Tailor-made Tests for Goodness-of-Fit to Semiparametric Hypotheses

Peter J. Bickel*
University of California at Berkeley

Ya’acov Ritov*
The Hebrew University of Jerusalem

Thomas M. Stoker
Massachusetts Institute of Technology

November 27, 2003

Abstract

We introduce a new framework for constructing tests of a general semiparametric hypotheses which
(i) have non-trivial power on the $n^{-1/2}$ scale in every direction.
(ii) can be tailored to put substantial power on alternatives of importance.
The approach is based on combining test statistics based on stochastic processes of score statistics with bootstrap critical values.

1. Introduction  The practice of statistical testing plays several roles in empirical research. These roles range from the careful assessment of the evidence against specific scientific hypotheses to the judgment of whether an estimated model displays decent goodness-of-fit to the empirical data. The paradigmatic situation we consider is one where the investigator views some departures from the hypothesized model as being of primary importance, with others of interest if sufficiently gross, but otherwise secondary. For instance, low frequency departures from a signal hypothesized to be constant might be considered of interest, even if of low amplitude; while high frequency departures are less so unless they are of high amplitude.

The optimal testing of a simple hypothesis against a simple alternative is the corner stone of modern statistical theory. However, there is no clear notion of optimality for more complicated situations. The Hajek-Le Cam asymptotic theory proved that there exists a strong concept of asymptotic efficiency in parametric estimation. This idea was extended to semiparametric models – see Pfanzagl and

*Research partially supported by NSF Grant FD01-04075.
Wefelmayer (1982), Ibragimov and Hasminskii (1981) and Bickel, Klassen, Ritov and Wellner (1993). However, there is no compelling sense of asymptotically optimum test, either in the parametric or semiparametric asymptotic theories, save in one parameter hypotheses.

We deal exclusively with the “elementary” case of i.i.d. data for ease of exposition and to avoid technicalities and all our considerations are asymptotic save for illustrative simulations. Generalization of this point of view to the independent non-identically distributed case, time series, etc., is conceptually not difficult. But the asymptotic theory requires all sorts of assumptions which obscure the basic heuristics we primarily want to transmit. We will focus on situations where the alternative is nonparametric, i.e., all departures from the hypothesis are potentially possible (within the i.i.d. framework).

The general types of tests that people have constructed fall into one of two classes

(i) Those which have non-negligible asymptotic power at the $n^{-\frac{1}{2}}$ scale in every possible direction. In the standard problems of testing goodness of fit to a single distribution against all alternatives, these are the classical tests of Kolmogorov and Cramer-von Mises and their classical extensions to the problem of testing fit to a parametric hypothesis on the one hand and independence on the other. These tests and their various modifications are all built around the empirical d.f.

(ii) Those which have trivial asymptotic power at the $n^{-\frac{1}{2}}$ scale in every direction. The $\chi^2$ tests with increasing number of cells as $n \to \infty$ are the preeminent example of this type, but a number of variants have recently been explored through devices such as empirical likelihood – see Fan, Zhang and Zhang (2001) for recent examples.

Tests of type (ii) have the feature that they have approximately equal power in all directions. As a consequence, they can enjoy minimax properties over suitable nonparametric families of alternatives – see Ingster (1993) for example. But, as we noted, they pay for this by not having power at rate $n^{-\frac{1}{2}}$ in any particular direction. The tests of type (i) considered so far have the weakness that they concentrate their power at the $n^{-\frac{1}{2}}$ scale in a very explicit alternative direction, dictated primarily by the metric used. For example, the Kolmogorov test for goodness of fit to the uniform $(0,1)$ distribution is well known to concentrate on alternatives such that $|P[X \leq \frac{1}{2}] - \frac{1}{2}|$ is large.

The principal reason for limiting oneself to tests of types (i) and (ii) appears to have been the need for simple approximations to the critical values under the null which need to be coupled with specification of a test statistic to implement a test. However, it has always been clear for a simple hypothesis that simulation could be used in conjunction with any test statistic. With composite parametric hypotheses simulation under a consistent estimate of the parameter governing the null (if $H$ is true) is a natural extension of this approach. More generally if the hypothesis is semiparametric and a suitably consistent estimate of the null distribution (if $H$ is true) can be constructed then again simulation becomes a
possibility. Of course, simulation from an arbitrary distribution is not simple. How-
ever, in these cases, as has been explored in Bickel, Götze and van Zwet (1997), Bickel and Ren (1996) it is also possible to use the \( m \) out of \( n \) bootstrap, simulate the distribution of the statistic for subsamples of size \( m \) drawn from the original sample where \( m \to \infty, \frac{m}{n} \to 0 \). Finally, for rather general classes of statistics Bickel and Ren (2000) show how the ordinary Efron (\( n \) out of \( n \)) bootstrap can also be used for the same purpose.

Our goal in this paper is to show that it is possible, not just in the classical problems, but in testing any semiparametric hypothesis, to construct tests which have as much power as possible at the \( n^{-1/2} \) scale in a few directions of interest, reserving some power for gross departures (in the \( n^{-\frac{3}{2}} \) scale) for other directions.

We also briefly compare and contrast our point of view with the minimax and adaptive minimax testing point of view of Ingster (1993). Our proposal does not aim at minimaxity and since we concentrate on the \( n^{-1/2} \) scale our tests do not have uniformity properties except over relatively small families. We believe that in testing, even more than in estimation, prior information or biases need to be paid attention to, since, as Janssen (2000) points out, achieving reasonable power over more than a few orthogonal directions is hopeless.

There has been another direction that we want to mention but do not develop in this paper. The idea is to construct a sequence of tests which are consistent against broader and broader classes of alternatives as one proceeds down the sequence and stopping testing at a data determined point on the sequence. A limited proposal of this type was made by Rayner and Best (1989) and developed more generally in Bickel and Ritov (1992). Some important special cases are discussed in Hart (1997), Chapter 7, in the context of testing the hypothesis of no effect in nonparametric regression.

Our general approach which is detailed in Sections 2 and 3 is to use as building blocks one-dimensional score (Rao) test statistics for simple hypotheses. For composite hypotheses we use their natural generalizations of Rao tests efficient scores. Then efficient scores are called, Neyman \( C(\alpha) \) test statistics in the literature or conditional moment restriction tests in the econometrics literature (see Bierens and Ploberger (1997)). These building blocks are discussed in Section 2 as are their relation to the classical statistics for goodness-of-fit to a simple hypothesis.

Conceptually, as we discuss in Section 3, our approach applies to general semiparametric hypotheses such as independence, the Cox model in survival analysis, and index models in econometrics. It also, as we demonstrate, guides us how to proceed when we test a parametric or semiparametric model within a semiparametric alternative, for instance, independence within copula models, simple index versus multiple index models. We view this as perhaps the most important nontechnical contribution of the paper.

In Section 3 we give some general conditions under which the asymptotic theory for the types of test statistics discussed in this section, and for appropriate bootstrap critical values, is valid. We also study the power behavior of these tests under these assumptions. In Section 4 we discuss the classical examples of goodness-of-fit to a parametric hypothesis and independence. We show how the classical type (ii) tests of Kolmogorov–Smirnov and Cramér–von Mises type fit into our framework, and
also derive a variety of new tests based on our principles. We indicate how the
general conditions of Section 3 are implied by mild and easily checkable conditions
in these classical situations. We have chosen to exhibit the approach in detail in two
situations, (Parametric hypothesis and independence) here, and one (Testing index
models) elsewhere, but, as we indicated, it is applicable in any of the hundreds of
goodness of fit problems that arise in semiparametric models.

In Section 5 we report the result of a small simulation study for the independence
case showing that the asymptotics accurately predict the qualitative behavior of the
tests we’ve constructed for moderate sample sizes and selected alternatives.

Our methods generalize fairly naturally to the two sample problems, \( X_1, \ldots, X_n \)
i.i.d. \( G, X_{n+1}, \ldots, X_{n+m} \) i.i.d. \( F \), with the hypothesis \( F = G \), see Hall and Tajvidi
(2001) for a recent reference. We note only that the choice of a critical value leading
to an exact level \( \alpha \) here is straightforward. We need only rely on the permutation
distribution of \( X_1, \ldots, X_m \) among the pooled \( X_i \)—see Bickel (1968) for instance.

2. Some Testing Heuristics

2.1. Simple Null Hypothesis Suppose that \( X_1, \ldots, X_n \) are a (i.i.d.) random
sample from the probability \( P \in \mathcal{Q} \), where \( P \ll \mu, \ p = \frac{dp}{d\mu} \). Suppose that
\( \mathcal{Q} = \{ P_\theta : \theta \in \mathbb{R} \} \) is a regular (one-dimensional) sub-model of probabilities. Con-
sider testing the hypothesis

\[ H : P = P_0 \]

against

\[ K : P = P_\theta \quad \text{where} \quad \theta > 0. \]

Denote the log likelihood of an observation and its derivative at \( \theta = 0 \), the
efficient score (see Rao (1947)) by,

\[ \ell (\cdot, P) \equiv \ln p, \quad \dot{\ell}_1 (\cdot, P_0) \equiv \frac{\partial \ell (\cdot, P_0)}{\partial \theta} \]

where \( \dot{\ell}_1 \in L_2 (P_0), \ E_0 \left( \dot{\ell}_1 (\cdot, P_0) \right) = 0. \) The familiar scoring test of Rao (1947)
which is locally and asymptotically most powerful is based on the mean score test
statistic which, since \( n^{-1/2} \sum_{i=1}^n \dot{\ell}_1 (X_i, P_0) \overset{D}{\sim} \mathcal{N} \left( 0, \| \dot{\ell}_1 \|_0^2 \right) \), under the null, uses
as asymptotic critical value \( z_{1-\alpha} \cdot \| \dot{\ell}_1 \|_0 \), where \( \| h \|_0^2 = \int h^2 dP_0 \) and \( z_{1-\alpha} \) is the
standard Gaussian \( 1 - \alpha \) quantile. Note that this test is consistent if and only if
\( E_0 \left( \dot{\ell}_1 (X, P_0) \right) > 0 \) for \( \theta > 0 \); namely if a nonzero \( \theta \) implies a positive mean score.

2.2. Composite Null Hypothesis The case of a simple null hypothesis is indeed
the simplest type of testing situation; the model under the null is specified by
setting \( \theta = 0 \) and the alternatives (within \( \mathcal{Q} \)) are associated with values \( \theta > 0. \) If
the null set \( \mathcal{P}_0 \) is composite, the score function depends on \( P \), and becomes the
family \( \dot{\ell}_1 (\cdot; P), \ P \in \mathcal{P}_0. \) In particular, one needs two ingredients:
(i) A “consistent” estimate $\hat{P}_n$ of $P$.

(ii) The tangent space $\hat{\mathcal{P}}_0 (P_0)$, for $P_0 \in \mathcal{P}_0$, $\hat{\mathcal{P}}_0 (P_0)$ is the linear closure (in $L_2 (P)$) of the set of all score functions of regular one-dimensional submodels of $P_0$ passing through $P_0$—see, for instance, Bickel, Klaassen, Ritov, and Wellner (1993) (hereafter, BKRW) Chapter 2 for details.

The space $\hat{\mathcal{P}}_0 (P_0)$ captures the directions of variation from $P_0$ that are consistent with the null hypothesis of interest. To test $\theta = 0$, we should remove from $\ell_1$ its component that is actually consistent with $P_0$. Therefore, the effective direction of interest for the alternative $Q$ is given by efficient score function

$$
\ell^*_1 (\cdot, P_0) \equiv \ell_1 (\cdot, P_0) - \Pi \left( \hat{\ell}_1 | \hat{\mathcal{P}}_0 (P_0) \right)
$$

where $\Pi$ denotes the projection operator in $L_2 (P_0)$.

A natural test statistic is then $\frac{1}{n} \sum_{i=1}^{n} \ell^*_1 (X_i, \hat{P})$ with asymptotic tests based on the critical value $z_{1-\alpha} I (\hat{P})^{1/2}$ where $\hat{P}$ is a consistent estimate of $P$ under $H$ and $I (P) \equiv E_P (\ell^*_1 (X, P)^2)$ is the variance of the efficient score. This statistic is analyzed by Choi, Hall, and Schick (1996).

For concreteness, we examine how these concepts arise in traditional tests of parametric models. Suppose that the general family is defined as $\mathcal{P} = \{ P_{\theta, \eta} : \theta \in \mathbb{R}, \eta \in \mathbb{R}^p \}$ so that the general log-density takes the form $\ell = \ell (\cdot, P_{\theta, \eta})$, which for simplicity we write $\ell (\theta, \eta)$.

The restricted family $\mathcal{P}_0$ is associated with the null hypothesis $H : \theta = 0$ so that $\mathcal{P}_0 \equiv \{ P_0 : P_0 = P_{(0, \eta_0)}^0, \eta_0 \in \mathbb{R}^p \}$. The score vector of an alternative direction is $\ell_1 \equiv \partial \ell (0, \eta_0) / \partial \theta$. and the tangent space associated with the null hypothesis is

$$
\hat{\mathcal{P}}_0 (P_0) = \text{Linear Span} \left\{ \frac{\partial \ell (0, \eta_0)}{\partial \eta_j} : 1 \leq j \leq p \right\}.
$$

Therefore, the efficient score is

$$
\ell^*_1 (\cdot, P_0) \equiv \ell_1 (\cdot, P_0) - \Pi \left( \hat{\ell}_1 | \hat{\mathcal{P}}_0 (P_0) \right)
$$

where $\Pi$ denotes the projection operator in $L_2 (P_0)$.

A natural test statistic is then $n^{-1/2} \sum_{i=1}^{n} \ell^*_1 (X_i, \hat{P})$ with asymptotic tests based on the critical value $z_{1-\alpha} I (\hat{P})^{1/2}$ where $\hat{P}$ is a consistent estimate of $P$ under $H$ and $I (P) \equiv E_P (\ell^*_1 (X, P)^2)$ is the variance of the efficient score. This statistic is analyzed by Choi, Hall, and Schick (1996).

For concreteness, we examine how these concepts arise in traditional tests of parametric models. Suppose that the general family is defined as $\mathcal{P} = \{ P_{\theta, \eta} : \theta \in \mathbb{R}, \eta \in \mathbb{R}^p \}$ so that the general log-density takes the form $\ell = \ell (\cdot, P_{\theta, \eta})$, which for simplicity we write $\ell (\theta, \eta)$.

The restricted family $\mathcal{P}_0$ is associated with the null hypothesis $H : \theta = 0$ so that $\mathcal{P}_0 \equiv \{ P_0 : P_0 = P_{(0, \eta_0)}^0, \eta_0 \in \mathbb{R}^p \}$. The score vector of an alternative direction is $\ell_1 \equiv \partial \ell (0, \eta_0) / \partial \theta$. and the tangent space associated with the null hypothesis is

$$
\hat{\mathcal{P}}_0 (P_0) = \text{Linear Span} \left\{ \frac{\partial \ell (0, \eta_0)}{\partial \eta_j} : 1 \leq j \leq p \right\}.
$$

Therefore, the efficient score is

$$
\ell^*_1 = \frac{\partial \ell (0, \eta_0)}{\partial \theta} - \sum_{j=1}^{p} a_j (0, \eta_0) \frac{\partial \ell (0, \eta_0)}{\partial \eta_j}
$$

where the $a_j$’s are projection (least squares) weights; namely $\{ a_j (0, \eta_0) \}$ minimize $\| \ell_1 - \sum_{j=1}^{p} a_j (0, \eta_0) \partial \ell (0, \eta_0) / \partial \eta_j \|_0^2$.

If $\hat{\eta}$ is a $\sqrt{n}$-consistent estimator of $\eta$ under $H$, then the Neyman (1959) $C (\alpha)$ test statistic is,

$$
T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell^*_1 (X_i, 0, \hat{\eta})
$$

Suppose $\hat{\eta} = \hat{\eta}_H$, the maximum likelihood estimator (MLE) of $\eta$ under $H$. In that case $T$ reduces to,

$$
T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_1 (X_i, 0, \hat{\eta}_H)
$$
since \( \hat{\eta}_H \) solves the likelihood equations:

\[
0 = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \eta_j}(X_i, 0, \hat{\eta}_H).
\]

In fact, under additional smoothness conditions on \( \ell(\cdot, \theta, \eta) \), any estimate \( \hat{\eta} \) with \( \hat{\eta} = \eta_0 + o_P(n^{-1/4}) \) (under \( P(\theta_0, \eta_0) \)) may be substituted into \( \ell_1^* \) and any efficient estimator \( \hat{\eta} \) may be substituted into \( \dot{\ell}_1 \).

### 2.3. Composite Alternatives

We now recall how to put score tests of one-dimensional alternatives together for testing composite alternatives. Consider first the situation where the null hypothesis is simple: \( P_0 = \{ P_0 \} \) and let \( Q = \{ P_\theta : \theta \in \mathbb{R} \} \) denote a regular submodel of \( P \) through \( P_0 \). Clearly, the score in the model \( Q \), \( \dot{\ell}_1 \equiv h_Q \) depends on \( Q \). A composite alternative \( P \) is the union of many, usually an infinite number of such \( Q \)'s. The relevant set of (alternative) scores is the tangent space \( \mathcal{T}(P_0) \) defined as the linear closure of all the associated scores \( h_Q \) – see BKRW.

A particular but standard way of combining the score tests associated with different one-dimensional submodels is classical. Consider the situation of a finite number of submodels (directions), with individual scores denoted \( h_1, \ldots, h_p \). Begin by assuming that the scores are an orthogonal system under \( P_0 \), i.e. \( \langle h_a, h_b \rangle_0 = \delta_{ab} \) for \( a, b = 1, \ldots, p \). In this case the Rao statistic (Rao (1973)) is,

\[
T = \sum_{j=1}^{p} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j(X_i) \right)^2 \overset{D}{\rightarrow} \chi^2_p
\]

under \( H \).

When the scores from different submodels are correlated, then their covariance structure can be diagonalized in the standard way to produce the general Rao statistic. That is, suppose that the matrix \( \Sigma_0 \) with entries \( \Sigma_{0ab} = \langle h_a, h_b \rangle_0 \) and \( \Sigma_0 \) is nonsingular, then the Rao statistic – see Rao (1973) for instance, is

\[
T = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma_0^{-1/2} \tilde{h}(X_i) \right\|^2
\]

\[
= \frac{1}{n} \left\| \sum_{i=1}^{n} \tilde{h}(X_i) \right\|_{\Sigma_0^{-1}} \sqrt{n} \left\| \sum_{i=1}^{n} \tilde{h}(X_i) \right\|_{\Sigma_0^{-1}} \overset{D}{\rightarrow} \chi^2_p
\]

where \( \tilde{h} \equiv (h_1, \ldots, h_p) \).

These formulae arise naturally in the parametric case as follows. If \( \mathcal{P} = \{ P_\theta : \theta \in \mathbb{R}^p \} \) then the tangent space \( \mathcal{T}(P_0) \) is the linear closure of \( \{ \partial \ell / \partial \theta_j : j = 1, \ldots, p \} \), and the test statistic is

\[
T = \frac{1}{n} \left\| \sum_{i=1}^{n} \frac{\partial \ell(X_i)}{\partial \theta} \right\|_{\mathcal{I}_0^{-1}} \overset{D}{\rightarrow} \chi^2_p
\]

where \( \mathcal{I}_0 \) is the information matrix.
The Rao tests which are called Lagrange multiplier tests in econometrics, have the advantage of making use of estimates of the statistical model only under the null hypothesis. In contrast, Wald tests or likelihood ratio tests are based on comparing estimates of the model under alternatives with those of the model estimated under the null. In the parametric context, to first order, both Wald and likelihood ratio tests are equivalent to score based tests, and are well known to have asymptotic minimax properties.

It is important to note that $T$ is just one way of combining the different test directions. There is nothing magic in the Mahalanobis distance. Suppose that we can rank the alternate one dimensional models for which $h_1, \ldots, h_p$ are score functions in order of plausibility. If $h_1, \ldots, h_p$ are orthogonal, it is plausible to replace $T$ in (2.1) by,

$$T_a \equiv \sum_{j=1}^{p} \lambda_j^2 \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j(X_i) \right)^2$$

where $0 < \lambda_1 < \ldots < \lambda_p$ reflect the relative importance of the $h_j$. In general we would arrive at,

$$T_a = \frac{1}{n} \left[ \sum_{i=1}^{n} \bar{h}(X_i) \right]' \Lambda \Sigma_0^{-1} \Lambda \left[ \sum_{i=1}^{n} \bar{h}(X_i) \right]$$

where $\Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_p)$.

Another alternative is to use the union-intersection principle of Roy (1957), to obtain

$$T_b = \max_{1 \leq j \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j(X_i) \right| .$$

We note that the $\lambda_j$ here are implicit by choosing $h_j$ with different norms.

This approach applies equally well to the situation where the alternative hypothesis contains an infinite number of directions, i.e. $\mathcal{P}(P_0)$ has infinite basis. In particular, suppose that the tangent space is the linear closure of the set of directions $\{h_\gamma : \gamma \in \mathbb{R}^q\}$ and $\mu$ is a measure on $\mathbb{R}^q$. Two forms of statistics arise in analogy to the finite dimensional case, namely the weighted squared average score

$$T_a = \int \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_\gamma (X_i) \right)^2 d\mu(\gamma)$$

(2.2)

and the maximum average score

$$T_b = \sup_{\gamma} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_\gamma (X_i) \right| .$$

(2.3)

Some familiar statistics from density estimation, as well as classical nonparametric goodness-of-fit tests, fall into this category of score tests.

**Example. (Goodness-of-Fit Statistics)** Consider testing the null hypothesis that a distribution on $\mathbb{R}$ is $P_0$ against “all” alternatives, namely where

$$\mathcal{P}(P_0) = \{ h \in L_2(P_0) : E_{P_0} h(X) = 0 \}.$$
We call \( \mathcal{P} (P_0) \) above \( L^2_0(P_0) \) in the sequel. If we consider the family of directions \( h_\gamma (\cdot) = 1 (\cdot \leq \gamma) - F_0 (\gamma); \gamma \in \mathbb{R} \) where \( F_0 \) is the cumulative distribution function of \( P_0 \), then the following two statistics arise in association with those above. Associated with (2.2) is the familiar Cramér–Von Mises (CvM) goodness-of-fit statistic
\[
T_a = n \int (F_n (\gamma) - F_0 (\gamma))^2 dF_0 (\gamma)
\]
where \( F_n \) is the empirical distribution function, and the weighting measure is \( \mu = F_0 \). Corresponding to (2.3) is the familiar Kolmogorov-Smirnov (KS) goodness-of-fit test statistic
\[
T_b = \sup_\gamma \left| \sqrt{n} (F_n (\gamma) - F_0 (\gamma)) \right|.
\]
Note that \( h_\gamma \) here are chosen because we start with the cumulative distribution function as our representation of the probability \( P \), not because of desire for power in clearly defined directions.

### 3. Score Tests

#### 3.1. The score process and the testing paradigm
The above ideas motivate our general testing paradigm. For a simple null hypothesis, we consider a collection of basic score statistics of the form
\[
Z_n (\gamma) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_\gamma (X_i) = \sqrt{n} P_n (h_\gamma)
\]
\( \gamma \in \Gamma \), where \( P_n \) is the empirical distribution and \( h_\gamma \) is a direction in the tangent space, \( \mathcal{H} \equiv \{ h_\gamma : \gamma \in \Gamma \} \subset \mathcal{P} (P_0) \) so that \( P_0 (h_\gamma) = 0, \gamma \in \Gamma \). \( \Gamma \) is simply an index set as in Example 2.1 such that \( \mathcal{H} \) is not too big, say a universal Donsker class. Our overall approach is to build test statistics using the appropriate generalizations of the process \( Z_n (\cdot) \).

**Notation:** Here and subsequently we write \( Q(h) \) for \( \int h dQ \).

For composite hypotheses we begin by considering the case, \( \mathcal{P} = \mathcal{M} \equiv \{ \text{All probabilities dominated by } \mu \} \) or at least \( \mathcal{P} \) such that the tangent space is saturated, \( \mathcal{P} (P) = L^2_0(P) = \{ h \in L_2(P) : P(h) = 0 \} \). Let the hypothesis \( \mathcal{P}_0 = \{ P_\alpha : \alpha \in A \} \) where \( A \) can be a function space. If \( \mathcal{P}_0 \) is the tangent space at \( P_\alpha \) of \( \mathcal{P}_0 \) we can write generalizing (3.1), the efficient score statistic at \( P_\alpha \) in a direction \( h_\gamma(\cdot, \alpha) \in L^2_2(P_\alpha) \), corresponding to a submodel of \( \mathcal{P} \) containing \( P_\alpha \) as, (see BKRW p.70)

\[
Z_n(\gamma, \alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (h - P_\alpha(h) - \Pi(h, \alpha))(X_i)
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Pi^{\perp}(h, \alpha)(X_i)
\]
\[
\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(\gamma, \alpha)(X_i)
\]
\[
= \sqrt{n} (P_n - P_\alpha)(S(\gamma, \alpha))
\]
for $h \equiv h_\gamma(\cdot, \alpha)$. Here, $\Pi(\cdot, \alpha)$ is the projection operator from $L_2(P_\alpha)$ to the subspace $\mathcal{P}_0(P_\alpha)$ of $L_2^0(P_\alpha)$ and $\Pi^\perp(\cdot, \alpha)$ is the projection to the orthocomplement of $\mathcal{P}_0(P_\alpha)$ within $L_2^0(P_\alpha)$. The identity uses $\Pi^\perp(h, \alpha) = \Pi^\perp(h + c, \alpha)$ for all $c$.

The index $(\gamma, \alpha)$ ranges over $\Gamma \times A$. As we shall see in examples, the reason for making $h$ depend on $\alpha$ also is that it is natural to have the family of scores considered depend on where we think we are in $A$ if $H$ is true. We assume $h_\gamma(x, \alpha) \in l_\infty(\Gamma)$, the space of all bounded real valued functions on $\Gamma$, for all $x, \alpha$, to avoid technicalities.

We write $h_\gamma(\cdot)$ suppressing dependence on $x$ usually.

Call $Z_n(\cdot, \cdot)$ from (3.2) the score process. If $\mathcal{P}_0 = \{P_0\}$ is simple then, since there is no dependence on $\alpha$, we obtain $Z_n(\gamma, \alpha) = Z_n(\gamma)$ as in (3.1).

In general, $Z_n(\gamma, \alpha)$ is not computable given the data, but if $\hat{\alpha} \in A$ is an estimate of $\alpha$ we can consider

$$\hat{Z}_n(\gamma) \equiv Z_n(\gamma, \hat{\alpha}) \quad (3.3)$$

defined on $\Gamma$. If $\hat{\alpha}$ is an MLE, $\hat{Z}_n$ simplifies as in the parametric case to,

$$\hat{Z}_n(\gamma) = \sqrt{n}(P_n - P_\alpha)(h_\gamma(\cdot, \hat{\alpha}))$$

since $P_n(v) = 0$ for all $v \in \mathcal{P}_0$ ($\hat{\alpha}$) is a restatement of the likelihood equations. In particular, $P_n(\Pi(h, \hat{\alpha})) = 0$. We will also consider $\hat{Z}_n$ more generally for $\hat{\alpha}$ an efficient estimate in the sense of BKRW, Chapter 5, pp 179–182.

We think of $\hat{Z}_n(\cdot)$, $\hat{Z}_n(\cdot)$ etc. as stochastic processes defined on $\Gamma$ related to empirical processes—see van der Vaart and Wellner (1996), for instance. We shall use $\hat{Z}_n$ and $\hat{Z}_n$ as $Z_n(\cdot)$ to construct tailor made tests.

Let $A_0 \subset A$ be a neighborhood of the true $\alpha_0$ where $A$ is a metric space with metric $\rho$. From now on we write $P_0$ for $P_{\alpha_0}$, etc.

We always require

$$Z_n(\cdot, \alpha_0) \Rightarrow Z(\cdot, \alpha_0) \quad (4.3)$$

under $P_0$ for all $\alpha_0 \in A_0$ in the sense of weak convergence for $l_\infty(\Gamma)$ valued variables where $Z(\cdot, \alpha_0)$ is a mean 0 Gaussian process with,

$$\text{cov}(Z(\gamma_1, \alpha_0), Z(\gamma_2, \alpha_0)) = \text{cov}(h_1(X) - \Pi(h_1, \alpha_0)(X), h_2(X) - \Pi(h_2, \alpha_0)(X))$$

$$= \text{cov}(h_1, h_2) - \text{cov}(\Pi(h_1, \alpha_0), \Pi(h_2, \alpha_0))$$

with the obvious convention in notation. (Strictly speaking, we should be speaking of outer probabilities since we interpret weak convergence in the sense of Hoffman–Jørgensen. But measurability issues can be dealt with easily in the situations we are interested in and we ignore them in future.) The property we want is,

$$\hat{Z}_n(\cdot) \Rightarrow Z(\cdot, \alpha_0) \quad (4.5)$$

(in the sense of weak convergence of $l_\infty(\Gamma)$ valued variables (respectively $\hat{Z}_n$) to a tight limit under the same circumstances).
We propose to base, at least conceptually, our tests on the score process. What (3.5) will give us is the weak convergence of statistics of the form \( T(\hat{Z}_n(\cdot)) \) where \( T : l_\infty(\Gamma) \to R \) continuously. Possibilities are
\[
T_\mu \equiv \int \hat{Z}_n^2(\gamma) d\mu(\gamma)
\]  
(3.6)
for \( \mu \) a finite measure on \( T \) i.e. generalizations of CvM or
\[
T_K \equiv \sup_{\Gamma} |\hat{Z}_n(\gamma)|
\]  
(3.7)
or more general \( \mu \) norms of \( |\hat{Z}_n| \), or even \( \alpha \) dependent \( \mu \)'s which are suitably continuous in \( \alpha \). By taking \( \{h_n(\cdot, \alpha) : \gamma \in \Gamma\} \) dense in \( L_2(P_\alpha) \) for all \( \alpha \), and \( \mu \) with support \( \Gamma \) we can expect consistency against all alternatives. We will illustrate further in the examples of the next section.

3.2. Other tests  There are many tests which a priori are not based on a \( \hat{Z}_n(\cdot) \) process at all. For instance, suppose \( P_0 = \{P : T(P) = 0\} \) where \( T : M \to T \) some linear function space. Then it is natural to consider statistics based on \( T(P_n) \). For instance, \( H : F = F_0 \) can be written \( F - F_0 = 0 \) and naturally leads to statistics based on \( F_n - F_0 \). Here the paradigms coincide but consider instead the quantile function \( F^{-1} \) and base tests on \( F_n^{-1}(\cdot) - F_0^{-1}(\cdot) \), deviations from the probability plot. In the Gaussian goodness-of-fit case, for instance,
\[
W = \int_{\frac{1}{\sqrt{n}}}^{1-\frac{1}{n}} n(F_n^{-1}(t) - \Phi^{-1}(t))^2 dt
\]
is equivalent asymptotically to the well-known Shapiro–Wilk statistic. See Del Barrio et al. (1999). But heuristically it’s clear that there are \( \hat{Z}_n \) based tests equivalent to such procedures if \( T \) is Fréchet differentiable, \( T_n = T(P_n) \), where \( T(P_n) = T(P) + \hat{T}(P)(P_n - P) + \text{lower order terms} \). For instance,
\[
W = \int_{-a_n}^{a_n} n \frac{(F_n(x) - \Phi(x))^2}{\varphi(x)} dx + o_P(1)
\]
where \( a_n = \Phi^{-1}(1 - n^{-1}) \) is also asymptotically equivalent to Shapiro–Wilk. This statistic too is implicitly based on \( h_l^{(n)} \).

3.3. Setting Critical Values  There is a novel issue that arises in the context of composite hypotheses. The statistic \( \hat{Z}_n(\gamma) \equiv \frac{1}{n} \sum_{i=1}^n l^*(X_i, P_\gamma) \) arising from the situation where there is only one direction of departure is, if \( P_\alpha \) is true, an approximation to \( Z_n(\alpha) \equiv \frac{1}{n} \sum_{i=1}^n l^*(X_i, P_\alpha) \). Since \( Z_n(\alpha) \) has a \( N(0, I(P_\alpha)) \) limiting distribution, a Gaussian critical value with \( I(P_\alpha) \) replaced by \( I(P_\alpha) \) is appropriate. On the other hand, if the hypothesis is simple, critical values for any statistic can be obtained by simulation. But in the general situation of composite hypotheses, that we now consider, unless there is invariance, the only plausible way of setting critical values is by a bootstrap. The natural choice is to simulate from \( P_\hat{\alpha} \). That is, let
\[
\hat{Z}_n(\gamma) = n^{-1/2} \sum_{i=1}^n S(\gamma, \hat{\alpha})(\hat{X}_i)
\]
where $\tilde{X}_i$’s are i.i.d. from $P_0$. We expect that if (3.5) holds $\tilde{Z}_n \Rightarrow Z(\cdot, \alpha_0)$ in $P_0$ probability. That is, the Prohorov distance between the $P_0$ distribution of $\tilde{Z}_n$ and the distribution of $Z$ tends to 0 in $P_0$ probability.

There is another way of bootstrapping discussed in Bickel and Ren (2000) which may be simpler since it only involves resampling.

Let

$$\tilde{Z}_n^*(\cdot) = n^{-1/2} \sum_{i=1}^{n} (S(\gamma, \hat{\alpha}^*)-S(\gamma, \hat{\alpha})(X_i))$$

where $\hat{\alpha}^* \equiv \hat{\alpha}(X_1^*, \ldots, X_n^*)$ and $X_1^*, \ldots, X_n^*$ are i.i.d. from the empirical distribution $P_n \equiv n^{-1} \sum_{i=1}^{n} \delta_{X_i}$. The appropriate heuristic is that again if (3.5) holds, $Z_n^* \Rightarrow Z(\cdot, \alpha_0)$ in $P_0$ probability. The condition for $\tilde{Z}_n$, $\tilde{Z}_n^*$, and their $\tilde{Z}$ counterparts for $Z_n$ of (3.3) to work will be discussed below.

3.4. Testing When $P \not\subseteq M$ For composite hypotheses, consider the space of alternatives $P = \{P(\alpha, \beta) : \alpha \in A, \beta \in B\}$ where $A, B$ are function spaces. The null hypothesis of interest is $H : \beta = 0$. Define the “full”, “null” and “alternative” tangent spaces: see BKRW pp. 70.

$$\hat{\mathcal{P}}(\alpha, \beta) = \text{Tangent Space of the Model } \mathcal{P} \text{ at } P(\alpha, \beta)$$

$$\hat{\mathcal{P}}_0(\alpha, 0) = \text{Tangent Space of } \{P(\alpha, 0) : \alpha \in A\} \text{ at } P(\alpha, 0)$$

$$\hat{\mathcal{P}}_0^\perp(\alpha, 0) = \text{Orthogonal Complement of } \hat{\mathcal{P}}_0(\alpha, 0) \text{ in } \hat{\mathcal{P}}(\alpha, 0)$$

That is, $\hat{\mathcal{P}}_0(\alpha, 0) \perp \hat{\mathcal{P}}_0^\perp(\alpha, 0)$ and $\hat{\mathcal{P}}(\alpha, 0) = \hat{\mathcal{P}}_0(\alpha, 0) \oplus \hat{\mathcal{P}}_0^\perp(\alpha, 0)$. For $h \in L^2_2(P_0)$ the correct score process is now,

$$Z_n(\gamma, \alpha) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Pi^\perp(h, \alpha)(X_i) \quad (3.8)$$

where $\Pi^\perp$ is defined here as the projection from $\hat{\mathcal{P}}(\alpha, 0)$ to $\hat{\mathcal{P}}_0^\perp(\alpha, 0)$ and $h \equiv h(\gamma, \alpha)$.

The process $Z_n(\cdot, \alpha)$ and the corresponding $\tilde{Z}_n(\cdot)$ etc. can readily be related to (3.2) by noting that if $\Pi_1(\cdot, \alpha, 0)$ is the projection operation from $L_2(\hat{\mathcal{P}}(\alpha, 0))$ to the tangent space $\hat{\mathcal{P}}(\alpha, 0)$ and $\Pi_2(\cdot, \alpha, 0)$ is the projection operator to $\hat{\mathcal{P}}_0(\alpha, 0)$, then, for $h \in L_2(\hat{\mathcal{P}}(\alpha, 0))$,

$$\Pi^\perp(h, \alpha) = (h - E_{(\alpha, 0)}h(X_1) - \Pi_2(h, \alpha, 0))$$

$$- (h - E_{(\alpha, 0)}h(X_1) - \Pi_1(h, \alpha, 0)).$$

Hence the study of the general score process (3.8) and the corresponding $\tilde{Z}_n(h)$ etc. reduce to the study of the processes (3.2) and (3.3). Thus, if $Z_n^{(1)}(\cdot, \alpha), Z_n^{(2)}(\cdot, \alpha)$ are defined by (3.2) using $\Pi_1, \Pi_2$ respectively then

$$Z_n(\gamma, \alpha) = Z_n^{(2)}(\gamma, \alpha) - Z_n^{(1)}(\gamma, \alpha).$$
We close this section with some general theorems giving essentially the minimal conditions under which our heuristics for test statistic construction and critical value setting are justified. Checking the conditions of these theorems is the major difficulty.

Here are two theorems.

3) (M5) precisely says that \( \hat{\alpha} \) is efficient under \( H \), the generalization of the requirement that \( \hat{\alpha} \) be (a regularly behaving) MLE – see BKRW pp 176–182.

Here are two theorems.

Notes:

1) (M3) and (M4) are automatically satisfied if \( h_\gamma(\cdot, \alpha) \) doesn’t depend on \( \alpha \).

2) If \( H_0 \equiv \{ h_\gamma(\cdot, \alpha_0) = P_{\alpha_0} h_\gamma(\cdot, \alpha_0) : \gamma \in \Gamma \} \) is a Donsker class showing that \( \{ \Pi(h, \alpha_0) : h \in \mathcal{H}_0 \} \) is also Donsker is usually not hard. For instance, if \( \Pi \) preserves order, \( \Pi(h_1, \alpha_0) \leq \Pi(h_2, \alpha_0) \) if \( h_1 \leq h_2 \) and \( \mathcal{H} \equiv \{ h_\gamma(\cdot, \alpha_0) : \gamma \in \Gamma \} \) satisfies the bracketing entropy condition of Theorem 2.8.4 p. 172 (van der Vaart and Wellner (1996)) then since \( \Pi(\cdot, \alpha_0) \) is \( L_2(P_0) \) norm reducing \( \{ \Pi(h, \alpha_0) : h \in \mathcal{H}_0 \} \) also satisfies the same condition. Thus (M0) is usually not difficult.

3) (M5) precisely says that \( \hat{\alpha} \) is efficient under \( H \), the generalization of the requirement that \( \hat{\alpha} \) be (a regularly behaving) MLE – see BKRW pp 176–182.

We will discuss concrete examples of test statistics based on \( \hat{Z}_n \), \( \hat{\tilde{Z}}_n \) such as those of Example 3.1 as well as a number of other types in the examples of the next section. For simplicity we mainly restrict ourselves to the case \( \mathcal{P} = \mathcal{M} \).
Theorem 3.1. If (M0), (M3), (M4), (M5) hold then for all $\alpha_0$,

$$Z_n(\cdot, \alpha_0) \Rightarrow Z(\cdot, \alpha_0)$$

(3.9)

and hence, $\hat{Z}_n(\cdot) = Z_n(\cdot, \alpha_0) + o_P(1)$

(3.10)

Proof. By construction and (M3)

$$\hat{Z}_n(\gamma) = n^{1/2} \{(P_n - P_\alpha)(h_{\gamma}(\cdot, \alpha)) - (P_\alpha - P_{\alpha_0})(h_{\gamma}(\cdot, \alpha))\}
= n^{1/2} \{(P_n - P_\alpha)(h_{\gamma}(\cdot, \alpha_0)) - (P_\alpha - P_{\alpha_0})(h_{\gamma}(\cdot, \alpha))\} + o_P(1)
= n^{1/2} \{(P_n - P_\alpha)(h_{\gamma}(\cdot, \alpha_0)) - (P_\alpha - P_{\alpha_0})(h_{\gamma}(\cdot, \alpha_0))\} + o_P(1)$$

by (M4). Finally by (M5),

$$\hat{Z}_n(\gamma) = n^{1/2} \{(P_n - P_\alpha)(h_{\gamma}(\cdot, \alpha_0)) - \Pi(h_{\gamma}(\cdot, \alpha_0), \alpha_0)\} + o_P(1)$$

which is just (3.10). Note that $o_P(1)$ is interpreted here in the sense of $| \cdot |_\infty$ on functions of $\gamma$. Conclusions (3.9) and (3.11) follow immediately from (M0) and (3.10).

Condition (M2) is implied by two more easily checkable conditions.

N1: (i) $\hat{\alpha}$ is consistent in the Hellinger metric $\rho_H$ given by $\rho^2_H(\alpha_1, \alpha_2) = \int (\sqrt{dP_{\alpha_1}} - \sqrt{dP_{\alpha_2}})^2$.

(ii) Let $A_0$ be a fixed Hellinger ball around $\alpha_0$ and suppose that $\mu \gg P_\alpha$, $\alpha \in A_0$ with $\mu$ a probability measure. Write $s(\alpha) = \sqrt{dP_{\alpha}/d\mu}$, $\hat{s} \equiv s(\hat{\alpha})$, $s_0 \equiv s(\alpha_0)$.

Assume $\| (\hat{s} - s_0)^2/\hat{s} \|^2_\mu = \int (s_0/\hat{s} - 1)^4 s_0^2 \hat{s}^2 d\mu = o_P(n^{-1})$.

(iii) Let $\Pi_\mu$ denote projection in $L_2(\mu)$ onto the tangent space at $s(\alpha_0)$ of $L = \{s(\alpha) : \alpha \in A\}$—see BKRW, p.50. If $\mu = P_{\alpha_0}$ then $s(\alpha_0) \equiv 1$ and $\Pi_\mu = \Pi(\cdot, \alpha_0)$. Let $\| \cdot \|_\mu$ be $L_2(\mu)$ norm.

Assume, $\| \hat{s} - s_0 - \Pi_\mu(\hat{s} - s_0) \|_\mu = o_P(n^{-1/2})$.

N2: (i) $\sup \{ \| h_{\gamma}(\cdot, \alpha) \|_\infty : \gamma \in \Gamma, \alpha \in A_0 \} < \infty$.

(ii) $\sup \{ \| \Pi(h_{\gamma}(\cdot, \alpha)) \|_\infty : \gamma \in \Gamma, \alpha \in A_0 \} < \infty$.

Notes:

1) Note that (N1)(ii) follows from the more standard $\| (p(\cdot, \hat{\alpha}) - p(\cdot, \alpha_0))^2 \|_\mu = O_P(n^{-\gamma})$, $\gamma > \frac{1}{2}$ and $p(\cdot, \alpha)$ is bounded away from 0 uniformly in $\alpha$.

2) In smooth parametric models (N1)(iii) holds if $\| \hat{s} - s_0 \|_\mu = o_P(n^{-1/4})$, $\| s - s_0 - \Pi_\mu(s - s_0) \|_\mu = O(\| s - s_0 \|_\mu^2)$.

Theorem 3.2. If (M0)–(M3) hold then

$$\hat{Z}_n(\gamma) = Z_n(\gamma) + o_P(1).$$
The proof appears in the appendix.

We may wish to consider (see below) statistics in which the averaging measure also depends on \( \alpha \) say \( \int \hat{Z}_n^\alpha(h_\gamma)d\mu(\gamma, \hat{\alpha}) \). This too can be dealt with by a condition such as \( \alpha \rightarrow \mu(\cdot, \alpha) \) is uniformly continuous on \( \mathcal{H} \) in the bounded variation topology on the finite signed measures on \( \Gamma \). More generally, we may simply consider any test statistic of the form \( F(\hat{Z}_n, \hat{\alpha}) \) where \( F : l_\infty(\Gamma) \times A \rightarrow R \) is continuous in the \( l_\infty \times \rho \) topology.

3.6. The Bootstraps

To ensure that the first bootstrap method works for both \( \hat{Z}_n \) and \( \hat{Z}_n \) we need to simply replace conditions (M0)–(M5) by versions uniform in \( \alpha_0 \in \mathcal{A}_0 \). We leave a formal statement to the reader.

The second bootstrapping method is more problematic to check. Essentially what is needed for \( \hat{Z}_n \) given by (3.5) are conditions given in Bickel and Ren (2000) if we make the following identification: Suppose \( \hat{\alpha} = \alpha(P_n) \) where \( \alpha : \mathcal{M} \rightarrow A, \alpha(P_\alpha) = \alpha \) for \( P_\alpha \in \mathcal{P} \) and \( \mathcal{M} \) is the set of all probabilities. Let \( T : \mathcal{M} \rightarrow L(H) \times L(H) \) be defined for \( \mathcal{H} \) a Banach space of functions containing \( \{h_\gamma : \gamma \in \Gamma\} \) and \( L(H) \) the set of bounded linear functionals on \( \mathcal{H} \) be defined by

\[
T(P)(h) = \left( \int \Pi(h, \alpha(P))dP, \int h \ d(P - P_\alpha(P)) \right).
\]

Note that \( T(P_\alpha) \equiv 0 \) so that the hypothesis is contained in \( \{P : T(P) \equiv 0\} \). Now put on \( T(P) \) the conditions specified by Bickel and Ren (2000).

3.7. Power

It is easiest to see what happens to the processes on which we build our tests in the case where alternatives converge to \( P_0 \) in the \( n^{-1/2} \) scale. Specifically suppose \( \{P_t : |t| < 1\} \) is a one dimensional regular parametric model through \( P_0 \) with score function \( g(\cdot) \) such that \( g \notin \mathcal{P}_0(P_0) \), i.e., it is possible to discriminate \( \{P_t, t \neq 0\} \) from \( P_0 \) at the \( n^{-1/2} \) scale. Then, let \( Z_\gamma(h_\gamma, 0) \) be the Gaussian process with the same covariance structure as \( Z(h_\gamma, 0) \) but with

\[
E_0Z_\gamma(h_\gamma, 0) = \int h_\gamma(g - \Pi(g, \alpha_0))dP_0
= \int (h_\gamma - \Pi(h_\gamma, \alpha_0))gdP_0.
\]

Evidently, \( Z_\gamma(h_\gamma, 0) = Z(h_\gamma, 0) \), if \( g \) does not have a component orthogonal to \( \mathcal{P}_0 \).

Define \( Z_\gamma(h_\gamma, \alpha) \) similarly for \( g \in L_2(P_\alpha), g \notin \mathcal{P}_0(P_\alpha) \). The following result is an immediate consequence of LeCam’s LAN theory, see for example his “third lemma” (Hajek and Sidak (1967)) and our theorems.

**Theorem 3.3.** Suppose \( g \in \mathcal{P}_0(P_\alpha) \) is, for each \( \alpha \), the score function of a regular model through \( P_\alpha \). Assume the sufficient conditions of any of Theorems 3.1–3.2 hold. Then,

\[
\hat{Z}_n(\cdot) \Rightarrow Z_\gamma(\cdot, \alpha)
\]
where $t_n \Rightarrow$ is weak convergence under $P_{t_n}$, where $\{P_t : |t| < 1\}$ is a regular model passing through $P_0 = P_\alpha$, and $t_n = tn^{-1/2}$ for fixed $t$.

Suppose $q$ is bowl shaped and symmetric and its discontinuity set has probability 0. That is, if $C \equiv \{z : q$ is continuous at $z\}$, $P[Z(\cdot, \alpha)$ or $Z_\gamma(\cdot, \alpha) \notin C] = 0$, and $q: l_\infty(\Gamma) \to \mathbb{R}$, $q(z) = q(-z)$.

$q(\lambda z) \nearrow$ strictly in $\lambda$ for $\lambda > 0$ all $z$. Then, if, as we assume, $Z(\cdot, \alpha)$ is tight

$$E(q(Z(\cdot, \alpha))) \leq E(q(Z_g(\cdot, \alpha))).$$

(3.12) follows from Anderson’s theorem if $\{G_1, \ldots, G_k\}$ forms a partition of $\Gamma$, and $Z(\gamma, \alpha)$ is replaced by $Z^{(k)}(\gamma, \alpha) \equiv \sum_{j=1}^k Z(\gamma_j, \alpha) 1(\gamma \in G_j)$, and $Z_j$ is similarly approximated. Now, $\|Z^{(k)}(\cdot, \alpha) - Z_g(\cdot, \alpha)\|_\infty \to 0$ as $k \to \infty$, for all $g$ including $g = 0$ and (3.12) follows in general.

Suppose first

$$\Pi(h_\gamma(\cdot, \alpha), \alpha) \neq h_\gamma(\cdot, \alpha)$$

for all $\gamma, \alpha$. Then, test statistics of the form (3.6)–(3.7) have

$$\lim_n P_{t_n}[T_n \geq c] > \lim_n P_0[T_n \geq c]$$

for $t_n = \lambda_n n^{-1/2}$, all $\lambda > 0$, all $c$ as desired. The requirement that (3.13) hold for all $\gamma, \alpha$ can obviously, for (3.6), be weakened to its holding, for all $\alpha$, for some $\gamma$, i.e., that there is discriminating power in at least one direction. For statistics of form (3.7) what is needed is that the limiting distribution under $P_0$ be nondegenerate since

$$h_\gamma(\cdot, \alpha_0) = \Pi(h_\gamma(\cdot, \alpha_0), \alpha)$$

for all $\gamma$ implies that $Z(h_\gamma(\cdot, \alpha_0), \alpha_0) = 0$.

3.8. Consistency  Consistency against fixed $P /\notin P_0$ can be obtained by a strengthening of conditions—though the strengthening we now give is overkill.

Suppose that $P$ is above. For a suitable $\alpha(P) \in A$ call $M_j$, $j = 0, \ldots, 5$, condition $M_j$ with $P_0$ replaced by $P$ and $\alpha_0$ replaced by $\alpha(P)$. Define the process $Z_P(\cdot)$ as the Gaussian process with mean 0 and the covariance structure given in (3.4) with $\Pi(\cdot, \alpha_0)$ replaced by $\Pi(\cdot, \alpha(P))$ and $P_0$ replaced by $P$. Then the conclusion of Theorems 3.1–3.2 holds if $P_0$ is replaced by $P$ with

(i) $\tilde{Z}_n(h_\gamma)$ replaced by

$$\tilde{Z}_n(h_\gamma) - \sqrt{n} \int h_\gamma(\cdot, \alpha(P))dP.$$ 

(ii) $Z(\cdot, \alpha_0)$ replaced by $Z_P(\cdot)$. 

We conclude that
\[ |\hat{Z}_n(h_\gamma)| \xrightarrow{P} \infty \]
if \( \int h_\gamma(\cdot, \alpha(P))dP \neq 0 \). Thus consistency holds for \( T_n \) given by (3.6) if \( \int h_\gamma(\cdot, \alpha(P))dP \neq 0 \) for some \( \gamma \) and all \( P \neq P_0 \). Consistency for other statistics can be reasoned analogously.

4. Examples

In this section we consider a few important examples in which we show how our notions produce tests which have appeared in the literature and some new ones. Our point is to illustrate the ideas of the score process and the tailor made tests.

4.1. Testing Goodness of Fit to a Composite Parametric Hypothesis

Let \( \{ P_\theta : \theta \in \Theta \subset \mathbb{R}^d \} \) be a regular parametric model and \( \hat{\theta}_n \) be a regular (BKRW p. 18–19) estimate of \( \theta \) under \( H \). Then \( P(\Theta) = \{ a \in L_2(\Theta) : E_p(a(X_1)) = 0 \} \)

\[ \bullet P(\Theta) = \left[ \frac{\partial l}{\partial \theta_1}(X_1, \theta), \ldots, \frac{\partial l}{\partial \theta_d}(X_1, \theta) \right] \]

where \([\ldots]\) denotes linear span in \( L_2(\Theta) \) and \( l = \log p(x, \theta) \) the log likelihood. Then,

\[ \bullet P(\Theta) = \left\{ a(X_1) - \sum_{j=1}^{d} c_j(a, \theta) \frac{\partial l}{\partial \theta_j}(X_1, \theta) : a \in L_2(\Theta), E_\theta a(X_1) = 0 \right\} \]

and \( c_j(a, \theta) \) is the coefficient of the projection of \( a \) on \( \bullet P(\Theta) \) i.e.

\[ E_\theta \left( a(X_1) - \sum_{j=1}^{d} c_j(a, \theta) \frac{\partial l}{\partial \theta_j}(X_1, \theta) \right)^2 = \min \]

for \( c_j = c_j(a, \theta), \ 1 \leq j \leq d \).

Identifying \( \alpha \) with \( \theta \), \( h = h_\gamma(\cdot, \theta) \) we obtain,

\[ S(\gamma, \theta) = h - E_\theta h(X_1) - \sum_{j=1}^{d} c_j(h, \theta) \frac{\partial l}{\partial \theta_j}(X_1, \theta). \]

The corresponding estimated score process is, for an estimate \( \hat{\theta} \), given by,

\[ \hat{Z}_n(h) = n^{-1/2} \sum_{i=1}^{n} \left\{ (h(X_i) - E_{\hat{\theta}} h(X_i)) - \sum_{j=1}^{d} c_j(h; \hat{\theta}) \frac{\partial l}{\partial \theta_j}(X_i, \hat{\theta}) \right\}. \]

Suppose \( P_0 \) is regular parametric, and more,

(R1) \( \hat{\theta} \) is regular on \( \Theta \).

(R2) Suppose \( \Gamma \) is compact \( \subset \bar{R}^p \), where \( \bar{R} = [-\infty, \infty] \), the processes \( (\gamma, \theta) \rightarrow n^{1/2}(P_n - P_0)h_\gamma(\cdot, \theta) \) are tight, and \( \sup_{x, \gamma, \theta} |h_\gamma(x, \theta)| < \infty \).
The map $\theta \to h(\cdot, \theta)$ is continuous in the norm on functions of $(\gamma, x)$ given by $\|\omega\|^2 = \sup_\gamma \int \omega^2(x, \gamma) dP_0(x)$.

The following proposition is a consequence of Theorem 3.2. We check its conditions using Lemma 3.1. (M0) and (M1) hold and (N1)(i) and (N2) are immediate. We can check (M2) via (N1) and (N2). Condition (N1)(iii) follows since

$\Pi_{\mu}(\hat{\theta} - s(\theta_0)) = s(\theta_0)(\hat{\theta} - \theta_0) + o_P(\hat{\theta} - \theta_0)$. Condition (N1)(ii) requires further conditions. For instance it follows if the likelihood ratio $s(\cdot, \theta)/s(\cdot, \theta')$ is uniformly bounded for $\theta, \theta'$ within $\epsilon$ of $\theta_0$ a case which unfortunately excludes the Gaussian but the condition can be checked directly fairly easily for suitable $\hat{\theta}$.

We have established

**Proposition 4.1.** If (R1)–(R3) and (N1)(ii) hold for a regular parametric hypothesis and given estimate $\hat{\alpha}$, then $\hat{Z}_n(\cdot) \Rightarrow Z(\cdot)$ where

$$
cov(Z(h_1), Z(h_2)) = \text{cov}_{\theta_0}(h_1(X_1), h_2(X_1)) - c^T (h_1, \theta_0) I(\theta_0) c(h_2, \theta_0)
$$

where $c = (c_1, \ldots, c_d)$ and $I$ is the Fisher Information.

Versions of a result such as this one appear in Durbin (1973), Khmaladze (1979) when the $h_\gamma$ are indicators of half lines.

The corresponding result for $\tilde{Z}_n$ is also valid as is that for $\hat{Z}_n^*$. That is both the parametric and the Bickel–Ren application of the nonparametric bootstrap to testing can be used to set critical values.

Suppose $h$ doesn’t depend on $\alpha$ and we weaken (R1) to,

(R1)” $\hat{\theta} = \theta + o_P(n^{-1/4})$ for all $\theta \in \Theta$.

The theorem will still hold provided that we have Cramér conditions on several derivatives of the likelihood ensuring that the remainders in (N1) are quadratic in $\hat{\theta} - \theta_0$. Note that Proposition 4.1 enables us to plug in subefficient estimates without affecting the properties of our tests.

**Proposition 4.2.** If the hypothesis is regular parametric, (R1), (R2), and (R3) hold and $\theta$ is efficient in the sense of BKRW, p.43 then the conclusion of Theorem 3.1 is valid.

**Proof.** We need only check (M5) and (M4). The former follows from

$$
\sup_\gamma \left| \int h(x)p(x, \hat{\theta})dx - \int h(x)p(x, \theta_0)dx - \int h(x)\ell(x, \theta)(\hat{\theta} - \theta_0)p(x, \theta_0)dx \right| = o_P(|\hat{\theta} - \theta_0|)
$$

and this requires in view of (R2) only that $\theta \to p(\cdot, \theta)$ is $L_1$ differentiable which is a consequence of regularity. The latter follows from regularity of $P_\theta$ and (R1) and (R3).
For \( \hat{\theta} \), the MLE, results such as this one appears in Durbin (1973) and Khmaladze (1979). The uniformity required for both versions of the bootstrap can easily be imposed.

Typical discussion in the literature start with a given test. We do not recommend a particular functional of the score function. Many functionals are reasonable, and the proper one should depend on the particular departures that are interesting. Are departures from the parametric model in the tail of the distribution more important than departures in the middle range? The answer depends on the specific application. Different answers should yield different tests. In our framework, this means that the particular functionals applied to the score process should be model dependent.

\section{The Gaussian Model}

We specialize to one of the most important parametric hypotheses \( P = \{ \mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \} \).

Here we naturally take \( \hat{\theta} = (\bar{X}, \hat{\sigma}^2) \), the MLE’s. It is convenient to use the invariance properties of the hypothesis and take

\[ h(\cdot, \gamma, \theta) = h_\gamma \left( \frac{x - \mu}{\sigma} \right) \quad \text{if} \quad \theta = (\mu, \sigma). \]

With this choice we are considering,

\[ \hat{Z}_n(\gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( h_\gamma \left( \frac{X_i - \bar{X}}{\hat{\sigma}} \right) - \int_{-\infty}^{\infty} h(z) \phi(z) dz \right). \]

If \( \{h_\gamma = 1(-\infty, \gamma), \gamma \in \mathbb{R}\} \equiv \mathcal{H}_0 \) then satisfaction of (R1)–(R3) is easy. Using \( \hat{Z}_n \) as above we arrive at the common test statistics, Kac, Kiefer and Wolfowitz (1955).

**Kolmogorov–Smirnov Type:**

\[ \sup_{\mathcal{H}_0} |\hat{Z}_n(\gamma)| = \sup_x \left| F_n \left( \frac{x - \bar{X}}{\hat{\sigma}} \right) - \Phi(x) \right| \]

**Cramér–von Mises Type:**

\[ \int_{-\infty}^{\infty} \hat{Z}_n^2(h_\gamma) d\Phi(x) = \int_{-\infty}^{\infty} \left( F_n \left( \frac{x - \bar{X}}{\hat{\sigma}} \right) - \Phi(x) \right)^2 d\Phi(x) \]

where \( F_n \) is the empirical \( df \). \( \mathcal{H}_0 \) is a well-known universal Donsker class and the classical limiting result for Cramér–von Mises given in Kac et al (1995) follows. Tests can be implemented for both statistics using either of the two bootstraps. Invariance here implies that only simulation under \( \mathcal{N}(0,1) \) is required.

Other classes of tests are covered, for example tests based on the empirical characteristic function (Feuerverger and Mureika (1977)).

We can also tailor statistics more carefully.

Consider the parametric mixture models

(i) \( (1 - \epsilon) \Phi \left( \frac{t - \mu}{\sigma} \right) + \epsilon \Phi \left( \frac{t - \mu - \Delta}{\sigma} \right), \mu, \Delta \in \mathbb{R}, \sigma > 0, 0 < \epsilon < 1 \)

(ii) \( (1 - \epsilon) \Phi \left( \frac{t - \mu}{\sigma} \right) + \epsilon \Phi \left( \frac{t - \mu - \Delta}{\tau} \right), \mu, \Delta, \epsilon \) as above \( \sigma > 0, \tau > 0 \).
At least formally the tangent set \( \mathcal{T}(P_0) \) (see BKRW p.50) at \( \epsilon = 0, \theta = (\mu, \sigma) \) is just the set \( \left\{ \left. e^{\frac{\Delta}{\sigma}}(X_1 - \mu) - \frac{\Delta^2}{2} \right| \Delta \in \mathbb{R} \right\} \cup [X_1 - \mu, (X_1 - \mu)^2] \). We are led to consider \( \mathcal{H}_0 = \{ e^{\lambda x - \frac{x^2}{2}} : \lambda \in \mathbb{R} \} \) and statistics such as

\[
T_n \equiv \sup_{\lambda} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( e^{\lambda (X_i - \bar{X}) - \frac{X_i^2}{2}} - 1 \right) \right|
\]

Unfortunately, see Bickel and Chernoff (1993), \( \mathcal{H}_0 \) is not a Donsker class and \( T_n \to \infty \) under \( H \). Our heuristics and Theorem 3.1 apply if we restrict \( \lambda \) to a compact. The power against \( n^{-1/2} \) alternatives of such \( T_n \) persists. Note that \( T_n \) can be viewed as a diagnostic since the maximizing value of \( \lambda \) indicates where a second component might be. Evidently (ii) simply leads to a similar two parameter process.

We can also consider versions of the Cramér-von Mises approach reflecting our goals more precisely. For instance, consider a wavelet basis for \([0,1]\) written lexicographically \( \omega_{ij} \) in order of scale and then within scale with \( \omega_{11} \equiv 1 \). Then given that we care more for departures at lower scales, consider \( \lambda_{ij} = \lambda_i \rho^j, 1 \leq j \leq 2^i, \rho < \frac{1}{4} \). Since \( \| \omega_{ij} \|_{\infty} = O(2^j) \), if we let \( h_{ij} = \omega_{ij}(\Phi(\cdot)) \) and,

\[
h_i(x) \equiv \sum_{i,j} \lambda_{ij} h_{ij}(x) h_{ij}(t)
\]

then,

\[
\|h_i\|_{\infty} \leq M < \infty
\]

and

\[
T = \int_0^1 \hat{Z}_n^2(t) dt = \sum_{i,j} \lambda_{ij}^2 [\hat{Z}_n(i,j)]^2
\]

where \( \hat{Z}_n(i,j) \leftrightarrow h_{ij} \) falls under the statistics covered by Theorem 3.2.

An interesting basis to consider is the normalized Hermite polynomials \( h_j(x) = (-1)^j \frac{d^j \varphi(x)}{dx^j} / \varphi(x), j \geq 3 \). Here \( h_3 \) and \( h_4 \) correspond to skewness and kurtosis so that it is attractive to make \( \lambda_3 = \lambda_4 = 1 \) and \( \lambda_j \) decrease rapidly further on.

We stress again that Propositions 4.1 and 4.2 can be applied to all these diverse tests.

4.3. Independence One of the most important semiparametric hypotheses corresponds to \( X = (U,V) \sim P, H : P = P_U \times P_V \), \( U \) and \( V \) are independent, \( U, V \in \mathbb{R} \). In this case the NPMLE of \( P \) under \( H \), known to be efficient—BKRW, Chapter 5—is

\[
P_n = P_{nU} \times P_{nV}
\]
where \( P_{nU} \) and \( P_{nV} \) are the empirical marginals of \( U \) and \( V \). Thus

\[
\hat{Z}_n(\gamma) = \sqrt{n}(P_n - (P_{nU} \times P_{nV}))(h)
\]

\[
= n^{1/2} \left\{ \frac{1}{n} \sum_{i=1}^{n} h(U_i, V_i, \gamma, \hat{P}_n) - \frac{1}{n^2} \sum_{i,j} h(U_i, V_j, \gamma, \hat{P}_n) \right\}.
\]

Natural \( h_j(u, v) \) here are of the form \( h_{1,\gamma}(u)h_{2,\gamma}(v) \). If we take,

\[
h_{\gamma}(u, v) = 1_{Q_1}(u, v)1_{Q_2}(v)
\]

where \( Q_j \) are of the form \( Q_1 \times Q_2 \), \( Q_j = (-\infty, \gamma_j], \) for \( j = 1, 2 \), we arrive at the familiar

\[
\hat{Z}_n(\gamma) = \sqrt{n}(F_{nU}(\gamma_1) - F_{nU}(\gamma_1))F_{nV}(\gamma_2))
\]

where \( F_n, F_{nU}, F_{nV} \) are the appropriate empirical d.f.'s.

Application of Theorem 3.1 here is appropriate and easy. (M0) simply says \( \gamma \to 1((x, y) \in Q_{\gamma}) - F_U(\gamma_1)1(y \leq \gamma_2) - F_V(\gamma_2)1(x \leq \gamma_1) \) is a Donsker class, essentially a statement about the bivariate empirical process. Since \( h_\gamma \) doesn't depend on \( \alpha \), (M3) and (M4) are immediate. Finally, (M5) is well known for this process—see, for instance, BKRW.

If we take \( \mu_\alpha(d\gamma) = dF_U(\gamma_1)F_V(\gamma_2) \) with \( \alpha = (F_U, F_V) \) we obtain the Kiefer–Wolfowitz statistic

\[
T = n \int \int (F_n(\gamma_1, \gamma_2) - F_{nU}(\gamma_1)F_{nV}(\gamma_2))^2 dF_{nU}(\gamma_1)dF_{nV}(\gamma_2).
\]

If we take \( T = \sup_\gamma |\hat{Z}_n(\gamma_1, \gamma_2)| \) we obtain the Kolmogorov–Smirnov version of the Kiefer–Wolfowitz statistic.

Invariance under monotone transformation of \( H \) suggests

\[
h(u, v, \gamma, P_U \times P_V) = 1_{Q_1}(F_U(u), F_V(v))
\]

where \( F_U, F_V \) are the c.d.f.'s of \( U, V \) and leads to \( \hat{Z}_n(\gamma) \), a linear rank test statistic

\[
\hat{Z}_n(\gamma) = n^{-1/2} \sum_{i=1}^{n} h_\gamma \left( \frac{R_i}{n}, \frac{S_i}{n} \right) - \frac{1}{n^2} \sum_{i,j} h_\gamma \left( \frac{i}{n}, \frac{j}{n} \right)
\]

where \( R_i \) is the rank of \( U_i \) among the \( U \)'s and \( S_i \) is the rank of \( V_i \) among the \( V \)'s.

These are the building blocks of the Kallenberg–Ledwina (1999) statistics, though the ones they propose are of non-\( n^{-1/2} \) consistent type. We leave it to the reader to construct tests with power against all \( n^{-1/2} \) alternatives and directions that (s)he prefers.

**Proposition 4.3.** Suppose \( h_\gamma \) is given by (4.1), \( \hat{\alpha} \) is the NPMLE as specified, and \( \alpha_0 \) has continuous marginals. Then, (M0), (M3), (M4), (M5) hold.
Proof. In this case

$$
\Pi(h, \alpha)(x, y) = \int h(x, v) dF_V(v) + \int h(u, y) dF_U(u)
$$

$$
-2 \int h(u, v) dF_U(u) dF_V(v)
$$

where $F_U, F_V \leftrightarrow \alpha$. So (M5) can be written

$$
\sup \{ \left| \int h_\gamma(x, y) d(F_{nU}(x) - F_U(x)) d(F_{nV}(y) - F_V(y)) \right| : \gamma \in \mathbb{R}^2 \} = o_P(n^{-1/2})
$$

where $F_{nU}$ is the empirical df of $U$, $F_{0U}$ corresponds to $\alpha_0$, etc. For this condition and all subsequent ones, we can assume, WLOG, that $\alpha_0$ is the uniform distribution on the unit square by making separate probability integral transforms. But (4.2) is just

$$
\sup_{0 \leq \gamma_1 \leq 1} |F_{nU}(\gamma_1) - \gamma_1| \sup_{0 \leq \gamma_2 \leq 1} |F_{nV}(\gamma_2) - \gamma_2| = O_P(n^{-1}),
$$

and (M5) follows. (M0) has been discussed in connection with $h_\gamma$. For (M3) write

$$(P_n - P_0)(h_\gamma(\cdot, \hat{\alpha}) - h_\gamma(\cdot, \alpha_0))
$$

$$
= (P_n - P_0) \left( 1(\{u, v\} \leq (F_{nU}^{-1}(\gamma_1), F_{nV}^{-1}(\gamma_2))) - 1(\{u, v\} \leq (\gamma_1, \gamma_2)) \right)
$$

where $(F_{nU}, F_{nV}) \leftrightarrow \hat{\alpha}$. Now, by Glivenko Cantelli, $\sup \left| F_{nU}^{-1}(\gamma_1) - \gamma_1 \right| \to 0$ and $\sup \left| F_{nV}^{-1}(\gamma_2) - \gamma_2 \right| \to 0$. So (M3) follows from the weak convergence of $n^{1/2}(P_n - P_0)(h_\gamma)$. (M4) is argued similarly.

Application of the type I bootstrap is straightforward since in view of the invariance we need only simulation under the uniform distribution on the unit square if we assume that $F_{U}$ and $F_{V}$ are continuous under $\alpha_0$. Resampling from the empirical (Type II) is also possible but the argument is more delicate.

This result can easily be extended to more general $h_{i1}(u)h_{i2}(v)$ and we can also tailor tests here. For instance consider the tensor wavelet basis on $[0, 1] \times [0, 1]$, $\{h_{i1, j_1, j_2}\}$ where $i$ corresponds to scale and $(j_1, j_2)$ to the location. We can again suppose that departures from independence at lower resolution are more significant and proceed as in Example 4.1 to form,

$$
T = \int_{I^2} \left( \int_{I^2} \sum_{i,j_1, j_2} \lambda_i h_{i1, j_1, j_2}(F_{nU}(x), F_{nV}(y)) d(P_n - \hat{P}_n)(x, y) h_{i1, j_1, j_2}(u, v) \right)^2 du dv
$$

$$
= \sum_{i,j_1, j_2} \lambda_i^2 \left( \int h_{i1, j_1, j_2}(F_{nU}(x), F_{nV}(y)) d(P_n - \hat{P}_n)(x, y) \right)^2
$$

where $I^2$ is the unit square and $\hat{P}_n = P_{nU} \times P_{nV}$. The $\lambda_i$ can be chosen so as to weight the lower resolution terms as one pleases.
4.4. Copula models The standard copula model is $X = (U, V)$, $U, V \in R$ as above, where for some monotone strictly increasing transformations, $a(\cdot): R \rightarrow R$, $b(\cdot): R \rightarrow R$ the vector $(a(U), b(V)) \sim P_\theta$, $\theta \in \Theta$, a regular parametric model. A natural model to consider here is the bivariate Gaussian copula, $P_\theta = N_2(0, 0, 1, 1, \theta)$, $-1 < \theta < 1$. (Assuming unknown means and variances adds nothing since making $a$ and $b$ arbitrary makes these parameters unidentifiable.) In such a model we consider two problems,

(i) $H: P \in P_0 = \{P_{\theta,a,b}: \theta \in \Theta, a, b \text{ general}\}$, the copula model hypothesis.

(ii) $H: \theta = 0$, $K: \theta \neq 0$ within $P_0$.

The first hypothesis requires use of efficient estimates of $(a, b, \theta)$. These are in general difficult to construct. Inefficient estimates are readily computable but application of Theorem 3.1 requires computation of $P(\cdot, \alpha)$ which can be characterized by Sturm-Liouville equations and computed numerically—see BKRW pp. 172–175. We do not pursue this interesting special case further.

On the other hand, by our assumptions in Section 3, tests of (ii) are naturally based on $h_\gamma(\cdot, \alpha) = h_\gamma(F_U(\cdot), F_V(\cdot))$, where the $\{h_\gamma(x, y), \gamma \in R^p\}$ are scores of the parametric model $\{P_{\theta}: \theta \in \Theta\}$ at $\gamma = 0$. If efficient estimates $\hat{F}_U$, $\hat{F}_V$ under $H$ are used, then $\hat{Z}_a(\gamma)$ is the asymptotically most powerful score test in direction $\gamma$. Otherwise, if say we use the empirical $F_{aU}$, $F_{aV}$, we can construct $P(\cdot, \gamma)$, the projection on the tangent space of $P_{\theta_0} = \{P_{(\theta_0, a, b)}: a, b \text{ arbitrary}\}$ and use $\hat{Z}_a$. Finally, if efficient estimates of $\theta$ under $P$ are available, these can be used in the obvious way, though in general such estimates will be difficult to obtain. These hypotheses of finite codimension are the subject of Choi, Hall, and Schick (1996). If we specialize to the Gaussian copula model and consider $H: \rho = 0$, the independence hypothesis, it is easy to see that the single asymptotically, most powerful test is to use the normal score rank statistic

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Phi^{-1}(R_i/n + 1)\Phi^{-1}(S_i/n + 1).$$

The reason here is that $F_{aU}$, $F_{aV}$ are efficient in this case—as we have already seen. Remarkably, Klaassen and Wellner (1997) show that $T$ is the asymptotically most powerful score statistic for $H: \rho = \rho_0$, any $\rho_0$, by showing effectively that

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(\Phi^{-1}F_U(U_i), \Phi^{-1}F_V(V_i), \rho_0) + o_P(1).$$

where $h(\Phi^{-1}F_U(U_i), \Phi^{-1}F_V(V_i), \rho_0)$ is orthogonal to the tangent space $P_{2}(\rho_0, (a_0, b_0))$, $a_0 = \Phi^{-1}F_U$, $b = \Phi^{-1}F_V$.

The development of tailored tests for independence in copula models in general should be the same as for independence in general.

5. Simulations We present power results for four tests of independence against departures in five different directions. The formal descriptions of the alternatives are given below but Figure 1 is more instructive. They have been chosen so
as to exhibit five qualitatively plausible types of departures. Two of the tests have been constructed using our principles to aim for some of these alternatives. The other two are a $\chi^2$ type of test and a Kolmogorov–Smirnov (Kiefer–Wolfowitz) type of test. All are presented in terms of the Haar tensor wavelets. Here are the distributions.

In all cases considered, the joint distribution of $(X, Y)$ is continuous. Under the null hypothesis, $X$ and $Y$ are independent. In all cases the sample size was 1024. The alternatives considered were:

(1) Let $(U, V)$ be two random variables distributed uniformly on a 64 by 64 grid. Let $W$ be a Bernoulli random variable, $P(W = 1) = 0.3$. Let

$$V' = V - 16W1(1 \leq U \leq 16, 49 \leq V \leq 64) + 16W1(16 \leq U \leq 32, 33 \leq V \leq 48).$$

Then, $X = U + \epsilon, Y = V' + \nu$, where $\epsilon, \nu$ are independent mean 0 normal random variables, with standard deviation equals 0.1 the grid spacing.

(2) $X$ has a uniform distribution on $(0, 1)$. Given $X = x$, $Y$ has a beta distribution with parameters $1 - 0.2x$ and $1 + 0.2x$.

(3) Again, there are latent variables $(U, V)$. With probability 0.9, $(U, V)$ are distributed uniformly on a 7 by 7 grid. Otherwise, they are distributed uniformly on the odd points of the grid. The observed variables $(X, Y)$ are given by
\[ X = U + \epsilon, \ Y = V + \nu, \] where \( \epsilon \) and \( \nu \) are independent mean 0 Gaussian variables with standard deviation equals to 0.45 times the grid spacing.

(4) \( X \) is uniform on the interval \((0, 1)\). Given \( X = x \), \( Y \) has a beta distribution with parameters \( 1 + 0.1 \sin(2\pi x) \) and \( 1 - 0.1 \sin(2\pi x) \).

(5) Similar to 3. The main difference is that the grid is 8 by 8 (and thus the distribution is more balanced). With probability 0.5 \((U, V)\) are sampled from the odd points, and the standard deviation of the normal distribution is 0.3 times the grid spacing.

The four tests are based on the ranks of the observations. To save computation time, the data were discretized to a 64 by 64 contingency table. The counts were then standardized to have mean 0 and variance 1. Let \( w = (w_1, w_2, \ldots) \) be the Haar wavelet decomposition of the resulting table, as given by the Matlab function \texttt{wavedec2} (level equals 5).

The tests are:

\begin{itemize}
  \item \textbf{[MAX]:} The maximal absolute value of one of the first 200 wavelets coefficients.
  \item \textbf{[L2]:} The L2 norm of \( w \) after soft-threshold according to the Matlab default.
  \item \textbf{[VM]:} The L2 norm, but with weights equal to \( i^{-2} \).
  \item \textbf{[KS]:} \( \max_{k,l} \sum_{i=1}^{k} \sum_{j=1}^{l} T_{i,j} \), where \( T_{i,j} \) are the entries of the 64 \times 64 contingency table.
\end{itemize}

The first of these tests evidently considers a “slippage” deviation with just one direction most important. The second is a proxy for the \chi^2 test. The third is a weighted \chi^2 placing more weight on lower frequency (lower level) departures. The fourth is a proxy for the Kolmogorov–Smirnov test.

Since, after reduction to the rank statistics, the null is simple, the tests were conducted as follows. 2000 Monte–Carlo simulations from the null were taken, and the test statistics were calculated. Rejection was declared if the test statistics were larger than 95% of the values calculated under the null. All powers estimated were based on 500 simulations.

As Table 1 demonstrates the tests behave more or less as expected. \( L_2 \) is never outstanding and does quite poorly against 2 and 4. \( \text{MAX} \) interestingly enough is best for 1, 3, and 5 but is very poor for 2 and 4. It is apparent that in accordance with Janssen (2000) having good power is possible only in a limited number of directions for any sample size and only extremely strong deviations in other directions.

6. Appendix

A. Proof of Theorem 3.2

We need

\textbf{Lemma A.1.} If \((N1)\) and \((N2)\) hold then so does \((M2)\),

\[ \sup \{ \| P_0(h) - P_0(h) + P_0(\Pi(h, \hat{a})) \|_{\infty} : h \in \mathcal{H} \} = o_P(n^{-1/2}). \]
Table 1: Simulations

* The power was increased to 0.682 when the threshold was reduced to half the Matlab default.

<table>
<thead>
<tr>
<th>Alternative</th>
<th>MAX</th>
<th>L2</th>
<th>VM</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7820</td>
<td>0.1480</td>
<td>0.0680</td>
<td>0.2440</td>
</tr>
<tr>
<td>2</td>
<td>0.0380</td>
<td>0.0740</td>
<td>0.5940</td>
<td>0.7480</td>
</tr>
<tr>
<td>3</td>
<td>0.5300</td>
<td>0.2960</td>
<td>0.0520</td>
<td>0.0500</td>
</tr>
<tr>
<td>4</td>
<td>0.0720</td>
<td>0.0860</td>
<td>0.7960</td>
<td>0.8480</td>
</tr>
<tr>
<td>5</td>
<td>0.9160</td>
<td>0.5500*</td>
<td>0.0680</td>
<td>0.0920</td>
</tr>
</tbody>
</table>

Proof. By (N1) we may W.L.O.G. assume \( \hat{\alpha} \in A_0 \). For simplicity let \( \mu \equiv P_0 \) so that \( s_0 \equiv 1 \). Obvious modifications suffice if \( \mu \gg P_0 \). Then

\[
P_{\hat{\alpha}}(h) - P_0(h) = \int h\hat{s}^2d\mu - \int hd\mu = 2\int h(\hat{s} - 1)d\mu + \int h(\hat{s} - 1)^2d\mu.
\]  

(A.1)

Since \( P_{\hat{\alpha}}\Pi(h, \hat{\alpha}) = 0 \),

\[
P_0(\Pi(h, \hat{\alpha})) = -\int \Pi(h, \hat{\alpha})(\hat{s}^2 - 1)d\mu = -2\int \Pi(h, \hat{\alpha})(\hat{s} - 1)d\mu + \int \Pi(h, \hat{\alpha})(\hat{s} - 1)^2d\mu.
\]  

(A.2)

But, since \( P_{\hat{\alpha}} \ll P_0 \) we have by BKRW formula (4b) on p. 50,

\[
\Pi(h, \hat{\alpha}) = \hat{s}^{-1}\Pi(h\hat{s}, \alpha_0).
\]  

(A.3)

Therefore, if \( \frac{\hat{s} - 1}{\hat{s}} \in L_2(\mu) \),

\[
\int \Pi(h, \hat{\alpha})(\hat{s} - 1)d\mu = \int \Pi(h\hat{s}, \alpha_0)\left(\frac{\hat{s} - 1}{\hat{s}}\right)d\mu = \int h\hat{s}\Pi\left(\frac{\hat{s} - 1}{\hat{s}}, \alpha_0\right)d\mu.
\]  

(A.4)
Substituting in (A.2) we get after some manipulation

\[ P_0(\Pi(h, \hat{\alpha})) = -2 \int h \Pi \left( \frac{\hat{s} - 1}{\hat{s}}, \alpha_0 \right) \hat{s} d\mu + \int \Pi(h, \hat{\alpha})(\hat{s} - 1)^2 d\mu \]

\[ = -2 \left\{ \int h \Pi (\hat{s} - 1, \alpha_0) d\mu + \int h(\hat{s} - 1) \Pi (\hat{s} - 1, \alpha_0) d\mu \right. \]

\[ \left. - \int \hat{s} \Pi \left( \frac{(\hat{s} - 1)^2}{\hat{s}}, \alpha_0 \right) d\mu \right\} + \int \Pi(h, \hat{\alpha})(\hat{s} - 1)^2 d\mu \]

\[ = -2 (I + II + III) + IV. \]  

(A.5)

We bound the last three terms in absolute value by

\[ |II| \leq M \int |\hat{s} - 1| \Pi(\hat{s} - 1, \alpha_0) d\mu \leq M \| \hat{s} - 1 \|_\mu \leq M \| (\hat{s} - 1)^2 \|_\mu \]

\[ |III| \leq M \| \Pi \left( \frac{(\hat{s} - 1)^2}{\hat{s}}, \alpha_0 \right) \|_\mu \leq M \| (\hat{s} - 1)^2 \|_\mu \]

where \( M = \sup_{H} \{ \| h \|_\infty + \| \Pi(h, \alpha_0) \|_\infty \} < \infty \) by (N2).

Again, using (N2),

\[ |IV| \leq M \| \hat{s} - 1 \|_\mu^2 \leq M \| (\hat{s} - 1)^2 \|_\mu. \]

Combining (A.1)–(A.5) we obtain

\[ P_0(\hat{h}) - P_0(h) + P_0(\Pi(h, \hat{\alpha})) = 2 \int h((\hat{s} - 1) - \Pi(\hat{s} - 1, \alpha_0)) d\mu \]

\[ + O_P \left( \| (\hat{s} - 1)^2 \|_\mu \right) \]

\[ = O_P \left( \| \hat{s} - 1 - \Pi(\hat{s} - 1, \alpha_0) \|_\mu + \| (\hat{s} - 1)^2 \|_\mu \right) \]

\[ = o_P(n^{-1/2}) \]

by (N1) and the lemma follows.

\textit{Proof of Theorem 3.2.} Write \( \hat{h} \) for \( h_\gamma(\cdot, \hat{\alpha}) \) and \( h \) for \( h_\gamma(\cdot, \alpha_0) \). Then,

\[ \hat{Z}_n(\gamma) = n^{1/2} \left\{ P_n \hat{h} - P_n \hat{h} - P_n \Pi(\hat{h}, \hat{\alpha}) \right\} \]

\[ = n^{1/2} \left\{ (P_n - P_0)(\hat{h}) - (P_n - P_0)(\hat{h}) - P_n \Pi(\hat{h}, \hat{\alpha}) \right\} \]

\[ = n^{1/2} (P_n - P_0)(\hat{h} - \Pi(\hat{h}, \hat{\alpha})) + o_P(1) \]

by (M2). But

\[ n^{1/2} (P_n - P_0)(\hat{h} - \Pi(\hat{h}, \hat{\alpha})) = n^{1/2} (P_n - P_0)(h - \Pi(h, \alpha_0)) + o_P(1) \]

\[ = Z_n(\gamma, \alpha_0) + o_P(1) \]

by (M3) and (M1). (The equivalences hold uniformly in \( \gamma \) by assumption.) \( \square \)
REFERENCES


