Goodness of fit for the constancy of a classical statistical model over time

Alex J. Koning

Econometric Institute
Erasmus University Rotterdam
P.O. Box 1738
NL-3000 DR Rotterdam
The Netherlands
koning@few.eur.nl

Econometric Institute Report EI-9959/A

Abstract

The classical statistical model relates to n independent random variables having a common distribution. In this paper we consider the situation where the common distribution involves an unknown parameter, and where at time $0 \leq t \leq 1$ only the first $[nt]$ random variables are observed. The innovation approach is used to derive goodness of fit processes which especially detect alternatives under which the unknown parameter does not remain constant, but varies over time.

The behaviour of these processes is investigated under the null hypothesis as well as under alternative hypotheses. Limiting Pitman efficacies of supremum type tests based on these processes are evaluated. Fixed change point alternative hypotheses and smooth alternative hypotheses receive additional treatment.

The methods are exemplified using covariance structure models, especially Gaussian graphical models.

Key Words: *Goodness of fit tests, parameter constancy, Pitman efficacies, Gaussian graphical models.*

1 Introduction

The classical statistical model relates to random variables $X_1, \ldots, X_n$ which are independent and have a common distribution. In most practical situations, the classical statistical model is not fully specified, but contains an unknown p-dimensional parameter, say θ, which is an element of the parameterspace Θ. In order to apply the model, the unknown parameter has to be estimated.

When the adequacy of the classical statistical model to describe a particular data set needs to be assessed, the presence of estimated parameters becomes a problem due to the complicating effect estimation has on the distribution of test statistics. The following quote is from Pollard (1984).

> The interest aroused when Durbin (1973) applied weak convergence methods to get limit distributions for statistics analogous to those of Kolmogorov and Smirnov, but with estimated parameters, died down when the intractable limit processes asserted themselves.

A way to avoid this problem was offered in Khmaladze (1981), which advocated in a different context the use of tests based on functionals applied to a martingale with respect to the so-called parametric filtration; we shall refer to such a martingale as a parametric martingale. The parametric filtration is the smallest filtration which contains the elements of the natural filtration as well as the events involving the estimator. Loosely speaking, the natural filtration describes for each t in some indexset $\mathcal{T}$ which events concerning the random variables $X_1, \ldots, X_n$ can be observed at time t; the parametric filtration allows some sort of "data-peeking", since at every moment in time we are also able to observe the estimator.

Khmaladze's exact innovation approach enables the use of martingale central limit theory to show that a functional of a parametric martingale converges weakly to the same functional applied to the Wiener process on $\mathcal{T}$. Subsequently, the rich theory of the well-studied Wiener process can be used to derive the limit distribution of the parametric martingale functional.

The "data-peeking" phenomenon inherent in Khmaladze's approach has met some criticism, since in some situations data-peeking seems very unnatural; for instance, in filtering. However, a vast majority of statistical analyses are performed after all data have been gathered. For these analyses the actual form of the filtration is merely a technical matter; in fact, even an artificial filtration may be used (Khmaladze, 1993).

A more serious criticism concerns the intractability of the parametric martingale, which often motivates the use of an approximation of the parametric martingale on the basis of a representation of its limit process. Unfortunately, the approximation often appears to be complex. This problem featured already prominently in Khmaladze (1981), where a situation not uncommon in survival analysis (Andersen et al.; 1993) was considered: the natural filtration only allows the observation at time t of those random variables which take value not exceeding t.

Several examples of Khmaladze's approximate innovation approach approach exist in literature, see for instance paragraph VI.3.3.4 in Andersen et al (1993). In contrast, Khmaladze's exact innovation approach has not received much attention. In Koning (1994) an example of the exact innovation approach is given; also Broström (1997) contains in effect an example, although the ideas of Khmaladze are not referred to. However, these examples are limited in scope since they only concern very specific models.

In this paper a new and more extensive example of Khmaladze's exact innovation approach is given, which may provide more insight into the basic ideas of Khmaladze (1981). The example concerns the classical statistical model, and involves a natural filtration that only allows the observation at time t of the random variables $X_1, \ldots, X_{[nt]}$, an observational scheme not uncommon in time series; we say that the natural filtration is generated by $X_1, \ldots, X_{[nt]}$. The parametric filtration allows at time t the observation of the estimator as well; we say that the parametric filtration is generated by the estimator and $X_1, \ldots, X_{[nt]}$.

The parametric filtration naturally leads to the consideration of the so-called innovation score process, which is a parametric martingale; that is, a martingale with respect to the parametric filtration. Although the innovation score process is the primary object of interest, this paper contains intermediate results concerning the behaviour of the maximum likelihood estimator and the estimated score process which have not appeared in literature before.

Apart from theoretical relevance, there is also direct practical relevance. The resulting new tests are in particular sensitive to alternatives under which the random variables $X_1, \ldots, X_n$ are still independent, but do not have a common distribution anymore. Thus, they can be used to assess whether the parameters of the statistical model remain constant over time. The test based on the statistic $T_\nu \circ \left(R * \breve{\Psi}^n\right)$, to be introduced in Section 6, turns out to be optimal under alternatives which earlier received attention in Page (1957) and Bissell (1984, 1986).

The structure of the paper is as follows. The null hypothesis and alternative hypothesis theory is developed in Section 2 and Section 3 respectively. Local alternatives receive extra attention in Section 4. The theory of Sections 2–4 is used in Section 5 to evaluate limiting Pitman efficacies of supremum type tests based on the processes under study. The behaviour of these tests under fixed change point alternatives and smooth alternatives is investigated in Section 6 and Section 7 respectively.

Section 8 describes the application to covariance structure models, which is exemplified in Section 9 by means of the mathematics marks data. Appendix A contains the proofs of two inequalities given in Section 3.

2 Null hypothesis theory

In this section a parametric martingale is derived by compensating the estimated score process under the assumption that the parameter of the statistical model remains constant; we shall refer to this situation as "the null hypothesis".

Consider a sample $X_1, \ldots, X_n$ of independent q-dimensional random variables having common density function $f_\theta(t)$. Suppose that at time $0 \leq t \leq 1$ we have only observed $X_1, \ldots, X_{[nt]}$, where $[nt]$ denotes the largest integer not exceeding nt.

The score process at time t is given by

$$\Psi_t^n = n^{-1/2} \left.\frac{\partial}{\partial\vartheta}\right|_{\vartheta=\theta} \sum_{i=1}^{[nt]} \log f_\vartheta(X_i) = n^{-1/2} \sum_{i=1}^{[nt]} \rho_\theta(X_i),$$

where

$$\sum_{i=1}^{[nt]} \log f_\theta(X_i)$$

is the log-likelihood at time t, and

$$\rho_\theta(x) = \left.\frac{\partial}{\partial\vartheta}\right|_{\vartheta=\theta} \log f_\vartheta(x)$$

is the classical score function. Observe that the classical score function and the score process are p-dimensional.

Let

$$\Sigma = \int_0^1 \rho_\theta(s) \left(\rho_\theta(s)\right)^T f_\theta(s) ds$$

be the Fisher information matrix evaluated at θ, and observe that the dependence of Σ on θ is not reflected in notation. If the Fisher information matrix Σ exists, then the Donsker theorem (Shorack and Wellner, 1986, p. 53) implies that Ψ_t^n converges weakly to a p-dimensional Gaussian process Ψ with expectation function identical to zero and covariance function satisfying

$$E\left(\Psi_s \Psi_t^T\right) = (s \wedge t)\Sigma. \tag{1}$$

The maximum likelihood estimator $\hat{\theta}^n$ is obtained by solving the likelihood equations

$$\left.\frac{\partial}{\partial\vartheta}\right|_{\vartheta=\hat{\theta}^n} \sum_{i=1}^{n} \log f_\vartheta(X_i) = 0. \tag{2}$$

Define the estimated score process $\hat{\Psi}^n$ by

$$\hat{\Psi}_t^n = n^{-1/2} \sum_{i=1}^{[nt]} \rho_{\hat{\theta}^n}(X_i)$$

then the likelihood equations (2) directly imply $\hat{\Psi}_1^n = 0$: maximum likelihood estimation actually selects the parameter value which makes the estimated score process ultimately return to zero. In Horváth and Parzen (1994) the estimated score process is called Fisher-score change process; see also Csörgő and Horváth (1997, p. 57). The use of the estimated score process in testing the constancy of model parameters is discussed in Hjort and Koning (1999).

As in Horváth and Parzen (1994), in typical cases one may show that the estimated score process $\hat{\Psi}^n$ converges weakly to the p-dimensional Gaussian process

$$\hat{\Psi}_t = \Psi_t - t\Psi_1.$$

From (1) we immediately derive that $\hat{\Psi}$ is a Gaussian process with expectation function identical to zero and covariance function

$$E\left(\hat{\Psi}_s \hat{\Psi}_t^T\right) = \{(s \wedge t) - st\}\,\Sigma.$$

Thus, the estimated score process $\hat{\Psi}^n$ has a more intricate asymptotic structure than the original score process Ψ^n. As in Durbin (1973), this may become a problem in the use of the estimated score process to assess the goodness of fit of a statistical model. As was mentioned in the introduction, this problem may be avoided by using the exact innovation approach proposed in Khmaladze (1981). In the situation considered in this paper, the exact innovation approach involves computing the conditional expectation

$$E\left(\rho_{\hat{\theta}^n}(X_i) \mid \hat{\theta}^n, X_1, X_2, \ldots, X_{i-1}\right).$$

Due to the fact that $X_1, \ldots, X_n$ are independent, the log-likelihood is symmetric in $X_1, \ldots, X_n$ for any θ, which implies that the maximum likelihood estimator $\hat{\theta}^n$ is also symmetric in $X_1, \ldots, X_n$. It follows that

$$E\left(\rho_{\hat{\theta}^n}(X_\ell) \mid \hat{\theta}^n, X_1, X_2, \ldots, X_{i-1}\right) = E\left(\rho_{\hat{\theta}^n}(X_i) \mid \hat{\theta}^n, X_1, X_2, \ldots, X_{i-1}\right)$$

for $\ell \geq i$, and hence

$$\begin{aligned}
&E\left(\rho_{\hat{\theta}^n}(X_i) \mid \hat{\theta}^n, X_1, X_2, \ldots, X_{i-1}\right) \\
&\quad = (n-i+1)^{-1} \sum_{\ell=i}^{n} E\left(\rho_{\hat{\theta}^n}(X_\ell) \mid \hat{\theta}^n, X_1, X_2, \ldots, X_{i-1}\right) \\
&\quad = (n-i+1)^{-1} E\left(\sum_{\ell=i}^{n} \rho_{\hat{\theta}^n}(X_\ell) \mid \hat{\theta}^n, X_1, X_2, \ldots, X_{i-1}\right) \\
&\quad = (n-i+1)^{-1} E\left(-\sum_{\ell=1}^{i-1} \rho_{\hat{\theta}^n}(X_\ell) \mid \hat{\theta}^n, X_1, X_2, \ldots, X_{i-1}\right) \\
&\quad = -\,(n-i+1)^{-1} \sum_{\ell=1}^{i-1} \rho_{\hat{\theta}^n}(X_\ell).
\end{aligned}$$

Subtracting the conditional expectation from $\rho_{\hat{\theta}^n}(X_i)$ yields the random variable

$$\breve{X}_i = \rho_{\hat{\theta}^n}(X_i) + (n-i+1)^{-1} \sum_{\ell=1}^{i-1} \rho_{\hat{\theta}^n}(X_\ell) = \rho_{\hat{\theta}^n}(X_i) - (n-i+1)^{-1} \sum_{\ell=i}^{n} \rho_{\hat{\theta}^n}(X_\ell),$$

the innovation part of $\rho_{\hat{\theta}^n}(X_i)$. Since

$$\text{cov}\left(\rho_{\hat{\theta}^n}(X_i), \rho_{\hat{\theta}^n}(X_j)\right) = \begin{cases} \text{var}\left(\rho_{\hat{\theta}^n}(X_1)\right) & \text{for } i = j, \\ \\ -(n-1)^{-1}\,\text{var}\left(\rho_{\hat{\theta}^n}(X_1)\right) & \text{for } i \neq j \end{cases}$$

from symmetry and the fact that $\sum_{i=1}^{n} \rho_{\hat{\theta}^n}(X_i)$ is equal to zero, it follows that

$$\text{cov}\left(\breve{X}_i, \breve{X}_j\right) = \begin{cases} \dfrac{n}{n-1}\dfrac{n-i}{n-i+1}\text{var}\left(\rho_{\hat{\theta}^n}(X_1)\right) & \text{if } i = j \\ \\ 0 & \text{if } i \neq j \end{cases}$$

The innovation score process is now defined by

$$\breve{\Psi}_t^n = n^{-1/2} \sum_{i=1}^{[nt]} c_{n,i} \breve{X}_i,$$

with

$$c_{n,i} = \left(\frac{n-i+1}{n-i}\right)^{1/2}.$$

The innovation score process is a p-dimensional martingale with respect to the parametric filtration, generated by the random variables $\hat{\theta}^n, X_1, X_2, \ldots, X_{[nt]}$. Moreover, we have

$$E\left(\breve{\Psi}_s^n \left(\breve{\Psi}_t^n\right)^T\right) = (s \wedge t)\frac{n}{n-1}\text{var}\left(\rho_{\hat{\theta}^n}(X_1)\right).$$

Typically, the right hand side of this equation tends to $(s \wedge t)\Sigma$ as n tends to infinity. It follows that the innovation score process $\breve{\Psi}^n$ converges weakly to a p-dimensional Gaussian process $\breve{\Psi}$ with expectation function identical to zero and covariance function

$$E\left(\breve{\Psi}_s \breve{\Psi}_t^T\right) = (s \wedge t)\Sigma.$$

Observe that the processes Ψ^n and $\breve{\Psi}^n$ have the same asymptotic distribution.

Now that we have found one example of a parametric martingale, many more examples are easily derived. Under conditions on the stochastic integrand, it follows from standard martingale theory that stochastic integrals with respect to the innovation score process are also parametric martingales. For instance, let L be a deterministic scalar function satisfying

$$\int_0^1 (L_s)^2\, ds = 1, \tag{3}$$

then the stochastic integral $L * \breve{\Psi}^n$ of L with respect to $\breve{\Psi}^n$, defined by

$$\left(L * \breve{\Psi}^n\right)_t = L_t \breve{\Psi}_t^n - \int_0^t \breve{\Psi}_s^n dL_s = n^{-1/2} \sum_{i=1}^{[nt]} L_{i/n} \breve{X}_i c_{n,i},$$

is also a martingale with respect to the parametric filtration generated by the random variables $\hat{\theta}^n, X_1, X_2, \ldots, X_{[nt]}$, and converges weakly to a p-dimensional Gaussian process $L * \breve{\Psi}$ with expectation function identical to zero and covariance function

$$E\left\{\left(L * \breve{\Psi}\right)_s \left(L * \breve{\Psi}\right)_t^T\right\} = \int_0^{s \wedge t} \left(L_u\right)^2 du \Sigma.$$

Thus, the simple asymptotic structure of the innovation score process easily extends to stochastic integrals with respect to the innovation score process. Stochastic integrals are convenient, since the freedom in the choice of the integrand L provides the opportunity to tune the behaviour of the resulting goodness of fit process. For instance, in Section 5 a choice of the integrand is given which maximizes the limiting Pitman efficacy of the supremum type goodness of fit test based on the stochastic integral with respect to a specific alternative.

Observe that that it is not difficult to compute the innovation parts once all $\rho_{\hat{\theta}^n}(X_i)$'s are known. Moreover, the $\rho_{\hat{\theta}^n}(X_i)$'s usually become available as a by-product of maximum likelihood estimation via solving the likelihood equations (2). This makes implementation of statistical techniques based on $\hat{\Psi}^n$ or $\breve{\Psi}^n$ rather easy.

If $X_1, \ldots, X_n$ are univariate and follow a normal distribution with expectation θ and variance 1, then $\rho_{\hat{\theta}^n}(X_i)$ coincides with the i^{th} least squares residual $X_i - n^{-1} \sum_{\ell=1}^n X_\ell$; refer to Section 8 for more details. It follows that both the least squares cumulative sum in Ploberger and Krämer (1992) and the standardized time series in Schruben (1982, 1983) coincide with our estimated score process $\hat{\Psi}^n$. Moreover, the right-hand side of the equation

$$c_{n,i} \breve{X}_i = \left(\frac{n-i}{n-i+1}\right)^{1/2} \left(X_i - (n-i+1)^{-1} \sum_{\ell=i}^n X_\ell\right).$$

may be interpreted as the $(n-i+1)^{th}$ recursive residual when the order of the sample is reversed. Hence, the recursive cumulative sum in Brown et al (1975) is strongly related to $\breve{\Psi}_1^n - \breve{\Psi}_t^n$, the increment of the innovation score process over the interval $(t, 1]$; see also Krämer et al (1988). However, it should be noted that the estimation method and general context in Krämer et al (1988) and Ploberger, Krämer (1992) differ from ours.

If $X_1, \ldots, X_n$ are univariate and follow a Bernoulli distribution with expectation θ, then $\breve{X}_i$ coincides with $Z_i^{(n)}$ in Broström (1997). Hence, the martingale approach in Broström (1997) is related to the exact innovation approach.

3 The alternative hypothesis

In this section we study the behaviour of the estimated and innovation score processes when the random variables $X_1, \ldots, X_n$ are still independent and belong to the same family of distributions, but do not necessarily have a common distribution anymore; we shall refer to this situation as "the alternative hypothesis", although the null hypothesis is actually a special case. The density function of the random variable X_i is assumed to be of the form $f_{\theta_i}(t)$.

To avoid technical details and additional notation, the results in this section are presented as descriptions of the rescaled estimated score process $\Sigma^{-1/2}\hat{\Psi}_t^n$ and the rescaled innovation score process $\Sigma^{-1/2}\breve{\Psi}_t^n$ rather than as descriptions of $\hat{\Psi}_t^n$ and $\breve{\Psi}_t^n$ themselves. Since the largest eigenvalue of $\Sigma^{1/2}$ is finite, approximations of the rescaled processes directly lead to approximations of the original processes. The proofs of Lemma 1 and Inequalities 1 and 2 are deferred to Appendix A.

Define $d_{\vartheta',\vartheta}$ as $\left|\Sigma^{1/2}(\vartheta' - \vartheta)\right|$, where Σ is as before the Fisher information matrix evaluated at θ; one may think of θ as the common value of the θ_i's under some null hypothesis in the vicinity of the alternative hypothesis. Assume that $\rho_\theta(x)$ satisfies the following condition.

Condition 1 *Let $\bar{\theta}$ denote $n^{-1}\sum_{\ell=1}^n \theta_\ell$. For every $r > 0$ and every $\theta \in \Theta$ satisfying $\max_{\ell=1,\ldots,n} d_{\theta,\theta_\ell} < r$, there exist a random variable $U_{n,r}$ such that p-dimensional vector*

$$R(X_i, \vartheta, \vartheta') = \Sigma^{-1/2}\rho_{\vartheta'}(X_i) - \Sigma^{-1/2}\rho_{\vartheta}(X_i) + \Sigma^{1/2}\left(\vartheta' - \vartheta\right)$$

satisfies

$$\max_{\ell=1,\ldots,n}\left| n^{-1}\sum_{i=\ell}^{n} R(X_i, \vartheta', \vartheta)\right| \leq U_{n,r} d_{\vartheta',\vartheta}, \tag{4}$$

for every ϑ, ϑ' satisfying $d_{\vartheta',\bar{\theta}} < r$ and $d_{\vartheta',\bar{\theta}} < r$, and

$$\max_{\ell=1,\ldots,n}\left| n^{-1}\sum_{i=\ell}^{n} R(X_i, \theta_i, \bar{\theta})\right| \leq U_{n,r} \max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}}. \tag{5}$$

As the next lemma illustrates, $U_{n,r}$ tyically converges in probability to a positive constant depending on r as n tends to infinity. Moreover, if r tends to zero, then the limit of $U_{n,r}$ tends to zero also. Regularity conditions (A)–(D) in Section 6.4 of Lehmann (1991) ensure consistency and asymptotic normality of maximum likelihood estimators.

Lemma 1 *Let $f_\vartheta(x)$ satisfies regularity conditions (A)–(D) in Section 6.4 of Lehmann (1991), and let $\dot{\rho}_{\theta_i}(X_i)$ denote the derivative of $\rho_{\vartheta'}(X_i)$ with respect to ϑ', evaluated in $\vartheta' = \theta_i$. Let $S_{1i,jk}$ and $S_{2i,jk}$ denote the (j,k) elements of $\Sigma^{-1/2}\left\{\dot{\rho}_{\theta_i}(X_i) - E\left(\dot{\rho}_{\theta_i}(X_i)\right)\right\}$ and $\Sigma^{-1/2}\left\{E\left(\dot{\rho}_{\theta_i}(X_i)\right) + \Sigma\right\}$, respectively. If there exists $0 \leq \epsilon \leq 1$ such that $n^{-1}\sum_{i=1}^n\sum_{j=1}^p\sum_{k=1}^p E\left(|S_{1i,jk}|^{1+\epsilon}\right)$ remains bounded for n tending to infinity, then Condition 1 holds with*

$$U_{n,r} = 3cr + n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p}|S_{2i,jk}| + U_n^*,$$

where c is a fixed constant, and U_n^* is a $O_P(n^{-\epsilon/(1+\epsilon)})$ random variable with distribution not depending on r.

Note that under the null hypothesis $-E\left(\dot{\rho}_{\theta_i}(X_i)\right)$ coincides with Σ, and hence $U_{n,r} = 3cr + U_n^*$. The well-known C_r-inequality (Shorack and Wellner, 1986, p. 843) may be helpful in establishing a bound on $E\left(|S_{1i,jk}|^{1+\epsilon}\right)$.

Inequality 1 *Suppose Condition 1 holds. Define W_n^* by*

$$W_n^* = n^{-1/2} \sum_{i=1}^{n} \Sigma^{-1/2} \rho_{\bar{\theta}}(X_i).$$

For any radius $r > 0$, given the event that

$$U_{n,r} < \tfrac{1}{3} \;\; \textit{and} \;\; |W_n^*| < \tfrac{1}{2} r n^{1/2} \tag{6}$$

both hold, there exists within the ball $\Theta_{n0} = \left\{\vartheta : d_{\vartheta,\bar{\theta}} < r\right\}$ a solution $\hat{\theta}^n$ to the likelihood equations (2) which satisfies

$$\left| n^{1/2} \Sigma^{1/2} \left(\hat{\theta}^n - \bar{\theta}\right) - W_n^* \right| < \tfrac{3}{2} U_{n,r} \left|W_n^*\right|.$$

Inequality 1 shows that the behaviour of the maximum likelihood estimator $\hat{\theta}^n$ is governed by the random variables $U_{n,r}$ and W_n^*. Note that if we apply Inequality 1 with a shrinking radius r_n, then the convergence in probability of U_{n,r_n} to zero yields that $n^{1/2}\Sigma^{1/2}\left(\hat{\theta}^n - \bar{\theta}\right)$ is asymptotically equivalent to W_n^*.

The random variable W_n^* is in turn related to the random variable

$$W_n = n^{-1/2} \sum_{i=1}^{n} \Sigma^{-1/2} \rho_{\theta_i}(X_i),$$

since (5) implies

$$\begin{aligned}
|W_n - W_n^*| &= \left| n^{-1/2} \sum_{i=1}^{n} \Sigma^{-1/2} \rho_{\theta_i}(X_i) - n^{-1/2} \sum_{i=1}^{n} \Sigma^{-1/2} \rho_{\bar{\theta}}(X_i) \right| \\
&= \left| n^{-1/2} \sum_{i=1}^{n} \left(\Sigma^{-1/2} \rho_{\theta_i}(X_i) - \Sigma^{-1/2} \rho_{\bar{\theta}}(X_i) + \Sigma^{1/2} \left(\theta_i - \bar{\theta}\right) \right) \right| \\
&= \left| n^{-1/2} \sum_{i=1}^{n} R(X_i, \theta_i, \bar{\theta}) \right| \\
&\leq n^{1/2} \max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}} U_{n,r}.
\end{aligned}$$

Usually, $\rho_{\theta_i}(X_i)$ has zero expectation vector; see Lindsey (1996, p. 188). It is assumed that W_n converges in distribution to some random variable. In most cases this can be shown by verifying the Lindeberg condition

$$\lim_{n\to\infty} n^{-1} E\left\{ \left(1_{\left\{h^T \rho_{\theta_i}(X_i) > \varepsilon n^{1/2}\right\}} h^T \rho_{\theta_i}(X_i) \right)^2 \right\} = 0$$

for every $\varepsilon > 0$ and every unit vector h. If this Lindeberg condition holds for every $\varepsilon > 0$ and every vector h, then W_n has a multivariate normal limit distribution with the null vector as expectation, and the identity matrix as covariance matrix. In other circumstances a multivariate stable limit distribution may arise.

The Lindeberg condition is implied by the existence of $0 < \epsilon \leq 1$ such that $n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p} E\left(\left|S_{1i,jk}\right|^{1+\epsilon}\right)$ remains bounded for n tending to infinity, where $S_{1i,jk}$ is as in Lemma 1.

Inequality 1 implies that

$$n^{1/2} d_{\hat{\theta}^n,\bar{\theta}} = n^{1/2}\left|\Sigma^{1/2}\left(\hat{\theta}^n - \bar{\theta}\right)\right|$$

is bounded by $\frac{2}{3}|W_n^*|$ if the event (6) holds; this result also follows from (16) in the proof of Inequality 1. By noting that $d_{\hat{\theta}^n,\theta_\ell}$ is bounded by $d_{\hat{\theta}^n,\bar{\theta}} + d_{\bar{\theta},\theta_\ell}$, we obtain the following corollary to Inequality 1.

Corollary 1 *Suppose Condition 1 holds. If the event (6) holds, then*

$$n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell} \leq \tfrac{2}{3}\left\{|W_n| + n^{1/2}\max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}}\left(1+U_{n,r}\right)\right\}.$$

Inequality 2 *Suppose Condition 1 holds, and the maximum likelihood estimator $\hat{\theta}^n$ exists. Let*

$$\hat{\mu}_i = \Sigma\left(\theta_i - \bar{\theta}\right), \text{ and } \breve{\mu}_i = \Sigma\left(\theta_i - \frac{1}{n-i+1}\sum_{\ell=i}^{n}\theta_\ell\right).$$

Then

$$\sup_{t\in[0,1]}\left|\Sigma^{-1/2}\hat{\Psi}_t^n - n^{-1/2}\sum_{i=1}^{[nt]}\Sigma^{-1/2}\left\{\hat{\mu}_i + \rho_{\theta_i}(X_i) - n^{-1}\sum_{\ell=1}^{n}\rho_{\theta_\ell}(X_\ell)\right\}\right|$$
$$\leq 4n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell}U_{n,r},$$

and

$$\sup_{t\in[0,1]}\left|\Sigma^{-1/2}\breve{\Psi}_t^n - n^{-1/2}\sum_{i=1}^{[nt]}c_{n,i}\Sigma^{-1/2}\left\{\breve{\mu}_i + \rho_{\theta_i}(X_i) - \frac{1}{n-i+1}\sum_{\ell=i}^{n}\rho_{\theta_\ell}(X_\ell)\right\}\right|$$
$$\leq 2(2+\log n)n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell}U_{n,r}.$$

One may view $n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell}$ as indicative of the order of magnitude of $\Sigma^{-1/2}\hat{\Psi}_t^n$ and $\Sigma^{-1/2}\breve{\Psi}_t^n$. Hence, the accuracy of the bound for the approximations for $\Sigma^{-1/2}\hat{\Psi}_t^n$ and $\Sigma^{-1/2}\breve{\Psi}_t^n$ in Inequality 2 is in essence given by $U_{n,r}$ and $(\log n)U_{n,r}$, respectively.

Since the largest eigenvalue of $\Sigma^{1/2}$ is finite, Inequality 2 implies that if $U_{n,r}$ and $(\log n)U_{n,r}$ are small, then

- the estimated score process $\hat{\Psi}_t^n$ is approximately equal to a random part

  $$n^{-1/2} \sum_{i=1}^{[nt]} \left\{ \rho_{\theta_i}(X_i) - n^{-1} \sum_{\ell=1}^{n} \rho_{\theta_\ell}(X_\ell) \right\}$$

  with zero expectation vector plus a deterministic drift $n^{-1/2} \sum_{i=1}^{[nt]} \hat{\mu}_i$,

- the innovation score process $\breve{\Psi}_t^n$ is approximately equal to a random part

  $$n^{-1/2} \sum_{i=1}^{[nt]} c_{n,i} \left(\rho_{\theta_i}(X_i) - \frac{1}{n-i+1} \sum_{j=i}^{n} \rho_{\theta_j}(X_j) \right)$$

  with zero expectation vector plus a deterministic drift $n^{-1/2} \sum_{i=1}^{[nt]} c_{n,i} \breve{\mu}_i$,

- the stochastic integral $\left(L * \breve{\Psi}^n \right)_t$ is approximately equal to a random part

  $$n^{-1/2} \sum_{i=1}^{[nt]} L_{i/n} \left(\rho_{\theta_i}(X_i) - \frac{1}{n-i+1} \sum_{j=i}^{n} \rho_{\theta_j}(X_j) \right) c_{n,i}$$

  with zero expectation vector plus a deterministic drift $n^{-1/2} \sum_{i=1}^{[nt]} L_{i/n} c_{n,i} \breve{\mu}_i$.

4 Local alternatives

Further insight into the behaviour of the estimated score process $\hat{\Psi}_t^n$ and the innovation score process $\breve{\Psi}_t^n$ can be gained by assuming that $X_1, \ldots, X_n$ is in fact the n^{th} row of a triangular scheme of independent random vectors, and that X_i has density $f_{\theta_i}(x)$ with

$$\theta_i = \theta_0 + c_n n^{-1/2} \xi \left(\frac{i}{n} \right) \tau, \tag{7}$$

where c_n a given scalar depending on n, $\xi(t)$ is a given scalar function defined on $[0, 1]$, and τ is a given p-dimensional vector.

If $c_n n^{-1/2}$ tends to zero, then we shall refer to (7) as a local alternative. For local alternatives, we may apply Inequality 2 with r replaced by r_n, where r_n tends to zero if n goes to infinity; we shall refer to r_n as a shrinking radius. In the light of (6) we should require that W_n is $o_P(n^{1/2} r_n)$. Thus, $n^{1/2} r_n$ should tend to infinity if W_n remains bounded in probability.

Under a local alternative, it follows from the martingale central limit theorem in Rootzén (1980) that if the Lindeberg type condition mentioned in Section 3 is satisfied, then the random part of the estimated score process $\hat{\Psi}^n$ converges in distribution to $\hat{\Psi}$, the limit in distribution of the estimated score process under the null hypothesis described in Section 2. Similarly, under a local alternative the random parts of $\breve{\Psi}^n$ and $L * \breve{\Psi}^n$ respectively converge in distribution to $\breve{\Psi}$ and $L * \breve{\Psi}$, where $\breve{\Psi}$ is the limit in distribution of $\breve{\Psi}^n$ under the null hypothesis. In Horváth and Parzen (1994) only the asymptotic null behaviour of $\hat{\Psi}^n$ is considered, assuming pointwise smoothness conditions. Pointwise smoothness conditions are less general than stochastic differentiability.

Now, let us use the behaviour of $n^{-1/2}\max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}}$ to distinguish between fast, contiguous and slow local alternatives. For fast local alternatives we have that $n^{-1/2}\max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}}$ tends to zero. Thus, if $(\log n)U_{n,r_n}$ tends to zero in probability and W_n remains bounded in probability, we have by Corollary 1 that $n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell}$ remains bounded in probability. It now follows by Inequality 2 that $\hat{\Psi}^n$, $\breve{\Psi}^n$ and $\left(L * \breve{\Psi}^n\right)_t$ exhibit the same limiting behavior as under the null hypothesis.

For contiguous alternatives we have that $n^{-1/2}\max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}}$ tends to a non-zero limit, δ say. If $(\log n)U_{n,r_n}$ tends to zero in probability and W_n remains bounded in probability, we again have by Corollary 1 that $n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell}$ remains bounded in probability. In this case the deterministic drifts of $\hat{\Psi}^n$, $\breve{\Psi}^n$ and $\left(L * \breve{\Psi}^n\right)_t$ tend to $\delta\hat{\Delta}_t\Sigma\tau$, $\delta\breve{\Delta}_t\Sigma\tau$ and $\delta\left(L * \breve{\Delta}\right)_t\Sigma\tau$ respectively, where

$$\hat{\Delta}_t = \int_0^t \left\{\xi(s) - \int_0^1 \xi(u)\,du\right\} ds,$$

$$\breve{\Delta}_t = \int_0^t \left\{\xi(s) - \frac{1}{1-s}\int_s^1 \xi(u)\,du\right\} ds$$

and

$$\left(L * \breve{\Delta}\right)_t = \int_0^t L_s \left\{\xi(s) - \frac{1}{1-s}\int_s^1 \xi(u)\,du\right\} ds.$$

For slow local alternatives we have that $n^{-1/2}\max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}}$ tends to infinity. If $(\log n)U_{n,r_n}$ tends to zero in probability and W_n remains bounded in probability, we have by Corollary 1 that $n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell}$ is $O_P(n^{-1/2}\max_{\ell=1,\ldots,n} d_{\theta_\ell,\bar{\theta}})$. The random fluctuations of the processes under consideration become negligible with respect to their deterministic drifts as n increases. Hence, one may show that the $\hat{\Psi}^n$, $\breve{\Psi}^n$ and $\left(L * \breve{\Psi}^n\right)_t$ are approximated by $c_n\hat{\Delta}_t\Sigma\tau$, $c_n\breve{\Delta}_t\Sigma\tau$ and $c_n\left(L * \breve{\Delta}\right)_t\Sigma\tau$, respectively. Plots of $\hat{\Delta}_t$ and $\breve{\Delta}_t$ versus t give an impression of the patterns showing up in graphical displays of the estimated and innovation score processes under a particular slow alternative hypothesis.

For $\hat{\Psi}^n$ the results in this section remain true if U_{n,r_n} tends to zero in probability. The requirement that Condition 1 holds with U_{n,r_n} tending to zero in probability is related to the notion of "stochastic differentiability"; see Pollard (1985). For ways of verifying stochastic differentiability the reader is referred to Pollard (1985) and Hoffmann-Jørgensen (1994b, Chapter 14).

5 Limiting Pitman efficacies of supremum type tests

Alternatives with $c_n \equiv 1$ are called contiguous, and are important in the evaluation of Pitman efficacies. Consequently, the functions $\hat{\Delta}$ and $\breve{\Delta}$ show up in the Pitman efficiencies of tests based on functionals of the estimated score process $\hat{\Psi}_t^n$ and the innovation score process $\breve{\Psi}_t^n$. In this section the limiting Pitman efficacies of the tests based on $T_\nu \circ \hat{\Psi}^n$, $T_\nu \circ \breve{\Psi}^n$ and $T_\nu \circ \left(L * \breve{\Psi}^n\right)$ are evaluated as an example. Here ν is a p-dimensional vector and the functional T_ν is defined by

$$T_\nu \circ \eta = \sup_{t \in [0,1]} \nu^T \eta_t$$

for any function η defined on $\mathcal{T}$. We shall refer to these tests as one-sided supremum type tests. For notational convenience we restrict ourselves to alternatives for which $\nu^T \Sigma \tau$ is equal to 1.

Supremum type tests are affiliated with a long tradition in goodness-of-fit; recall the quote from Pollard (1984): “statistics analogous to those of Kolmogorov and Smirnov”. Most popular are the two-sided versions, obtained by taking the absolute value of the inproduct $\nu^T \eta_t$ before suping. Since two-sided tests are less convenient for theoretical purposes due to problems related to their biasedness (Lehmann, 1986, Section 4.1), we concentrate in Sections 5–7 on one-sided tests. More on two-sided supremum type tests can be found in Section 8.

The rationale behind taking the inproduct $\nu^T \eta_t$ originates from the desire to construct p tests, each focusing on a different aspect of the model, which can be easily combined into one overall test. This is accomplished by setting $\nu_1, \nu_2, \ldots, \nu_p$ equal to the rows of some root of Σ^{-1}, the inverse of the Fisher information under the null hypothesis. The maximum of the resulting test statistics is a convenient overall test statistic: under the null hypothesis its limit distribution is easily derived, since the individual test statistics are asymptotically independent and have a common distribution. From a statistical perspective, the spectral root and the LU-root of Σ^{-1} are especially useful. A slight complication in this approach arises from the dependence of $\nu_1, \nu_2, \ldots, \nu_p$ on the unknown parameter; however, replacing the parameter by its estimator usually repairs this problem without affecting the asymptotic behaviour of the individual test statistics.

Our test statistics have asymptotic null hypothesis distributions which are nonnormal, but well-known. The asymptotic null hypothesis distribution of $T_\nu \circ \hat{\Psi}^n$ coincides with the distribution of the supremum of the one-dimensional standard Brownian bridge, which has cumulative distribution function

$$F_{\text{Weibull}(2^{-1/2},2)}(x) = 1 - \exp\left\{-2x^2\right\}$$

belonging to the Weibull distribution with scale parameter $2^{-1/2}$ and shape parameter 2. The asymptotic null hypothesis distribution of $T_\nu \circ \breve{\Psi}^n$ coincides with the distribution of the supremum of the one-dimensional standard Wiener process on the unit interval, which has cumulative distribution function

$$F_{\mathrm{HN}(0,1)}\left(x\right)=\left(2\pi\right)^{-1/2}\int_{-x}^{x}e^{-s^2/2}ds$$

belonging to the half-standardnormal distribution; the same holds true for the asymptotic null hypothesis distribution of $T_\nu\circ\left(L*\breve{\Psi}^n\right)$ if the integrand L satisfies (3).

Since supremum type test statistics are not asymptotically normal, their Pitman efficacy has the unfavourable property of depending on the size of the test. However, it can be shown that the Pitman efficacies of the tests based on $T_\nu\circ\hat{\Psi}^n$, $T_\nu\circ\breve{\Psi}^n$ and $T_\nu\circ\left(L*\breve{\Psi}^n\right)$ all tend from below to

$$\left\{2\sup_{t\in[0,1]}\hat{\Delta}_t\right\}^2,\quad\left\{\sup_{t\in[0,1]}\breve{\Delta}_t\right\}^2\quad\text{and}\quad\left\{\sup_{t\in[0,1]}\left(L*\breve{\Delta}\right)_t\right\}^2$$

respectively, as the sizes of the tests tend to zero. The factor 2 appearing in the asymptotic Pitman efficacy of the test based on $T_\nu\circ\hat{\Psi}^n$ is due to the fact that the maximum value of the variance function of the one-dimensional standard Brownian bridge is only a quarter of the maximum value of the variance function of the one-dimensional standard Wiener process on the unit interval, which is 1.

For a specific alternative, the efficacy of the test based on $T_\nu\circ\left(L*\breve{\Psi}^n\right)$ can be maximized by choosing L equal to

$$L_t^*=\frac{\xi\left(t\right)-\dfrac{1}{1-t}\displaystyle\int_t^1\xi\left(s\right)ds}{\left(\displaystyle\int_0^1\left\{\xi\left(s\right)-\dfrac{1}{1-s}\int_s^1\xi\left(u\right)du\right\}^2ds\right)^{1/2}};$$

an alternative expression for the optimal choice L^* of the integrand is obtained by using the identity

$$\xi\left(t\right)-\frac{1}{1-t}\int_t^1\xi\left(s\right)ds=-\frac{1}{1-t}\int_t^1\left(1-s\right)d\xi\left(s\right).$$

Observe that this choice satisfies (3); moreover, the limiting Pitman efficacy of the test based on $T_\nu\circ\left(L^**\breve{\Psi}^n\right)$ is equal to

$$\int_0^1\left(\xi\left(s\right)-\frac{1}{1-s}\int_s^1\xi\left(u\right)du\right)^2ds. \tag{8}$$

Using this expression, it can be shown that $T_\nu\circ\left(L^**\breve{\Psi}^n\right)$ has limiting Pitman efficiency 1 with respect to the generalized likelihood ratio test belonging to the specific alternative. Thus, for every specific alternative there exists a test based on the supremum of a certain parametric martingale which approaches the optimal parametric test in power.

The limiting Pitman efficiency results in this section generally extend to limiting approximate Bahadur efficiency results. If $U_{n,r}$ takes the form $cr + U_n^*$ where U_n^* is a $o_P((\log n)^{-1})$ random variable with distribution not depending on r or θ, then Condition III* in Wieand (1976) follows from Inequality 2. Hence, Theorem 1 in Kallenberg and Koning (1995) yields that for the test statistics under study the limiting Pitman and limiting approximate Bahadur concepts give the same answers. For $T_\nu \circ \hat{\Psi}^n$ this remains true if U_n^* is a $o_P(1)$ random variable with distribution not depending on r or θ. Under regularity conditions on $E\left(\dot{\rho}_{\theta_i}(X_i)\right)$, the special structure $cr + U_n^*$ of $U_{n,r}$ may be verified with the aid of Lemma 1.

The results above remain valid if T_ν is replaced by a Lipschitz functional; that is, a functional T for which there exists a finite constant C_T such that

$$|T \circ \eta_1 - T \circ \eta_2| \leq C_T \sup_{t\in[0,1]} |\eta_{1,t} - \eta_{2,t}|.$$

6 A fixed change point alternative hypothesis

Let R_t denote $2(1-t)$, and observe that $L = R$ satisfies (3). In this section we compare the behaviour of the supremum type test statistics $T_\nu \circ \hat{\Psi}^n$, $T_\nu \circ \breve{\Psi}^n$ and $T_\nu \circ \left(R * \breve{\Psi}^n\right)$ to the behaviour of the optimal supremum type test statistic $T_\nu \circ \left(L^* * \breve{\Psi}^n\right)$ under an alternative hypothesis under which the parameter of the classical model jumps suddenly from θ_0 to $\theta_0 - n^{-1/2} c_n \tau$ after $[nu]$ observations:

$$\xi\left(t\right) = \begin{cases} 0 & \text{if } t \leq u, \\ \\ -1 & \text{if } t > u; \end{cases}$$

see also Page (1957). As in the previous section, we restrict ourselves without loss of generality to alternatives for which $\nu^T \Sigma \tau = 1$ for notational convenience.

If the relative position u of the jump remains fixed, then using

$$\xi\left(t\right) - \frac{1}{1-t}\int_t^1 \xi\left(s\right) ds = \frac{1-u}{1-t} 1_{\{t\leq u\}},$$

it can be easily shown that

$$T_\nu \circ \hat{\Delta} = \left\{\sup_{0\leq t\leq u} (1-u)t\right\} + \left\{\sup_{u\leq t\leq 1} u(1-t)\right\} = u(1-u),$$

$$T_\nu \circ \breve{\Delta} = \sup_{0\leq t\leq u} \breve{\Delta}_t = \sup_{0\leq t\leq u} -(1-u)\log(1-t) = -(1-u)\log(1-u),$$

$$T_\nu \circ \left(R * \breve{\Delta}\right) = \sup_{0\leq t\leq u} \left(R * \breve{\Delta}\right)_t = \sup_{0\leq t\leq u} 2(1-u)t = 2u(1-u),$$

and, according to (8),

$$T_\nu \circ \left(L^* * \breve{\Delta}\right) = \left\{\int_0^u \left(-\frac{1-u}{1-s}\right)^2 ds\right\}^{1/2} = \left\{u(1-u)\right\}^{1/2}.$$

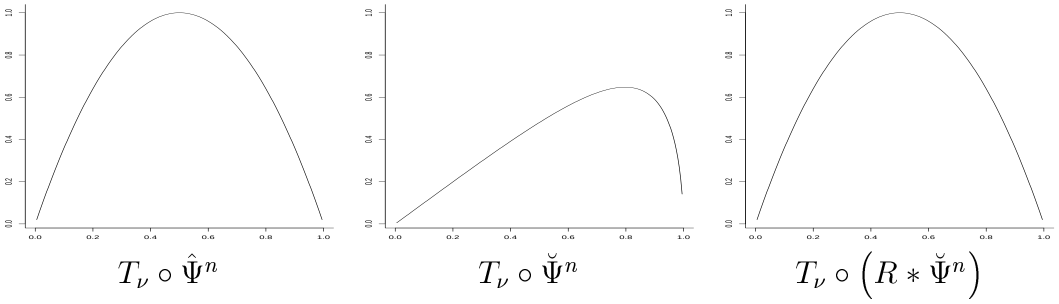


Figure 1: Fixed change point alternatives: limiting efficiencies of three supremum type tests as a function of the location u of the change point.

Figure 1 plots the efficiency of $T_\nu \circ \hat{\Psi}^n$ and the efficiency of $T_\nu \circ \breve{\Psi}^n$ with respect to the optimal supremum type test $T_\nu \circ \left(L^* * \breve{\Psi}^n\right)$ as a function of the location u of the change point.

Except for change points occurring in the last 20 percent of the sample, the supremum type test based on the estimated score process should be preferred to the supremum type test based on the innovation process. The supremum type test based on the estimated score process is optimal for change points occurring in exactly the middle of the the sample. The supremum type test based on the innovation process is never optimal.

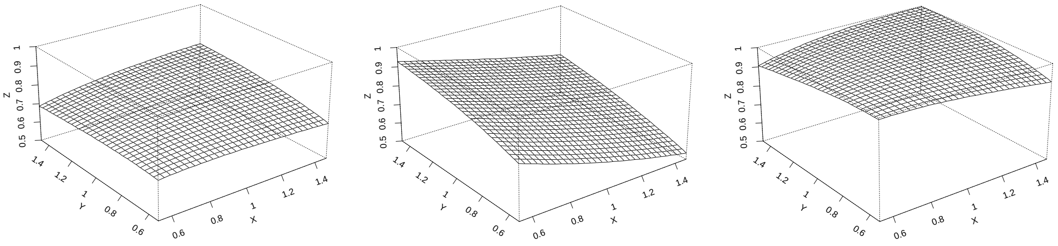
The conclusions with respect to $T_\nu \circ \hat{\Psi}^n$ also extend to $T_\nu \circ \left(R * \breve{\Psi}^n\right)$, since both test statistics share the same efficacy for every relative change point position u. It follows that the relative poor performance of the innovation process can be greatly improved by stochastic integration.

The two-sided version of $T_\nu \circ \hat{\Psi}^n$ appears as entry "LS-CUSUM" in Table 3 in Chu et al (1995), which present simulation results for fixed change points occurring in univariate normal distributions with variance 1. This table indeed underlines the high power of the supremum type test based on the estimated score process for change points alternatives with u between 30 and 70 percent.

7 A smooth alternative hypothesis

In this section we compare the behaviour of the supremum type test statistics $T_\nu \circ \hat{\Psi}^n$, $T_\nu \circ \breve{\Psi}^n$ and $T_\nu \circ \left(R * \breve{\Psi}^n\right)$ to the behaviour of the optimal supremum type test statistic $T_\nu \circ \left(L^* * \breve{\Psi}^n\right)$ under an alternative hypothesis under which the parameter of the classical model changes in a smooth way from θ_0 to $\theta_0 - n^{-1/2} c_n \tau$: we assume that $-\xi\left(\right)$ is equal to some cumulative distribution function G on the unit interval $(0, 1)$.

This situation may serve as a close approximation of the situation in which the distribution parameter changes suddenly from θ_0 to $\theta_0 - n^{-1/2} c_n \tau$ and the relative position π of the change point within the sample is random and follows G. To see this, observe that conditioning on the event that π takes the value u leads to the situation



$T_\nu \circ \hat{\Psi}^n$ $T_\nu \circ \breve{\Psi}^n$ $T_\nu \circ \left(R * \breve{\Psi}^n\right)$

Figure 2: "Beta" smooth alternatives: limiting efficiencies of three supremum type tests as a function of the parameters a and b.

considered in the previous example, with deterministic drifts of $\hat{\Psi}^n$, $\breve{\Psi}^n$ and $\left(L * \breve{\Psi}^n\right)_t$ depending on u. These deterministic drifts become random as the conditioning on the event that π takes the value u is removed, but the law of large numbers shows that it is still possible to approximate them by $c_n \hat{\Delta}_t \Sigma\tau$, $c_n \breve{\Delta}_t \Sigma\tau$ and $c_n \left(L * \breve{\Delta}\right)_t \Sigma\tau$ respectively.

The identity

$$\xi(t) - \frac{1}{1-t}\int_t^1 \xi(s)\, ds = \frac{1}{1-t}\int_t^1 (1-s)\, dG_s$$

facilitates the evaluation of $T_\nu \circ \hat{\Delta}$, $T_\nu \circ \breve{\Delta}$, $T_\nu \circ \left(R * \breve{\Delta}\right)$, and $T_\nu \circ \left(L^* * \breve{\Delta}\right)$.

Figure 2 assumes that G corresponds to a beta distribution with parameters a and b, and contains contour plots of the efficiencies of $T_\nu \circ \hat{\Psi}^n$, $T_\nu \circ \breve{\Psi}^n$ and $T_\nu \circ \left(R * \breve{\Psi}^n\right)$ with respect to the optimal supremum type test as a function of the parameters a and b of the beta distribution.

Observe that the efficiency of $T_\nu \circ \hat{\Psi}^n$ increases as a tends to ∞ while the ratio a/b tends to $\frac{1}{2}$. This is actually not very surprising, since in this situation G will become degenerate in $\frac{1}{2}$, corresponding to a fixed change point alternative with $u = \frac{1}{2}$; recall that $T_\nu \circ \hat{\Psi}^n$ was optimal with respect to this alternative.

The efficiency of $T_\nu \circ \breve{\Psi}^n$ increases as a becomes larger and b becomes smaller. This indicates the sensitivity of $T_\nu \circ \breve{\Psi}^n$ to change points occurring in the last part of the sample, since in this situation G will become more concentrated near 1.

The test statistic $T_\nu \circ \left(R * \breve{\Psi}^n\right)$ shows the strongest performance. It is optimal in the special case that $a = 1$ and $b = 1$, which corresponds to the situation where G is the uniform distribution on the unit interval $(0, 1)$, and the distribution parameter changes linearly from θ_0 to $\theta_0 - n^{-1/2} c_n \tau$; see also Bissel (1984, 1986). Moreover, in the range of values of a and b displayed in Figure 2 the efficiency of $T_\nu \circ \left(R * \breve{\Psi}^n\right)$ always exceeds the efficiency of $T_\nu \circ \hat{\Psi}^n$ as well as $T_\nu \circ \breve{\Psi}^n$.

8 Application to covariance structure models

In this section we consider the model in which the independent q-dimensional random vectors $X_1, \ldots, X_n$ have a common multivariate normal distribution with expectation vector μ and full rank covariance matrix Ω. Both μ and Ω depend on the unknown p-dimensional parameter vector θ. This model coincides with the model discussed in Magnus (1978), and is an extension of the covariance structure model in which the components of μ are all assumed to be zero. The classical theory for covariance structures was developed in Browne (1974, 1982, 1984) and Jöreskog (1981). The simultaneous equations model (Wold, 1954; Goldberger, 1964), the explanatory factor analysis model (Lawley and Maxwell, 1963), the higher order confirmatory factor analysis model (Jöreskog, 1971), the linear structural equations model (Jöreskog, 1977), and the Gaussian graphical model (Dempster, 1972; Whittaker 1990) are special cases of the covariance structure model. In Bentler and Dudgeon (1996) an overview of covariance structure analysis is given.

Let μ_i denote the i^{th} component of the expectation vector μ, and let ω_{ij} and ω^{ij} respectively denote the (i,j) elements of the covariance matrix Ω and its inverse Ω^{-1}. Observe that $\omega^{i\ell} = \sum_{j=1}^{p}\sum_{k=1}^{p} \omega^{ij}\omega_{jk}\omega^{k\ell}$. Moreover, introduce

$$\dot{\mu}_j = \begin{pmatrix} \dfrac{\partial \mu_j}{\partial \theta_1} \\ \dfrac{\partial \mu_j}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial \mu_j}{\partial \theta_m} \end{pmatrix}, \quad \dot{\omega}_{ij} = \begin{pmatrix} \dfrac{\partial \omega_{ij}}{\partial \theta_1} \\ \dfrac{\partial \omega_{ij}}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial \omega_{ij}}{\partial \theta_m} \end{pmatrix} \quad \text{and} \quad \dot{\omega}^{ij} = \begin{pmatrix} \dfrac{\partial \omega^{ij}}{\partial \theta_1} \\ \dfrac{\partial \omega^{ij}}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial \omega^{ij}}{\partial \theta_m} \end{pmatrix},$$

and observe that

$$\dot{\omega}_{i\ell} = -\sum_{j=1}^{p}\sum_{k=1}^{p} \omega_{ij}\dot{\omega}^{jk}\omega_{k\ell} \text{ and } \dot{\omega}^{ij} = -\sum_{k=1}^{p}\sum_{\ell=1}^{p} \omega^{ik}\omega^{j\ell}\dot{\omega}_{k\ell}.$$

According to the model the random vectors $X_1, \ldots, X_n$ have common density

$$f_\theta(x) = \{\det(2\pi\Omega)\}^{-1/2} \exp\left\{-\tfrac{1}{2}(x-\mu)^T \Omega^{-1}(x-\mu)\right\},$$

and hence the classical score function is given by

$$\rho_\theta(x) = \sum_{i=1}^{q}\sum_{j=1}^{q} \left\{\omega^{ij}(x_i-\mu_i)\dot{\mu}_j + \tfrac{1}{2}\{\omega_{ij} - (x_i-\mu_i)(x_j-\mu_j)\}\dot{\omega}^{ij}\right\};$$

see also (13) in Magnus (1978). Using the relations between $\omega^{i\ell}$, $\dot{\omega}_{i\ell}$ and $\dot{\omega}^{i\ell}$ just mentioned, we derive that the Fisher information matrix is given by

$$\begin{aligned} \Sigma &= \sum_{i=1}^{q}\sum_{k=1}^{q} \omega^{ik}\dot{\mu}_i(\dot{\mu}_k)^T + \tfrac{1}{4}\sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q} \{\omega_{ik}\omega_{j\ell} + \omega_{i\ell}\omega_{jk}\}\dot{\omega}^{ij}\left(\dot{\omega}^{k\ell}\right)^T \\ &= \sum_{i=1}^{q}\sum_{k=1}^{q} \left\{\omega^{ik}\dot{\mu}_i(\dot{\mu}_k)^T - \tfrac{1}{4}\dot{\omega}_{ik}\left(\dot{\omega}^{ik}\right)^T - \tfrac{1}{4}\dot{\omega}^{ik}(\dot{\omega}_{ik})^T\right\}. \end{aligned}$$

For every $0 < \epsilon \leq 1$ one may show that $n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{p} E\left(|S_{1i,jk}|^{1+\epsilon}\right)$ remains bounded for n tending to infinity, where $S_{1i,jk}$ is defined as in Lemma 1. Thus, if Σ is a differentiable function of θ, Lemma 1 implies that if r_n is of the order $n^{-1/2}$, the random variable U_{n,r_n} is $O_P(n^{-1/2})$.

Observe that if $X_1, \ldots, X_n$ are univariate and follow a normal distribution with expectation θ and variance 1, then $\rho_\theta(x)$ is equal to $x - \theta$, $\hat{\theta}^n$ is equal to the sample mean, and hence $\rho_{\hat{\theta}^n}(X_i)$ coincides with the i^{th} least squares residual.

9 An example: the mathematics marks data

In this section we analyze the mathematics marks data given in Table 1.2.1 in Mardia, Kent and Bibby (1979), by means of the overall test described in Section 5. The observations are marks in mechanics, vectors, algebra, analysis and statistics exams for 88 students. A “butterfly” Gaussian graphical model was fitted in Example 6.7.1 in Whittaker (1990) and in Example 3.4 in Edwards (1995); the parametrization

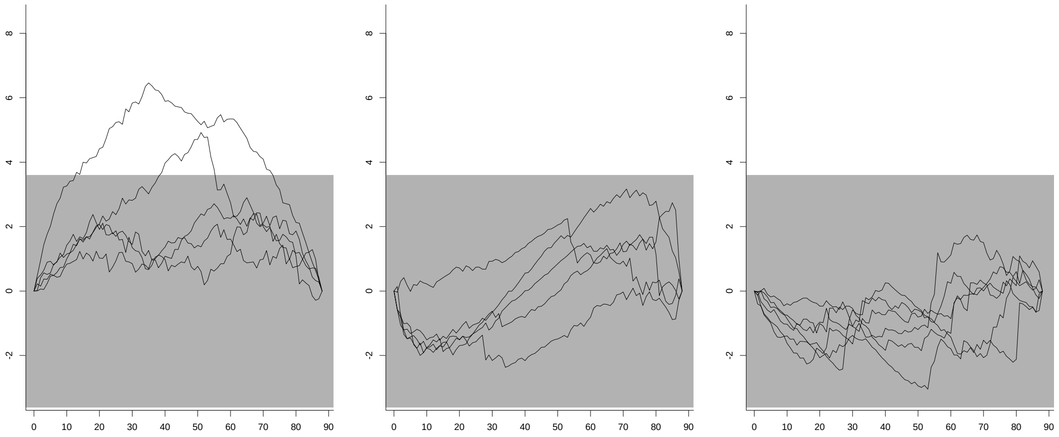
$$\mu = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{pmatrix}, \quad \Omega^{-1} = \begin{pmatrix} \theta_6 & \theta_{11} & \theta_{12} & 0 & 0 \\ \theta_{11} & \theta_7 & \theta_{13} & 0 & 0 \\ \theta_{12} & \theta_{13} & \theta_8 & \theta_{14} & \theta_{15} \\ 0 & 0 & \theta_{14} & \theta_9 & \theta_{16} \\ 0 & 0 & \theta_{15} & \theta_{16} & \theta_{10} \end{pmatrix}.$$

is in accordance with this model. Under the butterfly model the estimated expectation vector and the estimated inverse covariance matrix are given by

$$\begin{pmatrix} 3.896 \\ 5.059 \\ 5.060 \\ 4.674 \\ 4.226 \end{pmatrix}, \quad \begin{pmatrix} .530 & -.247 & -.291 & .000 & .000 \\ -.247 & 1.046 & -.567 & .000 & .000 \\ -.291 & -.567 & 2.895 & -.784 & -.493 \\ .000 & .000 & -.784 & 1.022 & -.210 \\ .000 & .000 & -.493 & -.210 & .647 \end{pmatrix},$$

Here we have divided each of the marks by ten, to avoid too many leading zeroes in our display of results. The information matrix depends on the unknown parameters; we estimate the information matrix by replacing the unknown parameters by their estimators. The LU-root of the inverse of the estimated information matrix is block-diagonal with two blocks. The first block concerns the estimation of the expectations $\theta_1, \ldots, \theta_5$, and is given by

$$\begin{pmatrix} 1.373 & & & & \\ .483 & 1.036 & & & \\ .606 & .508 & .673 & & \\ .371 & .311 & .412 & 1.160 & \\ .632 & .529 & .702 & .891 & 1.722 \end{pmatrix};$$



Expectations

Inverse covariances (diagonal)

Inverse covariances (off-diagonal)

Figure 3: Marks data: the processes $2\nu_1^T\hat{\Psi}^n, \ldots, 2\nu_{16}^T\hat{\Psi}^n$.

the entries not explicitly given are all equal to zero. The second block concerns the estimation of the inverse covariances $\theta_6, \ldots, \theta_{16}$, and is given by

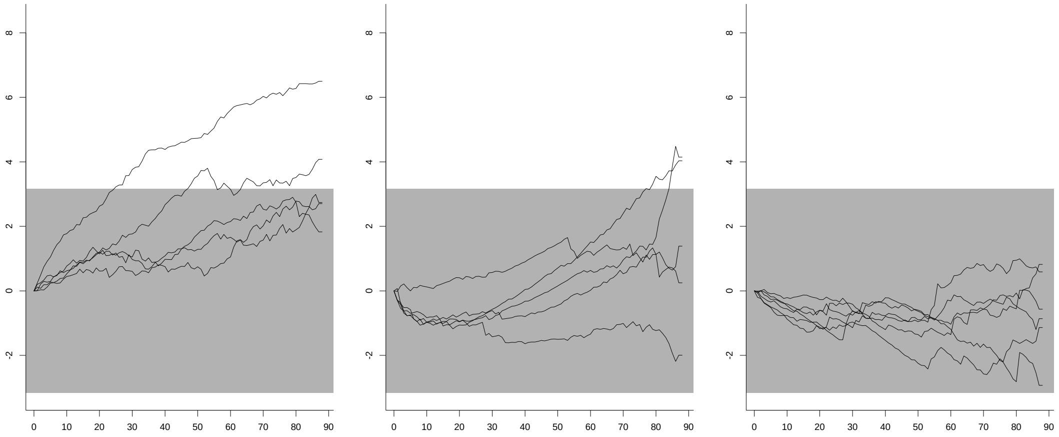
$$\left(\begin{array}{rrrrrrrrrrr}
.468 &&&&&&&&&& \\
-.150 & .869 &&&&&&&&& \\
-.105 & -.282 & 1.396 &&&&&&&& \\
 & & -.752 & .769 &&&&&&& \\
 & & -.490 & -.148 & .542 &&&&&& \\
-.452 & -.788 & -.178 & & & .731 &&&&& \\
-.308 & .152 & -1.189 & & & -.261 & .948 &&&& \\
.102 & -.843 & -1.278 & & & -.117 & -.161 & 1.407 &&& \\
 & & -2.093 & -1.101 & -.045 & & & & 1.629 && \\
 & & -1.444 & .069 & -.661 & & & & -.225 & 1.199 & \\
 & & .921 & -.510 & -.323 & & & & -.405 & -.481 & .840
\end{array}\right).$$

Figure 3 displays $2\nu_1^T\hat{\Psi}^n, \ldots, 2\nu_{16}^T\hat{\Psi}^n$, where ν_i^T is equal to the i^{th} row of the LU-root of the inverse of the estimated information matrix. Observe that every process starts and ends at value zero. The "two-sided version" of the overall test described in the beginning of Section 5 is based on the statistic

$$\max_{i=1,\ldots,16} \sup_{t\in[0,1]} \left|2\nu_i^T\hat{\Psi}_t^n\right|.$$

The asymptotic null hypothesis distribution of each of the individual test statistics $\sup_{t\in[0,1]} \left|2\nu_i^T\hat{\Psi}_t^n\right|$ coincides with the distribution of two times the absolute supremum of the one-dimensional standard Brownian bridge, which has cumulative distribution function

$$1 - 2\sum_{k=1}^{\infty} \left\{F_{\text{Weibull}(2^{-1/2},2)}\left(kx\right) - F_{\text{Weibull}(2^{-1/2},2)}\left((k-\tfrac{1}{2})x\right)\right\},$$



Expectations

Inverse covariances (diagonal)

Inverse covariances (off-diagonal)

Figure 4: Marks data: the processes $\nu_1^T \breve{\Psi}^n, \ldots, \nu_{16}^T \breve{\Psi}^n$.

where, as in Section 5, $F_{\text{Weibull}(2^{-1/2},2)}(x)$ is the cumulative distribution function belonging to the Weibull distribution with scale parameter $2^{-1/2}$ and shape parameter 2. As each of the 16 individual suprema are asymptotically independent, it follows that the asymptotic critical value at the 5 percent significance level of the overall test statistic is 3.60; see also the table in Smirnov (1948). The shaded area in Figure 3 represents the acceptance region of the overall test: the null hypothesis is rejected as soon as one of the 16 processes moves outside this area. The process $2\nu_5^T \hat{\Psi}^n$ reaches a maximal value of 6.46, which clearly exceeds the overall asymptotic critical value 3.60; moreover, the process $2\nu_4^T \hat{\Psi}^n$ reaches a maximal value of 4.78, which also exceeds the overall asymptotic critical value. Thus, the estimated score process indicates "time"-dependent behaviour of the statistics and analysis expectations.

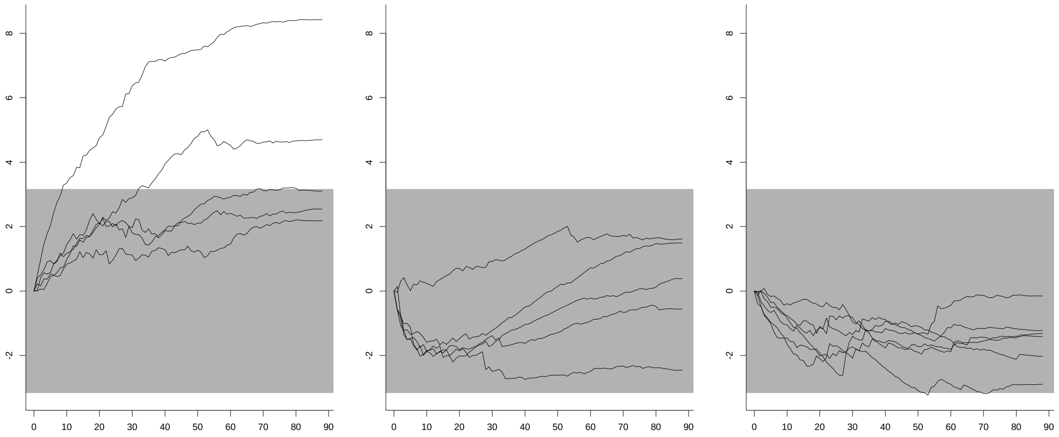
Figure 4 displays $\nu_1^T \breve{\Psi}^n, \ldots, \nu_{16}^T \breve{\Psi}^n$; observe the "free" behaviour of the processes at the end of the interval. The overall test is now based on the statistic

$$\max_{i=1,\ldots,16} \sup_{t\in[0,1]} \left| \nu_i^T \breve{\Psi}_t^n \right| .$$

The asymptotic null hypothesis distribution of each of the individual test statistics $\sup_{t\in[0,1]} \left| \nu_i^T \breve{\Psi}_t^n \right|$ coincides with the distribution of the absolute supremum of the one-dimensional standard Wiener process on the unit interval, which has cumulative distribution function

$$1 - 2\sum_{k=1}^{\infty} \left\{ F_{\text{HN}(0,1)}\left((4k-1)x\right) - F_{\text{HN}(0,1)}\left((4k-3)x\right) \right\},$$

where, as in Section 5, $F_{\text{HN}(0,1)}(x)$ is the cumulative distribution function belonging to the half-standardnormal distribution. One may show that the asymptotical critical value at the 5 percent significance level of the overall test statistic is equal to 3.16. The



Expectations

Inverse covariances (diagonal)

Inverse covariances (off-diagonal)

Figure 5: Marks data: the processes $\nu_1^T\left(R * \breve{\Psi}^n\right), \dots, \nu_{16}^T\left(R * \breve{\Psi}^n\right)$.

shaded area in Figure 4 represents the acceptance region of the overall test: the null hypothesis is rejected as soon as one of the 16 processes moves outside this area. The processes $\nu_4^T\breve{\Psi}^n$, $\nu_5^T\breve{\Psi}^n$, $\nu_6^T\breve{\Psi}^n$ and $\nu_8^T\breve{\Psi}^n$ reach maximal values of 4.08, 6.50, 4.04 and 4.48 respectively, which all exceed the overall asymptotic critical value 3.16. Thus, the innovation process indicates “time”-dependent behaviour of the statistics and analysis expectations, and of the mechanics and algebra diagonal inverse covariances.

Figure 5 displays $\nu_1^T\left(R * \breve{\Psi}^n\right), \dots, \nu_{16}^T\left(R * \breve{\Psi}^n\right)$; observe that due to the use of R as integrand, the volatility of these processes is relatively high at the beginning of the time interval and relatively low at the end. The overall test is based on the statistic

$$\max_{i=1,\dots,16} \sup_{t\in[0,1]} \left|\nu_i^T\left(R * \breve{\Psi}_t^n\right)\right| .$$

The asymptotic null hypothesis distribution of each of the individual test statistics $\sup_{t\in[0,1]}\left|\nu_i^T\left(R * \breve{\Psi}_t^n\right)\right|$ coincides with the distribution of the absolute supremum of the one-dimensional standard Wiener process on the unit interval, and hence the asymptotical critical value at the 5 percent significance level of the overall test statistic is equal to 3.16. The shaded area in Figure 5 represents the acceptance region of the overall test: the null hypothesis is rejected as soon as one of the 16 processes moves outside this area. The processes $\nu_2^T\left(R * \breve{\Psi}_t^n\right)$, $\nu_4^T\left(R * \breve{\Psi}_t^n\right)$, $\nu_5^T\left(R * \breve{\Psi}_t^n\right)$ and $\nu_{14}^T\left(R * \breve{\Psi}_t^n\right)$ reach maximal absolute values of 3.22, 5.00, 8.43 and 3.23 respectively, which all exceed the overall asymptotic critical value 3.16. Thus, the process $R * \breve{\Psi}^n$ indicates “time”-dependent behaviour of the statistics, analysis and vectors expectations, and of the algebra/analysis off-diagonal inverse covariance.

Our results are consistent with the ones obtained in Edwards (1995, Section 7.3) after adding a continuous latent variable to the butterfly model, yielding the conclusion “that the data have been sorted by some criterion”; observe that sorting creates time-dependence. However, the latter results were obtained ad hoc, and without a formal test. Moreover, attention was restricted to changes in the components of μ, since the continuous latent variable only affects the location of the five marks.

Acknowledgement

The author is grateful to Nils Lid Hjort and Ronald J.M.M. Does for their encouragement.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, Berlin.

[2] Bentler, P.M., Dudgeon, P. (1996). Covariance structure analysis: statistical practice, theory and directions. *Annual Review of Psychology* **47**, 563–592.

[3] Bissell, A.F. (1984). The performance of control charts and cusums under linear trend. *Applied Statistics* **33**, 145–151.

[4] Bissell, A.F. (1986). ”Corrigendum” (to Bissell, 1984). *Applied Statistics* **35**, 214.

[5] Broström (1997). A martingale approach to the changepoint problem. *Journal of the American Statistical Association* **92**, 1177–1183.

[6] Brown, R.L., Durbin, J., Evans, J.M. (1975). Techniques for testing the constancy of regression relationships over time. *J. R. Statist. Soc.* **B37**, 149-163.

[7] Browne, M.W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal* **8**, 1–24.

[8] Browne, M. W. (1982). Covariance structures. In: Hawkins, D.M. (Ed.) *Topics in applied multivariate analysis*. Cambridge University Press, Cambridge, 72–141.

[9] Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

[10] Chu, C.-S.J., Hornik, K., Kuan, C.-M. (1995). MOSUM tests for parameter constancy. *Biometrika* **82**, 603–617.

[11] Csörgő, M., Horvath, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, New York.

[12] Dempster, A.P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

[13] Durbin, J. (1973). *Distribution Theory for Tests based on the Sample Distribution Function*. SIAM, Philadelphia.

[14] Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer Texts in Statistics. Springer-Verlag, New York.

[15] Goldberger, A.S. (1964). *Econometric theory*. New York, Wiley.

[16] Hjort, N.L, Koning, A.J. (1999). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, accepted.

[17] Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika* **36**, 109–133.

[18] Jöreskog, K.G. (1977). Structural equation models in the social sciences: specification, estimation and testing. In: Krishnaiah, P.R. (Ed.): *Applications of Statistics*. North-Holland Publishing Co., Amsterdam, 265–287.

[19] Jöreskog, K.G. (1981). Analysis of covariance structures. *Scandinavian Journal of Statistics* **8**, 65–92.

[20] Hoffmann-Jorgensen, J. (1994a). *Probability with a View towards Statistics, Volume I*. Chapman and Hall, London.

[21] Hoffmann-Jorgensen, J. (1994b). *Probability with a View towards Statistics, Volume II*. Chapman and Hall, London.

[22] Horváth, L., Parzen, E. (1994). Limit theorems for Fisher score change processes. In: *Change Point Problems*. IMS lecture notes - monograph series **23**, 157–169.

[23] Khmaladze, E.V. (1981). Martingale approach in the theory of goodness of fit tests. *Theory Probab. Appl.* **26**, 246–265.

[24] Khmaladze, E.V. (1993). Goodness of fit problems and scanning innovation martingales. *Ann. Statist.* **21**, 798–829.

[25] Koning, A.J. (1994). Approximation of the basic martingale. *Ann. Statist.*, **22**, 565–579.

[26] Kallenberg, W.C.M., Koning, A.J. (1995) On Wieand's theorem. *Statistics & Probability Letters* **25**, 121–132.

[27] Krämer, W., Ploberger, W., Alt, R. (1988). Testing for structural change in dynamic models. *Econometrica* **56**, 1355–1369.

[28] Lehmann, E.L. (1986). *Testing Statistical Hypotheses, second edition*. Chapman and Hall, London.

[29] Lehmann, E.L. (1991). *Theory of Point Estimation*. Wadsworth, Belmont. Wiley, New York.

[30] Lindsey, J.K. (1996). *Parametric Statistical Inference*. Oxford University Press.

[31] Lawley, D.N., Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*. Butterworths.

[32] Magnus, J.R. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics* **7**, 281–312.

[33] Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, London.

[34] Page, E.S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika* **44**, 248–252.

[35] Ploberger, W. and Krämer, W. (1992). The CUSUM test with OLS residuals. *Econometrica* **60**, 271–85.

[36] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

[37] Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1**, 295–314.

[38] Rootzén, H. (1980). On the functional limit theorem for martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **51**, 79–93.

[39] Schruben, L.W. (1982). Detecting initialization bias in simulation output. *Oper. Res.* **30**, 569–590.

[40] Schruben, L.W. (1983). Confidence interval estimation using standardized time series. *Oper. Res.* **31**, 1090–1108.

[41] Shorack, G.R., Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

[42] Smirnov, N.V. (1948). Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.* **19**, 279–281.

[43] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

[44] Wieand, H.S. (1976). A condition under which the Pitman and Bahadur approaches to efficiency coincide. *Ann. Statist.* **4** , 1003-1011.

[45] Wold, H.O.A. (1954). Causality and Econometrics. *Econometrica* **28**, 443–463.

A Proofs

This section contains the proofs of Lemma 1, Inequality 1 and Inequality 2.

Proof of Lemma 1 A Taylor expansion of $\rho_{\vartheta}(X_i) - \rho_{\vartheta'}(X_i)$ around ϑ, immediately followed by a second Taylor expansion of $\dot{\rho}_{\vartheta}(X_i)$ around θ_i, yields

$$R(X_i, \vartheta'_i, \vartheta) = \sum_{j=1}^{4} R_j(X_i, \vartheta'_i, \vartheta),$$

with

$$\begin{aligned} R_1(X_i, \vartheta'_i, \vartheta) &= \Sigma^{-1/2} \left(\rho_{\vartheta'_i}(X_i) - \rho_{\vartheta}(X_i) - \dot{\rho}_{\vartheta}(X_i)\,(\vartheta'_i - \vartheta) \right), \\ R_2(X_i, \vartheta'_i, \vartheta) &= \Sigma^{-1/2} \left(\dot{\rho}_{\vartheta}(X_i) - \dot{\rho}_{\theta_i}(X_i) \right) (\vartheta'_i - \vartheta), \\ R_3(X_i, \vartheta'_i, \vartheta) &= \Sigma^{-1/2} \left\{ \dot{\rho}_{\theta_i}(X_i) - E\left(\dot{\rho}_{\theta_i}(X_i) \right) \right\} (\vartheta'_i - \vartheta), \\ R_4(X_i, \vartheta'_i, \vartheta) &= \Sigma^{-1/2} \left\{ E\left(\dot{\rho}_{\theta_i}(X_i) \right) + \Sigma \right\} (\vartheta'_i - \vartheta). \end{aligned}$$

It is easily seen that

$$\max_{\ell=1,\dots,n} n^{-1} \sum_{i=\ell}^{n} |R_4(X_i, \vartheta'_i, \vartheta)| \leq \max_{\ell=1,\dots,n} n^{-1} \sum_{i=\ell}^{n} \sum_{j=1}^{p} \sum_{k=1}^{p} |S_{2i,jk}| \left(\max_{\ell=1,\dots,n} d_{\vartheta'_\ell,\vartheta} \right). \qquad (9)$$

Furthermore, it follows by regularity condition (D) in Section 6.4 of Lehmann (1991) that there exists some constant c such that

$$\max_{\ell=1,\dots,n} n^{-1} \sum_{i=\ell}^{n} |R_1(X_i, \vartheta'_i, \vartheta)| \leq c \left(\max_{\ell=1,\dots,n} d_{\vartheta'_i,\vartheta} \right)^2 \leq 2cr \left(\max_{\ell=1,\dots,n} d_{\vartheta'_\ell,\vartheta} \right), \qquad (10)$$

$$\max_{\ell=1,\dots,n} n^{-1} \sum_{i=\ell}^{n} |R_2(X_i, \vartheta'_i, \vartheta)| \leq c \left(\max_{\ell=1,\dots,n} d_{\theta_\ell,\vartheta} \cdot d_{\vartheta'_\ell,\vartheta} \right) \leq cr \left(\max_{\ell=1,\dots,n} d_{\vartheta'_\ell,\vartheta} \right). \qquad (11)$$

In the context of (5), let δ_{ik} denote the k^{th} element of $\theta_i - \bar{\theta}$, and observe that $|\delta_{ik}|$ is bounded by $\max_{\ell=1,\dots,n} d_{\theta_\ell,\bar{\theta}}$. We may write

$$\max_{\ell=1,\dots,n} \left| n^{-1} \sum_{i=\ell}^{n} R_3(X_i, \theta_i, \bar{\theta}) \right| \leq \sum_{j=1}^{p} \sum_{k=1}^{p} \max_{\ell=1,\dots,n} \left| n^{-1} \sum_{i=\ell}^{n} S_{1i,jk} \delta_{ik} \right|.$$

For $0 \leq \epsilon \leq 1$, the maximal moment inequality (4.34.1) in Hoffmann-Jørgensen (1994a, p. 308) yields that

$$\begin{aligned} E\left\{ \left(\max_{\ell=1,\dots,n} \left| n^{-1} \sum_{i=\ell}^{n} S_{1i,jk} \delta_{ik} \right| \right)^{1+\epsilon} \right\} &\leq Cn^{-\epsilon} n^{-1} \sum_{i=1}^{n} E\left\{ |S_{1i,jk} \delta_{ik}|^{1+\epsilon} |\delta_{ik}|^{1+\epsilon} \right\} \\ &\leq Cn^{-\epsilon} \left(\max_{\ell=1,\dots,n} d_{\theta_\ell,\bar{\theta}} \right)^{1+\epsilon} \end{aligned}$$

for some constant C only depending on ϵ. Application of Markov's inequality (Shorack and Wellner, 1986, p. 842) shows that

$$n^{\epsilon/(1+\epsilon)} \frac{\max_{\ell=1,\dots,n} |n^{-1} \sum_{i=\ell}^{n} S_{1i,jk} \delta_{ik}|}{\max_{\ell=1,\dots,n} d_{\theta_\ell,\bar{\theta}}} = O_P(1).$$

Hence, there exists a random variable $U_n' = O_P(n^{-\epsilon/(1+\epsilon)})$ which satisfies

$$\max_{\ell=1,\dots,n} \left| n^{-1} \sum_{i=\ell}^{n} R_3(X_i, \theta_\ell, \bar{\theta}) \right| \le U_n' \max_{\ell=1,\dots,n} d_{\theta_\ell,\bar{\theta}}. \tag{12}$$

In the context of (4), define U_n'' by

$$U_n'' = \max_{\ell=1,\dots,n} \left| n^{-1} \sum_{i=\ell}^{n} \sum_{j=1}^{p} \sum_{k=1}^{p} S_{1i,jk} \right|,$$

and observe that

$$\max_{\ell=1,\dots,n} \left| n^{-1} \sum_{i=\ell}^{n} R_3(X_i, \vartheta', \vartheta) \right| \le U_n'' d_{\vartheta',\vartheta}. \tag{13}$$

Again combining the maximal moment inequality (4.34.1) in Hoffmann-Jørgensen (1994a) with Markov's inequality yields that U_n'' is $O_P(n^{-\epsilon/(1+\epsilon)})$, since

$$n^{\epsilon/(1+\epsilon)} \left| n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{p} S_{1i,jk} \right| = O_P(1).$$

Finally, take U_n^* equal to the maximum of U_n' and U_n'', and observe that U_n^* is $O_P(n^{-\epsilon/(1+\epsilon)})$. Combination of (9)–(11) and (13) yields (4), and combination of (9)–(11) and (12) yields (5). This concludes the proof of Lemma 1. □

Proof of Inequality 1 Define $\Gamma_\vartheta^n = \vartheta + n^{-1} \sum_{i=1}^{n} \Sigma^{-1} \rho_\vartheta(X_i)$, and observe that the likelihood equations (2) may be written in the form $\Gamma_{\hat{\theta}^n}^n = \hat{\theta}^n$. Moreover,

$$n^{1/2} \left(\Gamma_{\hat{\theta}^n}^n - \bar{\theta} \right) = n^{1/2} \left(\hat{\theta}^n - \bar{\theta} \right) \quad \text{and} \quad n^{1/2} \left(\Gamma_{\bar{\theta}}^n - \bar{\theta} \right) = n^{-1/2} \sum_{i=1}^{n} \Sigma^{-1} \rho_{\bar{\theta}}(X_i) = W_n^*.$$

The differentiability condition implies that

$$\begin{aligned} \left| \Sigma^{1/2} \left(\Gamma_{\vartheta'}^n - \Gamma_\vartheta^n \right) \right| &= \left| n^{-1} \sum_{i=1}^{n} \left(\Sigma^{-1/2} \rho_{\vartheta'}(X_i) - \Sigma^{-1/2} \rho_\vartheta(X_i) + \Sigma^{1/2} \left(\vartheta' - \vartheta \right) \right) \right| \\ &= \left| n^{-1} \sum_{i=1}^{n} R(X_i, \vartheta, \vartheta') \right| \le d_{\vartheta,\vartheta'} U_{n,r} \end{aligned} \tag{14}$$

for every $\vartheta, \vartheta' \in \Theta$ satisfying $d_{\vartheta',\theta} < r$ and $d_{\vartheta',\theta} < r$.

First, we shall prove the existence of a solution in Θ_{n0} to the likelihood equations given the event that (6) holds. Condition on this event, and consider $\vartheta \in \Theta_{n0}$. Since $U_{n,r} < \frac{1}{3}$, it follows that

$$\left|\Sigma^{1/2}\left(\Gamma^n_{\vartheta'} - \Gamma^n_{\vartheta}\right)\right| \leq \tfrac{1}{3}\left|\Sigma^{1/2}\left(\vartheta' - \vartheta\right)\right| = \tfrac{1}{3} d_{\vartheta',\vartheta} \tag{15}$$

for every $\vartheta, \vartheta' \in \Theta$ satisfying $d_{\vartheta',\theta} < r$ and $d_{\vartheta',\theta} < r$. Consequently, it follows that

$$\begin{aligned} \left|\Sigma^{1/2}\left(\Gamma^n_{\vartheta} - \bar{\theta}\right)\right| &\leq \left|\Sigma^{1/2}\left(\Gamma^n_{\bar{\theta}} - \bar{\theta}\right)\right| + \left|\Sigma^{1/2}\left(\Gamma^n_{\vartheta} - \Gamma^n_{\bar{\theta}}\right)\right| \\ &\leq n^{-1/2}\left|W^*_n\right| + \tfrac{1}{3} d_{\vartheta,\bar{\theta}} < \tfrac{1}{2} r + \tfrac{1}{3} r < r, \end{aligned}$$

and hence Γ^n maps Θ_{n0} into itself. Define the sequence $\{\hat{\theta}^n_i\}_{i=1}^{\infty}$ recursively by $\hat{\theta}^n_i = \Gamma^n_{\hat{\theta}^n_{i-1}}$, where $\hat{\theta}^n_0$ is equal to $\bar{\theta}$. From (15) it now follows that this sequence converges to a point $\hat{\theta}^n \in \Theta_{n0}$ which satisfies $\Gamma^n_{\hat{\theta}^n} = \hat{\theta}^n$.

Next, we prove the final statement of Inequality 1. Condition on the event that (6) holds. It immediately follows from (15) that

$$\left|\Sigma^{1/2}\left(\Gamma^n_{\hat{\theta}^n} - \Gamma^n_{\bar{\theta}}\right)\right| < \tfrac{1}{3}\left|\Sigma^{1/2}\left(\Gamma^n_{\hat{\theta}^n} - \bar{\theta}\right)\right|, \tag{16}$$

which implies

$$\left|\Sigma^{1/2}\left(\Gamma^n_{\hat{\theta}^n} - \bar{\theta}\right)\right| < \left(1 - \tfrac{1}{3}\right)^{-1}\left|\Sigma^{1/2}\left(\Gamma^n_{\bar{\theta}} - \bar{\theta}\right)\right| = \tfrac{3}{2}\left|\Sigma^{1/2}\left(\Gamma^n_{\bar{\theta}} - \bar{\theta}\right)\right|.$$

Combining this inequality with (16) yields

$$\left|\Sigma^{1/2}\left(\Gamma^n_{\hat{\theta}^n} - \Gamma^n_{\bar{\theta}}\right)\right| < \tfrac{1}{3}\left|\Sigma^{1/2}\left(\Gamma^n_{\hat{\theta}^n} - \bar{\theta}\right)\right| < \tfrac{3}{2} U_{n,r}\left|\Sigma^{1/2}\left(\Gamma^n_{\bar{\theta}} - \bar{\theta}\right)\right|,$$

which concludes the proof of Inequality 1. □

Proof of Inequality 2 Since we may write

$$\begin{aligned} &\Sigma^{-1/2}\rho_{\hat{\theta}^n}(X_i) - \Sigma^{-1/2}\hat{\mu}_i - \Sigma^{-1/2}\rho_{\theta_i}(X_i) + n^{-1}\sum_{\ell=1}^{n}\Sigma^{-1/2}\rho_{\theta_\ell}(X_\ell) \\ &\quad = R(X_i, \hat{\theta}^n, \theta_i) - n^{-1}\sum_{\ell=1}^{n} R(X_\ell, \hat{\theta}^n, \theta_\ell) \end{aligned}$$

the first statement follows from

$$\begin{aligned} &\left|\Sigma^{-1/2}\hat{\Psi}^n_t - n^{-1/2}\sum_{i=1}^{[nt]}\Sigma^{-1/2}\left\{\hat{\mu}_i + \rho_{\theta_i}(X_i) - n^{-1}\sum_{\ell=1}^{n}\rho_{\theta_\ell}(X_\ell)\right\}\right| \\ &\quad \leq \left|n^{-1/2}\sum_{i=1}^{[nt]}\left\{R(X_i, \hat{\theta}^n, \theta_i) - n^{-1}\sum_{\ell=1}^{n} R(X_\ell, \hat{\theta}^n, \theta_\ell)\right\}\right| \\ &\quad \leq 4\left(n^{1/2}\max_{\ell=1,\ldots,n} d_{\hat{\theta}^n,\theta_\ell}\right) U_{n,r} \end{aligned}$$

since $R(X_i, \hat{\theta}^n, \theta_i)$ equals $R(X_i, \hat{\theta}^n, \bar{\theta}) + R(X_i, \theta_i, \bar{\theta})$. Similarly, the second statement follows from

$$
\begin{aligned}
&\left| \Sigma^{-1/2} \breve{\Psi}_t^n - n^{-1/2} \sum_{i=1}^{[nt]} c_{n,i} \Sigma^{-1/2} \left\{ \breve{\mu}_i + \rho_{\theta_i}(X_i) - (n-i+1)^{-1} \sum_{\ell=i}^{n} \rho_{\theta_\ell}(X_\ell) \right\} \right| \\
&\quad \le \left| n^{-1/2} \sum_{i=1}^{[nt]} c_{n,i} \left\{ R(X_i, \hat{\theta}^n, \theta_i) - \frac{1}{n-i+1} \sum_{\ell=i}^{n} R(X_\ell, \hat{\theta}^n, \theta_\ell) \right\} \right| \\
&\quad \le \left| n^{1/2} \sum_{j=1}^{[nt]} (c_{n,j} - c_{n,j-1})\, n^{-1} \sum_{i=j}^{[nt]} R(X_i, \hat{\theta}^n, \theta_i) \right| \\
&\qquad + \left| n^{1/2} \sum_{i=1}^{[nt]} \frac{c_{n,i}}{n-i+1} n^{-1} \sum_{\ell=i}^{n} R(X_\ell, \hat{\theta}^n, \theta_\ell) \right| \\
&\quad \le 2 \left\{ \sum_{j=1}^{[nt]} (c_{n,j} - c_{n,j-1}) + \sum_{i=1}^{[nt]} \frac{c_{n,i}}{n-i+1} \right\} n^{1/2} \max_{\ell=1,\ldots,n} d_{\theta_\ell, \hat{\theta}^n} U_{n,r}.
\end{aligned}
$$

This concludes the proof of Inequality 2. □