Parker, Thomas

Working Paper

A comparison of alternative approaches to sup-norm goodness of fit tests with estimated parameters

cemmap working paper, No. CWP34/10

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Parker, Thomas (2010) : A comparison of alternative approaches to sup-norm goodness of fit tests with estimated parameters, cemmap working paper, No. CWP34/10, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2010.3410

This Version is available at:
https://hdl.handle.net/10419/64794

cemmap
centre for microdata methods and practice

A comparison of alternative approaches to sup-norm goodness of fit tests with estimated parameters

Thomas Parker

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP34/10

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

A Comparison of Alternative Approaches to Sup-Norm Goodness of Fit Tests with Estimated Parameters

Thomas Parker
Department of Economics, University of Illinois at Urbana-Champaign
Email: tmparker@illinois.edu

May 26, 2010

Abstract

Goodness of fit tests based on sup-norm statistics of empirical processes have nonstandard limiting distributions when the null hypothesis is composite — that is, when parameters of the null model are estimated. Several solutions to this problem have been suggested, including the calculation of adjusted critical values for these nonstandard distributions and the transformation of the empirical process such that statistics based on the transformed process are asymptotically distribution-free. The approximation methods proposed by Durbin (1985) can be applied to compute appropriate critical values for tests based on sup-norm statistics. The resulting tests have quite accurate size, a fact which has gone unrecognized in the econometrics literature. Some justification for this accuracy lies in the similar features that Durbin's approximation methods share with the theory of extrema for Gaussian random fields and for Gauss-Markov processes. These adjustment techniques are also related to the transformation methodology proposed by Khmaladze (1981) through the score function of the parametric model. Monte Carlo experiments suggest that these two testing strategies are roughly comparable to one another and more powerful than a simple bootstrap procedure.

Keywords: Goodness of fit test, Estimated parameters, Gaussian process, Gauss-Markov process, Boundary crossing probability, Martingale transformation

JEL Classification Code: C12, C14, C46

1 Introduction

Empirical processes are central to the theory of Kolmogorov-Smirnov-type specification tests, and it is a standard result that for simple null hypotheses, the empirical process $\sqrt{n}(\mathbb{F}_n - F_0)$ converges weakly to the Brownian bridge and the Kolmogorov-Smirnov statistic is distribution-free. General study of the convergence of empirical processes with estimated parameters was first conducted by Durbin (1973a) and Neuhaus (1976). The limiting distributions of these processes were found to be significantly more complex than the limiting distribution of the process for simple null hypotheses. As a result, the evaluation of sup-norm test statistics based on these processes has been an enduring problem.

The calculation of the null distribution or critical values for a statistic based on an empirical process with estimated parameters is accomplished quite often via simulation techniques. There are, however, alternatives. One solution to the problem of testing goodness of fit with estimated parameters is the martingale transform method proposed by Khmaladze (1981). This approach has received attention in the statistics and econometrics literature recently, notably in Koenker and Xiao (2002); Bai (2003); Khmaladze and Koul (2004); Delgado and Stute (2008) and Khmaladze and Koul (2009). The martingale transform method employs a Doob-Meyer decomposition to transform the empirical process so that it is asymptotically distribution-free, a property that test statistics, as functionals of the process, inherit. This method may be applied quite generally: see for example Song (2010) for its application to semiparametric models, or Li (2009), who analyzes this method as a technique of projection onto a series of orthogonal polynomials, drawing on the work of Bickel et al. (1993) and Cabaña and Cabaña (1997).

Another solution to this problem for Kolmogorov-Smirnov-type tests (parallel to techniques devised for example by Durbin et al. (1975) for Cramér-von Mises-type tests,) is to calculate appropriate distributionally dependent critical values for each test. For Kolmogorov-Smirnov-type tests, Durbin (1973b, 1975, 1985), explored a number of approaches to the calculation of critical values for tests based on processes with estimated parameters. These methods involve varying amounts of computational effort and deserve greater recognition as an alternative methodology. In particular, one result of this work is a collection of simple approximations that are accurate, generalizable, and involve only modest computation. The approximate boundary crossing probabilities first derived in Durbin (1985) are presented in Section 3, with particular focus on two approximations to the distribution of the Kolmogorov-Smirnov statistic when parameters are estimated. It will be shown that one of these approximations is identical to results derived using the theory of extrema of Gaussian fields and the other can be interpreted as a Gauss-Markov-process approximation to the distribution of the sup-norm statistic. The present work is then, in some sense, a continuation of Durbin's research in goodness of fit testing — even though a goodness of fit problem was the primary applied example of Durbin (1985), his boundary crossing results are used almost exclusively in other applications. The connections that Durbin's approximations have to these other well-developed fields of probability simultaneously provide justification for their great accuracy and offer a means of generalization perhaps not apparent in Durbin's original work.

Durbin's approximate boundary crossing probabilities are also compared with Khmaladze's martingale transform in a few simple situations. The essentials of each technique are presented and applied to the context of one-sample tests of normality and exponentiality, drawing some connections and elaborating upon the example given in Durbin (1985, p. 117). Finally, simulation experiments investigate the empirical size and power of a one-sided test of exponentiality. The adjusted critical values result in tests of approximately the same size and power as tests using a transformed process, although the experiments suggest differential power performance over the space of alternatives.

2 Parametric models

Consider a sample of size n from a random variable with distribution function F. A goodness-of-fit test is defined as a test of the hypothesis that F is a member of a parametric model; that is, $H_0: F \in \mathcal{F} := \{F(x,\theta); x \in \mathcal{X}, \theta \in \Theta\}$, with $\mathcal{X} \subseteq \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^p$. It is assumed that all members of $\mathcal{F}$ are absolutely continuous and mutually absolutely continuous. Process-based specification tests for F are typically based on one of the following empirical processes: the uniform empirical process

$$V_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x,\theta_0)), \qquad x \in \mathcal{X} \tag{1}$$

for simple null hypotheses, or the parametric empirical process

$$\hat{V}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x,\hat{\theta})) \qquad x \in \mathcal{X} \tag{2}$$

for composite null hypotheses, where $\hat{\theta}$ is some estimate of θ_0 and $\mathbb{F}_n$ is the empirical distribution function.

The uniform empirical process is convenient because under the above assumptions on $\mathcal{F}$ an inverse function F^{-1} is well defined and we can make the time transformation $t = F(x,\theta_0)$, which (with some abuse of notation, let $\mathbb{F}_n(t) = \frac{1}{n}\sum_i I(F(X_i,\theta_0) \le t)$ makes process (1) equivalent to

$$v_n(t) = \sqrt{n}(\mathbb{F}_n(t) - t), \quad t \in [0,1]. \tag{3}$$

That is, process (1) is equivalent to a process based on n iid realizations of a uniform random variable. Donsker's theorem implies that v_n converges weakly to a Brownian bridge on $[0,1]$ — in other words, V_n converges weakly to $B \circ F$, a time-changed Brownian bridge.

In many cases of practical interest the investigator is interested in the parametric model $\mathcal{F}$ but reluctant to specify θ_0. It may be hoped that similar calculations would work for both the uniform empirical process and the parametric empirical process. However, this is unfortunately not the case.

To explore this further, we make the following two assumptions:

A1 The model $\mathcal{F}$ satisfies the following condition (cf. Durbin (1973a)): that the function

$$g(t,\theta) = \nabla_\theta F(x,\theta)\big|_{x=F^{-1}(t,\theta_0)} \tag{4}$$

is almost everywhere finite and continuous in its arguments for all $(t,\theta) \in [0,1] \times \nu$, where ν is a closed neighborhood of θ_0 in Θ.

A2 There exists an estimator of the parameters $\hat{\theta}_n$ that satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_p(1) \tag{5}$$

Because the (uniform) $\sqrt{n}$ rate of convergence of $\mathbb{F}_n$ to F is the same as the rate of convergence

of the estimator $\hat{\theta}_n$ to θ_0, the effect of parameter estimation is not asymptotically negligible. When considering the convergence of $\hat{V}_n(t)$ to its limit, it was shown in Durbin (1973a) (and re-derived many times thereafter) that under the above assumptions the following informal analysis of the parametric empirical process is justified:

$$\begin{aligned}\hat{V}_n(x) &= \sqrt{n}(\mathbb{F}_n(x) - F(x,\hat{\theta}_n)) \\ &= \sqrt{n}(\mathbb{F}_n(x) - F(x,\theta_0)) - \sqrt{n}(F(x,\hat{\theta}_n) - F(x,\theta_0)) \\ &= V_n(x) - \sqrt{n}(\hat{\theta}_n - \theta_0)\nabla_\theta F(x,\theta_0) + o_p(1) \end{aligned} \tag{6}$$

where the $o_p(1)$ term is uniform in $x \in \mathcal{X}$. From (6) it is apparent that in general the distribution of $\hat{V}_n$ depends on the value of the parameter θ_0 and the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. Because of this, expression (6) can be very complicated, but in the limit it can be simplified if one assumes more about the estimator $\hat{\theta}_n$[1].

The most common simplifying assumption is that $\hat{\theta}_n$ is asymptotically efficient; that is, that it has an asymptotically linear (or Bahadur) representation:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i,\theta_0) + o_p(1) \tag{7}$$

where ψ is such that

$$\int \psi(x,\theta_0)\mathrm{d}F(x,\theta_0) = 0, \qquad \int \psi(x,\theta_0)\psi^\top(x,\theta_0)\mathrm{d}F(x,\theta_0) = J \tag{8}$$

and J is a finite $p \times p$ positive definite matrix. Many estimators satisfy this condition; see for example Khmaladze (1979, pp. 289-290). Under these conditions, it can be shown using (6) that (under the time change $t = F(x,\theta_0)$) the parametric empirical process converges weakly to a mean zero Gaussian process on $[0,1]$ with covariance function

$$\rho(s,t) = s \wedge t - st - g(s,\theta_0)^\top \int_0^t H(r)\mathrm{d}r - g(t,\theta_0)^\top \int_0^s H(r)\mathrm{d}r + g(s,\theta_0)^\top J g(t,\theta_0) \tag{9}$$

where $H(t) = \psi(x,\theta_0)\big|_{x=F^{-1}(t,\theta_0)}$. See for example the clever proof of Durbin (1973a, Lemma 3), or del Barrio (2007, Section 4.2) for an elegant derivation. As was shown in Durbin (1973a), when a maximum likelihood estimator exists and the model has a finite Fisher information matrix $I(\theta)$ (which requires more regularity conditions on F and its density f that will be discussed below,) we have $\psi(x,\theta_0) = I^{-1}(\theta_0)\nabla_\theta \log f(x,\theta_0)$, $\int_0^t H(r)\mathrm{d}r = I^{-1}(\theta_0)g(t,\theta_0)$ and $J = I^{-1}(\theta_0)$. Then the covariance function of the limiting Gaussian process is reduced to

$$\rho(s,t) = s \wedge t - st - g^\top(s,\theta_0)I^{-1}(\theta_0)g(t,\theta_0). \tag{10}$$

[1]It should be noted that for the purposes of hypothesis testing it is not strictly necessary that this relationship be known, if one employs the transformation technique of Khmaladze (1981) discussed in Section 4.

The above result (that is, the additional terms in expressions (9) and (10) as compared to the covariance function of the Brownian bridge,) is the source of what has been called the Durbin problem (Koenker and Xiao, 2002, p. 1589). In the of case the examples discussed in Section 5, a maximum likelihood estimator exists and so the covariance function takes the form of (10).

2.1 Location-scale and scale-shape models

Two classes of commonly used parametric models will be represented in the examples below, because when the hypothesized distribution is a member of one of these classes, the parametric empirical process does not depend on specific parameter values.

The first of these is the well-known class of location-scale models. Models in this class have distribution functions that take the form

$$F(x,\theta) = F_0\left(\frac{x-\theta_1}{\theta_2}\right); \qquad x \in \mathcal{X} \subseteq \mathbb{R}, \quad \theta \in \mathbb{R}\times(0,\infty) \tag{11}$$

for a fixed function F_0. Process-based goodness-of-fit tests for location models have analogs based on regression residuals. The earliest example of such tests is Loynes (1980). For a more recent treatment, see Koul (2002, Chapter 6), Koul (2006) or Khmaladze and Koul (2004).

The second class may be called scale-shape models: these models have distribution functions of the form

$$F(x,\theta) = F_0\left(\left(\frac{x}{\theta_1}\right)^{\theta_2}\right); \qquad x \in \mathcal{X} \subseteq [0,\infty), \quad \theta \in (0,\infty)\times(0,\infty). \tag{12}$$

Scale-shape models include the Weibull, Pareto and exponential models. These models have a natural connection to duration models — see, for example Hong and Liu (2007), Hong and Liu (2009) and the references cited therein. This invariance for scale-shape models was noted, with some examples, by Martynov (2009). See Appendix B for more on both model classes.

The null hypotheses considered in the examples will be that F is a certain location-scale or scale-shape model. It will be assumed that maximum likelihood estimators exist and the Fisher information matrix is finite. For these families, that is equivalent to the condition that F_0 has an absolutely continuous density f_0 that is positive on its support and has a derivative $\dot{f}_0$ almost everywhere, and such that

$$\sup_{x\in\mathbb{R}} |x| f_0(x) < \infty \quad \text{and} \quad \int (\dot{f}_0/f_0))^2(x) + (1 + x(\dot{f}_0/f_0)(x))^2 \mathrm{d}F_0(x) < \infty \tag{13}$$

for location-scale families (cf. Koul (2006, eq. (1.6))) or

$$\sup_{x\in\mathcal{X}} x\log x f_0(x) < \infty \quad \text{and} \quad \int (1 + x(\dot{f}_0/f_0)(x))^2 + (1 + \log x + x\log x(\dot{f}_0/f_0)(x))^2 \mathrm{d}F_0(x) \tag{14}$$

for scale-shape families[2] This simplifies the presentation while still including a large number of com-

[2]One might also consider a model in which a transformation of x is nested in a location-scale or scale-shape model, such as the lognormal model. As long as the transformation does not depend on parameters of the model in which it is nested, this invariance continues to hold.

monly specified parametric models; however, see Rabinovitz (1993) for examples in which the parametric empirical process depends on estimated quantities.

Kulinskaya (1995, Theorem 3) shows that the above model classes have this invariance essentially because the score function of the model can be decomposed into a part that is a function of θ and a second part that is a function of x. This decomposability leads to cancellation when constructing the covariance function (9).

3 Approximate boundary crossing probabilities

Asymptotic critical values for Kolmogorov-Smirnov-type tests are derived from boundary crossing probabilities for the weak (Gaussian) limit of the empirical process used in the construction of the test statistic. For example, letting v be the weak limit of v_n from equation (1), the standard one-sided Kolmogorov-Smirnov test relies on critical values derived from the distribution of $D_+ = \sup_{t\in[0,1]} v(t)$, or the probability that v crosses some horizontal boundary for t between 0 and 1. However, even for one-dimensional Gaussian processes, exact boundary crossing probabilities have analytic expressions for only a few special cases beyond Brownian motion and the Brownian bridge. The presence of the extra terms in (9) and (10) (compared to the covariance function of the Brownian bridge) implies that the weak limit of the parametric empirical process depends in general on the hypothesized parametric model and the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. These complicated covariance functions also imply that the limiting processes are non-Markovian and nonstationary. Faced with this challenge, Durbin (1985) proposed approximate boundary-crossing probabilities for Gaussian processes under very weak conditions and applied these results to the limits of parametric empirical processes.

Let y be a continuous mean-zero Gaussian process on $[0,1]$ starting at the origin. Assume y has a differentiable covariance function $\rho(s,t)$, let $a > 0$, and define $\tau_a = \inf\{t : y(t) = a\}$ — i.e., the first time y reaches the boundary $a(t) \equiv a$. In order to find the boundary crossing probability P defined by

$$P(a) = \mathrm{P}\left\{\sup_{t\in[0,1]} \{y(t)\} \geq a\right\}, \tag{15}$$

Durbin showed that $P(a)$ can be characterized by the integral of a boundary crossing density $p(t,a)$:

$$P(a) = \int_0^1 p(t,a)\mathrm{d}t = \int_0^1 b(t,a)f(t,a)\mathrm{d}t \tag{16}$$

where

$$b(t,a) = \lim_{s\to t} \frac{\mathrm{E}\left[\mathrm{I}(s < \tau_a)\left(a - y(s)\right)|y(t) = a\right]}{t-s} \tag{17}$$

and

$$f(t,a) = \frac{1}{\sqrt{2\pi\rho(t,t)}} \exp\left\{\frac{-a^2}{2\rho(t,t)}\right\}. \tag{18}$$

However, the function b is almost always intractable; this complication motivated three approximate boundary crossing probabilities.

3.1 The first approximation P_1

Durbin's first approximation, achieved simply through the removal of the indicator function from (17), was justified by the fact that the approximation holds exactly in the special case of Brownian motion and more generally by the fact that any Gaussian process satisfying Durbin's (mild) conditions "... behaves locally like Brownian motion and the boundary is locally linear[3]" (Durbin, 1985, p. 110-111). That is, approximation P_1 starts with the following approximation to the function b:

$$b_1(t,a) = \frac{\rho_1(t,t)}{\rho(t,t)} a \tag{19}$$

where $\rho_1(t,t) := \frac{\partial \rho(s,t)}{\partial s}\Big|_{s=t}$. This approximation to b owes its simple form to a hypothetical regression argument and the definition of a derivative[4]. Approximations to the first passage density for y and the boundary crossing probability are respectively

$$p_1(t,a) = b_1(t,a) f(t,a) \tag{21}$$

and

$$P_1(a) = \int_0^1 p_1(t,a)\mathrm{d}t. \tag{22}$$

Given ρ and ρ_1, $P_1(a)$ is easy to compute for simple parametric models. Since the difference between b and b_1 becomes smaller as $a \to \infty$, P_1 is a better approximation of P for small α.

3.2 The global approximation P_g and large deviations for Gaussian processes

Durbin also derived a "rough estimate" of P_1 that obviates the final integration step between p_1 and P_1 above. His "rough estimate" is remarkably accurate in most cases for quantiles of practical interest.

[3] Durbin (1985) considered differentiable boundaries, not just constant boundaries.

[4] After removing the indicator function from b, we have

$$b_1(t,a) = \lim_{s \nearrow t} \frac{a - \mathrm{E}\left[y(s)|y(t)=a\right]}{t-s}. \tag{20}$$

Imagine a hypothetical regression of $y(s)$ on $y(t)$, without an intercept. Then we would have $\mathrm{E}\left[y(s)|y(t)=a\right] = \frac{\rho(s,t)}{\rho(t,t)}a$. The rest is the definition of a derivative.

Interestingly, a separate branch of research on extrema of Gaussian processes and fields can be used to show that this estimate is identical to a crossing probability that becomes exact as the boundary diverges. The work of Tyurin (1985) was perhaps the first such derivation, but the focus here will be on the work of Fatalov (1992, 1993) and Piterbarg (1996). The reason for this is that the theory underlying the result is quite general and has roots that are more clearly probabilistic — the work of Tyurin relies on the solution to boundary value problems and can usually only be applied in the same situations as Durbin's P_g, while Fatalov's work allows one to compute probabilities not available via Durbin's or Tyurin's formulation of the solution. Fatalov's results are based on the theory of large deviations for Gaussian processes; all the relevant results can be found in self-contained form in Piterbarg (1996).

Let the variance function of the process be defined as $\sigma^2(t) := \rho(t,t)$ and the point of maximal variance $t_0 := \arg\max_t \sigma^2(t)$. Durbin's global approximation P_g is

$$P_g(a) = \frac{\rho_1(t_0,t_0)}{\sigma^2(t_0)} \left(\frac{-2\sigma^2(t_0)}{\frac{\mathrm{d}^2\sigma^2(t)}{\mathrm{d}t^2}\big|_{t=t_0}} \right)^{1/2} \exp\left\{ \frac{-a^2}{2\sigma^2(t_0)} \right\}. \tag{23}$$

This is achieved by starting with equation (21), evaluating all the non-exponential parts at t_0, and replacing the exponential part with a rough expansion to evaluate it. This formula is easily inverted for the purposes of calculating approximate critical values.

Some other important features of Durbin's P_g are contained in the following theorem.

Theorem 1. *Suppose the parametric family F is such that $\frac{\mathrm{d}^2}{\mathrm{d}x\mathrm{d}\theta} f(x,\theta)$ is finite for all (x,θ). Let y be the weak limit of the parametric empirical process for F, with covariance function ρ and variance function $\sigma^2(t) = \rho(t,t)$, and let $t_0 = \arg\max_{t\in[0,1]} \sigma^2(t)$. Then the approximate probability that y crosses the horizontal boundary a is*

$$P_g(a) = \frac{\exp\left\{ \frac{-a^2}{2\sigma^2(t_0)} \right\}}{2\sqrt{-\sigma^2(t_0)\left(\rho_{11}(t_0,t_0) + \rho_{12}(t_0,t_0)\right)}}. \tag{24}$$

Proof. Durbin's approximation P_g in (23) requires that $\frac{\mathrm{d}^2}{\mathrm{d}^2 t}\sigma^2(t)$ be finite. This is implied by the condition that $\frac{\mathrm{d}^2}{\mathrm{d}x\mathrm{d}\theta} f(x,\theta)$ is finite: the derivatives of the covariance function for the parametric empirical process are (letting $s \le t$ and suppressing dependence on θ as an argument in the functions g and I)

$$\rho_1(s,t) = 1 - t - \dot{g}^\top(s) I_\theta^{-1} g(t), \qquad \rho_2(s,t) = -s - g^\top(s) I_\theta^{-1} \dot{g}(t) \tag{25}$$

and the second derivatives are

$$\rho_{11}(s,t) = -\ddot{g}^\top(s) I_\theta^{-1} g(t), \qquad \rho_{12}(s,t) = -\dot{g}^\top(s) I_\theta^{-1} \dot{g}(t) \qquad \rho_{22}(s,t) = -g^\top(s) I_\theta^{-1} \ddot{g}(t). \tag{26}$$

When evaluated at $s = t$, we find that $\rho_{11}(t,t) = \rho_{22}(t,t)$, and their existence is implied by the existence of $\ddot{g}$, which in turn is implied by the above assumption on the density of the model, because

the second derivative of g involves derivative terms up to $\frac{\partial^3 F(x,\theta)}{\partial x^2 \partial \theta}\big|_{x=F^{-1}(t,\theta)}$.

By the definition of t_0,

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma^2(t)\Big|_{t=t_0} = \rho_1(t_0,t_0) + \rho_2(t_0,t_0) = 0. \tag{27}$$

We also have, from (25),

$$\rho_1(t,t) - \rho_2(t,t) = 1 \tag{28}$$

for all t. Putting these two equations together we find that at t_0,

$$\rho_1(t_0,t_0) = -\rho_2(t_0,t_0) = 1/2. \tag{29}$$

Inserting (29) and (26) into (23), we have the result. ■

A drawback to the use of P_g is that if $\frac{\mathrm{d}^2\sigma^2(t_0)}{\mathrm{d}t^2}\big|_{t=t_0} = 0$ (which occurs, e.g., when testing $\mathcal{N}(\mu,\sigma^2)$ with μ unspecified,) P_g does not exist. Some more explicit calculations of P_g for the normal and exponential distributions are presented in Appendix A. Furthermore, there is only a rough understanding that P_g becomes more accurate as the boundary diverges. Both of these issues are addressed formally in the following theorem, due originally to Fatalov (1992, 1993). Note that an attractive feature of Theorem 2 is that convergence to the true boundary crossing probability is at a relatively quick rate as the boundary diverges: in Durbin (1985, p. 113), it could only be estimated that P_g approaches P_1 at a polynomial rate.

Theorem 2. *Let y be the weak limit of the parametric empirical process, with variance function σ^2. Assume that σ^2 has a derivative of some order $2k$ ($k \in 1,2,\ldots$) that is nonzero (negative) at $t_0 = \arg\max_{t\in[0,1]} \sigma^2(t)$. Then the probability that the parametric empirical process crosses level a in $[0,1]$ is*

$$\mathrm{P}\left\{\sup_{t\in[0,1]} X(t) > a\right\} = H(A,C,k)\left(\frac{a}{\sigma(t_0)}\right)^{1-1/k} \phi\left(\frac{a}{\sigma(t_0)}\right)(1+o(1)), \quad a\to\infty \tag{30}$$

where ϕ is the standard normal density,

$$H(A,C,k) = \frac{C}{kA}\Gamma\left(\frac{1}{2k}\right) \tag{31}$$

and

$$A = \left(\frac{\left|\frac{\mathrm{d}^{(2k)}}{\mathrm{d}t^{(2k)}}\sigma^2(t_0)\right|}{2(2k)!\sigma^2(t_0)}\right)^{1/(2k)}, \qquad C = \frac{1}{2\sigma^2(t_0)}. \tag{32}$$

Proof. Because the third term of σ^2 is the quadratic form $g^\top(t)I^{-1}g(t)$, the first nonzero derivative at t_0 will be of even order $2k$. A Taylor expansion around t_0 shows that the standard deviation of y

locally about t_0 is

$$\sigma(t_0) = \sigma(t) + \frac{1}{2(2k)!\sigma(t_0)} \frac{\mathrm{d}^{(2k)}}{\mathrm{d}t^{(2k)}} \sigma^2(t) \Big|_{t=t_0} |t - t_0|^{(2k)} (1 + o(1)), \quad t \to t_0 \tag{33}$$

because all derivatives of order lower than $2k$ are zero. By Lemma 1, the correlation of y locally about t_0 is

$$r(s,t) = 1 - \frac{1}{2\sigma^2(t_0)} |t - s| (1 + o(1)), \quad s, t \to t_0. \tag{34}$$

These results, combined with Theorem 8.2 of Piterbarg (1996) imply the result. Specifically, the fact that the order of the first term in the expansion of the correlation is 1 and for the standard deviation the order is $2k > 1$ implies that case (i) of the theorem applies. Specialized to this context, we have

$$\mathrm{P}\left\{ \sup_{t \in [0,1]} y(t) > a \right\} = H(A, C, k) \left(\frac{a}{\sigma(t_0)} \right)^{2-1/k} \Psi\left(\frac{a}{\sigma(t_0)} \right) (1 + o(1)), \quad a \to \infty \tag{35}$$

where

$$H(A, C, k) = H_1 \int_{\mathbb{R}} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t. \tag{36}$$

The value H_1 is a constant special to this literature and is known to equal 1. Using the substitution $x = t^{2k}$, one finds

$$H(A, C, k) = \int_{\mathbb{R}} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t = 2 \int_{[0,\infty)} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t = \frac{C}{kA} \Gamma\left(\frac{1}{2k} \right) \tag{37}$$

Finally use the relation

$$a\Psi(a) = \phi(a)(1 + o(1)) \tag{38}$$

in (35) to establish the result. ■

Lemma 1. *Let y be the weak limit of the parametric empirical process with differentiable covariance function ρ. Then*

$$r(s,t) = 1 - \frac{1}{2\sigma^2(t_0)} |t - s| (1 + o(1)), \quad s, t \to t_0 \tag{39}$$

where $r(s,t) = \rho(s,t)/\sqrt{\sigma^2(s)\sigma^2(t)}$ is the correlation function of the process.

Proof. Expanding the squared covariance function $\rho^2(s,t)$ in s around t results in

$$\rho^2(s,t) = \rho^2(t,t) + 2\rho(t,t)\rho_1(t,t)(s - t)(1 + o(1)), \quad s \to t, \tag{40}$$

while an expansion of $\rho(s,s)$ in s around t implies

$$\rho(s,s) = \rho(t,t) + [\rho_1(t,t) + \rho_2(t,t)](s-t)(1+o(1)), \quad s \to t. \tag{41}$$

This implies that

$$\begin{aligned}
\rho^2(s,t) - \rho(s,s)\rho(t,t) &= \rho^2(t,t) + 2\rho(t,t)\rho_1(t,t)(s-t) \\
&\quad - \rho^2(t,t) - \rho(t,t)[\rho_1(t,t) - \rho_2(t,t)](s-t) + o(s-t), \quad s \to t \\
&= \rho(t,t)[\rho_1(t,t) - \rho_2(t,t)](s-t) + o(s-t) \\
&= \rho(t,t)(s-t)(1+o(1)), \quad s \to t,
\end{aligned} \tag{42}$$

this last equality occuring because $\rho_1(t,t) - \rho_2(t,t) = 1$ for all t. The Cauchy-Schwarz inequality and the fact that $\rho(t,t) = \rho(t_0,t_0) + o(1)$ imply we can rewrite the above as

$$= -\sigma^2(t_0)|t-s|(1+o(1)), \quad s,t \to t_0. \tag{43}$$

Then, using the definition of correlation and the expansion $\sqrt{1-x} = 1 - \frac{1}{2}x(1+o(1)),\ x \to 0$ we have that

$$\begin{aligned}
r(s,t) &= \sqrt{1 - \frac{\sigma^2(t_0)}{\sigma^2(s)\sigma^2(t)}|t-s|(1+o(1))} \\
&= 1 - \frac{1}{2\sigma^2(t_0)}|t-s|(1+o(1)), \quad s,t \to t_0.
\end{aligned} \tag{44}$$

■

Unfortunately, it is difficult to see how the steps Durbin took to arrive at P_g are at all related to the techniques developed in Piterbarg (1996). However, in the light of this other work one recognizes a few details that make Durbin's approximation a good one. First, Durbin conjectures that the point of maximal variance is the only point needed to compute his approximation, because for boundaries that are high enough, the probability that a crossing will occur anywhere else becomes extremely small[5]. This is formally justifiable; see for example Piterbarg (1996, "Stage 2", p. 21 or the corresponding part of Theorem 8.1, p. 120-121). Second, the assumption that the variance function is twice differentiable is satisfied in a great number of parametric models, so this is not a strong assumption. This is also assumed in the work of Tyurin (1985), and although not a formal assumption in the general theory of Piterbarg (1996), this differentiability is implicitly used in the calculation of the constants necessary for the asymptotic results given there.

[5]Tyurin and Fatalov both point out that the maximal variance need not occur at a single point — the variance of the process used to test the Cauchy distribution has two points of maximum, for example.

3.3 The Gauss-Markov approximation P_2

As noted above, the limit of the parametric empirical process is generally a non-Markovian, nonstationary Gaussian process. Because this limit is non-Markovian, its increments may be related in complicated ways. Durbin's suggestion was essentially to calculate boundary crossing probabilities as if this inconvenience did not exist. This final approximation improves upon P_1 and is the solution to a numerically evaluated integral equation. A great deal of tractability is gained through this simplification, and the examples below suggest that the results are, perhaps surprisingly, quite accurate.

Let y be a mean-zero Gaussian process with covariance function ρ. Define[6]

$$\begin{bmatrix} \beta_1(s,t) \\ \beta_2(s,t) \end{bmatrix} = \begin{bmatrix} \rho(s,s) & \rho(s,t) \\ \rho(t,s) & \rho(t,t) \end{bmatrix}^{-1} \begin{bmatrix} \rho_2(s,t) \\ \rho_1(t,t) \end{bmatrix}. \tag{45}$$

The approximate density $p_2(t,a)$ is the solution of the integral equation

$$p_2(t,a) = p_1(t,a) - a\int_0^t \left[\beta_1(s,t) + \beta_2(s,t)\right] f(t|s,a) p_2(s,a)\mathrm{d}s. \tag{46}$$

In (46), $p_1(t,a)$ is as in equation (21) and $f(t|s,a)$ is the value of the transition density of the process on the boundary a at time t given that the process is on the boundary at time $s \leq t$ and assuming it is Markovian; in the present case of a constant boundary, the transition distribution is

$$F(t|s,a) = F(y(t)|y(s) = a) = \mathcal{N}\left(\frac{\rho(s,t)}{\rho(s,s)}a, \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)}\right) \tag{47}$$

and the density is evaluated at a.

Equation (46) holds exactly for Gaussian processes that are also Markovian, and it was Durbin's suggestion to use this relation as an approximation method for non-Markovian processes as well. That is, given a Gaussian process $\hat{v}$ (the weak limit of $\hat{v}_n$) with covariance function ρ, the Gauss-Markov approximation to the probability that $\hat{v}$ crosses $a > 0$ in $[0,1]$ is given by

$$P_2(a) = \int_0^1 p_2(t,a)\mathrm{d}t \tag{48}$$

where p_2 has been computed as one would compute p_2 in (46). This disregards the potentially intractable autocovariance structure of the process but also delivers reasonable results, as will be seen in Section 6.

[6]This is similar to the linear estimate in the derivation of p_1 in that it comes from consideration of a hypothetical regression of $y(r)$ on $y(t)$ and $y(s)$, $s,t \leq r$.

3.3.1 Gauss-Markov processes

A Gauss-Markov process is a Gaussian process that also satisfies the Markov property[7] (it need not be stationary — for example, the Brownian bridge is nonstationary.) This immediately implies that a mean-zero process with covariance function ρ has transition probabilities that can be characterized as

$$(x,t)|(y,s) \sim \mathcal{N}\left(\frac{\rho(s,t)}{\rho(s,s)}y, \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)}\right). \tag{49}$$

Mehr and McFadden (1965) derive several important results for these processes. These results stem from the fact that the covariance functions of such processes must be triangular; that is, a Gaussian process is also Markovian if and only if its covariance function ρ satisfies, for all $0 \leq r \leq s \leq t$

$$\rho(r,t) = \frac{\rho(r,s)\rho(s,t)}{\rho(s,s)}. \tag{50}$$

Because of this, there must exist (differentiable) functions η and ζ such that $\rho(s,t) = \eta(s)\zeta(t)$. Furthermore, it can be shown (Doob, 1953; Mehr and McFadden, 1965) that all such processes are scaled, time-changed Brownian motions: that is, if y is a Gauss-Markov process and W is standard Brownian motion, then η/ζ is strictly increasing and we have the representation

$$y(t) = \zeta(t)W\left((\eta/\zeta)(t)\right). \tag{51}$$

Using these results, Di Nardo et al. (2001) have shown that Durbin's derivation is a special case of a result on boundary crossing probabilities for diffusion processes found in Buonocore et al. (1987). A mean-zero Gauss-Markov process is a diffusion process with a transition probability density function f that satisfies the Fokker-Planck equation

$$\frac{\partial}{\partial t}f(x,t|y,s) = -\frac{\partial}{\partial x}\left\{A_1(x,t)f(x,t|y,s)\right\} + \frac{A_2(t)}{2}\frac{\partial^2}{\partial x^2}f(x,t|y,s) \tag{52}$$

with $\lim_{s\to t} f(x,t|y,s) = \delta(x-y)$ (Di Nardo et al., 2001), and where

$$A_1(x,t) = \lim_{s\to t}\frac{\partial}{\partial t}\frac{\rho(s,t)}{\rho(s,s)}y = \frac{\rho_2(t,t)}{\rho(t,t)}y \tag{53}$$

and

$$A_2(t) = \lim_{s\to t}\frac{\partial}{\partial t}\rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)} = \rho_1(t,t) - \rho_2(t,t) \tag{54}$$

The function A_2 in particular is intimately connected to Durbin's approximation— see equation (47) above and equation (4) of Durbin (1985). The function A_1 is also strikingly similar to equation (19) above, especially given the fact that for the parametric empirical process, $\rho_1(t,t) - \rho_2(t,t) = 1$ for all

[7] That is, if a process y is defined on the filtration $\mathscr{F}$, it satisfies the Markov property if $\mathrm{E}\left[y_t|\mathscr{F}_s\right] = \mathrm{E}\left[y_t|y_s\right]$ for $s \leq t$.

t.

It may be noted that a Gauss-Markov process allows several integral equations involving the first passage density to be derived; for example, one may start with the Chapman-Kolmogorov equations that are so fundamental to Markov processes. In particular, one particularly simple formulation is the following, which uses an argument analogous to Peskir (2002, Theorem 2.2)[8]:

Theorem 3. *Let $y : T \to \mathbb{R}$, $T \subset [0,\infty)$ be a Gaussian Markov process with a.s.-continuous sample paths such that $\mathrm{P}\{y_0 = 0\} = 1$, mean function $m(t) \equiv 0$ and covariance function $\rho(s,t)$ such that y has regular conditional probabilities. Let $a > 0$, let*

$$\tau_a = \inf\{t > 0 : y_t \geq a\}$$

be the first exit time of y from the set $(-\infty, a)$, $a > 0$, and let F be the distribution function of τ_a. Then for all t for which y is well defined, the following integral equation is satisfied:

$$\Psi\left(\frac{a}{\sqrt{\rho(t,t)}}\right) = \int_0^t \Psi\left(\frac{a - m(s,t)}{\sqrt{V(s,t)}}\right) p(s,a)\mathrm{d}s \tag{55}$$

where

$$m(s,t) = \frac{\rho(s,t)}{\rho(s,s)}a \quad \text{and} \quad V(s,t) = \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)} \tag{56}$$

and $\Psi = 1 - \Phi$, where Φ denotes the standard normal cumulative distribution function.

Proof. The result follows from the combination of Peskir (2002, Theorem 2.2) and the transition distributions of Gauss-Markov processes, given above in (49). Namely, because y is Markovian,

$$\mathrm{P}\{y_t \in B\} = \int_0^t \mathrm{P}\{y_t \in B | y_s = a\}\mathrm{d}F(s) \tag{57}$$

for all measurable $B \subseteq [a,\infty)$. Given the distributions (49),

$$\mathrm{P}\{y_t \in [a,\infty)\} = \Psi\left(\frac{a}{\sqrt{\rho(t,t)}}\right) \tag{58}$$

because $\mathrm{P}\{y_0 = 0\} = 1$ and

$$\mathrm{P}\{y_t \in [a,\infty) | y_s = a\} = \Psi\left(\frac{a - m(s,t)}{\sqrt{V(s,t)}}\right) \tag{59}$$

where m and V are defined above. The distribution of τ_a has a density because of the relationship between Brownian motion and y, that is, equation (51). ∎

[8] One might also start with a similar equation due to Fortet; see Durbin (1971, Section 2) for a derivation.

The connection between the integral equations (55) and (46) is not as straightforward as it might seem. Differentiating equation (55) with respect to t results in another integral equation that is remarkably similar to equation (46). Despite the similarities, only a circuitous connection can be made[9]—see Di Nardo et al. (2001) and Buonocore et al. (1987). The decision regarding which integral equation to employ in computing the critical values presented in Section 5 was made on practical grounds: although equation (55) is slightly simpler to put into practice, Durbin's second-kind Volterra equation was more stable in numerical experiments.

3.3.2 Computation of p_2

Equation (46) is a nonseparable Volterra integral equation of the second kind and thus must be solved numerically. Elementary methods can be used to calculate the solution. Following Press et al. (2001, p. 786), one simple algorithm is a recursively computed numerical integral that steps forward from 0 to 1 on an equally spaced grid. The properties of ρ make this easy to accomplish: the kernel of the integral equation — $-a(\beta_1(s,t)+\beta_2(s,t))f(t|s,a)$, for $s \leq t$ — has a limiting value of 0 whenever t or s are 0, 1, or equal to each other. Given an equally-spaced partition $\{t_i = (i-1)/m,\ i = 1,2,\cdots m+1\}$ (the value of m is chosen by the researcher,) the integration algorithm simplifies to the following recursive rule: for $i = 0, 1$ (recall $t_0 = 0$),

$$p_2(0,a) = 0, \qquad p_2(t_2,a) = p_1(t_2,a) \tag{60}$$

and for $i \geq 3$

$$p_2(t_i,a) = p_1(t_i,a) + a\frac{1}{m}\sum_{j=2}^{i-1} K(t_j,t_i)p_2(t_j,a) \tag{61}$$

where $K(\cdot,\cdot)$ is the kernel of the integral equation.

A partition of $(0,1)$ using m subintervals for numerical integration results in accuracy of order $O(1/m^2)$ for any a; as it appeared that convergence was slower than theory predicted in small experiments, the value of m was set at $10,000$ to produce the results below. The weighting technique proposed by Di Nardo et al. (2001) did not appear to have an effect on final critical value estimates, and so was not used in the calculations.

This technique will be applied to specification tests in Section 5. As an alternative to working directly with distributionally dependent statistics, the next section explores a technique that is designed to bypass this dependence through the construction of a different process that results in asymptotically distribution-free statistics.

[9]Once again, this is because both equations can be related to the result of Fortet (cf. Durbin (1971).)

3.4 Discussion

The approximations discussed above are useful alternatives to simulation methods for sup-norm tests. Although there is no clear theoretical way to quantify the relationship between Durbin's approximations and the true boundary crossing probability for the limit of the parametric empirical process, the arguments above are strong evidence in support of their accuracy. In fact, Theorem 2 is strong evidence that all of the approximations perform quite well, since it applies to P_g, and Durbin's original intent was that this approximation be the roughest of the three. One possible drawback to the approach outlined below should be noted: since the approximates presented in this section are applied to the Gaussian limit of the parametric empirical process, there is no formal guarantee that they necessarily improve as the sample size of a given experiment increases. However, in the case examined in Section 6, performance is not affected as sample size increases. It seems likely that this is due to the accuracy of the approximations relative to small sample anomalies.

Furthermore, these methods are generalizable. While Subsection 3.2 may appear to be only a serendipitous confluence of results from some quite different theoretical starting points, it should be noted that the body of theory represented in Piterbarg (1996) is very general and applicable to a wide variety of Gaussian processes and fields, and as such may serve as a fruitful point of departure for solutions to more general problems, for example the extension of these techniques to test statistics that converge to Gaussian processes in higher dimensions. On the other hand, approximation P_2 is also very flexible — it may be applied to any sup-norm test for which the empirical process has a Gaussian limit, as is for example the case with the empirical characteristic function (Matsui and Takemura, 2005, Theorem 2.1). For goodness of fit tests based on regression residuals, very few modifications must be made; in the dynamic case, some regularity is required on the sequence of score functions to ensure weak convergence of the process — see Bai (2003). However, beyond these conditions, it is only required that the covariance function of the limiting process be tractable enough to be used in the formulas above. On the other hand, addressing problems for which estimators are not efficient is more challenging. If $\hat{\theta}$ only satisfies assumption **A2** above but is not efficient, the covariance function needs to be derived on a case-by-case basis. The method presented in the next Section can be used in such situations.

These approximations are attractive because they tie the adjusted critical values of a test to the parametric family being tested through the function g, the score function of the model. This makes this testing strategy more involved than simulation for the applied researcher, but it makes it possible to understand more about the test under consideration and the relationship between the model and the test statistic. In addition, as will be seen in Section 6, there is reason to believe that in more complicated settings, tests that use adjusted critical values can perform at least as well as tests that rely on simulation methods.

4 Khmaladze's martingale transform

An alternative approach to the problem of testing a statistical model with estimated parameters was suggested by Khmaladze (1979, 1981). He proposed a transformation of the empirical process that is not affected asymptotically by the estimation of model parameters, thereby avoiding the problems inherent in the use of the parametric empirical process. In the one-sample setting, some interesting connections can be made between the martingale transform, the parametric empirical process, and simple projection techniques.

Viewed as a real-valued random element of $L_2[0,1]$, $\mathbb{F}_n$ is a submartingale with respect to $\mathscr{F}^{\mathbb{F}_n} = \{\mathscr{F}_t^{\mathbb{F}_n}\}_{t\geq 0}$, the filtration of σ-algebras generated by $\mathbb{F}_n$. Therefore the Doob-Meyer decomposition implies a right-continuous increasing and predictable compensator K may be calculated that renders $\mathbb{F}_n - K$ a martingale with respect to $\mathscr{F}^{\mathbb{F}_n}$. At any point x in the support of F, the compensator $K(x, \mathbb{F}_n, \theta)$ is asymptotically equivalent to the conditional expectation $\mathrm{E}\left[\mathbb{F}_n(x) \middle| \mathbb{F}_n(y), y \leq x, \theta\right]$.

The process

$$\tilde{V}_n(x) = \sqrt{n}\left(\mathbb{F}_n(x) - K(x, \mathbb{F}_n, \hat{\theta}_n)\right) \tag{62}$$

is called the transformed empirical process, and Khmaladze (1981) showed that $\tilde{V}_n$ converges weakly in $L_2[0,1]$ to $W \circ F$, a time changed Brownian motion. This renders statistics based on process (62) asymptotically distribution-free.

The intimate relation between the function g defined in equation (4) and the compensator is somewhat more clear if one makes the time transformation $t = F(x, \theta)$ — in the sequel, $g(t, \theta)$ will generally be shortened to $g(t)$ when the parameters used in the transformation and the evaluation of the function are identical. The reason for this is that it can be shown that $\dot{g}$, the derivative of g with respect to t, satisfies the equation

$$\dot{g}(t) = \frac{\partial}{\partial t} g(t, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta)\Big|_{x = F^{-1}(t,\theta)} \tag{63}$$

implying that g is in effect the integrated score function for the model. The compensator $K(t, \mathbb{F}_n, \hat{\theta})$ is a projection of changes in the empirical distribution function onto the score of the null model. With this in mind, define the $p+1$ dimensional extended score function h and the $(p+1)\times(p+1)$-dimensional function Γ by

$$h(t,\theta) = \begin{bmatrix} 1 \\ \frac{\partial g(t,\theta)}{\partial t} \end{bmatrix} \qquad \text{and} \qquad \Gamma(t,\theta) = \int_t^1 h(s,\theta) h(s,\theta)^\top \mathrm{d}s. \tag{64}$$

Finally, let the compensator K be defined as follows: for any $t \in (0,1)$

$$K(t, \mathbb{F}_n, \theta) = \int_0^t h(s,\theta)^\top \Gamma^{-1}(s,\theta) \int_s^1 h(r,\theta) \mathrm{d}\mathbb{F}_n(r) \mathrm{d}s. \tag{65}$$

It is usually easier to perform computations using the following equivalent expression:

$$= \int_0^1 \int_0^{t\wedge r} h(s,\theta)^\top \Gamma^{-1}(s,\theta) \mathrm{d}s\, h(r,\theta) \mathrm{d}\mathbb{F}_n(r). \tag{66}$$

One may think of equation (65) as a functional analog to $\hat{y} = x\hat{\beta}$ familiar from usual regression analysis, with $h(t)$ playing the role of explanatory variable and the projection $\Gamma^{-1}(t)\int_t^1 h(s)\mathrm{d}\mathbb{F}_n(s)$ as $\hat{\beta}$. Note also the fact that $\Gamma(0,\theta)$ is simply an augmented version of the Fisher information matrix of the model: because of the similarities between h and the score, and Γ and the Fisher information, it can be shown that the compensator also has a form that does not depend on parameter values when the null model is a member of the special classes of parametric models discussed in Subsection 2.1. See Appendix B for more on this topic. For a more general interpretation of the martingale transform as a projection onto the score function of a parametric model, see Li (2009).

Although the compensator may be difficult to calculate analytically, it can be easily implemented using a projection technique employing recursive least squares and the score function from the null model. This ease of implementation is an attractive feature of the martingale transform method. The details are addressed in Subsection 4.1. It should also be noted that this technique need not be limited to tests of Kolmogorov-Smirnov type; after transformation of the empirical process, any functional can be used to derive an asymptotically distribution-free test statistic, for example an L^2 statistic like the Cramér-von Mises statistic. The approximation methods of Section 3 are approximate boundary crossing probabilities, and as such they only apply to sup-norm tests, although L^2 analogs exist (i.e. Durbin et al. (1975))

4.1 Computation of the compensator

Khmaladze's compensator can be calculated using standard recursive least squares and numerical integration methods on a finite partition of $[0,1]$ — see Bai (2003, Appendix B) for an alternate explanation. Its accuracy depends only on the fineness of the partition used for integration.

Suppose we have a partition $\{t_i\}$ of the unit interval. First, least squares coefficients $\{\hat{\beta}_i\}_{i=1}^m$ are generated at each t_i by projecting the empirical distribution function onto the score of the model for each $\{t_j\}_{j\geq i}$. Then, projections are integrated from 0 to each t_i to make a "prediction" of the score function integrated up to the t^{th} quantile of the null model.

This can be accomplished as follows. Suppose we use the same evenly spaced grid as in Subsection 3.3.2. The score and empirical distribution functions are evaluated at each t_i and then stacked into the following series of matrices of size $(m-i+2)\times 2$ and $(m-i+2)\times 1$ respectively:

$$X_i = \begin{bmatrix} \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_{m+1}) \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_m) \\ \vdots & \vdots \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_i) \end{bmatrix} \qquad y_i = \begin{bmatrix} \sqrt{m}\left(\mathbb{F}_n(t_{m+1}) - \mathbb{F}_n(t_m)\right) \\ \sqrt{m}\left(\mathbb{F}_n(t_m) - \mathbb{F}_n(t_{m-1})\right) \\ \vdots \\ \sqrt{m}\left(\mathbb{F}_n(t_i) - \mathbb{F}_n(t_{i-1})\right) \end{bmatrix} \tag{67}$$

Then, least squares coefficients for each t_i are calculated:

$$\begin{aligned}\hat{\beta}(t_i) &= (X_i^\top X_i)^{-1} X_i^\top y_i \\ &= \begin{bmatrix} \frac{1}{m}(m-j+2) & \frac{1}{m}\sum_{j=i}^{m+1} \dot{g}(t_j) \\ \frac{1}{m}\sum_{j=i}^{m+1} \dot{g}(t_j) & \frac{1}{m}\sum_{j=i}^{m+1} \dot{g}^2(t_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=i}^{m+1} [\mathbb{F}_n(t_j) - \mathbb{F}_n(t_{j-1})] \\ \sum_{j=i}^{m+1} \dot{g}(t_j)[\mathbb{F}_n(t_j) - \mathbb{F}_n(t_{j-1})] \end{bmatrix}. \end{aligned} \tag{68}$$

That is, for each t_i, $\hat{\beta}(t_i)$ is the projection of changes in $\{\mathbb{F}_n(t_j)\}_{j\geq i}$ onto $\{h(t_j)\}_{j\geq i}$. Given the form of $\{X_i\}_i$ and $\{y_i\}_i$ it can be seen that rather than generating $m-p+1$ very similar X and y matrices, an efficient way to calculate the sequence $\{\hat{\beta}(t_i)\}_i$ is via recursive least squares from t_{m-p+1} to t_1. Then for any t_i the compensator $\hat{K}(t_i)$ is obtained by integrating numerically:

$$\hat{K}(t_i) = \frac{1}{m}\sum_{j=1}^{i} h^\top(t_j)\hat{\beta}(t_j). \tag{69}$$

Here it can be seen why Bai (2003) called the martingale transform method a "continuous time detrending operation" using the score function of the model. The above algorithm is simply a discretized approximation to the operator K. As such, each estimate $\hat{K}$ is subject to some approximation error that shrinks as the size of the partition (m) increases. That is, because the inverted matrix term of equation (68) and h are continuous functions of t, the accuracy of the numerical integral (69) increases with m.

5 Examples

One-sample tests of exponentiality and normality with estimated parameters are simple examples with which one can compare the approaches proposed by Durbin and Khmaladze. For tests of exponentiality there is one parameter[10], while for tests of normality there are two parameters and therefore a greater variety of distributionally dependent boundary crossing probabilities. The martingale transform is illustrated analytically for the exponential case, a result first presented in Haywood and Khmaladze (2008) and developed here under the time transformation $t = F(x, \theta_0)$. Khmaladze and Koul (2004) and Khmaladze and Koul (2009) discuss some features of the compensator for the null hypothesis of normality, although it is tedious to express it analytically. Some other examples may be found in Koul and Sakhanenko (2005).

5.1 The exponential distribution

The exponential model has convenient distribution and quantile functions. The hypothesis of exponentiality is

$$H_0 : F(x,\lambda) = 1 - e^{-\lambda x}, \quad x \in [0,\infty),\ \lambda \in (0,\infty). \tag{70}$$

[10]Martynov (2009) shows that the calculation of the parametric empirical process for the Weibull model is only marginally more complicated than for the exponential model, but an analytic expression for the compensator is difficult to derive.

The function g for the exponential model is

$$g(s,\lambda) = \frac{-1}{\lambda_0}(1-s)\log(1-s)e^{\frac{\lambda}{\lambda_0}} \tag{71}$$

and the weak limit of the parametric empirical process for tests of exponentiality with efficiently estimated parameters is a mean-zero Gaussian process with covariance function

$$\rho(s,t) = s \wedge t - st - (1-s)(1-t)\log(1-s)\log(1-t). \tag{72}$$

which clearly does not depend on any parameter values (this distribution is a member of the scale-shape class discussed in Subsection 2.1.)

The methods of Section 3 were applied using (72) to produce the approximate critical values in Table 1 for testing the hypothesis of exponentiality. The corresponding standard Kolmogorov-Smirnov critical values are included in the last column to give an impression of the magnitude of the difference between them and the distributionally dependent critical values. Note that since the third term in equation (10) is positive definite, the covariance function of the parametric empirical process is smaller than that of the Brownian bridge for all t, and therefore critical values for the Kolmogorov-Smirnov test using the parametric empirical process should always be smaller than for the standard test (van der Vaart and Wellner, 1996, p. 441).

Table 1: Approximate critical values for the composite hypothesis of exponentiality and corresponding classical Kolmogorov-Smirnov critical values. These values are invariant to the value of the scale parameter.

Significance Level	P_1	P_g	P_2	K-S
10%	0.89401	0.88055	0.87726	1.07298
5%	1.00063	0.99105	0.98983	1.22387
2.5%	1.09766	1.09042	1.09013	1.35810
1%	1.21464	1.20930	1.20955	1.51743

Both P_g and P_2 adjust the first approximation P_1 downward slightly. Although it is a global approximation, the values of P_g are extremely close to those produced using P_1 and P_2: for purposes of quick approximation, P_g offers reasonable precision with very little computation.

5.1.1 The compensator for the exponential case

Khmaladze's compensator for the exponential distribution is presented here on $t \in [0,1]$. For the exponential distribution, straightforward computation reveals that

$$h(t,\lambda) = \begin{bmatrix} 1 \\ \frac{1}{\lambda}(1+\log(1-t)) \end{bmatrix} \tag{73}$$

and

$$\Gamma(t,\lambda)=\begin{bmatrix}1-t & \frac{1}{\lambda}(1-t)\log(1-t)\\ \frac{1}{\lambda}(1-t)\log(1-t) & \frac{1}{\lambda^2}(1-t)(1+\log^2(1-t))\end{bmatrix}. \tag{74}$$

From here one can compute the compensator for any t. Let $\{\hat{\varepsilon}_i\}_{i=1}^n = \{F(X_i,\hat{\lambda})\}_{i=1}^n$ for some appropriate estimator $\hat{\lambda}$. Then

$$\begin{aligned}K(t,\mathbb{F}_n,\hat{\lambda})=&\int_0^t \frac{1}{2}\log^2(1-\hat{\varepsilon})-2\log(1-\hat{\varepsilon})-\log^2(1-\hat{\varepsilon})\mathrm{d}\mathbb{F}_n(\hat{\varepsilon})\\ &+\int_t^1 \frac{1}{2}\log^2(1-t)-2\log(1-t)-\log(1-\hat{\varepsilon})\log(1-t)\mathrm{d}\mathbb{F}_n(\hat{\varepsilon}),\end{aligned} \tag{75}$$

or alternatively

$$\begin{aligned}K(t,\mathbb{F}_n,\hat{\lambda})=&\frac{1}{n}\sum_{i:\hat{\varepsilon}_i\le t}\left(\frac{-1}{2}\log^2(1-\hat{\varepsilon}_i)-2\log(1-\hat{\varepsilon}_i)\right)\\ &+\left(\frac{1}{2}\log^2(1-t)-2\log(1-t)\right)\big(1-\mathbb{F}_n(t)\big)-\frac{1}{n}\log(1-t)\sum_{i:\hat{\varepsilon}_i>t}\log(1-\hat{\varepsilon}_i),\end{aligned} \tag{76}$$

both of which depend only on the parameter estimate through $\{\hat{\varepsilon}_i\}_i$. Note that without making the transformation $t=F(x,\theta)$ Haywood and Khmaladze (2008) derive this compensator, which is

$$\begin{aligned}\tilde{K}(x,\mathbb{F}_n,\hat{\lambda})=&\frac{\hat{\lambda}}{n}\sum_{i:X_i\le x}\left(2X_i-\frac{\hat{\lambda}}{2}X_i^2\right)\\ &+\left(2\hat{\lambda}x+\frac{\hat{\lambda}^2}{2}x^2\right)(1-\mathbb{F}_n(x))-\frac{\hat{\lambda}^2}{n}x\sum_{i:X_i>x}X_i\end{aligned} \tag{77}$$

but from this expression it is not apparent that the form of the compensator is independent of the value of the estimate $\hat{\lambda}$.

5.2 The normal distribution

The normal model is also of interest. The hypothesis of normality is

$$H_0:\ F(x,\theta)=\int_{-\infty}^{x}\frac{e^{\frac{-1}{2\sigma^2}(y-\mu)^2}}{\sqrt{2\pi\sigma^2}}\mathrm{d}y,\quad x\in\mathbb{R},\ \theta=(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty). \tag{78}$$

All of Durbin's approximations are available for the $\mathcal{N}(0,\sigma^2)$ distribution with σ^2 unknown and with μ and σ^2 unknown, but P_g does not exist for the mean-unknown case (the result as computed in Appendix A is shown in Table 2.) Letting $\xi_{\mu\sigma}(s)$ be the s^{th} quantile of the $\mathcal{N}(\mu,\sigma^2)$ distribution, the

function g for the mean- and variance-unknown case is equal to

$$g(s,\theta) = \begin{bmatrix} \frac{\partial}{\partial\mu}\int_{-\infty}^{\xi_{\mu\sigma}(s)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}\,\mathrm{d}y \\ \frac{\partial}{\partial\sigma^2}\int_{-\infty}^{\xi_{\mu\sigma}(s)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}\,\mathrm{d}y \end{bmatrix} = \begin{bmatrix} \frac{-e^{\frac{(\xi_{\mu\sigma}(s)-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \\ \frac{-(\xi_{\mu\sigma}(s)-\mu)}{2\sigma^2} \frac{e^{\frac{-(\xi_{\mu\sigma}(s)-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \end{bmatrix}. \tag{79}$$

Since the normal model is in the location-scale class, specific parameter values can be ignored and standard normal quantiles can be used (see Appendix B.) Then, letting ξ and ϕ be respectively the quantile and density functions of the standard normal distribution, one finds that the limit of the parametric empirical process has the covariance function

$$\rho_{\mu\sigma}(s,t) = s \wedge t - st - \phi(\xi(s))\phi(\xi(t))\left(1 + \frac{1}{2}\xi(s)\xi(t)\right). \tag{80}$$

The global approximation in this case is $P_g(a) = \sqrt{\frac{2\pi}{\pi-2}}\exp\{-2\pi a^2/(\pi-2)\}$.

Table 2: Approximate critical values for the composite hypothesis of normality. These values are invariant to parameter values, although they change according to the combination of parameters left unspecified in the null hypothesis. The values of P_g are computed using the methods of Fatalov (1992, 1993); see Appendix A for more details.

Significance Level	P_1	P_g	P_2
Both parameters unspecified			
10%	0.76690	0.75716	0.74979
5%	0.84364	0.83620	0.83274
2.5%	0.91429	0.90839	0.90673
1%	1.00036	0.99581	0.99526
Mean unspecified			
10%	0.82311	0.82541	0.81305
5%	0.90099	0.90299	0.89410
2.5%	0.97198	0.97375	0.96690
1%	1.05786	1.05940	1.05421
Variance unspecified			
10%	1.04103	1.02466	1.03443
5%	1.19298	1.18174	1.18906
2.5%	1.32857	1.32026	1.32604
1%	1.48967	1.48365	1.48810

The diagonal nature of the information matrix for the normal model makes the third term of the covariance function additive in the two parameters. Therefore the covariance functions for the other two possible cases are immediate. For the mean-unknown case we have

$$\rho_{\mu}(s,t) = s \wedge t - st - \phi(\xi(s))\phi(\xi(t)) \tag{81}$$

As mentioned above, the approximation P_g does not exist in this case, because the second derivative of $\rho(t,t)$ evaluated at $t^* = 1/2$ is equal to zero. Similarly, the covariance function in the variance-

unspecified case is

$$\rho_\sigma(s,t) = s \wedge t - st - \frac{1}{2}\xi(s)\xi(t)\phi(\xi(s))\phi(\xi(t)) \tag{82}$$

and $P_g(a) = (2/3)^{1/2}\exp\{-2a^2\}$. Note that there is a small error in this expression in Durbin (1985, p. 117); a sketch of the derivations required appears in Appendix A.

Approximate critical values are presented in Table 2. The values are all quite close to one another; as in the exponential case, the values of P_g and P_2 are uniformly lower than those of P_1. Due to the invariance of the limiting process to parameter values in these model classes, one may use the standard normal distribution to compute the values given in all the tables of this section; the resulting values are the same for any configuration of parameter values.

6 Monte Carlo experiments

Table 3 presents the results of a small Monte Carlo experiment using the D^- statistic for testing the null hypothesis of exponentiality against stochastically dominant alternatives Both the Gauss-Markov approximation and the martingale transform were included. Because there is an analytic form for the compensator, the numerical approximation calculated as in Subsection 4.1 can be compared to the exact version. A partition of $m = 1.5n$ points in the interval was used for the recursive least squares algorithm for the compensator. This is meant to reflect the fact that in some cases (for example, quantile regression processes,) the total number of points in the partition has an upper limit.

Table 3: Sizes (in percent) of a one-sided sup-norm test (D^-) using adjusted critical values or a martingale transform for a test of exponentiality. Nominal sizes appear in the column header. 50,000 repetitions.

sample size	10	5	2.5	1
50				
G-M approximation	10.41	4.92	2.36	0.92
analytic transform	11.03	4.53	1.72	0.46
RLS transform	8.77	3.60	1.42	0.37
Standard K-S	2.70	0.81	0.23	0.05
100				
G-M approximation	10.52	5.15	2.48	0.95
analytic transform	10.54	4.56	1.87	0.50
RLS transform	9.26	4.02	1.66	0.48
Standard K-S	2.84	0.83	0.26	0.06
200				
G-M approximation	10.36	5.04	2.44	0.97
analytic transform	10.12	4.64	1.96	0.57
RLS transform	9.42	4.38	1.87	0.57
Standard K-S	2.77	0.87	0.26	0.05

As theory predicts, naively applied classical Kolmogorov-Smirnov critical values result in tests that have a size much lower than the nominal size. The exact compensator leads to inferences that improve as the sample size increases, as is to be expected, although the improvement is smaller at lower levels

(cf. Table 1 of Haywood and Khmaladze (2008)). At the 10% and 5% levels, the process using the exact compensator is clearly closer to the nominal level than its discretized counterpart, but this relationship reverses at the 2.5% and 1% levels. The Gauss-Markov approximation results in tests that are reasonably close to their nominal size, although they appear to do slightly better for smaller sample sizes and for smaller levels. The compensator computed using recursive least squares ("RLS transform" in Table 3,) typically the only feasible transformed process, performs roughly as well as the Gauss-Markov approximation in most cases.

The power of these tests is not often addressed; notable exceptions include Aki (1986), Haywood and Khmaladze (2008) and Koul and Sakhanenko (2005), with some results on power for the martingale transformation technique. A second small Monte Carlo experiment was conducted using smooth local alternatives to the null hypothesis of exponentiality. Stochastically dominant alternatives were constructed in one of two ways. First, local alternative mixture densities were generated using the following formula:

$$f_{mix}(x,n) = \left(1 - \frac{c}{\sqrt{n}}\right) f_{exp}(x) + \frac{c}{\sqrt{n}} f_{alt}(x) \tag{83}$$

where f_{exp} is the exponential density and f_{alt} is a different density. These alternative densities were arbitrarily chosen to be lognormal$(0, 1/2)$, or uniform $[0, 4]$, with the parameters and constants c chosen so as to achieve nontrivial (i.e., not 0 or 100%) power for all the tests. Two other convergent alternative models that nest the exponential were considered: the gamma and weibull models. These alternatives were set with scale parameters equal to 1 and shape parameters equal to $1 + c/\sqrt{n}$. The tests considered were Durbin's P_2 and P_g approximations, compensated empirical processes calculated both analytically and using recursive least squares, and a bootstrap test.

The bootstrap was conducted following Romano (1988). That is, each sample was used to generate a bootstrapped critical value by estimating $\hat{\lambda}$ in the given sample and then producing 200 random exponential$(\hat{\lambda})$ samples with the same sample size as the original. Although it would be more natural in this simple setting to generate a critical value by simply simulating the distribution of the supremum of the parametric empirical process, the above algorithm was chosen to reflect a setting in which such a strategy was not an option.

The results of the power experiment appear in Table 4. The first row simply repeats the size of the tests, and the remaining rows are the empirical power from 50,000 simulated samples for the alternatives described above. It can be seen that the classical Kolmogorov-Smirnov critical values result in tests that are uniformly less powerful than tests using adjusted values, which is to be expected since the adjusted values are always smaller than the unadjusted ones. This bootstrap technique tends to be less powerful than using tests with asymptotically derived critical values. However, it is also of interest to note that of the two alternative strategies — to test with either an adjusted critical value or a transformed process — neither is a uniformly better test. For example, tests based on the transformed process do extremely well against the uniform alternative. On the other hand, they do not seem to do quite as well as tests using the parametric empirical process against the lognormal and gamma alternatives. Evidently these tests have differential performance against alternatives from different

Table 4: Empirical size and power for the alternatives described in the text. All tests are intended to have a size of 5%; e.g. the first row shows that the last four methods are more or less conservative in this experiment. 50,000 repetitions.

sample size	P_2	P_g	analytic transform	RLS transform	bootstrap	K-S
null model						
50	5.0	4.9	4.4	3.5	1.2	0.8
100	5.0	4.9	4.6	4.1	1.2	0.8
200	5.1	5.0	4.7	4.3	1.2	0.8
uniform mixture						
50	83	83	99	99	55	49
100	71	71	98	97	37	32
200	57	57	97	96	22	18
lognormal mixture						
50	40	40	34	31	19	16
100	40	40	33	32	19	16
200	40	40	33	32	18	16
gamma alternative						
50	56	56	53	49	28	24
100	62	62	59	57	34	30
200	67	67	63	62	39	36
weibull alternative						
50	51	51	55	51	25	21
100	55	55	59	57	28	25
200	59	58	63	61	31	28

parts of the space of alternatives.

7 Conclusion

The techniques examined in this paper exploit the structure of the parametric empirical process, in particular the score function under the null model. This function is the common thread that connects Khmaladze's transformation to the covariance function underlying Durbin's approximations. Using the exponential model, the martingale transform method is compared with two critical value approximations for the one-sample sup-norm test with estimated parameters. Monte Carlo evidence suggests that the approximations proposed by Durbin result in tests that have a size comparable to tests based on the transformed empirical process. It is also apparent that neither method dominates the other in terms of power, although the experiment suggests that these tests are more powerful than bootstrap tests.

Appendix A: P_g

In order to clarify equation (23), Durbin's global approximation, some further details are presented for the specific cases mentioned in the examples.

First of all, the calculation of t_0 is straightforward: for the normal cases, simple optimization shows

$\arg\max_t \rho(t,t) = 1/2$, while for the exponential distribution, t_0 must satisfy the following equation:

$$1 - 2t_0 + 2(1-t_0)\left(\log(1-t_0) + \log^2(1-t_0)\right) = 0. \tag{84}$$

Using a numerical root-finding procedure, one finds that the value of t_0 is approximately 0.3398 for the exponential case. The rest of the calculations for the exponential case must also be done numerically. However, it is possible to calculate P_g analytically for the two normal cases mentioned above.

For the two computable normal cases (i.e., when both parameters or only the variance parameter are unspecified,) the second derivatives of each $\rho(t,t)$ are respectively

$$\frac{\mathrm{d}^2\rho_{\mu\sigma}(t,t)}{\mathrm{d}t^2} = -1 + \left(1 + \phi(\xi_0(t))\right)\xi_0^2(t) - \xi_0^4(t) \tag{85}$$

and

$$\frac{\mathrm{d}^2\rho_{\sigma}(t,t)}{\mathrm{d}t^2} = -3 + 4\xi_0^2(t) - \xi_0^4(t), \tag{86}$$

where ϕ is the standard normal density function and ξ_0 is the standard normal quantile function. When evaluated at $t_0 = 1/2$ we have -1 and -3 respectively.

Evaluating the above functions and the covariance functions together at the maximum $t_0 = 1/2$ (recall $\rho_1(t_0, t_0) = 1/2$ for all models) and putting everything together as in equation (23), we have

$$P_g(a) = \frac{1/2}{\frac{1}{4} - \frac{1}{2\pi}}\sqrt{\frac{-2\left(\frac{1}{4} - \frac{1}{2\pi}\right)}{-1}}\exp\left\{\frac{-a^2}{2\left(\frac{1}{4} - \frac{1}{2\pi}\right)}\right\} = \sqrt{\frac{2\pi}{\pi-2}}e^{\frac{-2\pi}{\pi-2}a^2} \tag{87}$$

and

$$P_g(a) = \frac{1/2}{1/4}\sqrt{\frac{-2/4}{-3}}\exp\left\{\frac{-a^2}{2/4}\right\} = \sqrt{2/3}e^{-2a^2}. \tag{88}$$

Large deviation approximations

The constants used in Fatalov's formulation of the boundary crossing probability for tests of normality, as presented in Theorem 1, are

$$(\hat{\mu}, \hat{\sigma}^2): \qquad \sigma^2(t_0) = \frac{\pi-2}{4\pi} \qquad A = \sqrt{\frac{\pi}{\pi-2}} \qquad C = \frac{2\pi}{\pi-2} \qquad k = 1 \tag{89}$$

$$(\mu, \hat{\sigma}^2): \qquad \sigma^2(t_0) = 1/4 \qquad A = \sqrt{3} \qquad C = 2 \qquad k = 1 \tag{90}$$

$$(\hat{\mu}, \sigma^2): \qquad \sigma^2(t_0) = \frac{\pi-2}{4\pi} \qquad A = \sqrt[4]{\frac{2\pi^2}{3(\pi-2)}} \qquad C = \frac{2\pi}{\pi-2} \qquad k = 2 \tag{91}$$

Note the value of A is different from what is printed in Piterbarg (1996) for two of three cases. Plugging these values into equation (30) results in

$$\mathrm{P}\left\{\sup_{t\in[0,1]} X(t) > a \middle| \hat{\mu}, \hat{\sigma}^2\right\} = \sqrt{\frac{2\pi}{\pi-2}}\, e^{\frac{-2\pi}{\pi-2}a^2} \tag{92}$$

$$\mathrm{P}\left\{\sup_{t\in[0,1]} X(t) > a \middle| \mu, \hat{\sigma}^2\right\} = \sqrt{2/3}\, e^{-2a^2} \tag{93}$$

$$\mathrm{P}\left\{\sup_{t\in[0,1]} X(t) > a \middle| \hat{\mu}, \sigma^2\right\} = \frac{\Gamma(1/4)}{\pi-2}\sqrt[4]{\frac{3\pi}{2}}\sqrt{a}\, e^{\frac{-2\pi}{\pi-2}a^2} \tag{94}$$

Appendix B: Location-scale and scale-shape families

These two classes of parametric families have the attractive feature that their score functions may be separated into two parts: one that contains parameter values and one that contains only functions that depend on the model. The location-scale case is very well-known (e.g. Shorack and Wellner (1986, Section 5.5),) the scale-shape case was noted as a general phenomenon by Martynov (2009), and both were noted as special cases in Kulinskaya (1995).

Members of the location-scale class, defined by the equivalence (11), have the following property:

$$g(t) = \nabla_\theta F(x,\theta)\big|_{x=F^{-1}(t,\theta)} = \frac{-1}{\theta_2}\begin{bmatrix} f_0(F_0^{-1}(t)) \\ F_0^{-1}(t) f_0(F_0^{-1}(t)) \end{bmatrix} \tag{95}$$

and the score function inherits this separability, since the derivative of g with respect to t is

$$\dot{g}(t) = \nabla_\theta \log f(x,\theta)\big|_{x=F^{-1}(t,\theta)} = \frac{-1}{\theta_2}\begin{bmatrix} (\dot{f}_0/f_0)(F_0^{-1}(t)) \\ 1 + F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t)) \end{bmatrix} \tag{96}$$

This in turn implies that the information matrix also has a separable structure: that is,

$$I(\theta) = \int_{[0,1]} \dot{g}(t)\dot{g}^\top(t)\mathrm{d}t = \frac{1}{\theta_2^2}\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \frac{1}{\theta_2^2} I_0 \tag{97}$$

where each σ_{ij} can be derived from equation (96) and I_0 is a fixed matrix depending only on the model.

The situation is similar for the scale-shape class. For members of this class we have

$$g(t) = \begin{bmatrix} \frac{-\theta_2}{\theta_1} F_0^{-1}(t) f_0(F_0^{-1}(t)) \\ \frac{1}{\theta_2}\log(F_0^{-1}(t)) F_0^{-1}(t) f_0(F_0^{-1}(t)) \end{bmatrix} \tag{98}$$

and

$$\dot{g}(t) = \begin{bmatrix} \frac{-\theta_2}{\theta_1}\left(1 + F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t))\right) \\ \frac{1}{\theta_2}\left(1 + \log(F_0^{-1}(t)) + \log(F_0^{-1}(t))F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t))\right) \end{bmatrix} \tag{99}$$

so that

$$I(\theta) = \begin{bmatrix} \frac{\theta_2^2}{\theta_1^2}\sigma_{11} & \frac{-1}{\theta_1}\sigma_{12} \\ \frac{-1}{\theta_1}\sigma_{12} & \frac{1}{\theta_2^2}\sigma_{22} \end{bmatrix} \tag{100}$$

Although the information matrix is not as simple as for the location-scale class, parameters cancel in the calculations described below. From the form that the third term takes in the covariance function of the parametric empirical process when efficient estimators have been used,

$$g^\top(s)\left(\int_0^1 \dot{g}(r)\dot{g}^\top(r)\mathrm{d}r\right)^{-1} g(t), \tag{101}$$

it is straightforward to show that the terms that depend on parameters cancel, for members of either the location-scale or scale-shape class. Note also that the conditions given for finite Fisher information in Subsection 2.1, equations (13) and (14), are the same as the assumption that $\dot{g}$ exists a.e. and $\int \dot{g}\dot{g}^\top < \infty$. The result is analogous for the compensator — it is constructed using only the augmented score function h, and as such, the parameter values in the integrand of the compensator,

$$h(s,\theta)^\top\left(\int_s^1 h(s,\theta)h^\top(s,\theta)\mathrm{d}s\right)^{-1}\int_s^1 h(r,\theta)\mathrm{d}\mathbb{F}_n(r) \tag{102}$$

can be factored out in the same way.

References

S. Aki. Some test statistics based on the martingale term of the empirical distribution function. *Annals of the Institute of Statistical Mathematics*, 38(1):1–21, 1986.

J. Bai. Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics*, 85(3):531–549, 2003.

P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

A. Buonocore, A. Nobile, and L. Ricciardi. A new integral equation for the evaluation of first-passage-time probability densities. *Advances in Applied Probability*, 19(4):784–800, 1987.

A. Cabaña and E. Cabaña. Transformed empirical processes and modified Kolmogorov-Smirnov tests for multivariate distributions. *Annals of Statistics*, 25(6):2388–2409, 1997.

E. del Barrio. *Lectures on Empirical Processes: Theory and Statistical Applications*, chapter Empirical and Quantile Processes in the Asymptotic Theory of Goodness-of-fit Tests, pages 1–92. EMS Series of Lectures in Mathematics. European Mathematical Society, 2007.

M. Delgado and W. Stute. Distribution-free specification tests of conditional models. *Journal of Econometrics*, 143(1):37–55, 2008.

E. Di Nardo, A. Nobile, E. Pirozzi, and L. Ricciardi. A computational approach to first-passage-time problems for Gauss-Markov processes. *Advances in Applied Probability*, 33(2):453–482, 2001.

J. Doob. *Stochastic Processes*. Wiley, 1953.

J. Durbin. Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test. *Journal of Applied Probability*, 8 (3):431–453, 1971.

J. Durbin. Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, 1(2):279–290, 1973a.

J. Durbin. *Distribution Theory for Tests Based on the Sample Distribution Function*. Number 9 in Regional Conference Series in Applied Mathematics. SIAM, 1973b.

J. Durbin. Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika*, 62(1):5–22, 1975.

J. Durbin. The first-passage density of a continuous Gaussian process to a general boundary. *Journal of Applied Probability*, 22(1):99–122, 1985.

J. Durbin, M. Knott, and C. Taylor. Components of the Cramér-von Mises statistics. II. *Journal of the Royal Statistical Society, Series B (Methodological)*, 37(2):216–237, 1975.

V. Fatalov. Asymptotics of large deviation probabilities for Gaussian fields. *Journal of Contemporary Mathematical Analysis*, 27(3):48–70, 1992.

V. Fatalov. Asymptotics of large deviation probabilities for Gaussian fields: Applications. *Journal of Contemporary Mathematical Analysis*, 28(5):21–44, 1993.

J. Haywood and E. Khmaladze. On distribution-free goodness-of-fit testing of exponentiality. *Journal of Econometrics*, 143(1):5–18, 2008.

Y. Hong and J. Liu. Generalized residual-based specification testing for duration models with censoring. Cornell University, 2007.

Y. Hong and J. Liu. Goodness-of-fit testing for duration models with censored grouped data. Cornell University, 2009.

E. Khmaladze. The use of ω^2 tests for testing parametric hypotheses. *Theory of Probability and its Applications*, 24(2):283–301, 1979.

E. Khmaladze. Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability and its Applications*, 26(2):240–257, 1981.

E. Khmaladze and H. Koul. Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics*, 32(3):995–1034, 2004.

E. Khmaladze and H. Koul. Goodness-of-fit problem for errors in nonparametric regression: Distribution free approach. *The Annals of Statistics*, 37(6A):3165–3185, 2009.

R. Koenker and Z. Xiao. Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612, 2002.

H. Koul. *Weighted Empirical Processes in Dynamic Nonlinear Models*, volume 166 of *Lecture Notes in Statistics*. Springer, 2nd edition, 2002.

H. Koul. Model diagnostics via martingale transforms: A brief review. In J. Fan and H. Koul, editors, *Frontiers in Statistics*, chapter 9, pages 183–206. Imperial College Press, 2006.

H. Koul and L. Sakhanenko. Goodness-of-fit testing in regression: A finite sample comparison of bootstrap methodology and Khmaladze transformation. *Statistics & Probability Letters*, 74(3):290–302, 2005.

E. Kulinskaya. Coefficients of the asymptotic distribution of the Kolmogorov-Smirnov statistic when parameters are estimated. *Journal of Nonparametric Statistics*, 5(1):43–60, 1995.

B. Li. Asymptotically distribution-free goodness-of-fit testing: A unifying view. *Econometric Reviews*, 28 (6):632–657, 2009.

R. Loynes. The empirical distribution function of residuals from generalised regression. *The Annals of Statistics*, 8(2):285–298, 1980.

G. Martynov. Goodness-of-fit tests for the Weibull and Pareto distributions. Paper presented at the Sixth International Conference on Mathematical Methods in Reliability, 2009.

M. Matsui and A. Takemura. Empirical characteristic function approach to goodness-of-fit tests for the Cauchy distribution with parameters estimated by MLE or EISE. *Annals of the Institute of Statistical Mathematics*, 57(1):183–199, 2005.

C. Mehr and J. McFadden. Certain properties of Gaussian processes and their first-passage times. *Journal of the Royal Statistical Society, Series B (Methodological)*, 27(3):505–522, 1965.

G. Neuhaus. *Weak Convergence Under Contiguous Alternatives when Parameters are Estimated: the D_k approach*, volume 566 of *Lecture Notes in Mathematics*, pages 68–82. Springer, 1976.

G. Peskir. On integral equations arising in the first-passage problem for Brownian motion. *Journal of Integral Equations and Applications*, 14(4):397–423, 2002.

V. Piterbarg. *Asymptotic Methods in the Theory of Gaussian Processes and Fields*, volume 148 of *Translations of Mathematical Monographs*. American Mathematical Society, 1996.

W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, 2001.

D. Rabinovitz. Estimating Durbin's approximation. *Biometrika*, 80(3):671–680, 1993.

J. Romano. A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708, 1988.

G. Shorack and J. Wellner. *Empirical Processes with Applications to Statistics*. Wiley, 1986.

K. Song. Testing semiparametric conditional moment restrictions using conditional martingale transforms. *Journal of Econometrics*, 154(1):74–84, 2010.

Y. Tyurin. On the limit distribution of Kolmogorov-Smirnov statistics for a composite hypothesis. *Mathematics of the USSR — Izvestiya*, 25(3):619–646, 1985.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.