Shrinkage of Variance for Minimum Distance Based Tests *

Saraswata Chaudhuri† and Eric Renault‡

Abstract

This paper promotes information theoretic inference in the context of minimum distance estimation. Various score test statistics differ only through the embedded estimator of the variance of estimating functions. We resort to implied probabilities provided by the constrained maximization of generalized entropy to get a more accurate variance estimator under the null. We document, both by theoretical higher order expansions and by Monte-Carlo evidence, that our improved score tests have better finite-sample size properties. The competitiveness of our non-simulation based method with respect to bootstrap is confirmed in the example of inference on covariance structures previously studied by Horowitz (1998).

Running Head: Variance Shrinkage for Minimum Distance Based Test

JEL Classification: C12; C13; C30

Keywords: Entropy; GMM; Implied probabilities; Score test; Shrinkage

*We wish to thank an anonymous associate editor and two anonymous referees for their comments and suggestions that helped to improve our paper.

†Department of Economics, CB 3305, University of North Carolina, Chapel Hill, NC 27519. Telephone: 919-966-3962. Email: saraswata_chaudhuri@unc.edu.

‡Department of Economics, Box B Brown University, Providence, RI 02912. Telephone: 401-863-3519. Email: eric_renault@brown.edu.

1 Introduction

The optimal minimum distance (OMD) estimator $\hat{\theta}_n$ of a vector θ of p unknown parameters identified by $K \geq p$ constraints

$$\lambda = g(\theta)$$

is the solution of the minimization problem

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \quad n \left(\hat{\lambda}_n - g(\theta)\right)' V_n^{-1} \left(\hat{\lambda}_n - g(\theta)\right), \tag{1}$$

where $\Theta \subset \mathbb{R}^p$ is the parameter space, $\hat{\lambda}_n$ is a $\sqrt{n}$-consistent asymptotically normal estimator with a positive definite asymptotic variance V and V_n is any consistent estimator of V.

The focus of our interest in this paper is the test of a null hypothesis

$$H_0 : \theta = \theta_0.$$

We study the dependence of the finite-sample properties of such a test on the choice of the asymptotic variance estimator V_n. We recommend a shrinkage estimator that leads to over-all superior performance of the test.

By doing so, we contribute to two strands of the econometrics literature.

First, we bring a new application of information theory in econometrics. All our shrinkage estimators are computed by using implied probabilities deduced from the minimization of some generalized entropy. While it has been known since Corcoran (1998) that the Bartlett adjustment derived by DiCiccio et al. (1991) for empirical likelihood does not work for other Cressie-Read discrepancy statistics, we are able to perform an adjustment that is similar in spirit to Bartlett adjustment for any generalized entropy function. The reason for this is the following. By setting the focus on testing of hypotheses, we do not need to bother with estimation of the unknown parameters θ. Our asymptotic theory of higher order improvements provided by information theoretic extensions of Generalized Method of Moments (GMM) is new because, by contrast with the extant literature (see, for example, Newey and Smith (2004), Guggenberger and Smith (2005), etc.), our higher order expansions deal with conditional distributions given that the critical value of a given test is reached.

Second, we bring some new light on alternatives to bootstrap for finite-sample improvements. We do not try to improve the critical value of a test based on a given test statistic but rather to improve the test statistic itself. The goal is to make this statistic as close as possible to the

infeasible one that is based on the known value of the asymptotic variance V of the sample mean of the moment vector. In this respect, our paper can be seen as an extension of the work of Rothenberg (1988) to non-linear settings in over-identified models.

We assume throughout that the consistent estimator $\hat{\lambda}_n$ in (1) is a sample mean of some known functions of the observations. For the sake of notational simplicity, we can write without loss of generality,

$$\hat{\lambda}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In other words, $\hat{\theta}_n$ can also be seen as an efficient GMM estimator associated to the moment conditions

$$E\left[X_i - g(\theta)\right] = 0. \tag{2}$$

The results of this paper could actually be partly extended to general non-separable moment conditions

$$E[\psi(X_i, \theta)] = 0. \tag{3}$$

While this general case is studied in a companion paper Chaudhuri and Renault (2011), we focus here on the specific conclusions that can be drawn from the particular form $\psi(X_i, \theta) := X_i - g(\theta)$ in (2), especially because, under the null hypothesis $H_0 : \theta = \theta_0$, it makes the expected Jacobian matrix of the moment function known

$$E\left[\frac{\partial}{\partial \theta'}\psi(X_i, \theta_0)\right] = -\frac{\partial}{\partial \theta'}g(\theta_0) \quad (=: G \text{ say}). \tag{4}$$

All the results of this paper are based on the maintained assumption that $G := E\left[\frac{\partial}{\partial \theta'}\psi(X_i, \theta_0)\right]$ is known, but does not further use the specific form of $\psi(X_i, \theta) = X_i - \theta$.

We maintain throughout the common assumption for asymptotic distributional theory of OMD estimators that the Jacobian matrix G (under the null) is of full column-rank p. While this Jacobian matrix is key for efficient estimation and score-type tests, its estimation is known to be an important source of poor finite-sample behavior of GMM-based inference due to a perverse correlation of its estimator with the moment function. The problem generally gets worse when the identification of θ_0 is not strong. By contrast, inference in the context of (2) will involve only the estimation of the asymptotic variance matrix V of the sample mean of the moment vector. For instance, the Newey and West (1987) score test of the null hypothesis $\theta = \theta_0$ will be simply

based on the test statistic

$$\xi_n = n\bar{\psi}_n'(V_n)^{-1}G\left(G'(V_n)^{-1}G\right)^{-1}G'(V_n)^{-1}\bar{\psi}_n,$$

where $\bar{\psi}_n := \sum_{i=1}^n \psi_i$ and $\psi_i := \psi(X_i, \theta_0)$ (for notational simplicity). Under the null, the score statistic ξ_n^S will asymptotically follow a $\chi^2(p)$ distribution if V_n is a consistent estimator of the asymptotic variance matrix V. A common practice is to take for V_n a moment-based estimator of V. For instance, in a case with observations without serial correlation, one would typically use a naive estimator like

$$\bar{V}_n^{\text{Center}} := \frac{1}{n}\sum_{i=1}^n(\psi_i - \bar{\psi}_n)\psi_i' \text{ (unconstrained), or } \bar{V}_n := \frac{1}{n}\sum_{i=1}^n \psi_i\psi_i' \text{ (constrained by } H_0).$$

The main thesis of this paper is that, by contrast with this common practice, the size of the score test would be better controlled by using as V_n an estimator of V that is asymptotically efficient under the null hypothesis. We will see that such an efficient estimation of the variance matrix V amounts to a shrinkage of the naive estimator that takes into account the information content of the moment conditions implied by the null hypothesis. The rationale for efficiently estimating V is that we try to mimic the behavior of the infeasible test statistic that would use the knowledge of the unknown asymptotic variance V, i.e., the statistic,

$$\xi_n^{\text{Infeas}} = n\bar{\psi}_n'V^{-1}G\left(G'^{-1}V^{-1}G\right)^{-1}G'^{-1}V^{-1}\bar{\psi}_n.$$

With this goal in mind, our methodology for comparing competing tests will be twofold. On the one hand, we will provide some compelling Monte-Carlo evidence that the targeted $\chi^2(p)$ distribution is better tracked by our proposed test statistics with variance shrinkage than by standard test statistics. On the other hand, we will display some rationale for this evidence through the asymptotic expansions of the distribution functions of the various test statistics. We refer to Cavanagh (1983) for technical results used for such expansions. We will typically show that we get with variance shrinkage a more accurate approximation of the distribution function of the infeasible score test statistic.

The reported Monte-Carlo illustration considers the estimation of covariance structures as commonly met in a variety of economic examples. Abowd and Card (1989) and Altonji and Segal (1996) have documented the poor finite-sample properties of OMD estimators and inference in this context. Horowitz (1998) have put forward bootstrap methods for finite-sample improvements.

It is worth realizing that the shrinkage strategy studied in this paper is not aimed at replacing bootstrap. First, it proposes a simple and user-friendly way to improve finite-sample size properties of score tests, without resorting to any simulation. Second, it could well be coupled with the bootstrap methods if necessary. Our approach is actually quite close in spirit to the bootstrapping for GMM as devised by Brown and Newey (2002). As in their work, we take advantage of the probabilities implied by the moment conditions (under the null hypothesis) for a proper re-weighting of the observations at hands. While they do that for the purpose of re-sampling, we just do it to find the efficient estimator V_n of the asymptotic variance matrix V.

The paper is organized as follows. In Section 2, we discuss the efficient estimators of the asymptotic variance matrix that are provided by a generalized maximum entropy approach to the moment conditions under the null. In Section 3, score statistics are compared through asymptotic expansions of their distribution functions. An extensive Monte-Carlo illustration is provided in Section 4 in the context of OMD inference on covariance structures. Our approach appears in many cases to be competitive with the more involved bootstrap approach. Section 5 concludes. All the proofs are collected in a technical appendix.

2 Efficient estimation of the variance matrix under the null hypothesis

The information theoretic approaches to inference in moment condition models have become popular in econometrics since the seminal papers by Kitamura and Stutzer (1997) and Imbens et al. (1998). The idea in the context of general moment conditions (3) is to look simultaneously for an estimator $\hat{\theta}_n^{\gamma}$ of θ and for the implied probabilities $\hat{\pi}_n^{(\gamma)} = (\hat{\pi}_{i,n}^{(\gamma)})_{1 \leq i \leq n}$ as solutions of

$$\min_{\theta \in \Theta, \pi} \frac{1}{\gamma(\gamma-1)} \sum_{i=1}^{n} \left[(n\pi_i)^{1-\gamma} - 1 \right] \tag{5}$$
$$\text{subject to} \quad \sum_{i=1}^{n} \pi_i = 1 \text{ and } \sum_{i=1}^{n} \pi_i \psi(X_i, \theta) = 0.$$

The objective function (5) is defined for any real γ, including the two limit cases $\gamma \to 0$ and $\gamma \to 1$. The family of these functions, indexed by γ, is generally referred to as the Cressie-Read family of power divergence statistics (see Kitamura and Stutzer (1997), Imbens et al. (1998) and the references therein). It is known that in the case of i.i.d. observations $X_i, i = 1, ...n$, and under standard regularity conditions, the estimator $\hat{\theta}_n^{\gamma}$ is asymptotically efficient (and asymptotically equivalent to efficient GMM) for any value of γ. In case of serially dependent observations, this

result can be extended by applying the above power divergence minimization to properly pre-averaged moments in the spirit of Kitamura and Stutzer (1997). All what is done in the following could be extended like that to time series models but will not be stated explicitly for the sake of expositional simplicity.

It is generally believed that implied probabilities are relevant for inference only in the case of over-identified moment conditions since, when $K = p$, one may generically find a method of moments estimator $\hat{\theta}_n$ such that

$$\sum_{i=1}^{n} \psi(X_i, \hat{\theta}_n) = 0$$
$$\text{and then,} \qquad \hat{\pi}_{i,n}^{(\gamma)} = \frac{1}{n}, \forall\ i = 1, ..., n,\ \forall\ \gamma \in \mathbb{R}. \tag{6}$$

Our use of implied probabilities in this paper is new since we want to devise the proper shrinkage implied by the null hypothesis $\theta = \theta_0$. In other words, in the context of separable moment conditions (2), we define the implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}$ as the solutions of

$$\min_{\pi} \frac{1}{\gamma(\gamma - 1)} \sum_{i=1}^{n} \left[(n\pi_i)^{1-\gamma} - 1 \right] \tag{7}$$
$$\text{subject to} \qquad \sum_{i=1}^{n} \pi_i = 1 \text{ and } \sum_{i=1}^{n} \pi_i \psi(X_i, \theta_0) = 0. \tag{8}$$

As a consequence, even in the just-identified case, implied probabilities do not coincide with the empirical distribution (6) because the null hypothesis is not exactly fulfilled with sample moments. The consistent estimators V_n of the variance matrix we promote in this paper are the ones associated to these implied probabilities,

$$V_n^{(\gamma)} = \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(\gamma)} \psi_i \psi_i'. \tag{9}$$

It is worth comparing the estimators $V_n^{(\gamma)}$ (for any choice of the power-divergence parameter γ) with the naive estimation principle based on the empirical probabilities (1/n) (and mentioned in the Introduction) that, under the null, i.e., when working with the constrained "estimator" would lead to consider

$$\bar{V}_n = \frac{1}{n} \sum_{i=1}^{n} \psi_i \psi_i'. \tag{10}$$

The key difference between (9) and (10) is that we have replaced the empirical distribution (6) by implied probabilities which make sure that the moment conditions (with the value of θ under

the null) are fulfilled in the sample. In yet other words, we have shrunk the variance estimator to take advantage of the information brought by the null hypothesis $H_0 : \theta = \theta_0$. This strategy is germane to pooling data to take advantage of invariance of parameter values across different samples (see, for example, Ziemer and Wetzstein (1983)). The shrinkage interpretation will be confirmed by the computation of the implied probabilities below.

Let us also note that this can be related to a point already made by Hall (2000) who recommends that variances be calculated using the data in mean deviation form for improved power properties of over-identification tests. It is precisely a way to acknowledge that the naive estimator must be shrunk in due proportion of the in-sample violation of the moment conditions. Hall's shrinkage would simply lead to replace $\bar{V}_n$ by

$$\bar{V}_n^{\text{Center}} = \frac{1}{n}\sum_{i=1}^{n} \psi_i \psi_i' - \bar{\psi}_n \bar{\psi}_n'. \tag{11}$$

However, under the null, $\bar{V}_n$ and $\bar{V}_n^{\text{Center}}$ are asymptotically equivalent estimators of V :

$$\sqrt{n}\left(\bar{V}_n - \bar{V}_n^{\text{Center}}\right) = \sqrt{n}\bar{\psi}_n \bar{\psi}_n' = o_P(1). \tag{12}$$

By contrast with (11), the shrinkage (9) makes an efficient use of the information content of the moment conditions. To see this, first note that the first order conditions of the minimization in (7) subject to the constraints in (8) gives, for a non-zero γ,

$$\begin{aligned} \hat{\pi}_{i,n}^{(\gamma)} &\propto \left[1 + \beta_\gamma' \psi_i\right]^{-1/\gamma} \\ &= 1 - \frac{1}{\gamma}\beta_\gamma' \psi_i + o_p(1/\sqrt{n}) \end{aligned} \tag{13}$$

where $\propto$ means "proportional to" and β_γ stands for a vector of re-scaled Lagrange multipliers. Note that, for the sake of expositional simplicity, we exclude the limit case $\gamma = 1$ which corresponds to the Kullback-Leibler Information Criterion estimator put forward by Kitamura and Stutzer (1997).

Since for any value of γ, the sequences $\sqrt{n}\beta_\gamma$ are asymptotically normal and asymptotically equivalent (see, for example, Imbens et al. (1998)), it is worth interpreting the implied probabilities in the particular case $\gamma = -1$, which corresponds to the Euclidean Empirical Likelihood (EEL) that is extensively documented in Antoine et al. (2007). However, it must be kept in mind that the score test extensively studied in the present paper, based on the estimator $V_n^{(-1)}$ for the variance matrix, is not the score test associated to EEL or used by Kleibergen (2005). Indeed, the first

order conditions of EEL, that deliver an estimator of θ numerically equal to the continuously updated GMM estimator of Hansen et al. (1996) do not resort to the efficient estimator for the variance matrix. It is actually the reason why Antoine et al. (2007) had proposed the 3-step EEL. The first two steps are not needed here since θ is known under the null.

The advantage of the Euclidean case $\gamma = -1$ is that the Taylor expansion (13) is actually exact, so that we get closed form formulas for the Lagrange multipliers and the implied probabilities:

$$\begin{aligned} \hat{\pi}_{i,n}^{(-1)} &\propto 1 + \beta'_{-1}\psi_i \\ \Rightarrow \hat{\pi}_{i,n}^{(-1)} &= \frac{1}{n} - \frac{1}{n}\bar{\psi}'_n \left(\bar{V}_n\right)^{-1} (\psi_i - \bar{\psi}_n). \end{aligned}$$

This closed form formula allows Antoine et al. (2007) to give a control variable interpretation of the constrained estimator of the expectation of any integrable function of the variables ψ_i.

Under the null, if the expectation with respect to the empirical distribution in (6) is denoted by $\hat{E}$ and that with respect to the implied probabilities $\hat{\pi}_{i,n}^{(\gamma)}$ by $\hat{E}^{(\gamma)}$, we get, for any scalar function $h(X_1)$:

$$\begin{aligned} \hat{E}^{(-1)}[h(X_1)] &= \sum_{i=1}^{n} \hat{\pi}_i^{-1} h(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} h(X_i) - \frac{1}{n}\bar{\psi}'_n(\bar{V}_n)^{-1}\sum_{i=1}^{n}\psi_i h(X_i) \\ &= \hat{E}[h(X_1)] - \hat{E}[\psi_1](\bar{V}_n)^{-1}\hat{E}[\psi_1 h(X_1)]. \end{aligned}$$

Therefore, in order to estimate $E[h(X_1)]$, we compute the sample mean of the residual of the regression of $h(X_i)$ on ψ_i. The control variable principle tells us that it is an efficient way to estimate $E[h(X_1)]$ while taking into account the information that $E[\psi_1] = 0$. In other words, as rigorously proved in Antoine et al. (2007), the estimator $\hat{E}^{(-1)}[h(X_1)]$ reaches the semi-parametric efficiency bound for the estimation of $E[h(X_1)]$ under the null hypothesis $H_0 : \theta = \theta_0$. The aforementioned first order equivalence implies that the semi-parametric efficiency bound is also reached under the null by a bunch of constrained estimators, associated to any value of the power γ:

$$\hat{E}^{(\gamma)}[h(X_1)] = \sum_{i=1}^{n} \hat{\pi}_i^{(\gamma)} h(X_i).$$

The focus of our interest in this paper will be the use, for the purpose of score testing, of two constrained estimators $V_n^{(\gamma)}$ of the variance matrix associated respectively to the values $\gamma = -1$ and $\gamma = 0$. As explained above, the use of $V_n^{(-1)}$ is in the line of score testing in the context of 3-

step EEL. The use of $V_n^{(0)}$ is in the line of score testing with Empirical Likelihood (EL) as studied by Guggenberger and Smith (2005) since the minimization of (5) in the limit case $\gamma \to 0$ amounts to the maximization of the empirical likelihood $\sum_{i=1}^{n} \log(\pi_i)$. The control variable interpretation above allows us to interpret these constrained estimators of the variance matrix as a result of a kind of shrinkage, that is replacing the cross products of components of $\psi_i := X_i - g(\theta_0)$ by the residual of their regression on the moment function. This interpretation is exact in the EEL case and asymptotic in the EL case (and all Cressie-Read cases as well).

Note also that the implied probabilities in the EEL case may take negative values in finite samples. It may be an issue for positive definite estimation of the variance matrix. Antoine et al. (2007) have proposed an additional shrinkage step to get rid of this non-positivity issue. They consider instead the following implied probabilities

$$\hat{\pi}_i^{(-1,p)} = \frac{1}{1+\varepsilon_n}\hat{\pi}_i^{(-1)} + \frac{\varepsilon_n}{1+\varepsilon_n}.\frac{1}{n}, \text{ where } \varepsilon_n = -n \times \min\left\{\min_{1\leq i\leq n} \hat{\pi}_i^{(-1)}, 0\right\}.$$

They show that this additional shrinkage does not prevent from reaching the semi-parametric efficiency bound under the null. We will denote:

$$V_n^{(-1,p)} = \sum_{i=1}^{n} \hat{\pi}_{i,n}^{(-1,p)} \psi_i \psi_i'.$$

Irrespective of the use of implied probabilities $\hat{\pi}_i^{(-1)}, \hat{\pi}_i^{(0)}$ or $\hat{\pi}_i^{(-1,p)}$, we end up with an estimator of the variance matrix that is asymptotically equivalent under the null to

$$V_n^{(-1)} = \bar{V}_n - \left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array}\right] V^{-1} \bar{\psi}_n \tag{14}$$

where, for $l = 1, 2, ..., K$, Ω_l is the square symmetric matrix of size K with coefficients $Cov\left[\psi_{ih}\psi_{ik}, \psi_{il}\right]$, for $h, k = 1, 2, ..., K$.

Of course, in practice, $V_n^{(-1)}$ is infeasible and the population moments Ω_l, for $l = 1, \ldots, K$, and V should be replaced by their sample counterparts. However, all these estimators would be asymptotically equivalent and, therefore, their differences are immaterial for us, as will be shown explicitly in the next section. Note that, as residuals of regressions on the moment function, these estimators are actually shrunk by comparison with the naive sample variance. Also see Brown and Newey (1998) and Antoine et al. (2007) for related discussion in the context of semiparametric efficiency.

We refer the reader to the companion paper Chaudhuri and Renault (2011) for the statement

of the regularity conditions that make all Cressie-Read variance estimators asymptotically equivalent. In particular, it may require the existence of the eight moment of ψ_i. In this paper, we maintain the assumption of asymptotic equivalence as a high level assumption:

Asymptotic equivalence of Cressie-Read Estimators:

$$\sqrt{n}(V_n^{(-1)} - V_n^{(\gamma)}) = o_P(1) = \sqrt{n}(V_n^{(-1)} - V_n^{(-1,p)}), \forall\ \gamma \in\ \mathbb{R} \setminus \{1\}. \tag{15}$$

A comparison of (14) and (15) clearly shows that the proposed improvement for estimation of the asymptotic variance matrix will matter when the moment function displays some kind of multivariate skewness. Of course, as exemplified by the recent literature on heteroskedasticity and autocorrelation consistent (HAC) estimation (see, for example, Andrews (1991), Sun et al. (2008), etc.) improving the variance estimator does not necessarily improve the finite-sample performance of inference. However, the next section of the paper will confirm that our proposed (i.e., the control-variable) improvement matters for testing of hypotheses. Even though the focus of our interest is more on size of the test, we suspect (see also Chaudhuri and Renault (2011)) that finite-sample improvements would also be noteworthy for power. It must be kept in mind that by contrast with Hall (2000), our focus of interest is not power of over-identification tests but on improved inference on the "structural" parameters.

3 Theoretical Analysis

The distribution of the score test statistic ξ_n^S defined in the Introduction obviously depends upon the choice of an estimator V_n of the unknown variance matrix V of the moment function ψ_i.

The score statistic, as a function of V_n, is defined as

$$\xi_n(V_n) := \quad n\bar{\psi}_n' V_n^{-1} G(G' V_n^{-1} G)^{-1} G' V_n^{-1} \bar{\psi}_n = t_n(V_n)' t_n(V_n) \text{ where} \tag{16}$$

$$t_n(V_n) := \quad (G' V_n^{-1} G)^{-1/2} G' V_n^{-1} \sqrt{n}\bar{\psi}_n. \tag{17}$$

The dependence on V_n is made explicit because the theme of the paper is the choice of V_n. We will assume throughout that our estimator sequence (V_n) is such that:

Assumption 1:

(i) For all n sufficiently large, with probability one V_n is a positive definite matrix of size K.

(ii) The sequence $Z_n(V_n) := \sqrt{n}\left(\bar{\psi}_n', Vec(V_n - V)'\right)' = (Z_{1n}', Z_{2n}(V_n)')'$ is asymptotically normal under the null hypothesis.

Note that this assumption will be fulfilled for all our estimators of interest when V is positive definite and a central limit theorem is valid for $(\bar{\psi}_n', Vec(\bar{V}_n)')$.

For our results, we will actually need to maintain a stronger assumption. To see that, it is useful to introduce the following two moment functions that define what Sowell (1996) had dubbed respectively the identifying and the over-identifying restrictions:

$$\begin{aligned}\sqrt{n}\bar{\psi}_n^I &:= P(G)\sqrt{n}\bar{\psi}_n = P(G)Z_{1n}, \text{ where } P(G) := G(G'V^{-1}G)^{-1}G'V^{-1},\\ \sqrt{n}\bar{\psi}_n^O &= M(G)\sqrt{n}\bar{\psi}_n = M(G)Z_{1n}, \text{ where } M(G) = I_K - G(G'V^{-1}G)^{-1}G'V^{-1}.\end{aligned}$$

Note that the infeasible score test statistic is nothing but

$$\xi_n^{\text{Infeas}} = \xi_n(V) = (\sqrt{n}\bar{\psi}_n^I)'V^{-1}(\sqrt{n}\bar{\psi}_n^I).$$

Hence the critical value of the score test will be defined by a quantile of $(P(G)Z_1)'V^{-1}(P(G)Z_1)$ where Z_1 is a Gaussian vector equal in distribution to the limit distribution of Z_{1n} under the null. For the sake of higher order asymptotic assessment of the size of the test, we will need to maintain an additional regularity condition for such a quantile.

Assumption CLT(a): The following hold for some given $a \in \mathbb{R}^K$:

(i) V is a positive definite matrix of size K.

(ii) The sequence $Z_n = (Z_{1n}', Z_{2n}')' := Z_n(\bar{V}_n) = (Z_{1n}', Z_{2n}'(\bar{V}_n))'$ is asymptotically normal under the null hypothesis, with an asymptotic distribution given by a Gaussian vector $(Z_1', Z_2')'$.

(iii) $\lim_{n=\infty} E[Z_{1n}|P(G)Z_{1n} = a] = a$, $\lim_{n=\infty} E[\tilde{Z}_{2n}|P(G)Z_{1n} = a] = E[\tilde{Z}_2|P(G)Z_1 = a]$ and $\lim_{n=\infty} E[(\tilde{Z}_{2n} - CZ_{1n})V^{-1}Z_{1n}|P(G)Z_{1n} = a] = 0$ with $Vec(\tilde{Z}_{2n}) = Z_{2n}$, $Vec(\tilde{Z}_2) = Z_2$ and $C = \left[\begin{array}{ccc} \Omega_1 & ... & \Omega_K \end{array}\right] V^{-1}$ where, for $l = 1, 2, ..., K$, Ω_l is the square symmetric matrix of size K with coefficients $Cov\left[\psi_{ih}\psi_{ik}, \psi_{il}\right]$, for $h, k = 1, 2, ..., K$.

Note that the assumption CLT(a) supersedes assumption 1. Besides the central limit theorem, we need convergence of a few additional moments $E[Z_{1n}|P(G)Z_{1n} = a]$, $E[\tilde{Z}_{2n}|P(G)Z_{1n} = a]$ and $E[(\tilde{Z}_{2n} - CZ_{1n})V^{-1}Z_{1n}|P(G)Z_{1n} = a]$. The last one is actually required only in the over-identified case. The first and last limit values, a and 0, respectively are implied by the fact that

$$E[Z_1|P(G)Z_1 = a] = E[P(G)Z_1|P(G)Z_1 = a] + E[M(G)Z_1|P(G)Z_1 = a] = a + 0,$$

and that $Z_2 - CZ_1$ is independent of Z_1 since Z_1 and Z_2 are jointly Gaussian.

Goggin (1994) gives some sufficient conditions for convergence in distribution of $E[h(Z_{1n}, Z_{2n})\,|P(G)Z_{1n}]$ towards $E[h(Z_1, Z_2)\,|P(G)Z_1]$. However this convergence in distribution does not imply almost sure convergence. This is the reason why we do not want to assume the validity of the limits for almost all $a \in \mathbb{R}^H$ but only for some given quantile at play in the proposed score test procedure.

Our assumption is more akin to assuming the following. First, the sequence of conditional distributions of (Z'_{1n}, Z'_{2n}) given $P(G)Z_{1n} = a$ converges weakly towards the normal conditional distribution of (Z'_1, Z'_2) given $P(G)Z_1 = a$. Second, some specific moments converge accordingly.

The required convergence in distribution is germane to an assumption of stable convergence in law (see, for example, Jacod and Shiryaev (2003)). The convergence of the specific moments would come with suitable uniform integrability conditions.

The quantile a of interest will actually be defined as $a = G(G'V^{-1}G)^{-1/2}\tau$ from a quantile τ of t_n, vector of coefficients of columns of $G(G'V^{-1}G)^{-1/2}$ in $P(G)Z_{1n}$:

$$P(G)Z_{1n} = G(G'V^{-1}G)^{-1/2}t_n, \text{ where } t_n = t_n(V) = (G'V^{-1}G)^{-1/2}G'V^{-1}Z_{1n}.$$

Of course, in practice and in this paper, τ will be defined as the quantile of the distribution of a standard normal vector.

The key idea is to use an approximation procedure derived in Cavanagh (1983) (Lemma A1, Chapter 2) (also see equations A1 and A3 in Rothenberg (1988)). We will start from an expansion:

$$t_n(V_n) = t_n + \frac{B_n}{\sqrt{n}} + \mathcal{O}_p\left(\frac{1}{n}\right)$$

and use Cavanagh's result to claim that $t_n(V_n)$ admits the same Edgeworth expansion to order $(1/\sqrt{n})$ as the variable

$$t_n^*(V_n) = t_n + \frac{E[B_n\,|t_n]}{\sqrt{n}}.$$

For the sake of expositional simplicity, it is convenient to set the focus on the case where $t_n(V_n) = t_n$ is a real random variable, that is $\dim(\theta) = p = 1$. This condition will be maintained throughout even though the approach is more generally valid, at the price of heavier notations. Note that the one-parameter setting allows us to even consider one-sided alternatives so that the focus of our interest will be an asymptotic approximation of probabilities like $P[t_n(V_n) \leq \tau]$. Using Cavanagh-Rothenberg-type approximation, we will get an expansion:

$$P[t_n(V_n) \leq \tau] = P[t_n \leq \tau] - \frac{E[B_n\,|t_n = \tau]}{\sqrt{n}} f_n(\tau) + O(1/n)$$

where $f_n(.)$ stands for the probability density function of t_n.

We first prove the following lemma.

Lemma 3.1 *Under assumption 1:*

$$t_n(V_n) \quad = \quad t_n + \frac{B_n(V_n)}{\sqrt{n}} + \mathcal{O}_p\left(\frac{1}{n}\right)$$

with:

$$B_n(V_n) \quad = \quad B_{1n}(V_n) + B_{2n}(V_n)$$

where $B_{1n}(V_n)$ and $B_{2n}(V_n)$ are such that

$$\begin{aligned} GB_{1n}(V_n) &= -\frac{1}{2}P(G)\tilde{Z}_{2n}(V_n)V^{-1}P(G)Z_{1n}\left(G'^{-1}G\right)^{1/2}, \\ GB_{2n}(V_n) &= -P(G)\tilde{Z}_{2n}(V_n)V^{-1}M(G)Z_{1n}\left(G'^{-1}G\right)^{1/2}, \end{aligned}$$

and $\tilde{Z}_{2n}(V_n)$ stands for the K-dimensional random square matrix such that $Vec(\tilde{Z}_{2n}(V_n)) = Z_{2n}(V_n)$.

Note that Cavanagh's approach can be applied since $B_n(V_n)$ is a smooth function of the asymptotically Gaussian vector $Z_n(V_n)$. It depends on our estimator V_n of V through the random matrix $\tilde{Z}_{2n}(V_n)$. The focus of our interest is to devise choices of V_n such that the first term in the expansion is equal to zero, that is $E[B_n(V_n)\,|t_n = \tau] = o(1)$. In order to do that, we will maintain the following assumption.

Assumption Edg (τ)**:**

(i) Assumption CLT(a) holds for $a = G(G'V^{-1}G)^{-1/2}\tau$.

(ii) For $B_n(V_n)$ given by Lemma 3.1,

$$P[t_n(V_n) \leq \tau] = P[t_n \leq \tau] - \frac{E[B_n(V_n)\,|t_n = \tau]}{\sqrt{n}} f_n(\tau) + O(1/n).$$

(iii) If $\psi_i = (\psi_{i1}, ..., \psi_{iK})'$, any sequence $(\psi_{ih}\psi_{ik}\psi_{il})_{i\in\mathbb{N}}$, fulfills a weak law of large numbers for all $h, k, l \in \{1, ..., K\}$.

We are then able to prove our main result:

Theorem 3.2 *Under assumptions 1 and Edg(τ), and under the null hypothesis H_0 :*

(i) If $V_n = \bar{V}_n$ or $V_n = \bar{V}_n^{Center}$:

$$E[B_{1n}\,|t_n = \tau] = -\frac{\tau^2}{2(G'V^{-1}G)^{3/2}} G'V^{-1}\left\{\left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array}\right] V^{-1}G\right\} V^{-1}G + o(1)$$

where, for $l = 1, 2, ..., K$, Ω_l is the square symmetric matrix of size K with coefficients $Cov\,[\psi_{ih}\psi_{ik}, \psi_{il}]$, for $h, k = 1, 2, ..., K$.

(ii) If $V_n = V_n^{(-1)}$:

$$E[B_n(V_n)\,|t_n = \tau] = o(1).$$

(iii) More generally for the Cressie-Read family:

$$E[B_n(V_n^{(\gamma)})\,|t_n = \tau] = o(1), \forall\ \gamma \in\ \mathbb{R} \setminus \{1\}.$$

Remarks:

(i) It is obvious from the expression of the matrices Ω_l for $l = 1, \ldots, K$ that what makes the first order bias $E[B_n(V_n)\,|t_n = \tau]$ non negligible in general when $V_n = \bar{V}_n$ or $V_n = \bar{V}_n^{\text{Center}}$ is the non-zero (multivariate) skewness of the moment function. The intuition is actually quite clear from the formula of the bias in Lemma 3.1. In both cases, B_{1n} and B_{2n}, the bias comes from the asymptotic correlation between the error $\tilde{Z}_{2n}$ in the estimation of the variance matrix V and the moment function $Z_{1n} = \sqrt{n}\bar{\psi}_n$. This correlation is typically akin to multivariate skewness in the moment function.

(ii) In the just-identified case $M(G) = 0$ and hence $B_{2n}(V_n) = 0$ for all choices of V_n. In addition, since $p = 1$, $E[B_{1n}(\bar{V}_n)|t_n = \tau] = -\frac{1}{2}E[\psi_i^3]\tau^2 + o(1)$. It is this distortion due to the skewness of the moment vector under H_0 that is being refined when one instead uses $V_n = V_n^{(-1)}$ or $V_n = V_n^{(-1,p)}$ or equivalently, up to the same order, some Cressie-Read family estimator $V_n^{(\gamma)}$. Of course, as stated above, refinements are also obtained in over-identified cases. The key intuition is provided by Lemma 3.1. jointly with formula (14) above. The proposed improvement in variance estimation based on the residuals of a regression on the moment function (according to the control variable interpretation) has gotten rid of the perverse correlation that is produced by multivariate skewness.

(iii) We stress that such refinements do not correct for all the skewness-related errors of approximation of the exact distribution of the test statistic. For example, in the case $p = K = 1$, as can be seen from a formal Edgeworth expansion of the infeasible statistic t_n, the effect of skewness is still present in the first-order approximation error of its exact distribution. To see

this, note that under the rotation such that $E[\psi^2] = 1$ and the assumptions $E[|\psi|^3] < \infty$, and $\sup_{|s|\geq\epsilon} |E[\exp(\iota s\psi)]| < 1$ for all $\epsilon > 0$ (Cramer's condition) with $\iota = \sqrt{-1}$:

$$P(t_n \leq \tau) = \Phi(\tau) + \frac{1}{6\sqrt{n}} E[\psi^3](2\tau^2 + 1)\phi(\tau) + o_p(1/\sqrt{n})$$

where $\phi(\tau)$ and $\Phi(\tau)$ are respectively the pdf and cdf of a N(0,1) distribution. Other existing methods of modifying the t-ratio without questioning the standard critical value, such as those proposed by Johnson (1978), Lyon et al. (1999) and Yanagihara and Yuan (2005), share with our approach a similar lack of complete skewness correction. It takes resampling methods, like bootstrap, to remove completely (up to order $o_p(1/\sqrt{n})$) the perverse effect of skewness on finite sample by modifying the critical value itself. This observation is confirmed by our Monte Carlo results in the next section.

(iv) More generally, our approach does not really try to make the behavior of the test statistic $\xi_n(V_n)$ the closest possible to $\chi^2(p)$ (or equivalently, $t_n(V_n)$ the closest possible to normal) under the null but rather the closest possible to the infeasible $\xi_n(V)$. In this respect, our approach can be seen as an extension of the work of Rothenberg (1988) to non-linear settings in over-identified models, although our focus is rather on size than on power. Extension of the results to $\sqrt{n}$-local alternatives is straightforward for parts (i) and (ii) of the proposition. For part (iii) it is also possible because the conditions in Chaudhuri and Renault (2011) allow for that.

(v) As already mentioned in the introduction, the proposed refinements are in the spirit of Bartlett correction and differ fundamentally from the other strand of the literature that seeks to refine the critical value for the test, for instance by resampling. While it has been shown that empirical likelihood is Bartlett-correctable (DiCiccio et al. (1991)) while other empirical discrepancy statistics are not in general (Corcoran (1998)), we circumvent this difficulty by focusing directly on implied probabilities (for improvement of variance estimation) and not on estimators provided by minimization of Cressie-Read discrepancies.

(vi) The key point is that, even though the first order conditions of efficient GMM amount to picking a subset of just-identified moment conditions, the efficient estimation of the variance matrix under the null will take advantage of the whole set of moment conditions. In fact, simulation results reported below show that the empirical size can be made closer to the nominal level (based on the first-order asymptotics) by the use of the score statistics that involve, respectively, the modified estimators $V_n^{(-1)}$ (i.e., EEL) and $V_n^{(-0)}$ (i.e., EL) of the asymptotic variance.

(vii) Higher order asymptotics derived in this paper had not been explicitly studied by the extant literature on higher order improvements of GMM provided by empirical likelihood (see, for

example, Newey and Smith (2004) or Guggenberger and Smith (2005)). Since we extend the use of the Cavanagh-Rothenberg-type approximations, we set the focus on conditional expectations $E[B_n|t_n = \tau]$ of the bias terms rather than just on their unconditional behavior.

4 Covariance structure model: A Monte-Carlo experiment

In this section we demonstrate the improvement in the finite-sample behavior of the score tests, in terms of closeness of size in finite samples to the nominal level, due to the use of the modified estimators $V_n^{(-1)}$ (EEL) and $V_n^{(0)}$ (EL) instead of the naive estimator $\bar{V}_n^{\text{Center}}$ of the asymptotic variance V.(Results with $\bar{V}_n$ are similar to that with $\bar{V}_n^{\text{Center}}$.) We also demonstrate that such improvements may often be comparable to that obtained by bootstrap.

The data generating process, originally due to Altonji and Segal (1996), is taken from Horowitz (1998). For $i = 1, \ldots, n$ and $j = 1, \ldots, J+1$ we generate $Y_{i,j} \overset{\text{i.i.d.}}{\sim} f(y, \theta)$ for various choices of f: uniform, normal, t_{10} and exponential(1), all of them standardized to have expectation 0 and variance θ. We further generate $Z_{i,j} = (Y_{i,j} + .5 \times Y_{i,j+1})/\sqrt{1 + (.5)^2}$ for $j = 1, \ldots, J$ and $i = 1, \ldots, n$. Therefore, for each $i = 1, \ldots, n$, and $j = 1, \ldots, J$ we have $E[Z_{i,j}] = 0$, $V(Z_{i,j}) = \theta$, $Cov(Z_{i,j}, Z_{i,j+1}) = .5 \times \theta/(1 + (.5)^2)$ and $Cov(Z_{i,j}, Z_{i,j+s}) = 0$ for $s \geq 2$. Hence, taking $X_i = (Z_{i,1}, \ldots, Z_{i,J})$ and $K = 2J - 1$, we have

$$\begin{aligned} \psi(X_i, \theta) \quad = \quad & \big[Z_{i,1}^2 - \theta, Z_{i,2}^2 - \theta, \ldots, Z_{i,J}^2 - \theta, Z_{i,1}Z_{i,2} - .5 \times \theta/(1 + (.5)^2), \\ & Z_{i,2}Z_{i,3} - .5 \times \theta/(1 + (.5)^2), \ldots, Z_{i,J-1}Z_{i,J} - .5 \times \theta/(1 + (.5)^2)\big]'. \end{aligned}$$

Since our focus is on size of the score tests, we take the true and hypothesized θ to be $\theta_0 = 1$.

Estimated size of the score tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, $H_2 : \theta < \theta_0$ and $H_3 : \theta \neq \theta_0$ at the 5% and 10% nominal levels are reported in Tables 1 and 2 respectively for sample sizes 500, 1000 and 5000. For the purpose of brevity and to keep the ratio of number of moment restrictions to observations relatively low, results are reported only for $J = 4, 8$ (i.e., $k = 7, 15$) based on 5000 Monte-Carlo trials.

As would be expected from Theorem 3.2(i), score tests based on the naive variance estimator $\bar{V}_n^{\text{Center}}$ is heavily size-distorted in small samples. The size distortion increases with the increase in skewness and kurtosis of the underlying distribution of the moment vector (due to the choice of the distributions – uniform, normal, t and exponential progressively).

On the other hand, supporting Theorem 3.2 (ii) and (iii), use of the modified estimators of variance corrects the size distortion substantially in all cases, although in the exponential case

such correction may not be deemed enough at least when the alternative is $H_1 : \theta > \theta_0$ and $H_3 : \theta \neq \theta_0$. Figures 1-4 contain plots for the kernel density estimators of the distribution of the three versions of the score statistic, and demonstrate the correction (towards normality) due to the use of the modified variance estimators. The density estimators are obtained based on 1000 Monte-Carlo replications of the score statistics under $H_0 : \theta = \theta_0$.

Size distortion of all tests diminishes when sample size increases, although the degree of it depends on the underlying distribution and the number of moments. This should not be surprising because all three versions of the score tests are (first order) asymptotically equivalent.

Incidentally the corrections in size are very similar to that obtained by bootstrap. This is evident from the results in Table 3 where we present the finite-sample rejection rate of bootstrap score tests for sample sizes $n = 500$ and $n = 1000$. The bootstrap method used is that proposed by Brown and Newey (2002) — we resample the observations by imposing the null hypothesis but instead of using the empirical probabilities to draw the bootstrap sample with replacement, we use the implied probabilities from EEL. We use the EEL implied probabilities, evaluated at $\theta = \theta_0$, for computational convenience and noting that, by construction, the EEL implied probabilities also impose the null hypothesis on the bootstrap population thus constructed. Precisely because we use the EEL implied probabilities to define the population distribution of the bootstrap sample, we use the shrinkage version $V_n^{(-1,p)}$ of Antoine et al. (2007) to ensure that the probabilities are non-negative.

$H_0 : \theta = \theta_0$ v/s (nominal level 5%)		$H_1 : \theta > \theta_0$			$H_2 : \theta < \theta_0$			$H_3 : \theta \neq \theta_0$		
DGP $\sim (0,1)$	Moments $= 2J-1$	Naive	EEL	EL	Naive	EEL	EL	Naive	EEL	EL
Uniform	7	8.04	5.26	4.74	3.08	5.12	4.64	5.92	5.62	4.84
	15	11.12	7.02	5.94	2.62	6.14	5.32	8.04	7.24	5.5
Normal	7	12.76	6.26	5.36	2.34	6.08	5.28	9.16	6.6	5.3
	15	18.38	7.94	6.5	1.62	6.44	5.12	13.02	8.12	6.44
t(10)	7	16.78	7.4	6.46	1.74	5.96	4.74	11.66	7.6	6.02
	15	26.8	10.3	8.5	0.54	5.84	4.28	18.92	9.26	6.92
Exp(1)	7	31.94	13.54	11.38	0.78	5.4	4.16	25.4	11.9	8.92
	15	52.82	19.56	15.82	0.22	3.6	2.94	43.72	15.6	11.7
Uniform	7	7.62	5.7	5.5	3.54	4.64	4.42	5.46	5.16	4.84
	15	8.52	5.44	4.96	2.9	5.26	4.78	5.78	5.38	4.78
Normal	7	9.98	5.88	5.26	2.94	5.8	5.34	6.96	5.94	5.08
	15	12.78	6.26	5.58	1.78	5.68	4.84	8.98	6.46	5.12
t(10)	7	13.96	7.26	6.34	1.86	5.02	4.38	9.34	6.38	5.5
	15	17.62	7	6.2	1.2	5.36	4.54	11.4	7.04	5.48
Exp(1)	7	23.82	9.76	8.52	0.92	5.76	4.6	17.72	8.56	6.74
	15	36.16	11.64	9.82	0.44	4.7	3.46	26.84	9.4	7.52
Uniform	7	6.16	5.26	5.2	3.72	4.4	4.28	5.08	4.82	4.78
	15	6.62	5.6	5.42	4	5.1	4.9	5.38	5.3	5.12
Normal	7	6.56	4.78	4.74	3.94	5.16	5.02	5.46	5.18	5.06
	15	7.68	5.2	5.02	3	5.08	4.94	5.4	5.06	4.9
t(10)	7	7.78	5.38	5.02	3.4	5.24	4.86	5.52	5.54	5.04
	15	9.12	5.24	4.94	2.88	5.58	5.2	6.72	5.94	5.6
Exp(1)	7	11.48	5.94	5.5	1.96	5.56	5.1	7.72	5.74	4.98
	15	16.24	6.56	5.98	1.48	5.36	4.54	10.66	6.34	5.58

Table 1: Finite-sample rejection rate (in %) of the true parameter value by score tests at the 5% nominal level. Results reported based on 5000 Monte Carlo trials. Naive, EEL and EL correspond to the score tests based on the statistics using $\bar{V}_n^{\text{Center}}$, $V_n^{(-1)}$ and $V_n^{(0)}$ respectively for V_n. Number of observations $n = 500$ in top panel, $n = 1000$ in middle panel, and $n = 5000$ in bottom panel.

$H_0:\theta=\theta_0$ v/s (nominal level 10%)		$H_1:\theta>\theta_0$			$H_2:\theta<\theta_0$			$H_3:\theta\neq\theta_0$		
DGP $\sim(0,1)$	Moments $=2J-1$	Naive	EEL	EL	Naive	EEL	EL	Naive	EEL	EL
Uniform	7	14.12	10.24	9.64	6.8	10.04	9.68	11.12	10.38	9.38
	15	18.66	11.9	10.62	6.12	11.64	10.24	13.74	13.16	11.26
Normal	7	20.7	11.82	10.76	5.24	11.1	10.04	15.1	12.34	10.64
	15	27.96	13.78	12.08	3.78	11.18	9.7	20	14.38	11.62
t(10)	7	25.58	13.46	12.38	3.94	10.68	8.94	18.52	13.36	11.2
	15	38.88	16.98	15.22	2.04	10.5	8.66	27.34	16.14	12.78
Exp(1)	7	41.88	20.76	19.02	1.98	9.94	7.88	32.72	18.94	15.54
	15	62.88	28.26	24.54	0.6	7	5.94	53.04	23.16	18.76
Uniform	7	13.96	10.98	10.48	7.54	10.36	10.02	11.16	10.34	9.92
	15	15.12	10.74	9.86	6.3	10.62	9.88	11.42	10.7	9.74
Normal	7	16.4	10.84	10.32	6.26	10.82	10.08	12.92	11.68	10.6
	15	21.44	11.68	10.74	4.24	11	10.24	14.56	11.94	10.42
t(10)	7	22.64	12.94	12.16	4.08	10.2	9.02	15.82	12.28	10.72
	15	27.94	12.64	11.44	2.8	10.58	9.2	18.82	12.36	10.74
Exp(1)	7	33	16.82	15.06	2.7	10.38	8.86	24.74	15.52	13.12
	15	47.68	18.38	16.36	1.08	8.66	7.52	36.6	16.36	13.28
Uniform	7	11.28	9.94	9.88	8.22	9.16	9.06	9.88	9.66	9.48
	15	12.84	10.76	10.64	8.48	10.1	9.98	10.62	10.7	10.32
Normal	7	12.7	10	9.88	7.72	10.2	10.02	10.5	9.94	9.76
	15	14.54	10.34	10.08	6.66	9.74	9.5	10.68	10.28	9.96
t(10)	7	14.38	10.32	10.04	6.68	9.92	9.6	11.18	10.62	9.88
	15	16.2	10.14	9.86	5.88	10.96	10.2	12	10.82	10.14
Exp(1)	7	19.08	11.54	11.12	4.96	9.94	9.3	13.44	11.5	10.6
	15	25.56	12.6	11.86	3.5	10.18	9.44	17.72	11.92	10.52

Table 2: Finite-sample rejection rate (in %) of the true parameter value by score tests at the 10% nominal level. Results reported based on 5000 Monte Carlo trials. Naive, EEL and EL correspond to the score tests based on the statistics using $\bar{V}_n^{\text{Center}}$, $V_n^{(-1)}$ and $V_n^{(0)}$ respectively for V_n. Number of observations $n = 500$ in top panel, $n = 1000$ in middle panel, and $n = 5000$ in bottom panel.

$H_0 : \theta = \theta_0$ tested at		Nominal level 5%			Nominal level 10%		
DGP	Moments	$H_1 : \theta > \theta_0$	$H_2 : \theta < \theta_0$	$H_3 : \theta \neq \theta_0$	$H_1 : \theta > \theta_0$	$H_2 : \theta < \theta_0$	$H_3 : \theta \neq \theta_0$
Uniform	7	4.74	4.64	4.92	9.54	9.5	9.38
	15	5.62	5.12	5.58	5.62	5.12	5.58
Normal	7	5.74	4.76	5.26	10.96	9.32	10.5
	15	6.1	4.7	5.44	11.46	9.22	10.8
t(10)	7	5.9	4.32	5.5	11.12	8.84	10.22
	15	6.62	3.68	4.82	12.58	7.38	10.3
Exp(1)	7	8.28	3.62	6.28	14.8	6.84	11.9
	15	8.56	2.4	5.58	15.56	5.7	10.96
Uniform	7	5.32	4.98	4.94	10.1	9.8	10.3
	15	4.54	4.82	4.72	9.58	10	9.36
Normal	7	5.28	5.04	4.88	10.32	9.88	10.32
	15	5.88	4.7	5.14	10.68	9.54	10.58
t(10)	7	5.86	4.36	5.1	11.14	8.86	10.22
	15	6.02	4.2	4.88	11.12	8.94	10.22
Exp(1)	7	7.62	4.64	6.68	13.32	9.18	12.26
	15	7.88	3.64	5.82	13.7	7.54	11.52

Table 3: Finite-sample rejection rate (in %) of the true parameter value by using EEL-Bootstrap critical values and the Naive score statistic. Critical values are computed based on 199 bootstrap resampling. Results are reported based on 5000 Monte-Carlo trials. Number of observations $n = 500$ in top panel and $n = 1000$ in bottom panel. The distributions are standardized (mean = 0 and variance = 1), and the number of moments is equal to $2J - 1$ as in Tables (2) and (2).

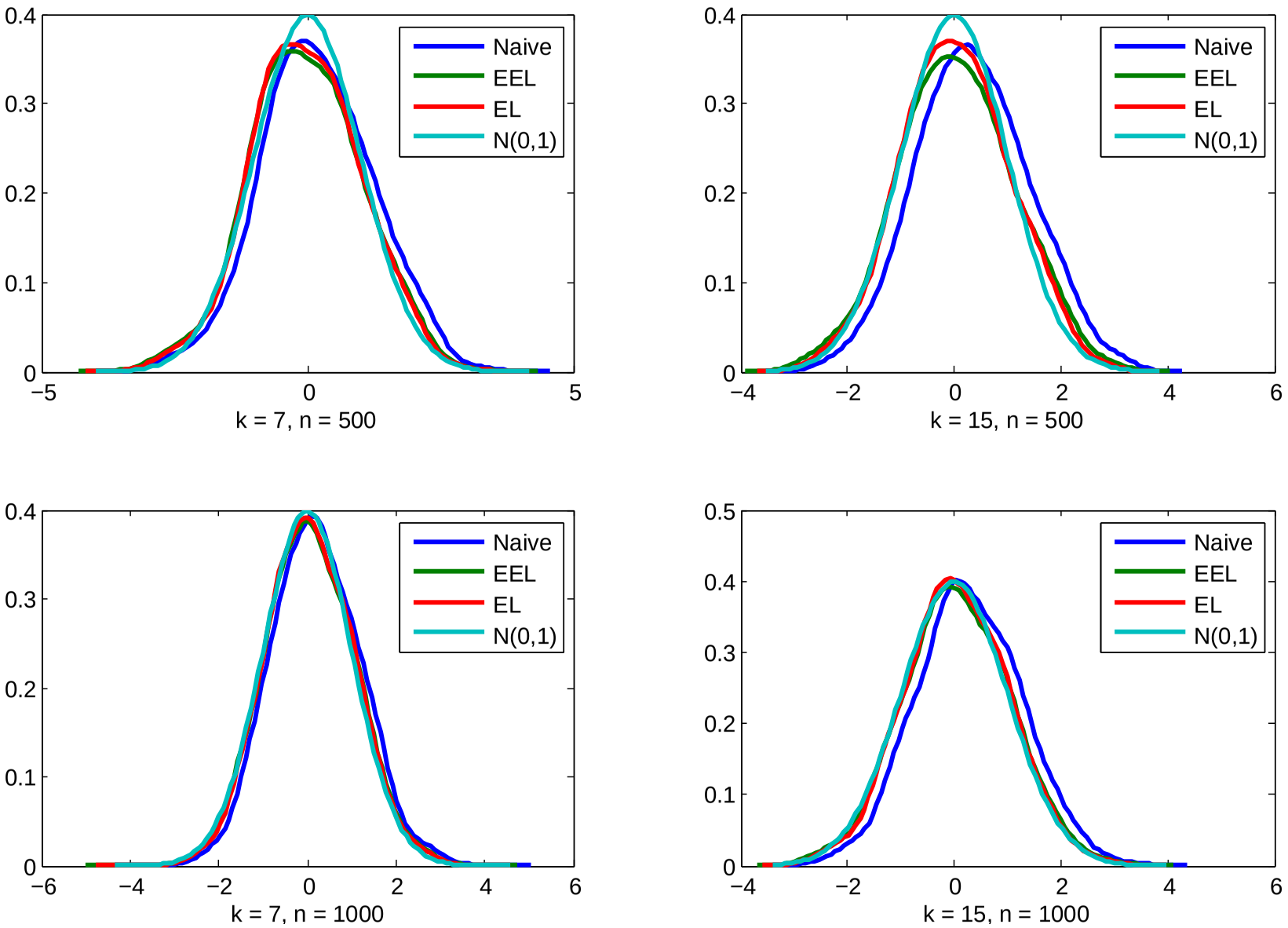

Figure 1: DGP-Uniform: Kernel density estimator of score statistics using Naive ($\bar{V}_n$), EEL ($V_n^{-1,0}$) and EL ($V_n^{0,0}$) estimators for V_n^0. Number of Monte-Carlo Trials is 1000.

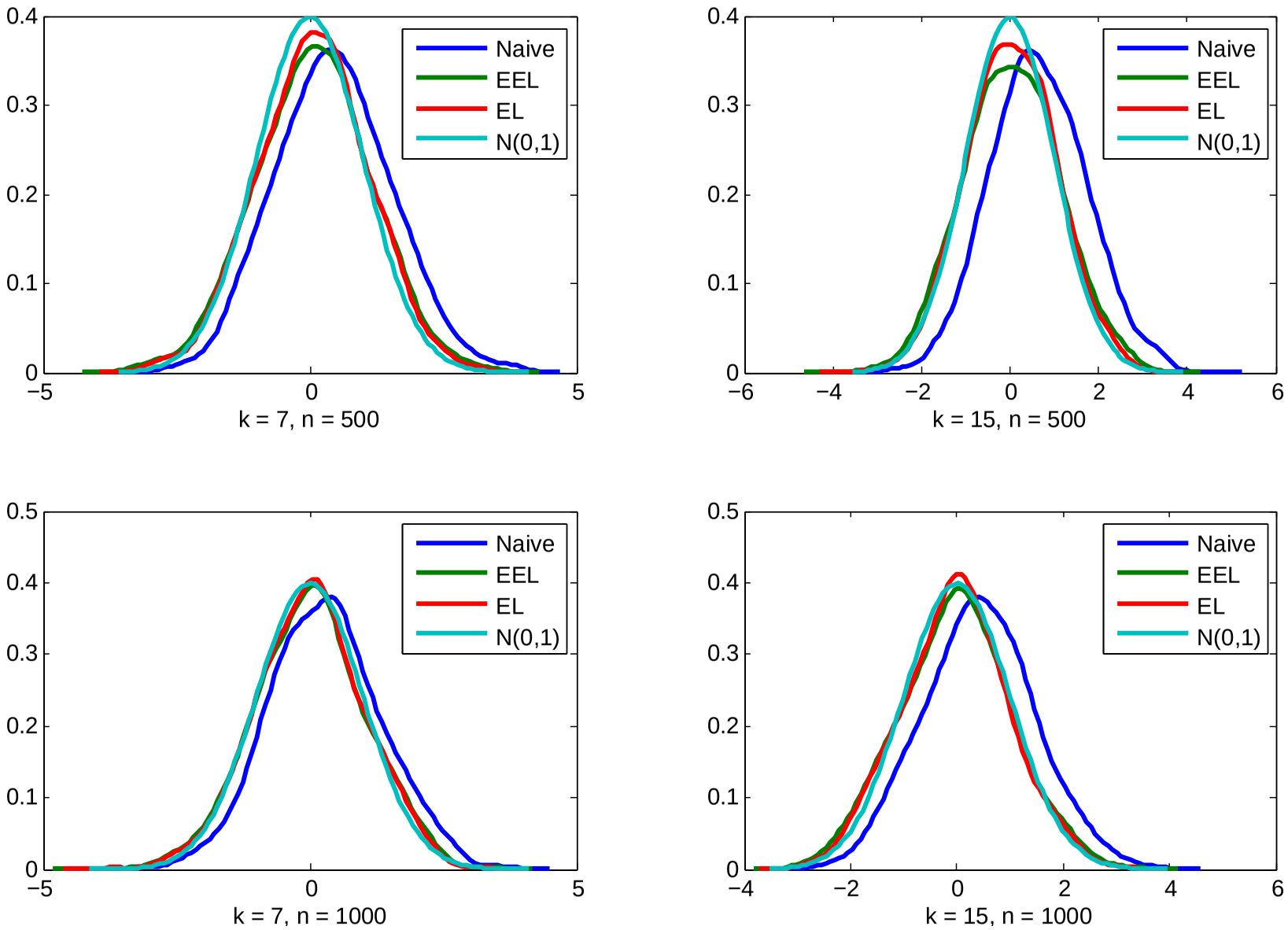

Figure 2: DGP-Normal: Kernel density estimator of score statistics using Naive ($\bar{V}_n$), EEL ($V_n^{-1,0}$) and EL ($V_n^{0,0}$) estimators for V_n^0. Number of Monte-Carlo Trials is 1000.

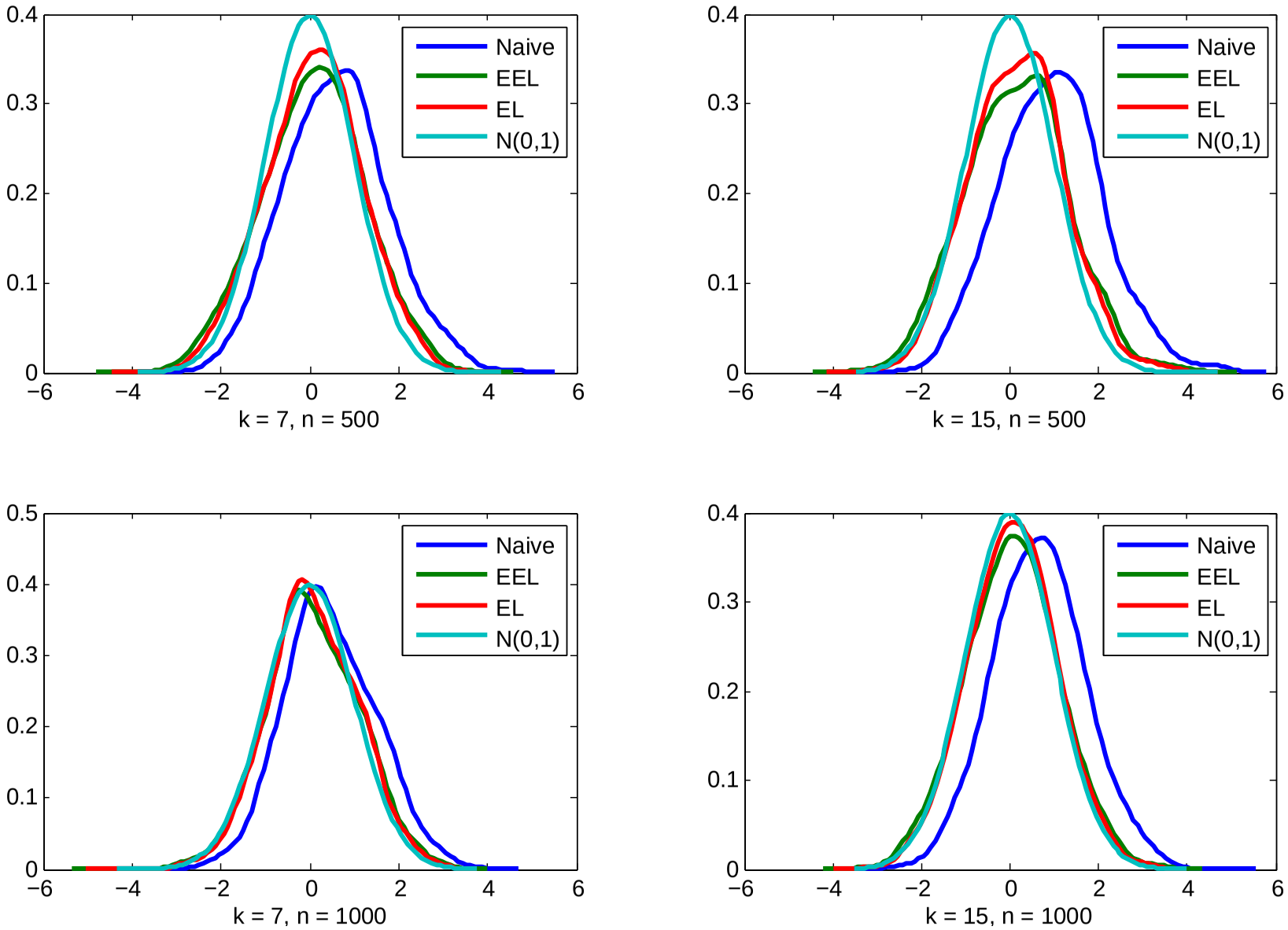

Figure 3: DGP-t(10): Kernel density estimator of score statistics using Naive ($\bar{V}_n$), EEL ($V_n^{-1,0}$) and EL ($V_n^{0,0}$) estimators for V_n^0. Number of Monte-Carlo Trials is 1000.

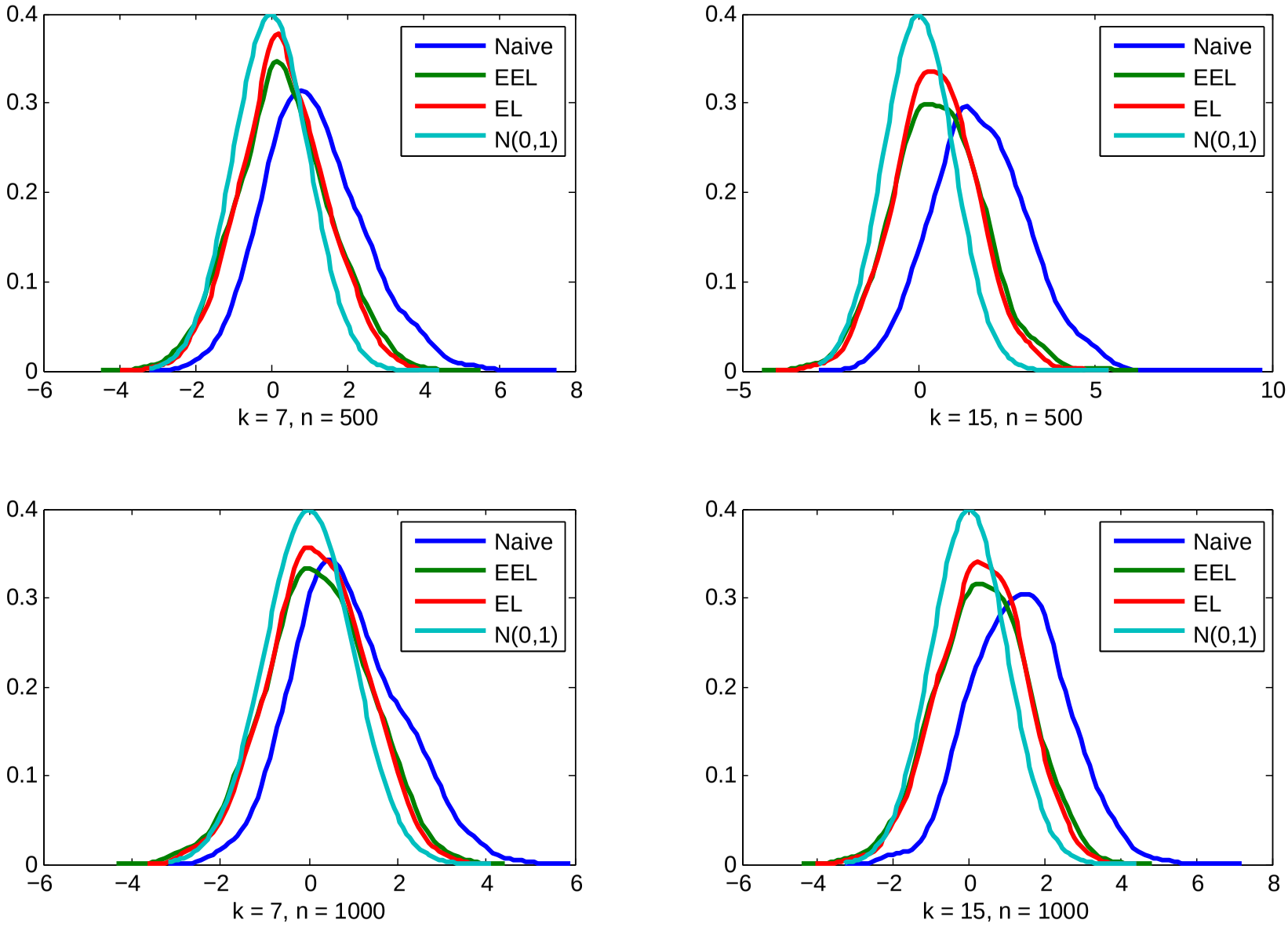

Figure 4: DGP-Exponential(1): Kernel density estimator of score statistics using Naive ($\bar{V}_n$), EEL ($V_n^{-1,0}$) and EL ($V_n^{0,0}$) estimators for V_n^0. Number of Monte-Carlo Trials is 1000.

5 Conclusion

Information theoretic approaches to GMM are known to bring finite-sample improvements that can be confirmed by the theoretical results of higher order asymptotic efficiency. These issues have been mainly addressed in the literature so far to show higher order efficiency of estimators (for example, Newey and Smith (2004)) and better performance of tests in the presence of weak identification (for example, Kleibergen (2005), Guggenberger and Smith (2005), Chaudhuri and Renault (2011), etc.). The weak identification literature typically sets the focus on near-rank deficiencies in Jacobian matrices.

In this paper, we consider situations where the Jacobian of the moment function is known under the null and is full column rank. The need for finite-sample improvements only comes from the sample uncertainty in the asymptotic variance of the moment function. It has been well documented (see, for example, Altonji and Segal (1996) for estimation of covariance structures) that a perverse correlation between the estimated asymptotic variance and the moment function is responsible for serious finite-sample bias in estimation and inference. Horowitz (1998) proposed a bootstrap approach for improved estimation and inference in covariance structures models.

The originality of the current paper is to propose a battery of possible adjustment procedures for score statistics that do not resort to any resampling strategy. The key idea is close to the well known Bartlett adjustment, a correction directly on the test statistic itself without modifying the usual chi square-based critical value. Moreover this adjustment pertains to the information theoretical methodology since it is obtained by using the implied probabilities derived from the minimization of any generalized entropy function.

While we provide evidence of improved performance of the score test, both by closed form higher order expansions and by Monte-Carlo experiments, this possibility of improvement may look at odds with the known impossibility results in the literature, at least for the generalized entropy approaches different from empirical likelihood. Corcoran (1998) had shown that Bartlett adjustments put forward by DiCiccio et al. (1991) for empirical likelihood cannot be extended to general Cressie-Read empirical discrepancies. Newey and Smith (2004) had shown that, in case of skewness in the moment function, only the empirical likelihood (and no other Cressie-Read empirical discrepancy) takes care of the efficient estimation of the asymptotic variance matrix. We actually circumvent these impossibility results by assuming that there is no unknown parameters θ under the null and then, the implied probabilities can be used to improve the variance estimation, irrespective of the user's preferred generalized entropy function. This is important since it allows in particular to use Euclidean likelihood which is more user-friendly, both for numerical and

analytical computations, and for interpretation as well (connection with continuously updated GMM). In this respect, our paper extends the work of Antoine et al. (2007) to issues related to testing of hypotheses. Their strategy of 3-step Euclidean likelihood based estimation could be used for the more general case where some nuisance parameters, other than the asymptotic variance matrix, would remain unknown under the null. Our paper can also be seen as an extension of the classical work of Rothenberg (1988) to nonlinear settings in over-identified models. Of course, much remains to be done for the sake of finite-sample improvements in GMM inference. To the best of our knowledge, the higher order expansions of power functions that we derive in this paper have not yet been obtained in more general circumstances with unknown Jacobian matrix and/or weak identification issues.

References

Abowd, J. M. and Card, D. (1989). On the Covariance Structure of Earnings and Hours Changes. *Econometrica*, 57: 411–445.

Altonji, J. and Segal, L. M. (1996). Small Sample Bias in GMM Estimation of Covariance Structures. *Journal of Business and Economic Statistics*, 14: 353–366.

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59: 817–854.

Antoine, B., Bonnal, H., and Renault, E. (2007). On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood. *Journal of Econometrics*, 138: 461–487.

Brown, B. and Newey, W. (1998). Efficient Semiparametric Estimation of Expectations. *Econometrica*, 66: 453–464.

Brown, B. and Newey, W. (2002). Generalised method of moments, efficient bootstrapping, and improved inference. *Journal of Business and Economic Statistics*, 20: 507–517.

Cavanagh, C. (1983). *Hypothesis Testing in Models with Discrete Dependent Variable*. PhD thesis, University of California, Berkeley.

Chaudhuri, S. and Renault, E. (2011). Finite-sample improvements of score tests by the use of implied probabilities from Generalized Empirical Likelihood. Technical report, University of North Carolina, Chapel Hill.

Corcoran, S. A. (1998). Bartlett Adjustment of Empirical Discrepancy Statistics. *Biometrika*, 85: 967–972.

DiCiccio, T., Hall, P., and Romano, J. (1991). Empirical Likelihood is Bartlett-Correctable. *Annals of Statistics*, 19: 1053–1061.

Goggin, E. (1994). Convergence in Distribution of Conditional Expectations. *Annals of Probability*, 22: 1097–1114.

Guggenberger, P. and Smith, R. (2005). Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification. *Econometric Theory*, 21: 667–709.

Hall, A. R. (2000). Covariance Matrix Estimation and the Power of the Overidentifying Restrictions Test. *Econometrica*, 68: 1517–1527.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business and Economic Statistics*, 14: 262–280.

Horowitz, J. (1998). Bootstrap Methods for Covariance Structures. *The Journal of Human Resources*, 33: 39–61.

Imbens, G. W., Spady, R. H., and Johnson, P. (1998). Information Theoretic Approaches to Inference in Moment Condition Models. *Econometrica*, 66: 333–357.

Jacod, J. and Shiryaev, A. N. (2003). *Limit Theorems For Stochastic Processes*. Springer.

Johnson, N. J. (1978). Modified t Tests and Confidence Intervals for Asymmetrical Populations. *Journal of the American Statistical Association*, 74: 536–544.

Kitamura, Y. and Stutzer, M. (1997). An Information-Theoretic Alternative to Generalized Method of Moments Estimation. *Econometrica*, 65: 861–874.

Kleibergen, F. (2005). Testing Parameters In GMM Without Assuming That They Are Identified. *Econometrica*, 73: 1103–1123.

Lyon, J. D., Barber, B. M., and Tsai, C.-L. (1999). Improved Methods for Tests of Long-Run Abnormal Stock Returns. *Journal of Finance*, LIV: 165–201.

Newey, W. K. and Smith, R. J. (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica*, 72: 219–255.

Newey, W. K. and West, K. D. (1987). Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review*, 28: 777–787.

Rothenberg, T. J. (1988). Power Functions for Some Robust Tests of Regression Coefficients. *Econometrica*, 56: 997–1019.

Sowell, F. (1996). Optimal Tests for Parameter Instability in the Generalized Method of Moments Framework. *Econometrica*, 64: 1085–1107.

Sun, Y., Phillips, P. C. B., and Jin, S. (2008). Optimal Bandwidth Selection in Heteroscedasicity-Autocorrelation Robust Testing. *Econometrica*, 76: 175–194.

Yanagihara, H. and Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British Journal of Mathematical and Statistical Psychology*, 58: 209–237.

Ziemer, R. F. and Wetzstein, M. E. (1983). A Stein-Rule Method for Pooling Data. *Economics Letters*, 11: 137–143.

A Appendix

Proof of Lemma 3.1: The approximation for

$$t_n(V_n) := (G'V_n^{-1}G)^{-1/2}G'V_n^{-1}Z_{1n}$$

is obtained as follows. Since $V_n - V = O_P(1/\sqrt{n}) = V_n^{-1} - V^{-1}$, it follows that

$$\begin{aligned} V_n^{-1} &= V^{-1} + V_n^{-1}(V - V_n)V^{-1} \\ &= V^{-1} + V^{-1}(V - V_n)V^{-1} + (V_n^{-1} - V^{-1})(V - V_n)V^{-1} \\ &= V^{-1} + V^{-1}(V - V_n)V^{-1} + O_P(1/n). \end{aligned} \tag{18}$$

Therefore, since $Z_{1n} = O_P(1)$, we obtain

$$\begin{aligned} t_n(V_n) &= t_{1n}(V_n) + t_{2n}(V_n) + O_P(1/n), \text{ where} \\ t_{1n}(V_n) &= (G'V_n^{-1}G)^{-1/2}G'V^{-1}Z_{1n}, \\ t_{2n}(V_n) &= (G'V_n^{-1}G)^{-1/2}G'V^{-1}(V - V_n)V^{-1}Z_{1n}. \end{aligned}$$

Moreover, using (18) again:

$$\begin{aligned} & G'V_n^{-1}G = G'V^{-1}G + G'V^{-1}(V - V_n)V^{-1}G + O_P(1/n) \\ \Rightarrow\ & (G'V_n^{-1}G)^{-1/2} = (G'V^{-1}G)^{-1/2}\left[1 - \frac{1}{2}(G'V^{-1}G)^{-1}G'V^{-1}(V - V_n)V^{-1}G\right] + O_P(1/n) \\ \Rightarrow\ & G\sqrt{n}(G'V_n^{-1}G)^{-1/2} = G\sqrt{n}(G'V^{-1}G)^{-1/2} + \frac{1}{2}P(G)\tilde{Z}_{2n}V^{-1}G(G'V^{-1}G)^{-1/2} + O_P\left(1/\sqrt{n}\right) \end{aligned}$$

where, for notational simplicity, we don't make explicit the dependence of $\tilde{Z}_{2n}$ on V_n:

$$\tilde{Z}_{2n} = \tilde{Z}_{2n}(V_n) = \sqrt{n}(V_n - V).$$

Therefore,

$$\begin{aligned} G\sqrt{n}t_{1n}(V_n) &= G\sqrt{n}t_n + \frac{1}{2}P(G)\tilde{Z}_{2n}V^{-1}P(G)Z_{1n}(G'V^{-1}G)^{1/2} + O_P\left(1/\sqrt{n}\right) \\ G\sqrt{n}t_{2n}(V_n) &= -P(G)\tilde{Z}_{2n}V^{-1}Z_{1n}(G'V^{-1}G)^{1/2} - \frac{1}{2}P(G)\tilde{Z}_{2n}V^{-1}P(G)\frac{\tilde{Z}_{2n}}{\sqrt{n}}V^{-1}Z_{1n}(G'V^{-1}G)^{1/2} + O_P\left(1/\sqrt{n}\right) \\ &= -P(G)\tilde{Z}_{2n}V^{-1}Z_{1n}(G'V^{-1}G)^{1/2} + O_P\left(1/\sqrt{n}\right) \\ &= -P(G)\tilde{Z}_{2n}V^{-1}P(G)Z_{1n}(G'V^{-1}G)^{1/2} - P(G)\tilde{Z}_{2n}V^{-1}M(G)Z_{1n}(G'V^{-1}G)^{1/2} + O_P\left(1/\sqrt{n}\right) \end{aligned}$$

since $\tilde{Z}_{2n} = O_P(1)$. Hence,

$$G\sqrt{n}t_n(V_n) = G\sqrt{n}t_n - \frac{1}{2}P(G)\tilde{Z}_{2n}V^{-1}P(G)Z_{1n}(G'V^{-1}G)^{1/2} - P(G)\tilde{Z}_{2n}V^{-1}M(G)Z_{1n}(G'V^{-1}G)^{1/2} + O_P\left(1/\sqrt{n}\right)$$

that gives the announced formula for $B_n(V_n)$. ■

Proof of Theorem 3.2: We first note that since $P(G)Z_{1n} = G(G'V^{-1}G)^{-1/2}t_n$:

$$\begin{aligned} & GB_{1n} = -\frac{1}{2}P(G)\tilde{Z}_{2n}V^{-1}P(G)Z_{1n}(G'V^{-1}G)^{1/2} \qquad (19) \\ \Rightarrow \quad & GE[B_{1n}\,|t_n = \tau] = -\frac{1}{2}P(G)E[\tilde{Z}_{2n}\,|P(G)Z_{1n} = a]\,V^{-1}a(G'V^{-1}G)^{1/2} \\ \Rightarrow \quad & GE[B_{1n}\,|t_n = \tau] = -\frac{1}{2}P(G)E[\tilde{Z}_{2n}\,|P(G)Z_{1n} = a]\,V^{-1}G\tau \end{aligned}$$

where $a = G(G'V^{-1}G)^{-1/2}\tau$.

The proof follows in three parts. In parts I and II we characterize the term $E[B_{1n}(V_n)\,|t_n = \tau]$ for $V_n = \bar{V}_n$ and $V_n = V_n^{(-1)}$ respectively. In part III, we obtain the announced results by using the asymptotic equivalence in (12) and (15) respectively.

[Part I] $V_n = \bar{V}_n$**:**

From assumption Edg(τ) (more specifically, CLT(a)(iii)) we know that:

$$\lim_{n=\infty} E[\tilde{Z}_{2n}\,|P(G)Z_{1n} = a] = E[\tilde{Z}_2\,|P(G)Z_1 = a]$$

where $Vec(\tilde{Z}_2) = Z_2$. Since the vector $(Z_1', Z_2')'$ is Gaussian with zero mean, we also have a Gaussian vector by considering $(t, Z_2')'$, when t is defined by $P(G)Z_1 = G(G'V^{-1}G)^{-1/2}t$. Then,

$$t = \frac{G'V^{-1}Z_1}{(G'V^{-1}G)^{1/2}}$$

is a standardized univariate normal and

$$E[Z_2\,|P(G)Z_1 = a] = E[Z_2\,|t = \tau] = -\tau Cov[Z_2, t]$$

where $Cov[Z_2, t]$ is the column vector of dimension $K(K+1)/2$ with coefficients:

$$\frac{1}{(G'V^{-1}G)^{1/2}}Cov\left[\psi_{ih}\psi_{ik}, \psi_i\right]V^{-1}G, h, k = 1, 2, ..., K.$$

Therefore,

$$E[\tilde{Z}_2 \,|P(G)Z_1 = a] = -\frac{\tau}{(G'V^{-1}G)^{1/2}} \left\{ \left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array} \right] V^{-1}G \right\}$$

where for $l = 1, 2, ..., K$, Ω_l is the square symmetric matrix of size K with coefficients $Cov\left[\psi_{ih}\psi_{ik}, \psi_{il}\right]$ for $h, k = 1, 2, ..., K$.

We can now conclude that

$$GE[B_{1n} \,|t_n = \tau] = -\frac{\tau}{2(G'V^{-1}G)^{1/2}} P(G) \left\{ \left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array} \right] V^{-1}G \right\} V^{-1}G\tau + o(1),$$

that is,

$$E[B_{1n} \,|t_n = \tau] = -\frac{\tau^2}{2(G'V^{-1}G)^{3/2}} G'V^{-1} \left\{ \left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array} \right] V^{-1}G \right\} V^{-1}G + o(1).$$

[Part II] $V_n = V_n^{(-1)}$**:**

From (14) we know that

$$\tilde{Z}_{2n}(V_n^{(-1)}) = \tilde{Z}_{2n}(\bar{V}_n) - \left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array} \right] V^{-1}Z_{1n}. \tag{20}$$

Therefore,

$$E[\tilde{Z}_{2n}(V_n^{(-1)}) \,|P(G)Z_{1n} = a] = E[\tilde{Z}_{2n}(\bar{V}_n) \,|P(G)Z_{1n} = a] - \left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array} \right] V^{-1}E[Z_{1n} \,|P(G)Z_{1n} = a]\,. \tag{21}$$

We already know that

$$\lim_{n=\infty} E[\tilde{Z}_{2n}(\bar{V}n) \,|P(G)Z_{1n} = a] = -\frac{\tau}{(G'V^{-1}G)^{1/2}} \left\{ \left[\begin{array}{cccc} \Omega_1 & ... & \Omega_K \end{array} \right] V^{-1}G \right\} \tag{22}$$

By assumption CLT(a)(iii) we also have

$$\lim_{n=\infty} E[Z_{1n} \,|P(G)Z_{1n} = a] = E[Z_1 \,|P(G)Z_1 = a] = a = G(G'V^{-1}G)^{-1/2}\tau. \tag{23}$$

Plugging (22) and (23) in (21), we conclude that

$$\lim_{n=\infty} E[\tilde{Z}_{2n}(V_n^{(-1)}) \,|P(G)Z_{1n} = a] = 0 \tag{24}$$

and, thus, by (19):

$$E[B_{1n}(V_n^{(-1)} \,|t_n = \tau] = o(1).$$

Moreover,

$$\begin{aligned} GB_{2n} &= -P(G)\tilde{Z}_{2n}V^{-1}M(G)Z_{1n}(G'V^{-1}G)^{1/2} \\ \Rightarrow GE[B_{2n}\,|t_n=\tau] &= -P(G)E[\tilde{Z}_{2n}V^{-1}M(G)Z_{1n}\,|P(G)Z_{1n}=a]\,(G'V^{-1}G)^{1/2} \\ &= -P(G)E[\tilde{Z}_{2n}V^{-1}Z_{1n}\,|P(G)Z_{1n}=a]\,(G'V^{-1}G)^{1/2} \\ &\quad + P(G)E[\tilde{Z}_{2n}\,|P(G)Z_{1n}=a]\,V^{-1}a(G'V^{-1}G)^{1/2}. \end{aligned}$$

Therefore, by (24), we get

$$\lim_{n=\infty} GE[B_{2n}(V_n^{(-1)}\,|t_n=\tau] = -P(G)(G'V^{-1}G)^{1/2}\lim_{n=\infty} E[\tilde{Z}_{2n}(V_n^{(-1)})V^{-1}Z_{1n}\,|P(G)Z_{1n}=a]$$

and this limit is zero by assumption CLT(a)(iii). Therefore, we also have

$$E[B_{2n}(V_n^{(-1)}\,|t_n=\tau] = o(1).$$

[Part III:] $V_n = \bar{V}_n^{\mathbf{Center}}$ **or** $V_n = V_n^{(\gamma)}$:

From (12) and (15) we only need to show that for V_n and $\tilde{V}_n$ that are consistent for V,

$$\sqrt{n}(V_n - \tilde{V}_n) = o_P(1) \Rightarrow \lim_{n=\infty} E[B_n(V_n) - B_n(\tilde{V}_n)\,|t_n=\tau] = 0.$$

Following the decomposition of $t_n(V_n)$ given by lemma 3.1., it is sufficient to show that

$$\sqrt{n}(V_n - \tilde{V}_n) = o_P(1) \Rightarrow \sqrt{n}(t_n(V_n) - t_n(\tilde{V}_n)) = o_P(1).$$

Let us write the following decomposition:

$$\begin{aligned} \sqrt{n}(t_n(V_n) - t_n(\tilde{V}_n)) &= A_{1n} + A_{2n} \\ \text{with } A_{1n} &= \sqrt{n}\left[(G'V_n^{-1}G)^{-1/2} - (G'\tilde{V}_n^{-1}G)^{-1/2}\right]G'V_n^{-1}Z_{1n}, \\ A_{2n} &= (G'\tilde{V}_n^{-1}G)^{-1/2}G'\left[\sqrt{n}(V_n^{-1} - \tilde{V}_n^{-1})\right]Z_{1n}. \end{aligned}$$

Since $Z_{1n} = O_P(1)$,we clearly have

$$\sqrt{n}(V_n - \tilde{V}_n) = o_P(1) \Rightarrow A_{1n} = o_P(1) = A_{2n}.$$

Hence the announced results (i), (ii) and (iii). ■