© 2011 by Thomas Michael Parker. All rights reserved.

INFERENCE FOR PARAMETRIC EMPIRICAL PROCESSES

BY

THOMAS MICHAEL PARKER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Roger Koenker, Chair
Professor Anil Bera
Professor Stephen Portnoy
Professor Richard Sowers

Abstract

Parametric empirical processes, empirical processes that incorporate parametric modeling components into their definition, play a natural role in many inferential settings. In this dissertation we illustrate their application, highlighting methods for inference that rely on supremum-norm test statistics. Chapter 1 illustrates the use of supremum-norm statistics for inference in simple parametric modeling situations. The estimation of parameters alters the distribution of commonly-used test statistics, but methods are explored that accommodate these differences using approximations based on features of the parametric model. Chapter 2 extends this methodology to tests based on a group of related processes, kernel-transformed empirical processes. It is shown that a martingale transform coupled with the kernel transformation results in processes that have simple limiting distributions. This makes them attractive for inference because their tractability facilitates straightforward calculation of approximate critical values in a variety of cases. Chapter 3 extends this methodology in a different direction, to two-sample tests of stochastic dominance. Tests for any order of dominance are considered, and the distribution of test statistics is shown to be related to the family of iteratively integrated Brownian bridges. A Rice series approximation, used to find boundary crossing probabilities for smooth Gaussian processes, is proposed as an inferential method. Tests that use residuals from conditional models are also considered and it is shown that their distribution is nonstandard due to the way in which the residuals are incorporated into tests. It is shown how approximate critical values can be altered to reflect this estimation effect.

Table of Contents

Chapter 1

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

1.1 Introduction

Empirical processes are central to the theory of supremum-norm specification tests. The standard Kolmogorov-Smirnov test statistic, used to test whether the data can be accurately represented using a certain parametric distribution, enjoys the property that its distribution under the null is invariant to the distribution that is hypothesized. This simplicity comes at a cost: the test is only applicable to simple null hypotheses, that is, for hypotheses that determine the parametric family of the distribution function *and* the parameters of the distribution. A general study of the convergence of empirical processes with estimated parameters was first conducted by Durbin (1973a) and Neuhaus (1976). The limiting distributions of these processes were found to be significantly more complex than the limiting distribution of the process for simple null hypotheses. As a result, the evaluation of sup-norm test statistics based on these processes has been an enduring problem. Given this difficulty, the calculation of critical values for tests 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.

One solution to this problem for supremum-norm tests (parallel to techniques devised for example by Durbin et al. (1975) for Cramér-von Mises-type tests,) is to calculate appropriate distributionally dependent critical values for each test. For sup-norm tests, Durbin (1973b, 1975, 1985), explored a number of approaches to the calculation of critical values and these results deserve greater recognition as an alternative methodology. In particular, Durbin (1985) provides a collection of simple approximations that are accurate, generalizable, and involve only modest computation. These approximate boundary crossing probabilities are analyzed in Section 1.3. Some justification for their great accuracy is given by links that the resulting expressions have to 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 as developed by 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 gone largely overlooked. The results presented here demonstrate the generality and accuracy of this methodology and indicate its usefulness as a model-checking device for semiparametric models.

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 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).

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

1.2 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

$$V_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x,\theta_0)), \qquad x \in \mathcal{X} \tag{1.1}$$

for simple null hypotheses, or the parametric empirical process

$$\hat{V}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x, \hat{\theta})) \qquad x \in \mathcal{X} \tag{1.2}$$

for composite null hypotheses, where $\hat{\theta}$ is some estimate of θ_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 (1.1) equivalent to

$$v_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(I(F(X_i, \theta_0) \leq t) - t \right), \quad t \in [0, 1]. \tag{1.3}$$

That is, under the null hypothesis, process (1.1) 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 $[0, 1]$ — in other words, V_n converges weakly to $B \circ F$, a time-changed Brownian bridge.

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

To explore this further, we make the following two assumptions, one with respect to the parametric model and one with respect to the parameter estimator:

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

$$g(t, \theta) = \nabla_\theta F(x, \theta)\big|_{x = F^{-1}(t, \theta_0)} \tag{1.4}$$

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

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

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_P(1) \tag{1.5}$$

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

the estimator $\hat{\theta}_n$ to θ_0, the effect of parameter estimation is not asymptotically negligible. Consider the following decomposition of $\hat{v}_n(t)$ (start here with the transformation $t = F(x, \hat{\theta})$):

$$\begin{aligned}\hat{v}_n(t) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(I(F(X_i, \hat{\theta}) \leq t) - t \Big) && (1.6) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(I(F(X_i, \theta_0) \leq t) - t \big) + \sqrt{n} \Big(F(F^{-1}(t, \theta_0), \hat{\theta}_n) - t \Big) \\ &+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \Big(I(F(X_i \leq \hat{\theta}) \leq t) - F(F^{-1}(t, \theta_0), \hat{\theta}_n) \Big) - \big(I(F(X_i, \theta_0) \leq t) - t \big) \Big\} && (1.7)\end{aligned}$$

Using assumption **A1**, **A2** and a one-term Taylor expansion, it can be shown[1] that the last term in (1.7) is $o_P(1)$ uniformly in $t \in [0, 1]$ and that

$$\hat{v}_n(t) = v_n(t) - \sqrt{n}(\hat{\theta}_n - \theta_0)^\top g(t, \theta_0) + o_P(1), \tag{1.8}$$

where the $o_P(1)$ term is uniform in $t \in [0, 1]$. Durbin (1973a) showed that $\hat{v}_n$ converges weakly to a mean-zero Gaussian process $\hat{v}$. From (1.8) it is apparent that in general the distribution of $\hat{v}_n$ may depend on the value of the parameter θ_0 and the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, and the distribution of the limit $\hat{v}$ may as well.

Because the parametric empirical process depends on the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$, the distribution of (1.6) can be complicated, but in the limit it can be simplified if more is assumed regarding the estimator $\hat{\theta}_n$[2]. The most common simplifying assumption is that $\hat{\theta}_n$ is asymptotically linear; that is,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, \theta_0) + o_P(1) \tag{1.9}$$

where ψ is such that

$$\int \psi(x, \theta_0) \mathrm{d}F(x, \theta_0) = 0, \qquad \int \psi(x, \theta_0) \psi^\top(x, \theta_0) \mathrm{d}F(x, \theta_0) = J \tag{1.10}$$

and J is a finite $p \times p$ positive definite matrix. Under these conditions, it can be shown[3] using (1.8) that (1.6) converges weakly to

$$\hat{v} \overset{D}{=} v - g^\top(t, \theta_0) \int \psi \mathrm{d}v \tag{1.11}$$

[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] It should be noted that for the purposes of hypothesis testing it is not strictly necessary that this relationship be known, if one employs the transformation technique of Khmaladze (1981) discussed in Section 1.4.

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

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

$$\rho(s,t) = s \wedge t - st - g(s,\theta_0)^\top \int_0^t H(r)\mathrm{d}r - g(t,\theta_0)^\top \int_0^s H(r)\mathrm{d}r + g(s,\theta_0)^\top J g(t,\theta_0) \tag{1.12}$$

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 $I(\theta)$ (which requires more regularity conditions on F and its density f,) 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

$$\rho(s,t) = s \wedge t - st - g^\top(s,\theta_0) I^{-1}(\theta_0) g(t,\theta_0). \tag{1.13}$$

The additional terms in expressions (1.12) and (1.13) as compared to the covariance function of the Brownian bridge ($\rho(s,t) = s \wedge t - st$) 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 1.5, a maximum likelihood estimator exists and so the covariance function takes the form of (1.13).

1.3 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. 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 $\hat{v}$.

1.3.1 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 pro-

cesses. Therefore, assume y has a covariance function $\rho(s,t)$ that is differentiable in both arguments for $0 \leq s \leq t \leq 1$. Note that this is weaker than full differentiability of ρ, because it is not necessary that ρ be differentiable on the diagonal (for such processes, other methods are available for the computation of boundary crossing probabilities). 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

$$\lim_{s \nearrow t} \frac{\mathrm{Var}\left(y(t) - y(s)\right)}{t-s} = \lim_{s \nearrow t} \left\{ \frac{\partial \rho(s,t)}{\partial s} - \frac{\partial \rho(s,t)}{\partial t} \right\} = \lambda_t \tag{1.14}$$

where $0 < \lambda_t < \infty$ for all t. For example, Brownian motion satisfies this condition with $\lambda_t \equiv 1$, as do processes with covariance functions (1.12) or (1.13), 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

$$P(a) = \mathrm{P}\left\{ \sup_{t \in [0,1]} y(t) \geq a \right\}, \tag{1.15}$$

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 τ_a:

$$P(a) = \int_0^1 p(t,a)\mathrm{d}t = \int_0^1 b(t,a) f(t,a)\mathrm{d}t \tag{1.16}$$

where

$$b(t,a) = \lim_{s \to t} \frac{\mathrm{E}\left[\mathrm{I}(s < \tau_a)\left(a - y(s)\right) | y(t) = a\right]}{t-s} \tag{1.17}$$

and

$$f(t,a) = \frac{1}{\sqrt{2\pi\rho(t,t)}} \exp\left\{ \frac{-a^2}{2\rho(t,t)} \right\}. \tag{1.18}$$

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

1.3.2 The first approximation P_1

Durbin's first approximation, achieved simply through the removal of the indicator function from (1.17), was justified by the fact that the approximation holds exactly in the special case of Brownian motion and more generally by the fact that any Gaussian process satisfying Durbin's (mild) conditions "...behaves locally like Brownian motion and the boundary is locally linear[4]" (Durbin, 1985, p. 110-111). That is, approximation P_1 starts with the following approximation to the function b:

$$b_1(t,a) = \frac{\rho_1(t,t)}{\rho(t,t)}a \tag{1.19}$$

where $\rho_1(t,t) := \frac{\partial \rho(s,t)}{\partial s}\Big|_{s=t}$. This approximation to b owes its simple form to a hypothetical regression argument and the definition of a derivative[5]. Approximations to the first passage density for y and the boundary crossing probability are respectively

$$p_1(t,a) = b_1(t,a)f(t,a) \tag{1.20}$$

and

$$P_1(a) = \int_0^1 p_1(t,a)\mathrm{d}t. \tag{1.21}$$

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

1.3.3 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 identical to an asymptotically exact (as the boundary $a \to \infty$) crossing probability. The results are based on the theory

[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.

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 $t_0 := \text{argmax}_t\, \sigma^2(t)$. Durbin's global approximation P_g is

$$P_g(a) = \frac{\rho_1(t_0,t_0)}{\sigma^2(t_0)} \left(\frac{-2\sigma^2(t_0)}{\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)} \right)^{1/2} \exp\left\{ \frac{-a^2}{2\sigma^2(t_0)} \right\}. \tag{1.22}$$

This is achieved by starting with equation (1.20), evaluating all the non-exponential parts at t_0, and replacing the exponential part with a rough expansion to evaluate it. This formula is easily inverted for the purposes of calculating approximate critical values, and therefore 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 (1.22).

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

$$P_g(a) = \frac{\exp\left\{\frac{-a^2}{2\sigma^2(t_0)}\right\}}{2\sqrt{-\sigma^2(t_0)\left(\rho_{11}(t_0,t_0)+\rho_{12}(t_0,t_0)\right)}}. \tag{1.23}$$

A drawback to the use of P_g is that if $\rho_{11}(t_0,t_0) = \rho_{12}(t_0,t_0) = 0$ (which occurs, e.g., when testing $\mathcal{N}(\mu,\sigma^2)$ with μ 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: given his original derivation, Durbin (1985, p. 113) could only claim that $P_g(a) = P_1(a) + O_P\left(a^{-2}\right)$ as $a \to \infty$.

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

$$\mathrm{P}\left\{ \sup_{t\in[0,1]} \hat{v}(t) > a \right\} = H(\sigma,k) \left(\frac{a}{\sigma(t_0)} \right)^{1-1/k} \phi\left(\frac{a}{\sigma(t_0)} \right) (1+o(1)), \quad a \to \infty \tag{1.24}$$

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

where ϕ is the standard normal density,

$$H(\sigma,k)=\frac{C}{kA}\Gamma\left(\frac{1}{2k}\right) \tag{1.25}$$

and

$$A=\left(\frac{-\frac{\mathrm{d}^{(2k)}}{\mathrm{d}t^{(2k)}}\sigma^2(t_0)}{2(2k)!\sigma^2(t_0)}\right)^{1/(2k)}, \qquad C=\frac{1}{2\sigma^2(t_0)}. \tag{1.26}$$

Note that setting $k=1$ is equivalent to the existence of $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)$ and (1.24) is identical to (1.22). This is because if $k=1$,

$$\mathrm{P}\left\{\sup_{t\in[0,1]}\hat{v}(t)>a\right\}\approx H(\sigma,1)\phi\left(\frac{a}{\sigma^2(t_0)}\right) \tag{1.27}$$

$$=\frac{1}{2\sigma^2(t_0)}\sqrt{\frac{4\sigma^2(t_0)}{-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)}}\sqrt{\pi}\frac{\exp\left\{\frac{a}{\sigma^2(t_0)}\right\}}{\sqrt{2\pi}}, \tag{1.28}$$

and because it can be shown that $\rho_1(t_0,t_0)=1/2$ (see the proof of Theorem 1),

$$=\frac{\rho_1(t_0,t_0)}{\sigma^2(t_0)}\left(\frac{-2\sigma^2(t_0)}{\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^2(t_0)}\right)^{1/2}\exp\left\{\frac{-a^2}{2\sigma^2(t_0)}\right\}=P_g(a). \tag{1.29}$$

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.

1.3.4 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

[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.

essentially to calculate boundary crossing probabilities as if this inconvenience did not exist. This final approximation improves upon P_1 and is the solution to a numerically evaluated integral equation. A great deal of 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 ρ. Define[9]

$$\begin{bmatrix} \beta_1(s,t) \\ \beta_2(s,t) \end{bmatrix} = \begin{bmatrix} \rho(s,s) & \rho(s,t) \\ \rho(t,s) & \rho(t,t) \end{bmatrix}^{-1} \begin{bmatrix} \rho_2(s,t) \\ \rho_1(t,t) \end{bmatrix}. \tag{1.30}$$

Durbin (1985) showed that the exact density $p_2(t,a)$ of the first passage time for Gauss-Markov process y is the solution to the integral equation

$$p_2(t,a) = p_1(t,a) - a\int_0^t \left[\beta_1(s,t) + \beta_2(s,t)\right] f(t|s,a) p_2(s,a)\mathrm{d}s. \tag{1.31}$$

Because (1.31) is a Volterra equation of the second kind, the solution p_2 is unique. In (1.31), $p_1(t,a)$ is as in (1.20) 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 $s \le t$, in the case of a constant boundary, the transition distribution is

$$F(t|s,a) = F(y(t)|y(s)=a) = \mathcal{N}\left(\frac{\rho(s,t)}{\rho(s,s)}a, \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)}\right) \tag{1.32}$$

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

$$P_2(a) = \int_0^1 p_2(t,a)\mathrm{d}t \tag{1.33}$$

Durbin (1985) showed that equation (1.31) 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 (1.33) where the covariance function of $\hat{v}$ is used to calculate (1.31) 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 1.6.

[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 $s \le t$.

[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 $y(s)$, $s,t \le r$.

Gauss-Markov processes

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

$$(x,t)|(y,s) \sim \mathcal{N}\left(\frac{\rho(s,t)}{\rho(s,s)}y, \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)}\right). \tag{1.34}$$

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

$$\rho(r,t) = \frac{\rho(r,s)\rho(s,t)}{\rho(s,s)}. \tag{1.35}$$

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

$$y(t) = \zeta(t)W\left((\eta/\zeta)(t)\right). \tag{1.36}$$

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

$$\frac{\partial}{\partial t}f(x,t|y,s) = -\frac{\partial}{\partial x}\left\{A_1(x,t)f(x,t|y,s)\right\} + \frac{A_2(t)}{2}\frac{\partial^2}{\partial x^2}f(x,t|y,s) \tag{1.37}$$

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

$$A_1(x,t) = \lim_{s\to t}\frac{\partial}{\partial t}\frac{\rho(s,t)}{\rho(s,s)}y = \frac{\rho_2(t,t)}{\rho(t,t)}y \tag{1.38}$$

and

$$A_2(t) = \lim_{s\to t}\frac{\partial}{\partial t}\rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)} = \rho_1(t,t) - \rho_2(t,t) \tag{1.39}$$

The function A_2 in particular is intimately connected to Durbin's approximation— see equation (1.32) above and equation (4) of Durbin (1985). The function A_1 is also strikingly similar to equation (1.19) above, especially given the fact that for the parametric empirical process, $\rho_1(t,t) - \rho_2(t,t) = 1$ 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]:

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 $\rho(s,t)$ such that y has regular conditional probabilities. Let $a > 0$, and define*

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

Then the density p of τ_a satisfies the following integral equation:

$$\Psi\left(\frac{a}{\sqrt{\rho(t,t)}}\right) = \int_0^t \Psi\left(\frac{a - m(s,t)}{\sqrt{V(s,t)}}\right) p(s,a)\mathrm{d}s \tag{1.40}$$

where

$$m(s,t) = \frac{\rho(s,t)}{\rho(s,s)}a \quad \text{and} \quad V(s,t) = \rho(t,t) - \frac{\rho^2(s,t)}{\rho(s,s)} \tag{1.41}$$

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

The connection between the integral equations (1.40) and (1.31) is not as straightforward as it might seem. Differentiating equation (1.40) with respect to t results in another integral equation that is remarkably similar to equation (1.31). 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 1.5 was made on practical grounds: although equation (1.40) is slightly simpler to put into practice, Durbin's equation (1.31) was more stable in numerical experiments.

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

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

Computation of p_2

Equation (1.31) 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 ρ make this easy to accomplish: the kernel of the integral equation — $-a(\beta_1(s,t)+\beta_2(s,t))f(t|s,a)$, for $s \leq t$ — has a limiting value of 0 whenever t or s are 0, 1, or equal to each other. Given an equally-spaced partition $\{t_i = (i-1)/m,\ i = 1,2,\cdots m+1\}$ (the value of m is chosen by the researcher,) the integration algorithm simplifies to the following recursive rule: for $i = 0, 1$ (recall $t_0 = 0$),

$$p_2(0,a) = 0, \qquad p_2(t_2,a) = p_1(t_2,a) \tag{1.42}$$

and for $i \geq 3$

$$p_2(t_i,a) = p_1(t_i,a) + a\frac{1}{m}\sum_{j=2}^{i-1} K(t_j,t_i)p_2(t_j,a) \tag{1.43}$$

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.

1.3.5 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. One possible drawback to the approach outlined below should be noted: since the approximates presented in this section are applied to the Gaussian limit of the parametric empirical process, there is no formal guarantee that they necessarily improve as the sample size of a given experiment increases. However, in the case examined in Section 1.6, perfor-

mance is not affected as sample size increases. It seems likely that this is due to the accuracy of the approximations relative to small sample anomalies.

Furthermore, these methods are generalizable. 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 (1.12) or (1.13)) and plug the covariance function and its derivatives into a computational formula. In addition, as will be seen in Section 1.6, tests that use adjusted critical values can perform at least as well as tests that rely on simulation methods.

1.4 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 problems inherent in the use of the parametric empirical process. In the one-sample setting, some interesting connections can be made between the martingale transform, the parametric empirical process, and projection techniques.

Viewed as a real-valued random element of $L_2[0,1]$, $\mathbb{F}_n$ is a submartingale with respect to $\mathscr{F}^{\mathbb{F}_n} = \{\mathscr{F}_t^{\mathbb{F}_n}\}_{t\geq 0}$, the filtration of σ-algebras generated by $\mathbb{F}_n$. Therefore the Doob-Meyer decomposition implies a right-continuous increasing and predictable compensator K may be calculated that renders $\mathbb{F}_n - K$ a martingale with respect to $\mathscr{F}^{\mathbb{F}_n}$. 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

$$\tilde{V}_n(x) = \sqrt{n}\left(\mathbb{F}_n(x) - K(x, \mathbb{F}_n, \hat{\theta}_n)\right) \tag{1.44}$$

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 (1.44) asymptotically distribution-free.

The function g defined in equation (1.4) 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

$$\dot{g}(t) = \frac{\partial}{\partial t} g(t,\theta) = \frac{\partial}{\partial \theta} \log f(x,\theta)\Big|_{x=F^{-1}(t,\theta)} \tag{1.45}$$

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 Γ by

$$h(t,\theta) = \begin{bmatrix} 1 \\ \frac{\partial g(t,\theta)}{\partial t} \end{bmatrix} \qquad \text{and} \qquad \Gamma(t,\theta) = \int_t^1 h(s,\theta)h(s,\theta)^\top \mathrm{d}s. \tag{1.46}$$

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

$$K(t, \mathbb{F}_n, \theta) = \int_0^t h(s,\theta)^\top \Gamma^{-1}(s,\theta) \int_s^1 h(r,\theta)\mathrm{d}\mathbb{F}_n(r)\mathrm{d}s. \tag{1.47}$$

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

$$= \int_0^1 \int_0^{t\wedge r} h(s,\theta)^\top \Gamma^{-1}(s,\theta)\mathrm{d}s\, h(r,\theta)\mathrm{d}\mathbb{F}_n(r). \tag{1.48}$$

One may think of equation (1.47) as a functional analog to $\hat{y} = x\hat{\beta}$ familiar from usual regression analysis, with $h(t)$ playing the role of explanatory variable and the projection $\Gamma^{-1}(t)\int_t^1 h(s)\mathrm{d}\mathbb{F}_n(s)$ as $\hat{\beta}$. Note also the fact that $\Gamma(0,\theta)$ is simply an augmented version of the Fisher information matrix of the model. Because of the similarities between h and the score, and Γ and the Fisher information, it can be shown that the compensator also has a form that does not 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 1.4.1. 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.

1.4.1 Computation of the compensator

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

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

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:

$$X_i = \begin{bmatrix} \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_{m+1}) \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_m) \\ \vdots & \vdots \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{g}(t_i) \end{bmatrix} \qquad y_i = \begin{bmatrix} \sqrt{m}\left(\mathbb{F}_n(t_{m+1}) - \mathbb{F}_n(t_m)\right) \\ \sqrt{m}\left(\mathbb{F}_n(t_m) - \mathbb{F}_n(t_{m-1})\right) \\ \vdots \\ \sqrt{m}\left(\mathbb{F}_n(t_i) - \mathbb{F}_n(t_{i-1})\right) \end{bmatrix} \tag{1.49}$$

Then, least squares coefficients for each t_i are calculated:

$$\begin{aligned} \hat{\beta}(t_i) &= (X_i^\top X_i)^{-1} X_i^\top y_i \\ &= \begin{bmatrix} \frac{1}{m}(m-j+2) & \frac{1}{m}\sum_{j=i}^{m+1}\dot{g}(t_j) \\ \frac{1}{m}\sum_{j=i}^{m+1}\dot{g}(t_j) & \frac{1}{m}\sum_{j=i}^{m+1}\dot{g}^2(t_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=i}^{m+1}[\mathbb{F}_n(t_j) - \mathbb{F}_n(t_{j-1})] \\ \sum_{j=i}^{m+1}\dot{g}(t_j)[\mathbb{F}_n(t_j) - \mathbb{F}_n(t_{j-1})] \end{bmatrix}. \end{aligned} \tag{1.50}$$

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:

$$\hat{K}(t_i) = \frac{1}{m}\sum_{j=1}^{i} h^\top(t_j)\hat{\beta}(t_j). \tag{1.51}$$

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.

1.4.2 Comparison with Wooldridge (1990)

Wooldridge (1990), extending the work of Davidson and MacKinnon (1985) in the context of robustifying regression specification tests, proposed a projection technique that achieves the same goal as the martingale transform — it accounts for the effect of estimation and leaves statistics asymptotically distribution-free. Khmaldaze'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 and they differ.

Wooldridge's proposal is moment-based: suppose given t we have the hypothesized conditional moment condition

$$\mathrm{E}\left[\phi(t, X_i, \theta)\right] = 0, \qquad \theta \in \Theta,\ i = 1, 2, \ldots \tag{1.52}$$

and let $\{\Lambda_i(t, X_i, \theta)\}_{i=1}^n$ be some vector of "misspecification indicators" used to robustify the test statistic against misspecifications of the model. Many test statistics can be defined as the inner product

$$\hat{T}_n = \frac{1}{\sqrt{n}}\Lambda^\top(t, X, \hat{\theta})\phi(t, X, \hat{\theta}) \tag{1.53}$$

where $\Lambda \in \mathbb{R}^{n\times d}$, $\phi \in \mathbb{R}^n$, or as some functional of T_n such as $T_n^\top T_n$. Define

$$\Phi(t, X_i, \theta) = \mathrm{E}\left[\nabla_\theta \phi(t, X_i, \theta)\right], \quad i = 1, 2, \ldots, n \tag{1.54}$$

Wooldridge (1990) noted that by using a mean-value expansion,

$$\hat{T}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda(t, X_i, \hat{\theta}) \phi(t, X_i, \hat{\theta}) \tag{1.55}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda(t, X_i, \theta) \phi(t, X_i, \theta) + \sqrt{n} \left(\hat{\theta} - \theta\right)^{\top} \frac{1}{n} \sum_{i=1}^{n} \Lambda(t, X_i, \theta) \nabla_{\theta} \phi(t, X_i, \theta) + o_P(1) \tag{1.56}$$

interpreting $\nabla_{\theta} \phi$ as an $n \times p$ matrix and assuming enough regularity that the $o_P(1)$ term is uniform in t. A modified version of Wooldridge's proposed test statistic is the following (simplified slightly by using identical weights for each observation):

$$\tilde{T}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Lambda(t, X_i, \hat{\theta}) - \Phi^{\top}(t, X_i, \hat{\theta}) \hat{\beta}(\hat{\theta}))^{\top} \phi(t, X_i, \hat{\theta}) \tag{1.57}$$

where

$$\hat{\beta}(\theta) = \hat{\beta}(t, X, \theta) = \left(\sum_{i=1}^{n} \Phi(t, X_i, \theta) \Phi^{\top}(t, X_i, \theta) \right)^{-1} \sum_{i=1}^{n} \Phi(t, X_i, \theta) \Lambda(t, X_i, \theta), \tag{1.58}$$

that is, $\hat{\beta}$ is the projection of Λ onto Φ. Wooldridge shows that a quadratic form using $\tilde{T}_n$ is equivalent to a Lagrange multiplier test that converges in distribution to a χ^2 distribution with degrees of freedom equal to the dimension of Λ_i. We rewrite this statistic as an inner product (compare with (1.53)):

$$\tilde{T}_n(t) = \frac{1}{\sqrt{n}} \Lambda^{\top}(t, X, \hat{\theta}) M_{\Phi} \phi(t, X, \hat{\theta}) \tag{1.59}$$

where $\phi \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{n \times d}$ and $M_{\Phi} \in \mathbb{R}^{n \times n}$ is defined by $M_{\Phi} = I_n - P_{\Phi}$, where

$$P_{\Phi} = P_{\Phi(t, X, \hat{\theta})} = \left(\Phi \left(\Phi^{\top} \Phi \right)^{-1} \Phi^{\top} \right) (t, X, \hat{\theta}) \tag{1.60}$$

Supposing that one desires to use this technique to test the hypothesis that the data is described by the distribution function $F \in \mathcal{F}$, we may choose

$$\phi(t, X_i, \theta) = I(X_i \leq t) - F(t, \theta) \tag{1.61}$$

which has zero expectation under the null hypothesis. For each observation, define Φ by

$$\Phi(t, X_i, \theta) = \nabla_{\theta} F(F^{-1}(t, \theta), \theta) = g(t, \theta). \tag{1.62}$$

This proposal is fundamentally different from that of Khmaladze (1981) because it is assumed that M_Φ defined above exists as a nontrivial finite-dimensional projection. Because for any t, $\Phi(t, X_i, \theta) = g(t, \theta), i = 1, 2, \ldots n$ does not depend on the observations, only the null hypothesis, the matrix

$$\Phi(t, X, \theta) = g^\top(t, \theta)\mathbf{1}_n \tag{1.63}$$

is an element of the space spanned by the unit vector $\mathbf{1}_n$, and therefore $I_n - P_\Phi = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ or $I_n - M_{\mathbf{1}_n}$, also a projection that is independent of the value of t. Khmaladze's projection is fundamentally different in that it projects into the space spanned by the *function* g, rather than onto the value of g evaluated at any specific t. The adaptation of Wooldridge's statistic by projecting the score function one point at a time results in a statistic that is identically zero whenever Λ is constant. When Λ is chosen to be nonconstant, it can be shown that

$$\tilde{T}_n(t) = \frac{1}{\sqrt{n}}\Lambda^\top(t, X, \hat{\theta})M_\Phi\phi(t, \hat{\theta}) = \sqrt{n}\Lambda^\top(t, X, \hat{\theta}) \begin{bmatrix} \mathbf{1}_{X_1 \le t} - \mathbb{F}_n(t) \\ \vdots \\ \mathbf{1}_{X_n \le t} - \mathbb{F}_n(t) \end{bmatrix}, \tag{1.64}$$

which could presumably be normalized to construct a χ^2 statistic for each t. The resulting statistic would have a marginal χ^2 distribution for each t (which does not immediately imply that this collection of statistics converges weakly to a χ^2 process in t). The problem with this statistic lies in the fact that $M_\Phi \equiv M_{\mathbf{1}_n}$ annihilates the imposition of the null hypothesis in ϕ — note that the effect of subtracting $F(t, \theta)$ from each observation does not appear in (1.64) because the matrix M_Φ projects into the space orthogonal to $\mathbf{1}_n$. This is another indication of the very different character of the two projections. 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 in this context.

1.5 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

[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.

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).

1.5.1 The exponential distribution

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

$$H_0 : F(x, \lambda) = 1 - e^{-\lambda x}, \quad x \in [0, \infty),\ \lambda \in (0, \infty). \tag{1.65}$$

The function g for the exponential model is

$$g(s, \lambda) = \frac{-1}{\lambda_0}(1-s)\log(1-s)e^{\frac{\lambda}{\lambda_0}}. \tag{1.66}$$

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

$$\rho(s,t) = s \wedge t - st - (1-s)(1-t)\log(1-s)\log(1-t). \tag{1.67}$$

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

$$1 - 2t_0 + 2(1-t_0)\log(1-t_0)\left(1 + \log(1-t_0)\right) = 0. \tag{1.68}$$

The methods of Section 1.3 were applied using (1.67) to produce the approximate critical values in Table 1.1 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 (1.13) is positive definite, the covariance function of the parametric empirical process is smaller than that of the Brownian bridge for all t, and therefore critical values for the Kolmogorov-Smirnov test using the parametric empirical process should always be smaller than for the standard test (van der

Vaart and Wellner, 1996, p. 441).

Table 1.1: 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

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.

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

$$h(t,\lambda) = \begin{bmatrix} 1 \\ \frac{1}{\lambda}(1+\log(1-t)) \end{bmatrix} \tag{1.69}$$

and

$$\Gamma(t,\lambda) = \begin{bmatrix} 1-t & \frac{1}{\lambda}(1-t)\log(1-t) \\ \frac{1}{\lambda}(1-t)\log(1-t) & \frac{1}{\lambda^2}(1-t)(1+\log^2(1-t)) \end{bmatrix}. \tag{1.70}$$

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

$$\begin{aligned} K(t,\mathbb{F}_n,\hat{\lambda}) = \int_0^t \frac{1}{2}\log^2(1-\hat{\varepsilon}) - 2\log(1-\hat{\varepsilon}) - \log^2(1-\hat{\varepsilon})\mathrm{d}\mathbb{F}_n(\hat{\varepsilon}) \\ + \int_t^1 \frac{1}{2}\log^2(1-t) - 2\log(1-t) - \log(1-\hat{\varepsilon})\log(1-t)\mathrm{d}\mathbb{F}_n(\hat{\varepsilon}), \end{aligned} \tag{1.71}$$

or alternatively

$$K(t,\mathbb{F}_n,\hat{\lambda}) = \frac{1}{n}\sum_{i:\hat{\varepsilon}_i\le t}\left(\frac{-1}{2}\log^2(1-\hat{\varepsilon}_i) - 2\log(1-\hat{\varepsilon}_i)\right) + \left(\frac{1}{2}\log^2(1-t) - 2\log(1-t)\right)\big(1-\mathbb{F}_n(t)\big) - \frac{1}{n}\log(1-t)\sum_{i:\hat{\varepsilon}_i>t}\log(1-\hat{\varepsilon}_i), \quad (1.72)$$

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

$$\tilde{K}(x,\mathbb{F}_n,\hat{\lambda}) = \frac{\hat{\lambda}}{n}\sum_{i:X_i\le x}\left(2X_i - \frac{\hat{\lambda}}{2}X_i^2\right) + \left(2\hat{\lambda}x + \frac{\hat{\lambda}^2}{2}x^2\right)(1-\mathbb{F}_n(x)) - \frac{\hat{\lambda}^2}{n}x\sum_{i:X_i>x}X_i \quad (1.73)$$

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

1.5.2 The normal distribution

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

$$H_0:\ F(x,\theta) = \int_{-\infty}^{x}\frac{e^{\frac{-1}{2\sigma^2}(y-\mu)^2}}{\sqrt{2\pi\sigma^2}}\mathrm{d}y = \int_{-\infty}^{\frac{x-\mu}{\sigma}}\phi(z)\mathrm{d}z, \quad x\in\mathbb{R}, \quad (1.74)$$

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 (1.13).

Letting $\xi(s)$ be the s^{th} quantile of the standard normal distribution, so that the s^{th} quantile of the $\mathcal{N}(\mu,\sigma^2)$ distribution is $\mu+\sigma\xi(s)$, the function g for the location- and scale-unknown case is equal to

$$g(s,\theta) = \left[\begin{array}{c}\frac{\partial}{\partial\mu}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\phi(z)\mathrm{d}z\\ \frac{\partial}{\partial\sigma}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\phi(z)\mathrm{d}z\end{array}\right]_{x=\mu+\sigma\xi(s)} = \frac{-1}{\sigma}\left[\begin{array}{c}\phi(\xi(s))\\ \xi(s)\phi(\xi(s))\end{array}\right]. \quad (1.75)$$

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 (1.13), one finds that $\hat{v}$ has covariance function

$$\rho_{\mu\sigma}(s,t) = s\wedge t - st - \phi(\xi(s))\phi(\xi(t))\left(1+\frac{1}{2}\xi(s)\xi(t)\right). \quad (1.76)$$

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 1.2: 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

$$\rho_\mu(s,t) = s \wedge t - st - \phi(\xi(s))\phi(\xi(t)) \tag{1.77}$$

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

$$\rho_\sigma(s,t) = s \wedge t - st - \frac{1}{2}\xi(s)\xi(t)\phi(\xi(s))\phi(\xi(t)), \tag{1.78}$$

$\rho_\sigma(t,t)$ 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 1.2. 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 1.2 are invariant to the true values of the parameters μ and σ.

1.6 Monte Carlo experiments

Table 1.3 presents the results of a small Monte Carlo experiment using the D^- statistic for testing the null hypothesis of exponentiality against alternatives that deviate from the null by placing some mass at higher values than the exponential (two-sided tests for the detection of any departure from the null are also available, but one-sided tests are considered in order to match the exposition above). 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 1.4.1 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 1.3: 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 1.3,) 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. A second small Monte Carlo experiment was conducted using smooth local alternatives to the null hypothesis of exponentiality. Stochastically dominant alternatives were constructed in one of two ways. First, local alternative mixture densities were generated using the following formula:

$$f_{mix}(x,n) = \left(1 - \frac{c}{\sqrt{n}}\right) f_{exp}(x) + \frac{c}{\sqrt{n}} f_{alt}(x) \tag{1.79}$$

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 1.4. 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 evidently mimics the distribution of the supremum quite well — testing using bootstrapped critical values and using adjusted

Table 1.4: Empirical size and power for local alternatives described in the text. All tests are intended to have a size of 5%; e.g. the first row shows that the methods investigated in the text are more or less conservative in this experiment, while the bootstrap overrejects slightly. 50,000 repetitions.

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

critical values result in almost identical inference in this experiment, although the bootstrap overrejects slightly. It is also of interest to note that neither of the two "analytic" strategies — to test with either an adjusted critical value or a compensated process — is a uniformly better test. 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 tests using the parametric empirical process against the lognormal and gamma alternatives. Evidently these tests have differential performance against alternatives from different parts of the space of alternatives.

1.7 Conclusion

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

power, although the experiment suggests that tests using adjusting critical values perform very similarly to tests using bootstrapped critical values.

Chapter 2

Goodness of Fit Tests Using Kernel-Transformed Empirical Processes with Estimated Parameters

2.1 Introduction

Test statistics for goodness of fit are commonly based on a measure of the discrepancy between the empirical distribution function and the hypothesized parametric model. Beyond the popular metrics derived from supremum or L_2 norms, several authors have considered distance functions applied to certain transformations of the empirical and hypothesized distribution functions. Extensive simulation experiments in Gürtler and Henze (2000) provide evidence, for example, that tests for the Cauchy distribution based on the L_2 norm of the difference between empirical and hypothesized characteristic functions may have significantly better power against some alternatives than those based on sup-norm or L_2 norms applied directly to the standard empirical process $\sqrt{n}(\mathbb{F}_n - F)$. In some cases working with density or distribution functions is not feasible, but a relatively simple expression exists when one uses the characteristic function of the data. In these cases working with test statistics derived from transformed processes may be the most natural approach.

When parameters are estimated, tests based on empirical processes (transformed or not) generally have nonstandard limiting distributions, which is a real issue for the evaluation of tests using asymptotic critical values. For example, it is shown in Matsui and Takemura (2005) that the choice of one of two different estimators has a nontrivial impact on the limiting distribution of the transformed empirical process and test statistics derived from it. Recently, Meintanis and Swanepoel (2007) proposed a bootstrap procedure to evaluate tests based on weighted L_2 distances between data and theory. Matsui and Takemura (2005) propose an analytic method to test a subset of the hypotheses addressed in Meintanis and Swanepoel (2007).

In this paper we propose another analytic method for inference. We propose a compounded transformation of the parametric empirical process, first using the martingale transform of Khmaladze (1981) and then the kernel transform. The resulting empirical processes have tractable limiting distributions that are related to the kernel function used to transform the data. In particular, using kernels that corre-

spond to moment generating functions and characteristic functions results in processes that have those functions as their covariance functions. Properties of these kernels dictate some special properties of the processes, which are explored in examples. Section 2.2 outlines the theoretical background and the proposed testing methods.

2.2 Kernel-transformed empirical processes

Given a random sample from random variable $X \in \mathbb{R}$, consider the null hypothesis that the distribution function of X is well-described by a continuous parametric distribution function F specified up to a p-dimensional parameter θ:

$$H_0 : X \sim F \in \mathcal{F} = \{F(\cdot, \theta); \theta \in \Theta\}, \tag{2.1}$$

against the alternative that the data have a distribution function that is not a member of $\mathcal{F}$:

$$H_1 : X \sim F \notin \mathcal{F}. \tag{2.2}$$

When the null distribution F is fully specified (that is, when there are no unspecified parameters in the model), tests can be based on a functional of the uniform empirical process

$$V_n(x) = \sqrt{n}\left(\mathbb{F}_n(x) - F(x, \theta)\right). \tag{2.3}$$

It is well-known that such tests are asymptotically distribution free under the null hypothesis because V converges weakly to a time-changed Brownian bridge $B_F = B \circ F(x)$, where B is a standard Brownian bridge on the unit interval (van der Vaart, 1998, Ch. 19). When parameters of F are unknown, estimates of θ must be made and the resulting parametric empirical process depending on estimated parameters,

$$\hat{V}_n(x) = \sqrt{n}\left(\mathbb{F}_n(x) - F(x, \hat{\theta})\right), \tag{2.4}$$

is in general affected by the hypothesized parametric model, the value of the parameters, and the choice of the estimator $\hat{\theta}$.

The tests considered in this paper are based on the observed difference between transformations of the data and the hypothesized parametric family. A wide variety of transformations might be used; however, due to their popularity we focus on the moment generating function process and the characteristic function process. These transformed processes were analyzed first in full generality by Csörgő (1983).

They are very tractable because they can be expressed as linear operators applied to V_n or $\hat{V}_n$. Given a kernel function k, define a kernel-transformed empirical process Y_n on some compact $\mathcal{T} \subset \mathbb{R}$ by

$$Y_n(t) := \int_{-\infty}^{\infty} k(t,x)\mathrm{d}V_n(x); \quad t \in \mathcal{T}. \tag{2.5}$$

Define analogously $\hat{Y}_n$ a kernel-transformed empirical process with estimated parameters by

$$\hat{Y}_n(t) := \int_{-\infty}^{\infty} k(t,x)\mathrm{d}\hat{V}_n(x); \quad t \in \mathcal{T}. \tag{2.6}$$

Note $\mathcal{T}$ may depend on the parametric model and the kernel used in each case. Below we describe the behavior of such kernel-transformed processes.

We make the following assumptions:

A1: The members of $\mathcal{F}$ are absolutely continuous with respect to Lebesgue measure and mutually absolutely continuous. For each $F \in \mathcal{F}$, the vector of derivatives $\nabla_\theta F$ exists, and in a neighborhood of θ, the model has finite Fisher information and

$$\mathrm{E}\left[\left(\nabla_\theta \log f(x,\theta) - \nabla_\theta \log f(x,\theta')\right)^2\right] = o(1) \quad \text{as} \quad \theta' \to \theta. \tag{2.7}$$

A2: The estimator $\hat{\theta}$ satisfies $\sqrt{n}\left(\hat{\theta} - \theta\right) = O_P(1)$.

A3: Assume

$$\sup_{t\in\mathcal{T}} \int_{-\infty}^{\infty} |k(t,x)|^{2+\delta} \mathrm{d}F(x) < \infty \tag{2.8}$$

and for some $0 < \alpha \leq 1$, $\nu : \mathcal{T} \times \mathcal{T} \to \mathcal{T}$ for all $X \in \mathcal{X}$

$$|k(s,x) - k(t,x)| \leq |s-t|^\alpha M(x, \nu(s,t)) \tag{2.9}$$

such that

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}} M^2(\cdot,t)\right] < \infty. \tag{2.10}$$

Assumption **A3** ensures that k is well-behaved enough that a limiting process exists, and is a sufficient, but not necessary assumption for this result. For more general assumptions see Csörgő (1983, (i)-(iii)); **A3** corresponds to the stronger assumptions (i)* and (ii)* of Csörgő (1983).

The main theorem (p. 527) of Csörgő (1983) describes the convergence of the transformed parametric empirical process $\hat{Y}_n$ to a limit $\hat{Y}$ in the space of continuous functions on $\mathcal{T}$. It is an extension of Durbin (1973a), (see also Meintanis and Swanepoel (2007, Theorem 3.1)). Under some regularity conditions (similar to those described in **A1** through **A3**), Csörgő's Theorem implies that uniformly in $t \in T$, when an asymptotically efficient estimator exists,

$$\hat{Y}_n(t) = Y_n(t) + \sqrt{n}\left(\hat{\theta} - \theta\right)^\top \nabla_\theta \int_{\mathcal{X}} k(t,x)\mathrm{d}F(x,\theta) + o_P(1). \tag{2.11}$$

Csörgő's Theorem as it is stated above is left intentionally vague concerning the exact distribution of $\hat{Y}_n$, Y_n and the limit $\hat{Y}$ (which exists). As described above, the limiting process $\hat{Y}$ will in general depend upon properties of the estimator used. However, an application of the martingale transform of Khmaladze (1981) would annihilate the effect of the estimation of parameters in $\hat{V}_n$ before the application of the kernel transformation. This results in more tractable limiting processes for which we can derive asymptotic critical values for testing.

Khmaladze's martingale transform is constructed using the score function of the parametric model. Given any $\theta \in \Theta$, define the extended score $\dot{g} : \mathcal{X} \to \mathbb{R}^{p+1}$ by $\dot{g}(x,\theta) = [1, \nabla_\theta^\top \log f(x,\theta)]^\top$; this will often be shortened to $\dot{g}(x)$ when it will cause no confusion. Define the operator $M : L_2(\mathcal{X}) \to \mathbb{R}$ by

$$(M \circ \varphi)(x) = \int_{-\infty}^{x} \dot{g}^\top(y)\Gamma^{-1}(y) \int_y^\infty \dot{g}(z)\mathrm{d}\varphi(z)\mathrm{d}F(y) \tag{2.12}$$

with

$$\Gamma(y) = \int_y^\infty \dot{g}(z)\dot{g}^\top(z)\mathrm{d}F(z) \tag{2.13}$$

and define the compensated empirical process

$$\tilde{V}_n(x) = \sqrt{n}\left(\mathbb{F}_n(x) - M \circ \mathbb{F}_n(x)\right). \tag{2.14}$$

Under the null hypothesis, Khmaladze (1981) showed, under assumptions **A1** and **A2**, that $\tilde{V}_n$ is a martingale[1]. Furthermore, under the null hypothesis this martingale converges weakly to a time-changed Brownian motion $W_F(t) = W \circ F(t)$, where W is a standard Brownian motion on the unit interval and

[1] More specifically, it was shown that $M \circ \mathbb{F}_n$ is the compensator in a Doob-Meyer decomposition of $\mathbb{F}_n$ when $\mathbb{F}_n$ is interpreted as a submartingale on $\left\{\mathscr{F}_t^{\mathbb{F}_n,\hat{\theta}}\right\}$, the filtration generated by $\mathbb{F}_n$ and $\hat{\theta}$.

F is the distribution function of the data.

Consider one additional transformation of the data: a kernel transformation applied to the process $\tilde{V}_n$. This function can be constructed by using a differential form of the compensator:

$$\tilde{Y}_n(t) := \int_{-\infty}^{\infty} k(t,x)\mathrm{d}\tilde{V}_n(x) \tag{2.15}$$

$$= \sqrt{n}\left(\int_{-\infty}^{\infty} k(t,x)\mathrm{d}\mathbb{F}_n(x) - \int_{-\infty}^{\infty} k(t,x)\dot{g}^{\top}(x)\Gamma^{-1}(x)\int_{x}^{\infty}\dot{g}(y)\mathrm{d}\mathbb{F}_n(y)\mathrm{d}F(x)\right). \tag{2.16}$$

These doubly-transformed processes in $\mathcal{T}$ possess many convenient properties that make them useful for testing. Rather than considering an abstract class of kernels we consider here a smaller yet manageable and practically relevant class of kernels that encompass the moment generating and characteristic functions for the model $\mathcal{F}$. We refer below to the complex generating function for the family $\mathcal{F}$ by

$$\varphi_F(t) = \int_{\mathcal{X}} e^{zx}\mathrm{d}F(x,\theta), \quad z \in \mathbb{C} \tag{2.17}$$

with kernel function e^{zx}. This kernel corresponds to the kernels corresponding to moment generating- and characteristic functions of the family $\mathcal{F}$ by letting $z = t \in \mathcal{T}$ or $z = it, t \in \mathcal{T}$. Theorems below apply to both sorts of generating function, after tailoring $\mathcal{T}$ to fit the chosen kernel so that (2.8) applies. In the sequel, $\overline{z}$ denotes the complex conjugate of z. This family of kernel functions is chosen for the wide familiarity that moment generating functions and characteristic functions enjoy and because interest in inference using kernel-transformed empirical processes has until now been focused primarily on the two special cases of the empirical moment generating function- and empirical characteristic function processes.

Theorem 4. *Let k be the kernel corresponding to the complex generating function φ_F. Then $\tilde{Y}_n$ converges weakly in $L_2(\mathcal{T})$ to a limit $\tilde{Y}$. $\tilde{Y}$ is a mean-zero Gaussian process with covariance function*

$$\rho(s,t) = \mathrm{E}\left[\tilde{Y}(s)\overline{\tilde{Y}(t)}\right] = \varphi_F(s+\overline{t}) \tag{2.18}$$

Proof. Khmaladze (1981) shows that $\tilde{V} \rightsquigarrow W_F$, and therefore the continuous mapping theorem and **A3** imply that $\tilde{Y}_n \rightsquigarrow \tilde{Y}$ where

$$\tilde{Y}(t) = \int_{\mathcal{X}} k(t,x)\mathrm{d}W_F(x). \tag{2.19}$$

Because the function k is deterministic, standard stochastic calculus (Potthoff, 2008) implies that

$$\mathrm{E}\left[\tilde{Y}(t)\right] = \mathrm{E}\left[\int_0^1 k(t,x)\mathrm{d}W_F(r)\right] = 0 \quad \text{and} \quad \mathrm{E}\left[\tilde{Y}(s)\overline{\tilde{Y}(t)}\right] = \int_0^1 k(s,x)\overline{k(t,x)}\mathrm{d}F(x,\theta). \tag{2.20}$$

Finally, for the kernels considered here, use the definition of the complex generating function to arrive at (2.18). ■

Note that for the empirical characteristic function process this is quite an appealing result: the covariance function of the process $\tilde{Y}$ is $\rho(s,t) = \varphi_F(s-t)$, implying the process $\tilde{Y}$ is stationary.

Also, weighting functions (in $\mathcal{T}$) can easily be incorporated into the above theorem because integration is with respect to x rather than t. For example, incorporating the weight function $\beta(t)$ in the characteristic function process,

$$\beta(t)\int_{-\infty}^{\infty} e^{itx}\mathrm{d}\tilde{V}_n(x), \tag{2.21}$$

the weighted process has zero mean and covariance function $\rho(s,t) = \beta(s)\beta(t)\varphi_F(s-t)$. Note however that stationarity will be affected.

Examples in Gürtler and Henze (2000) and in Matsui and Takemura (2005) show that using different estimators normally complicates the testing procedure, because different distributions result when choosing between alternative estimators. This is not an issue when using $\tilde{Y}_n$ for inference because the processes are asymptotically distribution-free.

A wider set of kernel functions is important to consider in future research. In particular, specific kernel functions may be chosen/constructed to "focus" tests on specific alternatives, perhaps by analyzing the properties of tests through component decomposition (Stute, 1997; Milbrodt and Strasser, 1990).

Finally, there is also one particularly simple transformation that, while not corresponding to anything that necessarily has a familiar name, has a distribution free of the parametric model[2]. Kernel functions $k(z,x)$ of the transformations of the type considered above can be replaced with corresponding kernels $k(z,F(x,\theta))$. In this way, taking the empirical characteristic function process as an example, one can show that $\tilde{Y}$ is simply a time-changed version of the standard Brownian motion W:

$$\tilde{Y}_0(t) = \int_0^1 e^{itr}\mathrm{d}W(r). \tag{2.22}$$

Due to its construction this process is free of the distribution F. The properties of this process will be considered in future research.

[2]The author wishes to thank Juan Carlos Escanciano for this suggestion.

2.3 Supremum norms

There are several functionals to choose from when creating goodness of fit statistics based on $\tilde{Y}_n$. Of course, [weighted] quadratic statistics such as the Cramér-von Mises statistic are other available options, as well as bootstrap approximations to the correct critical values for testing. See for example Meintanis and Swanepoel (2007) and references cited therein. Because they have received less attention, we focus here on sup-norm test statistics. We propose analytic approximations for the calculation of appropriate critical values.

First we recall some definitions; for more detailed treatments, see Azaïs and Wschebor (2009); Adler (1981) or Loève (1978). A random process x is continuous in quadratic mean at t if

$$\lim_{s \to t} x(s) \tag{2.23}$$

converges in quadratic mean to $x(t)$. A process is continuous in quadratic mean if and only if its covariance function is continuous in both of its arguments. A process is differentiable in quadratic mean at t if

$$\lim_{s \to t} \frac{x(s) - x(t)}{s - t} \tag{2.24}$$

converges in quadratic mean to a limit that we denote $x'(t)$. If a process x has a covariance function ρ that is differentiable with respect to each argument, then it is differentiable in quadratic mean and $\frac{\partial}{\partial s}\rho(s,t) = \mathrm{Cov}\left(x'(s), x(t)\right)$. If $\frac{\partial^2}{\partial s \partial t}\rho(s,t)$ exists, then it is equal to $\mathrm{Cov}\left(x'(s), x'(t)\right)$.

We note some convenient features of the process $\tilde{Y}$ in the following Theorem. Transformed processes using the kernels considered here can be relatively smooth compared to $\tilde{V}$. Under certain conditions they even have differentiable sample paths.

Theorem 5. *When the generating function φ_F is continuous for all $t \in \mathcal{T}$, $\tilde{Y}$ is continuous in quadratic mean and a version of $\tilde{Y}$ exists with continuous sample paths. If φ_F is differentiable with respect to t, then $\tilde{Y}$ has a derivative in quadratic mean and admits a version of $\tilde{Y}$ with differentiable sample paths; this function coincides with the derivative of $\tilde{Y}$ in quadratic mean.*

Proof. Continuity of $\varphi_F(z)$ implies continuity of both the moment generating function and the characteristic function, which in turn implies continuity in quadratic mean of $\tilde{Y}$ (Adler, 1981, Theorem 2.2.1). Azaïs and Wschebor (2009, pages 23-25) show that Itō integrals admit a version with continuous sample paths, and kernel-transformed processes may be considered a special case of this class, namely as Itō integrals with deterministic integrands.

If the complex generating function of $\mathcal{F}$ is differentiable, this implies differentiability of the covariance function of $\tilde{Y}$ and therefore differentiability in quadratic mean for $\tilde{Y}$. That this derivative is the same as a pathwise derivative can be found in Potthoff (2008, Theorem 3.2). ■

2.3.1 Critical values for sup-norm tests

When $\tilde{Y}$ is smooth, we propose the use of a Rice formula (Azaïs and Wschebor, 2009; Adler, 2000) to find approximate critical values for tests based on $\tilde{Y}_n$. The Rice formula links the number of level crossings of a stochastic process to the distribution of the supremum of the process. Define the number of upcrossings of $\tilde{Y}$ of a level c as

$$U_c := \#\left\{t \in T : \tilde{Y}(t) = c,\ \tilde{Y}'(t) > 0\right\} \tag{2.25}$$

and the number of downcrossings by

$$D_c := \#\left\{t \in T : \tilde{Y}(t) = c,\ \tilde{Y}'(t) < 0\right\} \tag{2.26}$$

Note that the complex generating function $\varphi_F(0) = 1$ for any $\mathcal{F}$, which implies that $\mathrm{P}\left\{\tilde{Y}(0) = 0\right\} = 1$ for any $\mathcal{F}$. Then from elementary considerations we have the one-sided bound

$$\mathrm{P}\left\{\sup_{t\in T} \tilde{Y}(t) > c\right\} = \mathrm{P}\left\{U_c \geq 1\right\} \leq \mathrm{E}\left[U_c\right] \tag{2.27}$$

and the two-sided bound

$$\mathrm{P}\left\{\sup_{t\in T} \left|\tilde{Y}(t)\right| > c\right\} = \mathrm{P}\left\{U_c \geq 1\right\} + \mathrm{P}\left\{D_{-c} \geq 1\right\} \leq \mathrm{E}\left[U_c\right] + \mathrm{E}\left[D_{-c}\right]. \tag{2.28}$$

These inequalities should provide roughly accurate critical values[3] for one- and two-sided testing situations. By taking φ_F as the kernel corresponding to the empirical moment generating function process, one finds $\rho(s,t) = \varphi_F(s+t)$, so these processes are differentiable whenever φ_F is differentiable.

On the other hand, the calculation of the expectations on the right-hand side of is rather convenient. The Rice formula[4] provides a method to calculate the required expectations. The general Rice formulas

[3] A two-term expansion would be better, and would provide conservative approximations from a size standpoint. Unfortunately, a two-term expansion is much more difficult to calculate than the one-term expansion.

[4] Rice appears to have extended a theorem of Kac for counting the crossings of a level by a nonstochastic function.

required are

$$\mathrm{E}\left[U_c\right] = \int_{\mathcal{T}} \mathrm{E}\left[(\tilde{Y}'(t))^+ \middle| \tilde{Y}(t) = c\right] p(c,t)\mathrm{d}t \tag{2.29}$$

$$\mathrm{E}\left[D_c\right] = \int_{\mathcal{T}} \mathrm{E}\left[(\tilde{Y}'(t))^- \middle| \tilde{Y}(t) = c\right] p(c,t)\mathrm{d}t \tag{2.30}$$

where $X^+ = X \vee 0$, $X^- = -(X \wedge 0)$ and p is the marginal density of $\tilde{Y}(t)$ evaluated at c:

$$p(c,t) = \frac{\exp\left\{\frac{-c^2}{2\rho(t,t)}\right\}}{\sqrt{2\pi\rho(t,t)}}. \tag{2.31}$$

It is possible to specify this formula more exactly using properties of the Gaussian distribution. We use the fact that when $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathrm{E}\left[X^+\right] = \mu\Phi(\mu/\sigma) + \sigma\phi(\mu/\sigma) \quad \text{and} \quad \mathrm{E}\left[X^-\right] = \mu\Phi(\mu/\sigma) - \mu + \sigma\phi(\mu/\sigma). \tag{2.32}$$

From these expressions, the appropriate formulas for the number of upcrossings (of a level $c > 0$ and downcrossings of a level $b < 0$ are

$$\mathrm{E}\left[U_c\right] = \int_{\mathcal{T}} \left(cM(t)\Phi\left(c\frac{M(t)}{\Sigma(t)}\right) + \Sigma(t)\phi\left(c\frac{M(t)}{\Sigma(t)}\right)\right) p(c,t)\mathrm{d}t \tag{2.33}$$

and

$$\mathrm{E}\left[D_b\right] = \int_{\mathcal{T}} \left(bM(t)\Phi\left(b\frac{M(t)}{\Sigma(t)}\right) - bM(t) + \Sigma(t)\phi\left(b\frac{M(t)}{\Sigma(t)}\right)\right) p(b,t)\mathrm{d}t \tag{2.34}$$

where a regression formula implies

$$M(t) = \frac{\mathrm{Cov}\left(\tilde{Y}(t), \tilde{Y}'(t)\right)}{\mathrm{Var}\left(\tilde{Y}(t)\right)} \tag{2.35}$$

$$\Sigma^2(t) = \mathrm{Var}\left(\tilde{Y}'(t)\right) - \frac{\mathrm{Cov}^2(\tilde{Y}(t), \tilde{Y}'(t))}{\mathrm{Var}\left(\tilde{Y}(t)\right)} \tag{2.36}$$

(note this is the expression for Σ^2, not Σ). For the approximation (2.28) one could alternatively use expressions for the absolute value of $\tilde{Y}'$ because $(\tilde{Y}')^+ + (\tilde{Y}')^- = \left|\tilde{Y}'\right|$. In this case the appropriate formula for $X \sim \mathcal{N}(\mu, \sigma^2)$ is $\mathrm{E}\left[|X|\right] = 2\mu\Phi(\mu/\sigma) - \mu + 2\sigma\phi(\mu/\sigma)$.

Taking φ_F as the kernel that makes $\tilde{Y}$ the empirical characteristic function process, we see that in

some cases the limit in (2.37) is zero and in others it is nonzero — smoothness of the process will depend on the $\mathcal{F}$ in question. When the process is not smooth, we suggest critical values based on an approximation suggested by Durbin (1985).

When $\tilde{Y}$ is nondifferentiable we assume that

$$\lim_{s \nearrow t} \frac{\operatorname{Var}\left(\tilde{Y}(t) - \tilde{Y}(s)\right)}{t - s} = \sigma^2(t), \quad 0 < \sigma^2(t) < \infty \tag{2.37}$$

for all $t \in \mathcal{T}$ and denote such $\tilde{Y}$ as locally Brownian (Durbin, 1985, p. 100). Note that condition (2.37) is equivalent to the condition that, for ρ the covariance function of $\tilde{Y}$,

$$\lim_{s \nearrow t} \left(\frac{\partial}{\partial s}\rho(s,t) - \frac{\partial}{\partial t}\rho(s,t) \right) = \sigma^2(t). \tag{2.38}$$

Durbin (1985) proposed a method to approximate critical values by using boundary crossing probabilities for locally Brownian processes. The proposal[5] used here is

$$\mathrm{P}\left\{ \sup_{t \in \mathcal{T}} \tilde{Y}(t) > c \right\} \approx \int_{\mathcal{T}} c \frac{\frac{\partial}{\partial s}\rho(s,t)|_{s=t}}{\rho(t,t)} p(c,t)\mathrm{d}t \tag{2.39}$$

where $p(c,t)$ is defined in (2.31). These critical values are easy to compute and accurate approximations in a wide variety of situations (Rabinowitz, 1993). For two-sided tests one can use the approximation $\mathrm{P}\left\{\sup_{t\in\mathcal{T}} \left|\tilde{Y}(t)\right| > c\right\} \approx 2\mathrm{P}\left\{\sup_{t\in\mathcal{T}} \tilde{Y}(t) > c\right\}$, accurate for large c (i.e., for tests with small size α).

2.4 Examples and applications

The examples below illustrate how the methods described above may be applied to simple goodness of fit tests for parametric models. Cauchy and exponential distributions are discussed, using respectively the characteristic function and moment generating function.

[5] In some ways this looks very similar to the Rice formula, because when $\tilde{Y}$ is differentiable, $\operatorname{Cov}\left(\tilde{Y}(t), \tilde{Y}'(t)\right) = \frac{\partial}{\partial s}\rho(s,t)|_{s=t}$. However, the considerations that led Durbin to (2.39) are substantially different from those that led Rice to propose his eponymous formula.

2.4.1 Example: testing for Cauchy and stable models with the characteristic function process

To test the hypothesis that the sample can be described by the location-scale Cauchy model,

$$H_0 : F \in \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{\cdot - \alpha}{\beta} \right) ; \alpha \in \mathbb{R}, \beta > 0 \right\} \tag{2.40}$$

one can use the characteristic function transform. The extended score function $\dot{g} : \mathbb{R} \to \mathbb{R}^3$, needed to build the compensator for the Cauchy location-scale model is

$$\dot{g}(x, \theta) = \begin{bmatrix} 1 \\ \frac{2(x-\alpha)}{\beta^2 + (x-\alpha)^2} \\ \frac{1}{\beta} - \frac{2\beta}{\beta^2 + (x-\alpha)^2} \end{bmatrix} \tag{2.41}$$

which can be used to calculate the compensator numerically.

The characteristic function transform is equivalent to using a transform with kernel $k(t, x) = e^{itx}$, letting $\mathcal{T} = \mathbb{R}$. In this case, the weighted process $\tilde{Y}_n$ may be expressed as

$$\tilde{Y}_n(t) = \beta(t)\sqrt{n} \left(\frac{1}{n} \sum_{j=1}^{n} e^{itX_j} - \tilde{K}_n(t) \right), \tag{2.42}$$

where $\tilde{K}_n$ is computed numerically using a discretization of the expression

$$\int_{-\infty}^{\infty} e^{itx} \dot{g}^{\top}(x) \Gamma^{-1}(x) \int_{x}^{\infty} \dot{g}(y) \mathrm{d}\mathbb{F}_n(y) \mathrm{d}F(x). \tag{2.43}$$

Details regarding computation can be found in Chapter 1. $\tilde{Y}_n$ converges weakly to the complex-valued Gaussian process $\tilde{Y}$, a mean-zero Gaussian process with covariance function

$$\mathrm{E}\left[\tilde{Y}(s) \overline{\tilde{Y}(t)} \right] = \beta(s)\beta(t) e^{-|t-s|}. \tag{2.44}$$

Note that with weight function $\beta(t) \equiv 1$ the function has the same distribution as a complex-valued Ornstein-Uhlenbeck process. Using the weight function from Matsui and Takemura (2005), $\beta(t) = e^{-a|t|}$, $a > 0$, we have the covariance function

$$\rho(s, t) = e^{-a(|s|+|t|)-|s-t|} \tag{2.45}$$

which is differentiable for all $s, t \neq 0$ and $s \neq t$; the fact that it is not differentiable everywhere implies that the function is continuous, but not differentiable in mean square.

The maximum modulus of the process $\tilde{Y}$ with a constant weight (set equal to 1 here) is related to the maximum of a Bessel process. The complex-valued Ornstein-Uhlenbeck process X is the solution to the stochastic differential equation

$$\begin{bmatrix} \mathrm{d}X_1(t) \\ \mathrm{d}X_2(t) \end{bmatrix} = \begin{bmatrix} -\lambda & -\omega \\ \omega & -\lambda \end{bmatrix} \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} + \begin{bmatrix} \mathrm{d}W_1(t) \\ \mathrm{d}W_2(t) \end{bmatrix}, \quad t \geq 0;\ \lambda > 0,\ \omega \in \mathbb{R} \tag{2.46}$$

where $(X_1, X_2) = (\mathrm{Re}\, X, \mathrm{Im}\, X)$, and $W = W_1 + iW_2$ is a standard complex-valued Wiener process. It can be calculated that its covariance function is

$$\rho(s,t) = \mathrm{E}\left[X(s)\overline{X(t)}\right] = \frac{1}{\lambda} e^{-(\lambda - \omega i)|s-t|} \tag{2.47}$$

and this covariance function is triangular — that is, for $r \leq s \leq t$,

$$\rho(r,t) = \frac{\rho(r,s)\rho(s,t)}{\rho(s,s)} \tag{2.48}$$

which means that X is a Gauss-Markov process. The covariance function of the Gaussian process $\tilde{Y}$ defined above is equal to the above definition with $\lambda = 1$ and $\omega = 0$. Therefore $\tilde{Y}$ is identical in distribution with the Ornstein-Uhlenbeck process.

It is also known that the Ornstein-Uhlenbeck process is related to Brownian motion in the following simple manner[6]

$$\tilde{Y}(t) = e^{-t/2} W(e^t) \tag{2.49}$$

where W is a standard complex-valued Brownian motion. The maximum modulus of a standard complex-valued Brownian motion is particularly tractable. This is because of the simple characteristics of the covariance function for the process: again using the notation $W = W_1 + iW_2$, where W_1 and W_2 are independent standard Brownian motions,

$$\mathrm{E}\left[W(s)\overline{W(t)}\right] = \mathrm{E}\left[W_1(s)W_1(t)\right] + \mathrm{E}\left[W_2(s)W_2(t)\right] = 2(s \wedge t). \tag{2.50}$$

Noting that $\mathrm{E}\left[|W(t)|\right] = \mathrm{E}\left[W(t)\overline{W(t)}\right]$, it can be seen that the maximum modulus is directly related to

[6] This is a consequence of the fact that the process is a Gauss-Markov process

the distribution of a two-dimensional Brownian motion composed of independent coordinate processes. The probability that such a Brownian motion leaves the region $A = \{z \in \mathbb{C} : |z| \leq a\}$ is described by the distribution of the supremum of a Bessel process. This implies that the maximum modulus of $\tilde{Y}$ can be calculated directly from some transformations of variables, because the first hitting time densities of Brownian motion and Gauss-Markov properties are linked (Doob, 1953; Mehr and McFadden, 1965; Di Nardo et al., 2001).

Stable models

The Cauchy model is one special case of a richer class of distributions, the symmetric stable model. Such models have characteristic functions

$$\varphi_F \in \left\{ e^{i\mu t - |\sigma t|^\alpha}; t \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0, \alpha \in (0, 2] \right\}. \tag{2.51}$$

This model nests the Cauchy ($\alpha = 1$) and normal ($\alpha = 2$) models. For this model, the limiting process $\tilde{Y}$ is stationary, but the limiting ECF process has the covariance function

$$\rho(s,t) = e^{|s-t|^\alpha}; \quad s, t \in \mathbb{R},\ \alpha \in (0, 2] \tag{2.52}$$

which is not a triangular function unless $\alpha = 1$, implying that except in that case the process $\tilde{Y}$ is non-Markovian. Matsui and Takemura (2008) have expressions for the covariance function of the parametric empirical process for two different $\sqrt{n}$-consistent estimators of the parameter α. Critical values for tests based on these limiting processes will necessarily rely on an estimate $\hat{\alpha}$. However, because the distribution and characteristic functions are bounded as functions of α, the divergence of $\tilde{V}$ will cause the test to be consistent (i.e., $\tilde{V}$ will diverge, resulting in large values for $\left|\tilde{Y}\right|$, beyond the bounded critical values coming from approximation (2.39)).

2.4.2 Example: testing exponentiality with the moment generating function process

The exponential model is important baseline model in several fields including survival analysis, and its simple moment generating function makes it a tractable example. The hypothesis that the data can be described using the exponential model

$$H_0 : F \in \left\{ 1 - e^{-\lambda \cdot}; \lambda > 0 \right\} \tag{2.53}$$

can be tested using a moment generating function transformation. The weight function $\beta(t) = e^{-a|t|}$, $a > 0$ is chosen again out of analytical convenience. For stochastic integrals to be well-defined, it is required that $t < \lambda/2$. So as not to let the estimate of λ affect the domain of the process, we set $\mathcal{T} = \{t : t < 0\}$, which also simplifies the weight function to $\beta(t) = e^{at}$. The compensator for the exponential model uses the extended score function $\dot{g}$ defined by

$$\dot{g}(x, \lambda) = \begin{bmatrix} 1 \\ \frac{1}{\lambda} - x \end{bmatrix}. \tag{2.54}$$

The proposed test statistic is the supremum of the process

$$\tilde{Y}_n(t) = \beta(t)\sqrt{n}\left(\mathbb{K}_n(t) - \tilde{K}(t)\right) = \beta(t)\sqrt{n}\left(\int_0^\infty e^{tx}\mathrm{d}\mathbb{F}_n(x) - \int_0^\infty e^{tx}\mathrm{d}(M \circ \mathbb{F}_n)(x)\right) \tag{2.55}$$

$$= \beta(t)\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^n e^{tX_j} + \left(\frac{2\hat{\lambda}}{t} - \frac{\hat{\lambda}^2}{t^2}\right)\left(1 - \frac{1}{n}\sum_{j=1}^n e^{tX_j}\right) - \frac{\hat{\lambda}^2}{t}\bar{X}\right). \tag{2.56}$$

When $\hat{\lambda} = \hat{\lambda}_{mle}$, it can be verified[7] that

$$\lim_{t\to 0}\tilde{Y}_n(t) = \frac{\hat{\lambda}^2}{2}\frac{1}{n}\sum_{j=1}^n X_j^2 - 1 \xrightarrow{p} 0. \tag{2.57}$$

Theorem 4 implies that $\tilde{Y}_n$ converges weakly to a mean-zero Gaussian process on $\mathcal{T}$ with covariance function

$$\rho(s, t) = \frac{\lambda e^{a(s+t)}}{\lambda - s - t}, \quad s, t < 0. \tag{2.58}$$

The sample paths of the limiting process $\tilde{Y}$ are infinitely differentiable (w.p. 1) with respect to t and the first derivative $\tilde{Y}$ is

$$\tilde{Y}'(t) = e^{at}\int_0^\infty (a + x)e^{tx}\mathrm{d}W_F(x), \quad t < 0 \tag{2.59}$$

which is a mean-zero Gaussian process with

$$\mathrm{Cov}\left(\tilde{Y}(s), \tilde{Y}'(t)\right) = \rho(s, t)\left(a + \frac{1}{\lambda - s - t}\right) \tag{2.60}$$

[7] It can also be verified that $\hat{Y}_n(t)$ and $\tilde{Y}_n(t)$ have the same roots.

and

$$\operatorname{Cov}\left(\tilde{Y}'(s), \tilde{Y}'(t)\right) = \rho(s,t)\left(a^2 + \frac{2a}{\lambda - s - t} + \frac{2}{(\lambda - s - t)^2}\right). \tag{2.61}$$

Approximate critical values for sup-norm tests using this process can be obtained using the Rice formula described in Subsection 2.3.1. In order to make the approximation, the expectation $\mathrm{E}\left[U_c\right]$ was calculated using the formula (2.33), noting that for this example

$$M(t) = a + \frac{1}{\lambda - 2t} \quad \text{and} \quad \Sigma^2(t) = \frac{\rho(t,t)}{\lambda - 2t}. \tag{2.62}$$

To derive approximate critical values for a (two-sided) level-α test, one can use these functions to solve the Rice formula

$$\alpha = \int_{\mathcal{T}} \left(2cM(t)\Phi\left(c\frac{M(t)}{\Sigma(t)}\right) - cM(t) + 2\Sigma(t)\phi\left(c\frac{M(t)}{\Sigma(t)}\right)\right) p(c,t)\mathrm{d}t \tag{2.63}$$

for c.

2.5 Application: Affine asset pricing models and discretely observed diffusions

The process $\tilde{Y}$ also has application in the evaluation of tests of fit for financial data. When observations are treated as realizations of the return diffusion sampled at discrete points in time, their distribution is most easily described using the characteristic function. Singleton (2001) describes affine asset pricing models and their estimation via conditional characteristic function. These conditional characteristic functions may also be used to test the adequacy of affine asset pricing models as a description of the data. For example, take a simple Cox-Ingersoll-Ross or square-root diffusion model (Cox et al., 1985) for some returns process:

$$\mathrm{d}r_t = \kappa(\theta - r_t)\mathrm{d}t + \sigma\sqrt{r}\mathrm{d}W_t \tag{2.64}$$

where W is a standard Brownian motion. Then because the conditional distribution is a non-central χ^2 distribution, the conditional characteristic function of returns of $r_{t+\Delta}$ given r_t is

$$\varphi_r(u|r_t) = \left(1 - \frac{iu}{c}\right)^{\frac{-2\kappa\theta}{\sigma^2}} \exp\left\{\frac{iue^{-\kappa\Delta}r_t}{1 - \frac{iu}{c}}\right\}, \quad c = \frac{2\kappa}{\sigma^2(1 - e^{-\kappa\Delta})} \tag{2.65}$$

(Singleton, 2001, p. 120).

Theorem 4 implies that $\tilde{Y}$ is a mean-zero Gaussian process with complex-valued covariance function given by (2.65). Finding the extended score function with which to calculate the compensated empirical process is tedious, but Bai (2003) shows that when one replaces the score function g with a uniformly consistent estimator $\hat{g}$, the limiting distribution of $\tilde{V}$, and thus $\tilde{Y}$, is unaffected.

2.6 Conclusion

Kernel transformations may be used in goodness of fit tests to improve the power of a testing procedure against some subspace of alternatives, or when it is difficult to work with densities. However, the estimation of the parameters of the null model causes complications in the limiting distributions of test statistics. Kernel transforms of the compensated empirical process result in processes that are convenient when kernels correspond to the moment generating function or characteristic function of the null model because the covariance functions of these processes are the characteristic function or moment generating function of the hypothesized model in the test.

Chapter 3

New Tests for Stochastic Dominance Via Gaussian Field Approximations

3.1 Introduction

The ordering of distributions via stochastic dominance has received considerable attention in the econometrics literature. McFadden (1989) and Davidson and Duclos (2000) propose basic methodology and provide economic motivation for inference in a variety of situations. Most recent articles propose simulation strategies for the evaluation of tests. These strategies are attractive because they allow the researcher to make few assumptions about dependence between populations and avoid the difficult limiting distributions that such statistics can have, thereby making tests applicable in a wide variety of situations. However, simulation can be time-consuming and may require the imposition of a null hypothesis that may be difficult to achieve. In some simple modeling situations, the empirical processes used for testing may have tractable limiting distributions that can be used to construct consistent tests. We propose tests based on approximate asymptotic critical values derived using a Rice formula. They are easy to compute and account for the distribution of the data, an issue that generally makes the calculation of critical values for stochastic dominance difficult.

There is a growing literature dealing with the technical issues surrounding tests for stochastic dominance. For example, Barrett and Donald (2003) and Linton et al. (2005) extend prior results to provide consistent tests of any order of dominance. The papers of Klecan et al. (1991) and Linton et al. (2005) extend the tests to an arbitrary number of samples that may be dependent upon one another. Horváth et al. (2006) and Linton et al. (2010) propose improvements to tests that improve the size and power of tests under general conditions. Linton et al. (2005) also deals with tests constructed from the residuals of conditional models and shows that the distribution of tests is affected by parameter estimation.

The present work differs from the existing literature by providing a detailed description of the family of processes used for testing stochastic dominance of any order and suggesting analytical techniques for inference that arise from the properties of these processes. It is shown that by using the properties of the integrated empirical processes commonly used for testing, very accurate approximate critical values

for supremum-norm tests can be derived that do not rely on simulation techniques. Furthermore, these critical values can be found for tests of stochastic dominance at any order. Results are divided into two parts: two-sample situations in which covariates are not used to condition the samples used in testing, and conditional model situations in which the samples used for testing are themselves residual values from regression models. It is shown that the integrated processes arising from residual empirical processes have an analogous relationship with integrated two-sample (i.e., unconditional) processes as do parametric and uniform empirical processes in one-sample tests (cf. (Durbin, 1973a)). The different character of these families of processes has not previously been explored in part because simulation techniques are used for inference. Simulations show that the two-sample tests proposed here perform well in relation to simulation methods proposed in other recent articles.

In the next Section the basic methodology is introduced. Sections 3.3 and 3.4 introduce proposals for the estimation of asymptotic critical values for tests of stochastic dominance on any order, in two-sample problems without any parameter estimation (Section 3.3) and when the data are themselves residuals from conditional models (Section 3.4). Section 3.5 investigates the performance of testing methods in small simulation experiments and compares their performance against simulation techniques proposed by Barrett and Donald (2003) and Linton et al. (2005).

3.2 Tests of j^{th} order stochastic dominance

Suppose F is a cumulative distribution function defined on $T \subseteq \mathbb{R}$, and define, for $j \geq 1$, the family of functionals

$$\mathcal{I}_j(t,F) = \begin{cases} F(t) & j = 1 \\ \int_{-\infty}^{t} \mathcal{I}_{j-1}(s,F)\mathrm{d}s & j = 2,3,\ldots \end{cases} \tag{3.1}$$

Note that we can also write, via integration by parts,

$$\mathcal{I}_j(t,F) = \int_{-\infty}^{t} \frac{(t-s)^{j-1}}{(j-1)!}\mathrm{d}F(s). \tag{3.2}$$

Suppose we have two real-valued random variables, X_1 and X_2 with distribution functions F_1 and F_2, and consider the following family of null hypotheses, indexed by order j: the hypothesis that X_2 dominates X_1 stochastically (at the j^{th} order) — this is equivalent to the formal hypotheses

$$H_0^j : \mathcal{I}_j(t,F_2) \leq \mathcal{I}_j(t,F_1) \quad \text{for all } t \in T$$

which are to be compared to the family (also indexed by j) of alternative hypotheses, that X_2 does not dominate X_1 stochastically:

$$H_1^j : \mathcal{I}_j(t, F_2) > \mathcal{I}_j(t, F_1) \quad \text{for some } t \in T.$$

Davidson and Duclos (2000) thoroughly discuss the motivations for testing the above hypotheses. The family of transformations (3.1) may be used to test stochastic dominance of any order $j = 1, 2, \ldots$ using empirical processes as follows. Let $\mathbb{F}_{1n}$ and $\mathbb{F}_{2m}$ be empirical distribution functions constructed from samples of size n and m from each population. Following Barrett and Donald (2003), we always assume below that $\lim_{m,n\to\infty} \frac{m}{n+m} = \lambda \in (0, 1)$. Define the process (which we refer to hereafter as the j^{th}-order process)

$$V_{jmn}(t) = \sqrt{\frac{nm}{n+m}} \left\{ \mathcal{I}_j(t, \mathbb{F}_{2m}) - \mathcal{I}_j(t, \mathbb{F}_{1n}) \right\}, \quad t \in T. \tag{3.3}$$

Convenient sup-norm test statistics for tests of H_0^j are of the form $\sup_{t\in T} V_{jmn}(t)$, and H_0^j is rejected when this statistic is greater than some critical value. It is well known that under the null hypothesis, $\sqrt{n}(\mathbb{F}_{kn} - F_k) \rightsquigarrow B_{F_k}$, $k = 1, 2$, where B_{F_k} are time-changed Brownian bridges; that is, satisfying $B_{F_k}(t) = B(F_k(t))$ and B is a standard Brownian bridge process. We further simplify the theoretical situation below by using the "least favorable" case — that is, the boundary case in which it would be hardest to reject the null hypothesis — for which $F_1 \equiv F_2$. Theorem 6 describes the behavior of V_{jmn} as $m, n \to \infty$, which determines the distribution of the supremum of this process. This Theorem collects together Lemma 1 of Barrett and Donald (2003) and Theorem 1 of Horváth et al. (2006); the focus of the present work is on the construction of asymptotic critical values, but Theorem 6 is the foundation of much of the theory guiding these constructions and therefore it is presented here for completeness.

For a nondegenerate limit (at any order $j \geq 1$), Horváth et al. (2006) note that it is sufficient to assume that

$$\int_{\mathbb{R}} (1 + t_-^{j-2}) \sqrt{F_k(t)(1 - F_k(t))} \mathrm{d}t < \infty, \quad k = 1, 2, \tag{3.4}$$

which we maintain everywhere below. We use "$\rightsquigarrow$" to denote weak convergence in the space of bounded functions on T with the supremum metric.

Theorem 6. *Suppose X_1 and X_2 are independent random variables with common cumulative distribution function $F_1 \equiv F_2 \equiv F$. Suppose j is 1 or 2. Then as $n, m \to \infty$,*

$$V_{jmn} \rightsquigarrow V_j \tag{3.5}$$

where V_j is a mean-zero Gaussian process with distribution identical to that of $\mathcal{I}_j(\cdot, B_F)$. Suppose $j \geq 3$ and assume there exists a weight function $w : \mathbb{R} \to [0, \infty)$ such that

$$\sup_{t \in \mathbb{R}} w(t)(1 + t_+)^{j-2} < \infty. \tag{3.6}$$

Then as $n, m \to \infty$,

$$w(\cdot)V_{jmn} \rightsquigarrow w(\cdot)V_j. \tag{3.7}$$

The linearity of the integral operator (inductively) implies many convenient properties for the limiting processes used in the construction of test statistics. Because, given (3.4), which ensures that limits exist, the process B_F is Gaussian and almost surely continuous, we have immediately (inductively) that the family of limiting processes $\{V_j\}_j$ are Gaussian processes (because Gaussian distributions are stable under passage to limits, (Loève, 1978)) and are absolutely continuous by definition. Barrett and Donald (2003) show that they are mean-zero and have covariance functions (indexed by order j)

$$\rho_j(s, t, F) = \sum_{\ell=0}^{j-1} \binom{2j-\ell-2}{j-1} \frac{|t-s|^\ell}{\ell!} \mathcal{I}_{2j-\ell-1}(s \wedge t, F) - \mathcal{I}_j(s, F)\mathcal{I}_j(t, F) \tag{3.8}$$

without weight function, and $w(s)w(t)\rho_j(t, t, F)$ when the process is weighted. We make one technical assumption below: we assume T, the sample space of t, is compact. Horváth et al. (2006) show that the supremum distributions of third- and higher-order tests are unbounded when T is unbounded and the process is not weighted, but we avoid issues of noncompactness for simplicity's sake, assuming that in practice researchers effectively truncate the empirical processes at the extreme observations (from the pooled samples).

Barrett and Donald (2003) also show that under H_0^j, when no parameters have been estimated,

$$\lim_{m,n \to \infty} \mathrm{P}\left\{ \sup_{t \in T} V_{jmn}(t) > c \right\} \leq \mathrm{P}\left\{ \sup_{t \in T} V_j(t) > c \right\} \tag{3.9}$$

for $c > 0$, with equality when $F_1 \equiv F_2$. Therefore asymptotic critical values for a test of size α can be found by finding an appropriate value c using the right side of (3.9). On the other hand, Linton et al. (2005) show that when the random variables in question are themselves residuals from conditional models,

$$\lim_{m,n \to \infty} \mathrm{P}\left\{ \sup_{t \in T} \hat{V}_{jmn}(t) > c \right\} \leq \mathrm{P}\left\{ \sup_{t \in T} \hat{V}_j > c \right\} \tag{3.10}$$

where $\hat{V}_j$ is a Gaussian process with a distribution related to, but not identical to V_j — namely, the

distribution of $\hat{V}_j$ is affected by the estimation of the parameters of the model assumed by the researcher. In this paper we investigate the circumstances under which one can determine the distribution of the suprema of processes such as V_j and $\hat{V}_j$ analytically, that is, without resorting to simulation techniques. We will deal with a simplified situation of independent samples and basic conditional models — for more complex semiparametric models and unmodeled dependence structures, simulation is probably still a preferable testing strategy.

McFadden (1989) shows that S_{1mn} has a distribution that is identical to that in the standard Kolmogorov-Smirnov test, for which critical values are readily available; however, for orders $j \geq 2$, and when model parameters have been estimated (for any order $j \geq 1$), the limiting processes have nonstandard distributions. Typically critical values have been simulated in some way to perform tests. In Section 3.3 a method is proposed to find asymptotic critical values for supremum-norm tests of stochastic dominance using features of the limiting processes V_j. The results rely theory of the distribution of the maximum of Gaussian processes and fields as developed in Azaïs and Wschebor (2009) or Adler (2000). This derivation is an alternative to the simulation methods used up to this point in the literature. The methods proposed here are flexible enough to accommodate tests of any order of dominance and tests based on residuals from conditional models, and approximate critical values using features of $\hat{V}_j$ are proposed in Section 3.4.

3.3 Two-sample tests

Unlike the $j = 1$ case, which is asymptotically distribution free, features of the distribution of the data affect integrated processes of order 2 and higher. However, typically researchers do not wish to specify a parametric family that governs each distribution — this is very restrictive and implies that a test of dominance could often be made through some comparison of parameters. In this Section we propose analytical methods to derive critical values for tests even though the distribution functions of the data are not known. Attention is focused on the $j \geq 2$ case, since it has already been mentioned above that the case of $j = 1$ can be tested using critical values from the standard Kolmogorov-Smirnov test. We apply a Rice formula that approximates the distribution function of the supremum, using consistent estimates of the quantities that depend on the shape of F.

The properties of the limiting process $V_j := \mathcal{I}_j(\cdot, B_F)$ that are expressed in its covariance function ρ_j are the source of all our asymptotic results. For $j \geq 2$, ρ_j is continuous and differentiable (see Theorem 12 in Appendix C for a proof of differentiability), implying that all V_j, $j \geq 2$ are continuous

and differentiable in mean square[1]. Furthermore this implies that each V_j has $(j-1)$-times-differentiable sample paths with $\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell} V_j(t) \stackrel{D}{=} V_{j-\ell}(t),\ \ell \leq (j-1)$, where " $\stackrel{D}{=}$ " means equality in distribution. Sample path differentiability is also apparent from the definition of V_j as the integral of V_{j-1}.

The sample path differentiability of V_j suggests that a method that takes advantage of this smoothness might be an effective means with which to derive asymptotic critical values. Below we propose a Rice formula to make approximations from bounds on level crossing probabilities. The Rice formula relates the expected number of level crossings of a stochastic process to the distribution of the supremum of that process. Specifically, define the number of upcrossings of c by the process V_j (which has derivative V_{j-1}) by

$$U_c := \#\left\{t \in T : V_j(t) = c,\ V_{j-1}(t) > 0\right\}. \tag{3.11}$$

Then, because V_j is equal to zero with probability one at the left endpoint of its domain,

$$\mathrm{P}\left\{\sup_{t\in T} V_j(t) > c\right\} = \mathrm{P}\left\{U_c \geq 1\right\} \leq \mathrm{E}\left[U_c\right]. \tag{3.12}$$

The simplest Rice formula is used to calculate the above expectation. In some cases it is possible to calculate a series that converges to the boundary crossing probability on the left-hand side of (3.12) (Adler, 2000), but these proved too complicated for the present case because of the nonstationarity of V_j. This implies, unfortunately, that the computed approximations overreject the null hypothesis[2]. Theorem 7 supplies a method to compute this expectation; it is adapted from Azaïs and Wschebor (2009, Theorem 3.2). Throughout, we use the notation $\phi(x)$ for the standard normal density, $\Phi(x) = \int_{-\infty}^{x} \phi(y)\mathrm{d}y$ for the standard normal distribution function and $\Psi(x) = 1 - \Phi(x)$.

Theorem 7. *Let $j \geq 2$. Then for $c > 0$,*

$$\mathrm{P}\left\{\sup_{t\in T} V_j(t) > c\right\} \leq \int_T \left(cM_j(t)\Phi\left(c\frac{M_j(t)}{\Sigma_j(t)}\right) + \Sigma_j(t)\phi\left(c\frac{M_j(t)}{\Sigma_j(t)}\right)\right) p_j(c,t)\mathrm{d}t \tag{3.13}$$

where

$$M_j(t) = \frac{\frac{\partial}{\partial s}\rho_j(s,t,F)|_{s=t}}{\rho_j(t,t,F)}, \tag{3.14}$$

$$\Sigma_j^2(t) = \rho_{j-1}(t,t,F) - \frac{\left(\frac{\partial}{\partial s}\rho_j(s,t,F)|_{s=t}\right)^2}{\rho_j(t,t,F)} \tag{3.15}$$

[1] It can also be shown that $\frac{\partial^2}{\partial s \partial t}\rho_j(s,t) = \rho_{j-1}(s,t)$.

[2] A lower bound to the level crossing probability resulting from just a two-term approximation should be closer to the true value than the upper bound is, and should also be conservative, which is an attractive feature. Unfortunately the nonstationarity of these processes makes the calculation of the second term difficult.

and p_j is the marginal density function of $V_j(t)$; that is,

$$p_j(c,t) = \frac{\exp\left\{\frac{-c^2}{2\rho_j(t,t,F)}\right\}}{\sqrt{2\pi\rho_j(t,t,F)}}, \quad c \in \mathbb{R},\ t \in T. \tag{3.16}$$

Recall that ρ_j is defined in (3.8), and it can be verified that (cf. Appendix C)

$$\frac{\partial}{\partial s}\rho_j(s,t,F)|_{s=t} = \binom{2j-3}{j-1}\mathcal{I}_{2j-2}(t,F) - \mathcal{I}_j(t,F)\mathcal{I}_{j-1}(t,F). \tag{3.17}$$

When the process is weighted, some care must be taken in defining the Rice approximation, because the pathwise derivative of the weighted function of order j is no longer an order-$(j-1)$ process. Assuming that the weight function w is differentiable, the derivative of the weighted process $w(t)V_j(t)$ is $w'(t)V_j(t) + w(t)V_j'(t)$. It has variance function

$$\mathrm{Var}\left(\frac{\mathrm{d}}{\mathrm{d}t}w(t)V_j(t)\right) = \left(w'(t)\right)^2\rho_j(t,t) + 2w'(t)w(t)\frac{\partial}{\partial s}\rho_j(s,t)|_{s=t} + w^2(t)\rho_{j-1}(t,t) \tag{3.18}$$

and the covariance between the weighted process and its derivative at t is

$$\mathrm{Cov}\left(\frac{\mathrm{d}}{\mathrm{d}t}w(t)V_j(t), w(t)V_j(t)\right) = w'(t)w(t)\rho_j(t,t) + w^2(t)\frac{\partial}{\partial s}\rho_j(s,t)|_{s=t}. \tag{3.19}$$

Using these definitions, we arrive at the following corollary for weighted processes.

Corollary 1. *Let $j \geq 2$, and assume w is a differentiable weight function that satisfies (3.4). Then for $c > 0$,*

$$\mathrm{P}\left\{\sup_{t\in T} w(t)V_j(t) > c\right\} \leq \int_T \left(c\tilde{M}_j(t)\Phi\left(c\frac{\tilde{M}_j(t)}{\tilde{\Sigma}_j(t)}\right) + \tilde{\Sigma}_j(t)\phi\left(c\frac{\tilde{M}_j(t)}{\tilde{\Sigma}_j(t)}\right)\right)p_j(c,t)\mathrm{d}t \tag{3.20}$$

where

$$\tilde{M}_j(t) = w'(t)w(t) + w^2(t)\frac{\frac{\partial}{\partial s}\rho_j(s,t)|_{s=t}}{\rho_j(s,t)} \tag{3.21}$$

$$\tilde{\Sigma}_j^2(t) = \mathrm{Var}\left(\frac{\mathrm{d}}{\mathrm{d}t}w(t)V_j(t)\right) - \frac{\left(\mathrm{Cov}\left(\frac{\mathrm{d}}{\mathrm{d}t}w(t)V_j(t), w(t)V_j(t)\right)\right)^2}{\rho_j(t,t,F)}. \tag{3.22}$$

Here, p_j is the marginal density function of $w(t)V_j(t)$ and the derivatives above are defined in (3.18) and (3.19).

The above corollary is necessary to accurately evaluate third-order processes; without weighing the process appropriately, integration up to the unknown upper end of the support T affects the estimation of p-values significantly.

Below we explore the performance of the approximation results using the natural benchmark of the Brownian bridge, and then the case of an arbitrary (continuous) distribution function F is considered. Subsection 3.3.1 may be of general interest because of the frequency with which the Brownian bridge is used in applications.

3.3.1 On integrated Brownian bridges

In this Section we write $\mathcal{I}_j(t)$ for $\mathcal{I}_j(t, \text{Unif}[0,1])$, B for a standard Brownian bridge process — that is, a mean-zero Gaussian process with covariance function $\mathrm{E}\left[B(s)B(t)\right] = s \wedge t - st$ — and the family of limiting processes $B^{(j)}(t) = \mathcal{I}_j(t, B)$. For example, $B^{(1)}$ is the standard Brownian bridge, $B^{(2)}$ is the integrated Brownian bridge $\int_0^t B(s)\mathrm{d}s$, and so on. The Rice method proposed in Theorem 7 will be applied to the distribution of $\sup_{t\in[0,1]} B^{(j)}(t)$.

Note that for the uniform distribution, $\mathcal{I}_j(t) = \frac{z^j}{j!}$. Utilizing this fact and specializing (3.8) to the uniform distribution, the covariance function of $B^{(j)}$ is

$$\mathrm{E}\left[B^{(j)}(s)B^{(j)}(t)\right] := r_j(s,t) = \sum_{\ell=0}^{j-1} \binom{2j-\ell-2}{j-1} \frac{|t-s|^\ell}{\ell!} \frac{(s\wedge t)^{2j-\ell-1}}{(2j-\ell-1)!} - \frac{s^j t^j}{(j!)^2} \tag{3.23}$$

and the variance function is

$$\sigma_j^2(t) = r_j(t,t) = \binom{2j-2}{j-1} \frac{t^{2j-1}}{(2j-1)!} - \frac{t^{2j}}{(j!)^2}. \tag{3.24}$$

The relevant properties of the processes $B^{(j)}$ can be checked using this covariance function.

We start with an implication of the properties of $B^{(j)}$ that were described above:

Theorem 8. *Let $j \geq 3$. Then*

$$\mathrm{P}\left\{\sup_{t\in[0,1]} B^{(j)}(t) \geq c\right\} = \Psi\left(\frac{j!\sqrt{2j-1}}{j-1}c\right)(1+o(1)), \quad c\to\infty. \tag{3.25}$$

Theorem 8 implies that for tests of order $j \geq 3$ we can find asymptotically exact critical values for tests, asymptotically exact in the sense that they become better as the level of the test becomes smaller. They rely on large deviation probabilities for Gaussian processes as developed in the monograph of Piterbarg

(1996). For the $j = 2$ case, this theorem does not work for technical reasons[3].

The boundary crossing probabilities described in Theorem 8 are remarkably simple to use to find critical values. For example, given any (small) α, asymptotic critical values for third-order tests can be found by solving

$$\alpha = \Psi\left(3\sqrt{5}c\right) \tag{3.26}$$

for c. It proved difficult to apply Theorem 8 to the case of $j = 2$, but below we give an approximation as a corollary of the Rice formula of Theorem 7. The convenient covariance structure of the family of $\{B^{(j)}\}_j$ makes an explicit characterization of the formula possible for the second-order process.

Corollary 2. *For $c > 0$*

$$\mathrm{P}\left\{\sup_{t\in[0,1]} B^{(2)}(t) > c\right\} \leq \int_0^1 \frac{6c(1-t)}{t(4-3t)}\Phi\left(\frac{6c}{t^{3/2}}\sqrt{\frac{1-t}{4-3t}}\right) + \sqrt{\frac{t(1-t)}{4-3t}}\phi\left(\frac{6c}{t^{3/2}}\sqrt{\frac{1-t}{4-3t}}\right) p_2(c,t)\mathrm{d}t, \tag{3.27}$$

where p_2 be the marginal density function of $B^{(2)}(t)$; that is,

$$p_2(c,t) = \frac{\exp\left\{\frac{-c^2}{2\sigma_2^2(t)}\right\}}{\sqrt{2\pi\sigma_2^2(t)}}, \quad t \in [0,1]. \tag{3.28}$$

The integral in (3.27) is very easy to invert to find approximate critical values. Some values corresponding to sizes used commonly for testing are given in Table 3.1. We also note that the approximation proposed in Corollary 2 above is not limited to use in second-order tests — Theorem 7 can be extended to processes of any order $j \geq 2$. This makes it possible to check the accuracy of the Rice formula by comparing it to large deviations approximation in the $j \geq 3$ cases, and for the uniform distribution the Rice approximation is quite accurate — see Table 3.2.

It is noted in Schmid and Trede (1998, Proof A.3, p. 193) that for $j = 2$ the approximations using the Brownian bridge are valid for a class of certain distribution functions beyond the uniform distribution. These are distribution functions for which $f > 0$ and $\dot{f} < 0$ — that is, distributions which have decreasing densities over their entire support. The reason for this lies with the major consequence of second-order stochastic dominance — invariance to concave transformations: one random variable dominates another stochastically at the second order if $\mathrm{E}[h(X)] \geq \mathrm{E}[h(Y)]$ for all concave increasing functions h, and in this case the transform $X \mapsto F(X)$ is one such function, which also conveniently makes the transformed samples uniformly distributed under the null hypothesis[4]. There are therefore

[3] It is difficult to establish local stationarity of $B^{(2)}$ at the right endpoint of $[0,1]$.

[4] Similarly, for any increasing function h, if X dominates Y stochastically at the first order, $\mathrm{E}[h(X)] \geq \mathrm{E}[h(Y)]$, which explains

several popular parametric distributions for which these critical values apply, and they should be roughly accurate for populations that exhibit everywhere-decreasing densities.

3.3.2 Performance of Rice formula approximations in the two-sample case

Some consequences of Theorem 8 and Corollary 2 are collected for the first three orders in Table 3.1, for some common test sizes. The first-order test statistics are simply the conventional Kolmogorov-Smirnov critical values, while the second- and third- order critical values are found by inverting the approximations (3.27) and (3.25) respectively.

Table 3.1: Approximate critical values for tests of stochastic dominance, orders 1, 2 and 3. The second line contains approximations derived from the Rice formula of Corollary 2, while the third line contains asymptotically exact values derived from Theorem 8.

	10%	5%	1%
j=1: Kolmogorov-Smirnov critical value	1.07298	1.22387	1.51743
j=2: Rice approximation	0.37772	0.47974	0.67390
j=3: Large deviations approximation	0.19104	0.24520	0.34679

We also note that the critical value approximations given by an application of Theorem 7 to processes $B^{(j)}$ for higher j are very close to the asymptotically exact values of Theorem 8. A comparison is presented in Table 3.2. Note that the first line of Table 3.2 is, practically speaking, identical to the simulation results of Schmid and Trede (1998), who obtained the following critical values for the sizes chosen in Table 3.2: for sizes 10%, 5% and 1% they found respectively 0.38, 0.48 and 0.68.

Table 3.2: A comparison of approximate critical values for the uniform distribution, for processes of different orders. For higher orders the Rice formula approximations (left half) are very close to the large deviations approximations (right half).

Order	**Rice formula approximations**			**Large deviation approximations**		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
j = 2	0.377751	0.479746	0.673877	-	-	-
j = 3	0.191047	0.245191	0.346762	0.191042	0.245200	0.346791
j = 4	0.060545	0.077714	0.109909	0.060548	0.077712	0.109910

3.4 Tests in conditional models

Conditional tests of stochastic dominance are also often of interest; for example, the distribution of stock returns conditional on covariates may be more relevant to researchers than the unconditional dis-

why it is possible to make the transformation $X \mapsto F(X)$ for any distribution function F and to use Kolmogorov-Smirnov critical values for testing.

tribution of returns — see for example the references cited in Linton et al. (2005, p. 736-37). The distribution of the empirical process constructed from regression residuals is affected by the estimation procedure. In this Section we analyze the effect that parametric estimation has on the empirical distributions used in tests and propose a method based on the Rice formula to construct critical values for tests. The assumption of a regression model implies that the only parametric assumption on the conditional distribution $Y|X$ is that it is a member of a location-shift family. The principal difference, then, between the critical values in this Section and those in Section 3.3 is to account for this feature as it affects the covariance functions of the limiting processes used in the Rice formula.

We assume the data are randomly sampled from two populations labeled $k = 1$ or 2, and the samples can also be assumed to be independent of one another, and the conditional models for each population are linear regression models. This is a simplified version of the modeling assumptions used in Linton et al. (2005). Assume for each model a probability space $(T \times \Theta_k, \Omega_k)$, data $Z_k = (Y_k, X_k) \in T \subseteq \mathbb{R}^{q+1}$ and $\theta_k = (\mu_k, \beta_k) \in \Theta_k \subseteq \mathbb{R}^{p+1}$. We assume the relationship between Y_k and X_k is modeled as

$$Y_k = \mu_k + X_k\beta_k + \varepsilon_k, \qquad k = 1, 2 \tag{3.29}$$

and we are concerned with the conditional distribution of the uncentered error terms $\varepsilon_k(Z_k, \theta_k) + \mu_k = Y_k - X_k\beta_k$, $k = 1, 2$. Below we write $\hat{\varepsilon}_k$ for $\varepsilon(Z_k, \hat{\theta}_k)$ and ε_k for $\varepsilon(Z_k, \theta_{k0})$ when it will cause no confusion.

Corresponding to the empirical distributions $\mathbb{F}_{1n}$ and $\mathbb{F}_{2m}$ defined above, in which parameters of the null distribution were assumed known, define empirical distribution functions depending on parameters by

$$\mathbb{F}_{kn}(t, \theta) = \frac{1}{n}\sum_{i=1}^{n} I(\varepsilon_{ki}(Z_{ki}, \theta) + \mu \le t), \quad k = 1, 2. \tag{3.30}$$

Then the object of primary interest for the development of approximate critical values is the process depending on both samples

$$V_{jmn}(t, \hat{\theta}_1, \hat{\theta}_2) = \sqrt{\frac{nm}{n+m}}\left(\mathcal{I}_j(t, \mathbb{F}_{2m}(\cdot, \hat{\theta}_2) - \mathcal{I}_j(t, \mathbb{F}_{1n}(\cdot, \hat{\theta}_1)\right) \qquad t \in T. \tag{3.31}$$

As before, we investigate sup-norm tests of H_0^j for which we reject the null hypothesis when $\sup_{t\in T} V_{jmn}(t, \hat{\theta}_1, \hat{\theta}_2)$ is large and assume the least favorable case $F_1 \equiv F_2 \equiv F$. More specifically, we make the following assumptions for each model — that is, for $k = 1, 2$, we assume:

A1: The data $\{Z_{ki}\}_{i=1}^n$ are iid and the samples are distributed independently of one another (for more on dependence, see Linton et al. (2005)). We also assume some regularity on the design: as-

sume $\lim_{n\to\infty} \frac{X_k^\top X_k}{n} = Q_k$ exists, $(X_k^\top X_k)^{-1}$ exists for all $n \geq p$, and $\max_i x_{ki}(X_k^\top X_k)^{-1}x_{ki}^\top = o(1)$. Throughout this Section, individual x_{ki} in X_k are interpreted as row vectors *without an intercept term*.

A2: The distribution of $Y_k|X_k$ is a member of a location-shift family; assume the distribution of the vector Y_k given X_k satisfies $F_{Y_k|X_k}(\cdot|\theta_k, X_k) = F_k(\cdot - \mu_k - X_k\beta_k)$ for some location-shift distribution function F_k. It is assumed that F_k is absolutely continuous with respect to Lebesgue measure and it has two derivatives f_k and $\dot{f}_k$ such that

$$0 < \Gamma_k := \int (\dot{f}_k/f_k)^2 \mathrm{d}F_k < \infty. \tag{3.32}$$

Note that this implies the vector of partial derivatives $\nabla_\beta F_{Y_k|X_k}(\cdot|\theta_k, X_k)$ is bounded because $\nabla_\beta F_{Y_k|X_k}(\cdot|\theta_k, X_k) = \nabla_\beta F_k(\cdot - \mu_k - x_{ki}\beta_k) = -x_{ki}^\top f_k(\cdot - \mu_k - x_{ki}\beta_k)$.

A3: The function $\mathcal{I}_j$ is differentiable with respect to β for all x in a neighborhood of β_{k0}. Assume that the vector $\nabla_\beta \mathcal{I}_j(\cdot, F(\cdot,\theta)) = \mathcal{I}_j(\cdot, \nabla_\beta F(\cdot,\theta))$.

A4: For each model the parameter estimates $\hat{\theta}_k$ are asymptotically efficient (see e.g. Durbin (1973a)); that is, letting $\tilde{X}_k = (\mathbf{1}_n, X_k)$ be the design including an intercept and $\tilde{Q}_k = \operatorname{plim} \frac{1}{n}\tilde{X}_k^\top \tilde{X}_k$,

$$\sqrt{n}\left(\hat{\theta}_k - \theta_{k0}\right) = \frac{1}{\sqrt{n}}\Gamma_k^{-1}\tilde{Q}_k^{-1}\sum_{i=1}^{n}\tilde{X}_{ki}^\top S_k(Y_{ki}, \tilde{X}_{ki}, \theta_{k0}) + o_P(1) \tag{3.33}$$

where $S_k(Y_{ki}, \tilde{X}_{ki}, \theta) = (-\dot{f}_k/f_k)(\varepsilon_{ki}(\theta))$. Note that this (and independence of the observations) implies $\mathrm{E}\left[S_k(Y_k, \tilde{X}_k, \theta_{k0})|\tilde{X}_k\right] = \mathbf{0}_n$ and $\mathrm{E}\left[S_k(Y_k, \tilde{X}_k, \theta_{k0})S_k^\top(Y_k, \tilde{X}_k, \theta_{k0})|\tilde{X}_k\right] = \Gamma_k I_{n\times n}$.

Assumption **A4** is important because the estimator is assumed to be $\sqrt{n}$-convergent. That is, the processes that depend on parameter estimates have distributions that are different from the two-sample case when the rate of convergence of the model parameters is the same as the rate of convergence of each empirical distribution function to the true distribution function of the data. When distribution functions are estimated in some other way, for example using kernel estimators, the distribution of the resulting $\hat{V}_j$ is affected — see Shen (2011).

Below, Assumption **A4** will only be explicitly used with the subset $\hat{\beta}_k$; that is, excluding the intercept estimate $\hat{\mu}_k$. Note that this assumption also implies we can write the distribution of $\hat{\beta}_k$ in a convenient way: $\sqrt{n}(\hat{\beta}_k - \beta_{k0}) = \frac{1}{\sqrt{n}}\Gamma_k^{-1}Q_k^{-1}X_k^\top S_k = \sqrt{n}\Gamma_k^{-1}(X_k^\top X_k)^{-1}X_k^\top S_k + o_P(1)$. This final expression is useful in the proof of Theorem 9.

We start by describing the distribution of each empirical process constructed using regression residuals from one model. We make one more convenient definition:

$$\overline{P}_k := \bar{X}_k^\top Q^{-1} \bar{X}_k = \lim_{n\to\infty} \frac{\mathbf{1}_n^\top X_k}{n} \left(\frac{X_k^\top X_k}{n} \right)^{-1} \frac{X_k^\top \mathbf{1}_n}{n} = \frac{1}{n} \mathbf{1}_n^\top P_{X_k} \mathbf{1}_n + o_P(1), \tag{3.34}$$

which is assumed to exist. This (scalar) value summarizes the impact of the design matrix X_k on the limiting covariance function. If an intercept were included in the design, this would be equal to 1, but without an intercept it is in $[0,1]$ because it is the projection of a unit vector on X_k via P_{X_k}. This also suggests an efficient way to calculate this quantity in practice: take the mean of the fitted values of a regression of 1 on X_k.

Theorem 9. *Assume **A1** through **A4**. For $k = 1$ or 2, define*

$$V_{1n}^k(t, \hat{\theta}_k) = \sqrt{n} \left(\mathbb{F}_{kn}(t, \hat{\theta}_k) - F_k(t - \mu_{k0}) \right), \qquad t \in T. \tag{3.35}$$

Then uniformly in $t \in T$,

$$V_{1n}^k(t, \hat{\theta}_k) = V_{1n}^k(t, \theta_{k0}) + \frac{1}{\sqrt{n}} f_k(t - \mu_{k0}) \Gamma_k^{-1} \mathbf{1}_n^\top P_{X_k} S_k + o_P(1) \tag{3.36}$$

where the $n \times 1$ vector $S_k = S_k(Y_k, X_k, \theta_{k0})$ and $P_{X_k} = X_k (X_k^\top X_k)^{-1} X_k^\top$ where X_k does not contain an intercept. *This implies that conditional on X_k, $V_{1n}^k(t, \hat{\theta}_k)$ converges weakly to a mean-zero Gaussian process $\hat{V}_1^k$ with covariance function*

$$\mathrm{Cov}\left(\hat{V}_1^k(s), \hat{V}_1^k(t) \right) = F_k(s \wedge t - \mu_{k0}) - F_k(s - \mu_{k0}) F_k(t - \mu_{k0}) - f_k(s - \mu_{k0}) f_k(t - \mu_{k0}) \Gamma_k^{-1} \overline{P}_k. \tag{3.37}$$

We remark that using the simple redefinition $t \in T \mapsto t - \mu_{k0} := t' \in T'$, we have that the limiting process in Theorem 9 is a mean-zero Gaussian process on T' with covariance function

$$\mathrm{Cov}\left(\hat{V}_1^k(s), \hat{V}_1^k(t) \right) = F_k(s \wedge t) - F_k(s) F_k(t) - f_k(s) f_k(t) \Gamma_k^{-1} \overline{P}_k. \tag{3.38}$$

This transformation will rarely affect any calculations and in any case will not affect the distribution of the supremum of $\hat{V}_1^k$. Henceforth we define limiting processes in this way.

Processes similar to (3.35) are dealt with extensively in Koul (2002). However, the addition of the estimated intercept parameter to the residuals makes a direct application of these results difficult. The process dealt with here is nonstandard for two reasons. First, the intercept estimate is added to the

residuals, changing the distribution of the process — note the process still depends on $\hat{\mu}_k$ because it is used to obtain estimated residuals. Second, estimated values are used in the empirical distribution function, while the true parameter values are used in the theoretical distribution function. In the proof, the elegant results of van der Vaart and Wellner (2007) are adapted to deal with these differences.

The final f_k terms in (3.38) reflect the effect of parameter estimation, although no actual parameter values enter the equation. This is a result of the fact that F is modeled as a location-shift model, which implies that specific parameter values do not affect the distribution of $\hat{V}_1^k$. This phenomenon can be seen in the distributions of all the integrated statistics to be seen below.

We now extend Theorem 9 to integrated processes of any order j. We note that the functional $\mathcal{I}_j$ need not only be defined for distribution functions, although this is how it has been applied thus far in practice. We define $\mathcal{I}_j(t, f_k)$ in a manner analogous to (3.1). These functions will play a major role in Theorems 10 and 11. To be clear, for some density function f we write

$$\mathcal{I}_j(\varepsilon, f) = \begin{cases} f(\varepsilon) & j = 1 \\ \int_{-\infty}^{\varepsilon} \mathcal{I}_{j-1}(\varepsilon', f) \mathrm{d}\varepsilon' & j = 2, 3, \ldots \end{cases} \tag{3.39}$$

$\mathcal{I}_j(\cdot, f)$ is conveniently related to $\mathcal{I}_j(\cdot, F)$: by the definition of any distribution function F with density f, $\mathcal{I}_2(\cdot, f) \equiv F$, and therefore inductively $\mathcal{I}_j(\cdot, f) \equiv \mathcal{I}_{j-1}(\cdot, F)$ for all higher j. In Theorems 10 and 11, we use $\mathcal{I}_j(t, f)$, but these functions may be substituted with $\mathcal{I}_{j-1}(t, F)$ when $j \geq 2$.

Using the above notation, Theorem 10 nests Theorem 9 as one member of a family of integrated processes. Extend (3.35) to any order j as follows:

$$V_{jn}^k(t, \hat{\theta}_k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathcal{I}_j \left(t, \mathbb{F}_{kn}(\cdot, \hat{\theta}_k) \right) - \mathcal{I}_j \left(t, F_k \right) \right), \qquad t \in T'. \tag{3.40}$$

Once again, estimated parameters are used to build the empirical distribution function but the theoretically true parameters are used in F.

Theorem 10. *Let $j \geq 1$. Assume **A1** to **A4** and that the model includes an intercept. Then uniformly in $t \in T'$,*

$$V_{jn}^k(t, \hat{\theta}_k) = V_{jn}^k(t, \theta_{k0}) + \frac{1}{\sqrt{n}} \mathcal{I}_j(t, f_k) \Gamma_k^{-1} \mathbf{1}_n^\top P_{X_k} S_k + o_P(1). \tag{3.41}$$

The limiting process $\hat{V}_j^k$ is a mean-zero Gaussian process on T' with covariance function

$$\mathrm{Cov}\left(\hat{V}_j^k(s), \hat{V}_j^k(t) \right) = \rho_j(s, t, F_k) - \mathcal{I}_j(s, f_k) \mathcal{I}_j(t, f_k) \Gamma_k^{-1} \overline{P}_k \tag{3.42}$$

where ρ_j is the covariance function defined in (3.8).

As in Theorem 9, the final term on the right hand side of the covariance function is a result of parameter estimation; the specific characterization shown above is only a convenient result of the fact that the distribution is modeled as a location-shift model and asymptotic efficiency of $\hat{\theta}_k$. The assumed efficiency of $\hat{\beta}_k$ in assumption **A4** makes the covariance function of the limiting process $\hat{V}_j^k$ tractable — without this assumption, $\hat{V}_j$ would have a more complicated covariance function.

Each process for $j \geq 2$ has $(j-1)$-times-differentiable sample paths, with derivatives $\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}\hat{V}_j^k(t) \stackrel{D}{=} \hat{V}_{j-\ell}^k(t)$, $\ell \leq j-1$. Intuitively, it is clear that integrated residual processes are smooth in the same way that the integrated two-sample processes were. To show this formally, consider the covariance function of $\hat{V}_j^k$ given in (3.42). It is apparent that

$$\lim_{s \nearrow t}\left(\frac{\partial}{\partial s}\mathrm{Cov}\left(\hat{V}_j^k(s), \hat{V}_j^k(t)\right) - \frac{\partial}{\partial t}\mathrm{Cov}\left(\hat{V}_j^k(s), \hat{V}_j^k(t)\right)\right) = 0 \tag{3.43}$$

because of the differentiability of (3.8) and the fact that the other term is symmetric in s and t. Furthermore, Theorem 10 also implies that the covariance functions of processes integrated to different orders satisfies the same relationship as in the case without estimated parameters:

$$\frac{\partial^2}{\partial s \partial t}\mathrm{Cov}\left(\hat{V}_j^k(s), \hat{V}_j^k(t)\right) = \mathrm{Cov}\left(\hat{V}_{j-1}^k(s), \hat{V}_{j-1}^k(t)\right). \tag{3.44}$$

The limiting distribution of $V_{jmn}(\cdot, \hat{\theta}_1, \hat{\theta}_2)$ can be expressed in a similar way, because it is a linear function of the two independent empirical processes V_{jn}^1 and V_{jm}^2.

Theorem 11. *Assume **A1** through **A4**. Then uniformly in $t \in T'$,*

$$\begin{aligned} V_{jmn}(t, \hat{\theta}_1, \hat{\theta}_2) = V_{jmn}(t, \theta_{10}, \theta_{20}) \\ + \frac{\sqrt{1-\lambda}}{\sqrt{m}}\mathcal{I}_j(t, f_2)\Gamma_2^{-1}\mathbf{1}_m^\top P_{X_2} S_2 \\ - \frac{\sqrt{\lambda}}{\sqrt{n}}\mathcal{I}_j(t, f_1)\Gamma_1^{-1}\mathbf{1}_n^\top P_{X_1} S_1 + o_P(1). \end{aligned} \tag{3.45}$$

Under the null hypothesis $F_1 \equiv F_2 \equiv F$, $V_{jmn}(\cdot, \hat{\theta}_1, \hat{\theta}_2)$ converges weakly to $\hat{V}_j$, a mean-zero Gaussian process on T' with covariance function

$$\mathrm{Cov}\left(\hat{V}_j(s), \hat{V}_j(t)\right) = \rho_j(s, t, F) - \mathcal{I}_j(s, f)\mathcal{I}_j(t, f)\Gamma^{-1}\left(\lambda\overline{P}_1 + (1-\lambda)\overline{P}_2\right). \tag{3.46}$$

3.4.1 Critical values for conditional tests

The distribution of $\hat{V}_j$ derived in Theorem 11 is the result from which we derive the critical value results in this Subsection. Corollary 3 is analogous to Theorem 7 and specifies the form that the Rice formula takes for conditional models. We also note that for first-order tests, a Rice formula may not be used but the method proposed in Durbin (1985) and Rabinowitz (1993) can be used with similar estimated quantities.

Corollary 3. *Let $j \geq 2$ and assume **A1** though **A4** and $\mathcal{I}_j(t,f) = \mathcal{I}_{j-1}(t,F)$. Then*

$$\mathrm{P}\left\{\sup_{t\in T} \hat{V}_j(t) > c\right\} \leq \int_T \left(c\hat{M}_j(t)\Phi\left(c\frac{\hat{M}_j(t)}{\hat{\Sigma}_j(t)}\right) + \hat{\Sigma}_j(t)\phi\left(c\frac{\hat{M}_j(t)}{\hat{\Sigma}_j(t)}\right)\right)\hat{p}_j(c,t)\mathrm{d}t \tag{3.47}$$

where

$$\hat{M}_j(t) = \frac{\frac{\partial}{\partial s}\rho_j(s,t,F)|_{s=t} - \mathcal{I}_{j-2}(t,F)\mathcal{I}_{j-1}(t,F)\Gamma^{-1}\overline{P}_\lambda}{\rho_j(t,t,F) - \mathcal{I}^2_{j-1}(t,F)\Gamma^{-1}\overline{P}_\lambda}, \tag{3.48}$$

$$\hat{\Sigma}^2_j(t) = \rho_{j-1}(t,t,F) - \mathcal{I}^2_{j-2}(t,F)\Gamma^{-1}\overline{P}_\lambda - \frac{\left(\frac{\partial}{\partial s}\rho_j(s,t,F)|_{s=t} - \mathcal{I}_{j-2}(t,F)\mathcal{I}_{j-1}(t,F)\Gamma^{-1}\overline{P}_\lambda\right)^2}{\rho_j(t,t,F) - \mathcal{I}^2_{j-1}(t,F)\Gamma^{-1}\overline{P}_\lambda} \tag{3.49}$$

where $\overline{P}_\lambda = \lambda\overline{P}_1 + (1-\lambda)\overline{P}_2$ and $\hat{p}_j$ is the marginal density function of $\hat{V}_j(t)$.

In order to apply the above Corollary to weighted processes, the distribution of the derivative of the process needs to be additionally derived. The distribution follows from the fact that

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t)\hat{V}_j(t) = w'(t)\hat{V}_j(t) + w(t)\hat{V}'_j(t) \tag{3.50}$$

and that the derivative is Gaussian.

Once again it is not realistic to assume that any more parametric features of F than that it is a member of a location-shift family. We propose to use consistent estimators of the appropriate functions involved in (3.46). The covariance function (3.46) and Corollary 3 rely on the unknown quantities

$$\mathcal{I}_j(t,F) = \int_{-\infty}^{t} \frac{(t-\varepsilon)^{j-1}}{(j-1)!}\mathrm{d}F(\varepsilon) \tag{3.51}$$

for several orders j and Γ the Fisher information from the location shift parameter. We propose to estimate these quantities in the following manner: for estimates of functions (3.51) we use the empirical

counterpart

$$\mathcal{I}_j(t,\hat{\mathbb{F}}_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{(t-\hat{\varepsilon}_i)^{j-1}}{(j-1)!}I(\hat{\varepsilon}_i \le t) \tag{3.52}$$

and for Γ we use a kernel estimator as described in Portnoy and Koenker (1989) to estimate the score function $-\dot{f}/f$, and integrate this to produce an estimate of the Fisher information. We use only one of the two samples to make these estimates. Given estimators for the above quantities, we propose approximate critical values for tests derived by following the formula in Theorem 3 but with these estimates in the place of their theoretical counterparts. Then under the null hypothesis $\sup_{t\in T}\hat{V}_{jmn}(t)$ converges in distribution to $\sup_{t\in T}\hat{V}_j(t)$, while under any alternative the difference between the two empirical distribution functions diverges, while the effect of estimated parameters converges its limit given F_1 only, making the test consistent against alternatives.

When $j=1$, the Rice formula given above does not apply to this process, because $\hat{V}_1$ does not have differentiable sample paths. This can be seen by noting that the derivatives with respect to each argument of the covariance function of $\hat{V}_1$ do not match when evaluated at the same point:

$$\lim_{s\nearrow t}\left\{\frac{\partial}{\partial s}\left[F(s)-F(s)F(t)-\Gamma^{-1}f(s)f(t)\right]-\frac{\partial}{\partial t}\left[F(s)-F(s)F(t)-\Gamma^{-1}f(s)f(t)\right]\right\} = f(t) > 0 \tag{3.53}$$

for any $t\in T$. Therefore, in the spirit of Durbin (1985), we propose an approximation using the same estimate for Γ as above and a kernel density estimate for f. Because in the $j=1$ case the test is invariant to monotone transformations, we make the substitution $t=F(x)$ to normalize the limiting process to the unit interval. Denote the covariance function of the normalized process by $\tilde{\rho}_1$:

$$\tilde{\rho}_1(s,t) = s\wedge t - st - f(F^{-1}(s))f(F^{-1}(t))\Gamma^{-1}\overline{P}_\lambda \tag{3.54}$$

Let $\hat{\tilde{\rho}}_j$ denote the estimate of this function using estimates of f, F, Γ and $\overline{P}_\lambda$. The Durbin-style approximation also requires the derivative of the covariance function with respect to s when $s\le t$. This function,

$$\frac{\partial}{\partial s}\tilde{\rho}_1(s,t) = 1 - t - (\dot{f}/f)(F^{-1}(s))f(F^{-1}(t))\Gamma^{-1}\overline{P}_\lambda \tag{3.55}$$

is easy to estimate because the estimated score function, $-\widehat{\dot{f}/f}$ is already used to estimate Γ. Putting these pieces together, the approximate level crossing probability of $\hat{V}_1$ is

$$\mathrm{P}\left\{\sup_{t\in T'}\hat{V}_1(t) > c\right\} \approx \int_{[0,1]} c\,\frac{\frac{\partial}{\partial s}\hat{\tilde{\rho}}_1(t,t)|_{s=t}}{\hat{\tilde{\rho}}_1(t,t)}\hat{p}_1(c,t)\mathrm{d}t. \tag{3.56}$$

where $\hat{p}_1(c,t)$ is the $\mathcal{N}(0,\hat{\tilde{\rho}}_1(t,t))$ density evaluated at c. Approximate critical values can be constructed by solving the following equation for the value of c that makes this probability equal to the desired size of the test.

3.5 Simulation study

In this section we present small simulation experiments corresponding to the tests described above. We include results on two-sample models, comparing them to previously proposed methods of testing. The Rice formula approximations proposed above are competitive with existing methods in terms of size and power. Simulations for conditional models are currently underway and will be added soon.

3.5.1 Rice formula approximations with known distribution functions

As a check on the methods proposed above, we simulated the null distribution using samples that were distributed uniformly, normally and exponentially. For these distributions $\mathcal{I}_j(\cdot,F)$ can be expressed analytically for the first 5 orders, allowing one to quickly and exactly compute values for $j=2$ and 3 using a Rice formula. These values are presented in Table 3.3.

Table 3.3: Approximate critical values derived from the Rice formula for three parametric models: uniform, exponential and normal. These critical values are for data that have not transformed onto the unit interval; if this transformation is used, the exponential distribution should have the same critical values as the uniform distribution when $j=2$.

	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.01$
Uniform			
$j=2$	0.37771	0.47974	0.67387
$j=3$	0.19102	0.24520	0.34680
Exponential			
$j=2$	1.32871	1.67277	2.33199
$j=3$	6.42263	8.24334	11.65867
Normal			
$j=2$	1.33875	1.68502	2.34874
$j=3$	4.06280	5.21455	7.37504

The uniform distribution is a natural benchmark for the behavior of the Rice formula because everything can be computed analytically and, as mentioned above, the approximations derived from this case should be accurate for distributions that appear to have a decreasing density function. To investigate the quality of the approximations, we simulated 10,000 uniform [0,1] samples of size 100 and tested them against the critical value in Table 3.3. This is done for the first three orders of dominance, and for the three typical test sizes that were presented above: for 10%, 5% and 1% level tests. The results are

presented in Table 3.4.

Table 3.4: Size of tests using simulated data and exact Rice formula. 10,000 samples of size 100 were drawn from the uniform, exponential and normal distributions and tested using theoretically derived critical values. These are the empirical sizes. Nominal sizes in the columns, orders on the left, and models also on the left. Note that first-order tests are checked against standard Kolmogorov-Smirnov values.

	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
Uniform			
$j = 1$	0.0962	0.0467	0.0091
$j = 2$	0.0942	0.0456	0.0094
$j = 3$	0.0961	0.0480	0.0096
Exponential			
$j = 1$	0.1003	0.0473	0.0112
$j = 2$	0.1024	0.0510	0.0111
$j = 3$	0.1220	0.0686	0.0234
Normal			
$j = 1$	0.0962	0.0472	0.0101
$j = 2$	0.0966	0.0471	0.0105
$j = 3$	0.1065	0.0602	0.0144

It can be seen in Table 3.4 that the second-order approximation does well for each model, while the third-order approximation is not as good. It is probable that this is the effect of the use of truncation instead of a more smooth method of weighting — changing the limits of integration in the normal case, in particular, can change the resulting critical values. To obtain the figures currently in the table, the exponential and normal cases were truncated (arbitrarily) at their 0.001^{th} and 0.999^{th} quantiles, since integrating these distributions over their whole (noncompact) domains would result in divergent values. Indeed, setting the limits of integration at more extreme (in absolute value) points, one has a number of difficulties with numerical integration due to this fact.

3.5.2 Two-sample models

In the last Subsection, the true distribution of the data was known, but in usual situations analysts will not be able to or desire to describe the data using a simple parametric model. Features of the distributions need to be estimated for tests of order 2 or higher (and order 1 or higher in the case of conditional models).

In order to investigate the performance of the Rice formula in the two-sample case when these features are estimated, a more extensive simulation study using lognormally-distributed data was made. In this study, it is not assumed anything is known about the distribution of the data. Features of the distribution function were estimated using the first of the two samples. Second-order test p-values were calculated on the region between the 1^{st} and 99^{th} quantiles of the data. However, performance in third-order

tests is severely affected by truncation without more smooth weighting of the data. The upper tail of the Gaussian distribution function, centered at the largest observation and with a standard deviation equal to a tenth of the range of the observed data, was chosen as a weight function. This was chosen arbitrarily (except that it is differentiable and familiar) with no attempt to optimize its performance. This is intended to illustrate the fact that it is necessary to weight the observations, but the specific choice of weight function is not as important as the fact that it does not simply take the form of truncation beyond the largest observation.

Several existing methods were chosen as simple benchmarks against which one can investigate the size and power of the methods proposed above. These methods are compared to the methods proposed here in the following simulation experiment. We briefly describe our implementation the procedures. Throughout, we assume that the sample size is n for both samples.

Barrett and Donald (2003) propose a simulated p-value approach using multiplier methods in the spirit of Hansen (1996). Create a sample of statistics

$$\sup_t \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathcal{I}_j(t, \mathbf{1}_{(-\infty, X_{1i}]}) - \mathcal{I}_j(t, \mathbb{F}_n) \right) \mathcal{N}_i \tag{3.57}$$

where $\{\mathcal{N}_i\}_{i=1}^n$ are artificially generated realizations of a Normal $(0,1)$ random variable. A multiplier critical value is given by the $(1-\alpha)^{\text{th}}$ quantile of this sample of artificial statistics, and the null hypothesis is rejected if the statistic from the data is larger than the multiplier critical value. In practice it is necessary to compute this statistic on a grid of values ranging over the domain of the pooled samples. 1000 repetitions were used.

Barrett and Donald (2003) also propose two bootstrap procedures, and one of these is used. This consists of comparing the test statistic to the appropriate quantile of a sample of statistics generated using bootstrap from one sample (that is, sampled from the realizations of X_1 with replacement). Samples of statistics are generated using

$$\sup_t \frac{1}{\sqrt{n}} \left(\mathcal{I}_j(t, \mathbb{F}_n^*) - \mathcal{I}_j(t, \mathbb{F}_{1n}) \right) \tag{3.58}$$

where $\mathbb{F}_n^*$ is the empirical distribution function from the bootstrapped sample X_1^*. The $(1-\alpha)^{\text{th}}$ quantile of this sample is taken as a bootstrap critical value, against which the test statistic is compared. As in Barrett and Donald (2003), 1000 bootstrap repetitions were used in this procedure.

Linton et al. (2005) propose a simple subsampling approach. Given any b, we generate test statistics from subsamples $X_{1i}, X_{1(i+1)}, \ldots, X_{1(i+b-1)}$ and $X_{2i}, X_{2(i+1)}, \ldots, X_{2(i+b-1)}$ — that is, for any b, compute all

$n-b+1$ possible subsamples of consecutive observations and for each pair of subsamples compute

$$\sup_t \frac{1}{\sqrt{b}} \left(\mathcal{I}_j(t, \mathbb{F}^*_{2b}) - \mathcal{I}_j(t, \mathbb{F}^*_{1b}) \right) \tag{3.59}$$

for comparison with the true test statistic, where F^*_{kb} is the empirical distribution function from the subsample of size b from population $k = 1$ or 2. Following their suggestion, we calculate samples for many[5] values of b. For each value of b, we calculate an empirical p-value for the test statistic and use the median of this collection to make a decision.

The results are summarized in Table 3.5. It can be seen that the Rice formula method works well for

Table 3.5: Empirical size and power of different testing methods. The columns labeled Alt. 1 are results against an alternative for which the null should be rejected at all orders $j = 1, 2, 3$, while in the columns labeled Alt. 2 the null should be rejected only for the $j = 1$ tests. The theoretical size of all tests is 5%, and 1,000 simulation repetitions were used.

	$n = \mathbf{50}$			$n = \mathbf{100}$			$n = \mathbf{200}$		
	Null	Alt. 1	Alt. 2	Null	Alt. 1	Alt. 2	Null	Alt. 1	Alt. 2
$j = \mathbf{1}$									
Kolmogorov-Smirnov	0.030	0.468	0.007	0.028	0.805	0.017	0.045	0.988	0.142
Simulated p-value	0.057	0.591	0.018	0.058	0.879	0.054	0.078	0.996	0.269
Bootstrap	0.048	0.562	0.014	0.051	0.878	0.044	0.052	0.990	0.177
Subsample	0.040	0.348	0.195	0.031	0.574	0.355	0.038	0.831	0.643
$j = \mathbf{2}$									
Rice formula	0.017	0.237	0.000	0.017	0.425	0.000	0.029	0.708	0.000
Simulated p-value	0.033	0.310	0.000	0.036	0.513	0.000	0.045	0.811	0.000
Bootstrap	0.027	0.286	0.000	0.034	0.488	0.000	0.041	0.784	0.000
Subsample	0.026	0.144	0.000	0.018	0.245	0.000	0.032	0.447	0.000
$j = \mathbf{3}$									
Rice formula	0.027	0.334	0.000	0.032	0.514	0.000	0.038	0.714	0.000
Simulated p-value	0.065	0.539	0.000	0.067	0.721	0.000	0.080	0.881	0.000
Bootstrap	0.065	0.531	0.000	0.058	0.716	0.000	0.075	0.878	0.000
Subsample	0.077	0.344	0.000	0.067	0.489	0.000	0.097	0.642	0.000

two-sample tests of dominance at all orders considered. The first-order results are simply those given by a standard Kolmogorov-Smirnov test as in McFadden (1989). The second- and third-order results in Table 3.5 are results of the Rice methods derived from Theorem 7. The null and alternative distributions were taken from Barrett and Donald (2003). Specifically, the distribution of X_1 in all three situations is the lognormal $(0.85, 0.6)$ distribution. The alternative distributions used in the table are the first two alternatives used in Barrett and Donald (2003), which are lognormal $(0.6, 0.85)$ and lognormal $(1.2, 0.2)$ distributions. The null hypothesis of $F_1 = F_2$ should be rejected for all three orders when X_2 is distributed according to the first alternative, and the first order null hypothesis should be rejected when the distribution of X_2 follows the second alternative (the difference is slight between the second

[5] Specifically, we let b range over values of the set $B = \{\underline{B}, \underline{B}+1, \ldots, \overline{B}\}$ with $\underline{B} = \max\{\lfloor \frac{n}{10} \rfloor, \lfloor \log\log n \rfloor\}$ and $\overline{B} = \min\{\lceil \frac{9n}{10} \rceil, \lceil \frac{n}{\log n} \rceil\}$. The complicated limits were used to accommodate small sample sizes as well as large.

alternative and the baseline distribution, as can be seen in the empirical power figures for the first order in the table). As the table shows, the tests are more conservative than the alternative methods at all three orders, but power is not severely affected.

Without weighting observations as described above, rejection probabilities of the third-order tests are too large and increase as the sample size increases. For example, when the sample size was 200, the observed rejection probability for unweighted third-order statistics was about 20%. This agrees entirely with the results of Horváth et al. (2006). Without a weight function, the third-order statistic should be (asymptotically) unbounded when integrals are computed using no weighting other than truncation over the observed range of the data. Using the arbitrarily-chosen function described above, rejection probabilities are kept under control.

It is also interesting that the other methods return what appear to be sensible results for the third-order test case. The power of the multiplier and the boostrap methods is greater than the Rice method tests; however, at the third order, the distribution of the statistic is not well-behaved and the simulation methods are applied to a degenerate asymptotic distribution. The Rice method accounts for this explicitly. In contrast, the focus on computational methodologies for inference makes it easy to ignore these somewhat pathological features of the distribution of the test statistic.

3.6 Conclusion

The Rice method for computing boundary crossing probabilities of smooth Gaussian processes can be used to derive tests of stochastic dominance for orders higher than the first order. In simulations, these tests are shown to have conservative size and power near the level of simulation-based tests for stochastic dominance in the two-sample case.

Appendix A

P_g and large deviation approximations

In order to clarify equation (1.22), 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:

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

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

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

and

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

where ϕ is the standard normal density function and ξ 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 (1.22), we have

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

for the model with both location and scale unspecified, and

$$P_g(a) = \frac{1/2}{1/4}\sqrt{\frac{-2/4}{-3}}\exp\left\{\frac{-a^2}{2/4}\right\} = \sqrt{2/3}e^{-2a^2} \tag{A.5}$$

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

$$(\hat{\mu},\hat{\sigma}): \qquad \sigma^2(t_0) = \frac{\pi-2}{4\pi} \qquad A = \sqrt{\frac{\pi}{\pi-2}} \qquad C = \frac{2\pi}{\pi-2} \qquad k = 1 \tag{A.6}$$

$$(\mu,\hat{\sigma}): \qquad \sigma^2(t_0) = 1/4 \qquad A = \sqrt{3} \qquad C = 2 \qquad k = 1 \tag{A.7}$$

$$(\hat{\mu},\sigma): \qquad \sigma^2(t_0) = \frac{\pi-2}{4\pi} \qquad A = \sqrt[4]{\frac{2\pi^2}{3(\pi-2)}} \qquad C = \frac{2\pi}{\pi-2} \qquad k = 2 \tag{A.8}$$

Note the value of A is different from what is printed in Piterbarg (1996) for two of three cases. Plugging these values into equation (1.24) results in

$$\mathrm{P}\left\{\sup_{t\in[0,1]} \hat{v}(t) > a \middle| \hat{\mu},\hat{\sigma}\right\} = \sqrt{\frac{2\pi}{\pi-2}}\, e^{\frac{-2\pi}{\pi-2}a^2} \tag{A.9}$$

$$\mathrm{P}\left\{\sup_{t\in[0,1]} \hat{v}(t) > a \middle| \mu,\hat{\sigma}\right\} = \sqrt{2/3}\, e^{-2a^2} \tag{A.10}$$

$$\mathrm{P}\left\{\sup_{t\in[0,1]} \hat{v}(t) > a \middle| \hat{\mu},\sigma\right\} = \frac{\Gamma(1/4)}{\pi-2}\sqrt[4]{\frac{3\pi}{2}}\sqrt{a}\, e^{\frac{-2\pi}{\pi-2}a^2} \tag{A.11}$$

Appendix 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

$$F(x,\theta) = F_0\left(\frac{x-\theta_1}{\theta_2}\right); \qquad x \in \mathcal{X} \subseteq \mathbb{R}, \quad \theta \in \mathbb{R}\times(0,\infty) \tag{B.1}$$

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

$$F(x,\theta) = F_0\left(\left(\frac{x}{\theta_1}\right)^{\theta_2}\right); \qquad x \in \mathcal{X} \subseteq [0,\infty), \quad \theta \in (0,\infty)\times(0,\infty). \tag{B.2}$$

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 (1.13). 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

$$\sup_{x\in\mathbb{R}} |x| f_0(x) < \infty \quad \text{and} \quad \int (\dot{f}_0/f_0))^2(x) + (1 + x(\dot{f}_0/f_0)(x))^2 \mathrm{d}F_0(x) < \infty \tag{B.3}$$

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

$$\sup_{x\in\mathbb{R}^+} x\log x f_0(x) < \infty \quad \text{and} \quad \int (1 + x(\dot{f}_0/f_0)(x))^2 + (1 + \log x + x\log x(\dot{f}_0/f_0)(x))^2 \mathrm{d}F_0(x) \tag{B.4}$$

for scale-shape families[1]. 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:

$$g(t) = \nabla_\theta F(x,\theta)\big|_{x=F^{-1}(t,\theta)} = \frac{-1}{\theta_2}\begin{bmatrix} f_0(F_0^{-1}(t)) \\ F_0^{-1}(t) f_0(F_0^{-1}(t)) \end{bmatrix} \tag{B.5}$$

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

$$\dot{g}(t) = \nabla_\theta \log f(x,\theta)\big|_{x=F^{-1}(t,\theta)} = \frac{-1}{\theta_2}\begin{bmatrix} (\dot{f}_0/f_0)(F_0^{-1}(t)) \\ 1 + F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t)) \end{bmatrix} \tag{B.6}$$

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

$$I(\theta) = \int_{[0,1]} \dot{g}(t)\dot{g}^\top(t)\mathrm{d}t = \frac{1}{\theta_2^2}\begin{bmatrix} \iota_{11} & \iota_{12} \\ \iota_{12} & \iota_{22} \end{bmatrix} = \frac{1}{\theta_2^2} I_0 \tag{B.7}$$

where each ι_{ij} can be derived from equation (B.6) and I_0 is a fixed matrix depending only on the model.

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

$$g(t) = \begin{bmatrix} \frac{-\theta_2}{\theta_1} F_0^{-1}(t) f_0(F_0^{-1}(t)) \\ \frac{1}{\theta_2}\log(F_0^{-1}(t)) F_0^{-1}(t) f_0(F_0^{-1}(t)) \end{bmatrix} \tag{B.8}$$

[1] 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.

and

$$\dot{g}(t) = \begin{bmatrix} \frac{-\theta_2}{\theta_1}\left(1 + F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t))\right) \\ \frac{1}{\theta_2}\left(1 + \log(F_0^{-1}(t)) + \log(F_0^{-1}(t))F_0^{-1}(t)(\dot{f}_0/f_0)(F_0^{-1}(t))\right) \end{bmatrix} \tag{B.9}$$

so that

$$I(\theta) = \begin{bmatrix} \frac{\theta_2^2}{\theta_1^2}\sigma_{11} & \frac{-1}{\theta_1}\sigma_{12} \\ \frac{-1}{\theta_1}\sigma_{12} & \frac{1}{\theta_2^2}\sigma_{22} \end{bmatrix} \tag{B.10}$$

Consider the third term in (1.13):

$$g^\top(s)\left(\int_0^1 \dot{g}(r)\dot{g}^\top(r)\mathrm{d}r\right)^{-1} g(t). \tag{B.11}$$

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 (B.3) and (B.4), 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,

$$h(s,\theta)^\top\left(\int_s^1 h(s,\theta)h^\top(s,\theta)\mathrm{d}s\right)^{-1}\int_s^1 h(r,\theta)\mathrm{d}\mathbb{F}_n(r) \tag{B.12}$$

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

Appendix C

Properties of the covariance function ρ_j

For all orders j, the covariance function is continuous. Here we show that for orders $j \geq 2$, the covariance function is also differentiable. Assuming $s \leq t$, the derivatives $\frac{\partial}{\partial s}\rho_j$ and $\frac{\partial}{\partial t}\rho_j$ exist and are given by

$$\begin{aligned}\frac{\partial}{\partial s}\rho_j(s,t,F) = &\binom{2j-2}{j-1}\mathcal{I}_{2j-2}(s,F) \\ &+\sum_{\ell=1}^{j-1}\binom{2j-\ell-2}{j-1}\left[\frac{(t-s)^\ell}{\ell!}\mathcal{I}_{2j-\ell-2}(s,F) - \frac{(t-s)^{\ell-1}}{(\ell-1)!}\mathcal{I}_{2j-\ell-1}(s,F)\right] - \mathcal{I}_{j-1}(s,F)\mathcal{I}_j(t,F)\end{aligned} \tag{C.1}$$

and

$$\frac{\partial}{\partial t}\rho_j(s,t,F) = \sum_{\ell=1}^{j-1}\binom{2j-\ell-2}{j-1}\frac{(t-s)^{\ell-1}}{(\ell-1)!}\mathcal{I}_{2j-\ell-1}(s,F) - \mathcal{I}_j(s)\mathcal{I}_{j-1}(t,F) \tag{C.2}$$

and it can be checked that these functions agree with one another when $s \nearrow t$. A more formal proof of differentiability everywhere is given in Theorem 12.

Theorem 12. *Let $j \geq 2$. $\rho_j(s,t,F)$ defined in (3.8) is differentiable in s and t.*

Proof. Note that

$$\begin{aligned}\rho_j(s+h,s,F) - \rho_j(s,s,F) = &\sum_{\ell=0}^{j-1}\binom{2j-\ell-2}{j-1}\frac{|h|^\ell}{\ell!}\mathcal{I}_{2j-\ell-1}(s\wedge(s+h),F) \\ &- \mathcal{I}_j(s,F)\mathcal{I}_j(s+h,F) - \binom{2j-2}{j-1}\mathcal{I}_{2j-1}(s,F) - \mathcal{I}_j(s,F)\mathcal{I}_j(s,F)\end{aligned} \tag{C.3}$$

which implies that

$$\begin{aligned}\lim_{h\to 0}\frac{\rho_j(s+h,s,F)-\rho_j(s,s,F)}{h} = &\lim_{h\to 0}\frac{1}{h}\sum_{\ell=1}^{j-1}\binom{2j-\ell-2}{j-1}\frac{|h|^\ell}{\ell!}\mathcal{I}_{2j-\ell-1}(s\wedge(s+h),F) \\ &+\binom{2j-2}{j-1}\lim_{h\to 0}\left[\frac{\mathcal{I}_{2j-1}(s\wedge(s+h),F)-\mathcal{I}_{2j-1}(s,F)}{h}\right] - \mathcal{I}_j(s,F)\lim_{h\to 0}\left[\frac{\mathcal{I}_j(s+h,F)-\mathcal{I}_j(s,F)}{h}\right].\end{aligned} \tag{C.4}$$

Consider the case $h > 0$: then only the term corresponding to $\ell = 1$ in the sum is nonzero, the second term in (C.4) is equal to 0 and the third term involves the derivative $\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{I}_j(s,F) = \mathcal{I}_{j-1}(s,F)$ by definition. This implies

$$\lim_{h\to 0^+} \frac{\rho_j(s+h,s,F) - \rho_j(s,s,F)}{h} = \binom{2j-3}{j-1}\mathcal{I}_{2j-2}(s,F) - \mathcal{I}_j(s,F)\mathcal{I}_{j-1}(s,F). \tag{C.5}$$

In case $h < 0$, the second term is also a derivative and we have

$$\lim_{h\to 0^-} \frac{\rho_j(s+h,s,F) - \rho_j(s,s,F)}{h} = -\binom{2j-3}{j-1}\mathcal{I}_{2j-2}(s,F) + \binom{2j-2}{j-1}\mathcal{I}_{2j-2}(s,F) - \mathcal{I}_j(s,F)\mathcal{I}_{j-1}(s,F) \tag{C.6}$$

$$= \binom{2j-3}{j-1}\mathcal{I}_{2j-2}(s,F) - \mathcal{I}_j(s,F)\mathcal{I}_{j-1}(s,F) \tag{C.7}$$

because

$$\binom{2j-2}{j-1} = \frac{2j-2}{j-1}\frac{(2j-3)!}{(j-1)!(j-2)!} = 2\binom{2j-3}{j-1}. \tag{C.8}$$

Because ρ_j is a covariance function, $\rho_j(s,t,F) = \rho_j(t,s,F)$ and we have the result. ■

Appendix D

Proof of results in the text

Proof of Theorem 1: Durbin's approximation P_g in (1.22) 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 θ as an argument in the functions g and I)

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

and the second derivatives are

$$\rho_{11}(s,t) = -\ddot{g}^\top(s) I_\theta^{-1} g(t), \qquad \rho_{12}(s,t) = -\dot{g}^\top(s) I_\theta^{-1} \dot{g}(t) \qquad \rho_{22}(s,t) = -g^\top(s) I_\theta^{-1} \ddot{g}(t). \tag{D.2}$$

When evaluated at $s = t$, we find that $\rho_{11}(t,t) = \rho_{22}(t,t)$, and their existence is implied by the existence of $\ddot{g}$, which in turn is implied by the above assumption on the density of the model, because the second derivative of g involves derivative terms up to $\frac{\partial^3 F(x,\theta)}{\partial x^2 \partial \theta}\Big|_{x=F^{-1}(t,\theta)}$.

By the definition of t_0,

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma^2(t)\Big|_{t=t_0} = \rho_1(t_0,t_0) + \rho_2(t_0,t_0) = 0. \tag{D.3}$$

We also have, from (D.1),

$$\rho_1(t,t) - \rho_2(t,t) = 1 \tag{D.4}$$

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

$$\rho_1(t_0,t_0) = -\rho_2(t_0,t_0) = 1/2. \tag{D.5}$$

Inserting (D.5) and (D.2) into (1.22), we have the result. ■

Proof of Theorem 2: Because θ is estimated by maximum likelihood, the covariance function of $\hat{v}$ is (1.13),

which implies that

$$\sigma^2(t) = t - t^2 - g^\top(t) I^{-1} g(t) \tag{D.6}$$

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

$$\sigma(t) = \sigma(t_0) + \frac{1}{2(2k)!\sigma(t_0)} \frac{\mathrm{d}^{(2k)}}{\mathrm{d}t^{(2k)}} \sigma^2(t_0) |t - t_0|^{(2k)} (1 + o(1)), \quad t \to t_0 \tag{D.7}$$

because all derivatives of order lower than $2k$ are zero by assumption. By Lemma 1, the correlation function of $\hat{v}$ locally about t_0 has a first-order expansion for all parametric models:

$$r(s,t) = 1 - \frac{1}{2\sigma^2(t_0)} |t - s| (1 + o(1)), \quad s, t \to t_0. \tag{D.8}$$

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

$$\mathrm{P}\left\{ \sup_{t \in [0,1]} \hat{v}(t) > a \right\} = H(\sigma, k) \left(\frac{a}{\sigma(t_0)} \right)^{2 - 1/k} \Psi\left(\frac{a}{\sigma(t_0)} \right) (1 + o(1)), \quad a \to \infty \tag{D.9}$$

where

$$H(\sigma, k) = \int_{\mathbb{R}} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t \tag{D.10}$$

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

$$H(\sigma, k) = \int_{\mathbb{R}} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t = 2 \int_{[0,\infty)} e^{-\left(\frac{A}{C} t\right)^{2k}} \mathrm{d}t = \frac{C}{kA} \Gamma\left(\frac{1}{2k} \right) \tag{D.11}$$

Finally use the relation

$$a\Psi(a) = \phi(a)(1 + o(1)) \tag{D.12}$$

in (D.9) to establish the result. ■

Lemma 1. *Let $\hat{v}$ have covariance function ρ as in (1.12) or (1.13) and correlation function $r(s,t) =$*

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

$$r(s,t) = 1 - \frac{1}{2\sigma^2(t_0)}|t-s|(1+o(1)), \quad s,t \to t_0 \tag{D.13}$$

Proof of Lemma 1: Expanding the squared covariance function $\rho^2(s,t)$ in s around t results in

$$\rho^2(s,t) = \rho^2(t,t) + 2\rho(t,t)\rho_1(t,t)(s-t)(1+o(1)), \quad s \to t, \tag{D.14}$$

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

$$\rho(s,s) = \rho(t,t) + [\rho_1(t,t) + \rho_2(t,t)](s-t)(1+o(1)), \quad s \to t. \tag{D.15}$$

This implies that

$$\begin{aligned}
\rho^2(s,t) - \rho(s,s)\rho(t,t) &= \rho^2(t,t) + 2\rho(t,t)\rho_1(t,t)(s-t) \\
&\quad - \rho^2(t,t) - \rho(t,t)[\rho_1(t,t) - \rho_2(t,t)](s-t) + o(s-t), \quad s \to t \\
&= \rho(t,t)[\rho_1(t,t) - \rho_2(t,t)](s-t) + o(s-t) \\
&= \rho(t,t)(s-t)(1+o(1)), \quad s \to t,
\end{aligned} \tag{D.16}$$

this last equality occurring because $\rho_1(t,t) - \rho_2(t,t) = 1$ 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

$$= -\sigma^2(t_0)|t-s|(1+o(1)), \quad s,t \to t_0. \tag{D.17}$$

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

$$\begin{aligned}
r(s,t) &= \sqrt{1 - \frac{\sigma^2(t_0)}{\sigma^2(s)\sigma^2(t)}|t-s|(1+o(1))} \\
&= 1 - \frac{1}{2\sigma^2(t_0)}|t-s|(1+o(1)), \quad s,t \to t_0.
\end{aligned} \tag{D.18}$$

■

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 (1.34). Namely, because y is Marko-

vian,

$$\mathrm{P}\left\{y_t \in B\right\} = \int_0^t \mathrm{P}\left\{y_t \in B | y_s = a\right\} \mathrm{d}F(s) \tag{D.19}$$

for all measurable $B \subseteq [a, \infty)$. Given the distributions (1.34),

$$\mathrm{P}\left\{y_t \in [a, \infty)\right\} = \Psi\left(\frac{a}{\sqrt{\rho(t,t)}}\right) \tag{D.20}$$

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

$$\mathrm{P}\left\{y_t \in [a, \infty) | y_s = a\right\} = \Psi\left(\frac{a - m(s,t)}{\sqrt{V(s,t)}}\right) \tag{D.21}$$

where m and V are defined above. The distribution of τ_a has a density because of the relationship between Brownian motion and y, that is, equation (1.36). ■

Proof of Theorem 6. Because $\mathcal{I}_j$ is a linear operator,

$$V_{jmn}(\cdot) = \mathcal{I}_j\left(\cdot, \sqrt{\frac{nm}{n+m}}(\mathbb{F}_{2m} - \mathbb{F}_{1n})\right) \tag{D.22}$$

$$= \mathcal{I}_j\left(\cdot, \sqrt{\frac{nm}{n+m}}\left((\mathbb{F}_{2m} - F) - (\mathbb{F}_{1n} - F)\right)\right). \tag{D.23}$$

The example in van der Vaart (1998) shows that the family $\mathcal{F} = \{(-\infty, t] : t \in \mathbb{R}\}$ is a Donsker class for the measure P associated with distribution function F. Therefore we have the weak convergence of $\sqrt{n}(\mathbb{F}_{1n} - F) \rightsquigarrow B_F^1$ and $\sqrt{m}(\mathbb{F}_{2m} - F) \rightsquigarrow B_F^2$, where the limiting processes are independent realizations of the Brownian bridge by assumed independence of X_1 and X_2. Given assumption (3.4) (which ensures that the mapping $w(\cdot)\mathcal{I}_j(\cdot, F)$ is bounded and thus continuous), the continuous mapping theorem implies

$$w(\cdot)\mathcal{I}_j\left(\cdot, \sqrt{\frac{nm}{n+m}}\left((\mathbb{F}_{1n} - F) - (\mathbb{F}_{2m} - F)\right)\right) \rightsquigarrow w(\cdot)\mathcal{I}_j\left(\cdot, \sqrt{\lambda}B_F^1 + \sqrt{1-\lambda}B_F^2\right) \tag{D.24}$$

$$\stackrel{D}{=} w(\cdot)\mathcal{I}_j(\cdot, B_F) \tag{D.25}$$

where "$\stackrel{D}{=}$" means equal in distribution. ■

Proof of Theorem 7. The differentiability of $\rho(s, t, F)$ implies (Azaïs and Wschebor, 2009, p. 29-30) the

variance function of the pair $(V_j(t), V_{j-1}(t))$ is given by

$$\mathrm{E}\left[\begin{bmatrix} V_j(t) \\ V_{j-1}(t) \end{bmatrix}\begin{bmatrix} V_j(t) & V_{j-1}(t) \end{bmatrix}\right] = \begin{bmatrix} \rho_j(t,t,F) & \frac{\partial}{\partial s}\rho_j(s,t,F)|_{s=t} \\ \frac{\partial}{\partial s}\rho_j(s,t,F)|_{s=t} & \rho_{j-1}(t,t,F) \end{bmatrix} \tag{D.26}$$

$$= \begin{bmatrix} \binom{2j-2}{j-1}\mathcal{I}_{2j-1} - \mathcal{I}_j^2 & \binom{2j-3}{j-2}\mathcal{I}_{2j-2} - \mathcal{I}_j\mathcal{I}_{j-1} \\ \binom{2j-3}{j-2}\mathcal{I}_{2j-2} - \mathcal{I}_j\mathcal{I}_{j-1} & \binom{2j-4}{j-2}\mathcal{I}_{2j-3} - \mathcal{I}_{j-1}^2 \end{bmatrix}(t,F) \tag{D.27}$$

The following Rice formula (Azaïs and Wschebor, 2009, Theorem 3.2, Example 1 on p. 79) can be used to evaluate the expected number of upcrossings of V_j to level $c > 0$:

$$\mathrm{E}\left[U_c\right] = \int_0^1 \mathrm{E}\left[V_{j-1}^+ | V_j = c\right] p_j(t,c)\mathrm{d}t \tag{D.28}$$

where $V_{j-1}^+ = 0 \vee V_{j-1}$ is the positive part of V_{j-1}. It can be verified that if $x \sim \mathcal{N}(\mu, \sigma^2)$ for any parameter pair (μ, σ^2),

$$\int_{\mathbb{R}} x^+ \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\mathrm{d}x = \mu\Psi(-\mu/\sigma) + \sigma\phi(-\mu/\sigma) \tag{D.29}$$

$$= \mu\Phi(\mu/\sigma) + \sigma\phi(\mu/\sigma). \tag{D.30}$$

Using a regression formula expression (cf. Azaïs and Wschebor (2009, Proposition 1.2, p. 15), Piterbarg (1996, p. 2-3)), we find that the conditional distribution of V_{j-1} given V_j is

$$V_{j-1}(t)\Big|\left\{V_j(t) = c\right\} \sim \mathcal{N}\left(c\frac{\mathrm{Cov}\left(V_{j-1}(t), V_j(t)\right)}{\mathrm{Var}\left(V_j(t)\right)}, \mathrm{Var}\left(V_{j-1}(t)\right) - \frac{\mathrm{Cov}^2\left(V_{j-1}(t), V_j(t)\right)}{\mathrm{Var}\left(V_j(t)\right)}\right). \tag{D.31}$$

For notational convenience, rewrite this distribution as

$$V_{j-1}(t)\Big|\left\{V_j(t) = c\right\} \sim \mathcal{N}\left(cM(t), \Sigma^2(t)\right). \tag{D.32}$$

Then we can write the Rice formula (D.28) as

$$\mathrm{E}\left[U_c\right] = \int_{-\infty}^{\infty}\left(cM(t)\Phi\left(c\frac{M(t)}{\Sigma(t)}\right) + \Sigma(t)\phi\left(c\frac{M(t)}{\Sigma(t)}\right)\right)p_j(c,t)\mathrm{d}t \tag{D.33}$$

which is the statement of the theorem. ■

Simple properties of the variance function (3.24) imply Lemma 2.

Lemma 2 (The point of maximal variance)**.** *Let $j \geq 2$. Then letting σ_j^2 be the variance function of $B^{(j)}$,*

$$\operatorname*{argmax}_{t\in[0,1]} \sigma_j^2(t) = 1 \qquad \textit{and} \qquad \sigma_j^2(1) = \frac{(j-1)^2}{(2j-1)(j!)^2}. \tag{D.34}$$

Furthermore, $\sigma_j^2(t)$ has the following expansion near $t = 1$:

$$\sigma_j^2(t) = \begin{cases} \frac{1}{12} - \frac{1}{2}(1-t)^2(1+o(1)), \quad t \nearrow 1 & \textit{if } j = 2 \\ \frac{(j-1)^2}{(j!)^2(2j-1)} - \frac{j(j-2)}{(j!)^2}(1-t)(1+o(1)) \quad t \nearrow 1 & \textit{if } j \geq 3. \end{cases} \tag{D.35}$$

Proof of Lemma 2. Note that

$$\sigma_j^2(t) = \binom{2j-2}{j-1} \frac{t^{2j-1}}{(2j-1)!} - \frac{t^{2j}}{(j!)^2} = \frac{j^2 t^{2j-1} - (2j-1)t^{2j}}{(j!)^2(2j-1)}. \tag{D.36}$$

Either by taking a derivative and showing that it is nonnegative on $[0,1]$ or recalling that the derivative $\frac{\mathrm{d}}{\mathrm{d}t}\sigma_j^2(t) = \sigma_{j-1}^2(t) \geq 0$, we find that the function is monotonically increasing and the maximum value must be achieved at $t = 1$; evaluating the function at 1 proves the first part of the Lemma. Taking a first-order expansion we find

$$\sigma_j^2(t) = \sigma_j^2(1) - \frac{j^2 t^{2j-2} - 2jt^{2j-1}}{(j!)^2}\Big|_{t=1} (1-t)(1+o(1)), \quad t \nearrow 1 \tag{D.37}$$

$$= \sigma_j^2(1) - \frac{j(j-2)}{(j!)^2}(1-t)(1+o(1)), \quad t \nearrow 1. \tag{D.38}$$

This implies a first-order expansion is sufficient to describe the function when j (an integer) is greater than 2. For the $j = 2$ case it is necessary to use a second order expansion, and one finds

$$\sigma_2^2(t) = \sigma_2^2(1) + \frac{1}{2}\frac{j^2(2j-2)t^{2j-3} - 2j(2j-1)t^{2j-2}}{(j!)^2}\Big|_{t=1,j=2} (1-t)^2(1+o(1)), \quad t \nearrow 1 \tag{D.39}$$

$$= \frac{1}{12} - \frac{1}{2}(1-t)^2(1+o(1)), \quad t \nearrow 1. \tag{D.40}$$

■

We also note the following lemma for the expansion of the correlation function for the case of $j \geq 3$. For $j = 2$ this does not apply, and because of this fact we give asymptotically exact critical values for $j \geq 3$ below but only a bound for the $j = 2$ case.

Lemma 3. *Let $j \geq 3$. The correlation function of $B^{(j)}$ can be expanded in a neighborhood of 1 as*

$$\operatorname{Corr}\left(B^{(j)}(s), B^{(j)}(t)\right) = 1 - K|t-s|^2(1+o(1)), \quad s,t \to 1, \tag{D.41}$$

with $K > 0$.

Proof of Lemma 3. The method of proof is the same as in Chapter 1.For the moment suppressing dependence on the order j, call r the covariance function of $B^{(j)}$. We expand $r^2(s,s)$ and $r(s,t)$ in s around t to two terms and subtract one from the other to arrive at the expression

$$\begin{aligned} r^2(s,t) - r(s,s)r(t,t) &= \left(\frac{\partial}{\partial s}r(s,t) - \frac{\partial}{\partial t}r(s,t)\right)\Big|_{s=t}(s-t) \\ &+ \left(\left(\frac{\partial}{\partial s}r(s,t)\right)^2 - r(s,t)\frac{\partial^2}{\partial s \partial t}r(s,t) + \frac{1}{2}r(s,t)\left(\frac{\partial^2}{\partial s^2}r(s,t) - \frac{\partial^2}{\partial t^2}r(s,t)\right)\right)\Big|_{s=t}(s-t)^2(1+o(1)) \end{aligned} \tag{D.42}$$

First note that

$$\frac{\partial}{\partial s}r(s,t)\Big|_{s=t} = \frac{\partial}{\partial t}r(s,t)\Big|_{s=t} \qquad \text{and} \qquad \frac{\partial^2}{\partial s^2}r(s,t)\Big|_{s=t} = \frac{\partial^2}{\partial t^2}r(s,t)\Big|_{s=t} \tag{D.43}$$

so that the above expression is equivalent to

$$r^2(s,t) - r(s,s)r(t,t) = \left(\left(\frac{\partial}{\partial s}r(s,t)\right)^2 - r(s,t)\frac{\partial^2}{\partial s \partial t}r(s,t)\right)\Big|_{s=t}(s-t)^2(1+o(1)) \tag{D.44}$$

Also note that

$$\binom{2j-2}{j-1} = 2\binom{2j-3}{j-1} \qquad \text{and} \qquad \binom{2j-3}{j-1} = \frac{2j-3}{j-2}\binom{2j-4}{j-1}, \tag{D.45}$$

which implies

$$\binom{2j-3}{j-1} - \binom{2j-4}{j-1} = \left(\frac{2j-3}{j-2} - 1\right)\binom{2j-4}{j-1} = \frac{j-1}{j-2}\binom{2j-4}{j-1}. \tag{D.46}$$

Then the coefficient in expansion (D.44) is

$$\left(\left(\frac{\partial}{\partial s}r(s,t)\right)^2 - r(s,t)\frac{\partial^2}{\partial s\partial t}r(s,t)\right)\Big|_{s=t} = \left(\binom{2j-3}{j-1}\mathcal{I}_{2j-2}(t) - \mathcal{I}_{j-1}(t)\mathcal{I}_j(t)\right)^2 - \left(\binom{2j-2}{j-1}\mathcal{I}_{2j-1}(t) - \mathcal{I}_j^2(t)\right)\left(\frac{j-1}{j-2}\binom{2j-4}{j-1}\mathcal{I}_{2j-3}(t) - \mathcal{I}_{j-1}^2(t)\right) \quad \text{(D.47)}$$

Then using the value of $\mathcal{I}_j(1) = 1/j!$ for the uniform distribution, we have

$$\left(\left(\frac{\partial}{\partial s}r(s,t)\right)^2 - r(s,t)\frac{\partial^2}{\partial s\partial t}r(s,t)\right)\Big|_{s=t=1} = \left(\binom{2j-3}{j-1}\frac{1}{(2j-2)!} - \frac{1}{(j-1)!}\frac{1}{j!}\right)^2 - \left(\binom{2j-2}{j-1}\frac{1}{(2j-1)!} - \frac{1}{(j!)^2}\right)\left(\frac{j-1}{j-2}\binom{2j-4}{j-1}\frac{1}{(2j-3)!} - \frac{1}{((j-1)!)^2}\right) \quad \text{(D.48)}$$

Using heroic algebra it can be verified that this may be expressed as a fraction with denominator $((j-1)!)^4$. The final expression is

$$\left(\left(\frac{\partial}{\partial s}r(s,t)\right)^2 - r(s,t)\frac{\partial^2}{\partial s\partial t}r(s,t)\right)\Big|_{s=t=1} = \frac{\left(\frac{1}{2}-\frac{1}{j}\right)^2 - \left(\frac{1}{2j-1}-\frac{1}{j^2}\right)\left(\frac{(j-1)^2}{2j-3}-1\right)}{((j-1)!)^4} \quad \text{(D.49)}$$

$$= \frac{\frac{(j-2)^2}{4j^2} - \frac{(j-1)^2(j-2)^2}{j^2(2j-1)(2j-3)}}{((j-1)!)^4} \quad \text{(D.50)}$$

Note that the numerator of this last expression is strictly negative for $j \geq 3$, because the inequality

$$\frac{(j-2)^2}{4j^2} < \frac{(j-1)^2(j-2)^2}{j^2(2j-1)(2j-3)} \quad \text{(D.51)}$$

is equivalent to

$$\frac{1}{4} < \frac{(j-1)^2}{(2j-1)(2j-3)} \quad \text{(D.52)}$$

and this can be verified by noting

$$\frac{(j-1)^2}{4j^2-8j+3} > \frac{(j-2)^2}{4j^2-8j+3} > \frac{(j-2)^2}{4j^2-8j+4} = \frac{(j-2)^2}{4(j-2)^2} = \frac{1}{4}. \quad \text{(D.53)}$$

Using this result in (D.44) we have

$$r^2(s,t) - r(s,s)r(t,t) = C(s-t)^2(1+o(1)), \quad s,t \to 1, \quad \text{(D.54)}$$

where

$$C = \frac{\frac{(j-2)^2}{4j^2} - \frac{(j-1)^2(j-2)^2}{j^2(2j-1)(2j-3)}}{((j-1)!)^4} < 0 \tag{D.55}$$

Using the same steps as above, it can then be shown that

$$\text{Corr}\left(B^{(j)}(s), B^{(j)}(t)\right) = 1 + \frac{C}{2\sigma^4(1)}|t-s|^2(1+o(1)), \quad s,t \to 1 \tag{D.56}$$

$$= 1 - K|t-s|^2(1+o(1)), \quad s,t \to 1 \tag{D.57}$$

where

$$K = \frac{j^2(j-2)^2(2j-1)}{8(j-1)^4(2j-3)} > 0 \tag{D.58}$$

■

Proof of Corollary 2. The variance function of $(B^{(2)}(t), B(t))$ is

$$\text{E}\left[\begin{bmatrix} B^{(2)}(t) \\ B(t) \end{bmatrix}\begin{bmatrix} B^{(2)}(t) & B(t) \end{bmatrix}\right] = \begin{bmatrix} \frac{1}{3}t^3 - \frac{1}{4}t^4 & \frac{t^2}{2}(1-t) \\ \frac{t^2}{2}(1-t) & t(1-t) \end{bmatrix}. \tag{D.59}$$

Specializing the conditional distribution formula in the proof of Theorem 7, to this pair, we find that

$$B(t)|\left\{B^{(2)}(t) = c\right\} \sim \mathcal{N}\left(\frac{6(1-t)}{t(4-3t)}c, \frac{t(1-t)}{4-3t}\right) \tag{D.60}$$

which implies the conditional expectation in the integral is

$$\text{E}\left[B^+(t)\middle|B^{(2)}(t) = c\right] = \frac{6c(1-t)}{t(4-3t)}\Phi\left(\frac{6c}{t^{3/2}}\sqrt{\frac{1-t}{4-3t}}\right) + \sqrt{\frac{t(1-t)}{4-3t}}\phi\left(\frac{6c}{t^{3/2}}\sqrt{\frac{1-t}{4-3t}}\right). \tag{D.61}$$

Using this expression in Theorem 7, we have the statement (3.27). ■

Proof of Theorem 8. The proof relies on methods that are exemplified by Piterbarg (1996). Note that by Lemma 2 the variance has a first-order expansion in a neighborhood of 1 and by Lemma 3 the correlation has a second-order expansion near 1. This implies that part (iii) of Theorem 8.2 of Piterbarg (1996) (equivalently, part (iii) of Theorem D.3) applies to the process $B^{(j)}$. Therefore by normalizing the process by its standard deviation evaluated at 1 we have the result. ■

Proof of Theorem 9. Make the time transformation $r = F_{Y_k|X_k}(t + X_k\beta_{k0}|\theta_{k0}, X_k)$, so that (3.35) is equiv-

alent to the process

$$v_{1n}^k(r,\hat{\theta}_k) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(F_{Y_k|X_k}(y_{ki} - X_{ki}(\hat{\beta}_k - \beta_{k0})|\theta_{k0}, X_k) \leq r) - r\right), \qquad r \in [0,1]. \tag{D.62}$$

In the spirit of van der Vaart and Wellner (2007), define

$$\eta_n(z) = F_{Y_k|X_k}(y - x(\hat{\beta}_k(z) - \beta_{k0})|\hat{\theta}(z), x), \qquad \eta_0(z) = F_{Y_k|X_k}(y|\theta_{k0}, x) \tag{D.63}$$

$$g_{\eta_n,r}(z) = I(\eta_n(z) \leq r), \qquad g_{\eta_0,r}(z) = I(\eta_0(z) \leq r) \tag{D.64}$$

interpreting x as a row vector throughout the proof. For this proof, we follow common notation in the empirical process literature and denote the expectation, empirical measure given a sample Z_k and empirical process at f by

$$Pf = \int f \, dP, \qquad \mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(Z_{ki}), \qquad \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f \tag{D.65}$$

(note that $Pg_{\eta_0,r} = r$ conditional on x). We can then rewrite (D.62) as

$$v_{1n}^k(r,\hat{\theta}_k) = \sqrt{n}\left(\mathbb{P}_n g_{\eta_n,r} - P g_{\eta_0,r}\right). \tag{D.66}$$

Now, adding and subtracting terms (specifically, $\pm\sqrt{n}\mathbb{P}_n g_{\eta_0,r}$, $\pm\sqrt{n}P g_{\eta_0,r}$ and $\pm\sqrt{n}P g_{\eta_n,r}$)

$$= \mathbb{G}_n\left(g_{\eta_n,r} - g_{\eta_0,r}\right) + \mathbb{G}_n g_{\eta_0,r} + \sqrt{n}P\left(g_{\eta_n,r} - g_{\eta_0,r}\right) \tag{D.67}$$

This corresponds to equation (2) of van der Vaart and Wellner (2007). We deal with each term in order.

Example 1 of van der Vaart and Wellner (2007) shows that because the model is assumed to be polynomial in x, the class of functions $g_{\eta,r}$ is a Donsker class, and furthermore that

$$\sup_{r\in[0,1]}\left|\mathbb{G}_n\left(g_{\eta_n,r} - g_{\eta_0,r}\right)\right| \xrightarrow{p} 0. \tag{D.68}$$

The second term of (D.67) can be rewritten as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(F_{Y_k|X_k}(Y_{ki}|\theta_{k0}, X_{ki}) \leq r) - r\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(F_k(\varepsilon_{ki}) \leq r) - r\right) \tag{D.69}$$

which is a uniform empirical process that we call $v(\cdot, \theta_{k0})$. The final term of (D.67), upon rewriting, is

$$\sqrt{n}\left(\mathrm{E}_x\left[F_{Y_k|x}\left(F^{-1}_{Y_k|x}(r|\theta_{k0}, x) + x(\hat{\beta}_k - \beta_{k0})|\theta_{k0}, x\right)\right] - r\right)$$
$$= \sqrt{n}(\hat{\beta}_k - \beta_{k0})^\top \mathrm{E}_x\left[\nabla_\beta F_{Y_k|X_k}(F^{-1}_{Y_k|X_k}(r))\right] + o_P\left(\sqrt{n}\left\|\hat{\beta}_k - \beta_{k0}\right\|\right) \quad \text{(D.70)}$$

using a one-term Taylor expansion about β_{k0}. This equals

$$\sqrt{n}(\hat{\beta}_k - \beta_{k0})^\top \mathrm{E}\left[x^\top f_{Y_k|X_k}(F^{-1}_{Y_k|X_k}(r))\right] + o_P(1) = \sqrt{n}(\hat{\beta}_k - \beta_{k0})^\top \mathrm{E}^\top[x] f_k(F_k^{-1}(r)) + o_P(1), \quad \text{(D.71)}$$

making use of the fact that F_k is assumed to be a location-shift model. Putting together the parts of (D.67) and using assumption **A1** we have that (transforming back using $t = F_k^{-1}(r) + \mu_{k0}$, which is equivalent to the transformation made before)

$$V^k_{1n}(t, \hat{\theta}_k) = V^k_{1n}(t, \theta_{k0}) + \sqrt{n}(\hat{\beta}_k - \beta_{k0})^\top \frac{1}{n} X_k^\top \mathbf{1}_n f_k(t - \mu_{k0}) + o_P(1). \quad \text{(D.72)}$$

Now using assumption **A4**,

$$= V^k_{1n}(t, \theta_{k0}) + \frac{1}{\sqrt{n}} f_k(t - \mu_{k0}) \Gamma_k^{-1} \mathbf{1}_n^\top X_k (X_k^\top X_k)^{-1} X_k^\top S_k + o_P(1) \quad \text{(D.73)}$$

which is (3.35).

The covariance function of the limiting process can be found by direct calculation. In this regard it is helpful to note the following:

$$\mathrm{E}\left[V^k_{1n}(s, \theta_{k0}) V^k_{1n}(t, \theta_{k0})\right] = F_k(s \wedge t - \mu_{k0}) - F_k(s - \mu_{k0}) F_k(t - \mu_{k0}) \quad \text{(D.74)}$$

$$\mathrm{E}\left[S_k S_k^\top\right] = \Gamma_k I_{n \times n} \quad \text{(D.75)}$$

$$\mathrm{E}\left[S_k V^k_{1n}(t, \theta_{k0})\right] = -\frac{1}{\sqrt{n}} f_k(t - \mu_{k0}) \mathbf{1}_n \quad \text{(D.76)}$$

■

Proof of Theorem 10. The first part of the statement is a result of the continuous mapping theorem applied to the corresponding processes in Theorem 9. The covariance function of the process can be found via direct calculation. Recall that $\mathrm{Cov}\left(V^k_{jn}(s, \theta_{k0}), V^k_{jn}(t, \theta_{k0})\right) = \rho_j(s, t, F)$ and $\mathrm{E}\left[S_k S_k^\top\right] = \Gamma_k I_{n \times n}$. Finally, we calculate $\mathrm{E}\left[S_k V^k_{jn}(t, \theta_{k0})\right]$: by independence of the observations, we show the calculations

below with a single observation ε_{ki} for some i and k. The required expectation is

$$\mathrm{E}\left[-\frac{\dot{f}_k(\varepsilon_{ki})}{f_k(\varepsilon_{ki})}V_{jn}^k(t,\theta_{k0})\right] = \mathrm{E}\left[-\frac{\dot{f}_k(\varepsilon_{ki})}{f_k(\varepsilon_{ki})}\frac{1}{\sqrt{n}}\left(\frac{(t-\varepsilon_{ki})^{j-1}}{(j-1)!}I(\varepsilon_{ki}\leq t) - \int_{-\infty}^{t}\frac{(t-s)^{j-1}}{(j-1)!}\mathrm{d}F(s)\right)\right] \quad \text{(D.77)}$$

Because the score function has expectation zero,

$$= -\frac{1}{\sqrt{n}}\mathrm{E}\left[\frac{\dot{f}_k(\varepsilon_{ki})}{f_k(\varepsilon_{ki})}\frac{(t-\varepsilon_{ki})^{j-1}}{(j-1)!}I(\varepsilon_{ki}\leq t)\right] \quad \text{(D.78)}$$

$$= -\frac{1}{\sqrt{n}}\int_{-\infty}^{t}\frac{(t-\varepsilon_{ki})^{j-1}}{(j-1)!}\dot{f}_k(\varepsilon_{ki})\mathrm{d}\varepsilon_{ki} \quad \text{(D.79)}$$

$$= -\frac{1}{\sqrt{n}}\mathcal{I}_j(t,f_k) \quad \text{(D.80)}$$

and in some situations (outlined in the text)

$$= -\frac{1}{\sqrt{n}}\mathcal{I}_{j-1}(t,F_k). \quad \text{(D.81)}$$

Therefore, we have that

$$\mathrm{E}\left[S_k V_{jn}^k(t,\beta_{k0})\right] = -\frac{1}{\sqrt{n}}\mathcal{I}_j(t,f_k)\mathbf{1}_n. \quad \text{(D.82)}$$

We note that this is very similar to the result found by Linton et al. (2005), although they express it in terms of $\nabla_\beta \mathcal{I}_j(t,F_{Y_k|X_k})$. In this case they are equivalent because we assume that integration (in t) and differentiation (in β) can be interchanged. ■

Proof of Theorem 11. Under the hypothesis that $F_1 \equiv F_2$, we can rewrite the process V_{jmn} by adding and subtracting $\sqrt{\frac{nm}{n+m}}\mathcal{I}_j(t,F_k)$,

$$V_{jmn}(t,\hat{\theta}_1,\hat{\theta}_2) = \sqrt{\frac{nm}{n+m}}\left(\left(\mathcal{I}_j(t,\mathbb{F}_{2m}(\cdot,\hat{\theta}_2) - \mathcal{I}_j(t,F)\right) - \left(\mathcal{I}_j(t,\mathbb{F}_{1n}(\cdot,\hat{\theta}_1)) - \mathcal{I}_j(t,F)\right)\right) \quad \text{(D.83)}$$

$$= \sqrt{\frac{n}{n+m}}V_{jm}^2(t,\hat{\theta}_2) - \sqrt{\frac{m}{n+m}}V_{jn}^1(t,\hat{\theta}_1). \quad \text{(D.84)}$$

Applying Theorem 10 to both of the above terms, this implies

$$\begin{aligned} V_{jmn}(t,\hat{\theta}_1,\hat{\theta}_2) = {} & \sqrt{\frac{n}{n+m}}V_{jm}^2(t,\theta_{20}) - \sqrt{\frac{m}{n+m}}V_{jn}^1(t,\theta_{10}) \\ & + \frac{1}{\sqrt{m}}\sqrt{\frac{n}{n+m}}\mathcal{I}_j(t,f)\Gamma^{-1}\mathbf{1}_m^\top P_{X_2}S_2 - \frac{1}{\sqrt{n}}\sqrt{\frac{m}{n+m}}\mathcal{I}_j(t,f)\Gamma^{-1}\mathbf{1}_n^\top P_{X_1}S_1 + o_P(1). \end{aligned} \quad \text{(D.85)}$$

Using $\frac{m}{m+n} = \lambda + o(1)$ we have (3.45). Once again the covariance function can be checked directly using the results noted in the proofs of Theorems 9 and 10. However, the expectation of $V_{jmn}(t, \theta_{10}, \theta_{20})$ multiplied with each score vector is different:

$$\mathrm{E}\left[S_1 V_{jmn}(t, \theta_{10}, \theta_{20})\right] = \frac{\sqrt{\lambda}}{\sqrt{n}} \mathcal{I}_j(t, f)\mathbf{1}_n \tag{D.86}$$

and

$$\mathrm{E}\left[S_2 V_{jmn}(t, \theta_{10}, \theta_{20})\right] = -\frac{\sqrt{1-\lambda}}{\sqrt{m}} \mathcal{I}_j(t, f)\mathbf{1}_m. \tag{D.87}$$

■

Proof of Corollary 3. This result is not technically a direct result of Theorem 7; rather, the calculations are the same but use the correlation functions for conditional models given in (3.46) instead of the functions (3.8). We note that the variance function of the pair $(\hat{V}_j(t), \hat{V}_{j-1}(t))$ is

$$\mathrm{E}\left[\begin{bmatrix} \hat{V}_j(t) \\ \hat{V}_{j-1}(t) \end{bmatrix} \begin{bmatrix} \hat{V}_j(t) & \hat{V}_{j-1}(t) \end{bmatrix}\right] = \begin{bmatrix} \binom{2j-2}{j-1}\mathcal{I}_{2j-1} - \mathcal{I}_j^2 - \mathcal{I}_{j-1}^2\Gamma^{-1}\overline{P}_\lambda & \binom{2j-3}{j-2}\mathcal{I}_{2j-2} - \mathcal{I}_j\mathcal{I}_{j-1} - \mathcal{I}_{j-1}\mathcal{I}_{j-2}\Gamma^{-1}\overline{P}_\lambda \\ \binom{2j-3}{j-2}\mathcal{I}_{2j-2} - \mathcal{I}_j\mathcal{I}_{j-1} - \mathcal{I}_{j-1}\mathcal{I}_{j-2}\Gamma^{-1}\overline{P}_\lambda & \binom{2j-4}{j-2}\mathcal{I}_{2j-3} - \mathcal{I}_{j-1}^2 - \mathcal{I}_{j-2}^2\Gamma^{-1}\overline{P}_\lambda \end{bmatrix}(t, F) \tag{D.88}$$

where $\overline{P}_\lambda = \lambda\overline{P}_1 + (1-\lambda)\overline{P}_2$. From this one can construct analogous conditional mean and variance functions for $\hat{V}_{j-1}$ given $\hat{V}_j$ and plug them into the Rice formula to arrive at the result. The definition of the marginal density is slightly different than in the model that uses V_j: the marginal variance of $\hat{V}_j$ is $\rho_j(t, t, F) - \mathcal{I}_{j-1}^2(t, F)\Gamma^{-1}\overline{P}_\lambda$ as specified in Theorem 11. ■

References

R. Adler. *The Geometry of Random Fields*. Wiley, 1981.

R. Adler. On excursion sets, tube formulas and maxima of random fields. *Annals of Applied Probability*, 10(1):1–74, 2000.

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

J.-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.

G. Barrett and S. Donald. Consistent tests for stochastic dominance. *Econometrica*, 71:71–104, 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.

J. Cox, J. Ingersoll, and S. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2): 385–407, 1985.

S. Csörgő. Kernel-transformed empirical processes. *Journal of Multivariate Analysis*, 13(4):517–533, 1983.

R. Davidson and J.-Y. Duclos. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica*, 68(6):1435–1464, 2000.

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.

N. Gürtler and N. Henze. Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Annals of the Institute of Statistical Mathematics*, 52(2):267–286, 2000.

B. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430, 1996.

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.

L. Horváth, P. Kokoszka, and R. Zitikis. Testing for stochastic dominance using the weighted McFadden-type statistic. *Journal of Econometrics*, 133:191–205, 2006.

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.

L. Klecan, R. McFadden, and D. McFadden. A robust test for stochastic dominance. Working paper, Department of Economics, MIT, 1991.

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.

O. Linton, E. Maasoumi, and Y.-J. Whang. Consistent testing for stochastic dominance under general sampling schemes. *Review of Economic Studies*, 72:735–765, 2005.

O. Linton, K. Song, and Y.-J. Whang. An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, 154:186–202, 2010.

M. Loève. *Probability Theory*, volume II. Springer, 1978.

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.

M. Matsui and A. Takemura. Goodness-of-fit tests for symmetric stable distributions — empirical characteristic function approach. *Test*, 17(3):546–566, 2008.

D. McFadden. Testing for stochastic dominance. In T. Fomby and T.K. Seo, editors, *Studies in the Economics of Uncertainty*, pages 113–134. Springer, 1989. In honor of J. Hadar.

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.

S. Meintanis and J. Swanepoel. Bootstrap goodness-of-fit tests with estimated parameters based on empirical transforms. *Statistics & Probability Letters*, 77:1004–1013, 2007.

H. Milbrodt and H. Strasser. On the asymptotic power of the two-sided Kolmogorov-Smirnov test. *Journal of Statistical Planning and Inference*, 26(1):1–23, 1990.

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.

S. Portnoy and R. Koenker. Adaptive L-estimation for linear models. *The Annals of Statistics*, 17(1): 362–381, 1989.

J. Potthoff. Sample properties of random fields III — differentiability. E-Print, University of Mannheim, 2008.

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

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

F. Schmid and M. Trede. A Kolmogorov-type test for second-order stochastic dominance. *Statistics and Probability Letters*, 37:183–193, 1998.

S. Shen. Tests for distributional partial effects. Unpublished Ph.D. thesis, 2011.

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

K. Singleton. Estimation of affine asset pricing models using the empirical characteristic function. *Journal of Econometrics*, 102(1):111–141, 2001.

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

W. Stute. Nonparametric model checks for regression. *The Annals of Statistics*, 25(2):613–641, 1997.

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

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

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.