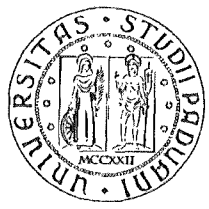

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIV

Pseudo-likelihoods from unbiased estimating functions in complex models

Direttore della Scuola: Prof.ssa Alessandra Salvan
Supervisore: Prof.ssa Laura Ventura

Dottorando: Nicola Lunardon

31 Gennaio 2012

Acknowledgements

There are few days to go and and I would like to thank and to say goodbye to some people who shared with me these years of PhD.

I thank my supervisor prof. Laura Ventura, for the help, the advices and the opportunities that she gave to me in these years. I know that it is not that easy to work with me (my moods) and I want also to thank her for having been patient and for giving me room to develop my skills.

I want to thank prof. Elvezio Ronchetti for giving me the opportunity to work with him in the Department of Econometrics at the University of Geneva. Although we spent only three weeks together, we worked hard and that has been a very fruitful period. He showed me home and provided me a very good environment during those weeks.

I am grateful to Luca Greco for the time spent togheter in Benevento, Padova, and London. Especially I want to thank him for the help in developing all the robust methods included in my thesis.

The mates of the XXIV cycle: Checca, Maestro, Monjed, Ricky and Tony. We have been together the first year and we shared a lot, not only statistical thoughts. Our roads are going to split in a few days and I know that these lines are not enough and there is still much left to say to each of you.

The other component of the "likelihood team": Ricky. We supported each other and talked about expansions, orders of magnitude and in general all those things that seemed incomprehensible to us (not only at first sight) and that seemed, seem and will seem almost useless to other people.

The post-doctoral guys: Michele, Giovanna, Manuela, Dario, Fany. You took me under your wings while I have been almost alone during the second and the third year, we had fun together and we enjoyed endless nights; you have been my other cycle. There is only one thing that I must say to all of you: do not give up.

In the last place here, but in the first one every day: you, that supported me when I wanted to give up being on my side and cutting down to size all my big deals. You pushed me ahead to try to give the best of me not only to you, but also to other people.

Abstract

The notion of likelihood function plays a central role in classical statistical inference, in particular from a Fisherian perspective, and represents an essential concept for Bayesian inference. Modern high-dimensional data, such as spatial data or complex structured longitudinal data, have generated challenges to the use of likelihood-based methods. These challenges involve both theoretical and computational difficulties that can be encompassed by specifying suitable pseudo-likelihoods, and in particular this thesis focuses on the class of composite likelihood functions.

In the present thesis three issues regarding composite likelihood inference will be discussed. The first one concerns the non-standard asymptotic distribution of composite log likelihood ratios. In order to recover the standard chi-square asymptotic distribution, an empirical log likelihood ratio test statistic derived from the composite score function is proposed. The second one is the possible lack of accuracy of composite likelihood test statistics. Accurate estimates of tail area probabilities can be obtained by using the proposed non-parametric saddlepoint test statistic, which is based on the density of the maximum composite likelihood estimator. The third one concerns the lack of robustness of the maximum composite likelihood estimator. A robust maximum composite likelihood estimator with a high breakdown point is derived by exploiting the idea of the minimum covariance determinant estimator.

Sommario

La nozione di funzione di verosimiglianza gioca un ruolo fondamentale nell'inferenza statistica, in modo particolare nell'approccio fisheriano, e rappresenta un concetto centrale nell'inferenza bayesiana. Dati con elevata dimensionalità, come dati spaziali o dati longitudinali con struttura di dipendenza complessa, hanno generato nuove sfide nell'utilizzo di procedure inferenziali basate sulla funzione di verosimiglianza. Queste sfide coinvolgono sia aspetti teorici che computazionali, che possono essere affrontati mediante la specificazione di opportune funzioni di pseudo-verosimiglianza. In particolare, questa tesi è incentrata sulla classe delle funzioni di verosimiglianza composite.

In questa tesi si discuteranno tre problemi che riguardano l'utilizzo delle funzioni di verosimiglianza composite. Il primo problema riguarda la distribuzione asintotica non standard del test log rapporto di verosimiglianza composito. Al fine di recuperare l'usuale distribuzione chi-quadrato, viene proposto un rapporto di verosimiglianza empirico derivato dalla funzione punteggio della verosimiglianza composita. Il secondo tema affrontato riguarda la possibile inacuratezza delle statistiche test ricavate dalle funzioni di verosimiglianza composite. Per ottenere stime accurate delle probabilità sulle code della distribuzione viene proposta una statistica test basata sull'approssimazione del punto di sella. Infine, il terzo problema riguarda la non robustezza dello stimatore di massima verosimiglianza composita. A tal fine viene proposta una versione robusta di tale stimatore che è basata sull'idea dello stimatore robusto "MCD" (minimum covariance determinant).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The notion of likelihood function plays a central role in classical statistical inference, in particular from a Fisherian perspective, and represents an essential concept for Bayesian inference. In the classical approach to inference, procedures based on the likelihood function are general and have optimal sampling properties, at least asymptotically, under relatively weak assumptions.

The computation of the likelihood function is often at odds with the need of introducing statistical models with highly structured dependencies or with the necessity of dealing with very large datasets. Therefore, it may not be convenient to specify, or compute, directly the likelihood function. Hence, it may be relevant to define suitable notions of approximate likelihoods, that are useful for inference. Such approximate likelihoods often belong to the wide class of pseudo-likelihoods, which includes, for instance, marginal and conditional likelihoods, the integrated likelihood, the partial likelihood, the profile likelihood and its modifications, the quasi-likelihood, the pairwise likelihood and the composite likelihood.

The approximate nature of pseudo-likelihoods must be considered when studying sampling properties of the related inferential procedures. Indeed, in many cases it is necessary to introduce suitable modifications that allow one to recover the usual asymptotic results of the proper likelihood. Several pseudo-likelihoods have been introduced in the literature in order to deal with complex models. However, the study of theoretical properties of the corresponding inferential procedures is still in progress.

A brief review on the literature that partially deals with this topic is given in Section 1.1, while in Section 1.2 the main contributions of the thesis are described.

1.1 Overview

When analyzing data on the basis of a statistical model, the availability of a likelihood function for the quantities of interest leads to inferential procedures that are both general and simple to apply. Consider, for instance, confidence regions or testing procedures based on usual first-order approximations for the distribution of the maximum likelihood estimator or of the log likelihood ratio statistic.

However, in the presence of complex models or in the presence of small deviations from the assumed model, likelihood inference may encounter some theoretical and computational difficulties. For instance, in models with complicated temporal and/or spatial dependence structures, a likelihood function based on the joint distribution of the observable data might even be unavailable. In other circumstances, the specification of the joint distribution can be straightforward, but the evaluation of the expression of the likelihood function could be computationally rather cumbersome. For instance, modeling a spatial process with a Gaussian random field requires the determinant and the inverse of the process' covariance matrix, whose dimension grows as the number of observed sites increases (Stein *et al.*, 2004). Finally, particularly in complex models, possible deviations from model assumptions, and/or the presence of outliers or influential observations, could produce unstable inferences.

In order to take into proper account the above difficulties, one could consider inference methods based upon unbiased estimating functions, whose specification only require mild assumptions concerning the random data generating mechanism, and on the associated pseudo-likelihood functions, such as quasi-likelihoods (see, *e.g.*, McCullagh, 1991), empirical likelihoods (see, *e.g.*, Owen, 2001) and composite likelihoods (see, among others, Lindsay, 1988; Varin *et al.*, 2011). These pseudo-likelihoods enjoy some properties of the full likelihood and prove useful in several context of practical interest. For instance, the corresponding estimators are consistent and asymptotically normally distributed, and the pseudo-score function has zero mean. Moreover, if the pseudo-score function is bounded, then the inferential procedures are robust with respect to outliers and influential observations (Hampel *et al.*, 1986). However, pseudo-log likelihood ratio statistics are not, in general, asymptotically chi-square distributed (Kent, 1982) and their distribution may depend on the elements of the Godambe information (Godambe, 1960).

In the above general context, this thesis focuses on some specific issues related to a particular class of composite likelihoods, namely the pairwise likelihood functions (Cox and Reid, 2004). Pairwise likelihoods are obtained by combining either marginal or conditional distributions specified for pairs of observations. Pairwise likelihoods have received an increasing interest in the last decade and there are a number of applications in statistical literature.

To quote just a few instances, we recall spatial data analysis (Hjort and Omre, 1994; Heagerty and Lele, 1998; Varin *et al.*, 2005), generalized linear mixed models (Renard *et al.*, 2004; Bellio and Varin, 2005), and longitudinal models (Fieuws and Verbeke, 2006).

In spite of the high flexibility and multiplicity of applications of pairwise likelihood functions, some concerns about the accuracy and the stability of the associated inferential procedures need further investigation.

A first drawback emerges in hypothesis testing, when the pairwise analogue of the log likelihood ratio statistic is considered. Indeed, the distribution of the pairwise log likelihood ratio statistic does not converge to the standard chi-square distribution, but to a linear combination of independent chi-square variates, with coefficients given by the eigenvalues of a matrix related to the Godambe information (Kent, 1982). Analytical expressions for this matrix are available in rather simple cases only, and to resort to its empirical counterparts may lead either to a slowdown or to a failure of the convergence to the distribution of the test statistic.

Another drawback with pairwise likelihood-based inference is related to the robustness of the resulting procedures (Hampel *et al.*, 1986). Although the pairwise likelihood function is by construction obtained from a misspecified model, it is, in general, not robust to outliers and influential observations. Indeed, pairwise score functions are combinations of genuine likelihood scores that may be unbounded. As a consequence, the resulting maximum pairwise likelihood estimator is, in general, not robust and this carries over to the pairwise counterparts of the Wald, score and log likelihood ratio test statistics, by affecting the stability of their coverage levels.

1.2 Summary and main contributions of the thesis

The main contributions of the present thesis are developed by exploiting the high versatility of the estimating function theory, which is outlined in Chapter 2. Chapter 3 introduces the pairwise likelihood functions, as a member of the more general class of the composite likelihood functions. Their main properties are reviewed and the two problems highlighted in the previous Section, arising in making inference based on pairwise likelihood functions, are discussed and outlined through simulation studies. These two problems are then treated in the three main contributions of this thesis, which are summarized below.

In Chapter 4 an empirical log likelihood ratio test statistic based on the pairwise score function is provided, whose asymptotic distribution is standard chi-square. Several adjustments to the pairwise log likelihood ratio to approximate the usual chi-square distribution have been already proposed (see, *e.g.*, Geys *et al.*, 1999; Chandler and Bate, 2007; Pace *et al.*, 2011), but they all require the computation of the elements of the Godambe information

which may be in some cases both computationally intensive and inaccurate. In this thesis an empirical log likelihood ratio test statistic derived from the pairwise score function is proposed, using the results in Adimari and Guolo (2010). The empirical log likelihood ratio test statistic enjoys some of the properties of the full likelihood one: the Bartlett-correctability (DiCiccio *et al.*, 1991); the derived confidence regions capture skewness and kurtosis; the ability to Studentize internally. In particular, the latter property leads to a non-parametric version of the Wilk's theorem and hence the distribution of the test statistic is standard chi-square while overcoming the estimation of the elements of the Godambe information.

Chapter 5 discusses a proposal which aims at overcoming the problem of poor accuracy of the pairwise log likelihood ratio statistic. A main concern in the pairwise likelihood framework is to find the best design to build the pairwise likelihood function (see, *e.g.*, Lindsay *et al.*, 2011; Davis and Yau, 2011). Indeed, the way the likelihood contributions are combined affects the asymptotic variance of the maximum pairwise likelihood estimator. We show that this problem may be circumvented by providing a test statistic whose behavior does not rely on the particular design chosen to give raise to the pairwise likelihood function. In particular, a non-parametric saddlepoint test statistic (Robinson *et al.*, 2003; Ma and Ronchetti, 2011) based on the pairwise score function is derived, which enjoys some desirable properties: it is asymptotically chi-square distributed and the approximation has a relative error of second order. Hence, the proposed test statistic claims a high level of accuracy but, nevertheless, it does not depend on the elements of the Godambe information.

The aim of Chapter 6 is twofold: (i) to provide robust maximum pairwise likelihood estimators and the related robust test statistics for the estimation of multivariate location and scatter; (ii) to exploit the simplifications provided by the use of the pairwise likelihood function in order to provide general robust procedures to deal with complex models. In particular, the focus is on mixed linear models and time series models where classical robust estimating functions are difficult to apply (see, *e.g.*, Maronna *et al.*, 2006; Heritier *et al.*, 2009). This research has moved towards the direction of providing a robust maximum pairwise likelihood estimator with a high breakdown point by exploiting the idea of the minimum covariance determinant estimator (Rousseeuw, 1984). The robust maximum pairwise likelihood estimator inherits from the minimum covariance determinant estimator both the difficulty in deriving the asymptotic properties and a high breakdown point. However, its computation is based on the pairwise score function and does not involve the specification of a new estimating function, as it is common in order to obtain robust M- or S-estimators. Furthermore, the proposed robust approach requires only mild assumptions about the shape of the underlying distribution and the computation of the estimator can be performed with a minor modification of the existing algorithm for the minimum covariance

determinant estimator.

Several examples and simulation studies are reported in all the chapters of the thesis.

Chapter 2

Estimating functions and pseudo-likelihoods

2.1 Introduction

Estimating functions provide a very general framework for statistical inference, both from a theoretical and a practical point of view. In many applications of practical interest, to rely on the assumptions that the specification of a likelihood function requires turns out to be either impossible or too stringent. This happens, for instance, in the robust framework, when the stability of the inferential procedures with respect to model misspecification or contamination is required, or in the context of generalized linear models, when over-dispersion occurs. Estimating functions are of intrinsic value themselves, since they play a central role in motivating inference based on both the likelihood function and several pseudo-likelihoods.

This Chapter aims at providing an overview about the theory of estimating functions, which are introduced along with their basic properties. The main ingredients useful to understand and encompass the subsequent developments of this thesis are introduced. In particular, optimal estimating functions are outlined and their role is shown in the context of robust inference. Moreover, a comprehensive view of the quasi-likelihoods and the empirical likelihoods is provided, as specific classes of pseudo-likelihood functions which are directly derivable from unbiased estimating functions. Additionally, two simulation studies are included in order to compare the finite sample accuracy of confidence intervals based on quasi and empirical log likelihood ratio statistics.

2.2 Estimating functions and M-estimators

Let $\mathcal{F} = \{f(y;\theta);\, y \in \mathcal{Y} \subseteq \mathbb{R}^q,\, \theta \in \Theta \subseteq \mathbb{R}^p,\, q,p \geq 1\}$ be a parametric statistical model for the random vector Y, and let $F_\theta = F(y;\theta)$ be the distri-

bution function associated to $f(y;\theta)$. Suppose to observe a random sample $y = (y_1, \ldots, y_n)$ of size n from F_θ.

The likelihood function for θ is

$$L(\theta) = L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta),$$

the log likelihood function is $\ell(\theta) = \ell(\theta; y) = \log L(\theta)$ and the score function is

$$\ell_*(\theta) = \ell_*(\theta; y) = \sum_{i=1}^{n} \frac{\partial \ell(\theta; y_i)}{\partial \theta} = \sum_{i=1}^{n} \ell_*(\theta; y_i). \tag{2.1}$$

Under regularity assumptions always assumed in the following (see, *e.g.*, Pace and Salvan, 1997, Sect. 1.4), the maximum likelihood estimator $\hat{\theta}$ is defined as the solution of the score equation $\ell_*(\theta) = 0$.

Basic properties of the score function (2.1) are the first and second Bartlett's identities. The first Bartlett's identity gives the unbiasedness of the score function, i.e.

$$\mathbb{E}\left(\ell_*(\theta; Y)\right) = 0, \text{ for all } \theta \in \Theta,$$

while the second Bartlett's identity gives the information identity

$$\mathrm{Var}\left(\ell_*(\theta; Y)\right) = \mathbb{E}\left(\ell_*(\theta; Y)\ell_*(\theta; Y)^\top\right) = -\mathbb{E}\left(\frac{\partial \ell_*(\theta; Y)}{\partial \theta^\top}\right) = I(\theta),$$

where $I(\theta)$ denotes the expected Fisher information matrix for one observation, and $\mathbb{E}(\cdot)$ and $\mathrm{Var}(\cdot)$ indicate expectation and variance with respect to F_θ.

A generalization of the score function is given by the broad class of *estimating functions* (see, *e.g.*, Godambe and Kale, 1991; Desmond, 1997), that do not require to be the gradient of an objective function. Estimating functions are functions both of the parameter θ and of the data y, of the form

$$\Psi_\theta = \Psi(\theta; y) = \sum_{i=1}^{n} \psi(\theta; y_i), \tag{2.2}$$

where $\psi(\cdot) : \Theta \times \mathcal{Y} \to \mathbb{R}^p$ is a given function.

An estimator $\tilde{\theta}$ for θ, obtained as the root of the estimating equation $\Psi_\theta = 0$, is called *M-estimator* (Huber, 1964, 1967) and can be viewed as a generalization of the maximum likelihood estimator. The properties of both estimating functions and M-estimators are reviewed in the following.

An estimating function is unbiased if

$$\mathbb{E}\left(\psi(\theta; Y)\right) = 0, \text{ for all } \theta \in \Theta. \tag{2.3}$$

Condition (2.3) does not necessarily imply the unbiasedness of the corresponding M-estimator $\tilde{\theta}$, unless Ψ_θ is linear in θ. However, under suitable regularity conditions, the unbiasedness of the estimating function implies the consistency of $\tilde{\theta}$ (Clarke, 1983, 1986; Huber, 1967, 1981). Moreover, M-estimators are approximately normal, with mean θ and covariance matrix $V(\theta) = H(\theta)^{-1}J(\theta)\left(H(\theta)^{-1}\right)^\top$, where

$$J(\theta) = \mathrm{Var}(\Psi_\theta) = \mathbb{E}\left(\Psi_\theta \Psi_\theta^\top\right) \quad \text{and} \quad H(\theta) = -\mathbb{E}\left(\frac{\partial \Psi_\theta}{\partial \theta^\top}\right). \tag{2.4}$$

Consistent estimates of $J(\theta)$ and $H(\theta)$ are

$$\hat{J}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi(\theta; y_i)\psi(\theta; y_i)^\top, \; \hat{H}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \psi(\theta; y_i)}{\partial \theta^\top}, \tag{2.5}$$

The matrix $G(\theta) = V(\theta)^{-1}$ is known as the expected *Godambe information* (Godambe, 1960), and its sandwich form is due to the failure of the second Bartlett's identity. When the score function $\ell_*(\theta)$ is considered, the asymptotic normality for the maximum likelihood estimator holds with the Godambe information replaced by the Fisher information.

2.2.1 Profile estimating functions

In parametric statistical models, very often only few components of the parameter are of interest. The remaining components are needed to increase model adequacy and may be considered as nuisance, i.e. they describe additional aspects of the specified family of probability distributions. Paralleling likelihood procedures, the estimating function framework offers the opportunity to make inference in the presence of nuisance parameters, by replacing them with suitable consistent estimates.

Suppose that θ is partitioned as $\theta = (\tau, \lambda)$, i.e. into a component of interest τ and a nuisance component λ, whose dimensions are p_0 and $(p - p_0)$, respectively. Similarly, the estimating function is partitioned as $\Psi_\theta = (\Psi_\tau, \Psi_\lambda)$, where $\Psi_\tau = \Psi_\tau(\theta; y)$ and $\Psi_\lambda = \Psi_\lambda(\theta; y)$ are the partial estimating functions corresponding to τ and λ, respectively. This means that if λ was known, then Ψ_τ may be used as an estimating function for τ. Let $\tilde{\lambda}_\tau$ denotes the constrained estimate of λ when τ is considered fixed, that is, the solution of $\Psi_\lambda(\tau, \lambda_\tau; y) = 0$, and let $\tilde{\theta}_\tau = (\tau, \tilde{\lambda}_\tau)$.

The elements (2.4) of the Godambe information are partitioned as well as

$$J(\theta) = \begin{pmatrix} J_{\tau\tau} & J_{\tau\lambda} \\ J_{\lambda\tau} & J_{\lambda\lambda} \end{pmatrix} \quad \text{and} \quad H(\theta) = \begin{pmatrix} H_{\tau\tau} & H_{\tau\lambda} \\ H_{\lambda\tau} & H_{\lambda\lambda} \end{pmatrix},$$

where, for instance, $J_{\tau\lambda} = J_{\tau\lambda}(\theta) = \mathbb{E}\left(\Psi_\tau(\theta)\Psi_\lambda(\theta)^\top\right)$ and $H_{\tau\lambda} = H_{\tau\lambda}(\theta) = -\mathbb{E}\left(\partial\Psi_\tau(\theta)/\partial\lambda^\top\right)$.

The profile M-estimator $\tilde{\tau}$ for τ is defined as the root of the profile estimating equation

$$\tilde{\Psi}_\tau = \Psi_\tau(\tau, \tilde{\lambda}_\tau; y) = \Psi_\tau(\tilde{\theta}_\tau; y) = 0.$$

Its asymptotic variance corresponds to the $\tau\tau$-block of $V(\tilde{\theta}_\tau)$, which is given by

$$\begin{aligned} V(\tilde{\theta}_\tau)_{\tau\tau} = \; & H(\tilde{\theta}_\tau)^{\tau\tau} J(\tilde{\theta}_\tau)_{\tau\tau} (H(\tilde{\theta}_\tau)^{\tau\tau})^\top + \\ & + 2H(\tilde{\theta}_\tau)^{\tau\lambda} J(\tilde{\theta}_\tau)_{\lambda\tau} H(\tilde{\theta}_\tau)^{\tau\tau} + \\ & + H(\tilde{\theta}_\tau)^{\tau\lambda} J(\tilde{\theta}_\tau)_{\lambda\lambda} (H(\tilde{\theta}_\tau)^{\tau\lambda})^\top \end{aligned} \tag{2.6}$$

where, for instance, $J(\tilde{\theta}_\tau)^{\tau\lambda}$ is the $\tau\lambda$-block of $J(\tilde{\theta}_\tau)^{-1}$, and $H(\tilde{\theta}_\tau)^{\tau\lambda}$ is the $\tau\lambda$-block of $H(\tilde{\theta}_\tau)^{-1}$. The first term in (2.6) is the asymptotic variance corresponding to τ if λ was known, whereas the remaining part reflects the cost of estimating the nuisance parameter. To study the asymptotic properties of $\tilde{\tau}$, useful references are Gong and Samaniego (1981), Pierce (1982), Parke (1986) and Jørgensen and Knudsen (2004).

The profile estimating function $\tilde{\Psi}_\tau$ suffers from bias like the ordinary profile score function, that is

$$\mathbb{E}\left(\Psi_\tau(\theta; Y)\right)|_{\theta=\tilde{\theta}_\tau} = O(1).$$

However, $\tilde{\Psi}_\tau$ can be adjusted in order to alleviate the effect of nuisance parameters, thus obtaining an approximately unbiased estimating function for τ (see Severini, 2002; Wang and Hanfelt, 2003; Jørgensen and Knudsen, 2004). Adjustments to eliminate the bias of $\tilde{\Psi}_\tau$ can also be found in Adimari and Ventura (2002) and Bellio *et al.* (2008).

2.3 Optimal estimating functions

The comparison between the Godambe and the Fisher information matrices highlights that the particular choice of an estimating function affects the efficiency of the corresponding estimator.

Assume, without loss of generality, that θ is scalar and let $\mathcal{U}$ be the class of all unbiased estimating functions with respect to the model $\mathcal{F}$. An estimating function Ψ_θ^* is *optimal* if

$$\mathrm{Var}(\Psi_\theta^*) = \min_{\Psi_\theta \in \mathcal{U}} \mathrm{Var}\left(\Psi_\theta\right). \tag{2.7}$$

The criterion (2.7) must be refined since Ψ_θ and $c\Psi_\theta$, where c is an arbitrary constant, define the same estimator, but the variance of $c\Psi_\theta$ is $c^2\mathrm{Var}\left(\Psi_\theta\right)$. In view of this, a standardized estimating function can be considered in (2.7), whose standardization depends on θ, of the form $\Psi_\theta^s = \Psi_\theta / H(\theta)$. By defining

$\mathcal{U}^s$ as the class of all the standardized estimating functions, the criteria (2.7) becomes

$$\mathrm{Var}\left(\Psi_\theta^{s*}\right) = \min_{\Psi_\theta \in \mathcal{U}^s} \mathrm{Var}\left(\Psi^s(\theta)\right). \tag{2.8}$$

Under mild conditions on the class $\mathcal{U}$ and on the model $\mathcal{F}$, Godambe (1960) showed that the score function $\ell_*(\theta)$ is optimal. This result provides a generalization of the Cramer-Rao inequality for standardized unbiased estimating functions (Kale, 1962), i.e.

$$\mathrm{Var}\left(\ell_*(\theta)\right) \leq \mathrm{Var}\left(\Psi_\theta\right). \tag{2.9}$$

Godambe and Thompson (1974) showed the uniqueness of the optimal estimating function Ψ_θ^*, meaning that if two estimating functions $\Psi_{1\theta}^s$ and $\Psi_{2\theta}^s$ are both optimal according to (2.8), then $\Psi_{1\theta}^s = \Psi_{2\theta}^s$.

The above results extend to the multidimensional setting, that is for $\theta \subseteq \mathbb{R}^p$, $p > 1$. There are several criteria in order to establish the optimality of Ψ_θ^s and the most common are

1. $\mathrm{tr}\left(\mathrm{Var}\left(\Psi_\theta^s\right)\right) > \mathrm{tr}(\mathrm{Var}\left(\Psi_\theta^{s*}\right))$ (Chandrasekar and Kale, 1984), where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix;

2. $\mathrm{Var}\left(\Psi_\theta^s\right) - \mathrm{Var}\left(\Psi_\theta^{s*}\right)$ is non negative definite (Kale, 1962);

3. $|\mathrm{Var}\left(\Psi_\theta^s\right)| > |\mathrm{Var}\left(\Psi_\theta^{s*}\right)|$, where $|\cdot|$ denotes the determinant of a matrix.

A notable application of the above results can be found in the context of generalized linear models, where the estimators of the regression parameters are derived from the quasi-score functions (Wedderburn, 1974). Quasi-score functions have properties similar to those of the genuine scores and can be related to the theory of optimal estimating functions. For a more detailed discussion the reader can refer to Desmond (1997).

2.4 Estimating functions and robustness

Robust statistic is concerned with the fact that many common assumptions made about randomness, independence, distributional models, and so on, are at most approximations to reality. Classical statistical procedures appear very sensitive even to slight departures from the assumptions under which they are developed.

Robust statistic aims at preserving inference from the possible misspecifications of the parametric family $\mathcal{F}$. Since the hypothesized model F_θ is only an approximation of the true underlying distribution G_θ, it can be supposed that G_θ lies in a neighborhood of F_θ, i.e. $\mathcal{P}_\epsilon(F_\theta) = \{G_\theta | d(G_\theta, F_\theta) < \epsilon\}$, where $d(\cdot)$ is a suitable measure of distance which describes how a small change in the underlying distribution can affect inference. This idea was

formalized by Huber (1964) by introducing a contamination neighborhood of F_θ, also known as *gross-error model*, defined as

$$\mathcal{P}_\epsilon(F_\theta) = \{F_\epsilon | F_\epsilon = (1-\epsilon)F_\theta + \epsilon H\}, \tag{2.10}$$

where H an arbitrary unknown distribution. The model $\mathcal{P}_\epsilon(F_\theta)$ assumes that a fraction $\epsilon \in (0,1)$ of the data may consist of gross errors coming from the distribution H.

The use of estimating functions and M-estimators to deal with models of the form (2.10) may prove useful. Historically, the first approach to robust inference is Huber's minimax (Huber, 1964, 1981). However, in the following, the infinitesimal approach (Hampel *et al.*, 1986) is considered. The infinitesimal approach is mainly based on the concepts of influence function and breakdown point. The influence function and the breakdown point measure different aspects of robustness of the inferential procedures: the former refers to local aspects, whereas the latter to global ones.

2.4.1 The infinitesimal approach

Let us consider an estimating function of the form (2.2), reformulated as

$$\mathbb{E}_{\hat{F}_n}(\Psi_\theta) = \int \Psi(\theta; y)\, d\hat{F}_n,$$

where $\hat{F}_n$ is the empirical distribution function assigning mass $1/n$ to each observation and $\mathbb{E}_{\hat{F}_n}$ denotes expectation with respect to $\hat{F}_n$. The root of the estimating equation $\mathbb{E}_{\hat{F}_n}(\Psi_\theta) = 0$ is the M-estimator $\tilde{\theta}$, represented as a function of $\hat{F}_n$, i.e. $\tilde{\theta} = \tilde{\theta}(\hat{F}_n)$. Since $\hat{F}_n$ is a nonparametric estimate of F_θ, the parameter θ can be defined as a functional as well, that is $\theta = \theta(F_\theta)$, and is obtained as the root of the equation

$$\mathbb{E}(\Psi_\theta) = \int \Psi(\theta; y)\, dF_\theta = 0. \tag{2.11}$$

Definition (2.11) points out the dependence of the parameter θ on the assumed model F_θ and helps to understand the behaviour of the M-estimator $\tilde{\theta}$ when small departures from the assumed model occur.

One way to assess the robustness of an estimator $T = T(hatF)$ is to consider the *influence function* (IF), defined as (Hampel *et al.*, 1986)

$$\begin{aligned} \mathrm{IF}(y; T, F_\theta) &= \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)F_\theta + \epsilon \Delta_y\right) - T(F_\theta)}{\epsilon} \\ &= \frac{\partial}{\partial \epsilon}\left[T\left((1-\epsilon)F_\theta + \epsilon \Delta_y\right)\right]\Big|_{\epsilon=0}, \end{aligned}$$

where Δ_y is a point mass in y. The IF measures the relative change on the estimator T provided by a infinitesimal contamination at y (Hampel,

1974). Moreover, the IF provides information about the properties of the estimator: the asymptotic bias of T caused by an infinitesimal contamination in y is given by the linear approximation $\epsilon\,\mathrm{IF}(y;T,F_\theta)$, while its asymptotic variance is

$$\mathrm{V}(T,F_\theta)=\int \mathrm{IF}(y;T,F_\theta)\mathrm{IF}(y;T,F_\theta)^\top dF_\theta.$$

There are at least three important summary values of the IF. The most important is the *gross error sensitivity* of T with respect to the model F_θ, defined as

$$\gamma(T,F_\theta)=\sup_y ||\mathrm{IF}(y;T,F_\theta)||, \tag{2.12}$$

where $||\cdot||$ denotes the Euclidean norm. The quantity in (2.12) measures the worst impact on the standardized asymptotic bias of the estimator T produced by an infinitesimal contamination at y. A desiderable property of an estimator is to have a finite $\gamma(T,F_\theta)$ and estimators with this property are said to be B-robust (Rousseeuw, 1981).

A second measure derived from the IF is the *local shift sensitivity* of T at F_θ, that measures the effect on the estimator of changing the value y with another one z. Finally, the *rejection point* represents the point after which the effect of the possible contamination on the value of the estimator is null.

The IF of an M-estimator $\tilde{\theta}$ is given by (Hampel *et al.*, 1986)

$$\mathrm{IF}(y;\tilde{\theta},F_\theta)=H(\theta)^{-1}\psi(\theta;y). \tag{2.13}$$

Expression (2.13) highlights that the IF of an M-estimator is proportional to the estimating function $\psi(\cdot)$. Thus, the gross error sensitivity of an M-estimator is finite if and only if $\psi(\cdot)$ is bounded. Hence, to check whether $\tilde{\theta}$ is B-robust, it is sufficient to look at $\psi(\cdot)$.

The IF is an useful tool, giving rise to important robustness measures, but there is one limitation: by construction, it is an entirely local concept. Therefore, it must be complemented by a measure of global reliability of the estimator, such as the *breakdown point*. The breakdown point of the estimator T is defined as (see, *e.g.*, Heritier *et al.*, 2009, pag. 20)

$$\epsilon^*(T,F_\theta)=\inf\left\{\epsilon:\epsilon\sup_H ||T(F_\epsilon)-T(F_\theta)||=\infty\right\},$$

with H defined in (2.10), and it measures the distance from the model distribution beyond which the statistic becomes totally unreliable and uninformative.

The link between the IF and the breakdown point is given by the requirement that an estimator should have a high breakdown point and a low gross error sensitivity. A high breakdown point is often easy to obtain. A low gross error sensitivity, however, is in conflict with the efficiency requirement of a low asymptotic variance with respect to the central model. Both $\gamma(T,F_\theta)$ and

$\mathrm{V}(T, F_\theta)$ have positive lower bounds, but in general these bounds cannot be reached simultaneously. Estimators which are optimal under the constraint of a bounded $\gamma(T, F_\theta)$ were proposed by Hampel (1974) and are presented in the following Section. These estimators are obtained by bounding the gross error sensitivity and, in general, have a low breakdown point. In view of this, a robust estimator with bounded IF can become useless in practice if its breakdown point is too small. Nevertheless, it is possible to combine the request of an estimator with a high breakdown point and with a low asymptotic variance, obtaining the so called MM-estimator (Yohai, 1987).

2.4.2 Robustness and optimality

Hampel (1974) considered the possibility of finding an optimal estimator under the constraint of bounded gross error sensitivity. Expression (2.13), togheter with (2.9), gives some clues in order to find such an estimator. The corresponding estimating function must be searched within a subclass of $\mathcal{U}$, denoted with $\mathcal{U}_c$, which contains all the unbiased estimating functions with respect to the model $\mathcal{F}$, which satisfies

$$\gamma(T, F_\theta) \leq c.$$

Indeed, if there were no constraints on the IF, then the class $\mathcal{U}_c$ would coincide with $\mathcal{U}$, and the optimality problem is solved by picking the score function. Once an upper bound c is set, the optimal bounded estimating function within $\mathcal{U}_c$ will coincide, for the majority of the observations, with the optimal estimating function in $\mathcal{U}$.

The optimal B-robust estimator (OBRE) $\tilde{\theta}^*$ is derived from the optimal bounded estimating function within the class $\mathcal{U}_c$. For a fixed upperbound c, the estimating function which defines the OBRE has the general form

$$\sum_{i=1}^{n} \psi_c^*(\theta; y_i) = \sum_{i=1}^{n} \left[\ell_*(\theta; y_i) - a(\theta)\right] w_c(\theta; y_i), \tag{2.14}$$

where $w_c(\theta; y_i) = \min\left(1, c/||A(\theta)\left[\ell_*(\theta; y_i) - a(\theta)\right]||\right)$ are suitable weights (Hampel *et al.*, 1986), $i = 1, \ldots, n$. The p-dimensional vector $a(\theta)$ and the $p \times p$ matrix $A(\theta)$ are determined implicitly by the equations

$$a(\theta) = \frac{\int \ell_*(\theta; y) w_c(\theta; y) dF_\theta}{\int w_c(\theta; y) dF_\theta},$$

and

$$\left[A(\theta)^\top A(\theta)\right]^{-1} = \int \left[\ell_*(\theta; y) - a(\theta)\right] \left[\ell_*(\theta; y) - a(\theta)\right]^\top w_c^2(\theta; y) dF_\theta.$$

A key quantity in (2.14) is the centering function $a(\theta)$ that ensures the unbiasedness of the resulting optimal bounded estimating function. The

computation of $\tilde{\theta}^*$ as the root of $\sum_{i=1}^{n} \psi_c^*(\theta; y_i) = 0$ can be carried out following an iterative procedure (see, for instance, Hampel *et al.*, 1986; Carroll and Ruppert, 1988; Ronchetti and Trojani, 2001; Bellio, 2007). As pointed out by several authors (see, *e.g.*, Dupuis and Field, 1998), the vector $a(\theta)$ should be computed via numerical integration.

2.5 Hypothesis testing and confidence regions

This Section focuses on testing hypothesis or setting confidence regions about θ. There are several test statistics available. The first one is the Wald-type test statistic, defined as

$$W_w(\theta) = (\tilde{\theta} - \theta)^\top V(\tilde{\theta})^{-1}(\tilde{\theta} - \theta). \tag{2.15}$$

As in classical inference, the asymptotic null distribution of (2.15) is chi-square with p degrees of freedom (Heritier and Ronchetti, 1994).

Despite Wald-type statistics are simple to compute, they lack invariance under reparametrization and force confidence regions to have an elliptical shape. If more accurate confidence regions are required, score-type test statistics may be considered, given by

$$W_s(\theta) = \Psi_\theta^\top J(\theta)^{-1} \Psi_\theta, \tag{2.16}$$

whose asymptotic null distribution is chi-square with p degrees of freedom.

Consider the partition of the parameter θ in the two components τ and λ. For testing hypothesis or setting confidence regions about τ, the Wald- and the score-type test statistics turn out to be

$$W_{wp}(\tau) = (\tilde{\tau} - \tau)^\top V(\tilde{\theta})^{\tau\tau} (\tilde{\tau} - \tau), \tag{2.17}$$

and

$$W_{sp}(\tau) = \Psi_\tau(\tilde{\theta}_\tau)^\top H(\tilde{\theta}_\tau)_{\tau\tau} V(\tilde{\theta}_\tau)^{\tau\tau} H(\tilde{\theta}_\tau)_{\tau\tau} \Psi_\tau(\tilde{\theta}_\tau). \tag{2.18}$$

The asymptotic null distributions of (2.17) and (2.18) are chi-square with p_0 degrees of freedom (Heritier and Ronchetti, 1994).

There is also the possibility to derive log likelihood ratio-type tests based of the function $\rho(\theta; y)$, defined as $\partial \rho(\theta; y)/\partial \theta = \Psi_\theta$. The function $\rho(\cdot)$ and the associated log likelihood ratio-type tests as well, will be pursued in Section 2.6.1.

The above results still hold and carry over the framework of bounded estimating functions. In particular, when optimal bounded estimating functions are considered (see Section 2.4.2), the statistics (2.15), (2.16), (2.17) and (2.18) bound the effect of a small amount of contamination on the asymptotic level and power of the derived test, while retaining the optimality properties (Heritier and Ronchetti, 1994).

2.6 Pseudo-likelihood functions

In the previous Sections it has been shown how to generalize likelihood-based inferential procedures by using general unbiased estimating functions and the associated M-estimators. In particular, the estimating function framework may address at obtaining sensible estimators when either a semiparametric model is specified or to accommodate for small departures from the assumed model. The term pseudo-likelihoods in general refers to functions of the parameter of interest θ, as well of the observations y, that resemble for some respects the behaviour of a genuine likelihood: zero mean pseudo-score function, maximum pseudo-likelihood estimators with asymptotic normal distribution, pseudo-log likelihood ratio test statistics with the standard chi-square limiting distribution.

In the following two particular pseudo-likelihoods derived from unbiased estimating functions are presented: the quasi-likelihood and the empirical likelihood functions. Their properties and the connections with the estimating function framework are reviewed.

2.6.1 Quasi-likelihoods

An estimating function can be specified avoiding the full specification of a parametric statistical model F_θ. For this reason, Ψ_θ frequently fails to be the gradient of an objective function called quasi-log likelihood. Quasi-log likelihood functions are obtained as the line integral of the estimating function Ψ_θ, i.e.

$$\ell_Q(\theta) = \sum_{i=1}^{n} \int_c^\theta \psi(t; y_i) dt, \tag{2.19}$$

with c arbitrary constant. When θ is a vector parameter, it is not always possible to uniquely define a quasi-log likelihood function because the line integral (2.19) depends on the path chosen. More precisely, it may happen that the matrix $H(\theta)$ is symmetric, while its observed counterpart $\hat{H}(\theta)$ is not. Hence, a necessary and sufficient condition for the existence of a quasi-log likelihood function is the symmetry of the matrix $\hat{H}(\theta)$.

The problem of the nonexistence of (2.19) can be overcomed when the parameter is scalar or when interest lies on a scalar component of θ. In the latter situation θ is partitioned as (τ, λ) and a profile estimating function for τ of the form $\tilde{\Psi}_\tau = \sum_{i=1}^n \psi_\tau(\tilde{\theta}_\tau; y_i)$ is considered. Then, the profile quasi-log likelihood function is

$$\ell_{QP}(\tau) = \sum_{i=1}^{n} \int_c^\theta \psi_\tau(\tilde{\theta}_t; y_i) dt.$$

When $\ell_Q(\theta)$ and $\ell_{QP}(\theta)$ exist, the corresponding quasi-log likelihood ratios are defined as

$$W_Q(\theta) = 2\left\{\ell_Q(\tilde{\theta}) - \ell_Q(\theta)\right\}$$

and

$$W_{QP}(\tau) = 2\left\{\ell_{QP}(\tilde{\tau}) - \ell_{QP}(\tau)\right\}.$$

These statistics do not present the classical chi-square asymptotic distribution (see, *e.g.*, Kent, 1982). In particular,

$$W_Q(\theta) \xrightarrow{d} \sum_{j=1}^{p} \lambda_j(\theta) Z_j^2,$$

and

$$W_{QP}(\tau) \xrightarrow{d} Z^2 \frac{H_{\tau\tau}(\tilde{\theta}_\tau)}{J_{\tau\tau}(\tilde{\theta}_\tau)},$$

with $\lambda_j(\theta)$ eigenvalues of $H(\theta)^{-1}J(\theta)$, $j = 1, \ldots, p$, and Z_j indipendent random variables having a standard normal distribution, and Z standard normal random variable.

Quasi-log likelihood ratios with the standard chi-square limiting distribution can be obtained in two different ways. One possibility is to consider the scaled test statistics proposed by Hanfelt and Liang (1995)

$$W_Q^{hl}(\theta) = \frac{W_Q(\theta)}{\kappa_1}$$

and

$$W_{QP}^{hl}(\tau) = \frac{W_{QP}(\tau)}{\kappa_{1p}}, \tag{2.20}$$

with $\kappa_1 = \sum_{j=1}^{p} \lambda_j(\theta)/p$ and $\kappa_{1p} = J_{\tau\tau}(\theta)/H_{\tau\tau}(\theta)$. The second one is to consider a linear trasformation of the estimating function Ψ_θ of the form $\bar{\Psi}_\theta = A(\theta)\Psi(\theta)$, with

$$A^\top(\theta) = J(\theta)^{-1}H(\theta). \tag{2.21}$$

Since $A(\theta)$ is non-singular, $\bar{\Psi}_\theta$ has the same solution as Ψ_θ, but the former satisfies the second Bartlett's identity as a genuine score does. Therefore, quasi-log likelihood ratios obtained by considering $\bar{\Psi}_\theta$ have the standard asymptotic chi-square distribution.

When a profile estimating function is considered, some conceptual problems may arise in applying a linear transformation to $\Psi_\theta = (\Psi_\tau, \Psi_\lambda)$. Barndorff-Nielsen (1995) defined a quasi-profile log likelihood function for τ based on $\bar{\Psi}_\theta$, but this estimating function mixes the components relative to τ and λ, i.e.

$$\bar{\Psi}_\theta = \left(A_{\tau\tau}\Psi_\tau + A_{\tau\lambda}\Psi_\lambda, A_{\lambda\tau}\Psi_\tau + A_{\lambda\lambda}\Psi_\lambda\right). \tag{2.22}$$

In (2.22) the interpretation of the components of the new estimating function cannot be clear, since the original partition of Ψ_θ is no longer respected. However, it is possible to define an alternative quasi-profile log likelihood for τ following Adimari and Ventura (2002). The idea is to adjust the profile

estimating function $\tilde{\Psi}_\tau$ so that its bias and information bias are of order $O(1)$. The corresponding quasi-profile log likelihood is given by

$$\ell_{QP}^{av}(\tau) = \int_c^\tau w(\tilde{\theta}_t)\psi_\tau(\tilde{\theta}_t; y_i)dt,$$

where $w(\cdot)$ is a suitable correction term that depends on the elements of $J(\theta)$ and $H(\theta)$ (Adimari and Ventura, 2002). The corresponding quasi-profile log likelihood ratio statistic

$$W_{QP}^{av} = 2\left\{\ell_{QP}^{av}(\tilde{\tau}) - \ell_{QP}^{av}(\tau)\right\} \tag{2.23}$$

is approximately χ_1^2 distributed.

2.6.2 Empirical likelihoods

The empirical likelihood function is a nonparametric tool introduced by Owen (1988, 1990) and subsequently applied to several contexts, including linear models (Owen, 1991; Chen, 1993), generalized linear models (Kolaczyk, 1994) and inference with dependent observations (see, *e.g.*, Kitamura, 1997; Monti, 1997; Nordman and Lahiri, 2006; Nordman, 2008). The empirical likelihood function has received great interest because it can be defined starting from a very general estimating function (Qin and Lawless, 1994; Owen, 2001).

Consider an estimating function Ψ_θ for the parameter θ. The empirical likelihood function can be defined by considering the class $\mathcal{D}$ of all the distribution functions having support on the sample $y = (y_1, \ldots, y_n)$. The generic element in $\mathcal{D}$ is

$$\hat{F}_\theta = \left\{ w_i(\theta) : \sum_{i=1}^n w_i(\theta) = 1, \sum_{i=1}^n w_i(\theta)\psi(\theta; y_i) = 0 \right\},$$

and the empirical likelihood function for θ is defined as

$$L_e(\theta) = \sup_{\hat{F}_\theta \in \mathcal{D}} \prod_{i=1}^n w_i(\theta). \tag{2.24}$$

Let $\ell_e(\theta) = \log L_e(\theta)$ be the empirical log likelihood function and let

$$W_e(\theta) = 2\left\{\ell_e(\tilde{\theta}) - \ell_e(\theta)\right\} = 2\sum_{i=1}^n \log \frac{w_i(\tilde{\theta})}{w_i(\theta)} \tag{2.25}$$

be the empirical log likelihood ratio. To compute (2.25), the distributions $\hat{F}_\theta$ and $\hat{F}_{\tilde{\theta}}$ are needed. When $\theta = \tilde{\theta}$, the distribution function $\hat{F}_{\tilde{\theta}}$ is nothing but the empirical distribution function $\hat{F}_n = \left\{w_i(\tilde{\theta}) = 1/n,\, i = 1, \ldots, n\right\}$

(Owen, 1988). For generic $\theta \in \Theta$, $\hat{F}_\theta$ is obtained by optimizing in $\delta(\theta)$ and $\lambda(\theta)$ the Lagrange function

$$\sum_{i=1}^{n} w_i(\theta) + \delta(\theta)\left(\sum_{i=1}^{n} w_i(\theta) - 1\right) + \lambda(\theta)^\top \sum_{i=1}^{n} \psi(\theta; y_i),$$

where $\delta(\theta) \in \mathbb{R}$ and $\lambda(\theta) \in \mathbb{R}^p$ are Lagrange multipliers. Following Owen (1988, 1990), $\delta(\theta) = 1$ and the elements of $\hat{F}_\theta$ have the following expression

$$w_i(\theta) = \frac{1}{n\left(1 + \lambda(\theta)^\top \psi(\theta; y_i)\right)}.$$

The Lagrange multiplier $\lambda(\theta)$ solves the equation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\psi(\theta; y_i)}{\left(1 + \lambda(\theta)^\top \psi(\theta; y_i)\right)} = 0.$$

Since $\hat{F}_{\tilde{\theta}}$ is already known, the empirical log likelihood ratio $W_e(\theta)$ can be rewritten in terms of forward Kullback-Leibler divergence (Kullback, 1997) between $\hat{F}_{\tilde{\theta}}$ and $\hat{F}_\theta$, that is

$$W_e(\theta) = 2n\mathrm{KL}(\hat{F}_\theta; \hat{F}_{\tilde{\theta}}) = -2n\frac{1}{n}\sum_{i=1}^{n} \log\left(n w_i(\theta)\right). \tag{2.26}$$

Hence, the distribution function $\hat{F}_\theta$ that maximizes (2.24) is the closest distribution function to $\hat{F}_{\tilde{\theta}}$ in the sense of the forward Kullback-Leibler divergence, i.e. $\hat{F}_\theta$ minimizes (2.26). Deeper insights about the relation between the empirical likelihood function and the forward Kullback-Leibler divergence will be discussed in Chapter 5.

Although the empirical log likelihood ratio (2.25) is based on nonparametric estimates of empirical distributions in the class $\mathcal{D}$, it can be used both to set confidence regions and to test hypothesis about θ as with a parametric likelihood ratio. Indeed, there are some notable properties that relates $W_e(\theta)$ to the classical log likelihood ratio.

The asymptotic null distribution of the empirical log likelihood ratio $W_e(\theta)$ is chi-square with p degrees of freedom. Hence, Owen (1988, 1990) obtained the same limiting result obtained by Wilks (1938), but in a nonparametric setting. This result is due to the empirical likelihood's ability to Studentize internally. Indeed, by taking Taylor series expansion of (2.25), the following first-order equivalence relations hold

$$\begin{aligned} W_e(\theta) &= (\tilde{\theta} - \theta)^\top V(\tilde{\theta})^{-1}(\tilde{\theta} - \theta) + O_p(n^{-1/2}), \\ W_e(\theta) &= \Psi_\theta^\top J(\theta)^{-1}\Psi_\theta + O_p(n^{-1/2}), \end{aligned} \tag{2.27}$$

as $\ell_e(\theta)$ was a genuine log likelihood function.

When $\theta = (\tau, \lambda)$, with τ p_0-dimensional parameter of interest, a profile version of (2.25) can be considered. Following the same notation of Section 2.2.1, the profile empirical log likelihood ratio is

$$W_{ep}(\tau) = \inf_{\lambda} W_e(\tau, \lambda), \tag{2.28}$$

whose asymptotic null distribution is chi-square with p_0 degrees of freedom. The numerical optimization of (2.28) might be computational intensive when the dimension of λ is large. Approximations to the profile empirical log likelihood ratio for a scalar parameter of interest are derived in DiCiccio and Monti (2001).

First-order theory highlights the relationships between empirical and parametric likelihoods. Furthermore, higher-order asymptotics developed by DiCiccio *et al.* (1991) show that, in the smooth function of means model (Bhattacharya and Ghosh, 1978), empirical log likelihood ratios are Bartlett-correctable. This is appealing since the Bartlett correction had previously been available only for parametric likelihoods. Indeed, $W_e(\theta)$ is asympotically chi-square distributed up to an error of magnitude $O_p(n^{-1})$, but due to its nonparametric nature the convergence to its distribution may be slow when the sample size is moderate or small. The use of a Bartlett correction means that a simple adjustment for the expected value of $W_e(\theta)$ improves the error of the approximation up to $O_p(n^{-2})$. Bartlett corrections in the presence of nuisance parameters are available for the profile empirical log likelihood ratio; the reader can refer to Chen and Cui (2006).

Further discussions about the methodology and algorithms for empirical likelihood can be found in Hall and La Scala (1990) and Owen (1990). Topics related to Bartlett correctability in the presence of nuisance parameters, and higher-order asymptotics can be found in DiCiccio and Romano (1989), Chen (1993) and Lazar and Mykland (1999). Extensions of the empirical likelihood methodology are developed in Hjort *et al.* (2009).

2.6.3 Numerical examples

Section 2.6.1 reviewed two approaches that lead to quasi-log likelihood ratios having the usual chi-square limiting distribution. In particular, the first approach aims at correcting directly the quasi-log likelihood ratio, whereas the second one focuses on recovering, at least approximately, the second Bartlett's identity for Ψ_θ. On the other hand, in Section 2.6.2 the empirical log likelihood ratio has been shown to be asymptotically chi-square distributed, without resorting to any kind of adjustment.

Quasi- and empirical log likelihoods are derived following two very different paths. The examples considered in this Section focus on a scalar parameter of interest and are intented to compare the coverage accuracy of quasi-log likelihood ratios given in (2.20) and (2.23), and empirical log likelihood ratios given in (2.28). In particular, we discuss two examples in order

to compare the finite-sample behaviour of $W_{QP}^{hl}(\tau)$, $W_{QP}^{av}(\tau)$ and $W_{ep}(\tau)$ for well-known robust M-estimators in the context of linear transformation models, which includes location and scale models and linear regression models. Monte Carlo studies have been performed in order to evaluate both the accuracy of confidence regions when the model is correctly specified, and to assess the stability of the coverage levels under small departures from the assumed model. We consider a contamination model of the form $F_\epsilon = (1-\epsilon)F_\theta + \epsilon H$, where $H(\cdot)$ denotes the contamination distribution and ϵ the contamination percentage, set at 5%.

Example 2.1: Location and scale model

Let $\theta = (\mu, \sigma)$, where $\mu \in \mathbb{R}$ is a location parameter and $\sigma > 0$ a scale parameter. In this framework, $F(y;\mu,\sigma) = F_0((y-\mu)/\sigma)$ and $\psi(y;\mu,\sigma) = \psi((y-\mu)/\sigma)$, with $F_0(\cdot)$ standard element of the family, assumed symmetric around the origin.

Consider inference about μ when σ is the nuisance parameter. The M-estimators used are the well-known Huber estimator for location and scale, and the biweight estimator (Hampel *et al.*, 1986, Sections 4.2 and 2.6) obtained, respectively, as solutions of the estimating functions

$$\Psi_{HF}(\theta) = \left(\sum_{i=1}^{n} \psi_{k_1}(\theta; r_i), \sum_{i=1}^{n} \psi_{k_2}(\theta; r_i)^2 - n\kappa(k_2) \right)$$

and

$$\Psi_{BIW}(\theta) = \left(\sum_{i=1}^{n} \psi_{\nu}^{b}(\theta; r_i), \sum_{i=1}^{n} \psi_{k_2}(\theta; r_i)^2 - n\kappa(k_2) \right),$$

with $r_i = (y_i - \mu)/\sigma$, $\psi_c(x) = \min\{c, \max\{-c, x\}\}$, $\psi_\nu^b(x) = x(\nu^2 - x^2)^2 I_{[-\nu,\nu]}(x)$, and $\kappa(k_2)$ consistency factor (see, *e.g.* Huber and Ronchetti, 2009). For this example, the quasi-profile and the empirical profile log likelihoods for μ are given in Adimari and Ventura (2002). The quasi-profile log likelihood ratio statistic of Hanfelt and Liang (2.20) is given by

$$W_{QP}^{hl}(\mu) = 2\frac{A_{\mu\mu}}{\tilde{\sigma}} \sum_{i=1}^{n} \int_{\mu}^{\tilde{\mu}} \psi_\mu \left(\frac{y_i - t}{\tilde{\sigma}_t} \right) dt,$$

where $\tilde{\sigma}_t$ is the estimate of σ when μ is considered known and set equal to t. When the central model is the normal one and the Huber estimator is used, the factor $A_{\mu\mu}$ is given by

$$A_{\mu\mu} = \frac{\Phi(k_1) - \Phi(-k_1)}{2(k_1^2\Phi(-k_1) - k_1\phi(k_1) + \Phi(k_1) - 1/2)},$$

for a given tunning constant $k_1 > 0$.

Table 2.1 gives the results of a Monte Carlo experiment, based on 10.000 trials, that compares confidence intervals for μ based on $W^{hl}_{QP}(\mu)$, $W^{av}_{QP}(\mu)$ and $W_{ep}(\mu)$ when the central model is the normal one. Data are generated from three different distributions: N(0,1), N(0,1) contaminated by a N(0,25), and N(0,1) contaminated by a half normal with mean 4. Huber and biweight estimators are used with $k_1 = 1.1$, $k_2 = 0.6$ and $v = 4$, respectively. From Table 2.1 it can be noted that inference based on quasi-likelihoods seems to be satisfactory, although inference on $W^{av}_{QP}(\mu)$ is slightly preferable than that based on $W^{hl}_{QP}(\mu)$. The empirical likelihood method seems to yield intervals with coverage closer to the nominal one and similar to $W^{av}_{QP}(\mu)$. Finally, we note that simulation studies, not reported here, show that the three methods are equivalent for $n \geq 50$.

distribution	n	$1-\alpha$	0.999	0.950	0.900	0.999	0.950	0.900
				Ψ_{HF}			Ψ_{BIW}	
N(0,1)	10	W^{av}_{QP}	0.971	0.917	0.863	0.979	0.928	0.877
		W^{hl}_{QP}	0.958	0.904	0.853	0.959	0.904	0.855
		W_{ep}	0.975	0.938	0.886	0.974	0.941	0.898
	20	W^{av}_{QP}	0.985	0.932	0.887	0.984	0.938	0.896
		W^{hl}_{QP}	0.974	0.928	0.879	0.975	0.930	0.883
		W_{ep}	0.988	0.946	0.894	0.989	0.949	0.897
N(0,1) cont. by half normal	10	W^{av}_{QP}	0.971	0.912	0.856	0.979	0.922	0.872
		W^{hl}_{QP}	0.957	0.897	0.843	0.955	0.897	0.846
		W_{ep}	0.976	0.937	0.884	0.973	0.942	0.894
	20	W^{av}_{QP}	0.980	0.929	0.878	0.983	0.939	0.888
		W^{hl}_{QP}	0.974	0.922	0.869	0.974	0.926	0.874
		W_{ep}	0.986	0.942	0.887	0.988	0.945	0.887
N(0,1) cont. by N(0,25)	10	W^{av}_{QP}	0.971	0.914	0.861	0.980	0.933	0.886
		W^{hl}_{QP}	0.958	0.904	0.852	0.959	0.911	0.859
		W_{ep}	0.975	0.943	0.889	0.976	0.946	0.898
	20	W^{av}_{QP}	0.979	0.932	0.878	0.985	0.945	0.898
		W^{hl}_{QP}	0.972	0.922	0.872	0.975	0.931	0.885
		W_{ep}	0.988	0.942	0.889	0.988	0.945	0.894

Table 2.1: Simulation study: empirical coverage probabilities in the location and scale model.

Example 2.2: Linear model

A regression and scale model has the form $y = X\beta + \sigma\epsilon$, where X is a fixed $n \times p$ matrix, $\beta \in \mathbb{R}^p$ an unknown regression coefficient, $\sigma > 0$ a scale parameter and ϵ an n-dimensional vector of random errors from a known distribution $F_0(\cdot)$ symmetric around 0. Let $\theta = (\beta, \sigma)$. A wide class of M-estimators for regression and scale parameters can be obtained by generalizing the Huber estimator and it includes the Hampel-Krasker estimator (see Maronna *et al.*, 1979). If interest is about a scalar component β_j, $1 \leq j \leq p$, the quasi-profile

log likelihood ratio $W_{QP}^{av}(\beta_j)$ for β_j is given in Adimari and Ventura (2002), while $W_{QP}^{hl}(\beta_j)$ is given by

$$W_{QP}^{hl}(\beta_j) = \frac{w}{\tilde{\sigma}} \sum_{i=1}^{n} \int_{\beta_j}^{\tilde{\beta}_j} \frac{x_{ij}}{||x_i||} \psi_{k_1}\left(||x_i|| \tilde{r}_{it}\right) dt,$$

with $\tilde{r}_{it} = (y_i - \tilde{\beta}_{1t} x_{i1} - \ldots - t x_{ij} - \ldots - \tilde{\beta}_{dt} x_{id})/\tilde{\sigma}_t$, $\tilde{\beta}_{kt}$, $k \neq j$, and $\tilde{\sigma}_t$ estimates of β_k and σ, respectively, when β_j is set equal to t. The correction term w has the general form

$$w = \frac{H_{\beta_j,\beta_j} - \gamma^\top H_{(-j)}^{-1} \gamma_{\beta_j}}{J_{\beta_j,\beta_j} - 2\gamma_{\beta_j}^\top H_{(-j)}^{-1} \eta_{\beta_j} + \gamma_{\beta_j}^\top H_{(-j)}^{-1} J_{(-j)} (H_{(-j)}^{-1})^\top \gamma_{\beta_j}},$$

where H_{β_j,β_j} is the jth diagonal element of H, γ_{β_j} is the jth column of H without its jth element, $H_{(-j)}$ denotes H without the jth column and the jth row, and η_{β_j} is the jth column of Ω without its jth element. Furthermore, it is possible to consider a profile empirical log likelihood ratio given by

$$W_{ep}(\beta_j) = \inf_{\beta_j} W_e(\beta, \sigma),$$

where $W_e(\beta, \sigma)$ is given in (2.25) for $\theta = (\beta, \sigma)$.

Consider the model $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, $i = 1, \ldots, n$, computed from Draper and Smith data (see Hampel *et al.*, 1986, Section 7.5d). The variables considered are the number of pounds of steam used per month (y_i), the average atmospheric temperature (x_{i2}) and the number of operating days in the month (x_{i3}). The sample size is $n = 25$, the normal model is assumed as the central one and the Hampel-Krasker estimator is used with $k_1 = 1.1$ and $k_2 = 0.6$. Data are generated from three different distributions: N(0,1), N(0,1) contaminated with N(4,1), and N(0,1) contaminated with N(0,25). Table 2.2 gives the results of a Monte Carlo experiment based on 5000 trials that compares confidence intervals for β_3 based on $W_{QP}^{av}(\beta_3)$ and $W_{QP}^{hl}(\beta_3)$. From Table 2 we can see that, for $n = 25$, $W_{QP}^{av}(\beta_3)$ and $W_{ep}(\beta_3)$ are preferable than $W_{QP}^{hl}(\beta_3)$ in order to construct robust confidence intervals for all the scenarios considered, even if $W_{QP}^{av}(\beta_3)$ seems to be slightly preferable

2.7 Final remarks

In this Chapter an overview of the estimating function theory has been summarized, in particular by focusing on some aspects related to optimal estimating functions and to their role in the context of robust statistic. Moreover, associated pseudo-likelihood functions, namely the quasi-likelihood and the empirical likelihood functions, have been reviewed and compared through

distribution	$1-\alpha$	0.999	0.950	0.900
N(0,1)	W_{QP}^{av}	0.992	0.958	0.911
	W_{QP}^{hl}	0.968	0.930	0.873
	W_{ep}	0.996	0.969	0.928
N(0,1) cont. by N(4,1)	W_{QP}^{av}	0.995	0.962	0.916
	W_{hl}	0.967	0.929	0.871
	W_{ep}	0.996	0.971	0.941
N(0,1) cont. by N(0,25)	W_{QP}^{av}	0.996	0.964	0.914
	W_{QP}^{hl}	0.971	0.922	0.872
	W_{ep}	0.997	0.971	0.937

Table 2.2: Simulation study: empirical coverage probabilities in the linear model.

a small simulation study. While the arisen issues do not claim to cover the considered topic in detail, they represent a useful background to better understand the subsequent developments of the present thesis. As a matter of fact, next Chapter is devoted to introduce composite likelihood functions, as a general framework to encompass the various issues treated in this work.

Chapter 3

Composite likelihoods

3.1 Introduction

The likelihood function has become a centerpiece of statistical inference since it was turned into a powerful tool by Fisher. Nevertheless, likelihood inference may be sometimes troublesome. For instance, it may happen that the joint distribution is difficult to be specified and the resulting likelihood function is awkward due to a complex dependence structure of the data. An example is provided by the use of max-stable processes for spatial multivariate extremes (Padoan *et al.*, 2010). On the other hand, the specification of the joint distribution can be straightforward, but the evaluation of the likelihood function might lead to computational burden. For instance, modeling a spatial process with a Gaussian random field requires the determinant and the inverse of the process' covariance matrix, whose dimension grows as the number of observed sites increases (Stein *et al.*, 2004).

In order to cope with these difficulties both in model specification or in computations, the use of a surrogate for the ordinary likelihood may prove useful for inferential purposes. Preliminary formulations of this idea date back to Besag (1974) who used pseudo-likelihoods to model spatial processes and to Cox (1975), that introduced the partial likelihood to fit proportional hazards models. These solutions fall within the general class of composite likelihood functions (Lindsay, 1988), that includes the full likelihood as a special case. In principle, composite likelihoods arise from the use of a partially misspecified model, and hence a partially misspecified likelihood function.

From a theoretical point of view, composite likelihoods are appealing since the validity of the derived inferential procedures can be assessed both from the standpoint of unbiased estimating functions and of the Kullback-Leibler criterion (Varin and Vidoni, 2005; Lindsay *et al.*, 2011; Varin *et al.*, 2011). Their use has been widely advocated by several authors both in the frequentist domain (Varin, 2008; Varin *et al.*, 2011) and, more recently, also in the Bayesian setting (Smith and Stephenson, 2009; Pauli *et al.*, 2011).

Recent reviews on composite likelihoods may be found in Varin (2008) and Varin *et al.* (2011). Composite likelihoods have shown a great impact also in practical applications where complex models are involved. Some examples are spatial processes (Hjort and Omre, 1994; Heagerty and Lele, 1998; Varin *et al.*, 2005), generalized linear mixed models (Renard *et al.*, 2004; Bellio and Varin, 2005), multivariate extremes (Padoan *et al.*, 2010), longitudinal models (Fieuws and Verbeke, 2006), time series (Davis and Yau, 2011), and genetics (Hudson, 2001; McVean *et al.*, 2002, 2004).

In this chapter the main properties of composite likelihood functions are reviewed by focusing, in particular, on a specific instance of composite likelihoods, namely the pairwise likelihood functions. Then, some issues emerging in inference based on composite likelihood functions are discussed. Especially, the interest is addressed on the problems related to the asymptotic distribution of the composite log likelihood ratio statistic and on the lack of robustness of the maximum composite likelihood estimator. The problems raised in the following sections will be then faced in the subsequent chapters, where some possible solutions are proposed.

3.2 Composite likelihoods

Let $Y \in \mathbb{R}^q$ be a random vector with probability distribution $F(y;\theta)$, $\theta \subseteq \mathbb{R}^p$, density function $f(y;\theta)$, and associated full log likelihood function $\ell(\theta) = \log f(y;\theta)$. Consider a set of marginal or conditional measurable events $\{\mathcal{E}_r \in \mathcal{Y},\, r = 1,\dots,m\}$ and let $f_r(y;\theta) = f(y \in \mathcal{E}_r;\theta)$ be the likelihood contribution generated from $f(y;\theta)$ by considering the set $\mathcal{E}_r$. Having observed a random sample $(y_1,\dots,y_n)$ of size n from Y, the composite likelihood is defined as the product

$$cL(\theta) = \prod_{i=1}^{n}\prod_{r=1}^{m} f_r(y_i;\theta)^{\omega_{ir}},$$

where ω_{ir} are non-negative weights, $i = 1,\dots,n$, $r = 1,\dots,m$. The composite log likelihood is

$$c\ell(\theta) = \log cL(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{m} \omega_{ir} \log f_r(y_i;\theta).$$

The former definitions are rather general, and the particular specification of the events $\mathcal{E}_r$ allows the combination of both marginal and conditional densities. Conditional composite log likelihoods are obtained by a suitable specification of the events $\mathcal{E}_r$ defining the conditional densities $f(y \in \mathcal{E}_r;\theta) = f_{X|Z}(x|z;\theta) = f(x|z;\theta)$, with X and Z sub-components of Y. For instance,

by pooling together all the possible pairwise conditional densities, we obtain

$$c\ell(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{m}\sum_{s=1}^{m} \omega_{irs} \log f(y_{ir}|y_{is};\theta),$$

or all the full conditional densities

$$c\ell(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{m} \omega_{ir} \log f(y_{ir}|y_{i(-r)};\theta),$$

where w_{irs} are the weights for the (r,s)-component of unit i and $y_{i(-r)}$ denotes the vector of all the observations but y_{ir}, $i = 1,\ldots,n$, $r = 1,\ldots,m$.

Marginal composite likelihoods are defined by considering events $\mathcal{E}_r$ giving rise to marginal densities of the form $f(y \in \mathcal{E}_r;\theta) = f_{X,Z}(x,z;\theta) = f(x,z;\theta)$. The special case of the independence log likelihood is derived under working independence assumptions, i.e.

$$c\ell_{ind}(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{m} \omega_{ir} \log f(y_{ir};\theta),$$

while the pairwise log likelihood is

$$c\ell_{pw}(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{m-1}\sum_{s=r+1}^{m} \omega_{irs} \log f(y_{ir},y_{is};\theta). \tag{3.1}$$

Recently, there have been some proposals aimed to build improved marginal composite likelihoods by combining the information supplied from likelihood contributions suited for the marginal and for the association parameters (Cox and Reid, 2004; Kuk, 2007).

The validity of using composite likelihoods to perform inference about θ can be assessed throughout the theory of unbiased estimating functions. The maximum composite likelihood estimator $\hat{\theta}_c$ is defined implicitly as the solution of the composite score equation

$$cs(\theta) = cs(\theta;y) = \sum_{i=1}^{n}\sum_{r=1}^{m} \omega_{ir} \frac{\partial \log f_r(y_i;\theta)}{\partial\theta} = 0.$$

Since $cs(\theta)$ belongs to the class of unbiased estimating functions, $\hat{\theta}_c$ inherits the properties of M-estimators, reviewed in Chapter 2. Under regularity conditions assumed throughout this chapter (see, *e.g.*, Molenberghs and Verbeke, 2005), the maximum composite likelihood estimator is consistent and asymptotically normal, with mean θ and covariance matrix given by the inverse of the Godambe information, i.e.

$$\begin{aligned} V(\theta) &= G(\theta)^{-1} \\ &= \mathbb{E}\left(-\frac{cs(\theta;Y)}{\partial\theta^\top}\right)^{-1} \mathbb{E}\left(cs(\theta;Y)cs(\theta;Y)^\top\right) \left[\mathbb{E}\left(-\frac{cs(\theta;Y)}{\partial\theta^\top}\right)^{-1}\right]^\top \\ &= H(\theta)^{-1}J(\theta)\left(H(\theta)^{-1}\right)^\top. \end{aligned}$$

In the composite likelihood framework, hypothesis testing and confidence regions can be obtained by using the analogous of the Wald, the score and the log likelihood ratio tests. The composite likelihood counterparts of the Wald and score tests are, respectively, given by

$$cW_w(\theta) = (\hat{\theta}_c - \theta)^\top V(\hat{\theta}_c)^{-1}(\hat{\theta}_c - \theta)$$

and

$$cW_s(\theta) = cs(\theta)^\top J(\theta)^{-1} cs(\theta),$$

with a standard χ^2_p asymptotic distribution. Instead, the composite log likelihood ratio

$$cW(\theta) = 2\left\{c\ell(\hat{\theta}_c) - c\ell(\theta)\right\}$$

converges in distribution to $\sum_{j=1}^{p} \lambda_j(\theta) Z_j^2$, with $\lambda_j(\theta)$ eigenvalues of $H(\theta)^{-1}J(\theta)$ and Z_j independent random variables having a standard normal distribution (see also Section 2.6.1).

In the presence of nuisance parameters, the profile versions of the aforementioned statistics can be considered. Consider the partition of θ into a component of interest τ and a nuisance component λ, whose dimensions are p_0 and $(p - p_0)$, respectively. The composite score function is similarly partitioned as $cs(\theta) = (cs_\tau(\theta), cs_\lambda(\theta))$. Let $\hat{\lambda}_{c\tau}$ be the root in λ of the equation $cs_\lambda(\tau, \lambda) = 0$ for fixed τ, and let $\hat{\tau}_c$ be the solution of the profile composite score equation

$$cs_\tau(\tau, \hat{\lambda}_{c\tau}) = 0.$$

The profile versions of the composite Wald and score test statistics are, respectively,

$$cW_{wp}(\tau) = (\hat{\tau}_c - \tau)^\top V(\hat{\theta}_{c\tau})^{\tau\tau} (\hat{\tau}_c - \tau) \tag{3.2}$$

and

$$cW_{sp}(\tau) = cs_\tau(\hat{\theta}_{c\tau})^\top H(\hat{\theta}_{c\tau})_{\tau\tau} V(\hat{\theta}_{c\tau})^{\tau\tau} H(\hat{\theta}_{c\tau})_{\tau\tau} cs_\tau(\hat{\theta}_{c\tau}), \tag{3.3}$$

with $\hat{\theta}_\tau = (\tau, \hat{\lambda}_{c\tau})$. Both (3.2) and (3.3) have a null asymptotic $\chi^2_{p_0}$ distribution. The profile composite log likelihood ratio is

$$cW_p(\tau) = 2\left\{c\ell(\hat{\theta}_c) - c\ell(\hat{\theta}_{c\tau})\right\} \tag{3.4}$$

and converges to $\sum_{j=1}^{p_0} \lambda_j(\theta) Z_j^2$, with $\lambda_j(\theta)$ eigenvalues of $\{H(\theta)^{\tau\tau}\}^{-1} G(\theta)^{\tau\tau}$ (Molenberghs and Verbeke, 2005).

3.3 Pairwise likelihood

This section is devoted to a particular composite likelihood, i.e. the pairwise likelihood function (3.1), since it will be considered in the main contributions of the present thesis. In the following, the paper by Cox and Reid (2004) is reviewed, since some theory for the pairwise likelihood function is developed.

Cox and Reid (2004) define a pairwise log likelihood function for θ by pooling together both bivariate and univariate densities, i.e.

$$\begin{aligned} p\ell(\theta) &= \sum_{i=1}^{n} p\ell(\theta; y_i) \\ &= \sum_{i=1}^{n} \left\{ \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} \log f(y_{ir}, y_{is}; \theta) - a \sum_{r=1}^{q} \log f(y_{ir}; \theta) \right\} \\ &= c\ell_{pw}(\theta) - a \cdot c\ell_{ind}(\theta), \end{aligned}$$

where a is a suitable constant. Note that in the above definition the weights w_{irs} in (3.1) are taken all equal to one. The choice $a = 0$ corresponds to take all possible bivariate marginal distributions, that is (3.1). This choice is appropriate if $c\ell_{ind}(\theta)$ is independent of θ, i.e. when the one-dimensional marginal distributions contain no information about θ. In those rare cases where most of the information is in $c\ell_{ind}(\theta)$ and none or relatively little in $c\ell_{pw}(\theta)$, a negative value of a would be needed, although in most cases a null or positive value of a is appropriate. In the following, it is shown that the choice of a can be crucial in order to guarantee the consistency of the maximum pairwise likelihood estimator.

The pairwise score function is given by

$$\begin{aligned} ps(\theta) &= \sum_{i=1}^{n} ps(\theta; y_i) = \\ &= \sum_{i=1}^{n} \left\{ \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} \frac{\partial \log f(y_{ir}, y_{is}; \theta)}{\partial \theta} - a \sum_{r=1}^{q} \frac{\partial \log f(y_{ir}; \theta)}{\partial \theta} \right\} \end{aligned} \tag{3.5}$$

The maximum pairwise likelihood estimator $\hat{\theta}_p$ is defined implicitly throughout the pairwise score equation $ps(\theta) = 0$. In the standard setting, where n diverges and q is fixed, the maximum pairwise likelihood estimator shares the same properties of the maximum composite likelihood estimator, i.e. it is consistent and asymptotically normally distributed with mean θ and covariance matrix given by the inverse of the Godambe information. In this context, consistent estimates of $J(\theta)$ and $H(\theta)$ are, respectively,

$$\hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^{n} ps(\theta; y_i) ps(\theta; y_i)^\top, \tag{3.6}$$

and

$$\hat{H}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial ps(\theta; y_i)}{\partial \theta^\top}. \tag{3.7}$$

Interesting results emerge when a small number n of individually large sequences is available, i.e. in the setting where q diverges and n is fixed. In

this context, the pairwise score function $ps(\theta)$ is still an unbiased estimating function, but this no longer implies satisfactory properties of the resulting estimator. For practical purposes, an important situation is when $n = 1$, that is when dealing with time series or spatial processes.

Consider a Taylor series expansion for $ps(\hat{\theta}_p)$ around θ. The expansion, up to first order, gives

$$\begin{aligned} ps(\hat{\theta}_p) &= q^{-2}\left\{\sum_{r=1}^{q-1}\sum_{s=r+1}^{q}\frac{\partial \log f(y_r, y_s;\theta)}{\partial\theta} - aq\sum_{r=1}^{q}\frac{\partial \log f(y_r;\theta)}{\partial\theta}\right\} + \\ &+ q^{-2}(\hat{\theta}_p-\theta)^\top\left\{\sum_{r=1}^{q-1}\sum_{s=r+1}^{q}\frac{\partial^2 \log f(y_r, y_s;\theta)}{\partial\theta\partial\theta^\top} + \right. \\ &- \left. aq\sum_{r=1}^{q}\frac{\partial^2 \log f(y_r;\theta)}{\partial\theta\partial\theta^\top}\right\} + O_p(q^{-1/2}). \end{aligned}$$

The second term is typically $O_p(1)$, whereas the first term has zero mean and variance Var $(ps(\theta; Y))$ which is rather complicated and it is not reported here (the reader can refer to Cox and Reid (2004)).TheA consistent estimator of θ can be obtained only if $q^{-4}\text{Var}\,(ps(\theta; Y)) \to 0$ as $q \to \infty$, and a sufficient and necessary condition for this is that there exists a real root a of the equation

$$\begin{aligned} &\mathbb{E}\left[\frac{\partial \log f(Y_r, Y_s;\theta)}{\partial\theta}\left(\frac{\partial \log f(Y_w, Y_u;\theta)}{\partial\theta}\right)^\top\right] + \\ - \;& 2a\mathbb{E}\left[\frac{\partial \log f(Y_r, Y_s;\theta)}{\partial\theta}\left(\frac{\partial \log f(Y_w;\theta)}{\partial\theta}\right)^\top\right] + \\ + \;& 2a^2\mathbb{E}\left[\frac{\partial \log f(Y_r;\theta)}{\partial\theta}\left(\frac{\partial \log f(Y_r;\theta)}{\partial\theta}\right)^\top\right] = 0, \end{aligned}$$

with the indexes $r, s, w, u = 1, \ldots, q$ all different.

Although the results of Cox and Reid (2004) give some insights about the consistency of $\hat{\theta}_p$ in the case of increasing q with fixed or slowly increasing sample size n, there does not seem to be a rigorous and general proof about the consistency of composite maximum likelihood estimators under various conditions on q and n.

As a final remark, in the following chapters the pairwise log likelihood function is considered with $a = 0$.

3.4 Some issues

In this Section two particular issues related with composite likelihood-based inference are reviewed. They will be pursued in Chapters 4, 5, and 6, in which new theoretical results and insights are provided.

3.4.1 Test statistics

In Section 3.2, the asymptotic distribution of composite likelihood test statistics has been given. In particular, the composite Wald and score test statistics keep the standard asymptotic behavior. On the contrary, the composite log likelihood ratio is asymptotically distributed as a linear combination of independent chi-square random variables, because of the failure of the second Bartlett's identity for the composite score function.

At a first glance, composite log likelihood ratios with a standard asymptotic distribution can be obtained, in principle, following the developments applied in the quasi-likelihood framework (see Section 2.6.1). The first possibility is to apply a linear transformation to $cs(\theta)$, of the from (2.21), but this will lead to a composite score function that might not be the primitive of any composite log likelihood function. For this reason, in the composite likelihood framework the approach used by Hanfelt and Liang (1995) is preferred (see Section 2.6.1). The idea is to correct directly $cW(\theta)$ by means of suitable scaling factors in order to preserve the objective function $c\ell(\theta)$. Adjusted versions of composite log likelihood ratios with a standard limiting distribution can be derived from three different approaches.

The first approach is based on moment adjustments to $cW(\theta)$, which lead to composite log likelihood ratios whose moments match some of those of the chi-square distribution. For instance, an adjusted composite log likelihood ratio, whose asymptotic distribution can be approximated by a χ^2_p, is obtained by considering a matching of the first moment, i.e.

$$cW_1(\theta) = \frac{cW(\theta)}{\kappa_1},$$

with $\kappa_1 = \sum_{j=1}^{p} \lambda_j(\theta)/p$ (see also Hanfelt and Liang, 1995). This test statistic was suggested by Rotnitzky and Jewell (1990) for the independence likelihood, and by Geys *et al.* (1999) for general pseudo-likelihoods. The statistic $cW_1(\theta)$ is simple to compute, but the chi-square approximation might be inaccurate since κ_1 corrects only the first moment of $cW(\theta)$. Other moment-based adjustments can be considered. For instance, first and second moment matching gives the Satterthwaite-type (Satterthwaite, 1946) adjustment suggested in Varin *et al.* (2011), whereas matching of moments up to higher-order have been considered in Wood (1989) and Lindsay *et al.* (2000).

The second class of adjusted composite likelihood statistics are given by the proposals of Chandler and Bate (2007) and Pace *et al.* (2011). The first one has the following expression

$$cW_{cb}(\theta) = cW(\theta)\frac{(\hat{\theta}_c - \theta)^\top V(\hat{\theta}_c)^{-1}(\hat{\theta}_c - \theta)}{(\hat{\theta}_c - \theta)^\top H(\hat{\theta}_c)(\hat{\theta}_c - \theta)}, \tag{3.8}$$

while the second is given by

$$cW_{inv}(\theta) = cW(\theta)\frac{cs(\theta)^\top J(\theta)^{-1} cs(\theta)}{cs(\theta)^\top H(\theta)^{-1} cs(\theta)}.$$

The test statistic (3.8) essentially stretches the composite log likelihood on the θ-axis about $\hat{\theta}_c$ to ensure that at least approximately, the second Bartlett's identity holds. The test statistic $cW_{inv}(\theta)$ can be derived from (3.8) by considering the formal relation $(\hat{\theta}_c - \theta) = H(\theta)^{-1} cs(\theta) + O_p(n^{-1/2})$. Both the test statistics are asymptotically χ^2_p distributed. Note that $cW_{cb}(\theta)$ and $cW_{inv}(\theta)$ are equivalent, up to first order, to $cW_w(\theta)$ and $cW_s(\theta)$, respectively.

Finally, the third way to obtain adjusted composite log likelihood ratios is by resorting to the bootstrap. The parametric bootstrap is used in order to approximate the distribution of $cW(\theta)$ without any additional estimation of the eigenvalues $\lambda_j(\theta)$, $j = 1, \ldots, p$. Indeed, the bootstrap automatically corrects for the misspecification of the joint distribution. An example of this technique is given in Aerts and Claeskens (1999). However, this approach can be quite demanding from a computational point of view and it can be applied in a limited range of applications, since the specification of the joint distribution is required.

The statistics $cW_1(\theta)$ and $cW_{inv}(\theta)$, can be used in the presence of nuisance parameters. In particular, it is possible to derive

$$cW_{1p}(\tau) = \frac{cW_p(\tau)}{\kappa_{1p}}, \tag{3.9}$$

with $\kappa_{1p} = \sum_{j=1}^{p_0} \lambda_j(\theta)$, or

$$cW_{invp}(\tau) = cW_p(\tau)\frac{cW_{sp}(\tau)}{cs_\tau(\hat{\theta}_\tau)^\top H(\hat{\theta}_\tau)^{\tau\tau} cs_\tau(\hat{\theta}_\tau)}. \tag{3.10}$$

The asymptotic null distribution of (3.9) and (3.10) is $\chi^2_{p_0}$.

The computation of the quantiles of the asymptotic distribution of $cW(\theta)$ relies both on the theory and the algorithms provided by Imhof (1961). The problem to face is related to the dependence of almost all the composite likelihood-based test statistics on the elements of the expected Godambe information. Analytic expressions for $J(\theta)$ and $H(\theta)$ can be worked out when the joint distribution is specified and this is usually done in simple cases only, in order to compare the composite likelihood procedures with those based on the likelihood function. Furthermore, the specification of the joint distribution is not affordable in the composite likelihood context, because it is too complex to deal with, thereby inference is based on an approximate model. It follows that, for practical purposes, the expected Godambe information must be replaced with its observed counterpart.

All the test statistics derived from the composite log likelihood function take into an account for the uncertainty of the estimation of $J(\theta)$ and $H(\theta)$. Indeed, inaccurate estimates may lead either to a slowdown or to a failure of the convergence to the distribution of the test statistics.

The matrix $H(\theta)$ is easy to estimate. In particular, it is possibile to provide an alternative expression to (2.5), that does not require the computation of the second derivatives of $c\ell(\theta)$, given by

$$\hat{H}(\theta) = \sum_{i=1}^{n} \sum_{r=1}^{m} \omega_{ir}^2 \left[\left(\frac{\partial \log f_r(y_i;\theta)}{\partial \theta} \right) \left(\frac{\partial \log f_r(y_i;\theta)}{\partial \theta} \right)^\top \right],$$

since the second Bartlett's identity still holds for each contribution to the composite log likelihood function.

The estimation of $J(\theta)$ might be a hard task. Indeed, in the standard setting where $n \to \infty$ and q is fixed, the estimate proposed in (2.5), i.e.

$$\hat{J}(\theta) = \sum_{i=1}^{n} \left[\sum_{r=1}^{m} \omega_{ir} \frac{\partial \log f_r(y_i;\theta)}{\partial \theta} \right] \left[\sum_{r=1}^{m} \omega_{ir} \frac{\partial \log f_r(y_i;\theta)}{\partial \theta} \right]^\top,$$

is consistent. In longitudinal studies, where there are short time series for each subject, this estimate of $J(\theta)$ may be improved in terms of accuracy by using the bootstrap or the jackknife (see, *e.g.*, Lipsitz *et al.*, 1994). The jackknife estimate of $J(\theta)$ is

$$\hat{J}(\hat{\theta}_c)_{jack} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_c^{(-i)} - \hat{\theta}_c)(\hat{\theta}_c^{(-i)} - \hat{\theta}_c)^\top, \tag{3.11}$$

where $\hat{\theta}_c^{(-i)}$ is the maximum composite likelihood estimate with y_i removed from the sample, $i = 1, \ldots, n$. The estimate (3.11) can be computationally time demanding if the maximum composite likelihood estimate $\hat{\theta}_c$ is expensive to obtain. In these circumstances, (3.11) can be obtained by considering a first-order approximation, in which $\hat{\theta}_c^{(-i)}$ is approximated with a single step of the Newton-Raphson algorithm (Varin *et al.*, 2011).

When $q \to \infty$ and n is fixed, or when $q \gg n$, the estimation of $J(\theta)$ gives rise to some issues. This setting is common, for instance, in genetics in which q and n are the number of genes and subjects, respectively. The extreme case is for $n = 1$, that is when a time series or a spatial process is considered. The estimate of $J(\theta)$ depends on the mixing properties of the random field or on the possibility to obtain internal replications. The available estimators of $J(\theta)$ in this context are based on window subsampling (Heagerty and Lele, 1998; Heagerty and Lumley, 2000; Caragea and Smith, 2006). Broadly speaking, these estimators are based on the idea to create pseudo-independent sub-samples from y, $\mathcal{S}_b$, accordingly to some criterion. For instance, in time series the sub-samples S_b can be obtained by considering

observations that are contiguous. The general expression for window subsampling estimators is given by

$$\hat{J}(\theta)_{wsub} = \sum_{b=1}^{B} |\mathcal{S}_b| \left(\frac{\partial \log f(y \in \mathcal{S}_b; \theta)}{\partial \theta} \right) \left(\frac{\partial \log f(y \in \mathcal{S}_b; \theta)}{\partial \theta} \right)^{\top},$$

where $\mathcal{S}_b$ are suitable sub-regions, $|\mathcal{S}_b|$ denotes the cardinality of the set $\mathcal{S}_b$ and B is the number of sub-regions.

As a final remark, if (3.11) is too expensive to obtain, or if the conditions for ensuring the validity of window subsampling estimators are not satisfied, the estimate of $J(\theta)$ can be obtained via the parametric bootstrap, that however requires the specification of a joint distribution for the data.

3.4.2 Robustness

Most of the research in the composite likelihood framework is concerned in providing a reasonable model specification that leads to sensible inferential procedures, when dealing with complex models. Despite this is still a challenge, the robustness aspects of the composite likelihood-based procedures have not yet been deeply investigated. The only exception is given by the paper of Xu and Reid (2011) and is related to the robustness of the maximum composite likelihood estimator. To review this contribution, the discussion is restricted to marginal composite likelihoods.

The asymptotic behavior of the maximum composite likelihood estimator has been justified in Section 3.2, by using the theory of unbiased estimating functions. Deeper insights about the consistency of $\hat{\theta}_c$ can be achieved by considering the composite Kullback-Leibler criterion. Varin and Vidoni (2005) define the composite Kullback-Leibler divergence between the model $f(y;\theta)$ and the true (unknown) $h(y)$, as a linear combination of the Kullback-Leibler divergences for each component of the composite log likelihood, i.e.

$$\mathrm{KL}_c(f, h; \theta) = \sum_{i=1}^{n} \sum_{r=1}^{m} \mathbb{E} \left(\log h_r(Y_i) - \log f_r(Y_i; \theta) \right) w_{ir},$$

where $h_r(y) = h(y \in \mathcal{E}_r)$ and $\mathbb{E}(\cdot)$ is taken with respect to the true model $h(y)$. The composite Kullback-Leibler divergence preserves the non-negativity as does the ordinary one. This ensures, under some regularity conditions (Varin and Vidoni, 2005), that $\tilde{\theta}_c$ is consistent for the parameter value minimizing $\mathrm{KL}_c(\cdot)$, defined as

$$\theta^* = \arg\min_{\theta} \mathrm{KL}_c(f, h; \theta). \tag{3.12}$$

To understand the implications of (3.12), it is necessary to distinguish between the full and the marginal correct specification of the model. The

former states that model $f(y;\theta)$ is correctly specified if there exists a $\theta_0 \in \Theta$ such that $f(y;\theta_0) = h(y)$. The latter focuses on all the component families $\{f_r(y;\theta)\}$ and requires $f_r(y;\theta_0) = h_r(y)$ for all $r = 1,\ldots,m$, for some $\theta_0 \in \Theta$ (Xu and Reid, 2011).

The interesting case for studying the robustness of $\hat{\theta}_c$ is when the components of the composite likelihood are correctly specified, but the joint model is misspecified. This kind of robustness is referred as robustness to consistency by Xu and Reid (2011). From this point of view, by definition $f(y;\theta)$ is misspecified and the maximum composite likelihood estimator converges to θ^*. On the other hand, if $\hat{\theta}_c$ is calculated from the composite likelihood making use of the correctly specified lower dimensional margins only, then it still converges to the true parameter value without depending on the joint model. In these cases, the maximum composite likelihood estimator might be more reliable than a maximum likelihood estimator, since misspecifying a high dimensional complex joint density may be much more likely than misspecifying some simpler lower dimensional densities (see, *e.g.*, Varin, 2008; Xu and Reid, 2011).

Although the contribution of Xu and Reid (2011) establishes the robustness of maximum composite likelihood estimators with respect to model misspecifications, it is not clear how to link this result to the classical theory of robustness (see Section 2.4). Roughly speaking, robustness usually means obtaining the same inferential results under small deviations from the assumed model. The range of models is often considered to be small-probability perturbations of the assumed model, to reflect the sampling notion of occasional outliers. In the composite likelihood framework a first issue arises: it is not clear which the central model is, since the range of models to be considered are those consistent with the specified set of sub-models $\{f_r(y;\theta),\, r = 1,\ldots,m\}$. Hence, the definition of a gross error model of the form (2.10) is not straightforward. Nevertheless, something can be said about the B-robustness of the maximum composite likelihood estimator, since the composite score function is an unbiased estimating function and $\hat{\theta}_c$ is an M-estimator. In the following, the existence of a gross error model $\mathcal{P}_\epsilon(F_\theta^c)$ is assumed, where F_θ^c is supposed to be the central model that includes all the models consistent with the marginal densities $\{f_r(y;\theta),\, r = 1,\ldots,m\}$. Then, the influence function of $\hat{\theta}_c$ is

$$\mathrm{IF}(y;\hat{\theta}_c, F_\theta^c) = H(\theta)^{-1}\omega_r \frac{\partial \log f_r(y;\theta)}{\partial\theta}.$$

Each contribution $\partial \log f_r(y;\theta)/\partial\theta$ is a genuine score and, in general, it is unbounded, leading to an unbounded gross-error sensitivity. This implies the lack of robustness of $\hat{\theta}_c$. Hence, regardless the marginal correct specification of the model, $\hat{\theta}_c$ might not be consistent.

Example 3.1: Multivariate normal distribution

This example aims to show the lack of B-robustness of the maximum pairwise likelihood estimator. The model considered is the multivariate normal, with vector of means $(\mu, \ldots, \mu) \in \mathbb{R}^q$ and compound symmetric matrix Σ, having diagonal elements σ^2 and off-diagonal elements $\sigma^2\rho$, with $\rho \in (-1/(q-1), 1)$. This model has been widely studied in the composite likelihood framework by several authors. For instance, Cox and Reid (2004) compare the asymptotic variance of the maximum pairwise likelihood estimator with the maximum likelihood one, whereas (Pace *et al.*, 2011) compare the accuracy of the composite log likelihood ratios presented in Section 3.4.1.

This example focuses on the marginal pairwise log likelihood function given in (3.5), with $a = 0$, for the parameter $\theta = (\mu, \sigma^2, \rho)$. The multivariate normal distribution is particularly appealing in order to study the robustness of the maximum pairwise likelihood estimator. Indeed, in this case a marginal correct specification of the model is achieved. Therefore, by the results stated above, $\hat{\theta}_p$ will converge to the true parameter value rather than to a pseudo-true one.

Given a random sample $y = (y_1, \ldots, y_n)$ the pairwise log likelihood function is

$$\begin{aligned} p\ell(\theta) &= -\frac{nq(q-1)}{2}\log\sigma^2 - \frac{nq(q-1)}{4}\log(1-\rho^2) - \frac{q-1+\rho}{2\sigma^2(1-\rho^2)}SS_W + \\ &- \frac{q(q-1)SS_B + nq(q-1)(\bar{y}-\mu)^2}{2\sigma^2(1+\rho)}, \end{aligned}$$

where

$$SS_B = \sum_{i=1}^{n}\sum_{h=1}^{q}(y_{ih}-\bar{y}_i)^2 \ , \qquad SS_W = \sum_{i=1}^{n}(\bar{y}_i-\bar{y})^2,$$

with $\bar{y}_i = \sum_{h=1}^{q} y_{ih}/q$ and $\bar{y} = \sum_{i=1}^{n}\sum_{h=1}^{q} y_{ih}/nq$. The pairwise score function has components

$$\begin{aligned} ps_\mu(\theta; y) = \frac{\partial p\ell(\theta)}{\partial\mu} &= \frac{nq(q-1)(\bar{y}-\mu)}{\sigma^2(1-\rho)}, \\ ps_{\sigma^2}(\theta; y) = \frac{\partial p\ell(\theta)}{\partial\sigma^2} &= \frac{q-1+\rho}{2(\sigma^2)^2(1-\rho^2)}SS_W + \\ &+ \frac{q(q-1)\left\{SS_B + n(\bar{y}-\mu)^2\right\}}{2(\sigma^2)^2(1+\rho)^2} - \frac{nq(q-1)}{2\sigma^2}, \\ ps_\rho(\theta; y) = \frac{\partial p\ell(\theta)}{\partial\rho} &= -\frac{\rho^2+2\rho(q-1)+1}{2\sigma^2(1-\rho^2)^2}SS_W + \\ &+ \frac{q(q-1)\left\{SS_B + n(\bar{y}-\mu)^2\right\}}{2\sigma^2(1+\rho)^2} + \frac{nq(q-1)\rho}{2(1-\rho^2)}. \end{aligned}$$

To see whether the influence function of $\hat{\theta}_p$ is bounded or not, it is sufficient to study the behavior of $ps(\theta)$ (see Section 2.4.1).

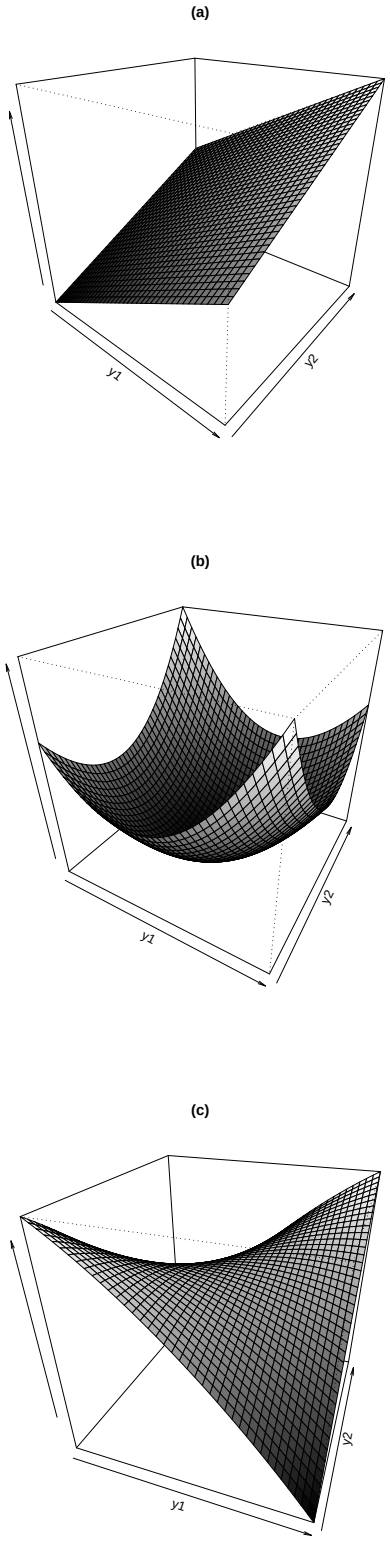

Figure 3.1: Multivariate normal distribution. Plots of the components of the pairwise score function. In panel (a) $ps_{\mu}(\theta; y)$; (b) $ps_{\sigma^2}(\theta; y)$; (c) $ps_{\rho}(\theta; y)$.

In particular, if the pairwise score function is bounded, this implies the B-robustness of the maximum pairwise likelihood estimator. It is straightforward to see that $ps(\theta)$ is not bounded, since it is a function of $\bar{y}$, SS_B and SS_W, that are linear and quadratic functions in the observations, respectively. The unboundedness of the pairwise score function can be seen from Figure 3.1, where the components of $ps(\theta)$ are plotted for $q = 2$ and $y = (y_1, y_2) \in [-3, 3] \times [-3, 3]$. In all the panels we see that when either $|y_1| \to \infty$ or $|y_2| \to \infty$ the components of $ps(\theta)$ are not bounded. This implies that the maximum pairwise likelihood estimator is not robust.

3.5 Final remarks

In this Chapter, some problems related to the use of pairwise likelihood functions for inferential purposes have been discussed.

Interest has been first focused on hypothesis testing, when the pairwise analogue of the log likelihood ratio test statistic is considered. It has been highlighted that the pairwise log likelihood ratio statistic does not converge to a standard limiting distribution, but to a linear combination of independent chi-square random variables with parameters depending on the Godambe information. The need of resorting to empirical expressions of the elements of this matrix turns out to worsen the convergence to the asymptotic distribution of pairwise likelihood test statistics.

Then, the attention has been moved towards a second problem, caused by the lack of robustness of the maximum pairwise likelihood estimator. This problem is particularly critical since it may affects all the inferential procedures based on the pairwise likelihood functions as, for instance, the power and the coverage levels of the derived test statistics.

These two problems are faced in the following chapters where some original solutions are proposed.

Chapter 4

Empirical pairwise log likelihood ratios

4.1 Introduction

The aim of this Chapter is to discuss a solution to cope with confidence regions and testing hypothesis in the pairwise likelihood framework. In Section 3.4.1, it has been shown that pairwise likelihood test statistics require the computation of the elements $J(\theta)$ and $H(\theta)$ of the Godambe information. As pointed out in Section 3.4.1, analytical expressions of these matrices are available in rather simple cases only, and often their empirical counterparts must be computed. These estimates, given in (3.6) and (3.7), may be inaccurate or even non-consistent. This happens, for instance, when dealing with times series or spatial processes. Some alternative estimators of $J(\theta)$ and $H(\theta)$ are available, but they suffer from practical limitations. The lack of accuracy of such estimates affects the convergence of the pairwise log likelihood test statistics to their asymptotic distributions.

These reasons motivate the contribution proposed in this Chapter, where a computationally and theoretically appealing approach is developed based on empirical log likelihood ratios derived from pairwise score functions. The proposed test statistics are attractive in the following situations:

1. the first one is related to the difficulties that might arise when an estimate of $J(\theta)$ is needed. For instance, when asymptotics are in n, the straightforward estimator $\hat{J}(\theta)$ given in (3.6) might be inaccurate when the sample size is moderate to small and this carries over the accuracy of all the pairwise likelihood test statistcs. On the other hand, the improved jackknife estimator $\hat{J}_{jack}(\theta)$ given in (3.11) can be computationally demanding to obtain. Instead, when asymptotics are in q, estimators of $J(\theta)$ based on window subsampling are strictly depending on the mixing conditions of the random field;

2. the second one is concerned with those applications where the computation of the maximum pairwise likelihood estimate requires a huge computational effort. As a matter of fact pairwise empirical likelihood ratios (and in general empirical likelihood ratios) are computed without requiring the knowledge of the maximum pairwise likelihood estimate (see Section 2.6.2).

The behavior of the proposed test statistics is also illustrated through two simulation studies in Section 4.3.

4.2 Pairwise score-based empirical log likelihood ratios

Let $Y \in \mathbb{R}^q$ be a random vector with probability distribution $F(y;\theta)$ and density function $f(y;\theta)$, $\theta \subseteq \mathbb{R}^p$. Consider a random sample $y = (y_1, \ldots, y_n)$ from Y, and define a set of measurable events $\{\mathcal{E}_r : r = 1, \ldots, nq(q-1)/2\}$ in terms of pairs of observations (y_{ih}, y_{ik}), $i = 1, \ldots, n$, $h \neq k = 1 \ldots, q$. In what follows, the pairwise score function given in (3.5) is considered by letting $a = 0$.

The pairwise score function is given by

$$ps(\theta) = ps(\theta; y) = \sum_{i=1}^{n} \sum_{h=1}^{q-1} \sum_{k=h+1}^{q} \frac{\partial \log f(y_{ih}, y_{ik}; \theta)}{\partial \theta} = \sum_{i=1}^{n} ps_i(\theta), \qquad (4.1)$$

with $ps_i(\theta) = \sum_{h=1}^{q-1} \sum_{k=h+1}^{q} \partial \log f(y_{ih}, y_{ik}; \theta)/\partial\theta$.

An empirical pairwise log likelihood ratio statistic for θ, derived from (4.1), can be expressed as

$$pW_e(\theta) = 2 \sum_{i=1}^{n} \log \left\{ 1 + \xi(\theta)^\top ps_i(\theta) \right\} , \qquad (4.2)$$

where the Lagrangian multiplier $\xi(\theta)$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \frac{ps_i(\theta)}{(1 + \xi(\theta)^\top ps_i(\theta))} = 0. \qquad (4.3)$$

The following Proposition states that, starting from the pairwise score function, it is possible to obtain a pseudo log likelihood ratio test with standard limiting distribution. This task is accomplished by deriving the empirical log likelihood ratio from the estimating function (4.1).

Proposition 1. *Consider the pairwise score function* (4.1) *and the pairwise empirical log likelihood ratio* (4.2)*. Then*

$$pW_e(\theta) \xrightarrow{d} \chi^2_p.$$

Proposition 1 states the same asymptotic result of Owen (1988, 1990) for the empirical pairwise log likelihood ratio statistic $pW_e(\theta)$. The proof can be easily sketched by exploiting the general results for unbiased estimating functions. In particular, it is required $ps(\theta) = O_p(n^{1/2})$, $\hat{J}(\theta) = O_p(n)$, and $\hat{H}(\theta) = O_p(n)$.

Proof. A McLaurin series expansion of (4.3) yields

$$\xi(\theta) = \hat{J}(\theta)^{-1} ps(\theta) + O_p(n^{-1}) = O_p(n^{-1/2}).$$

The expansion for $pW_e(\theta)$ is

$$\begin{aligned} pW_e(\theta) &= 2\sum_{i=1}^{n} \log\left\{1 + \xi(\theta)^\top ps_i(\theta)\right\} = \\ &= 2\sum_{i=1}^{n} \left\{\xi(\theta)^\top ps_i(\theta) - \frac{1}{2}\xi(\theta)^\top ps_i(\theta) ps_i(\theta)^\top \xi(\theta) + O_p(n^{-3/2})\right\} = \\ &= 2\left\{\xi(\theta)^\top ps(\theta) - \frac{1}{2}\xi(\theta)^\top \hat{J}(\theta)^{-1}\xi(\theta)\right\} + O_p(n^{-1/2}) = \\ &= ps(\theta)^\top \hat{J}(\theta)^{-1} ps(\theta) + O_p(n^{-1/2}) = \\ &= pW_u(\theta) + O_p(n^{-1/2}). \end{aligned}$$

□

A more rigorous proof can be easily obtained following the theory in Adimari and Guolo (2010). The chi-square approximation still holds when θ is partitioned as $\theta = (\tau, \lambda)$, where τ is a parameter of interest and λ is a nuisance parameter, i.e. for the profile version of $pW_e(\theta)$. In particular, $pW_{ep}(\tau) = \inf_\lambda pW_e(\tau, \lambda)$ still converges in distribution to a chi-square random variable.

The asymptotic behavior of $pW_e(\theta)$ is determined by the relations in (2.27). This highlights the empirical likelihood's ability to Studentize internally and, roughly speaking, this means that, up to first order, the empirical log likelihood ratio resembles Wald or score statistics without explicit estimation of $J(\theta)$, $H(\theta)$ of either $\hat{\theta}_p$.

Although the chi-square approximation for $pW_e(\theta)$ is in error by order $O_p(n^{-1})$ (Hall and La Scala, 1990; DiCiccio *et al.*, 1991), the empirical log likelihood ratio statistic (4.2) may lead to unsatisfactory inferences when the sample size is relative small. Nevertheless, it is possible to derive an empirical likelihood ratio that, in some circumstances, enhances the accuracy of the approximation.

An alternative version of (4.2) can be obtained by rewriting (4.1) as

$$ps(\theta) = p\bar{s}(\theta) = p\bar{s}(\theta; y) = \sum_{r=1}^{m} \frac{\partial \log f(y \in \mathcal{E}_r; \theta)}{\partial \theta} = \sum_{r=1}^{m} ps_r(\theta), \qquad (4.4)$$

where, without loss of generality, $f(y \in \mathcal{E}_1;\theta) = f(y_{11}, y_{12};\theta)$, $f(y \in \mathcal{E}_2;\theta) = f(y_{11}, y_{13};\theta), \ldots, f(y \in \mathcal{E}_m;\theta) = f(y_{n(q-1)}, y_{nq};\theta)$, with $m = nq(q-1)/2$. The main difference between (4.1) and (4.4) is that the latter does not group the score contributions derived from the same unit. In other words, using $\bar{ps}(\theta)$ is like observing a random sample of dimension $nq(q-1)/2$ from a bivariate distribution.

The pairwise empirical log likelihood ratio derived from $\bar{ps}(\theta)$ is

$$\bar{p}W_e(\theta) = 2\sum_{r=1}^{m} \log\left\{1 + \bar{\xi}(\theta)^\top ps_r(\theta)\right\}, \tag{4.5}$$

where the Lagrangian multiplier $\bar{\xi}(\theta)$ satisfies

$$\frac{1}{m}\sum_{r=1}^{m} \frac{ps_r(\theta)}{\left(1 + \bar{\xi}(\theta)^\top ps_r(\theta)\right)} = 0. \tag{4.6}$$

The asymptotic behavior of $\bar{p}W_e(\theta)$ is stated in the following Proposition.

Proposition 2. *Consider the pairwise score function* (4.4) *and and the pairwise empirical log likelihood ratio* (4.5)*. Then*

$$\bar{p}W_e(\theta) \xrightarrow{d} \sum_{j=1}^{p} \lambda_j(\theta) Z_j^2,$$

with $\lambda_j(\theta)$ eigenvalues of $H(\theta)^{-1}J(\theta)$ and Z_j independent standard normal random variables, $j = 1, \ldots, p$.

Proof. Following Cox and Reid (2004), we formally expand $ps_i(\hat{\theta}_p)$ around θ, up to the first order, i.e.

$$ps_i(\theta) - (\hat{\theta}_p - \theta)^\top \frac{\partial ps_i(\theta)}{\partial \theta^\top} \doteq 0.$$

The second term is $O_p(q^2)$, while the order of the first term is $O_p(q^k)$. The constant $k \in [1, 2]$ accommodates for the dependence structure of the data (see Cox and Reid, 2004).

Furthermore, we have

$$\begin{aligned} &\sum_{i=1}^{n} \left\{ps_i(\theta) - (\hat{\theta}_p - \theta)^\top \frac{\partial ps_i(\theta)}{\partial \theta^\top}\right\} = \\ &= ps(\theta) + (\hat{\theta}_p - \theta)^\top \frac{\partial ps(\theta)}{\partial \theta^\top} = \\ &= O_p(n^{1/2}q^k) + O_p(n^{-1/2}q^{k-2})O_p(nq^2). \end{aligned} \tag{4.7}$$

The estimator of $J(\theta)$ supplied by (4.4) is $\tilde{J}(\theta) = m^{-1}\sum_{r=1}^{m} ps_r(\theta)ps_r(\theta)^\top$, whereas that from (4.1) is $n^{-1}\hat{J}(\theta) = \sum_{i=1}^{n} ps_i(\theta)ps_i(\theta)^\top$. Thus,

$$\tilde{J}(\theta) \xrightarrow{p} H(\theta)$$

and

$$\hat{J}(\theta) \xrightarrow{p} J(\theta).$$

A McLaurin series expansion of (4.6) yields

$$\bar{\xi}(\theta) = \tilde{J}(\theta)^{-1}ps(\theta) + O_p(n^{-1}q^{2k-4}) = O_p(n^{-1/2}q^{k-2}),$$

where the order of $\tilde{J}(\theta)$ and $ps(\theta)$ can be derived from (4.7).

The expansion for $\bar{p}W_e(\theta)$ is

$$\begin{aligned}
\bar{p}W_e(\theta) &= 2\sum_{r=1}^{m}\log\left(1+\bar{\xi}(\theta)^\top ps_r(\theta)\right) = \\
&= 2\left(\bar{\xi}(\theta)^\top ps(\theta) - \frac{1}{2}\bar{\xi}(\theta)^\top\tilde{J}(\theta)\bar{\xi}(\theta)\right) + O_p\left(n^{-1/2}q^{3k-4}\right) = \\
&= ps(\theta)^\top\tilde{J}(\theta)^{-1}ps(\theta) + O_p\left(n^{-1/2}q^{3k-4}\right) = \\
&= \left\{\left(\hat{J}(\theta)^{-1/2}\right)^\top ps(\theta)\right\}^\top \hat{J}(\theta)^{1/2}\tilde{J}(\theta)^{-1}\left(\hat{J}(\theta)^{1/2}\right)^\top\left\{\left(\hat{J}(\theta)^{-1/2}\right)^\top ps(\theta)\right\} + \\
&+ O_p\left(n^{-1/2}q^{3k-4}\right)\ , \qquad (4.8)
\end{aligned}$$

where $\hat{J}(\theta)^{1/2}\left(\hat{J}(\theta)^{1/2}\right)^\top = \hat{J}(\theta)$. The following equality

$$\left|\hat{J}(\theta)^{1/2}\tilde{J}(\theta)^{-1}\left(\hat{J}(\theta)^{1/2}\right)^\top\right| = \left|\tilde{J}(\theta)^{-1}\hat{J}(\theta)\right|$$

implies that the eigenvalues of the two matrices are equal. Moreover

$$\left|\tilde{J}(\theta)^{-1}\hat{J}(\theta)\right| \xrightarrow{p} \left|H(\theta)^{-1}J(\theta)\right|.$$

Finally, the result follows since in (4.8) we have a quadratic form in normal random variables. □

Proposition 2 states that the asymptotic distribution of $pW(\theta)$ and $\bar{p}W_e(\theta)$ are the same. Hence, $\bar{p}W_e(\theta)$ should be scaled with the same scaling factors used for $pW(\theta)$ (see Section 3.4.1) in order to obtain the standard chi-square limiting distribution. Although $\bar{p}W_e(\theta)$ must be computed along with the elements of the Godambe information, in some circumstances its usage is

preferable than to $pW_e(\theta)$ and $pW(\theta)$. For instance, consider the pairwise empirical log likelihood ratio adjusted by matching the first moment

$$\bar{p}W_{e1}(\theta) = \frac{\bar{p}W_e(\theta)}{\kappa_1},$$

with $\kappa_1 = \sum_{j=1}^{p} \lambda_j(\theta)/p$. It is easy to show that $\kappa_1 = \mathrm{tr}(H(\theta)^{-1}J(\theta))/p = O_p(q^{2k-2})$ and hence the remainder term of the scaled statistic $\bar{p}W_{e1}(\theta)$ is $O_p\left(n^{-1/2}q^{k-2}\right)$. Thus, it is worth to use $\bar{p}W_{e1}(\theta)$ when the correlation is moderate. Indeed, as the correlation strengthens, hence k moves from 1 to 2, the convergence will be slower. For instance if $k = 1$ and $q = O(n)$, then the remainder term is bounded by $O_p(n^{-3/2})$.

4.3 Numerical examples

In this Section two examples are discussed in order to compare the finite-sample behavior of the inferential procedures based on the test statistics presented in Sections 3.4.1 and 4.2. The first example deals with the equicorrelated multivariate normal distribution and the second one considers correlated binary data. The first example considers a vector parameter and is feasible to do closed form calculations both for complete and pairwise likelihood quantities. The second example provides a framework of practical interest, where the pairwise likelihood function is not in closed form and the estimation of the matrices $H(\theta)$ and $J(\theta)$ is needed and can be computationally intensive.

4.3.1 Multivariate normal distribution

Let us focus on the mean μ, variance σ^2, and on the correlation coefficient ρ of an equicorrelated multivariate normal distribution (see Example 3.1). In this case, the full log likelihood function $\ell(\theta)$, with $\theta = (\mu, \sigma^2, \rho)$, is available and it is possible to compare the full log likelihood ratio statistic $W(\theta)$, based on $\ell(\theta)$, with the scaled versions of $pW\theta)$ presented in Section 3.4.1, and with the proposed pairwise empirical log likelihood ratios $pW_e(\theta)$ and $\bar{p}W_e(\theta)$. The pairwise log likelihood function and the pairwise score function for this model have been given in Section 3.4.2.

In order to assess the behavior of $pW_e(\theta)$, $\bar{p}W_{e1}(\theta)$, $pW_1(\theta)$, $pW_w(\theta)$, $pW_s(\theta)$, $pW_{cb}(\theta)$ and $pW_{inv}(\theta)$, we ran a simulation experiment, with $n = 15$, 30 and $q = 30$, for three values of ρ, ranging from a moderate to a strong correlation. The analysis is restricted to positive values of ρ as in Cox and Reid (2004) and Pace *et al.* (2011). In order to guarantee that Σ is positive definite it must be $-1/(q-1) < \rho < 1$.

Table 4.3.1 gives the empirical coverages of confidence regions. Note that both $pW_1(\theta)$ and $\bar{p}W_{e1}(\theta)$ are multiplied by the same scale factor $1/\hat{\kappa}_1$, with

$q = 30$	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.9$		
$n = 15$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
$W(\theta)$	0.890	0.941	0.986	0.887	0.939	0.986	0.887	0.940	0.987
$pW_1(\theta)$	0.819	0.874	0.942	0.808	0.868	0.940	0.841	0.896	0.961
$pW_w(\theta)$	0.711	0.773	0.856	0.708	0.768	0.848	0.592	0.641	0.712
$pW_s(\theta)$	0.791	0.866	0.967	0.788	0.863	0.964	0.785	0.859	0.964
$pW_{inv}(\theta)$	0.845	0.931	0.996	0.890	0.960	0.995	0.904	0.952	0.990
$pW_{cb}(\theta)$	0.716	0.785	0.873	0.728	0.791	0.879	0.606	0.659	0.737
$pW_e(\theta)$	0.815	0.876	0.937	0.826	0.883	0.941	0.855	0.903	0.951
$\bar{p}W_{e1}(\theta)$	0.894	0.951	0.992	0.880	0.943	0.991	0.798	0.869	0.951
$n = 30$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
$W(\theta)$	0.891	0.943	0.988	0.893	0.943	0.987	0.892	0.862	0.885
$pW_1(\theta)$	0.848	0.899	0.957	0.842	0.896	0.958	0.862	0.913	0.969
$pW_w(\theta)$	0.797	0.857	0.928	0.794	0.853	0.923	0.698	0.749	0.821
$pW_s(\theta)$	0.827	0.886	0.954	0.831	0.886	0.953	0.828	0.888	0.952
$pW_{inv}(\theta)$	0.851	0.915	0.978	0.878	0.936	0.989	0.897	0.947	0.989
$pW_{cb}(\theta)$	0.803	0.867	0.940	0.809	0.872	0.941	0.711	0.767	0.844
$pW_e(\theta)$	0.886	0.930	0.976	0.884	0.935	0.949	0.889	0.934	0.969
$\bar{p}W_{e1}(\theta)$	0.895	0.946	0.989	0.885	0.944	0.989	0.849	0.910	0.971

Table 4.1: Multivariate normal distribution: empirical coverage probabilities of confidence regions for θ based on 20.000 Monte Carlo trials.

$\hat{\kappa}_1 = \mathrm{tr}(\hat{H}(\hat{\theta}_p)^{-1}\hat{J}(\hat{\theta}_p))/p$ and the results show that the proposed pairwise empirical log likelihood statistic $\bar{p}W_{e1}(\theta)$ has a reasonably performance in terms of coverage and is close to $W(\theta)$, $pW_s(\theta)$ and $pW_{inv}(\theta)$ when the correlation is less than 0.9. For $n = 15$ and $n = 30$, $\bar{p}W_{e1}(\theta)$ outperforms $pW_e(\theta)$, $pW_w(\theta)$ and $pW_{cb}(\theta)$. On the other hand, $pW_e(\theta)$ performs as well as $pW_1(\theta)$ when $n = 15$, but for $n = 30$ the empirical coverages are closer to the nominal levels than those of the scaled versions of $pW(\theta)$. Larger sample sizes (results not reported here) give, as one would expect, rather little differences between the results of all the test statistics.

4.3.2 Binary data

The pairwise likelihood is particularly useful for modeling correlated binary outcomes, as discussed in Le Cessie and Van Houwelingen (1994). This kind

of data arises, *e.g.*, in the context of repeated measurements on the same subject, where a standard likelihood analysis involves multivariate integrals whose dimension equals the cluster sizes.

Let us focus on a multivariate probit model with constant cluster sizes. In this case, the pairwise log likelihood is

$$p\ell(\theta) = \sum_{i=1}^{n} \sum_{h=1}^{q-1} \sum_{k=h+1}^{q} \log P(Y_{ih} = y_{ih}, Y_{ik} = y_{ik}; \theta) \tag{4.9}$$

(see Le Cessie and Van Houwelingen, 1994; Kuk and Nott, 2000). Pairwise likelihood inference is much simpler than using the full likelihood since it involves only bivariate normal integrals. For instance (see also Renard *et al.*, 2004), we have $P(Y_{ih} = 1, Y_{ik} = 1; \theta) = \Phi_2(\gamma_{ih}, \gamma_{ik}; \rho)$, where $\Phi_2(\cdot, \cdot; \rho)$ denotes the standard bivariate normal distribution function with correlation coefficient ρ and $\gamma_{ih} = x_{ih}\beta/\sigma$ is the component of place h of $\gamma_i = X_i\beta/\sigma^2$, with β unknown p-dimensional regression coefficient, σ known scale parameter and X_i design matrix for unit i with ones in the first column, $i = 1, \ldots, n$, $h, k = 1, \ldots, q$.

In our simulation setting, we have $\beta = (\beta_0, \beta_1)$ and the covariate in X_i, $i = 1, \ldots, n$, is generated considering q independent trials from a uniform random variable on the interval $[-1, 1]$. The binary outcomes for unit i are obtained simulating from a q-variate normal Z having vector of means γ_i, covariance matrix Σ with $\Sigma_{hh} = \sigma^2$, $\Sigma_{hk} = \sigma^2\rho$, $h \neq k$, and then dichotomizing the result according to $Y_{ih} = 1$ if $Z_{ih} \geq 0$, as described in Section 2 of Renard *et al.* (2004).

Simulation results for the overall parameter $\theta = (\beta_0, \beta_1, \rho)$ are summarized in Table 4.3.2, which gives the empirical coverages for confidence regions for θ. The derivatives of (4.9) are not available in closed form, and their numerical evaluation has been carried out using the `R` library `numDeriv`, while the maximization step has been performed using functions in standard `R` libraries. The results in Table 4.3.2 show that the pairwise empirical log likelihood statistic $\bar{p}W_{e1}(\theta)$ gives quite good results for moderate sample sizes, but in this example the statistic $pW_e(\theta)$ slightly improves on all the statistics. This example highlights that the use of the statistic $pW_e(\theta)$ might be preferable and of practical interest than the scaled versions of $pW(\theta)$ when the matrices $H(\theta)$ and $J(\theta)$ must be computed numerically.

$q = 20$	$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.75$		
$n = 50$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
$pW_1(\theta)$	0.872	0.919	0.969	0.869	0.915	0.966	0.867	0.915	0.967
$pW_w(\theta)$	0.845	0.903	0.963	0.854	0.913	0.970	0.866	0.920	0.973
$pW_s(\theta)$	0.862	0.915	0.970	0.869	0.921	0.974	0.876	0.929	0.978
$pW_{cb}(\theta)$	0.844	0.902	0.963	0.854	0.912	0.969	0.864	0.921	0.974
$pW_{inv}(\theta)$	0.869	0.924	0.977	0.878	0.933	0.983	0.888	0.942	0.987
$pW_e(\theta)$	0.875	0.927	0.977	0.886	0.936	0.981	0.892	0.941	0.983
$\bar{p}W_{e1}(\theta)$	0.880	0.926	0.974	0.876	0.919	0.970	0.872	0.920	0.970
$n = 100$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
$pW_1(\theta)$	0.878	0.924	0.974	0.875	0.920	0.968	0.869	0.918	0.970
$pW_w(\theta)$	0.872	0.927	0.979	0.878	0.931	0.981	0.879	0.935	0.983
$pW_s(\theta)$	0.881	0.931	0.980	0.883	0.934	0.980	0.887	0.938	0.984
$pW_{cb}(\theta)$	0.872	0.926	0.979	0.876	0.931	0.980	0.880	0.935	0.983
$pW_{inv}(\theta)$	0.884	0.937	0.984	0.887	0.939	0.985	0.893	0.943	0.989
$pW_e(\theta)$	0.894	0.946	0.987	0.894	0.944	0.987	0.893	0.946	0.988
$\bar{p}W_{e1}(\theta)$	0.882	0.928	0.976	0.878	0.922	0.970	0.872	0.920	0.971

Table 4.2: Binary data: empirical coverage probabilities of confidence regions based on 20.000 Monte Carlo trials, with $\beta_0 = 1/2$ and $\beta_1 = 1$.

4.4 Final remarks

In this Chapter, the possibility of deriving empirical likelihoods from a pairwise score function has been investigated. The simulation results in Section 4.3 indicate that the proposed $\bar{p}W_{e1}(\theta)$ and $pW_e(\theta)$ can be useful to make inference in complex models, and they offer a new appealing computational method to derive likelihood ratio-type test statistics in this framework.

The pairwise empirical log likelihood ratio $pW_e(\theta)$ provides several advantages over $pW(\theta)$ and its scaled versions, in some respects. First, the computation of the elements of the Godambe information is avoided, and in particular that of $J(\theta)$ that can be troublesome (see Section 3.4.1). In the equicorrelated multivariate normal example $J(\theta)$ is available, but we used $\hat{J}(\theta)$ in order to perform a fair and more realistic comparison of $pW_e(\theta)$ with the scaled versions of $pW(\theta)$ and $\bar{p}W_e(\theta)$. Second, to perform hypothesis testing the statistic $pW_e(\theta)$ does not require the knowledge of the pairwise

maximum likelihood estimate $\hat{\theta}_p$ that can be computationally demanding to obtain (such as in the context of max-stable processes; see, *e.g.*, Padoan *et al.*, 2010).

For what concerns the statistic $\bar{p}W_{e1}(\theta)$, it does not overcome the estimation of $J(\theta)$ but provides coverages that are even better than those of the scaled versions of $pW(\theta)$.

In general, to compute the proposed empirical pairwise likelihoods we have to solve equations (4.3) and (4.6) and the computational demand, to find the roots $\xi(\theta)$ and $\bar{\xi}(\theta)$, is negligible compared to that for $pW(\theta)$, since we need to compute only once the pairwise score function at θ. Furthermore, the algorithm described in Owen (1990) reformulates the problem of solving (4.3) and (4.6) into a minimization problem, providing fast and reliable roots for these equations.

As a final remark, we note that the proposed pairwise empirical likelihoods may be readily extended to general composite score functions, providing inferential tools alternative to composite likelihood functions.

Chapter 5

Saddlepoint test based on the maximum pairwise likelihood estimator

5.1 Introduction

In the previous Chapter, empirical log likelihood ratio statistics based on the pairwise score function were developed and discussed. The proposals aim at providing test statistics that may prove useful in the composite likelihood framework. In particular, it has been shown that the test statistic $pW_e(\theta)$ behaves as it was a parametric log likelihood ratio, i.e. its limiting distribution is standard chi-square without any additional estimation of the elements of the Godambe information. This feature is particularly appealing and provides a notable improvement with respect to composite likelihood-based test statistics in those situations where the computation of the elements of the Godambe information is somehow troublesome (see Section 3.4.1).

The nominal levels of confidence regions based on $pW_e(\theta)$ are asymptotically correct, and in finite samples the empirical levels may be far away from the nominal ones, especially when the sample size n is small. To cope with the possible lack of accuracy of $pW_e(\theta)$, an alternative pairwise empirical log likelihood ratio, namely $\bar{p}W_e(\theta)$, is proposed. It has been shown that the convergence of $\bar{p}W_e(\theta)$ to its asymptotic distribution depends on the sample size n, the dimensionality of the data q as well as on the dependence among the observations. Hence, in some situations the use of $\bar{p}W_e(\theta)$ may be preferable than that of both $pW_e(\theta)$ and composite likelihood-based test statistics. However, the asymptotic distribution of $\bar{p}W_e(\theta)$ is no longer standard chi-square and it turns out to be the same asymptotic distribution of the composite log likelihood ratio (3.4). Hence, inference based on $\bar{p}W_e(\theta)$ requires the computation of the elements of the Godambe information.

In this Chapter, the use of a test statistic based on the saddlepoint ap-

proximation to the density of M-estimators is discussed in the pairwise likelihood framework. There are at least three notable features provided by the use of the nonparametric saddlepoint test statistic in the pairwise likelihood framework:

1. despite saddlepoint approximations were originally proposed in a fully parametric setting, the proposed test statistic is based on nonparametric saddlepoint approximations. This is very relevant in the pairwise likelihood framework since the specification of the joint distribution of the data is avoided and inference is based on an approximate model;

2. the proposed test statistic has a standard asymptotic behavior and the error in the approximation is of order $O_p(n^{-1})$. In particular, the error is relative rather than absolute, meaning that the approximation provided is very accurate also in the tails of the distribution. This is an important property of the proposed test statistic since the main alternatives, such as bootstrapped test statistics, claim absolute errors;

3. the proposed test statistic possesses the aforementioned nice properties in a general setting: the accuracy of the approximation and the standard asymptotic behavior hold under the assumptions required for composite likelihood methods.

The behavior of the proposed nonparametric saddlepoint test statistic is illustrated through two simulation studies in Section 5.5.

5.2 Background on saddlepoint approximations

Edgeworth expansions are used to approximate the density and distribution functions of standardized sums of random variables. Several statistics can be expressed in such a form, and a notable example of the use of the Edgeworth formula is to approximate the density of maximum likelihood estimators. Although Edgeworth expansions have practical disadvantages compared to saddlepoint approximations, they play a central role in theoretical discussions of small-sample inference. Edgeworth approximations are not pursued in the remaining part of the Chapter and the reader can refer to Hall (1997).

Saddlepoint approximations in statistic date back to Daniels (1954). Broadly speaking, they are an improvement over Edgeworth expansions that give rise to highly accurate density estimates. Saddlepoint approximations control the relative error of the approximation rather than the absolute one (as Edgeworth approximations do), thereby they are usually very accurate when the sample size is small. Therefore, their use is relevant when accurate estimates of tail area probabilities are required.

In order to sharpen the scope of this Section, the saddlepoint approximation for the density of multivariate M-estimators is presented. The main

references for saddlepoint approximations for the density and distribution function of sums of independent random variables are Daniels (1954) and Lugannani and Rice (1980), whereas a general reference is Butler (2007).

Let $\mathcal{F} = \{f(y;\theta);\, y \in \mathcal{Y} \subseteq \mathbb{R}^q,\, \theta \in \Theta \subseteq \mathbb{R}^p,\, q, p \geq 1\}$ be a parametric statistical model for the random vector Y, and let $F_\theta = F(y;\theta)$ be the distribution function associated to $f(y;\theta)$. Consider a random sample $y = (y_1, \ldots, y_n)$ of size n from F_θ. An M-estimator $\tilde{\theta}$ is defined implicitly as the solution of the estimating equation

$$\Psi_\theta = \Psi(\theta; y) = \sum_{i=1}^{n} \psi(\theta; y_i) = 0,$$

with $\psi(\cdot)$ known function (see Section 2.2).

In this context, attention is restricted to situations where the cumulant generating function of Ψ_θ, defined as

$$K_\Psi(\lambda(\theta); \theta) = \log \mathbb{E}\left(\exp\left\{\lambda(\theta)^\top \psi(\theta; Y)\right\}\right),$$

with $\lambda(\theta) \in \mathbb{R}^p$, exists. Under this assumption, the saddlepoint approximation to the density of $\tilde{\theta}$ is given by (Field, 1982)

$$f_{\tilde{\theta}}(t) = \left(\frac{2\pi}{n}\right)^{p/2} \exp\left\{n K_\Psi(\lambda(t); t)\right\} |B(t)||\Sigma(t)|^{-1/2}(1 + O_p(n^{-1})), \quad (5.1)$$

where the saddlepoint $\lambda(t) = \lambda$ satisfies the saddlepoint equation

$$\frac{\partial}{\partial \lambda} K_\Psi(\lambda(t); t) = 0.$$

Further, in (5.1)

$$B(t) = \exp\left\{-K_\Psi(\lambda(t); t)\right\} \mathbb{E}\left(\exp\left\{\lambda(t)^\top \psi(t; Y)\right\} \frac{\partial}{\partial t^\top} \psi(t; Y)\right)$$

and

$$\Sigma(t) = \exp\left\{-K_\Psi(\lambda; t)\right\} \mathbb{E}\left(\psi(t; Y)\psi(t; Y)^\top \exp\left\{\lambda(t)^\top \psi(t; Y)\right\}\right).$$

The saddlepoint approximation (5.1) was given in Field (1982), and has been subsequently considered by Skovgaard (1990), Jensen and Wood (1998) and Almudevar *et al.* (2000). Conditions which imply the existence of $f_{\tilde{\theta}}(t)$, and which cover cases with $\psi(\cdot)$ not differentiable, are given in Almudevar *et al.* (2000).

The saddlepoint approximation (5.1) can be extended to the situation in which the true underlying distribution function F_θ is replaced by its empirical counterpart, i.e. $\hat{F}_n$. More precisely, Ronchetti and Welsh (1994) define

the empirical saddlepoint approximation for the density of multivariate M-estimators as

$$\tilde{f}_{\tilde{\theta}}(t) = \left(\frac{2\pi}{n}\right)^{p/2} \exp\left\{n\hat{K}_{\Psi}(\hat{\lambda}(t);t)\right\} |\hat{B}(t)||\hat{\Sigma}(t)|^{-1/2} \tag{5.2}$$

where

$$\hat{K}_{\Psi}(\hat{\lambda}(t);t) = \log\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}(t)^{\top}\psi(t;y_i)\right),$$

$$\hat{B}(t) = \exp\left\{-\hat{K}_{\Psi}(\hat{\lambda}(t);t)\right\} \frac{1}{n}\sum_{i=1}^{n} \exp\left\{\hat{\lambda}(t)^{\top}\psi(t;y_i)\right\} \frac{\partial}{\partial t^{\top}}\psi(t;y_i),$$

$$\hat{\Sigma}(t) = \exp\left\{-\hat{K}_{\Psi}(\hat{\lambda}(t);t)\right\} \frac{1}{n}\sum_{i=1}^{n} \psi(t;y_i)\psi(t;y_i)^{\top} \exp\left\{\hat{\lambda}(t)^{\top}\psi(t;y_i)\right\},$$

and the saddlepoint $\hat{\lambda}(t) = \hat{\lambda}$ solves the equation

$$\frac{\partial}{\partial\lambda}\hat{K}_{\Psi}(\lambda(t);t) = 0.$$

To evaluate the error of (5.2), it is necessary to consider the density of $n^{1/2}(\tilde{\theta} - \theta)$. It may be shown that as n diverges, the following result holds

$$f_{\tilde{\theta}}(\theta + n^{-1/2}u) = \tilde{f}_{\tilde{\theta}}(\tilde{\theta} + n^{-1/2}u)\left\{1 + \frac{a(u)}{\sqrt{n}} + O_p(n^{-1})\right\},$$

where $a(u) = O_p(1)$ and its expression, as well as u, are given in Ronchetti and Welsh (1994).

5.3 Saddlepoint test based on multivariate M-estimators

Consider the null hypothesis $H_0 : \theta = \theta_0$. The parametric saddlepoint test is (Robinson *et al.*, 2003)

$$h(\tilde{\theta}) = -2nK_{\Psi}(\lambda(\tilde{\theta});\theta_0) = -2n\log\mathbb{E}_{\theta_0}\left(\exp\left\{\lambda(\tilde{\theta})^{\top}\psi(\tilde{\theta};Y)\right\}\right), \tag{5.3}$$

where E_{θ_0} is the expected value with respect to F_{θ_0}, and its asymptotic null distribution is chi-square with p degrees of freedom. In particular, under the assumptions given in Robinson *et al.* (2003), $h(\tilde{\theta})$ is asymptotically pivotal, and the following result holds

$$p = P_{\theta_0}\left[h(\tilde{\theta}(Y)) \geq h(\tilde{\theta}(y))\right] = P_{\theta_0}\left[\chi^2_p \geq h(\tilde{\theta}(y))\right]\left\{1 + O_p(n^{-1})\right\}, \tag{5.4}$$

where $P_{\theta_0}(\cdot)$ denotes the probability under H_0, $\tilde{\theta}(Y)$ is the M-estimator, $\tilde{\theta}(y)$ is the M-estimate, and χ^2_p is a chi-square random variable with p degrees of

freedom. Relation (5.4) states that the error in the approximation is relative and of second order. It is worth to note that if the underlying distribution of the observations belongs to a full exponential family with score function $\psi(\theta; y) = y - \theta$, where θ is the mean parameter, then the statistic (5.3) is the log likelihood ratio.

In practice, the distribution F_θ underlying the data may be unknown. The conventional approach to obtain an estimate of F_θ, which puts mass on the observed data points, is to follow Owen (2001), which results in maximizing the empirical likelihood under suitable constraints. As it has been shown in Section 2.6.2, this approach is equivalent to minimize the constrained forward Kullback-Leibler divergence between $\hat{F}_n$ and

$$\hat{F}_\theta = \left\{ w_i(\theta) : \sum_i w_i(\theta) = 1, \sum_i w_i(\theta)\psi(\theta; y_i) = 0 \right\}.$$

The idea is to apply the parametric saddlepoint test (5.3) by replacing F_{θ_0}, i.e. the distribution under H_0, by a suitable nonparametric estimate $\hat{F}_{\theta_0}$, in order to retain the second-order property (5.4). This task is accomplished by minimizing rather than the constrained forward Kullback-Leibler divergence the backward's one, that is

$$\begin{aligned} &\sum_{i=1}^{n} w_i(\theta_0) \log\left(\frac{w_i(\theta_0)}{n}\right) + \\ &+\delta(\theta_0)\left(\sum_{i=1}^{n} w_i(\theta_0) - 1\right) + \beta(\theta_0)^\top \sum_{i=1}^{n} \psi(\theta_0; y_i) \end{aligned} \tag{5.5}$$

where $\delta(\theta_0) \in \mathbb{R}$ and $\beta(\theta_0) \in \mathbb{R}^p$ are Lagrange multipliers. The minimization in $w_i(\theta_0)$ of (5.5) is equivalent to the unconstrained minimization of

$$\log\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left\{\beta(\theta_0)^\top \psi(\theta_0; y_i)\right\}\right)$$

in $\beta(\theta_0)$. It may be shown that the elements of $\hat{F}_{\theta_0}$ have the following exponential analytical form (Robinson *et al.*, 2003)

$$w_i(\theta_0) = \frac{\exp\left\{\beta(\theta_0)^\top \psi(\theta_0; y_i)\right\}}{\sum_{i=1}^{n} \exp\left\{\beta(\theta_0)^\top \psi(\theta_0; y_i)\right\}}, \; i = 1, \ldots, n, \tag{5.6}$$

and it turns out that $\hat{F}_{\theta_0}$ is the nonparametric tilted distribution used by Efron (1981) in the bootstrap framework.

Following the basic idea of the saddlepoint test, and using $\hat{F}_{\theta_0}$ as the

underlying distribution, a nonparametric version of (5.3) is given by

$$\begin{aligned}\hat{h}(\tilde{\theta}) &= -2n \log \hat{\mathbb{E}}_{\theta_0}\left[\exp\left\{\lambda(\tilde{\theta})^\top \psi(\tilde{\theta}; Y)\right\}\right] = \\ &= -2n \log\left[\sum_{i=1}^{n} w_i(\theta_0) \exp\left\{\lambda(\tilde{\theta})^\top \psi(\tilde{\theta}; y_i)\right\}\right] = \\ &= -2n \log\left[\frac{\sum_{i=1}^{n} \exp\left\{\beta(\theta_0)^\top \psi(\theta_0; y_i) + \lambda(\tilde{\theta})^\top \psi(\tilde{\theta}; y_i)\right\}}{\sum_{i=1}^{n} \exp\left\{\beta(\theta_0)^\top \psi(\theta_0; y_i)\right\}}\right],\end{aligned} \tag{5.7}$$

where $\hat{\mathbb{E}}_{\theta_0}(\cdot)$ denotes the expected value with respect to $\hat{F}_{\theta_0}$. The statistic $\hat{h}(\tilde{\theta})$ retains the desired second-order property. This means that, even when the sample size n is small, the distribution of (5.7) is very close to a χ^2_p since the error in the approximation is relative and not absolute. Indeed, the following result holds (Ma and Ronchetti, 2011)

$$\begin{aligned}p &= P_{\theta_0}\left[\hat{h}(\tilde{\theta}(Y)) \geq \hat{h}(\tilde{\theta}(y))\right] = \\ &= P\left[\chi^2_p \geq \hat{h}(\tilde{\theta}(y))\right]\left\{1 + O(n^{-1})\right\}.\end{aligned} \tag{5.8}$$

It is worth to outline some points about the relation in (5.8):

1. the derivation of (5.8) stems from the bootstrap, since $\hat{F}_{\theta_0}$ is the nonparametric tilted distribution under H_0. In particular, the result is obtained: *i*) by using some results about second-order relative errors on the accuracy of bootstrap tests, and *ii*) by linking the bootstrap p-value to the p-value given in (5.8). An heuristic discussion is given in the following.

   Let y^* be a sample drawn from $\hat{F}_{\theta_0}$, and let $\tilde{\theta}^*$ denote the solution of $\Psi(\theta; y^*) = 0$. In the bootstrap framework, when the sample space is $y = (y_1, \ldots, y_n)$, the true distribution is $\hat{F}_{\theta_0}$. Under mild regularity conditions (see Field *et al.*, 2008; Ma and Ronchetti, 2011), there exists a saddlepoint approximation to the distribution of $\tilde{\theta}^*(Y^*)$. Therefore, for fixed $\hat{h}(\tilde{\theta}(y))$, the bootstrap p-value satisfies

$$\begin{aligned}p^* &= P^*_{\hat{F}_{\theta_0}}\left[\hat{h}(\tilde{\theta}^*(Y^*)) \geq \hat{h}(\tilde{\theta}(y))\right] = \\ &= P\left[\chi^2_p \geq \hat{h}(\tilde{\theta}(y))\right]\left\{1 + O(n^{-1})\right\},\end{aligned} \tag{5.9}$$

   where $P^*_{\hat{F}_{\theta_0}}(\cdot)$ denotes the probability under the bootstrap distribution. Hence, (5.9) shows that the distribution of the statistic $\hat{h}(\tilde{\theta}^*(Y^*))$ is second order accurate, in relative terms. Moreover, following Field *et al.* (2008) it is possible to link the bootstrap p-value to the p-value given in (5.8), i.e.

$$
\begin{aligned}
p^* &= P^*_{\hat{F}_{\theta_0}}\left[\hat{h}(\tilde{\theta}^*(Y^*)) \geq \hat{h}(\tilde{\theta}(y))\right] = \\
&= P_{F_{\theta_0}}\left[\hat{h}(\tilde{\theta}(Y)) \geq \hat{h}(\tilde{\theta}(y))\right]\left\{1 + O(n^{-1})\right\} = \\
&= p(1 + O(n^{-1})).
\end{aligned} \tag{5.10}
$$

2. the result in (5.8) holds under mild conditions given in Field *et al.* (2008). Here, it is stressed that the agreement in (5.10) between p^* and p holds if the estimating function Ψ_θ is bounded. The boundedness of Ψ_θ is not required for robustness purposes, but, loosely speaking, to retain the second-order property (5.7) without resorting to the bootstrap. The importance of this assumption will be outlined in Section 5.5.2 through a simulation study.

In the following, it is sketched the computation of the bootstrap distribution of $\hat{h}(\tilde{\theta}(Y))$, although relation (5.10) shows that it is possible to achieve the same level of accuracy by using the asymptotic distribution of $\hat{h}(\tilde{\theta}(Y))$.

Given the observed sample, compute $\tilde{\theta} = \tilde{\theta}(y)$, $w_i(\theta_0)$, $i = 1, \ldots, n$, and $\hat{h}(\tilde{\theta})$. Then, draw B samples from $\hat{F}_{\theta_0}$, denoted by $y^*_b = (y^*_{1b}, \ldots, y^*_{nb})$, for $b = 1, \ldots, B$. For each sample compute $\tilde{\theta}^*_b$ by solving the equation $\Psi(\theta; y^*_b) = 0$, and compute

$$
\hat{h}(\tilde{\theta}^*_b) = -2n \log\left[\sum_{i=1}^{n} w_i(\theta_0) \exp\left\{\lambda(\tilde{\theta}^*_b)^\top \psi(\tilde{\theta}^*_b; y_i)\right\}\right].
$$

The bootstrap distribution of $\hat{h}(\tilde{\theta}(Y))$ is given by the elements $\hat{h}(\tilde{\theta}^*_b)$. Therefore the bootstrap p-value can be computed as

$$
p^* = \frac{1}{B+1} \sum_{b=1}^{B} I\left\{\hat{h}(\tilde{\theta}^*_b) \geq \hat{h}(\tilde{\theta})\right\},
$$

where $I(\cdot)$ is the indicator function.

5.4 Nonparametric saddlepoint test based on the maximum pairwise likelihood estimator

The nonparametric saddlepoint statistic (5.7) is derived for general multivariate M-estimators. In this Section, we outline the possibility to apply it in the composite likelihood framework. Indeed, it has been shown that maximum composite likelihood estimators belong to the general class of M-estimators (see Section 3.2).

In the following, we give the formulation of $\hat{h}(\tilde{\theta})$ for the pairwise likelihood setting, since the numerical examples in Section 5.5 are devoted to the pairwise likelihood function.

Let $\hat{\theta}_p$ be the maximum pairwise likelihood estimator, defined as the solution of the pairwise score equation

$$ps(\theta) = \sum_{i=1}^{n} ps(\theta; y_i) = \sum_{i=1}^{n} \left\{ \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} \frac{\partial \log f(y_{ir}, y_{is}; \theta)}{\partial \theta} \right\} = 0\,.$$

Assuming that the density of $\hat{\theta}_p$ admits the saddlepoint approximation (5.1), the nonparametric saddlepoint pairwise test statistic is defined as

$$\begin{aligned} \hat{h}(\hat{\theta}_p) &= -2n \log \left[\frac{\sum_{i=1}^{n} \exp\left\{ \beta(\theta_0)^\top ps(\theta_0; y_i) + \lambda(\hat{\theta}_p)^\top ps(\hat{\theta}_p; y_i) \right\}}{\sum_{i=1}^{n} \exp\left\{ \beta(\theta_0)^\top ps(\theta_0; y_i) \right\}} \right] \\ &= -2n \log \left[\sum_{i=1}^{n} w_i(\theta_0) \exp\left\{ \lambda(\hat{\theta}_p)^\top ps(\hat{\theta}_p; y_i) \right\} \right], \end{aligned} \tag{5.11}$$

where

$$w_i(\theta_0) = \frac{\exp\left\{ \beta(\theta_0)^\top ps(\theta_0; y_i) \right\}}{\sum_{i=1}^{n} \exp\left\{ \beta(\theta_0)^\top ps(\theta_0; y_i) \right\}}, \quad i = 1, \ldots, n.$$

It is worth to outline the advantages provided by the use of (5.7) in the pairwise likelihood framework:

1. the nonparametric saddlepoint test statistic does not require the specification of the underlying distribution F_θ, since it is based on a suitable nonparametric estimate $\hat{F}_\theta$, given in (5.6). Hence, it is possible to make accurate inference regardless the specification of the joint distribution;

2. the nonparametric saddlepoint test statistic is asymptotically chi-square distributed up to a relative error of order $O_p(n^{-1})$, without any additional estimation of the elements of the Godambe information. Indeed, as it is outlined in Section 3.4.1, the lack of accuracy of composite likelihood-based tests might be ascribed to an inaccurate estimate of $J(\theta)$ and/or $H(\theta)$.

5.5 Numerical examples

This section aims at showing some numerical evidence about the behavior of the pairwise nonparametric saddlepoint test statistic (5.11). Two examples are illustrated, each of them highlighting a different feature of $\hat{h}(\hat{\theta}_p)$.

In the first example (5.11) is compared to the pairwise likelihood-based test statistics presented in Section 3.4.1. In particular, the finite sample accuracy of the χ^2 approximation is analized in the context of a multivariate normal model (see Section 3.4.2 and Section 4.3.1).

In the second example, a first order autoregressive model is considered in order to apply $\hat{h}(\hat{\theta}_p)$ in a context of robust regression based on the pairwise

likelihood function. It is shown that the use of bounded estimating functions in order to compute $\hat{h}(\hat{\theta}_p)$ is recommended not merely for robustness purposes. Indeed, the boundedness of the estimating function is required to have a second order agreement, in relative terms, between the distributions of $\hat{h}(\hat{\theta}_p^*)$ and $\hat{h}(\hat{\theta}_p)$ (Ma and Ronchetti, 2011).

In both examples the full log likelihood function $\ell(\theta)$ is available and this allows to set the log likelihood ratio test $W(\theta)$ as a benchmark.

5.5.1 Multivariate normal distribution

Let Y be a normally distributed random vector, with q-dimensional vector of means $(\mu, \ldots, \mu)^\top$ and symmetric covariance matrix Σ, having diagonal elements σ^2 and off-diagonal elements $\sigma^2\rho$, with $\rho \in (-1/(q-1), 1)$. The pairwise log likelihood for $\theta = (\mu, \sigma^2, \rho)$ and the pairwise score function have been given in Example 3.1 of Section 3.4.2.

In order to compare the accuracy of confidence regions based on $\hat{h}(\hat{\theta}_p)$ with the accuracy of the confidence regions based on composite likelihood test statistics, a simulation study has been performed by generating 100.000 samples of size $n = 10$ from $Y \in \mathbb{R}^{30}$, with $\mu = 0$, $\sigma^2 = 1$, and ρ ranging from a moderate to a strong correlation. For each sample, the nonparametric saddlepoint test statistic, as well as the statistics given in Section 3.4.1, have been computed. Both the observed and the expected elements of the Godambe information have been used to compute the pairwise likelihood test statistics, being the latter available in this example (Pace *et al.*, 2011). The statistics denoted with the superscript "e" are computed using the elements of the expected Godambe information.

Table 5.1 reports the empirical coverage probabilities of confidence regions for θ. As expected, the best results are obtained when the elements of the expected Godambe information have an analytical expression and in particular when $pW_u^e(\theta)$ and $pW_{inv}^e(\theta)$ are used. However, when $\hat{J}(\theta)$ and $\hat{H}(\theta)$ are used, the pairwise likelihood test statistics have coverage probabilities far from the nominal levels. Instead, the bootstrap distribution of the nonparametric saddlepoint test statistic $\hat{h}(\hat{\theta}_p^*)$ approximates quite well the χ^2_3, and the approximation is close to that provided by the gold standard $W(\theta)$. From other simulation studies not reported here, it emerges that pairwise likelihood test statistics achieve the nominal levels when either the sample size is increased or when resampling-based estimates of $J(\theta)$ and $H(\theta)$ are computed.

5.5.2 Robust first order autoregression

Consider a stationary first order autoregressive model, of the form

$$y_j = \phi_0 + \phi_1 y_{j-1} + \epsilon_j, \quad j = 2, \ldots, q, \tag{5.12}$$

	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.9$		
$1-\alpha$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
$W(\theta)$	0.8802	0.9375	0.9858	0.8795	0.9367	0.9858	0.8800	0.9365	0.9859
$\hat{h}(\hat{\theta}_p^*)$	0.8644	0.9282	0.9820	0.8722	0.9300	0.9833	0.8650	0.9254	0.9809
$pW_w(\theta)$	0.5215	0.5855	0.6842	0.3273	0.3733	0.4567	0.1280	0.1466	0.1815
$pW_u(\theta)$	0.7733	0.8826	1.0000	0.7727	0.8826	1.0000	0.7747	0.8826	1.0000
$pW_1(\theta)$	0.7847	0.8442	0.9194	0.7505	0.8179	0.9058	0.7540	0.7823	0.8197
$pW_{cb}(\theta)$	0.5570	0.6250	0.7286	0.4201	0.4829	0.5906	0.1689	0.1991	0.2581
$pW_{inv}(\theta)$	0.7955	0.8950	0.9786	0.7980	0.8791	0.9516	0.9122	0.9462	0.9758
$pW_w^e(\theta)$	0.7618	0.8155	0.8840	0.7286	0.7853	0.8601	0.5758	0.6194	0.6865
$pW_u^e(\theta)$	0.9051	0.9443	0.9805	0.9038	0.9435	0.9807	0.9040	0.9433	0.9807
$pW_1^e(\theta)$	0.8133	0.8673	0.9336	0.8136	0.8692	0.9361	0.8407	0.8983	0.9613
$pW_{cb}^e(\theta)$	0.7885	0.8459	0.9126	0.7858	0.8463	0.9190	0.6296	0.6836	0.7610
$pW_{inv}^e(\theta)$	0.9080	0.9528	0.9883	0.8940	0.9477	0.9889	0.8699	0.9276	0.9802

Table 5.1: Multivariate normal model: empirical coverage probabilities of confidence regions for θ based on 100.000 Monte Carlo trials.

with $\phi_0 \in \mathbb{R}$, $\phi_1 \in (-1, 1)$, and ϵ_j independent normal random variables with mean 0 and variance $\sigma^2 > 0$. Under these assumptions, the process can be described by a q-variate normal random variable Y, with vector of means $(\phi_0/(1-\phi_1), \ldots, \phi_0/(1-\phi_1))^\top \in \mathbb{R}^q$ and covariance matrix Σ, having generic element $\Sigma_{jk} = \sigma^2\phi_1^{|j-k|}/(1-\phi_1^2)$, $j, k = 1, \ldots, q$.

Instead of considering bivariate marginal distributions for pairs of contiguous observations (Pace *et al.*, 2011), the pairwise log likelihood function for $\theta = (\phi_0, \phi_1, \sigma^2)$ is derived by considering the univariate conditional distributions $Y_j|Y_{j-1} = y_{j-1} \sim N(\phi_0 + \phi_1 y_{j-1}, \sigma^2)$, and it has the following expression

$$p\ell(\theta) = -\frac{(q-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{r=2}^{q}(y_r - \phi_0 - \phi_1 y_{r-1})^2\,. \tag{5.13}$$

The resulting pairwise score function leads to the ordinary least squares estimate of θ, that can be easily robustified using a Mallows'-type estimate for ϕ_0 and ϕ_1 and the Huber's Proposal 2 for σ. This is accomplished by solving the system of estimating equations

$$\begin{aligned}
&\sum_{j=2}^{q}\psi_{k_1}(r_j) = 0\\
&\sum_{j=2}^{q}\psi_{k_2}(r_j)\psi_b(y_{j-1}) = 0\\
&\sum_{j=2}^{q}\psi_{k_3}(r_j)^2 - (q-1)\beta(k_3) = 0,
\end{aligned} \tag{5.14}$$

where $r_j = (y_j - \phi_0 - \phi_1 y_{j-1})/\sigma$, $\psi_k(r) = \min\{k, \max(-k, r)\}$, $k > 0$, and $\beta(k_3)$ is a consistency factor (see, *e.g.,* Huber and Ronchetti, 2009).

The purpose of this example is to point out the importance of using bounded estimating functions instead of unbounded ones. Indeed, the reason for using the former is threefold: to show that the χ^2 approximation to the distribution of $\hat{h}(\hat{\theta}_p)$ has a second order relative error; to show that the second order agreement also holds between the asymptotic distributions of $\hat{h}(\hat{\theta}_p)$ and $\hat{h}(\hat{\theta}_p^*)$; to provide versions of $\hat{h}(\hat{\theta}_p)$ whose accuracy remains stable under small contaminations of the model.

In order to take into account for contaminated and non-contaminated series, an additive outlier model has been considered (Maronna *et al.*, 2006). Hence (5.12) becomes

$$y_j = \phi_0 + \phi_1 y_{j-1} + \epsilon_j + u_j, \tag{5.15}$$

with $u_j \sim (1-\xi)\delta_0 + \xi N(\mu_u, \sigma_u^2)$, $\xi \in [0, 1]$, δ_0 point mass distribution located at zero, $\mu_\nu \in \mathbb{R}$, and $\sigma_\nu^2 > 0$.

The simulation study is based on 100.000 Monte Carlo trials and series of length $q = 50$ have been generated according to (5.15). The true parameter value θ is set to $(\phi_0, \phi_1, \sigma^2) = (0, 0.5, 1)$. We consider both non-contaminated series and series where at most 5% of data points are contaminated, with $\mu_\nu = \phi_0/(1-\phi_1)$ and $\sigma_\nu^2 = 25\sigma^2$. For each replication, $\hat{h}(\hat{\theta}_p)$ and $\hat{h}(\hat{\theta}_p^*)$ have been computed using the estimating functions in (5.14), and these statistics are denoted, respectively, with $\hat{h}(\hat{\theta}_p)_\gamma$ and $\hat{h}(\hat{\theta}_p^*)_\gamma$, where $\gamma = (k_1, k_2, k_3)$. We use two values of γ, $\gamma_1 = (1.3, 1.3, 1.3)$ and $\gamma_2 = (\infty, \infty, \infty)$ leading, respectively, to a bounded and to an unbounded estimating function. It is worth to note that in this example some care is needed to evaluate $\hat{h}(\hat{\theta}_p^*)_\gamma$: in order to preserve the dependence structure of the series and to be consistent with the specification of (5.13), pairs of data points (y_{j-1}, y_j) must be resampled instead of single observations y_j.

In Table 5.2 the empirical coverage probabilities of confidence regions for θ are reported. When $\xi = 0$, the comparison between $\hat{h}(\hat{\theta}_p)_{\gamma_1}$ and $\hat{h}(\hat{\theta}_p)_{\gamma_2}$ highlights that the use of a bounded estimating function improves the accuracy of the χ^2 approximation. Moreover, the distributions of $\hat{h}(\hat{\theta}_p^*)_{\gamma_1}$ and $\hat{h}(\hat{\theta}_p)_{\gamma_1}$ are very close, and the accuracy of the approximation is comparable to that of the full log likelihood ratio $W(\theta)$. When contamination occurs, the coverage levels of the nonparametric saddlepoint test statistics computed with $\Psi(\theta)_{\gamma_1}$ remain quite stable, while those of the log likelihood ratio $W(\theta)$ and of $\hat{h}(\hat{\theta}_p)_{\gamma_2}$ drop, as one would expect.

5.6 Final remarks

In this Chapter a nonparametric saddlepoint test statistic based on the maximum pairwise likelihood estimator has been discussed. The use of the pro-

	$\xi = 0$			$\xi = 0.05$		
$1-\alpha$	0.90	0.95	0.99	0.90	0.95	0.99
$W(\theta)$	0.8888	0.9432	0.9882	0.2967	0.3363	0.3999
$\hat{h}(\hat{\theta}_p^*)_{\gamma_1}$	0.8914	0.9447	0.9892	0.8885	0.9434	0.9886
$\hat{h}(\hat{\theta}_p)_{\gamma_1}$	0.9007	0.9512	0.9875	0.8573	0.9298	0.9872
$\hat{h}(\hat{\theta}_p)_{\gamma_2}$	0.8232	0.8822	0.9534	0.3623	0.4356	0.5621

Table 5.2: First order autoregressive model: coverage probabilities of confidence regions for θ based on 100.000 Monte Carlo trials.

posed test statistic $\hat{h}(\hat{\theta}_p)$ is particularly appealing in the pairwise and more generally in the composite likelihood framework for the following motivations:

1. $\hat{h}(\hat{\theta}_p)$ can be derived under mild assumptions that do not involve the specification of the joint distribution. Indeed, as pointed out in Section 5.3, the proposed test statistic is based on a suitable nonparametric estimate of the underlying distribution under the null hypothesis. The only assumptions needed involve some smoothness conditions on the estimating function in order to justify the validity of formal Edgeworth expansions. For more details see Field *et al.* (2008) and Ma and Ronchetti (2011);

2. the computation of $\hat{h}(\hat{\theta}_p)$, as well as its asymptotic distribution, do not depend on the elements of the Godambe information. In particular, its asymptotic distribution is standard chi-square. As highlighted in Section 3.4.1 the computation of the elements of the Godambe information can be troublesome in some circumstances. For instance, when a spatial process or a time series are observed, the estimate of $J(\theta)$ must be obtained by resorting to resample methods, such as the jackknife and windows subsampling (see Section 3.4.1). The example of Section 5.5.2 shows that the proposed $\hat{h}(\hat{\theta}_p)$ is very close to the likelihood ratio $W(\theta)$, even at the central model, while overcoming the estimation of $J(\theta)$ and $H(\theta)$;

3. the test statistic $\hat{h}(\hat{\theta}_p)$ claims a high level of accuracy and it is not related to a specific setting. Indeed, in the pairwise likelihood framework it is not possible to derive general results, since the specification of the pairwise likelihood depends on the structure of the data. Therefore, the inferential procedures are affected by the particular specification of the pairwise likelihood function and the properties, such as the efficiency of the estimator, must be investigated case by case.

Chapter 6

Robust pairwise likelihood estimation of multivariate location and scatter

6.1 Introduction

Chapters 4 and 5 discussed two solutions to cope with the possible lack of accuracy of pairwise likelihood test statistics, due to the estimation of the elements of the Godambe information. However, any attempt to improve the convergence of the considered statistics to their asymptotic distribution would become a wild-goose chase in the presence of small deviations from the assumptions under which they are developed. Indeed, the solutions proposed in the previous Chapters do not take into account for the possible occurrence of outliers and influential observations, and their potential effect on composite likelihood inference. For instance, the stability of the level and the power of these test statistics relies on the robustness of the maximum composite likelihood estimator.

As far as we know, the study of the robustness of composite likelihood-based procedures has been largely neglected by the statistical literature. As outlined in Section 3.4.2, Xu and Reid (2011) provide results about the consistency of maximum composite likelihood estimators derived from marginal composite likelihoods, when there is a correct marginal specification of the model regardless the correct specification of the model. Despite this is a valuable result, it is not possible to measure, for instance, the bias of the maximum composite likelihood estimator caused by outliers and influential observations. Indeed, this would require a rigorous definition of a gross error model in the composite likelihood framework.

The results discussed in Chapter represent a first attempt to provide a robust maximum composite likelihood estimator for multivariate location and covariance, and some preliminary results are presented for some spe-

cific models. The proposed robust pairwise likelihood estimator stems from the idea of minimum covariance determinant estimator (Rousseeuw, 1984). Notably, the proposed estimator claims a high breakdown point (see Section 2.4) and, in particular, it has the following advantages over the existing robust estimators for multivariate location and covariance:

1. it is based on the pairwise score function, that does not need to be modified in order to bound the gross-errors. The computation of robust M-estimators usually requires either a modification of the score function in order to bound the gross-error sensitivity or the specification a new estimating function.

2. the simplifications provided by the use of pairwise likelihood functions are appealing when robust inference is needed for complex models. Indeed, there are situations where classical robust estimators are difficult to obtain due to the complex dependence structure of the data. Examples where robust inference is still challenging are mixed linear models (Heritier *et al.*, 2009) and time series models (Maronna *et al.*, 2006).

   The behavior of the proposed robust maximum pairwise likelihood estimator is illustrated in two examples in Section 6.4, and is compared with some existing robust estimators in the context of mixed linear models and time series models.

6.2 Minimum covariance determinant estimators

Consider a random vector $Y \in \mathbb{R}^q$, with vector of means $\mu \in \mathbb{R}^q$ and covariance matrix Σ, belonging to the set of all positive definite matrices of size q. Here, the density function of Y is assumed to be of the form

$$f(y; \mu, \Sigma) = |\Sigma|^{-1/2} g(d(y; \mu, \Sigma)), \tag{6.1}$$

where

$$d(y; \mu, \Sigma) = \sqrt{(y-\mu)^\top \Sigma^{-1} (y-\mu)}$$

is the Mahalanobis distance between y and μ. In (6.1) the function $g(\cdot)$ is known and it is assumed to have a strictly negative first derivative, so that the density function (6.1) belongs to the parametric class of elliptically symmetric distributions (Croux and Haesbroeck, 1999). For instance, the multivariate normal density is obtained by setting $g(x) \propto e^{-\frac{1}{2}x^2}$.

Let $y = (y_1, \ldots, y_n)$ be a random sample of size n from Y, and let H be the set including all the possible subsets of size $h = \lfloor \delta n \rfloor$, $0 < \delta < 1$, that can be obtained from $\{1, \ldots, n\}$, where $\lfloor x \rfloor$ denotes the integer part of x. The minimum covariance determinant (MCD) estimators of location and

covariance are, respectively,

$$\hat{\mu}(\mathcal{S}) = \frac{1}{h}\sum_{j\in\mathcal{S}} y_j \tag{6.2}$$

and

$$\hat{\Sigma}(\mathcal{S}) = \frac{1}{h}\sum_{j\in\mathcal{S}} (y_j - \hat{\mu}(\mathcal{S}))\,(y_j - \hat{\mu}(\mathcal{S}))^\top, \tag{6.3}$$

where $\mathcal{S} \in H$ is such that

$$|\hat{\Sigma}(\mathcal{S})| \leq |\hat{\Sigma}(\mathcal{L})|, \quad \text{for all } \mathcal{L} \in H.$$

Hence, MCD estimators select the subset of h observations out of n, whose covariance matrix has the lowest determinant.

The size of the subset $\mathcal{S}$ is crucial in determining the breakdown point of MCD estimators, and to balance the trade-off between robustness and efficiency. The choice $h = \lfloor (n+q+1)/2 \rfloor$ yields to the highest possible breakdown point (Lopuhaa and Rousseeuw, 1991). However, since the breakdown point of MCD estimators is $\min(\delta, 1-\delta)$, the value $h = \delta n = 0.75n$ yields to a better compromise between efficiency/stability and high breakdown (Croux and Haesbroeck, 1999).

MCD estimators (6.2) and (6.3) are $n^{1/2}$ consistent and asymptotically normally distributed (Butler *et al.*, 1993), i.e.

$$n^{1/2}(\hat{\mu}(\mathcal{S}) - \mu) \xrightarrow{d} N(0, \kappa(\delta)\Sigma)$$

and

$$n^{1/2}(\hat{\Sigma}(\mathcal{S}) - \kappa(\delta)\Sigma) \xrightarrow{d} N(0, \Omega(\delta)), \tag{6.4}$$

where the constant $\kappa(\delta)$ can be chosen in order to obtain consistency with respect to the assumed model. In particular, for elliptically symmetric unimodal distributions, the consistency factor is (Butler *et al.*, 1993)

$$\kappa(\delta) = \int_0^{\sqrt{k_\delta}} r^{q+1} g(r^2) dr, \tag{6.5}$$

where $k_\delta = F^{-1}(1-\delta)$ and $F(\cdot)$ is the distribution function associated to $f(\cdot)$. The matrix $\Omega(\delta)$ in (6.4) gives the asymptotic variance of the MCD estimator of covariance and its expression is given in Croux and Haesbroeck (1999, page 169). The influence function of $\hat{\mu}(\mathcal{S})$ and $\hat{\Sigma}(\mathcal{S})$ are given, respectively, in Butler *et al.* (1993) and Croux and Haesbroeck (1999).

When the sample size is small, the consistency factor (6.5) is not sufficient to make the MCD estimator of covariance unbiased. In this respect, Pison *et al.* (2002) provide some finite sample correction factors when the underlying distribution is the multivariate normal.

6.2.1 Computation of MCD

The major drawback of the MCD estimator is its computation time. Indeed, the best subset $\mathcal{S}$ must be searched in the set H, whose cardinality increases with the sample size. Thus, its computation turns out to be unfeasible for large samples. To cope with this issue, Rousseeuw and Van Driessen (1999) proposed the so called FAST-MCD algorithm, that can handle samples of size with order of magnitude tens of thousands.

The MCD solution $\mathcal{S}$ is obtained by repeating selective iterations, called C-steps. Below, a sketch of the algorithm is reported.

Compute preliminary estimates $\hat{\mu}(\mathcal{L}_0)$ and $\hat{\Sigma}(\mathcal{L}_0)$ by using a subset $\mathcal{L}_0$ of size $q+1$. Then, for $b = 1, \ldots, B$:

1. calculate the distances

$$d_i^b = \sqrt{(y_i - \hat{\mu}(\mathcal{L}_b))^\top \hat{\Sigma}(\mathcal{L}_b)^{-1}(y_i - \hat{\mu}(\mathcal{L}_b))},\ i = 1, \ldots, n;$$

2. take a new subset $\mathcal{L}_{b+1}$, by keeping those indices corresponding to those y_i leading to the h lowest distances;

3. obtain new estimates based on $\mathcal{L}_{b+1}$, i.e.

$$\hat{\mu}(\mathcal{L}_{b+1}) = \frac{1}{h} \sum_{j \in \mathcal{L}_{b+1}} y_j$$

and

$$\hat{\Sigma}(\mathcal{L}_{b+1}) = \frac{1}{h} \sum_{j \in \mathcal{L}_{b+1}} (y_j - \hat{\mu}(\mathcal{L}_{b+1}))\,(y_j - \hat{\mu}(\mathcal{L}_{b+1}))^\top .$$

4. If $|\hat{\Sigma}(\mathcal{L}_{b+1}))| < |\hat{\Sigma}(\mathcal{L}_b))|$ return to Step 1, otherwise stop.

This algorithm (Rousseeuw and Van Driessen, 1999) requires to repeat steps 1-4 for different starting subsets $\mathcal{L}_0$.

6.3 Robust maximum pairwise likelihood estimator

In this Section a robust maximum pairwise likelihood estimator is proposed, by using the idea of the minimum covariance determinant estimator.

Before starting, recall that the maximum pairwise likelihood estimator $\hat{\theta}_p$ is defined implicitly through the pairwise score equation

$$ps(\theta) \quad = \quad \sum_{i=1}^{n} ps(\theta; y_i) = \sum_{i=1}^{n} \left\{ \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} \frac{\partial \log f(y_{ir}, y_{is}; \theta)}{\partial \theta} \right\} = 0.$$

The basic idea of MCD estimators is to compute the sample mean and sample covariance matrix with a suitable subset of observations. However, pairwise likelihood functions are usually defined for a parameter θ, which does not necessarily coincide with the vector of means μ and the covariance matrix Σ of the underlying distribution. Instead, θ is usually in some relation with μ and Σ, i.e. $\mu = \mu(\theta)$ and $\Sigma = \Sigma(\theta)$. Hence, steps 1 and 3 of the FAST-MCD algorithm need a slight modification, since it is necessary to compute the maximum pairwise likelihood estimate $\hat{\theta}_p(\mathcal{L}_b)$ by solving

$$\sum_{i \in \mathcal{L}_b} \left\{ \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} \frac{\partial \log f(y_{ir}, y_{is}; \theta)}{\partial \theta} \right\} = 0,$$

and then to take $\hat{\mu}(\mathcal{L}_b) = \mu(\hat{\theta}_p(\mathcal{L}_b))$ and $\hat{\Sigma}(\mathcal{L}_b) = \Sigma(\hat{\theta}_p(\mathcal{L}_b))$.

Theoretically, the robust maximum pairwise likelihood estimator can be simply obtained with the above minor modification to the FAST-MCD algorithm. However, as outlined in Section 3.4.2, in the composite likelihood framework it is not easy to define a gross-error model of the form

$$\mathcal{P}_\epsilon(F_\theta^c) = \{F_\epsilon^c | F_\epsilon^c = (1-\epsilon)F_\theta^c + \epsilon G\},$$

where F_θ^c is supposed to be the central model that includes all the models consistent with the marginal specification $f_{Y_r,Y_s}(\cdot)$, in order to justify the results reviewed in Section 2.4 and Section 6.2. Roughly speaking, without the definition of a gross error model it is not clear how "far" from the central model it is possible to go while keeping the estimator reliable.

Results in Xu and Reid (2011) reveal that the maximum composite likelihood estimator is consistent if a correct marginal specification of the model is achieved regardless the correct specification of the model (see Section 3.4.2). Hence, the proposed robust maximum pairwise likelihood estimator is derived following this idea. In place of considering n q-dimensional observations, $m = nq(q-1)/2$ pairs of data points are considered as they were a sample from a bivariate distribution. Then, the pairwise score function, to be used in the FAST-MCD algorithm, has the following alternative expression (see Section 4.2)

$$\bar{ps}(\theta) = \bar{ps}(\theta; y) = \sum_{h=1}^{m} \frac{\partial \log f(y \in \mathcal{E}_h; \theta)}{\partial \theta} = \sum_{h=1}^{m} ps_h(\theta),$$

where, without loss of generality $f(y \in \mathcal{E}_1; \theta) = f(y_{11}, y_{12}; \theta)$, $f(y \in \mathcal{E}_2; \theta) = f(y_{11}, y_{13}; \theta), \ldots, f(y \in \mathcal{E}_m; \theta) = f(y_{n(q-1)}, y_{nq}; \theta)$.

The FAST-MCD algorithm needs two further modifications in order to accommodate for $\bar{ps}(\theta)$. First, in Step a, the Mahalanobis distances are evaluated for pairs of data points $y_{irs} = (y_{ir}, y_{is})$, i.e.

$$d_{irs}^b = \sqrt{(y_{irs} - \hat{\mu}(\mathcal{L}_b)_{rs})^\top \hat{\Sigma}(\mathcal{L}_b)_{rs}^{-1} (y_{irs} - \hat{\mu}(\mathcal{L}_b)_{rs})},$$

where $\hat{\mu}(\mathcal{L}_b)_{rs} = (\hat{\mu}(\mathcal{L}_b)_r, \hat{\mu}(\mathcal{L}_b)_s)$ and $\hat{\Sigma}(\mathcal{L}_b)_{rs}$ is a 2×2 matrix whose diagonal elements are $\hat{\Sigma}(\mathcal{L}_b)_{rr}$, $\hat{\Sigma}(\mathcal{L}_b)_{ss}$, respectively, while the off-diagonal ones are $\hat{\Sigma}(\mathcal{L}_b)_{rs}$. Second, the consistency factor (6.5) has to be evaluated with $q = 2$, and it is given by

$$\kappa(\delta) = \int_0^{\sqrt{k_\delta}} r^3 g(r^2) dr.$$

6.3.1 Computation of the robust maximum pairwise likelihood estimator

In the following, the main steps of the FAST-MCD algorithm, to compute the robust maximum pairwise likelihood estimator, are summarized.

Compute a preliminary estimate $\hat{\theta}(\mathcal{L}_0)$. Then, take

$$\hat{\mu}(\mathcal{L}_0) = \mu(\hat{\theta}(\mathcal{L}_0))$$

and

$$\hat{\Sigma}(\mathcal{L}_0) = \Sigma(\hat{\theta}(\mathcal{L}_0)).$$

For $b = 1, \ldots, B$:

1. calculate the distances

$$d^b_{irs} = \sqrt{(y_{irs} - \hat{\mu}(\mathcal{L}_b)_{rs})^\top \hat{\Sigma}(\mathcal{L}_b)^{-1}_{rs} (y_{irs} - \hat{\mu}(\mathcal{L}_b)_{rs})},$$
$$i = 1, \ldots, n,\ r \neq s = 1, \ldots, q;$$

2. take a new subset $\mathcal{L}_{b+1}$, by keeping those indices corresponding to those pairs (y_{ir}, y_{is}) leading to the $h = \lfloor \delta m \rfloor$ lowest distances;

3. obtain $\hat{\theta}_p(\mathcal{L}_{b+1})$, and then compute

$$\hat{\mu}(\mathcal{L}_{b+1}) = \mu(\mathcal{L}_{b+1})$$

and

$$\hat{\Sigma}(\mathcal{L}_{b+1}) = \Sigma(\mathcal{L}_{b+1});$$

4. If $|\hat{\Sigma}(\mathcal{L}_{b+1}))| < |\hat{\Sigma}(\mathcal{L}_b))|$ return to Step 2, otherwise stop.

6.4 Numerical examples

6.4.1 MCD in mixed linear models

The general formulation of a mixed linear model (MLM) is

$$Y = X\beta + \sum_{j=1}^{r} Z_j \gamma_j + \epsilon \ ,$$

where X and Z_j are individual and cluster level design matrices, respectively, β is the $q\times 1$ vector of fixed effects and γ_j are n-dimensional vectors of random effects, $r \geq 1$. It is assumed that the random effects γ_j are independent each other and are normally distributed as $N(0, \sigma^2_{\gamma_j})$, that the error terms ϵ are independent and $N(0, \sigma^2_\epsilon)$, and that γ_j are independent of ϵ and the overall parameter vector is identifiable. In particular, the marginal distribution of Y_i is $N(X_i\beta, \Sigma_i)$ at the cluster level, with $\Sigma_i = \sum_{j=1}^r \sigma^2_{\gamma_j} \left[Z_j \Psi Z_j^T\right]_{(ii)} + \sigma^2_\epsilon I_q$, where $\left[Z_j \Psi Z_j^T\right]_{(ii)}$ stands for the ith block-diagonal element of $Z_j \Psi Z_j^T$; whereas $Y \sim N(X\beta, \Sigma)$, with $\Sigma = \sum_{j=1}^r \sigma^2_{\gamma_j} Z_j \Psi Z_j^T + \sigma^2_\epsilon I_N, N = nq$.

The multivariate normal formulation of MLMs has been the starting point in Copt and Feser (2006) in order to develop robust techniques based on constrained S-estimators for μ and Σ. By adopting the same model formulation, we aim at studying the behavior of the MCD solution based on the maximum likelihood (ML), restricted maximum likelihood (REML) and pairwise maximum likelihood (PML) estimators, respectively.

In order to investigate the behavior of the proposed method, we consider a real example with data coming from an experiment in which 5 types of electrodes were applied to the arms of 16 subjects and their skin resistance is measured. This example was also considered by Copt and Feser (2006); see references therein for the original source of the data.

The model is a one-way within factor design, given by

$$y_{ir} = \mu + \beta_r + \gamma_i + \epsilon_{ir},\ i = 1, 2, \ldots, 16,\ r = 2, \ldots, 5.$$

The skin resistance is the response, that is assumed to depend on the electrode type acting here as a fixed effect and the random effect has one level. This formulation leads to a multivariate normal model with vector of means $\mu = vec(\mu + \beta_j)$ and compound symmetric covariance matrix $\Sigma = \sigma^2 R$, where $\sigma^2 = \sigma^2_\gamma + \sigma^2_\epsilon$ and R has unit diagonal values and off-diagonal elements $\rho = \sigma^2_\gamma/(\sigma^2_\gamma + \sigma^2_\epsilon)$. As the pairwise likelihood estimator of $\theta = (\mu, \sigma^2_\gamma, \sigma^2_\epsilon)$ coincides with its maximum likelihood version (see Pace *et al.* 2011), this example is rather peculiar and it is considered as a toy example to asses the reliability of the pairwise MCD with respect to the other robust and non robust methods.

The standard errors for the regression parameters of the proposed MCD estimators are evaluated as

$$se(\hat{\beta}) = \left[\kappa(\delta) \left(\sum_{i=1}^n X_i^T \hat{\Sigma}_i^{-1} X_i\right)^{-1}\right]^{1/2}, \tag{6.6}$$

where $\hat{\Sigma}$ is the MCD estimate.

Table 6.1 gives the fitted model by classical ML and REML, the proposed MCD estimators, evaluated with $\delta = 0.75$, and the constrained S-estimator (CS) of Copt and Feser (2006) (evaluated with a 75%-breakdown point).

The MCD based techniques prevent against the overestimation of the (6.6), and of the variance components provided by ML and REML, in the same fashion as the constrained-S approach of Copt and Feser (2006), even if the resulting values are larger with respect to the latter method.

The employ of the pairwise likelihood equations leads to larger $se(\hat{\beta})$ and to larger estimates of variance components than the use of (restricted) likelihood equations in the MCD machinery. However, this behavior has to be expected and comes from the fact that the full likelihood has been approximated starting from a misspecified model.

	MCD-ML	MCD-PML	MCD-REML
β_1	1.288 (0.386)	1.456 (0.405)	1.288 (0.403)
β_2	0.625 (0.406)	0.686 (0.443)	0.625 (0.424)
β_3	0.320 (0.406)	0.352 (0.443)	0.320 (0.424)
β_4	-0.104 (0.406)	0.028 (0.443)	-0.104 (0.424)
β_5	-0.158 (0.406)	-0.195 (0.443)	-0.158 (0.424)
σ^2_γ	0.422	0.562	0.460
σ^2_ϵ	0.523	0.835	0.571
	CS	ML	REML
β_1	1.279 (0.281)	1.817 (0.463)	1.817 (0.478)
β_2	0.564 (0.276)	1.056 (0.512)	1.056 (0.529)
β_3	0.404 (0.276)	0.763 (0.512)	0.763 (0.529)
β_4	-0.008 (0.276)	-0.313 (0.512)	-0.313 (0.529)
β_5	-0.155 (0.276)	-0.438 (0.512)	-0.438 (0.529)
σ^2_γ	0.710	1.329	1.418
σ^2_ϵ	0.579	2.098	2.238

Table 6.1: Skin resistance data: estimates (standard errors) by MCD-ML, MCD-PML, MCD-REML, CS, ML and REML.

6.4.2 The case of one observation: first order autoregression

Consider a normal autoregressive process of order one, of the form

$$y_{ir} - \mu = \rho(y_{ir-1} - \mu) + \epsilon_{ir},\ i = 1, 2, \ldots, n,\ r = 2, \ldots, q,$$

where ϵ_{ir} are independently normal distributed with zero mean, variance σ^2 and covariance $(Y_{ir}, Y_{is}) = \sigma^2\rho^{|r-s|}/(1-\rho^2), r, s = 1, \ldots, q$. Assume that a single observed series is available. The pairwise log likelihood for $\theta = (\mu, \sigma^2, \rho)$ is derived using only pairs of contiguous components, i.e. for

$r-s=1$, $r>s$, as outlined in Pace *et al.* (2011). The total number of pairs is $N=q-1$.

When it is of interest to perorm robust inference in this setting, the use of the pairwise-MCD based estimator appears particularly appealing, since it can be based on a set of N bivariate observations and represent an alternative to existing robust methods (see Maronna *et al.*, 2006, for a detailed account).

In order to asses the finite sample behavior of the pairwise-MCD estimator, a simulation study based on 1000 Monte Carlo trials has been performed. The contaminated scenario is based on an additive outlier model, according to which the observed value at time r is $y_r-\mu+\nu_r$, with

$$\nu_r \sim (1-\kappa)\delta_0 + \kappa N(\mu_\nu, \sigma_\nu^2),$$

where δ_0 is a point mass distribution located at zero and $0 \leq \kappa \leq 1$. The setting of the numerical study is as follows: $\kappa = 0, 0.1$, $\mu = 0$, $\sigma^2 = 1$, $\rho = 0.5$, $\mu_\nu = \mu$ and $\sigma_\nu^2 = 10\sigma^2$. The size of the subsets for the MCD-pairwise likelihood procedure is $h = \delta N = 0.75N$.

Table 6.2 provides the means of the maximum likelihood, pairwise maximum likelihood and MCD-pairwise maximum likelihood estimates for the parameter θ. It can be observed that, when the model is not contaminated, the pairwise-MCD estimator behaves closely to the ML and the PML estimators. On the contrary, under 10% contamination the pairwise-MCD estimator provides accurate and reliable estimation.

		$q = 200$		
		ML	PMLE	MCD-PMLE
μ	$\kappa = 0$	0.003 (0.137)	0.001 (0.138)	0.001 (0.147)
	$\kappa = 0.1$	0.006 (0.177)	0.005 (0.178)	0.001 (0.153)
σ^2	$\kappa = 0$	0.987 (0.101)	0.987 (0.101)	0.971 (0.116)
	$\kappa = 0.1$	2.371 (0.610)	2.371 (0.612)	1.150 (0.181)
ρ	$\kappa = 0$	0.494 (0.063)	0.493 (0.063)	0.486 (0.081)
	$\kappa = 0.1$	0.264 (0.081)	0.264 (0.081)	0.448 (0.084)
		$q = 500$		
μ	$\kappa = 0$	0.005 (0.089)	0.003 (0.088)	0.001 (0.093)
	$\kappa = 0.1$	0.013 (0.012)	0.006 (0.121)	0.002 (0.096)
σ^2	$\kappa = 0$	0.994 (0.065)	0.993 (0.064)	0.995 (0.078)
	$\kappa = 0.1$	2.480 (0.386)	2.481 (0.386)	1.171 (0.111)
ρ	$\kappa = 0$	0.496 (0.040)	0.496 (0.040)	0.492 (0.049)
	$\kappa = 0.1$	0.273 (0.051)	0.273 (0.051)	0.455 (0.049)

Table 6.2: First order autoregression: mean (standard errors) of ML, PML, MCD-PML estimators of θ for $q = 200$, 500, and $k = \{0, 0.1\}$

6.5 Final remarks

In this Chapter, a robust maximum pairwise likelihood estimator with a high breakdown point has been provided by exploiting the idea of the minimum covariance determinant estimator. The proposed robust procedure does not need to modify the given pairwise score function, neither to specify a new estimating function, as it is common in order to obtain robust M- or S-estimators. Furthermore, the proposed robust approach requires only mild assumptions about the shape of the underlying distribution and the computation of the estimator can be performed with a minor modification of the existing algorithm for the minimum covariance determinant estimator.

Some room for further investigation is left. Indeed, the robust maximum pairwise likelihood estimator has been derived for two specific models. The hard task to accomplish with the proposed robust estimator is to derive its standard errors. To this end, the influence function is needed, but its computation requires the specification of a gross error model. As mentioned in Section 3.4.2, the definition of a gross error model is not straightforward since it is not clear what the central model is.

Bibliography

Adimari, G. and Guolo, A. (2010). A note on the asymptotic behaviour of empirical likelihood statistics. *Stat. Meth. Appl.*, **19**, 463–476.

Adimari, G. and Ventura, L. (2002). Quasi-profile log likelihoods for unbiased estimating functions. *Ann. Inst. Statist. Math.*, **54**, 235–244.

Aerts, M. and Claeskens, G. (1999). Bootstrapping pseudolikelihood models for clustered binary data. *Ann. Inst. Statist. Math.*, **51**, 515–530.

Almudevar, A., Field, C., and Robinson, J. (2000). The density of multivariate M-estimates. *Ann. Statist.*, **28**, 275–297.

Barndorff-Nielsen, O. E. (1995). Quasi profile and directed likelihoods from estimating functions. *Ann. Inst. Statist. Math.*, **47**, 461–464.

Bellio, R. (2007). Algorithms for bounded-influence estimation. *Comput. Statist. Data Anal.*, **51**, 2531–2541.

Bellio, R. and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Stat. Model.*, **5**, 217–227.

Bellio, R., Greco, L., and Ventura, L. (2008). Modified quasi-profile likelihoods from estimating functions. *J. Statist. Plann. Inf.*, **138**, 3059–3068.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. B*, **36**, 192–236.

Bhattacharya, R. and Ghosh, J. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.*, **6**, 434–451.

Butler, R. (2007). *Saddlepoint approximations with applications*. Cambridge University Press.

Butler, R., Davies, P., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.*, **21**, 1385–1400.

Caragea, P. and Smith, R. (2006). Approximate likelihoods for spatial processes. *www.stat.unc.edu/postscript/rs/caragea.pdf*.

Carroll, R. and Ruppert, D. (1988). *Transformation and weighting in regression*. Chapman & Hall.

Chandler, R. and Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, **94**, 167–183.

Chandrasekar, B. and Kale, B. (1984). Unbiased statistical estimation functions for parameters in presence of nuisance parameters. *J. Statist. Plann. Inference*, **9**, 45–54.

Chen, S. (1993). On the accuracy of empirical likelihood confidence regions for linear regression model. *Ann. Inst. Statist. Math.*, **45**, 621–637.

Chen, S. and Cui, H. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika*, **93**, 215–220.

Clarke, B. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.*, **11**, 1196–1205.

Clarke, B. (1986). Nonsmooth analysis and Fréchet differentiability of M-functionals. *Prob. Theory Relat. Fields*, **73**, 197–209.

Cox, D. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91**, 729–737.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multivar. Anal.*, **71**, 161–190.

Daniels, H. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.*, **25**, 631–650.

Davis, R. and Yau, C. (2011). Comments on pairwise likelihood in time series models. *Statist. Sinica*, **21**, 255–277.

Desmond, A. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling. *J. Statist. Plann. Inf.*, **60**, 77–104.

DiCiccio, T. and Monti, A. (2001). Approximations to the profile empirical likelihood function for a scalar parameter in the context of M-estimation. *Biometrika*, **88**, 337–351.

DiCiccio, T. and Romano, J. (1989). On adjustments based on the signed root of the empirical likelihood ratio statistic. *Biometrika*, **76**, 447–456.

DiCiccio, T., Hall, P., and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.*, **19**, 1053–1061.

Dupuis, D. and Field, C. (1998). Robust estimation of extremes. *Can. J. Statist.*, **26**, 199–215.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canad. J. Statist.*, **9**, 139–158.

Field, C. (1982). Small sample asymptotic expansions for multivariate M-estimates. *Ann. Statist.*, **10**, 672–689.

Field, C., Robinson, J., and Ronchetti, E. (2008). Saddlepoint approximations for multivariate M-estimates with applications to bootstrap accuracy. *Ann. Inst. Statist. Math.*, **60**, 205–224.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, **62**, 424–431.

Geys, H., Molenberghs, G., and Ryan, L. (1999). Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *J. Amer. Statist. Assoc.*, **94**, 734–745.

Godambe, V. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208–1211.

Godambe, V. and Kale, B. (1991). Estimating functions: an overview. In *Estimating functions*, volume 7 of *Oxford Statist. Sci. Ser.*, pages 3–20. Oxford Univ. Press, New York.

Godambe, V. and Thompson, M. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Statist.*, **2**, 568–571.

Gong, G. and Samaniego, F. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.*, **9**, 861–869.

Hall, P. (1997). *The bootstrap and Edgeworth expansion*. Springer, Verlag.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *Int. Statist. Rev.*, **58**, 109–127.

Hampel, F. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust statistics*. Wiley, New York.

Hanfelt, J. and Liang, K. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461–477.

Heagerty, P. and Lele, R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.*, **93**, 1099–1111.

Heagerty, P. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *J. Amer. Statist. Assoc.*, **95**, 197–211.

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.*, **89**, 897–904.

Heritier, S., Cantoni, E., and Copt, S. (2009). *Robust methods in biostatistics*. Wiley, New York.

Hjort, N. and Omre, H. (1994). Topics in spatial statistics. *Scand. J. Statist.*, **21**, 289–357.

Hjort, N., McKeague, I., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.*, **37**, 1079–1111.

Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif.

Huber, P. (1981). *Robust statistics*. Wiley, New York.

Huber, P. and Ronchetti, E. (2009). *Robust statistics*. Wiley, New York.

Hudson, R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.

Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.

Jensen, J. and Wood, A. (1998). Large deviation and other results for minimum contrast estimators. *Ann. Inst. Statist. Math.*, **50**, 673–695.

Jørgensen, B. and Knudsen, S. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scand. J. Statist.*, **31**, 93–114.

Kale, B. (1962). On the solution of likelihood equations by iteration processes. The multiparametric case. *Biometrika*, **49**, 479–486.

Kent, J. (1982). Robust properties of likelihood ratio tests. *Biometrika*, **69**, 19–27.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Ann. Statist.*, **25**, 2084–2102.

Kolaczyk, E. (1994). Empirical likelihood for generalized linear models. *Statist. Sinica*, **4**, 199–218.

Kuk, A. (2007). A hybrid pairwise likelihood method. *Biometrika*, **94**, 939–952.

Kuk, A. and Nott, D. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statist. Prob. Lett.*, **47**, 329–335.

Kullback, S. (1997). *Information theory and statistics*. Dover.

Lazar, N. and Mykland, P. (1999). Empirical likelihood in the presence of nuisance parameters. *Biometrika*, **86**, 203–211.

Le Cessie, S. and Van Houwelingen, J. (1994). Logistic regression for correlated binary data. *Appl. Statist.*, **43**, 95–108.

Lindsay, B. (1988). Composite likelihood methods. In *Statistical inference from stochastic processes (Ithaca, NY, 1987)*, volume 80 of *Contemp. Math.*, pages 221–239. Amer. Math. Soc., Providence, RI.

Lindsay, B., Pilla, R., and Basak, P. (2000). Moment-based approximations of distributions using mixtures: Theory and applications. *Ann. Inst. Statist. Math.*, **52**, 215–230.

Lindsay, B., Yi, G., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica*, **21**, 71–105.

Lipsitz, S., Dear, K., and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics*, **50**, 842–846.

Lopuhaa, H. and Rousseeuw, P. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, **19**, 229–248.

Lugannani, R. and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Adv. Appl. Prob.*, **12**, 475–490.

Ma, Y. and Ronchetti, E. (2011). Saddlepoint test in measurement error models. *J. Amer. Statist. Assoc.*, **106**, 147–156.

Maronna, R., Bustos, O., and Yohai, V. (1979). Bias and efficiency-robustness of general M-estimators for regression with random carriers. *Smoothing Tech. Curve Estim.*, **757**, 91–116.

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust statistics*. Wiley, New York.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

McVean, G., Myers, S., Hunt, S., Deloukas, P., Bentley, D., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer, New York.

Monti, A. (1997). Empirical likelihood confidence regions in time series models. *Biometrika*, **84**, 395–405.

Nordman, D. (2008). A blockwise empirical likelihood for spatial lattice data. *Statist. Sinica*, **18**, 1111–1129.

Nordman, D. and Lahiri, S. (2006). A frequency domain empirical likelihood for short-and long-range dependence. *Ann. Statist.*, **34**, 3019–3050.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.*, **19**, 1725–1747.

Owen, A. (2001). *Empirical likelihood*. Chapman & Hall.

Pace, L. and Salvan, A. (1997). *Principles of statistical inference: from a Neo-Fisherian perspective*. World Scientific, Singapore.

Pace, L., Salvan, A., and Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statist. Sinica*, **21**, 129–148.

Padoan, S., Ribatet, M., and Sisson, S. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.*, **105**, 263–277.

Parke, W. (1986). Pseudo maximum likelihood estimation: the asymptotic distribution. *Ann. Statist.*, **14**, 355–357.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statist. Sinica*, **21**, 149–164.

Pierce, D. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Statist.*, **10**, 475–478.

Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, **55**, 111–123.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.

Renard, D., Molenberghs, G., and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comput. Statist. Data Anal.*, **44**, 649–667.

Robinson, J., Ronchetti, E., and Young, G. (2003). Saddlepoint approximations and tests based on multivariate M-estimates. *Ann. Statist.*, **31**, 1154–1169.

Ronchetti, E. and Trojani, F. (2001). Robust inference with GMM estimators. *J. Econ.*, **101**, 37–69.

Ronchetti, E. and Welsh, A. (1994). Empirical saddlepoint approximations for multivariate M-estimators. *J. Roy. Statist. Soc. B*, **56**, 313–326.

Rotnitzky, A. and Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485–497.

Rousseeuw, P. (1981). A new infinitesimal approach to robust estimation. *Prob. Theory Relat. Fields*, **56**, 127–132.

Rousseeuw, P. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871–880.

Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, **2**, 110–114.

Severini, T. (2002). Modified estimating functions. *Biometrika*, **89**, 333–343.

Skovgaard, I. (1990). On the density of minimum contrast estimators. *Ann. Statist.*, **18**, 779–789.

Smith, E. and Stephenson, A. (2009). An extended Gaussian max-stable process model for spatial extremes. *J. Statist. Plann. Inf.*, **139**, 1266–1275.

Stein, M., Chi, Z., and Welty, L. (2004). Approximating likelihoods for large spatial data sets. *J. Roy. Statist. Soc. B*, **66**, 275–296.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Stat. Anal.*, **92**, 1–28.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, **92**, 519–528.

Varin, C., Høst, G., and Skare, Ø. (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Comput. Statist. Data Anal.*, **49**, 1173–1191.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica*, **21**, 5–42.

Wang, M. and Hanfelt, J. (2003). Adjusted profile estimating function. *Biometrika*, **90**, 845–858.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **9**, 60–62.

Wood, A. (1989). An F approximation to the distribution of a linear combination of chi-squared variables. *Comm. Statist. - Simul. Comput.*, **18**, 1439–1456.

Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *www.utstat.utoronto.ca/ ximing/Robustness.pdf*.

Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, **20**, 642–656.

Nicola Lunardon

CURRICULUM VITAE

Personal Details

Date of Birth: May 20, 1984
Place of Birth: Maracaibo, Venezuela
Nationality: Italian

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4111
e-mail: lunardon@stat.unipd.it

Current Position

Since January 2009;
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Pseudo-likelihoods from unbiased estimating functions in complex models
Supervisor: Prof. Laura Ventura

Research interests

- Pseudo-likelihoods
- Saddlepoint approximations
- Robust statistics

Education

September 2007 – July 2008
Master (*laurea specialistica*) degree in Statistics for the Demographic and Social Sciences.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Model selection in estimating equations" (in Italian)
Supervisor: Prof. Laura Ventura
Final mark: 104

September 2004 – February 2007
Bachelor degree (*laurea triennale*) in Statistics and Management.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Sport e multiculturalità in Veneto: il punto di vista di allenatori e atleti."
Supervisor: Prof. Maria Cristiana Martini
Final mark: 101

Visiting periods

August 2011 – September 2011
University of Geneva
Geneva.
Supervisor: Prof. Elvezio Ronchetti

June 2011
Università degli Studi del Sannio
Italy.
Supervisor: Dott. Luca Greco

Awards and Scholarship

2009
PhD scholarship (University of Padova)

Computer skills

- Operative System: Linux, Windows, OSX
- Programming: C, R, Matlab
- Markup Languages: LaTeX, HTML
- Other skills: Parellel computing

Language skills

Italian/Spanish: native; English: fluent (written/spoken);

Publications

Lunardon, N., Pauli, F., Ventura, L. (2012). A note on empirical likelihoods derived from pairwise score functions, *J. Stat. Comput. Simul.*, to appear (DOI: 10.1177/0962280210385865)

Grilli, L., Rampichini, C., Salmaso, L., Lunardon, N. and Samuh, M. (2012). The use of permutation tests for variance components in linear mixed models, *Comm. Statist.*, to appear.

Lunardon, N., Greco, L. and Ventura, L. (2011) Pairwise robust estimation of multivariate location and covariance. In S.Co. 2011, Padova, September 19-21, USB stick (ISBN: 9788861297531), 1-6.

Lunardon, N., Pauli, F. and Ventura, L. (2010) On empirical composite likelihoods. In COMPSTAT 2010, Paris, August 22-27, e-book (ISBN: 9781605584959), 1319-1326.

Lunardon, N. and Ventura, L. (2009) A comparison of quasi-likelihood ratios for general estimating functions. In S.Co. 2009, Milano, September 14-16, 263-268.

Conference presentations

17-19/12/11: "Pairwise robust estimation of multivariate location and covariance", ERCIM 2011, London (contributed poster).

19-21/9/11: "Pairwise robust estimation of multivariate location and covariance", S.Co. 2011, Padova (contributed talk).

14-16/6/10: "A comparison of quasi-likelihood ratios for general estimating functions", S.Co. 2009, Milano (contributed talk).