Reconciling Model Selection and Prediction

George Casella

Department of Statistics, University of Florida, Gainesville, FL 32611.

casella@stat.ufl.edu

Guido Consonni

Dipartimento di Economia Politica e Metodi Quantitativi

University of Pavia, 27100 Pavia, Italy.

guido.consonni@unipv.it

March 20, 2009

Abstract

It is known that there is a dichotomy in the performance of model selectors. Those that are consistent (having the "oracle property") do not achieve the asymptotic minimax rate for prediction error. We look at this phenomenon closely, and argue that the set of parameters on which this dichotomy occurs is extreme, even pathological, and should not be considered when evaluating model selectors. We characterize this set, and show that, when such parameters are dismissed from consideration, consistency and asymptotic minimaxity can

be attained simultaneously.

1 Introduction

Model selection is an important area of statistical practice and research. However, model selection often represents a first step towards our main goal, which may be estimation or prediction. A stylized scheme is the following: first use a model-selection procedure to select a model, and then proceed with inference conditionally on the chosen model. This method leads to so-called "post-model-selection" estimators (or predictors). Our concern here is with the asymptotic risk of such procedures.

The review paper Leeb & Pötscher (2005) argues that the (data-driven) model selection step typically has dramatic effect on the sampling properties of the estimators; see also Leeb & Pötscher (2006). These properties are quite different from their single-model counterpart, and cannot be ignored even when the sample size is large and when the model selector is consistent.

Although interest in the performance of post-model selection estimators has gained momentum over the last years, the problem has been around for a few decades and can be traced backed, in its essence, to Hodges' estimator. For an interesting discussion of Hodges' estimator see van der Vaart (1998, Example 8.1).

Within the context of regression functions, and for squared-error loss, Yang (2005) has shown that consistent model selection procedures, such as BIC, produce estimators which cannot attain the asymptotic minimax rate. Failure to attain this optimal rate extends also to model combination, or Bayesian model averaging with subjectively specified priors. On the other hand, it is known that AIC, which is inconsistent, does attain the minimax rate; see Yang (2005, Proposition 1). This tension between

model consistency and inference-optimality is sometimes referred to as the "AIC-BIC dilemma".

Attempts at overcoming this dilemma include adaptive model selection, which, unlike AIC or BIC, employs a data-driven penalty to achieve both consistency and optimality; for a brief account see Yang (2005, sec. 1.4). Recently, van Erven, Grünwald and Rooij introduced the notion of "switch distribution", as an alternative to standard model selection methods, such as the Bayes factor and leave-one-out cross-validation, in an effort to combine the strengths of AIC and BIC; see their 2008 Technical Report entitled Catching Up Faster by Switching Sooner: a Prequential Solution to the AIC-BIC Dilemma (arXiv:0807.1005v1 [math.ST]).

The remainder of the paper is organized as follows. In Section 2 we review the criteria for asymptotic comparison of tests based on power against local alternatives, and show how it relates to the prediction problem in linear regression. In Section 3 we revisit the result of Yang (2005) for the simple linear regression example. In particular we provide an evaluation of the proof he gave for his Theorem 1, and we link his sequence of alternatives to the categorization given in Lemma 2.1. This puts into perspective, and actually explains why the failure to attain the minimax rate occurs; we also show that the minimax rate is achieved for sequences of type 3 in Lemma 2.1, which are recognized as the only reasonable sequences for the asymptotic comparison of tests. In Section 4 we comment on the use of *contiguity* for proving lack of minimax rate for consistent model selectors. Finally, Section 5 offers some concluding remarks.

2 Asymptotic Comparison of Tests

In this section we review some results on asymptotic test comparison. In particular we are interested in the categorization of local alternatives, and how their convergence to the null value interacts with the power of a test. We then look at the simple example of model selection in linear regression.

2.1 Categorizing Alternatives

Consider a sequence of statistical models $\{P_{n,\theta}, \theta \in \Theta\}$ for observations $\mathbf{y}_n := (y_1, \ldots, y_n)$, $n = 1, 2, \ldots$, where we want to test $H_0 : \theta = \theta_0 \; vs \; H_1 : \theta > \theta_0$. If $\pi_n(\theta)$ is the power function of a test, for most reasonable tests it holds true that $\lim_{n \to \infty} \pi_n(\theta) = 0$ if $\theta = \theta_0$, and $\lim_{n \to \infty} \pi_n(\theta) = 1$ if $\theta > \theta_0$. This is to be expected because, with arbitrarily many observations, it should be possible to tell the null and the alternative apart with complete accuracy. This fact means that to compare tests asymptotically we should make the problem "harder". One way to do this is to consider a sequence of testing problems

$$H_0 : \theta = \theta_0 \; vs \; H_{1n} : \theta = \theta_n, \tag{1}$$

where $\theta_n > \theta_0$ and $\theta_n \to \theta_0^+$ in specific ways.

The L_1-distance between P_{n,θ_0} and P_{n,θ_n} provides a useful characterization of the power function associated with (1). Denote the L_1-distance between two probability measures P and Q (having density p, respectively q, with respect to a common

measure μ) by $||P - Q||$. Then

$$||P - Q|| \stackrel{def}{=} \int |p - q| d\mu = 2 \sup_A |P(A) - Q(A)| \stackrel{def}{=} 2||P - Q||_{TV},$$

where the second equality follows from the well-known relationship between L_1 distance and *total variation norm* $||P - Q||_{TV}$.

The following lemma relates the L_1-distance to the power function.

Lemma 2.1 (VAN DER VAART, 1998, LEMMA 14.30)

The power function π_n of any test satisfies

$$\pi_n(\theta) - \pi_n(\theta_0) \leq \frac{1}{2}||P_{n,\theta} - P_{n,\theta_0}||. \tag{2}$$

For any θ and θ_0 there exists a test whose power function attains equality.

The implications of Lemma 2.1 are

1. If $||P_{n,\theta_n} - P_{n,\theta_0}|| \to 2$, then the sequence θ_n is converging to θ_0 at a *slow* rate, so that the two hypotheses are *strongly separated*. In this case the *difference* $\pi_n(\theta) - \pi_n(\theta_0)$ tends to 1, which means that we can get all sort of tests; in particular, since equality can be attained, there exist a sequence of tests with power tending to 1 and size tending to 0 (a *perfect* sequence of tests).

2. If $||P_{n,\theta_n} - P_{n,\theta_0}|| \to 0$, then the sequence θ_n is converging to θ_0 at a *fast* rate, so that the two hypotheses are *weakly separated*. In this case the power of any sequence of tests is asymptotically less than the level (every sequence of tests is *worthless*).

6

3. If $||P_{n,\theta_n} - P_{n,\theta_0}||$ is bounded away from 0 and 2, then the sequence θ_n is con-

   verging to θ_0 at a rate such that the two hypotheses are *well separated*. In this

   case, there exists no perfect sequence of tests, but not every test is worthless

   either.

The consensus in the literature, see for example Lehmann & Romano (2005, sec.

13.1) or van der Vaart (1998, sec. 14.5), is that situation 3 is the only reasonable one

for the comparison of tests, otherwise the problem is "asymptotically degenerate".

For iid observations from smooth models, case 3 occurs when θ_n converges to θ_0 at

rate $1/\sqrt{n}$.

Easier calculation often results when using *Hellinger distance* rather than the L_1-

distance. The Hellinger distance between P and Q is $H(P,Q) = \left\{ \int (\sqrt{p} - \sqrt{q})^2 d\mu \right\}^{1/2}$,

and we can rewrite its square as $H^2(P,Q) = 2 - 2A(P,Q)$, where $A(P,Q) = \int \sqrt{pq} d\mu$

is called the *Hellinger affinity*. The following inequality holds (van der Vaart, 1998):

$$H^2(P,Q) \leq ||P - Q|| \leq \min\{2 - A^2(P,Q), 2H(P,Q)\}. \tag{3}$$

2.2 Example: Simple Linear Regression

Here we look in detail at the simple testing problem considered by Yang (2005). For

$i = 1, \ldots, n$ let

$$H_0 : y_i = \epsilon_i \text{ and } H_1 : y_i = \beta x_i + \epsilon_i; \quad \beta > 0, \tag{4}$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$.

Recall that two sequences a_n and b_n are said to be of the same order, written $a_n \asymp$

b_n, when there exist constants $0 < r < R < \infty$ and an integer n_0 such that, for $n > n_0$, $r < |a_n/b_n| < R$. We start with a simple lemma whose proof is straightforward.

Lemma 2.2 *Let $P_{n,0}$ be the probability measure associated with H_0 and $P_{n,\beta}$ that associated with H_1 in (4). Assume that $\lim_{n\to\infty} \sum_{i=1}^n x_i^2/n$ is a strictly positive constant, so that $\sum_{i=1}^n x_i^2 \asymp n$. The Hellinger affinity is*

$$A(P_{n,0}, P_{n,\beta}) = \exp\left\{ -\frac{\beta^2}{8} \sum_{i=1}^n x_i^2 \right\}. \tag{5}$$

As a direct consequence of this lemma, we can characterize sequences β_n as follows.

1. if $\beta_n = c_n/\sqrt{n}$, with $c_n \to \infty$, then $||P_{n,0} - P_{n,\beta_n}|| \to 2$ and the two models are strongly separated;

2. if $\beta_n = c_n/\sqrt{n}$, with $c_n \to 0^+$, then $||P_{n,0} - P_{n,\beta_n}|| \to 0$ and the two models are weakly separated;

3. if $\beta_n \asymp (1/\sqrt{n})$, then $||P_{n,0} - P_{n,\beta_n}||$ is bounded away from 0 and 2, so that the two hypotheses are well separated.

Following Lehmann & Romano (2005, p. 498) we conclude that also for this regression problem the problem of testing $P_{n,0}$ versus P_{n,β_n} is degenerate unless $\beta_n \asymp 1/\sqrt{n}$. These are therefore the only meaningful local alternative sequences for evaluating tests.

2.3   Consistency and Prediction

The simple testing problem described in (4) can be cast as a model selector by defining $A_n = \{(x_i, y_i), i = 1, \ldots n : H_1 \text{ is selected }\}$, and estimating β with the post-model-

8

selection estimator $\hat{\beta}I(A_n)$, where $I(A_n)$ is the indicator function of the set A_n. For simplicity we take $\hat{\beta}$ to be the least squares estimator under H_1.

Given $(x_i, y_i), i = 1, \ldots n$, we run our model selection procedure and then we predict at values $x_i^*, i = 1, \ldots, m$, where the x_i^* may be the same as the original x_i, or not. The average prediction error is

$$\frac{1}{m}\sum_{i=1}^{m}(\beta x_i^* - \hat{\beta}I(A_n)x_i^*)^2 = \left(\frac{1}{m}\sum_{i=1}^{m}x_i^{*\,2}\right)(\beta - \hat{\beta}I(A_n))^2,$$

which shows why it doesn't matter whether we predict x_i^* or x_i, or how many x_i^* we predict (as long as m/n is finite as $n \to \infty$). Taking expectations gives the predictive risk function

$$
\begin{aligned}
R_m(\beta, A_n) &= \left(\frac{1}{m}\sum_{i=1}^{m}(x_i^*)^2\right)\mathbb{E}_\beta(\beta - \hat{\beta}I(A_n))^2 \\
&= \left(\frac{1}{m}\sum_{i=1}^{m}(x_i^*)^2\right)[\mathbb{E}_\beta\{(\beta - \hat{\beta})^2 I(A_n)\} + \beta^2 P_\beta(A_n^c)].
\end{aligned}
$$

The predictor attains the asymptotic minimax rate if $n \sup_\beta R_m(\beta, A_n) \to$ constant as $n \to \infty$. The quantity $nR_m(\beta, A_n)$ is called the *scaled* risk (function). Recalling that $\sum_{i=1}^{m}(x_i^*)^2 \asymp m$, we only need to be concerned with

$$n \sup_\beta R_m(\beta, A_n) \asymp n[\mathbb{E}_\beta\{(\beta - \hat{\beta})^2 I(A_n)\} + \beta^2 P_\beta(A_n^c)]. \tag{6}$$

The first term is bounded by $n\mathbb{E}_\beta\{(\beta - \hat{\beta})^2 = n\sigma^2/\sum_{i=1}^{n}x_i^2$, so its limit is a positive constant. Thus, the model selector *achieves* the minimax rate if and only if

$$\sup_\beta n\beta^2 P_\beta(A_n^c) \to \text{constant as } n \to \infty. \tag{7}$$

Yang (2005) found a sequence β_n converging to zero, equivalently a sequence of alternative models converging to the null-model, *slowly enough* to have $n\beta_n^2 \to \infty$, but at

9

the same time *fast enough* to "confuse" the model selector, leading it to choose the null model; this keeps $P_{\beta_n}(A_n^c)$ away from zero, and actually arbitrarily close to one, so that the minimax rate is not achieved.

3 The Prediction/Minimax Rate Conflict

In this section we describe the sequence β_n, chosen by Yang, to establish his result. We also show that, with a minor modification, a similar sequence attains the minimax rate. We then look a bit closer at Yang's sequence, showing how it unfairly "confuses" the model selector.

3.1 Yang's Sequence

For the model selection problem described in (4), the model selector is *consistent* if

$$P(A_n) \to 0 \text{ if } H_0 \text{ is true and } P(A_n) \to 1 \text{ if } H_1 \text{ is true.}$$

Yang (2005) shows that a consistent model selector cannot achieve the minimax prediction rate using the following argument. He considers a sequence of UMP tests for the hypotheses (4) having power function

$$\pi_n(\beta) = \Pr_\beta\{\sum_{i=1}^n x_i y_i \geq d_n\} = \Pr_\beta\{Z \geq \frac{d_n - \beta \sum_{i=1}^n x_i^2}{\sqrt{\sum_{i=1}^n x_i^2}}\}, \tag{8}$$

where $Z \sim N(0,1)$, and equates $\Pr_{\beta=0}(A_n)$ with $\pi_n(0)$, for each n. Note that the set $A_n^* = \{(x_i, y_i) : \sum_i x_i y_i > d_n\}$ defines a model selector based on the UMP test, where we would estimate β_n with the least squares estimator if this set occurred.

10

The requirement of consistency means that we must have $\pi_n(0) \to 0$ which implies

$$\frac{d_n}{\sqrt{\sum_{i=1}^n x_i^2}} \to \infty. \tag{9}$$

Yang's objective is to find a sequence of alternatives $\beta_n \to 0^+$ such that

$$C1: \quad n\beta_n^2 \to \infty \quad \text{and} \quad C2: (1 - \pi_n(\beta_n)) \to \text{constant} > 0, \tag{10}$$

which would imply that $n\beta_n^2 \text{Pr}_{\beta_n}(A_n^c) \to \infty$ (because the test is UMP), and hence lead to the conclusion that a consistent model selector does *not* attain the minimax rate, that is, it violates (7). Specifically, Yang's choice for β_n is

$$\beta_n = \frac{1}{2} \frac{d_n}{\sum_{i=1}^n x_i^2}, \tag{11}$$

which satisfies $C1$ and $C2$. (There is a typo in Yang (2005), as noted by the author on his webpage: on lines 4 and 6 of page 947 the factor 2 should be in the denominator.)

We now look at sequence (11) a bit more closely. Write

$$\beta_n = \frac{c_n}{\sqrt{\sum_{i=1}^n x_i^2}}, \quad \text{with } c_n = \frac{1}{2} \frac{d_n}{\sqrt{\sum_{i=1}^n x_i^2}} \to \infty, \tag{12}$$

and, recalling that $\sum_{i=1}^n x_i^2 \asymp n$, we immediately conclude that we are in scenario 1 of Lemma 2.2, namely that the two hypotheses are strongly separated. In this case any type of test can occur, and with Yang's choice the power function is asymptotically zero:

$$\pi_n(\beta_n) = \text{Pr}_{\beta_n}\{\sum_{i=1}^n x_i y_i \geq d_n\} = \text{Pr}\{Z \geq \frac{1}{2} \frac{d_n}{\sqrt{\sum_{i=1}^n x_i^2}}\} \to 0,$$

which follows from (9) and the fat that $\sum_i x_i y_i \sim \text{N}(\beta_n \sum_i x_i^2, \sum_i x_i^2)$.

Yang's sequence for β_n thus produces a worthless test, since both the size (by assumption) and the power (by construction) tend to zero. In this case we know that

11

there even exists a perfect sequence of tests (we will construct one below). Clearly

Yang's argument holds for any sequence $\beta_n = b\frac{d_n}{\sum_{i=1}^n x_i^2}$, $0 < b < 1$. If $b = 1$ we get

$\pi_n(\beta_n) \to 1/2$, which would still support Yang's argument, although the test is no

longer worthless (but actually rather poor because its power is fixed at $1/2$ for each

n).

Looking at Yang's result from a testing perspective reveals one of its weak points.

His result holds because he chooses a sequence of alternatives that, while producing

asymptotically a strong separation between the two models, converges to the null

along a path leading to a worthless, or at best a mediocre, test. This happens despite

the test being UMP, and despite the fact that, by a minor modification, one could

get a perfect sequence of tests, as the following example shows.

Example 3.1 *Consider the UMP test for the hypotheses (4). Choose*

$$\beta_n = (1 + b')\frac{d_n}{\sum_{i=1}^n x_i^2}, \; b' > 0. \tag{13}$$

Then

$$\pi_n(\beta_n) = \Pr\{Z \geq -b'\frac{d_n}{\sqrt{\sum_{i=1}^n x_i^2}}\} \to 1,$$

which follows from (9) and the fact that $b' > 0$. As a consequence, by choosing a

sequence of alternatives which is structurally equivalent to Yang's, although uniformly

larger by a factor $(1 + b')/b$, $b' > 0$, $0 < b < 1$, we get a perfect sequence of tests.

$\square$

With the choice of sequence (13), $1 - \pi_n(\beta_n)$ goes to zero exponentially fast, and

it is easy to show that $n\beta_n^2(1 - \pi_n(\beta_n)) \to 0$, instead of going to ∞ as under Yang's

choice (11). Thus, under this sequence of alternatives, we even beat the minimax rate! Example 3.1 reinforces the view that Yang's result is based on a rather artificial sequence. Next we provide further insight into his choice of sequence β_n.

3.2 Confusing the Model Selector

In Section 2.3 we noted that Yang's sequence β_n was constructed in such a way as to "confuse" the model selector. We now look a bit more closely at this claim.

What happens with Yang's sequence is that, as $n \to \infty$, all of the mass of the distribution of $\sum_i x_i y_i$ is concentrated in the acceptance region of the test (that is, in A_n^c), even when $\beta_n > 0$, making the "correct" decision that of accepting H_0. To see this, recall that, for $\beta = \beta_n$, the UMP test-statistic $\sum_i x_i y_i$ is distributed as $N(\beta_n \sum_i x_i^2, \sum_i x_i^2)$. Consider the probability

$$\Pr\left(\sum_i x_i y_i < \beta_n \sum_i x_i^2 + M\sqrt{\sum_i x_i^2}\right),$$

which grows arbitrarily close to 1 as M increases. Now set $\beta_n = b\frac{d_n}{\sum_{i=1}^n x_i^2}$, $0 < b < 1$ as in Yang's choice. Then, for any $M > 0$, $\beta_n \sum_{i=1}^n x_i^2 + M\sqrt{\sum_{i=1}^n x_i^2} < d_n$ eventually (that is, as n grows), because the previous inequality is equivalent to

$$M < (1 - b)\frac{d_n}{\sqrt{\sum_{i=1}^n x_i^2}}, \tag{14}$$

which holds true since the right-hand-side tends to infinity because of (9). But this means that the support of the UMP-test statistic under this sequence of alternatives is eventually disjoint from the H_1-acceptance region of the test: this is why the test is fooled and chooses H_0 incorrectly, with probability tending to one. Notice that if $b = 1$ then (14) does not hold.

13

Finally, for the sequence with $\beta_n = (1 + b') \frac{d_n}{\sum_{i=1}^n x_i^2}$, $b' > 0$, then $\beta_n \sum_{i=1}^n x_i^2 - M \sqrt{\sum_{i=1}^n x_i^2} > d_n$, eventually, because the previous inequality is equivalent to

$$- M > -b' \frac{d_n}{\sqrt{\sum_{i=1}^n x_i^2}}, \tag{15}$$

which holds true since the right-hand-side tends to $-\infty$ because of (9). This means that the support of the UMP-test statistic under this sequence of alternatives is eventually contained in the H_1-acceptance region of the test: the test correctly chooses H_1 with probability tending to one, and thus chooses H_0 with probability tending to zero. The model-selector estimator attains the minimax rate.

4   Contiguous Sequences

Leeb & Pötscher (2005, Appendix C, Proposition C.1) present a result which is comparable to that of Theorem 1 in Yang (2005). They consider a linear regression model with mean structure $\alpha x_{1i} + \beta x_{2i}$ under the unrestricted case, and mean structure αx_{1i} under the reduced model. They deal, among other things, with the scaled risk, under squared-error loss, of the post-model selection estimator (least squares) of α. As in Yang (2005), they claim that its supremum diverges to infinity whenever the model selection procedure is consistent. Although Yang is concerned with prediction and not estimation, the connection between the two results is apparent in the case of normal errors. Yet, Leeb and Pötscher's argument is quite different from Yang's, because it relies on the notion of *contiguity*. For an introduction to the notion of contiguity, see van der Vaart (1998, sec. 6.2) and Lehmann & Romano (2005, sec. 12.3). Here we revisit Yang's problem using the notion of contiguity. We provide an evaluation of

this technique for the problem at hand and raise some critical issues.

Let P_n and Q_n be measures on a measurable spaces $(\Omega_n, \mathcal{A}_n)$.

Definition 4.1 *The sequence Q_n is contiguous with respect to the sequence P_n if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ for every sequence of measurable sets A_n. This is denoted $Q_n \triangleleft P_n$.*

One can regard contiguity as the asymptotic analogue of the classic notion of absolute continuity of measures. The strength of contiguity stems from the fact that Q_n-limit law of random vectors $U_n : \Omega_n \mapsto \mathbb{R}^k$ can be obtained from suitable P_n-limit laws; the usefulness of such result is apparent when the latter calculations are much easier than the former. If Q_n is contiguous with respect to P_n, and *viceversa*, then we write $Q_n \triangleleft \triangleright P_n$.

The following result considers the model selection problem discussed by Yang, and relates it to the notion of contiguity.

Proposition 4.1 *Consider the problem described in (4). Let $P_{n,0}$ be the sequence of probability measures under the null model H_0, and P_{n,β_n} be the sequence of probability measures corresponding to the local alternative models $H_{1n} : y_i = \beta_n x_i + \epsilon_i$, $\beta_n > 0$, $\beta_n \to 0^+$. Then $P_{n,\beta_n} \triangleleft P_{n,0}$ if and only if $\beta_n = O(1/\sqrt{n})$; additionally $P_{n,\beta_n} \triangleleft \triangleright P_{n,0}$ under the same condition.*

Proof. The proof is essentially the same as that for proving contiguity of the joint distribution of n iid observations from a Normal with mean ξ_n and variance 1 with respect to the joint distribution of n iid with observations from a Normal with mean

15

0 and variance 1; see Lehmann & Romano (2005, examples 12.3.3 and 12.3.6). To see why, simply notice that the likelihood ratio is

$$dP_{n,\beta_n}/dP_{n,0} = \exp\{\beta_n \sum_{i=1}^{n} x_i y_i - (\beta_n^2/2) \sum_{i=1}^{n} x_i^2\}.$$

Under $P_{n,0}$, $\sum_{i=1}^{n} x_i y_i \sim N(0, \sum_{i=1}^{n} x_i^2)$, and thus,

$$\beta_n \sum_{i=1}^{n} x_i y_i - (\beta_n^2/2) \sum_{i=1}^{n} x_i^2 \sim N((-\beta_n^2/2) \sum_{i=1}^{n} x_i^2, \beta_n^2 \sum_{i=1}^{n} x_i^2),$$

again under $P_{n,0}$. So the only difference between the simple regression case we are discussing and the iid case from a Normal is that the former has $\sum_{i=1}^{n} x_i^2$ while the latter has n. Since these two quantities are asymptotically of the same order, one can use the same argument in either case. ∎

From Proposition 4.1 it appears that $P_{n,\beta_n} \lhd P_{n,0}$ if and only if the sequence $n\beta_n^2$ remains bounded; under the same condition mutual contiguity holds. In particular even if $\beta_n \to 0$, but at a slower rate than $1/\sqrt{n}$, as in Yang's case, see (12), then contiguity of P_{n,β_n} fails. Notice that contiguity holds also if $\beta_n = o(1/\sqrt{n})$, because $\beta_n^2 \sum_{i=1}^{n} x_i^2$ goes to zero and hence is bounded. However this case is of no interest for proving failure to attain the minimax rate, because condition C1 in (10) is not satisfied (that is, $n\beta_n^2$ does not diverge but actually goes to zero).

How can we use contiguity to obtain a result similar to Yang's? Here is the idea. Yang's result obtains if we show that

$$\lim_{n\to\infty} n\beta_n^2 \mathrm{Pr}_{P_{n,\beta_n}} \{A_n^c\} = \infty$$

To exploit contiguity, β_n must be of order $1/\sqrt{n}$. Set for definiteness $\beta_n = r/\sqrt{n}$, for

some positive *fixed* r. We get

$$\lim_{n\to\infty} \mathrm{Pr}_{P_{n,r/\sqrt{n}}}\{A_n^c\} = \lim_{n\to\infty} \mathrm{Pr}_{P_{n,0}}\{A_n^c\} = 1,$$

where the first equality sign follows from contiguity of $P_{n,r/\sqrt{n}}$ with respect to $P_{n,0}$, while the second is a consequence of the assumed consistency of the model selector (recall that A_n^c means accepting $H_0 : \beta = 0$). Therefore

$$\lim_{n\to\infty} nr^2(1/\sqrt{n})^2 \mathrm{Pr}_{P_{n,r/\sqrt{n}}}\{A_n^c\} = r^2. \tag{16}$$

At this stage it would seem that the minimax rate *is* attained under this sequence, because the limit, however large, is finite. To circumvent (16), and get the opposite conclusion that the rate is actually infinite, one ought to apply the argument in Leeb & Pötscher (2005, p. 59), and let r grow *arbitrarily large* (technically this amounts to take a further limit $r \to \infty$). However, there is a subtle difference between this argument and Yang's result.

First of all, to conclude that $\lim_{n\to\infty} n\beta_n^2 \mathrm{Pr}_{P_{n,\beta_n}}\{A_n^c\} = \infty$ one should prove that, for any $R > 0$, there exists an n_0 such that

$$n > n_0 \Rightarrow n\beta_n^2 \mathrm{Pr}_{P_{n,\beta_n}}\{A_n^c\} > R. \tag{17}$$

Having set $\beta_n = r/\sqrt{n}$, condition (17) translates to

$$n > n_0 \Rightarrow \mathrm{Pr}_{P_{n,r/\sqrt{n}}}\{A_n^c\} > R/r^2. \tag{18}$$

Since r is *fixed*, condition (18) can be easily violated choosing for instance $R > r^2$.

Secondly, and possibly more importantly, the argument based on contiguity conveys the misconception that sequences of alternatives β_n of order $1/\sqrt{n}$ can fail to

attain the minimax rate. This would be quite surprising because, on the contrary, it is well known that this type of sequences is the *only* one which makes sense for asymptotic comparison of tests, as lucidly remarked, for instance, in Lehmann & Romano (2005, example 12.3.6).

In the light of the above remarks, and of (16), it should be clear that sequences of alternatives of order $1/\sqrt{n}$ *do* achieve the minimax rate.

5   Concluding Remarks

In this paper we have cast the so-called AIC-BIC dilemma into perspective. On the one hand it is true that estimators and predictors based on consistent model selection procedures may lead to an infinite scaled risk, thus failing to attain the usual minimax rate, as Yang (2005) showed. On the other hand, this phenomenon occurs only for sequences of alternatives which are strongly separated from the null model (this inflates the bias when the null model is chosen, while the alternative holds). But such sequences are well known to be of no use in asymptotic comparison of testing procedures, because they always admit a perfect sequence of tests (the power goes to 1 while the size goes to zero).

Additionally, the non-attainment of the minimax rate takes place only for a specific subset of these sequences, namely those whose support under the alternative is eventually fully contained in the null-acceptance region. This explains why the problem occurs: the selector (not surprisingly!) chooses the null model, although the alternative holds. Finally we have argued that contiguity arguments have little to

say with regard to the AIC-BIC dilemma: contiguity is synonymous with sequences of alternatives converging to the null at the appropriate rate $1/\sqrt{n}$: no pathological behaviour can occur in this case.

ACKNOWLEDGEMENT

References

LEEB, H. & PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* 21 21–59.

LEEB, H. & PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics* 34 2554–2591.

LEHMANN, E. L. & ROMANO, J. P. (2005). *Testing Statistical Hypotheses, Third Edition.* Springer.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92 937–950.