

TESTING FOR DIFFERENCES IN PREDICTIVE ACCURACY

Eva Ferreira¹ and Winfried Stute²

¹ Departamento de Economía Aplicada III, Univ. del País Vasco, Bilbao, Spain.
Email: eva.ferreira@ehu.es

² Mathematical Institute, University of Giessen, Arndtstr. 2, D-35392 Giessen, Germany.
Email: Winfried.Stute@math.uni-giessen.de

ABSTRACT

In this paper we provide new tests for the difference in predictive accuracy of two prognostic factors X_1 and X_2 on a common output Y . Given a set of independent replicates of (X_1, X_2, Y) , we split this sample into a learning part for estimating the unknown regression functions, and a validation part for which the residuals need to be computed. We show that the null distributions of our test statistics may be approximated by a normal. In simulations, the power is promising already for small to moderate sample sizes. Extensions to the time series context are briefly outlined.

KEYWORDS

Predictive accuracy, residuals, nonparametric test, data split

2000 Mathematics Subject Classification: Primary 62H15, Secondary 62F05, 62G10

1 INTRODUCTION

Suppose that Y is an unknown variable of interest. For example, Y could measure the future status of a company or a patient. Rather than Y , what might be available already now, is a vector X_1 of covariates which may be helpful to predict the value of Y . In a regression context we decompose Y into a term depending only on X_1 and a noise variable ε_1 orthogonal to X_1 , i.e.,

$$Y = m_1(X_1) + \varepsilon_1 \text{ such that } \mathbb{E}(\varepsilon_1 | X_1) = 0 \text{ a.s.}$$

The function m_1 is the regression function of Y w.r.t. X_1 . When Y is unknown, $m_1(X_1)$ is the best predictor based on X_1 , i.e., among all functions $\varphi(X_1)$ of X_1 , $m_1(X_1)$ is the one minimizing the expected squared prediction error. Unfortunately, in a practical situation, m_1 is unknown and needs to be estimated from a learning sample $(X_{11}, Y_1), \dots, (X_{1T}, Y_T)$ of independent replicates of (X_1, Y) . Denote with \hat{m}_1 any of such estimators, and let

$$e_1 := Y - \hat{m}_1(X_1)$$

be the associated residual. The availability of such data may be limited by the sampling costs for the covariate X_1 . Therefore it makes sense to also take into account an alternative covariate X_2 . Denote with

$$Y = m_2(X_2) + \varepsilon_2 \text{ such that } \mathbb{E}(\varepsilon_2|X_2) = 0 \text{ a.s.}$$

the corresponding decomposition. While $m_1(X_1)$ and $m_2(X_2)$ are optimal within their classes, it remains open which of $m_1(X_1)$ and $m_2(X_2)$ outperforms the other.

It is the purpose of this paper to provide some methodology for comparing the predictive accuracy of two prognostic factors X_1 and X_2 w.r.t. a common dependent variable Y . The quality of the fit is measured through $g(\varepsilon_1)$ and $g(\varepsilon_2)$, where g is a weight function chosen by the statistician. We already mentioned the quadratic loss associated with $g(u) = u^2$. In Robust Statistics, a popular weight is $g(u) = |u|$ leading to the mean absolute deviations $|\varepsilon_1|$ and $|\varepsilon_2|$. Another possibility would be

$$g(u) = u^2 1_{\{|u| > \delta\}}.$$

When one applies this g , one neglects deviations which fall below the threshold δ .

Our statistical analysis will be based on a learning (or estimation) sample $(X_{1t}, X_{2t}, Y_t), 1 \leq t \leq T$, of independent replicates of (X_1, X_2, Y) . The covariates X_{1t} and X_{2t} can be quite different. They may coincide in some of their coordinates but not in others. Their dimensions d_1 and d_2 may also differ. Later we shall in detail discuss the case when X_1 is a subvector of X_2 so that $d_1 < d_2$. The learning sample will be used to estimate the unknown regression functions m_1 and m_2 through \hat{m}_1 and \hat{m}_2 , say. In a parametric framework, m_1 and m_2 are of the type

$$m_1(x_1) = m_1(x_1, \beta) \text{ and } m_2(x_2) = m_2(x_2, \gamma).$$

In such a situation we have to estimate β and γ by the Least Squares Estimator or robust alternatives $\hat{\beta}$ and $\hat{\gamma}$. For \hat{m}_1 and \hat{m}_2 we then take the plug-in estimators

$$\hat{m}_1(x_1) = m_1(x_1, \hat{\beta}) \text{ and } \hat{m}_2(x_2) = m_2(x_2, \hat{\gamma}).$$

In a nonparametric framework one may take for \hat{m}_1 and \hat{m}_2 any nonparametric smoother. See Stone (1977) for general conditions on such smoothers to obtain universal consistency. Spiegelman and Sacks (1980) is a relevant reference for the consistency of the Nadaraya-Watson estimator.

Now, after having obtained \hat{m}_1 and \hat{m}_2 , the associated residuals are computed for a validation sample $(X_{1t}, X_{2t}, Y_t), T + 1 \leq t \leq T + n$, being independent of the first:

$$e_{it} = Y_t - \hat{m}_i(X_{it}), T + 1 \leq t \leq T + n, i = 1, 2.$$

For a given weight function g , a comparison of the predictive accuracy will now be based on

$$\bar{d} = n^{-1} \sum_{t=T+1}^{T+n} [g(e_{1t}) - g(e_{2t})].$$

Typically, a large value of \bar{d} indicates that X_1 has less predictive accuracy than X_2 . A test for

$$H_0 : \mathbb{E}g(\epsilon_1) = \mathbb{E}g(\epsilon_2)$$

versus

$$H_1 : \mathbb{E}g(\epsilon_1) > \mathbb{E}g(\epsilon_2) \tag{1.1}$$

rejects H_0 in favor of H_1 when \bar{d} exceeds a critical value. In general, the null distribution is very complicated. A quantity which is much easier to handle is one which is obtained after replacing the residuals by the true errors:

$$\bar{d}_1 = n^{-1} \sum_{t=T+1}^{T+n} [g(\epsilon_{1t}) - g(\epsilon_{2t})].$$

Under H_0 , this is a sum of centered independent identically distributed summands to which the Central Limit Theorem (CLT) applies. In our main results we show that

$$n^{1/2}\bar{d} = n^{1/2}\bar{d}_1 + o_{\mathbb{P}}(1) \quad \text{as } n \rightarrow \infty \tag{1.2}$$

under appropriate conditions so that by Slutsky's theorem the distribution of $n^{1/2}\bar{d}$ can in fact be approximated by a normal. Before we come to the main results, some further comments are in order.

Remark 1. Suppose that all components of X_1 are also included in X_2 . For $g(u) = u^2$ we then have

$$\begin{aligned} \mathbb{E}\epsilon_1^2 - \mathbb{E}\epsilon_2^2 &= \mathbb{E}[Y - m_1(X_1)]^2 - \mathbb{E}[Y - m_2(X_2)]^2 \\ &= \mathbb{E}m_1^2(X_1) - \mathbb{E}m_2^2(X_2) - 2\mathbb{E}[Ym_1(X_1)] + 2\mathbb{E}[Ym_2(X_2)] \\ &= \mathbb{E}m_2^2(X_2) - \mathbb{E}m_1^2(X_1) = \mathbb{E}[m_2(X_2) - m_1(X_1)]^2 = c \geq 0, \end{aligned}$$

where the third but last equality follows from the facts that $m_1(X_1)$ and $m_2(X_2)$ are conditional expectations of Y w.r.t. X_1 and X_2 . The second but last equality utilizes that X_1 is a subvector of X_2 . In most situations c will be strictly positive so that (1.1) holds true and no extra test is necessary. In such a situation one may be interested to know whether the inclusion of more covariables would increase the predictive accuracy by an amount of at least c_0 . In other words we want to test

$$H_0 : \mathbb{E}g(\epsilon_1) = \mathbb{E}g(\epsilon_2) + c_0$$

versus

$$H_1 : \mathbb{E}g(\epsilon_1) > \mathbb{E}g(\epsilon_2) + c_0.$$

Our approach also applies here. Just replace \bar{d} by $\bar{d} - c_0$.

Of course there may be situations where the augmented X_1 will not improve the predictive accuracy at all. Consider, e.g., the two linear models

$$Y = X_1' \beta + \epsilon_1 \quad \text{and} \quad Y = X_1' \gamma + U_2' \delta + \epsilon_2. \quad (1.3)$$

Hence $X_2 = (X_1', U_2)'$. In this case two situations are possible. If $\delta = 0$, then $\epsilon_1 = \epsilon_2$ whence $c = 0$. In other words, the augmentation of X_1 has no effect on the predictive accuracy. In general, we have

$$c = \mathbb{E}[X_1'(\gamma - \beta) + U_2' \delta]^2.$$

If, e.g., X_1 and U_2 are centered and uncorrelated, then

$$c = \mathbb{E}[X_1'(\gamma - \beta)]^2 + \mathbb{E}[U_2' \delta]^2$$

so that the overall difference may be attributed to the variability contained in U_2 weighted by δ and a (possibly reduced) variability contained in X_1 .

Remark 2. Since the test statistic \bar{d} depends on estimated residuals, a crucial role in our setup will be played by \hat{m}_1 and \hat{m}_2 . For the replacement of the e 's by the ϵ 's the following error bounds on $\hat{m}_i - m_i$ will be needed:

$$\mathbb{E} \left\{ [\hat{m}_i(X_{i,T+1}) - m_i(X_{i,T+1})]^2 \right\} = O(T^{-\beta_i}) \quad (1.4)$$

$i = 1, 2$. Generally the constant β_i depends on whether we are in a parametric or nonparametric framework. In the parametric case we typically can estimate unknown parameters at the rate $T^{-1/2}$ so that under smoothness of the model the bound (1.4) holds with $\beta_i = 1$. In the nonparametric framework the quality of the estimators deteriorates as the dimension

of X_1 and X_2 gets large, a consequence of the so-called curse of dimensionality. E.g., for the Nadaraya-Watson estimator, Spiegelman and Sacks (1980) showed that (1.4) holds true with

$$\beta_i = \frac{2}{2 + d_i},$$

under surprisingly weak regularity assumptions on the m_i 's. Here again d_i is the dimension of $X_i, i = 1, 2$.

Remark 3. Since (1.2) is concerned with standardized variables, the estimation error encountered in (1.4) has to become negligible compared with the sampling variances of the validation part. This may be achieved if the sample size of the learning sample, T , is large enough compared with n , the size of the validation sample.

Remark 4. Another issue is the choice of the weight function g . More or less this is up to the applicant of the statistical methodology. If one prefers to downweight large residuals, the absolute deviation function (or a robust variant) might be appropriate. Statistical inference is then based on the Mean Absolute Deviation (MAD), a popular means for goodness-of-fit in Robust Regression. If large deviations are to be upweighted large powers of e are in order. Note, however, that such a choice also requires higher moments for the errors ε . Therefore, in this paper, we shall focus only on g 's which increase at most as fast as $g(u) = u^2$. In particular, g' and g'' (if they exist) are assumed to be Lipschitz and bounded, respectively. While Taylor expansion is an appropriate tool in the smooth case, g 's with possible discontinuities need to be studied separately.

Remark 5. Prediction accuracy in regression has been often discussed in the context of model selection. See Efron (2004) for a discussion and review. For a given (nested) family of models one adds a penalty for the complexity of the model to the residual sum of squares. The resulting objective function is then minimized for a data set at hands. In our approach no penalty is considered and the regression need not be parametrically specified. Rather we study the unconditional distribution of the relevant quantity \bar{d} .

Remark 6. Concerning prediction, Granger and Newbold (1978) were pioneers in designing formal tests. Their procedure was based on the correlation of some combination of the residuals. Since then other authors extended their work into various directions. A simple but often applied test is due to Diebold and Mariano (1995). They compare two fully specified parameter-free models and compare the known errors. Typically, however, the assumed models are more complex to the effect, that the true errors are unknown and parameters need to be estimated. This changes the distributional character of the DM-test and the distribution presented in their work is not applicable.

Hence it is not surprising that Clark and West (2004) found some bias in the DM-test when applied to residuals. In the context of classical goodness-of-fit tests based on the empirical distribution function, this is known at least since Durbin (1973). For the corresponding discussion in regression, see Stute (1997). To circumvent these problems one may, as in the present paper, split the whole sample of size $T + n$ into a learning sample of size T and a validation sample of size n . At the same time the regularity assumptions on m_1 and m_2 are weak since only (1.4) is required.

The rest of the paper is organized as follows. In Section 2 we present the main results. In Section 3 we report on a simulation study. Proofs are postponed to the Appendix.

2 MAIN RESULTS

As in the first section, let $(X_{1t}, X_{2t}, Y_t), 1 \leq t \leq T + n$, be a sample of independent random vectors with the same distribution as (X_1, X_2, Y) , and let g be a given weight function. Recall \bar{d} and \bar{d}_1 .

Theorem 1. *Assume that g is differentiable such that g' is Lipschitz of order one, and let \hat{m}_1 and \hat{m}_2 be such that (1.4) is satisfied for $i = 1, 2$. Set $\beta = \min(\beta_1, \beta_2)$. Then, if $n = o(T^\beta)$, as n and T tend to infinity,*

$$n^{1/2}\bar{d} = n^{1/2}\bar{d}_1 + o_{\mathbb{P}}(1). \quad (2.1)$$

Furthermore, if g is twice continuously differentiable with

$$\mathbb{E}[g'(\varepsilon_i)|X_i] = 0 \text{ a.s. for } i = 1, 2 \quad (2.2)$$

then the assertion (2.1) holds true under the weaker condition $n = o(T^{2\beta})$.

Note that (2.2) is satisfied for $g(u) = u^2$.

While Theorem 1 covers the case of a differentiable g , in the next Theorem we consider the important special cases $g_1(u) = |u|$ and $g_2(u) = u^2 1_{\{|u| > \delta\}}$.

Theorem 2. *The expansion (2.1) holds true*

- for g_1 whenever $n = o(T^\beta)$
- for g_2 whenever $n \ln n = o(T^{\beta/2})$ and

the distribution function of $|\varepsilon|$ is differentiable at δ .

A careful check of the proof for g_2 shows that one may extend Theorem 1 to functions g with finitely many jumps but which are twice continuously differentiable in between. Details are omitted.

Our conditions on n and T show that $g(u) = u^2$ requires the smallest T . This is due to the fact that with this g the orthogonality of ε_i and $m_i(X_i)$ can be effectively used. On the other hand, discontinuities of g require a larger T and some smoothness of the distribution function of $|\varepsilon|$ in order to cope with the jump of g at δ .

Corollary 2.1. Under the conditions of Theorem 1 or 2, we have

$$\frac{n^{1/2}\bar{d}}{\hat{\sigma}_{\bar{d}}^2} \rightarrow \mathcal{N}(0, 1) \text{ in distribution.}$$

Here $\hat{\sigma}_{\bar{d}}^2$ is a consistent estimator of the variance of $g(\varepsilon_1) - g(\varepsilon_2)$.

For example, we could take for $\hat{\sigma}_{\bar{d}}^2$ the sample variance of the $g(e_{1t}) - g(e_{2t}), T + 1 \leq t \leq T + n$.

Remark 7. Our results may be extended to the dependent case, i.e., when the regression models allow for time series errors. In such a situation the asymptotic normality of \bar{d} and \bar{d}_1 may not be inferred from the classical CLT. Rather, one needs to apply proper versions of the CLT for dependent variables. E.g., we obtain Corollary 2.1 under the assumption that, under H_0 , the $g(\varepsilon_{1t}) - g(\varepsilon_{2t})$ form a martingale difference sequence. See Hall and Heyde (1980).

Instead of splitting the data into two parts, one could also study a \bar{d} based on the full sample with the t -th residuals computed from the cross-validated data. Following the proof of Theorem 1 one finds out that the terms d_2 and d_3 there are now nonnegligible and yield a more complicated limit variance, under further regularity assumptions on the model. If the function g is non-differentiable and/or the m 's are nonsmooth, things are becoming even less obvious. These are the main reasons why in this paper we stick to data splitting.

3 SIMULATIONS

In this section we report on several simulation results designed to support our theoretical findings. Only the weight function $g(u) = u^2$ will be considered. Also only parametric models will be studied so that in each case $\beta = 1$. In each scenario empirical frequencies were obtained from 1000 runs.

By Theorem 1 the sample sizes n and T need to satisfy $n = o(T^2)$. In the tables to follow we demonstrate among other things that such a condition is indispensable in that the

level of the tests, α , is attained only when n is small compared with T^2 . From a technical point of view, this is not surprising since, as our proofs will show, in comparing \bar{d} with \bar{d}_1 , the error terms are not necessarily negligible if $n = o(T^2)$ is violated.

Example 3.1. The data are coming from the model $Y_t = X_{1t} + X_{2t} + \varepsilon_t$, where the $(X_{1t}, X_{2t}, \varepsilon_t)$ are cross- and serially independent. Moreover,

$$\mathbb{E}X_{1t} = \mathbb{E}X_{2t} = \mathbb{E}\varepsilon_t = 0$$

$$\text{Var}X_{1t} = 1 = \text{Var}\varepsilon_t \quad \text{Var}X_{2t} = \sigma^2.$$

Consequently

$$Y_t = X_{1t} + \varepsilon_{1t} \quad Y_t = X_{2t} + \varepsilon_{2t}$$

with

$$\mathbb{E}\varepsilon_{1t}^2 = \sigma^2 + 1 \quad \mathbb{E}\varepsilon_{2t}^2 = 2.$$

Hence the null hypothesis is true when $\sigma^2 = 1$ and (1.1) is satisfied if and only if $\sigma^2 > 1$. Significance levels were $\alpha = 0.05$ and $\alpha = 0.10$. We also report on the attained levels for \bar{d}_1 though in a practical situation \bar{d}_1 will not be available. In a simulation study, however, they are known and helpful to demonstrate the quality of the approximation of \bar{d} through \bar{d}_1 . For example, in Table 1, when $T = 50$ and $n = 2.500$ so that $n = T^2$, the attained level is 0.19 while for the true errors, i.e., for \bar{d}_1 , the attained level equals the true α . Residuals were computed by fitting the data to the marginal linear models

$$Y_t = a_1 + b_1X_{1t} + \varepsilon_{1t} \quad Y_t = a_2 + b_2X_{2t} + \varepsilon_{2t}.$$

Tables 2 and 3 deal with power. We see that as σ increases, so does power. For T and n we always have $T = n$ with n ranging from 20 to 200. When n is small compared with T^2 , the percentages are acceptable already for moderate n .

Example 3.2. In this example the variables were generated from

$$Y_t = X_{1t} + bX_{2t} + \varepsilon_t,$$

where X_{1t}, X_{2t} and ε_t are standard normal and independent. In this case we analyze the effect of adding the second covariate to the first. In other words, the second covariable equals (X_{1t}, X_{2t}) while the first is X_{1t} . There is no doubt that adding a new variable improves the fit, but there may be doubts that this would significantly improve the predictive accuracy.

Table 1: $T = 50, \varepsilon_{1t}, \varepsilon_{2t} \sim NI(0, 1)$ Empirical frequencies of rejection

Using residuals			Using errors	
n	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
10	0.06	0.12	0.07	0.13
30	0.04	0.09	0.05	0.10
100	0.06	0.11	0.06	0.11
200	0.07	0.13	0.05	0.10
1000	0.16	0.22	0.05	0.10
2500	0.19	0.22	0.05	0.10

Table 2: Empirical frequencies of rejection for the case $\sigma = 1.2$

Using residuals			Using errors	
$T = n$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
20	0.15	0.28	0.10	0.45
50	0.21	0.32	0.17	0.28
100	0.30	0.43	0.26	0.39
200	0.48	0.61	0.39	0.54

In Table 4 we present, for different scenarios, summary statistics \bar{d} and \bar{d}_1 . The number of runs in each case was again 1000. Both estimate the increase in predictive accuracy c from Remark 1, which in the present case equals b^2 . The residuals were computed after fitting a Linear Model. The results are promising already for small sample sizes. That the two values of \bar{d} for $b = 0.1, T = n = 20$ and 50 are slightly negative is due to sampling errors.

Table 3: Empirical frequencies of rejection for the case $\sigma = 1.5$

Using residuals			Using errors	
$T = n$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
20	0.34	0.47	0.27	0.45
50	0.57	0.72	0.57	0.71
100	0.83	0.92	0.83	0.91
200	0.98	0.99	0.98	0.99

Table 4: Summary statistics for \bar{d} and \bar{d}_1

\bar{d}				\bar{d}_1			
	$T = n$	$b = 0.1$	$b = 0.5$		$b = 1$	$b = 0.1$	$b = 0.5$
20		-.0588	.2128	1.0190	.0101	.2613	.9869
50		-.0084	.2381	1.0267	.0100	.2500	1.0075
100		.00063	.2464	.9969	.0103	.2506	.9856
200		.0053	.2469	1.0018	.0095	.2495	.9983

4 APPENDIX

Proof of Theorem 1. In the smooth case we have

$$\begin{aligned} \bar{d} &= n^{-1} \sum_{t=T+1}^{T+n} [g(\varepsilon_{1t}) - g(\varepsilon_{2t})] + n^{-1} \sum_{t=T+1}^{T+n} g'(v_{1t})(e_{1t} - \varepsilon_{1t}) \\ &\quad - n^{-1} \sum_{t=T+1}^{T+n} g'(v_{2t})(e_{2t} - \varepsilon_{2t}) \equiv \bar{d}_1 + d_2 - d_3, \end{aligned}$$

where v_{it} is between e_{it} and ε_{it} , $i = 1, 2$ and $t = T + 1, \dots, T + n$. Therefore it suffices to show that $n^{1/2}d_2$ and $n^{1/2}d_3$ converge to zero. Now, since g' is Lipschitz of order one, we

have

$$\begin{aligned} |d_2| &\leq n^{-1} \sum_{t=T+1}^{T+n} |g'(v_{1t})| |e_{1t} - \epsilon_{1t}| \leq n^{-1} \sum_{t=T+1}^{T+n} C |v_{1t}| |e_{1t} - \epsilon_{1t}| \\ &\leq \frac{C}{n} \sum_{t=T+1}^{T+n} [|\epsilon_{1t}| + |e_{1t} - \epsilon_{1t}|] |e_{1t} - \epsilon_{1t}| \\ &\leq \frac{C}{n} \sum_{t=T+1}^{T+n} |\epsilon_{1t}| |e_{1t} - \epsilon_{1t}| + \frac{C}{n} \sum_{t=T+1}^{T+n} |e_{1t} - \epsilon_{1t}|^2. \end{aligned}$$

Apply the Cauchy-Schwarz inequality to each term in the first sum and use the fact that the ϵ_{1t} come from the same distribution and hence have identical second moments to get that, by (1.4),

$$\mathbb{E}|d_2| = O(T^{-\beta/2}).$$

Since $n^{1/2}T^{-\beta/2} \rightarrow 0$, we obtain $n^{1/2}d_2 \rightarrow 0$ in probability. Similarly for d_3 .

When g is twice differentiable and

$$\mathbb{E}(g'(\epsilon_{1t})|X_{1t}) = 0 = \mathbb{E}(g'(\epsilon_{2t})|X_{2t}),$$

one has to decompose \bar{d} as

$$\begin{aligned} \bar{d} &= n^{-1} \sum_{t=T+1}^{T+n} (g(\epsilon_{1t}) - g(\epsilon_{2t})) + n^{-1} \sum_{t=T+1}^{T+n} g'(\epsilon_{1t})(e_{1t} - \epsilon_{1t}) \\ &\quad + \frac{1}{2n} \sum_{t=T+1}^{T+n} g''(v_{1t})(e_{1t} - \epsilon_{1t})^2 - n^{-1} \sum_{t=T+1}^{T+n} g'(\epsilon_{2t})(e_{2t} - \epsilon_{2t}) \\ &\quad - \frac{1}{2n} \sum_{t=T+1}^{T+n} g''(v_{2t})(e_{2t} - \epsilon_{2t})^2 \equiv \bar{d}_1 + d_2 + d_3 - d_4 - d_5. \end{aligned}$$

As to d_2 , note that

$$d_2 = n^{-1} \sum_{t=T+1}^{T+n} g'(\epsilon_{1t})(m_1(X_{1t}) - \hat{m}_1(X_{1t}))$$

and that each ϵ_{1t} is independent of the estimation sample and, at the same time, we have

$$\mathbb{E}[g'(\epsilon_{1t})|X_{1t}] = 0.$$

From this we obtain $\mathbb{E}d_2 = 0$. Moreover, conditioning on the estimation sample and on $X_{1,T+1}, \dots, X_{1,T+n}$ and applying (2.2) again we see that the summands in d_2 are uncorrelated. Hence

$$\text{Var } d_2 = n^{-1} \mathbb{E} \{ \mathbb{E} [g'^2(\epsilon_{1,T+1})|X_{1,T+1}] (m_1(X_{1,T+1}) - \hat{m}_1(X_{1,T+1}))^2 \}.$$

The consistency of the \hat{m}_1 implies that the expectation tends to zero. Hence $n^{1/2}d_2 \rightarrow 0$ in probability. Similarly for d_4 . The bounds for d_3 and d_5 follow the previous pattern. E.g.,

$$\begin{aligned}\mathbb{E}|d_3| &\leq \mathbb{E}\{ |g''(\mathbf{v}_{1,T+1})|(e_{1,T+1} - \varepsilon_{1,T+1})^2 \} \\ &= O(T^{-\beta}), \text{ whenever } g'' \text{ is bounded.}\end{aligned}$$

Summarizing, we have obtained

$$n^{1/2}\bar{d} = n^{1/2}\bar{d}_1 + o_{\mathbb{P}}(1)$$

under $n^{1/2}T^{-\beta} \rightarrow 0$. □

Proof of Theorem 2. We first deal with the weight function $g(u) = |u|$. From the triangle inequality we get

$$\begin{aligned}\bar{d} &= n^{-1} \sum_{t=T+1}^{T+n} [|e_{1t}| - |e_{2t}|] \\ &= n^{-1} \sum_{t=T+1}^{T+n} [|\varepsilon_{1t} + e_{1t} - \varepsilon_{1t}| - |\varepsilon_{2t} - e_{2t} - \varepsilon_{2t}|] \\ &\leq n^{-1} \sum_{t=T+1}^{T+n} [|\varepsilon_{1t}| + |e_{1t} - \varepsilon_{1t}| - |\varepsilon_{2t}| - |e_{2t} - \varepsilon_{2t}|] \\ &= n^{-1} \sum_{t=T+1}^{T+n} [|\varepsilon_{1t}| - |\varepsilon_{2t}|] + n^{-1} \sum_{t=T+1}^{T+n} [|e_{1t} - \varepsilon_{1t}| + |e_{2t} - \varepsilon_{2t}|].\end{aligned}$$

As in the proof of Theorem 1 we obtain that the expectation of the second sum is $O(T^{-\beta/2})$. Since similar arguments also yield a lower bound we have

$$n^{1/2}\bar{d} = n^{1/2}\bar{d}_1 + o_{\mathbb{P}}(1) \text{ under } n^{1/2}T^{-\beta/2} \rightarrow 0.$$

To deal with the weight function

$$g(u) = u^2 1_{\{|u|>\delta\}},$$

the arguments are more involved. Denote with F_1 and F_2 the distribution functions of ε_{1t} and ε_{2t} , respectively. Since, under the null hypothesis,

$$\int_{\{|u|>\delta\}} u^2 F_1(du) = \int_{\{|u|>\delta\}} u^2 F_2(du),$$

we may write

$$n^{1/2}\bar{d} = z_1(\delta) - z_2(\delta)$$

with

$$z_i(\delta) = n^{-1/2} \sum_{t=T+1}^{T+n} [e_{it}^2 1_{\{|e_{it}|>\delta\}} - \int_{\{|u|>\delta\}} u^2 F_i(du)]$$

for $i = 1, 2$. We expand each $z_i(\delta)$ separately. For ease of notation we delete the sample index. Also set

$$R(y) = n^{-1/2} \sum_{t=T+1}^{T+n} \left[\varepsilon_t^2 1_{\{|\varepsilon_t|>y\}} - \int_{\{|u|>y\}} u^2 F(du) \right].$$

Clearly, to show the assertion, it remains to prove

$$z(\delta) = R(\delta) + o_{\mathbb{P}}(1).$$

Now,

$$\begin{aligned} z(\delta) &= n^{-1/2} \sum_{t=T+1}^{T+n} \left[\varepsilon_t^2 1_{\{|\varepsilon_t|>\delta\}} - \int_{\{|u|>\delta\}} u^2 F(du) \right] \\ &+ \frac{2}{n^{1/2}} \sum_{t=T+1}^{T+n} (e_t - \varepsilon_t) \varepsilon_t 1_{\{|\varepsilon_t|>\delta\}} + n^{-1/2} \sum_{t=T+1}^{T+n} (e_t - \varepsilon_t)^2 1_{\{|\varepsilon_t|>\delta\}} \\ &\equiv R_1(\delta) + R_2 + R_3. \end{aligned}$$

As in the first part of the proof of Theorem 1 we get

$$\mathbb{E}|R_2| = O(n^{1/2}T^{-\beta/2})$$

and

$$\mathbb{E}|R_3| = O(n^{1/2}T^{-\beta}).$$

The handling of $R_1(\delta)$ is a little more tricky since the replacement of e_t by ε_t takes place in the term causing the discontinuity. For this, write

$$\begin{aligned} R_1(\delta) &= n^{-1/2} \sum_{t=T+1}^{T+n} \left[\varepsilon_t^2 1_{\{|\varepsilon_t|>\delta, |e_t - \varepsilon_t| < x\}} - \int_{\{|u|>\delta\}} u^2 F(du) \right] \\ &+ n^{-1/2} \sum_{t=T+1}^{T+n} \varepsilon_t^2 1_{\{|\varepsilon_t|>\delta, |e_t - \varepsilon_t| \geq x\}}, \end{aligned}$$

where $x = n^{-1/2}/\ln n$. From the Cauchy-Schwarz and the Chebychev inequalities the expectation of the second sum is less than or equal to

$$\begin{aligned} n^{1/2} [\mathbb{E}(\epsilon_{T+1}^4)]^{1/2} \mathbb{P}^{1/2}(|e_{T+1} - \epsilon_{T+1}| \geq x) \\ = O(n \ln n T^{-\beta/2}) = o(1). \end{aligned}$$

As to the first sum in $R_1(\delta)$, recall the definition of $R(y)$. Under a finite fourth-moment assumption on the ϵ 's, this process converges, as $n \rightarrow \infty$, to a centered Gaussian process in the Skorokhod-space $D[0, \infty]$, the limit having continuous sample paths whenever F is continuous. The convergence of the finite dimensional distributions follows from the multivariate CLT while tightness follows from adapting the arguments in Billingsley (1968), p. 106. As a conclusion we may infer that $R(y)$ is close to $R(\delta)$ with high probability for all large n when y tends to δ . Since

$$\begin{aligned} \mathbf{1}_{\{|\epsilon_t| > \delta + x\}} - \mathbf{1}_{\{|\epsilon_t - \epsilon_t| \geq x\}} &\leq \mathbf{1}_{\{|\epsilon_t| > \delta + x, |\epsilon_t - \epsilon_t| < x\}} \\ &\leq \mathbf{1}_{\{|\epsilon_t| > \delta, |\epsilon_t - \epsilon_t| < x\}} \leq \mathbf{1}_{\{|\epsilon_t| > \delta - x\}}, \end{aligned}$$

we obtain

$$R_1(\delta) \leq R(\delta - x) + o_{\mathbb{P}}(1) + n^{1/2} \left[\int_{\{|u| > \delta - x\}} u^2 F(du) - \int_{\{|u| > \delta\}} u^2 F(du) \right]$$

and

$$R_1(\delta) \geq R(\delta + x) + o_{\mathbb{P}}(1) + n^{1/2} \left[\int_{\{|u| > \delta + x\}} u^2 F(du) - \int_{\{|u| > \delta\}} u^2 F(du) \right].$$

Since $x \rightarrow 0$, we have by the aforementioned asymptotic continuity of R :

$$R(\delta + x) = R(\delta) + o_{\mathbb{P}}(1) = R(\delta - x).$$

If the distribution function of $|\epsilon|$ is differentiable in a neighborhood of δ , each of the two differences of the integrals is $O(x) = O(n^{-1/2}/\ln n)$. Conclude that

$$R_1(\delta) = R(\delta) + o_{\mathbb{P}}(1),$$

as desired. □

4.1 Acknowledgements

The first author was supported by the MEC grant SEJ2005-05549 and the BBVA grant 1/BBVA 00038.16421/2004.

REFERENCES

1. Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
2. Clark, T. and West, K. (2004). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. Working Paper 04-03, FRB of Kansas City, <http://ssm.com/abstract=557101>.
3. Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253-263.
4. Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics*, **1**, 279-290.
5. Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, **99**, 619-632.
6. Granger, C. and Newbold, P. (1977). *Forecasting economic time series*, Academic Press, Orlando, FL.
7. Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its applications*, Academic Press, New York.
8. Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Annals of Statistics*, **8**, 240-246.
9. Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, **25**, 613-641.
10. Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, **25**, 613-641.