which is unique so Lemma 1.A.2 and Lemma 1.30 apply. We then demonstrate that the obtained cumulative distribution function evaluated at $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is a realization from a standard uniform random variable. The conclusion is eventually reached by the Proposition 1.28.

Let $\pi_1 : \boldsymbol{\Theta}_n \times W_n \to \boldsymbol{\Theta}_n$ and $\pi_2 : \boldsymbol{\Theta}_n \times W_n \to W_n$ be the projections defined by $\pi_1(\boldsymbol{\theta}, \mathbf{w}) = \boldsymbol{\theta}$ and $\pi_2(\boldsymbol{\theta}, \mathbf{w}) = \mathbf{w}$ if $(\boldsymbol{\theta}, \mathbf{w}) \in \boldsymbol{\Theta}_n \times W_n$. By Assumption 1.36 the conditions of the global implicit function theorem of [Cri17, Theorem 1] are satisfied, so it holds that there exists a unique (global) continuous implicit function $\mathbf{a} : W_n \to \boldsymbol{\Theta}_n$ such that $\mathbf{a}(\mathbf{w}_0) = \boldsymbol{\theta}_0$ and $\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\mathbf{w}, \mathbf{a}(\mathbf{w})) = \mathbf{0}$ for every $\mathbf{w} \in W$. In addition, the mapping is continuously differentiable on $W_n \setminus \pi_2(K_n)$ with derivative given by

$$D_{\mathbf{w}}\mathbf{a} = -\left[D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_p|_{\boldsymbol{\theta}=\mathbf{a}(\mathbf{w})}\right]^{-1} D_{\mathbf{w}}\boldsymbol{\varphi}_p$$

for every $\mathbf{w} \in W_n \setminus \pi_2(K_n)$. Clearly the map $\mathbf{a}$ is invertible with a continuous inverse. Since the derivative $D_{\mathbf{w}}\boldsymbol{\varphi}_p$ is continuous and invertible for $(\boldsymbol{\theta}, \mathbf{w}) \in \boldsymbol{\Theta}_n \times W_n \setminus K_n$, we immediately have that $\mathbf{a}$ is a C^1-diffeomorphism with deriative of the inverse given by

$$D_{\boldsymbol{\theta}}\mathbf{a}^{-1} = -\left[D_{\mathbf{w}}\boldsymbol{\varphi}_p|_{\mathbf{w}=\mathbf{a}^{-1}(\boldsymbol{\theta})}\right]^{-1} D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_p$$

for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_n \setminus \pi_1(K_n)$. The conditions of Lemma 1.A.2 are satisfied and we obtain the cumulative distribution function

$$F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n} = \int_{W_n} f_{\mathbf{w}}\left(\mathbf{a}^{-1}(\boldsymbol{\theta})|\hat{\boldsymbol{\pi}}_n\right) \frac{\det\left(D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_p\right)}{\det\left(D_{\mathbf{w}}\boldsymbol{\varphi}_p\right)}\, \mathrm{d}\mathbf{w} \ = F_{\mathbf{w}|\hat{\boldsymbol{\pi}}_n},$$

proving point *(i)*. Since $\hat{\boldsymbol{\pi}}_n$ is the unique zero of $\boldsymbol{\Phi}_n(\boldsymbol{\theta}_0, \mathbf{u}_0, \boldsymbol{\pi})$, and hence of $\boldsymbol{\varphi}_p(\boldsymbol{\theta}_0, \mathbf{w}_0, \boldsymbol{\pi})$, and $\boldsymbol{\theta} = \mathbf{a}(\mathbf{w})$ is the unique zero of $\boldsymbol{\varphi}_p(\boldsymbol{\theta}, \mathbf{w}, \hat{\boldsymbol{\pi}}_n)$, we have by Lemma 1.30 that $\boldsymbol{\theta}_0 = \mathbf{a}(\mathbf{w}_0)$, and therefore that $\mathbf{w}_0 = \mathbf{a}^{-1}(\boldsymbol{\theta}_0)$. In consequence, evaluating the above distribution at $\boldsymbol{\theta}_0$ leads to

$$F_{\hat{\boldsymbol{\theta}}_n|\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}_0) = F_{\mathbf{w}|\hat{\boldsymbol{\pi}}_n}(\mathbf{w}_0) = u \sim \mathcal{U}(0,1),$$

that is, the distribution evaluated at $\boldsymbol{\theta}_0$ is a realization from a standard uniform random variable. The conclusion follows by the Proposition 1.28.

2. Fix $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_n$ and $\mathbf{w}_0 \in W_n$. Fix $\hat{\boldsymbol{\pi}}_n \in \boldsymbol{\Pi}_n$, the point such that $\boldsymbol{\varphi}_p(\boldsymbol{\theta}_0, \mathbf{w}_0, \hat{\boldsymbol{\pi}}_n) = \mathbf{0}$. Let $\pi_1 : W_n \times \boldsymbol{\Pi}_n \to W_n$ and $\pi_2 : W_n \times \boldsymbol{\Pi}_n \to \boldsymbol{\Pi}_n$ be the projections such that $\pi_1(\mathbf{w}, \boldsymbol{\pi}) = \mathbf{w}$ and $\pi_2(\mathbf{w}, \boldsymbol{\pi}) = \boldsymbol{\pi}$ if $(\mathbf{w}, \boldsymbol{\pi}) \in W_n \times \boldsymbol{\Pi}_n$. By Assumption 1.37 (*(i), (iii), (v)*), the Theorem 1 in [Cri17] is satisfied, as a consequence it holds that $\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}$ admits a unique global implicit function $\boldsymbol{\pi}_{\boldsymbol{\theta}_0} : W_n \to \boldsymbol{\Pi}_n$ such that $\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}(\mathbf{w}, \boldsymbol{\pi}_{\boldsymbol{\theta}_0}(\mathbf{w})) = \mathbf{0}$ for every $\mathbf{w} \in W_n$, $\boldsymbol{\pi}_{\boldsymbol{\theta}_0}(\mathbf{w}_0) = \hat{\boldsymbol{\pi}}_n$, and $\boldsymbol{\pi}_{\boldsymbol{\theta}_0}$ is once continuously differentiable on $W_n \setminus \pi_1(K_{1n})$ with derivative given by

$$D_{\mathbf{w}}\boldsymbol{\pi}_{\boldsymbol{\theta}_0} = -[D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}]^{-1} D_{\mathbf{w}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}.$$

Clearly $\mathbf{w} \mapsto \boldsymbol{\pi}_{\boldsymbol{\theta}_0}$ is a homeomorphism. Since $D_{\mathbf{w}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}$ is continuous and invertible on $W_n \times \boldsymbol{\Pi} \setminus K_{1n}$, we have that $\boldsymbol{\pi}_{\boldsymbol{\theta}_0}$ is a C^1-diffeomorphism with differentiable inverse function on $\boldsymbol{\Pi} \setminus \pi_2(K_{1n})$ given by Lemma 1.A.1:

$$D_{\boldsymbol{\pi}}\boldsymbol{\pi}_{\boldsymbol{\theta}_0}^{-1} = [D_{\mathbf{w}}\boldsymbol{\pi}_{\boldsymbol{\theta}_0}]^{-1} = -[D_{\mathbf{w}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}]^{-1} D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}.$$

Let $\pi_3 : \boldsymbol{\Theta}_n \times \boldsymbol{\Pi}_n \to \boldsymbol{\Theta}_n$ and $\pi_4 : \boldsymbol{\Theta}_n \times \boldsymbol{\Pi}_n \to \boldsymbol{\Pi}_n$ denotes the projections such that $\pi_3(\boldsymbol{\theta}, \boldsymbol{\pi}) = \boldsymbol{\theta}$ and $\pi_4(\boldsymbol{\theta}, \boldsymbol{\pi}) = \boldsymbol{\pi}$. By using the same argument presented above, the Assumption 1.37 (*(ii), (iv), (vi)*) permits us to have an implicit C^1-diffeomorphism $\boldsymbol{\pi}_{\mathbf{w}_0} : \boldsymbol{\Theta}_n \to \boldsymbol{\Pi}_n$ with the following continuous derivatives:

$$\begin{aligned} D_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\mathbf{w}_0} &= -[D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\mathbf{w}_0}]^{-1} D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\mathbf{w}_0}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \pi_3(K_2), \\ D_{\boldsymbol{\pi}}\boldsymbol{\pi}_{\mathbf{w}_0}^{-1} &= -[D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\mathbf{w}_0}]^{-1} D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\mathbf{w}_0}, \quad \boldsymbol{\pi} \in \boldsymbol{\Pi} \setminus \pi_4(K_2). \end{aligned}$$

Now define the function $\boldsymbol{\xi}(\boldsymbol{\theta}) = \boldsymbol{\pi}_{\boldsymbol{\theta}_0}^{-1} \circ \boldsymbol{\pi}_{\mathbf{w}_0}(\boldsymbol{\theta})$. It is trivial to show that this mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\xi}$ is a C^1-diffeomorphism. We have from the preceding results and the chain rule that

$$D_{\boldsymbol{\theta}}\boldsymbol{\xi} = [D_{\mathbf{w}_0}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}]^{-1} D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0} [D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\mathbf{w}_0}]^{-1} D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\mathbf{w}_0}.$$

We make the following remarks. First, note that all these derivatives are square matrices of dimension $p \times p$. Second, we have that $D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}(\mathbf{w}_0, \hat{\boldsymbol{\pi}}_n) = D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_p(\boldsymbol{\theta}_0, \mathbf{w}_0, \hat{\boldsymbol{\pi}}_n) = D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\mathbf{w}_0}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\pi}}_n)$ so $D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}[D_{\boldsymbol{\pi}}\boldsymbol{\varphi}_{\mathbf{w}_0}]^{-1} = \mathbf{I}_p$. Third, it holds that $D_{\mathbf{w}}\boldsymbol{\varphi}_{\boldsymbol{\theta}_0}(\mathbf{w}_0, \hat{\boldsymbol{\pi}}_n) = D_{\mathbf{w}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}_0, \mathbf{w}_0)$ and $D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\mathbf{w}_0}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\pi}}_n) = D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}_0, \mathbf{w}_0)$. As a consequence, we obtain that

$$\det\left(D_{\boldsymbol{\theta}}\boldsymbol{\xi}\right) = \frac{\det\left(D_{\boldsymbol{\theta}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\mathbf{w}_0, \boldsymbol{\theta}_0)\right)}{\det\left(D_{\mathbf{w}}\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\mathbf{w}_0, \boldsymbol{\theta}_0)\right)}.$$

Using Lemma 1.A.2 ends the proof of point *(i)* in Theorem 1.38. From the above display, we have that the relation $\boldsymbol{\pi}_{\boldsymbol{\theta}_0}(\mathbf{w}_0) = \hat{\boldsymbol{\pi}}_n = \boldsymbol{\pi}_{\mathbf{w}_0}(\boldsymbol{\theta}_0)$ is uniquely defined, so $\boldsymbol{\xi}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_{\boldsymbol{\theta}_0}^{-1}(\hat{\boldsymbol{\pi}}_n) = \mathbf{w}_0$. Since $\boldsymbol{\xi}$ is a diffeomorphism, then $\boldsymbol{\xi}^{-1}(\mathbf{w}_0) = \boldsymbol{\theta}_0$, which finishes the proof. □

Proof of Proposition 1.41. This is a special case of the Theorem 1.38. Let define $\boldsymbol{\varphi}_{\hat{\boldsymbol{\pi}}_n}(\mathbf{w},\boldsymbol{\theta}) = \mathbf{h}(\mathbf{x}_0) - \boldsymbol{g}(\boldsymbol{\theta},\mathbf{w})$, where $\mathbf{h}(\mathbf{x}_0) = \hat{\boldsymbol{\pi}}_n$ is fixed. Following the proof of Theorem 1.38, we have by assumption that $\mathbf{a} : W_n \to \boldsymbol{\Theta}_n$ is a C^1-diffeomorphism with derivatives

$$\begin{aligned} D_{\mathbf{w}}\mathbf{a} &= -\left[D_{\boldsymbol{\theta}}\boldsymbol{g}|_{\boldsymbol{\theta}=\mathbf{a}(\mathbf{w})}\right]^{-1} D_{\mathbf{w}}\boldsymbol{g}, \quad \mathbf{w} \in W_n \setminus \pi_2(K_n), \\ D_{\boldsymbol{\theta}}\mathbf{a}^{-1} &= -\left[D_{\mathbf{w}}\boldsymbol{g}|_{\mathbf{w}=\mathbf{a}^{-1}(\boldsymbol{\theta})}\right]^{-1} D_{\boldsymbol{\theta}}\boldsymbol{g}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_n \setminus \pi_1(K_n). \end{aligned}$$

The rest of the proof is identical to the proof of Theorem 1.38. □

1.C Asymptotics

Proof of Theorem 1.45. For any estimator, we proceed by verifying the assumptions for the weak consistency result of Lemma 3.1. We start by showing the claim 1: the pointwise convergence of $\hat{\boldsymbol{\pi}}_n$. Then we demonstrate the claim 2 with two different approaches corresponding respectively to the Assumptions 1.43 and 1.44.

1. Fix $\boldsymbol{\pi}_0 \in \boldsymbol{\Pi}$. Since $\{\boldsymbol{\Phi}_n(\boldsymbol{\theta},\mathbf{u},\boldsymbol{\pi})\}$ is stochastically Lipschitz in $\boldsymbol{\pi}$, it is stochastically equicontinuous by the Lemma 3.4. In addition, $\boldsymbol{\Pi}$ is compact and $\{\boldsymbol{\Phi}_n\}$ is pointwise convergent by assumption, so by the Lemma 3.3 $\{\boldsymbol{\Phi}_n\}$ converges uniformly and the limit $\boldsymbol{\Phi}$ is uniformly continuous. By $\boldsymbol{\Pi}$ compact and the continuity of the norm, the infimum of the norm of $\boldsymbol{\Phi}$ exists. The infimum of $\boldsymbol{\Phi}$ is well-separated by the bijectivity of the function. Therefore, all the conditions of Lemma 3.1 are satisfied and $\{\hat{\boldsymbol{\pi}}_n\}$ converges pointwise to $\boldsymbol{\pi}_0$.

2 (i). For this proof, we consider $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ jointly. Let $\mathscr{K} = \boldsymbol{\Theta} \cap \boldsymbol{\Pi}$ be the set for both $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. Fix $(\boldsymbol{\theta}_0,\boldsymbol{\pi}_0) \in \mathscr{K}$. Since $\boldsymbol{\Pi} \subset \mathrm{I\!R}^p$ and $\boldsymbol{\Theta} \subset \mathrm{I\!R}^p$ are compact subsets of a metric space, they are closed (see the Theorem 2.34 in [Rud76]), and $\mathscr{K}$ is compact (see the Corollary to the Theorem 2.35 in [Rud76]) and nonempty (Theorem 2.36 in [Rud76]). Having $\mathscr{K}$ compact, it is now sufficient to show that $\{\boldsymbol{\Phi}_n\}$ is jointly stochastically Lipschitz as the rest of the proof follows exactly the same steps as the claim 1.

For every $(\boldsymbol{\theta}_1,\boldsymbol{\pi}_1),(\boldsymbol{\theta}_2,\boldsymbol{\pi}_2) \in \mathscr{K}$, n and $\mathbf{u} \sim F_{\mathbf{u}}$, we have by the triangle inequality that

$$\begin{aligned} \|\boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}_1) - \boldsymbol{\Phi}_n(\boldsymbol{\theta}_2,\mathbf{u},\boldsymbol{\pi}_2)\| &= \Big\|\boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}_1) - \boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}_2) \\ &\quad + \boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}_2) - \boldsymbol{\Phi}_n(\boldsymbol{\theta}_2,\mathbf{u},\boldsymbol{\pi}_2)\Big\| \\ &\leq \|\boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}_1) - \boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}_2)\| \\ &\quad + \|\boldsymbol{\Phi}_n(\boldsymbol{\theta}_1,\mathbf{u},\boldsymbol{\pi}_2) - \boldsymbol{\Phi}_n(\boldsymbol{\theta}_2,\mathbf{u},\boldsymbol{\pi}_2)\| \\ &\leq D_n\left(\|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\| + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\right), \end{aligned}$$

where for the last inequality we make use of the marginal stochastic Lipschitz assumptions and $D_n = \max(A_n, B_n)$. Let $a = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$ and $b = \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\|$. Now remark that for the ℓ_2-norm we have

$$\left\|\begin{pmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\pi}_1\end{pmatrix} - \begin{pmatrix}\boldsymbol{\theta}_2 \\ \boldsymbol{\pi}_2\end{pmatrix}\right\| = \sqrt{a^2+b^2}.$$

Since a, b are positive real numbers, a direct application of the inequality of arithmetic and geometric means gives

$$\sqrt{2}\sqrt{a^2+b^2} \geq a + b.$$

Therefore, we have that

$$D_n\left(\|\boldsymbol{\pi}_1-\boldsymbol{\pi}_2\|+\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|\right)\leq D_n^\star\left\|\begin{pmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\pi}_1\end{pmatrix}-\begin{pmatrix}\boldsymbol{\theta}_2\\ \boldsymbol{\pi}_2\end{pmatrix}\right\|,$$

where $D_n^\star=\sqrt{2}D_n$. Consequently, $\{\boldsymbol{\Phi}_n\}$ is jointly stochastically Lipschitz, and following the proof of claim 1 we have that $\hat{\boldsymbol{\theta}}_n\xrightarrow{p}\boldsymbol{\theta}_0$. More precisely, we even have that $(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\pi}}_n)\xrightarrow{p}(\boldsymbol{\theta}_0,\boldsymbol{\pi}_0)$.

2 (ii). This proof is different from 2 (i) since $\hat{\boldsymbol{\pi}}_{\text{II},n}$ is considered as a function of $\boldsymbol{\theta}$. Fix $\boldsymbol{\pi}_0\in\boldsymbol{\Pi}$. Since $\{\hat{\boldsymbol{\pi}}_{\text{II},n}\}$ is stochastically Lipschitz in $\boldsymbol{\theta}$, it is stochastically equicontinuous by the Lemma 3.4. In addition, $\boldsymbol{\Theta}$ is compact and $\{\hat{\boldsymbol{\pi}}_{\text{II},n}\}$ is pointwise convergent by the claim 1, so by the Lemma 3.3 $\{\hat{\boldsymbol{\pi}}_{\text{II},n}\}$ converges uniformly and the limit $\boldsymbol{\pi}$ is uniformly continuous in $\boldsymbol{\theta}$. Let the stochastic and deterministic objective functions be $Q_n(\boldsymbol{\theta})=\|\hat{\boldsymbol{\pi}}_n-\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta})\|$ and $Q(\boldsymbol{\theta})=\|\boldsymbol{\pi}_0-\boldsymbol{\pi}(\boldsymbol{\theta})\|$, for any norms. Now, we have by using successively the reverse and the regular triangle inequalities

$$\begin{aligned}|Q_n(\boldsymbol{\theta})-Q(\boldsymbol{\theta})|&=\left|\|\hat{\boldsymbol{\pi}}_n-\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta})\|-\|\boldsymbol{\pi}_0-\boldsymbol{\pi}(\boldsymbol{\theta})\|\right|\\&\leq\|\hat{\boldsymbol{\pi}}_n-\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta})-\boldsymbol{\pi}_0+\boldsymbol{\pi}(\boldsymbol{\theta})\|\\&\leq\|\hat{\boldsymbol{\pi}}_n-\boldsymbol{\pi}_0\|+\|\boldsymbol{\pi}(\boldsymbol{\theta})-\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta})\|.\end{aligned}$$

By the convergence of $\{\hat{\boldsymbol{\pi}}_n\}$ and the uniform convergence of $\{\hat{\boldsymbol{\pi}}_{\text{II},n}\}$, we have

$$\lim_{n\to\infty}\Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|Q_n(\boldsymbol{\theta})-Q(\boldsymbol{\theta})|\right)=o_p(1).$$

By $\boldsymbol{\Pi}$ compact and the continuity of the norm, the infimum of the norm of $\boldsymbol{\Phi}$ exists. The infimum of $\boldsymbol{\Phi}$ is well-separated by the bijectivity of the function. Therefore, all the conditions of Lemma 3.1 are satisfied and $\{\hat{\boldsymbol{\pi}}_n\}$ converges pointwise to $\boldsymbol{\pi}_0$. □

Proof of Theorem 1.49. We first demonstrate the asymptotic distribution of the auxiliary estimator, then separately shows the result for $\hat{\boldsymbol{\theta}}_n$ using independentely the Assumption 1.47 and 1.48.

1. The result on $\hat{\boldsymbol{\pi}}_n$ is a special case of $\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta})$. Fix $\boldsymbol{\theta}_0\in\boldsymbol{\Theta}^\circ$ and denote $\boldsymbol{\pi}(\boldsymbol{\theta}_0)\equiv\boldsymbol{\pi}_0$. By assumptions, the conditions for the delta method in Lemma 3.8 are satisfied so we have

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta}_0)\right)-\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\boldsymbol{\pi}_0\right)=D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\boldsymbol{\pi}_0\right)\cdot(\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta}_0)-\boldsymbol{\pi}_0)+o_p\left(\|\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta}_0)-\boldsymbol{\pi}_0\|\right).\tag{1.9}$$

By the Definition 1.4, we have $\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta}_0)\right)=\mathbf{0}$. By the Theorem 1.45, $o_p\left(\|\hat{\boldsymbol{\pi}}_{\text{II},n}(\boldsymbol{\theta}_0)-\boldsymbol{\pi}_0\|\right)=o_p(1)$. By assumptions, $D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\boldsymbol{\pi}_0\right)\xrightarrow{p}\mathbf{K}$, $\mathbf{K}$ nonsingular. Multiplying by square-root n, the proof results from the central limit theorem assumption on $\boldsymbol{\Phi}_n$ and the Slutsky's lemma.

2 *(i)*. From the delta method in Lemma 3.8, we obtain

$$\boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\theta}}_n,\mathbf{u}_s,\hat{\boldsymbol{\pi}}_n\right)-\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\hat{\boldsymbol{\pi}}_n\right)=D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\hat{\boldsymbol{\pi}}_n\right)\cdot\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)+o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right).$$

By definition we have $\boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\theta}}_n,\mathbf{u}_s,\hat{\boldsymbol{\pi}}_n\right)=\mathbf{0}$. Using again the delta method on the non-zero left-hand side element, we obtain from (1.9)

$$\begin{aligned}&\mathbf{0}-\left[\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\boldsymbol{\pi}_0\right)+D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\boldsymbol{\pi}_0\right)\cdot(\hat{\boldsymbol{\pi}}_n-\boldsymbol{\pi}_0)+o_p\left(\|\hat{\boldsymbol{\pi}}_n-\boldsymbol{\pi}_0\|\right)\right]\\&\quad=D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0,\mathbf{u}_s,\hat{\boldsymbol{\pi}}_n\right)\cdot\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)+o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right).\end{aligned}$$

Since $\{D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\boldsymbol{\theta}_0,\mathbf{u}_s,\boldsymbol{\pi})\}$ is stochastically Lipschitz in $\boldsymbol{\pi}$, it is stochastically equicontinuous by the Lemma 3.4. In addition, $\boldsymbol{\Pi}$ is compact and $\{D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n\}$ is pointwise convergent by assumption, so by the Lemma 3.3 $\{D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n\}$ converges uniformly and the limit $\mathbf{J}$ is uniformly continuous in $\boldsymbol{\pi}$.

Next, we obtain the following

$$\begin{aligned}\left\|D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\hat{\boldsymbol{\pi}}_n)-\mathbf{J}(\boldsymbol{\pi}_0)\right\| &\leq \left\|D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\hat{\boldsymbol{\pi}}_n)-\mathbf{J}(\hat{\boldsymbol{\pi}}_n)\right\|+\left\|\mathbf{J}(\hat{\boldsymbol{\pi}}_n)-\mathbf{J}(\boldsymbol{\pi}_0)\right\|\\ &\leq \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\left\|D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\boldsymbol{\pi})-\mathbf{J}(\boldsymbol{\pi})\right\|+\left\|\mathbf{J}(\hat{\boldsymbol{\pi}}_n)-\mathbf{J}(\boldsymbol{\pi}_0)\right\|.\end{aligned}$$

By uniform convergence $\sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\left\|D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\boldsymbol{\pi})-\mathbf{J}(\boldsymbol{\pi})\right\|=o_p(1)$ and by the continuous mapping theorem $\left\|\mathbf{J}(\hat{\boldsymbol{\pi}}_n)-\mathbf{J}(\boldsymbol{\pi}_0)\right\|=o_p(1)$.

The central limit theorem is satisfied for the estimating equation thus $n^{1/2}\boldsymbol{\Phi}_n \rightsquigarrow \mathcal{N}\left(\mathbf{0},\mathbf{Q}\right)$. Let $\mathbf{y}$ be a random variable identically and independently distributed according to $\mathcal{N}(\mathbf{0},\mathbf{Q})$. Therefore, by multiplying by square-root n we obtain

$$-\mathbf{y}-\mathbf{K}n^{1/2}\left(\hat{\boldsymbol{\pi}}_n-\boldsymbol{\pi}_0\right)-o_p\left(\left\|\hat{\boldsymbol{\pi}}_n-\boldsymbol{\pi}_0\right\|\right)=\mathbf{J}n^{1/2}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)+o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right).$$

By the Theorem 1.45, we have $o_p\left(\left\|\hat{\boldsymbol{\pi}}_n-\boldsymbol{\pi}_0\right\|\right)=o_p(1)$ and $o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right)=o_p(1)$. By the result of the claim 1 and the nonsingularity of $\mathbf{J}$, we have

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)=-\mathbf{J}^{-1}\left(\mathbf{y}+\mathbf{K}\cdot\mathbf{K}^{-1}\mathbf{y}+o_p(1)\right)+o_p(1).$$

Slutsky's lemma ends the proof.

2 (ii). Let $\mathbf{g}_n(\boldsymbol{\theta})=\hat{\boldsymbol{\pi}}_n-\hat{\boldsymbol{\pi}}_{\mathrm{II},n}(\boldsymbol{\theta})$. The conditions for the delta method in Lemma 3.8 are satisfied by assumption so we have

$$\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)-\mathbf{g}_n(\boldsymbol{\theta}_0)=D_{\boldsymbol{\theta}}\mathbf{g}_n(\boldsymbol{\theta}_0)\cdot\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)+o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right). \tag{1.10}$$

Since $\hat{\boldsymbol{\theta}}_n=\operatorname{argzero}_{\boldsymbol{\theta}} d(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\theta}}_{\mathrm{II},n}(\boldsymbol{\theta}))$, we have $\hat{\boldsymbol{\theta}}_n-\hat{\boldsymbol{\theta}}_{\mathrm{II},n}(\hat{\boldsymbol{\theta}}_n)=\mathbf{0}$ and thus $\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)=\mathbf{0}$. By the Theorem 1.45, we have $o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\right)=o_p(1)$. We have $D_{\boldsymbol{\theta}}\mathbf{g}_n(\boldsymbol{\theta}_0)=-D_{\boldsymbol{\theta}}\hat{\boldsymbol{\pi}}_{\mathrm{II},n}(\boldsymbol{\theta}_0)$ which, by assumption converges pointwise to $D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta}_0)$. By the claim 1, we have $n^{1/2}(\hat{\boldsymbol{\pi}}_n-\boldsymbol{\pi}_0)\overset{d}{=}n^{1/2}(\hat{\boldsymbol{\pi}}_{\mathrm{II},n}(\boldsymbol{\theta})-\boldsymbol{\pi}_0)\overset{d}{=}\mathbf{K}^{-1}\mathbf{y}$ as $n\to\infty$. Hence, multiplying the Equation 1.10 by square-root n, gives the following

$$\mathbf{K}^{-1}(\mathbf{y}-\mathbf{y})=D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta}_0)\cdot n^{1/2}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)+o_p(1),$$

for sufficiently large n. Remark that the mapping $\boldsymbol{\theta}\mapsto\boldsymbol{\pi}$ is implicitely defined by

$$\boldsymbol{\Phi}\left(\boldsymbol{\theta},\boldsymbol{\pi}(\boldsymbol{\theta})\right)=\mathbf{0}.$$

Since $\boldsymbol{\Phi}$ is once continuously differentiable in $(\boldsymbol{\theta},\boldsymbol{\pi})$ and the partial derivatives are invertibles, the conditions for invoking an implicit function theorem are satisfied (see for example the Theorem 9.28 in [Rud76]) and one of the conclusion is that

$$D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta}_0)=-\mathbf{K}^{-1}\mathbf{J}.$$

Since $\mathbf{J}$ is invertible, the conclusion follows by Slutsky's lemma. □

Proof of Proposition 1.50. The proof follows essentially the same steps as the proof of Theorem 1.49. From the proof of Theorem 1.49, the following holds: $n^{1/2}\left(\hat{\boldsymbol{\pi}}_n - \boldsymbol{\pi}_0\right) \overset{\mathrm{d}}{=} \mathbf{K}^{-1}\mathbf{y}_0$ and $n^{1/2}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0, \mathbf{u}_s, \boldsymbol{\pi}_0\right) \overset{\mathrm{d}}{=} \mathbf{y}_s$ as $n \to \infty$ where $\mathbf{y}_j \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}\right)$, $j \in \mathbb{N}^+$, $D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0, \mathbf{u}_0, \boldsymbol{\pi}_0\right)$ converges in probability to $\mathbf{K}$ and $D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0, \mathbf{u}_s, \boldsymbol{\pi}\right)$ converges uniformly in probability to $\mathbf{J}$. The $\{\mathbf{u}_j : j \in \mathbb{N}_S\}$ are assumed independent and so are $\{\mathbf{y}_j : j \in \mathbb{N}_S\}$.

From the delta method in Lemma 3.8, we obtain

$$\frac{1}{S}\sum_{s\in\mathbb{N}_S^+} \boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\theta}}_n^{(s)}, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) - \frac{1}{S}\sum_{s\in\mathbb{N}_S^+} \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) = \frac{1}{S}\sum_{s\in\mathbb{N}_S^+} D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right)\cdot\left(\hat{\boldsymbol{\theta}}_n^{(s)} - \boldsymbol{\theta}_0\right) + o_p(1).$$

By definition $\frac{1}{S}\sum_{s\in\mathbb{N}_S^+} \boldsymbol{\Phi}_n\left(\hat{\boldsymbol{\theta}}_n^{(s)}, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right) = \mathbf{0}$. Using the delta method on $\frac{1}{S}\sum_{s\in\mathbb{N}_S^+} \boldsymbol{\Phi}_n\left(\boldsymbol{\theta}_0, \mathbf{u}_s, \hat{\boldsymbol{\pi}}_n\right)$, multiplying by square-root n, we obtain from the results of Theorem 1.49:

$$-\frac{1}{S}\sum_{s\in\mathbb{N}_S^+} \mathbf{y}_s - \mathbf{K}\mathbf{K}^{-1}\mathbf{y}_0 - o_p(1) = \mathbf{J}n^{1/2}\left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) + o_p(1).$$

Clearly $\frac{1}{S}\sum_{s\in\mathbb{N}_S^+} \mathbf{y}_s \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{S}\mathbf{Q}\right)$. The conclusion follows from Slutsky's lemma. □

1.D Additional simulation results

1.D.1 Lomax distribution

	SwiZs	Boot	AB	RSwiZs	RBoot
$n = 35$	0.1430	0.0222	0.0197	0.5613	0.0998
$n = 50$	0.2002	0.0293	0.0268	0.7889	0.1320
$n = 100$	0.3826	0.0526	0.0504	1.3520	0.2314
$n = 150$	0.5580	0.0753	0.0736	1.7792	0.3291
$n = 250$	0.8998	0.1228	0.1211	2.3141	0.5174
$n = 500$	1.7763	0.2364	0.2398	3.2132	0.9848

Table 1.5: Average computationnal time in seconds to approximate a distribution on $S = 10,000$ points.

	SwiZs		Boot		BA		RSwiZs		RBoot	
α	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
$n = 35$										
50%	49.48	50.07	43.10	44.26	0.00	0.00	42.73	44.07	36.72	36.84
75%	74.49	75.14	65.82	65.39	0.00	0.00	65.84	66.59	55.00	55.06
90%	89.31	89.39	80.64	78.74	0.00	0.00	81.41	81.97	64.47	64.26
95%	94.27	94.34	86.71	84.28	0.03	0.00	87.58	87.41	67.33	67.13
99%	98.26	98.43	91.23	91.07	0.75	0.00	93.84	93.53	69.64	70.39
$n = 50$										
50%	49.59	49.88	44.48	45.30	0.01	0.00	45.70	46.93	37.37	37.64
75%	74.73	76.67	68.43	67.84	0.08	0.00	67.40	68.21	57.44	56.73
90%	89.89	90.62	83.15	81.57	0.76	0.00	82.51	82.75	69.52	68.81
95%	94.67	94.94	89.26	87.11	1.92	0.00	88.47	88.35	73.01	72.49
99%	98.40	98.46	95.19	93.69	10.86	0.00	94.79	94.80	75.97	76.43
$n = 100$										
50%	49.86	49.95	47.52	48.04	20.52	27.75	49.44	49.80	36.19	35.48
75%	75.37	75.88	72.00	71.59	44.13	57.82	73.07	74.32	57.01	55.61
90%	90.20	90.42	86.69	85.86	69.68	81.85	86.54	86.83	73.68	71.96
95%	95.41	95.67	92.06	90.96	81.89	91.13	91.69	91.52	80.75	79.17
99%	98.85	98.91	97.32	96.42	94.93	98.74	96.85	96.79	86.96	86.38
$n = 150$										
50%	50.12	49.80	48.36	48.58	47.05	49.78	49.80	49.82	33.94	33.00
75%	74.85	75.32	72.41	72.63	70.68	72.58	74.44	74.69	55.12	53.45
90%	90.31	90.32	87.58	86.85	86.94	89.18	88.95	89.22	72.14	70.01
95%	95.08	95.35	93.03	92.11	93.26	94.89	93.60	93.74	80.17	78.15
99%	99.08	99.10	97.92	97.43	98.72	99.28	97.81	97.69	90.07	88.56
$n = 250$										
50%	49.46	49.84	48.60	49.01	47.61	47.09	49.55	49.90	29.16	28.45
75%	75.02	74.49	73.59	72.75	72.09	72.63	74.83	74.80	49.94	47.56
90%	89.55	89.81	88.05	88.11	89.54	90.13	89.56	89.58	67.50	65.25
95%	94.77	94.79	93.56	93.34	94.79	95.68	94.50	94.70	76.90	74.39
99%	99.02	99.03	98.46	97.92	99.18	99.50	98.61	98.70	89.37	87.24
$n = 500$										
50%	50.08	49.89	49.29	49.81	48.76	48.67	50.26	49.64	20.51	18.95
75%	74.73	74.36	73.90	73.64	73.68	73.85	74.55	74.68	37.76	34.96
90%	89.53	89.75	88.86	88.69	89.03	89.22	89.45	89.80	56.15	52.68
95%	94.92	94.86	94.11	94.22	94.33	94.77	94.92	94.80	66.89	63.51
99%	98.97	98.99	98.62	98.40	99.01	99.07	98.94	99.03	83.63	80.06

Table 1.6: Estimated coverage probabilities.

	SwiZs	Boot	BA	RSwiZs	RBoot
α			Gini index		
			$n = 35$		
50%	50.22	44.26	0.02	44.27	36.84
75%	76.03	65.44	0.72	67.12	55.06
90%	91.07	78.96	68.11	83.07	64.36
95%	96.76	84.35	100.00	89.43	67.19
99%	98.84	91.10	100.00	93.88	70.41
			$n = 50$		
50%	49.89	45.30	0.00	46.94	37.64
75%	76.86	67.84	0.00	68.26	56.73
90%	90.83	81.58	41.20	82.68	68.82
95%	95.17	87.16	71.42	88.40	72.49
99%	98.92	93.76	99.82	95.14	76.45
			$n = 100$		
50%	49.95	48.04	32.96	49.80	35.48
75%	75.88	71.59	59.90	74.32	55.61
90%	90.42	85.86	82.63	86.83	71.96
95%	95.74	90.98	91.44	91.64	79.19
99%	98.85	96.46	98.73	96.83	86.43
			$n = 150$		
50%	49.80	48.58	46.30	49.82	33.00
75%	75.32	72.63	72.68	74.69	53.45
90%	90.32	86.85	89.18	89.22	70.01
95%	95.35	92.12	94.87	93.73	78.15
99%	99.06	97.47	99.27	97.71	88.60
			$n = 250$		
50%	49.84	49.01	46.99	49.90	28.45
75%	74.49	72.75	72.41	74.80	47.56
90%	89.81	88.11	88.95	89.58	65.25
95%	94.81	93.34	94.99	94.69	74.43
99%	99.04	97.93	99.48	98.68	87.34
			$n = 500$		
50%	49.89	49.81	48.67	49.64	18.95
75%	74.36	73.64	73.85	74.68	34.96
90%	89.75	88.69	89.22	89.80	52.68
95%	94.86	94.22	94.77	94.79	63.57
99%	98.98	98.41	99.03	99.02	80.28

Table 1.7: Estimated coverage probabilities of Gini index.

α	SwiZs	Boot	BA	RSwiZs	RBoot
			95% value-at-risk		
			$n = 35$		
50%	47.30	46.08	20.92	45.34	41.13
75%	73.76	67.53	55.77	70.38	61.00
90%	90.05	80.35	93.73	88.08	73.92
95%	95.67	85.36	98.92	94.80	79.41
99%	99.17	91.63	99.97	99.25	87.26
			$n = 50$		
50%	48.14	47.23	31.76	46.40	41.27
75%	73.39	69.40	63.30	70.22	61.47
90%	89.63	82.24	91.60	87.07	74.72
95%	94.89	87.41	97.72	93.60	80.20
99%	99.23	93.17	99.90	99.27	87.87
			$n = 100$		
50%	49.75	48.90	48.33	49.18	39.94
75%	74.68	72.61	75.68	72.93	61.39
90%	89.48	86.38	91.97	87.16	75.97
95%	95.07	91.17	96.79	94.17	82.45
99%	99.23	96.31	99.75	99.11	90.45
			$n = 150$		
50%	50.10	49.19	49.47	49.91	37.43
75%	74.13	73.17	75.42	73.57	59.31
90%	89.77	87.25	91.21	88.49	75.26
95%	94.76	92.57	96.18	93.31	81.76
99%	98.89	97.34	99.61	98.46	91.00
			$n = 250$		
50%	50.28	49.52	50.02	50.24	34.09
75%	75.29	74.25	74.87	74.75	55.55
90%	89.43	88.10	90.27	89.13	72.28
95%	94.66	93.26	95.15	94.14	80.35
99%	98.89	97.85	99.10	98.67	90.11
			$n = 500$		
50%	49.15	48.63	49.00	49.22	27.45
75%	74.88	74.01	74.63	74.53	45.61
90%	90.02	89.46	90.37	89.93	62.84
95%	94.97	94.45	95.18	94.85	72.65
99%	98.92	98.32	98.87	98.96	86.63

Table 1.8: Estimated coverage probabilities of value-at-risk at 95%.

	SwiZs	Boot	BA	RSwiZs	RBoot
α		95% expected shortfall			
$n = 35$					
50%	50.33	48.55	0.02	50.08	47.38
75%	74.97	72.60	0.72	74.70	71.28
90%	89.61	87.63	68.11	89.24	86.35
95%	94.65	92.87	100.00	94.37	92.23
99%	98.80	97.97	100.00	98.72	97.48
$n = 50$					
50%	49.48	48.24	0.00	49.28	47.06
75%	74.81	72.74	0.00	74.45	71.28
90%	89.76	88.07	41.20	89.25	86.85
95%	94.74	93.32	71.42	94.48	92.16
99%	98.89	97.92	99.82	98.62	97.48
$n = 100$					
50%	49.94	49.16	32.96	49.64	47.22
75%	74.47	74.12	59.90	74.37	72.21
90%	90.13	89.15	82.63	89.99	87.57
95%	95.10	94.23	91.44	95.00	93.13
99%	98.98	98.55	98.73	98.91	98.10
$n = 150$					
50%	49.91	49.49	46.30	49.81	48.13
75%	75.03	74.25	72.68	74.95	72.45
90%	89.82	89.31	89.18	89.74	87.76
95%	95.05	94.37	94.87	94.98	93.15
99%	98.91	98.62	99.27	98.86	98.14
$n = 250$					
50%	50.53	50.64	46.99	50.44	47.94
75%	75.01	74.97	72.41	74.91	72.31
90%	89.96	89.72	88.95	89.98	87.75
95%	95.11	94.58	94.99	95.13	93.16
99%	99.04	98.70	99.48	99.06	98.14
$n = 500$					
50%	49.25	49.34	48.67	49.48	46.61
75%	74.50	74.29	73.85	74.28	70.91
90%	90.02	89.56	89.22	89.99	86.47
95%	95.05	94.77	94.77	95.13	92.52
99%	99.01	99.01	99.03	99.04	98.23

Table 1.9: Estimated coverage probabilities of expected shortfall at 95%.

	SwiZs		Boot		BA		RSwiZs		RBoot	
α	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
$n = 35$										
50%	2.19	1.85	7.52	6.04	0.26	0.34	1.89	1.64	7.97	6.73
75%	4.79	3.92	216.08	179.86	0.46	0.54	3.84	3.37	27.20	23.62
90%	11.18	8.56	9710.48	8673.53	1.31	1.09	6.96	5.97	86.42	75.85
95%	24.30	18.00	2.55×10^4	2.18×10^4	8.99	8.89	9.89	7.92	161.13	142.10
99%	2488.08	1849.66	1.19×10^5	1.05×10^5	3.18×10^9	3.30×10^9	22.62	17.10	435.99	401.28
$n = 50$										
50%	1.78	1.51	3.61	2.98	0.39	0.42	1.56	1.34	5.04	4.20
75%	3.60	2.97	10.55	8.78	0.66	0.68	3.11	2.65	14.89	12.37
90%	6.78	5.41	642.67	551.95	1.22	0.94	5.57	4.83	44.67	38.37
95%	10.78	8.38	7.40×10^3	6.31×10^3	6.13	5.27	7.80	6.70	84.42	73.24
99%	54.20	39.06	5.57×10^4	4.82×10^4	1.09×10^7	1.04×10^7	15.60	12.65	231.61	202.96
$n = 100$										
50%	1.26	1.06	1.69	1.39	0.64	0.60	1.19	1.01	2.73	2.27
75%	2.32	1.92	3.32	2.74	1.08	1.02	2.23	1.87	6.01	5.01
90%	3.74	3.04	6.28	5.20	1.55	1.36	3.67	3.03	13.00	10.88
95%	4.92	3.94	10.30	8.58	1.93	1.54	4.89	4.00	22.10	18.69
99%	8.58	6.63	181.34	153.63	20.11	16.79	8.41	6.95	64.18	55.35
$n = 150$										
50%	1.02	0.86	1.21	1.01	0.71	0.62	1.00	0.85	2.02	1.68
75%	1.82	1.52	2.24	1.88	1.23	1.08	1.80	1.52	4.00	3.35
90%	2.78	2.30	3.71	3.11	1.78	1.59	2.80	2.32	7.50	6.28
95%	3.52	2.89	5.05	4.26	2.12	1.90	3.58	2.95	11.12	9.34
99%	5.38	4.35	10.59	8.97	2.86	2.27	5.62	4.52	26.58	22.47
$n = 250$										
50%	0.78	0.66	0.85	0.72	0.64	0.55	0.79	0.66	1.45	1.21
75%	1.36	1.15	1.52	1.29	1.13	0.96	1.38	1.16	2.68	2.24
90%	2.01	1.69	2.34	1.99	1.68	1.44	2.07	1.72	4.41	3.68
95%	2.48	2.08	2.97	2.52	2.07	1.78	2.56	2.12	5.94	4.97
99%	3.56	2.92	4.72	4.01	2.96	2.55	3.69	3.01	10.84	9.10
$n = 500$										
50%	0.55	0.46	0.57	0.48	0.50	0.42	0.56	0.47	0.97	0.81
75%	0.94	0.80	0.99	0.84	0.87	0.74	0.96	0.81	1.71	1.43
90%	1.37	1.16	1.47	1.25	1.27	1.08	1.41	1.18	2.63	2.20
95%	1.66	1.40	1.80	1.53	1.54	1.32	1.71	1.43	3.31	2.78
99%	2.27	1.90	2.55	2.16	2.16	1.83	2.35	1.95	5.05	4.22

Table 1.10: Estimated median interval length.

	SwiZs: mean		SwiZs: median		MLE		AB		RSwiZs: mean		RSwiZs: median		WMLE	
	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
Mean bias														
$n = 35$	2511.13	2226.09	2504.27	2230.19	2492.15	2241.82	-1.38×10^{12}	-1.34×10^{12}	13.33	11.50	13.38	11.53	13.78	12.10
$n = 50$	832.02	739.28	829.87	739.77	827.45	742.50	-1.54×10^{11}	-1.55×10^{11}	5.99	5.19	6.07	5.22	6.52	5.70
$n = 100$	45.96	37.47	45.71	37.28	45.81	37.48	-6.65×10^{8}	-5.22×10^{8}	1.20	1.03	1.26	1.05	1.72	1.47
$n = 150$	1.03	0.91	0.96	0.82	1.06	0.92	-1.60×10^{4}	-1.48×10^{4}	0.48	0.42	0.52	0.43	0.96	0.82
$n = 250$	0.17	0.15	0.15	0.12	0.21	0.18	-0.02	-0.02	0.20	0.18	0.21	0.17	0.62	0.53
$n = 500$	0.08	0.07	0.07	0.06	0.10	0.08	0.00	0.00	0.08	0.08	0.08	0.06	0.45	0.39
Median bias														
$n = 35$	0.4583	0.4894	0.0538	0.0276	0.5885	0.4654	-1.5551	-1.2966	0.2523	0.3257	0.0561	0.0309	0.9571	0.7846
$n = 50$	0.2083	0.2374	0.0250	0.0197	0.3684	0.3008	-1.1319	-0.9168	0.1691	0.2039	0.0335	0.0213	0.7112	0.5986
$n = 100$	0.0801	0.0824	0.0191	0.0135	0.1770	0.1389	-0.4093	-0.3267	0.0813	0.0905	0.0228	0.0195	0.5025	0.4289
$n = 150$	0.0358	0.0434	0.0051	0.0021	0.1011	0.0851	-0.2259	-0.1848	0.0385	0.0470	0.0063	0.0041	0.4140	0.3623
$n = 250$	0.0151	0.0265	-0.0022	0.0028	0.0541	0.0521	-0.1255	-0.1011	0.0184	0.0268	-0.0017	0.0029	0.3686	0.3268
$n = 500$	0.0129	0.0150	0.0050	0.0046	0.0331	0.0275	-0.0560	-0.0473	0.0145	0.0163	0.0049	0.0034	0.3449	0.3056
Root mean squared error														
$n = 35$	17263.26	15552.08	17223.34	15587.83	17137.54	15667.69	2.97×10^{13}	2.95×10^{13}	59.16	50.35	59.00	50.44	58.45	50.95
$n = 50$	7996.07	7382.94	7982.45	7395.00	7957.28	7418.68	5.15×10^{12}	5.62×10^{12}	27.55	24.08	27.52	24.13	27.32	24.35
$n = 100$	1331.57	1055.16	1330.24	1056.18	1328.51	1057.59	4.41×10^{10}	3.36×10^{10}	6.15	5.21	6.22	5.27	6.26	5.37
$n = 150$	36.30	32.42	36.27	32.44	36.24	32.48	1.11×10^{6}	1.06×10^{6}	2.46	2.13	2.56	2.20	2.70	2.34
$n = 250$	0.77	0.66	0.75	0.63	0.78	0.66	0.58	0.49	0.92	0.79	1.01	0.85	1.26	1.07
$n = 500$	0.46	0.39	0.46	0.38	0.47	0.40	0.42	0.35	0.49	0.41	0.50	0.42	0.77	0.66
Mean absolute deviation														
$n = 35$	2.1893	2.0002	1.5119	1.2537	2.0914	1.7082	0.5845	0.3672	1.7446	1.4744	1.5891	1.2890	2.5445	2.0878
$n = 50$	1.5636	1.4044	1.2510	1.0720	1.5649	1.3200	0.4261	0.3293	1.3908	1.2241	1.2901	1.0831	1.9384	1.6231
$n = 100$	0.9693	0.8220	0.8979	0.7479	1.0042	0.8306	0.5443	0.4800	0.9576	0.8300	0.9091	0.7685	1.2615	1.0552
$n = 150$	0.7571	0.6546	0.7291	0.6191	0.7807	0.6627	0.5752	0.4942	0.7685	0.6633	0.7396	0.6308	0.9975	0.8454
$n = 250$	0.5871	0.4942	0.5737	0.4782	0.5991	0.4995	0.5058	0.4256	0.5959	0.4984	0.5810	0.4827	0.7737	0.6368
$n = 500$	0.4084	0.3440	0.4041	0.3390	0.4130	0.3456	0.3818	0.3200	0.4127	0.3516	0.4076	0.3452	0.5295	0.4502

Table 1.11: Performances of point estimators.

1.D.2 Random intercept and random slope linear mixed model

	SwiZs	Parametric bootstrap
$N = 25$	1.87	0.20
$N = 100$	6.49	0.73
$N = 400$	35.60	4.58
$N = 1,600$	245.59	37.80

Table 1.12: Average computational time in seconds to approximate a distribution on $S = 10,000$ points.

			SwiZs					parametric bootstrap		
α	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ
					$n = 5$ $m = 5$					
50%	51.78	53.87	48.54	54.18	70.38	42.37	43.61	44.60	32.27	28.10
75%	76.89	78.87	73.58	81.67	89.09	64.17	66.19	66.20	48.35	41.80
90%	91.87	92.93	88.89	94.10	98.80	78.38	81.94	81.07	61.72	46.87
95%	96.45	97.04	94.32	97.83	99.98	84.58	88.45	86.61	68.68	47.30
99%	99.54	99.71	98.73	99.87	100.00	91.93	95.40	93.54	79.03	47.61
					$n = 10$ $m = 10$					
50%	50.10	51.20	50.70	50.65	62.48	46.25	45.37	50.05	40.01	39.84
75%	75.16	77.08	74.92	75.64	85.74	69.81	68.68	74.48	60.54	59.68
90%	90.38	92.03	90.20	90.61	95.49	84.81	84.32	88.65	75.01	73.29
95%	95.23	96.40	95.23	94.96	97.86	90.71	90.32	93.95	81.30	79.29
99%	99.16	99.54	99.25	99.09	99.64	96.45	96.76	98.41	89.37	84.71
					$n = 20$ $m = 20$					
50%	50.78	49.10	49.97	49.74	49.85	49.03	47.58	49.63	45.40	45.75
75%	75.28	74.45	75.24	74.89	75.88	73.08	71.87	75.06	67.66	66.98
90%	90.06	89.79	89.95	90.28	90.75	87.59	87.02	89.73	81.76	81.83
95%	95.05	94.83	94.79	95.06	95.97	93.10	92.69	94.59	87.48	87.52
99%	98.96	98.97	98.93	98.90	99.50	97.77	97.82	98.75	94.20	94.15
					$n = 40$ $m = 40$					
50%	49.52	48.48	49.80	52.42	53.19	49.41	48.92	49.94	47.47	47.95
75%	74.70	72.86	75.27	77.89	78.39	74.22	73.34	75.63	70.93	71.46
90%	90.07	88.10	89.69	91.81	92.46	89.30	87.99	89.70	85.62	86.34
95%	95.15	94.09	94.71	96.27	96.59	94.37	93.65	94.82	91.29	91.82
99%	99.01	98.62	98.99	99.37	99.43	98.56	98.39	98.90	96.80	96.67

Table 1.13: Estimated coverage probabilities.

	SwiZs					parametric bootstrap				
α	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ
$n = 5$ $m = 5$										
50%	0.3303	0.2243	0.4976	1.2050	0.1755	0.2712	0.1728	0.4453	1.5575	0.0005
75%	0.5940	0.3882	0.8552	2.0974	0.4491	0.4606	0.2947	0.7607	3.5624	0.0012
90%	0.9314	0.5682	1.2436	3.1286	1.1761	0.6577	0.4217	1.0909	12.9753	0.0024
95%	1.1956	0.6934	1.5222	3.9149	3.7094	0.7845	0.5031	1.3051	13.9626	0.0036
99%	1.8698	1.0031	2.3468	9.8944	8.6739	1.0290	0.6623	1.7335	15.3409	0.0070
$n = 10$ $m = 10$										
50%	0.2230	0.1198	0.2136	0.7311	1.0080	0.2038	0.1069	0.2099	0.7676	1.6745
75%	0.3902	0.2068	0.3638	1.2540	1.8614	0.3471	0.1818	0.3594	1.3370	8.6134
90%	0.5817	0.3008	0.5210	1.8131	2.9290	0.4953	0.2601	0.5144	1.9844	11.7988
95%	0.7162	0.3658	0.6218	2.1764	3.9196	0.5887	0.3097	0.6140	2.4462	12.6107
99%	1.0284	0.5130	0.8177	2.8992	7.9667	0.7745	0.4075	0.8055	3.6688	13.8600
$n = 20$ $m = 20$										
50%	0.1547	0.0699	0.1006	0.4750	0.5665	0.1482	0.0674	0.0998	0.4733	0.6557
75%	0.2672	0.1205	0.1718	0.8065	0.9934	0.2530	0.1149	0.1708	0.8102	1.1462
90%	0.3900	0.1752	0.2455	1.1499	1.4857	0.3622	0.1643	0.2447	1.1655	1.7189
95%	0.4718	0.2117	0.2926	1.3701	1.8096	0.4311	0.1957	0.2918	1.3964	2.1535
99%	0.6436	0.2894	0.3833	1.8121	2.4686	0.5645	0.2569	0.3825	1.8686	3.4277
$n = 40$ $m = 40$										
50%	0.1056	0.0452	0.0490	0.2816	0.1124	0.1056	0.0451	0.0493	0.3194	0.3628
75%	0.1810	0.0772	0.0834	0.4466	0.3469	0.1804	0.0770	0.0839	0.5429	0.6249
90%	0.2596	0.1107	0.1191	0.6923	0.6031	0.2576	0.1102	0.1197	0.7759	0.9014
95%	0.3100	0.1323	0.1420	0.8523	0.7672	0.3070	0.1313	0.1423	0.9257	1.0804
99%	0.4094	0.1747	0.1870	1.1467	1.1309	0.4020	0.1724	0.1864	1.2163	1.4420

Table 1.14: Estimated median interval length.

	SwiZs: mean					SwiZs: median					Maximum likelihood				
	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ
Mean bias×100															
$N = 25$	-0.0647	-0.3827	-3.1193	1.8554	1.5502	-0.0761	-0.3732	-2.4630	6.4149	3.3175	-0.0708	-0.4203	-1.2985	-5.8224	-0.3807
$N = 100$	0.2843	-0.0320	-0.2911	2.4583	0.6119	1.6374	-0.1452	0.7182	-1.8475	1.8127	0.0685	0.0314	-0.0166	-2.8806	-0.6425
$N = 400$	0.0163	0.0374	0.0739	1.2927	0.0944	0.0149	0.0386	0.0514	0.9056	0.1565	0.0245	0.0417	0.0133	-1.3425	-0.2785
$N = 1,600$	0.0010	0.0385	0.0183	-0.9811	-0.2965	-0.0011	0.0394	0.0120	-1.1600	-0.2121	0.0130	0.0343	-0.0021	-0.6265	-0.1253
Median bias×100															
$N = 25$	-0.0341	-0.2171	-3.8669	-6.5130	-0.0876	-0.0018	-0.2114	-3.3736	-0.8483	0.0121	0.0327	-0.2932	-2.1012	-10.1138	-3.9990
$N = 100$	0.4345	0.0289	-0.4759	0.1208	-0.0951	5.3959	-1.4459	0.5589	-0.7598	0.0354	0.1838	0.0069	-0.1815	-4.8730	-1.2975
$N = 400$	0.0020	-0.0378	0.0422	0.4196	-0.1116	0.0149	-0.0286	0.0211	-0.0405	-0.0068	-0.0140	-0.0261	-0.0220	-2.1176	-0.4517
$N = 1,600$	-0.0332	0.0500	0.0082	-1.0639	-0.1813	-0.0060	0.0543	0.0041	-0.0818	-0.0021	-0.0098	0.0480	-0.0098	-1.1378	-0.1833
Root mean squared error×100															
$N = 25$	24.6914	16.0625	9.2357	27.0499	6.2389	24.7198	16.0766	8.6916	24.2014	8.3432	24.7291	16.0853	8.1605	18.5249	6.8108
$N = 100$	16.4663	8.8542	3.9374	14.7976	3.5251	14.3449	7.7017	4.1388	11.5080	3.3680	16.5630	8.7967	3.8703	12.0714	3.1774
$N = 400$	11.4174	5.2549	1.8779	9.1330	1.8623	11.4174	5.2550	1.8752	8.9859	1.7515	11.4182	5.2554	1.8689	8.2404	1.7092
$N = 1,600$	7.8721	3.4528	0.9119	4.7681	0.6698	7.9083	3.4524	0.9117	4.4706	0.5759	7.8981	3.4532	0.9110	5.7583	1.0216
Mean absolute deviation×100															
$N = 25$	24.4139	15.8780	8.2892	23.3025	0.6468	24.4872	15.9113	8.1000	17.1094	0.2293	24.4752	15.9014	7.8528	15.1530	0.0015
$N = 100$	16.7958	8.9936	3.8427	13.3232	2.8386	13.0610	6.2264	2.9059	8.0151	1.4453	16.9915	8.9079	3.8351	10.8654	3.0194
$N = 400$	11.2283	5.3202	1.8651	8.8004	1.8018	11.2417	5.3225	1.8695	8.8299	1.4204	11.2634	5.3160	1.8653	7.8895	1.6541
$N = 1,600$	7.9220	3.4259	0.9115	4.3804	0.5033	7.9954	3.4277	0.9108	0.2978	0.0214	7.9745	3.4325	0.9082	5.7040	0.9952

Table 1.15: Performances of point estimators

	Coverage probability					Median interval length				
α	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ	β_0	β_1	σ^2_ϵ	σ^2_α	σ^2_γ
$n = 5$ $m = 5$										
50%	43.16	44.94	48.66	40.42	36.49	0.2770	0.1791	0.1043	0.1868	0.0375
75%	67.51	69.17	73.83	64.17	70.73	0.4942	0.3180	0.1836	0.3625	0.0945
90%	83.68	86.75	88.79	81.88	96.33	0.7612	0.4897	0.2764	0.6358	0.2010
95%	90.37	93.23	93.83	88.93	98.88	0.9671	0.6226	0.3431	0.8982	0.3095
99%	97.04	98.93	98.54	96.95	99.75	1.4991	0.9746	0.4982	1.8138	0.7069
$n = 10$ $m = 10$										
50%	46.38	45.98	50.75	45.86	44.91	0.2060	0.1082	0.0525	0.1422	0.0383
75%	70.85	71.03	75.36	70.65	68.84	0.3591	0.1888	0.0901	0.2583	0.0690
90%	87.23	87.08	90.04	86.58	85.82	0.5321	0.2806	0.1304	0.4088	0.1078
95%	93.20	93.27	95.12	92.37	93.09	0.6534	0.3449	0.1569	0.5299	0.1392
99%	98.41	98.53	98.95	98.02	99.59	0.9264	0.4903	0.2111	0.8593	0.2265
$n = 20$ $m = 20$										
50%	49.20	47.62	49.92	48.00	47.31	0.1491	0.0677	0.0251	0.1048	0.0216
75%	73.66	72.54	75.09	72.49	72.86	0.2571	0.1168	0.0429	0.1845	0.0381
90%	88.70	88.34	89.97	88.33	88.10	0.3742	0.1700	0.0616	0.2774	0.0573
95%	94.09	94.02	94.81	93.80	93.72	0.4524	0.2055	0.0735	0.3445	0.0712
99%	98.56	98.61	98.94	98.40	98.59	0.6167	0.2801	0.0972	0.5019	0.1038
$n = 40$ $m = 40$										
50%	49.46	49.32	49.79	48.67	49.01	0.1060	0.0452	0.0122	0.0748	0.0136
75%	74.46	73.78	75.28	73.52	74.77	0.1819	0.0776	0.0209	0.1295	0.0236
90%	89.88	88.76	89.70	88.89	89.83	0.2623	0.1119	0.0299	0.1899	0.0346
95%	94.95	94.28	94.85	94.22	94.71	0.3148	0.1343	0.0356	0.2310	0.0420
99%	98.98	98.86	98.99	98.77	98.82	0.4212	0.1797	0.0468	0.3194	0.0582

Table 1.16: Asymptotic results

1.D.3 M/G/1 queueing model

	SwiZs			Indirect inference			Parametric bootstrap		
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
50%	46.92	38.68	56.73	40.59	9.95	54.59	18.31	10.23	20.96
75%	71.56	55.41	81.80	68.01	34.11	84.50	32.70	20.96	37.38
90%	87.55	67.77	94.47	87.62	57.13	96.04	48.62	35.24	53.71
95%	93.16	74.78	97.97	94.66	70.22	98.75	57.05	46.03	63.21
99%	98.17	90.06	99.90	98.84	94.89	99.94	71.99	65.43	77.64

Table 1.17: Estimated coverage probabilities.

	SwiZs			Indirect inference			Parametric bootstrap		
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
50%	0.0235	0.0805	0.1379	0.0382	0.0468	0.1368	0.0263	0.0420	0.1134
75%	0.0404	0.1467	0.2357	0.0911	0.0978	0.2389	0.0460	0.0757	0.2051
90%	0.0585	0.2207	0.3378	0.1563	0.1914	0.3835	0.0708	0.1185	0.3131
95%	0.0705	0.2733	0.4032	0.2225	0.2952	0.5432	0.0895	0.1533	0.3855
99%	0.0952	0.3934	0.5407	0.5331	0.7152	1.6084	0.1327	0.2514	0.5562

Table 1.18: Estimated median interval length.

	SwiZs: starting value is $\boldsymbol{\theta}_0$			SwiZs: sample size is $n = 1,000$.		
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
50%	50.22	58.64	49.98	50.07	46.06	49.37
75%	75.24	91.25	74.24	75.24	71.82	74.77
90%	90.52	99.82	89.55	89.73	89.84	89.49
95%	95.37	100.00	94.87	94.81	95.41	94.69
99%	99.09	100.00	99.02	98.95	99.28	99.10

Table 1.19: Estimated coverage probabilities under different conditions than Table 1.17.

	SwiZs: mean			SwiZs: median			Indirect inference			Indirect inference: mean			Indirect inference: median		
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
Mean bias	0.0037	-0.0149	0.0006	0.0057	-0.0096	0.0002	2×10^{90}	3×10^{90}	1.6107	0.0309	0.0254	3×10^{89}	0.0157	0.0297	0.0201
Median bias	0.0026	-0.0219	-0.0044	0.0046	-0.0157	-0.0041	0.0135	0.0270	0.0181	0.0295	0.0235	0.0772	0.0150	0.0257	0.0200
RMSE	0.0197	0.0764	0.0890	0.0200	0.0762	0.0888	2×10^{92}	3×10^{92}	135.72	0.0451	0.0976	3×10^{91}	0.0254	0.1041	0.0851
MAD	0.0192	0.0705	0.0884	0.0190	0.0718	0.0882	0.0307	0.1069	0.1405	0.0365	0.0918	0.1109	0.0182	0.0968	0.0823

Table 1.20: Performances of point estimator.

2

Bounded-Influence Robust Estimation of Copulae

Always look on the bright side of life.

– Monty Python, *Life of Brian*

2.1 Introduction

Copula functions are very convenient for modeling multivariate observations. One of their clear advantage is that they offer to the modeler the possibility to focus separately on the marginal distributions and the multivariate model. Moreover, a wide variety of copula functions exist (see [Nel06; Joe14] for monographs) that are convenient for modeling specific characteristics of the the joint model. This flexibility brought by copula models has made it a popular tool in fields that require specific care for tails and asymmetric dependencies of the joint distribution, a situation very often encountered in application to real data.

In welfare economics, well-being consists of many dimensions such as income, health and education. Even within such dimensions, modifications in time of their distribution is of interest, as is done in, for example, [DL01] who measure pre- and post-tax living standard distributions using copulae, [BR09] who study earnings mobility in France by proposing a model of earnings dynamics including a transition probability modeled by means of a copula, [VGC10] who use copulae for income mobility (see also [CG17] who propose a joint modeling of income and consumption with a copula function to capture the dependence structure) or [DG12] who take advantage of a copula model for measuring household financial fragility. In financial economics, for example [Pat06] use copula to model asymmetric exchange rate dependence and [CHV08] to model international financial returns. Inequality indices, which are functions of the distributions, have also been extended to multivariate versions, using copulae, either between different dimensions or within the same dimensions. For example, [ANG06] propose multivariate measures of inequality, [Qui09] measures income-related inequalities in health using copulae (see also [JSVK15]), [FL12] use copulae in multidimensional poverty evaluation and [Dec14] propose more generally a class of dependence measures between well-being dimensions (see also [Atk11]).

Copula based models have also been used, for example in health sciences by [SL95] to study survival risk factors in AIDS, by [SLY09] to study treatment effect in multiple sclerosis, by [DW10] to study burn injuries, by [He+12] for the analysis of secondary phenotypes in case-control genetic association studies, by [PCJ12] to study headache severity, by [St15] to study comorbidity of chronic diseases in the elderly patients, in natural and engineering sciences by [Smi+10] for modeling and forecasting electricity load at an intraday resolution, by [GS11] to measure spatial dependence in wind for optimal wind power allocation and by [ZK16] for predicting disruption length in public transportation.

Multivariate generalized linear models can also be modeled using copulae as is done for example in [Son00; Joe05; SLY09; DW10; NJR11; MF11; He+12] and in the survey of [Nik13] for the discrete case. Copulas have also been proposed as an alternative to the covariance matrix in factor (correlation structure) analysis for dimension reduction; see e.g. [KK09; KJ13; KHK15; OP17; KHG18].

Given the complexity of (parametric) copula model, several estimation methods have been proposed. An alternative to the maximum likelihood estimator (MLE) is the method of inference functions for margins (IFM), a two-step MLE proposed by [JX96; Joe97]. Indeed, even though the IFM has a reduced (asymptotic) efficiency ([Joe05]), direct optimization of the likelihood function can be numerically difficult and a sequential approach is often preferred. In the first step, a parametric (or non-parametric) estimator of marginal distributions is computed and in the second step, given this estimator, the parameters of the copula are estimated by the MLE. With a non-parametric approach the empirical distribution function (EDF) is used [GGR95; SL95; CFT06], essentially to prevent the risk of misspecification of margins (see e.g. [KSS07]) at a cost of an efficiency loss compared to the IFM (when the model is correctly specified).

Even so, multivariate copula estimation can become computationally challenging in even moderate dimensions (e.g. more than three). The extension to the multivariate setting can easily be made using elliptical copulae (e.g. Gausssian or Student), but the number of parameters can become very large and the postulated dependence is symmetric. To avoid this curse of dimensionality, a possible extension to high dimensional settings is the construction of multivariate model using only bivariate copulae, which is called a pair-copula construction (PCC). This approach makes use of vine families proposed in [Joe96] and [BC02] (see [KJ11] for a monograph), and [PCJ12] sets a general estimation framework (see also [KC06; Aas+09; Fis+09; BCA12; Hob13]). A similar approach is the use of Hierarchical Archimedean copulae (HAC) [Joe97]; for simulation and inferential methods, see e.g. [McN08; Hof08; MN09; Hof10; HMM12; Hof12; AGN12; OOS13; Rez15; GHH17; DC17; Uyt18]; see also [Bre14] hierarchical Kendall copulae, a similar concept. Alternatively, [Son+05] proposes an algorithm for the MLE based on a convenient decomposition of the likelihood function and [OP13] extends the simulated method of moments ([McF89; PP89]) to estimate copula models based on empirical bivariate dependence measures (that do not depend on the specification of the model).

An important drawback of most of the estimation methods proposed so far is that they rely on the exact specification of the (parametric) model. In other terms, if the data shows small deviations from the assumed model, these estimators can be dramatically biased. These small deviations take the form of data contamination (e.g. outliers) and robust statistics can limit their influence on the estimation and testing procedures ([Hub+64; Ham+86]). At the marginal level, the benefits of a robust estimation approach has been shown with parametric estimation of densities (see e.g. [VFR94a; AVH05; DVF06; Van+07; Bí14; Brz16]). For generalized linear models (GLM), a robust estimation approach has been motivated for example in [CP93; Chr94; CR01; VF02; RC03;

MY04; CR06; C08́; BY11; BBR13; VY14; ACH14; Her+09; HFZ05; MVF06].

Hence, at the marginal level robust estimation is available while at the copula level, surprisingly, robust estimation approaches have received little attention. [MMN07] define estimator as the solution of weighted goodness-of-fit measures between empirical and estimated parametric copula, yet asymptotic properties of the resulting estimator remain unclear. [DM11] approach is based on the method of likelihood depth (see also [RH99]), but is limited to two bivariate copula and shows difficulties to generalize to other models. [KL13] concentrate their work on copula-based dynamic models, they proposed to replace the MLE of copula by minimum density power divergence estimator (see also [Bas+98]). However, none of the previous work treats issues with model contamination affecting both copula and margins.

We propose here to generalize the IFM estimator to a two-steps M-estimator [Hub+64; HR09] that includes bounded Influence Function (IF) estimators [Ham74; Ham+86]. Indeed, as shown in [ZGR12], a two-step estimator has a bounded IF only if the estimation equations at each stage are bounded. Bounded IF estimators are usually numerically challenging, except for symmetric models, since they need to include, in the estimating equations, a centrality quantity (for Fisher consistency) that consists of multiple integrals. In order to alleviate the numerical aspects, so that robust copula estimation can be performed in high dimensional settings, we use the indirect inference framework to provide consistent robust estimators that are simple to compute as proposed in [Gue+18b].

This paper is organized as follows. In Section 2.2, we set a general framework for joint dependence modeling, within which we propose a multi-step indirect inference estimator with focus on the dependence parameters. In Section 2.3, we set the general conditions under which the consistency and asymptotic normality of the proposed estimator are demonstrated. Some conditions are particularly hard to verify, therefore this section is also ponctuated by alternative propositions. The Section 2.4 covers practical aspects of implementation, in particular, it gives computationally efficient algorithm for solving a point estimator and a first-order approximation to facilitate bootstrap procedure. In Section 2.5, an IF is developed that permits to appreciate the robustness of estimators in this framework. Inline with these findings, in Section 2.6 is discussed the weighted maximum likelihood estimator as the most reasonable approach in terms of bias and mean squared errors when the data generating mechanism deviates from the assumed model. In Section 2.7, we present a simulation study to compare performances of the proposed estimators in two situations: first, when outlying observations contaminate the data, second, when the assumed model is complex, that is the likelihood function has no known analytical expression. Eventually, in Section 2.8 an application to real data is presented. The income mobility (in time) is studied for the Swiss Household Panel, by means of related inequality measures derived from classical and robust copula estimators.

2.2 A general indirect inference framework for multivariate models with joint dependence

We consider the same general class of data-generating process used by [CF06; Cha+09; OP13; Rém17]. This class permits to model in an unified fashion time-varying conditional mean and variance together with any parametric marginal models. Thus, it covers many commonly used multivariate models such as copula GARCH, multivariate ARMA models, multivariate stochastic volatility models, multivariate regression models, and so on. Let the d-dimensional multivariate random sequence

$$\mathbf{y}_t = \boldsymbol{\sigma}_t^{-1}(\boldsymbol{\nu})\{\mathbf{x}_t - \boldsymbol{\mu}_t(\boldsymbol{\nu})\}, \quad t \in \mathcal{T} = \{1, \ldots, n\}, \tag{2.1}$$

to be jointly independent and identically distributed according to an assumed probability model G_Y. More specifically, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1 \ \ldots \ \boldsymbol{\mu}_n]$ is a $\mathbb{R}^d \times \mathbb{R}^n$ matrix where each column $\boldsymbol{\mu}_t$ is $\mathcal{F}_{t-1}$-measurable and independent of $\mathbf{y}_t$. $\mathcal{F}_t$ denotes the sigma field containing all information from the past sequences up to t and possibly depending on fixed covariates (omitted in the notation). $\boldsymbol{\sigma}$ is a $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^n$ tensor where each slice $\boldsymbol{\sigma}_t = \operatorname{diag}(\sigma_{1t}, \ldots, \sigma_{dt})$ is a diagonal square matrix also assumed $\mathcal{F}_{t-1}$-measurable and independent of $\mathbf{y}_t$. The vector parameter $\boldsymbol{\nu} \in \mathcal{V} \subset \mathbb{R}^r$ characterises both the dynamic of each univariate random process $\{x_{jt} : t \in \mathcal{T}, j \in \mathcal{J}\}$, $\mathcal{J} = \{1, \ldots, d\}$, and the marginal cumulative distribution function (*cdf*) of the innovation, $F_Y(\mathbf{y}, \boldsymbol{\nu}) = [F_1(\mathbf{y}_1, \boldsymbol{\nu}_1)^T \ \ldots \ F_d(\mathbf{y}_d, \boldsymbol{\nu}_d)^T]^T$, where $\mathbf{y} = [y_{jt}]_{j \in \mathcal{J}, t \in \mathcal{T}} \in \mathbb{R}^n \times \mathbb{R}^d$ is a real-valued matrix. This setup allows conveniently to separate between the marginal sequential dependence and the joint multivariate dependence of the processes. Indeed, we have by Sklar's Theorem ([Skl59]) that $G_Y(\mathbf{y}) = C(F_Y(\mathbf{y}, \boldsymbol{\nu}), \boldsymbol{\theta})$, where $C = \{C_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is a copula model indexed by parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. Our interest is in the estimation and the inference of $\boldsymbol{\theta}$. However, if $\boldsymbol{\nu}$ is unknown, $F_Y(\mathbf{y}, \boldsymbol{\nu})$ depends upon the estimation of $\boldsymbol{\nu}$. In view of the foregoing, we consider the situation, usually encountered in practice, of estimating $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ separately in a multistep procedure.

We denote generically a real-valued random matrix that is an assumely known measurable function of the observations, the marginal parameters and the dependence paramters by $\mathbf{u} : \mathbf{X} \times \mathcal{V} \times \boldsymbol{\Theta} \to [0,1]^d$. In particular, we have in mind that $\mathbf{u}(\mathbf{x}, \boldsymbol{\nu}; \boldsymbol{\theta}) = F_Y(\mathbf{y}, \boldsymbol{\nu})$ is a matrix of *cdf* and $\mathbf{u}(\mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\nu})$ is a matrix of jointly dependent standard uniform variates identically and idenpendently distributed according to $C_{\boldsymbol{\theta}}$. For convenience, the function takes the index from the observation so $\mathbf{u}_t(\boldsymbol{\nu}) = \mathbf{u}(\mathbf{x}_t, \boldsymbol{\nu}; \boldsymbol{\theta}_1)$, for a point $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$, and $\mathbf{u}_t(\boldsymbol{\theta}) = \mathbf{u}(\mathbf{x}_t, \boldsymbol{\theta}; \boldsymbol{\nu}_1)$, for a point $\boldsymbol{\nu}_1 \in \mathcal{V}$.

Our inferential problem concerns estimating sequentially $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$. Within the scope of this chapter, our focus is driven by $\boldsymbol{\theta}$, and thereby $\boldsymbol{\nu}$, considered as "nuisance parameters", receive a less exhaustive treatment. It is nonetheless not taken as known, the fact that the nuisance parameters need to be estimated is of primordial importance for example when considering the robust properties of the estimator of $\boldsymbol{\theta}$ in Section 2.5.

We focus our attention to the situation where estimating $\boldsymbol{\theta}$ is a complex problem, as for example the likelihood function is intractable or more generally the estimating equation has no analytic expression. This situation is in fact unrestrictive and broader than usually admitted. To tackle these issues, we use the general framework of indirect inference ([GMR93; Smi93]) refined for our purpose. The basis of this method consists of two successive steps. First, an auxiliary parameter $\boldsymbol{\pi} \in \boldsymbol{\Pi} \subset \mathbb{R}^q$, $q \geq p$, is obtained by finding the roots of the following Z-estimating equation

$$\hat{\boldsymbol{\pi}}_{n^*} = \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}}{\operatorname{argzero}} \frac{1}{n^*} \sum_{t \in \mathcal{T}^*} \boldsymbol{\phi}\left(\mathbf{u}_t\left(\hat{\boldsymbol{\nu}}_n\right), \boldsymbol{\pi}\right) = \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}}{\operatorname{argzero}} \, \boldsymbol{\Phi}_{n^*}\left(\hat{\boldsymbol{\nu}}_n, \boldsymbol{\pi}\right), \tag{2.2}$$

where $\mathcal{T}^* \subseteq \mathcal{T}$ denotes the set of jointly observed variables and is assumed non-empty with cardinality $n^* = |\mathcal{T}^*| > 0$. More precisely, $n^* = \lfloor n\rho^* \rfloor$ with $0 < \rho^* \leq 1$, the least positive integer smaller than $n\rho^*$. There are several reasons why the joint sample size n^* might be smaller than n (and thus $\rho^* < 1$): marginally or jointly missing data, data considered as fixed for estimation purposes, and so on. With this point of view, $1-\rho^*$ may be interpreted as the percentage loss of information due to disjoint observation. Second, the same auxiliary parameter may be obtained on an independent copy of $\mathbf{u}_t(\boldsymbol{\nu})$, say $\mathbf{u}_t(\boldsymbol{\theta})$, and the solution is obtained by matching both auxiliary estimators:

$$\hat{\boldsymbol{\theta}}_{n^*} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left\|\hat{\boldsymbol{\pi}}_{n^*} - \bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta})\right\|^2_{\boldsymbol{\Omega}}, \tag{2.3}$$

where for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\pi}}_m^{(b)}(\boldsymbol{\theta}),$$

and

$$\hat{\boldsymbol{\pi}}_m^{(b)}(\boldsymbol{\theta}) = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\operatorname{argzero}} \frac{1}{m}\sum_{t=1}^{m} \boldsymbol{\phi}\left(\mathbf{u}_t^{(b)}(\boldsymbol{\theta}),\boldsymbol{\pi}\right) = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_m(\boldsymbol{\theta},\boldsymbol{\pi}),$$

where $m = \lceil Hn^*\rceil$, $H \in \mathrm{I\!R}^+$, is the least positive integer greater than Hn^*, $B \in \mathbb{N}^+$ and $\mathbf{u}^{(b)}$ is the bth independent sample with a length m. The matrix $\boldsymbol{\Omega} \in \mathrm{I\!R}^q \times \mathrm{I\!R}^q$ in (2.3) is assumed symmetric and positive-definite with finite elements. It possibly needs to be estimated. Although not essential in theory, the function $\boldsymbol{\phi}$ is assumed to be known in analytical form, it echos the practical interest of the method.

Remark 2.1. *Several indirect inference estimators exist in the literature (see [GMR93; FZ14]): one auxiliary estimator on one large simulated sample, several auxiliary estimators on simulated samples of size n^*. Here we combine them in an unified fashion and study the consequence in next section.*

Remark 2.2. *Z-estimating equations are usually defined with ψ-function notation in the robust statistical literature (see [Hub+64; Ham+86; HR09]). Here, we are purposely using a different notation with $\boldsymbol{\phi}$ instead as we allow $\hat{\boldsymbol{\pi}}_{n^*}$ to be an inconsistent estimator of $\boldsymbol{\theta}_0$. This difference aims at emphasing the gain in generality of the proposed method.*

2.3 Asymptotic results

The following conditions are sufficient to prove the consistency of $\hat{\boldsymbol{\theta}}_{n^*}$. We start by characterizing the estimating equation $\boldsymbol{\Phi}_n$. In the next assumption we use n instead of m or n^* for simplicity. The statement should nonetheless be understood as *for all* $\rho^* \in (0,1]$ and *for all* $H \in \mathrm{I\!R}^+$ where pertinent.

Assumption 2.3 (characterization of $\boldsymbol{\Phi}_n$)**.** *The following holds:*

i. For all $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, for a sufficiently large n, $\boldsymbol{\Phi}_n(\boldsymbol{\nu},\boldsymbol{\pi})$ is continuous at each $\boldsymbol{\nu} \in \mathcal{V}$ with probability one.

ii. For all $(\boldsymbol{\theta}\times\boldsymbol{\nu}) \in \boldsymbol{\Theta}\times\mathcal{V}$, for a sufficiently large n there exist random values $A_n = \mathcal{O}_p(1)$ and $B_n = \mathcal{O}_p(1)$ such that for every $\boldsymbol{\pi}_1,\boldsymbol{\pi}_2 \in \boldsymbol{\Pi}$

$$\left\|\boldsymbol{\Phi}_n(\boldsymbol{\theta},\boldsymbol{\pi}_1) - \boldsymbol{\Phi}_n(\boldsymbol{\theta},\boldsymbol{\pi}_2)\right\| \leq A_n\left\|\boldsymbol{\pi}_1-\boldsymbol{\pi}_2\right\|, \left\|\boldsymbol{\Phi}_n(\boldsymbol{\nu},\boldsymbol{\pi}_1) - \boldsymbol{\Phi}_n(\boldsymbol{\nu},\boldsymbol{\pi}_2)\right\| \leq B_n\left\|\boldsymbol{\pi}_1-\boldsymbol{\pi}_2\right\|.$$

iii. For all $(\boldsymbol{\theta},\boldsymbol{\nu},\boldsymbol{\pi}) \in \boldsymbol{\Theta}\times\mathcal{V}\times\boldsymbol{\Pi}$, the following expectations exist and are finite

$$\mathbb{E}\|\boldsymbol{\Phi}_n(\boldsymbol{\nu},\boldsymbol{\pi})\| < \infty, \quad \mathbb{E}\|\boldsymbol{\Phi}_n(\boldsymbol{\theta},\boldsymbol{\pi})\| < \infty,$$

when n is sufficiently large.

iv. *For all* $(\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\pi}) \in \boldsymbol{\Theta} \times \mathcal{V} \times \boldsymbol{\Pi}$*, the following non-stochastic limits exists*

$$\lim_{n\to\infty} \boldsymbol{\Phi}_n(\boldsymbol{\nu}, \boldsymbol{\pi}) = \boldsymbol{\Phi}(\boldsymbol{\nu}, \boldsymbol{\pi}), \quad \lim_{n\to\infty} \boldsymbol{\Phi}_n(\boldsymbol{\theta}, \boldsymbol{\pi}) = \boldsymbol{\Phi}(\boldsymbol{\theta}, \boldsymbol{\pi}).$$

v. *For all* $(\boldsymbol{\theta}, \boldsymbol{\nu}) \in \boldsymbol{\Theta} \times \mathcal{V}$*, we have for every* $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \boldsymbol{\Pi}$ *that*

$$\boldsymbol{\Phi}(\boldsymbol{\theta}, \boldsymbol{\pi}_1) = \boldsymbol{\Phi}(\boldsymbol{\theta}, \boldsymbol{\pi}_2), \quad \boldsymbol{\Phi}(\boldsymbol{\nu}, \boldsymbol{\pi}_1) = \boldsymbol{\Phi}(\boldsymbol{\nu}, \boldsymbol{\pi}_2),$$

if and only if $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$*.*

The first condition states that $\boldsymbol{\Phi}_n$ is continuous at every $\boldsymbol{\nu} \in \mathcal{V}$. This condition is very mild and is typically verified in application. Together with Assumption 2.7 (i), it permits to employ the continuous mapping theorem ([Vaa98]) thereby faciliating the handling of the two-steps estimators $\hat{\boldsymbol{\nu}}_n$ and $\hat{\boldsymbol{\pi}}_{n^*}$ in (2.2). All the next conditions, namely stochastic Lipschitz, finite expectation, deterministic limit and identifiability are regular ones. We now charaterize the stochastic mapping $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{\pi}}_m$ in (2.3) which is specific to indirect inference. This mapping, once made deterministic by either taking the expectation or the limit, is refered in the literature to as the "binding" function ([GMR93]) or the "bridge relationship" ([JT04]) and is a key ingredient of the method.

Assumption 2.4 (characterization of $\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta})$)**.** *We have the followings:*

i. *For all* $(B, H, \rho^*) \in \mathbb{N}^+ \times \mathbb{R}^+ \times (0, 1]$*, for a sufficiently large* n *there exists a random value* $C_m = \mathcal{O}_p(1)$*, where* $m = \mathcal{O}(n)$*, such that for every* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$

$$\|\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}_1) - \bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}_2)\| \leq C_m \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| .$$

ii. *For all* $(\boldsymbol{\theta}, B, H, \rho^*) \in \boldsymbol{\Theta} \times \mathbb{N}^+ \times \mathbb{R}^+ \times (0, 1]$*, the following non-stochastic limit exists*

$$\lim_{n\to\infty} \bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}) = \boldsymbol{\pi}(\boldsymbol{\theta}).$$

iii. *For every* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$*, we have*

$$\boldsymbol{\pi}(\boldsymbol{\theta}_1) = \boldsymbol{\pi}(\boldsymbol{\theta}_2),$$

if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$*.*

The first condition requires $\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta})$ to be stochastically Lipschitz. It does not usually appear in the literature (see [GM96]) as it is a lower level condition that implies the uniform convergence of the stochastic function (see Lemma 3.4). Since the mapping $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{\pi}}_m$ is known only implicitly, this condition can in general not be verified. The next Propoposition 2.5 gives the mean to verify this assumption.

Proposition 2.5 (Lipschitz)**.** *If the implicit mapping* $\boldsymbol{\theta} \mapsto \hat{\boldsymbol{\pi}}_n$ *is continuously once differentiable and the Jacobian is bounded, i.e.* $D_{\boldsymbol{\theta}}\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}) < \infty$*, then it is Lipschitz.*

The second hypothesis of Assumption 2.4 is regular. The third condition is on the identifiability and is related to the choice of the auxiliary parameter. It is commonly assumed to hold (see [GM96]) but it is typically hard to verify since the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}$ is unknown in an explicit form. The next proposition provides a mean to imply this condition.

Proposition 2.6 (local identifiability)**.** *Let $\boldsymbol{\Theta}$ be a open convex subset of $\mathrm{I\!R}^p$. If the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}$ is continuously once differentiable in $\boldsymbol{\Theta}$ and the Jacobian $D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta})$ is nonvanishing and of full column rank at a point $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then there exists a neighborhood of $\boldsymbol{\theta}_0$ on which $\boldsymbol{\pi}$ is injective.*

Before stating the consistency of $\hat{\boldsymbol{\theta}}_{n^*}$, we require a last assumption to assess the asymptotic behaviour of the marginal esitmator $\hat{\boldsymbol{\nu}}_n$ and matrix of weight $\boldsymbol{\Omega}$ if estimated.

Assumption 2.7 (asymptotics)**.** *The following asymptotic results hold:*

i. If $\boldsymbol{\Omega} = [\omega_{kl}]_{k,l=1,\dots,q}$ is estimated by $\widehat{\boldsymbol{\Omega}}$, we have

$$\hat{\omega}_{kl} = \omega_{kl} + o_p(1),$$

elementwise, for $k, l = 1, \dots, q$.

ii. For the marginal estimators, we have

$$\hat{\boldsymbol{\nu}}_n = \boldsymbol{\nu} + o_p(1).$$

The first condition is about the consistency of $\widehat{\boldsymbol{\Omega}}$ and is a regular one. The second condition is on the consistency of the marginal estimator. Because such hypothesis highly depends on the marginal models at hand, it is simply assumed here for the sake of generality. There is a large body of literature for most models.

Theorem 2.8 (consistency)**.** *Let $\boldsymbol{\Theta}$ be compact subset of $\mathrm{I\!R}^p$. Let $\boldsymbol{\Pi}$ be a compact subset of $\mathrm{I\!R}^q$. Let $\{Q_{n^*}(\boldsymbol{\theta}) = \|\hat{\boldsymbol{\pi}}_{n^*} - \bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta})\|^2_{\widehat{\boldsymbol{\Omega}}}\}$ be sequence of real-valued function. Let $\{\hat{\boldsymbol{\theta}}_{n^*}\}$ be a sequence that nearly minimises $\{Q_{n^*}(\boldsymbol{\theta})\}$. If the Assumptions 2.3 to 2.7 hold, then any sequence $\{\hat{\boldsymbol{\theta}}_{n^*}\}$ converges weakly in probability to $\boldsymbol{\theta}_0$.*

We now turn our attention to the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{n^*}$. We start by defining quantities of interest for latter convenience. We use the following notation for the Jacobian matrices of the estimating equation with respect to $\boldsymbol{\theta}$, $\boldsymbol{\pi}$ and $\boldsymbol{\nu}$:

$$\begin{aligned}
&\mathbf{J}_n(\boldsymbol{\theta}, \boldsymbol{\pi}) \equiv D_{\boldsymbol{\theta}}\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \boldsymbol{\pi}), \\
&\mathbf{K}_n(\boldsymbol{\nu}, \boldsymbol{\pi}) \equiv D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n(\boldsymbol{\nu}, \boldsymbol{\pi}), \quad \mathbf{K}_n(\boldsymbol{\theta}, \boldsymbol{\pi}) \equiv D_{\boldsymbol{\pi}}\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \boldsymbol{\pi}), \\
&\mathbf{L}_n(\boldsymbol{\nu}, \boldsymbol{\pi}) \equiv D_{\boldsymbol{\nu}}\boldsymbol{\Phi}_n(\boldsymbol{\nu}, \boldsymbol{\pi}).
\end{aligned}$$

By convention, when one of the above quantities is evaluated to one of the following points $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, $\boldsymbol{\pi}_0 \in \boldsymbol{\Pi}$ or $\boldsymbol{\nu}_0 \in \boldsymbol{\mathcal{V}}$, the argument is omitted from the notation. For example, $\mathbf{J}_n$ means the Jacobian of $\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \boldsymbol{\pi})$ with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_0$ and $\boldsymbol{\pi}_0$. Having defined these Jacobian matrices, we now impose some restrictions on them.

Assumption 2.9 (characterization of $\mathbf{J}$, $\mathbf{K}$ and $\mathbf{L}$)**.** *We have the following:*

i. The matrices $\mathbf{J}_n(\boldsymbol{\theta}, \boldsymbol{\pi})$, $\mathbf{K}_n(\boldsymbol{\theta}, \boldsymbol{\pi})$, $\mathbf{K}_n(\boldsymbol{\nu}, \boldsymbol{\pi})$ and $\mathbf{L}_n(\boldsymbol{\nu}, \boldsymbol{\pi})$ exist and are continous.

ii. The matrices $\mathbf{J}_n$, $\mathbf{K}_n$ and $\mathbf{L}_n$ converges pointwise to $\mathbf{J}$, $\mathbf{K}$ and $\mathbf{L}$ respectively.

iii. The matrices $\mathbf{K}$ and $\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\mathbf{J}$ are nonsingular.

iv. *For a sufficiently large n there exists a random value $D_n = \mathcal{O}_p(1)$ such that for every $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{V}$ we have*

$$\|\mathbf{K}_n(\boldsymbol{\nu}_1) - \mathbf{K}_n(\boldsymbol{\nu}_2)\| \leq D_n \|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|.$$

v. *For a sufficiently large n there exists a random value E_n with $\mathbb{E}[E_n] = \mathcal{O}(1)$ and such that for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ we have*

$$\left\|\mathbf{K}_n^{-1}(\boldsymbol{\theta}_1)\mathbf{J}_n(\boldsymbol{\theta}_1) - \mathbf{K}_n^{-1}(\boldsymbol{\theta}_2)\mathbf{J}_n(\boldsymbol{\theta}_2)\right\| \leq E_n \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

The first three conditions are regular ones. The fourth condition accounts for the fact that we require $\mathbf{K}_n$ to converge uniformly because of the marginal estimator appearing in the two-step procedure. As with the fourth condition, the fifth permits to imply the uniform convergence of the Jacobian matrix of $\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})$. As remarked by [Phi12], contrary to usual estimating methods such maximum likelihood or generalized method of moments, the indirect inference uses a stochastic mapping, $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{\pi}}_m$ in our notation in (2.3), thereby requiring a more involved treatment. The reason for the form of this last condition will become clearer after the next proposition. As already remarked, the binding function is unknown in an explicit form, however its Jacobian may be derived explicitely.

Proposition 2.10 (Jacobian)**.** *If Assumption 2.9 (i,ii,iii) hold, then*

$$D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\mathbf{K}^{-1}\mathbf{J}.$$

Moreover, if we strengthen Assumption 2.9 (i,ii,iii) to hold for all n, then

$$D_{\boldsymbol{\theta}}\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\mathbf{K}_n^{-1}(\boldsymbol{\pi}_1)\mathbf{J}_n(\boldsymbol{\pi}_1),$$

where $\boldsymbol{\pi}_1 \equiv \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}_0)$.

Proposition 2.10 delivers the Jacobian matrix of the binding function but also of the stochastic mapping $\boldsymbol{\theta} \mapsto \hat{\boldsymbol{\pi}}_n$. As demonstrated in Propositions 2.5 and 2.6, it is of theoretical interest to know these Jacobian matrices. On a more practical aspect, it is also interesting to have the Jacobian of $\boldsymbol{\theta} \mapsto \hat{\boldsymbol{\pi}}_n$ as it allows to obtain the gradient of the indirect inference estimator (2.3) that may useful for any gradient-based optimization routines.

The next assumption characterizes the asymptotic distribution of the marginal estimators and restricts the asymptotic behavior of the estimating equation.

Assumption 2.11 (asymptotics II)**.** *The followings hold:*

i. *For the marginal estimators, we have that*

$$\hat{\boldsymbol{\nu}}_n = \boldsymbol{\nu}_0 + n^{-1/2}\boldsymbol{\Lambda}^{-1}\mathbf{z} + o_p(1),$$

where $\boldsymbol{\Lambda}^T\boldsymbol{\Lambda} = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 \oplus \cdots \oplus \boldsymbol{\Sigma}_d$ is a d-blocks-diagonal matrix such that $\|\boldsymbol{\Sigma}\|_\infty < \infty$ and $\mathbf{z}$ is a multivariate standard Gaussian random variable.

ii. *For all $(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\nu}) \in \boldsymbol{\Theta} \times \boldsymbol{\Pi} \times \mathcal{V}$, we have*

$$\lim_{n\to\infty} \mathbb{E}\|\boldsymbol{\Phi}_n(\boldsymbol{\nu}, \boldsymbol{\pi})\|^2\mathbf{1}\{\|\boldsymbol{\Phi}_n(\boldsymbol{\nu}, \boldsymbol{\pi})\| > \varepsilon\} = 0, \quad \lim_{n\to\infty} \mathbb{E}\|\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \boldsymbol{\pi})\|^2\mathbf{1}\{\|\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \boldsymbol{\pi})\| > \varepsilon\} = 0,$$

for every $\varepsilon > 0$, and

$$\mathrm{Cov}\left(\boldsymbol{\Phi}_n(\boldsymbol{\nu}, \boldsymbol{\pi})\right) = \mathbf{Q}, \quad \mathrm{Cov}\left(\boldsymbol{\Phi}_n(\boldsymbol{\theta}, \boldsymbol{\pi})\right) = \mathbf{Q}, \quad \|\mathbf{Q}\|_\infty < \infty,$$

when n is sufficiently large.

For the same reasons invoked after Assumption 2.7, the asymptotic normality of $\hat{\boldsymbol{\nu}}_n$ is simply assumed here for marginal estimators. We let the reader refer to the above literature. Assumption 2.11 (*ii*) is regular as it permits to invoke Lindeberg-Feller central limit theorem (see [Vaa98] for instance). Next, we present the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{n^*}$.

Theorem 2.12 (asymptotic normality)**.** *Let $\boldsymbol{\Theta}, \boldsymbol{\Pi}$ be as in Theorem 2.8 and denote by $\boldsymbol{\Theta}^\circ, \boldsymbol{\Pi}^\circ$ the interior sets assumely convex. Suppose that $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^\circ$ and $\boldsymbol{\pi}_0 \in \boldsymbol{\Pi}^\circ$. If Assumptions 2.3, 2.4, 2.7, 2.9 and 2.11 hold, then*

$$n^{1/2}\boldsymbol{\chi}^{-1/2}\left(\hat{\boldsymbol{\theta}}_{n^*} - \boldsymbol{\theta}_0\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right),$$

where

$$\boldsymbol{\chi} = \left(\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\left[\gamma^*\mathbf{Q} + \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^T\right]\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\mathbf{J}\left(\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\mathbf{J}\right)^{-T},$$

and

$$\gamma^* = \frac{1 + BH}{\rho^* BH} \geq 1.$$

This result is remarkable as all sources responsible for the loss of efficiency of $\hat{\boldsymbol{\theta}}_{n^*}$ appear in the asymptotic variance $\boldsymbol{\chi}$. First, the term $\mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^T$ that inflates the variance is due to the two-steps procedure (see [Joe05] for similar result on IFM). Second, the fact the binding function is unknown and requires to be estimated is reflected in γ^* with constants B and H increasing the variance. Individual effects of B or H were known in the literature, what may be more surprising here is that both have a multiplicative effect on the variability of $\hat{\boldsymbol{\theta}}_{n^*}$. Third, losing information with $\rho^* < 1$ inevitably increases the variability of $\hat{\boldsymbol{\theta}}_{n^*}$.

If in addition to Assumption 2.9 (*iii*) we suppose $\mathbf{J}$ to be invertible, and thereby implicitly supposing that $\dim(\boldsymbol{\pi}) = \dim(\boldsymbol{\theta})$, an optimal choice of weight function could be

$$\boldsymbol{\Omega}_{\text{opt}} = \mathbf{J}^{-1}\mathbf{K},$$

as it simplifies the asymptotic variance to

$$\mathbf{J}^{-1}\left[\gamma^*\mathbf{Q} + \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^T\right]\mathbf{J}^{-T}.$$

An other choice, which is more general as $\boldsymbol{\pi}$ can be of a larger dimension than $\boldsymbol{\theta}$, is to select $\boldsymbol{\Omega}_{\text{opt}} = \mathbf{K}^T\mathbf{K}$ to simplify the asymptotic variance. Note that with both choices, Assumption 2.7 (*i*) is satisfied by Assumption 2.9 (*iii*). If one is willing to strengthen slightly the above conditions, a general optimal weight function is found and given in the next proposition.

Proposition 2.13 (optimal $\boldsymbol{\Omega}$)**.** *If in addition to the conditions under which Theorem 2.12 is derived, the matrices $\mathbf{Q}$ and $\mathbf{J}^T\mathbf{Q}^{-1}\mathbf{J}$ are nonsingular, then an optimal weight function is*

$$\boldsymbol{\Omega}_{opt} = \mathbf{K}^T\mathbf{Q}^{-1}\mathbf{K}.$$

The asymptotic variance matrix in Theorem 2.12 simplifies to

$$\boldsymbol{\chi}_{opt} = \left(\mathbf{J}^T\mathbf{Q}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{Q}^{-1}\left[\gamma^*\mathbf{Q} + \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^T\right]\mathbf{Q}^{-1}\mathbf{J}\left(\mathbf{J}^T\mathbf{Q}^{-1}\mathbf{J}\right)^{-1}.$$

Supposing the covariance $\mathbf{Q}$ and the squared symmetric matrix $\mathbf{J}^T\mathbf{Q}^{-1}\mathbf{J}$ to be invertible seems reasonable in general. The optimal weight function in Proposition 2.13 corresponds in fact to the inverse of the asymptotic variance matrix of the auxiliary estimator $\hat{\boldsymbol{\pi}}_{n^*}$ (see the appendix for more details, especially Lemma 2.A.2). Eventually, note that if the marginal estimators are fixed, the asymptotic variance in Theorem 2.12 further simplifies to

$$\boldsymbol{\chi}_{\text{opt}} = \gamma^* \left(\mathbf{J}^T\mathbf{Q}^{-1}\mathbf{J}\right)^{-1}.$$

Heretofore, we demonstrated that the indirect inference estimator in (2.3) is consistent and we gave the distribution towards which it weakly converges. But is this estimator and its distribution obtainable has remained a silent question. The next section aims at shedding lights on such feasibility.

2.4 Some practical aspects for indirect inference procedure

Indirect inference procedures are not in particular computationally easy ones. In this section we discuss and propose some practical techniques to obtain a point estimator and make inference for indirect inference procedure.

2.4.1 Point estimator

Finding a point estimator for indirect inference procedure is usually very demanding in computational power. The main reason for this drawback is that every step for solving $\hat{\boldsymbol{\theta}}_{n^*}$ in an optimization procedure requires B optimizations for approximating the binding function. As discussed in the first Chapter in Theorem 1.8, if one considers only $B = 1$ and $\dim(\boldsymbol{\pi}) = \dim(\boldsymbol{\theta})$, then it becomes equivalent to solve the problem directly within the Z-estimating equation (the SwiZs), a solution much more computationally efficient (see Example 1.63 on M/G/1 queue of the first Chapter).

Another reason is that the gradient of the objective function in (2.3) is usually unknown, thereby requiring further costly numerical approximations. Proposition 2.10 offers in this view an interesting solution. For example, the kth step of a gradient descent algorithm may be expressed as

$$\hat{\boldsymbol{\theta}}_{n^*}^{(k)} = \hat{\boldsymbol{\theta}}_{n^*}^{(k-1)} + \mathbf{J}^T\mathbf{Q}^{-1}\mathbf{K}\left[\hat{\boldsymbol{\pi}}_{n^*} - \bar{\boldsymbol{\pi}}_m\left(\hat{\boldsymbol{\theta}}_{n^*}^{(k-1)}\right)\right],$$

when n is sufficiently large and $\boldsymbol{\Omega}_{\text{opt}}$ in Propostion 2.13 is used. When $\dim(\boldsymbol{\pi}) = \dim(\boldsymbol{\theta})$, a further simplification is the iterative bootstrap as proposed by [Gue+18b]:

$$\hat{\boldsymbol{\theta}}_{n^*}^{(k)} = \hat{\boldsymbol{\theta}}_{n^*}^{(k-1)} + \left[\hat{\boldsymbol{\pi}}_{n^*} - \bar{\boldsymbol{\pi}}_m\left(\hat{\boldsymbol{\theta}}_{n^*}^{(k-1)}\right)\right].$$

Under some general assumptions on the form of the bias of the auxiliary estimator, [Gue+18a] show that the iterative bootstrap procedure indeed converges to the indirect inference estimator. The aforementioned solutions induced no approximation of any sort (except numerical). Other authors proposed to approximate directly the binding function to avoid its computation (see e.g. [AD15]).

2.4.2 Inference

Closed-form expression for the asymptotic variance in Theorem 2.12 is almost impossible to obtain. As an illustration, in Figure 2.1 we show the asymptotic variance corresponding to different estimators of a Gumbel-Hougaard copula ([Gum60; Hou86]), with survival Weibull marginal distributions whose joint bivariate survival 5-parameters distribution is given by:

$$\exp\left\{-\left([\eta_1\mathbf{y}_1]^{\kappa_1/\theta}+[\eta_2\mathbf{y}_2]^{\kappa_2/\theta}\right)^{\theta}\right\},\quad 0\leq\theta\leq 1,$$

where θ is the copula dependence parameter corresponding to Kendall's tau ([Ken38]) $\tau = 1-\theta$. [OM92] derived Fisher information matrix for the maximum likelihood estimator when both marginal parameters are assumed fixed or random. In Appendix 2.C, we derived the additional quantities to obtain the asymptotic variance of the two-steps maximum likelihood estimator, or IFM; see [Joe05] for the general form of the asymptotic covariance matrix. Two elements from this derivation are worth noticing: first, we could not find a closed-form solution (some integrals need to be evaluated numerically), second, the expression is long and tedious. Numerical solutions are therefore in general necessary.

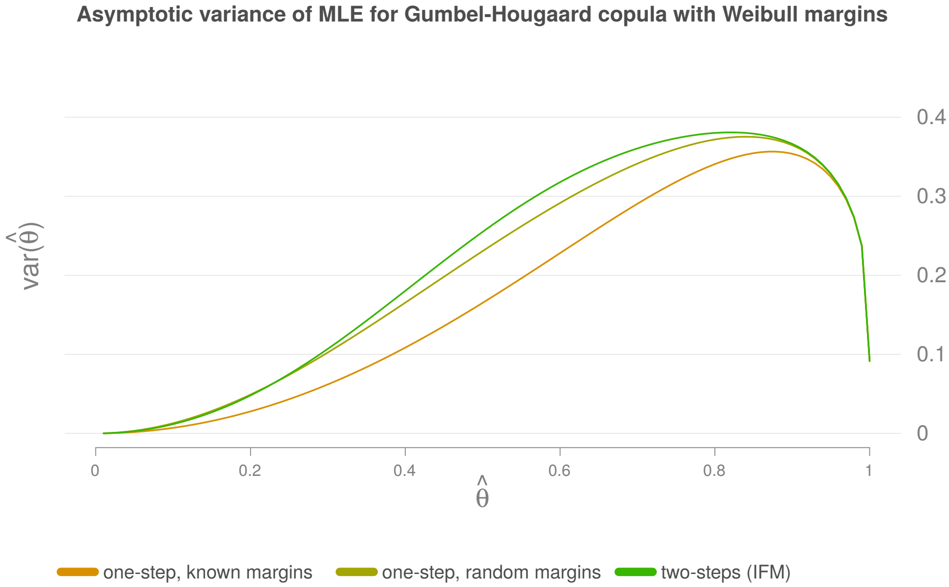

Figure 2.1: Asymptotic variance of different maximum likelihood estimators of Gumbel-Hougaard copula dependence parameter with survival Weibull marginal distributions: one-step MLE when marginal Weibull parameters are fixed or random (due to [OM92]) and two-steps MLE (IFM). When $\hat{\theta}$ tends to 1 (independence) or to 0 (comonotonicity), the asymptotic variances take the same values.

The bootstrap ([Efr79]) has been extensively used for the purpose of inference. If a point estimate has been computationally hard to obtain, as it is generally the case for indirect inference estimator, it is however unthinkable to use this method. The first Chapter provides an alternative solution which is computationally efficient and remarkable in terms of the quality of the inference, but it is derived under the restrictive situation where $\dim(\boldsymbol{\pi}) = \dim(\boldsymbol{\theta})$ and $B = 1$. We therefore propose a fast bootstrap strategy for the general indirect inference estimator in (2.3). The main idea is to bypass the repetitive optimization of $\boldsymbol{\theta}$ by a first order approximation. Higher order of approximations may be developped but has not been considered here. Define the following bootstrap random variables

$$\hat{\boldsymbol{\theta}}^{\star}_{n^*} = \hat{\boldsymbol{\theta}}_{n^*} - \left(\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\left[\hat{\boldsymbol{\pi}}^{\star}_{n^*} - \bar{\boldsymbol{\pi}}^{\star}_{m}\right],$$

where "$\star$" is for the bootstrap, the rest of the quantities being fixed. If $\sqrt{n}(\hat{\boldsymbol{\pi}}^{\star}_{n^*} - \bar{\boldsymbol{\pi}}^{\star}_{m})$ converges to the same distribution as $\sqrt{n}(\hat{\boldsymbol{\pi}}_{n^*} - \bar{\boldsymbol{\pi}}_{m}(\boldsymbol{\theta}_0))$, then, in view of Theorem 2.12, it is clear that this bootstrap converges to the same distribution as $\sqrt{n}(\hat{\boldsymbol{\theta}}_{n^*} - \boldsymbol{\theta}_0)$. The demonstration of the convergence $\sqrt{n}(\hat{\boldsymbol{\pi}}^{\star}_{n^*} - \bar{\boldsymbol{\pi}}^{\star}_{m})$ seems straightforward following [BF81] but the formal treatment is left for further research.

Different strategies may be adopted for obtaining the bootstrap distribution of the auxiliary estimators. We only present the "naive" version. In order to gain numerical speed, one may consider for example the bootstrap of [HK00]; see also [CB+05] for an alternative approach. Supposing the auxiliary and marginal estimators to be numerically easy to obtain, one can use the usual bootstrap procedure:

$$\hat{\boldsymbol{\pi}}^{\star}_{n^*} = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_{n^*}\left(\hat{\boldsymbol{\nu}}^{\star}_{n}, \boldsymbol{\pi}\right),$$

where $\hat{\boldsymbol{\nu}}^{\star}_{n}$ is the bootstrap estimate of marginal estimator, and

$$\hat{\boldsymbol{\pi}}^{\star}_{m} = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\operatorname{argzero}}\, \boldsymbol{\Phi}_{m}\left(\hat{\boldsymbol{\theta}}_{n^*}, \boldsymbol{\pi}\right)$$

to form the average $\bar{\boldsymbol{\pi}}^{\star}_{m}(\hat{\boldsymbol{\theta}}_{n^*})$. If using the parametric bootstrap, the bootstrapped observations are obtained by first generating dependent uniform variates $\mathbf{u}^{\star}$ from model $C_{\hat{\boldsymbol{\theta}}_{n^*}}$, then using distribution inverse $F_Y^{-1}(\mathbf{u}^{\star}; \hat{\boldsymbol{\nu}}_n)$ or equivalent to get $\mathbf{y}^{\star}_t$, and eventually set $\mathbf{x}^{\star}_t = \boldsymbol{\mu}_t(\hat{\boldsymbol{\nu}}_n) + \boldsymbol{\sigma}_t(\hat{\boldsymbol{\nu}}_n)\mathbf{y}^{\star}_t$. Note that the $\mathbf{u}^{\star}$ employed to generate $\mathbf{x}^{\star}$ and to estimate $\bar{\boldsymbol{\pi}}^{\star}_{m}$ should be independently sampled to avoid an unwanted dependence.

2.5 Bounding the Influence Function

We use here the general framework presented in Section 2.2 to develop a bounded IF of the two-steps indirect estimator in (2.3). Several results are useful in the construction of our argument. First, as shown in [GDL00; GR03], the indirect inference estimator $\hat{\boldsymbol{\theta}}_{n^*}$ in (2.3) has a bounded IF only if the auxiliary estimator $\hat{\boldsymbol{\pi}}_{n^*}$ in (2.2) has a bounded IF. Second, as demonstrated by [ZGR12; ZGR16] in the context of Heckman's two steps method ([Hec79]), the IF of first step estimators must be bounded for the second step to be robust. Third, in the context of a multivariate location model, [Alq+09] shows that the influence of outliers on the location estimator is more severe depending on data generating mechanism.

Our result presented in the next theorem is essentially a combination of the aforementioned literature whereas the implications discussed after Corollary 2.16 are surprisingly not concommitant. To this end, let $\Delta_{\mathbf{z}}$ be the multivariate Dirac distribution with point-mass one at $\mathbf{z}$ and suppose that the data generating mechanism is the following deviation model:

$$C_{\varepsilon} = (1 - \delta_{\varepsilon})\, C_{\boldsymbol{\theta}_0} + \delta_{\varepsilon}\Delta_{\mathbf{z}}, \tag{2.4}$$

where ε is the marginal probability of observing an outlier from the multivariate Dirac model, common to each margin, and δ_ε is a proportion reflecting the overall probability of having at least one outlier appearing in one of the d dimensions.

Theorem 2.14 (influence function)**.** *Suppose the binding function $\boldsymbol{\pi}(\boldsymbol{\theta})$ is Hadamard differentiable. If the conditions of Theorem 2.12 hold, then the influence function of $\hat{\boldsymbol{\theta}}_{n^*}$ at the point $\mathbf{z} \in \mathbb{R}^d$ is given by*

$$\mathfrak{I}\left(\hat{\boldsymbol{\theta}}_{n^*}, \mathbf{z}\right) = \left(\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\left[\kappa\boldsymbol{\phi}\left(F_Y(\mathbf{z}, \boldsymbol{\nu}_0)\right) + \mathbf{L}\mathfrak{I}\left(\hat{\boldsymbol{\nu}}_n, \mathbf{z}\right)\right],$$

where $\mathfrak{I}(\hat{\boldsymbol{\nu}}_n, \mathbf{z})$ is the influence function of $\hat{\boldsymbol{\nu}}_n$ at the points $\mathbf{z}$ and κ is a factor that accounts for the dependence among $\mathbf{z}$. In particular, $\kappa = 1$ if $\mathbf{z}$ are comonotonic and $\kappa = -d$ if $\mathbf{z}$ are independent.

Theorem 2.14 may be derived under the additional condition that the binding function, seen as a functional, is Hadamard differentiable. In general in the literature, Gâteaux differentiability, a weaker form of functional derivative, is usually sufficient for the purpose of deriving the IF (see e.g. [Ham+86] and references therein). However, here we require a stronger concept as, because of the two-steps procedure, the chain rule needs to be defined (see [Vaa98], Chapter 20). Note that Hadamard differentiability is weaker than Fréchet differentiability in our context.

Remark 2.15. *The deviation model in (2.4) assumes the proportion of outliers ε to be the same in all the d dimensions of the multivariate model. One may consider a different proportion for each dimension as is done for example in [PB02; Ors13] in the bivariate case. The conclusions would be essentially the same.*

An important implication of Theorem 2.14 is formalized in the following Corollary:

Corollary 2.16. *If and only if the marginal estimators have a bounded IF, $\mathfrak{I}(\hat{\boldsymbol{\nu}}_n, \mathbf{z}) < \infty$, then $\hat{\boldsymbol{\theta}}_{n^*}$ has a bounded IF.*

Several important messages stem from Theorem 2.14 and Corollary 2.16. First, as studied by [Alq+09], we conclude that if outliers appear independently in the multivariate model, they have a much sever effect than if they are totally dependent. The intuition behind is that, the number of outliers in the data being equal, independent outliers affect a larger proportion of data row-wise thus influencing more importantly the dependence estimator than totally dependent outliers. This has been shown, for example, in Lorenz curve comparisons, which are based on quantiles, in [CVF02]. Second, and maybe most importantly, only the marginal estimators need a bounded IF for the copula dependence esitmator $\hat{\boldsymbol{\theta}}_{n^*}$ to be robust to outliers. As an interesting consequence, the IFM (two-steps maximum likelihood, see [Xu96; Joe97]) is not robust to outliers, but the semi-parametric maximum likelihood estimator is ([GGR95; SL95; CFT06; KSS07]) as indeed the empirical distribution function has a bounded IF (see [CVF02; HR09]). This result may be qualified of “counter-intuitive” and it is an apparent contradiction of the results of [GDL00; GR03] on the IF of indirect inference estimator and [ZGR12; ZGR16] on the robustness of two-steps procedure. However, the intuition on why this phenomena happens is quite simple: copula parameters are estimated on a compact set, and thus, the IF being a function of this set, it is inevitably bounded (see the proof in the appendix).

In next Section 2.6, we give more substance to the estimating procedure proposed in Section 2.2 by proposing specific forms of auxiliary esitmators in the light of the preceding findings.

2.6 A weighted maximum likelihood indirect estimator

In the light of Corollary 2.16, one may wonder why several authors proposed robust procedure for copulae estimation while assuming marginal parameters to be fixed (see for instance [MMN07; DM11; KL13]). In fact, simulation studies show evidences that outliers may drastically bias the copula estimator if not robust (see [Ors13]). Bounding the IF is a concept too weak to be useful. Indeed, the literature indicates that having a bounded bias is a protection not sufficient enough, researchers seem to be willing to obtain a bias due to outliers as close to 0 as possible while maintaining the efficiency of the estimating procedure. With this purpose in mind, we propose to use the weighted maximum likelihood estimator ([FS94]) as the auxiliary estimator $\hat{\boldsymbol{\pi}}_{n^*}$ in our indirect inference procedure in (2.3).

The weighted maximum likelihood is indeed a straightforward method to modify the maximum likelihood estimator to gain robustness against data contamination while suffering a small loss of efficiency, controlled under model $C_{\boldsymbol{\theta}_0}$. The likelihood score function is simply multiplied by a weight function. Typical weight functions are Huber's function [Hub+64]

$$w_c(\mathbf{s}) = \min\left(\frac{c}{\|\mathbf{s}\|}, 1\right),$$

where $\mathbf{s}$ represents the q-dimensional vector of likelihood score function, or *redescending* weight function such as the Tukey biweighted function [BT74]

$$w_c(\mathbf{s}) = \begin{cases} \left[1 - \left(\frac{\|\mathbf{s}\|}{c}\right)^2\right]^2 & \text{if } \|\mathbf{s}\| \leq c, \\ 0 & \text{if } \|\mathbf{s}\| > c. \end{cases}$$

The biweight function has the particularity that it entirely discards extreme scores, *i.e.* $\|\mathbf{s}\| > c$, whereas Huber's function takes the value of 0 only at limits. The tuning parameter c plays the role of trade-off between robustness and efficiency when there is no model contamination (i.e. $\varepsilon = 0$).

More general weighted estimating equations could also be used, as is proposed for example with generalized linear models in [PQ99; CR01] where the quasi-likelihood estimating equations ([Wed74]) are weighted; see also [BGT+97] for a broad discussion on estimating equation. Other robust approaches are identical to the weighted maximum likelihood in the sense that estimating equations are downweighted according to the value they take: large values of the estimating equation results in weights close to 0.

Despite being widespread and simple in idea, weighting an estimation equation is not an easy task as, very often, the resulting estimator is not consistent to the target quantity ([DM02]). The indirect inference proposition in (2.3) thus allows for targeting the correct quantity; see for instance [MVF06; Gue+18b].

2.7 Simulation Study

In this section we experiment the findings of the preceeding sections by simulations. In particular, in subsection 2.7.1 we illustrate how the maximum likelihood and the indirect inference estimators in 2.3 with a weighted maximum likelihood as the auxiliary estimator respond to outliers for a bivariate copula when the marginal distribution are known, thereby indicating that both approaches are robust (see Corollary 2.16). The interest is therefore on the bias and the mean squared errors. We let the reader refer to the exhaustive simulation studies in [Ors13] for scenarii where marginal parameters need also to be estimated under contaminations and the contamination model takes several forms. In subsection 2.7.2, we study a particular model called *factor copula* for which the likelihood function has no known closed-form. We illustrate the benefit of our approach in comparison to an alternative by measuring the performance of the estimators in terms of bias, mean squared errors and computational efficiency.

2.7.1 Bivariate Clayton copula: influence of contamination on bias and mean squared error

Clayton's copula ([Cla78]), also called *Mardia-Takahasi-Clayton-Cook-Johnson* copula in [Joe14] in reference to the several authors who independently discovered this dependence model, is one of the most studied and used copula model. In its bivariate form, the cumulative distribution function is given by

$$\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}, \quad 0 \leq \theta < \infty, \quad 0 \leq u_1, u_2 \leq 1.$$

The dependence parameter θ corresponds to Kendall's tau $\tau = \theta/\theta+2$. When $\theta \downarrow 0$, Clayton's copula is the independent copula; when θ diverges, Clayton's copula is the comonotonic copula. Interestingly, this copula model has a lower tail dependence but is upper tail independent. The density is illustrated in Figures 2.2 and 2.3 with $\theta = 4$ ($\tau = 2/3$).

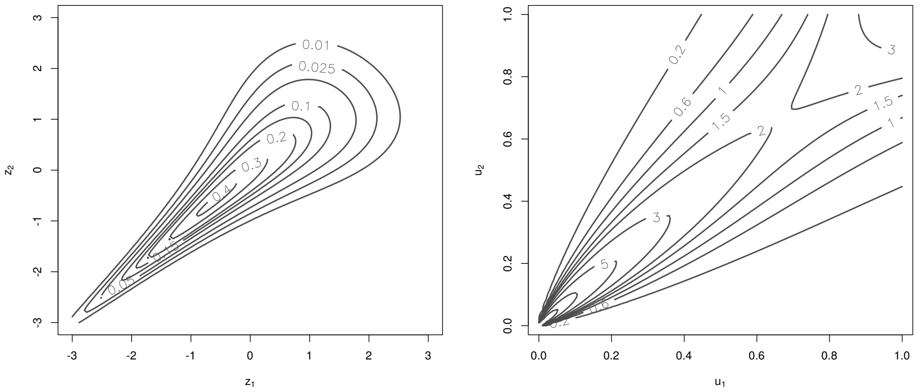

Figure 2.2: Clayton's copula density plots with $\theta = 4$, which corresponds to a Kendall's tau of $\tau = 2/3$. *On the left panel*: the uniform margins are transformed to standard Gaussian, i.e. $z_j \equiv \Phi^{-1}(u_j)$, $j = 1, 2$. *On the right panel*: same representation as the *left panel*, but with the original uniform margins.

We study the impact of contaminations as expressed in (2.4) on the dependence estimator of the bivariate Clayton copula. More specifically, the Dirac distribution $\Delta_{\mathbf{z}}$ takes two dependence structure:

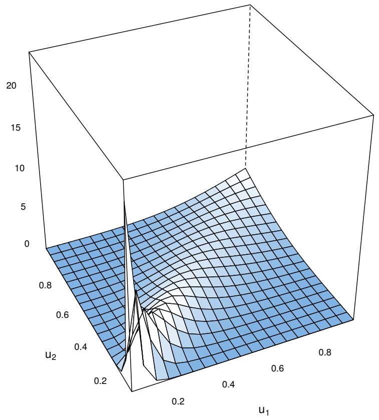

Figure 2.3: Same density illustration as the *right panel* of Figure 2.2 but here the density is represented on the z-axis.

($\mathcal{M}_1$) Outliers are independent: $\Pr(u_1 = z_1,\ u_2 = z_2) = \Pr(u_1 = z_1) \Pr(u_2 = z_2)$.

($\mathcal{M}_2$) Outliers are comonotone: $\Pr(u_1 = z_1,\ u_2 = z_2) = \Pr(u_1 = z_1)$.

We use the following setting: $n = 10^3$ and $\theta_0 = 4$. For the weighted maximum likelihood estimator, we consider Tukey's biweighted function with three tuning constants $c = (5, 10, 20)$ which corresponds roughly to $60\%, 85\%$ and 95% of efficiency compared to the maximum likelihood estimators. We study the model under no contaminations and 10% of outliers taking values in the regular grid from 0 to 1. For each value of $\mathbf{z}$ considered, 100 simulations are performed. For indirect inference estimator in (2.3), we consider for simplicity $H = 1$, $B = 1$ and $\Omega = 1$.

In Figure 2.4, the estimators are illustrated when there is no contamination ($\delta_\varepsilon = 0$ in (2.4)). We clearly vizualize that the weighted maximum likelihood estimators have important bias increasing when c decreases. The indirect inference estimators that correspond to these weighted maximum likelihood estimators have however nearly no bias, which shows well the ability of the method to correct bias. The variability of the indirect inference estimators are sensibly larger than their auxiliary estimators counterparts. This is no surprise in view of Theorem 2.12, as we set $B = 1$ and $H = 1$ the asymptotic variance of $\hat{\boldsymbol{\theta}}_{n^*}$ should be roughly two times larger than the asymptotic variance of $\hat{\boldsymbol{\pi}}_{n^*}$. Larger values of these parameters will surely reduce this dispersion.

Figures 2.5 and 2.6 illustrate the bias and mean squared errors under contamination design $\mathcal{M}_1$. The effect of contamination is striking: no estimators may pretend to be unbiased. However, the weighted maximum likelihood and the indirect inference estimators constantly outperform the maximum likelihood estimator in both terms of bias and mean squared error. Interestingly, almost at every contamination point $\mathbf{z}$ the bias is negative, indicating thereby that the dependence parameter is closer to 0, and thus to the independence copula, than what it actually is. This seems no surprise as outliers are generated independently.

Figures 2.7 and 2.8 show the bias and mean squared errors under contamination design $\mathcal{M}_2$. The effect of contamination is more surprising than what we observed on Figures 2.5

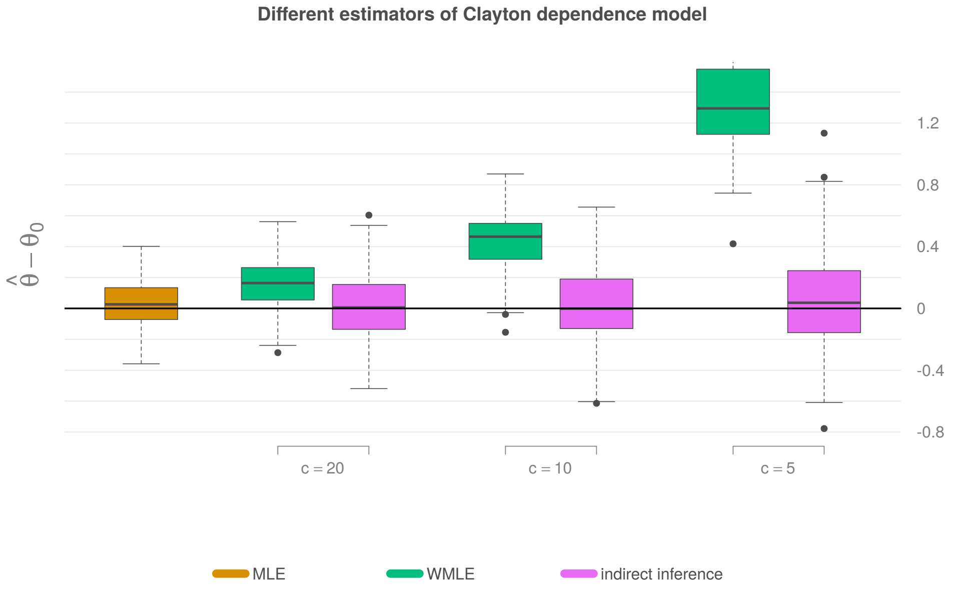

Figure 2.4: Centered boxplots on 100 maximum likelihood, weighted maximum likelihood and indirect inference estimators of dependence parameter of a bivariate Clayton copula when $\theta_0 = 4$. The weight function is Tukey's biweighted function. The tuning constants c corresponds to approximatively 60% ($c = 5$), 85% ($c = 10$) and 95% ($c = 20$) of relative efficiency to the maximum likelihood estimator. $B = 1$, $H = 1$ and $\Omega = 1$ are used for the indirect inference estimator in (2.3).

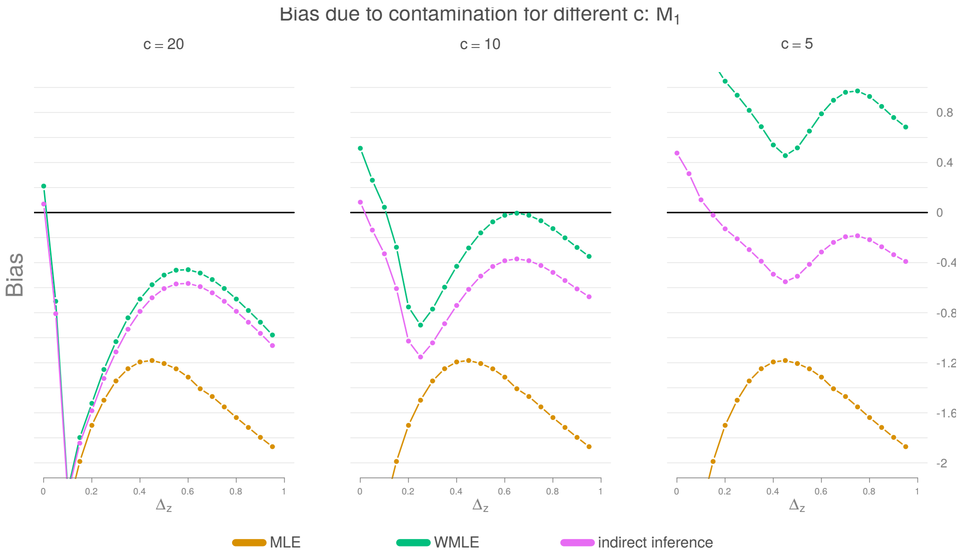

Figure 2.5: Empirical bias of the maximum likelihood, weighted maximum likelihood and indirect inference estimators for different values of outliers $\mathbf{z}$, assuming $z_1 = z_2$, when they represent 10% of the data and they are generated independently and different values of the tuning constant c. Each dot represent the average of 100 estimators minus the true value $\theta_0 = 4$. The maximum likelihood estimator is the same in the three figures.

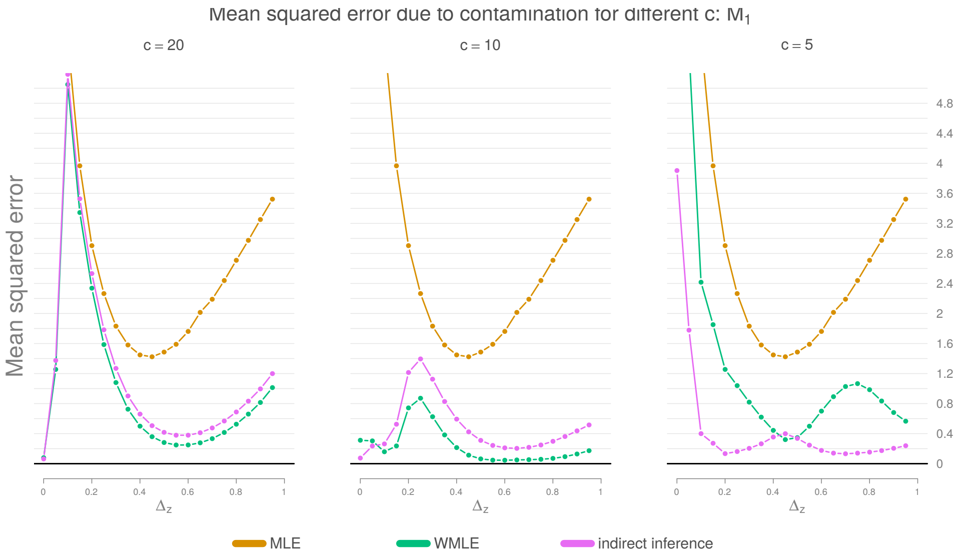

Figure 2.6: Empirical mean squared error of the maximum likelihood, weighted maximum likelihood and indirect inference estimators for different values of outliers $\mathbf{z}$, assuming $z_1 = z_2$, when they represent 10% of the data and they are generated independently and different values of the tuning constant c. Each dot represent the average of 100 of the square of estimators minus the true value $\theta_0 = 4$. The maximum likelihood estimator is the same in the three figures.

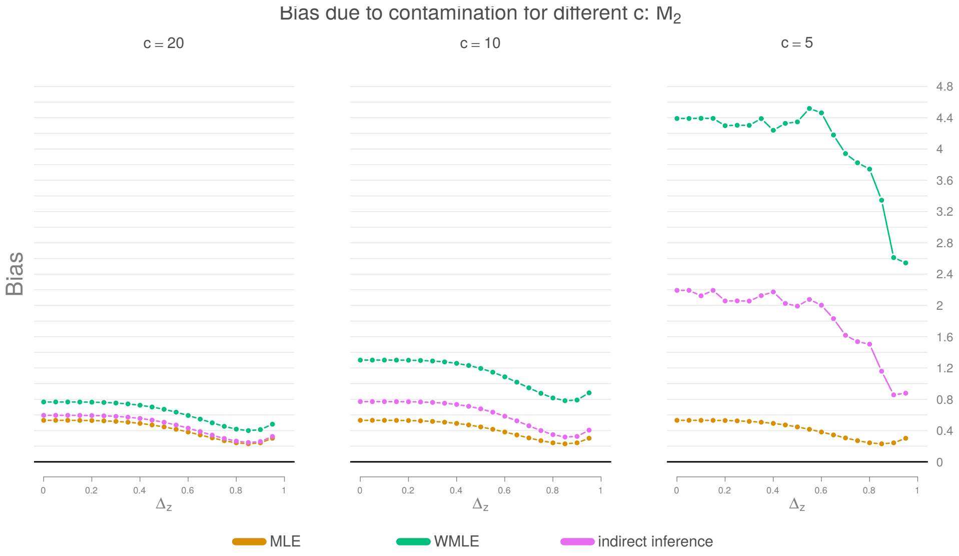

Figure 2.7: Empirical bias of the maximum likelihood, weighted maximum likelihood and indirect inference estimators for different values of outliers $\mathbf{z}$, assuming $z_1 = z_2$, when they represent 10% of the data and they are generated totally dependently and different values of the tuning constant c. Each dot represent the average of 100 estimators minus the true value $\theta_0 = 4$. The maximum likelihood estimator is the same in the three figures.

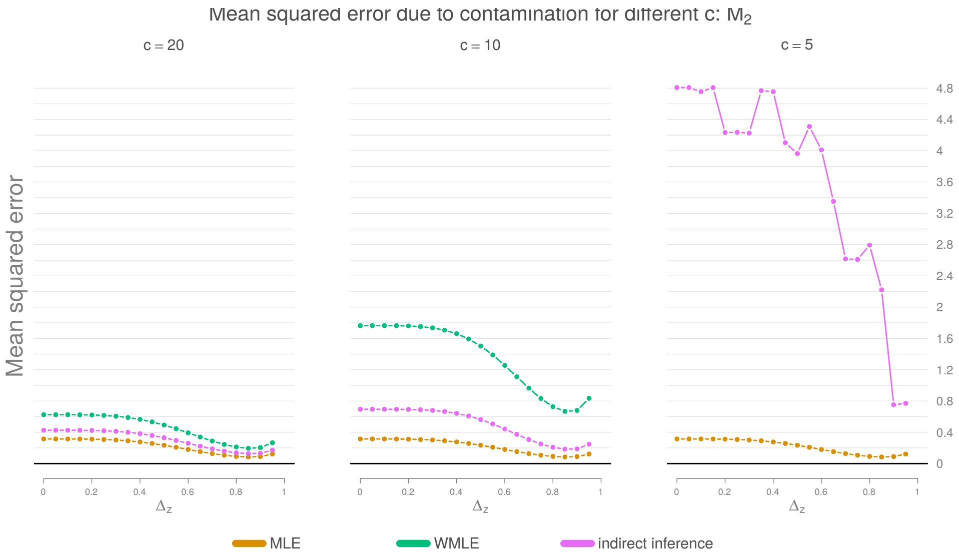

Figure 2.8: Empirical mean squared error of the maximum likelihood, weighted maximum likelihood and indirect inference estimators for different values of outliers $\mathbf{z}$, assuming $z_1 = z_2$, when they represent 10% of the data and they are generated totally dependently and different values of the tuning constant c. Each dot represent the average of 100 of the square of estimators minus the true value $\theta_0 = 4$. The maximum likelihood estimator is the same in the three figures.

and 2.6: on all account the maximum likelihood estimators show better performances than the weighted maximum likelihood and indirect inference estimators. These figure are counter-intuitives as robust estimators are always perceived to outperform classical estimator in cases of data contamination. The explanation is quite simple though: we forced $z_1 = z_2$, the contaminants thus appear in the lower-left to upper-right diagonal, that is where the majority of the density is (see *right panel* of Figure 2.2), so $\mathbf{z}$ are what may be qualified of “inliers”. Interestingly, at every contamination point $\mathbf{z}$ the bias is positive, indicating thereby that the dependence parameter is larger than θ_0, that is closer to the comonotonic copula. Again, this seems no surprise as outliers are generated totally dependently.

Figures 2.9 represents the absolute bias under contamination design $\mathcal{M}_2$ for every coordinates in the unit square (as opposed as Figures 2.7 and 2.8 for which only the down-left to upper-right diagonal is illustrated). The density of the copula (Figure 2.2) is illustrated on top of the absoulte bias. The message is clear: wherever the density is important, outliers have a mild effect of the same order of magnitude on both the maximum likelihood and the robust indirect inference estimators, whereas in the bottom and left areas where the density is small, the robust estimator have no bias, the maximum likelihood estimator have an important bias.

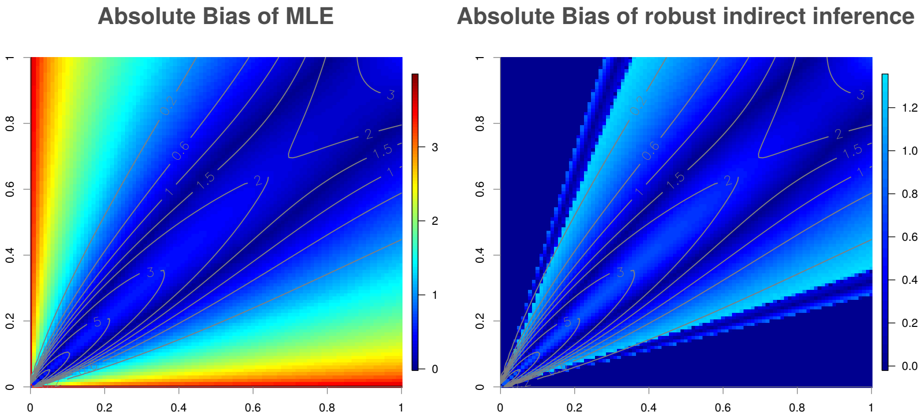

Figure 2.9: Effect of the location of outliers $\mathbf{z}$ on the absolute bias evaluated by 100 simulations for each point of the maximum likelihood (*left panel*) and indirect inference estimators (*right panel*) when using the weighted maximum likelihood as auxiliary estimator with Tukey's biweighted function and tuning constant $c = 10$. The closer to 0 is the better.

2.7.2 Large dimension with factor copula: a time-constrained study

Many situations occur in which the likelihood function has no known analytical expressions. A famous example is the stable distribution which has no analytical expression for its density depending on the parametrization (see e.g. [PSF12]). Indirect inference is a natural candidate for trying to solve this estimation problem, the auxiliary estimator do not need to be consistent to $\boldsymbol{\theta}_0$. The choice of the auxiliary estimator is however delicate (see Example 1.63 of the first Chapter) and many proposition may be valid. For the stable distribution, [GRV11] for example proposed a skewed-t distribution as the auxiliary model. This situation is also common when modeling latent variables. In this section, we focus on a specific latent model, the *factor copula* of [OP17] (see also [MFE05; KJ13; KHG18] for alternatives). The authors developed a factor copula model tailored for the modeling of high-dimensional dependent random variables that requires particular care for tail events and asymmetrical dependence structure and showed the benefits of such approach in applications. The main idea behind the factor copula model is to separate the modeling of the marginal distributions from the joint distribution through a copula function, but nonetheless uses a factor model for the joint model to benefit from the dimension-reduction capacity. Specifically, we study the following factor copula model:

$$\mathbf{y}_j = \mathbf{w} + \boldsymbol{\epsilon}_j, \quad j = 1, \cdots, d, \tag{2.5}$$

where the latent variable $\mathbf{w}$, common to all d dimensions, follows a centered Hansen's skewed-t distribution ([Han94]) with unknown parameters $\boldsymbol{\theta} = (\sigma^2, \lambda, \eta)^T$, and $\boldsymbol{\epsilon}_j$ are identically and independently distributed according to a standardized t with the same unknown parameter η. (Note that there exist several proposal in the statistical literature to introduce skewness in the t distribution and we therefore always mention the author's name to avoid ambiguity). Hansen's skewed-t density for $i = 1, \cdots, n$ is given by:

$$f(w_i, \boldsymbol{\theta}) = \begin{cases} bc\left(1 + \frac{1}{\eta-2}\left(\frac{bw_i+a}{\sigma(1-\lambda)}\right)^2\right)^{-(\eta+1)/2}, & \text{if } w_i < -{}^{a}\!/\!{}_{b}, \\ bc\left(1 + \frac{1}{\eta-2}\left(\frac{bw_i+a}{\sigma(1+\lambda)}\right)^2\right)^{-(\eta+1)/2}, & \text{if } w_i \geq -{}^{a}\!/\!{}_{b}, \end{cases}$$

where

$$a \equiv 4\lambda c \frac{\eta - 2}{\eta - 1}, \quad b^2 \equiv 1 + 3\lambda^2 - a^2, \quad c \equiv \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\Gamma\left(\frac{\eta}{2}\right)\sqrt{(\eta-2)\pi\sigma}},$$

and $\sigma^2 > 0$ is the scale parameter, $-1 \leq \lambda \leq 1$ is the parameter that gauges the asymmetry of the distribution ($\lambda = 0$ means no asymmetry) and $\eta > 2$ denotes the degree-of-freedom (see also [JR03] for interesting properties of Hansen's skewed-t).

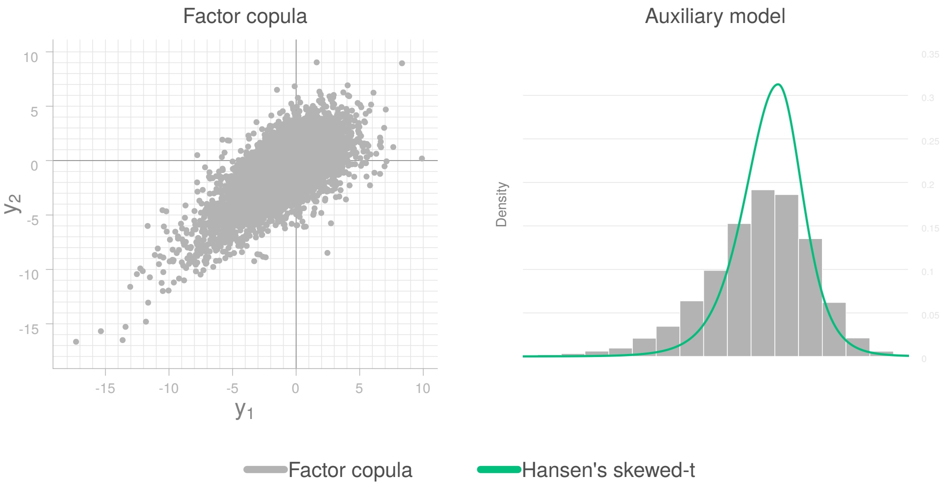

Figure 2.10: *On the left panel*: sampled factor copula in (2.5) with $d = 2$, $\boldsymbol{\theta}_0 = (\sigma_0^2 = 2, \lambda_0 = -0.5, \eta_0 = 6)^T$ and $n = 10^4$. *On the right panel*: the histogram represents the sample in the *left panel* vectorized. The solid line is the density of a Hansen's skewed-t distribution with parameters estimated on the vectorized sample, we obtained $\hat{\boldsymbol{\pi}}_{n^*} = (2.29, -0.14, 6.35)^T$.

As simple as it looks, this factor model has no known closed-form expression for the likelihood function and is therefore complicated to estimate. Indeed, already the convolution of two independent t variables is known only on restrictive cases where the degree-of-freedom are odd numbers ([ND05]). A similar model is studied in [OP13] under a more complex scenario where each marginal observation are assumed to follow an AR(1)-GARCH(1,1) process. Here we focus our interest on $\boldsymbol{\theta}$ and assume the margins to be known. [OP13] propose to estimate the above model by using bivariate measures of dependence as the auxiliary estimator $\hat{\boldsymbol{\pi}}_{n^*}$ in (2.3), namely Spearman's rank correlation ρ_{ij} ([Spe04]) and quantile dependence q_{ij}^{α}, whose sample version, for $i, j = 1, \cdots, d$, are defined as:

$$\begin{aligned}
\hat{\rho}_{ij} &\equiv \frac{12}{n^*}\sum_{t\in\mathcal{T}^*}\widehat{F}_i\left(y_{it}\right)\widehat{F}_j\left(y_{jt}\right)-3,\\
\hat{q}_{ij}^{\alpha} &\equiv \begin{cases} \frac{1}{n^*\alpha}\sum_{t\in\mathcal{T}^*}\mathbf{1}\left\{\widehat{F}_i\left(y_{it}\right)\le\alpha,\widehat{F}_j\left(y_{jt}\le\alpha\right)\right\}, & \alpha\in(0,0.5],\\ \frac{1}{n^*(1-\alpha)}\sum_{t\in\mathcal{T}^*}\mathbf{1}\left\{\widehat{F}_i\left(y_{it}\right)>\alpha,\widehat{F}_j\left(y_{jt}>\alpha\right)\right\}, & \alpha\in(0.5,1),\end{cases}
\end{aligned}$$

where $\widehat{F}_i(y_0)\equiv {}^1/_{(n^*+1)}\sum_{t\in\mathcal{T}^*}\mathbf{1}\{y_{it}\le y_0\}$ is the empirical cumulative distribution.

There is at least one clear advantage in [OP13]'s approach: these dependence measures do not depend on the chosen factor copula model, so it is easy to use the same measures across different models, and also they offer robustness agains model misspecification. On the other hand, we see two disadvantages to their proposal. First, the proposed measures of dependence are not smooth and Proposition 2.10 do not hold for finite n, as a consequence the numerical optimization for this problem may be cumbersome unless for example considering the iterative bootstrap procedure ([Gue+18b; Gue+18a]), but this requires to restrict the number of moment to satisfy the constraint $\dim(\boldsymbol{\pi})=\dim(\boldsymbol{\theta})$, which is impossible here as we discuss next. [OP13] in their supplementary material mentioned this problem, they had to opt for a derivative-free algorithm. Second, since these measures are bivariate, the number of moments q increases with d, the dimension of the problem. If q_2 denotes the number of moments when $d=2$, there are $q_2d(d-1)/2$ moments when $d>2$. A solution to this problem is discussed in the appendix of [OP17]. They propose to average certain measure to keep the dimension of $\boldsymbol{\pi}$ constant. This maybe simplify the optimization procedure, but the $q_2d(d-1)/2$ dependence measures still need to be computed, which can be a tremendous effort in high-dimensions.

As an alternative, we propose an auxiliary model that ignores the latent structure and look at the data as if it was one large sample of size $d\times n$ identically and independently distributed according to Hansen's skewed-t. This model seems not far in idea to the data generating process but clearly there is no reason to believe for the resulting estimators to be consistent, and we thus correct them via indirect inference as in (2.3). In Figure 2.10, we illustrate the "closeness" between the factor copula and the auxiliary model when $d=2$. We simulate a sample of size $n=10^4$ with $\boldsymbol{\theta}_0=(\sigma_0^2=2,\lambda_0=-0.5,\eta_0=6)^T$. We find $\hat{\boldsymbol{\pi}}_{n^*}=(2.29,\ -0.14,\ 6.35)^T$.

For the simulations, we set the followings: several dimensions with $d=\{2,\ 10,\ 50,\ 100,\ 1,000\}$, a (marginal) sample size $n=200$, $M=1,000$ Monte Carlo replicates for each dimension and $\boldsymbol{\theta}_0=(\sigma_0^2=2,\lambda_0=-0.5,\eta_0=6)^T$. [OP17] gives the upper and lower tail dependence measures of the copula factor model in (2.5) when σ^2 is fixed to 1 (see Proposition 2 in [OP17]). Supposing $\boldsymbol{\theta}_0=(1,-0.5,6)^T$, the upper tail coefficient is close to 0 whereas the lower tail is stronger, about 0.2. These tail coefficients translate into dependent lower tail event and (quasi) independent upper tail events (see the *left panel* of Figure 2.10 for a representation). We compare our proposed auxiliary model to the proposition of [OP13]. We use their suggested bivariate dependence measures: Spearman's rank correlation and quantile dependence with $\alpha=\{0.05,0.1,0.9,0.95\}$, so $q_2=5$. We also follow [OP17] recommendation and averaged each dependence measures separately when $d>2$ so $q=5$. For the parameters of indirect inference in (2.3), we select them as follows: we fix $H=1$ and choose B such that it takes about 60 seconds to obtain one estimator with [OP13] proposition; if this time limit is exceeded when $B=1$, we reduce $H\in(0,1]$ until approximatively respecting the constraint. The value we obtained on 5 simulations are given in Table 2.1. Concerning the weighting matrix, we estimated $\widehat{\boldsymbol{\Omega}}=\mathbf{Q}^{-1}$ once on large sample for our approach and set $\mathbf{\Omega} = \mathbf{I}_{q_2}$ for the approach on dependence measures. For both [OP13] and our proposal, we use our proposed auxiliary estimator as starting value for finding $\hat{\boldsymbol{\theta}}_{n^*}$.

	Parameters for indirect inference				
d	2	10	50	100	1,000
B	1,000	400	35	9	1
H	1	1	1	1	0.1
$\bar{s}_1$	29.36	43.79	20.04	11.94	1.32
$\bar{s}_2$	22.40	51.30	53.59	53.86	69.05

Table 2.1: Parameters H and B for indirect inference in (2.3) depending on the dimension of the problem used for the simulation study for [OP13] and our proposals. These values target a time of approximatively 60 seconds of computations for [OP13]'s proposal. We report the average computational time in seconds over the 1,000 Monte Carlo replicates: $\bar{s}_1$ is for our proposed auxiliary model and $\bar{s}_2$ is for [OP13]'s proposal.

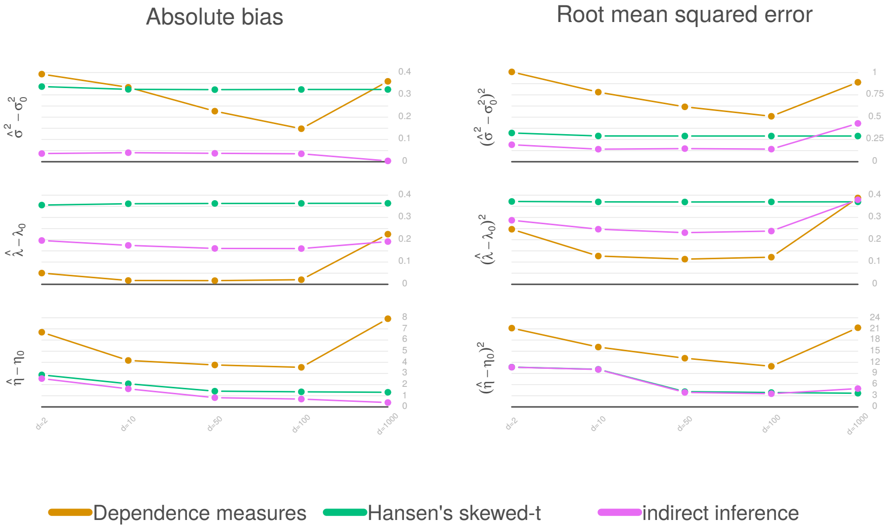

Figure 2.11: *On the left panel*: the absolute bias on $\hat{\boldsymbol{\theta}}_{n^*}$ of [OP13]'s proposal with dependence measures as auxiliary estimators, our proposition to use Hansen's skewed-t distribution as the auxiliary model and the indirect inference estimator. *On the right panel*: likewise *left panel* but using the root mean squared error as measure of performance.

We report the absolute bias and root mean squared error in Figure 2.11. In terms of both performance measures, our approach systematically outperforms [OP13]'s proposal for the scale and degree-of-freedom estimators $\hat{\sigma}^2$ and $\hat{\eta}$, yet the opposite can be observed for the skewness estimator $\hat{\lambda}$, except when $d = 1,000$. Increasing the dimension d seems to increase the performance of the estimators, regardless of the value that B takes, except at the highest dimension considered when $d = 1,000$. This seems to indicate that taking $H = 0.1$ inflates drastically the variability of the estimators. The choice for this value was considered such that [OP13]'s approach takes on average 60 seconds for one estimation. We can clearly see in Table 2.1 that our proposal is faster, the difference in time between the two approaches escalates when the dimensions of the problem increases. Eventually, not that under the same constraint of 60 seconds, our approach could have benefit from larger values for both B and H and thereby increased the performance of the resulting estimators.

2.8 Application to income mobility

We are interested in the income mobility in time of unit-record Swiss households between 2011 and 2012. The data is provided by the Swiss Household Panel, a longitudinal study of Swiss households. For this application, we use the OECD equivalence scale yearly household net income in order to take into account the household's size and composition. We give the descriptive statistics in Table 2.2 after having removed missing data and scaled it by a factor of 10^3.

We follow [VGC10] analysis and model the incomes with Singh-Maddala ([SM76]) distribution (aslo known as Burr XII distribution) and the bivariate distribution with Frank ([Fra79]), Clayton ([Cla78]) and Gumbel-Hougaard ([Gum60; Hou86]) copulae. The Singh-Maddala density function is

$$\frac{aqx_i^{a-1}}{b^a\left[1+\left(\frac{x_i}{b}\right)^a\right]^{q+1}}, \quad x_i > 0,$$

where a, b, q are positive parameters, b is a scale parameters and the two others are shape parameters. The bivariate Frank copula distribution is given by

$$-\theta^{-1}\log\left(\frac{1-e^{-\theta}-(1-e^{-\theta u_1})(1-e^{-\theta u_2})}{1-e^{-\theta}}\right), \quad -\infty < \theta < \infty,$$

where the dependence parameter θ corresponds to Kendall's tau $\tau = 1-{}^4/_\theta+{}^4/_{\theta^2}\int_0^\theta {}^t/_{(e^1-1)}\,\mathrm{d}t$. The countermonotonic copula is obtained as $\theta \to -\infty$, the independence copula when $\theta \to 0$ and the comonotonic copula as $\theta \to \infty$. From the modeler's point of view, Clayton's copula is interesting for its lower tail dependence as opposite to Gumbel-Hougaard's copula which permits to model upper tail dependent events. Frank's copula is useful as it has both negative and positive dependence (see Figure 2.14).

Estimators that are robust to outliers are particularly important, it has been outlined on many occasions in the literature of income distribution in both practical and theoretical situations (see e.g. [CVF96b; CVF96a; RVF97; VFR97; CVF02; CVF06; CVF08; ATF13; Rob+15; PS18]). We therefore compare the performance of the maximum likelihood against robust indirect inference when using the weighted maximum likelihood estimator with Tukey's biweighted function as the auxiliary model.

For these estimators, we perform grid searches in order to obtain the starting values for the estimation procedures and set the tuning constants c so the robust estimator achieves roughly 90% of relative efficiency compared to the maximum likelihood estimator. We report the estimators and standard errors in Table 2.3. The standard errors are obtained using the bootstrap scheme presented in Section 2.4 using a parametric bootstrap, therefore there are three possibilities, each one corresponding to one the three multivariate models considered.

It is clear from the descriptive statistics in Table 2.2 and Figure 2.12 that the data have a heavy right tails, with some incomes very far from the averages. Hence we expect the maximum likelihood estimator to be biased. Surprisingly, the maximum likelihood and robust indirect inference estimators have very comparable values for both marginal parameters suggesting that data is not subject to contaminations. However, for the copula dependence parameter, there is a clearer distinction between the two methods of estimation regardless of the choice of copula. Robust estimators indicate stronger dependencies than classical estimators. The parameters may be related to Kendall's tau and as shown

Yearly net income, OECD equivalised		
	2011	2012
observations	4103	4065
min	1	1
median	58.80	59.10
mean	66.79	67.67
max	2062	5503
skewness	17.27	46.42
kurtosis	504.81	2621.47
linear correlation	0.3120	
Spearman's correlation	0.8467	
Kendall's tau	0.6885	
Gini's index	0.2192	

Source: Swiss Household Panel (SHP)

Table 2.2: Descriptive statistics of Swiss households. There are 3692 households present in both 2011 and 2012 surveys. Bivariate Gini's index is given in (2.6).

in Table 2.3 the estimators provide quite different values which might lead to a different interpretation of the degree of dependence between the two income cohorts. This analysis is supported by the bivariate plots in Figure 2.14. This empirical application shows that even if margins do not suffer from contamination, the dependency may be impacted.

To have a better understanding of the implication of the difference in estimates, we compute the bivariate Gini (distance-Gini) index [KM97] as is done in e.g. [JSVK15]. The population version for a d-dimensional parametric distribution $G_Y(\mathbf{y})$ (which corresponds to $C_{\boldsymbol{\theta}}(F_Y(\mathbf{y},\boldsymbol{\nu}),\boldsymbol{\theta})$) is

$$\frac{1}{2d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\mathbf{y}-\mathbf{y}'\|\,\mathrm{d}\widetilde{G}_Y(\mathbf{y})\,\mathrm{d}\widetilde{G}_Y(\mathbf{y}')$$

where $\widetilde{G}_Y$ is the *relative distribution function* of G_Y, that is the joint distribution of

$$\left(\frac{\mathbf{y}_1}{\mathbb{E}_{F_1}[\mathbf{y}_1]},\dots,\frac{\mathbf{y}_d}{\mathbb{E}_{F_d}[\mathbf{y}_d]}\right).$$

The sample version of this bivariate Gini index is given by:

$$\frac{1}{2dn^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\sum_{s=1}^{d}\frac{(y_{is}-y_{js})^2}{\bar{y}_s^2}\right)^{1/2}, \tag{2.6}$$

where $\bar{y}_s = {}^1/_n\sum_{i=1}^{n} y_{is}$ is the average value of the sth dimension.

In general, there is no closed-form expression for this bivariate Gini index, but one can easily simulate a large sample from the copula model and then approximate the population index with the sample version given in (2.6). In Figure 2.13, we illustrate 1,000 such simulations using $n = 3,692$ as the sample size and Clayton's copula. It appears clearly that the Gini index issued from maximum likelihood and indirect inference estimators almost do not overlap.

	MLE		Robust	
Parameters	Estimates	$\tau(\hat{\theta})$	Estimates	$\tau(\hat{\theta})$
$\hat{a}_{(2011)}$	3.2039		3.1530	
	(GH : .0696, FR : .0698, CL : .0719)		(GH : .0032, FR : .0040, CL : .0029)	
$\hat{b}_{(2011)}$	64.6926		67.1316	
	(GH : 1.8050, FR : 1.8194, CL : 1.9929)		(GH : 4.9892, FR : 5.2536, CL : 5.0833)	
$\hat{q}_{(2011)}$	1.2696		1.3803	
	(GH : .0752, FR : .0757, CL : .0830)		(GH : .0298, FR : .0305, CL : .0299)	
$\hat{a}_{(2012)}$	3.2993		3.2259	
	(GH : .0754, FR : .0788, CL : .0779)		(GH : .0024, FR : .0029, CL : .0022)	
$\hat{b}_{(2012)}$	63.8496		66.3655	
	(GH : 1.7531, FR : 1.8836, CL : 1.9320)		(GH : 2.9393, FR : 3.2197, CL : 3.0343)	
$\hat{q}_{(2012)}$	1.2188		1.3312	
	(GH : .0716, FR : .0777, CL : .0786)		(GH : .0440, FR : .0521, CL : .0448)	
$\hat{\theta}^{\text{GH}}$	2.9759	.6640	3.1858	.6861
	(.0514)		(.0224)	
$\hat{\theta}^{\text{FR}}$	11.2032	.6953	11.5488	.7030
	(.2021)		(.3170)	
$\hat{\theta}^{\text{CL}}$	2.5820	.5635	3.2734	.6207
	(.0750)		(.0116)	

Source: Swiss Household Panel (SHP)

Table 2.3: Maximum likelihood and robust indirect inference estimators of the Singh-Maddala distribution and Gumbel-Hougaard (GH), Frank (FR) and Clayton (CL) copulae. Standard errors are in parenthesis, they are estimated by bootstrap with $B = 500$ replicates (see Section 2.4). Kendall's tau are calculated from the estimated models.

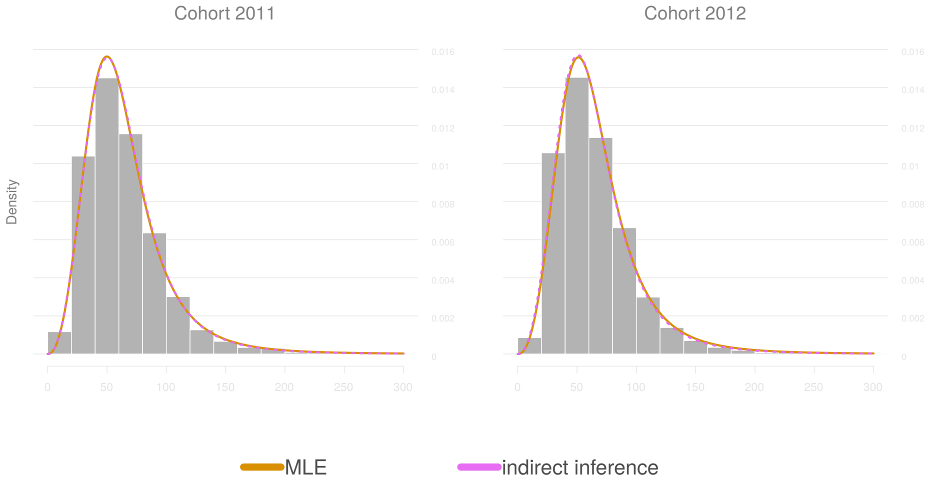

Figure 2.12: *On the left panel*: histogram representation of income for 2011. Solid lines correspond to Singh-Maddala density with maximum likelihood and robust indirect inference estimators. *On the right panel*: likewise *left panel* but for 2012.

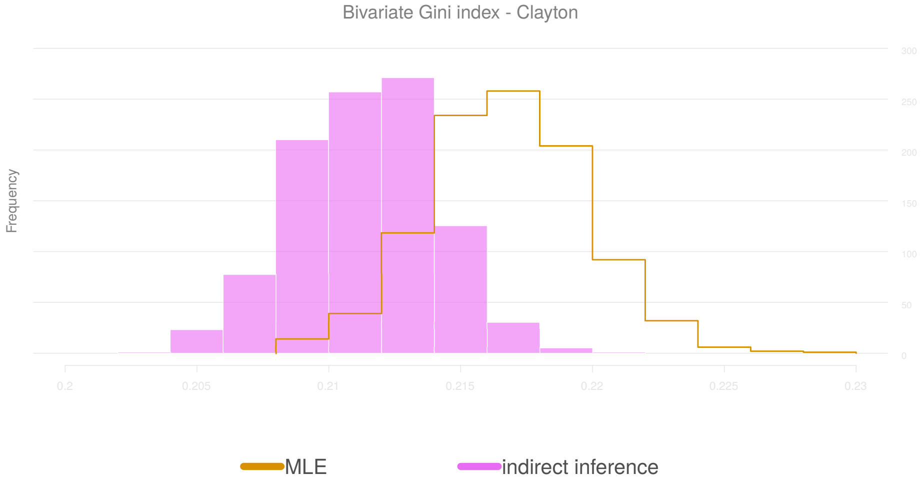

Figure 2.13: Histogram representation of 1,000 simulated multivariate Gini's indices based on a Clayton copula with Singh-Maddala marginal distributions. The parameters are set to estimates given in Table 2.3.

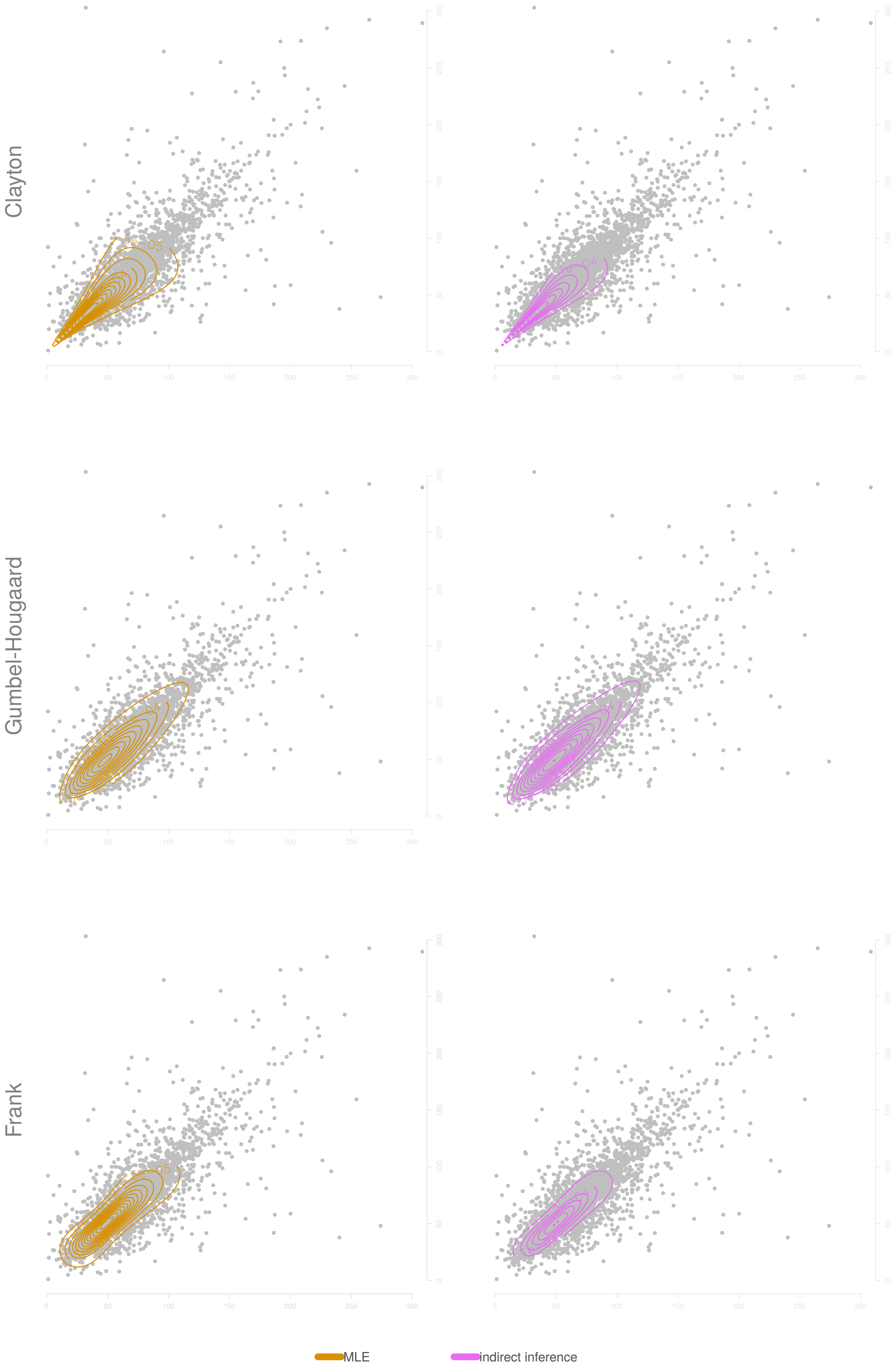

Figure 2.14: Representation of the bivariate incomes (dots) with 2011 cohort in the x-axis and 2012 cohort in the y-axis on top of which densities (solid lines) implied by the copula models with Singh-Maddala margins with parameters set at the estimates in Table 2.3 are illustrated.

References for Chapter 2

[Aas+09] K. Aas et al. "Pair-copula constructions of multiple dependence". In: *Insurance: Mathematics and Economics* 44 (2009), pp. 182–198.

[ACH14] W. H. Aeberhard, E. Cantoni, and S. Heritier. "Robust inference in the negative binomial regression model with an application to falls data". In: *Biometrics* 70 (2014), pp. 920–931.

[AD15] Stelios Arvanitis and Antonis Demos. "A class of indirect inference estimators: higher-order asymptotics and approximate bias correction". In: *The Econometrics Journal* 18.2 (2015), pp. 200–241.

[AGN12] E. F. Acar, C. Genest, and J. Neslehová. "Beyond simplified pair-copula constructions". In: *Journal of Multivariate Analysis* 110 (2012). Special Issue on Copula Modeling and Dependence, pp. 74–90.

[Alq+09] F. Alqallaf et al. "Propagation of outliers in multivariate data". In: *The Annals of Statistics* 37.1 (2009), pp. 311–331.

[ANG06] R. H. Abul Naga and P.-Y. Geoffard. "Decomposition of bivariate inequality indices by attributes". In: *Economics Letters* 90.3 (2006), pp. 362–367.

[ATF13] A. Alfons, M. Templ, and P. Filzmoser. "Robust estimation of economic indicators from survey samples based on Pareto tail modelling". In: *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 62.2 (2013), pp. 271–286.

[Atk11] A. B. Atkinson. "On lateral thinking". In: *Journal of Economic Inequality* 9 (2011), pp. 319–328.

[AVH05] E. S. Ahmed, A. I. Volodin, and A. A. Hussein. "Robust Weighted Likelihood Estimation of Exponential Parameters". In: *IEEE Transactions on Reliability* 54 (2005), pp. 389–395.

[Bas+98] A. Basu et al. "Robust and efficient estimation by minimising a density power divergence". In: *Biometrika* 85.3 (1998), pp. 549–559.

[BBR13] A. M. Bianco, G. Boente, and I. M. Rodrigues. "Resistant estimators in Poisson and Gamma models with missing responses and an application to outlier detection". In: *Journal of Multivariate Analysis* 114 (2013), pp. 209 –226.

[BC02] T. Bedford and R.M. Cooke. "Vines - a new graphical model for dependent random variables". In: *Annals of Statistics* 30 (2002), pp. 1031–1068.

[BCA12] E. C. Brechmann, C. Czado, and K. Aas. "Truncated regular vines in high dimensions with application to financial data". In: *Canadian Journal of Statistics* 40.1 (2012), pp. 68–85.

[BF81] Peter J Bickel and David A Freedman. "Some asymptotic theory for the bootstrap". In: *The annals of statistics* (1981), pp. 1196–1217.

[BGT+97] Ishwar V Basawa, VP Godambe, Robert Lee Taylor, et al. "Selected proceedings of the Symposium on Estimating Functions". In: IMS. 1997.

[BR09] S. Bonhomme and J.-M. Robin. "Assessing the Equalizing Force of Mobility Using Short Panels: France, 1990–2000". In: *Review of Economic Studies* 76 (2009), pp. 63–92.

[Bre14] E. C. Brechmann. “Hierarchical Kendall copulas: Properties and inference”. In: *Canadian Journal of Statistics* 42 (2014), pp. 78–108.

[Brz16] M. Brzezinski. “Robust estimation of the Pareto tail index: a Monte Carlo analysis”. In: *Empirical Economics* 51 (2016), pp. 1–30.

[BT74] Albert E Beaton and John W Tukey. “The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data”. In: *Technometrics* 16.2 (1974), pp. 147–185.

[BY11] A. Bergesio and V. J. Yohai. “Projection Estimators for Generalized Linear Models”. In: *Journal of the American Statistical Association* 106 (2011), pp. 661–671.

[Bí14] D. Bílková. “Robust Parameter Estimations Using L-Moments, TL-Moments and the Order Statistics”. In: *American Journal of Applied Mathematics* 2 (2014), pp. 36–53.

[Ć08] P. Cíźek. “Robust and Efficient Adaptive Estimation of Binary-Choice Regression Models”. In: *Journal of the American Statistical Association* 103 (2008), pp. 687–696.

[CB+05] Snigdhansu Chatterjee, Arup Bose, et al. “Generalized bootstrap for estimating equations”. In: *The Annals of Statistics* 33.1 (2005), pp. 414–436.

[CF06] Xiaohong Chen and Yanqin Fan. “Estimation of copula-based semiparametric time series models”. In: *Journal of Econometrics* 130.2 (2006), pp. 307–335.

[CFT06] X. Chen, Y. Fan, and V. Tsyrennikov. “Efficient Estimation of Semiparametric Multivariate Copula Models”. In: *Journal of the American Statistical Association* 101 (2006), pp. 1228–1240.

[CG17] F. Clementi and L. Gianmoena. “Chapter 9 - Modeling the Joint Distribution of Income and Consumption in Italy: A Copula-Based Approach With κ-Generalized Margins”. In: *Introduction to Agent-Based Economics*. Ed. by M. Gallegati, A. Palestrini, and A. Russo. Academic Press, 2017, pp. 191–228.

[Cha+09] Ngai-Hang Chan et al. “Statistical inference for multivariate residual copula of GARCH models”. In: *Statistica Sinica* (2009), pp. 53–70.

[Chr94] A. Christmann. “Least Median of Weighted Squares in Logistic Regression with Large Strata”. In: *Biometrika* 81 (1994), pp. 413–417.

[CHV08] L. Chollete, A. Heinen, and A. Valdesogo. “Modeling International Financial Returns with a Multivariate Regime Switching Copula”. In: *Journal of Financial Econometrics* 7 (2008), pp. 437–480.

[Cla78] D. G. Clayton. “A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence”. In: *Biometrika* 65.1 (1978), pp. 141–151.

[CP93] R. J. Carroll and S. Pederson. “On Robustness in the Logistic Regression Model”. In: 55 (1993), pp. 693–706.

[CR01] E. Cantoni and E. Ronchetti. “Robust Inference for Generalized Linear Models”. In: *Journal of the American Statistical Association* 96 (2001), pp. 1022–1030.

[CR06] E. Cantoni and E. Ronchetti. “A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures”. In: *Journal of Health Economics* 25 (2006), pp. 198–213.

[CVF02] F. A. Cowell and M.-P. Victoria-Feser. “Welfare Rankings in the Presence of Contaminated Data”. In: *Econometrica* 70 (2002), pp. 1221–1233.

[CVF06] F. A. Cowell and M.-P. Victoria-Feser. “Distributional Dominance with Trimmed Data”. In: *Journal of Business & Economic Statistics* 24 (2006), pp. 291–300.

[CVF08] F. A. Cowell and M.-P. Victoria-Feser. “Modelling Lorenz Curves: robust and semi-parametric issues”. In: *Modelling Income Distributions and Lorenz Curves*. Ed. by D. Chotikapanich. Springer, 2008, pp. 241–254.

[CVF96a] F. A. Cowell and M.-P. Victoria-Feser. “Poverty measurement with contaminated data: A robust approach”. In: *European Economic Review* 40 (1996), pp. 1761–1771.

[CVF96b] F. A. Cowell and M.-P. Victoria-Feser. “Robustness Properties of Inequality Measures”. In: *Econometrica.* 64 (1996), pp. 77–101.

[DC17] Y. Deng and N. R. Chaganty. “Hierarchical Archimedean copula models for the analysis of binary familial data”. In: *Statistics in Medicine* 37 (2017), pp. 590–597.

[Dec14] K. Decancq. “Copula-based measurement of dependence between dimensions of well-being”. In: *Oxford Economic Papers* 66 (2014), pp. 681–701.

[DG12] F. Domma and S. Giordano. “A stress–strength model with dependent variables to measure household financial fragility”. In: *Statistical Methods and Applications* 21 (2012), pp. 375–389.

[DL01] V. Dardanoni and P. J. Lambert. “Horizontal inequity comparisons”. In: *Social Choice and Welfare* 18 (2001), pp. 799–816.

[DM02] Debbie J Dupuis and Stephan Morgenthaler. “Robust weighted likelihood estimators with an application to bivariate extreme value problems”. In: *Canadian Journal of Statistics* 30.1 (2002), pp. 17–36.

[DM11] L. Denecke and C. H. Müller. “Robust estimators and tests for bivariate copulas based on likelihood depth”. In: *Computational Statistics and Data Analisys* 55 (2011), pp. 2724–2738.

[DVF06] D. J. Dupuis and M.-P. Victoria-Feser. “A Robust Prediction Error Criterion for Pareto Modeling of Upper Tails”. In: *The Canadian Journal of Statistics* 34 (2006), pp. 639–658.

[DW10] A. R. De Leon and B. Wu. “Copula-based regression models for a bivariate mixed discrete and continuous outcome”. In: *Statistics in Medicine* 30 (2010), pp. 175–185.

[Efr79] B. Efron. “Bootstrap Methods: Another Look at the Jackknife”. In: *The Annals of Statistics* 7.1 (1979), pp. 1–26.

[Fis+09] M. Fischer et al. “An empirical analysis of multivariate copula models”. In: *Quantitative Finance* 9 (2009), pp. 839–854.

[FL12] F. H. G. Ferreira and M. A. Lugo. *Multidimensional Poverty Analysis: Looking for a Middle Ground*. Tech. rep. IZA Policy Paper Series, 2012.

[Fra79] Maurice J Frank. "On the simultaneous associativity ofF (x, y) andx+y- F (x, y)". In: *Aequationes mathematicae* 19.1 (1979), pp. 194–226.

[FS94] C Field and B Smith. "Robust estimation: A weighted maximum likelihood approach". In: *International Statistical Review/Revue Internationale de Statistique* (1994), pp. 405–424.

[FZ14] Peter Fuleky and Eric Zivot. "Indirect inference based on the score". In: *The Econometrics Journal* 17.3 (2014), pp. 383–393.

[GDL00] M. G. Genton and X. De Luna. "Robust simulation-based estimation". In: *Statistics & probability letters* 48.3 (2000), pp. 253–259.

[GGR95] C. Genest, K. Ghoudi, and L.-P. Rivest. "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions". In: *Biometrika* 82.3 (1995), pp. 543–552.

[GHH17] J. Górecki, M. Hofert, and M. Holena. "On structure, family and parameter estimation of hierarchical Archimedean copulas". In: *Journal of Statistical Computation and Simulation* 87 (2017), pp. 3261–3324.

[GM96] Christian Gourieroux and Alain Monfort. *Simulation-based econometric methods*. Oxford university press, 1996.

[GMR93] Christian Gourieroux, Alain Monfort, and Eric Renault. "Indirect inference". In: *Journal of applied econometrics* 8.S1 (1993).

[GR03] M. G. Genton and E. Ronchetti. "Robust indirect inference". In: *Journal of the American Statistical Association* 98.461 (2003), pp. 67–76.

[GRV11] René Garcia, Eric Renault, and David Veredas. "Estimation of stable distributions by indirect inference". In: *Journal of Econometrics* 161.2 (2011), pp. 325–337.

[GS11] O. Grothe and J. Schnieders. "Spatial dependence in wind and optimal wind power allocation: A copula-based analysis". In: *Energy Policy* 39.9 (2011), pp. 4742–4754.

[Gue+18a] Stéphane Guerrier et al. "On the Properties of Simulation-based Estimators in High Dimensions". In: *arXiv preprint arXiv:1810.04443* (2018).

[Gue+18b] Stéphane Guerrier et al. "Simulation-Based Bias Correction Methods for Complex Models". In: *Journal of the American Statistical Association* (2018), pp. 1–12.

[Gum60] E. J. Gumbel. "Distributions des valeurs extrêmes en plusieurs dimensions". In: *Publ. Inst. Statist. Univ. Paris* 9 (1960), pp. 171–173.

[Ham+86] F. R. Hampel et al. *Robust statistics: the approach based on influence functions*. Wiley, 1986.

[Ham74] F. R. Hampel. "The influence curve and its role in robust estimation". In: *Journal of the American Statistical Association* 69.346 (1974), pp. 383–393.

[Han94] Bruce E Hansen. "Autoregressive conditional density estimation". In: *International Economic Review* (1994), pp. 705–730.

[He+12] J. He et al. “A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies”. In: *Biostatistics* 13 (2012), pp. 497–508.

[Hec79] James J Heckman. “Sample Selection Bias as a Specification Error”. In: *Econometrica* 47.1 (1979), pp. 153–161.

[Her+09] S. Heritier et al. *Robust methods in Biostatistics.* Vol. 838. Wiley, 2009.

[HFZ05] X. He, W. K. Fung, and Z. Zhu. “Robust Estimation in Generalized Partial Linear Models for Clustered Data”. In: *Journal of the American Statistical Association* 100 (2005), pp. 1176–1184.

[HK00] Feifang Hu and John D Kalbfleisch. “The estimating function bootstrap”. In: *Canadian Journal of Statistics* 28.3 (2000), pp. 449–481.

[HMM12] M. Hofert, M. Mächler, and A. J. McNeil. “Likelihood inference for Archimedean copulas in high dimensions under known margins”. In: *Journal of Multivariate Analysis* 110 (2012). Special Issue on Copula Modeling and Dependence, pp. 133–150.

[Hof08] M. Hofert. “Sampling Archimedean copulas”. In: *Computational Statistics and Data Analysis* 52 (2008), pp. 5163–5174.

[Hof10] M. Hofert. “Construction and sampling of nested Archimedean copulas”. In: *Copula Theory and Its Applications.* Ed. by P. Jaworski et al. Berlin: Springer, 2010, pp. 147–160.

[Hof12] M Hofert. “A stochastic representation and sampling algorithm for nested Archimedean copulas”. In: *Journal of Statistical Computation and Simulation* 82 (2012), pp. 1239–1255.

[Hou86] P. Hougaard. “A class of multivariate failure time distributions”. In: *Biometrika* 73.3 (1986), pp. 671–678.

[HR09] P. J. Huber and E. Ronchetti. *Robust Statistics.* Wiley, 2009.

[Hub+64] Peter J Huber et al. “Robust estimation of a location parameter”. In: *The annals of mathematical statistics* 35.1 (1964), pp. 73–101.

[Joe05] H. Joe. “Asymptotic efficiency of the two-stage estimation method for copula-based models”. In: *Journal of Multivariate Analysis* 94.2 (2005), pp. 401–419.

[Joe14] Harry Joe. *Dependence modeling with copulas.* CRC Press, 2014.

[Joe96] H. Joe. “Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters”. In: *Distributions with fixed marginals and related topics.* Ed. by L. Rüschendorf, B. Schweizer, and M. D. Taylor. Hayward, CA: Institute of Mathematical Statistics, 1996, pp. 120–141.

[Joe97] H. Joe. *Multivariate models and multivariate dependence concepts.* Vol. 73. Chapman & Hall/CRC, 1997.

[JR03] Eric Jondeau and Michael Rockinger. “Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements”. In: *Journal of Economic dynamics and Control* 27.10 (2003), pp. 1699–1737.

[JSVK15] M. Jäntti, E. Sierminska, and P. Van Kerm. *Modelling the joint distribution of income and wealth.* IZA Discussion paper No. 9190. 2015.

[JT04] Wenxin Jiang and Bruce Turnbull. “The indirect method: inference based on intermediate statistics—a synthesis and examples”. In: *Statistical Science* 19.2 (2004), pp. 239–263.

[JX96] H. Joe and J. J. Xu. *The estimation method of inference functions for margins for multivariate models*. Technical Report 166. Department of Statistics, University of British Columbia, 1996.

[KC06] D. Kurowicka and R. Cooke. *Uncertainty analysis with high dimensional dependence modelling*. Chichester: Wiley, 2006.

[Ken38] M. G. Kendall. “A New Measure Of Rank Correlation”. In: *Biometrika* 30 (1938), pp. 81–93.

[KHG18] Pavel Krupskii, Raphaël Huser, and Marc G Genton. “Factor copula models for replicated spatial data”. In: *Journal of the American Statistical Association* 113.521 (2018), pp. 467–479.

[KHK15] C. Kluppelberg, S. Haug, and G. Kuhn. “Copula structure analysis based on extreme dependence”. In: *Statistics and Its Interface* 8 (2015), pp. 93–107.

[KJ11] D. Kurowicka and H. Joe. *Dependence Modeling: Vine Copula Handbook*. World Scientific, 2011.

[KJ13] Pavel Krupskii and Harry Joe. “Factor copula models for multivariate data”. In: *Journal of Multivariate Analysis* 120 (2013), pp. 85–101.

[KK09] C. Kluppelberg and G. Kuhn. “Copula structure analysis”. In: *Journal of the Royal Statististical Society, Series B* 71 (2009), pp. 737–753.

[KL13] B. Kim and S. Lee. “Robust Estimation for Copula Parameter in SCOMDY Models”. In: *Journal of Time Series Analysis* 34.3 (2013), pp. 302–314.

[KM97] G. A. Koshevoy and K. Mosler. “Multivariate Gini indices”. In: *Journal of Multivariate Analysis* 60 (1997), pp. 252–276.

[KSS07] G. Kim, M. J. Silvapulle, and P. Silvapulle. “Comparison of semiparametric and parametric methods for estimating copulas”. In: *Computational Statistics & Data Analysis* 51.6 (2007), pp. 2836–2850.

[McF89] D. McFadden. “A method of simulated moments for estimation of discrete response models without numerical integration”. In: *Econometrica* (1989), pp. 995–1026.

[McN08] A. J. McNeil. “Sampling nested Archimedean copulas”. In: *Journal of Statistical Computation and Simulation* 78 (2008), pp. 567–581.

[MF11] L. Madsen and Y. Fang. “Joint Regression Analysis for Discrete Longitudinal Data”. In: *Biometrics* 67 (2011), pp. 1171–1175.

[MFE05] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: concepts, techniques, and tools*. Princeton university press, 2005.

[MMN07] B. V. M. Mendes, E. F. L. de Melo, and R. B. Nelsen. “Robust fits for copula models”. In: *Communications in Statistics - Simulation and Computation* 36.5 (2007), pp. 997–1017.

[MN09] A. J. McNeil and J. Neslehová. “Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions”. In: *The Annals of Statistics* 37 (2009), pp. 3059–3097.

[MVF06] Irini Moustaki and Maria-Pia Victoria-Feser. “Bounded-influence robust estimation in generalized linear latent variable models”. In: *Journal of the American Statistical Association* 101.474 (2006), pp. 644–653.

[MY04] A. Marazzi and V. J. Yohai. “Adaptively truncated maximum likelihood regression with asymmetric errors”. In: *Journal of Statistical Planning and Inference* 122 (2004), pp. 271–291.

[ND05] S Nadarajah and DK Dey. “Convolutions of the T distribution”. In: *Computers & Mathematics with Applications* 49.5-6 (2005), pp. 715–721.

[Nel06] R. B. Nelsen. *An introduction to copulas*. Springer, 2006.

[Nik13] A.K. Nikoloulopoulos. “Copula-Based Models for Multivariate Discrete Response Data”. In: *Copulae in Mathematical and Quantitative Finance*. Ed. by Jaworski P., Durante F., and Härdle W. Vol. 213. Lecture Notes in Statistics. Berlin, Heidelberg: Springer, 2013, pp. 231–249.

[NJR11] A. K. Nikoloulopoulos, H. Joe, and N. R. Rao Chaganty. “Weighted scores method for regression models with dependent data”. In: *Biostatistics* 12 (2011), pp. 653–665.

[OM92] David Oakes and Amita K Manatunga. “Fisher information for a bivariate extreme value distribution”. In: *Biometrika* 79.4 (1992), pp. 827–832.

[OOS13] O. Okhrin, Y. Okhrin, and W. Schmid. “On the structure and estimation of hierarchical Archimedean copulas”. In: *Journal of Econometrics* 173 (2013), pp. 189–204.

[OP13] D. W. Oh and A. J. Patton. “Simulated Method of Moments Estimation for Copula-Based Multivariate Models”. In: *Journal of the American Statistical Association* 108 (2013), pp. 689–700.

[OP17] Dong Hwan Oh and Andrew J Patton. “Modeling dependence in high dimensions with factor copulas”. In: *Journal of Business & Economic Statistics* 35.1 (2017), pp. 139–154.

[Ors13] Samuel Orso. “Robust Estimation for Bivariate Distribution”. eng. In: *Master Thesis* (2013). URL http://archive-ouverte.unige.ch/unige:33672. URL: http://archive-ouv

[Pat06] A. J. Patton. “Modelling Asymmetric Exchange Rate Dependence”. In: *International Economic Review* 47 (2006), pp. 527–556.

[PB02] A. M. Pires and J. A. Branco. “Partial influence functions”. In: *Journal of Multivariate Analysis* 83.2 (2002), pp. 451–468.

[PCJ12] A. Panagiotelis, C. Czado, and H. Joe. “Pair copula constructions for multivariate discrete data”. In: *Journal of the American Statistical Association* 107 (2012), pp. 1063–1072.

[Phi12] Peter CB Phillips. “Folklore theorems, implicit maps, and indirect inference”. In: *Econometrica* 80.1 (2012), pp. 425–454.

[PP89] A. Pakes and D. Pollard. “Simulation and the Asymptotics of Optimization Estimators”. In: *Econometrica* 57 (1989), pp. 1027–1057.

[PQ99] J. S. Preisser and B. F. Qaqish. “Robust Regression for Clustered Data with Application to Binary Responses”. In: *Biometrics* 55 (1999), pp. 574–579.

[PS18] L. A. Prendergast and R. G. Staudte. “A Simple and Effective Inequality Measure”. In: *The American Statistician* (2018). online publication.

[PSF12] Gareth W Peters, Scott A Sisson, and Yanan Fan. “Likelihood-free Bayesian inference for α-stable models”. In: *Computational Statistics & Data Analysis* 56.11 (2012), pp. 3743–3756.

[Qui09] C. Quinn. *Measuring income-related inequalities in health using a parametric dependence function.* HEDG Working Paper no 09/24. 2009.

[RC03] P. Rousseeuw and A. Christmann. “Robustness against separation and outliers in logistic regression”. In: *Computational Statistics and Data Analysis* 43 (2003), pp. 315–332.

[Rez15] M. Rezapour. “On the construction of nested Archimedean copulas for d-monotone generators”. In: *Statistics and Probability Letters* 101 (2015), pp. 21–32.

[RH99] P. J. Rousseeuw and M. Hubert. “Regression Depth”. In: *Journal of the American Statistical Association* 94 (1999), pp. 388–402.

[Rob+15] T. Robertson et al. “The role of material, psychosocial and behavioral factors in mediating the association between socioeconomic position and allostatic load (measured by cardiovascular, metabolic and inflammatory markers)”. In: *Brain, Behavior, and Immunity* 45 (2015), pp. 41–49.

[RVF97] E. Ronchetti and M.-P. Victoria-Feser. “Resistant Modelling of Income Distribution and Inequality Measures”. In: *The Practice of Data Analysis: Festschrift in Honour of John W. Tukey for his 80th Birthday.* Ed. by D. R. Brillinger, L. T. Fernholz, and Stephan Morgenthaler. Princeton: Princeton University Press, 1997.

[Rém17] Bruno Rémillard. “Goodness-of-fit tests for copulas of multivariate time series”. In: *Econometrics* 5.1 (2017), p. 13.

[Skl59] M. Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Université Paris 8, 1959.

[SL95] J. H. Shih and T. A. Louis. “Inferences on the association parameter in copula models for bivariate survival data”. In: *Biometrics* (1995), pp. 1384–1399.

[SLY09] P. X. K. Song, M. Li, and Y. Yuan. “Joint Regression Analysis of Correlated Data Using Gaussian Copulas”. In: *Biometrics* 65 (2009), pp. 60–68.

[SM76] S. K. Singh and G. S. Maddala. “A Function for Size Distribution of Incomes”. In: *Econometrica* 44.5 (1976), pp. 963–70.

[Smi+10] M. Smith et al. “Modeling Longitudinal Data Using a Pair-Copula Decomposition of Serial Dependence”. In: *Journal of the American Statistical Association* 105 (2010), pp. 1467–1479.

[Smi93] Anthony A Smith. “Estimating nonlinear time-series models using simulated vector autoregressions”. In: *Journal of Applied Econometrics* 8.S1 (1993).

[Son+05] P. X.-K. Song et al. “Maximization by Parts in Likelihood Inference [With Comments, Rejoinder]”. In: *Journal of the American Statistical Association* 100 (2005), pp. 1145–1167.

[Son00] P. X. K. Song. “Multivariate Dispersion Models Generated From Gaussian Copula”. In: *Scandinavian Journal of Statistics* 27 (2000), pp. 305–320.

[Spe04] C. Spearman. “General Intelligence Objectively Determined and Measured”. In: *American Journal of Psychology* 15 (1904), pp. 201–293.

[St15] J. Stöber et al. “Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses”. In: *Computational Statististics and Data Analysis* 88 (2015), pp. 28–39.

[Uyt18] N. Uyttendaele. “On the estimation of nested Archimedean copulas: a theoretical and an experimental comparison”. In: *Computational Statistics”* 33 (2018), pp. 1047–1070.

[Vaa98] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 1998.

[Van+07] B. Vandewalle et al. “A robust estimator for the tail index of Pareto-type distributions”. In: *Computational Statistics and Data Analysis* 51 (2007), pp. 6252–6268.

[VF02] M.-P. Victoria-Feser. “Robust Inference with Binary Data”. In: *Psychometrika* 67 (2002), pp. 21–32.

[VFR94a] M.-P. Victoria-Feser and E. Ronchetti. “Robust Methods for Personal-Income Distribution Models”. In: *Canadian Journal of Statistics* 22 (1994), pp. 247–258.

[VFR97] M.-P. Victoria-Feser and E. Ronchetti. “Robust Estimation for Grouped Data”. In: *Journal of the American Statistical Association* 92 (1997), pp. 333–340.

[VGC10] A. Vinh, W. E. Griffiths, and D. Chotikapanich. “Bivariate income distributions for assessing inequality and poverty under dependent samples”. In: *Economic Modelling* 27.6 (2010), pp. 1473–1483.

[VY14] M. Valdora and V. J. Yohai. “Robust estimators for generalized linear models”. In: *Journal of Statistical Planning and Inference* 146 (2014), pp. 31–48.

[Wed74] R. W. M. Wedderburn. “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method”. In: *Biometrika* 61 (1974), pp. 439–447.

[Xu96] J. J. Xu. “Statistical modelling and inference for multivariate and longitudinal discrete response data”. PhD thesis. University of British Columbia, 1996.

[ZGR12] M. Zhelonkin, M. G. Genton, and E. Ronchetti. “On the robustness of two-stage estimators”. In: *Statistics & Probability Letters* 82 (2012), pp. 726–732.

[ZGR16] Mikhail Zhelonkin, Marc G Genton, and Elvezio Ronchetti. “Robust inference in sample selection models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.4 (2016), pp. 805–827.

[ZK16] A. A. Zilko and D. Kurowicka. “Copula in a Multivariate Mixed Discrete-continuous Model”. In: *Computational Statistics and Data Analalysis* 103 (2016), pp. 28–55.

[Hob13] I. Hobaek Haff. “Parameter estimation for pair-copula constructions”. In: *Bernoulli* 19 (2013), pp. 462–491.

Appendix

2.A Intermediate results

Lemma 2.A.1 (asymptotic distribution of $\hat{\boldsymbol{\pi}}_{n^*}$ in (2.3)). *Let $\boldsymbol{\Pi}$ be a convex open set in $\mathbb{R}^q$. Let $\mathcal{V}$ be a compact convex open set in $\mathbb{R}^r$. If Assumptions 2.9 and 2.11 hold, and $\hat{\boldsymbol{\pi}}_{n^*}$ is pointwise convergent, then*

$$n^{1/2}\left(\hat{\boldsymbol{\pi}}_{n^*} - \boldsymbol{\pi}_0\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{K}^{-1}\left[({}^{1}\!/\!{}_{\rho^*})\mathbf{Q} + \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^T\right]\mathbf{K}^{-T}\right).$$

Proof. Fix $\boldsymbol{\nu}_0 \in \mathcal{V}$, $\boldsymbol{\pi}_0 \in \boldsymbol{\Pi}$ and $\rho^* \in (0,1]$. From the mean value theorem stated in Lemma 3.8 we have

$$\boldsymbol{\Phi}_{n^*}\left(\hat{\boldsymbol{\nu}}_n, \hat{\boldsymbol{\pi}}_{n^*}\right) - \boldsymbol{\Phi}_{n^*}\left(\hat{\boldsymbol{\nu}}_n, \boldsymbol{\pi}_0\right) = \mathbf{K}_{n^*}\left(\hat{\boldsymbol{\nu}}_n\right) \cdot \left(\hat{\boldsymbol{\pi}}_{n^*} - \boldsymbol{\pi}_0\right) + o_p\left(\left\|\hat{\boldsymbol{\pi}}_{n^*} - \boldsymbol{\pi}_0\right\|\right).$$

Using again Lemma 3.8 and multiplying by $n^{1/2}$ leads to

$$\begin{aligned} n^{1/2}\boldsymbol{\Phi}_{n^*}\left(\hat{\boldsymbol{\nu}}_n, \hat{\boldsymbol{\pi}}_{n^*}\right) - & n^{1/2}\left[\boldsymbol{\Phi}_{n^*}\left(\boldsymbol{\nu}_0, \boldsymbol{\pi}_0\right) + \mathbf{L}_{n^*} \cdot \left(\hat{\boldsymbol{\nu}}_n - \boldsymbol{\nu}_0\right) + o_p\left(\left\|\hat{\boldsymbol{\nu}}_n - \boldsymbol{\nu}_0\right\|\right)\right] \\ & = \mathbf{K}_{n^*}(\hat{\boldsymbol{\nu}}_n) \cdot n^{1/2}\left(\hat{\boldsymbol{\pi}}_{n^*} - \boldsymbol{\pi}_0\right) + n^{1/2} o_p\left(\left\|\hat{\boldsymbol{\pi}}_{n^*} - \boldsymbol{\pi}_0\right\|\right). \end{aligned}$$

By definition $\boldsymbol{\Phi}_{n^*}\left(\hat{\boldsymbol{\nu}}_n, \hat{\boldsymbol{\pi}}_{n^*}\right) = \mathbf{0}$. By assumption $n^{*1/2}\boldsymbol{\Phi}_{n^*}(\boldsymbol{\nu}_0, \boldsymbol{\pi}_0)$ satisfies the Lindeberg-Feller central limit theorem, $n^{*1/2}\boldsymbol{\Phi}_{n^*}(\boldsymbol{\nu}_0, \boldsymbol{\pi}_0) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{Q}\right)$. Note that we have $n^{1/2} \sim n^{*1/2}/\sqrt{\rho^*}$, thus $n^{1/2}\boldsymbol{\Phi}_{n^*} \rightsquigarrow \mathcal{N}(\mathbf{0}, ({}^{1}\!/\!{}_{\rho^*})\mathbf{Q})$. By assumption $\mathbf{L}_{n^*}$ converges in probability to $\mathbf{L}$. By hypothesis, $n^{1/2}(\hat{\boldsymbol{\nu}}_n - \boldsymbol{\nu}_0) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

We need to demonstrate the uniform convergence in probability of $\mathbf{K}_{n^*}(\hat{\boldsymbol{\nu}}_n)$. By the continuity of $\mathbf{K}_{n^*}$, the continuous mapping theorem is satisfied (see [Vaa98]) so $\mathbf{K}_{n^*}(\hat{\boldsymbol{\nu}}_n) = \mathbf{K}_{n^*} + o_p(1)$. Moreover, by assumption $\mathbf{K}_{n^*}$ converges in probability to $\mathbf{K}$, so $\mathbf{K}_{n^*}(\hat{\boldsymbol{\nu}}_n) \xrightarrow{p} \mathbf{K}$ pointwise. By the additional assumption that $\mathbf{K}_{n^*}(\boldsymbol{\nu})$ is stochastically Lipschitz and $\mathcal{V}$ is compact, Lemma 3.6 is satisfied and as a conclusion $\mathbf{K}_{n^*}(\boldsymbol{\nu}) \xrightarrow{p} \mathbf{K}(\boldsymbol{\nu})$ uniformly for all $\boldsymbol{\nu} \in \mathcal{V}$, and $\mathbf{K}(\boldsymbol{\nu})$ is uniformly continuous. Following, this result, we obtain

$$\begin{aligned} \left\|\mathbf{K}_{n^*}(\hat{\boldsymbol{\nu}}_n) - \mathbf{K}\right\| & \leq \left\|\mathbf{K}_{n^*}(\hat{\boldsymbol{\nu}}_n) - \mathbf{K}(\hat{\boldsymbol{\nu}}_n)\right\| + \left\|\mathbf{K}(\hat{\boldsymbol{\nu}}_n) - \mathbf{K}\right\| \\ & \leq \sup_{\boldsymbol{\nu} \in \mathcal{V}} \left\|\mathbf{K}_{n^*}(\boldsymbol{\nu}) - \mathbf{K}(\boldsymbol{\nu})\right\| + \left\|\mathbf{K}(\hat{\boldsymbol{\nu}}_n) - \mathbf{K}\right\|. \end{aligned}$$

The first term of the last inequality converges to zero by the uniform convergence result, and the second term converge to zero by the continuous mapping theorem. Consequently, $\mathbf{K}_{n^*}(\hat{\boldsymbol{\nu}}_n) \xrightarrow{p} \mathbf{K}$ uniformly. Since $\mathbf{K}$ is invertible, we obtain by Slutsky's lemma

$$n^{1/2}\left(\hat{\boldsymbol{\pi}}_{n^*} - \boldsymbol{\pi}_0\right) = -\mathbf{K}^{-1}\left[\rho^{*-1/2}\mathbf{Q}^{1/2}\mathbf{z}_q + \mathbf{L}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_r + o_p(1)\right] - \mathbf{K}^{-1}o_p(1),$$

where $\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. The proof follows from well-known properties of Gaussian distributed random variables. □

Lemma 2.A.2 (asymptotic distribution of $\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}_0)$ in (2.3))**.** *Let* $\boldsymbol{\Pi}$ *be convex open set in* $\mathbb{R}^q$ *and suppose* $\boldsymbol{\theta}_0$ *is an interior point of* $\boldsymbol{\Theta}$*. If Assumptions 2.9 and 2.11 hold, and* $\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}_0)$ *is pointwise convergent to, say* $\boldsymbol{\pi}_0$*, then*

$$n^{1/2}\left(\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}_0)-\boldsymbol{\pi}_0\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, (1/\rho^*)\gamma\mathbf{K}^{-1}\mathbf{Q}\mathbf{K}^{-T}\right),$$

where $\gamma = 1/BH$*.*

Proof. Let $\boldsymbol{\pi}_0 \equiv \boldsymbol{\pi}(\boldsymbol{\theta}_0)$ and fix $H \in \mathbb{R}^+$, $B \in \mathbb{N}^+$ and $\rho^* \in (0,1]$. Let us pick the bth auxiliary estimator $\hat{\boldsymbol{\pi}}_m^{(b)}$ from the sequence $\{\hat{\boldsymbol{\pi}}_m^{(b)} : b = 1,\cdots,B\}$. From Lemma 3.8 we have

$$\boldsymbol{\Phi}_m\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\pi}}_m^{(b)}\right)-\boldsymbol{\Phi}_m\left(\boldsymbol{\theta}_0,\boldsymbol{\pi}_0\right)=\mathbf{K}_m\cdot\left(\hat{\boldsymbol{\pi}}_m^{(b)}-\boldsymbol{\pi}_0\right)+o_p\left(\left\|\hat{\boldsymbol{\pi}}_m^{(b)}-\boldsymbol{\pi}_0\right\|\right)$$

By definition $\boldsymbol{\Phi}_m\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\pi}}_m^{(b)}\right)=\mathbf{0}$. By assumption $\mathbf{K}_m$ is pointwise convergent to $\mathbf{K}$. Since $\mathbf{K}$ is invertible, multiplying the above by $n^{1/2}$ yields

$$n^{1/2}\left(\hat{\boldsymbol{\pi}}_m^{(b)}-\boldsymbol{\pi}_0\right)=-(H\rho^*)^{-1/2}\mathbf{K}^{-1}\mathbf{Q}^{1/2}\mathbf{z}_q-n^{1/2}\mathbf{K}^{-1}o_p(1),$$

where $\mathbf{z}_q \sim \mathcal{N}(\mathbf{0},\mathbf{I}_q)$. Indeed, by assumption $m^{1/2}\boldsymbol{\Phi}_m(\boldsymbol{\theta}_0,\boldsymbol{\pi}_0)$ satisfies the Lindeberg-Feller central limit theorem so it converges in distribution to $\mathcal{N}(\mathbf{0},\mathbf{Q})$. Remark that $\sqrt{n} \sim \sqrt{m}/\sqrt{H\rho^*}$. The rest of the above result follows from Slutsky's lemma. To conclude the proof, one can remark that the distribution for the average $\bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta}_0) \equiv \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\pi}}_m(\boldsymbol{\theta}_0)$ follows directly by property of the variance of the averaged of identically and independently distributed Gaussian random variable. □

2.B Main proofs

Proof of Proposition 2.5. The proof results directly from the mean value inequality stated in Lemma 3.7. □

Proof of Proposition 2.6. Fix $\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ and define the line segment $\lambda = \boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2$, $t \in [0,1]$. From the mean value inequality in Lemma 3.7 we have

$$\left\|\boldsymbol{\pi}\left(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2\right)-\boldsymbol{\pi}\left(\boldsymbol{\theta}_1\right)-D\boldsymbol{\pi}(\boldsymbol{\theta}_0)\cdot\boldsymbol{\theta}_2\right\| \leq \sup_{\lambda}\left\|D\boldsymbol{\pi}(\lambda)-D\boldsymbol{\pi}(\boldsymbol{\theta}_0)\right\|\cdot\left\|\boldsymbol{\theta}_2\right\|.$$

Let $L \equiv \|D\boldsymbol{\pi}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_2\|$. Since $D\boldsymbol{\pi}(\boldsymbol{\theta}_0)$ is not vanishing and full column rank, $L > 0$. Let $\mathcal{B}(\boldsymbol{\theta}_0,\|\boldsymbol{\theta}_2\|)$ be a closed neighborhood of $\boldsymbol{\theta}_0$ with radius $\|\boldsymbol{\theta}_2\|$. Let choose $\boldsymbol{\theta}_2$ such that

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_0,\|\boldsymbol{\theta}_2\|)}\|D\boldsymbol{\pi}(\boldsymbol{\theta})-D\boldsymbol{\pi}(\boldsymbol{\theta}_0)\| \leq \frac{L}{2}.$$

By the triangle inequality we obtain

$$\begin{aligned}\|\boldsymbol{\pi}(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2)-\boldsymbol{\pi}(\boldsymbol{\theta}_1)\| &\geq \|D\boldsymbol{\pi}(\boldsymbol{\theta}_0)\cdot\boldsymbol{\theta}_2\|-\|\boldsymbol{\pi}(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2)-\boldsymbol{\pi}(\boldsymbol{\theta}_1)-D\boldsymbol{\pi}(\boldsymbol{\theta}_0)\cdot\boldsymbol{\theta}_2\|\\ &\geq L-\sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_0,\|\boldsymbol{\theta}_2\|)}\|D\boldsymbol{\pi}(\boldsymbol{\theta})-D\boldsymbol{\pi}(\boldsymbol{\theta}_0)\|\cdot\|\boldsymbol{\theta}_2\|\\ &\geq \frac{L}{2}\|\boldsymbol{\theta}_2\|\end{aligned}$$

which concludes the proof. □

Proof of Theorem 2.8. We proceed by verifying the assumptions for the weak consistency result of Lemma 3.1. We separate the proof in four parts. We start by showing (*i*) the pointwise convergence of the auxiliary estimator for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then we demonstrate (*ii*) the uniform convergence of the auxiliary estimator, followed by (*iii*) the uniform convergence of the stochastic objective function. Eventually we show that (*iv*) the non-stochastic objective function has a unique minimum.

(*i*). We proceed by verifying the assumptions of Lemma 3.1. Fix $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$. Fix also $H \in \mathbb{R}^+$ and $B \in \mathbb{N}^+$ in (2.3) so m diverges at the rate of n and $\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}_1)$ and $\bar{\boldsymbol{\pi}}_n(\boldsymbol{\theta}_1)$ are alike whenever n diverges. The n^* diverges at the rate of n so we use only the notation with n in the proof. By assumption $\hat{\boldsymbol{\nu}}_n = \boldsymbol{\nu}_0 + o_p(1)$ and $\boldsymbol{\Phi}_n(\boldsymbol{\nu}, \boldsymbol{\pi})$ is continuous at every $\boldsymbol{\nu} \in \mathcal{V}$, so by the continuous mapping theorem ([Vaa98]) we have

$$\boldsymbol{\Phi}_n(\hat{\boldsymbol{\nu}}_n, \boldsymbol{\pi}) = \boldsymbol{\Phi}_n(\boldsymbol{\nu}_0, \boldsymbol{\pi}) + o_p(1).$$

Since the expectation of $\{\boldsymbol{\Phi}_n\}$ exists, $\{\boldsymbol{\Phi}_n\}$ converges pointwise in probability to $\boldsymbol{\Phi}$ by the weak law of large numbers. By assumption $\boldsymbol{\Phi}_n$ is globally Lipschitz, the condition of Lemma 3.5 is verified so $\{\boldsymbol{\Phi}_n\}$ is stochastically uniformly equicontinuous. By compactness of $\boldsymbol{\Pi}$, Lemma 3.3 yields the uniform convergence of $\{\boldsymbol{\Phi}_n\}$ and the uniform continuity of $\boldsymbol{\Phi}$. By compactness and continuity of $\boldsymbol{\Phi}$, the infinum of the norm of $\boldsymbol{\Phi}$ exists. The minimum of $\boldsymbol{\Phi}$ is well-separated by the bijectivity of the function. Therefore, we have by Lemma 3.1 that the sequence $\{\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}_1)\}$ converges pointwise to $\boldsymbol{\pi}(\boldsymbol{\theta}_1)$.

(ii). We have by assumption that $\{\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})\}$ is globally Lipschitz so by Lemma 3.4 it is also stochastically uniformly equicontinuous. Since $\boldsymbol{\Theta}$ is compact, we have by Lemma 3.3 that $\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})$ is uniformly convergent and $\boldsymbol{\pi}(\boldsymbol{\theta})$ is uniformly continuous.

(iii). By assumption the matrix $\boldsymbol{\Omega}$ is symmetric positive-definite. A direct application of the Courant-Fischer minimax theorem ([GVL12]) gives the following upper bound

$$Q(\boldsymbol{\theta}) = \|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Omega}} \leq \lambda_{\max} \|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\|^2 ,$$

where $\boldsymbol{\pi}_0 \equiv \boldsymbol{\pi}(\boldsymbol{\theta}_0)$ and $\lambda_{\max}$ is the maximum eigenvalue of $\boldsymbol{\Omega}$. The Gershgorin circle theorem ([GVL12]) gives an upper bound on the eigenvalues, so

$$\lambda_{\max} \leq \max_i \sum_{j=1}^{q} |\omega_{ij}| = k,$$

the largest eigenvalue is bounded by the maximum row sum of absolute elements. By assumption k is finite. Similarly, denote $\hat{\lambda}_{\max}$ the largest eigenvalue of $\widehat{\boldsymbol{\Omega}}$. We have by assumption that $\hat{\lambda}_{\max} \leq k + \gamma_n$ where γ_n is $o_p(1)$. Without loss of generality, let assume that γ_n is positive so that

$$\begin{aligned} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| &\leq \left| (k+\gamma_n) \|\hat{\boldsymbol{\pi}}_n - \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})\|^2 - k \|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\|^2 \right| \\ &\leq k \left| \|\hat{\boldsymbol{\pi}}_n - \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})\|^2 - \|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\|^2 \right| + \gamma_n \|\hat{\boldsymbol{\pi}}_n - \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})\|^2 , \end{aligned} \tag{2.7}$$

where we use the triangle inequality for the second inequality. For the left-hand side of (2.7), we obtain from the reverse triangle inequality and the triangle inequality

$$\begin{aligned} \left| \|\hat{\boldsymbol{\pi}}_n - \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})\|^2 - \|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\|^2 \right| &\leq \|\hat{\boldsymbol{\pi}}_n - \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}) - \boldsymbol{\pi}_0 + \boldsymbol{\pi}(\boldsymbol{\theta})\|^2 \\ &\leq \|\hat{\boldsymbol{\pi}}_n - \boldsymbol{\pi}_0\|^2 + \|\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}) - \boldsymbol{\pi}(\boldsymbol{\theta})\|^2 . \end{aligned}$$

For the right-hand side of (2.7), by using the triangle inequality we have

$$\left\|\hat{\boldsymbol{\pi}}_n - \hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta})\right\|^2 \leq \left\|\hat{\boldsymbol{\pi}}_n - \boldsymbol{\pi}_0\right\|^2 + \left\|\hat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}) - \boldsymbol{\pi}(\boldsymbol{\theta})\right\|^2 + \left\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\right\|^2 .$$

From the results of parts (i) and (ii), we obtain the following

$$\lim_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| > \varepsilon\right) \leq k o_p(1) + k o_p(1) + \gamma_n o_p(1) + \gamma_n o_p(1) + \gamma_n \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\right\|^2 .$$

Since the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}$ is unifomly continuous, continuity preserved by the norm, and $\boldsymbol{\Theta}$ is compact, the supremum exists so $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})\right\|^2 < \infty$. Since γ_n is $o_p(1)$, the objective function converges uniformly.

(iv). It follows from the uniform continuity of $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}$ that $Q(\boldsymbol{\theta})$ is uniformly continuous. Since $\boldsymbol{\Theta}$ is compact, the infimum exists. By the injectivity of $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}$, the infimum is wel-separated, which concludes the proof. □

Proof of Proposition 2.10. The proof results directly from the implicit function theorem stated in Lemma 3.10. □

Proof of Theorem 2.12. Fix $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^\circ$, $\boldsymbol{\pi}_0 \in \boldsymbol{\Pi}^\circ$, $H \in \mathbb{R}^+$, $B \in \mathbb{N}^+$ and $\rho^* \in (0,1]$. Let $\mathbf{g}_{(n^*,m)}(\boldsymbol{\theta}) \equiv \hat{\boldsymbol{\pi}}_{n^*} - \bar{\boldsymbol{\pi}}_m(\boldsymbol{\theta})$. From Proposition 2.10, we have the following when n is sufficiently large

$$D_{\boldsymbol{\theta}} \mathbf{g}_{(n^*,m)}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \equiv \mathbf{H}_m(\boldsymbol{\theta}_0) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{K}_m^{-1}\left(\boldsymbol{\pi}_1^{(b)}\right) \mathbf{J}_m\left(\boldsymbol{\pi}_1^{(b)}\right),$$

where $\boldsymbol{\pi}_1^{(b)} \equiv \hat{\boldsymbol{\pi}}_m^{(b)}(\boldsymbol{\theta}_0)$. Under the condition of this theorem, it is equivalent to solve

$$\hat{\boldsymbol{\theta}}_{n^*} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}^\circ}{\operatorname{argmin}} \left\|\mathbf{g}_{(n^*,m)}(\boldsymbol{\theta})\right\|^2_{\widehat{\boldsymbol{\Omega}}} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}^\circ}{\operatorname{argzero}}\, \mathbf{H}_m^T(\boldsymbol{\theta}) \widehat{\boldsymbol{\Omega}} \mathbf{g}_{(n^*,m)}(\boldsymbol{\theta}).$$

By the mean value theorem (Lemma 3.8) we have the following

$$\mathbf{g}_{(n^*,m)}\left(\hat{\boldsymbol{\theta}}_{n^*}\right) - \mathbf{g}_{(n^*,m)}\left(\boldsymbol{\theta}_0\right) = \mathbf{H}_m(\boldsymbol{\theta}_0)\cdot\left(\hat{\boldsymbol{\theta}}_{n^*} - \boldsymbol{\theta}_0\right) + o_p\left(\left\|\hat{\boldsymbol{\theta}}_{n^*} - \boldsymbol{\theta}_0\right\|\right).$$

Using this result, the above equivalence and multiplying by square-root n leads to

$$\begin{aligned} & n^{1/2}\mathbf{H}_m^T\left(\hat{\boldsymbol{\theta}}_{n^*}\right)\widehat{\boldsymbol{\Omega}}\mathbf{g}_{(n^*,m)}\left(\hat{\boldsymbol{\theta}}_{n^*}\right) - n^{1/2}\mathbf{H}_m^T\left(\hat{\boldsymbol{\theta}}_{n^*}\right)\widehat{\boldsymbol{\Omega}}\mathbf{g}_{(n^*,m)}\left(\boldsymbol{\theta}_0\right) \\ &= n^{1/2}\mathbf{H}_m^T\left(\hat{\boldsymbol{\theta}}_{n^*}\right)\widehat{\boldsymbol{\Omega}}\mathbf{H}_m(\boldsymbol{\theta}_0)\cdot\left(\hat{\boldsymbol{\theta}}_{n^*} - \boldsymbol{\theta}_0\right) + n^{1/2}\mathbf{H}_m^T\left(\hat{\boldsymbol{\theta}}_{n^*}\right)\widehat{\boldsymbol{\Omega}} o_p\left(\left\|\hat{\boldsymbol{\theta}}_{n^*} - \boldsymbol{\theta}_0\right\|\right). \end{aligned}$$

By definition $\mathbf{g}_{(n^*,m)}(\hat{\boldsymbol{\theta}}_{n^*}) = \mathbf{0}$. By assumption, all the quantities $\hat{\boldsymbol{\theta}}_{n^*}$, $\widehat{\boldsymbol{\Omega}}$, $\mathbf{K}_m$, $\mathbf{J}_m$, and thus $\mathbf{H}_m$, are pointwise convergent. From Lemma 2.A.1 and 2.A.2, we have straightforwardly that

$$n^{1/2}\mathbf{g}_{(n^*,m)}(\boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{K}^{-1}\left[(\gamma^*\mathbf{Q} + \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^T\right]\mathbf{K}^{-T}\right),$$

when n is large enough.

It remains to demonstrate that $\mathbf{H}_m(\hat{\boldsymbol{\theta}}_{n^*})$ converges to some quantity $\mathbf{H} \equiv \mathbf{K}^{-1}\mathbf{J}$. By the continuity of $\mathbf{K}_m$ and $\mathbf{J}_m$, $\mathbf{H}_m$ is continuous and thus the continuous mapping theorem is satisfied (see [Vaa98]) so $\mathbf{H}_m(\hat{\boldsymbol{\theta}}_{n^*}) = \mathbf{H}_m(\boldsymbol{\theta}_0) + o_p(1)$. As already stated, $\mathbf{H}_m$ is pointwise convergent so $\mathbf{H}_m(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{H}(\boldsymbol{\theta}_0)$. Since $\mathbf{K}_m^{-1}(\hat{\boldsymbol{\pi}}_1^{(b)})\mathbf{J}_m(\hat{\boldsymbol{\pi}}_1^{(b)})$ is stochastically Lipschitz, we have by Lemma 3.5 that $\mathbf{H}_m(\boldsymbol{\theta})$ is also stochastically Lipschitz when n is sufficiently large. By the additional assumption that $\boldsymbol{\Theta}$ is compact, Lemma 3.6 is satisfied and as a conclusion $\mathbf{H}_m(\boldsymbol{\theta}) \xrightarrow{p} \mathbf{H}(\boldsymbol{\theta})$ uniformly for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $\mathbf{H}(\boldsymbol{\theta})$ is uniformly continuous. Following, this result, we obtain

$$\begin{aligned}\left\|\mathbf{H}_m(\hat{\boldsymbol{\theta}}_{n^*}) - \mathbf{H}\right\| &\leq \left\|\mathbf{H}_m(\hat{\boldsymbol{\theta}}_{n^*}) - \mathbf{H}(\hat{\boldsymbol{\theta}}_{n^*})\right\| + \left\|\mathbf{H}(\hat{\boldsymbol{\theta}}_{n^*}) - \mathbf{H}\right\| \\ &\leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\mathbf{H}_m(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})\right\| + \left\|\mathbf{H}(\hat{\boldsymbol{\theta}}_{n^*}) - \mathbf{H}\right\|.\end{aligned}$$

The first term of the last inequality converges to zero by the uniform convergence result, and the second term converge to zero by the continuous mapping theorem. Consequently, $\mathbf{H}_m(\hat{\boldsymbol{\theta}}_{n^*}) \xrightarrow{p} \mathbf{H}$ uniformly. The rest of the proof results from Slutksy's lemma. □

Proof of Theorem 2.14. Let the data generating mechanism be the following deviation model

$$C_\varepsilon = (1-\delta_\varepsilon)\, C_{\boldsymbol{\theta}} + \delta_\varepsilon \Delta_{\mathbf{z}},$$

where ε is the marginal probability of observing an outlier from the multivariate Dirac model $\Delta_{\mathbf{z}}$, common to each margin, and δ_ε is the proportion of having overall at least one outlier. Clearly $\delta_\varepsilon \downarrow 0$ as $\varepsilon \downarrow 0$.

Let (R, T, S) be the functional estimator corresponding to $(\hat{\boldsymbol{\nu}}_n, \hat{\boldsymbol{\theta}}_{n^*}, \hat{\boldsymbol{\pi}}_{n^*})$ and assume for convenience that they are Fisher consistent, e.g. $T(C_{\boldsymbol{\theta}_0}) = \boldsymbol{\theta}_0$. We have

$$\left[D_{\boldsymbol{\theta}}\boldsymbol{\pi} \circ T(C_{\boldsymbol{\theta}_0})\right]^T \boldsymbol{\Omega} \left[S(C_{\boldsymbol{\theta}_0}) - \boldsymbol{\pi} \circ T(C_{\boldsymbol{\theta}_0})\right] = \mathbf{0}.$$

Note that in the previous equation, we use T with two meanings: a functional and the matrix transpose. The functional is never used as an exponent so we keep this apparent ambiguity in the rest of the proof as we think the distinction is clear from the context. Replacing $C_{\boldsymbol{\theta}_0}$ by C_ε and taking the Hadamard derivative yields

$$\begin{aligned}&\left[\frac{\partial}{\partial\varepsilon} D_{\boldsymbol{\theta}}\boldsymbol{\pi} \circ T(C_\varepsilon)\right]^T \boldsymbol{\Omega} \left[\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta}_0)\right] \\ &\quad + \left[D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta}_0)\right]^T \boldsymbol{\Omega} \left[\frac{\partial}{\partial\varepsilon} S(C_\varepsilon) - D_{\boldsymbol{\theta}}\boldsymbol{\pi}\Big|_{\boldsymbol{\theta}_0} \circ \frac{\partial}{\partial\varepsilon} T(C_\varepsilon)\right] = \mathbf{0}, \qquad \text{as } \varepsilon \downarrow 0.\end{aligned}$$

The influence functions are $\mathfrak{I}(S) = \partial S(C_\varepsilon)/\partial\varepsilon$ and $\mathfrak{I}(T) = \partial T(C_\varepsilon)/\partial\varepsilon$ as $\varepsilon \downarrow 0$. By construction $\boldsymbol{\pi}_0 - \boldsymbol{\pi}(\boldsymbol{\theta}_0) = \mathbf{0}$ so the first term disappear. By Proposition 2.10 $D_{\boldsymbol{\theta}}\boldsymbol{\pi}(\boldsymbol{\theta}_0) = -\mathbf{K}^{-1}\mathbf{J}$. Rearranging yields

$$\mathfrak{I}(T) = \left(\mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathbf{K}^{-1}\mathbf{J}\right)^{-1} \mathbf{J}^T\mathbf{K}^{-T}\boldsymbol{\Omega}\mathfrak{I}(S).$$

It remains to demonstrate the explicit form of $\mathfrak{I}(S)$. By definition we have

$$\int \boldsymbol{\phi}\left(\mathbf{u}, R(C_{\boldsymbol{\theta}_0}), S(C_{\boldsymbol{\theta}_0})\right) \mathrm{d}C_{\boldsymbol{\theta}_0} = \mathbf{0}.$$

Replacing $C_{\boldsymbol{\theta}_0}$ by C_ε and taking the Hadamard derivative yields

$$\begin{aligned}&\int D_{\boldsymbol{\nu}}\boldsymbol{\phi}\left(\mathbf{u}, \boldsymbol{\nu}_0, \boldsymbol{\pi}_0\right) \mathrm{d}C_{\boldsymbol{\theta}_0} \cdot \left[\frac{\partial}{\partial\varepsilon} R(C_\varepsilon)\right] \\ &\quad + \int D_{\boldsymbol{\pi}}\boldsymbol{\phi}\left(\mathbf{u}, \boldsymbol{\nu}_0, \boldsymbol{\pi}_0\right) \mathrm{d}C_{\boldsymbol{\theta}_0} \cdot \left[\frac{\partial}{\partial\varepsilon} S(C_\varepsilon)\right] \\ &\quad\quad + \left[\frac{\partial}{\partial\varepsilon}\delta_\varepsilon\right] \boldsymbol{\phi}\left(\mathbf{v}, \boldsymbol{\pi}_0, \boldsymbol{\theta}_0\right) = \mathbf{0}, \quad \text{as } \varepsilon \downarrow 0.\end{aligned}$$

Let $\kappa = \partial\delta_\varepsilon/\partial\varepsilon$, $\varepsilon \downarrow 0$. Rearranging the terms yields the result.

For the specific values that κ takes, let $\{B_j : j = 1, \dots, d\}$ denotes a sequence of Bernoulli random variables that takes values one with probability ε and 0 with probability $1 - \varepsilon$. By definition $\delta_\varepsilon = 1 - \Pr(B_1 = 0, \dots, B_d = 0)$. If the B are independent, we have $\delta_\varepsilon = 1 - \prod_{j=1}^{d}(1 - \varepsilon) = 1 - (1 - \varepsilon)^d$, and thus $\kappa = -d$. If the B are comonotonic, we have $\delta_\varepsilon = 1 - (1 - \varepsilon) = \varepsilon$, and thus $\kappa = 1$. □

Proof of Corollary 2.16. Let $\mathbf{u} \equiv F_Y(\mathbf{z}, \boldsymbol{\nu}_0)$. Note that $\mathbf{u}$ takes values in the unit simplex $[0, 1]^d$. Since $[0, 1]^d$ is compact, the mapping $\mathbf{u} \mapsto \boldsymbol{\phi}$ is bounded (see Theorem 4.16 in [Rud76]). Thus, the only source of unboundess for the influence function of $\hat{\boldsymbol{\theta}}_{n^*}$ is $\mathfrak{I}(\hat{\boldsymbol{\nu}}_n, \mathbf{z})$, *i.e.* the influence function of the marginal estimators. □

2.C Additional results

[OM92] derived the asymptotic variance of the MLE of a Gumbel-Hougaard copula with survival Weibull margins, *i.e.* $\mathbb{P}(X_j > x_j) = \exp(-(\eta_j x_j)^{\kappa_j})$, in both cases where margins are known and unknown. In order to derive the asymptotic covariance matrix of the IFM, we need in addition to derive the covariances between the log-likelihood score functions of the two marginal distribution (see [Joe05] for more details). Let $\mathcal{I}_{\eta_1\eta_2}$, $\mathcal{I}_{\eta_j\kappa_k}$ $(j, k = 1, 2)$ and $\mathcal{I}_{\kappa_1\kappa}$ denote these covariances. Using the approach proposed by [OM92], we were able to find the following quantities:

$$\mathcal{I}_{\eta_1\eta_2} = \kappa_1\kappa_2(\eta_1\eta_2)^{-1}\Big\{2(2\alpha+1)\mathcal{B}(\alpha+1, \alpha+1) - 1\Big\},$$

$$\begin{aligned}\mathcal{I}_{\eta_j\kappa_k} = \frac{\kappa_j}{\eta_j\kappa_k 2\Gamma(\alpha+\frac{3}{2})}\Big\{ &-\alpha\sqrt{\pi}\Gamma(\alpha)(\alpha^2\log(2) + \alpha\gamma - 1)2^{-2\alpha+2} \\ &+ 2(\alpha + {}^1\!/\!{}_2)(\alpha\gamma + \alpha\Psi(\alpha) + \gamma - 1)\Gamma(\alpha + {}^1\!/\!{}_2) - 2\alpha\Gamma(\alpha)4^{-\alpha}\sqrt{\pi} \\ &\times\Big[(\alpha^2 + {}^\alpha\!/\!{}_2)\Psi(\alpha + {}^1\!/\!{}_2) + (-\alpha^2 - {}^\alpha\!/\!{}_2)\Psi(\alpha) + \alpha\log(2) - 4\alpha + \gamma\Big]\Big\} \qquad (1 \le j, k \le 2),\end{aligned}$$

$$
\begin{aligned}
\mathcal{J}_{\kappa_1\kappa_2} = (\kappa_1\kappa_2)^{-1}\Bigg\{ & \Big[3(\alpha + 1/2)(\alpha+1)^2 2^{2\alpha}(\pi^2 + 9\gamma^2 - 12\gamma - 6)\Gamma(\alpha + 3/2)\Big]^{-1} \\
& \Big[- 3(\alpha + 1/2)\Big(\alpha(\alpha+1) \times (\alpha(-2+\gamma) - 1 + \gamma)\Psi(\alpha) + (-1+\gamma)\alpha^4 \\
& + (\pi^2/6 + 2\gamma^2 - 3\gamma)\alpha^3 + (\pi^2/2 + 5\gamma^2 - 13\gamma + 3)\alpha^2 \\
& + (\pi^2/2 + 4\gamma^2 - 12\gamma + 4)\alpha + \pi^2/6 + \gamma^2 - 3\gamma\Big) 2^{2\alpha}(\pi^2 + 9\gamma^2 - 12\gamma - 6)\Gamma(\alpha + 3/2) \\
& + \Big(3(\alpha + 1/2)^2(\alpha(-2+\gamma) + \pi^2/6 + \gamma^2 - 4\gamma + 3)2^{2\alpha}(\pi^2 + 9\gamma^2 - 12\gamma - 6)\Gamma(\alpha + 1/2) \\
& + \Gamma(\alpha)(3\alpha(\alpha(-2+\gamma) + (1/2)\gamma - 3/4)((9\gamma^2 - 12\gamma - 6)\sqrt{\pi} + \pi^{5/2})(\alpha + 1/2)\Psi(\alpha + 1/2) \\
& + (54(-\alpha/2(\alpha(-2+\gamma) + \gamma/2 - 3/4)(\alpha + 1/2)\Psi(\alpha) + \alpha(\alpha(-2+\gamma) + \gamma/2 - 3/4)(\alpha + 1/2)\log(2) \\
& + (\gamma^2 - 4\gamma + 5/2)\alpha^2 + (\gamma^2 - 15\gamma/4 + 9/4)\alpha + \gamma^2/4 - 7\gamma/8 + 7/16))(\gamma^2 - 4\gamma/3 - 2/3)\sqrt{\pi} \\
& + (-3\alpha(\alpha(-2+\gamma) + \gamma/2 - 3/4)(\alpha + 1/2)\Psi(\alpha) + 6\alpha(\alpha(-2+\gamma) + \gamma/2 - 3/4)(\alpha + 1/2)\log(2) \\
& + (\pi^2 + 15\gamma^2 - 36\gamma + 9)\alpha^2 + (15\gamma^2 - 69\gamma/2 + 15/2 + \pi^2)\alpha + 15\gamma^2/4 - 33\gamma/4 \\
& + \pi^2/4 + 9/8)\pi^{5/2}))\alpha(\alpha+1)^2)/3(-(3(\alpha + 1/2))(\alpha(\alpha+1)(\alpha(-2+\gamma) - 1 + \gamma)\Psi(\alpha) \\
& + (-1+\gamma)\alpha^4 + (\pi^2/6 + 2\gamma^2 - 3\gamma)\alpha^3 + (\pi^2/2 + 5\gamma^2 - 13\gamma + 3)\alpha^2 + (\pi^2/2 + 4\gamma^2 - 12\gamma + 4)\alpha \\
& + \pi^2/6 + \gamma^2 - 3\gamma)2^{2\alpha}(\pi^2 + 9\gamma^2 - 12\gamma - 6)\Gamma(\alpha + 3/2) + (3(\alpha + 1/2)^2(\alpha(-2+\gamma) + \pi^2/6 \\
& + \gamma^2 - 4\gamma + 3)2^{2\alpha}(\pi^2 + 9\gamma^2 - 12\gamma - 6)\gamma(\alpha + 1/2) + \Gamma(\alpha)(3\alpha(\alpha(-2+\gamma) + \gamma/2 - 3/4) \\
& \times ((9\gamma^2 - 12\gamma - 6)\sqrt{\pi} + \pi^{5/2})(\alpha + 1/2)\Psi(\alpha + 1/2) + (54(-(1/2)\alpha(\alpha(-2+\gamma) + \gamma/2 - 3/4) \\
& \times (\alpha + 1/2)\Psi(\alpha) + \alpha(\alpha(-2+\gamma) + \gamma/2 - 3/4)(\alpha + 1/2)\log(2) + (\gamma^2 - 4\gamma + 5/2)\alpha^2 \\
& + (\gamma^2 - (15/4)\gamma + 9/4)\alpha + (1/4)\gamma^2 - (7/8)\gamma + 7/16))(\gamma^2 - (4/3)\gamma - 2/3)\sqrt{\pi} + (-3\alpha(\alpha(-2+\gamma \\
& + (1/2)\gamma - 3/4)(\alpha + 1/2)\Psi(\alpha) + 6\alpha(\alpha(-2+\gamma) + (1/2)\gamma - 3/4)(\alpha + 1/2)\log(2) + \\
& (\pi^2 + 15\gamma^2 - 36\gamma + 9)\alpha^2 + (15\gamma^2 - (69/2)\gamma + 15/2 + \pi^2)\alpha + (15/4)\gamma^2 - (33/4)\gamma \\
& + (1/4)\pi^2 + 9/8)\pi^{5/2})\Big)\alpha(\alpha+1)^2\Big] \\
& + (1/6)\alpha(-1+\alpha)(6 - 6\gamma + \alpha(\pi^2 - 12)) + (\alpha^2 + 2\alpha + 1)^{-1}\Big[- \alpha^5(2K_1 + 2K_3 + K_4) \\
& + (-2K_1\gamma + 3K_2\gamma + 3K_1 - 5K_2 + 2K_3 - K_4 - \gamma + 3)\alpha^2 + (2K_1\gamma + K_2\gamma - 5K_1 - 2K_2 + 2K_3 - \\
& + \alpha^4(2K_1\gamma - 7K_1 - 2K_3 - 3K_4) + \alpha\gamma K_2 - \alpha\gamma - \alpha K_2 + \gamma K_2 + 2\alpha - K_2 - \alpha(2\gamma - 3)(K_1 - K_2) \\
& + (1/6(-12 + 6\gamma - 6K_1 + \alpha(-\pi^2 + 6K_1 + 12) + 36\alpha K_3 + (6(11 - 6\gamma))K_1))\alpha^2 \Bigg\}
\end{aligned}
$$

where $\mathcal{B}(\cdot,\cdot)$ is the beta function, $\Psi(\cdot)$ is the digamma function, $\gamma = -\Psi(1)$ is Euler's constant. The functions $K_1 = \int_0^1 u^\alpha \bar{u}^\alpha \log(\bar{u})\mathrm{d}u$, $K_2 = \int_0^1 u^\alpha \log(\bar{u})\mathrm{d}u$, $K_3 = \int_0^1 u^\alpha \bar{u}^\alpha \log(\bar{u})\log(u)\mathrm{d}u$, $K_4 = \int_0^1 \bar{u}^\alpha \log(\bar{u})\log(u)\mathrm{d}u$ and $K_5 = \int_0^1 u^\alpha \log(\bar{u})\log(u)\mathrm{d}u$ have no analytical expression but may be easily computed by numerical integration. We intentionally ignored the dependence with α in their notation to avoid confusion.

3

Generic results

This chapter assembles some generic theoretical results useful for the other chapters.

We generically denote $\{\mathbf{g}_n : n \geq 1\}$ a sequence of a random vector-valued function and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ a vector of parameters.

The next Lemma is Theorem 5.9 in [Vaa98]. The proof is given for the sake of completeness.

Lemma 3.1 (weak consistency). *Let $\{\mathbf{g}_n(\boldsymbol{\theta})\}$ be sequence of a random vector-valued function of vector parameter $\boldsymbol{\theta}$ with a deterministic limit $\mathbf{g}(\boldsymbol{\theta})$. If $\boldsymbol{\Theta}$ is compact, if the random function sequence converges uniformly as $n \to \infty$*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta})\| \xrightarrow{p} 0, \tag{3.1}$$

and if there exist $\delta > 0$ such that

$$\inf_{\boldsymbol{\theta}\notin\mathscr{B}(\boldsymbol{\theta}_0,\delta)} \|\mathbf{g}(\boldsymbol{\theta})\| > 0 = \|\mathbf{g}(\boldsymbol{\theta}_0)\|\,, \tag{3.2}$$

then any sequence of estimators $\{\hat{\boldsymbol{\theta}}_n\}$ converges weakly in probability to $\boldsymbol{\theta}_0$.

Proof. Choose $\hat{\boldsymbol{\theta}}_n$ that nearly minimises $\|\mathbf{g}_n(\boldsymbol{\theta})\|$ so that

$$\left\|\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)\right\| \leq \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{g}_n(\boldsymbol{\theta})\| + o_p(1)$$

Clearly we have $\inf_{\boldsymbol{\theta}}\|\mathbf{g}_n(\boldsymbol{\theta})\| \leq\|\mathbf{g}_n(\boldsymbol{\theta}_0)\|$, and by (3.1) $\|\mathbf{g}_n(\boldsymbol{\theta}_0)\| \xrightarrow{p} \|\mathbf{g}(\boldsymbol{\theta}_0)\|$ so that

$$\left\|\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)\right\| \leq \|\mathbf{g}(\boldsymbol{\theta}_0)\| + o_p(1)$$

Now, substracting both sides by $\|\mathbf{g}(\hat{\boldsymbol{\theta}}_n)\|$, we have by the reverse triangle inequality

$$-\left\|\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) - \mathbf{g}(\hat{\boldsymbol{\theta}}_n)\right\| \leq \|\mathbf{g}(\boldsymbol{\theta}_0)\| - \left\|\mathbf{g}(\hat{\boldsymbol{\theta}}_n)\right\| + o_p(1)$$

The left-hand side is bounded by the negative supremum, thus

$$\|\mathbf{g}(\boldsymbol{\theta}_0)\| - \left\|\mathbf{g}(\hat{\boldsymbol{\theta}}_n)\right\| \geq -\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta})\| - o_p(1)$$

It follows from (3.1) that the limit in probability of the right-hand side tends to 0. Let $\varepsilon > 0$ and choose a $\delta > 0$ as in (3.2) so that

$$\|\mathbf{g}(\boldsymbol{\theta})\| > \|\mathbf{g}(\boldsymbol{\theta}_0)\| - \varepsilon$$

for every $\boldsymbol{\theta} \notin \mathscr{B}(\boldsymbol{\theta}_0, \delta)$. If $\hat{\boldsymbol{\theta}}_n \notin \mathscr{B}(\boldsymbol{\theta}_0, \delta)$, we have

$$\|\mathbf{g}(\boldsymbol{\theta}_0)\| - \left\|\mathbf{g}(\hat{\boldsymbol{\theta}}_n)\right\| < \varepsilon$$

The probability of this event converges to 0 as $n \to \infty$. □

The next definition is taken from [And92] (see also [Pol84, Chapter 7.1])

Definition 3.2. *$\{\mathbf{g}_n(\boldsymbol{\theta})\}$ is stochastically uniformly equicontinuous on $\boldsymbol{\Theta}$ if for every $\varepsilon > 0$ there exist a real $\delta > 0$ such that*

$$\limsup_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}'\in\mathscr{B}(\boldsymbol{\theta},\delta)} \|\mathbf{g}_n(\boldsymbol{\theta}') - \mathbf{g}_n(\boldsymbol{\theta})\| > \varepsilon\right) < \varepsilon \tag{3.3}$$

Lemma 3.3 (uniform consistency)**.** *If $\boldsymbol{\Theta}$ is compact, if the sequence of random vector-valued function $\{\mathbf{g}_n(\boldsymbol{\theta})\}$ is pointwise convergent for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and is stochastically uniformly equicontinuous on $\boldsymbol{\Theta}$, then*

i. $\{\mathbf{g}_n(\boldsymbol{\theta})\}$ converges uniformly,

ii. $\mathbf{g}$ is uniformly continuous.

Proof. (i) (Inspired from [Rud76, Theorem 7.25(b)]). Let $\varepsilon > 0$, choose $\delta > 0$ so to satisfy stochastic uniform equicontinuity in (3.3). Let $\mathscr{B}(\boldsymbol{\theta}, \delta) = \{\boldsymbol{\theta}' \in \boldsymbol{\Theta} : d(\boldsymbol{\theta}, \boldsymbol{\theta}') < \delta\}$. Since $\boldsymbol{\Theta}$ is compact, there are finitely many points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ in $\boldsymbol{\Theta}$ such that

$$\boldsymbol{\Theta} \subset \mathscr{B}(\boldsymbol{\theta}_1, \delta) \cup \cdots \cup \mathscr{B}(\boldsymbol{\theta}_k, \delta)$$

Since $\{\mathbf{g}_n(\boldsymbol{\theta})\}$ converges pointwise for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have

$$\limsup_{n\to\infty} \Pr\left(\|\mathbf{g}_n(\boldsymbol{\theta}_l) - \mathbf{g}(\boldsymbol{\theta}_l)\| > \varepsilon\right) < \varepsilon,$$

whenever $1 \leq l \leq k$. If $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, so $\boldsymbol{\theta} \in \mathscr{B}(\boldsymbol{\theta}_l, \delta)$ for some l, so that

$$\limsup_{n\to\infty} \Pr\left(\|\mathbf{g}_n(\boldsymbol{\theta}_l) - \mathbf{g}_n(\boldsymbol{\theta})\| > \varepsilon\right) \leq \limsup_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}'\in\mathscr{B}(\boldsymbol{\theta},\delta)} \|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}_n(\boldsymbol{\theta}')\|\right) < \varepsilon$$

Then, by the triangle inequality we have

$$\begin{aligned}
&\limsup_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta})\| > \varepsilon\right) \\
&\quad\leq \limsup_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}'\in\mathscr{B}(\boldsymbol{\theta},\delta)} \|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}_n(\boldsymbol{\theta}')\| > \varepsilon\right) \\
&\qquad+ \limsup_{n\to\infty} \Pr\left(\|\mathbf{g}_n(\boldsymbol{\theta}') - \mathbf{g}(\boldsymbol{\theta}')\| > \varepsilon\right) + \Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}'\in\mathscr{B}(\boldsymbol{\theta},\delta)} \|\mathbf{g}(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta}')\| > \varepsilon\right) < 3\varepsilon
\end{aligned}$$

(ii). The proof follows the same steps. □

The next Lemma is similar to [And92, Lemma 1]. The result of [And92] is on the difference between a random and a nonrandom functions and requires the extra assumption of absolute continuity of the nonrandom function. The proof provided here is also different.

Lemma 3.4. *If for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}_n(\boldsymbol{\theta}')\| \leq B_n d(\boldsymbol{\theta}, \boldsymbol{\theta}')$ with $B_n = \mathcal{O}_p(1)$, then $\{\mathbf{g}_n(\boldsymbol{\theta})\}$ is stochastically uniformly equicontinuous.*

Proof. By $B_n = \mathcal{O}_p(1)$, there is $M > 0$ such that for all n, $\Pr(|B_n| > M) < \varepsilon$. Let $\varepsilon > 0$ and choose a sufficiently small $\delta > 0$ such that for all $\boldsymbol{\theta}', \boldsymbol{\theta} \in \boldsymbol{\Theta}$, $d(\boldsymbol{\theta}, \boldsymbol{\theta}') < \varepsilon/M = \tau$, $\delta \leq \tau$. Let $\mathcal{B}(\boldsymbol{\theta}, \delta) = \{\boldsymbol{\theta}' \in \boldsymbol{\Theta} : d(\boldsymbol{\theta}, \boldsymbol{\theta}') < \delta\}$. Then, we have

$$
\begin{aligned}
&\limsup_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}'\in\mathcal{B}(\boldsymbol{\theta},\delta)} \|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}_n(\boldsymbol{\theta}')\| > \varepsilon \right) \\
&\quad \leq \limsup_{n\to\infty} \Pr\left(B_n \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}'\in\mathcal{B}(\boldsymbol{\theta},\delta)} d(\boldsymbol{\theta}, \boldsymbol{\theta}') > \varepsilon \right) \\
&\qquad \leq \limsup_{n\to\infty} \Pr\left(B_n \tau > \varepsilon \right) \leq \limsup_{n\to\infty} \Pr\left(|B_n| > M \right) < \varepsilon
\end{aligned}
$$

□

The next Lemma is a special case of [New91, Corollary 3.1].

Lemma 3.5. *Let $\{\mathbf{x}_i : i \geq 1\}$ be an i.i.d. sequence of random variable and let $\mathbf{g}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta})$. If for all $i = 1, \ldots, n$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\|\mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}')\| \leq b_n(\mathbf{x}_i) d(\boldsymbol{\theta}, \boldsymbol{\theta}')$ with $\mathbb{E}[b_n(\mathbf{x}_i)] = \mu_n = \mathcal{O}(1)$, then $\{\mathbf{g}_n(\boldsymbol{\theta})\}$ is stochastically uniformly equicontinuous.*

Proof. Let $B_n = n^{-1}\sum_{i=1}^{n} b_n(\mathbf{x}_i)$, so $\mathbb{E}[B_n] = \mathcal{O}(1)$. We have by triangle inequality

$$
\|\mathbf{g}_n(\boldsymbol{\theta}) - \mathbf{g}_n(\boldsymbol{\theta}')\| \leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}')\| \leq B_n d(\boldsymbol{\theta}, \boldsymbol{\theta}')
$$

The rest of the proof follows from Lemma 3.4. □

Lemma 3.6 (uniform weak law of large number)**.** *If, in addition to Lemma 3.5, for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbf{g}_n(\boldsymbol{\theta})$ is pointwise convergent, then $\{\mathbf{g}_n(\boldsymbol{\theta})\}$ converges uniformly.*

Proof. The proof is an immediat consequence of Lemma 3.5 and Lemma 3.3. □

The next Lemma is essentially a combination of Theorem 4.2 and Corollary 4.3 in [Lan93]. The proof is given for the sake of completeness.

Lemma 3.7 (mean value inequality)**.** *Let U be a convex open set in $\boldsymbol{\Theta}$. Let $\boldsymbol{\theta}_1 \in U$ and $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$. If $\mathbf{g} : U \to F$ is a $\mathcal{C}^1$-mapping, then*

i. $\mathbf{g}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) - \mathbf{g}(\boldsymbol{\theta}_1) = \int_0^1 D\mathbf{g}(\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2)dt \cdot \boldsymbol{\theta}_2$

ii. $\|\mathbf{g}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) - \mathbf{g}(\boldsymbol{\theta}_1)\| \leq \sup_{0\leq t\leq 1}\|D\mathbf{g}(\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2)\| \cdot \|\boldsymbol{\theta}_2\|$

Proof. (i). Fix $\boldsymbol{\theta}_1 \in U$, $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$. Let $\boldsymbol{\theta}_3 = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2$ and $\lambda_t = (1-t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_3$. For $t \in [0,1]$ we have by the convexity of U that $\lambda_t \in U$, and so $\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2$ is in U as well. Put $\mathbf{h}(t) = \mathbf{g}(\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2)$, so $D\mathbf{h}(t) = D\mathbf{g}(\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2) \cdot \boldsymbol{\theta}_2$. By the fundamental theorem of calcul we have that

$$
\int_0^1 D\mathbf{h}(t)\,\mathrm{d}t = \mathbf{h}(1) - \mathbf{h}(0)
$$

Since $\mathbf{h}(1) = \mathbf{g}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)$, $\mathbf{h}(0) = \mathbf{g}(\boldsymbol{\theta}_1)$, and $\boldsymbol{\theta}_2$ is allowed to be pulled out of the integral, part *(i)* is proven.
(ii). We have that

$$\begin{aligned}\|\mathbf{g}(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2)-\mathbf{g}(\boldsymbol{\theta}_1)\| &\le \left\|\int_0^1 D\mathbf{g}(\boldsymbol{\theta}_1+t\boldsymbol{\theta}_2)\,\mathrm{d}t\right\|\cdot\|\boldsymbol{\theta}_2\|\,,\\ &\le |(1-0)|\sup_{0\le t\le 1}\|D\mathbf{g}(\boldsymbol{\theta}_1+t\boldsymbol{\theta}_2)\|\cdot\|\boldsymbol{\theta}_2\|\,,\end{aligned}$$

where we use the Cauchy-Schwarz inequality for the first inequality, and the upper bound of integral for the second. The supremum of the norm exists because the affine line $\boldsymbol{\theta}_1+t\boldsymbol{\theta}_2$ is compact and the Jacobian is continuous. □

Lemma 3.8 (delta method). *If conditions of Lemma 3.7 holds, then*

$$\mathbf{g}(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2)-\mathbf{g}(\boldsymbol{\theta}_1) = D\mathbf{g}(\boldsymbol{\theta}_1)\cdot\boldsymbol{\theta}_2 + o\left(\|\boldsymbol{\theta}_2\|\right)$$

Proof. Fix $\boldsymbol{\theta}_1 \in U$ and $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$. By Lemma 3.7, we have

$$\left\|\int_0^1 D\mathbf{g}(\boldsymbol{\theta}_1+t\boldsymbol{\theta}_2)\,\mathrm{d}t\right\| \le \sup_{0\le t\le 1}\|D\mathbf{g}(\boldsymbol{\theta}_1+t\boldsymbol{\theta}_2)\|$$

Let $\boldsymbol{\theta}_3 = \boldsymbol{\theta}_1+\boldsymbol{\theta}_2$ so $\lambda_t = (1-t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_3$, $t\in[0,1]$, is in U and $\boldsymbol{\theta}_1+t\boldsymbol{\theta}_2$ as well. Let $\mathscr{B}^c(\boldsymbol{\theta}_1,\|\boldsymbol{\theta}_2\|) = \{\|\boldsymbol{\theta}_1-\boldsymbol{\theta}\|\le\|\boldsymbol{\theta}_2\|\}$. We have

$$\begin{aligned}\|t\boldsymbol{\theta}_1+(1-t)\boldsymbol{\theta}_3-\boldsymbol{\theta}\| &\le t\,\|\boldsymbol{\theta}_1-\boldsymbol{\theta}\| + (1-t)\,\|\boldsymbol{\theta}_3-\boldsymbol{\theta}\|\\ &\le t\|\boldsymbol{\theta}_2\| + (1-t)\|\boldsymbol{\theta}_2\| = \|\boldsymbol{\theta}_2\|,\end{aligned}$$

so the line segment λ_t is in the closed ball. Hence, we have

$$\left\|\int_0^1 D\mathbf{g}(\boldsymbol{\theta}_1+t\boldsymbol{\theta}_2)\,\mathrm{d}t\right\| \le \sup_{\boldsymbol{\theta}\in\mathscr{B}(\boldsymbol{\theta}_1,\|\boldsymbol{\theta}_2\|)}\|D\mathbf{g}(\boldsymbol{\theta})\|$$

Eventually, we have by continuity of the Jacobian in a neighborhood of $\boldsymbol{\theta}_1$ that

$$\sup_{\boldsymbol{\theta}\in\mathscr{B}(\boldsymbol{\theta}_1,\|\boldsymbol{\theta}_2\|)}\|D\mathbf{g}(\boldsymbol{\theta})-D\mathbf{g}(\boldsymbol{\theta}_1)\|\to 0$$

as $\|\boldsymbol{\theta}_2\|\to 0$. □

Lemma 3.9 (asymptotic normality). *Let U be a convex open set in $\boldsymbol{\Theta}$. Let $\{\hat{\boldsymbol{\theta}}_n\}$ be a sequence of estimator (roots of) the mapping $\mathbf{g}_n : U \to F$. If*

i. *$\hat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0 \in U$,*

ii. *$\{\mathbf{g}_n\}$ is a $\mathscr{C}^1$-mapping,*

iii. *$n^{1/2}\mathbf{g}_n(\boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}(\mathbf{0},\mathbf{V})$,*

iv. *$D\mathbf{g}_n(\boldsymbol{\theta}_0)$ converges in probability to $\mathbf{M}$,*

v. *$D\mathbf{g}_n(\boldsymbol{\theta}_0)$ is nonsingular,* *then*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-T}$.

Proof. Fix $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0$, from Lemma 3.7 and Lemma 3.8 we have

$$\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{g}_n(\boldsymbol{\theta}_0) + D\mathbf{g}_n(\boldsymbol{\theta}_0) \cdot (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|)$$

By definition $\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Multiplying by square-root n leads to

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\left[D\mathbf{g}_n(\boldsymbol{\theta}_0)\right]^{-1} n^{1/2}\mathbf{g}_n(\boldsymbol{\theta}_0) - n^{1/2}\left[D\mathbf{g}_n(\boldsymbol{\theta}_0)\right]^{-1} o_p\left(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|\right)$$

By the continuity of the matrix inversion $[D\mathbf{g}_n(\boldsymbol{\theta}_0)]^{-1} \xrightarrow{p} \mathbf{M}^{-1}$. Since the central limit theorem holds for $n^{1/2}\mathbf{g}_n(\boldsymbol{\theta}_0)$, the proof results from Slutsky's lemma. □

The next Lemma is Theorem 9.4 in [LS68] and is given without proof.

Lemma 3.10 (implicit function theorem)**.** *Let* $\boldsymbol{\Xi} \times \boldsymbol{\Theta}$ *be an open subset of* $\mathbb{R}^m \times \mathbb{R}^p$*. Let* $\mathbf{g} : \boldsymbol{\Xi} \times \boldsymbol{\Theta} \to \mathbb{R}^p$ *be a function of the form* $\mathbf{g}(\boldsymbol{\xi}, \boldsymbol{\theta}) = k$*. Let the solution at the points* $(\boldsymbol{\xi}_0, \boldsymbol{\theta}_0) \in \boldsymbol{\Xi} \times \boldsymbol{\Theta}$ *and* $k_0 \in \mathbb{R}^p$ *be*

$$\mathbf{g}(\boldsymbol{\xi}_0, \boldsymbol{\theta}_0) = k_0$$

If

i. $\mathbf{g}$ *is differentiable in* $\boldsymbol{\Xi} \times \boldsymbol{\Theta}$*,*

ii. The partial derivative $D_{\boldsymbol{\xi}}\mathbf{g}$ *is continuous in* $\boldsymbol{\Xi} \times \boldsymbol{\Theta}$*,*

iii. The partial derivative $D_{\boldsymbol{\theta}}\mathbf{g}$ *is invertible at the points* $(\boldsymbol{\xi}_0, \boldsymbol{\theta}_0) \in \boldsymbol{\Xi} \times \boldsymbol{\Theta}$*,*

then, there are neighborhoods $X \subset \boldsymbol{\Xi}$ *and* $O \subset \boldsymbol{\Theta}$ *of* $\boldsymbol{\xi}_0$ *and* $\boldsymbol{\theta}_0$ *on which the function* $\hat{\boldsymbol{\theta}} : O \to X$ *is uniquely defined, and such that:*

1. $\mathbf{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}(\boldsymbol{\xi})) = k_0$ *for all* $\boldsymbol{\xi} \in X$*,*

2. *For each* $\boldsymbol{\xi} \in X$*,* $\hat{\boldsymbol{\theta}}(\boldsymbol{\xi})$ *is the unique solution lying in* O *such that* $\hat{\boldsymbol{\theta}}(\boldsymbol{\xi}_0) = \boldsymbol{\theta}_0$*,*

3. $\hat{\boldsymbol{\theta}}$ *is differentiable on* X *and*

$$D_{\boldsymbol{\xi}}\hat{\boldsymbol{\theta}} = -\left[D_{\boldsymbol{\theta}}\mathbf{g}\right]^{-1} D_{\boldsymbol{\xi}}\mathbf{g}$$

References for Chapter 3

[And92] Donald WK Andrews. “Generic uniform convergence”. In: *Econometric theory* 8.02 (1992), pp. 241–257.

[Lan93] Serge Lang. *Real and functional analysis.* 3rd. Springer-Verlag New York, Inc., 1993.

[LS68] Lynn H Loomis and Shlomo Sternberg. *Advanced Calculus.* Reading, Massachussets: Addison-Wesley, 1968.

[New91] Whitney K Newey. “Uniform convergence in probability and stochastic equicontinuity”. In: *Econometrica: Journal of the Econometric Society* (1991), pp. 1161–1167.

[Pol84] David Pollard. *Convergence of stochastic processes.* Springer series in statistics, 1984.

[Rud76] Walter Rudin. *Principles of mathematical analysis.* 3rd. McGraw-Hill, Inc., 1976.

[Vaa98] Aad W Van der Vaart. *Asymptotic statistics.* Vol. 3. Cambridge university press, 1998.