UNIVERSITA' DELLA SVIZZERA ITALIANA – LUGANO

PhD in Economics

Contributions to Robustness Theory

Supervisors:
Prof. Elvezio Ronchetti and Prof. Fabio Trojani

Tesi di:
Davide A. La Vecchia

UNIVERSITA' DELLA SVIZZERA ITALIANA
FACOLTA' DI ECONOMIA

The thesis "**Contributions to Robustness Theory**" by **Davide A. La Vecchia** is recommended for acceptance by the members of the delegated committee, as stated by the enclosed reports, in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dated: March 2011

Research Supervisor:	**Elvezio Ronchetti and Fabio Trojani**
External Examiner:	**Patrick Gagliardini**
	March Hallin

UNIVERSITA' DELLA SVIZZERA ITALIANA

Author: **Davide A. La Vecchia**

Title: **Contributions to Robustness Theory**

Department: **Facolta' di Economia**

History doesn't repeat itself, but it does rhyme.

Mark Twain

Table of Contents

Abstract

The goal of this PhD Thesis is the definition of new robust estimators, thereby extending the available theory and exploring new directions for applications in finance. The Thesis contains three papers, which analyze three different types of estimators: M-, Minimum Distance- and R-estimators. The focus is manly of their infinitesimal robustness, but global robustness properties are also considered.

The first paper (*"Higher-order infinitesimal robustness"*) studies M-estimators and it is a joint work with Elvezio Ronchetti and Fabio Trojani. Using the higher-order von Mises expansion, we go beyond the Influence Function and we extend Hampel's paradigm of robustness, introducing higher-order infinitesimally robust M-estimators. We show that a bounded estimating function having also bounded gradient with respect to the parameter ensures, at the same time, the stability of the: (i) second-order approximated bias (B-robustness); (ii) asymptotic variance (V-robustness) and (iii) saddlepoint density approximation. An application in finance (risk management) concludes the paper.

The second paper (*"On robust estimation via pseudo-additive information measures"*) is jointly written with Davide Ferrari and it studies a new class of Minimum Divergence (in the following, MD) estimators. The theoretical contribution of the paper is to show that robustness is dual to information theory. Information theory plays a crucial role in statistical inference: Maximum Likelihood estimators are related to it through the minimization of Shannon entropy (namely, minimization of the Kullback-Leibler divergence). The fundamental axiom characterizing Shannon entropy is additivity. Relaxing this assumption, we obtain a generalized entropy (called q-entropy) which exploits the link between information theory and infinitesimal robustness. Minimizing the q-entropy, we define a new class of MD robust re-descending estimators, featuring B-, V-robustness and that have also good global robustness properties in terms of high-breakdown.

The third paper (*"Semi-parametric rank-based tests and estimators for Markov processes"*) contains the preliminary results of a working paper that I have started in Princeton, working with Marc Hallin. The paper deals with R-estimators and rank-based tests. Precisely, combining the flexibility of the semi-parametric approach with the distribution-freeness of rank statistics, we define R-estimators and tests for stationary Markov processes. An application for inference and testing in stochastic volatility (SV) models concludes the paper.

Acknowledgement

I wish to thank my supervisors Elvezio Ronchetti and Fabio Trojani for their guidance and for the long time spent with me, discussing ideas, proofs, and problems related to my research. Their precious comments and their criticism have been very helpful for the success of this Thesis. Moreover, I'd like to thank them for the unfailing patience that they have had in reading this document.

I also thank Patrick Gagliardini and Marc Hallin for their comments on my PhD Thesis. Additionally, I warmly thank Marc for the interesting discussions about the ranks and for the moral support that he gave to me, in Princeton. Thanks to our discussions about *"Il gattopardo"*, by Giuseppe Tomasi di Lampedusa, and about the Italian political environment, I was able to overcome some cold and rainy afternoons in Princeton, avoiding to think of my family, so far away, in Lugano.

Finally, I thank my parents, and my PhD colleagues in Lugano (Diego, Elisa, Fulvio, Ilaria, Lorenzo, Nicola, Hackim, and Thomas) for their unconditional encouragement.

Introduction

In statistics, econometrics and other fields, many papers have studied the robustness properties of estimators and tests under different forms of deviations from ideal model assumptions. Nowadays, the need for a robust statistical approach which limits the extreme sensitivity of classical procedures (e.g., Maximum Likelihood, Ordinary Least Squares, Pseudo Maximum Likelihood, Method of Moments) to deviations from the theoretical model, is well recognized. Robust statistics can be viewed as an extension of parametric statistics, taking into account that parametric models are only approximations to reality. Corresponding theories for robust estimation and testing procedures have been developed and, historically, there exist three main lines of research formalizing the robustness problem.

The first line of research was Huber's (1964) minimax theory, where the statistical problem is viewed as a game between the Nature (which chooses a distribution in the neighborhood of the model) and the statistician (who chooses a statistical procedure in a given class). The statistician achieves robustness by constructing a minimax procedure which minimizes a loss criterion (e.g. the asymptotic variance) at the least favorable distribution in the neighborhood.

The second line of research relies on the concept of breakdown points (see Hampel (1968), Donoho and Huber (1983)) and studies global robustness features, i.e., the potential impact of large contaminations by outliers on a statistical procedure.

The third line of research, opened by Hampel (1968), considers local robustness, or infinitesimal robustness, i.e., the impact of moderate distributional deviations from ideal models on a statistical procedure.

Here the quantities of interest (e.g., the bias or the variance of an estimator) are viewed as functionals of the underlying distribution and their linear approximation is used to study their behavior in a neighborhood of the ideal model. A key tool is the first-order derivative of such a functional, the influence function (IF), which describes its local (linear) stability. The generality of this approach coupled with a weaker robustness requirement has enabled the authors to develop robust methodologies applicable to a wide variety of settings and models, both in iid and time series contexts.

The goal of this PhD Thesis is the definition of new robust estimators, thereby extending the available theory and exploring new directions for applications in finance. The Thesis contains three papers, which analyze three different types of estimators: M-, Minimum Distance- and R-estimators. The focus is manly of their infinitesimal robustness, but global robustness properties are also considered.

The first paper (*"Higher-order infinitesimal robustness"*) studies M-estimators and it is a joint work with Elvezio Ronchetti and Fabio Trojani. Using the higher-order von Mises expansion, we go beyond the IF and we extend Hampel's paradigm of robustness, introducing higher-order infinitesimally robust M-estimators. We show that a bounded estimating function having also bounded gradient with respect to the parameter ensures, at the same time, the stability of the: (i) second-order approximated bias (B-robustness); (ii) asymptotic variance (V-robustness) and (iii) saddlepoint density approximation. An application in finance concludes the paper. Specifically, our analysis deals with maximal losses of Nikkei 225 index returns for static risk-management purposes. We show that both Maximum Likelihood and first-order robust estimators can be badly attracted by anomalous negative shocks in the Japanese market, leading to higher measures of risk. In contrast, our second-order roust M-estimator improves on both Maximum Likelihood and first-order robust estimators, down-weighting the abnormal negative returns and yielding more stable estimates of the market risk.

The second paper (*"On robust estimation via pseudo-additive information measures"*) is jointly written with Davide Ferrari and it studies a new class of Minimum Divergence (in the following, MD) estimators.

The theoretical contribution of the paper is to show that robustness is dual to information theory. Information theory plays a crucial role in statistical inference: Maximum Likelihood estimators are related to it through the minimization of Shannon entropy (namely, minimization of the Kullback-Leibler divergence). The fundamental axiom characterizing Shannon entropy is additivity. Relaxing this assumption, we obtain a generalized entropy (called q-entropy) which exploits the link between information theory and infinitesimal robustness. Minimizing the q-entropy, we define a new class of MD robust re-descending estimators, featuring B-, V-robustness and that have also good global robustness properties in terms of high-breakdown. Besides the theoretical aspects, from a practical stand point, the MD-estimators defined in the paper are easy-to-implement, since they are a kind of weighted-likelihood M-estimators that can be applied for general inferential problems (e.g., location, scale, shape estimation). The paper is submitted to *Biometrika*, since 2009 (fourth round).

The third paper (*"Semi-parametric rank-based tests and estimators for Markov processes"*) contains the main theoretical results of a working paper started in Princeton, under the supervision of Marc Hallin. The paper deals with R-estimators and rank-based tests. Precisely, combining the flexibility of the semi-parametric approach with the distribution-freeness of rank statistics, we define R-estimators and tests for stationary Markov chains. The paper has two different grounds.

The first ground is methodological. We introduce a rank-based semi-parametric efficient (at a given reference model) score function and we apply it to define: (i) a root-n consistent R-estimator and (ii) the most stringent test for a null semi-parametric hypothesis. The paper relies on the results in Hallin and Werker (2003, 2006), who define R-estimators for AR(p) processes. We adapt these results to general (linear and non linear, homo- and hetero-schedastic), stationary Markov chains. The semi-parametric approach and the distribution freeness of the derived rank-based statistical procedures imply good global robustness properties. Nevertheless, we show that semi-parametric efficient R-estimators and tests are not infinitesimally robust. To overcome this problem, bounded rank-based scores must be introduced: The new scores implies infinitesimal

robustness, but typically leads to efficiency losses.

The second ground is related to an application in finance. We define the setting for making inference and testing in stochastic volatility (SV) models. Our statistical procedures rely on the Two Scale Realized Volatility (TSRV), applied as a proxy for the unobserved asset volatility. The advantage related to the TSRV id twofold. First, the TSRV simplifies the estimation procedure, since it avoids MCMC and other computational intensive filtering procedures . Second, the TSRV allows us to recover the Markovian structure of the joint process of returns and volatility, justifying the application of our rank-based procedures for Markov chains. Both static and dynamic risk-management applications are going to be considered in the next future.

Chapter 1

Higher-order infinitesimal robustness

1.1 Introduction

Many authors in statistics, econometrics and other fields have studied the robustness properties of estimators and test statistics in a variety of settings, under different forms of deviations from ideal model assumptions. Nowadays, the need for a robust statistical approach, one which limits the sensitivity to small and moderate deviations from the theoretical model, is well recognized. Robust statistics can then be viewed as an extension of parametric statistics, taking into account that parametric models are only approximations of reality. Corresponding theories for robust estimation and testing have been developed. Tukey [79], Huber [54] and Hampel [48] laid the foundations of modern robust statistics. Book-length expositions can be found in Huber [55] (1981, 2nd edition by Huber and Ronchetti [56]), Hampel et al. [49] and Maronna et al. [68].

Historically, the first line of research formalizing the robustness problem was Huber's minimax theory [54]. In this context, the statistical problem is formulated as a game between the Nature, which chooses a distribution in the neighborhood of the model, and the statistician, who chooses a statistical procedure in a given class. The statistician achieves robustness by constructing a minimax procedure that minimizes a loss criterion (e.g., the asymptotic variance) at the least favorable distribution in the neighborhood.

A second line of research opened by Hampel has considered local robustness, or infinitesimal robustness,

i.e., the impact of moderate distributional deviations from ideal models on a statistical procedure. Here the quantities of interest (for instance the bias or the variance of an estimator) are viewed as functionals of the underlying distribution and their linear approximation is used to study the behavior in a neighborhood of the ideal model. In this setting, a key tool is a functional derivative, the Influence Function (IF), which describes the local stability of the functional. The generality of this approach has enabled authors to develop robust methodologies applicable to a wide variety of settings and models, both in iid and time series contexts. An overview of the theory in the iid setting is provided by Hampel et al. [49]. In the time series framework, Martin and Yohai [69] characterize the influence of different types of contamination outliers for estimators of linear ARMA processes, while Künsch [59] constructs optimal robust estimators for linear autoregressive processes. More recently, Ronchetti and Trojani [72], Mancini et al. [65], and La Vecchia and Trojani [60] propose a class of M-type robust statistical procedures that are broadly applicable to a variety of strictly stationary, potentially non-linear, time-series models.

Finally, using the concept of breakdown point (see Hampel [48], Donoho and Huber [23]), a third stream of literature has focused on global robustness features, i.e., the potential impact of large contaminations by outliers on a statistical procedure. In this literature, several so-called high-breakdown estimators have been proposed, in order to cope with a large fraction of observations not consistent with the ideal model. An overview of this approach for the linear regression model can be found in Maronna et al. [68, Chapter 4-5].

Infinitesimal robustness properties are characterized by the smoothness properties of a statistical functional in neighborhoods of a given reference model P_0. A minimal requirement is qualitative robustness, i.e., weak continuity of the functional. Alternatively, different notions of differentiability, i.e., Gâteaux or Fréchet differentiability, can be considered. So far, the literature has focused on first-order differentiable robust functionals. Hampel [46] introduces bounded-influence robust estimators, implying a uniformly bounded first-order Gâteaux derivative. These estimators are characterized by a bounded first-order von Mises [83]

kernel given by the estimator's IF. Such infinitesimally robust estimators are first-order robust. The first-order von Mises [83] kernel is the key instrument to robustify estimators with unbounded IF, like, e.g., many Maximum Likelihood estimators: Optimal (first-order) bounded-influence estimators are obtained by down-weighting observations that are measured as too influential with respect to the unbounded first-order kernel of a non-robust estimator. Thus, the first-order description of the local estimator's behavior has important implications for the definition and construction of optimal (first-order) bounded-influence M-estimators, which are defined by a bounded estimating function.

An important aspect for the motivation of our higher-order infinitesimally robust approach is the fact that the characterization of the local functional's behavior by means of the first-order von Mises [83] kernel can rapidly become inaccurate, even for small deviations from the ideal model. This happens when the functional behavior is sufficiently nonlinear. A simple example within a linear regression model, $y = \beta_0 + \beta_1 x + u$, where $x \sim N(0, 10.5)$ and $u \sim N(0, 1)$ under the ideal model, illustrates this point. In this context, we can consider a bivariate outlier in (x, y) coordinates, for an increasing probability of contamination ε. In Figure 1.1, the continuous line plots the "asymptotic bias" (henceforth called the bias, for the sake of brevity) of the Least Squares estimator as ε increases. The dashed-dotted curve plots the (linear) approximation of the bias provided by the first-order von Mises [83] kernel, while the dashed line plots the bias according to a second-order von Mises [83] approximation. Even in this simple setting, the local functional approximation provided by first-order von Mises kernels can produce quantitatively large approximation errors. For instance, while for a (quite small) degree of contamination $\varepsilon = 5\%$, the second-order functional approximation for the intercept is virtually exact, the first order approximation implies a relative error of about 30%. Such approximation errors can have quantitatively relevant implications for the optimality properties of first-order infinitesimally robust estimators.

As a result of the motivation and intuition evident in the above example, we study systematically the higher-order infinitesimal robustness properties of a general M-functional. To achieve this goal, we rely on

higher-order von Mises [83] expansions and characterize in detail the second-order robustness features of M-functionals. This analysis provides a number of novel findings. First, we show that second-order robustness is equivalent to the boundedness of both the estimating function and its derivative with respect to the parameter. In location models, boundedness of the derivative of the estimating function implies a bounded local-shift sensitivity, i.e., a bounded influence of grouping and rounding effects on the estimator; see Hampel et al. [49], p. 88. Second, we prove that second-order robustness not only implies the robustness of the second-order bias of an estimator (second-order B-robustness), but it also yields the robustness of its asymptotic variance functional (V-robustness). This last property is not shared in general by first-order robust functionals; see Hampel et al. [49]. Therefore, taking into account the second-order robustness features of an estimator reconciles B- and V-robustness aspects. Third, we find that second-order robustness allows one to go beyond bias and variance characterizations, making it possible to analyze in detail the robustness of procedures used to obtain refined finite sample approximations to the density of an estimator. We show that second-order robustness of an M-estimator implies the robustness of the corresponding saddlepoint density approximation (see, e.g., Daniels [21] and Field and Ronchetti [31]) and its relative errors. Fourth, since many infinitesimally (first-order) robust estimators are not second-order robust, we introduce a new class of second-order robust M-estimators, show that their estimating function can be redescending, and provide an algorithm for their implementation. This opens up a wide field of possible applications, which include a large number of models for which a first- order B-robust estimator is available. Finally, we study the finite sample properties of our second-order robust M-estimators by Monte Carlo simulation and in a real-data application to the estimation of the tail of maximal losses of Nikkei 225 index returns. Monte Carlo simulations, in the linear regression and the Generalized Extreme Value (GEV) setting, show that the second-order robust M-estimator (i) has a better control on the damaging effect of outliers and (ii) produces a lower Mean Squared Error than Maximum Likelihood and optimal first-order robust estimators. The real-data application shows additional interesting features. We find that second-order robust M-estimators

can control better the influence of observations that artificially inflate the point estimate of the tail index. This leads to a more accurate quantification of tail risk and implies more accurate estimated maximal losses in one-out-of k-years.

This paper is organized as follows. Section 1.2 introduces the higher-order von Mises approximation of a statistical functional. Section 1.3 characterizes second-order infinitesimally robust M-functionals, derives their main properties and explains their construction. Section 1.4.1 proves the admissibility of our class of second-order robust M-estimators, while Section 1.5 provides an algorithm for their implementation. Section 1.6 applies our second-order robust methodology in Monte Carlo simulations and to real-data. Assumptions and proofs are in Appendix 1.9 and Appendix 1.10, respectively, while Figures and Tables are collected in Appendix 1.11.

1.2 Higher-order Expansion for Statistical Functionals

Let $\mathcal{M}$ be the family of all probability measures on $\mathcal{Z} \subset \mathbb{R}^m$ and let $T : dom(T) \to \mathbb{R}^p$ be a statistical functional, defined on $dom(T) \subset \mathcal{M}$ and taking values in $\mathbb{R}^p$, with $p \geq 1$. For $P \in \mathcal{M}$, the functional value $T(P)$ can represent any characteristic of P, e.g., the location, the scale, some quantile, a tail area or, more generally, a quantity of interest depending on P. Robust statistics is concerned with the smoothness properties of $T(\cdot)$ in neighborhoods of a fixed ideal model $P_0 \in \mathcal{M}$. Intuitively, small deviations from P_0 should imply small variations in the value of the functional. To formalize this idea, we consider the contamination neighborhood:

$$\mathcal{U}_\eta(P_0) := \{P_{\varepsilon,G} := (1-\varepsilon)P_0 + \varepsilon G \ : \ 0 \leq \varepsilon < \eta, \eta \ll 1 \text{ and } G \in \mathcal{M}\}, \tag{1.2.1}$$

and study the behavior of $T(P_{\varepsilon,G})$ as a function of $P_{\varepsilon,G} \in \mathcal{U}_\eta(P_0)$. The standard robust analysis of $T(P_{\varepsilon,G})$ relies on its linearization, as given by the first-order von Mises expansion. The linear term is uniquely identified by the first-order von Mises kernel, computed using the first-order Gâteaux derivative:

$$\varphi_1(z_1; P_0) = \frac{\partial}{\partial \varepsilon_1} T(P_0 + \varepsilon_1(\delta_{z_1} - P_0))|_{\varepsilon_1 = 0}, \tag{1.2.2}$$

where δ_{z_1} is a Dirac mass at $z_1 \in \mathcal{Z}$. By definition, the first order kernel $\varphi_1(z_1; P_0)$ is the IF of $T(\cdot)$ at P_0 (see Hampel [46] and Hampel et al. [49]) and a bounded kernel $\varphi_1(\cdot; P_0)$ implies a first-order robust statistical functional. Higher-order terms are related to the higher-order Gâteaux derivatives of $T(\cdot)$; for a formal definition, see von Mises [83] and Fernholz [28]. The general k–th order kernel is:

$$\varphi_k(z_1, z_2, ..., z_k; P_0) = \frac{\partial^k}{\partial\varepsilon_1\partial\varepsilon_2...\partial\varepsilon_k} T((1 - \sum_{i=1}^{k}\varepsilon_i)P_0 + \sum_{i=1}^{k}\varepsilon_i\delta_{z_i})|_{\varepsilon_1=0,...,\varepsilon_k=0}. \tag{1.2.3}$$

All kernels are symmetric and characterize the von Mises expansion of $T(\cdot)$ at the k–th order:

$$T(P_{\varepsilon,G}) = T(P_0) + \varepsilon T_1(P_{\varepsilon,G} - P_0) + \varepsilon^2 T_2(P_{\varepsilon,G} - P_0) + ... + \varepsilon^k T_k(P_{\varepsilon,G} - P_0) + O\left(\|P_{\varepsilon,G} - P_0\|^{k+1}\right), \tag{1.2.4}$$

where:

$$\begin{aligned}
T_1(P_{\varepsilon,G} - P_0) &= \int_{\mathcal{Z}} \varphi_1(z_1; P_0) d(P_{\varepsilon,G} - P_0)(z_1) \\
T_2(P_{\varepsilon,G} - P_0) &= \frac{1}{2}\int_{\mathcal{Z}^2} \varphi_2(z_1, z_2; P_0) d(P_{\varepsilon,G} - P_0)(z_1) d(P_{\varepsilon,G} - P_0)(z_2) \\
&\vdots \\
T_k(P_{\varepsilon,G} - P_0) &= \frac{1}{k!}\int_{\mathcal{Z}^k} \varphi_k(z_1, z_2, ..., z_k; P_0) \prod_{i=1}^{k} d(P_{\varepsilon,G} - P_0)(z_i).
\end{aligned}$$

If ε is small, the first-order kernel might provide a good approximation of $T(P_{\varepsilon,G})$. In general, higher-order kernels can be considered, in order to better capture potential (local) nonlinearities. Therefore, the analysis of higher-order terms in the von Mises expansion can give additional insights about the local robustness properties of functional $T(\cdot)$. The next example illustrates this point.

Example 1: Logistic regression. Consider for simplicity a logistic regression model with one covariate: $Y|X = x \sim Bin(1, \Lambda(x\beta))$, where $\Lambda(x\beta) = \exp(x\beta)/(1 + \exp(x\beta))$, $\beta \in \mathbb{R}$ and $x \in \mathbb{R}$. The Maximum Likelihood estimator of β_0 is the M-functional $T(\cdot)$ such that $E_{P_0}[\psi(Y, X; \beta_0)] = 0$ iff $T(P_0) = \beta_0$, where $\psi(y, x; \beta_0) := (y - \Lambda(x\beta_0))x$. It is well-known that the first-order kernel is $\varphi_1(y_1, x_1; P_0) = M^{-1}(\psi; P_0)\psi(y_1, x_1; \beta_0)$, where $M(\psi; P_0) = E_{P_0}[\Lambda(X\beta_0)\,(1 - \Lambda(X\beta_0))\,X^2]$. After some algebra, we obtain

the second-order von Mises kernel as:

$$\begin{aligned}\varphi_2(y_1, y_2, x_1, x_2; P_0) &= M^{-1}(y_1 - \Lambda(x_1\beta_0))x_1 + M^{-1}(y_2 - \Lambda(x_2\beta_0))x_2 \\ &\quad + WM^{-2}\left(y_1 - \Lambda(x_1\beta_0)\right)x_1x_2\left(y_2 - \Lambda(x_2\beta_0)\right) \\ &\quad - M^{-2}\Lambda(x_1\beta_0)\left(1 - \Lambda(x_1\beta_0)\right)x_1^2x_2\left(y_2 - \Lambda(x_2\beta_0)\right) \\ &\quad - M^{-2}\Lambda(x_2\beta_0)\left(1 - \Lambda(x_2\beta_0)\right)x_2^2x_1\left(y_1 - \Lambda(x_1\beta_0)\right), \qquad (1.2.5)\end{aligned}$$

where $M = M(\psi; P_0)$ and $W = -E_{P_0}\left[\Lambda(X\beta_0)\left[1 - \Lambda(X\beta_0)\right]\left[1 - 2\Lambda(X\beta_0)\right]X^3\right]$. Note that, in contrast to $\varphi_1(y_1, x_1; P_0)$, the dependence on y and x in equation (1.2.5) is quadratic and cubic, respectively, highlighting a potentially more pronounced nonlinearity pattern in the second-order approximation of $T(P_{\varepsilon,G})$.

Remark A special case of $T(P_{\varepsilon,G})$ arises for $P_{\varepsilon,G} \equiv P_n$, where P_n is the empirical distribution of $z_1, \ldots, z_n$. Then, equation (1.2.4) is for $k = 1$ the first-order approximation of $T(P_n) - T(P_0)$, i.e.:

$$T(P_n) - T(P_0) = \frac{1}{n}\sum_{i=1}^{n}\varphi_1(z_i; P_0) + O_p(n^{-1}). \qquad (1.2.6)$$

In this case, the condition $\|\varphi_1(z; P_0)\| < b < \infty$, for all $z \in \mathcal{Z}$ and for some positive constant b, ensures a bounded sensitivity of the linearized empirical statistical functional to perturbations of observations $z_1, \ldots, z_n$. For a generic vector $r \in \mathbb{R}^p$, $\|r\|$ represents the Euclidean norm, while for a matrix $M = (M)_{1\le i,j\le p}$, we have $\|M\| := (\sum_{i,j} M_{i,j}^2)^{1/2}$. Using the theory of U-statistics (see Hoeffding [53]), the second-order kernel can be used to shed additional light on the behavior of $T(P_n)$. Note that Equation (1.2.4) for $k = 2$ gives:

$$T(P_n) - T(P_0) = \frac{1}{n}\sum_{i=1}^{n}\varphi_1(z_i; P_0) + \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\varphi_2(z_i, z_l; P_0) + O_p(n^{-3/2}). \qquad (1.2.7)$$

Thus, the joint analysis of φ_1 and φ_2 can provide a more accurate description of the finite sample behavior of $T(P_n)$; cf. also Mallows [64].

Higher-order von Mises expansions like (1.2.4) can be applied to study several statistical functionals. Among the functionals we are interested in, M-functionals play a crucial role in statistics and econometrics. For

these functionals, the behavior of $T(P_{\varepsilon,G})$ in terms of the first-order kernel approximation has been analyzed in great detail in the literature; see, among others, Hampel et al. [49]. However, the simple example in the Introduction and Example 1 suggest that additional information can be obtained by a robust analysis involving $\varphi_2(z_1, z_2; P_0)$. Equations (1.2.2) and (1.2.3), with $k = 2$, are the general definitions of $\varphi_1(z_1; P_0)$ and $\varphi_2(z_1, z_2; P_0)$, but they do not provide explicit characterizations for more concrete applications. In the next section, we provide such a characterization for the class of M-functionals.

1.3 Second-order Robust M-Functionals

We study higher order robustness properties of M-statistical functionals and focus on second-order robustness. We first introduce the relevant definition. In a second step, we study the link between second-order robustness and the robustness of additional statistical functionals related to M-estimators, including, e.g., the estimator's asymptotic variance and the saddlepoint approximation to its finite sample density.

1.3.1 Definition

Given an estimating function $\psi : \mathcal{Z} \times \mathbb{R}^p \to \mathbb{R}^p$, the M-functional $T(\cdot)$ is implicitly defined as the unique functional root of the system of moment conditions: $E_P[\psi(Z; T(P))] = 0$. Given observations $z_1, z_2, ..., z_n$, a M-estimator is the sample counterpart of the M-functional, i.e., the implicit solution of the finite sample equations $\sum_{i=1}^n \psi(z_i; \hat{\theta}_n) = 0$, where $\hat{\theta}_n := T(P_n)$. When the function ψ satisfies the assumptions in Appendix 1.9, $\hat{\theta}_n$ is root-n consistent and asymptotically normal.

M-functionals provide a convenient setting for the computation of first and second-order kernels. It is well-known that the first-order kernel is proportional to the estimating function: $\varphi_1(z; \theta_0) = M^{-1}(\psi; \theta_0)\psi(z; \theta_0)$, where $M(\psi; \theta_0) := E_{P_0}[-\nabla_{\theta'}\psi(Z; \theta_0)]$ and $\theta_0 := T(P_0)$. By definition, B-robust M-functionals have a bounded first-order von Mises kernel or, equivalently, a bounded estimating function.

Using results in Gatto and Ronchetti [32], Appendix A, the second-order kernel is given by:

$$\begin{aligned}\varphi_2(z_1,z_2;\theta_0) &= \varphi_1(z_1;\theta_0)+\varphi_1(z_2;\theta_0)+M^{-1}(\psi;\theta_0)\gamma(z_1,z_2;\theta_0)\\ &\quad +M^{-1}(\psi;\theta_0)\left[\nabla_{\theta'}\psi(z_2;\theta_0)\varphi_1(z_1;\theta_0)+\nabla_{\theta'}\psi(z_1;\theta_0)\varphi_1(z_2;\theta_0)\right], \qquad (1.3.1)\end{aligned}$$

where

$$\gamma(z_1,z_2;\theta_0)' = \begin{pmatrix} \varphi_1'(z_2;\theta_0)E_{P_0}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\psi^{(1)}(Z;\theta_0)\right]\varphi_1(z_1;\theta_0) \\ \vdots \\ \varphi_1'(z_2;\theta_0)E_{P_0}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\psi^{(p)}(Z;\theta_0)\right]\varphi_1(z_1;\theta_0) \end{pmatrix}, \qquad (1.3.2)$$

where $\psi^{(j)}$ is the j-th component of the vector ψ.

Remark. If the estimating function ψ is continuos, $\nabla_{\theta'}\psi$ and the derivatives in (1.3.2) are classical derivatives. Under the weaker assumption that ψ is only P_0-a.s. continuous, then $\nabla_{\theta'}\psi$ and $\partial^2\psi^{(j)}/\partial\theta\partial\theta'$, for $j=1,...,p$, must be interpreted as distributional derivatives.

Higher-order kernels can be computed with the recursive formula in Withers [89]. Note that the kernel $\varphi_2(z_1,z_2;\theta_0)$ in equation (1.3.1) depends on $\psi(z;\theta_0)$ and its first derivative with respect to θ. More generally, the $k-$th order kernel $\varphi_k(z_1,...,z_k;\theta_0)$ depends on the estimating function and its derivatives up to order $k-1$. Consequently, a bounded (continuous) estimating function $\psi(z;\theta_0)$ with bounded (continuous) derivatives up to order $(k-1)$ ensures that the k-th order von Mises approximation (1.2.4) of functional $T(P_{\epsilon,G})$ is a bounded function of $G\in\mathcal{M}$. Therefore, we call k-th order B-robust M-functionals, the M-functionals having bounded continuous von Mises kernels up to order $k-1$. For brevity, we also denote by $BiasII(\varepsilon;\psi,G;\theta_0)$ the second-order von Mises approximation of $T(P_{\epsilon,G})-T(P_0)$, implied by von Mises expansion (1.2.4) for $k=2$.

From the above discussion, it follows that second-order B-robust M-functionals are characterized by a bounded second-order von Mises expansion (1.2.4), i.e., a bounded (continuous) estimating function with bounded derivative. We summarize this finding in the next proposition.

Proposition 1.3.1. *Let $T(\cdot)$ be a Fisher consistent M-functional defined by a continuous estimating function $\psi:\mathcal{Z}\times\mathbb{R}^p\to\mathbb{R}^p$ satisfying the assumptions in Appendix 1.9 and satisfying the two following conditions:*

$$(i)\ \sup_{z\in\mathcal{Z}}\|\psi(z;\theta_0)\|<\infty\ ;\ (ii)\ \sup_{z\in\mathcal{Z}}\|\nabla_{\theta'}\psi(z;\theta_0)\|<\infty. \qquad (1.3.3)$$

Then, $BiasII(\varepsilon;\psi,\cdot,\theta_0)$ is a bounded function of $G \in \mathcal{M}$.

Remark If the estimating function ψ is P_0-a.s. continuous, then some additional care is needed in Proposition 1.3.1. Let $\mathcal{D}(\psi)$ denote the set of points where ψ is not continuous. Then, $\varphi_2(z_1, z_2; \theta_0)$ has an irregular part consisting of delta functions. Thus, $BiasII(\varepsilon;\psi,\cdot,\theta_0)$ is unbounded for contaminations with positive mass on $\mathcal{D}(\psi)$. In these cases, a smooth version of the estimating function ψ can be defined in order to ensure the continuity. In Example 2 and 6, we introduce explicit examples of such a regularized estimating function for specific models.

The class of estimating functions in Proposition 1.3.1 defines the family of second-order robust M-estimators. By definition, second-order B-robustness implies first-order B-robustness. However, since $\varphi_2(z_1, z_2; P_0)$ typically depends on $\varphi_1(z_1; P_0)$, it is interesting to study whether there are situations in which first order robustness implies second-order robustness. The next example introduces a class of M-functionals where this is the case.

Example 2: Location models. Let $Z \sim \phi(z - \theta_0)$, where ϕ is the standard normal density. Then, the log-likelihood score function is $s(z;\theta_0) = z - \theta_0$ and the Maximum Likelihood M-functional (the arithmetic mean) has an unbounded first-order kernel. First-order robustness can be ensured by a M-functional, defined by a bounded estimating function satisfying the assumptions in Appendix 1.9, which implies a bounded first-order von Mises kernel $\varphi_1(z_1;\theta_0)$. The discussion of the second-order kernel properties requires additional assumptions on the smoothness of the estimating function $\psi(z-\theta_0)$. Let $\mathcal{D}(\psi)$ be the finite set of the points in $\mathcal{Z}$ where the function ψ is not continuous. Then, the derivative $\psi' := \partial_\theta \psi$ must be interpreted as a Schwartz distribution (see, e.g., Hampel et al. [49, page 127]):

$$\psi' = \psi' I_{\mathcal{Z}\setminus\mathcal{D}(\psi)} + \sum_{c\in\mathcal{D}(\psi)} (\psi(c+) - \psi(c-))\Delta_c, \tag{1.3.4}$$

which is the sum of a regular part and a linear combination of delta functions Δ_c. Then, formula (1.3.1) reads:

$$\varphi_2(z_1, z_2; \theta_0) = \left(1 - M^{-1}(\psi_b; \theta_0)\psi'(z_2 - \theta_0)\right) \varphi_1(z_1; \theta_0) + \left(1 - M^{-1}(\psi_b; \theta_0)\psi'(z_1 - \theta_0)\right) \varphi_1(z_2; \theta_0).$$

If ψ is a continuous bounded function (e.g., Tukey's biweight, Andrew's sine and Huber-estimating function), kernel φ_2 is regular. Thus, the first-order robust location estimator is also second-order robust. If ψ is only P_0-a.s. continuous (e.g., in the case of skipped median or skipped mean), the kernel φ_2 contains irregular parts. Then, contaminations with non zero mass on $\mathcal{D}(\psi)$ imply an infinite $BiasII(\varepsilon; \psi, \cdot, \theta_0)$, even when ψ is bounded. This problem can be avoided using a regularized version of the estimating function. Precisely, given an estimating function ψ discontinuous on a set of zero P_0-measure, a regularized version $\tilde{\psi}$ can be obtained, e.g., by the convolution with a symmetric kernel, as illustrated in Hampel et al. [47]:

$$\tilde{\psi}(x) = \int_{\mathcal{Z}} \psi(x + u) dQ(u) \tag{1.3.5}$$

where $Q(u)$ is $N(0, V)$ and V is the smoothing parameter.

Beyond location functionals, there is a wide class of first-order robust M-estimators that are not second-order robust. In order to explore the link between first and second-order robustness more broadly, it is convenient to start from the standard construction of optimal first-order robust M-estimators; see, e.g., Hampel et al. [49]. Given an unbounded estimating function $\psi(z; \theta_0)$ such that $E_{P_{\theta_0}}[\psi(Z; \theta_0)] = 0$, an optimal first-order B-robust M-estimator is defined by a weighted estimating function given by:

$$\psi_b(z; \theta_0) := v \min(1; b/\|v\|) \ ; \ v := A(\theta_0)[\psi(z; \theta_0) - \tau(\theta_0)] \ , \tag{1.3.6}$$

where the matrix $A(\theta_0)$ is such that either

$$E_{P_0}[-\nabla_{\theta'} \psi_b(Z; \theta_0)] = I \quad \textit{(unstandardized case)} \tag{1.3.7}$$

or

$$E_{P_0}[\psi_b(Z; \theta_0)\psi_b'(Z; \theta_0)] = I \quad \textit{(self-standardized case)} \tag{1.3.8}$$

and vector $\tau(\theta_0)$ satisfies the moment condition $E_{P_{\theta_0}}(\psi_b(Z;\theta_0)) = 0$ at the reference (parametric) model.

In order to study the second-order robustness properties of this estimator, we can study the derivative $\nabla_{\theta'}\psi_b(z;\theta_0)$, which is piecewise given by:

$$\nabla_{\theta'}\psi_b(z;\theta_0) = \begin{cases} D_{\theta'}(A,v) + A\left[\nabla_{\theta'}\psi - \nabla_{\theta'}\tau\right] & \text{for } \|v\| \leq b \\ \frac{b}{\|v\|}\left(I - \frac{v}{\|v\|}\frac{v'}{\|v\|}\right)\left(D_{\theta'}(A,v) + A\left[\nabla_{\theta'}\psi - \nabla_{\theta'}\tau\right]\right) & \text{for } \|v\| > b \end{cases}, \tag{1.3.9}$$

where $D_{\theta'}(A,v)$ is a $p \times p$ matrix with il-th component given by $\sum_{j=1}^{p} \frac{\partial A_{ij}}{\partial \theta_l}(\psi^{(j)} - \tau^{(j)})$; see also Lô and Ronchetti [63] for an application in the context of empirical likelihood-type estimators. Therefore, $\nabla_{\theta'}\psi_b(z;\theta_0)$ is bounded provided that:

$$\begin{cases} \|\nabla_{\theta'}\psi\| & \text{is bounded for } \|v\| \leq b \\ \|\nabla_{\theta'}\psi\|\,\|v\|^{-1} & \text{is bounded for } \|v\| > b \end{cases}. \tag{1.3.10}$$

The following examples illustrate the implications of condition (1.3.10) for some well-known first-order robust M-functionals.

Example 3: Linear regression and GLM. Let $y = x'\beta + u$, where $u \sim N(0,1)$ and $x \in \mathbb{R}^p$. In this setting, $z = (y,x)$ and the maximum likelihood score function is $s(y,x;\beta) = (y - x'\beta)\,x$. The (optimal) first-order B-robust estimator is obtained using the estimating function (1.3.6) with $\psi(z;\theta) := s(y,x;\beta)$, where in this case $\tau = 0$ and A is constant. It is easy to see that condition (1.3.10) is violated, since the derivative of estimating function $\psi(z;\theta)$ is $\nabla_{\beta'}s(y,x;\beta) = -xx'$. Thus, this estimator is not second-order robust. For instance, for a sequence (z_n) such that $\|x_n\| \to \infty$ and $\|As(y_n,x_n;\beta)\|$ is bounded from below by b and from above by a second constant, we find that $\|\nabla_{\beta'}s(y_n,x_n;\beta)\|/\|s(y_n,x_n;\beta)\| = O(\|x_n\|^2)$, which violates the second requirement in condition (1.3.10). Analogous calculations hold for Generalized Linear Models; See also Example 1.

Example 4: Generalized Extreme Value (GEV) distribution. The GEV distribution is parameterized by a parameter vector $\theta_0 = (\xi_0, \mu_0, \sigma_0)'$ and has a density given by:

$$p_0(z;\xi_0,\mu_0,\sigma_0) = \frac{1}{\sigma_0}\exp\left[-\left(1 + \xi_0\left(\frac{z-\mu_0}{\sigma_0}\right)\right)^{-\frac{1}{\xi_0}}\right]\left(1 + \xi_0\left(\frac{z-\mu_0}{\sigma_0}\right)\right)^{-\frac{1}{\xi_0}-1}. \tag{1.3.11}$$

For $\xi \to 0$, $\xi > 0$ and $\xi < 0$ equation (1.3.11) corresponds, respectively, to the Gumbel, Fréchet and Weibull family. In this model, the log-likelihood score function $s(z;\theta)$ is unbounded and the classical Maximum Likelihood estimator is not robust. An optimal first-order B-robust estimator of form (1.3.6) was proposed in Dupuis [25] and Dupuis and Field [26] for the choice $\psi(z;\theta) = s(z;\theta)$. Analytical calculations show that the second requirement in condition (1.3.10) is not satisfied. Therefore, this estimator is not second-order robust. This finding can be illustrated numerically for the parameter choice $\mu_0 = 0$, $\sigma_0 = 1$ and $\xi_0 = -0.1$. The function $\|\nabla_{\theta'} s\|$ is continuous over the support of $GEV(\xi_0, \mu_0, \sigma_0)$. Thus, it has a finite maximum on the compact set $|z| \leq 9$ (namely when $\|s\| \leq 2500$). Therefore, the first requirement of (1.3.10) is satisfied. In contrast, the function $\|\nabla_{\theta'} s\| \|s\|^{-1}$ is unbounded, when $|z| > 9$ (namely for $\|s\| > 2500$). This violates the second requirement of condition (1.3.10).

1.3.2 Main properties

In this section, we derive additional robustness properties of second-order robust M-estimators, which are more directly related to the mean square error and finite sample features of these estimators. For the sake of brevity, from now on we assume that the function ψ satisfies the assumptions in Appendix 1.9, implying that ψ is continuous.

B-robustness and V-robustness

According to Proposition 1.3.1, second-order robust M-estimators feature a bounded second-order bias functional. In contrast to first-order robust estimators, they also feature the robustness of their asymptotic variance functional. To highlight this point, let $V(\psi; P_{\varepsilon,G}) := E_{P_{\varepsilon,G}}[\varphi_1(Z; T(P_{\varepsilon,G}))\varphi_1'(Z; T(P_{\varepsilon,G}))]$ and consider (i) the Change of Variance Function (CVF):

$$CVF(z;\psi,P_0) := \frac{\partial}{\partial\varepsilon} V(\psi; P_{\varepsilon,\delta_z})|_{\varepsilon=0}, \tag{1.3.12}$$

as well as (ii) the standardized Change of Variance Sensitivity (CVS):

$$\kappa^*(\psi;P_0) := \sup_{z\in\mathcal{Z}} \frac{trCVF(z;\psi,P_0)}{trV(\psi;P_0)}. \tag{1.3.13}$$

An M-functional such that $\kappa^*(\psi; P_0) < \infty$ is called V-robust. By definition, V-robustness is the robustness of the (asymptotic) variance functional of an estimator. It is known that V-robust M-functionals are first-order B-robust, but in general the converse implication does not hold; see again Hampel et al. [49]. The next Proposition states that, instead, second-order robustness is a sufficient condition for V-robustness.

Proposition 1.3.2. *Let $\psi(z; \theta_0)$ be an estimating function satisfying condition (1.3.3). A second-order B-robust M-functional is V-robust.*

An immediate important consequence of Proposition 1.3.2 is that second-order robust M-functionals also imply a robust Mean Squared Error (MSE) functional, to the first-order in ε. To see this, note that the first two von Mises kernels and the CVF can be applied to approximate the worst-case MSE of an estimator over the neighborhood $\mathcal{U}_\eta(P_0)$ (see Hampel et al. [49]). To this end, define the Maximal MSE (MMSE) over the neighborhood $\mathcal{U}_\eta(P_0)$ as:

$$\text{MMSE}(\varepsilon, n; \psi, P_0) := \sup_G \|BiasII(\varepsilon; \psi, G; P_0)\|)^2 + (1/n) trV(\psi; P_0) \exp[\varepsilon \kappa^*(\psi; P_0)], \tag{1.3.14}$$

where the second term on the right hand side provides a first-order approximation of the asymptotic estimator's variance, since:

$$\sup_G trV(\psi; P_{\varepsilon,G}) = trV(\psi; P_0) \exp[\varepsilon \kappa^*(\psi; P_0)] + O(\varepsilon^2).$$

Proposition 1.3.1 states that second-order robust estimators have a bounded $BiasII(\varepsilon; \psi, \cdot; P_0)$, while Proposition 1.3.2 implies a bounded variance kernel $CVF(\cdot; \psi, P_0)$. Together, this implies that the MMSE of second-order robust functionals in equation (2.4.3) is finite. This property is violated by first-order B-robust M-functionals with unbounded CVF.

Robust saddlepoint density approximation

Bias and variance are two characteristics providing important information about the location and the dispersion of the distribution of an M-estimator. Together, they characterize the first-order asymptotic (normal) distribution of the estimator. More accurate approximations of the finite sample distribution of an estimator can be obtained by means of Edgeworth or similar expansions. In this section, we focus on saddlepoint

techniques, which are known to provide excellent small sample approximations of finite sample densities and tail probabilities of M-estimators, exhibiting relative error properties that improve on the absolute errors of Edgeworth expansions; see Field and Hampel [30] and Field and Ronchetti [31], among others.

When the data are generated from a distribution $P_{\varepsilon,G} \in \mathcal{U}_\eta(P_0)$, satisfying the Assumptions in Field and Ronchetti [31, page 62], the exact finite sample density $f(t; n, \varepsilon, G)$ of the M-functional can be approximated by a saddlepoint approximation of the form

$$f(t; n, \varepsilon, G) = g(t; n, \varepsilon, G)\{1 + a_1(t; P_{\varepsilon,G})n^{-1} + O(n^{-2})\}, \tag{1.3.15}$$

where

$$g(t; n, \varepsilon, G) = (n/2\pi)^{p/2} c^{-n}(\alpha(t; P_{\varepsilon,G}); P_{\varepsilon,G}) \left|\det \tilde{M}(t; P_{\varepsilon,G})\right| \left|\det \tilde{Q}(t; P_{\varepsilon,G})\right|^{-1/2}, \tag{1.3.16}$$

where $\tilde{M}(t; P_{\varepsilon,G}) = E_{H_{\varepsilon,G;t}}[-\nabla_{t'}\psi(Z; t)]$ and $\tilde{Q}(t; P_{\varepsilon,G}) = E_{H_{\varepsilon,G;t}}[\psi(Z; t)\psi'(Z; t)]$.

In this approximation, the term $a_1(t; P_{\varepsilon,G})$ depends, among other things, on the standardized third and fourth cumulants of ψ under $P_{\varepsilon,G}$ (see the Proof of Lemma 1.3.3 in Appendix 1.10). The vector $\alpha(t; P_{\varepsilon,G})$ is the saddlepoint, i.e., the solution of the saddlepoint equation

$$E_{H_{\varepsilon,G;t}}[\psi(Z; t)] = 0, \tag{1.3.17}$$

with $H_{\varepsilon,G;t}(\cdot)$ the conjugate measure of $P_{\varepsilon,G}$, defined by the density:

$$\frac{dH_{\varepsilon,G;t}}{dP_{\varepsilon,G}}(z) = c(\alpha(t; P_{\varepsilon,G}); P_{\varepsilon,G}) \exp\{\alpha'(t; P_{\varepsilon,G})\psi(z; t)\}, \tag{1.3.18}$$

where $c(\alpha(t; P_{\varepsilon,G}); P_{\varepsilon,G})^{-1} = E_{P_{\varepsilon,G}}[\exp\{\alpha'(t; P_{\varepsilon,G})\psi(Z; t)\}]$.

For any given underlying distribution $P_{\varepsilon,G}$, note that the leading term $g(t; n, \varepsilon, G)$ in equation (1.3.16) provides a very accurate approximation, implying a relative error of order $O(n^{-1})$. We are interested in the robustness properties of functionals $g(t; n, \varepsilon, G)$ and $a_1(t; P_{\varepsilon,G})$, and in features of an M-estimator that guarantee robustness of these functionals. To this end, consider the following von Mises expansion of $g(t; n, \varepsilon, G)$ (and similarly for $a_1(t; P_{\varepsilon,G})$):

$$g(t; n, \varepsilon, G) = g(t; n, 0, P_0) + \varepsilon \frac{\partial}{\partial \varepsilon} g(t; n, \varepsilon, G)|_{\varepsilon=0} + O(\varepsilon^2).$$

Note that $\frac{\partial}{\partial\varepsilon}g(t;n,\varepsilon,G)|_{\varepsilon=0}/g(t;n,0,P_0)$ is a standardized sensitivity of the saddlepoint approximation $g(t;n,\varepsilon,G)$ to ε contaminations of the reference model P_0 in direction $G \in \mathcal{M}$. The next Proposition shows that such sensitivities are bounded for second-order robust M-estimators.

Proposition 1.3.3. *Let $\psi(z;\theta_0)$ be an estimating function satisfying the conditions in (1.3.3), then*

$$(i)\sup_G \left|\frac{\partial}{\partial\varepsilon}g(t;n,\varepsilon,G)|_{\varepsilon=0}\right| < \infty \quad ; \quad (ii)\sup_G \left|\frac{\partial}{\partial\varepsilon}a_1(t;P_{\varepsilon,G})|_{\varepsilon=0}\right| < \infty.$$

Proposition 1.3.3 implies an additional property of second-order robust M-functionals: The saddlepoint approximation and its relative error of order $O(n^{-1})$ are uniformly bounded over the contamination neighborhood $\mathcal{U}_\eta(P_0)$. A similar result does not hold in general for first-order robust M-functionals.

Remark Ronchetti and Ventura [73] analyze the effects of a model misspecification on the accuracy of saddlepoint density approximations for univariate estimators of location. They find that small deviations from the parametric model can easily wipe out the improvements in finite sample accuracy, obtained by saddlepoint density approximations, for classical estimators like the arithmetic mean. In contrast, the saddlepoint approximations implied by Huber-type M-estimator of location provide a robust second-order accuracy. According to Example 2, this robust location estimator is second-order robust. Therefore, Proposition 1.3.3 provides a theoretical justification for these findings.

1.4 Construction of Second-order Robust M-functionals

1.4.1 Admissible M-functionals

Let us define the class Ψ of continuous functions $\psi(z;\theta) : \mathcal{Z}\times\mathbb{R}^p \to \mathbb{R}^p$, satisfying Assumptions in Appendix 1.9, such that $E_{P_0}[\psi(z;\theta_0))] = 0$. An admissible second-order robust M-estimator in Ψ is an estimator defined by an estimating function $\psi^* \in \Psi$ which minimizes $trV(\psi,\theta_0)$, subject to a robustness constraint on the estimating function and on its derivative with respect to θ (see Proposition 1.3.1). Thus, an admissible second-order M-functional is the solution to the following minimization problem:

$$\min_{\psi\in\Psi} trV(\psi;\theta_0), \quad s.t. \quad \sup_{z\in\mathcal{Z}}\|\psi(z;\theta_0)\| \le b \quad and \quad \sup_{z\in\mathcal{Z}}\|\nabla_\theta\psi^{(j)}(z;\theta_0)\| \le c^{(j)}, \quad j=1,...,p \tag{1.4.1}$$

for b and $c^{(1)}, ..., c^{(p)}$ positive constants.

Remark If $c^{(j)} \to \infty$ for each j, the problem (1.4.1) becomes the standard first-order robust optimization problem, with one robustness constraint on the Euclidean norm of the estimator IF. The solution to this problem is the well-known estimator implied by: $\psi^* = A(s-\tau)\min(1; b/\|A(s-\tau)\|)$, with A such that $E_{P_0}[-\nabla_{\theta'}\psi^*] = I$. In order to derive an admissible estimator solution to (1.4.1) for a well-known benchmark, we consider the one-dimensional location problem.

Proposition 1.4.1. *Let $Z \sim \phi(z-\theta_0)$, where ϕ is the standard normal density, $\mathcal{Z} \subset \mathbb{R}$ and $\theta_0 \in \Theta \subset \mathbb{R}$. The likelihood score function is $s(z;\theta_0) := \partial_\theta \log \phi(z-\theta_0) = z - \theta_0$. For $b \in \mathbb{R}^+$, define*

$$\tilde{\psi}_{b,c}(z-\theta_0) = \begin{cases} c(z-\theta_0)\min\left(1; \frac{b}{|Ac(z-\theta_0)|}\right) & 0 < c < 1 \\ (z-\theta_0)\min\left(1; \frac{b}{|A(z-\theta_0)|}\right) & c \geq 1. \end{cases} \tag{1.4.2}$$

The M-estimator implied by the estimating function $\psi_{b,c}(z-\theta_0) = A\tilde{\psi}_{b,c}(z-\theta_0)$, where A is such that $E_{P_0}[-\partial_\theta \psi_{b,c}]$=1, is second-order robust and minimizes $V(\psi;\theta_0)$ among all functions $\psi \in \Psi$, such that both $|\psi(z;\theta_0)| \leq b$ and $|\partial_\theta \psi(z;\theta_0)| \leq Ac$, for all $z \in \mathcal{Z}$.

Since in the location problem the estimating function has form $\psi(z-\theta_0)$, it follows that $|\partial_z \psi(z-\theta_0)| = |\partial_\theta \psi(z-\theta_0)|$. Eq. (1.4.2) points out that the constraint on the second-order kernel is binding for $c \in (0,1)$. In this case, Eq. (1.4.2) defines a new estimator having a tighter bound on the derivative. This bound implies a stronger control for the impact that contaminations around the symmetry center θ_0 can have on the second-order von Mises kernel. For $c \geq 1$, since $\sup_z |\partial_\theta \tilde{\psi}_{b,c}(z-\theta_0)| = 1$, the solution to the minimization problem (1.4.1) is the first-order unstandardized robust M-functional of Huber-type (see Eq. (1.3.7)).

For general estimation problems, both in the univariate and multivariate settings, the explicit computation of the solution to (1.4.1) is a difficult task. In the next section, we propose second-order robust estimating functions that can be applied also in these general settings. The implied M-estimators feature second-order robustness, but they are not necessarily admissible in the class Ψ.

1.4.2 M-functionals for general settings

Second-order robust M-functionals are characterized by a bounded estimating function having a bounded derivative with respect to the parameter. The construction of such robust functionals can be based on a

new class of robust weights, which are designed to bound the impact of influential observations on both the first and second-order von Mises kernels.

Let us start from an (unbounded) estimating function $\psi(z;\theta)$ and assume the existence of two functions $h_i : \mathcal{Z} \to \mathbb{R}^+$, $i = 1, 2$, such that (i) $\|\nabla_{\theta'}\psi\|\|\psi\|^{-1} = O(h_1)$ and (ii) $\|\nabla_{\theta'}\psi\| = O(h_2)$. We denote by $g := \max(h_1, h_2)$ the maximum of these two functions. An estimating function implying bounded first and second-order von Mises kernels can be then constructed according to the next proposition.

Proposition 1.4.2. *Given positive constants b and c, define the bounded estimating function:*

$$\psi_{b,c} := A(\psi - \tau) \min\left(1; \frac{b}{\|A(\psi - \tau)\|}\right) \min\left(1; \frac{c}{g}\right), \tag{1.4.3}$$

where the matrix A and the vector τ are solutions of the equations:

$$E_{P_{\theta_0}}[\psi_{b,c}] = 0\ , \tag{1.4.4}$$

$$E_{P_0}[-\nabla_{\theta'}\psi_{b,c}] = I\ . \tag{1.4.5}$$

Then, the function $\psi_{b,c}$ satisfies condition (1.3.3).

Remark In the location case, the M-estimator defined by $\psi_{b,c}$ coincides with the admissible estimator defined in Proposition 1.4.1, e.g., for: $c < 1$, $g \equiv 1$, and b/A in (1.4.3) equal to $b/(Ac)$ in (1.4.2). See, also Example 6 below for a graphical illustration.

Remark By construction, as $b, c \to \infty$ the estimating function $\psi_{b,c}$ in Eq. (1.4.3) converges to the estimating function ψ. For fixed b and as $c \to \infty$ the bound on the second-order kernel is relaxed and the function $\psi_{b,c}$ converges to ψ_b, i.e., the estimating function of the standard first-order B-robust M-estimator. Note that the definition of matrix A depends on the choice of the norm used to measure the bound on the first order kernel $\varphi_1(z_1; P_0)$. Equation (1.4.5) implies a bound defined by the Euclidean norm, but other standardizations are possible. For instance, a condition of the form

$$E_{P_0}\left[\psi_{b,c}\psi'_{b,c}\right] = I \tag{1.4.6}$$

implies a bound with respect to a self-standardized metric; see also equation (1.3.8) and section 4.3 in Hampel et al. [49]. Overall, Proposition 1.3.1 and 1.4.2 give rise to a new class of second-order robust M-estimators, as summarized by the next corollary.

Corollary 1.4.3. *The M-estimator $\hat{\theta}_n$ defined by the solution to $\sum_{i=1}^{n} \psi_{b,c}(z_i; \hat{\theta}_n) = 0$, where $\psi_{b,c}$ is defined by equation (1.4.3), is second-order robust.*

Example 6: Second-order robustness in location models. Consider again the one-dimensional location problem in Example 2, in which the standard (optimal) first-order robust M-estimator of location was shown to be second-order robust. In this setting, the second-order robust estimator of Proposition 1.4.2 can be used to impose a tighter bound on the second-order von Mises kernel, in dependence of the choice of function g and tuning constant c. Additionally, a regularized version of the Huber estimator can be implemented in order to eliminate the problem of non-differentiability, while controlling the bound on the first derivative. Figure 1.2 illustrates these features, by plotting the estimating function of the following M-estimator: the standard first-order robust M-estimator (dashed line), and two second-order robust M-estimators, obtained for $g(z) = z^2$, $b = 2.5$, $c = 6.5$ (continuous line) and for $g(z) = 1$, $b = 4.545$ and $c = 0.55$ (dot-dashed line). For the sake of comparison, we also plot the regularized version of Huber M-estimator (1.3.5), with $V = 2.046$, as proposed in Hampel et al. [47] (dotted line). Note that for $g(z) \equiv c$ our second-order robust M-estimator of location coincides with the standard first-order robust M-estimator. For $c = 0.85$ we obtain an estimating function thath behaves like the Huber estimating function in the extreme regions of the state space, but with a smaller first derivative over the compact set $[-b, b]$. Finally, a polynomial function $g(z)$ implies M-estimators of redescending type, which strongly downweight anomalous observations (continuous line); see, e.g., Hampel et al. [49, Ch. 2.6].

Example 7: Second-order robustness in linear regression models. Consider again the linear regression model of Example 3 with $\psi(z;\theta) = s(y, x; \beta) = (y - x'\beta)x$ and $\tau = 0$. Condition (1.3.10) is violated for $\|v\| > b$ when $\|x\| \to \infty$ and the residual $(y - x'\beta)$ is bounded. The largest speed at which the ratio $\|\nabla_{\beta'} s(y, x; \beta)\| \|v\|^{-1}$ can diverge as $\|x\| \to \infty$ is $\|x\|^2$. This suggests that we can set $h_1(x) = \|x\|^2$. Similarly, condition (1.3.10) can be violated for $\|v\| \leq b$ when $\|x\| \to \infty$ and the likelihood score $(y - x'\beta)x$ is bounded. In this case, the speed of divergence of $\|\nabla_{\beta'} s(y, x; \beta)\|$ is $\|x\|^2$. This suggests that we can set $h_2(x) = \|x\|^2$

as well, i.e., $g(z) = \|x\|^2$. The resulting second-order robust M-estimator (1.4.3) has an estimating function given by:

$$\psi_{b,c}(y, x; \beta) = Axu \min\left(1; \frac{b}{|u|\|Ax\|}\right) \min\left(1; \frac{c}{\|x\|^2}\right), \tag{1.4.7}$$

where $u = y - x'\beta$ and matrix A is determined either by equation (1.3.7) or equation (1.3.8). Note that estimating function (1.4.7) corresponds to the estimating function of an optimal B-robust Hampel-Krasker (unstandardized) or Krasker-Welsch (self-standardized) estimator, with additional *Mallows-type weights* on the x variables. It follows that large x observations are down-weighted more than in the optimal first-order B-robust estimator.

Remark Different criteria can be defined to select the tuning constants b and c. A first selection criterion is related to the ratio of the MSE of the most efficient benchmark (i.e., the MLE) and the MSE of the second-order robust M-functional, under P_0. This ratio gives information about the cost (in terms of efficiency loss) that we have to pay for robustness. Defining:

$$\text{rMSE}(\psi_{b,c}; \theta_0) = \text{MSE}(s; \theta_0)/\text{MSE}(\psi_{b,c}; \theta_0), \tag{1.4.8}$$

we can select the pair (b, c) in order to achieve a pre-specified level of efficiency loss.

A second approach can be based on the selection of a pair (b, c) which minimizes the MMSE defined in (2.4.3). Since the MMSE$(\varepsilon, n; \psi, P_0)$ depends both on the degree of ε–contamination and on the sample size n, the selected values (b, c) will automatically take into account the interplay between bias and variance, in the presence of deviations from P_0, for any given sample size.

1.5 Algorithm

To compute the second-order robust M-estimator in (1.4.3), an iterative algorithm is applied, because the robust weights appearing in (1.4.3), the matrix $A(\theta_0)$, and the vector $\tau(\theta_0)$ all depend on the estimator itself in a nonlinear way. Here we implement the self–standardized version for the matrix $A(\theta)$ given by (1.3.8).

Given a constant $b \geq \sqrt{p}$ (see, e.g., Hampel et al [49], p. 228), the M-estimator is computed by the following algorithm.

Initialize the algorithm with $\omega_{ci} := \min(1; c/g_i)$, and $g_i = g(z_i)$. Then, perform the following steps.

1. Set initial values $\theta^{(0)}$ and $\tau^{(0)} = 0$. For the given (unbounded) score estimating function ψ, set:

$$\left(A^{(0)}A^{(0)\prime}\right)^{-1} = \frac{1}{n}\sum_{i=1}^{n}\psi_i^{(0)}\psi_i^{(0)\prime}$$

where, here for brevity, $\psi_i^{(0)} := \psi\left(z_i;\theta^{(0)}\right)$. Morever, we set:

$$\omega_{bi}^{(0)} := \min\left(1; \frac{b}{\left\|A^{(0)}\left(\psi_i^{(0)} - \tau^{(0)}\right)\right\|}\right).$$

2. Compute $\tau^{(1)}$ as:

$$\tau^{(1)} = \frac{E_{\theta^{(0)}}\left(\psi_i^{(0)}\omega_{bi}^{(0)}\omega_{ci}\right)}{E_{\theta^{(0)}}\left(\omega_{bi}^{(0)}\omega_{ci}\right)} \tag{1.5.1}$$

and $A^{(1)}$ as

$$\left(A^{(1)}A^{(1)\prime}\right)^{-1} = \frac{1}{n}\sum_{i=1}^{n}\left[(\omega_{bi}^{(0)}\omega_{ci})^2\left(\psi_i^{(0)} - \tau^{(0)}\right)\left(\psi_i^{(0)} - \tau^{(0)}\right)'\right].$$

3. Given $\tau^{(1)}$ and $A^{(1)}$, compute the parameter $\theta^{(1)}$ as the solution of the implicit equation:

$$0 = \sum_{i=1}^{n} A^{(1)}\left(\psi\left(z_i;\theta^{(1)}\right) - \tau^{(1)}\right)\min\left(1; \frac{b}{\left\|A^{(1)}\left(\psi_i^{(0)} - \tau^{(1)}\right)\right\|}\right)w_{ci} \tag{1.5.2}$$

Given $\theta^{(1)}$, set $\psi_i^{(1)} := \psi(z_i;\theta^{(1)})$ and:

$$\omega_{bi}^{(1)} := \min\left(1; \frac{b}{\left\|A^{(1)}\left(\psi_i^{(1)} - \tau^{(1)}\right)\right\|}\right).$$

4. Go back to Step 2 and replace $\omega_{bi}^{(0)}$ by $\omega_{bi}^{(1)}$, $\psi_i^{(0)}$ by $\psi_i^{(1)}$ and $\tau^{(0)}$ by $\tau^{(1)}$. Then iterate Steps 2. and 3. until convergence of the sequences $\{\theta^{(j)}\}$, $\{A^{(j)}\}$ and $\{\tau^{(j)}\}$.

In Step 2. of the algorithm, we obtain $\tau^{(j)}$ by computing the two expectations in the numerator and the denominator of equation (1.5.1). Typically those integrals cannot be computed in closed form and Monte Carlo methods or Gaussian quadrature techniques have to be applied.

1.6 Monte Carlo Evidence

In this section, we first study by Monte Carlo simulation the finite sample behavior of our second-order robust M-estimator in two distinct model settings.

1.6.1 Linear regression model

Consider the linear regression model

$$y = \alpha + x'\beta + u, \tag{1.6.1}$$

where $u \sim N(0,1)$, $\alpha = 0.5$ and $\beta = (3,1)'$. In our Monte Carlo experiments, we consider both clean and contaminated samples. Clean samples are generated according to the model (1.6.1), where $x \sim N(0, I)$ and I is the 2×2 identity matrix. Contaminated samples feature contaminations in (x, y)-space. The y-space contamination is obtained using a contaminated response variable $\tilde{y} = H_y y + (1 - H_y)\zeta_y$, where H_y is a Bernoulli random variable such that $P(H_y = 0) = 0.05$ and $\zeta_y \sim \delta_y$, where δ_y is a Dirac mass at -5. For each j-th component $x^{(j)}$ of the vector x, contaminations in the x-direction are obtained as $\tilde{x}^{(j)} = H_x x^{(j)} + (1 - H_x)\zeta_x$, where $\zeta_x \sim N(5, 0.01)$ and H_x is a Bernoulli random variable such that $P(H_x = 0) = 0.10$. The Monte Carlo size is 1000.

We study the finite sample properties of Maximum Likelihood, first-order and second-order robust M-estimators. Sample sizes are $n = 15, 30, 90$. The implemented first-order robust M-estimator is the Hampel-Krasker estimator with $b = 4$. The implemented second-order robust M-estimator is the estimator in (1.4.7), with $A(\theta)$ as in Eq. (1.3.7) and tuning constants $b = 4$ and $c = 3.5$. For both robust estimators, the choice of the tuning constants ensures a good performance in the presence of contamination, while preserving a reasonable Asymptotic Relative Efficiency (ARE) at the reference model. This feature is illustrated in Table 1, which reports the MSE implied by the different M-estimators in clean and contaminated samples, respectively. We find that for $\varepsilon_x = \varepsilon_y = 0$ and a sample size $n = 90$ the first-order and the second-order robust M-estimators have a similar ARE of about 87% and 83%, respectively. For $\varepsilon_x, \varepsilon_y > 0$, we find across

all sample sizes that the second-order robust M-estimator produces a more robust inference and the smallest MSE. For instance, while for $n = 30$ the MSE of the Hampel-Krasker M-estimator is about 75% the MSE of the MLE, the MSE of the second-order M-estimator is about 50% the MSE of the MLE.

1.6.2 Generalized extreme value estimation

In a second Monte Carlo example, we estimate the parameters of a Generalized Extreme Value distribution, where clean samples are generated according to $z \sim GEV(\xi_0, \mu_0, \sigma_0)$, with $\xi_0 = -0.1, \mu_0 = 0, \sigma_0 = 1$, as in Example 4. Contaminated samples are generated by the replacement model $\tilde{z} = zH_z + (1 - H_z)\zeta_q$, where H_z is a Bernoulli random variable such that $\epsilon_z = P(H_z = 0) = 0.09$ and $\zeta_q \sim \delta_q$, with δ_q the Dirac mass at $q = 15$. We compute the MSE of MLE, first- and second-order robust M-estimators for sample sizes $n = 27, 30, 45, 54$ such that $n/p = 9, 10, 15, 18$. The Monte Carlo size is 2500 and tuning constants $b = 2.2$, $c = 15$ are selected to ensure an efficiency loss of about 10% at the reference model. We implement self-standardized versions of first- and second-order robust M-estimators in Eq. (1.4.3), with $\psi(z; \theta) = s(z; \theta)$ and with matrix $A(\theta)$ as in Eq. (1.3.8). Table 2 presents the results. We find that all estimators have a similar performance when $\epsilon_x = 0$ (clean sample), but quite different properties in the presence of contaminations. For each sample size considered, the second-order robust M-estimator has a lower MSE than the MLE and the first-order robust M-estimator, when data are contaminated.

1.7 Empirical Illustration

1.7.1 Application to static risk management

To test empirically our robust estimation procedure, we model the maximal losses of the returns of the Nikkei 225 index using a GEV distribution, as is standard in the static risk management literature; see for instance, McNeil, Frey and Embrechts [70], Chapter 7. We estimate the tail of the stationary distribution of an underlying stationary heteroskedastic process for returns (e.g. ARCH or GARCH). By construction, this approach focuses only on the properties of the tails of the stationary distribution of returns, without explicitly

modeling dynamic volatility features estimated using robust estimation methods. A natural extension of our approach could follow Mancini and Trojani [67], who combine (first-order) robust estimators of GARCH parameters with (first-order) robust estimators of the tails of standardized GARCH residuals.

The sample period is from 7-Jan-70 to 17-Aug-04 and contains 8567 daily returns. Data are available online from the web page www.unige.ch/ses/metri/gilli/evtrm/ and have been analyzed by Gilli and Këllezi [34] using standard (Pseudo) Maximum Likelihood estimators. The sample has been split into 35 non-overlapping sub-samples, each containing the daily returns of the successive calendar year. Figure 1.3 presents the return time series (top panel), together with the histogram for their minima, i.e., the block maximal losses over the 35 sub-samples (bottom panel). The histogram shows that the majority of maximal losses is in the interval $[2\%, 8\%]$, with a right tail starting approximatively at 8%. To infer the distribution of maximal losses, we apply the Fisher-Tippet theorem, which states that the limit distribution of maximal losses is a GEV distribution, whose parameters can be estimated by means of the robust estimation approach developed in the previous sections.

1.7.2 Estimation results

We compare the performance of our second-order robust M-estimator with the performance of MLE and first-order robust M-estimator. The robust tuning constants have been set at $b = 3.9$ and $c = 20$. Table 3 summarizes the results.

We find that while all methods estimate very similar location and scale parameter, they imply a quite different estimated shape parameter. The MLE estimate of the parameter ξ is about 2 times larger than the estimate of the first-order robust M-estimator. The second-order robust M-estimator gives a fitted value of ξ which is very close to zero (-0.0086), implying a distribution similar to a Gumbel or a Weibull. As a result, both the MLE and first-order robust M-estimator imply a heavier estimated right tail for the loss distribution. This feature is shown in the right panel of Figure 1.5. In order to understand in more detail why

different estimators perform differently, Figure 1.4 plots the robust weights estimated by first- and second-order robust M-estimators. The top and the bottom panels of Figure 1.4 show that (i) both estimators clearly downweight observation z_{18} and (ii) the second-order robust M-estimator additionally downweights the observations z_1 and z_3, which are found to be influential with respect to the estimator's second-order kernel. Such influential observations can largely inflate the estimated shape parameter of classical and first-order robust estimators, largely determining the estimated tail properties.

Our findings have potentially important implications for estimated risk measures derived from this analysis. To illustrate this point, we compute a well-known risk measure, the return-level over k periods of length one year. If F is the distribution function of maximal losses over successive non-overlapping periods, the return level $R^k = F^{-1}(1 - 1/k)$ is expected to be exceeded in one out of k periods of length one year. This quantity can be used as a measure of the maximum loss of a portfolio. We estimate this risk measure using the GEV parameters estimated by the three different estimation methods. Figure 1.5 (left panel) shows that the higher tail probability estimated by the MLE and first-order robust M-estimator yields a clearly higher estimated value for R^k. For instance, when $k = 10$, R^{10} is about 8.24% for MLE, about 7.84% for the first-order robust M-estimator and about 7.27% for the second-order robust M-estimator.

1.8 Conclusions

We introduce higher-order infinitesimally robust M-estimators and characterize in detail second-order robustness properties. First, we show that second-order robust estimators are characterized by a bounded estimating function having bounded derivative with respect to the parameter of interest. Second, we show that second-order robustness implies, at the same time, (i) a robust second-order bias (second-order B-robustness), (ii) a robust asymptotic variance functional (V-robustness) and (iii) a robust saddlepoint approximation functional to the estimator's finite sample density. Third, we introduce a new class of second-order robust M-estimators and find that their estimating function can be redescending. Finally, we study the finite sample

properties of second-order robust estimators, by Monte Carlo simulation and in a risk management application to the estimation of the tail of maximal losses of stock market index returns. Our findings indicate that second-order robust estimators can improve on Maximum Likelihood and first-order robust estimators, in terms of efficiency and robustness, for moderate to small sample sizes and in the presence of deviations from ideal parametric models. Given our general construction of second-order robust M-estimators, these findings indicate the potential usefulness and applicability of second-order robust estimation for a variety of relevant fields and model settings.

1.9 Appendix: assumptions

Root-n consistency and asymptotic normality. In order to ensure the root-n consistency and asymptotic normality of the M-estimators defined in the paper, we rely on the assumptions B1-B4 and N1-N4 in Huber [55, page 129-131].

Saddlepoint density approximation. The assumptions A4.1M-A4.5M in Field and Ronchetti [31, page 62] ensures the existence of the saddlepoint density approximation for $P_{\varepsilon,G} \in \mathcal{U}_\eta(P_0)$.

1.10 Appendix: proofs

In all proofs of this Appendix, derivatives of generic functions $f(\varepsilon)$ evaluated at $\varepsilon = 0$ are denoted by $\partial_\varepsilon f(\varepsilon)$.

Proof of Proposition 1.3.1. Kernels (1.2.2)–(1.2.3) can be applied to compute von Mises expansion (1.2.4). Since for every $k \in \mathbb{N}$ term $T_k(P_{\varepsilon,G} - P_0)$ is a function of ψ and its derivatives with respect to θ up to $(k-1)$-th order (provided that they exist), a bounded estimating function having bounded derivatives up to $(k-1)$-th order implies that every summand in (1.2.4) can be uniformly bounded by a single constant for all $G \in \mathcal{M}$. Setting $k = 2$ concludes the proof.

Proof of Proposition 1.3.2. Let ψ be an estimating function satisfying conditions (1.3.3) . The asymptotic variance functional at $P_{\varepsilon,G}$ is given by:

$$V(\psi; P_{\varepsilon,G}) = M(\varepsilon)^{-1} Q(\varepsilon) M(\varepsilon)'^{-1},$$

where

$$\begin{aligned} M(\varepsilon) &:= [M_{ij}(P_{\varepsilon,G})]_{1\leq i,j\leq p} := E_{P_{\varepsilon,G}}\left[-\nabla_{\theta'}\psi(Z; T(P_{\varepsilon,G}))\right] , \\ Q(\varepsilon) &:= [Q_{ij}(P_{\varepsilon,G})]_{1\leq i,j\leq p} := E_{P_{\varepsilon,G}}[\psi(Z, T(P_{\varepsilon,G}))\psi'(Z, T(P_{\varepsilon,G}))] . \end{aligned}$$

Boundedness of $\partial_\varepsilon V(\psi; P_{\varepsilon,G})$ in G is equivalent to the boundedness of $\partial_\varepsilon M(\varepsilon)$ and $\partial_\varepsilon Q(\varepsilon)$. Therefore, we just need to compute these (functional) derivatives and show that they are bounded under the given assumptions. We first have:

$$\begin{aligned} \partial_\epsilon M_{ij}(P_{\varepsilon,G}) &= \partial_\epsilon E_{P_{\varepsilon,G}}\left[-\partial_{\theta_j}\psi^{(i)}(Z; T(P_{\varepsilon,G}))\right] \\ &= -\left\{E_G[\partial_{\theta_j}\psi^{(i)}(Z; T(P_0))] + E_{P_0}[\partial_{\theta_j}\psi^{(i)}(Z; T(P_0))]\right\} \\ &\quad -E_{P_0}[\nabla_{\theta'}\partial_{\theta_j}\psi^{(i)}(Z, T(P_0))]E_G[IF(Z; \psi; \theta_0)]. \end{aligned} \tag{1.10.1}$$

This derivative is bounded in G since $\partial_{\theta_j}\psi^{(i)}$ is bounded for each $1 \leq i, j \leq p$ and $IF(\cdot; \psi; \theta_0)$ is proportional to bounded estimating function ψ. Moreover,

$$\begin{aligned} \partial_\epsilon Q_{ij}(P_{\varepsilon,G}) &= \partial_\epsilon E_{P_{\varepsilon,G}}\left[\psi^{(i)}(Z; T(P_{\varepsilon,G}))\psi^{(j)}(Z; T(P_{\varepsilon,G}))\right] \\ &= E_G[\psi^{(i)}(Z; T(P_0))\psi^{(j)}(Z; T(P_0))] - E_{P_0}[\psi^{(i)}(Z; T(P_0))\psi^{(j)}(Z; T(P_0))] \\ &\quad +E_{P_0}[\nabla_{\theta'}(\psi^{(i)}(Z; T(P_0)) + \psi^{(j)}(Z; T(P_0)))]E_G[IF(Z; \psi; \theta_0)]. \end{aligned} \tag{1.10.2}$$

This derivative is bounded in G since the estimating function ψ is bounded. This concludes the proof.

Proof of Proposition 1.3.3. Without loss of generality we consider $p = 1$. The proof for the multivariate case is similar, but requires a more involved notation. In the following we assume that G in $P_{\varepsilon,G}$ is: (i) absolutely continuous w.r.t. the Lebesgue measure, with Radon-Nykodym derivative g; (ii)

such that the implied M-functional $T_n(P_{\varepsilon,G})$ admits a finite sample density $f(t;n,\varepsilon,G)$; (iii) such that the assumptions $S1-S3$ hold under $P_{\varepsilon,G}$, thus the saddlepoint in Eq. (1.3.15) exists under $P_{\varepsilon,G}$.

Proof of (i): The saddlepoint density approximation functional $g(t;n,\varepsilon,G)$ is explicitly given by:

$$g(t;n,\varepsilon,G) = (n/2\pi)^{1/2} c^{-n}(\alpha(t;P_{\varepsilon,G});P_{\varepsilon,G}) \left|\tilde{M}(t;P_{\varepsilon,G})\right| \tilde{Q}^{-1/2}(t;P_{\varepsilon,G}) \ .$$

For brevity, we introduce the following simplified functional notation:

$$\alpha(\varepsilon) = \alpha(t;P_{\varepsilon,G}) \ , \ \tilde{Q}(\varepsilon) = \tilde{Q}(t;P_{\varepsilon,G}) \ , \ \tilde{M}(\varepsilon) = \tilde{M}(t;P_{\varepsilon,G}) \ , \ c(\varepsilon) = c(\alpha(t;P_{\varepsilon,G});P_{\varepsilon,G}) \ .$$

Given the smooth dependence of $g(t;n,\varepsilon,G)$ on these functionals, in order to show that $\partial_\varepsilon g(t,n,\varepsilon,G)$ is bounded in G, it is sufficient to show that the derivatives $\partial_\varepsilon \alpha(\varepsilon)$, $\partial_\varepsilon c^{-1}(\varepsilon)$, $\partial_\varepsilon \tilde{M}(\varepsilon)$ and $\partial_\varepsilon \tilde{Q}(\varepsilon)$ are bounded in G. In the sequel, we compute these derivatives. First, the saddlepoint functional is defined as the solution of the implicit equation:

$$E_{P_{\varepsilon,G}}\left[\exp\{\alpha(t;P_{\varepsilon,G})\psi(Z;t)\}\psi(Z;t)\right] = 0 \ . \tag{1.10.3}$$

Therefore, implicit differentiation of this equation immediately yields:

$$\partial_\varepsilon \alpha(\varepsilon) = -E_{P_0}\left[\exp\{\alpha(t;P_0)\psi(Z;t)\}\psi^2(Z;t)\right]^{-1} E_G\left[\exp\{\alpha(t;P_0)\psi(Z;t)\}\psi(Z;t)\right] \ . \tag{1.10.4}$$

This derivative is bounded in G, provided that estimating function ψ is bounded. Second, we obtain, using the saddlepoint property (1.10.3) for $P_{\varepsilon,G} = P_0$:

$$\begin{aligned}
\partial_\varepsilon c^{-1}(\varepsilon) &= E_G[\exp\{\alpha(t;P_0)\psi(Z;t)\}] - E_{P_0}[\exp\{\alpha(t;P_0)\psi(Z;t)\}] \\
&\quad + E_{P_0}[\exp\{\alpha(t;P_0)\psi(Z;t)\}\,\psi(Z;t)]\partial_\varepsilon\alpha(\varepsilon) \\
&= E_G[\exp\{\alpha(t;P_0)\psi(Z;t)\}] - E_{P_0}[\exp\{\alpha(t;P_0)\psi(Z;t)\}] \ .
\end{aligned} \tag{1.10.5}$$

This derivative is bounded in G, provided that estimating function ψ is bounded. Third, we get:

$$\begin{aligned}
\partial_\varepsilon \tilde{Q}(\varepsilon) &= \partial_\varepsilon E_{P_{\varepsilon,G}}[\psi^2(Z;t)\exp\{\alpha(t,P_{\varepsilon,G})\psi(Z,t)\}c(\alpha(t,P_{\varepsilon,G}),P_{\varepsilon,G})] \\
&= \partial_\varepsilon c(\varepsilon)\frac{\tilde{Q}(0)}{c(0)} + E_{P_0}[\psi^3(Z;t)\exp\{\alpha(t;P_0)\psi(Z;t)\}c(0)]\partial_\varepsilon\alpha(\epsilon) \\
&\quad + E_G[\psi^2(Z;t)\exp\{\alpha(t;P_0)\psi(Z;t)\}c(0)] - \tilde{Q}(0) \ .
\end{aligned}$$

This derivative is bounded in G, provided that the estimating function ψ is bounded, since by identities (1.10.4) and (1.10.5) the derivatives $\partial_\varepsilon \alpha(\epsilon)$ and $\partial_\varepsilon c^{-1}(\varepsilon) = -\partial_\varepsilon c(\varepsilon)/c(0)^2$ are bounded. Finally, we have:

$$\begin{aligned} -\partial_\varepsilon \tilde{M}(\varepsilon) &= \partial_\varepsilon E_{P_{\varepsilon,G}} \left[\partial_\theta \psi(Z;t) \exp\left\{\alpha(t;P_{\varepsilon,G})\psi(Z;t)\right\} c(\alpha(t;P_{\varepsilon,G}),P_{\varepsilon,G})\right] \\ &= \partial_\varepsilon c(\varepsilon)\frac{\tilde{M}(0)}{c(0)} + E_{P_0}\left[\partial_\theta \psi(Z;t)\psi(Z;t)\exp\left\{\alpha(t;P_0)\psi(Z;t)\right\} c(0)\right]\partial_\epsilon \alpha(\epsilon) \\ &\quad + E_G[\partial_\theta \psi(Z;t)\exp\left\{\alpha(t;P_0)\psi(Z;t)\right\} c(0)] - \tilde{M}(0). \end{aligned} \tag{1.10.6}$$

Similarly to the above arguments, this derivative is bounded in G, provided that both estimating function ψ and its derivative $\partial_\theta \psi$ are bounded.

Overall, we obtain that an estimating function satisfying condition (1.3.3) implies a bounded derivative for each term appearing in the definition of $g(t;n,\varepsilon,G)$. Consequently, given the smooth dependence of $g(t;n,\varepsilon,G)$ on all these terms, we conclude $|\partial_\varepsilon g(t;n,\varepsilon,G)| < \infty$, for all G.

Proof of (ii): The statement for the functional $a_1(t;P_{\varepsilon,G})$ is obtained along similar lines as for statement (i), by noting that a_1 is a differentiable function of the $P_{\varepsilon,G}$-expectation of functions $\partial\psi/\partial\theta$, ψ^2, $\partial\psi/\partial\theta\psi$ and $\partial\psi/\partial\theta\psi^2$ under the ε, G-contaminated conjugate measure $H_{\varepsilon,G;t}$; see, for instance, equation (4.1) in Field and Hampel [30]. Under the given assumptions, these functions are all bounded functions of $z \in \mathcal{Z}$. Moreover, the density $dH_{\varepsilon,G;t}/dP_{\varepsilon,G}$ is a bounded function of $z \in \mathcal{Z}$ that features also a bounded dependence with respect to G, because $\partial_\varepsilon \alpha(\epsilon)$ is a bounded function of G, as was shown in (i). This concludes the proof.

Proof of Proposition 1.4.1. Consider the class Ψ of functions $\psi(z-\theta_0)$, $\Psi = \{\psi | \psi : \mathcal{Z}\times\mathbb{R} \to \mathbb{R}, \quad E_{P_0}[\psi(Z-\theta_0)] = 0\}$. Without loss of generality, we write the estimating functions in the class Ψ in canonical form, namely each ψ is such that $E_{P_0}[-\partial_\theta \psi(z;\theta_0)] = 1$. The last standardization implies that $\|IF\| = \|\psi\|$. We consider the following minimization problem

$$\min_{\psi\in\Psi} V(\psi;\theta_0), \quad s.t. \quad \sup_{z\in\mathcal{Z}} |\psi(z;\theta_0)| \le b \quad and \quad \sup_{z\in\mathcal{Z}} |\partial_\theta \psi(z;\theta_0)| \le Ac, \tag{1.10.7}$$

for $b \in \mathbb{R}^+$, where A is such that $E_{P_0}[-\partial_\theta \tilde{\psi}_{b,c}] = A^{-1}$, with $\tilde{\psi}_{b,c}$ defined in Eq. (1.4.2). From $E_{P_0}[\psi(Z-\theta_0)] = 0$, it follows $E_{P_0}[\psi(Z-\theta_0)s(Z-\theta_0)] = E_{P_0}[-\partial_\theta \psi(Z-\theta_0)] = 1$. Then, in short notation, we have

$$E_{P_0}\left\{[A(s-\tau)-\psi]^2\right\} = \delta + V(\psi;\theta_0) - 2A,$$

where $\delta := E_{P_0}\left[A^2(s-\tau)^2\right]$ is a constant independent of ψ. Therefore, solving (1.10.7) is equivalent to solving

$$\min_{\psi\in\Psi} E_{P_0}\left[A(s-\tau)-\psi\right]^2, \quad s.t. \quad \sup_{z\in\mathcal{Z}} |\psi(z;\theta_0)| \leq b \quad and \quad \sup_{z\in\mathcal{Z}} |\partial_\theta \psi(z;\theta_0)| \leq Ac.$$

Since $|\partial_z \psi| = |\partial_\theta \psi|$, the solution to this optimization problem can be obtained by minimizing pointwise the following Lagrangian function

$$\Lambda = |\psi - A(s-\tau)|^2 - \lambda\left(|\psi|^2 - b^2\right) - \mu\left(|\partial_z\psi|^2 - A^2c^2\right). \tag{1.10.8}$$

For $c \geq 1$, the solution to (1.10.8) is the Huber M-estimator for location implied by $\psi_b(z-\theta_0) = A(z-\theta_0 - \tau(\theta_0))\min(1; b/|A(z-\theta_0-\tau(\theta_0))|)$. Additionally, from symmetry considerations, it follows that $\tau(\theta_0) \equiv 0$, so that the solution to (1.10.8) is $\psi_b = A\tilde{\psi}_{b,c}$. The absolute value of the derivative $|\partial_z\psi|$ can be either 0 (for $|\psi_b|^2 = b$) or A (for $|\psi_b|^2 < b$). Thus, the constraint $|\partial_z\psi_b|^2 \leq A^2c^2$ is never binding, when $c \geq 1$.

For $c \in (0,1)$, we consider a class of ψ-functions having derivative bounded by Ac. Thus for z values such that the constraint is binding, we obtain $|\partial_z\psi(z-\theta_0)| = Ac$. This implies that $\psi(z-\theta_0) = Ac(s-\kappa(\theta_0)) = Ac(s-\tau(\theta_0))$, where $\tau(\theta_0)$ preserves the Fisher consistency, i.e., $E_{P_0}[\psi(z-\theta_0)] = 0$. To take into account the bound on the first-order kernel, we can distinguish two cases. (i) $|\psi|^2 < b^2$, then the first constraint in (1.10.8) is not binding and the solution to minimization problem is $\psi(z-\theta_0) = Ac(s-\tau(\theta_0))$. This solution holds for all z such that $|(s-\tau(\theta_0))| \leq b/(Ac)$. (ii) $|\psi|^2 = b^2$, i.e. $|(s-\tau(\theta_0))| > b/(Ac)$. Therefore, $\psi = b$ for $(s-\tau(\theta_0)) > b/(Ac)$ and $\psi = -b$ for $(s-\tau(\theta_0)) < -b/(Ac)$. As a result, the solution to (1.10.8) has form $\psi_{b,c}(z-\theta_0) = Ac(s-\tau(\theta_0))\min\left(1; b/|Ac(s-\tau(\theta_0))|\right)$. From symmetry considerations, we set $\tau(\theta_0) \equiv 0$. Thus, the solution to (1.10.8) is $\psi_b = A\tilde{\psi}_{b,c}$. This concludes the proof.

Proof of Proposition 1.4.2. Define $\mathcal{Z}(\theta_0) := \{z \in \mathcal{Z} : ||A(\theta_0)(\psi(z;\theta_0) - \tau(\theta_0))|| \leq b\}$, $\mathcal{Z}_1(\theta_0) := \{z : g(z) \leq c\}$ and $v := A(\psi - \tau)$. Using these definitions, we partition the state space $\mathcal{Z}$ into four subsets:

$$S_1 := \mathcal{Z}_1(\theta_0) \cap \mathcal{Z}^c(\theta_0) \ , \ S_2 := \mathcal{Z}_1(\theta_0)^c \cap \mathcal{Z}(\theta_0) \ , \ S_3 := \mathcal{Z}_1(\theta_0) \cap \mathcal{Z}(\theta_0) \ , \ S_4 := \mathcal{Z}_1(\theta_0)^c \cap \mathcal{Z}(\theta_0)^c \ .$$

Since by definition estimating function $\psi_{b,c}(z;\theta_0)$ is bounded by b, we only need to show that its gradient $\nabla_{\theta'}\psi_{b,c}(z;\theta_0)$ is bounded on each set $S_1, \ldots, S_4$.

(i) Set S_1: For $z \in S_1$, $\nabla_{\theta'}\psi_{b,c}$ is given by equation (1.3.9), where $\|v\| > b$. Note that $\lambda_{\min}\|\psi - \tau\| \leq \|v\| \leq \lambda_{\max}\|\psi - \tau\|$, where $\lambda_{\min}, \lambda_{\max}$ are the square roots of the smallest and largest eigenvalues of the symmetric positive definite matrix $A'A$, respectively. Recall that $D_{\theta'}(A, v)$ is proportional to $\psi - \tau$. Therefore, we obtain for some constant $K \geq 0$:

$$\frac{\|D_{\theta'}(A,v)\|}{\|v\|} \leq K\frac{\|\psi - \tau\|}{\|v\|} \leq \frac{K}{\lambda_{\min}} \ . \tag{1.10.9}$$

Moreover, for any $z \in S_1$ and some constant $K_1 \geq 0$:

$$\frac{\|\nabla_{\theta'}\psi\|}{\|v\|} \leq \frac{K_1}{\lambda_{\min}}\frac{g}{\|\psi - \tau\|} \leq \frac{K_1 c\lambda_{\max}}{b\lambda_{\min}} \ . \tag{1.10.10}$$

Overall, this shows that $\nabla_{\theta'}\psi_{b,c}$ given by equation (1.3.9) is bounded for $\|v\| > b$. Therefore, $\nabla_{\theta'}\psi_{b,c}$ is bounded on S_1.

(ii) Set S_2: When $z \in S_2$, explicit computation of $\nabla_{\theta'}\psi_{b,c}$ using (1.3.9) for $\|v\| \leq b$ gives:

$$\nabla_{\theta'}\psi_{b,c} = (D_{\theta'}(A,v) + A(\nabla_{\theta'}\psi - \nabla_{\theta'}\tau))\,\frac{c}{g} \ .$$

Moreover,

$$\|D_{\theta'}(A,v)\|\frac{c}{g} \leq K\|\psi - \tau\| \leq \frac{K}{\lambda_{min}}\|v\| \leq \frac{Kb}{\lambda_{min}} \ .$$

Similarly, $\|\nabla_{\theta'}\psi/g\| \leq K_1$ and $c/g \leq 1$. Overall, this shows that gradient $\nabla_{\theta'}\psi_{b,c}$ is bounded on set S_2.

(iii) Set S_3: For $z \in S_3$, $\|v\| \leq b$ and gradient

$$\nabla_{\theta'}\psi_{b,c}(z;\theta_0) = D_{\theta'}(A,v) + A(\nabla_{\theta'}\psi - \nabla_{\theta'}\tau)$$

is again bounded, because $\|\nabla_{\theta'}\psi\| \leq K_1 g \leq K_1 c$ and

$$\|D_{\theta'}(A,v)\| \leq K\|\psi - \tau\| \leq \frac{K}{\lambda_{\min}}\|v\| \leq \frac{Kb}{\lambda_{\min}} \ .$$

(iv) Set S_4: When $z \in S_4$, $\|v\| > b$ and the gradient is given explicitly by:

$$\nabla_{\theta'}\psi_{b,c} = \left(I - \frac{vv'}{\|v\|^2}\right)\left(D_{\theta'}(A,v) + A(\nabla_{\theta'}\psi - \nabla_{\theta'}\tau)\right)\frac{bc}{g\|v\|} \tag{1.10.11}$$

Moreover, $\|\nabla_{\theta'}\psi\|/(g\|v\|) \leq K_1/b$, $bc/(g\|v\|) \leq 1$ and

$$\|D_{\theta'}(A,v)\|\frac{bc}{g\|v\|} \leq \frac{Kb}{\lambda_{\min}}\frac{c}{g} \leq \frac{Kb}{\lambda_{\min}} \ . \tag{1.10.12}$$

Overall, we obtain that $\nabla_{\theta'}\psi$ is bounded on $\bigcup_{i=1}^4 S_i = \mathcal{Z}$. This concludes the proof.

1.11 Appendix: figures and tables

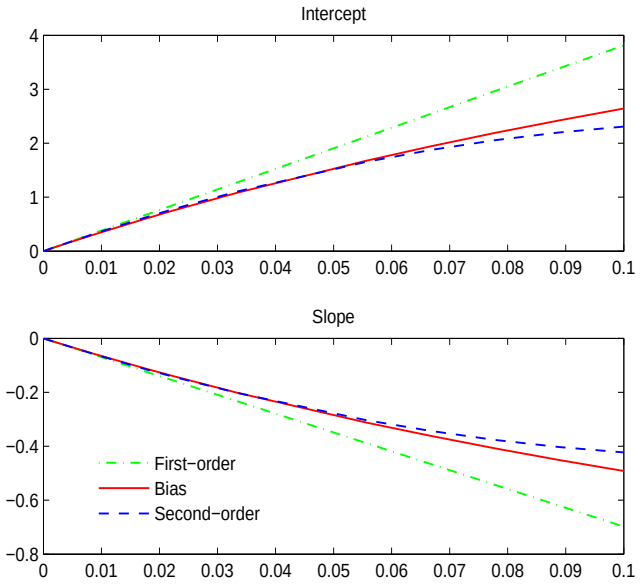

Figure 1.1: True bias (continuous line), first-order bias approximation (dashed-dot line) and second-order bias approximation (dashed line) of LS estimator for the intercept (top panel) and the slope (bottom panel) in the regression model $y = \beta_0 + \beta_1 x + u$, with $x \sim N(0, 10.5)$ and $u \sim N(0, 1)$. The ε contaminated data are in the (x, y)-space. The regressor is contaminated by a Dirac mass at -20.5 and the response variable is contaminated by a Dirac mass at -0.5. The contaminated samples are obtained by replacing an ε-percentage of the observations ranging from 0 to 0.1 in the clean sample. The true parameters are $\beta_0 = 4$ and $\beta_1 = 2$. The Monte Carlo size is 3000, and samples have size $n = 900$. The explicit expression of the bias under the model $P_\varepsilon = (1-\varepsilon)P_0 + \varepsilon\delta_{(\tilde{x_0}, y_0)}$ is given by a simple computation: $T(P_\varepsilon) - T(P_0) = \varepsilon[I + \varepsilon(S_0^{-1}(\tilde{x_0}\tilde{x_0}') - I)]^{-1} S_0^{-1}\tilde{x_0}[y_0 - \tilde{x_0}' S_0^{-1} d_0]$, where $T(P_0) = S_0^{-1} d_0$ is the least squares functional, $S_0 = E_{P_0}[\tilde{x}\tilde{x}']$, $d_0 = E_{P_0}[\tilde{x}y]$, $\tilde{x_0} = (1, x_0)'$ and $\tilde{x} = (1, x)'$.

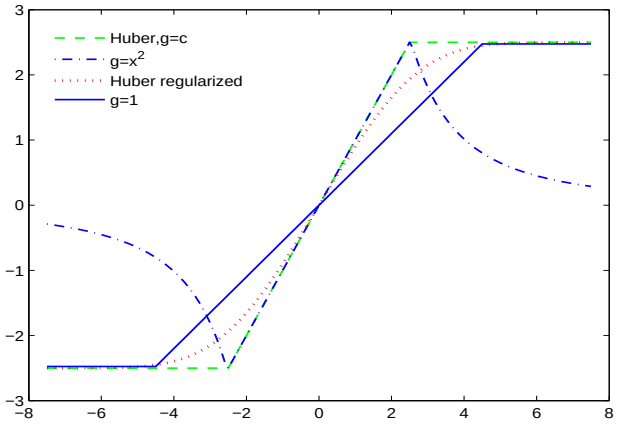

Figure 1.2: Estimating function for symmetric location, with $\theta_0 = 0$. The dashed line is for the Huber M-estimator having $b = 2.5$. This estimator coincides with the second-order robust estimator in (1.4.3) with $\mathrm{g}(z) = c$. The dash-dotted line represents the second-order robust estimator in (1.4.3), with $\mathrm{g}(z) = z^2$, $b = 2.5$ and $c = 6.5$. The continuous line corresponds to (1.4.3) with $g(z) = 1$, $b = 4.545$ and $c = 0.55$. For illustration purposes, we set $A = 1$ The dotted line corresponds to regularized first-order M-estimator in (1.3.5), with $V = 2.046$.

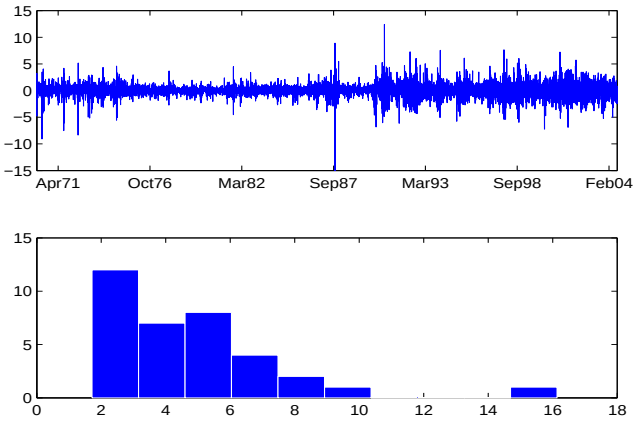

Figure 1.3: Top panel: Time series of daily returns of the Nikkey 225 index. Bottom panel: histogram for the maximal losses observed in the 35 non-overlapping sub-samples, containing the daily returns of the successive calendar year.

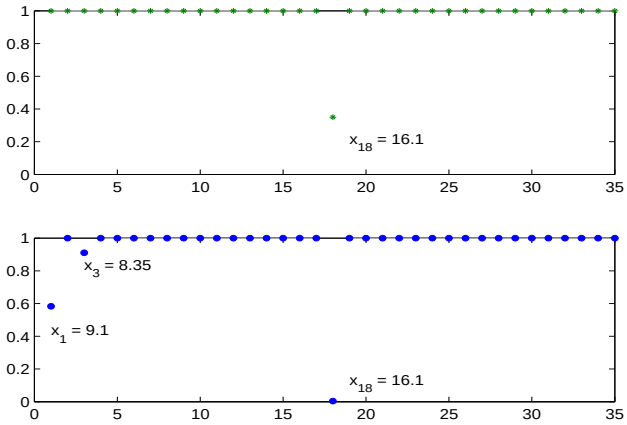

Figure 1.4: Weights for the First-order robust (top panel) and for the Second-order robust (bottom panel) M-estimator of the parameters of the GEV distribution for the maximal losses for Nikkey 225 index returns.

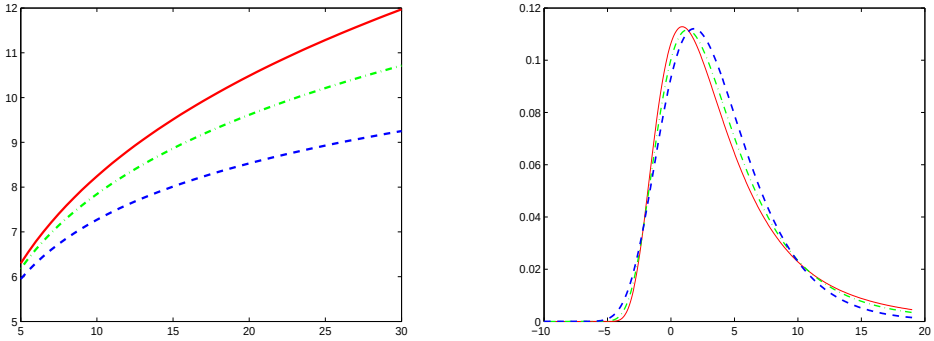

Figure 1.5: Left Panel: Nikkey 225 yearly return level for $k \in [5, 30]$ on the x-axis, obtained using three different estimation methods. Right panel: GEV density implied by the three methods. In both panels, the continuous line represents the MLE, the dash-dotted and the dotted line are for first- and second-order robust M-estimators, respectively.

	$n = 15$	$p = 3$	$n/p = 5$
	MLE	First-Order rob.	Second-Order rob.
Clean	0.253	0.216	0.288
Cont	0.497	0.326	0.301

	$n = 30$	$p = 3$	$n/p = 10$
	MLE	First-Order rob.	Second-Order rob.
Clean	0.109	0.113	0.131
Cont	0.257	0.195	0.122

	$n = 90$	$p = 3$	$n/p = 30$
	MLE	First-Order rob.	Second-Order rob.
Clean	0.036	0.042	0.043
Cont	0.123	0.067	0.041

LTable 1: MSE for MLE, first-order and second-order robust M-estimators for linear regression. Clean samples are generated according to the model (1.6.1), with $x \sim N(0, I)$. Contaminated samples are generated by a replacement model in the (x, y)-space. This contamination model generates: (i) leverage points having large residuals; (ii) leverage points having small residuals; (iii) points with large residuals and no leverage on the factors space. The sample sizes are $n = 15, 30, 90$ and the tuning constants for First-order (Hampel-Krasker) robust estimator is $b = 4$. The second-order robust M-estimator has estimating function as in Eq. (1.4.7), with $b = 4$, $c = 3.5$ and $A(\theta)$ as in Eq. (1.3.7). The Monte Carlo size is 1000.

	$n = 27$	$p = 3$	$n/p = 9$
	MLE	First-Order rob.	Second-Order rob.
Clean	0.107	0.113	0.116
Cont	0.688	0.193	0.137

	$n = 30$	$p = 3$	$n/p = 10$
	MLE	First-Order rob.	Second-Order rob.
Clean	0.093	0.100	0.103
Cont	0.573	0.295	0.138

	$n = 45$	$p = 3$	$n/p = 15$
	MLE	First-Order rob.	Second-Order rob.
Clean	0.059	0.065	0.069
Cont	0.359	0.130	0.075

	$n = 54$	$p = 3$	$n/p = 18$
	MLE	First-Order rob.	Second-Order rob.
Clean	0.048	0.053	0.055
Cont	0.388	0.270	0.103

LTable 2: MSE for different M-estimators of the $GEV(-0.1, 0, 1)$ distribution. Clean samples are generated from a GEV. Contaminated samples are obtained by a replacement model with replacement probability 9% and with contaminations in $z = 15$. The clipping constants are $b = 2.2$ and $c = 15$. The second-order robust M-estimator has estimating function as in Eq. (1.4.3), with $A(\theta)$ as in Eq. (1.3.8). The Monte Carlo size is 2500. The sample sizes are $n = 27, 30, 45, 54$, so that $n/p = 9, 10, 15, 18$.

	MLE	First-order rob.	Second-order rob.
ξ	0.251 (0.195)	0.138 (0.059)	-0.008 (0.103)
μ	3.356 (0.129)	3.326 (0.130)	3.283 (0.098)
σ	1.615 (0.233)	1.712 (0.168)	1.789 (0.098)

LTable 3: Parameter estimates (and standard errors in parenthesis) of the GEV distribution modeling the maximal losses of Nikkei 225 daily returns. The GEV distribution models the maximum daily loss over the 35 yearly sub-samples. Point estimates are obtained using three different estimation methods: ML, first- and second-order robust M-estimators. The tuning constants for robust M-estimators are $b = 3.9$ and $c = 20$.

Chapter 2

On robust estimation via pseudo-additive information measures

2.1 Introduction

Let $\mathcal{F}_\Theta = \{F_t, t \in \Theta \subseteq \mathbb{R}^p\}$, $p \geq 1$ be a family of parametric distributions with densities f_t with respect to the Lebesgue measure and let $\mathcal{G}$ be the class of all distributions G having density g with respect to the Lebesgue measure. Assume f_t and g have support $\mathcal{X} \in \mathbb{R}^k$, $k \geq 1$. Although we focus on continuous variables, our arguments apply to the discrete case as well.

Shannon pioneered the idea of measuring the amount of uncertainty (entropy) inherent in a probability distribution by $H(X) = H(g) = -E_G \log\{g(X)\}$, interpreted as the average gain of information after the actual outcome of X is revealed (e.g., see [19] and references therein). Shannon entropy plays a central role in inference. Classic maximum likelihood estimation is related to it through the minimisation of the Kullback-Leibler divergence

$$D(f_t||g) = E_G \log\{g(X)/f_t(X)\} = H(f_t||g) - H(g), \tag{2.1.1}$$

where only $H(f_t||g) = -E_G \log\{f_t(X)\}$ depends on t, whereas the second term does not affect minimisation [2]. Given independent observations $X_1, \ldots, X_n$ from a certain $G \in \mathcal{G}$, one can approximate G by the empirical distribution G_n and minimise the negative log-likelihood function $-E_{G_n} \log\{f_t(X)\}$. If the model

is correctly specified, i.e., $G = F_\theta \in \mathcal{F}_\Theta$ for some $\theta \in \Theta$, the law of large numbers yields a consistent M-estimator with well-known optimality. In the presence of model mis-specifications, however, such a procedure is not trustworthy, as each observation receives the same weight, whether or not it agrees with the assumed model. Classic infinitesimally robust statistics deals with this issue by imposing Huber-type weights, which reduce the impact of anomalous data on the log-likelihood score and result in M-estimators with bounded estimating functions [50].

Another stream of literature explored the achievement of robustness by changing the notion of divergence between f_t and g. [22] studied the asymptotic stability of minimum distance estimators for multivariate location/scatter. [10] put forward an estimator based on minimisation of the power-density divergence, a Bregman-type divergence which includes (2.1.1) and ℓ_2-distance, or integrated square error, as special cases. [75] investigates the behavior of the ℓ_2-distance for parameter estimation. In a different direction, much attention has been devoted to the class of power-divergences defined by

$$D_q(f_t||g) = -\frac{1}{q} E_G L_q \left\{ \frac{f_t(X)}{g(X)} \right\} = -\frac{1}{q} \int_{\mathcal{X}} L_q \left\{ \frac{f_t(x)}{g(x)} \right\} g(x) dx, \tag{2.1.2}$$

where

$$L_q(u) = (u^{1-q} - 1)/(1-q) \tag{2.1.3}$$

if $q \neq 1$ and $L_q(u) = \log(u)$ if $q = 1$. The class (2.1.2) includes notable divergences as special cases, such as the Kullback-Leibler divergence ($q \to 1$) and twice the Hellinger distance ($q = 1/2$). [20] considered (2.1.2) in the context of goodness-of-fit testing. [12] first introduced an estimator based on minimisation of Hellinger distance, which affords a large fraction of bad data, yet maintains full efficiency. [61] and [11] extended Beran's approach to the case $q \neq 1/2$. Current approaches to empirical minimisation of (2.1.2), however, require kernel density estimation with non-trivial complications in multivariate problems: bandwidth selection is difficult and the accuracy of the estimator rests on the convergence of the kernel density smoother, which suffers from the curse of dimensionality.

Similarly to the relationship between the Kullback-Leibler and Shannon entropy in (2.1.1), we show that

power-divergences are closely related to a generalized class of information measures, sometimes referred to as q-entropies, defined by $H_q(X) = H_q(g) = -E_G L_q\{g(X)\}$ [51, 78]. We write $H_q(f_t||g) = -E_G L_q\{f_t(X)\}$ when the argument of expectation involves f_t. For two independent variables, H_q uniquely satisfies the pseudo-additivity rule $H_q(X_1, X_2) = H_q(X_1) + H_q(X_2) + (1-q)H_q(X_1)H_q(X_2)$, which relaxes the usual additivity assumption underlying Shannon information. Note that $L_q(\cdot) \to \log(\cdot)$, as $q \to 1$, recovering the fully additive Shannon entropy. We show that minimisation of (2.1.2) is equivalent to minimising $H_q(f_t||g)$, conditional to a power-transformation on f_t needed to preserve Fisher-consistency (see §2). The key advantage of working with $H_q(f_t||g)$ instead of $D_q(f_t||g)$, however, is that the former can be easily estimated from data averages. This motivates an estimator based on minimisation of the empirical q-entropy

$$-E_{G_n} L_q\{f_t(X)\} = -n^{-1}\sum_{i=1}^{n} L_q\{f_t(X_i)\}. \tag{2.1.4}$$

Differently from other classic minimisers of (2.1.2), this approach avoids the difficulties of kernel smoothing, thus allowing for the treatment of cases where $\dim(\mathcal{X})$ is moderate or large. Moreover, for common models, the tilting transformation needed to obtain Fisher-consistency is usually simple and fully analytic. For the practitioner, this is a clear advantage over other robust methods, for which "re-centering " is usually burdensome when $\dim(\mathcal{X})$ and $\dim(\Theta)$ are large (e.g. see [76], Section 2.4).

Finally, we establish connections between q-entropy minimisation and traditional literature on robustness in terms of infinitesimal stability properties. In principle, any value of the tuning constant q is admissible. Typically, however, only values $q < 1$ are useful for robust estimation in the presence of bad data, as they translate into re-descending estimators with stable asymptotic bias and variance (§§2.3, 2.4). The constant q controls the trade-off between efficiency and robustness: when $q \to 1$, we obtain maximum likelihood; if q is set far from 1, the estimator gains robustness. Here we focus our analysis on values $0 < q < 1$, which preserve the convexity of L_q and ensure a wide range of efficiency/robustness combinations. In §2.4, we devise an approximation of the mean squared error under ϵ-contamination based on a multi-parameter expression of the change of variance function. Variance stability as given by the change of variance function

has been studied only in one-parameter location and/or scale problems for M-estimators [50, 33]. We extend those results to general estimation problems in order to provide a tool for the selection of the tuning constant q by a min-max approach.

2.2 Link between power divergences and q-entropies

We consider the family of power divergences of f_t with respect to g, defined in (2.1.2). If $q \to 1$, $L_q(\cdot) \to \log(\cdot)$, the additivity of $\log(\cdot)$ implies $D_1(f_t||g) = H_1(f_t||g) - H_1(g||g)$ and minimisation on Θ depends only on Shannon entropy $H_1(f_t||g) = -E_G \log\{f_t(X)\}$. Hence, estimation of the integral $H_1(f_t||g)$ can be done by appealing to the law of large numbers. In particular, if G is replaced by G_n, minimisation of the entropy yields the maximum likelihood estimator. For $q \neq 1$, however, $L_q(\cdot)$ is non-additive. Hence, we cannot proceed as for $q = 1$ and g needs to be replaced by some non-parametric estimate with the drawbacks previously described. Next, we point out an alternative strategy.

Lemma 2.2.1. *Let $f^{(\alpha)}(x) = f^{\alpha}(x)/\int f^{\alpha}(x)dx$, $\alpha > 0$. Then, for all f, $g \in \mathcal{G}$, such that $(\int g(x)^{1/q})dx < \infty$,*

$$D_q(f||g^{(1/q)}) = q^{-1}\left\{E_G g(X)^{1/q-1}\right\}^{-q}\left\{H_q(f||g) - H_q(g^{(1/q)}||g)\right\}. \tag{2.2.1}$$

Lemma 2.2.1 shows that up to a constant not depending on t, $D_q(f_t||g^{(1/q)})$ can be split into the difference between the q-entropy for f_t and that for $g^{(1/q)}$. Moreover, for any f_t, $g \in \mathcal{G}$, $D_q(f_t||g) \geq 0$, where $D_q(f_t||g) = 0$ is attained if and only if $f_t = g$ almost everywhere (see Proof of Proposition 2.2.2). Such a discrimination property combined with Lemma 2.2.1 implies Fisher-consistency of the minimiser of $H_q(f_t||g)$, when the latter is properly rescaled.

Equation (2.2.1) implies that minimisation of $H_q(f_t||g)$ and $D_q(f_t||g^{(1/q)})$ are equivalent. Let θ and $\theta^* = T_q^*(G)$ be the minimisers of $D_q(f_t||g)$ and $H_q(f_t||g)$, respectively, whereas t is a generic element of Θ. Throughout the paper, we assume uniqueness of θ and θ^*. Occasionally, the minimiser of the q-entropy, θ^*, will be referred to as the surrogate parameter.

Proposition 2.2.2. *Let $\tau_\alpha : \Theta \mapsto \Theta$ be a continuous mapping on a compact Θ such that $\tau_\alpha(t) = \{t' \in \Theta : f_{t'}(x) = f_t^{(\alpha)}(x)\}$ and assume $f_t \in \mathcal{F}_\Theta$ is such that $\int f_t(x)^{1/q}dx < \infty$ for all $t \in \Theta$. Then, $\tau_q(T_q^*(F_\theta)) = \theta$.*

To clarify the role played by $\tau_q(\cdot)$, consider differentiating $H_q(f_t||g)$ under F_θ. If differentiation and integration can be exchanged, $\nabla_t H_q(f_t||f_\theta) = -\int \nabla_t f_t(x) f_\theta(x) f_t(x)^{-q} dx$. If t is such that $f_t(x) = f_\theta(x)^{1/q} / \int f_\theta(x)^{1/q} dx$, or equivalently $t = \tau_q^{-1}(\theta) = \tau_{1/q}(\theta)$,

$$\nabla_t H_q(f_t||f_\theta) = -c_q(\theta) \int \nabla_t f_t(x) dx = -c_q(\theta) \nabla_t \int f_t(x) dx = 0, \tag{2.2.2}$$

where $c_q(\theta) = \{\int f_\theta^{1/q}(x) dx\}^q$, i.e., $\tau_q^{-1}(\theta)$ is the root of $\nabla_t H_q(f_t||f_\theta) = 0$. Note that $\tau_q(t)$ is simply the parameter of the density proportional to $f_t^q(x)$, which is usually straightforward to compute analytically. For the univariate exponential distribution with density $f_t(x) = t \exp\{-tx\}$, $x > 0$, $t > 0$, we have $\tau_q(t) = qt$. For the multivariate normal distribution with mean vector μ and covariance matrix Σ, we have $\tau_q(\mu^T, \text{vech}^T \Sigma) = (\mu^T, q^{-1} \text{vech}^T \Sigma)$. Lemma 2.2.1 points out the role played by the transformation $g^{(1/q)}$, which is the target density when minimising $H_q(f_t||g)$. Sometimes, $g^{(1/q)}$ is called zooming transformation of g as it enhances certain parts of g. If $0 < q < 1$, parts of g with larger density values are emphasized. Such values of q reduce the importance of tails – which are usually most severely affected by the presence of contamination – and increase the relevance of the majority of the data instead. Setting $q > 1$ increases small density values, which is helpful for tail inference [29].

2.3 Parameter estimator, asymptotics and link with other procedures

2.3.1 A fully parametric approach to q-entropy minimisation

The considerations in §2.2 motivate the following estimation strategy. Let $X_1, \ldots, X_n$ be independent observations from G and define the estimator

$$\hat{\theta}_{q,n} = T_q(G_n) = \tau_q \left\{ T_q^*(G_n) \right\} = \tau_q \left[\text{argmin}_{t \in \Theta} - \frac{1}{n} \sum_{i=1}^n L_q\{f_t(X_i)\} \right]. \tag{2.3.1}$$

In general, computing (2.3.1) involves two steps: (i) Solving the estimating equations

$$\sum_{i=1}^n u_q(X_i, t) = \sum_{i=1}^n u(X_i, t) f_t(X_i)^{1-q} = 0, \tag{2.3.2}$$

where $u(x,t) = \nabla_t \log\{f_t(x)\}$ is the usual maximum likelihood score function ans (ii) since the solution of (2.3.2), $T^*(G_n)$, is consistent for $\tau_q^{-1}(\theta)$, we rescale it as $\hat{\theta}_{q,n} = \tau_q(T^*(G_n))$. Eq. (2.3.2) outlines the self-weighting nature of this estimator: the usual score function receives weights depending on the model itself and q. Particularly, if $0 < q < 1$, extreme observations in the tails receive small weights. Different values of q in Eq. (2.3.2) change the trade-off between robustness and efficiency, thus characterizing the impact of anomalous observations. Setting $q \to 1$ gives the maximum likelihood (equal weights), whereas $q = 1/2$ gives empirical minimisation of Hellinger distance between $f_t^{(1/2)}$ and $g^{(1/2)}$.

Note that some care is needed in order to choose Θ, the family $\mathcal{F}_\Theta$ and a range for q. Existence of (2.3.2) is typically ensured by imposing standard regularity conditions on $E_{G_n} L_q\{f_t(X_i)\}$ and $H(f_t||g)$, which are usually needed for consistency of M-estimators. A possible set of conditions are compactness of Θ, uniqueness of $\theta^* \in \Theta$ and existence of an integrable function dominating $L_q\{f_t(x)\}$ for all $t \in \Theta$ (see [81]).

2.3.2 Relationship with other robust estimators

The strategy of setting weights proportional to the assumed model has appeared for different reasons in various contexts. The procedure described here shares some of the appealing features of the minimum power density divergence estimator put forward by [10]. Both approaches are fully parametric, as they do not require kernel smoothing and are applicable to a wide range of models. When F_θ is a location family, the estimation equations (2·4) p.551, in [10] are basically the same as our (2.3.2) for $\alpha = 1 - q$. In general, however, the two methods rely on two different families of divergences, which overlap only for the special case of the Kullback-Leibler divergence: [10] consider a Bregman-type divergence which generalizes the ℓ_2-distance; instead, our information-theory approach leads to a generalization of the Hellinger distance. Consequently, outside the location family, the trade-off between efficiency and robustness is not necessarily the same for the two estimators and depends both on the form of F_θ and the degree of contamination. This is illustrated by an Example in §2.4.3. Finally, the approach of [10] preserves the Fisher consistency using the typical re-centering of the estimating function, by computing $\int f_t(x)^c u(x,t)dx$, $c > 0$. In practice, the

computation of this quantity can be cumbersome, especially when considering multivariate models with many parameters. Instead, our re-centering by the zooming transformation from Lemma 2.2.1 offers a convenient escape from this. Tilting by zooming is typically fully analytic, thus easier to compute than the re-centering of other classic robust estimators, which require numerical integration.

Our approach is broad, but in some specific instances coincides with known re-descending M-estimators of location and scatter, say μ and Σ, of an elliptic density $\phi(s)$ with $s = (x-\mu)^T\Sigma^{-1}(x-\mu)$. In this setting, the solutions of the weighted likelihood (2.3.2) take the form

$$\hat{\mu} = E_{G_n} w(S)X, \quad \hat{\Sigma} = E_{G_n} w(S)(X-\mu)(X-\mu)^T \tag{2.3.3}$$

where $w(s) = \phi(s)^{1-q}/E_{G_n}\phi(S)^{1-q}$. [58] put forward constrained M-estimates by minimising $E_G\rho_k(S) + \log\{\det(\Sigma)\}/2$, subject to $E_G\rho_k(S) \leq \epsilon\rho_k(\infty)$. When ρ_k is the exponentially weighted function $\rho_k(s) = 1-\exp\{(-ks)\}$ and ϕ is the multivariate normal density, minimising (2.3.2) with $k = 1-q$ is equivalent to their approach for estimating μ. Their scatter matrix estimate, however, differs from ours as their weights in (2.3.3) take form $w(s) = 2k\exp\{(-ks)\}$. The different weight specification plays a relevant role in the trade-off between robustness and efficiency: while the re-descending estimator in (2.3.3) can be more efficient, Kent and Tyler's estimator allows for more pronounced control on the gross-error sensitivity (see §2.4.3).

The estimator defined in (2.3.1) is related to the robustification procedure proposed by [88], where the model-based re-weighting is applied to a general estimating function and the procedure can be interpreted as a generalized method of moments. Here, we focus on the particular case where the estimating function is actually the score function. [17], identify a tilting procedure of the likelihood equations, which, in a special case, corresponds to Eq. (2.3.2). These robustification strategies are introduced as direct manipulations of a set of estimating equations. Here, we provide an information theory justification for those re-weighting schemes.

2.4 Infinitesimal robustness

2.4.1 Asympotics, influence and change-of-variance function

Eq. (2.3.2) defines an M-estimator and asymptotics of $T_q^*(G_n)$ and $T_q(G_n)$ can be treated using existing theory. Define $p \times p$ matrices $K_q(t,G) = E_G u_q(x,t) u_q(x,t)^T$ and $J_q(t,G) = E_G \nabla_t u_q(x,t)$ and write $K_q(t) = K_q(t,F_t)$ and $J_q(t) = J_q(t,F_t)$, if $G = F_t$. One can show that $n^{1/2}\hat{\theta}_{q,n}$ converges to a multivariate normal with mean θ and variance

$$V_q(\theta,G) = \overline{J}_q(\theta,G)^{-1}\overline{K}_q(\theta,G)\overline{J}_q(\theta,G)^{-1} \tag{2.4.1}$$

where $\overline{J}_q(t,G) = \{\nabla_t \tau_q(t)\}^{-1} J_q\{\tau^{-1}(t),G\}$ and $\overline{K}_q(t,G) = K_q\{\tau^{-1}(t),G\}$. In the rest of the paper we also use the notations: $V_t(t) = V_q(t,F_t)$, $\overline{J}_q(t) = \overline{J}_q(t,F_t)$ and $\overline{K}_q(t) = \overline{K}_q(t,F_t)$.

We consider deviations from $\mathcal{F}_\Theta$ defined by the neighborhood $F_\epsilon = (1-\epsilon)F_\theta + \epsilon\delta_x$, $0 \le \epsilon \le 1/2$, where δ_x, $x \in \mathcal{X}$ is Dirac's delta, interpreted as worst-case contamination. In what follows, $T_q(\epsilon) = T_q(F_\epsilon)$ and $V_q(\epsilon) = V_q\{T_q(\epsilon), F_\epsilon\}$ denote the estimating functional and the asymptotic variance evaluated under the mis-specified model, respectively.

A standard calculation shows that the influence function for the original functional $T(\cdot)$ is $IF_q(x,\theta) = \nabla_\theta \tau(\theta) IF_q^*(x,\theta)$, where $IF_q^*(x,\theta) = J_q^{-1}(\theta^*, F_\theta) u_q(x,\theta^*)$ is the influence function for the surrogate functional $T_q^*(\cdot)$. If $0 < q < 1$, the influence function is proportional to $f_{\theta^*}^{1-q}(x) u(x,\theta^*)$, where the term $f_{\theta^*}^{1-q}(x)$ corrects for the unboundedness of the score function, implying a smoothly re-descending estimator. In this case, the estimator has finite bias and is said to be bias-robust (B-robust). For instance, this is the case when F_θ belongs to an exponential family.

The influence function alone does not provide direct information on the stability of the asymptotic variance of $T_q(\cdot)$. To this end, we use the change-of-variance-function for $T_q(\cdot)$ defined as the mapping $CVF_q : \mathcal{X} \times \Theta \mapsto \mathbb{R}^{p\times p}$ such as $\partial\{V_q(\epsilon)\}_{ij}/\partial\epsilon|_{\epsilon=0} = \{CVF_q(x,\theta)\}_{ij}$ $\quad i,j = 1,\ldots,p$ and the change-of-variance sensitivity is $\kappa_q(\theta) = \sup_{x\in\mathcal{X}} trCVF_q(x,\theta)/trV_q(\theta)$. The change-of-variance function measures the

influence of a small amount of contamination on the asymptotic variance and $\kappa_q(\theta)$ represents the worst-scenario variability change under infinitesimal contamination. If $\kappa_q(\theta) < \infty$, the estimator is said to be variance robust (V-robust).

Proposition 2.4.1. *Let u_q as in (2.3.2) and J_q,$\overline{J}_q$,$\overline{K}_q$, V_q $p \times p$ matrices as in (2.4.1). Assume $E_{F_\theta}(u_q)_i < \infty$, $E_{F_\theta}(\nabla_\theta u_q)_{ij} < \infty$, $E_{F_\theta}\frac{\partial}{\partial\theta_k}(\nabla_\theta u_q)_{ij} < \infty$, $i,j,k = 1,\ldots,p$. Then, under the contaminated distribution F_ϵ:*

$$\begin{aligned} CVF(x,\theta) &= \overline{J}_q(\theta)^{-1}\widetilde{K}_q(x,\theta)\overline{J}_q(\theta)^{-1} + \overline{J}_q(\theta)^{-1}\widetilde{K}_q(x,\theta)^T\overline{J}_q(\theta)^{-1} - \widetilde{J}_q(x,\theta)^{-1}\overline{K}_q(\theta)\overline{J}_q(\theta)^{-1} \\ &\quad -\overline{J}_q(\theta)^{-1}\overline{K}_q(\theta)\widetilde{J}_q(x,\theta)^{-T} + IF_q(x,\theta)IF_q(x,\theta)^\top - V_q(\theta), \end{aligned} \tag{2.4.2}$$

where

$$\begin{aligned} \widetilde{K}_q(x,\theta) &= IF_q(x,\theta)E_{F_\theta}u_q(X,\theta)^T\nabla_\theta u_q(X,\theta), \\ \widetilde{J}_q(x,\theta)^{-1} &= \overline{J}_q^{-1}(\theta)\left(D_q(x,\theta)\nabla_\theta\tau_q(\theta)^{-1} + \nabla_{\theta^*}u_q(x,\theta^*)\nabla_\theta\tau_q(\theta)^{-1} - \overline{J}_q(\theta)\right)\overline{J}_q^{-1}(\theta) \end{aligned}$$

and $\{D_q(x,\theta)\}_{i,j} = E_{F_\theta}\sum_{k=1}^p \frac{\partial}{\partial\theta_k}\{\nabla_\theta u_q\}_{i,j}IF_k^*(x)$, with $i,j = 1,\ldots,p$.

The expression in (2.4.2) is a simplified expression holding for the distributions belonging to the exponential family. The general expression for the CVF_q is provided in Appendix. We remark that $E_{F_\theta}IF_q(X,\theta) = 0$ and, if all expectations in (2.4.2) are well-defined, we also have $E_{F_\theta}CVF_q(X,\theta) = 0$, since $E_{F_\theta}IF_q(X,\theta)IF_q(X,\theta)^T = V_q(\theta)$ and for the mixed terms $E_{F_\theta}\widetilde{K}_q(X,\theta) = 0$ and $E_{F_\theta}\widetilde{J}_q(X,\theta)^{-1} = 0$. If $u_q(x,t)$ is replaced by a generic estimating functional, Proposition 2.4.1 provides a useful generalization of known results derived for the one-parameter case. Setting $p = 1$ in (2.4.2) gives formulas matching those in [50] and [33] for univariate scale and location. Finally, from an inspection of the expressions in the above formulas, sufficient conditions for B- and V-robustness are boundedness of u_q and its first and second derivatives. These are satisfied for common families and can be verified on a case-by-case basis.

2.4.2 Worst-case mean squared error and min-max selection of q

In this section, we study the mean squared error of $\hat{\theta}_{q,n}$, under ϵ-contamination. The first-order approximation of the bias for the M-functional $T_q(\cdot)$ is $T_q(\epsilon) - \theta \approx \epsilon\,\partial T_q(\epsilon)/\partial\epsilon|_{\epsilon=0} = \epsilon IF_q(x,\theta)$. The gross-error sensitivity is defined by $\gamma_q(\theta) = \sup_{x\in\mathcal{X}}\|IF_q(x;\theta)\|$ and represents the worst influence of a small amount of contamination on the estimator. If $\gamma_q(\theta) < \infty$, we say that $T_q(\cdot)$ is B-robust. An extrapolation of the

asymptotic variance can be obtained using the approximation

$$trCVF_q(x,\theta)/trV_q(\theta) = \partial \log trV_q(\epsilon)/\partial\epsilon|_{\epsilon=0} \approx \epsilon^{-1}\left[\log\{trV_q(\epsilon)\} - \log\{trV_q(\theta)\}\right].$$

From the above expression, $tr\{V_q(\theta)\}\exp(\epsilon\kappa_q(\theta))$ is an approximate upper bound to $tr\{V_q(\epsilon)\}$. By combining the information about the worst-case bias, $\gamma_q(\theta)$, with the expression above, we obtain an approximated upper bound for the mean squared error:

$$MMSE\,(q,\theta;n,\epsilon) = \epsilon^2\gamma_q(\theta)^2 + n^{-1}tr\{V_q(\theta)\}\exp\ (\epsilon\kappa_q(\theta)), \tag{2.4.3}$$

which generalizes the notion in Hampel et al. (1986) for the function u_q to the multi-parameter situation. Eq. (2.4.3) can be used as a criterion for choosing q. For a given sample size n and a nominal contamination level $0 \leq \epsilon \leq 1/2$, we choose a grid of tuning parameters and compute corresponding estimates. Then, we choose the value minimising $MMSE(q,\hat{\theta}_q;n,\epsilon)$. The selected value of q will automatically take care of the interplay between bias and variance, as a function of the ϵ-contamination and n.

2.4.3 Trade-off between robustness and efficiency

Exponential distribution. Consider estimating an exponential distribution Exp(λ) with density $\lambda\exp(-x\lambda)$, $x > 0$, $\lambda > 0$. A lengthy calculation shows that the asymptotic variance and gross error sensitivity are

$$V_q(\lambda) = \lambda^2\frac{(2q-2-q^2)}{(q-2)^3q^3}, \quad \text{and} \quad \gamma_q(\lambda) = \lambda q\left\{1+\frac{1}{(1-q)^{1/2}}\right\}. \tag{2.4.4}$$

Note that $V_q \to \infty$ and $\gamma_q \to 0$, as $q \to 0$. For small values of q, the worst-case bias under ϵ-contamination is small and the estimator is expected to be remarkably robust. This advantage, however, comes with large efficiency losses compared to maximum likelihood. Conversely, if $q \to 1$, the estimator approaches full efficiency as $V_q \to V_1 = \lambda^2$, i.e., the variance for maximum likelihood, but $\gamma_q \to \infty$. Intermediate choices of q balance those two limit scenarios. The trade-off robustness/efficiency is illustrated in more detail by computing numerically the worst-case mean squared error (2.4.3). In Figure 2.1, we plot the maximal mean squared error ($\lambda = 1$ and $n = 150$) against q for various nominal contamination levels ϵ. For small

contaminations, a fairly ample interval for q, from about $0 \cdot 3$ to $0 \cdot 95$, ensure small errors. Choices of q closer to 1 in that range are preferred, since they imply small efficiency losses. When the contamination level increases, the interval of safe choices for q narrows down and moves away from 1.

Multivariate normal distribution. For estimating the mean μ of a multivariate normal $N_p(\mu, \Sigma)$, the asymptotic variance is $V_q = \{q(2-q)\}^{-p/2-1}\,\Sigma$, when no contamination occurs. In the presence of contamination, both influence and the change-of-variance functions are bounded for $0 < q < 1$. The former exhibits the typical shape of re-descending estimators. In Figure 2.2, we plot the worst-case mean squared error for estimating a mean component of $N_p(0, I)$ for $p = 1, 2, 4$ and 8 when $\epsilon = 0 \cdot 05$ and $n = 100$. We also report the corresponding optimal values of q. When p increases, two simultaneous (and non-obvious) effects occur: the optimal value of the tuning constant q gets closer to 1, and the global maxima of both influence and change-of-variance functions decrease. Figure 2.2 shows that our M-estimator performs well in terms of the maximal mean squared error also when n/p is small. Interestingly, the values of q minimising the worst-case error also correspond to estimators with higher nominal efficiency. In Table 2.6, we report the relative efficiencies (at the model) with respect to the maximum likelihood estimator corresponding to optimal values of q selected by (2.4.3) for $\epsilon = 0 \cdot 05, 0 \cdot 15$ and $n = 100, 1000$.

For estimating μ, the procedure is the same as that of [10] when their tuning parameter is $\alpha = 1 - q$. Thus, the trade-off between robustness and efficiency illustrated above holds equally for both estimators. However, outside the location case this is not true and typically the two estimators yield different efficiencies for a given robustness level. In Table 2.6, we compare the asymptotic relative efficiencies for estimating the scale σ of a univariate normal $N_1(0, \sigma^2)$ of the two estimators for different choices of α and q. For any α, we compute the gross-error sensitivity γ_α of Basu's estimator and set q to yield the same gross-error sensitivity $\gamma_q = \gamma_\alpha$ for our estimator. For estimating σ, we found a pronounced non-linear relationship between $1 - q$ and α. When α is close to zero ($\alpha \leq 0 \cdot 05$), the two estimators have similar efficiency close to maximum likelihood. For $0 \cdot 1 \leq \alpha \leq 0 \cdot 5$, the estimator of [10] does a slightly better job, while choices

$\alpha > 0 \cdot 5$ imply a better trade–off between robustness and efficiency of our estimator. Finally, we considered the estimator of [58], with choice of weights as in §2.3.2. Differently from the two minimum divergence estimators, we did not observe a monotone relationship between k and the efficiency/robustness trade-off. In general, Kent and Tyler's estimator has a more remarked robustness and the gross-error sensitivity gets as low as $0 \cdot 77$. This advantage, however, comes at some expense of efficiency: the minimal efficiency loss compared to maximum likelihood is about 35%. These differences are not surprising and are due to the different nature of the methods. While Kent and Tyler's method is designed to merge global breakdown and infinitesimally robustness properties, ours is a purely infinitesimally robust estimator, which smoothly deviates from maximum likelihood depending on the value of q.

2.5 Appendix: proofs

Proof of Lemma 1

For any $u_1, u_2 > 0$, $L_q(\cdot)$ obeys the following pseudo-additivity rule:

$$L_q(u_1 u_2) = L_q(u_1) + L_q(u_2) + (1-q) L_q(u_1) L_q(u_2). \tag{2.5.1}$$

Let $f = f(x)$ and $h = h(x)$ be two densities with support on R^k, $k \geq 1$. By applying property (2.5.1) to $D_q(f||h)$, we obtain

$$\begin{aligned} -qD_q(f||h) &= \int \left\{ \frac{f^{1-q}-1}{1-q} + \frac{h^{q-1}-1}{1-q} + (1-q)\frac{f^{1-q}-1}{1-q}\frac{h^{q-1}-1}{1-q} \right\} h dx \\ &= \int \left(\frac{h^{q-1}-1}{1-q} + \frac{f^{1-q}h^{q-1} - h^{q-1}}{1-q} \right) h dx \\ &= \int L_q(f)\ h^q dx - \int L_q(h)\ h^q dx. \end{aligned} \tag{2.5.2}$$

Finally, setting $h(x) = g^{(1/q)}(x) = g^{1/q}(x) / \int g^{1/q}(x) dx$ in Equation (2.5.2) gives

$$q \left\{ \int g^{1/q}(x) dx \right\}^q D_q(f||g^{(1/q)}) = \int L_q\{g^{(1/q)}(x)\} g(x) dx - \int L_q\{f(x)\} g(x) dx. \tag{2.5.3}$$

Proof of Proposition 1

Note that $q^{-1}\partial^2 L_q(u)/\partial u^2 < 0$. Thus, by Jensen's inequality we have

$$-\frac{1}{q}\int g(x) L_q\left\{\frac{f(x)}{g(x)}\right\} dx \geq -\frac{1}{q} L_q\left\{\int g(x)\frac{f(x)}{g(x)}dx\right\} = 0. \tag{2.5.4}$$

Therefore, for all $f,\ g \in G,\ D_q(f||g) \geq 0$, where $D_q(f||g) = 0$ is attained if and only if $f = g$ almost everywhere. Particularly, this implies that $D_q(f_{T_q(F_\theta)}||f_\theta) = 0$ is equivalent to $T_q(F_\theta) = \theta$. Let $\theta^* = T^*(F_\theta)$ be the value such that $H_q(f_{\theta^*}||f_\theta) = \min_{t\in\Theta^*} H_q(f_t||f_\theta)$. Lemma 1 implies that θ^* also satisfies

$$D_q(f_{\theta^*}||f_\theta^{(1/q)}) = \min_{t\in\Theta} D_q(f_t||f_\theta^{(1/q)}), \tag{2.5.5}$$

which is zero if and only if

$$f_{\theta^*}(x) = f_\theta^{(1/q)}(x) = f_{\tau_q^{-1}(\theta)}(x). \tag{2.5.6}$$

Therefore, we have $T_q^*(F_\theta) = \theta^* = \tau_q^{-1}(\theta)$, which implies $\tau_q\{T_q^*(F_\theta)\} = \theta$.

Derivation of the influence function

Let $F_\theta \in F_\theta$. By definition, for a generic M-functional $T(\cdot)$, the influence function is obtained by the first Gâteaux derivative in the direction of contamination $F_\epsilon = (1-\epsilon)F_\theta + \epsilon\delta_x$, where δ_x is a Dirac delta function in $x \in \mathcal{X}$. The influence function is defined by

$$\mathrm{IF}(x,\theta) = \lim_{\epsilon\to 0}\frac{T\{(1-\epsilon)F_\theta + \epsilon\delta_x\} - T(F_\theta)}{\epsilon}. \tag{2.5.7}$$

For the surrogate parameter θ^*, $E_{F_\theta}\{u_q(X,\theta^*)\} = 0$ or equivalently $E_{F_\epsilon}(u_q[X,\tau_q^{-1}\{T_q(\epsilon)\}]) = 0$, if $\epsilon = 0$. Implicit differentiation of the last expectation gives

$$\begin{aligned} 0 \quad &= \quad u_q\left[x,\tau_q^{-1}\left\{T_q(\theta)\right\}\right] - E_{F_\theta}\left(u_q\left[X,\tau_q^{-1}\left\{T_q(\theta)\right\}\right]\right) + E_{F_\theta}\left(\frac{\partial u_q\left[X,\tau_q^{-1}\left\{T_q(\epsilon)\right\}\right]}{\partial\epsilon}\right)\Bigg|_{\epsilon=0}. \end{aligned} \tag{2.5.8}$$

The second term in (2.5.8) is zero. Moreover, chain-differentiating the third term gives

$$\begin{aligned} 0 \quad &= \quad u_q\left[x,\tau_q^{-1}\left\{T_q(\theta)\right\}\right] + \nabla_\theta\tau_q^{-1}\left\{T_q(\theta)\right\}\frac{\partial T_q(F_\epsilon)}{\partial\epsilon}|_{\epsilon=0}\ E_{F_\theta}\left(\nabla_\theta u_q\left[X,\tau_q^{-1}\left\{T_q(\theta)\right\}\right]\right) \\ &= \quad u_q\left(x,\theta^*\right) + E_{F_\theta}\left\{\nabla_\theta u_q\left(X,\theta^*\right)\right\}\nabla_\theta\tau_q^{-1}(\theta)\mathrm{IF}_q(x,\theta). \end{aligned}$$

Finally, since $\nabla_\theta \tau_q \nabla_\theta \tau_q^{-1} = I$, or $\nabla_\theta \tau_q^{-1} = (\nabla_\theta \tau_q)^{-1}$, we have

$$\mathrm{IF}_q(x,\theta) = -\nabla_\theta \tau_q(\theta) \left[E_{F_\theta}\{\nabla_\theta u_q(X,\theta^*)\}\right]^{-1} u_q(x,\theta^*) = \nabla_\theta \tau_q(\theta) \mathrm{IF}_q^*(x,\theta), \tag{2.5.9}$$

where we denote influence functions for $T_q(\cdot)$ and $T_q^*(\cdot)$ by $\mathrm{IF}_q(x,\theta)$ and $\mathrm{IF}_q^*(x,\theta)$, respectively.

Proof of Proposition 2

Define the following quantities computed under the contaminated distribution F_ϵ: $u_q(\epsilon) = u_q\{x, T^*(\epsilon)\}$, $J_q^{-1}(\epsilon) = J_q^{-1}\{T^*(\epsilon), F_\epsilon\}$, $\overline{J}_q^{-1}(\epsilon) = \overline{J}_q^{-1}\{T(\epsilon), F_\epsilon\}$. Define analogous quantities computed under the reference model F_θ: $u_q = u_q\{x, T^*(F_\theta)\}$, $\nabla_\theta u_q = \nabla_\theta u_q\{x, T^*(F_\theta)\}$, $J_q^{-1} = J_q(\theta)^{-1}$, $\overline{J}_q^{-1} = \overline{J}_q(\theta)^{-1}$, and $V_q(\epsilon) = V_q(F_\epsilon)$. Moreover, to simplify the notation, we denote the influence functions at x for $T_q(\cdot)$ and $T_q^*(\cdot)$ by $\mathrm{IF}(x,\theta)$ and $\mathrm{IF}^*(x,\theta)$, respectively.

Computing the Gâteaux derivative of $V_q(\epsilon)$ in $\epsilon = 0$ yields

$$\begin{aligned}
&\frac{\partial}{\partial \epsilon} E_{F_\epsilon}\left\{\bar{J}_q^{-1}(\epsilon) u_q(\epsilon) u_q(\epsilon) \bar{J}_q^{-}(\epsilon)\right\}\Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\bar{J}_q^{-1}(\epsilon)\Big|_{\epsilon=0} E_{F_\theta}\left(u_q u_q \bar{J}_q^{-}\right) \\
&+\bar{J}_q^{-1} E_{F_\theta}\left\{\frac{\partial}{\partial \epsilon} u_q(\epsilon)\Big|_{\epsilon=0} u_q \bar{J}_q^{-}\right\} + \bar{J}_q^{-1} E_{F_\theta}\left\{u_q \frac{\partial}{\partial \epsilon} u_q(\epsilon)\Big|_{\epsilon=0} \bar{J}_q^{-}\right\} \\
&+\bar{J}_q^{-1} E_{F_\theta}\left\{u_q u_q \frac{\partial}{\partial \epsilon}\bar{J}_q^{-}(\epsilon)\Big|_{\epsilon=0}\right\} - \overline{J}_q^{-1}\overline{K}_q(\theta)\overline{J}_q^{-} + \overline{J}_q^{-1} u_q u_q^{\overline{J}_q^{-}}.
\end{aligned} \tag{2.5.10}$$

To compute (2.5.10), we first need $\partial \overline{J}_q^{-1}(\epsilon)/\partial\epsilon|_{\epsilon=0}$. Since $\overline{J}_q(\theta) = J_q \nabla_\theta \tau(\theta)^{-1}$, we have

$$\frac{\partial}{\partial \epsilon}\bar{J}_q^{-1}(\epsilon)\Big|_{\epsilon=0} = \nabla_\theta \tau_q(\theta) \frac{\partial}{\partial \epsilon} J_q^{-1}(\epsilon)\Big|_{\epsilon=0} + \frac{\partial}{\partial \epsilon}\nabla_\theta \tau_q\{T(\epsilon)\}\Big|_{\epsilon=0} J_q^{-1}. \tag{2.5.11}$$

Now, we introduce the notation $Q_q(x,\theta) = \partial \nabla_\theta \tau_q(T(\epsilon))/\partial\epsilon|_{\epsilon=0}$, $\tilde{Q}_q(x,\theta) = Q(x,\theta) J_q^{-1}(\theta)$, where $Q_q(x,\theta)$ is a $p \times p$-matrix with entries

$$\{Q_q(x,\theta)\}_{i,j} = \sum_{k=1}^{p} \frac{\partial}{\partial \theta_k}(\nabla_\theta \tau_q)_{i,j}\{\mathrm{IF}(x,\theta)\}_k, \tag{2.5.12}$$

where $\{\mathrm{IF}(x,\theta)\}_k$ is the k-th component of $\mathrm{IF}(x,\theta)$. Write $J_q(\epsilon) J_q^{-1}(\epsilon) = I$ and differentiate both sides at $\epsilon = 0$, obtaining

$$\frac{\partial}{\partial \epsilon} J_q^{-1}(\epsilon)\Big|_{\epsilon=0} = -J_q^{-1} \frac{\partial}{\partial \epsilon} J_q(\epsilon)\Big|_{\epsilon=0} J_q^{-1}, \tag{2.5.13}$$

where

$$\frac{\partial}{\partial\epsilon}J_q(F_\epsilon)\bigg|_{\epsilon=0} = \frac{\partial}{\partial\epsilon}E_{F_\epsilon}\left\{\nabla_\theta u_q(\epsilon)\right\}\bigg|_{\epsilon=0} = D_q(x,\theta) - J_q + \nabla_\theta u_q(x), \tag{2.5.14}$$

and $D(x,\theta)$ is the $p \times p$ matrix with elements

$$\left\{D_q(x,\theta)\right\}_{i,j} = E_{F_\theta}\left\{\sum_{k=1}^{p}\frac{\partial}{\partial\theta_k}(\nabla_\theta u_q)_{i,j}\right\}\left\{\mathrm{IF}^*(x,\theta)\right\}_k, \tag{2.5.15}$$

Finally, in the second and third terms of (2.5.10) we have

$$\frac{\partial u_q(\epsilon)}{\partial\epsilon}\bigg|_{\epsilon=0} = \nabla_\theta u_q \mathrm{IF}^*(x,\theta). \tag{2.5.16}$$

Re-organizing all the terms as in Equation (2.5.10) gives the final expression of the change of variance function in Proposition 2.

Remark 1. If $\nabla^2_\theta \tau_q(\theta) = 0$, then $Q_q(x,\theta) = 0$, which simplifies the expression of $\mathrm{CVF}_q(x,\theta)$ in certain cases. For example, this occurs if the reference density is a canonical exponential family with density $f_t(x) = \exp\{\eta(t)^a(x) - b(t)\}$.

2.6 Appendix: figures and tables

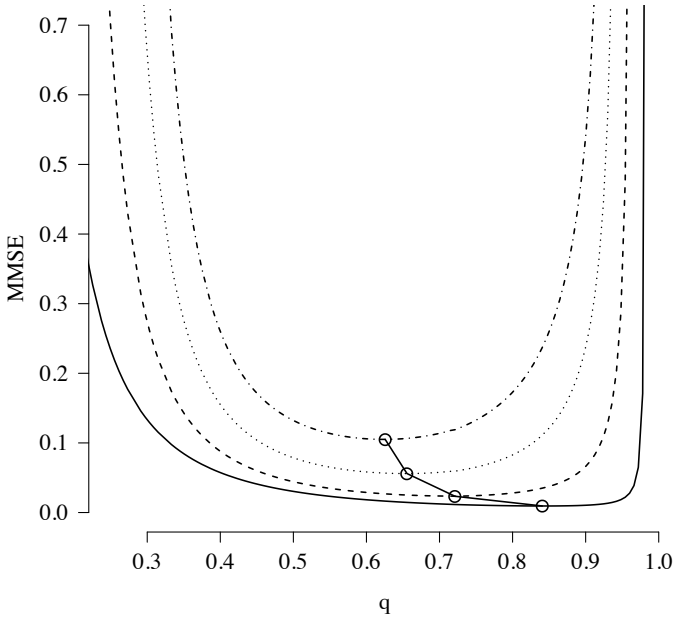

Figure 2.1: MMSE for Exp(1) with $n = 150$ and $\epsilon = 1\%$ (solid), $\epsilon = 5\%$ (dashed), $\epsilon = 10\%$ (dotted) and $\epsilon = 15\%$ (dot-dashed). The circles represent the optimal trajectory for q.

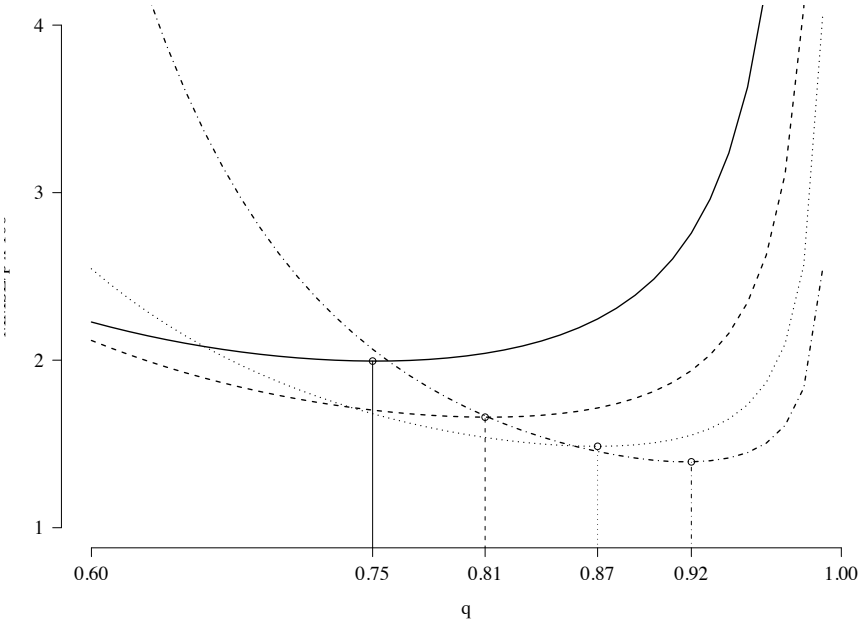

Figure 2.2: Element-wise MMSE for the mean of a $N_p(0, I)$ ($\epsilon = 0{\cdot}05$, $n = 100$) and optimal q for $p = 1$ (solid), $p = 2$ (dashed), $p = 4$ (dotted) and $p = 8$ (dot-dashed)

		$\epsilon = 0 \cdot 05$				$\epsilon = 0 \cdot 15$			
n	$p =$	1	5	15	30	1	5	15	30
100		0.9077	0.9583	0.9789	0.9857	0.8594	0.9331	0.9698	0.9857
		[0.75]	[0.89]	[0.95]	[0.97]	[0.69]	[0.86]	[0.94]	[0.97]
1000		0.8504	0.9331	0.9698	0.9857	0.8412	0.93302	0.9698	0.9857
		[0.68]	[0.86]	[0.94]	[0.97]	[0.67]	[0.86]	[0.94]	[0.97]

Table 2.1: Efficiency for the mean of a p-valued normal distribution, under min-max selection of q via MMSE minimisation (in squared brackets) for $n = 100, 1000$ and $\epsilon = 0{\cdot}05, 0{\cdot}15$.

$\gamma_\alpha = \gamma_q$	∞	$8 \cdot 12$	$4 \cdot 47$	$2 \cdot 35$	$1 \cdot 74$	$1 \cdot 58$	$1 \cdot 55$	$1 \cdot 51$	$1 \cdot 46$	$1 \cdot 44$
α	$0 \cdot 00$	$0 \cdot 05$	$0 \cdot 10$	$0 \cdot 25$	$0 \cdot 50$	$0 \cdot 75$	$0 \cdot 85$	$1 \cdot 00$	$1 \cdot 25$	$1 \cdot 35$
ARE_α	$1 \cdot 00$	$0 \cdot 99$	$0 \cdot 98$	$0 \cdot 89$	$0 \cdot 73$	$0 \cdot 62$	$0 \cdot 58$	$0 \cdot 54$	$0 \cdot 49$	$0 \cdot 48$
q	$1 \cdot 00$	$0 \cdot 98$	$0 \cdot 95$	$0 \cdot 90$	$0 \cdot 85$	$0 \cdot 82$	$0 \cdot 81$	$0 \cdot 80$	$0 \cdot 79$	$0 \cdot 78$
ARE_q	$1 \cdot 00$	$0 \cdot 98$	$0 \cdot 93$	$0 \cdot 81$	$0 \cdot 70$	$0 \cdot 63$	$0 \cdot 62$	$0 \cdot 60$	$0 \cdot 57$	$0 \cdot 56$

Table 2.2: Asymptotic relative efficiency for $N_1(0, \sigma^2)$ of our estimator (ARE_q) and [10] estimator (ARE_α) computed for different tuning constants q and α yielding the same gross-error sensitivity $\gamma_\alpha = \gamma_q$.

Chapter 3

Semi-parametric rank-based tests and estimators for Markov processes

3.1 Introduction

Time series analysis plays an important role in several scientific fields, like biology, physics, economics and finance. Among the stochastic processes applied for modeling the behavior of time-dependent data, Markov processes are a fundamental tool that has been deserving special attention. Many mathematical aspects of Markov processes have been studied in great details (see, e.g., Karlin and Taylor [57, Chapters 14-15]), but the statistical procedures for Markov models still present several challenges. For instance, quite often there is not a closed-form expression for the transition density, the data contain contamination or the model is mispecified, since the actual density is more leptokurtic than the reference density. In order to overcome many of these issues, we exploit distribution-free rank-based procedures for inference and testing, in a semi-parametric framework. Our setting includes both continuous-time Markov processes with discrete-time observations (e.g., daily or weakly observations of diffusions), and discrete-time Markov processes (e.g., AR(p) or ARCH(p) processes).

The main tool that one should apply for inference and testing in Markov processes is the score function of the Maximum Likelihood Estimator (MLE). MLE relies on the complete parametric specification of the

transition density of the process and it conveys two problems. First, the reference transition density is often too rigid and unrealistic. This implies that statistical models are generally mispecified, and even a small degree of mispecification can yield unreliable estimates and tests. Second, even if the transition dynamics is well specified, for many Markov processes (e.g. diffusions or affine diffusions), the transition density does not admit a closed-form expression: only the first conditional moments can be specified. Both those aspects have led to a well recognized need for more flexible procedures, alternative to MLE. The idea of all these procedures is to exploit the information conveyed by some characteristics of the process, avoiding the complete specification of the transition density.

Exactly in this spirit, a possible alternative to MLE relies on the specification of the conditional (or unconditional) moments of the process. This is the so-called Method of Moments (MM), that yields root-n consistent M-estimators neglecting a large part of the information contained in the transition density. As a result, the gain in flexibility comes at a price payed in terms of efficiency loss wrt MLE. Another drawback of MM is that the higher is the dimension of the parameter space, the larger is the number of moment conditions to be specified. This aspect leads to high mispecification risk, which could imply unreliable testing and inference.

Another alternative to MLE is the Pseudo Maximum Likelihood (PML) method based on the Gaussian transition density. As in the MM, only the first two conditional moments are specified in closed-form, but the PML method additionally relies on the assumption that the transition density is Gaussian. If the first two conditional moments are correctly specified, PML method defines a root-n consistent and asymptotically normal estimator (see Gourieroux et al. [36]). The artificial assumption of a Gaussian transition density plays a central role also in testing, in particular in the Lagrange Multiplier (LM) test. A drawback of the statistical procedures based on Gaussian score is that they can be extremely inefficient (in terms of variance of estimates and power of tests), when the sample size is small and when the actual distribution is not Gaussian. The latter is a crucial aspect, for instance, in the estimation for financial data, where the actual

density is leptokurtic (see, e.g., Engle and Gonzalez-Rivera [27], and Linton [62]), or when the real-data contain some departures from the Gaussian assumption (see Mancini et al. [66]).

A further way to obtain flexible estimators is provided by the Martingale Estimating Functions (in the following MEF) method; see Heyde [52, Chapter 6]. The method of MEF has a semi-parametric flavor, since it avoids the complete specification of the transition density. The idea is to define a class of estimating equations based on linear combinations of the first and second conditional moments. The weights of these linear combinations efficiently exploit the information about the transition density contained in the third and fourth moments of the process. Optimal MEF define root-n consistent estimators, and typically improve upon the efficiency of the MM or PMLE. Clearly, the specification of the higher-order conditional moments can be a difficult task, which can introduce large bias, if the first four conditional moments are mispecified.

In a similar direction, Wefelmeyer (see [85], [86], [87]) have proposed a semi-parametric inferential method for Markov processes. Wefelmeyer's approach does not specify the transition density, and only the first two conditional moments of the process have a parametric form. Thanks to the Local Asymptotic Normality (LAN), the semi-parametric efficient score is defined by the tangent space projection. The method requires a kernel density estimation of the transition density in order to estimate the third and fourth conditional moment of the process (see Wefelmeyer [86]). This device avoids the mispecification risk characterizing the MM or the MEF, but it is a non-trivial task, even in the simple AR(1) model; see Wefelmeyer [85]. Moreover, Wefelmeyer does not discuss explicitly how to derive semi-parametric testing procedures.

Our approach goes in the same direction of Wefelmeyer, but it relies on a simplified method. We also work in a semi-parametric framework, since we specify only the first two conditional moments of the Markov process, while the entire transition density is left unspecified. Then, we define a score function which is semi-parametrically efficient at a given reference distribution, and rank-based. Efficiency must be understood á la Le Cam, in a local and asymptotic sense: given a reference class of densities $\mathcal{G}$, we say that a method is semi-parametrically efficient at a given $g \in \mathcal{G}$ if it is semi-parametrically efficient at some reference density

g (see, e.g., Hallin and Werker [43]). Beside being semi-parametric efficient at g, our score functions are distribution-free. This important feature implies that, even if the actual distribution is different from the reference distribution, rank-based scores define tests having the right α-level, and the implied R-estimators are always root-n consistent. Differently from classical semi-parametric approach, rank-based scores achieve these goals at a smaller computation cost, since a kernel density estimation of the actual density is not needed. An additional important point of our method is that its implementation does not suffer from the typical computational issues (e.g., non-convexity of the estimating function) that discourage the application of R-estimators. Indeed, we follow Hallin and Paindavaine [40] and we propose a one-step procedure, which is easy-to-implement.

The paper has the following structure. In Section 2, and Section 3, we specify the assumptions characterizing our framework. In Section 4, we introduce the rank-based statistical procedures for inference and testing, and we discuss the main properties of R-estimators and tests. Some illustrative theoretical examples are provided. In Section 5, we discuss three different numerical exercises about a specific Markov model: the AR(3) process. Finally, in Section 7, we apply our method to the estimation and testing of stochastic volatility models. Proofs, tables and figures are in Appendix.

3.2 Model setting

Assume that we have observations $(X_{-q+1}, ..., X_0, X_1, .., X_t, ..., X_n)$ of some covariates $X_t \in \mathbb{R}^k$ and observations $(Y_{-q+1}, ..., Y_0, Y_1, .., Y_t, ..., Y_n)$ of the response variable $Y_t \in \mathbb{R}$. We assume that Y_t, is a time-homogeneous, stationary and ergodic Markov process, with invariant measure $\pi(y, \theta)$ (unspecified). Sufficient conditions for this are provided in Tweedie [80]. Let $\mathbf{Y}_{t-1} := (Y_{t-1}, ..., Y_{t-q})$ and $\mathbf{y}_{t-1} := (y_{t-1}, ..., y_{t-q})$. We label $S \subset \mathbb{R}$ the state space of the Markov process.

Assume that we do not specify the transition distribution $(Q_{\theta,g})$ of the process, whereas we know in closed-form only the first and the second conditional moment. We consider the information conveyed by

$\mathbf{Y}_{t-1}$ and by some exogenous covariates X_t, so the conditional mean reads:

$$m(\mathbf{y}_{t-1}, x_t, \theta) := E_\theta[Y_t|Y_{t-1} = y_{t-1}, ..., Y_{t-p} = y_{t-p}, X_t = x_t] = \int_S y_t Q_{\theta,g}(dy_t; \mathbf{y}_{t-1}; x_t), \tag{3.2.1}$$

while the conditional variance is

$$V(\mathbf{y}_{t-1}, x_t, \theta) := V_\theta[Y_t|Y_{t-1} = y_{t-1}, ..., Y_{t-p} = y_{t-p}, X_t = x_t] = \int_S (y_t - m(\mathbf{y}_{t-1}, x_t, \theta))^2 Q_{\theta,g}(dy_t; \mathbf{y}_{t-1}; x_t). \tag{3.2.2}$$

In Eq. (3.2.1) and Eq. (3.2.2), $Q_{\theta,g}$ represents the transition distribution of the Markov model and $\theta \in \Theta \subset \mathbb{R}^p$. For the sake of notational convenience in the following we drop the dependence of the response variable from the covariates. Clearly all our results can be extended to include the covariates just by means of some small and merely notational changes.

Following Wefelmeyer ([86]), we propose a method for drawing inference about θ avoiding a complete parametric specification of the transition distribution. To this end, we cast the quantities of Eq. (3.2.1)-(3.2.2) in the following dynamics:

$$Y_t = m(\mathbf{Y}_{t-1}, \theta) + v(\mathbf{Y}_{t-1}, \theta)\varepsilon_t \tag{3.2.3}$$

where $\{\varepsilon_t\}$, for $t = 0, \pm1, \pm2, ...$, is an iid sequence with unknown density g, zero mean and unitary variance. For the sake of simplicity, we label $v(\cdot, \theta) := V(\cdot, \theta)^{1/2}$.

The Eq. (3.2.3) implies the following semi-parametric probability model for the transition distribution of $Y_t|\mathbf{Y}_{t-1}$, when the sample size is given by n:

$$\mathcal{Q} = \left\{ Q^{(n)}_{\theta,g}, \theta \in \Theta, g \in \mathcal{G} \right\}. \tag{3.2.4}$$

In Eq. (3.2.4), the parameter θ is the Euclidean parameter in Θ, while function g denotes the unknown nuisance infinite dimensional parameter, belonging to a specific class $\mathcal{G}$. We restrict ourself to the case where $\mathcal{G}$ is a subset of the class $\mathcal{G}_0$ of all non–vanishing densities over the real line and having zero mean and unitary variance. Specifically, we assume that $\mathcal{G}$ contains densities g such that:

A. For $x \in \mathbb{R}$, function $g(x)$ is strictly positive, with $E_g(X) = 0$ and $E_g(X^2) = 1$;

B. Function g is absolutely continuous on finite intervals, i.e. there exists a derivative (wrt x) $\dot{g}$ such that for all $-\infty < a < b < \infty$, $g(a) - g(b) = \int_a^b \dot{g}(x)dx$;

C. The Fisher information for location is $I_{ll}(g) := \int_{\mathbb{R}} (\dot{g}(x)/g(x))^2 g(x)dx < \infty$ and the Fisher information for scale is $I_{ss}(g) := \int_{\mathbb{R}} (1 + x\dot{g}(x)/g(x))^2 g(x)dx < \infty$.

The model in Eq. (3.2.3) is called *regression-autoregression model* for the Markovian dynamics of the response variable (see, e.g., Wefelmeyer [85]). Its specification includes both the case of heteroscedastic (when the variance depends on $\mathbf{Y}_{t-1}$) and homoscedastic (when $v(\mathbf{Y}_{t-1}, \theta) \equiv v(\theta)$) regression-autoregression model, and it is related to following semi-parametric model for the transition density:

$$Q_{\theta,g}^{(n)}(dy_t; \mathbf{Y}_{t-1}) = \frac{1}{v(\mathbf{Y}_{t-1}, \theta)} g\left(\frac{y_t - m(\mathbf{Y}_{t-1}, \theta)}{v(\mathbf{Y}_{t-1}, \theta)}\right) dy_t. \tag{3.2.5}$$

For a given $g \in \mathcal{G}$ and a given θ, let $H_g^{(n)}(\theta)$ denote the hypothesis that $\mathbf{Y}_n := (Y_1, ..., Y_n)$ is generated according to Eq. (3.2.3). We denote the semi-parametric hypothesis by $\mathcal{H}^{(n)}(\theta) := \cup_{g \in \mathcal{G}} H_g^{(n)}(\theta)$, with $\mathcal{G}$ as in Assumption 3.2, and we introduce the residual function:

$$Z_t(\theta) := \frac{Y_t - m(\mathbf{Y}_{t-1}, \theta)}{v(\mathbf{Y}_{t-1}, \theta)}. \tag{3.2.6}$$

Comparing Eq. (3.2.3) and Eq. (3.2.6), the semi-parametric hypothesis $\mathcal{H}^{(n)}(\theta)$ holds true iff $Z_t(\theta) \equiv \varepsilon_t$.

3.3 Uniform Local Asymptotic Normality

In this section we introduce our methodology. In the first subsection we derive the ULAN for the model in Eq. (3.2.3). Our results are based on Drost et al. [24] Theorem 2.1. In the second subsection, following Hallin and Werker [43], we derive the rank-based version of the central sequence.

3.3.1 Specification

To state the ULAN in our setting, we rely on standard assumptions. Let us consider the regression-autoregression model in Eq. (3.2.3). As in the classical semi-parametric setting, we embed g into some

parametric family $\mathcal{U} := \{U_\eta : \eta \in [-1,1]^m \subset \mathbb{R}^m\}$, with dominating measure μ (typically the Lebesgue measure) and density $u(\eta) = dU_\eta/d\mu$, such that $g = u(0)$. Define $l(\eta) := \log u(\eta)$. To derive the ULAN we introduce the following:

D. The vector $(Y_{-q+1}, ..., Y_0)$ and $\{\varepsilon_t, t \geq 1\}$ are independent;

E. Let k_θ the joint probability density of $(Y_{-q+1}, ..., Y_0)$. Then,

$$\log(k_\theta(Y_{-q+1}, ..., Y_0)) - \log(k_{\theta+n^{-1/2}\tau_n}(Y_{-q+1}, ..., Y_0)) = o_P(1),$$

as $n \to \infty$, for every $\tau_n \in \mathbb{R}^p$ such that $\tau_n'\tau_n$ is uniformly bounded. Under Assumption D the unconditional log-likelihood is given by:

$$\log L_{g,\theta}^{(n)}(Y_{-q+1}, ..., Y_0, Y_1, ..., Y_n) = \log k_\theta(Y_{-q+1}, ..., Y_0) + \sum_{t=1}^{n} \log g(Z_t(\theta))$$

Assumption **D** is related to the paper of Swenson [77] and it guarantees that the influence of $(Y_{-q+1}, ..., Y_0)$ on $\log L_{g,\theta}^{(n)}(Y_{-q+1}, ..., Y_0, Y_1, ..., Y_n)$ is asymptotically negligible. Thus, we state the following:

Proposition 3.3.1. ***(Drost et al. [24], Theorem 2.1.)*** *For a fix $g \in \mathcal{G}$ and under Assumptions **A**-**E**, we have:*

$$\Lambda_n := \log \frac{L_{\theta+\tau_n n^{-1/2},g}^{(n)}}{L_{\theta,g}^{(n)}} = \tau_n' \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \dot{\mathbf{l}}_\mathbf{t}(\theta, g) - \frac{1}{2n}\tau_n'\Gamma(\theta, g)\tau_n + R_n, \tag{3.3.1}$$

where $\dot{\mathbf{l}}_\mathbf{t}(\theta_0, g) := \dot{\mathbf{l}}(\theta, g, \mathbf{Y}_{t-1}, z_t) = \nabla_\theta \mathbf{l}(\theta, g, \mathbf{Y}_{t-1}, z_t)$ and

$$\Gamma(\theta, g) := E_{Q_{\theta,g}^{(n)}}[\dot{\mathbf{l}}_t(\theta, g)\dot{\mathbf{l}}_t(\theta, g)']$$

represents the Information Matrix. Under the distribution $Q_{\theta,g}^{(n)}$, we have:

$$R_n \xrightarrow{P} 0 \quad \textit{and} \quad \Lambda_n \xrightarrow{D} N\left(-\frac{1}{2}\tau_n'\Gamma(\theta, g)\tau_n, \tau_n'\Gamma(\theta, g)\tau_n\right).$$

From Proposition 3.3.1, it easily follows:

Corollary 3.3.2. *Let Y_i be a strictly stationary and ergodic Markov model as in the regression-autoregression model of Eq. (3.2.3). Under the same Assumptions as in Proposition 3.3.1, the central sequence in Eq. (3.3.1) specifies as:*

$$\mathbf{\Delta}^{(n)}(\theta, g) := \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \dot{\mathbf{l}}_\mathbf{t}(\theta, g), \tag{3.3.2}$$

where $\dot{\mathbf{l}}_\mathbf{t}(\theta, g) \in \mathbb{R}^p$ has expression

$$\dot{\mathbf{l}}_\mathbf{t}(\theta, g) := \frac{\dot{\mathbf{m}}(\mathbf{Y}_{t-1}, \theta)}{v(\mathbf{Y}_{t-1}, \theta)} \frac{\dot{g}}{g}(Z_t(\theta)) - \frac{\dot{\mathbf{v}}(\mathbf{Y}_{t-1}, \theta)}{v(\mathbf{Y}_{t-1}, \theta)} \left(1 + Z_t(\theta)\frac{\dot{g}}{g}(Z_t(\theta))\right), \tag{3.3.3}$$

where $\dot{\mathbf{m}}(\mathbf{Y}_{t-1},\theta)=\nabla_\theta \mathbf{m}(\mathbf{Y}_{t-1},\theta)$ *and* $\dot{\mathbf{v}}(\mathbf{Y}_{t-1},\theta)=\nabla_\theta \mathbf{v}(\mathbf{Y}_{t-1},\theta)$. *Moreover:*

$$\boldsymbol{\Delta}^{(n)}(\theta,g)\xrightarrow{D}\boldsymbol{\Delta}\sim N(\mathbf{0},\Gamma(\theta,g)) \tag{3.3.4}$$

under $Q^{(n)}_{\theta,g}$, *as* $n\to\infty$.

Proof. See Appendix A. □

For $\theta\in\Theta$ and for a fixed $g\in\mathcal{G}$, the ULAN property implies the weak convergence of the sequence of local experiments:

$$\mathcal{E}^{(n)}_{g,\theta}=\left\{S^{(n)},\mathcal{F}^{(n)},Q^{(n)}_{\theta+n^{-1/2}\tau_n,g};\tau_n\in\mathbb{R}^p\right\},n\in\mathbb{N},$$

to the p-dimensional Gaussian shift experiment:

$$\mathcal{E}_{g,\theta}=\left\{S^{(p)},\mathcal{F}^{(p)},N(\Gamma(\theta,g)\tau_n,\Gamma(\theta,g));\tau_n\in\mathbb{R}^p\right\}. \tag{3.3.5}$$

As far as the Information Matrix $\Gamma(\theta,g)$ is concerned, we introduce the following **F.** The Information Matrix in Eq. (3.3.1) is such that:

F1.

$$\Gamma(\theta,g)=\left(\begin{array}{c|c} I_{ll}(g)\Gamma_1(\theta) & \mathbf{0} \\ \hline \mathbf{0} & I_{ss}(g)\Gamma_2(\theta)\end{array}\right) \tag{3.3.6}$$

where $\Gamma_1(\theta)$ is a $(p_1\times p_1)-$matrix and $\Gamma_2(\theta)$ is a $(p_2\times p_2)-$matrix, with $p_1+p_2=p$. Neither $\Gamma_1(\theta)$ nor $\Gamma_2(\theta)$ depend on the nuisance parameter. The scalar quantities $I_{ll}(g)$ and $I_{ss}(g)$ have been defined in Assumption 3.2 and they are independent from the Euclidean parameter.

F2. $\Gamma(\theta,g)$ is full-rank and $\theta\mapsto\Gamma(\theta,g)$ is continuous on Θ, for a given g.

The block-diagonal structure in Eq. (3.3.6) can be rewritten as $\Gamma(\theta,g)=\mathcal{J}(g)\Upsilon^{-1}(\theta)$, where $\mathcal{J}(g)$ depends only on g and $\Upsilon^{-1}(\theta)$ depends only on θ. See Cassart et al. [16] for additional discussions. Assumption **F2** is fairly standard and assumption **F1** is satisfied for several well-known models. For the sake of illustration, we provide three examples satisfying it.

Example 3.3.3. *Let us consider the AR(q) process. The processes satisfies the following stochastic difference equation:*

$$Y_t=\sum_{i=1}^{q}\theta_i Y_{t-i}+\varepsilon_t,$$

*The ε_t, for $t = 0, \pm 1, ..., \pm 1$, are iid with probability g, such that $E_g(\varepsilon_t) = 0$ and $V_g(\varepsilon_t) = 1$, and satisfying Assumption **B**,**C**,**D**. The model has the form of Eq. (3.2.3), where $m(\mathbf{y}_{t-1}; \theta) = \theta' \mathbf{y}_{t-1}$, and $v(\mathbf{y}_{t-1}; \theta) = 1$. Under the standard stationarity condition about $\theta := (\theta_1, ..., \theta_p)$, Corollary 3.3.2 implies that the process admits a ULAN representation, with central sequence:*

$$\mathbf{\Delta}^{(\mathbf{n})}(\theta, g) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\dot{g}}{g}(Z_t(\theta)) \begin{pmatrix} Y_{t-1} \\ \vdots \\ Y_{t-q} \end{pmatrix},$$

where, from Eq.(3.2.6) $Z_t(\theta) = Y_t - \sum_{i=1}^{q} \theta_i Y_{t-i}$. Hallin and Werker in [44] show that the Information Matrix has form:

$$\Gamma(\theta, g) := \int_{-\infty}^{\infty} \left(\frac{\dot{g}}{g}(z_t) \right)^2 g(z_t) dz_t \Gamma_G(\theta) = I_{ll} \Gamma_G(\theta)$$

where $\Gamma_G(\theta)$ is the auto-covariance matrix of order q of the stationary AR(q) processes (see, e.g., Hamilton [45, p.58-59]).

Example 3.3.4. *Let us consider the ARCH(1):*

$$\begin{cases} Y_t = \sigma_t \varepsilon_t \\ \sigma_t = \sqrt{1 + \theta Y_{t-1}^2}, \end{cases}$$

with ε_t iid having density $g(0,1)$ as in Assumption 3.2. The variance σ_t^2 is $\mathcal{F}_{t-1}-$measurable and the process is stationary if $\theta \in \left(0; \left(\int_{-\infty}^{\infty} x^2 g(x) dx\right)^{-1}\right)$. ARCH(1) model dynamics can be cast into the Eq. (3.2.3) just setting: $m(\mathbf{y}_{t-1}; \theta) \equiv 0$ and $V(\mathbf{y}_{t-1}; \theta) := 1 + \theta y_{t-1}^2$. From Eq. (3.3.2) and Eq. (3.3.3) we get:

$$\Delta^{(n)}(\theta, g) = \frac{1}{2\sqrt{n}} \sum_{t=1}^{n} \frac{Y_{t-1}^2}{1 + \theta Y_{t-1}^2} \left(1 + Z_t(\theta) \frac{\dot{g}}{g}(Z_t(\theta)) \right),$$

where $Z_t(\theta) = \sigma_t^{-1} Y_t$. The Information Matrix is

$$\Gamma(\theta, g) := \int_{-\infty}^{\infty} \left(1 + z_t \frac{\dot{g}}{g}(z_t) \right)^2 g(z_t) dz_t \Xi(\theta) = I_{ss} \Xi(\theta),$$

with $\Xi(\theta)$ that does not depend upon g.

Example 3.3.5. *Let us consider the following discrete-time stochastic volatility model:*

$$\begin{cases} Y_t = \sigma_t \varepsilon_t \\ \log \sigma_t^2 = \mu + \rho(\log \sigma_{t-1}^2 - \mu) + \kappa v_t, \end{cases}$$

*where $\rho \in (-1, 1)$, $\mu \in \mathbb{R}$, and $\kappa > 0$. Moreover, we assume ε_t iid having density $N(0,1)$, v_t iid with density $g(0,1)$, symmetric and satisfying Assumption 3.2. Finally, we assume that the errors in the two equations are independent. Y_t represents the log-return at time t, whereas $\log \sigma_t^2$ is the log-volatility. Differently from the time-varying volatility in Example 7, in this model the (log)volatility is stochastic, since $\log \sigma_t^2$ is not $\mathcal{F}_{t-1}-$measurable. Let us consider the log-volatility process and define $Z_t(\theta) := \kappa^{-1}(\log \sigma_t^2 - (\mu + \rho(\log \sigma_{t-1}^2 - \mu)))$, where $\theta = (\mu, \rho, \kappa)$. Clearly the AR(1) dynamics of $\log \sigma_t^2$ can be cast in Eq. (3.2.3). If $|\rho| < 1$, the log-volatility process is stationary and under Assumptions **A**-**E**, the process is ULAN. Using Eq. (3.3.2) and Eq. (3.3.3), the central sequence reads:*

$$\mathbf{\Delta}^{(n)}(\theta, g) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{1}{\kappa} \begin{pmatrix} (1 - \rho) \frac{\dot{g}}{g}(Z_t(\theta)) \\ (\log \sigma_{t-1}^2 - \mu) \frac{\dot{g}}{g}(Z_t(\theta)) \\ Z_t(\theta) \left(1 + Z_t(\theta) \frac{\dot{g}}{g}(Z_t(\theta)) \right) \end{pmatrix}.$$

The information matrix is

$$\Gamma(\theta, g) = \left(\begin{array}{c|c} I_{ll}(g) \Gamma_1(\theta) & \mathbf{0} \\ \hline \mathbf{0} & I_{ss}(g) \Gamma_2(\theta) \end{array} \right),$$

where,

$$\Gamma_1(\theta) = \begin{pmatrix} 1-\rho^2 & 0 \\ 0 & \frac{\kappa^2}{1-\rho^2} \end{pmatrix}, \quad \textit{and} \quad \Gamma_2(\theta) = 1,$$

and $I_{ll}(g)$ *and* $I_{ss}(g)$ *are given in Assumption* ***C****.*

3.3.2 Semi-parametric central sequence and its rank-based version

In the classical semi-parametric approach, the efficient central sequence is the tool that one needs in order to define inference and testing procedures; see, e.g, Bickel et al. [13] for the iid framework, and Hallin and Werker [43] for time-series. To define this tool, let us consider the sequence of parametric experiments

$$\mathcal{E}^{(n)}_{u,\theta} = \left(S^{(n)}, \mathcal{F}^{(n)}, Q^{(n)}_{\theta,u(\eta)} : \theta \in \Theta, \eta \in [-1,1]^m\right), n \in \mathbb{N}.$$

We write $\left(\mathbf{\Delta}^{(n)}(\theta,g)', \mathbf{H}^{(n)}_{\eta}(\theta,g)'\right)$ for the central sequences related to both the Euclidean (θ) and to the nuisance (η) parameter. The efficient central sequence $\mathbf{\Delta}^{*(n)}(\theta,g)$ is obtained by regressing the θ-part of the central sequence $\mathbf{\Delta}^{(n)}(\theta,g)$ on $\mathbf{H}^{(n)}_{\eta}(\theta,g)$.

The computation of the efficient central sequence for models as in Eq. (3.2.3) requires to go through the L_2-projection of the central sequence for θ onto the sub-space of L_2 containing the Hellinger derivatives of the transition density; see Wefelmeyer [85] and [86]. The computation of this projection is typically a tough task in standard semi-parametric contexts, and it could discourage the application of semi-parametric procedures.

Beside this aspect, there is another drawback of the tangent space approach. Consider the case where f represents the true unknown actual density, while g represents a fixed reference distribution. Classical parametric approach fix $g \equiv f$ and they have a strong limitation: they yield estimators (tests) that lose their consistency (validity), when $g \neq f$. This implies that parametric estimators are no longer root-n consistent and the tests do not have α-validity. The same concern applies to $\mathbf{\Delta}^{*(n)}(\theta,g)$, since the semi-parametric central sequence is unaffected by local perturbations of g, but it remains sensitive the the nonlocal ones. The main reason is that

$$E_{Q^{(n)}_{\theta,f}}[\mathbf{\Delta}^{*(n)}(\theta,g)] \neq 0. \tag{3.3.7}$$

when $g \neq f$. To overcome this problem, semi-parametric procedures rely on a consistent estimator of the actual density, obtained using complicated technicalities like the sample-splitting. The invariance property of the ranks offers an alternative solution: the ranks can be applied to define a distribution-free central sequence, obtained by the L_2-projection onto the maximal invariant σ-field generated by the ranks. This yields root-n consistent R-estimators and tests with the right α-level, even when $g \neq f$. The ranks achieve this goal without kernel density estimation of f, and they achieve semi-parametric efficiency at a fix reference density g.

Let us denote by $R_t^{(n)}(\theta)$ the rank of $Z_t^{(n)}(\theta)$ among $Z_1^{(n)}(\theta), ..., Z_n^{(n)}(\theta)$ and by $\mathbf{R}^{(n)}$ the vector of ranks $(R_1^{(n)}(\theta), ..., R_n^{(n)}(\theta))$. To shorten the notation, we write $R_1^{(n)}, ..., R_n^{(n)}$ and $Z_1^{(n)}, ..., Z_n^{(n)}$, dropping the dependence on θ. Under Assumption **A**, for fixed θ and n, the nonparametric model $\mathcal{E}_{\eta,\theta}^{(n)}$, for $\eta \in [-1,1]^m$, is generated by the group $(\mathcal{W}_{z^{(n)}}, \cdot) = (\{\mathcal{W}_h^{(n)}, h \in \mathcal{H}\}, \cdot)$ of continuous order-preserving transformations acting on $Z_1^{(n)}, ..., Z_t^{(n)}$. The transformations are such that:

$$\mathcal{W}_h^{(n)}(Z_1^{(n)}, ..., Z_t^{(n)}) = (h(Z_1^{(n)}), ..., h(Z_t^{(n)})),$$

for functions $h \in \mathcal{H}$, where $h : \mathbb{R} \to \mathbb{R}$ is monotone, strictly increasing and $\lim_{x\to\pm\infty} h(x) = \pm\infty$. The maximal invariant[1] σ-field is given by

$$\mathcal{B}^{(n)}(\theta) := \sigma(R_1^{(n)}, ..., R_n^{(n)}) = \sigma(\mathbf{R}^{(n)}). \tag{3.3.8}$$

Proposition 3.1 in Hallin and Werker [43] shows that under some regularity conditions on the score, there exists a central sequence $\mathbf{\Delta}^{(n)}(\theta, g)$ providing a rank-based approximation to $\mathbf{\Delta}^{*(n)}(\theta, g)$.

To compute such an approximation in our setting, we rewrite the central sequence $\mathbf{\Delta}^{(n)}(\theta, g)$ as a

[1] If we assume f to be symmetric, the maximal invariant σ-field is given by the ranks and the signs. In particular, let us denote by $R_{+,t}^{(n)}$ the rank of $|Z_t^{(n)}|$ among the absolute values $|Z_1^{(n)}|, ..., |Z_n^{(n)}|$, by $\mathbf{R}_+^{(n)}$ the vector of ranks $(R_{+,1}^{(n)}, ..., R_{+,n}^{(n)})$, by $sgn(Z_t^{(n)})$ the sign of $Z_t^{(n)}$ and by $\mathbf{sgn}(\mathbf{Z^{(n)}})$ the vector of signs $(sgn(Z_1^{(n)}, ..., sgn(Z_n^{(n)}))$. To shorten the notation we write $R_{+,1}^{(n)}, ..., R_{+,n}^{(n)}$ and $(sgn(Z_1^{(n)}, ..., sgn(Z_n^{(n)}))$, dropping the θ. Then, the maximal invariant σ-field is given by

$$\mathcal{B}^{(n)}(\theta) := \sigma(R_{+,1}^{(n)}, ..., R_{+,n}^{(n)}, sgn(Z_1^{(n)}), ..., sgn(Z_n^{(n)})) = \sigma(\mathbf{R}_+^{(n)}; \mathbf{sgn}(\mathbf{Z^{(n)}})).$$

function of the innovations only. To this end, we label $\mathbf{Z}_{t-1} := (Z_{t-1}, ..., Z_0, Z_{-1}, ...)$ the vector containing the innovations up to time $t-1$. The regression-autoregressive structure of the Markov model implies that $Y_{t-1} = Y(\theta, \mathbf{Z}_{t-1})$. Thus, every element $\dot{\mathbf{l}}_t(\theta, g)$ of the central sequence in Eq. (3.3.3) is a function of the past values of the innovations up to time t. This implies that the conditional mean and the conditional variance of Y_t can be rewritten as functions of the past values of the innovations using the vector $\mathbf{Z}_{t-1}$:

$$\begin{aligned}\dot{\mathbf{l}}_t(\theta, g) &= \frac{\dot{\mathbf{m}}(\theta, Y_{t-1})}{v(\theta, Y_{t-1})} \frac{\dot{g}}{g}(Z_t) - \frac{\dot{\mathbf{v}}(\theta, Y_{t-1})}{\tilde{v}(\theta, Y_{t-1})} \left(1 + Z_t \frac{\dot{g}}{g}(Z_t)\right) \\ &= \frac{\dot{\mathbf{m}}(\theta, \mathbf{Z}_{t-1})}{v(\theta, \mathbf{Z}_{t-1})} \frac{\dot{g}}{g}(Z_t) - \frac{\dot{\mathbf{v}}(\theta, \mathbf{Z}_{t-1})}{\tilde{v}(\theta, \mathbf{Z}_{t-1})} \left(1 + Z_t \frac{\dot{g}}{g}(Z_t)\right). \end{aligned} \tag{3.3.9}$$

The definition of a rank-based version of $\dot{\mathbf{l}}_t(\theta, g)$ requires an approximation of the score by another score function involving only a finite number $s(n)$ of lags. To this end, we notice that, under the stationarity and ergodicity assumptions of the Markov process, it is always possible to truncate the dependence on the past up to a lag $s(n) \leq (n-1)$, such that $s(n) \to \infty$ as $n \to \infty$. This implies that the central sequence can be re-written as the sum of a quantity involving a finite number of lagged innovations (say $\mathbf{Z}_t^{(s)} := (Z_t, Z_{t-1}, ..., Z_{t-s})$) plus an $o_P(1)$. With a small abuse of notation, we still label $\mathbf{\Delta}^{(n)}(\theta, g)$ the central sequence obtained reconstructing the response variable Y_t by means of $\mathbf{Z}_i^{(s)}$.

Example 3.3.6. *The truncation of the central sequence is illustrated by several examples in Hallin and Werker [43] for different kinds of stochastic processes. We here briefly recall the case of a stationary AR(1) process: $Y_t = \theta Y_{t-1} + \varepsilon_t$, with iid innovations ε_t, having unspecified density, with mean zero and variance one. Under the stationary condition, $Y_{t-1} := \sum_{j=0}^{\infty} \theta^j \varepsilon_{t-1-j}$. The central sequence is obtained by Eq. (3.3.2), with the likelihood given Eq. (3.3.9). Since the likelihood involves infinitely many lags of the innovations, the central sequence inherits this feature:*

$$\begin{aligned}\Delta^{(n)}(\theta, g) &= \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} -\frac{g'}{g}(Z_t(\theta)) Y_{t-1} = \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} -\frac{g'}{g}(Z_t(\theta)) \sum_{j=0}^{\infty} \theta^j Z_{t-1-j}(\theta) \\ &= \sum_{j=0}^{\infty} \theta^j \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} -\frac{g'}{g}(Z_t(\theta)) Z_{t-1-j}(\theta). \end{aligned} \tag{3.3.10}$$

The stationarity condition $|\theta| < 1$ implies that there exists a number $s(n) \to \infty$ as $n \to \infty$ and such that the lags of higher-order than $s(n)$ have a negligible impact on the expression of Y_{t-1}. As a result, the summation in Eq. (3.3.10) can be truncated to the $s(n)$-th term. After some algebra, we get

$$\Delta^{(n)}(\theta, g) = \sum_{i=1}^{s(n)} \frac{1}{\sqrt{n-i}} \sum_{t=i+1}^{n} -\frac{g'}{g}(Z_t(\theta)) \theta^j Z_{t-i}(\theta) + Rem$$

where Rem converges to zero in quadratic mean, so that it is a $o_p(1)$ term (see Hallin and Werker [43]). We notice that analogous calculations hold true also when the $\sqrt{Var(\varepsilon_t)} = v(\theta) \neq 1$.

The truncation of the central sequence allows us to define a rank-based efficient central sequence for regression-autoregression Markov models. To this end, we introduce the following assumptions: **J.** The function $\dot{\mathbf{l}}_t$ is such that $0 < E[\|\dot{\mathbf{l}}_z(\theta, g, Z_0, ..., Z_s)\|^2] < \infty$ and

$$E_g[\dot{\mathbf{l}}(\theta, g, Z_0, ..., Z_s)|z_1, ..., z_s] = 0;$$

K. The function $\dot{\mathbf{l}}_t$ is component-wise monotone wrt to all its arguments or it is a linear combinations of such functions. Then, we state the following:

Proposition 3.3.7. *Let Assumptions **A**-**K** be satisfied for a stationary Markov chain, having dynamics as in Eq. (3.2.3), with central sequence $\mathbf{\Delta}^{(n)}(\theta, g)$ and with efficient central sequence $\mathbf{\Delta}^{*(n)}(\theta, g)$. Then, the following approximation holds:*

$$\mathbf{\Delta}^{*(n)}(\theta, g) = \mathbf{\Delta}^{(n)}(\theta, g) + o_P(1), \tag{3.3.11}$$

where

$$\mathbf{\Delta}^{(n)}(\theta, g) = E_{Q^{(n)}_{\theta,g}}[\mathbf{\Delta}^{(n)}(\theta, g)\|\mathcal{B}^{(n)}(\theta)]. \tag{3.3.12}$$

Being $\mathcal{B}^{(n)}(\theta)$ measurable, $\mathbf{\Delta}^{(n)}(\theta, g)$ is distribution-free.

Proof. See Appendix A. □

The distribution-freeness in Proposition 3.3.7 has important consequences for inference and testing. Specifically, let us consider the case where the actual distribution f is different from the reference distribution g. If f and g have the same maximal invariant σ-field, then it follows (up to $o_P(1)$):

$$E_{Q^{(n)}_{\theta,g}}[\mathbf{\Delta}^{(n)}(\theta, g)] \quad = E_{Q^{(n)}_{\theta,f}}[\mathbf{\Delta}^{(n)}(\theta, g)] \quad = 0. \tag{3.3.13}$$

Remark 3.3.8. *The Eq. (3.3.13) implies that $\mathbf{\Delta}^{(n)}(\theta, g)$ is insensitive to non-local perturbations of g, irrespectively of the sample size n. In that sense, rank-based procedures are globally resistant to model mispecification, since if $g \neq f$, and the actual density f belongs to $\mathcal{G}$, R-estimators are root-n consistent and tests have the right α-level. The property in Eq. (3.3.13) does not hold for the semi-parametric central sequence $\mathbf{\Delta}^{*(n)}(\theta, g)$ (see Eq. (3.3.7)).*

From Eq. (3.3.13) it follows that the LAN results of the Section 3.3.1 read as:

$$\mathbf{\Delta}^{(n)}(\theta, g) \xrightarrow{D} N(\mathbf{0}, \Gamma(\theta, g)) \tag{3.3.14}$$

under $Q^{(n)}_{\theta,f}$. Moreover from the Third Le Cam's Lemma it follows that

$$\mathbf{\Delta}^{(n)}(\theta, g) \xrightarrow{D} N(\Gamma(\theta, f, g)\tau_n, \Gamma(\theta, g)) \tag{3.3.15}$$

under the contiguous $Q^{(n)}_{\theta+\tau n^{-1/2},f}$. In Eq. (3.3.15), we have:

$$\Gamma(\theta,f,g) := E_{Q^{(n)}_{\theta,f}}\left[\mathbf{\Delta}^{(n)}(\theta,g)\mathbf{\Delta}^{(n)}(\theta,f)'\right]. \tag{3.3.16}$$

Computation of the rank-based central sequence. Let us label by $\dot{\mathbf{l}}^{(s)}_t(\theta,g)$ the score function computed using only a finite number $s(n)$ of innovations lags. Proposition 3.3.7 implies that the rank-based score is obtained by $E_{Q^{(n)}_{\theta,g}}\left[\dot{\mathbf{l}}^{(s)}_t(\theta,g)|\mathcal{B}^{(n)}(\theta)\right]$. If the score function is square-integrable, for every t there exist functions $\mathbf{a}^{(n)}_g : (R^{(n)}_t,...,R^{(n)}_{t-s},\theta)\to\mathbb{R}^p$, the so-called *exact scores*, such that:

$$\lim_{n\to\infty} E_g\left[||\dot{\mathbf{l}}^{(s)}_t(\theta,g,Z_t,...,Z_{t-s}) - \mathbf{a}^{(n)}_g(R^{(n)}_t,...,R^{(n)}_{t-s},\theta)||^2\right] = 0. \tag{3.3.17}$$

Lemma 1 and Theorem 1 in Hájek, Šidák and Sen [37, pag 187-188] show that the exact score function $\mathbf{a}^{(n)}_g$ can be computed as

$$\mathbf{a}^{(n)}_g(\mathbf{R^{(n)}}(\mathbf{z}^{(s)}_t);\theta) := E_g\left[\dot{\mathbf{l}}^{(s)}_t(\theta,g,Z_t,...,Z_{t-s})|R^{(n)}_t,...,R^{(n)}_{t-s})\right],$$

which satisfies the L_2 approximation in Eq. (3.3.17).

Nevertheless, quite often the conditional expectations involved in the expression of $\mathbf{a}^{(n)}_g(\mathbf{R^{(n)}}(\mathbf{z}^{(s)}_t);\theta)$ do not admit a closed-form. If it is the case, the exact scores must be replaced by the so called *approximate scores*, given by:

$$\mathbf{a}^{(n)}_g(\mathbf{R^{(n)}}(\mathbf{Z}^{(s)}_t);\theta) := \dot{\mathbf{l}}^{(s)}_t\left(\theta,g,G^{-1}\left(\frac{R^{(n)}_t}{n+1}\right),...,G^{-1}\left(\frac{R^{(n)}_{t-s}}{n+1}\right)\right), \tag{3.3.18}$$

where G is the CDF of $g\in\mathcal{G}$.

The approximated scores are introduced in Lemma 1 of Hájek, Šidák and Sen's book [37, pag 195], for monotone square-integrable functions. We notice that Hájek, Šidák and Sen Lemma deals with a score function having only one argument. In contrast, we are dealing with functions containing $s(n)$ lagged innovations. This implies that to go further in our construction, we need the following straightforward multivariate version of the Lemma 1 of Hájek, Šidák and Sen:

Lemma 3.3.9. *If the function* $\dot{\mathbf{l}}_t^{(s)}(\theta, g)$ *is a.e. continuous, square-integrable and monotone wrt all its arguments and if* $\mathbf{a}_g^{(n)}(\mathbf{R}^{(\mathbf{n})}(\mathbf{Z}_t^{(s)}); \theta)$ *are as in Eq. (3.3.18), then the condition*

$$\lim_{n\to\infty} E_g\left[||\dot{\mathbf{l}}_t^{(s)}(\theta, g, Z_t, ..., Z_{t-s}) - \mathbf{a}_g^{(n)}(\mathbf{R}^{(\mathbf{n})}(\mathbf{Z}_t^{(s)}); \theta)||^2\right] = 0. \tag{3.3.19}$$

is satisfied.

Proof. See Appendix A. □

A more intuitive and heuristic justification of Eq. (3.3.18) is that, when $Z_1, ..., Z_n$ are iid, with distribution function G, then

$$\frac{R_t^{(n)}}{n+1} \approx G(Z_t) \quad \Leftrightarrow \quad Z_t \approx G^{-1}\left(\frac{R_t^{(n)}}{n+1}\right), \tag{3.3.20}$$

where $G(Z_1), ..., G(Z_n)$ are iid, uniformly over $[0, 1]$.

Example 3.3.10. *For the AR(1) of Example 3.3.6, the rank-based central sequence can be derived along the same lines as in Hallin and Puri [41]. In particular* $\Delta^{(n)}(\theta, g) = \sqrt{n}\sum_{i=1}^{n-1}\theta^i r_{g,i}^{(n)}$, *where* $r_{g,i}^{(n)} := 1/n\sum_{t=i+1}^{n}\frac{\dot{g}}{g}(Z_t)Z_{t-i}(\theta)$. *The function* $r_{g,i}^{(n)}$ *can be rewritten as a function of the ranks as:*

$$r_{g,i}^{(n)} := \frac{1}{n-i}\sum_{t=i+1}^{n}\frac{\dot{g}}{g}\left(G^{-1}\left(\frac{R_t^{(n)}}{n+1}\right)\right)G^{-1}\left(\frac{R_{t-i}^{(n)}}{n+1}\right). \tag{3.3.21}$$

Thus,

$$\Delta^{(n)}(\theta, g) := \sqrt{n}\sum_{i=1}^{n-1}\theta^{i-1}r_{g,i}^{(n)}.$$

Different specifications for the reference density g lead to different rank-base scores.

Example 3.3.11. *Let us consider the class of non-gaussian Ornstein-Uhlenbeck processes defined by Barndorff–Nielsen and Shephard [8] through the stochastic differential equation:*

$$dX_t = -\lambda X_t dt + dL_t(\lambda) \tag{3.3.22}$$

where $\lambda \in \mathbb{R}^+$, *and* L_t *is the so called Background Driving Lévy Process (BDLP). We assume that daily observations of* X_t *are given. For some BDLP specifications, the SDE in Eq. (3.3.22) admits an exact discrete-time representation and MLE can be applied for drawing inference about* λ. *In a different direction, we here follow Hallin at al. [38], and we introduce a semi-parametric extension which takes the regression-autoregression form:*

$$X_t = \exp(-\lambda t)X_{t-1} + \sqrt{\frac{1-\exp(-2\lambda)}{\lambda}}\varepsilon_t, \tag{3.3.23}$$

where $\varepsilon_t \sim g \in \mathcal{G}$. *In this much more general discrete-time model, the BDLP is not specified and it can have any density*[2] *belonging to the class* $\mathcal{G}$. *The central sequence for the Markov process in Eq. (3.3.23) is given in formula (3.4) of Hallin et al. [38]. Then, a rank-based version of the central sequence can be obtained using the rank-based score in Eq. (3.3.21).*

[2] *We remark that the semi-parametric model contains all discretized versions of Eq. (3.3.22), including those for which the Lèvy process* L_t *contains jumps, such as the compound Poisson process. However, it is not guaranteed that for every density g there exists a Lèvy process such that discretizing Eq. (3.3.22) leads to innovations* ε_t *that have density g. See Hallin et al. [38] for a discussion.*

3.4 Semi-parametric rank-based procedures

3.4.1 Rank-based test

The null hypotheses we are interested in are general linear hypotheses, under which θ belongs to some linear restriction of Θ. More precisely, the null hypotheses are characterized by a $(p\times r)-$matrix Ω having full rank $r \leq p$, and by an element θ_0 in $\mathbb{R}^p$. Let us denote by $\mathcal{M}(\Omega)$ the $r-$dimensional subspace of $\mathbb{R}^p$ spanned by the columns of Ω. We consider the hypothesis under which $\theta - \theta_0$ belongs to $\Theta \cap \mathcal{M}(\Omega)$, namely it satisfies a set of $p-r$ linearly independent constraints on θ. For a given $g \in \mathcal{G}$, the null is thus given by:

$$H_g^{(n)}(\theta_0;\Omega) := \left\{ Q_{\theta,g}^{(n)} | g \in \mathcal{G}, \theta - \theta_0 \in \Theta \cap \mathcal{M}(\Omega) \right\}.$$

In our semi-parametric framework the innovation density belongs to $\mathcal{G}$, thus we consider the following semi-parametric null hypothesis:

$$\mathcal{H}^{(n)}(\theta_0;\Omega) := \bigcup_{g\in\mathcal{G}} H_g^{(n)}(\theta_0;\Omega). \tag{3.4.1}$$

The goal of this section is to define the most stringent rank-based test for $\mathcal{H}^{(n)}(\theta_0;\Omega)$. To this end, we recall the following:

Definition 3.4.1. *Let ϕ_0 belong to thee class $\mathcal{C}$ of tests having α-level over the null $\mathcal{H}^{(n)}(\theta_0;\Omega)$. Let $\mathcal{K}$ denote the alternative hypothesis. The regret is defined as:*

$$r_{\phi_0}(\theta) = \sup_{\phi\in\mathcal{C}} E_\theta(\phi) - E_\theta(\phi_0) \tag{3.4.2}$$

and a test $\psi^* \in \mathcal{C}$ is defined most stringent within $\mathcal{C}$ and against $\mathcal{K}$, if its maximal (over $\mathcal{K}$) regret is minimal (over $\mathcal{C}$):

$$\sup_{\theta\in\mathcal{K}} r_{\phi^*}(\theta) \leq \sup_{\theta\in\mathcal{K}} r_\phi(\theta), \forall \phi \in \mathcal{C}.$$

The last definition simply implies that the most stringent test has the minimal maximal loss of power resulting from using ϕ^* rather than any other test in $\mathcal{C}$; see Wald [84].

To illustrate the idea of most stringent test, let us consider the typical problem of testing $\Gamma(\theta_0, f, g)\tau = 0$ in Eq. (3.3.15), which corresponds to the null $\mathcal{H}^{(n)}(\theta_0;\Omega) : \tau \in \mathcal{M}(\Omega)$ versus the alternative $\mathcal{K} : \tau \notin \mathcal{M}(\Omega)$. In the exactly specified parametric setting (i.e. for a fix $f \equiv g$), the ULAN can be applied in order to find the most stringent test, which consists in rejecting the null whenever:

$$T_g^{(n)} = \mathbf{\Delta}^{(n)}(\theta, g)' \left[\Gamma(\theta, g) - \Omega \left(\Omega'\Gamma(\theta, g)\Omega \right)^{-1} \Omega' \right] \mathbf{\Delta}^{(n)}(\theta, g) > \chi^2_{p-r;1-\alpha} \tag{3.4.3}$$

where $\mathbf{\Delta}^{(n)}(\theta, g) \sim N(\mathbf{0}; \Gamma(\theta, g))$ (under the null) and where $\chi^2_{p-r;1-\alpha}$ is the $(1-\alpha)-$percentile of a chi-square distribution having $(p-r)$-degrees of freedom.

When g is the Gaussian, Eq. (3.4.3) yields the classical Gaussian Lagrange Multiplier (LM) test, which has α-validity and is locally and asymptotically most stringent against $\mathcal{H}^{(n)}(\theta_0; \Omega)$ (see Hallin and Werker [44]). When g is different form the Gaussian, the ULAN approach allows for well identified asymptotic optimality properties, and provides explicit informations about local powers and asymptotic relative efficiency. Moreover, it opens the door to rank-based tests which can be more efficient than LM test.

A natural rank-based version of the Gaussian LM test in Eq. (3.4.3), is clearly obtained replacing the central sequence $\mathbf{\Delta}^{(n)}(\theta, g)$ by $\mathbf{\Delta}^{(n)}(\theta, g)$:

$$T^{(n)}_g(\theta) = \mathbf{\Delta}^{(n)}(\theta, g)' \left[\Gamma(\theta, g) - \Omega \left(\Omega' \Gamma(\theta, g) \Omega\right)^{-1} \Omega'\right] \mathbf{\Delta}^{(n)}(\theta, g) > \chi^2_{p-r;1-\alpha}. \tag{3.4.4}$$

Since θ remains unspecified, the test has no practical interest, and we should replace θ by an appropriate estimate. To this end, we introduce the following additional assumptions: **H.** There exists a root-n consistent, locally discrete and constrained (such that $\hat{\theta}^{(n)} - \theta_0 \in \mathcal{M}(\Omega)$) estimator $\hat{\theta}^{(n)}$ of θ. Assumption **H.** is very mild, since in our Markov setting, we have a plenty of root-n consistent estimators, like the M-estimators implied by PML, MM, and MEF. Thus, the definition of such an estimator is not a theoretical issue.

Under the additional assumption that $\|\tau_n\| < M$, from the ULAN and from the Third Le Cam's Lemma, it follows that:

$$\sup_{\|\tau_n\| < M} |\mathbf{\Delta}^{(n)}(\theta + n^{-1/2}\tau_n, g) - \mathbf{\Delta}^{(n)}(\theta, g) + \Gamma(\theta, g, f)\tau_n| = o_p(1) \tag{3.4.5}$$

as $n \to \infty$, with $M < \infty$ and under $H^{(n)}_f(\theta)$.

In spite of the mild assumptions **A.**-**H.**, the replacement of θ in Eq. (3.4.4) by a root-n consistent estimator determines the so called aligment problem (see Hallin and Puri [42]). The aligned ranks are the ranks obtained from the estimated residuals $\hat{Z}_t := Z_t(\hat{\theta}^{(n)})$. Since $\hat{\theta}^{(n)}$ is a function of all estimated residuals, the main trouble with aligned ranks is that the iid structure of $\hat{Z}_t$, $t = 1, ..., n$ is lost. This implies that

also the invariance property is lost and there is no theoretical reason for using the ranks. Nevertheless, in our setting, is possible to solve this problem, showing that the test based on aligned-ranks is *asymptotically invariant*, that is it is asymptotically equivalent to a distribution-free and genuinely invariant rank-based test. To this end we first need the following:

Lemma 3.4.2. *Under Assumption **A**-**H**, we have:*

$$\mathbf{\Delta}^{(n)}(\hat{\theta}^{(n)}, g) - \mathbf{\Delta}^{(n)}(\theta, g) + \Gamma(\theta, g, f)\sqrt{n}(\hat{\theta}^{(n)} - \theta) = o_p(1) \tag{3.4.6}$$

under $H_f^{(n)}(\theta)$ with $\Gamma(\theta, g, f)$ as in Eq. (3.3.16).

Proof. See Appendix A. □

The previous Lemma shows that $\mathbf{\Delta}^{(n)}(\hat{\theta}^{(n)}, g) - \mathbf{\Delta}^{(n)}(\theta, g)$ is not distribution-free, since the quantity $\Gamma(\theta, g, f)$ depends both upon g and f. Nevertheless is possible to apply the result of Lemma 3.4.2 in order to define a test, based on the aligned-ranks and that is asymptotically equivalent to the test in Eq. (3.4.4). To achieve this goal, we need an additional assumption: **F1'.** In the Gaussian shift experiment, $\Gamma(\theta, g, f)$ can be rewritten as

$$\Gamma(\theta, g, f) = \left(\begin{array}{c|c} I_1(g, f)\Gamma_1(\theta) & \mathbf{0} \\ \hline \mathbf{0} & I_2(g, f)\Gamma_2(\theta) \end{array}\right) \tag{3.4.7}$$

where $I_1(g, f)$ and $I_2(g, f)$ are positive scalars depending only on g and f, whereas $\Gamma_1(\theta)$ and $\Gamma_2(\theta)$ are functions of θ, as in Assumption **F1**. Then, we can show the following:

Proposition 3.4.3. *Let $f \in \mathcal{G}$ and let us consider the semi-parametric hypothesis $\mathcal{H}^{(n)}(\theta_0; \Omega)$. We reject it whenever:*

$$T_g^{(n)}(\hat{\theta}^{(n)}) := \Delta^{(n)}(\hat{\theta}^{(n)}, g)' \left[\Gamma(\hat{\theta}^{(n)}, g) - \Omega \left(\Omega'\Gamma(\hat{\theta}^{(n)}, g)\Omega\right)^{-1} \Omega'\right] \Delta^{(n)}(\hat{\theta}^{(n)}, g) > \chi^2_{p-r;1-\alpha}. \tag{3.4.8}$$

The test $T_g^{(n)}(\hat{\theta}^{(n)})$:

- *(i) is asymptotically invariant (distribution-free) with size α;*
- *(ii) is asymptotically and locally most stringent in the class of tests having validity α;*
- *(iii) under the contiguous alternative $H_f^{(n)}(\theta + n^{-1/2}\tau_n)$, with $\theta - \theta_0 \in \mathcal{M}(\Omega)$ and $\tau_n \notin \mathcal{M}(\Omega)$, is asymptotically non-central chi-square, with non-centrality parameter ρ:*

$$\left(\begin{array}{c|c} I_{ll}^{-1}(g) I_1(g, f)^2 \Gamma_1^{-1}(\theta) & \mathbf{0} \\ \hline \mathbf{0} & I_{ss}^{-1}(g) I_2(g, f)^2 \Gamma_2^{-1}(\theta) \end{array}\right) \tag{3.4.9}$$

Proof. See Appendix A. □

The last Proposition shows that although $\Delta^{(n)}(\hat{\theta}^{(n)}, g)$ fails to be either asymptotically invariant or asymptotically distribution-free, the quadratic form $T_g^{(n)}(\hat{\theta}^{(n)})$ defined by means of the aligned ranks features all the desirable properties of the genuine rank-based test $T_g^{(n)}(\theta)$.

3.4.2 R-estimator

Construction

The rank-based test defined in the previous Section can be applied to define a R-estimator for θ. In particular, given $\mathbf{\Delta}^{(n)}(\theta, g)$, an estimator for θ can be found solving the minimization problem:

$$\tilde{\theta}_n := Argmin_{\theta\in\Theta}\|\mathbf{\Delta}^{(n)}(g,\theta)\| \quad \textit{or equivalently} \quad \mathbf{\Delta}^{(n)}(g,\tilde{\theta}_n) = 0. \tag{3.4.10}$$

where $\|\cdot\|$ represents the Euclidean norm.

In spite of the simplicity of their definition, R-estimators are unfortunately infamous. Rank-based tests are commonly applied in statistics and they are extremely attractive for their distribution-freeness and for their optimality properties. In contrast, R-estimators are almost unknown. The main reason is related to the non-convex form of the estimating equations derived from Eq. (3.4.10), a feature that can cause practical implementation problems (e.g., multiple solutions or inconsistent roots). We here propose a more convenient way to define an R-estimator.

One-step estimator. Our goal is to derive a root-n consistent estimator $\tilde{\theta}_n$ such that $n^{1/2}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(\mathbf{0}, \Gamma^{-1}(\theta,g,f)\Gamma(\theta,g)\Gamma^{-1}(\theta,g,f))$. To define such an estimator, we propose an approach that goes back to Le Cam and that relies on the so-called one-step estimation procedure.

Proposition 3.4.4. *Under the Assumptions **A**-**H**, the estimator defined by:*

$$\tilde{\theta}_n = \hat{\theta}^{(n)} + n^{-1/2}\Gamma^{-1}(\hat{\theta}^{(n)}, g, f)\mathbf{\Delta}^{(n)}(\hat{\theta}^{(n)}, g) \tag{3.4.11}$$

is root-n consistent and asymptotically normal, with distribution $N(\mathbf{0}, \Gamma^{-1}(\theta,g,f)\Gamma(\theta,g)\Gamma^{-1}(\theta,g,f))$.

Proof. See Appendix A. □

The one-step M-estimator defined in Eq. (3.4.11) is an *Asymptotic Generalized M-estimator* (see Bickel et al. [13]). Proposition 3.4.4 shows that $\tilde{\theta}_n$ is root-n consistent and asymptotically normal. Notice that, by construction, $\tilde{\theta}_n$ achieves the semi-parametric efficiency bound (at a given g) implied by the ULAN.

Cross-information quantities. An important task related to the implementation of our estimation and testing method is that to compute the power of the test in Eq. (3.4.9) and to calculate the one-step estimator in Eq. (3.4.11), we need a consistent estimator of $\Gamma(\theta, g, f)$. This is typically a complicated issue, since the computation involves the integration wrt f which is unspecified. Nevertheless, from Assumption **F1'**, we know that the matrix can be alternative re-written as:

$$\left(\begin{array}{c|c} I_1(g,f)\Gamma_1(\theta) & \mathbf{0} \\ \hline \mathbf{0} & I_2(g,f)\Gamma_2(\theta) \end{array}\right) := \left(\begin{array}{c|c} I_1(g,f)I_{p_1\times p_1} & \mathbf{0} \\ \hline \mathbf{0} & I_2(g,f)I_{p_2\times p_2} \end{array}\right) \Upsilon(\theta) \tag{3.4.12}$$

where $I_{p_1\times p_1}$ and $I_{p_2\times p_2}$ are the identity matrix with dimension $p_1 \times p_1$ and $p_2 \times p_2$ respectively. If a consistent estimator for the scalars $I_1(g,f)$ and $I_2(g,f)$ ara available, then the one-step estimator would be defined by:

$$\tilde{\theta}_n = \hat{\theta}^{(n)} + n^{-1/2}\Upsilon(\theta)^{-1} \left(\begin{array}{c|c} \hat{I}_1^{-1}(g,f)I_{p_1\times p_1} & \mathbf{0} \\ \hline \mathbf{0} & \hat{I}_2^{-1}(g,f)I_{p_2\times p_2} \end{array}\right) \mathbf{\Delta}^{(n)}(\hat{\theta}^{(n)}, g) \tag{3.4.13}$$

Following Cassart et al. [16], a consistent estimator of $\hat{I}_1^{-1}(g,f)$ and $\hat{I}_2^{-1}(g,f)$ can be obtained as follows:

- Writing

$$\tilde{\theta}_n(\lambda_1,\lambda_1) = \hat{\theta}^{(n)} + n^{-1/2}\Upsilon(\theta)^{-1} \left(\begin{array}{c|c} \lambda_1 I_{p_1\times p_1} & \mathbf{0} \\ \hline \mathbf{0} & \lambda_2 I_{p_2\times p_2} \end{array}\right) \mathbf{\Delta}^{(n)}(\hat{\theta}^{(n)}, g). \tag{3.4.14}$$

- Let $(\hat{I}_1^{-1}(g,f), \hat{I}_2^{-1}(g,f)) = (\lambda_{1,*}^n, \lambda_{2,*}^n)$, where

$$(\lambda_{1,*}^n, \lambda_{2,*}^n) := \inf_{(\lambda_1,\lambda_2)\in\mathbb{R}^+\times\mathbb{R}^+} \left\{\lambda_1,\lambda_2 | \mathbf{\Delta}^{(n)}(\hat{\theta}^{(n)})'\Upsilon(\hat{\theta}^{(n)})\Upsilon(\tilde{\theta}_n(\lambda_1,\lambda_2))\mathbf{\Delta}^{(n)}(\tilde{\theta}_n(\lambda_1,\lambda_1)) < 0\right\}. \tag{3.4.15}$$

Remark 3.4.5. *In order to compute the one-step estimator or the cross-information quantities, we need a preliminary root-n consistent estimator of θ. This estimator can be considered as a starting point for a Newton-Raphson procedure and it should be as close as possible to θ. It is a good idea to use as starting point a robust estimator of θ. In our setting, where the first two conditional moments (location and scale) are specified, a preliminary M-estimator could obtained by (Robust) PMLE in dynamic location and scale models. See Gourieroux et al. [36], for PMLE, and see Mancini et al. [66] for its robust version.*

Robustness features

The distribution freeness of the ranks implies that $\mathbf{\Delta}^{(n)}(\theta, g)$ defines a root-n consistent estimator also when the actual density is different from the reference density. This is due to fact that the rank-based central

sequence has expectation equal to zero, even if $g \neq f$, and $f, g \in \mathcal{G}$ (see Eq. (3.3.13)). Besides that global robustness feature, it is interesting to understand also the local robustness properties of $\tilde{\theta}_n$. To this end, we study the infinitesimal robustness of the estimator in Eq. (3.4.11).

Our R-estimator is derived under the model specification in Eq. (3.2.3) and under the Assumptions **A**-**K**. In spite of the generality of these assumptions, it is possible that the actual density does not match exactly the theoretical requirements. Indeed, we can observe some failures either in the parametric assumptions or in the non-parametric part of the model. For instance, violations of the parametric assumptions arise when a fraction of the data does not follow exactly the conditional moments in Eq. (3.2.1) and Eq. (3.2.2). Violations of the non-parametric assumptions can be due to some anomalous observations generated according to a distribution not satisfying Assumptions **A**-**K**. Thus, it is possible that for the majority of the data have a density belonging to $\mathcal{G}$, while a small number of observations has a different distribution. The goal of infinitesimal robustness is to understand the impact that these outliers have on $\tilde{\theta}_n$.

To formalize mathematically the ideas, let us consider the joint marginal density:

$$\mathbb{Q}_{\theta,g}(\mathbf{dy_{t-1}}, dy_t) := \pi(\mathbf{dy_{t-1}}, \theta) Q_{\theta,g}(dy_t; \mathbf{y_{t-1}}). \tag{3.4.16}$$

Following the infinitesimal approach of Hampel et al. [49], we define the contamination neighborhoods:

$$\mathcal{U}_\varepsilon(\mathbb{Q}_{\theta,g}) = \{(1-\varepsilon)\mathbb{Q}_{\theta,g} + \varepsilon\mathbb{P}, \varepsilon < b, b \in [0,1]\} \tag{3.4.17}$$

for $\theta \in \Theta$, $g \in \mathcal{G}$ and $\mathbb{P}$ is a contaminating measure belonging to the class $\mathcal{M}_{stat}$, representing the family of marginals of stationary Markov processes (see Künsch [59] and La Vecchia and Trojani [60] for a discussion about neighborhoods as in Eq. (3.4.17) in Markov setting). In order to study the robustness properties of our rank-based procedures, we assume that the actual density f belongs to the neighdorhoods $\mathcal{U}_\varepsilon(\mathbb{Q}_{\theta,g})$. The interpretation is that f is a contaminated version of a distribution having density $g \in \mathcal{G}$ and conditional moments as in Eq. (3.2.1) and Eq. (3.2.2).

The main tool that we need to study the robustness property of R-estimators implied by rank-based

scores is the Influence Function (IF), which provides us with a first order approximation of the asymptotic bias of estimators, when the actual distribution belongs to $\mathcal{U}_\varepsilon(\mathbb{Q}_{\theta,g})$. In the iid setting the IF is proportional to the score a_g; see Bickel et al. [13, page 19]. A similar result holds in time-series, nevertheless we recall that differently from the iid setting, in time-series the IF rises uniqueness problems. See, e.g., Künsch [59]. To overcome this problem, we consider the version of IF satisfying the additional requirement: $E_g[IF^{cond}(\mathbf{a}_g^{(n)}(\mathbf{R}^{(n)}(\mathbf{Z_t}^{(s)});\theta))|\mathbf{Z_{t-1}^{(s)}}] = 0$ and label this version: Conditional Influence Function (IF^{cond}). Künsch shows existence and uniqueness of IF^{cond} for strictly stationary AR(q) processes. La Vecchia and Trojani [60] extend Künsch's result to general (linear and non linear) stationary Markov processes. The latter result can be here applied to the estimator in Eq. (3.4.11). In particular, the rank-based score function $\mathbf{\Delta}^{(n)}(\theta,g)$ is a L_2-martingale difference (up to $o_p(1)$, see, Eq. (3.3.11) and Example 3.3.6), satisfying $E_g[IF^{cond}(\mathbf{a}_g^{(n)}(\mathbf{R}^{(n)}(\mathbf{Z_t}^{(s)});\theta))|\mathbf{Z_{t-1}^{(s)}}] = 0$. Thus, the existence and uniqueness of $IF^{cond}(\mathbf{a}_g^{(n)}(\mathbf{R}^{(n)}(\mathbf{z_t}^{(s)});\theta))$ follows directly from Proposition 4, in La Vecchia and Trojani [60].

When $\mathbb{P} = \delta_{\mathbf{z_t^{(s)}}}$, where $\mathbf{z_t^{(s)}} \in \mathbb{R}^{t-s}$ as in Eq. (3.3.18), a standard computation implies that the R-estimator defined by $\mathbf{\Delta}^{(n)}(\theta,g)$ has:

$$IF^{cond}(\mathbf{z}_t^{(s)};\theta) := IF^{cond}(\mathbf{a}_g^{(n)}(\mathbf{R}^{(n)}(\mathbf{z}_t^{(s)});\theta)) = -M(\theta,g)\mathbf{a}_g^{(n)}(\mathbf{R}^{(n)}(\mathbf{z}_t^{(s)});\theta), \tag{3.4.18}$$

where $M(\theta,g) = E_g[\nabla_\theta \mathbf{a}_g^{(n)}(\mathbf{R}^{(n)}(\mathbf{Z}_t^{(s)});\theta)]$. Now, Eq. (3.4.18) implies that the IF is uniformly bounded over S iff the score generating function $\mathbf{a}_g^{(n)}$ is uniformly bounded. A similar result has been obtained by Ronchetti and Yen [74], for R-estimators in the iid one-dimensional pure location setting.

Analogous considerations hold also for the test $T_g^{(n)}(\hat{\theta}^{(n)})$. The test is globally robust for $g \neq f$ and $g, f \in \mathcal{G}$, nevertheless it could be not infinitesimally robust when $f \in \mathcal{U}_{\varepsilon/\sqrt{n}}(\mathbb{Q}_{\theta,g})$: unbounded scores can determine arbitrarily large distortions in the size and/or in the power. See Rieder for an analysis of infinitesimally robust rank-based tests [71].

3.5 Numerical analysis

We illustrate the performance of our testing and estimation procedure described in the previous sections, by means of several numerical analyses. The data generating process is an AR(3):

$$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \theta_3 Y_{t-3} + \varepsilon_t, \tag{3.5.1}$$

where ε_t, iid $\sim g(0, \sigma^2)$. We assume that g satisfies Assumptions **A**-**C**, thus the model in Eq. (3.5.1) has dynamics as in Eq. (3.2.3). For estimation and testing about θ_1, θ_2, and θ_3, we consider four rank-based scores:

- van der Waerden (associated with a Gaussian reference density f for the errors):

$$r^{(n)}_{\mathrm{vdW};i} = \left(s^{(n)}_{vdW}\right)^{-1} \left\{ (n-i)^{-1} \sum_{t=i+1}^{(n)} \Phi^{-1}\left(\frac{R^{(n)}_t}{n+1}\right) \Phi^{-1}\left(\frac{R^{(n)}_{t-i}}{n+1}\right) - m^{(n)}_{vdW} \right\}, \tag{3.5.2}$$

  where Φ is the usual notation for the standard normal distribution function;

- Wilcoxon (associated with a logistic reference density f for the errors):

$$r^{(n)}_{W;i} = \left(s^{(n)}_{W}\right)^{-1} \left\{ (n-i)^{-1} \sum_{t=i+1}^{(n)} \left(\frac{R^{(n)}_t}{n+1} - \frac{1}{2}\right) \log\left(\frac{R^{(n)}_{t-u}}{n+1-R^{(n)}_{t-i}}\right) - m^{(n)}_{W} \right\}; \tag{3.5.3}$$

- Laplace (associated with a double-exponential reference density f for the errors):

$$\begin{aligned} r^{(n)}_{L;i} = \left(S^{(n)}_{L}\right)^{-1} \Bigg\{ (n-i)^{-1} \sum_{t=i+1}^{(n)} \mathrm{sign}\left(\frac{R^{(n)}_t}{n+1} - \frac{1}{2}\right) \left[\log\left(2\frac{R^{(n)}_{t-u}}{n+1}\right) I\left[\frac{R^{(n)}_{t-u}}{n+1} \le \frac{1}{2}\right]\right] \\ - \left[\log\left(2 - 2\frac{R^{(n)}_{t-u}}{n+1}\right) I\left[\frac{R^{(n)}_{t-u}}{n+1} > \frac{1}{2}\right]\right] - m^{(n)}_{L} \Bigg\} \end{aligned} \tag{3.5.4}$$

  where $I[\cdot]$ is the indicator function;

- Sign autocorrelations:

$$r^{(n)}_{\mathrm{SA};i} = \left(s^{(n)}_{SA}\right)^{-1} \left\{ (n-i)^{-1} \sum_{t=i+1}^{(n)} \mathrm{sign}\left(\frac{R^{(n)}_t}{n+1} - \frac{1}{2}\right) \mathrm{sign}\left(\frac{R^{(n)}_{t-i}}{n+1} - \frac{1}{2}\right) - m^{(n)}_{SA} \right\}, \tag{3.5.5}$$

In Eq. (3.5.2)-(3.5.5), the quantities $m_i^{(n)}$, and $(s_i^{(n)})^{-1}$ are exact centering and scaling constants, whose explicit expressions can be found in Hallin and Melard [39] . The resulting rank-based central sequence is:

$$\Delta_k^{(n)}(\theta, g) = \sqrt{n} \begin{pmatrix} \sum_{i=1}^{n-1} g_{i-1}(\theta) r_{k;i}^{(n)} \\ \sum_{i=2}^{n-1} g_{i-2}(\theta) r_{k;i}^{(n)} \\ \sum_{i=3}^{n-1} g_{i-3}(\theta) r_{k;i}^{(n)} \end{pmatrix}, \tag{3.5.6}$$

where $g_i(\theta)$ denote the Green's functions associated with Eq. (3.5.1) and $r_{k;i}^{(n)}(\theta)$ are given in Eq. (3.5.2)-(3.5.5), for $k = \text{vdW}, W, L, \text{SA}$. We study the behavior of the implied R-estimators and tests in different contests.

3.5.1 Finite-sample analysis

Robust estimation

In our first Monte Carlo experiment, we analyze the robustness features of a bounded rank-based score in the presence of contamination. Eq. (3.4.18) implies that an infinitesimally robust R-estimator is characterized by a bounded IF^{cond}, namely by a bounded rank-based score. For instance, we notice that the van der Waerden scores in Eq. (3.5.2) are proportional to $y_t = \Phi^{-1}(R_t/n+1)$, so they are defined by an unbounded score. In contrast, the sign autocorrelation scores $r_{\text{SA};i}^{(n)}$ in Eq. (3.5.5) are bounded. Thus, we conclude that the R-estimator implied by $r_{\text{SA};i}^{(n)}$ is infinitesimally robust, while the R-estimator implied by $r_{\text{vdW};i}^{(n)}$ does not have such a feature. Similar conclusions have been recently obtained in the AR(1) case by Boldin [14].

An interesting question remains to compare the stability of the R-estimator implied by $r_{\text{SA};i}^{(n)}$ with the stability yielded by the Gaussian Pseudo Maximum Likelihood estimator (we label it PMLE, Gourieroux et al. [36]), and by its robust version (we label it Robust PMLE). We select these competitors since: (i) PMLE is widely applied in time-series, and it is considered a (globally) robust estimator, because it does not completely specify the transition density; (ii) Robust PMLE is based on a truncated (i.e., bounded) score, and it is widely applied in robust estimation for time-series since it is an infinitesimally robust M-estimator[3].

We investigate numerically the degree of robustness of the three different estimators, and we deal with

[3]To compute Robust PMLE, we apply a huberized version of Gaussian PMLE, using a bounding constant of 1, which ensures an efficiency loss at the reference model of about 10%.

three kinds of AR-process. The A-type process is the clean process generated according to Eq. (3.5.1), where $\theta_1 = 0.3, \theta_2 = 0.2, \theta_3 = 0.1$, and $\varepsilon_t^A \sim N(0,1)$. The B-type process has a A-type trajectory contaminated by ten Replacement Outliers (RO), having values $2 + \varepsilon_t^B$, with $\varepsilon_t^B = 12\varepsilon_t^A$. Being located around the point 2, these patchy outliers stress mainly the mass in the right tails of the transition density and increase the variability. Finally, the C-type process has also a A-type trajectory, but it has different contamination: the trajectories have ten patchy outliers generated as in Eq. (3.5.1), where $\theta_1 = \theta_2 = \theta_3 = 0$, and errors $\varepsilon_t^C = 2\varepsilon_t^A$. These contaminations have probability mass around 0, thus they change the long-run mean and the variance of the AR(3) process. For each type of process, the sample size is $n = 150$. In Figure 3.2, we plot one simulated trajectory for each type of process. We notice that the trajectories of B- and C-type processes are characterized by some more pronounced spikes in the first half of the path, but they are identical to the A-type trajectory from observation 75 to 150.

In Table 3.1, we show the empirical MSE for a Monte Carlo simulation having size 4000. The last column shows the MSE for the PMLE for A-, B-, and C-type process. The other columns give the ratio between the MSE of the estimator implied by $r_{\mathrm{SA};i}^{(n)}$ (first column) and by Robust PML (second column), versus the MSE of the PMLE. Robust PMLE has been applied as preliminary root-n consistent estimator in the one-step formula.

The first row refers to A-type process. In this setting, the Gaussian PMLE coincides with the classical Maximum Likelihood Estimator (MLE). The analysis of this case is important in order to gauge the trade-off robustness/efficiency. We notice that the efficiency loss of the R-estimator is about 41%, since the MSE for $r_{\mathrm{SA};i}^{(n)}$ is about 159% the MSE of the MLE. The Robust (P)MLE implies a smaller efficiency loss: about 10%. The higher cost in term of efficiency for the R-estimator yields a larger degree of robustness for the C-type process: the MSE of $r_{\mathrm{SA};i}^{(n)}$ is about one-third the MSE of the (P)MLE and about two-third the MSE of Robust (P)MLE. For B-type process, Robust PLE and R-estimator have similar gain in robustness, with a small advantage (about 10%) for the Robust (P)MLE.

To visualize our findings, Figure 3.2 shows the boxplot for the three estimators for A- (first row), B- (second row), and C-type (third row) process. The first column refers to θ_1, the second column is for θ_2 and the third column is for θ_3. The dotted line in each boxplot represents the true value of the parameter. Looking at the first row, we notice that for A-type process all the estimators are unbiased, but the R-estimator has a larger variability than the others estimators: the larger variance represents the efficiency cost. However, an inspection of the third row (C-type process) shows that both (P)MLE and Robust (P)MLE have typically larger bias and larger variance than the R-estimator. For instance, this can be noticed looking at the first plot in the third row: for the estimation of θ_1, both (P)MLE and Robust (P)MLE show a large downward bias and a large variability. In contrast, the R-estimator implied by $r^{(n)}_{\mathrm{SA};i}$ has a negligible bias and the smallest variability. Analogous considerations hold for the estimation of θ_2, in the middle panel of the third row. Finally, looking at θ_3 (third panel of the third row), we notice that: (i) Robust (P)MLE shows a small bias, while the R-estimator is almost unbiased; (ii) Robust (P)MLE has a slightly smaller variance than the R-estimator; (iii) (P)MLE has a small bias, but the largest variance.

Summarizing, we notice that in the presence of contamination, the R-estimator implied by $r^{(n)}_{\mathrm{SA};i}$ and Robust (P)MLE achieve a better performance than the (P)MLE. The latter is typically biased and shows a large variability. For some kinds of contamination (e.g., C-type process), $r^{(n)}_{\mathrm{SA};i}$ scores imply a better stability, yielding a MSE lower that the one of Robust (P)MLE. For some other kinds of contamination (e.g., B-type process), Robust (P)MLE achieves a MSE slightly better than the one of $r^{(n)}_{\mathrm{SA};i}$.

Estimation for leptokurtic distributions

In a second Monte Carlo experiment, we illustrate the finite performance of the R-estimators implied by $r^{(n)}_{\mathrm{vdW};i}$, $r^{(n)}_{W;i}$, and $r^{(n)}_{L;i}$, when the actual error distribution is leptokurtic. We assume that the errors in the Eq. (3.5.1) have: Laplace distribution, t-distribution with 6, and with 8 degrees-of-freedom. The sample size is $n = 100$ and the Monte Carlo size is 4000. The values of the parameters are $\theta_1 = 0.3, \theta_2 = 0.2$, and $\theta_3 = 0.1$.

In Table 3.2, we show the empirical MSE for the considered R-estimators and for other two widely applied benchmarks: Gaussian PMLE and Robust PMLE. From Table 3.2, we notice that the Wilcoxon R-estimator over-performs all the other estimators, when the underlying error distribution is either a Laplace (first column) or a t-distribution with 8 degrees-of-freedom (last column). Additionally, we notice that, for any considered error distribution, both Wilcoxon and van der Waerden R-estimators have a smaller MSE than Robust PMLE.

Figure 3.3 shows the corresponding boxplots, which allow to disentangle the bias and the variance component in the MSE of each estimator. We notice that, when the actual density is a Laplace (first row), van der Waerden R-estimator has a smaller variance than Robust PMLE and similar bias to Robust PMLE. Differently, when the actual distribution is a t-distribution with 8 degrees-of-freedom (third row), the R-estimator implied by van der Waerden scores has both smaller bias and smaller variance than Robust PMLE, for each estimated parameter.

3.5.2 Asymptotic analysis: ARE of rank-based tests vs LM test

In the third numerical exercise, we compute the Asymptotic Relative Efficiency (ARE) of the scores $r^{(n)}_{\mathrm{vdW};i}, r^{(n)}_{W;i}$, and $r^{(n)}_{L;i}$, versus the Gaussian score-based test. The latter is one of the most classical ways for testing in AR(p) models and yields the so-called LM method, based on:

$$r^{(n)}_{\mathcal{N};i} = (n-i)^{-1} \sum_{t=u+1}^{(n)} \frac{Z_t(\theta) Z_{t-i}(\theta)}{n^{-1} \sum_{t=1}^{n} Z_t^2(\theta)}. \tag{3.5.7}$$

In our AR(3) setting, the Gaussian central sequence is obtained as in Eq. (3.5.6), replacing $r^{(n)}_{k;i}$ by $r^{(n)}_{\mathcal{N};i}$. Moreover, the non-centrality parameter of the test $T^{(n)}_g(\hat{\theta}^{(n)})$ as in Eq. (3.4.9) reads:

$$\left[\sigma(f;g)^2 \mathcal{I}_f^2(g;f)/\mathcal{I}_f(g)\right] \mathcal{Q}(\theta,\tau). \tag{3.5.8}$$

The cross-information quantity is given by the product of

$$\sigma(f;g) = \int_0^1 F^{-1}(u) G^{-1}(u) du,$$

and

$$\mathcal{I}_f(g;f) \quad = \int_{-\infty}^{\infty} \frac{\dot{g}}{g}(y)\,\frac{\dot{f}}{f}(y)dy = \int_0^1 \frac{\dot{g}}{g}\left(F^{-1}(u)\right)\frac{\dot{f}}{f}\left(F^{-1}(u)\right)du.$$

With obvious notation, we have that $\mathcal{I}_f(g) = \mathcal{I}_f(g;g)$. The quantity $\mathcal{Q}(\theta,\tau)$ in Eq. (3.5.8) is a positive definite quadratic form depending on θ and τ, but not on f and g and it coincides with the non-centrality parameter of the asymptotic chi-square distribution of the Gaussian LM test under local alternatives of the form $H_f^{(n)}(\theta + n^{-1/2}\tau_{\mathbf{n}})$. Thus, it follows that the ARE, with respect to LM under innovation density f, of $T_g^{(n)}(\hat{\theta}^{(n)})$ is simply $\left[\sigma(f;g)^2\mathcal{I}_f^2(g;f)/\mathcal{I}_f(g)\right]$.

In Table 3.3, we compute the AREs for $r_{\mathrm{vdW};i}^{(n)}, r_{W;i}^{(n)}, r_{L;i}^{(n)}$. We use different actual densities f. The middle column refers to $N(0,1)$. The first four columns consider Leptokurtic distributions: Laplace (first column), and three t-distribution with degrees-of-freedom equal to 4 (second column), 8 (third column), and 20 (fourth column). The last four columns refer to the class of skewed-gaussian distributions introduced by Azzalini [6]. A skewed-gaussian distribution is characterized by a parameter a controlling the degree of skewness: values of $a > 0$ imply a right skewed density, while values of $a < 0$ imply left skewed density. The symmetric gaussian case is recovered when $a = 0$. In our calculations, we consider: $a = \pm 1$ (sixth column), $a = \pm 3$ (seventh column), $a = \pm 6$ (eighth column), $a = \pm 12$ (ninth column). From an analysis of the first column (the actual density is a Laplace), we notice that all the three rank-based tests over-perform LM test. When the actual density is a t-distribution with 4 degrees-of-freedom (second column), both Wilcoxon (34%) and Laplace (about 11%) scores have an improvement with respect to LM test, while the van der Waerden scores imply a small efficiency loss of about 7%. For higher degrees-of-freedom (implying smaller values of kurtosis), the improvement with respect to LM test decreases. In a different direction, no improvements with respect to LM test are observable for the considered class of skewed-gaussian distributions (see last four columns).

A joint analysis for both leptokurtic and skewed distributions highlights that Wilcoxon scores (second row) imply a good efficiency gain (about 67%), when the actual distribution is a Laplace, while the maximal

efficiency loss is about 16%, when the actual distribution is a strongly skewed-normal with $a = \pm 12$.

3.6 Real-data: FX-stochastic volatility model

3.6.1 Model setting for the USD/CHF exchange rate

In this last section we implement the R-estimators described in Section 3.5. Our aim is to draw inference for the parameters of a discrete-time stochastic volatility (SV) model for the USD/CHF exchange rate.

SV models are characterized by two equations: the first one describes the log-return dynamics, while the second one is for the volatility dynamics. One of the main issue in SV models is that the volatility is a latent factor. This determines two problems. First, the log-return process is not Markovian (see e.g. Cont and Tankov [18]). Second, the inference and testing procedures for SV models are based on computationally intensive MCMC or Kalman filtering methods (see, e.g., Barndorff-Nielsen and Shephard [9] for a related discussion). To overcome both those problems, we here adopt a different estimation strategy. Following a stream of literature which is becoming popular in financial econometrics (see, e.g., Andersen et al. [5], Barndorff-Nielsen and Shephard [9], Bollerslev and Zhou [15]), we proxy the unobserved volatility by the Two Scales Realized Volatility (TSRV) (see Aït-Sahlia et al. [1]), computed using high-frequency data. Indeed, the TSRV yields a model-free unbiased proxy for the asset volatility. In the following, we call the TSRV, simply Realized Volatility, or in short RV. The benefit of using RV in SV models is two-fold. First, we recover the Markovian structure of the joint process. Second, the volatility process is observable, thus we avoid MCMC or filtering methods.

Now, let us consider the stochastic volatility model in Example 3.3.5. We generalize the model as:

$$Y_t = \sigma_t \varepsilon_t \tag{3.6.1}$$

$$K_t = K(\sigma_t) = m_1(\mathbf{K}_{t-1}, \vartheta) + v_1(\mathbf{K}_{t-1}, \gamma)\omega_t, \tag{3.6.2}$$

where $\theta = (\vartheta, \gamma) \in \mathbb{R}^p$, $\varepsilon_t \sim g_\varepsilon(0,1)$ and $\omega_t \sim g_\omega(0,1)$. The vector $\mathbf{Y_{t-1}}$ contains the past values of Y_t. We additionally assume that $\varepsilon_t = y_t/RV_t \sim N(0,1)$. This assumption can be justified following the

argument in Andersen et al. [4]. More precisely, we notice that the log-returns standardized by the RV are (nearly) Gaussian. This feature is illustrated in the bottom panel of Figure 3.5, which shows that the $N(0,1)$ assumption is appropriate for y_t/RV_t.

As far as the Eq. (3.6.2) is concerned, the vector $\mathbf{K}_{t-1}$ contains the past values of the transformed volatility process. These values are computed approximating σ_t by RV_t, and then applying the transformation

$$K(\sigma_t) \approx K(RV_t) = \begin{cases} (RV_t^{\beta} - 1)\beta^{-1} & \text{if } \beta \neq 0 \\ \log RV_t & \text{if } \beta = 1. \end{cases} \tag{3.6.3}$$

Gonçalves and Meddhai [35] have recently shown that the Box-Cox transformation in (3.6.3) is able to reduce skewness and heteroscedasticity of the transformed process.

In our real-data analysis, we apply (3.6.3) with both $\beta = 0$ (log-transform), and $\beta = -1$ (the so-called RV precision, suggested by Gonçalves and Meddhai [35]). Additionally, we assume that $g_\omega \in \mathcal{G}$, that $\vartheta \in \Theta$ and that the errors in the Eq. (3.6.1) and Eq. (3.6.2) are independent. The latter independence assumption is quite common in SV models on exchange rates, since forex data show a negligible leverage effect (see, e.g., Andersen et al. [5]). The density g_ω is left unspecified and it is treated as a nuisance infinite-dimensional parameter.

Our program is to estimate ϑ in Eq. (3.6.2), using PMLE, Robust PMLE and the rank-based scores in Eq. (3.5.2)-(3.5.5).

Remark 3.6.1. *A class of models similar to Eq. (3.6.1) and Eq. (3.6.2) has been introduced by Andersen [3], with the name of Stochastic Autoregressive Volatility. We notice that, when K is the identity and g_ε is the Gaussian density, the model features a conditional distribution structure similar to the normal variance-mean mixture introduced by Barndorff-Nielsen et al. [7].*

3.6.2 Data description and estimation results

In Figure 3.4, we plot the times-series of log-returns (top panel) and annualized RV (bottom panel), for the period 25 November 1996 to 19 December 2003. In Figure 3.6, we plot also the log-transformed annualized RV (bottom panel, obtained for $\beta = 0$ in Eq. (3.6.3)), and the precision of the RV (top panel, obtained for $\beta = -1$).

We notice that both the transformed volatility series show some anomalous spikes (e.g., around the beginning of 1999 or in the beginning of 2003), due to periods of very high and very low volatility. Moreover, an autoregressive structure is visible for both the time-series. Indeed, a Partial Auto Correlation analysis highlights that both the time-series can be modeled as AR(3) processes. Thus, we set $m(\mathbf{K}_{t-1};\vartheta) = \vartheta_1 K_{t-1} + \vartheta_2 K_{t-2} + \vartheta_3 K_{t-3}$ and $v(\mathbf{K}_{t-1};\vartheta_1) = \gamma$. The object of our inference is the parameter $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3) \in \mathbb{R}^3$ in the conditional mean.

In Table 3.4 we show the estimates for ϑ_1, ϑ_2, and ϑ_3, and their standard errors (in brackets). Six different estimators are considered: PMLE (first row), Robust PMLE (second row), and four R-estimators obtained using the scores in Eq. (3.5.2)-(3.5.5). We notice that all the estimation methods yield similar results for the log-transformed volatility (first three columns). In contrast, a remarkable difference can be noticed for the precision of the RV (last three columns): Robust PMLE, and the four kinds of R-estimator yield very similar estimates, but the PML method yields anomalously small values for all the parameters. Additionally, we notice that Robust PMLE and rank-based methods estimate a decreasing impact on Y_t of its lags, since $\vartheta_1 > \vartheta_2 > \vartheta_3$. In contrast, this relationship is violated by PMLE on the precision of RV. Our conjecture for these anomalous estimates is that PMLE is badly attracted (downward biased) by some outliers in the tail of the transition density. This result is in line with our simulation study.

3.7 Conclusion

The paper introduces a class of rank-based scores and applies them for inference and testing purposes. The developed R-estimators and tests achieve semi-parametric efficiency at a reference model, but they simplify the classical semi-parametric approach, since they avoid the projection onto tangent space. Being distribution-free, define root-n consistent R-estimators and tests having the right power, even if the actual density does not coincide with the reference density. In that sense, rank-based procedures are "globally" robust. Nevertheless, they are not necessarily also infinitesimally robust. This implies that even small

violations of some model assumptions can lead to unreliable estimates and tests. We show that rank-based procedures based on bounded scores feature both infinitesimal and global robustness. Numerical examples illustrate the implementation of our theory. A first Monte Carlo experiment provides evidence of the stability of robust R-estimators in finite sample and in the presence of contamination. A second numerical illustration shows that our rank-based tests can have a better asymptotic efficiency than LM test, when the actual density is lepotokurtic. Finally, a real-data application to the RV of the USD/CHF exchange rate concludes the paper. The real-data example confirms the findings of our Monte Carlo analysis, since Gaussian PMLE seems to be badly attracted by some anomalous observations of the transformed RV process. Robust PMLE and rank-based R-estimators do not show the same weakness.

3.8 Appendix: proofs

Proof of Corollary 3.3.2. The proof follows form Drost et al. [24]. In particular, we must check that the Assumption A-E in Drost et al. [24] hold and the ULAN follows from Theorem 2.1 in Drost et al. [24] (Proposition 3.3.1). The likelihood ratio

$$\Lambda_n := \log \frac{L^{(n)}_{\theta+\tau_n n^{-1/2},g}}{L^{(n)}_{\theta,g}}$$

can be written as

$$\Lambda_n = \sum_{t=1}^{n} l(\eta_0 + W'_{nt}(\tilde{\theta}-\theta))(z_t(\theta)) - l(\eta_0)(z_t(\theta)) + \Lambda^s_n. \tag{3.8.1}$$

where

$$W'_{nt}(\tilde{\theta}-\theta) = \frac{1}{v(\mathbf{y}_{t-1},\theta)}\left(m(\mathbf{y}_{t-1},\tilde{\theta}) - m(\mathbf{y}_{t-1},\theta), v(\mathbf{y}_{t-1},\tilde{\theta}) - v(\mathbf{y}_{t-1},\theta)\right)', \tag{3.8.2}$$

$\eta_0 := (0,1)'$ and

$$\Lambda^s_n = \log k_{\theta+\tau_n n^{-1/2}}(Y_{-q+1},...,Y_0) - \log k_\theta(Y_{-q+1},...,Y_0).$$

For our regression-autoregression model, the expression of $l(\eta)(\cdot)$ in Eq. (3.8.1) is

$$l(\eta)(\cdot) = \log v(\mathbf{y}_{t-1},\theta)^{-1} + \log g((\cdot - m(\mathbf{y}_{t-1},\theta))/v(\mathbf{y}_{t-1},\theta)). \tag{3.8.3}$$

Our Assumption E implies that the impact of the last summand in Eq. (3.8.1) is asymptotically negligible. Thus, Assumption A in Drost et al. [24] is satisfied. Moreover, from Eq. (3.8.3) we notice that $\eta \in \mathbb{R}^2$ and $g(\cdot)$ is a location/scale model, so that also Assumption B in Drost et al. [24] is satisfied, with $m = 2$. Setting $W'_{nt}(\tilde{\theta}-\theta)$ as in Eq. (3.8.2), we obtain the following $p \times 2$ matrix of weights:

$$\begin{aligned} W_t &= \frac{1}{v(\mathbf{y}_{t-1},\vartheta)}\frac{\partial}{\partial_{\vartheta'}}(m(\mathbf{y}_{t-1},\vartheta), v(\mathbf{y}_{t-1},\vartheta))\Big|_{\theta=\theta} \\ &= \frac{1}{v(\mathbf{y}_{t-1},\theta)}(\dot{\mathbf{m}}(\mathbf{y}_{t-1},\theta), \dot{\mathbf{v}}(\mathbf{y}_{t-1},\theta)) \end{aligned} \tag{3.8.4}$$

Thus, we have a $\mathcal{F}_{t-1}$-measurable function satisfying Assumption C and D in Drost at al. [24]. Finally, setting

$$\psi(z_t(\theta)) = -\left(\frac{\dot{g}}{g}(z_t(\theta)), 1 + z_t(\theta)\frac{\dot{g}}{g}(z_t(\theta))\right)' \tag{3.8.5}$$

from Eq. (2.5) in Drost et al. [24], it follows that $\dot{l}_t(\theta, g) = W_t(\theta)\psi(z_t(\theta))$. Finally, the ULAN and Eq. (3.3.4) follow from Proposition 3.3.1.

Proof of Proposition 3.3.7 The proof follows with minor notational changes from Hallin and Werker [43]. Let us label $Y_1^{(n)}, ..., Y_n^{(n)}$ a finite realization of the regression-autoregression model in Eq. (3.2.3). We indicate by $Z_0, ..., Z_s$ arbitrary $(s+1)$-tuple of iid r.v. having density g. Let be

$$Z_t^{(n)}(\theta) := (Y_t^{(n)} - m(Y_{t-1}^{(n)}, ..., Y_1^{(n)}; \theta))v((Y_{t-1}^{(n)}, ..., Y_1^{(n)}; \theta))^{-1}.$$

We assume that $Z_0, ..., Z_s$ are independent from the residuals $Z_t^{(n)}(\theta)$. Moreover, let us define:

$$\mathbf{i}^*(\theta, g, z_t^{(n)}, ..., z_{t-s}^{(n)}) := \mathbf{i}(z_t^{(n)}, ..., z_{t-s}^{(n)}) - E_g\left[\mathbf{i}(\theta, g, Z_t^{(n)}, Z_1, ..., Z_s | Z_t^{(n)} = z_t^{(n)})\right].$$

From Corollary 3.2 in Hallin and Werker [43], it follows that:

$$\Delta^{(n)}(\theta, g) = \frac{1}{\sqrt{n-s}} \sum_{t=s+1}^{n} \mathbf{i}^*(\theta, g, z_t^{(n)}, ..., z_{t-s}^{(n)}) + o_P(1).$$

Furthermore, from Proposition 3.3. in Hallin and Werker [43], it follows:

$$\Delta^{*(n)}(\theta, g) := \frac{1}{\sqrt{n-s}} \sum_{t=s+1}^{n} \mathbf{i}^*(\theta, g, z_t^{(n)}, ..., z_{t-s}^{(n)}). \tag{3.8.6}$$

Proof of Lemma 3.3.9 The proof readily follows from the proof of Lemma 1, in Háyek and Šidák ([37]), page 195 and page 196.

Proof of Lemma 3.4.2 From Assumption **H** it follows that $\hat{\theta}^{(n)} = \theta + O_p(n^{-1/2})$. Then, the proof readily follows from the third Le Cam's Lemma as in Eq. (3.3.15) and from Eq. (3.4.5), simply replacing $\theta + \tau_n n^{-1/2}$ by $\hat{\theta}^{(n)}$.

Proof of Proposition 3.4.3 Point (i). First notice that $T_g^{(n)}(\theta)$ is the square of Euclidean norm of:

$$\left[I_{p\times p} - \Pi_{\Gamma^{1/2}(\theta,g)}\right]\Gamma^{-1/2}(\theta, g)\mathbf{\Delta}^{(n)}(\theta, g), \tag{3.8.7}$$

where $\Pi_{\Gamma^{1/2}(\theta,g)}$ represents the orthogonal Euclidean projection matrix onto $\mathcal{M}(\Gamma^{1/2}(\theta,g)\Omega)$, where $\Gamma^{1/2}(\theta,g)$ is the symmetric squared-root matrix of the positive definite $\Gamma(\theta,g)$. The projection error matrix $[I_{p\times p} - \Pi_{\Gamma^{1/2}(\theta,g)}]$ is idempotent. Under the null, we have that $\Gamma(\theta,g)\sqrt{n}(\hat{\theta}^{(n)} - \theta) \in \mathcal{M}(\Gamma(\theta,g)\Omega)$ (see Lemma 3.4.2), so that:

$$T_g^{(n)}(\hat{\theta}^{(n)}) - T_g^{(n)}(\theta) = \left[I_{p\times p} - \Pi_{\Gamma^{1/2}(\hat{\theta}^{(n)},g)}\right] \Gamma^{-1/2}(\hat{\theta}^{(n)},g)(\Delta^{(n)}(\hat{\theta}^{(n)},g) - \mathbf{\Delta}^{(n)}(\theta,g)).$$

The continuity of $\Gamma(\theta,g)$ in Assumption **F2** implies that $\Gamma(\hat{\theta}^{(n)},g)$ converges to $\Gamma(\theta,g)$. Moreover, from Lemma 3.4.2, it follows

$$\begin{aligned} T_g^{(n)}(\hat{\theta}^{(n)}) - T_g^{(n)}(\theta) &= \left[I_{p\times p} - \Pi_{\Gamma^{1/2}(\hat{\theta}^{(n)},g)}\right] \Gamma^{-1/2}(\hat{\theta}^{(n)},g)(\Delta^{(n)}(\hat{\theta}^{(n)},g) - \mathbf{\Delta}^{(n)}(\theta,g)) \\ &= \left[I_{p\times p} - \Pi_{\Gamma^{1/2}(\theta,g)}\right] n^{1/2}\Gamma^{-1/2}(\theta,g)\Gamma(\theta,g,f)(\hat{\theta}^{(n)} - \theta) + o_p(1), \end{aligned} \tag{3.8.8}$$

where, considering Assumption **F1** and **F1'**, we have:

$$\Gamma^{-1/2}(\theta,g)\Gamma(\theta,g,f) = \left(\begin{array}{c|c} I_{ll}^{-1/2}(g)\Gamma_1^{-1/2}(\theta) & \mathbf{0} \\ \hline \mathbf{0} & I_{ss}^{-1/2}(g)\Gamma_2^{-1/2}(\theta) \end{array}\right) \left(\begin{array}{c|c} I_1(g,f)\Gamma_1(\theta) & \mathbf{0} \\ \hline \mathbf{0} & I_2(g,f)\Gamma_2(\theta) \end{array}\right). \tag{3.8.9}$$

Since $\Gamma^{-1/2}(\theta,g)\Gamma(\theta,g,f)(\hat{\theta}^{(n)} - \theta)$ belongs to the linear subspace of $\mathbb{R}^p$ spanned by the columns of $\Gamma_1(\theta)$ and $\Gamma_2(\theta)$, from Eq. (3.8.8) and Eq. (3.8.9), we conclude that $T_g^{(n)}(\hat{\theta}^{(n)}) - T_g^{(n)}(\theta) = o_p(1)$, thus $T_g^{(n)}(\hat{\theta}^{(n)})$ is asymptotically invariant.

Point (ii) follows by a straightforward application of Theorem III in Wald [84].

Point (iii) follows from standard theory about chi-square tests and applying Eq. (3.3.6) and Eq. (3.4.7).

Proof of Proposition 3.4.4 The one-step estimator in Eq. (3.4.11) is root-n consistent since $\hat{\theta}^{(n)}$ it is; see van der Vaart [82, p. 71]. Moreover, the asymptotic normality follows from Theorem 1 page 312 of Bickel et al. [13].

3.9 Appendix: figures and tables

Process type	Sign autocorrelation	Robust PMLE	MSE for PMLE
A	159.62%	110.32%	0.0214
B	92.09%	82.87%	0.0418
C	32.04%	48.07%	0.1512

Table 3.1: Ratios, in percentage, for the empirical MSE of signs autocorrelation R-estimator (first column) and of Robust Pseudo Maximum Likelihood Estimator (Robust PMLE, second column), versus the empirical MSE of Gaussian Pseudo Maximum Likelihood Estimator (PMLE). The last column shows the MSE for the PMLE for the considered AR(3). The comparison is dealt under the three types of simulated processes A, B, and C (see Figure 3.1). The sample size is $n = 150$ and the Monte Carlo size is 4000. The values of θ are $\theta_1 = 0.3, \theta_2 = 0.2$, and $\theta_3 = 0.1$. The MSE ratios for van der Waerden R-estimator are: 121.49% for A-type, 92.12% for B-type, and 38.12% for C-type.

	Actual Density f		
Estimator	Laplace	t-dist(6)	t-dist(8)
PMLE	0.0214	0.0216	0.0216
Robust PMLE	0.0220	0.0229	0.0229
$\mathbf{\Delta}^{(n)}_{\text{vdW}}$	0.0216	0.0223	0.0223
$\mathbf{\Delta}^{(n)}_{W}$	0.0207	0.0225	0.0204
$\mathbf{\Delta}^{(n)}_{L}$	0.0291	0.0278	0.0289

Table 3.2: Empirical MSE for the Gaussian PMLE, Robust PMLE, and the R-estimators implied by van der Waerden, Wilcoxon, and Laplace scores. The actual distribution for the errors in the Eq. (3.5.1) are leptokurtic distributions: Laplace (first column), t-distribution with 6 degrees-of-freedom, and t-distribution with 8 degrees-of-freedom. The sample size is $n = 100$ and the Monte Carlo size is 4000. The values of θ are $\theta_1 = 0.3, \theta_2 = 0.2$, and $\theta_3 = 0.1$.

	Actual Density f								
	Leptokurtic					Skewed Normal			
Rank-score	Laplace	t-dist(4)	t-dist(6)	t-dist(8)	Gaussian	SN($a=\pm1$)	SN($a=\pm3$)	SN($a=\pm6$)	SN($a=\pm12$)
$\mathbf{\Delta}^{(n)}_{\text{vdW}}$	122.62%	93.53%	99.06%	99.89%	100.00%	99.89%	97.45%	94.92%	93.51%
$\mathbf{\Delta}^{(n)}_{W}$	167.56%	134.01%	108.87%	98.69%	94.11%	94.04%	93.76%	90.02%	84.56%
$\mathbf{\Delta}^{(n)}_{L}$	200.00%	110.64%	78.84%	67.20%	63.14%	61.99%	56.90%	48.32%	45.27%

Table 3.3: Asymptotic Relative Efficiency (ARE) as in Eq. (3.5.8) of test defined in Eq. (3.4.8), in percent versus LM test. The underlying process in an AR(3). The comparison is for three rank-based scores: van der Waerden, Wilcoxon, and Laplace scores, where different densities for the errors are specified. The first four columns refer to leptokurtic densities: the Laplace density (first column), and the t-distribution (from second to fourth column), with different degrees-of-freedom (in bracket). The fifth column is for the standard Gaussian $N(0,1)$ density. Finally, the last four columns are for Skewed-Normal ($SN(a)$), having zero mean, unitary variance, and with different skewness parameter: $a=\pm1$, $a=\pm3$, $a=\pm6$, $a=\pm12$.

	$\beta = 1$			$\beta = -1$		
	ϑ_1	ϑ_2	ϑ_3	ϑ_1	ϑ_2	ϑ_3
PMLE	0.4044	0.1491	0.1097	0.1245	0.0945	0.0949
	(0.0234)	(0.0250)	(0.0234)	(0.0352)	(0.0352)	(0.0352)
Robust PMLE	0.3986	0.1445	0.1047	0.3309	0.1550	0.0873
	(0.0210)	(0.0225)	(0.0210)	(0.0096)	(0.0096)	(0.0096)
$\boldsymbol{\Delta}^{(n)}_{\mathrm{vdW}}$	0.4074	0.1716	0.0899	0.3748	0.1919	0.0581
	(0.0214)	(0.0228)	(0.0214)	(0.0214)	(0.0225)	(0.0214)
$\boldsymbol{\Delta}^{(n)}_{W}$	0.3978	0.1670	0.0934	0.3959	0.1911	0.0618
	(0.0166)	(0.0176)	(0.0166)	(0.0148)	(0.0157)	(0.0148)
$\boldsymbol{\Delta}^{(n)}_{L}$	0.4102	0.1466	0.1161	0.4086	0.1909	0.088
	(0.0261)	(0.0280)	(0.0261)	(0.0214)	(0.0227)	(0.0214)
$\boldsymbol{\Delta}^{(n)}_{\mathrm{SA}}$	0.4375	0.1769	0.1065	0.4061	0.1952	0.1050
	(0.0205)	(0.0221)	(0.0205)	(0.0223)	(0.0237)	(0.0223)

Table 3.4: Estimated parameters for the annualized RV process of the USD/CHF exchange rate model in Eq. (3.6.2). The first two rows refer to Gaussian PMLE and to Robust PMLE, while the last four rows are for the rank-based scores: van der Waerden, Wilcoxon, Laplace and sign autocorrelation. The Robust PMLE has been applied as preliminary root-n consistent estimator in the one-step formula as in Eq. (3.4.11). Standard errors are given in brackets.

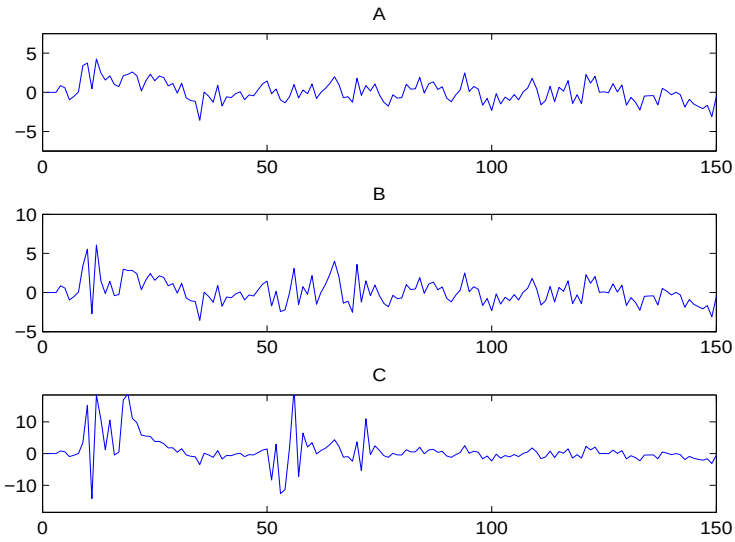

Figure 3.1: Simulated trajectory for A-, B- and C-type processes. The top panel represents the trajectory of process of A-type, which follows Eq. (3.5.1), with $\varepsilon_t^A \sim N(0,1)$. The middle panel shows a trajectory for process of B-type, which has a A-type trajectory contaminated by 10 Replacement Outliers (RO) having values $2+\varepsilon_t^B$, with $\varepsilon_t^B = 12\varepsilon_t^A$ (in patch of different length), stressing the tails of the error distribution. The bottom panel shows a trajectory for process of C-type, which has also a A-type trajectory, contaminated by 10 patchy outliers, generated as in Eq. (3.5.1), with $\theta_1 = \theta_2 = \theta_3 = 0$ and errors $\varepsilon_t^C = 2\varepsilon_t^A$.

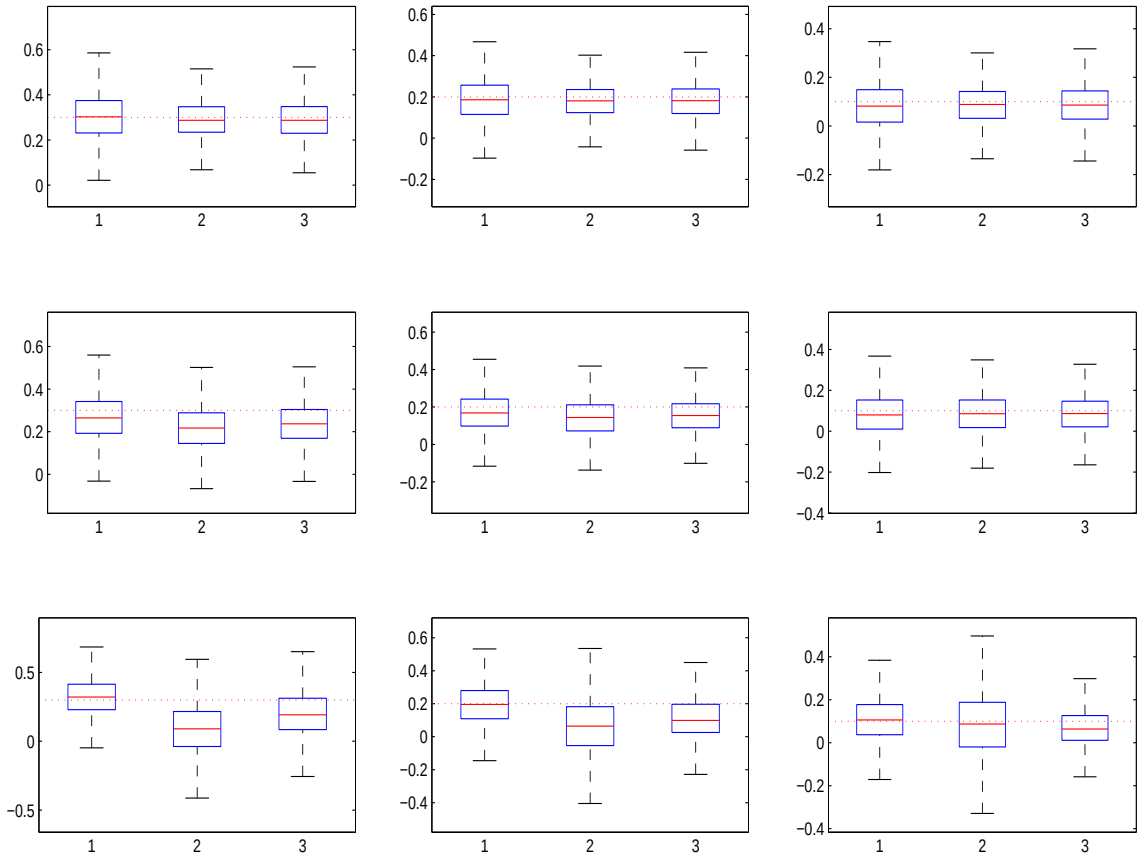

Figure 3.2: Boxplots for the R-estimator implied by the sign autocorrelation (1), the M-estimator implied by PML (2), and the M-estimator defined by Robust PML (3). The estimators are for $\theta_1 = 0.3$ (first column), $\theta_2 = 0.2$ (second comlumn), $\theta_3 = 0.1$ (third column), for A- (first row), B- (second row) and C-type (third row) processes. Monte Carlo size is 4000 and the sample size of each simulated trajectory is $n = 150$.

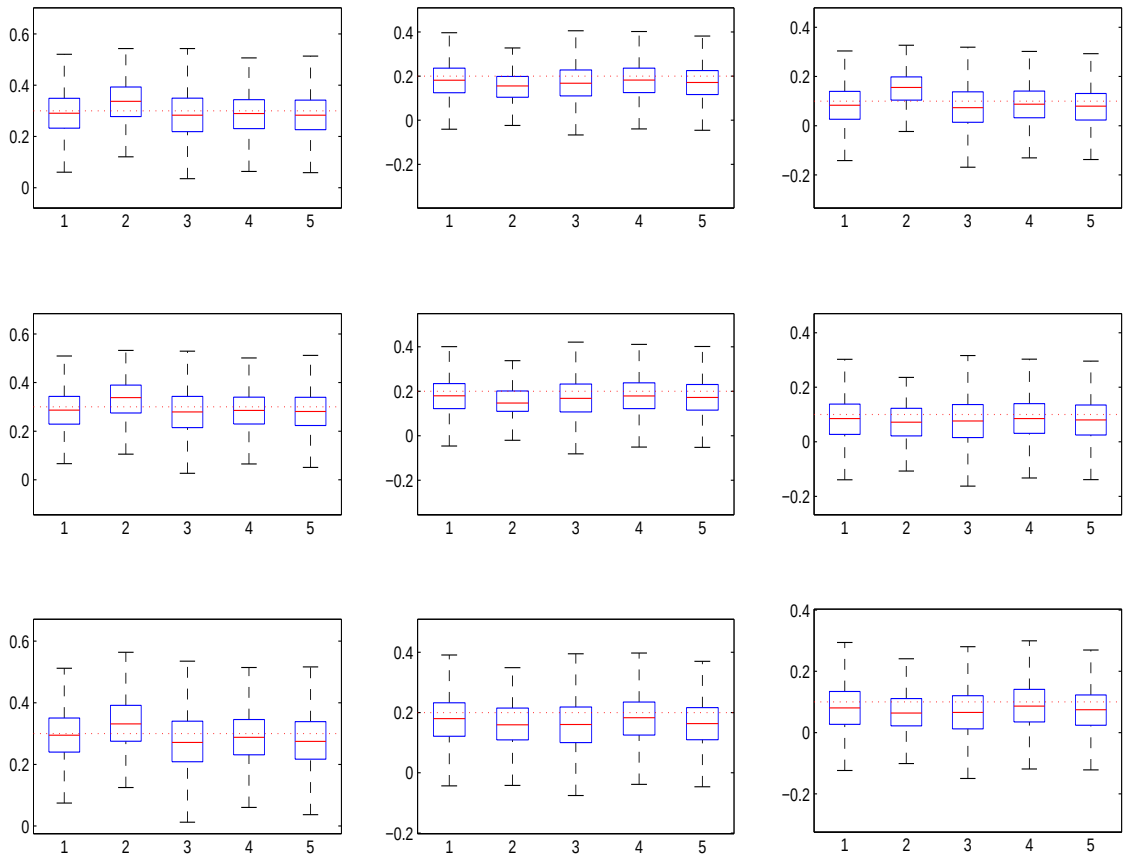

Figure 3.3: Boxplots for the the estimators implied by the van der Waerden (1), Wilcoxon (2), Laplace (3), PMLE (4), and Robust PMLE (5). The estimators are for $\theta_1 = 0.3$ (first column), $\theta_2 = 0.2$ (second comlumn), $\theta_3 = 0.1$ (third column), for different actual error distributions: Laplace (first row), t-distribution with 6 degrees-of-freedom (second row), and t-distribution with 8 degrees-of-freedom (third row). Monte Carlo size is 4000 and the sample size of each simulated trajectory is $n = 100$.

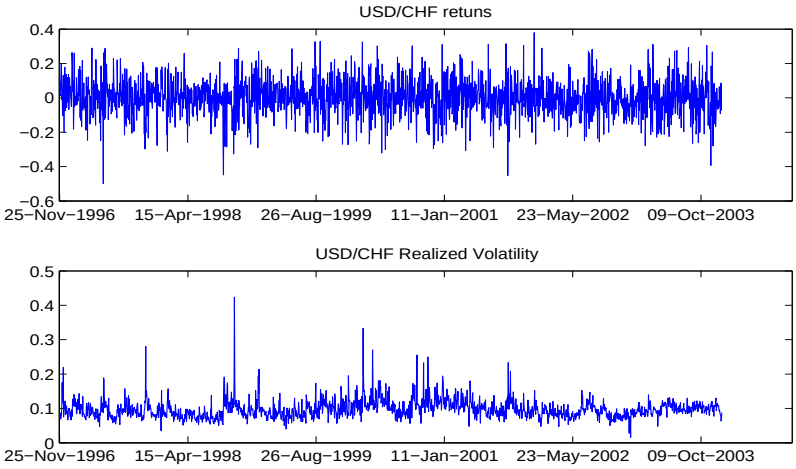

Figure 3.4: Time series for annualized log-returns (top panel) and their RV (bottom panel) for the USD/CHF exchange rate. The sample goes from 25 November 1996 to 19 December 2003.

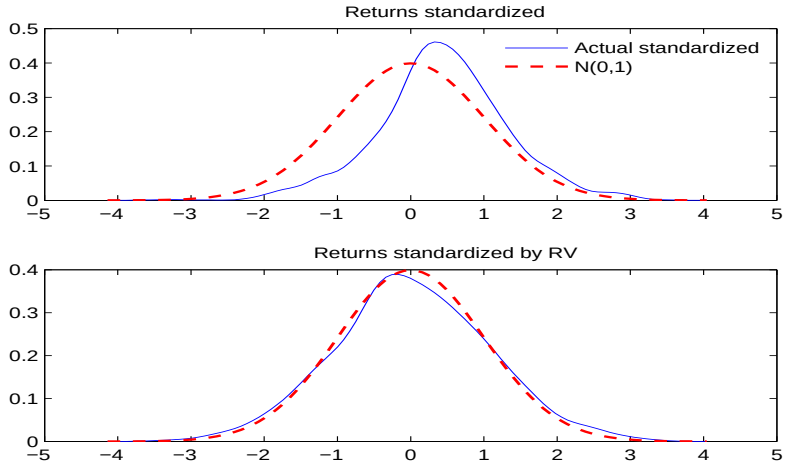

Figure 3.5: Kernel density for the standardized USD/CHF exchange rate log-returns. Top panel: log-returns standardized by their mean and standard deviation (solid line) and N(0,1) pdf (dotted line). Bottom panel: log-returns standardized by the annualized RV (solid line) and N(0,1) pdf (dotted line).

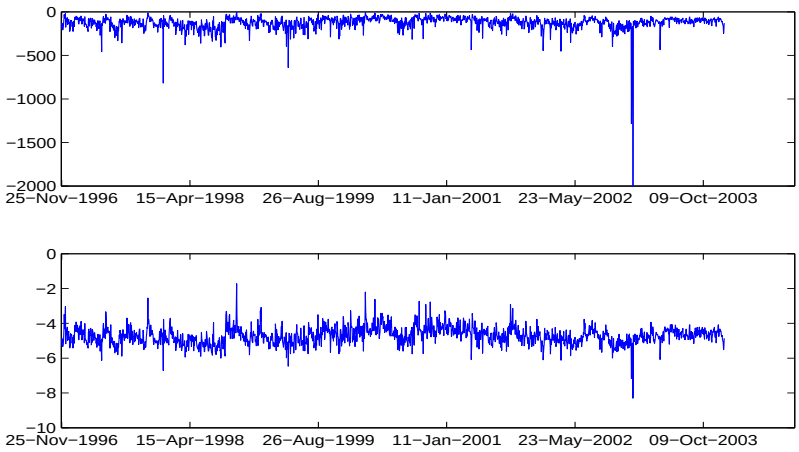

Figure 3.6: Box-Cox transformation as in Eq. (3.6.3) of the RV for the annualized log-returns of USD/CHF exchange rate. Top panel is for $\beta = -1$ (RV precision). The bottom panel is for $\beta = 0$ (log-transform)

Bibliography

[1] Y. Aït-Sahalia, P. Mykland, and L. Zhang. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100:1394–1411, 2005.

[2] H. Akaike. Information theory and an extension of the likelihood principle, in: 2nd international symposium of information theory. Mar 1973.

[3] T.G Andersen. *Stochastic Autoregressive Volatility: A framework for Volatility Modeling*. Stochastic Volatility: Selected Readings, Oxford University Press, Editor: Shephard, N., 2005.

[4] T.G. Andersen, T. Bollerslev, F.X. Diebold, and P. Labys. Exchange rate returns standardized by realized volatility are (nearly) gaussian. *Multinational Finance Journal*, 4:159–179, 2000.

[5] T.G Andersen, T. Bollerslev, F.X. Diebold, and P. Labys. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96:42–55, 2001.

[6] A. Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188, 2005.

[7] O. Barndorff-Nielsen, J. Kent, and M. Sørensen. Normal variance-mean mixtures and z-distribution. *International Statistical Review*, 50:145–159, 1982.

[8] O. Barndorff-Nielsen and N. Shephard. Non-gaussian ornstein-uhlenbeck-based models and of their uses in financial economics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63:167–241, 2001.

[9] O. Barndorff-Nielsen and N. Shephard. Econometrics analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62:253–280, 2002.

[10] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85:549–559, 1998.

[11] Ayanendranath Basu and Bruce G. Lindsay. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46:683–705, 1994.

[12] R. Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5:445–463, 1977.

[13] P. J. Bickel, C. A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive estimation for Semi-Parametric models*. Johns Hopkins Univ. Press, 1993.

[14] MV Boldin. Local robustness of sign tests in ar (1) against outliers. *Mathematical Methods of Statistics*, 20(1):1–13, 2011.

[15] T. Bollerslev and H. Zhou. Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Journal of Econometrics*, 109:33–65, 2002.

[16] D. Cassart, M. Hallin, and D. Paindaveine. On the estimation of cross-information quantities in rank-based inference. Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jureckova, 2010. Vol. 7, pages 35-45.

[17] E Choi, P Hall, and B Presnell. Rendering parametric procedures more robust by empirically tilting the model. *Biometrika*, 87(2):453–465, 2000.

[18] R. Cont and P. Tankov. *Financial Modelling with Jump Processes*. Chapman & Hall, CRC Press, 2004.

[19] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications, 2006.

[20] N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B: Methodological*, 46:440–464, 1984.

[21] H. E. Daniels. Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25:631–650, 1954.

[22] D.L. Donho and R.C. Liu. The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16:552–586, 1988.

[23] D. L. Donoho and P. J. Huber. The notion of breakdown point. *In A Festschrift for Erich L. Lehmann*, P.J. Bickel, K.A. Doksum and J.L. Hodges eds., Wadsworth, Belmont (CA):157–184, 1983.

[24] F.C. Drost, C. Klaassen, and B. Werker. Adaptive estimation in time-series models. *The Annals of Statistics*, 25:786–817, 1997.

[25] D. J. Dupuis. Extreme value theory based on the r largest annual events: a robust approach. *Journal of Hydrology*, 200:295–306, 1997.

[26] D. J. Dupuis and C. A. Field. Robust estimation of extremes. *The Canadian Journal of Statistics*, 26:199–215, 1998.

[27] R. Engle and G. Gonzalez-Rivera. Semiparametric arch. *Journal of Business and Economics Statistics*, 9:345–359, 1991.

[28] L. T. Fernholz. On multivariate higher-order von mises expansions. *Metrika*, 53:123–140, 2001.

[29] Davide Ferrari and Yuhong Yang. Maximum Lq-Likelihood Estimation. *The Annals of Statistics*, 38(2):753–783, 2010.

[30] C. A. Field and F. R. Hampel. Small sample asymptotic distributions of m–estimators of location. *Biometrika*, 69:29–46, 1982.

[31] C. A. Field and E. Ronchetti. *Small Sample Asymptotics*. IMS, Lecture notes-monograph series, 1990.

[32] R. Gatto and E. Ronchetti. General saddlpeoint approximations of marginal densities and tail probabilities. *Journal of the American Statistical Association*, Vol. 91, No. 434:666–673, 1996.

[33] Marc Genton and P.J. Rouseeuw. The change-of-variance function of M-estimators of scale under general contaminations. *Journal of Computational and Applied Mathematics*, 64:69–80, 1995.

[34] M. Gilli and E. Këllezi. An application of extreme value theory for measuring financial risk. *Computational Economics*, 27:207–228, 2006.

[35] S. Gonçalves and N. Meddhai. Box - cox transforms for realized volatility. *Journal of Econometrics*, In press, 2010.

[36] C. Gourieroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica*, 52:681–700, 1984.

[37] J. Hájek, Z. Šidák, and P. Sen. *Theory of Rank Tests*. Academic Press, San Diego, 1999.

[38] M. Hallin, C. Koell, and B. Werker. Optimal inference for discretely observed semiparametric ornstein-uhlenbeck processes. *Journal of Statistical Planning and Inference.*, 91:323–340, 2000.

[39] M. Hallin and G. Melard. Rank-based tests for randomness against first-order serial dependence. *Journal of the American Statistical Association*, 83(404):1117–1128, 1988.

[40] M. Hallin and D. Paindaveine. Semi-parametric efficient one-step r-estimators. 2010, Working paper, ECARE, Univeriste' Libre de Bruxelles.

[41] M. Hallin and M.L. Puri. Time-series analysis via rank-order theory: signed-rank tests for arma models. *Journal of Multivariate Analysis*, 39:175–237, 1991.

[42] M. Hallin and M.L. Puri. Aligned rank tests for linear models with autocorrelated error terms. *Journal of Multivariate Analysis*, 50:175–237, 1994.

[43] M. Hallin and B. Werker. Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli*, 9:137–165, 2003.

[44] M. Hallin and B. Werker. Optimal testing for semi-parametric ar-models: from gaussian lagrange multipliers to autoregression rank scores and adaptive tests. Working paper, ISRO and ECARE, Univerisite' Libre de Bruxelles, Belgium, July 2006.

[45] J.D. Hamilton. *Time-series analysis*. Princeton Univ Pr, 1994.

[46] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.

[47] F. R. Hampel, Henning C., and E. Ronchetti. A smoothing principle for the huber and other location m-estimators. *Computational Statistics and Data Analysis*, 55:324–337, 2011.

[48] F.R. Hampel. Contribution to the theory of robust estimation. Ph.D Thesis, University of California, Berkeley, 1968.

[49] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: The approach based on Influence Functions*. Wiley, 1986.

[50] Frank R. Hampel, E.M. Ronchetti, and W.A. Rousseeuw, P.J. Stahel. *Robust Statistics: the Approach Based on Influence Functions*. Wiley, New York, 1986.

[51] J. Havrda and F. Charvát. Quantification method of classification processes: Concept of structural entropy. *Kibernetika*, 3:30–35, 1967.

[52] C. C. Heyde. *Quasi-likelihood and Its Applications*. Springer-Verlag, New York, 1997.

[53] W. Hoeffding. A class of statistics with asymptotically normal distributions. *The Annals of Mathematical Statistics*, Vol. 19, No. 3:293–325, 1948.

[54] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

[55] P. .J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[56] P. J. Huber and E. Ronchetti. *Robust Statistics*. Wiley, New York, Second Edition, 2009.

[57] S. Karlin and H.W. Taylor. *A Second Course in Stochastic Processes*. Academic Press, New York, 1981.

[58] John T. Kent and David E. Tyler. Constrained M-estimation for multivariate location and scatter. *The Annals of Statistics*, 24(3):1346–1370, 1996.

[59] H. Künsch. Infinitesimal robustness for autoregressive processes. *The Annals of Statistics*, 12:843–863, 1984.

[60] D. La Vecchia and F. Trojani. Infinitesimal robustness for diffusions. *Journal of the American Statistical Association*, 105:703–712, 2010.

[61] B. G. Lindsay. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22:1081–1114, 1994.

[62] O. Linton. Adaptive estimation in arch models. *Econometric Theory*, 9:539–569, 1993.

[63] S. Lô and E. Ronchetti. Robust small sample accurate inference in moment condition models. *Computational Statistics and Data Analysis*, to appear, 2011.

[64] C. L. Mallows. On some topics in robustness. Technical report, Bell Telephone Laboratories, Murray Hill, NJ, 1975.

[65] L. Mancini, E. Ronchetti, and F. Trojani. Optimal conditionally unbiased bounded-influence inference in dynamic location and scale models. *Journal of the American Statistical Association*, 100(470):628–641, 2005.

[66] L. Mancini, E. Ronchetti, and F. Trojani. Optimal conditionally unbiased bounded-influence inference in dynamic location and scale models. *Journal of the American Statistical Association*, 100(470):628–641, 2005.

[67] L. Mancini and F Trojani. Robust value at risk prediction. *Journal of Fiancial Econometrics*, 9(2):281–313, 2011.

[68] R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.

[69] D. R. Martin and V. .J. Yohai. Influence functionals for time series. *Annals of Statistics*, 14:781–818, 1986.

[70] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management*. Princeton Series in Finance, 2005.

[71] H. Rieder. Robustness of one-and two-sample rank tests against gross errors. *The Annals of Statistics*, 9:245–265, 1981.

[72] E. Ronchetti and F. Trojani. Robust inference with gmm estimators. *Journal of Econometrics*, 101:37–69, 2001.

[73] E. Ronchetti and L. Ventura. Between stability and higher-order asymptotics. *Statistics and Computing*, 11:67–73, 2001.

[74] E. Ronchetti and J. Yen. Variance stable r-estimators. *Statistics,*, 17:189–199, 1986.

[75] David W Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):274–285, 2001.

[76] L. A. Stefanski, R. J. Carroll, and D. Ruppert. Optimallly bounded score functions for generalized linear modes with applications to logistic regression. *Biometrika*, 73:413–425, 1986.

[77] A.R. Swensen. The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. *Journal of Multivariate Analysis*, 16:54–70, 1985.

[78] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, July 1988.

[79] J.W. Tukey. A survey of sampling from contaminated distributions. *In Contributions to Probability and Statistics*, I. Olkin ed., Stanford University Press:448–485, 1960.

[80] R.L. Tweedie. Sufficient conditions for ergodicity and recurrence of markov chains on a general state space. *Stochastic Processes and their Applications*, 13:385–403, 1975.

[81] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, New York, 1998.

[82] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[83] R. von Mises. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348, 1947.

[84] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 53:426–482, 1943.

[85] W. Wefelmeyer. Quasi-likelihood models and optimal inference. *The Annals of Statistics*, 24:405–422, 1996.

[86] W. Wefelmeyer. Quasi-likelihood regression models for markov chains. *Lecture notes–Monograph Series*, 32, Selected Proceedings of the Symposium on Estimating Functions:149–173, 1997.

[87] W. Wefelmeyer. *Efficient estimation in Markov chain models: an introduction*. Statistics: Textbooks and Monographs. Asymptotics, Nonparametrics, and Time Series, Editor: S. Ghosh, 158, Dekker, New York, 1999.

[88] Michael P. Windham. Robustifying model fitting. *Journal of the Royal Statistical Society, Series B: Methodological*, 57:599–609, 1995.

[89] C. S. Withers. Expansion for the distributions and quantiles of a regular functional of the empirical distribution, with applications to nonparametric confidence intervals. *The Annals of Statistics*, 11:577–567, 1983.