Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Composite Likelihood Inference by Nonparametric Saddlepoint Tests

NICOLA LUNARDON

Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Piazzale Europa 1, 34127, Trieste, Italy.

ELVEZIO RONCHETTI

Research Center for Statistics and Dept. of Economics, University of Geneva, 1211, Geneva, Switzerland.

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Composite Likelihood Inference
by Nonparametric Saddlepoint Tests

NICOLA LUNARDON

Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Piazzale Europa 1, 34127, Trieste, Italy.

ELVEZIO RONCHETTI

Research Center for Statistics and Dept. of Economics, University of Geneva, 1211, Geneva, Switzerland.

ABSTRACT[1]

The class of composite likelihood functions provides a flexible and powerful toolkit to carry out approximate inference for complex statistical models when the full likelihood is either impossible to specify or unfeasible to compute. However, the strenght of the composite likelihood approach is dimmed when considering hypothesis testing about a multidimensional parameter because the finite sample behavior of likelihood ratio, Wald, and score-type test statistics is tied to the Godambe information matrix. Consequently inaccurate estimates of the Godambe information translate in inaccurate p-values. In this paper it is shown how accurate inference can be obtained by using a fully nonparametric saddlepoint test statistic derived from the composite score functions. The proposed statistic is asymptotically chi-square distributed up to a relative error of second order and does not depend on the Godambe information. The validity of the method is demonstrated through simulation studies.

[1]**Corresponding author:** Nicola Lunardon, Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Piazzale Europa 1, 34127, Trieste, Italy, email: nicola.lunardon@econ.units.it

1. Introduction

The likelihood function plays a central role in statistical inference. However, with statistical models becoming increasingly complex in many fields such as genetics and finance, the full likelihood function is often not available in closed form or is too difficult to specify. This can be due for instance to a complex dependence structure of the data. Examples include e.g. the estimation of diffusion models in finance and models based on max-stable processes for spatial multivariate extremes (Padoan *et al.* (2010), Thibaud *et al.* (2013)). Even when the specification of the full likelihood is straightforward, its evaluation can be computationally awkward. 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 these cases and in the frequentist setting, one can rely on indirect inference techniques (see the surveys by Heggland and Frigessi (2004) and Jiang and Turnbull (2004)), whereas in the Bayesian framework one can use sequential Monte Carlo methods for approximate Bayesian computations (see, for instance Del Moral *et al.* (2006), Beaumont *et al.* (2009)).

An attractive alternative which has gained popularity in the past few years is the approach based on composite likelihood functions originally proposed by Lindsay (1988). The basic idea is to approximate the unknown full likelihood by a sum of likelihood components obtained e.g. by combining either marginal or conditional densities. An important special case is the pairwise likelihood constructed using pairs of components; see Cox and Reid (2004). Although the resulting combined function is no longer a proper likelihood, the derived inferential procedures are M-estimators and tests based on unbiased estimating functions. From a theoretical point of view this is an appealing property because their asymptotic theory is readily available; cf. e.g. Heritier and Ronchetti (1994) in the context of robust tests. Specifically, Wald and score test statistics for pairwise likelihoods are asymptotically χ^2 distributed, whereas the asymptotic distribution of the pairwise log-likelihood ratio test statistic is a linear combination of independent χ_1^2 random variables.

The use of composite likelihoods has been advocated by several authors both in the frequentist setting (see the good review paper by Varin *et al.* (2011) in a special issue devoted to this topic in *Statistica Sinica*) and also in the Bayesian framework (Pauli *et al.*, 2011; Ribatet *et al.*, 2011). Successful use of this approach in fairly complex models include applications in spatial processes (Heagerty and Lele (1998), Varin *et al.* (2005)), generalized linear mixed models (Renard *et al.* (2004), Bellio and Varin (2005)), longitudinal models (Fieuws and Verbeke, 2006), and genetics (Hudson (2011), McVean *et al.* (2004)).

In spite of the availability of standard asymptotic theory for Wald, score, and likelihood ratio tests based on pairwise likelihoods, their actual computation requires the evaluation of the expectations of minus the derivative and of the square of the pairwise likelihood score which, as opposite to the full likelihood score, are not equal. Their estimation in this case is akward and the corresponding p-values and coverage probabilities based on the asymptotic distribution become inaccurate when the sample size is moderate or when

small tail probabilities are required; cf. Section 2 and 4. To improve the accuracy, the test statistics could be adjusted as in the classical case by means of Barlett corrections and related methods. However, these methods would provide only improvements in terms of the absolute error of the approximation which would still be inaccurate in the tails.

In this paper we consider an alternative test for pairwise likelihood defined by (4). It is a nonparametric test derived by building on the results by Robinson *et al.* (2003). It enjoys the following desirable properties: i) the test statistic is asymptotically χ^2 distributed; ii) the χ^2 approximation to the exact distribution has a *relative error* of order $O(n^{-1})$; iii) the test is fully nonparametric; iv) the test can combine accuracy and robustness by an appropriate choice of the pairwise likelihood score; v) the test does not require the computation of elements of the asymptotic covariance matrix of M-estimators (so-called sandwich formula or Godambe information); vi) the test statistic is parametrization invariant.

These properties will be discussed in detail in Section 3 and make this test an attractive alternative for inference with pairwise likelihoods.

The rest of the paper is organized as follows. In Section 2 we define the pairwise likelihood and discuss the available test procedures. In Section 3 we introduce the new test and discuss its properties. Section 4 present three examples that show the excellent finite sample behavior of the new test. Finally, some conluding remarks and an outlook are given in Section 5.

2. Pairwise Likelihood

Let $y = (y_1, \ldots, y_n)^{\mathrm{T}}$, be a random sample of independent realizations of the q-dimensional random vector Y having probability distribution $F(\cdot; \theta)$ and density function $f(\cdot; \theta)$, $\theta \subseteq \mathbb{R}^p$. The full log-likelihood function and ratio are respectively $\ell(\theta) = \log f(y; \theta)$ and $w(\theta) = 2[\ell(\hat{\theta}) - \ell(\theta)]$, with $\hat{\theta}$ the maximum likelihood estimate. Consider a set of measurable events $\{\mathcal{E}_r \in \mathcal{Y}, r = 1, \ldots, m\}$ on the sample space $\mathcal{Y}$, defined for pairs of components (y_{ij}, y_{ik}), $j \neq k = 1, \ldots, q$, 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 event $\mathcal{E}_r$. Then the pairwise log-likelihood is defined as

$$p\ell(\theta) = \sum_{i=1}^{n} \sum_{r=1}^{m} \omega_{ir} \log f_r(y_i; \theta), \tag{1}$$

where ω_{ir} are weights not depending on θ nor y. In general these weights are chosen both to improve the efficiency of the maximum pairwise likelihood estimator and to reduce the computational effort (Lindsay *et al.*, 2011). The pairwise score function associated to (1) is

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

Since it is a combination of genuine scores, $ps(\theta)$ is an unbiased estimating function, that is $\mathbb{E}_F [ps(\theta)] = 0$, where the notation $\mathbb{E}_F$ is used to highlight that expectation is taken with respect to the full model.

The maximum pairwise likelihood estimator $\hat{\theta}_p$ belongs to the class of M-estimators and is implicitly defined through the equation

$$ps(\theta) = 0.$$

Under broad conditions (see, e.g., Molenberghs and Verbeke, 2005), the maximum pairwise likelihood estimator is consistent and asymptotically normal, with covariance matrix given by the so-called sandwich formula or expected Godambe information

$$V(\theta) = H(\theta)^{-1} J(\theta) H(\theta)^{-1},$$

where $J(\theta) = \mathbb{E}_F\left[ps(\theta; Y)ps(\theta; Y)^{\mathrm{T}}\right]$, $H(\theta) = -\mathbb{E}_F\left[\partial ps(\theta; Y)/\partial \theta^{\mathrm{T}}\right]$.

In the context of hypothesis testing, the pairwise likelihoods allow to perform the analogous of the Wald, the score and the likelihood ratio tests. The pairwise likelihood counterparts of the Wald and score test statistics are

$$pw_w(\theta) = n(\hat{\theta}_p - \theta)^{\mathrm{T}} V(\hat{\theta}_p)^{-1}(\hat{\theta}_p - \theta) \quad \text{and} \quad pw_s(\theta) = n^{-1} ps(\theta)^{\mathrm{T}} J(\theta)^{-1} ps(\theta),$$

respectively. Under the hypothesis $H_0 : \theta = \theta_0$ both $pw_w(\theta_0)$ and $pw_s(\theta_0)$ converge to a chi-square distribution with p degrees of freedom. Instead, the pairwise log-likelihood ratio

$$pw(\theta) = 2\left\{p\ell(\hat{\theta}_p) - p\ell(\theta)\right\}$$

converges in distribution to $\sum_{j=1}^{p} \lambda_j(\theta) Z_j^2$, where $\lambda_1(\theta), \ldots, \lambda_p(\theta)$ are the eigenvalues of $H(\theta)^{-1} J(\theta)$ and the $Z_j's$ independent random variables with a standard normal distribution (see, e.g., Kent, 1982). Adjustments to $pw(\theta)$ have been proposed to provide a pairwise log-likelihood ratio with the usual asymptotic chi-square distribution. The simplest adjustment is based on first moment matching

$$pw_1(\theta) = \frac{pw(\theta)}{\kappa_1},$$

where $\kappa_1 = \mathbb{E}\left[\sum_{j=1}^{p} \lambda_j(\theta) Z_j^2\right]/p = \sum_{j=1}^{p} \lambda_j(\theta)/p$. A χ_p^2 approximation is used for the distribution of $pw_1(\theta)$ (see, e.g. Rotnitzky and Jewell, 1990). Alternatively, Chandler and Bate (2007) propose the so-called vertical scaling to $pw(\theta)$

$$pw_{cb}(\theta) = \frac{pw_w(\theta)}{\kappa_{cb}}, \tag{2}$$

where $\kappa_{cb} = n(\hat{\theta}_p - \theta)^{\mathrm{T}} H(\hat{\theta}_p)(\hat{\theta}_p - \theta)/pw(\theta)$. Finally, Pace *et al.* (2011) propose a parametrization invariant adjustment

$$pw_{inv}(\theta) = \frac{pw_s(\theta)}{\kappa_{inv}}, \tag{3}$$

where $\kappa_{inv} = n^{-1} ps(\theta)^{\mathrm{T}} H(\theta)^{-1} ps(\theta)/pw(\theta)$. Test statistics (2) and (3) are first order equivalent to $pw_w(\theta)$ and $pw_s(\theta)$ respectively and are asymptotically χ_p^2 distributed. Even with these adjustments, the χ^2 approximation for the distribution of these test statistics

may be inaccurate in moderate sample sizes or when small tail probabilities are required. The accuracy of the approximation mostly depends on the Godambe information matrix, as can be seen from the definition of the test statistics. To better understand this statement it is important to distinguish two relevant settings in the pairwise likelihood framework. In the first one, pairwise likelihoods replace the full likelihood function for computational convenience. Therefore, either analytic expressions or (parametric) bootstrap estimates for $J(\theta)$ and $H(\theta)$ can be worked out under the assumed $F(;\theta)$. In the second one, pairwise likelihoods are used as an approximation to $\ell(\theta)$ and in this case only empirical counterparts of such matrices can be computed. In the case of independent observations the estimates

$$\hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^{n} ps(\theta; y_i) ps(\theta; y_i)^{\mathrm{T}} \quad \text{and} \quad \hat{H}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial ps(\theta; y_i)}{\partial \theta^{\mathrm{T}}},$$

are consistent for $J(\theta)$ and $H(\theta)$, respectively. However, depending on the application area, $\hat{J}(\theta)$ may not be appropriate and a consistent estimate should be obtained by using resampling methods (see Varin *et al.*, 2011, and references therein).

The second setting is the most likely to occur in real applications and it is the most critical. Indeed, the estimation of $J(\theta)$ and $H(\theta)$ introduces additional variability and deteriorates the accuracy of the χ^2 approximation in finite samples. In the next section we present an alternative test which avoids these problems.

3. Saddlepoint Test

Consider for simplicity of notation the case of a simple hypothesis. The new test statistic is

$$pw_{sp}(\theta) = -2n \log \left\{ \sum_{i=1}^{n} w_i(\theta) \exp\{\lambda(\hat{\theta}_p)^{\mathrm{T}} ps(\hat{\theta}_p; y_i)\} \right\}, \tag{4}$$

where

$$w_i(\theta) = \exp\{\beta(\theta)^{\mathrm{T}} ps(\theta; y_i)\} / \sum_{j=1}^{n} \exp\{\beta(\theta)^{\mathrm{T}} ps(\theta; y_j)\},$$

$\beta(\theta)$ is the root of the equation

$$\sum_{i=1}^{n} w_i(\theta) ps(\theta; y_i) = 0, \tag{5}$$

and $\lambda(\hat{\theta}_p)$ satisfies the equation

$$\sum_{i=1}^{n} ps(\hat{\theta}_p; y_i) \exp\{\lambda(\hat{\theta}_p)^{\mathrm{T}} ps(\hat{\theta}_p; y_i)\} = 0.$$

The following theorem states the large sample properties of p-values obtained from test statistic (4). The proof is provided in the Appendix.

Theorem. *Suppose that conditions (A.1), (A.2), and (A.3) in the Appendix hold. Then under the null hypothesis $H_0 : \theta = \theta_0$*

$$P_{H_0}[pw_{sp}(\theta_0) \geq pw_{sp}(\theta_0)^{obs}] = (1 - Q_p(pw_{sp}(\theta_0)^{obs}))(1 + O_p(n^{-1}))$$

where $pw_{sp}(\theta_0)^{obs}$ is the observed value of the statistic and $Q_p(\cdot)$ is the distribution function of a chi-square random variable with p degrees of freedom.

The test statistic (4) can be rewritten as $pw_{sp}(\theta) = -2n\hat{K}_w(\lambda(\hat{\theta}_p), \hat{\theta}_p)$, where $\hat{K}_w(\cdot; \cdot)$ is the cumulant generating function of $ps(\cdot; Y)$ under the discrete distribution defined by $\{w_i\}$, with $w_i = w_i(\theta)$. The latter is the discrete distribution which is closest to the empirical one $\{\frac{1}{n}\}$ with respect to the *backward* Kullback- Leibler divergence

$$d_{KL}(\{w_i\}, \{\frac{1}{n}\}) = \sum_{i=1}^{n} w_i \log\left[\frac{w_i}{1/n}\right] = \sum_{i=1}^{n} w_i \log w_i + \log n$$

and which makes $ps(\theta)$ unbiased (see equation (5)). Notice that the use of the *forward* Kullback-Leibler divergence

$$d_{KL}(\{\frac{1}{n}\}, \{w_i\}) = \sum_{i=1}^{n} \frac{1}{n} \log\left[\frac{1/n}{w_i}\right] = -\frac{1}{n}\sum_{i=1}^{n} \log w_i - \log n$$

would lead to the classical empirical log-likelihood ratio test statistic (Owen , 2001) which is also asymptotically χ_p^2 distributed, but which does not enjoy the second-order relative error property of the present test.

Let us now discuss in more details the properties of this test which are summarized in the Introduction.

The new test statistic is asymptotically χ^2 distributed, therefore it is, up to first-order, equivalent to the standard tests but it differs for the following relevant features. Firstly, $pw_{sp}(\theta)$ is asymptotically pivotal and the result does not depend on suitable scaling factors, contrasted to the approximate pivots proposed by Rotnitzky and Jewell (1990), Chandler and Bate (2007), and Pace *et al.* (2011). Secondly, as $pw_{sp}(\theta)$ stems from a small sample asymptotics framework, it introduces an unexplored stream in the pairwise likelihood setting concerning the accuracy of tests statistics. In particular, the exact distribution of our test proposal is χ^2 up to a relative error of magnitude $O(n^{-1})$. This provides an excellent accuracy uniformly in the tails for the approximation obtained by using the asymptotic distribution. Thirdly, the asymptotic approximation can not be enhanced by bootstrap calibration as the actual distribution of $pw_{sp}(\theta)$ and its bootstrap counterpart $pw_{sp}^*(\theta)$ are also distant by a relative error of order $O(n^{-1})$. In contrast, resorting to a computationally expensive resampling procedure is the only viable path either to estimate the quantiles of $pw(\theta)$ without computing the elements of the Godambe information (see, e.g. Aerts and Claeskens, 2001) or to obtain refined estimates of $J(\theta)$ and $H(\theta)$ (see Varin *et al.*, 2011, Section 5.1). Fourthly, the test is fully nonparametric and depends only on the function $ps(\theta; y)$. Therefore, it does not require the specification of the full model $F(\cdot; \theta)$ which is clearly a key issue in this setup (see Section 2). Furthermore, as it solely depends on $ps(\theta; y)$, by choosing the latter bounded with respect to y we can combine accuracy in small samples and resistance with respect to potential outliers; see (Lô and

Ronchetti, 2012) in the GMM framework and the second example in Section 4 below. Finally, $pw_{sp}(\theta)$ enjoys the desirable property of invariance under reparametrization as well as $pw(\theta)$, $pw_s(\theta)$, and $pw_{inv}(\theta)$. However, the latter lose exact invariance once the empirical estimates $\hat{J}(\hat{\theta}_p)$ and $\hat{H}(\hat{\theta}_p)$ are used.

4. Numerical Examples

This section aims at showing some numerical evidence about the behaviour of the nonparametric saddlepoint test statistic in the pairwise likelihood framework. Three examples will be illustrated, each of them enlightening a different feature of the test.

In the first example, the new test is compared to the pairwise likelihood ones presented in Section 2, and their finite sample accuracy to the χ^2 approximations is analized in the context of a multivariate normal model.

In the second and third example, we consider a first-order autoregressive and a geostatistical model, respectively. The purpose of these examples is twofold. In first place we want to point out that the use of bounded estimating functions to compute $pw_{sp}(\theta)$ is recommended not only to provide versions of $pw_{sp}(\theta)$ whose accuracy remains stable under contaminations of the model. Indeed, we will provide empirical evidence that supports, in this setup, the following results outlined in the Appendix: a) $pw_{sp}(\theta)$ converges to the χ^2 distribution and the approximation has a relative error of second order; b) a second order agreement also holds between the asymptotic distribution of $pw_{sp}(\theta)$ and its bootstrap distribution $pw_{sp}^*(\theta)$. In second place, these models provide a challenging setting in which $n = 1 \ll q$ and consequently a suitable definition of the pairwise likelihood function is needed.

In the first two examples the full log-likelihood function $\ell(\theta)$ is available and this allows us to set the log-likelihood ratio test $w(\theta)$ as a benchmark. In the third example this is not possible because the evaluation of the likelihood function is computationally prohibitive.

The statistical environment R (R Core Team, 2012) was used to carry out all the computations in this paper.

a. Multivariate Normal Model

Let Y be a normally distributed random vector, with expectation $(\mu, \ldots, \mu)^{\mathrm{T}} \in \mathbb{R}^q$ and covariance matrix Σ having diagonal elements σ^2 and off-diagonal ones $\sigma^2\rho$, $\rho \in (-1/(q-1), 1)$. The pairwise log-likelihood for the parameter $\theta = (\mu, \sigma^2, \rho)$ is

$$pl(\theta) = -\frac{nq(q-1)}{2}\left[\log\sigma^2 + \frac{\log(1-\rho^2)}{2}\right] - \frac{1}{2\sigma^2(1-\rho^2)}\sum_i (y_i - \mu)^{\mathrm{T}}\Gamma(\theta)(y_i - \mu),$$

with $y_{i\cdot} = \sum_j y_{ij}$, $\Gamma_{jj}(\theta) = (q-1)$, $\Gamma_{jk}(\theta) = -\rho$, $j \neq k = 1, \ldots, q$.

We run simulations by generating 100000 samples of size $n = 10$ from $Y \in \mathbb{R}^{30}$, with $\mu = 0$, $\sigma^2 = 1$, and ρ ranging from moderate to strong correlation values.

For each sample we computed the nonparametric saddlepoint statistic as well as those discussed in Section 2. As for this example, $J(\theta)$ and $H(\theta)$ are available (see, Pace *et al.*, 2011), this allows us to compare also the finite sample behavior among pairwise likelihood

test statistics computed by using the exact matrices and their empirical counterparts $\hat{J}(\theta)$ and $\hat{H}(\theta)$. In the following, the superscript e will refer to statistics evaluated using $J(\theta)$ and $H(\theta)$.

Table 1 reports empirical coverage probabilities for three dimensional confidence regions for θ. As expected, the best results are obtained when the elements of the expected Godambe information are used and, in particular, when one considers $pw_s^e(\theta)$ and $pw_{inv}^e(\theta)$. However, it should be stressed that, in most real applications, only the observed Godambe information is available. In this case, pairwise likelihood statistics have empirical coverages far from the nominal levels. Instead, the bootstrap distribution of the nonparametric saddlepoint test statistic $pw_{sp}^*(\theta)$ is approximated quite well by the χ_3^2 and the approximation is close to the one provided by the gold standard $w(\theta)$. From simulation studies (not reported here) it is shown that confidence sets based on pairwise likelihood statistics achieve the nominal levels either by increasing the sample size or by using resampling-based estimates of $J(\theta)$ and $H(\theta)$.

In order to investigate the reliability of the proposed test and the ones based on pairwise likelihood statistics, it is useful to analyze the shape of the associated confidence sets and to compare them with the one provided by the full log-likelihood ratio. In Fig. 1 we display confidence sets for (σ^2, ρ) with nominal level $1 - \alpha = 0.95$, based on statistics of Table 1, from a simulated sample with $n = 10$, $q = 30$, $\mu = 0$, $\sigma^2 = 1$, and $\rho = 0.9$. For this analysis, the location parameter μ is considered as known. Although all confidence sets cover the true parameter value, the ones provided by $pw_1(\theta)$, $pw_w(\theta)$, and $pw_{cb}(\theta)$ depart remarkably from that of $w(\theta)$. In particular, $pw_1(\theta)$ generates a confidence set that is quite inflated and almost includes the one of $w(\theta)$, whereas Wald-type confidence sets are narrow and elliptically shaped. On the other hand, confidence sets provided by $pw_{sp}(\theta)$, $pw_s(\theta)$, and $pw_{inv}(\theta)$ resemble the gold standard. It is also worth to note how the shape of confidence sets derived from pairwise likelihood statistics is affected by the use of $J(\theta)$, $H(\theta)$ and $\hat{J}(\theta)$, $\hat{H}(\theta)$.

b. Robust First Order Autoregression

We consider a stationary process $\{Y_j\}_{j \in \mathbb{Z}}$, modeled as a first order autoregressive model

$$Y_j = \phi_0 + \phi_1 Y_{j-1} + \epsilon_j, \tag{6}$$

$\phi_0 \in \mathbb{R}$, $\phi_1 \in (-1, 1)$ and ϵ_j independent and normally distributed with mean 0 and variance σ^2. Under these assumptions any trajectory of length q can be thought of as a normal random vector with expectation $(\phi_0/(1 - \phi_1), \ldots, \phi_0/(1 - \phi_1))^{\mathrm{T}} \in \mathbb{R}^q$ 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 here by means of univariate conditional distributions $Y_j|Y_{j-1} = y_{j-1} \sim N(\phi_0 + \phi_1 y_{j-1}, \sigma^2)$, and is:

$$pl(\theta) = -\frac{(q-1)}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{r=2}^{q} (y_r - \phi_0 - \phi_1 y_{r-1})^2. \tag{7}$$

The resulting pairwise score function leads to the ordinary least squares estimate of θ that can be easily robustified by using a Mallows-type estimate for ϕ_0 and ϕ_1 and Huber's

TABLE 1. Multivariate normal model: empirical coverage probabilities of three dimensional confidence regions for $\theta = (\mu, \sigma^2, \rho)$. The superscript e refers to statistics computed by using the elements of the expected Godambe information.

	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.9$		
$1 - \alpha$	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
$pw_{sp}^*(\theta)$	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
$pw_s(\theta)$	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
$pw_s^e(\theta)$	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

Proposal 2 for σ^2. This is obtained by solving the system of estimating equations

$$\sum_{j=2}^{q} \psi_a(r_j) = 0$$
$$\sum_{j=2}^{q} \psi_a(r_j)\psi_b(y_{j-1}) = 0 \tag{8}$$
$$\sum_{j=2}^{q} \psi_c(r_j)^2 - (q-1)\beta(c) = 0,$$

where $r_j = (y_j - \phi_0 - \phi_1 y_{j-1})/\sigma$, $\psi_k(r) = \min\{k, \max(-k, r)\}$, $k > 0$ and $\beta(k)$ is a factor to ensure consistency at the model; see Huber (1981), Huber and Ronchetti (2009).

In order to consider both contaminated and non-contaminated series, we included an additive outlier term in (6), that becomes:

$$Y_j = \phi_0 + \phi_1 Y_{j-1} + \epsilon_j + u_j, \tag{9}$$

where $u_j \sim (1-\xi)\delta_0 + \xi N(\mu_u, \sigma_u^2)$, $\xi \in [0,1]$ and δ_0 is a point mass distribution located at zero.

We performed the simulation study by drawing 100000 series of length $q = 50$ from model (9). We set the true parameter value to have components $\phi_0 = 0$, $\sigma^2 = 1$, and $\phi_1 = \{0.2, 0.5, 0.9\}$ and we generated contaminated series by letting $\xi = 0.05$, $\mu_u = \phi_0/(1-\phi_1)$ and $\sigma_u^2 = 25\sigma^2$. $\xi = 0$ corresponds to the case of non-contaminated series. For each replication we computed the nonparametric saddlepoint test statistic as well as its bootstrap version using the estimating equations in (8). They are denoted by $pw_{sp}(\theta; \gamma)$
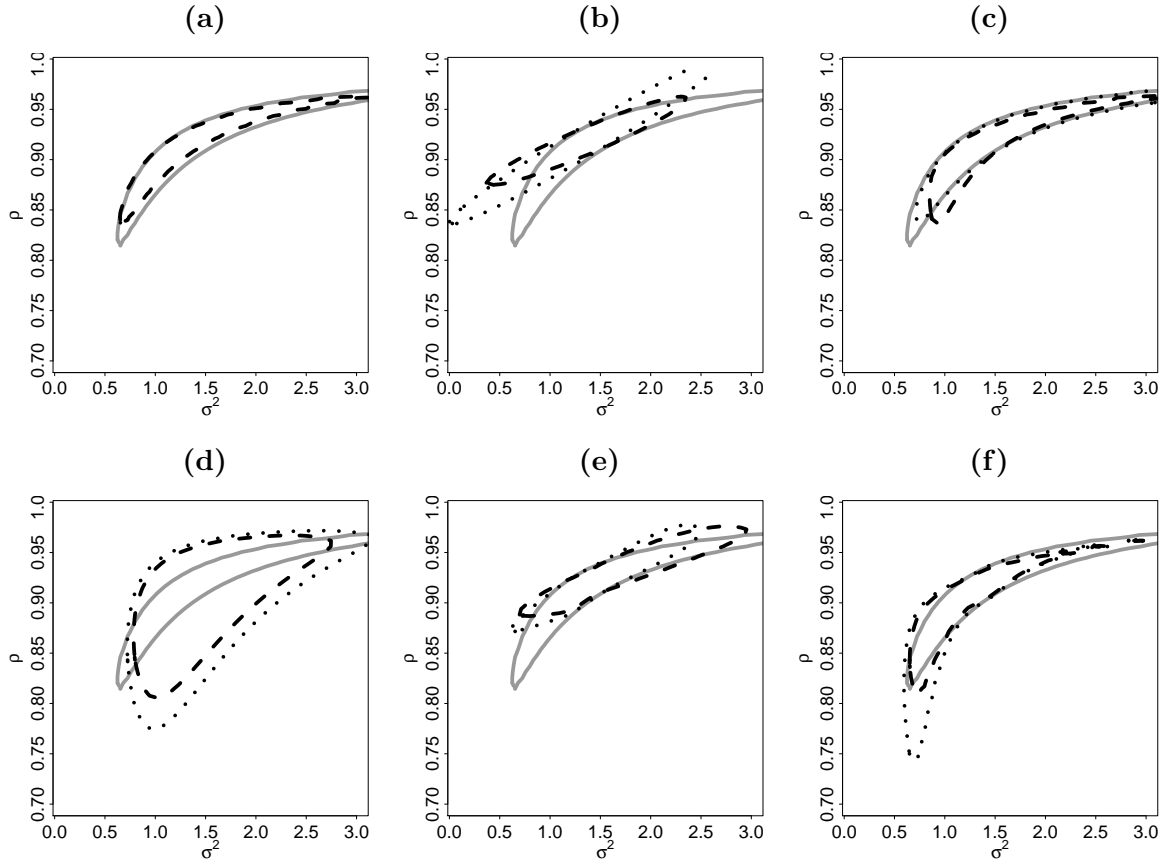
FIG. 1. Multivariate normal model: confidence regions for (σ^2, ρ) with nominal level $1 - \alpha = 0.95$, with known $\mu = 0$ from a simulated sample with $n = 10$ and $q = 30$. In each plot confidence regions in gray solid line is obtained from $w(\theta)$. Confidence regions in dashed and dotted lines derive from pairwise likelihood statistics computed by using $J(\hat{\theta}_p)$ and $H(\hat{\theta}_p)$ and $\hat{J}(\hat{\theta}_p)$ and $\hat{H}(\hat{\theta}_p)$, respectively. In particular: (a) $pw_{sp}^*(\theta)$; (b) $pw_w(\theta)$, $pw_w^e(\theta)$; (c) $pw_s(\theta)$, $pw_s^e(\theta)$; (d) $pw_1(\theta)$, $pw_1^e(\theta)$; (e) $pw_{cb}(\theta)$, $pw_{cb}^e(\theta)$; (f) $pw_{inv}(\theta)$, $pw_{inv}^e(\theta)$

and $pw_{sp}^*(\theta; \gamma)$ respectively, with $\gamma = (a, b, c)$. The choice $\gamma_1 = (1.3, 1.3, 1.3)$ gives a bounded estimating function and leads to a robust estimator with high efficiency at the normal model. The choice $\gamma_2 = (\infty, \infty, \infty)$ defines the classical unbounded estimating function and leads to a non-robust estimator.

It is worth noticing that in order to preserve the dependence structure of the series and to be consistent with the specification of (6), pairs of data points (y_{j-1}, y_j) must be resampled instead of single observations y_j for the evaluation of $pw_{sp}^*(\theta; \gamma)$.

In Table 2 we report empirical coverage probabilities of confidence regions for θ. When $\xi = 0$, the comparison between $pw_{sp}(\theta; \gamma_1)$ and $pw_{sp}(\theta; \gamma_2)$ shows that the use of a bounded estimating function speeds up the convergence to the χ^2 distribution. Moreover, empirical coverages of $pw_{sp}(\theta; \gamma_1)$ and $pw_{sp}^*(\theta; \gamma_1)$ are very close and their accuracy is comparable to the one of the full log-likelihood ratio $w(\theta)$. When contamination occurs, the coverage levels of nonparametric saddlepoint test statistics, computed with a bounded estimating

function, remain quite stable, while those of the log-likelihood ratio and $pw_{sp}(\theta; \gamma_2)$ drop away, as one would expect.

In Fig. 2 we display Q-Q plots for some statistics in Table 2 when $\theta = (0, 0.5, 1)$. The χ^2 approximation for $pw_{sp}(\theta; \gamma_1)$ is quite accurate, even when considering contaminated series, up to $\chi^2_{3;0.99} \approx 11$.

TABLE 2. First order autoregressive model: empirical coverage probabilities of three dimensional confidence regions for $\theta = (\phi_0, \phi_1, \sigma^2)$ by considering non-contaminated ($\xi = 0$) and contaminated series ($\xi = 0.05$).

	$\phi_1 = 0.2$			$\phi_1 = 0.5$			$\phi_1 = 0.9$		
$1 - \alpha$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
	$\xi = 0$								
$w(\theta)$	0.8915	0.9432	0.9876	0.8879	0.9403	0.9873	0.8478	0.9165	0.9792
$pw^*_{sp}(\theta; \gamma_1)$	0.8914	0.9447	0.9892	0.8911	0.9447	0.9892	0.8911	0.9436	0.9881
$pw_{sp}(\theta; \gamma_1)$	0.9007	0.9512	0.9885	0.9007	0.9512	0.9885	0.8946	0.9503	0.9898
$pw_{sp}(\theta; \gamma_2)$	0.8232	0.8822	0.9534	0.8232	0.8822	0.9534	0.7764	0.8548	0.9376
	$\xi = 0.05$								
$w(\theta)$	0.3441	0.3901	0.4641	0.2942	0.3365	0.4034	0.2315	0.2702	0.3236
$pw^*_{sp}(\theta; \gamma_1)$	0.8818	0.9411	0.9877	0.8918	0.9456	0.9873	0.8902	0.9422	0.9869
$pw_{sp}(\theta; \gamma_1)$	0.8921	0.9517	0.9917	0.8976	0.9508	0.9907	0.8728	0.9410	0.9915
$pw_{sp}(\theta; \gamma_2)$	0.4612	0.5413	0.6599	0.3591	0.4328	0.5608	0.2659	0.3215	0.4251

c. Geostatiscal model

Let $\{Y(s), s = (s_1, \ldots, s_q)\}$, be a stationary Gaussian random field with zero mean and exponential covariogram

$$\text{cov}\,[Y(s_j), Y(s_k); \theta] = \sigma^2 \exp\left(-3||h_{jk}||/\phi\right) = \sigma^2 \rho_{jk}(\phi)$$

where, $h_{jk} = (s_j - s_k)$, $j, k = 1, \ldots, q$, $\theta = (\sigma^2, \phi)$, $|| \cdot ||$ is the Euclidean norm. The process is supposed to be observed on a regular lattice and we assume that the sites $s'_j s$ are coordinates in $\mathbb{N}^2$. In the following the discussion is developed in an increasing domain rather than an infill framework (see, e.g., Zhang and Zimmerman , 2005) but this choice does not affect the validity of our results.

The pairwise log-likelihood function for θ is obtained by specifying univariate conditional distributions $Y_j|Y_k = y_k \sim N(\rho_{jk}(\phi)y_k, \sigma^2)$ and is given by

$$pl(\theta) = -\frac{1}{2} \sum_{j=1}^{q} \sum_{\substack{k=1 \\ k \neq j}}^{q} \left\{ \log \sigma^2 + \frac{1}{\sigma^2} \left(y_j - \rho_{jk}(\phi)y_k\right)^2 \right\} \omega(h_{jk}), \tag{10}$$
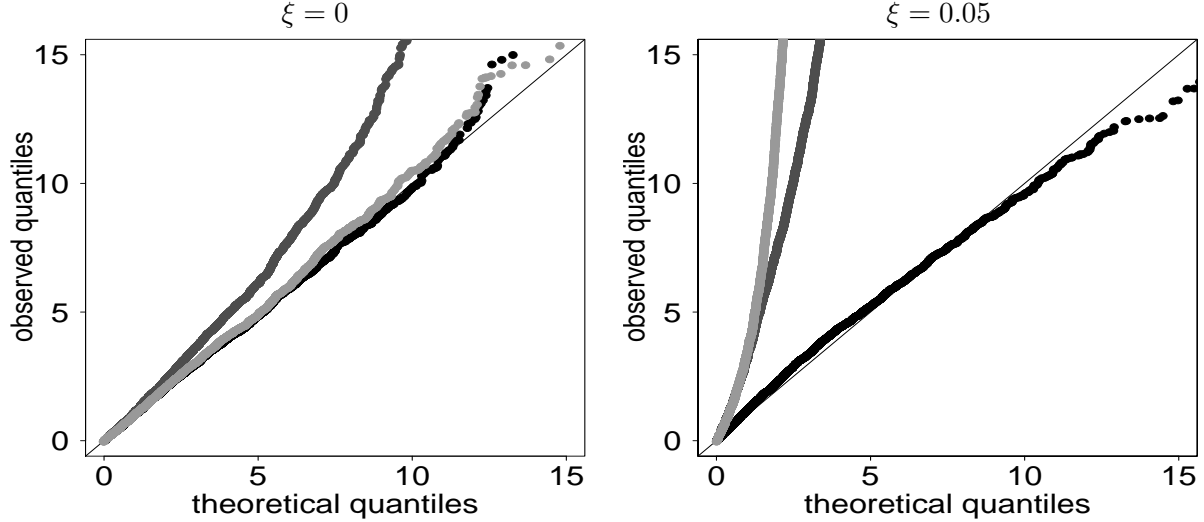
Fig. 2. First order autoregressive model: Q-Q plots for some statistics against theoretical quantiles of the χ_3^2. In black $pw_{sp}(\theta; \gamma_1)$, in dark grey $pw_{sp}(\theta; \gamma_2)$, and in light grey $w(\theta)$

where $y_j = y(s_j)$. The weights $\omega(h_{jk})$ are defined to form a disjoint partition of the sampling region in block of observations. Loosely speaking, the weights are chosen to form $N = [q/(1+l)]^2$ squared blocks B_u, $u = 1, \ldots, N$, each containing $(1+l)^2$ sites, where l is the side length of the square. Inside each block only $(1+l)^2 - 1$ pairs are considered to compute $pl(\theta)$. Therefore, (10) becomes the sum of N pseudo-independent blocks each of them summarizing $(1+l)^2 - 1$ likelihood contributions. In Fig. 3 we display how the blocks and the pairs are defined in a 6×6 sampling region by considering squares with sides of length 1 and 2.

It is worth to point out that the sampling region could be partitioned by constructing overlapping blocks each of them centred on a specific observation, e.g. $B_j = \{(y_j, y_k) : ||h_{jk}|| < d, d > 0, j \neq k = 1, \ldots, q$, and by considering different schemes to form the pairs inside each block. For our purposes the rationale behind the splitting rule is to obtain blocks which are as uncorrelated as possible, this condition being crucial to compute both $pw_{sp}(\theta)$ and a window subsampling estimate for $J(\theta)$.

Also in this example, $pw_{sp}(\theta)$ is computed by using a set of bounded estimating functions. From (10) it is easily seen that the resulting score function for a single pair is

$$
\begin{aligned}
\ell_{\sigma^2}(\theta) &= -\frac{1}{2(\sigma^2)^2}(y_j - \rho_{jk}(\phi)y_k)^2 \\
\ell_{\phi}(\theta) &= \frac{\partial \rho_{jk}(\phi)}{\partial \phi} \frac{1}{\sigma^2}(y_j - \rho_{jk}(\phi)y_k)y_k,
\end{aligned}
\tag{11}
$$

which can be bounded by using the same arguments as in Example b. In particular, we substitute (11) by the third and the second estimating functions in (8), respectively.

Simulations have been run by generating 10000 spatially correlated data from three different scenarios, corresponding to increasing levels of spatial correlation, by setting $\sigma^2 = 1$ and $\phi = \{5, 7, 9\}$. The sampling region $\{1, \ldots, q\} \times \{1, \ldots, q\}$ have been increased
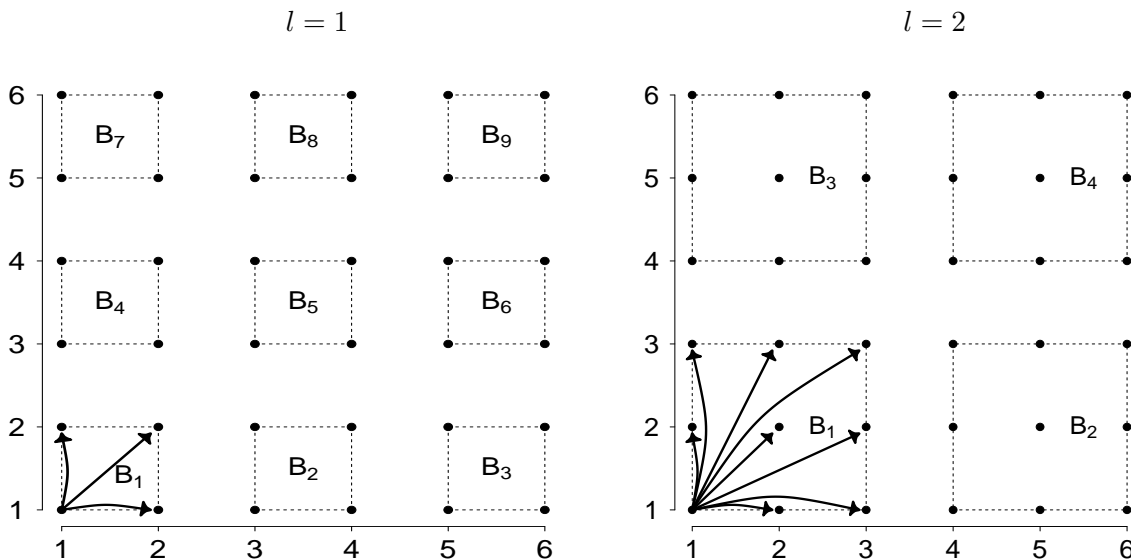
FIG. 3. Partition of a 6×6 sampling region in block of observations. Dashed lines connect observations belonging to a specific block, whereas the arrows indicate which pairs are considered to compute the pairwise likelihood function

accordingly to increasing values of ϕ as well as the side length of the squares defining the blocks. In particular, $q = \{35, 42, 54\}$ and $l = \{5, 7, 9\}$, which means setting l to the effective range, i.e. the distance beyond which the correlation between pairs is less or equal to 0.05. As a guideline we suggest to set l greater or equal to the effective range, and in practical applications this can be obtained by using an empirical estimate of the correlogram.

For each replication we computed the statistics presented in Section 2 as well as $pw_{sp}(\theta)$ by using the bounded counterparts of (11) with $\gamma_1 = (1.3, 1.3, 1.3)$. The full log-likelihood ratio has not been considered in our simulations as its computation is prohibitive for the chosen values of q.

In Fig. 4(a, b, c) we plot the actual sizes against the nominal sizes of tests for the three settings considered. Overall, the actual distribution of $pw_{sp}(\theta; \gamma_1)$ is closer to the χ_2^2 than the ones of the other statistics. In panel (d) of Fig. 4 we display the relative error for the tail area probabilities defined as $(P\left[pw_{sp}(\theta; \gamma) \geq \chi_{2;1-\alpha}^2\right] - \alpha)/\alpha$, for $\alpha \in (0.01, 0.1)$. The plot confirms that the approximation is quite accurate uniformly regardless the strength of the spatial dependence.

5. Concluding Remarks

We introduced in the pairwise likelihood framework a second-order accurate test statistic derived by using saddlepoint techniques. The new test is appealing as it circumvent the specification of the joint density and only requires the availability of the pairwise score function. Moreover, it exhibits several desirable properties which are not shared by the available tests. In particular, it does nor require the availability of the Godambe
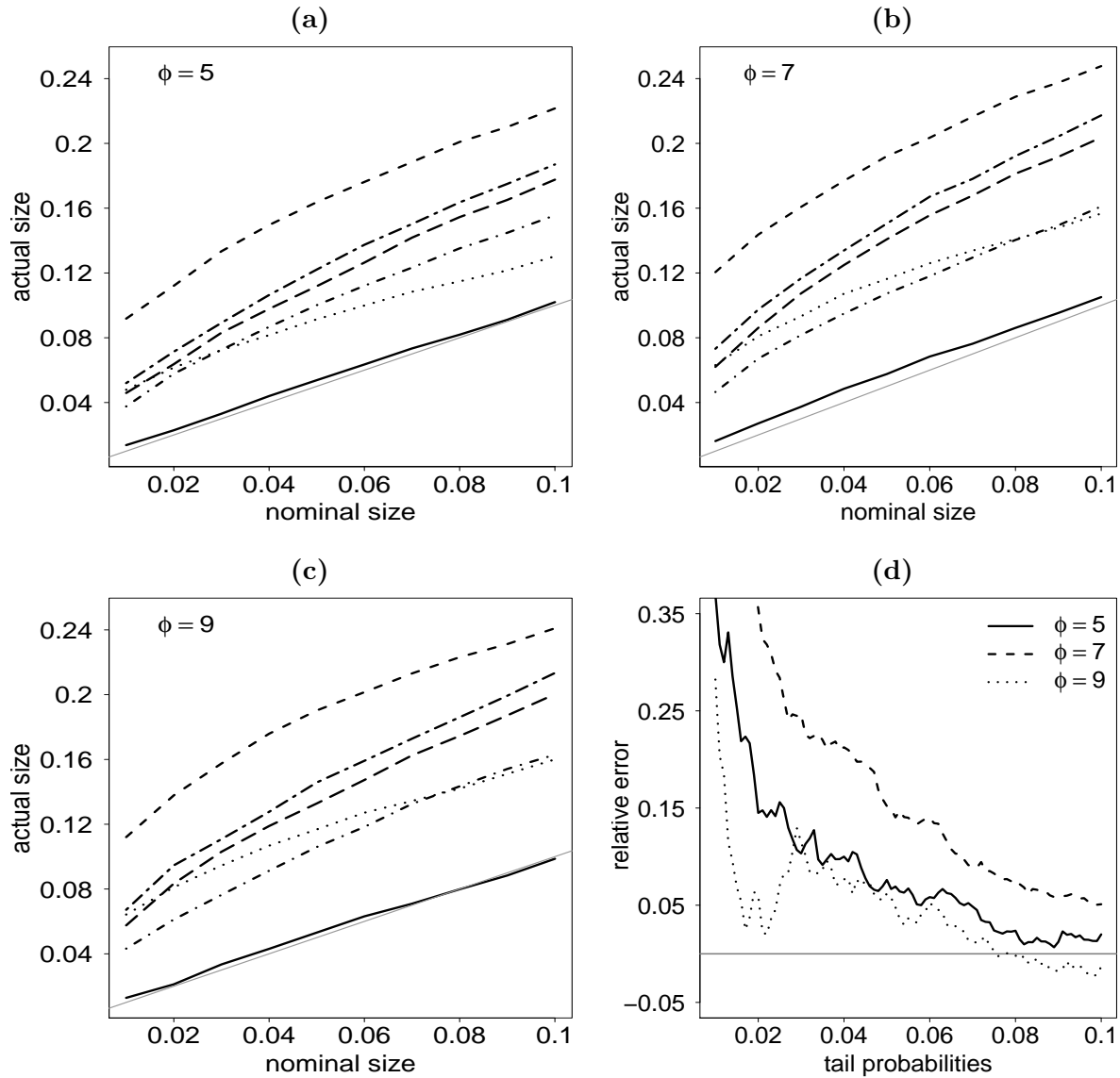
FIG. 4. Geostatistical model: in panel (a), (b), (c) actual size is plotted against nominal size for the following test statistics: $(-)\, pw_{sp}(\theta; \gamma)$, $(\text{- -})\, pw_w(\theta)$, $(\cdots)\, pw_s(\theta)$, $(\cdot - \cdot)\, pw_1(\theta)$, $(- -)\, pw_{cb}(\theta)$, $(\cdot - - \cdot)\, pw_{inv}(\theta)$. In panel (d) approximation of the relative error for tail area probabilities provided by $pw_{sp}(\theta; \gamma)$

information matrix of the full model, which is the case for other standard tests. This opens up the actual possibility to perform small sample asymptotics's inference in rather complex, yet little explored, frameworks.

Acknowledgements

Appendix

Conditions

(A.1): $H(\theta)$ is continuous in θ and $|H(\theta_0)| \neq 0$;

(A.2): The components in $ps(\theta; y)$ as well as their first four derivatives with respect to θ exists and are bounded and continuous;

(A.3): The cumulant generating function of $ps(\theta; Y)$ exists and the distribution function of the random vector $U = (ps(\theta; Y), S(\theta), Q(\theta))$ admits an Edgeworth expansion, where $S(\theta)$ is formed by the elements of $ps(\theta; Y)ps(\theta; Y)^{\mathrm{T}}$ and $\partial ps(\theta; Y)/\partial \theta^{\mathrm{T}}$, whereas $Q(\theta)$ has components $\partial S(\theta)/\partial \theta^{\mathrm{T}}$.

Condition (A.1) essentially ensures that there exists a compact subset of $\mathbb{R}^p$, θ_0 being an interior point of it, in which θ_0 is the unique solution to $\mathbb{E}[ps(\theta)] = 0$. Concerning condition (A.3), the reader may refer to Field *et al.* (2008) for a detailed account of this technical condition.

Proof of Theorem. Let y^* be a bootstrap version of y obtained by sampling according to the set of probabilities $\{w_i(\theta_0)\}$, $\hat{\theta}_p^*$ be the solution to $\sum w_i(\theta_0)ps(\theta; y_i^*) = 0$, and finally denote by $P_w[\cdot]$ the probability under the discrete distribution defined by $\{w_i(\theta_0)\}$. The proof proceeds along the lines of that of Theorem 1 in Ma and Ronchetti (2011) and is splitted into two steps: first the size of the error of the bootstrap p-value $P_w[pw_{sp}^*(\theta_0) \geq pw_{sp}(\theta_0)^{obs}]$ is established, then it is linked to the p-value $P[pw_{sp}(\theta_0) \geq pw_{sp}(\theta_0)^{obs}]$.

From Robinson *et al.* (2003) we have

$$P_w[pw_{sp}^*(\theta_0) \geq pw_{sp}(\theta_0)^{obs}] = [1 - Q_p(pw_{sp}(\theta_0)^{obs})](1 + O(n^{-1})),$$

and from this relation it is easily seen that bootstraping the proposed statistic according to $\{w_i(\theta_0)\}$ leads to a p-value which error size is relative and of second-order. Then, from the results in Field *et al.* (2008) about second-order bootstrap tests, we obtain

$$
\begin{aligned}
P_{H_0}[pw_{sp}(\theta_0) \geq pw_{sp}(\theta_0)^{obs}] &= P_w[pw_{sp}^*(\theta_0) \geq pw_{sp}(\theta_0)^{obs}](1 + O(n^{-1})) \\
&= [1 - Q_p(pw_{sp}(\theta_0)^{obs})](1 + O(n^{-1})),
\end{aligned}
$$

and this proves the theorem.

REFERENCES

Aerts, M. and Claeskens, G. (2001). Bootstrap tests for misspecified models, with application to clustered binary data. *Comput. Statist. Data Anal.*, **36**, 383–401.

Beaumont, M.A., Cornuet, J.-M., Marin, J.-M., Robert, C.P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983–990.

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

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

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

Del Moral, P., Doucet, A., Jasra, A.(2006). Sequential Monte Carlo samplers. *J. Roy. Statist. Soc. B*, **68**, 411–436.

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

Field, C., Robinson, J., Ronchetti, E.(2008). Saddlepoint approximations for multivariate M-estimates with applications to bootstrap accuracy. *Ann. Inst. Statist. Math.*, **60**, 205–224/225–227.

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

Heggland, K., Frigessi, A. (2004). Estimating functions in indirect inference. *J. Roy. Statist. Soc. B*, **66**, 447–462.

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

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

Huber, P., Ronchetti, E. (2009). *Robust Statistics.* 2nd edition, New York: Wiley.

Hudson, R. (2011). Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.

Jiang,W., Turnbull, B. (2004). The indirect method: inference based on intermediate statistics — A synthesis and examples. *Statistical Science*, **19**, 239–263.

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

Lindsay, B. (1988). Composite likelihood methods. *Contemp. Math.*, **80**, 221–240.

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

Lô, S. N., Ronchetti, E. (2012). Robust Small Sample Accurate Inference in Moment Condition Models. *Comput. Statist. Data Anal.*, **56**, 3182-3197.

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

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

Molenberghs, G., Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Owen, A. (2001). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

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

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

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

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ribatet, M., Cooley, D., Davison, A. C. (2011). Bayesian inference for composite likelihood models and an application to spatial extremes. *Statist. Sinica*, (doi: 10.5705/ss. 2009.248).

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

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

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

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

Thibaud, E., Davison, A., Huser, R. (2013). Composite likelihood inference for complex extremes. ENAR Spring Meeting, Orlando (FL).

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

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

Zhang, H., Zimmerman, D. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, **92**, 921–936.