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

Thomas Parker*

Department of Economics, University of Waterloo

Email: tmparker@uwaterloo.ca

December 3, 2012

**Abstract**

Goodness of fit tests based on parametric empirical processes have nonstandard limiting distributions when the null hypothesis is composite — that is, when parameters of the null model are estimated. Several analytic 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 conduct inference for tests based on supremum-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. Simulation experiments suggest that in small samples, Durbin-style adjustments result in tests that have higher power than tests based on transformed processes, and in some cases have higher power than parametric bootstrap procedures.

**Keywords:** Goodness of fit test, Estimated parameters, Gaussian process, Gauss-Markov process,

*The author wishes to express great appreciation to Roger Koenker for countless helpful discussions and able direction. This research also benefited from the valuable comments of Juan Carlos Escanciano, Andreas Hagemann, Marie Hušková, and Kyungchul Song. The author also wishes to thank Peter C. B. Phillips, the Coeditor and referees at Econometric Theory for their helpful suggestions. Finally, the author wishes to thank the late James Durbin, who in many ways inspired this entire project.

Boundary crossing probability, Martingale transformation

**JEL Classification Code:** C12, C14, C46

# 1 Introduction

Empirical processes are central to the theory of supremum-norm specification tests. The analysis of the empirical process $\sqrt{n}(\mathbb{F}_n - F_0)$ when $F_0$ is a fixed distribution function is quite well established, but a general study of the convergence of empirical processes when $F$ has estimated parameters was first conducted by Durbin (1973a) and Neuhaus (1976). The limiting distributions of these processes is significantly more complex than the limiting distribution of the simpler process. As a result, the evaluation of sup-norm test statistics based on these processes has been an enduring problem. Given this difficulty, inference based on an empirical process when parameters have been estimated is quite often accomplished via simulation techniques. There are, however, alternative solutions that can be derived analytically. In this paper, two such solutions are compared with each other and with some other proposed tests of model specification.

Parametric models, when reasonably accurate, help analysts interpret data in a manner that is easy to understand and communicate. However, model misspecification can result in incorrect inferences and policy decisions. The tests discussed here illustrate some difficulties in the evaluation of specification tests when the relevant hypothesis is with respect to a parametric model — a family of curves indexed by a finite-dimensional parameter. No matter what the actual value of the parameters of the model, the assumption that the data may be reasonably described by some member of the hypothesized parametric model dictates which estimators may be considered optimal and how to conduct inference.

For example, it is often convenient to use an exponential model when modeling duration data, due to its simple distribution, density and hazard functions. However, in order to confidently use this model in applications, one would like to be sure that the exponential model provides a reasonable approximation to the real distribution of the data. Suppose an analyst is confronted with a sample $\{X_i\}_{i=1}^n$ that are iid with distribution $F$. If the analyst is only concerned with testing the adequacy of

the exponential model, that is, that the $F$ describing the distribution of the sample is

$$F_{exp}(x, \lambda) = 1 - e^{-\lambda x}, \quad x \in \mathbb{R}_+, \, \lambda > 0 \tag{1}$$

for *some* value of $\lambda$ (i.e., but without the hypothesis that $\lambda = \lambda_0$). A convenient and powerful test for this hypothesis uses a supremum-norm statistic; that is, the largest difference between the empirical cumulative distribution function $\mathbb{F}_n$ and the theoretical model $F_{exp}$:

$$T_n = \sup_{x \in \mathbb{R}_+} \sqrt{n} \left| \mathbb{F}_n(x) - F_{exp}(x, \lambda^*) \right|, \tag{2}$$

where a candidate member of the exponential model, corresponding to a specific value $\lambda^*$, must be used in the test statistic. Were there also a reasonable hypothesized value $\lambda^* = \lambda_0$, then the typical Kolmogorov-Smirnov asymptotic distribution can be used for inference. However, the distribution of $T_n$ and corresponding critical values or p-values are affected by the value of $\lambda^*$ used by the analyst when no $\lambda_0$ is given by hypothesis. Subsection 5.1 discusses methods for drawing inferences from $T_n$ when the candidate $\lambda^*$ is the efficient (under the null hypothesis) maximum likelihood estimator $\hat{\lambda} = \bar{X}^{-1}$.

The linear models often used in applied work offer another example of the applications of the methods discussed in this paper; the most basic of such models assumes that the joint distribution of a response $y$ and covariates $X$ is well-described by the model

$$y_i = X_i^\top \beta + \sigma \varepsilon, \quad (\beta, \sigma) \in \mathbb{R}^p \times \mathbb{R}_+. \tag{3}$$

The use of this model implicitly assumes the distribution of $\varepsilon$ is a member of a location-scale family (i.e., one such that the cumulative distribution function of $\varepsilon$ satisfies $F(e) = F_0((e - \mu)/\sigma)$ for a fixed function $F_0$ and some $\mu$ and $\sigma$). Furthermore, when the distribution of $\varepsilon$ is assumed to be Gaussian, ordinary least squares is the optimal estimator, regardless of the specific values of $\beta$ and $\sigma$. When the error distribution is wrongly assumed to be the Gaussian model, the least squares estimator may be inefficient and inference regarding the magnitudes and statistical significance of $\beta$ and $\sigma$ can be incorrect. Under the assumption that the model is correctly specified — both the linear regression form and the Gaussianity of the error term — a sup-norm test statistic used to test the hypothesis $F_0 \equiv \Phi$,

the standard normal distribution function, is

$$T_n^{'} = \sup_{e \in \mathbb{R}} \sqrt{n} \left| \mathbb{F}_n(e) - \Phi\left(e/\hat{\sigma}\right) \right| \tag{4}$$

where now $\mathbb{F}_n$ is the empirical distribution function of the least squares residuals $\{y_i - X_i^\top \hat{\beta}\}_{i=1}^n$. Once again, the use of $(\hat{\beta}, \hat{\sigma})$ in the construction of the test statistic causes the distribution of $T_n^{'}$ to be nonstandard. This example is discussed more in Subsection 5.3.

One solution to the problem of nonstandard distributions for supremum-norm tests (parallel to techniques devised for example by Durbin et al. (1975) for Cramér-von Mises-type tests,) is to conduct distributionally dependent inference. For sup-norm tests, Durbin (1973b, 1975, 1985), explored a number of such inferential methods and these results deserve greater recognition as an alternative methodology. In particular, it is demonstrated below that Durbin (1985) provides a collection of simple approximations that are accurate, generalizable, and involve only modest computation. These rely on approximate boundary crossing probabilities that are analyzed in Section 3. Some justification for their great accuracy is provided by links that the approximations have to approximation results from other areas of probability theory. One of Durbin's approximations is a special case of results derived using the theory of extrema of Gaussian fields (see, for example, Piterbarg (1996)). Another is an approximation to the distribution of the statistic using a simplification that arises for Gauss-Markov processes. The present work supports and refines Durbin's research in the methodology of goodness of fit testing in econometrics — even though a goodness of fit problem was the primary applied example of Durbin (1985), his boundary crossing results have been largely overlooked.

Another 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 is convenient because the resulting statistics are asymptotically pivotal, implying that drawing inferences using (asymptotic) p-values or critical values is the same procedure, regardless of the hypothesized parametric model. 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). Wooldridge (1990) also proposes a similar testing strategy, using orthogonal projections to remove the effect of parameter estimation in moment-based specification tests. It is shown in Subsection 4.2 that when adapted to the relevant setting, Wooldridge's test statistic is different because it does not rely on all of the information contained in the null hypothesis (which is not finite-dimensional) in the same way as Khmaladze's projection.

Durbin's approximate boundary crossing probabilities are 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 these methods and compare them to bootstrap-based procedures. The adjusted inferential procedures result in approximately the same size and power as tests using a transformed process, although the experiments suggest differential power performance over the space of alternatives.

Adjusted inferential procedures appear attractive on several dimensions. Their size appears roughly comparable to simulation-based tests, and they have better power in some cases, while they often appear to have power at least as good as that of tests based on transformed empirical processes. On the other hand, their implementation is quite straightforward — transformation of the process is not required and a simple formula (presented in Theorem 1 below) is often sufficient to conduct accurate inference.

## 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 $\Theta \subseteq \mathbb{R}^p$. Process-based specification tests for $F$ are typically based on one of the following empirical processes: the uniform empirical process

$$V_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x,\theta_0)), \qquad x \in \mathcal{X} \tag{5}$$

for simple null hypotheses, or the parametric empirical process

$$\hat{V}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x, \hat{\theta})) \qquad x \in \mathcal{X} \tag{6}$$

for composite null hypotheses, where $\hat{\theta}$ is some estimate of $\theta_0$ and $\mathbb{F}_n$ is the empirical distribution function.

It is assumed that all members of $\mathcal{F}$ are absolutely continuous and mutually absolutely continuous. The uniform empirical process is convenient because under these 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 makes process (5) equivalent to

$$v_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( I(F(X_i, \theta_0) \leq t) - t \right), \quad t \in [0, 1]. \tag{7}$$

That is, under the null hypothesis, process (5) is equivalent to a process based on $n$ iid realizations of a uniform random variable and the value of $V_n$ (or $v_n$) measures the difference between the empirical distribution of $\{F(X_i, \theta_0)\}_i$ and the uniform distribution function. Donsker's theorem implies that $v_n$ converges weakly to $v$, a Brownian bridge on $[0, 1]$ — equivalently, $V_n$ 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 $\theta_0$. 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, one with respect to the parametric model and one with respect to the parameter estimate:

**A1** The model $\mathcal{F}$ satisfies the following condition: the function

$$g(t, \theta) = \nabla_\theta F(x, \theta)\big|_{x = F^{-1}(t, \theta_0)} \tag{8}$$

is bounded and continuous in its arguments for all $(t, \theta) \in [0, 1] \times \nu$, where $\nu$ is a closed neighborhood of $\theta_0$ in $\Theta$.

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

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_P(1). \tag{9}$$

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 $\theta_0$, the effect of parameter estimation is not asymptotically negligible. Consider the following decomposition of $\hat{v}_n(t)$ (start here with the transformation $t = F(x, \hat{\theta})$):

$$\begin{aligned}
\hat{v}_n(t) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big( I(F(X_i, \hat{\theta}) \leq t) - t \Big) && (10) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big( I(F(X_i, \theta_0) \leq t) - t \big) + \sqrt{n} \Big( F(F^{-1}(t, \theta_0), \hat{\theta}_n) - t \Big) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \Big( I(F(X_i \leq \hat{\theta}) \leq t) - F(F^{-1}(t, \theta_0), \hat{\theta}_n) \Big) - \big( I(F(X_i, \theta_0) \leq t) - t \big) \Big\} && (11)
\end{aligned}$$

Using assumptions **A1** and **A2** with a one-term Taylor expansion, it can be shown[1] that the last term in (11) is $o_P(1)$ uniformly in $t \in [0, 1]$ and that the following asymptotic linearity result holds:

$$\sup_{t \in [0,1]} \left| \hat{v}_n(t) - v_n(t) + \sqrt{n}(\hat{\theta}_n - \theta_0)^\top g(t, \theta_0) \right| = o_P(1). \tag{12}$$

Durbin (1973a) showed that $\hat{v}_n$ converges weakly to a mean-zero Gaussian process $\hat{v}$. From (12) it is apparent that in general the distribution of the limit $\hat{v}$ may depend on the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and even on the value of the parameter $\theta_0$ (through the function $g$).

Because the parametric empirical process depends on the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$, the distribution of (10) can be complex, but it can be simplified if more is assumed regarding the estimator $\hat{\theta}_n$[2].

**A3** Assume that $\hat{\theta}_n$ is asymptotically linear; that is,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, \theta_0) + o_P(1) \tag{13}$$

[1] See van der Vaart and Wellner (2007) for a general and elegant proof, which also applies to tests based on regression residual processes.

[2] Note that it is not necessary that this relationship be known if one employs the transformation technique of Khmaladze (1981) discussed in Section 4.

where $\psi$ is such that

$$\int \psi(x,\theta_0)\mathrm{d}F(x,\theta_0) = 0, \qquad \int \psi(x,\theta_0)\psi^\top(x,\theta_0)\mathrm{d}F(x,\theta_0) = J \tag{14}$$

and $J$ is a finite $p \times p$ positive definite matrix.

Under **A1**-**A3**, it can be shown[3] using (12) that (10) converges weakly to

$$\hat{v} \stackrel{D}{=} v - g^\top(t,\theta_0)\int \psi \mathrm{d}v \tag{15}$$

which is a mean-zero Gaussian process on $[0,1]$ with covariance function

$$\rho(s,t) = s \wedge t - st - g(s,\theta_0)^\top \int_0^t H(r)\mathrm{d}r - g(t,\theta_0)^\top \int_0^s H(r)\mathrm{d}r + g(s,\theta_0)^\top J g(t,\theta_0) \tag{16}$$

where $H(t) = \psi(x,\theta_0)\big|_{x=F^{-1}(t,\theta_0)}$. As was shown in Durbin (1973a), when a maximum likelihood estimator exists and the model has a finite Fisher information matrix $I(\theta)$ 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 process $\hat{v}$ is reduced to

$$\rho(s,t) = s \wedge t - st - g^\top(s,\theta_0)I^{-1}(\theta_0)g(t,\theta_0). \tag{17}$$

Expressions (16) and (17) are relatively complicated: recall that $v_n$ converges weakly to $v$, a Brownian bridge that has covariance function $\rho(s,t) = s \wedge t - st$. The extra terms in (16) and (17) reflect the effect of parameter estimation, and are the source of what has been called the Durbin problem (Koenker and Xiao, 2002, p. 1589). In the examples discussed in Section 5, a maximum likelihood estimator exists and so the covariance function takes the form of (17).

## 3 Approximate boundary crossing probabilities

Asymptotic critical values for Kolmogorov-Smirnov tests (i.e., tests using the process $v_n$) are derived from known formulas for boundary crossing probabilities of the limiting Brownian bridge $v$. For example, the standard one-sided Kolmogorov-Smirnov test relies on critical values derived from the distribution of $D_+ = \sup_{t\in[0,1]} v(t)$; equivalently, the probability that $v$ crosses some horizontal boundary.

[3] See for example the proof of Durbin (1973a, Lemma 3), or del Barrio (2007, Section 4.2) for an elegant derivation.

However, analytic expressions for boundary crossing probabilities have been found for only a few special Gaussian processes besides the Brownian motion and Brownian bridge. As described above, the distribution of the limiting process $\hat{v}$ depends in general on the hypothesized parametric model in a nontrivial way, and the distribution of $\sup_{t\in[0,1]} \hat{v}(t)$ is affected as well. Faced with this challenge, Durbin (1985) proposed approximate boundary-crossing probabilities for Gaussian processes under very weak conditions and applied these results to the process $\hat{v}$.

### 3.1 The exact boundary crossing probability $P$

Let $y$ be a continuous mean-zero Gaussian process on $[0,1]$ starting at the origin. The original motivation of Durbin (1985) was the analysis of boundary crossing probabilities for locally Brownian processes. Therefore, assume $y$ has a covariance function $\rho(s,t)$ that is differentiable in both arguments for $0 \le s \le t \le 1$. Note that this is weaker than full differentiability of $\rho$, because it is not necessary that $\rho$ be differentiable on the diagonal (for such processes, other methods are available for the computation of boundary crossing probabilities — see Azaïs and Wschebor (2009), for example). As an example, Brownian motion, with covariance function $\rho(s,t) = s \wedge t$, satisfies this assumption. The second assumption on $y$ is what makes the process locally Brownian: Durbin assumed that

$$\lim_{s\nearrow t} \frac{\mathrm{V}\big(y(t)-y(s)\big)}{t-s} = \lim_{s\nearrow t} \left\{ \frac{\partial \rho(s,t)}{\partial s} - \frac{\partial \rho(s,t)}{\partial t} \right\} = c_t \tag{18}$$

where $0 < c_t < \infty$ for all $t$. For example, Brownian motion satisfies this condition with $c_t \equiv 1$, as do processes with covariance functions (16) or (17), but the "incremental variance" need not be constant. Let $a > 0$, and define the first passage time $\tau_a = \inf\{t : y(t) = a\}$ — i.e., the first point at which $y$ reaches the boundary $a(t) \equiv a$. Considering the boundary crossing probability $P$ defined by

$$P(a) = \mathrm{P}\left\{ \sup_{t\in[0,1]} y(t) \ge a \right\}, \tag{19}$$

Durbin (1985) showed that $P(a)$ can be characterized by the integral of the boundary crossing density $p(t,a)$ of the first passage time $\tau_a$, which can be decomposed into two functions:

$$P(a) = \int_0^1 p(t,a)\mathrm{d}t = \int_0^1 b(t,a) f(t,a)\mathrm{d}t \tag{20}$$

where

$$b(t,a) = \lim_{s \to t} \frac{\mathrm{E}\left[\mathrm{I}(s < \tau_a)\left(a - y(s)\right) | y(t) = a\right]}{t - s} \tag{21}$$

and

$$f(t,a) = \frac{1}{\sqrt{2\pi\rho(t,t)}} \exp\left\{\frac{-a^2}{2\rho(t,t)}\right\}. \tag{22}$$

However, $b$ is almost always intractable; this complication motivated Durbin to propose three approximate boundary crossing probabilities.

### 3.2 The first approximation $P_1$

Durbin's first approximation, achieved simply through the removal of the indicator function from (21), 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 the mild conditions outlined above "...behaves locally like Brownian motion and the boundary is locally linear[4]" (Durbin, 1985, p. 110-111). That is, approximation $P_1$ starts with the following approximation to the function $b$:

$$b_1(t,a) = \frac{\rho_{10}(t,t)}{\rho(t,t)} a \tag{23}$$

using the convention here and below that $\rho_{ij}(s,t) := \frac{\partial^{i+j}\rho(s,t)}{\partial s^i \partial t^j}$. This approximation to $b$ owes its simple form to a hypothetical regression argument[5]. Approximations to the first passage density for $y$ and the boundary crossing probability are respectively

$$p_1(t,a) = b_1(t,a) f(t,a) \tag{24}$$

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

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

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

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.

and

$$P_1(a) = \int_0^1 p_1(t,a)\mathrm{d}t. \tag{25}$$

Given $\rho$ and $\rho_{10}$, $P_1(a)$ is easy to compute for simple parametric models. Since the difference between $b$ and $b_1$ becomes smaller as $a \to \infty$, it is clear that $P_1$ is an accurate approximation of $P$ for relevant testing situations because large values of $a$ correspond to low values of $\alpha$.

### 3.3 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. This estimate is remarkably accurate for situations of practical interest. Interestingly, research on the extrema of Gaussian processes and fields can be used to show that this estimate is asymptotically exact as the boundary $a \to \infty$. The results are based on the theory of large deviations for Gaussian processes which can be found in the monograph of Piterbarg (1996).

Let the variance function of a Gaussian process $y$ be defined as $\sigma^2(t) := \rho(t,t)$ and the point of maximal variance $t_0 := \mathrm{argmax}_t\, \sigma^2(t)$. Durbin's global approximation $P_g$ is

$$P_g(a) = \frac{\rho_{10}(t_0,t_0)}{\sigma^2(t_0)} \left( \frac{-2\sigma^2(t_0)}{\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)} \right)^{1/2} \exp\left\{ \frac{-a^2}{2\sigma^2(t_0)} \right\}. \tag{26}$$

This is achieved by starting with equation (24), evaluating all the non-exponential parts at $t_0$, and replacing the exponential part with an expansion to evaluate it. This formula is easy to use for the purposes of calculating approximate critical values or p-values, and can be used without the step of numerically integrating a boundary crossing density.

Some important features of Durbin's $P_g$ when applied to $\hat{v}$ are contained in the following theorem. This form of $P_g$ may sometimes be easier to compute than (26).

**Theorem 1.** *Suppose that* $\frac{\partial^2}{\partial x \partial \theta} f(x,\theta)$ *is bounded for all* $(x,\theta)$. *Then the approximation* $P_g$ *to the probability* $\mathrm{P}\left\{\sup_t \hat{v}(t) > a\right\}$ *is*

$$P_g(a) = \frac{\exp\left\{\frac{-a^2}{2\sigma^2(t_0)}\right\}}{2\sqrt{-\sigma^2(t_0)\left(\rho_{20}(t_0,t_0)+\rho_{11}(t_0,t_0)\right)}}. \tag{27}$$

A drawback to the use of $P_g$ is that if $\rho_{20}(t_0,t_0)=\rho_{11}(t_0,t_0)=0$ (which occurs, e.g., when testing $\mathcal{N}(\mu,\sigma^2)$ with $\mu$ unspecified,) $P_g$ does not exist[6]. Furthermore, it is not very clear that $P_g$ becomes more accurate as the boundary diverges. Both of these issues are addressed formally in the following theorem. It is due originally to Fatalov (1992, 1993) and is part of the literature on large deviations for Gaussian processes and fields. 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 — Durbin's original approximation was made without theoretical guarantee of its accuracy, only empirical evidence that it worked well.

**Theorem 2.** *Suppose $\theta$ is estimated by maximum likelihood and $\sigma^2$, the variance function of $\hat{v}$, has a derivative of some order $2k$ ($k\in 1,2,\ldots$) that is nonzero at $t_0=\operatorname{argmax}_{t\in[0,1]}\sigma^2(t)$. Then*

$$\mathrm{P}\left\{\sup_{t\in[0,1]}\hat{v}(t)>a\right\}=H(\sigma,k)\left(\frac{a}{\sigma(t_0)}\right)^{1-1/k}\phi\left(\frac{a}{\sigma(t_0)}\right)(1+o(1)),\quad a\to\infty \tag{28}$$

*where $\phi$ is the standard normal density function,*

$$H(\sigma,k)=\frac{C}{kA}\Gamma\left(\frac{1}{2k}\right), \tag{29}$$

*$\Gamma(\cdot)$ is the standard gamma function and*

$$A=\left(\frac{-\frac{\mathrm{d}^{(2k)}}{\mathrm{d}t^{(2k)}}\sigma^2(t_0)}{2(2k)!\sigma^2(t_0)}\right)^{1/(2k)},\qquad C=\frac{1}{2\sigma^2(t_0)}. \tag{30}$$

Note that setting $k=1$ is equivalent to the existence of $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)$ and (28) is identical to (26). This is because if $k=1$,

$$\mathrm{P}\left\{\sup_{t\in[0,1]}\hat{v}(t)>a\right\}\approx H(\sigma,1)\phi\left(\frac{a}{\sigma^2(t_0)}\right) \tag{31}$$

$$=\frac{1}{2\sigma^2(t_0)}\sqrt{\frac{4\sigma^2(t_0)}{-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)}}\sqrt{\pi}\frac{\exp\left\{\frac{a}{\sigma^2(t_0)}\right\}}{\sqrt{2\pi}}, \tag{32}$$

[6]Some more explicit calculations of $P_g$ for the normal and exponential distributions are presented in Appendix A.

and because it can be shown that $\rho_{10}(t_0, t_0) = 1/2$ (see the proof of Theorem 1),

$$= \frac{\rho_{10}(t_0, t_0)}{\sigma^2(t_0)} \left( \frac{-2\sigma^2(t_0)}{\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)} \right)^{1/2} \exp\left\{ \frac{-a^2}{2\sigma^2(t_0)} \right\} = P_g(a). \tag{33}$$

Theorem 2 indicates some features that make Durbin's $P_g$ a good approximation. First, Durbin conjectured 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 negligible[7]. 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.

### 3.4 The Gauss-Markov approximation $P_2$

The limiting process $\hat{v}$ is generally a non-Markovian, nonstationary Gaussian process. Because this limit is non-Markovian, its increments may be related in complicated ways. Durbin's final suggestion was essentially to calculate boundary crossing probabilities as if this inconvenience were negligible. This final approximation improves upon $P_1$ and is the solution to a numerically evaluated integral equation. A great deal of mathematical tractability is gained through this simplification, and the examples below suggest that the results are quite accurate.

Let $y$ be a mean-zero Gauss-Markov process (that is, a Gaussian process that also satisfies the Markov property[8]) with covariance function $\rho$. Define[9]

$$\begin{bmatrix} \beta_1(s,t) \\ \beta_2(s,t) \end{bmatrix} = \begin{bmatrix} \rho(s,s) & \rho(s,t) \\ \rho(t,s) & \rho(t,t) \end{bmatrix}^{-1} \begin{bmatrix} \rho_{01}(s,t) \\ \rho_{10}(t,t) \end{bmatrix}. \tag{34}$$

Durbin (1985) showed that the exact density $p_2(t,a)$ of the first passage time for Gauss-Markov process

[7] Note 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.

[8] 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 $s \leq t$.

[9] 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 $y(s)$, $s, t \leq r$.

$y$ is the solution to the integral equation

$$p_2(t,a) = p_1(t,a) - a\int_0^t \left[\beta_1(s,t) + \beta_2(s,t)\right] f(t|s,a) p_2(s,a)\mathrm{d}s. \tag{35}$$

Because (35) is a Volterra equation of the second kind, the solution $p_2$ is unique. In (35), $p_1(t,a)$ is as in (24) 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 $s \leq t$, in the case of a constant boundary, the transition distribution is

$$F(t|s,a) = F(y(t)|y(s) = a) = \mathcal{N}\left(\frac{\rho(s,t)}{\rho(s,s)}a, \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)}\right) \tag{36}$$

and the density is evaluated at $a$. Then the probability $\mathrm{P}\left\{\sup_t y(t) > a\right\}$ is given by

$$P_2(a) = \int_0^1 p_2(t,a)\mathrm{d}t \tag{37}$$

Durbin (1985) showed that equation (35) holds exactly for Gauss-Markov processes, and he suggested to use this relation as an approximation method for non-Markovian processes as well. That is, the Gauss-Markov approximation to $\mathrm{P}\left\{\sup_t \hat{v}(t) > a\right\}$ is given by (37) where the covariance function of $\hat{v}$ is used to calculate (35) despite the fact that $\hat{v}$ is not Markovian. This disregards the intractable autocovariance structure of $\hat{v}$ but also delivers reasonable results, as will be seen in Section 6.

### 3.4.1 Gauss-Markov processes

A mean-zero Gauss-Markov process with covariance function $\rho$ has transition probabilities that can be characterized as

$$(x,t)|(y,s) \sim \mathcal{N}\left(\frac{\rho(s,t)}{\rho(s,s)}y, \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)}\right). \tag{38}$$

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 $\rho$ satisfies, for all $0 \leq r \leq s \leq t$

$$\rho(r,t) = \frac{\rho(r,s)\rho(s,t)}{\rho(s,s)}. \tag{39}$$

Because of this, there must exist (differentiable) functions $\eta$ and $\zeta$ 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 $\eta/\zeta$ is strictly increasing and we have the representation

$$y(t) = \zeta(t)W\left((\eta/\zeta)(t)\right). \tag{40}$$

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

$$\frac{\partial}{\partial t}f(x,t|y,s) = -\frac{\partial}{\partial x}\left\{A_1(x,t)f(x,t|y,s)\right\} + \frac{A_2(t)}{2}\frac{\partial^2}{\partial x^2}f(x,t|y,s) \tag{41}$$

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

$$A_1(x,t) = \lim_{s\to t}\frac{\partial}{\partial t}\frac{\rho(s,t)}{\rho(s,s)}y = \frac{\rho_{01}(t,t)}{\rho(t,t)}y \tag{42}$$

and

$$A_2(t) = \lim_{s\to t}\frac{\partial}{\partial t}\rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)} = \rho_{10}(t,t) - \rho_{01}(t,t) \tag{43}$$

The function $A_2$ in particular is intimately connected to Durbin's approximation— see equation (36) above and equation (4) of Durbin (1985). The function $A_1$ is also strikingly similar to equation (23) above, especially given the fact that for the parametric empirical process, $\rho_{10}(t,t) - \rho_{01}(t,t) = 1$ for all $t$ (see the proof of Theorem 1).

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)[10]:

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

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

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

*Then the density $p$ of $\tau_a$ satisfies the following integral equation:*

$$\Psi\left(\frac{a}{\sqrt{\rho(t,t)}}\right) = \int_0^t \Psi\left(\frac{a - m(s,t)}{\sqrt{V(s,t)}}\right) p(s,a)\mathrm{d}s \tag{44}$$

*where*

$$m(s,t) = \frac{\rho(s,t)}{\rho(s,s)}a \quad \text{and} \quad V(s,t) = \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)} \tag{45}$$

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

The connection between the integral equations (44) and (35) is not as straightforward as it might seem. Differentiating equation (44) with respect to $t$ results in another integral equation that is remarkably similar to equation (35). Despite the similarities, in general only a circuitous connection can be made[11]— 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 (44) is slightly simpler to put into practice, Durbin's equation (35) was more stable in numerical experiments.

### 3.4.2 Computation of $p_2$

Equation (35) is a nonseparable Volterra integral equation of the second kind and thus must be solved numerically, but 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 $\rho$ 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 $s \le t$ — 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

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

value of $m$ is chosen by the researcher,) the integration algorithm simplifies to the following recursive rule: for $i = 0, 1$ (recall $t_0 = 0$),

$$p_2(0, a) = 0, \qquad p_2(t_2, a) = p_1(t_2, a) \tag{46}$$

and for $i \geq 3$

$$p_2(t_i, a) = p_1(t_i, a) + a\frac{1}{m}\sum_{j=2}^{i-1} K(t_j, t_i)p_2(t_j, a) \tag{47}$$

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.

### 3.5 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 least accurate of the three. In the simulation experiments examined in Section 6, performance is quite competitive with other methods.

Furthermore, these methods are generalizable. 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. Approximation $P_2$ is also quite 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 — see van der Vaart and Wellner (2007). 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 asymptotically linear, the covariance function needs to be derived on a case-by-case basis. The method presented in the next Section may be very useful in such situations.

These approximations are attractive because the adjusted critical values are tied to the parametric family being tested through computable features of the model. They require only that the researcher can derive a few functions related to the model (as required in (16) or (17)) and plug the covariance function and its derivatives into a relatively simple formula. In contrast, as will be seen below, Khmaladze's martingale transform can at times seem relatively complicated. In addition, as will be seen in Section 6, 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 (1981). He proposed a transformation of the empirical process that is not affected asymptotically by the estimation of model parameters, thereby avoiding the problem that statistics are not pivotal, a problem 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 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 $\sigma$-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 $\mathscr{F}^{\mathbb{F}_n}$. 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

$$\tilde{V}_n(x) = \sqrt{n}\left(\mathbb{F}_n(x) - K(x, \mathbb{F}_n, \hat{\theta}_n)\right) \tag{48}$$

is called the compensated 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 (48) asymptotically distribution-free.

The function $g$ defined in equation (8) is intimately related to the score function of the parametric model. The reason for this is that it can be shown that $\dot{g}$, the derivative of $g$ with respect to $t$, satisfies the equation

$$\dot{g}(t) = \frac{\partial}{\partial t} g(t,\theta) = \frac{\partial}{\partial \theta} \log f(x,\theta)\Big|_{x=F^{-1}(t,\theta)} \tag{49}$$

implying that $g$ is in effect the integrated score function for the model. 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 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 $\Gamma$ by

$$h(t,\theta) = \begin{bmatrix} 1 \\ \frac{\partial g(t,\theta)}{\partial t} \end{bmatrix} \qquad \text{and} \qquad \Gamma(t,\theta) = \int_t^1 h(s,\theta)h(s,\theta)^\top \mathrm{d}s. \tag{50}$$

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

$$K(t,\mathbb{F}_n,\theta) = \int_0^t h(s,\theta)^\top \Gamma^{-1}(s,\theta) \int_s^1 h(r,\theta)\mathrm{d}\mathbb{F}_n(r)\mathrm{d}s. \tag{51}$$

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

$$= \int_0^1 \int_0^{t\wedge r} h(s,\theta)^\top \Gamma^{-1}(s,\theta)\mathrm{d}s\, h(r,\theta)\mathrm{d}\mathbb{F}_n(r). \tag{52}$$

One may think of equation (51) 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 $\Gamma$ and the Fisher information, it can be shown that the compensator also has a form that does not always depend on parameter values when the null model is a member of special classes of parametric models (location-scale models, for example); 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.

### 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 $t^{\text{th}}$ quantile of the null model.

Suppose we once again use an evenly spaced partition (with $m$ points) of $[0,1]$. The score and empirical distribution functions are evaluated at each point in the partition and then stacked into the following sequence of matrices of size $(m-i+2)\times 2$ and $(m-i+2)\times 1$ respectively:

$$X_i = \begin{bmatrix} \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_{m+1}) \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_m) \\ \vdots & \vdots \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_i) \end{bmatrix} \qquad y_i = \begin{bmatrix} \sqrt{m}\left(\mathbb{F}_n(t_{m+1}) - \mathbb{F}_n(t_m)\right) \\ \sqrt{m}\left(\mathbb{F}_n(t_m) - \mathbb{F}_n(t_{m-1})\right) \\ \vdots \\ \sqrt{m}\left(\mathbb{F}_n(t_i) - \mathbb{F}_n(t_{i-1})\right) \end{bmatrix} \tag{53}$$

Then, least squares coefficients for each $t_i$ are calculated:

$$\begin{aligned}\hat{\beta}(t_i) &= (X_i^\top X_i)^{-1} X_i^\top y_i \\ &= \begin{bmatrix} \frac{1}{m}(m-j+2) & \frac{1}{m}\sum_{j=i}^{m+1} \dot{g}(t_j) \\ \frac{1}{m}\sum_{j=i}^{m+1} \dot{g}(t_j) & \frac{1}{m}\sum_{j=i}^{m+1} \dot{g}^2(t_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=i}^{m+1} [\mathbb{F}_n(t_j) - \mathbb{F}_n(t_{j-1})] \\ \sum_{j=i}^{m+1} \dot{g}(t_j)[\mathbb{F}_n(t_j) - \mathbb{F}_n(t_{j-1})] \end{bmatrix}. \end{aligned} \tag{54}$$

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:

$$\hat{K}(t_i) = \frac{1}{m}\sum_{j=1}^{i} h^\top(t_j)\hat{\beta}(t_j). \tag{55}$$

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.

### 4.2 Comparison with Wooldridge (1990)

Wooldridge (1990), extending the work of Davidson and MacKinnon (1985) in the context of robustifying regression specification tests, proposed an orthogonal projection that achieves the same goal as the martingale transform — it accounts for the effect of estimation and leaves statistics asymptotically distribution-free. Khmaladze's martingale transform bears a good deal of similarity to Wooldridge's proposal. However, these proposals are fundamentally different with regard to the transformation that is made to the data. Here we adapt Wooldridge's test statistics to the one-sample case to facilitate comparison with Khmaladze's transformation. As the analysis below shows, Wooldridge's idea applies to finite-dimensional features of a density, while Khmaladze's applies to the shape of the entire density function.

Suppose for a given $x \in \mathcal{X}$, we have a hypothesized vector of conditional moment restrictions

$\phi \in \mathbb{R}^L$ satisfying

$$\mathrm{E}\left[\phi(x, X_i, \theta)\right] = 0, \quad \forall i,\ \forall x \in \mathcal{X},\ \text{for some } \theta_0 \in \Theta \tag{56}$$

and let $\{\Lambda(x, X_i, \theta)\}_{i=1}^n$ be some "misspecification indicators" used to robustify the test statistic against model misspecification. Many test statistics can be defined by

$$\hat{T}_n(x, \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda(x, X_i, \hat{\theta}) \phi(x, X_i, \hat{\theta}) \tag{57}$$

or as some functional of $\hat{T}_n$. Define

$$\Phi(x, X_i, \theta) = \mathrm{E}\left[\nabla_\theta \phi(x, X_i, \theta)\right], \quad i = 1, 2, \ldots, n \tag{58}$$

Wooldridge (1990) noted that by using a mean-value expansion, under some regularity conditions,

$$\hat{T}_n(x, \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda(x, X_i, \hat{\theta}) \phi(x, X_i, \hat{\theta}) \tag{59}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda(x, X_i, \theta_0) \phi(x, X_i, \theta_0) + \sqrt{n} \left(\hat{\theta} - \theta_0\right)^\top \frac{1}{n} \sum_{i=1}^{n} \Lambda(x, X_i, \theta_0) \nabla_\theta \phi(x, X_i, \theta_0) + o_P(1) \tag{60}$$

uniformly in $x$. This statistic is similar to the parametric empirical process in that its distribution is affected by $\hat{\theta}$ and the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$. Wooldridge showed that by using an orthogonal projection of $\Lambda$ on $\Phi$, it is possible to define statistics that do not depend on these unknowns:

$$\tilde{T}_n(x, \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Lambda(x, X_i, \hat{\theta}) - \Phi^\top(x, X_i, \hat{\theta}) \hat{\beta}(\hat{\theta}))^\top \phi(x, X_i, \hat{\theta}) \tag{61}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Lambda(x, X_i, \theta_0) - \Phi^\top(x, X_i, \theta_0) \hat{\beta}(\theta_0))^\top \phi(x, X_i, \theta_0) + o_P(1), \tag{62}$$

where

$$\hat{\beta}(\theta) = \hat{\beta}(x, X, \theta) = \left( \sum_{i=1}^{n} \Phi(x, X_i, \theta) \Phi^\top(x, X_i, \theta) \right)^{-1} \sum_{i=1}^{n} \Phi(x, X_i, \theta) \Lambda(x, X_i, \theta). \tag{63}$$

Therefore it is possible to use $\tilde{T}_n$ with an estimator that satisfies $\sqrt{n}(\hat{\theta} - \theta_0) = O_P(1)$, and Wooldridge shows that a quadratic form using $\tilde{T}_n$ is equivalent to a Lagrange multiplier test that converges in distribution to a $\chi^2$ distribution with degrees of freedom equal to the dimension of $\Lambda_i$. That is, Wooldridge's

modified test statistic uses an orthogonal projection to remove the effect of parameter estimation.

The difference between this approach and Khmaladze's can be illustrated by adapting it to test the hypothesis that the data is described by the distribution function $F \in \mathcal{F}$. Start by transforming the problem to the unit interval, and define

$$\phi(t, X_i, \theta) = I(F(X_i, \theta) \leq t) - t \tag{64}$$

which has zero expectation for all $i$ under the null hypothesis. For each observation $\Phi$ is

$$\Phi(x, \theta) = \nabla_\theta F(x, \theta)|_{x=F^{-1}(t,\theta_0)} = g(t, \theta). \tag{65}$$

Notably, $\Phi$ does not depend on the observed data. Letting $\lambda(t, X_i, \theta) = \Lambda(F^{-1}(t, \theta_0), X_i, \theta)$ results in

$$\hat{T}_n(t, \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \lambda(t, X_i, \hat{\theta}) \left( I(F(X_i, \hat{\theta}) \leq t) - t \right) \tag{66}$$

a weighted parametric empirical process evaluated at $t$ (cf. Koul (2002) for conditions under which this would also converge weakly to a limiting function that is continuous in $t$). By letting $\lambda \equiv 1$ one obtains $\hat{T}_n(t, \hat{\theta}) = \hat{v}_n(t)$.

Consider now to the specialization of (61) to this setting. If $g(t, \theta) g^\top(t, \theta)$ were invertible, we could rewrite $\tilde{T}_n$ defined in (61) as

$$\tilde{T}_n(t, \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \lambda(t, X_i, \hat{\theta}) - g^\top(t, \hat{\theta}) \left( g(t, \hat{\theta}) g^\top(t, \hat{\theta}) \right)^{-1} g(t, \hat{\theta}) \frac{1}{n} \sum_{i=1}^{n} \lambda(t, X_i, \hat{\theta}) \right) \left( I(X_i \leq t) - F(t, \hat{\theta}) \right). \tag{67}$$

Lemma 1 shows that it is indeed possible to define $\tilde{T}_n$ in this way, although the value of $g(t, \hat{\theta})$ is irrelevant. A more precise characterization is given in Lemma 1.

**Lemma 1.** *$\tilde{T}_n(t, \hat{\theta})$ is well-defined when $\phi$ and $\Phi$ are defined as in (64) and (65), and*

$$\tilde{T}_n(t, \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \lambda(t, X_i, \hat{\theta}) - \overline{\lambda}(t, \hat{\theta}) \right) \left( I(F(X_i, \hat{\theta}) \leq t) - t \right) \tag{68}$$

*where* $\overline{\lambda}(t,\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\lambda(t,X_i,\hat{\theta})$. *In particular,*

$$g^{\top}(t,\hat{\theta})\hat{\beta}(\hat{\theta}) = \overline{\lambda}(t,\hat{\theta}), \tag{69}$$

*where* $\hat{\beta}$ *is defined in (63).*

This proposal is different from that of Khmaladze (1981) because here $\hat{\beta}(\hat{\theta})$ is an orthogonal projection onto the space spanned by $g(t,\hat{\theta})$ — that is, the function evaluated at $t$. Khmaladze's compensator is a projection into the space spanned by the *function* $g$. Statistics derived by using Wooldridge's projection in this way are uninformative because the null hypothesis is functional in nature and the value of the function evaluated at a single point is not informative. The statistic $\tilde{T}_n(t,\hat{\theta})$ may indeed have some interesting properties, but it is beyond the scope of this paper to extend this test statistic to a process in $t$ and to examine conditions under which $\sup_t \tilde{T}_n(t,\hat{\theta})$ has a tractable limiting distribution (cf. Koul (2002, Section 6.6.2) for some results that might be applied here); Lemma 1 reveals that, as defined here, $g$ plays no part in $\tilde{T}_n$ and $\tilde{T}_n$ is not generally a distribution-free test statistic.

## 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[12], while for tests of normality there are two parameters and therefore a greater variety of boundary crossing probabilities to compute. 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).

[12]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.

## 5.1 The exponential distribution

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

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

The function $g$ for the exponential model is

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

A maximum likelihood estimate $\hat{\lambda}_n = \bar{x}^{-1}$ exists, and therefore $\hat{v}$ for a hypothesis of exponentiality 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 Appendix B.) For computation of $P_g$ the point of maximal variance must be solved numerically as the solution to

$$1 - 2t_0 + 2(1-t_0)\log(1-t_0)\left(1+\log(1-t_0)\right) = 0. \tag{73}$$

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 (17) 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).

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.

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 | 0.88054 | 0.87726 | 1.07298 |
| 5% | 1.00063 | 0.99105 | 0.98983 | 1.22387 |
| 2.5% | 1.09766 | 1.09041 | 1.09013 | 1.35810 |
| 1% | 1.21464 | 1.20930 | 1.20955 | 1.51743 |

### 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

$$h(t,\lambda) = \begin{bmatrix} 1 \\ \frac{1}{\lambda}(1+\log(1-t)) \end{bmatrix} \tag{74}$$

and

$$\Gamma(t,\lambda) = \begin{bmatrix} 1-t & \frac{1}{\lambda}(1-t)\log(1-t) \\ \frac{1}{\lambda}(1-t)\log(1-t) & \frac{1}{\lambda^2}(1-t)(1+\log^2(1-t)) \end{bmatrix}. \tag{75}$$

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

$$K(t,\mathbb{F}_n,\hat{\lambda}) = \int_0^t \frac{1}{2}\log^2(1-\hat{\varepsilon}) - 2\log(1-\hat{\varepsilon}) - \log^2(1-\hat{\varepsilon})\mathrm{d}\mathbb{F}_n(\hat{\varepsilon}) \\ + \int_t^1 \frac{1}{2}\log^2(1-t) - 2\log(1-t) - \log(1-\hat{\varepsilon})\log(1-t)\mathrm{d}\mathbb{F}_n(\hat{\varepsilon}), \tag{76}$$

or alternatively

$$K(t,\mathbb{F}_n,\hat{\lambda}) = \frac{1}{n}\sum_{i:\hat{\varepsilon}_i\le t}\left(\frac{-1}{2}\log^2(1-\hat{\varepsilon}_i) - 2\log(1-\hat{\varepsilon}_i)\right) \\ + \left(\frac{1}{2}\log^2(1-t) - 2\log(1-t)\right)\left(1-\mathbb{F}_n(t)\right) - \frac{1}{n}\log(1-t)\sum_{i:\hat{\varepsilon}_i>t}\log(1-\hat{\varepsilon}_i), \tag{77}$$

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

$$\begin{aligned}\tilde{K}(x,\mathbb{F}_n,\hat{\lambda}) = &\frac{\hat{\lambda}}{n}\sum_{i:X_i\le x}\left(2X_i - \frac{\hat{\lambda}}{2}X_i^2\right)\\ &+\left(2\hat{\lambda}x + \frac{\hat{\lambda}^2}{2}x^2\right)(1-\mathbb{F}_n(x)) - \frac{\hat{\lambda}^2}{n}x\sum_{i:X_i>x}X_i\end{aligned} \tag{78}$$

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

$$H_0:\ F(x,\theta) = \int_{-\infty}^{x}\frac{e^{\frac{-1}{2\sigma^2}(y-\mu)^2}}{\sqrt{2\pi\sigma^2}}\mathrm{d}y = \int_{-\infty}^{\frac{x-\mu}{\sigma}}\phi(z)\mathrm{d}z, \quad x\in\mathbb{R}, \tag{79}$$

where $\theta = (\mu,\sigma)\in\mathbb{R}\times(0,\infty)$ and $\phi(z) = \frac{\exp\{-z^2/2\}}{\sqrt{2\pi}}$. Maximum likelihood estimators exist for the parameters of the model, so the covariance function generally takes the form of (17).

Letting $\Phi$ be the distribution function of the standard normal distribution, the location-scale invariance of the normal model implies that $F^{-1}(s,\theta) = \mu + \sigma\Phi^{-1}(s)$, and the function $g$ for the location- and scale-unknown case is equal to

$$g(s,\theta) = \left[\begin{array}{c}\frac{\partial}{\partial\mu}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\phi(z)\mathrm{d}z\\ \frac{\partial}{\partial\sigma}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\phi(z)\mathrm{d}z\end{array}\right]_{x=\mu+\sigma\Phi^{-1}(s)} = \frac{-1}{\sigma}\begin{bmatrix}\phi(\Phi^{-1}(s))\\ \Phi^{-1}(s)\phi(\Phi^{-1}(s))\end{bmatrix}. \tag{80}$$

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.) Using (17), one finds that $\hat{v}$ has covariance function

$$\rho_{\mu\sigma}(s,t) = s\wedge t - st - \phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))\left(1+\frac{1}{2}\Phi^{-1}(s)\Phi^{-1}(t)\right). \tag{81}$$

The function $\rho_{\mu\sigma}(t,t)$ is maximized at $t_0 = \frac{1}{2}$, and 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. For the location-unspecified case, 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 location-unknown case we have

$$\rho_\mu(s,t) = s \wedge t - st - \phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t)) \tag{82}$$

The function $\rho_\mu(t,t)$ is maximized at $t_0 = \frac{1}{2}$; however, $P_g$ does not exist in this case, because the second derivative of $\rho(t,t)$ evaluated at $t_0$ is equal to zero. We can, however, use Theorem 2 to find that $P_g = \frac{\Gamma(1/4)}{\pi-2}\sqrt[4]{\frac{3\pi}{2}}\sqrt{a}\exp\{-2\pi a^2/(\pi-2)\}$ (cf. Appendix A).

Similarly, the covariance function in the scale-unspecified case is

$$\rho_\sigma(s,t) = s \wedge t - st - \frac{1}{2}\Phi^{-1}(s)\Phi^{-1}(t)\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t)), \tag{83}$$

$\rho_\sigma(t,t)$ is maximized at $t_0 = \frac{1}{2}$ and $P_g(a) = (2/3)^{1/2}\exp\{-2a^2\}$. Note that there is a small typographical 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 fact that the normal distribution is a location-scale class, the critical values tabulated in Table 2 are invariant to the true values of the parameters $\mu$ and $\sigma$.

### 5.3 Regression residual processes

Suppose that the distribution of $y_i \in \mathbb{R}$ conditional on $X_i \in \mathbb{R}^p$ may be specified as

$$y_i = X_i^\top \beta + \sigma \varepsilon_i, \quad \varepsilon_i \sim F_0, \quad i = 1, 2, \ldots n, \tag{84}$$

where $\varepsilon_i$ are iid, mean-zero and independent of $\{X_i\}$. The linear form of the conditional mean can be relaxed; see Khmaladze and Koul (2009). The null hypothesis is that the distribution function of $y_i$ conditional on $X_i$ is a member of a parametric model — for example, the normal model. This is equivalent to the hypothesis that $\varepsilon_i$ are distributed according to a location-scale model, because the model implies that the distribution function for each error $\varepsilon_i$ satisfies

$$F(e, \theta | X_i) = F_0\left(\frac{e - X_i^\top \beta}{\sigma}\right), \qquad \theta = (\beta, \sigma) \in \mathbb{R}^p \times \mathbb{R}_+. \tag{85}$$

Define the parametric empirical process of regression residuals ($\hat{\varepsilon}_i = (y_i - X_i^\top \hat{\beta})/\hat{\sigma}$) by

$$\hat{v}_n(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( I(F_0(\hat{\varepsilon}_i) \le t) - t \right), \tag{86}$$

and let $v_n$ be the empirical process of the true errors — that is, analogously to $\hat{v}_n$ but with null value $\theta_0$. The function $g : \mathbb{R} \to \mathbb{R}^{p+1}$ must be defined conditional on $X_i$, and is analogous to (8): let

$$g(t|X_i) = \nabla_\theta F(F^{-1}(t|X_i, \theta)|X_i, \theta) = \frac{-1}{\sigma} \begin{bmatrix} X_i f_0(F_0^{-1}(t)) \\ F_0^{-1}(t) f_0(F_0^{-1}(t)) \end{bmatrix}. \tag{87}$$

Koul (2002, Theorem 6.4.1) implies that $\hat{v}_n$ satisfies the following asymptotic linearity characterization, analogous to (12):

$$\sup_{t \in [0,1]} \left| \hat{v}_n(t) - v_n(t) + \sqrt{n}(\hat{\theta} - \theta_0)^\top \frac{1}{n} \sum_{i=1}^{n} g(t, \theta_0) \right| = o_P(1). \tag{88}$$

Assuming an asymptotically efficient $\hat{\theta}$ exists, the distribution of the supremum norm test statistic can be written down more explicitly. Using (87), it can be verified that the covariance function of the limiting process $\hat{v}$ is

$$\mathrm{E}\left[\hat{v}(s)\hat{v}(t)|X\right] = s \wedge t - st - \frac{1}{\sigma^2} f_0(F_0^{-1}(s)) f_0(F_0^{-1}(t)) \left[\mathrm{E}\left[X^\top\right] F_0^{-1}(s)\right] I^{-1}(\theta_0) \begin{bmatrix} \mathrm{E}[X] \\ F_0^{-1}(t) \end{bmatrix}. \tag{89}$$

For the purpose of testing normality note that the information matrix for the regression model with normal errors is

$$I(\theta) = \frac{1}{\sigma^2} \begin{bmatrix} Q & \mathbf{0}_{p\times 1} \\ \mathbf{0}_{1\times p} & \frac{1}{2} \end{bmatrix} \tag{90}$$

where $Q = \operatorname{plim} \frac{1}{n} X^\top X$. Along with (87) specialized to the normal distribution, it can be verified that equation (89) becomes

$$\mathrm{E}\left[\hat{v}(s)\hat{v}(t)|X\right] = s \wedge t - st - \phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))\overline{P} - \frac{1}{2}\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))\Phi^{-1}(s)\Phi^{-1}(t), \tag{91}$$

where

$$\overline{P} = \operatorname{plim} \frac{1}{n} \mathbf{1}_n^\top X (X^\top X)^{-1} X^\top \mathbf{1}_n \tag{92}$$

summarizes the effect that the design matrix $X$ has on the limiting covariance function. When the design includes an intercept term, $\frac{1}{n}\mathbf{1}_n^\top P_X \mathbf{1}_n = \frac{1}{n}\mathbf{1}_n^\top \mathbf{1}_n = 1$ and the process of regression residuals has the same asymptotic distribution as the one-sample process, unaffected by the distribution of $X$.

# 6 Simulation experiments

## 6.1 The exponential distribution

Table 3 presents the results of a small simulation experiment using the $D^-$ statistic for testing the null hypothesis of exponentiality against one-sided 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** | | | | |
| $P_2$ | 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 |
| Kolmogorov-Smirnov | 2.70 | 0.81 | 0.23 | 0.05 |
| **100** | | | | |
| $P_2$ | 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 |
| Kolmogorov-Smirnov | 2.84 | 0.83 | 0.26 | 0.06 |
| **200** | | | | |
| $P_2$ | 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 |
| Kolmogorov-Smirnov | 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 compensated process, performs roughly as well as the Gauss-Markov approximation in most cases.

The power of these tests has been addressed in a few papers, notably Aki (1986), Haywood and Khmaladze (2008) and Koul and Sakhanenko (2005), with some results on power for the martingale transformation technique. Another experiment was conducted using smooth local alternatives to the null hypothesis of exponentiality. Alternatives were constructed in one of two ways. First, local alter-

native mixture densities were generated using the following formula:

$$f_{mix}(x,n) = \left(1 - \frac{c}{\sqrt{n}}\right) f_{exp}(x) + \frac{c}{\sqrt{n}} f_{alt}(x) \tag{93}$$

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 Stute et al. (1993). 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. Stute et al. (1993) show that a bootstrapped empirical process converges in distribution to the parametric empirical process, implying that the supremum statistic also converges in distribution to the distribution of the supremum of the parametric empirical process.

The results of the power experiment appear in Table 4. The first row simply repeats the size of the tests, and the remaining rows report the empirical power from 50,000 simulated samples for the local 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 lower than the unadjusted ones. The bootstrap technique and Durbin's approximations are strikingly similar to one another, which is to be expected because in this simple setting, the bootstrap is effectively a simulation of the distribution described by the approximations. It is also of interest to note that no one method has uniformly better performance than all the others. For example, tests based on the compensated process do extremely well against the uniform alternative. On the other hand, they do not seem to do quite as well as other tests under lognormal and gamma alternatives. Evidently these tests have differential performance against alternatives from different parts of the space of alternatives.

Table 4: Empirical size and power for the local alternatives described in the text. All tests are intended to have a size of 5%. 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 | 5.5 | 0.8 |
| 100 | 5.0 | 4.9 | 4.6 | 4.1 | 5.4 | 0.8 |
| 200 | 5.1 | 5.0 | 4.7 | 4.3 | 5.4 | 0.8 |
| **uniform mixture** | | | | | | |
| 50 | 83 | 83 | 99 | 99 | 84 | 49 |
| 100 | 71 | 71 | 98 | 97 | 71 | 32 |
| 200 | 57 | 57 | 97 | 96 | 58 | 18 |
| **lognormal mixture** | | | | | | |
| 50 | 40 | 40 | 34 | 31 | 42 | 16 |
| 100 | 40 | 40 | 33 | 32 | 41 | 16 |
| 200 | 40 | 40 | 33 | 32 | 41 | 16 |
| **gamma alternative** | | | | | | |
| 50 | 56 | 56 | 53 | 49 | 57 | 24 |
| 100 | 62 | 62 | 59 | 57 | 63 | 30 |
| 200 | 67 | 67 | 63 | 62 | 68 | 36 |
| **weibull alternative** | | | | | | |
| 50 | 51 | 51 | 55 | 51 | 53 | 21 |
| 100 | 55 | 55 | 59 | 57 | 56 | 25 |
| 200 | 59 | 58 | 63 | 61 | 59 | 28 |

## 6.2 The normal distribution

We illustrate the performance of tests for normality using the regression residual process. Consider the simple model

$$y_i = X_i^\top \beta + \sigma \varepsilon_i, \tag{94}$$

for all $i$, where $\{\varepsilon_i\}$ are independent of $\{X_i\}$, and suppose $H_0 : \varepsilon_i \sim \mathcal{N}(0,1)$ for all $i$. Two natural test statistics that arise as generalizations of the one-sample statistics used above are the supremum statistic $\hat{D}_n = \sup_{t\in[0,1]} \left|\hat{v}_n(t)\right|$, where $\hat{v}_n$ is defined in (88), and the statistic $\tilde{D}_n = \sup_t \left|\tilde{v}_n(t)\right|$ created by applying the martingale transformation (assuming normality) to the regression residual process. For the experiment shown in Table 5, a grid of $3n$ points on the unit interval was used, to increase the precision of the compensator.

The conditional Kolmogorov (CK) test proposed by Andrews (1997) can be specialized to this situation and used to test this hypothesis. That test statistic is defined as

$$CK_n = \max_{j\le n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( I(y_i \le y_j) - F(y_j|X_i, \hat{\theta}) \right) I(X_i \le X_j) \right|. \tag{95}$$

This test statistic is calculated using only the points in the sample because of the computational burden that would be imposed by maximization over the entire sample space. Andrews proposed the following bootstrap procedure for inference: fix the covariates and create $b$ bootstrap samples of size $n$ by resampling ($n$ times) from $y$ to create $y^*$ and constructing the sample $\{y_i^*, X_i\}_{i=1}^n$. Following this suggestion, we repeated this 299 times to find a bootstrap distribution of $CK_n$. Note that the regression residual process (86) is very similar to the process appearing in Andrews' CK statistic — because the model asserts that $y_i|X_i$ is a member of a location-scale model for all $i$, $F(y_i|X_i, \theta) = F_0((y_i - X_i^\top \beta)/\sigma)$, the only difference is the addition of the indicators $I(X_j \leq X_i)$ in the definition of the CK statistic.

In order to produce the results in Table 5, regression models were generated with correct specification of the conditional expectation, but with error terms following different distributions. Specifically, all models were linear models

$$y_i = 1 + x_i + .5\varepsilon_i, \tag{96}$$

where $x_i \sim \mathcal{N}(0,1)$ for all $i$ and $\varepsilon_i$ following another "local mixture" distribution with the following density:

$$f_{mix}(e) = \left(1 - \frac{c}{\sqrt{n}}\right)\phi(e) + \frac{c}{\sqrt{n}} t_\nu(e) \tag{97}$$

where $\phi$ is the standard normal density, $t_\nu$ is the density associated with the (Student's) $t$-distribution, and where $\nu$, the degrees of freedom associated with the $t$ distribution in the mixture, were chosen to be infinite or one of $\nu = 10, 4$, and 2. Naturally, the infinite value for $\nu$ is chosen to examine the performance of the tests under the null hypothesis. The value $c = 6$ was chosen so as to avoid trivial powers.

Table 5 shows the results of a simulation experiment comparing Andrews' conditional Kolmogorov test, a test utilizing Durbin's $P_g$ approximation, and the test based on a transformed process under the null hypothesis of normality and under the local $t$ deviations from normality. The size of Andrews' bootstrap-based test is quite close to the intended 5%, better than the other two methods, especially in the smallest sample size, although for similar sample sizes this test had a tendency to overreject the null hypothesis in Andrews' Table 1 (p. 1111). It should be noted however, that the conditional model discussed in the simulation results of Andrews (1997) is different and more complex than the model considered here — in this paper only univariate response variables have been examined. Andrews' method can easily be extended to conditional models with a multivariate response variable (as in

Table 5: Empirical size and power for locally $t$-distributed alternatives described in the text. All tests are intended to have a size of 5%. 25,000 repetitions.

| sample size | Andrews' CK | $P_g$ | RLS transform |
|---|---|---|---|
| **null model** | | | |
| 50 | 4.4 | 3.7 | 1.8 |
| 100 | 4.5 | 4.1 | 3.1 |
| 200 | 4.8 | 4.6 | 3.6 |
| $t_{10}$ **mixture** | | | |
| 50 | 8.0 | 6.7 | 4.3 |
| 100 | 9.6 | 7.0 | 8.1 |
| 200 | 11.1 | 7.0 | 10.7 |
| $t_4$ **mixture** | | | |
| 50 | 13.8 | 23.7 | 13.4 |
| 100 | 19.4 | 30.3 | 26.4 |
| 200 | 25.6 | 36.5 | 38.4 |
| $t_2$ **mixture** | | | |
| 50 | 16.3 | 68.8 | 42.1 |
| 100 | 24.3 | 85.0 | 68.0 |
| 200 | 34.6 | 93.1 | 85.9 |

the example of a trivariate logit model) because inference is carried out using a parametric bootstrap technique. It is more difficult to apply the analytic techniques considered here to such settings.

For small departures from normality (represented by the $t_{10}$ mixture density), the bootstrap procedure also appears to have somewhat better power than the other two methods. However, against the heavier-tailed alternatives — those mixtures using $t_4$ and $t_2$ distributions — inference using adjusted critical values appears to be the most powerful. Inference based on the transformed empirical process appears to be as powerful as that using Durbin-style adjustments, but this only becomes apparent when the sample size is large. Tests using the transformed process were also rather under-sized, especially for small samples. The average time used to conduct each test was different between the methods and much higher for Andrews' test — for example, for samples of size 200, the average time to compute Andrews' statistic was 6.7 seconds, while the compensated empirical process statistic took .03 seconds and the parametric empirical process statistic took .006 seconds on average. These times are trivial for researchers who only wish to conduct the test once for an analysis; on the other hand, the relative simplicity of the Durbin-style adjustments may sometimes be an argument in their favor. The unmodified parametric empirical process can be used, and (assuming the hypothesized model is simple enough) the covariance function can be quickly calculated and used in a formula like that in Theorem 1 to conduct inference.

# 7 Conclusion

Durbin (1985) proposed several very accurate approximations to the boundary crossing probability for a class of Gaussian processes, of which the standard parametric empirical process is a leading example. In this paper I show that it is simple to conduct sup-norm inference for empirical processes based on Durbin's approximations, and that its performance in finite samples is competitive with two other empirical-process-based inferential methods — the martingale transformation proposed by Khmaladze (1981) and parametric bootstrap techniques. The score function of the null parametric model is the common thread that connects Khmaladze's transformation to Durbin's approximations. Evidence from simulation experiments suggests that Durbin's approximations result in tests that have a size comparable to tests based on the compensated empirical process. Simulation suggests that Durbin-style adjustments may offer a power advantage over the other inferential methods.

# A $P_g$ and large deviation approximations

In order to clarify equation (26), Durbin's global approximation, some further details are presented for the specific cases mentioned in the examples. For the exponential distribution, $t_0$ must satisfy the following equation:

$$1 - 2t_0 + 2(1 - t_0)\left(\log(1 - t_0) + \log^2(1 - t_0)\right) = 0. \tag{98}$$

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 be done numerically because of the lack of a convenient value of $t_0$. However, it is possible to calculate $P_g$ analytically for the two normal cases mentioned above. Note that for all normal distribution cases, $t_0 = 0$.

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

$$\frac{\mathrm{d}^2\rho_{\mu\sigma}(t,t)}{\mathrm{d}t^2} = -1 + \left(1 + \phi(\xi(t))\right)\xi^2(t) - \xi^4(t) \tag{99}$$

and

$$\frac{\mathrm{d}^2\rho_\sigma(t,t)}{\mathrm{d}t^2} = -3 + 4\xi^2(t) - \xi^4(t), \tag{100}$$

where $\phi$ is the standard normal density function and $\xi$ 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 (26), we have

$$P_g(a) = \frac{1/2}{\frac{1}{4} - \frac{1}{2\pi}} \sqrt{\frac{-2\left(\frac{1}{4} - \frac{1}{2\pi}\right)}{-1}} \exp\left\{\frac{-a^2}{2\left(\frac{1}{4} - \frac{1}{2\pi}\right)}\right\} = \sqrt{\frac{2\pi}{\pi - 2}} e^{\frac{-2\pi}{\pi-2}a^2} \tag{101}$$

for the model with both location and scale unspecified, and

$$P_g(a) = \frac{1/2}{1/4} \sqrt{\frac{-2/4}{-3}} \exp\left\{\frac{-a^2}{2/4}\right\} = \sqrt{2/3} e^{-2a^2} \tag{102}$$

for the scale-unspecified case.

## A.1 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

$$(\hat{\mu}, \hat{\sigma}): \qquad \sigma^2(t_0) = \frac{\pi - 2}{4\pi} \qquad A = \sqrt{\frac{\pi}{\pi - 2}} \qquad C = \frac{2\pi}{\pi - 2} \qquad k = 1 \tag{103}$$

$$(\mu, \hat{\sigma}): \qquad \sigma^2(t_0) = 1/4 \qquad A = \sqrt{3} \qquad C = 2 \qquad k = 1 \tag{104}$$

$$(\hat{\mu}, \sigma): \qquad \sigma^2(t_0) = \frac{\pi - 2}{4\pi} \qquad A = \sqrt[4]{\frac{2\pi^2}{3(\pi - 2)}} \qquad C = \frac{2\pi}{\pi - 2} \qquad k = 2 \tag{105}$$

Note the value of $A$ is different from what is printed in Piterbarg (1996) for two of three cases. Plugging these values into equation (28) results in

$$\mathrm{P}\left\{\sup_{t\in[0,1]} \hat{v}(t) > a \middle| \hat{\mu}, \hat{\sigma}\right\} = \sqrt{\frac{2\pi}{\pi-2}}\, e^{\frac{-2\pi}{\pi-2}a^2} \tag{106}$$

$$\mathrm{P}\left\{\sup_{t\in[0,1]} \hat{v}(t) > a \middle| \mu, \hat{\sigma}\right\} = \sqrt{2/3}\, e^{-2a^2} \tag{107}$$

$$\mathrm{P}\left\{\sup_{t\in[0,1]} \hat{v}(t) > a \middle| \hat{\mu}, \sigma\right\} = \frac{\Gamma(1/4)}{\pi-2}\sqrt[4]{\frac{3\pi}{2}}\sqrt{a}\, e^{\frac{-2\pi}{\pi-2}a^2} \tag{108}$$

# B Location-scale and scale-shape models

Two classes of commonly used parametric models are represented in the examples. 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 classes is the well-known class of location-scale models. Models in this class have distribution functions that take the form

$$F(x,\theta) = F_0\left(\frac{x-\theta_1}{\theta_2}\right); \qquad x \in \mathcal{X} \subseteq \mathbb{R}, \quad \theta \in \mathbb{R}\times(0,\infty) \tag{109}$$

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

$$F(x,\theta) = F_0\left(\left(\frac{x}{\theta_1}\right)^{\theta_2}\right); \qquad x \in \mathcal{X} \subseteq [0,\infty), \quad \theta \in (0,\infty)\times(0,\infty). \tag{110}$$

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).

We assume that efficient estimates exist for the parameters, so that the covariance function of $\hat{v}$ takes the form described in (17). For these families, the assumptions that maximum likelihood estimators exist and the Fisher information matrix is finite are 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

$$\sup_{x\in\mathbb{R}} |x| f_0(x) < \infty \quad \text{and} \quad \int (\dot{f}_0/f_0))^2(x) + (1 + x(\dot{f}_0/f_0)(x))^2 \mathrm{d}F_0(x) < \infty \tag{111}$$

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

$$\sup_{x\in\mathbb{R}^+} x \log x f_0(x) < \infty \quad \text{and} \quad \int (1 + x(\dot{f}_0/f_0)(x))^2 + (1 + \log x + x \log x(\dot{f}_0/f_0)(x))^2 \mathrm{d}F_0(x) \tag{112}$$

for scale-shape families[13]. 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 have the following property:

$$g(t) = \nabla_\theta F(x,\theta)\big|_{x=F^{-1}(t,\theta)} = \frac{-1}{\theta_2} \begin{bmatrix} f_0(F_0^{-1}(t)) \\ F_0^{-1}(t) f_0(F_0^{-1}(t)) \end{bmatrix} \tag{113}$$

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

$$\dot{g}(t) = \nabla_\theta \log f(x,\theta)\big|_{x=F^{-1}(t,\theta)} = \frac{-1}{\theta_2} \begin{bmatrix} (\dot{f}_0/f_0)(F_0^{-1}(t)) \\ 1 + F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t)) \end{bmatrix} \tag{114}$$

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

$$I(\theta) = \int_{[0,1]} \dot{g}(t)\dot{g}^\top(t)\mathrm{d}t = \frac{1}{\theta_2^2} \begin{bmatrix} \iota_{11} & \iota_{12} \\ \iota_{12} & \iota_{22} \end{bmatrix} = \frac{1}{\theta_2^2} I_0 \tag{115}$$

where each $\iota_{ij}$ can be derived from equation (114) and $I_0$ is a fixed matrix depending only on the

[13]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.

model.

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

$$g(t) = \begin{bmatrix} \frac{-\theta_2}{\theta_1} F_0^{-1}(t) f_0(F_0^{-1}(t)) \\ \frac{1}{\theta_2} \log(F_0^{-1}(t)) F_0^{-1}(t) f_0(F_0^{-1}(t)) \end{bmatrix} \tag{116}$$

and

$$\dot{g}(t) = \begin{bmatrix} \frac{-\theta_2}{\theta_1} \left(1 + F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t))\right) \\ \frac{1}{\theta_2} \left(1 + \log(F_0^{-1}(t)) + \log(F_0^{-1}(t)) F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t))\right) \end{bmatrix} \tag{117}$$

so that

$$I(\theta) = \begin{bmatrix} \frac{\theta_2^2}{\theta_1^2}\sigma_{11} & \frac{-1}{\theta_1}\sigma_{12} \\ \frac{-1}{\theta_1}\sigma_{12} & \frac{1}{\theta_2^2}\sigma_{22} \end{bmatrix} \tag{118}$$

Consider the third term in (17):

$$g^\top(s) \left( \int_0^1 \dot{g}(r) \dot{g}^\top(r) \mathrm{d}r \right)^{-1} g(t). \tag{119}$$

Given the above expressions for $g$ and $\dot{g}$, it is straightforward to show that the terms that depend on parameters cancel for members of either the location-scale or scale-shape class. Therefore the distribution of the parametric empirical process does not depend on specific parameter values for members of these model classes. Note also that because $\dot{g}$ is the score function of the model, the conditions given for finite Fisher information, equations (111) and (112), are equivalent to the assumptions that $\dot{g}$ exists a.e. and $\int \dot{g}\dot{g}^\top < \infty$, assumptions that are needed for a well-behaved compensator. Invariance of the compensator to parameter values for either of these classes is analogous — the compensator is constructed using only the augmented score function $h$, and as such, the parameter values in the integrand of the compensator,

$$h(s,\theta)^\top \left( \int_s^1 h(s,\theta) h^\top(s,\theta) \mathrm{d}s \right)^{-1} \int_s^1 h(r,\theta) \mathrm{d}\mathbb{F}_n(r) \tag{120}$$

can be factored out in the same way using the above calculations and partitioned matrices.

## C Proof of results in the text

*Proof of Theorem 1:* Durbin's approximation $P_g$ in (26) requires that $\frac{\mathrm{d}^2}{\mathrm{d}^2 t}\sigma^2(t)$ be finite for all $t$. This is implied by the condition that $\frac{\partial^2}{\partial x \partial \theta} f(x,\theta)$ is finite: the derivatives of the covariance function for the parametric empirical process are (letting $s \leq t$ and suppressing dependence on $\theta$ as an argument in the functions $g$ and $I$)

$$\rho_{10}(s,t) = 1 - t - \dot{g}^\top(s) I_\theta^{-1} g(t), \qquad \rho_{01}(s,t) = -s - g^\top(s) I_\theta^{-1} \dot{g}(t) \tag{121}$$

and the second derivatives are

$$\rho_{20}(s,t) = -\ddot{g}^\top(s) I_\theta^{-1} g(t), \qquad \rho_{11}(s,t) = -\dot{g}^\top(s) I_\theta^{-1} \dot{g}(t) \qquad \rho_{02}(s,t) = -g^\top(s) I_\theta^{-1} \ddot{g}(t). \tag{122}$$

When evaluated at $s = t$, we find that $\rho_{20}(t,t) = \rho_{02}(t,t)$, 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 $t_0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma^2(t)\Big|_{t=t_0} = \rho_{10}(t_0,t_0) + \rho_{01}(t_0,t_0) = 0. \tag{123}$$

We also have, from (121),

$$\rho_{10}(t,t) - \rho_{01}(t,t) = 1 \tag{124}$$

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

$$\rho_{10}(t_0,t_0) = -\rho_{01}(t_0,t_0) = 1/2. \tag{125}$$

Inserting (125) and (122) into (26), we have the result. ■

*Proof of Theorem 2:* Because $\theta$ is estimated by maximum likelihood, the covariance function of $\hat{v}$ is (17),

which implies that

$$\sigma^2(t) = t - t^2 - g^\top(t) I^{-1} g(t) \tag{126}$$

and a Taylor expansion around $t_0$ shows that the standard deviation of $\hat{v}$ locally about $t_0$ is

$$\sigma(t) = \sigma(t_0) + \frac{1}{2(2k)!\sigma(t_0)} \frac{\mathrm{d}^{(2k)}}{\mathrm{d}t^{(2k)}} \sigma^2(t_0) |t - t_0|^{(2k)} (1 + o(1)), \quad t \to t_0 \tag{127}$$

because all derivatives of order lower than $2k$ are zero by assumption. By Lemma 2, the correlation function of $\hat{v}$ locally about $t_0$ has a first-order expansion for all parametric models:

$$r(s,t) = 1 - \frac{1}{2\sigma^2(t_0)} |t - s| (1 + o(1)), \quad s, t \to t_0. \tag{128}$$

These results, combined with Theorem 8.2 of Piterbarg (1996) imply the result. Specifically, because the correlation function admits a first-order expansion, while for the standard deviation the order of the expansion is $2k > 1$, case $(i)$ of the theorem applies. Specialized to this context, we have

$$\mathrm{P}\left\{ \sup_{t \in [0,1]} \hat{v}(t) > a \right\} = H(\sigma, k) \left( \frac{a}{\sigma(t_0)} \right)^{2 - 1/k} \Psi\left( \frac{a}{\sigma(t_0)} \right) (1 + o(1)), \quad a \to \infty \tag{129}$$

where

$$H(\sigma, k) = \int_{\mathbb{R}} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t \tag{130}$$

and $A$ and $C$ as described in the statement of the theorem (which come from the leading terms in the expansions of the variance and covariance functions above). Using the substitution $x = t^{2k}$, one finds

$$H(\sigma, k) = \int_{\mathbb{R}} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t = 2 \int_{[0,\infty)} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t = \frac{C}{kA} \Gamma\left( \frac{1}{2k} \right) \tag{131}$$

Finally use the relation

$$a\Psi(a) = \phi(a)(1 + o(1)) \tag{132}$$

in (129) to establish the result. ■

**Lemma 2.** *Let $\hat{v}$ have covariance function $\rho$ as in (16) or (17) and correlation function $r(s,t) =$*

$\rho(s,t)/\sqrt{\sigma^2(s)\sigma^2(t)}$. *Then*

$$r(s,t) = 1 - \frac{1}{2\sigma^2(t_0)}|t-s|(1+o(1)), \quad s,t \to t_0 \tag{133}$$

*Proof of Lemma 2:* Expanding the squared covariance function $\rho^2(s,t)$ in $s$ around $t$ results in

$$\rho^2(s,t) = \rho^2(t,t) + 2\rho(t,t)\rho_{10}(t,t)(s-t)(1+o(1)), \quad s \to t, \tag{134}$$

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

$$\rho(s,s) = \rho(t,t) + [\rho_{10}(t,t) + \rho_{01}(t,t)](s-t)(1+o(1)), \quad s \to t. \tag{135}$$

This implies that

$$\begin{aligned}
\rho^2(s,t) - \rho(s,s)\rho(t,t) &= \rho^2(t,t) + 2\rho(t,t)\rho_1(t,t)(s-t) \\
&\quad - \rho^2(t,t) - \rho(t,t)[\rho_{10}(t,t) - \rho_{01}(t,t)](s-t) + o(s-t), \quad s \to t \\
&= \rho(t,t)[\rho_{10}(t,t) - \rho_{01}(t,t)](s-t) + o(s-t) \\
&= \rho(t,t)(s-t)(1+o(1)), \quad s \to t,
\end{aligned} \tag{136}$$

this last equality occurring because $\rho_{10}(t,t) - \rho_{01}(t,t) = 1$ for all $t$. Continuity of $\sigma^2(t) = \rho(t,t)$ implies that $\rho(t,t) = \rho(t_0,t_0) + o(1)$ so we can rewrite the above as

$$= -\sigma^2(t_0)|t-s|(1+o(1)), \quad s,t \to t_0. \tag{137}$$

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

$$\begin{aligned}
r(s,t) &= \sqrt{1 - \frac{\sigma^2(t_0)}{\sigma^2(s)\sigma^2(t)}|t-s|(1+o(1))} \\
&= 1 - \frac{1}{2\sigma^2(t_0)}|t-s|(1+o(1)), \quad s,t \to t_0.
\end{aligned} \tag{138}$$

■

*Proof of Theorem 3:* The result follows from the combination of Peskir (2002, Theorem 2.2) and the transition distributions of Gauss-Markov processes, given above in (38). Namely, because $y$ is Markovian,

$$\mathrm{P}\left\{y_t \in B\right\} = \int_0^t \mathrm{P}\left\{y_t \in B | y_s = a\right\} \mathrm{d}F(s) \tag{139}$$

for all measurable $B \subseteq [a, \infty)$. Given the distributions (38),

$$\mathrm{P}\left\{y_t \in [a, \infty)\right\} = \Psi\left(\frac{a}{\sqrt{\rho(t,t)}}\right) \tag{140}$$

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

$$\mathrm{P}\left\{y_t \in [a, \infty) | y_s = a\right\} = \Psi\left(\frac{a - m(s,t)}{\sqrt{V(s,t)}}\right) \tag{141}$$

where $m$ and $V$ are defined above. The distribution of $\tau_a$ has a density because of the relationship between Brownian motion and $y$, that is, equation (40). ■

*Proof of Lemma 1:* This proof is adapted and extended slightly from Khmaladze and Koul (2009, proof of Lemma 2.1, p. 3169). For notational simplicity, let $g(t,\theta) = g_t = [g_{t1}, g_{t2}, \ldots g_{tp}]^\top$, and define $G_t = g_t g_t^\top$. The image and kernel of $G_t$ are

$$\mathcal{I}(G_t) = \left\{c \in \mathbb{R}^p : c = G_t a,\ a \in \mathbb{R}^p\right\} \tag{142}$$

$$= \left\{c : c = b g_t,\ b \in \mathbb{R}\right\} \tag{143}$$

and

$$\mathcal{K}(G_t) = \left\{a \in \mathbb{R}^p :\ G_t a = \mathbf{0}\right\} \tag{144}$$

$$= \left\{a :\ g_t^\top a = 0\right\}. \tag{145}$$

Notably, $g_t \in \mathcal{I}(G_t)$. For any $c = b g_t \in \mathcal{I}(G_t)$,

$$G_t c = b g_t g_t^\top g_t = b g_t \sum_{i=1}^p g_{ti}^2 = c \left\|g\right\|^2 \tag{146}$$

which implies that $G_t^{-1}$ is any matrix such that for $c \in \mathcal{I}(G_t)$,

$$G_t^{-1} c = \frac{c}{\|g\|^2} + a, \qquad a \in \mathcal{K}(G_t) \tag{147}$$

Now, for any $d \in \mathcal{I}(G_t)$,

$$g_t^\top G_t^{-1} d = \frac{g_t^\top d}{\|g\|^2} \tag{148}$$

because $g_t^\top a = 0$ for any $a \in \mathcal{K}(G_t)$. Therefore the above quantity $g_t^\top G_t^{-1} d$ is defined uniquely for any $d \in \mathcal{I}(G_t)$. In particular,

$$g_t^\top G^{-1} g_t = \frac{g_t^\top g_t}{\|g\|^2} = 1. \tag{149}$$

This means

$$g_t^\top \hat{\beta}(t, X, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \lambda(t, X_i, \hat{\theta}) = \overline{\lambda}(t, \hat{\theta}) \tag{150}$$

and implies the result (68). ■

## 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.

D. Andrews. A conditional Kolmogorov test. *Econometrica*, 65(5):1097–1128, Sep. 1997.

J.-M. Azaïs and M. Wschebor. *Level Sets and Extrema of Random Processes and Fields*. Wiley, 2009.

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.

R. Davidson and J. MacKinnon. Heteroskedasticity-robust tests in regression directions. *Annales de l'INSEE*, (59/60):183–218, 1985.

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. 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.

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.

W. Stute, W. Gonzáles Manteiga, and M. Presedo Quindmill. Bootstrap based goodness-of-fit-tests. *Metrika*, 40:243–256, 1993.

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

A. van der Vaart and J. Wellner. Empirical processes indexed by estimated functions. In E. Cator, G. Jongbloed, C. Kraaikamp, H. Lopuhaä, and J. Wellner, editors, *Asymptotics: Particles, Processes and Inverse Problems*, volume 55 of *IMS Lecture Notes — Monograph Series*, pages 234–252. Institute of Mathematical Statistics, 2007.

J. Wooldridge. A unified approach to robust, regression-based specification tests. *Econometric Theory*, 6(1):17–43, 1990.