# Inference About Predictive Ability When There Are Many Predictors

Stanislav Anatolyev

# Inference about predictive ability when there are many predictors

Stanislav Anatolyev*

New Economic School, Moscow

January 2007

### Abstract

We enhance the theory of asymptotic inference about predictive ability by considering the case when a set of variables used to construct predictions is sizable. To this end, we consider an alternative asymptotic framework where the number of predictors tends to infinity with the sample size, although more slowly. Depending on the situation the asymptotic normal distribution of an average prediction criterion either gains additional variance as in the few predictors case, or gains non-zero bias which has no analogs in the few predictors case. By properly modifying conventional test statistics it is possible to remove most size distortions when there are many predictors, and improve test sizes even when there are few of them.

---

*Address: Stanislav Anatolyev, New Economic School, Nakhimovsky Prospekt, 47, Moscow, 117418 Russia. E-mail: sanatoly@nes.ru.

# 1 Introduction

The theory of asymptotic inference about predictive ability (IPA) put forward in West (1996), West and McCracken (1998, 2002), and other works, has proved to be useful in asymptotically correct testing of various predictive qualities of models when forecasts are regression-based and hence flawed by parameter estimation noise. In a nutshell, given that the sizes of regression and prediction intervals asymptotically grow with the same rate, this noise may (or may not) inflate the asymptotic variance of a t-test statistic based on an average prediction criterion. The IPA theory, however, is developed for the case when the dimensionality of regression parameters is asymptotically fixed in theory and small in practice relative to the regression and prediction sample sizes. Sometimes, however, researchers use large sets of predictors, each accompanied by an own unknown parameter. In a recent survey, Stock and Watson (2006) mention macroeconomic studies that use 20, 30, 33, 43, 66, and even 135, 147, 215, or 447 predictors. As a result, the parameter estimation noise may be less than that innocuous, and may asymptotically have a larger impact on the asymptotic distribution of the statistic of interest than predicted by the IPA theory.

In this paper, we take a close look at the situation when many predictors are used, and determine how asymptotic distributions change and how test statistics have to be modified. We develop an alternative "moderately many predictors" asymptotic framework when the number of predictors is let go to infinity with the sample size, but more slowly than proportionately. Indeed, as noted in Stock and Watson (2006, section 2), equality of growth rates in the number of predictors and the sample size leads to forecasts being overwhelmed by estimation noise. Therefore, to keep testing meaningful, the growth rate of number of predictors must be set at least smaller than the growth rate of regression sample size, hence the qualifier "moderately many".[1] We find the bounds for the relative growth of number of predictors and regression and prediction sample sizes when the asymptotic distribution does not change at all, or, alternatively, when testing stays meaningful but the asymptotic distribution changes. In cases when the asymptotic

---

[1] When there are "very many" predictors, dimension-reduction tools (e.g., Galbraith and Zinde-Walsh, 2006) seem more appropriate.

distribution does change, we compute the changes and provide the ways to suitably modify the test statistic so that it has correct asymptotic size. The asymptotic framework we consider is similar to one used in the recent literature on estimation with many instruments or many moment conditions considered by various authors, e.g., Bekker (1994), Koenker and Machado (1999), Newey and Windmeijer (2005), Chao and Swanson (2005), Stock and Yogo (2005), and others.

To have a quick preview of asymptotic results, let us denote by $R$ the length of the regression sample, by $P$ the length of the prediction sample, by $m$ the number of predictors. Asymptotically, $R$, $P$ and $m$ all diverge to infinity, and the growth of $m$ is restricted by a condition such as $m^\kappa/R \to 0$ for $\kappa$ equaling 2 or 3, reflecting the "moderately many predictors" paradigm.[2] It turns out that the average prediction criterion exhibits qualitatively different asymptotic behavior depending on whether the expected derivative of the tested criterion with respect to parameters (the "expected score") is zero or non-zero. As a result, this factor determines the direction in which asymptotic distributions change and test statistics should be modified.

When the expected score is zero (which includes, in particular, an empirically interesting case of mean squared prediction errors from a regression; e.g., Diebold and Mariano, 1995), the original IPA theory predicts no change in the asymptotic distribution, and no correction of test statistics is required. In our asymptotic framework, with $Pm^2/R^2$ converging to a finite positive constant, the asymptotic normal distribution stays normal with the same variance, but gains a deterministic bias. This phenomenon has no analogs in the original IPA theory. It is easy to correct for the asymptotic bias by subtracting its empirical analog to have asymptotically correct inference. In practice, this bias correction removes an overwhelming portion of size distortions.

When the expected score is non-zero (which includes, in particular, the case of mean squared prediction errors when regressors are endogenous), the original IPA theory predicts no change in asymptotics in case $P/R \to 0$, and inflation of asymptotic variance in case $P$ and $R$ grow at the same rate; in the latter case estimation of the inflated variance

---

[2]Koenker (1988) estimates that the number of predictors used in the (cross-sectional) wage determination literature is related to the sample size approximately as $m \sim R^{1/4}$ which satisfies the requirement of "moderately many predictors".

3

constitutes the necessary adjustment. In our asymptotic framework, with $Pm/R$ converging to a finite positive constant, there is an analogous phenomenon that the asymptotic variance increases, although formally the asymptotics is different. Moreover, the variance adjustment may be done in the same way that the original IPA theory prescribes, provided that a researcher exercises care in balancing the values of $P$ and $R$ in face of large $m$. In a way, $R$ is rescaled by division by $m$, and the rescaled regression sample size $R/m$ is balanced against $P$.

We illustrate our proposed framework with an application to determination of housing starts in the US. We consider a variety of structural models with a sizable number of predictors on the one hand, and a simple autoregression with a small number of lags as predictors on the other hand. We test the hypotheses of forecast unbiasedness and perform comparison of models on the basis of out-of-sample prediction errors using both the original IPA theory and our many predictor modification. The results indicate that sometimes the two theories yield different inference outcomes.

In order to concentrate on effects caused by numerosity of predictors, we assume that the estimated model is linear and the hypothesis of interest contains one restriction, and consider a "fixed scheme" of obtaining parameter estimates when the regression is run once on the regression sample. Some remarks on notation used now follow. For any square matrix $B$, denote by $\lambda_i(B)$ its $i^{\text{th}}$ eigenvalue, and by $\underline{\lambda}(B)$ and $\bar{\lambda}(B)$ its minimal and maximal eigenvalues; $\left|\bar{\lambda}(B)\right|$ is understood as an absolute value of a maximal *in absolute value* eigenvalue of $B$. For any matrix $A$ define its minimal and maximal singular values as $\underline{\sigma}(A) = \sqrt{\underline{\lambda}(A'A)}$ and $\bar{\sigma}(A) = \sqrt{\bar{\lambda}(A'A)}$. Unless otherwise noted, we work with the notion of spectral matrix norm $\|A\| = \bar{\sigma}(A)$ induced by the Euclidean vector norm.[3] Next, $\text{tr}(A)$ denoted the rank of $A$, and $\iota_m$ is an $m$-vector of ones. Finally, $c$ and $C$ are generic constants that do not depend on $m$ and $R$. "MN" and "GV" are abbreviations for Magnus and Neudecker (1988) and Golub and Van Loan (1996), respectively.

The paper is structured as follows. In section 2 the setup is presented, and assumptions

---

[3] We use the spectral norm in place of the more widely used Frobenius norm $\|A\|_F = \sqrt{\text{tr}(A'A)}$ because it is more convenient to use in the context of asymptotically expanding matrices, and because it is more directly linked to eigenvalues.

are discussed. In section 3 we tackle the case of zero expected score. In section 4, we handle the case when some of elements of the expected score vector are not zeros. Section 5 contains an illustrative empirical application. We conclude in section 6.

## 2   Setup and asymptotic framework

Suppose we are interested in testing a null hypothesis about $E\left[f_t\right]$, where $f_t$ is some criterion of prediction quality, depending on prediction errors $u_t$, $u_{t+1}$, $\cdots$, $u_{t+\tau-1}$, where $\tau$ is the prediction horizon. Because the prediction errors are unobservable, they are estimated from a parametric model. Let the estimated model be linear:

$$y_t = x_t'\beta + u_t,$$

where $x_t$ and $\beta \in B \subseteq \mathbb{R}^m$ are $m \times 1$. The $u_t$ is an error which has mean zero conditional on $z_t$, the vector of instrumental variables (most often, $z_t = x_t$). For simplicity, we use the fixed scheme is generating predictions, i.e. $\beta$ is estimated once using the "regression sample" $t = 1, \cdots, R$.

Let the dimensionality of $z_t$ be $m$, which trivially holds when $x_t$ is exogenous and OLS is used. When there is endogeneity, equality of dimensionality of $x_t$ and $z_t$ is a restriction made for reducing algebra; the underlying motivating scenario is that for endogenous right side variables (of which there may be many) a researcher (desperately) searches for a minimally necessary number of extraneous instruments.[4] Hence, $\beta$ is estimated as

$$\hat{\beta} = \left(\sum_{t=1}^{R} z_t x_t'\right)^{-1} \sum_{t=1}^{R} z_t y_t.$$

This estimate is used in making estimates of prediction errors $\hat{u}_t = y_t - x_t'\hat{\beta}$ for the "prediction sample" $t = R+1, \cdots, R+P$, which are converted into values of a forecast criterion of interest $\hat{f}_t$, $t = R+1, \cdots, R+P$ (the total number of observations is larger than $R+P$ because $\tau > 0$ and $x_t$ and/or $z_t$ may include lags of $y_t$). These values are then

---

[4]The case of overidentification and 2SLS estimation does not provide new insights. The instrument $E\left[x_t z_t'\right] E\left[z_t z_t'\right]^{-1} z_t$ implied by 2SLS replaces the original overidentifying instrument $z_t$.

collected into the average criterion

$$\overline{\hat{f}} = \frac{1}{P} \sum_{t=R+1}^{R+P} \hat{f}_t.$$

This average is used to test the hypothesis about the value of $Ef$ by constructing a t test statistic from the difference $\overline{\hat{f}} - Ef$. Note that for simplicity we treat the case of a scalar criterion; vector criteria may well be allowed at the expense of more complicated proofs with no new insights.

We consider the asymptotic framework where both $P$ and $R$ tend to infinity, possibly with different rates, and simultaneously $m \to \infty$, with a smaller rate than $R \to \infty$, i.e. $m = o(R)$, which we name the "moderately many predictors" framework. The relative growth rates will be discussed later more precisely. In practical applications, it is hoped that the modified tests will be advantageous even when $m$ is rather small.

We make the following assumptions about properties of the data.

**Assumption 1** *For some $\nu > 1$,*

(i) *the sequence $\{(x_t, z_t, u_t)\}$ is strictly stationary and strongly mixing with mixing coefficients $\alpha_i$ satisfying $\sum_{i=1}^{\infty} i\alpha_i^{1-1/\nu} < C$,*

(ii) *$E\left[u_t^{8\nu}\right] < C$, $\max_{1 \leq i \leq m} E\left[z_{t,i}^{8\nu}\right] < C$, $\max_{1 \leq i \leq m} E\left[x_{t,i}^{4\nu}\right] < C$,*

(iii) *$E\left[u_t | z_t, u_{t-1}, z_{t-1}, u_{t-2}, \cdots\right] = 0$.*

While assumptions 1(i,ii) are pretty standard, the martingale difference structure imposed in assumption 1(iii) seems necessary in the framework with growing $m$.[5] Assumption 1(iii) concerns the serial correlation properties of the error term in the predictive regression, and not those of the prediction criterion $f_t$ which may well be serially correlated. Introduce the familiar quantities

$$\begin{aligned} Q_{zx} &= E\left[z_t x_t'\right], \\ V_{zu} &= var\left[z_t u_t\right], \end{aligned}$$

---

[5]We conjecture that if the prediction error is serially corelated of finite order, the results in the paper are valid after obvious corrections, particularly in the definition of $V_{zu}$ below.

assuming that these objects exist and are finite, and

$$\Sigma_\beta = Q_{zx}^{-1} V_{zu} Q_{zx}'^{-1}.$$

Note that we do not impose conditional homoskedasticity. In conditional homoskedasticity does take place, moment assumptions in 1(ii) may be relaxed.

**Assumption 2** *For some c and C,*

(i) $c < \underline{\sigma}(Q_{zx})$ *and* $\bar{\sigma}(Q_{zx}) < C,$

(ii) $\bar{\lambda}(V_{zu}) < C.$

The first condition of assumption 2(i) says that all incoming instruments are relevant for incoming predictors so that the inverse of the matrix of their cross-products is uniformly separated from zero.[6] The second condition precludes trends in predictors and/or instruments as $m$ grows; if instruments are the same as predictors, this is equivalent to the condition $\bar{\lambda}(E[x_t x_t']) < C$. Assumption 2(ii) imposes a similar restriction on the variance of $z_t u_t$.

**Assumption 3** *The criterion $f_t$ is a Borel measurable function of b, $x_t$ and $y_t$ for all $b \in B$ and continuously differentiable in b as many times as needed for all $b \in B$ and for all $x_t$ and $y_t$ in their support.*

Assumptions on moments of various derivatives of $f_t$ will be imposed later. Now define the "expected score"

$$Q_{\partial f} = E\left[\frac{\partial f_t}{\partial \beta}\right],$$

and also introduce

$$Q_{\partial^2 f} = E\left[\frac{\partial^2 f_t}{\partial \beta \partial \beta'}\right],$$

$$V_f = \lim_{P \to \infty} var\left[\frac{1}{\sqrt{P}} \sum_{t=R+1}^{R+P} f_t\right] = \sum_{k=-\infty}^{+\infty} cov\left[f_t, f_{t-k}\right],$$

---

[6]This precludes use of weak instruments in the sense of asymptotically zero correlation between some of instruments and some of predictors. However, this has nothing to do with the strength of predictors which is allowed to be weak in the sense that the $\mathbb{R}^2$ of the predictive regression may be small.

assuming that these objects exist and are finite. Note that the latter two are symmetric by construction.

Given the fixed scheme, the original IPA theory concludes that the additional error resulting from the estimation step asymptotically leads or does not lead to an increase of asymptotic variance:

$$\sqrt{P}(\overline{\hat{f}} - Ef) \xrightarrow{d} \mathcal{N}\left(0, V_f + \pi Q'_{\partial f}\Sigma_\beta Q_{\partial f}\right), \tag{1}$$

where $\pi = \lim_{P,R\to\infty} P/R$. The phenomenon of "asymptotic irrelevance" (West, 1996) occurs when the additional variance equals zero, which is possible either when $\pi = 0$, or $Q_{\partial f} = 0$. The former case means that the researcher sets the prediction interval to be a negligible part of the whole sample, although still to a large number (as formally $P \to \infty$). The latter condition is a property of the problem at hand. Below, when we take $m$ to grow asymptotically, whether this condition is satisfied or not will result in different asymptotic distributions both differing from $\mathcal{N}(0, V_f)$. In the original IPA framework, when there is no asymptotic irrelevance, a t-statistic for testing the null can be properly constructed as

$$t_0 = \frac{\sqrt{P}\left(\overline{\hat{f}} - Ef\right)}{\sqrt{\hat{V}_f + (P/R)\hat{Q}'_{\partial f}\hat{\Sigma}_\beta \hat{Q}_{\partial f}}},$$

where $\hat{Q}_{\partial f}$ and $\hat{\Sigma}_\beta$ are consistent (e.g., analog) estimators of $Q_{\partial f}$ and $\Sigma_\beta$. The t-statistic $t_0$ is asymptotically $\mathcal{N}(0, 1)$ under the null.

Within our "moderately many predictors" asymptotic framework, the asymptotics will be markedly different depending on whether $Q_{\partial f} = 0$ or $Q_{\partial f} \neq 0$.

# 3 Problem with zero expected score

In this section we consider such problems where the expected score is zero:

$$Q_{\partial f} = 0.$$

The leading example is the mean squared prediction error (MSPE) criterion $f_t = u_t^2$ when right side variables are exogenous. In this case $Q_{\partial f} = -2E\left[u_t x_t\right] = 0$. Arguably the most interesting application of this criterion is testing for equal forecasting accuracy

in the style of Diebold and Mariano (1995). The latter test will be considered at some length in empirical section 5. Using the mean quartic prediction error criterion $f_t = u_t^4$ with exogenous right side variables and conditionally symmetric errors so that $Q_{\partial f} = -4E\left[u_t^3 x_t\right] = 0$ also fits this framework.

Define the matrix

$$V_{\partial f} = \lim_{P \to \infty} var\left[\frac{1}{\sqrt{P}} \sum_{t=R+1}^{R+P} \frac{\partial f_t}{\partial \beta}\right] = \sum_{k=-\infty}^{+\infty} cov\left[\frac{\partial f_t}{\partial \beta}, \frac{\partial f_{t-k}}{\partial \beta}\right],$$

**Assumption 4** *In addition to $Q_{\partial f} = 0$, the derivatives of $f_t$ satisfy:*

(i) *$Q_{\partial^2 f}$ is such that $\|Q_{\partial^2 f}\| < C$,*

(ii) *for some stationary series $d_t$ with finite $E\left[d_t^2\right]$, $\|\partial^2 f_t^*/\partial\beta\partial\beta' - \partial^2 f_t/\partial\beta\partial\beta'\| \leq \sqrt{m} d_t$*
*$\times \|\beta^* - \beta\|$ for all $\beta^* \in B$,*

(iii) *$\left\|\hat{Q}_{\partial^2 f} - Q_{\partial^2 f}\right\|$ is $O_p\left(m/\sqrt{P}\right)$, where $\hat{Q}_{\partial^2 f} = P^{-1} \sum_{t=R+1}^{R+P} \partial^2 f_t/\partial\beta\partial\beta'$,*

(iv) *$\max_{1 \leq i \leq m} E\left[(\partial f_t/\partial\beta_i)^{2\nu}\right] < C$ for $\nu$ of assumption 1,*

(v) *$\bar{\lambda}(V_{\partial f}) < C$.*

The importance of assumption 4(i) will be discussed shortly. This requirement usually (as in the examples above) reduces to an analogous condition placed on the variance of $x_t$, $x_t u_t$, or the like, and essentially restricts predictors to be uniformly bounded in variance. The conditions in assumptions 4(ii, iii, iv, v) are technical.

Define

$$\psi_1 = \lim_{m \to \infty} \frac{\text{tr}\left(Q_{\partial^2 f}\Sigma_\beta\right)}{m},$$

assuming that the limit exists. Note that

$$\frac{\left|\text{tr}\left(Q_{\partial^2 f}\Sigma_\beta\right)\right|}{m} = \frac{\left|\sum_{i=1}^m \lambda_i\left(Q_{\partial^2 f}\Sigma_\beta\right)\right|}{m}$$
$$\leq \left|\bar{\lambda}\left(Q_{\partial^2 f}\Sigma_\beta\right)\right|$$
$$\leq \left|\bar{\lambda}\left(Q_{\partial^2 f}\right)\right| \bar{\lambda}\left(\Sigma_\beta\right) < \infty,$$

because the trace of a square matrix equals a sum of its eigenvalues (MN, thm.17, p.19), and by assumption 4(i) and Lemma 2(c). Therefore, the expression under the limit sign in $\psi_1$ is uniformly bounded.

9

We start from the following important observation that sets the relative rate of divergence of $P$, $R$ and $m$. This result follows from the proof of Theorem 1 below.

**Proposition 1** *Suppose that $P \to \infty$, $R \to \infty$, $m \to \infty$, and $\psi_1 \neq 0$. (a) If $Pm^2/R^2 \to \infty$, then in $\sqrt{P}(\bar{\hat{f}} - Ef)$ the estimation error noise asymptotically dominates the signal from $\sqrt{P}(\bar{f} - Ef)$. (b) If $Pm^2/R^2 \to 0$, then $\sqrt{P}(\bar{\hat{f}} - Ef)$ has the same distribution as if there is no estimation error noise.*

Thus, for an emerging test to be asymptotically meaningful and non-trivial and the asymptotic results to provide a better approximation, $Pm^2/R^2$ has to converge to a non-zero constant. Hence, we make the following assumption.

**Assumption 5** *As $P \to \infty$, $R \to \infty$ and $m \to \infty$, we have*

$$\frac{Pm^2}{R^2} \to \mu_1 > 0$$

*and*

$$\frac{m^2}{R} \to 0.$$

Because of the second condition in Assumption 5, $P$ has to grow faster than $R$, which means that the prediction sample should be significantly bigger than the regression sample, and the number of predictors, albeit large, should be small relative to the these sizes. For example, we may have $P \propto R^{1+\delta}$ and $m^2 \propto R^{1-\delta}$ for some $0 < \delta < 1$ so that asymptotically $Pm^2/R^2 = \text{const} \neq 0$.

**Theorem 1** *Under the asymptotics of assumption 5 and conditions of assumptions 2, 3 and 4,*

$$\sqrt{P}(\bar{\hat{f}} - Ef) \xrightarrow{d} \mathcal{N}\left(\frac{\sqrt{\mu_1}}{2}\psi_1, V_f\right)$$

As follows from this result, the presence of many predictors induces bias provided that $\psi_1 \neq 0$. This asymptotic bias appears from the *second order* term in the Taylor expansion of the average prediction criterion around the true value of the parameter (recall that the

original IPA theory utilizes the first order Taylor expansion), while the first order term is asymptotically negligible:

$$\underbrace{\bar{f} - Ef}_{\substack{\text{relevant} \\ \text{uncertainty} \\ \text{(contributes } V_f)}} + \underbrace{\frac{1}{P} \sum_{t=R}^{R+P} \frac{\partial f_t}{\partial \beta'} \left(\hat{\beta} - \beta\right)}_{\substack{\text{estimation noise} \\ \text{negligible term} \\ \text{(contributes nothing)}}} + \underbrace{\frac{1}{2} \left(\hat{\beta} - \beta\right)' \frac{1}{P} \sum_{t=R}^{R+P} \frac{\partial^2 f_t}{\partial \beta \partial \beta'} \left(\hat{\beta} - \beta\right)}_{\substack{\text{estimation noise} \\ \text{leading term} \\ \text{(contributes } \sqrt{\mu_1}\psi_1/2)}}$$

In the original IPA theory, the second order term would yield a second-order bias which is bound to be a higher order asymptotic phenomenon. In our asymptotic framework, thanks to multiplicity of predictors, this term is of the same order as the zeroth order, "relevant noise", term.

Note that the asymptotic bias is necessarily positive or negative if $f_t$ is convex or concave in parameters[7] (as in the case of MSPE), and zero if $f_t$ is linear in parameters. Technically, this is because (using a Choleski decomposition of the positive definite $\Sigma_\beta$)

$$\operatorname{tr}\left(Q_{\partial^2 f} \Sigma_\beta\right) = \operatorname{tr}\left(Q_{\partial^2 f} \Lambda \Lambda'\right) = \operatorname{tr}\left(\Lambda' Q_{\partial^2 f} \Lambda\right) = \sum_{i=1}^{m} e_i' \Lambda' Q_{\partial^2 f} \Lambda e_i \gtrless 0.$$

**Remark**. Note that $\psi_1$, and thus the asymptotic bias, may be zero even when $Q_{\partial^2 f} \neq 0$ (of course, it is zero when $Q_{\partial^2 f} = 0$). This may happen if the rank of $Q_{\partial^2 f}$ grows slowlier than $m$, so that

$$\begin{aligned}
|\psi_1| &= \lim_{m\to\infty} \frac{\left|\sum_{i=1}^{m} \lambda_i \left(Q_{\partial^2 f} \Sigma_\beta\right)\right|}{m} \\
&\leq \left|\bar{\lambda}\left(Q_{\partial^2 f} \Sigma_\beta\right)\right| \lim_{m\to\infty} \frac{\operatorname{rk}\left(Q_{\partial^2 f} \Sigma_\beta\right)}{m} \\
&\leq \left|\bar{\lambda}\left(Q_{\partial^2 f}\right)\right| \bar{\lambda}\left(\Sigma_\beta\right) \lim_{m\to\infty} \frac{\operatorname{rk}\left(Q_{\partial^2 f}\right)}{m} \\
&= 0,
\end{aligned}$$

where it is used that the number of non-zero eigenvalues does not exceed the rank (MN, thm.18, p.19), and that $\operatorname{rk}(AB) = \operatorname{rk}(A)$ if $B$ is square of full rank (MN, eqn.5, p.8).

---

[7]This is often the case because the criterion $f_t$ is used not only for prediction evaluation, but also for parameter estimation in the same problem, and is concave or convex for the latter reason.

**Remark**. Note that it is important to have eigenvalues of $Q_{\partial^2 f}$ uniformly bounded from above. If this was not the case, contrary to assumption 4(i), the additional bias might be stochastic. For instance, if $Q_{\partial^2 f} = \iota_m \iota_m'$ with $\bar{\lambda}(Q_{\partial^2 f}) = m$, we will have asymptotically the stochastic bias having the same mean:

$$\chi^2_{(1)} \frac{\sqrt{\mu_1}}{2} \lim_{m \to \infty} \frac{\iota_m' \Sigma_\beta \iota_m}{m}$$

(cf. Lemma 8(a)). This scenario is, however, unrealistic; in examples at the beginning of this section $Q_{\partial^2 f}$ is proportional to the (positive definite) mean squared error matrix of some variable like $x_t$ or $x_t u_t$, in which case assumption 4(i) amounts to the requirement that the latter matrix have uniformly bounded eigenvalues (which may be already guaranteed by assumption 2 if $z_t = x_t$).

The asymptotic result of Theorem 1 suggests that the test statistic can be constructed by proper recentering and scaling the usual criterion. Let $\hat{V}_f$ and $\hat{\psi}_1$ be usual analog estimators for $V_f$ and $\hat{\psi}_1$, then one may use the t statistic

$$t_1 = \frac{\sqrt{P}\left(\bar{\hat{f}} - Ef - \frac{m}{2R}\hat{\psi}_1\right)}{\sqrt{\hat{V}_f}}.$$

**Theorem 2** *Under the asymptotics of assumption 5,*

$$t_1 \xrightarrow{d} \mathcal{N}(0,1).$$

Note that if $\hat{\psi}_1$ is constructed using sample analogs $\hat{Q}_{\partial^2 f}$ and $\hat{\Sigma}_\beta$ of $Q_{\partial^2 f}$ and $\Sigma_\beta$, the numerator of $t_1$ can be rewritten as $\sqrt{P}\left(\bar{\tilde{f}} - Ef\right)$, where $\tilde{f} = \hat{f} - \text{tr}\left(\hat{Q}_{\partial^2 f}\hat{\Sigma}_\beta\right)/2R$ is the original estimated criterion adjusted for the bias caused by estimation of a large number of parameters.

Next we look at an example where we can observe actual distributions, and rejection frequencies in particular, of original and modified test statistics. In this and subsequent examples we set some of unknown population quantities at their actual values rather than estimate from a sample, in order to isolate size distortions associated with a large number of predictors. Thus, in a full sense these are not Monte–Carlo experiments. The computations are performed from 100,000 simulations.

12

**Example.** Consider the MSPE criterion $f_t = u_t^2$ from the linear regression $y_t = x_t'\beta + u_t$ with conditionally homoskedastic normal disturbance having variance $\sigma_u^2$, so $z_t = x_t$. Let for simplicity $x_t \sim \mathcal{N}(0, I_m)$. Then $Q_{\partial f} = 0$, $V_f = 2\sigma_u^4$, $Q_{\partial^2 f} = 2I_m$, $\Sigma_\beta = \sigma_u^2 I_m$, $\psi_1 = 2\sigma_u^2$, so

$$\overline{\hat{u}^2} - 1 \overset{A}{\sim} \mathcal{N}\left(\frac{m}{R}\sigma_u^2, \frac{2\sigma_u^4}{P}\right),$$

and

$$t_0 = \frac{\sqrt{P}\left(\overline{\hat{u}^2} - 1\right)}{\sqrt{2}\sigma_u^2},$$

$$t_1 = \frac{\sqrt{P}\left(\overline{\hat{u}^2} - 1 - \frac{m}{R}\sigma_u^2\right)}{\sqrt{2}\sigma_u^2}.$$

The additional bias term induced by multiplicity of predictors is able to seriously distort inference. For example, when the square of the t-statistic is used for two-sided testing, neglecting multiplicity of predictors by using standard normal instead of the biased asymptotic distribution leads to a shift in the concentration point from the value 1 by $\Delta = Pm^2/(2R^2) \to \frac{1}{2}\mu_1$.

We set $\beta = \varrho\iota_m$, where $\varrho$ is set so that $R^2 = 50\%$ (identical figures result for other values of $R^2$. This indicates that the theory does not depend of whether the predictors are weak or strong). The following table presents actual rejection rates based on the nominal rate of 5%. We in addition report $\mu_1$, $m^2/R$, and the concentration point shift $\Delta$. The accompanying graph presents the distributions of $t_0$ and $t_1$ for the case $R = P = 200$ and $m = 16$. The positive bias in $t_0$ is apparent from the figure, as well as its absence in $t_1$ and approximately normal shapes of the distributions.

| $m$ | $\mu_1$ | $m^2/R$ | $\Delta$ | $t_0^{\neq}$ | $t_1^{\neq}$ | $t_0^{<}$ | $t_1^{<}$ | $t_0^{>}$ | $t_1^{>}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $R = P = 200$ | | | | | |
| 2 | 0.02 | 0.02 | 0.01 | 5.45 | 5.29 | 3.67 | 4.66 | 6.97 | 5.81 |
| 4 | 0.08 | 0.08 | 0.04 | 6.35 | 5.70 | 2.97 | 4.82 | 8.71 | 6.22 |
| 8 | 0.32 | 0.32 | 0.16 | 8.74 | 6.55 | 1.99 | 5.19 | 12.59 | 6.89 |
| 16 | 1.28 | 1.28 | 0.64 | 16.96 | 8.36 | 0.80 | 5.84 | 24.03 | 8.72 |
| 32 | 5.12 | 5.12 | 2.56 | 46.80 | 13.83 | 0.11 | 5.70 | 56.47 | 15.33 |
| | | | | $R = 300, P = 100$ | | | | | |
| 8 | 0.07 | 0.21 | 0.04 | 6.18 | 5.59 | 2.75 | 4.56 | 8.66 | 6.44 |
| 16 | 0.28 | 0.85 | 0.14 | 8.46 | 6.50 | 1.86 | 5.00 | 12.32 | 7.05 |
| 32 | 1.14 | 3.41 | 0.57 | 16.46 | 8.40 | 0.74 | 5.44 | 23.16 | 9.25 |
| | | | | $R = 100, P = 300$ | | | | | |
| 4 | 0.48 | 0.16 | 0.24 | 10.63 | 7.49 | 1.86 | 5.63 | 15.13 | 7.65 |
| 8 | 1.92 | 0.64 | 0.96 | 22.82 | 10.79 | 0.70 | 6.79 | 30.52 | 10.60 |
| 16 | 7.68 | 2.56 | 3.84 | 57.84 | 19.18 | 0.08 | 7.46 | 66.31 | 19.35 |
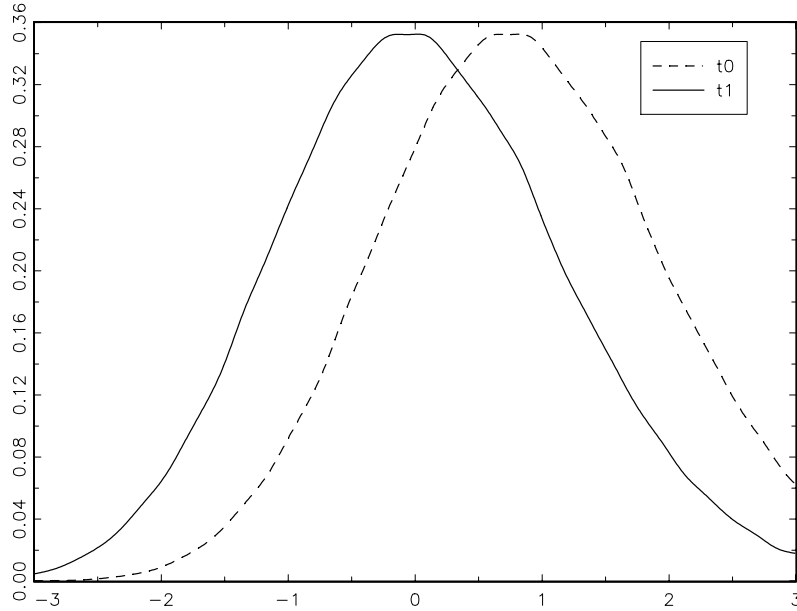


Figure 1: Densities of unadjusted and adjusted t-statistics

From the first panel of the table one can see that the unadjusted test statistic exhibits

large size distortions, especially for one-sided tests,[8] most of which can be removed by removing the asymptotic bias. The asymptotic theory gives good approximations when $m^2/R$ is small, say for $m = 16$ and smaller when $P = R = 200$, and worse approximations when $m^2/R$ is big as in the case $m = 16$ when $R = 100$ and $P = 300$, although it is this case when adjusting for the bias gives the maximal improvement in actual rejection frequencies as $\Delta$ is largest. The other two panels of the table present some evidence when $P$ differs from $R$ significantly. Some tendencies are similar, while the quality of approximations is better when $R > P$ than when $R < P$, which may seem in contradiction with the asymptotic presumption that $P$ grows faster than $R$. The explanation is that one should in fact compare lines with comparable levels of $\mu_1$ rather than with the same values of $m$. For example, in the last line $\mu_1 > 7$ which is of a similar magnitude as $m = 16$, although asymptotically $m$ tends to infinity while $\mu_1$ stays fixed. Again, in this case adjusting for the bias gives great improvement in actual rejection frequencies because $\Delta$ is very large. Note also that often rejection frequencies improve as a result of bias adjustment even in case $m$ is small, like 2 and 4, when no researcher thinks about such asymptotic effects.

To summarize, even though the bias-corrected $t_1$ still exhibits some overrejection, size distortions are much smaller than those displayed by $t_0$. For the degree of overrejection, the values of $P/R$ and $m^2/R$ are not as critical as values of $\mu_1$ which should not be large (larger than 2, say) for this degree to be moderate. However, the value $P/R$ seems to be most critical for size improvement: the improvement is more impressive the bigger $P/R$ is, other things equal.

## 4    Problem with non-zero expected score

Now we consider a problem that has non-zero "expected score"

$$Q_{\partial f} \neq 0.$$

The leading example is the mean prediction error (MPE) criterion $f_t = u_t$ when (some of) regressors are not centered, in which case $Q_{\partial f} = -E\left[x_t\right]$. This choice corresponds to

---

[8]In their simulations, West (1996) and West and McCracken (1998, 2002) study the behavior of only squared t-statistics, and therefore, only two-sided tests.

the empirically interesting test of forecast unbiasedness, which will be considered at some length in empirical section 5. Other examples are the MSPE criterion $f_t = u_t^2$ when $x_t$ is endogenous in which case $Q_{\partial f} = -2E\left[u_t x_t\right]$, and mean linear-exponential prediction error criterion $f_t = \exp\left(\alpha u_t\right) - \alpha u_t - 1$, $\alpha \neq 0$, in which case $Q_{\partial f} = -\alpha E\left[\left(\exp\left(\alpha u_t\right) - 1\right) x_t\right]$.

Let us denote by $\mathring{m}$ the number of non-zero elements in $Q_{\partial f}$. In the MPE example above, $m$ is the dimension of $x_t$, while $\mathring{m}$ is the number of non-centered variables in $x_t$. In the MSPE example, $m$ is the dimension of $x_t$, while $\mathring{m}$ is the number of endogenous variables in $x_t$. Of course, $1 \leq \mathring{m} \leq m$. We are most interested in the case where $\mathring{m} \to \infty$ asymptotically, because the case of fixed $\mathring{m}$ is very close to the original IPA theory even when $m \to \infty$.

**Assumption 6** *The derivatives of $f_t$ satisfy:*

(i) $Q_{\partial f}$ *is such that* $c\sqrt{\mathring{m}} < \|Q_{\partial f}\| < C\sqrt{\mathring{m}}$,

(ii) *for some stationary series $d_t$ with finite $E\left[d_t^2\right]$, $\|\partial f_t^*/\partial\beta - \partial f_t/\partial\beta\| \leq \sqrt{m} d_t \|\beta^* - \beta\|$ for all $\beta^* \in B$,*

(iii) $\left\|\hat{Q}_{\partial f} - Q_{\partial f}\right\|$ *is* $O_p\left(\mathring{m}/\sqrt{P}\right)$, *where* $\hat{Q}_{\partial f} = P^{-1}\sum_{t=R+1}^{R+P}\partial f_t/\partial\beta$.

Assumption 6(i) reflects the previous discussion that when $Q_{\partial f} \neq 0$, its "volume" (i.e. the effective number of non-zero elements) asymptotically grows at rate $\sqrt{m}$ (at which, for instance, $\|\iota_m\|$ grows). The conditions in assumptions 6(ii, iii) are technical.

Define
$$\psi_2 = \lim_{\mathring{m}\to\infty} \frac{Q_{\partial f}'\Sigma_\beta Q_{\partial f}}{\mathring{m}},$$
assuming that the limit exists. Note that by construction $\psi_2 \geq 0$, and that
$$\left|\frac{Q_{\partial f}'\Sigma_\beta Q_{\partial f}}{\mathring{m}}\right| \leq \bar{\lambda}\left(\Sigma_\beta\right)\frac{\|Q_{\partial f}\|^2}{\mathring{m}} < \infty,$$
so the expression under the limit sign in $\psi_2$ is uniformly bounded.

**Remark**. Most likely, $\psi_2$ is positive because of scaling by $\mathring{m}$ instead of $m$, but still may be zero because of a special structure of $\Sigma_\beta$. Consider the case when $z_t = x_t \sim \mathcal{N}\left(\iota_m, I_m\right)$ and $u_t$ is independent of $x_t$, then $E[x_t x_t'] = I_m + \iota_m\iota_m'$, $\Sigma_\beta \propto E[x_t x_t']^{-1} = I_m - \iota_m\iota_m'/\left(m+1\right)$,

and if $Q_{\partial f} \propto E[x_t] = \iota_m$, then $\mathring{m} = m$ and $Q'_{\partial f}\Sigma_\beta Q_{\partial f} \propto \iota'_m (I_m - \iota_m\iota'_m/(m+1)) \iota_m = m/(m+1)$, so $\psi_2 = 0$. Here $\bar{\lambda}(\Sigma_\beta) = 1$ and $\|Q_{\partial f}\| = \sqrt{m}$, but the structure of $\Sigma_\beta$ makes $\psi_2$ zero. Interestingly, this phenomenon is relevant to the case of MPE and corresponingly testing for forecast unbiasedness; see also section 5.

We start from the following important observation that sets the relative rate of divergence of $P$, $R$ and $m$. This result follows from the proof of Theorem 3 below.

**Proposition 2** *Suppose that $P \to \infty$, $R \to \infty$, $\mathring{m} \to \infty$, and $\psi_2 \neq 0$. (a) If $P\mathring{m}/R \to \infty$, then in $\sqrt{P}(\bar{\mathring{f}} - Ef)$ the estimation error noise asymptotically dominates the signal from $\sqrt{P}(\bar{f} - Ef)$. (b) If $P\mathring{m}/R \to 0$, then $\sqrt{P}(\bar{\mathring{f}} - Ef)$ has the same distribution as if there is no estimation error noise.*

Thus, for an emerging test to be asymptotically meaningful and non-trivial and the asymptotic results to provide a better approximation, $P\mathring{m}/R$ has to converge to a non-zero constant. Intuitively, to have a balance between the uncertainty in mean criterion and the estimation noise, the regression sample must be much larger than the prediction sample. The balance is achieved when the following assumption holds.

**Assumption 7** *As $P \to \infty$, $R \to \infty$ and $\mathring{m} \to \infty$, we have*

$$\frac{P\mathring{m}}{R} \to \mu_2 > 0$$

*and*

$$\frac{m^3}{R} \to 0.$$

Because of the first condition in Assumption 7, $R$ has to grow faster than $P$ (note that the relation between relative growth rates of $P$ and $R$ is opposite to the case of zero expected score), i.e. the regression sample should significantly exceed the prediction sample compensating for the numerosity of parameters to be estimated, and the number of predictors, albeit large, should be quite small relative to these sizes. For example, we may have $P \propto R^{1-\delta}$ and $\mathring{m} = m \propto R^\delta$ for some $0 \le \delta < \frac{1}{3}$ so that eventually $P\mathring{m}/R = \text{const}$, the case $\delta = 0$ representing the original IPA.

As in West (1996) and West and McCracken (2002), the estimation error noise inflates the asymptotic variance.

**Theorem 3** *Under the asymptotics of assumption 7*

$$\sqrt{P}(\bar{\hat{f}} - Ef) \xrightarrow{d} \mathcal{N}\left(0, V_f + \mu_2 \psi_2\right).$$

The additional asymptotic bias appears from the *first* term in the Taylor expansion of the average prediction criterion around the true value of the parameter, as in the original IPA theory:

$$\underbrace{\bar{f} - Ef}_{\substack{\text{relevant} \\ \text{uncertainty} \\ \text{(contributes } V_f)}} + \underbrace{\frac{1}{P}\sum_{t=R}^{R+P}\frac{\partial f_t}{\partial \beta'}\left(\hat{\beta} - \beta\right)}_{\substack{\text{estimation noise} \\ \text{leading term} \\ \text{(contributes } \mu_2\psi_2)}} + \underbrace{\frac{1}{2}\left(\hat{\beta} - \beta\right)'\frac{1}{P}\sum_{t=R}^{R+P}\frac{\partial^2 f_t}{\partial \beta \partial \beta'}\left(\hat{\beta} - \beta\right)}_{\substack{\text{estimation noise} \\ \text{negligible term} \\ \text{(contributes nothing)}}}$$

Let now $\hat{V}_f$ and $\psi_2$ be usual analog estimators for $V_f$ and $\psi_2$, then one may use the t statistic

$$t_1 = \frac{\sqrt{P}(\bar{\hat{f}} - Ef)}{\sqrt{\hat{V}_f + (P\mathring{m}/R)\psi_2}}.$$

**Theorem 4** *Under the asymptotics of assumption 7,*

$$t_1 \xrightarrow{d} \mathcal{N}\left(0, 1\right).$$

Suppose $\hat{Q}_{\partial f}$ and $\hat{\Sigma}_\beta$ are constructed as consistent sample analogs of $Q_{\partial f}$ and $\Sigma_\beta$ From a practical perspective, one can construct a t statistic

$$t_1 = \frac{\sqrt{P}(\bar{\hat{f}} - Ef)}{\sqrt{\hat{V}_f + (P/R)\hat{Q}'_{\partial f}\hat{\Sigma}_\beta\hat{Q}_{\partial f}}},$$

as $\mathring{m}$ cancels out. Note now that $t_1$ is exactly the t-statistic that would be constructed by a researcher following West (1996) and West and McCracken (2002) relying on the asymptotics with $\pi = \lim_{P,R\to\infty}(P/R) \neq 0$ (cf. (1)).

We can conclude that our adjustment is equivalent to that suggested by original IPA theory. Thus, a practitioner is welcome to use the original IPA theory when $Q_{\partial f} \neq 0$ despite the multiplicity of predictors, provided that care is exercised in that (a) $P\mathring{m}/R$

should not be allowed to be too large, and (b) even though $P/R$ may be tiny, variance adjustment still should be performed as $P\mathring{m}/R$ may be large.

The equivalence of modifications prescribed by the original IPA theory and by our asymptotic framework in case $Q_{\partial f} \neq 0$ can be interpreted in the following way. According to the original IPA theory, asymptotic variance inflates when $P$ and $R$ grow at the same rate and $\mathring{m}$ is constant, implying in particular that $P$ and $R/\mathring{m}$ grow at the same rate. In our asymptotic framework, even though $\mathring{m}$ increases, $P$ and $R/\mathring{m}$ still grow at the same rate (see Assumption 7). That is, in all circumstances it is the balance of growth rates of $P$ and $R/\mathring{m}$ that matters asymptotically. In a way, the ratio $R/\mathring{m}$ measures the degree of parameter estimation uncertainty, and higher numerosity of predictors should be compensated by a proportionately larger regression sample used to form parameter estimates.

**Example.** Consider the MSPE criterion $f_t = u_t^2$ from the linear model $y_t = x_t'\beta + u_t$ with conditionally homoskedastic normal disturbance having variance $\sigma_u^2$ and endogenous regressors

$$x_t = \frac{\gamma}{\sigma_u} u_t \iota_m + z_t,$$

where $z_t$ is $m \times 1$ vector of instruments independent of $u_t$, and for simplicity $z_t \sim \mathcal{N}(0, I_m)$. Then $Q_{\partial f} = -2\gamma\sigma_u\iota_m$, $\mathring{m} = m$, $V_f = 2\sigma_u^4$, $\Sigma_\beta = \sigma_u^2 I_m$, $\psi_2 = 4\gamma^2\sigma_u^4$, so

$$\overline{\hat{u}^2} - 1 \overset{A}{\sim} \mathcal{N}\left(0, \frac{2\sigma_u^4}{P}\left(1 + 2\mu_2\gamma^2\right)\right),$$

and

$$
\begin{aligned}
t_0 &= \frac{\overline{\hat{u}^2} - 1}{\sqrt{2/P}\sigma_u^2}, \\
t_1 &= \frac{\overline{\hat{u}^2} - 1}{\sqrt{2/P + 4(m/R)\gamma^2}\sigma_u^2}.
\end{aligned}
$$

In the first experiment we demonstrate that comparably long regression and prediction intervals leads to testing failure when there are many predictors. We set $R = P = 200$ so that $\pi = 1$, $\beta = \varrho\iota_m$ with the value of $\varrho$ from the example on page 13, and $\gamma = 0.5$. The following table shows rejection frequencies (RF) at the 5% nominal size using the t-statistic $t_1$.

| $m$ | $\mu_2$ | $t_1^{\neq}$ | $t_1^{<}$ | $t_1^{>}$ |
|---|---|---|---|---|
| 2 | 2 | 5.96 | 2.47 | 8.13 |
| 4 | 4 | 6.95 | 1.67 | 9.68 |
| 8 | 8 | 9.19 | 0.78 | 12.51 |
| 16 | 16 | 12.77 | 0.16 | 16.45 |
| 32 | 32 | 19.73 | 0.00 | 23.73 |

One can see that both the two-sided and especially one-sided RF differ appreciably from the nominal levels even when $m = 8$, with severe underrejection at the left tail and severe overrejection at the right tail.[9] The reason is that $\mu_2$ which is supposed to stay fixed asymptotically is too large and comparable (equal in this design) to $m$ which is supposed to grow asymptotically.

In the next experiment, we set $R = 2000$, $P = 100$ so that $\pi = 0.05$, $\beta = \varrho \iota_m$ with the value of $\varrho$ from the example on page 13, and $\gamma = 0.5$. Here, the prediction sample size is negligible relative to the regression sample size ($\pi = 0.05$), and for that reason a researcher may decide to use the conventional unadjusted t-statistic $t_0$. Alternatively, a researcher may adjust for the extra variance and use $t_1$. The following table presents actual rejection rates based on the nominal rate of 5%.

| $m$ | $\mu_2$ | $t_0^{\neq}$ | $t_1^{\neq}$ | $t_0^{<}$ | $t_1^{<}$ | $t_0^{>}$ | $t_1^{>}$ |
|---|---|---|---|---|---|---|---|
| 2 | 0.1 | 5.34 | 4.79 | 4.18 | 3.84 | 6.26 | 5.85 |
| 4 | 0.2 | 6.14 | 5.08 | 4.73 | 3.80 | 6.96 | 6.12 |
| 8 | 0.4 | 7.11 | 5.01 | 4.98 | 3.41 | 8.03 | 6.31 |
| 16 | 0.8 | 10.06 | 5.60 | 5.84 | 2.89 | 10.76 | 7.37 |
| 32 | 1.6 | 14.96 | 6.17 | 7.40 | 2.23 | 14.98 | 8.61 |

One can see that while the variance adjustment for estimation error noise does not change the situation significantly when $m$ is small, it does make good for the actual rejection frequencies, especially two- and right-sided, when there are many predictors.

---

[9]Note that there are problems with one-sided testing even when $m$ is tiny. See footnote 8.

The left-sided tests, however, turn from being oversized to being undersized for large $m$, which is a sort of undershooting phenomenon. This is, however, a purely finite sample issue: values 32 and 16 and even 8 for $m$ are not very much smaller that the value 100 of $P$. In experiments where $P$ gets larger and larger, the left-sided rejection frequency straighten out eventually. The following graph presents the distributions of $t_0$ and $t_1$ when $m = 16$. Apparently, variance adjustment is necessary when there are many predictors because $\mu_2$ is appreciable even though $\pi$ is negligible.
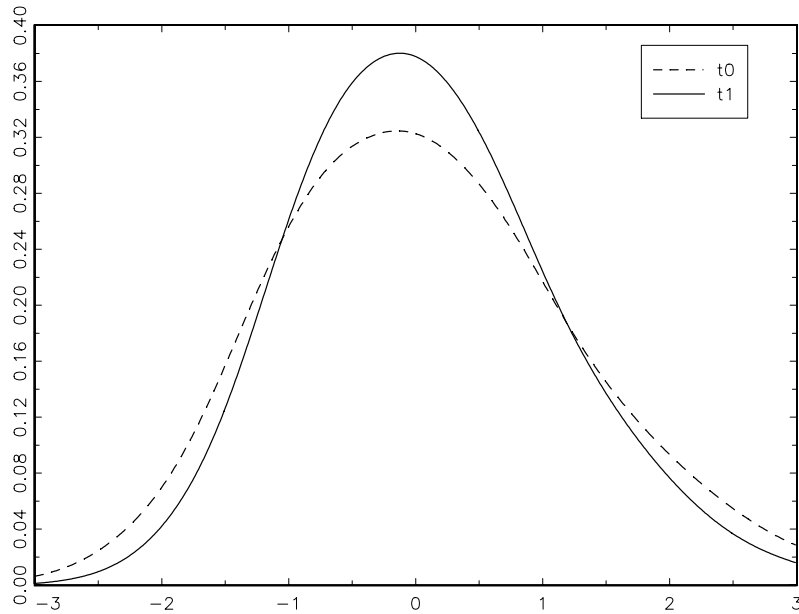


Figure 2: Densities of unadjusted and adjusted t-statistics

# 5   Application

To illustrate, we compare two models for determination of housing starts in the US using the dataset from Galbraith and Zinde-Walsh (2006) originaly drawn from the St. Louis Federal Reserve Bank FRED database.[10] All data are monthly dated from Jan 69 to Mar 04, seasonally adjusted, used in the first differences form, and contain 421 observations in the raw form. The dependent variable is the first difference of housing starts (private including farm, in $mln).

---

[10]I thank John Galbraith for sharing this dataset with me.

Model 1 is a pure autoregressive model of order 2 having $m = 3$ parameters, with $R^2 = 9.3\%$ and $\bar{R}^2 = 8.9\%$. Model 2 is structural, with the predictors (presumed exogenous) representing various sectors of the US economy: national accounts, consumption, real output, wholesale, retail & inventories, money and credit aggregates, price indexes, employment, average hourly earnings, stock market, exchange rates, and interest rates. The sets of predictors are found using the general-to-specific methodology by removing regressors (apart from the constant term) having t-ratios smaller than 1.00, 1.50, or 2.00 (hence 3 sets of predictors) from the full sample regression estimated by OLS[11]. When the t-ratio threshold is 1.00, the selection procedure resulted in $m = 24$ predictors, with $R^2 = 22.3\%$ and $\bar{R}^2 = 17.8\%$. When the t-ratio threshold is 1.50, the selection procedure resulted in $m = 16$ predictors, with $R^2 = 19.7\%$ and $\bar{R}^2 = 16.8\%$. When the t-ratio threshold is 2.00, the selection procedure resulted in $m = 12$ predictors, with $R^2 = 17.5\%$ and $\bar{R}^2 = 15.3\%$. These values of $m$ may well be qualified as "moderately many" relative to the given sample size. Note that the criteria based on the in-sample fit would prefer the structural model with 24 predictors.

The first battery of tests verifies the hypothesis of unbiased one-step-ahead prediction, where the null is $H_0 : E\left[u_t\right] = 0$. The expected score is $Q_{\partial f} = -E\left[x_t\right]$, and all elements of $Q_{\partial f}$ are non-zero, hence the results of section 4 apply, with $\mathring{m} = m$. The uncorrected t-statistic equals

$$t_0 = \frac{\sqrt{P}\,\overline{\hat{u}}}{\sqrt{\overline{\hat{u}^2}}},$$

where $\overline{\hat{u}}$ is an average of prediction errors over the prediction sample, and $\overline{\hat{u}^2}$ is an average of squared regression erros over the regression sample.

It is straightforward to derive using partioned matrix algebra and the fact of the presence of the constant term that $Q_{\partial f}'\Sigma_\beta Q_{\partial f} = \sigma^2$ (a similar relationship also holds at the polulation level, i.e. $\hat{Q}_{\partial f}'\hat{\Sigma}_\beta \hat{Q}_{\partial f} = \overline{\hat{u}^2}$ with $\hat{Q}_{\partial f} = -\bar{x}$, $\hat{\Sigma}_\beta = \overline{\hat{u}^2}(\overline{xx'})^{-1}$). Then the variance-corrected t-statistic equals

$$t_1 = \frac{\sqrt{P}\,\overline{\hat{u}}}{\sqrt{\overline{\hat{u}^2}\,(1 + P/R)}} = \frac{t_0}{\sqrt{1 + P/R}}.$$

[11] We do not view this procedure as a rigorous tool of model selection, but rather use it as an appealing in practice algorithm in order to end up with several "structural" models.

Interestingly, in this class of problems the value of $m$ does not affect inference. This is because $\psi_2 = \lim(\sigma^2/m) = 0$, and the results of section 4 hold trivially. Recall that Assumption 7 requires $P/R \to 0$, so under it $t_0$ and $t_1$ are asymptotically equivalent. However, in the original IPA theory framework $P/R \to \pi$, and if $\pi \neq 0$, the statistic $t_1$ is asymptotically correct while the statistic $t_0$ is not.

The following tables contains the results for $P/R \approx \frac{1}{9}, \frac{1}{3}, 1, 3$ and $9$ (smaller or larger values of $P/R$ are hardly justifiable given the sample size). The slight difference in values of $R$ and $P$ across the models is due to a different number of lags employed. All inference conclusions are made at the 5% significance level.

| $R$ | $P$ | $P/R$ | $t_0$ | $t_1$ |
|---|---|---|---|---|
| 376 | 42 | $\approx \frac{1}{9}$ | 0.93 | 0.89 |
| 313 | 105 | $\approx \frac{1}{3}$ | 0.90 | 0.78 |
| 209 | 209 | 1 | $-0.03$ | $-0.02$ |
| 104 | 314 | $\approx 3$ | $-0.86$ | $-0.44$ |
| 42 | 376 | $\approx 9$ | $-5.10$ | $-1.67$ |

(A) Autoregressive model

| | | | $m = 24$ | | $m = 16$ | | $m = 12$ | |
|---|---|---|---|---|---|---|---|---|
| $R$ | $P$ | $P/R$ | $t_0$ | $t_1$ | $t_0$ | $t_1$ | $t_0$ | $t_1$ |
| 378 | 42 | $\approx \frac{1}{9}$ | $-0.66$ | $-0.63$ | 0.52 | 0.49 | 0.32 | 0.31 |
| 315 | 105 | $\approx \frac{1}{3}$ | $-2.01$ | $-1.74$ | 0.16 | 0.14 | 0.09 | 0.08 |
| 210 | 210 | 1 | $-2.36$ | $-1.67$ | $-1.13$ | $-0.80$ | $-1.83$ | $-1.29$ |
| 105 | 315 | $\approx 3$ | $-2.67$ | $-1.34$ | $-1.89$ | $-0.95$ | $-0.93$ | $-0.46$ |
| 42 | 378 | $\approx 9$ | $-4.57$ | $-1.46$ | $-14.27$ | $-4.56$ | $-10.61$ | $-3.39$ |

(B) Structural model

We classify the values $\frac{1}{9}$, $\frac{1}{3}$, 1, 3 for $P/R$ as compatible with the original IPA theory (where $P/R \to \pi < \infty$, possibly $\pi = 0$), the value $\frac{1}{9}$ compatible with our asymptotic framework (where $P/R \to 0$, $Pm/R \to \mu_2 < \infty$, $\mu_2 \neq 0$), and the value 9 incompatible with both.

For the autoregressive model with few predictors, the values of $t_1$ indicate that the data support that the model generates unbiased forecasts, and this conclusion is consistent across combinations of $R$ and $P$. It can also be seen that the use of $t_0$ may lead to the wrongful rejection of the unbiasedness hypothesis when $P/R$ is large (line 5) because of a big estimation noise unaccounted for.

For the structural model with many predictors, in lines 1–4 compatible with the original IPA theory, the variance-adjusted statistic $t_1$ leads to the acceptance of the null, while the unadjusted statistic $t_0$ may reject it for a variety of combinations of $P$ and $R$ when $m = 24$. In line 1 compatible with our framework both statistics agree on the outcome for all three sets of predictors. Note that in line 5 incompatible with both frameworks even the adjusted statistic $t_1$ may lead to a wrong outcome.

The second battery of tests verifies the hypothesis of equal accuracy of one-step-ahead prediction across the models (Diebold and Mariano, 1995), where the null is $H_0 :$ $E\left[u_{2,t}^2 - u_{1,t}^2\right] = 0$. Here $u_{s,t}^2$ is the one-step-ahead prediction error from model $s$ $(s = 1, 2)$, and let us also denote by $x_{s,t}$, $m_s$ and $\beta_s$ the predictors, their number, and the parameters from model $s$. The full parameter vector is $\beta = (\beta_1', \beta_2')'$, and the expected score is $Q_{\partial f} = 2E\left[(u_{1,t}x_{1,t}, -u_{2,t}x_{2,t})'\right] = 0$, hence the results of section 3 apply. The uncorrected t-statistic equals

$$t_0 = \frac{\sqrt{P}\,\overline{\hat{u}_2^2 - \hat{u}_1^2}}{\sqrt{\hat{V}_{\Delta u^2}}},$$

where $\hat{u}_s$ are estimated prediction errors from model $s$, $\overline{\hat{u}_1^2 - \hat{u}_2^2}$ is an average of the difference of their squares over the prediction sample, and $\hat{V}_{\Delta u^2}$ is the HAC esimate of the variance of $u_{2,t}^2 - u_{1,t}^2$. This statistic is simply computed as a Newey–West corrected t-ratio from a regression of $\hat{u}_2^2 - \hat{u}_1^2$ on a constant.

Now observe that $Q_{\partial^2 f}$ is block-diagonal with blocks equalling $-2E\left[x_{1,t}x_{1,t}'\right]$ and $2E\left[x_{2,t}x_{2,t}'\right]$, and the diagonal blocks of $\Sigma_\beta$ equal $\sigma_1^2 E\left[x_{1,t}x_{1,t}'\right]^{-1}$ and $\sigma_2^2 E\left[x_{2,t}x_{2,t}'\right]^{-1}$,

thus $\text{tr}\,(Q_{\partial^2 f}\Sigma_\beta) = 2\,(\sigma_2^2 m_2 - \sigma_1^2 m_1)$, and the finite-sample version of $\psi_1$ is $\hat\psi_1 = 2(\overline{\hat u_2^2}m_2 - \overline{\hat u_1^2}m_1)/(m_1 + m_2)$. If by coincidence $\sigma_2^2 m_2 = \sigma_1^2 m_1$, then $\lim_{m\to\infty}\hat\psi_1 = 0$, and no bias correction is needed. Bias correction is obviously needed in our case when one of the two models (model 1) contains few predictors, and the other (model 2) contains many predictors, in which case the appropriate asymptotics is $m_1 = \text{const}$, $m_2 \to \infty$ so that $\lim_{m\to\infty}\hat\psi_1 = 2\sigma_2^2 \neq 0$. The bias-corrected t-statistic equals

$$t_1 = \frac{\sqrt{P}\left(\overline{\hat u_2^2 - \hat u_1^2} - \hat b_{\Delta u^2}\right)}{\sqrt{\hat V_{\Delta u^2}}}, \quad \text{where } \hat b_{\Delta u^2} = \frac{\overline{\hat u_2^2}m_2 - \overline{\hat u_1^2}m_1}{R}.$$

The statistic $t_1$ is simply computed as a Newey–West corrected t-ratio from a regression of $\hat u_2^2 - \hat u_1^2 - \hat b_{\Delta u^2}$ on a constant.

The following tables contains the results for $P/R \approx \frac{1}{9}, \frac{1}{3}, 1, 3$ and $9$. All inference conclusions are made at the 5% significance level. We in addition report values of $Pm^2/R^2$.

| | | | $m_2 = 24$ | | | $m_2 = 16$ | | | $m_2 = 12$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R$ | $P$ | $P/R$ | $Pm^2/R^2$ | $t_0$ | $t_1$ | $Pm^2/R^2$ | $t_0$ | $t_1$ | $Pm^2/R^2$ | $t_0$ | $t_1$ |
| 376 | 42 | $\approx\frac{1}{9}$ | 0.22 | 0.59 | 0.11 | 0.11 | $-0.44$ | $-0.82$ | 0.07 | $-1.68$ | $-2.05$ |
| 313 | 105 | $\approx\frac{1}{3}$ | 0.78 | 3.88 | 3.21 | 0.39 | 2.68 | 2.07 | 0.24 | 2.04 | 1.54 |
| 209 | 209 | 1 | 3.49 | 2.69 | 1.40 | 1.73 | 0.84 | $-0.27$ | 1.08 | 0.61 | $-0.32$ |
| 104 | 314 | $\approx 3$ | 21.16 | 2.83 | 1.49 | 10.48 | 1.98 | 1.05 | 6.53 | 0.86 | 0.08 |
| 42 | 376 | $\approx 9$ | 155.39 | 7.09 | 6.64 | 76.95 | 7.25 | 6.36 | 47.96 | 3.86 | 2.52 |

We classify the values $\frac{1}{9}, \frac{1}{3}, 1, 3$ for $P/R$ as compatible with the original IPA theory (where $P/R \to \pi < \infty$, possibly $\pi = 0$), and recall that there is no need to adjust the t-statistic. From the viewpoint of the original IPA theory using the unadjusted $t_0$, the results regarding the null hypothesis of equal prediction accuracy are contradictory for all three sets of predictors: the null is rejected for some combinations of $R$ and $P$ and is not rejected for others. There is no futher guide which of the results are more reliable.

From the viewpoint of our theory, we cannot trust line 1 because of too low value of $P/R$ (recall that Assumption 5 requires $P/R \to \infty$) and lines 4 and 5 (recall that

Assumption 5 requires $Pm^2/R^2 \to \mu_1 < \infty$, and that simulations reported in section 3 show that for large values of $\mu_1$ there is severe overrejection). Therefore, when $m_2 = 24$, only line 2 can be trusted, line 3 being excluded for the reason of too big $Pm^2/R^2$, and the bias-adjusted t-statistic $t_1 = 3.21$ is able to reject equal prediction accuracy. Because $t_1 > 0$, the largest structural model may be deemed less accurate in terms of predictive ability than the autoregressive model (recall though that the in-sample criterion $\bar{R}^2$ favors it among all four models). When $m_2 = 16$, most trustable is line 3, with a statistically insignificant bias-adjusted t-statistic $t_1 = 0.84$. Even in the less credible line 2, even though there is rejection at the 5% level, it is marginal. Hence, for the medium structural model we cannot reject the hypothesis that it is as accurate as the autoregressive model. When $m_2 = 12$, most trustworthy is line 3, with a statistically insignificant bias-adjusted t-statistic $t_1 = -0.32$; using line 2 leads to the same outcome that for the smallest structural model we cannot reject the hypothesis of equal prediction accuracy.

# Conclusion

This paper complements the theory of asymptotic inference about predictive ability of West (1996) and West and McCracken (1998, 2002) by considering the case when a set of variables used to construct predictions is sizable. Depending on the situation the asymptotic normal distribution of an average prediction criterion either gains additional variance as in the few predictors case, or gains non-zero bias which has no analogs in the few predictors case. By properly modifying conventional test statistics it is possible to remove most size distortions when there are many predictors, and improve test sizes even when there are few of them.

One of methodological implications of our results for the time series analysis is that testing of out-of-sample qualities of semi-nonparametric models such as ANN (intrinsically having many parameters to estimate) often observed in empirical literature may not be valid in their classical unadjusted form. This is one of potential topics of future research. Another line of research may be devoted to comparison of nested models as in Clark and McCracken (2001), Clark and West (2006) and McCracken (2006).

# 6 References

Bekker, P.A. (1994) Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica* 62, 657–681.

Chao, J.C. and N.R. Swanson (2005) Consistent Estimation with a Large Number of Weak Instruments. *Econometrica* 73, 1673–1692.

Clark, T.E. and M.W. McCracken (2001) Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105, 85–110.

Clark, T.E. and K.D. West (2006) Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* forthcoming.

Davidson, J. (2000) *Econometric Theory*. UK: Blackwell Publishers.

Diebold, F.X. and R.S. Mariano (1995) Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13, 253–263.

Galbraith J.W. and V. Zinde-Walsh (2006) Reduced-Dimension Control Regression. Manuscript, McGill University.

Golub, G.H. and C.F. Van Loan (1996) *Matrix Computations*. Baltimore: Johns Hopkins University Press.

Ibragimov, I.A. (1962) Some limit theorems for stationary processes. *Theory of Probability and its Applications* 7, 349–382.

Koenker, R. (1988) Asymptotic Theory and Econometric Practice. *Journal of Applied Econometrics* 3, 139–147.

Koenker, R. and J.A.F. Machado (1999) GMM Inference When the Number of Moment Conditions is Large. *Journal of Econometrics* 93, 327–344.

Magnus, J.R and H. Neudecker (1988) *Matrix Differential Calculus with Applications in Statistics and Econometrics*. UK: John Wiley and Sons.

McCracken, M.W. (2006) Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics* forthcoming.

Newey, W.K. and F. Windmeijer (2005) GMM Estimation with Many Weak Moment Conditions. Manuscript, MIT.

Stock, J.H. and M.W. Watson (2006) Forecasting with many predictors. In: G. Elliott,

C. Granger, A. Timmermann, K. Arrow, M.D. Intriligator, eds. *Handbook of economic forecasting*, Elsevier: North Holland.

Stock, J.H. and M. Yogo (2005) Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments. In: D.W.K. Andrews and J.H. Stock, eds. *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas J. Rothenberg.* Cambridge, UK: Cambridge University Press.

West, K.D. (1996) Asymptotic inference about predictive ability. *Econometrica* 64, 1067–1084.

West, K.D. and M.W. McCracken (1998) Regression-based tests of predictive ability. *International Economic Review* 39, 817–840.

West, K.D. and M.W. McCracken (2002) Inference about predictive ability. In: M. Clements and D. Hendry (eds.) *Companion to Economic Forecasting*, Oxford: Blackwell.

# A   Appendix: auxiliary results

Let $A$ be $m \times m$ matrix. By $\|A\|_\infty$ we denote the $L^\infty$ matrix norm $\max_{1 \leq i \leq m} \left( \sum_{j=1}^m |a_{ij}| \right)$, by $\|A\|_1$ – the $L^1$ norm $\max_{1 \leq j \leq m} \left( \sum_{i=1}^m |a_{ij}| \right)$, and by $\|A\|_F$ – the Frobenius norm $\sqrt{\operatorname{tr}(A'A)}$.

**Lemma 1** *Let $m$ be the dimension of square matrices $A$ and $B$.*

(a) *Suppose $B$ is symmetric. Then $\|B\| = |\bar{\lambda}(B)|$.*

(b) *Suppose $B_1$ and $B_2$ are symmetric. Then $|\bar{\lambda}(B_1 B_2)| \leq |\bar{\lambda}(B_1)| |\bar{\lambda}(B_2)|$.*

(c) *For any $m \times m$ matrix $B$, $\sum_i \sum_j |b_{ij}| \leq m^{3/2} \|B\|$.*

(d) *For any $m \times m$ matrix $B$, $\|B\|_F \leq \sqrt{m} \|B\|$.*

**Proof.** (a) Because $B$ is symmetric, all eigenvalues of $B'B$ are squared eigenvalues of $B$ because $\lambda(B'B) = \lambda(B^2) = \lambda(B)^2$, so $\|B\| = \sqrt{\bar{\lambda}(B)^2} = |\bar{\lambda}(B)|$. (b) Take unit norm eigenvector $v(B_1 B_2)$ of matrix $B_1 B_2$ corresponding to eigenvalue $\lambda(B_1 B_2)$, then $\|B_1 B_2 v(B_1 B_2)\| = \|\lambda(B_1 B_2) v(B_1 B_2)\| = |\lambda(B_1 B_2)|$. On the other hand, $\|B_1 B_2 v(B_1 B_2)\| \leq \|B_1\| \|B_2\| = |\bar{\lambda}(B_1)| |\bar{\lambda}(B_2)|$ by (a). Thus, we obtain $|\lambda(B_1 B_2)| \leq |\bar{\lambda}(B_1)| |\bar{\lambda}(B_2)|$, which holds for any eigenvalue $\lambda(B_1 B_2)$, hence for the maximal in absolute value too. (c) $\sum_i \sum_j |b_{ij}| \leq \sum_i \max_i \sum_j |b_{ij}| = m \|B\|_1 \leq m\sqrt{m} \|B\|$ (GV, eqn. 2.3.12). (d) See GV, eqn. 2.3.7. ∎

**Lemma 2** *Under assumption 2,*

(a) $c < \|Q_{zx}\| < C$ *and* $C^{-1} < \|Q_{zx}^{-1}\| < c^{-1}$,

(b) $\left\| V_{zu}^{1/2} \right\| < C^{1/2}$,

(c) $\bar{\lambda}(\Sigma_\beta) = \|\Sigma_\beta\| < C/c^2$.

**Proof.** (a) Trivially, $\|Q_{zx}\| = \bar{\sigma}(Q_{zx}) < C$ and $\|Q_{zx}\| \geq \underline{\sigma}(Q_{zx}) > c$. From the properties of matrix norms, $\|Q_{zx}^{-1}\| \geq \|Q_{zx}\|^{-1} > C^{-1}$. Next, $\|Q_{zx}^{-1}\| = \sqrt{\bar{\lambda}(Q_{zx}^{-1\prime} Q_{zx}^{-1})} = \sqrt{\bar{\lambda}((Q_{zx} Q_{zx}')^{-1})} = \sqrt{\underline{\lambda}(Q_{zx} Q_{zx}')^{-1}} = \underline{\sigma}(Q_{zx})^{-1} < c^{-1}$. (b) $\left\| V_{zu}^{1/2} \right\| = \sqrt{\bar{\lambda}\left( V_{zu}^{1/2\prime} V_{zu}^{1/2} \right)} = \sqrt{\bar{\lambda}(V_{zu})} < C^{1/2}$. (c) By Lemma 1(a) and because $\Sigma_\beta$ is symmetric and positive definite, $\bar{\lambda}(\Sigma_\beta) = \|\Sigma_\beta\|$. Next, $\|\Sigma_\beta\| \leq \bar{\lambda}(V_{zu}) \|Q_{zx}^{-1}\|^2 < C/c^2$ using (a). ∎

**Lemma 3** *Let $x_t$ and $y_t$ be $\Im_t$-measurable stationary $\alpha$-mixing scalar processes with mixing coefficients $\alpha_i$, $E\left[|x_t|^{2\nu}\right], E\left[|y_t|^{2\nu}\right] < \infty$ for some $\nu > 1$, and let $x_t$ have zero mean. Then for all $j > 0$,*

$$\left|E\left[x_t x_{t+j}\right]\right| \le 8\alpha_j^{1-1/\nu}\left(E\left[|x_t|^{2\nu}\right]\right)^{1/\nu}.$$

*and*

$$\left|E\left[x_t y_{t+j}\right]\right| \le 8\alpha_j^{1-1/\nu}\left(E\left[|x_t|^{2\nu}\right]\right)^{1/2\nu}\left(E\left[|y_t|^{2\nu}\right]\right)^{1/2\nu}.$$

**Proof.** Using the Cauchy–Schwartz and Ibragimov (1962) inequalities,

$$
\begin{aligned}
\left|E\left[x_t y_{t+j}\right]\right| &= \left|E\left[x_t\left(E\left[y_{t+j}|\Im_t\right] - E\left[y_t\right]\right)\right]\right| \\
&\le \left(E\left[|x_t|^{2\nu}\right]\right)^{1/2\nu}\left(E\left[\left|E\left[y_{t+j}|\Im_t\right] - E\left[y_t\right]\right|^{2\nu/(2\nu-1)}\right]\right)^{1-1/2\nu} \\
&\le \left(E\left[|x_t|^{2\nu}\right]\right)^{1/2\nu} \cdot 8\alpha_j^{1-1/\nu}\left(E\left[|y_t|^{2\nu}\right]\right)^{1/2\nu}.
\end{aligned}
$$

The first inequality is a special case with $y_t = x_t$. ∎

Denote

$$
\begin{aligned}
\xi_{zu} &= \frac{1}{\sqrt{R}}\sum_{t=1}^{R} z_t u_t, \\
\hat{Q}_{zx} &= \frac{1}{R}\sum_{t=1}^{R} z_t x_t'.
\end{aligned}
$$

**Lemma 4** *Under Assumption 1 and 2, as $m \to \infty$, $R \to \infty$ and $m^2/R \to 0$,*

(a) $\left\|\hat{Q}_{zx} - Q_{zx}\right\|$ and $\left\|\hat{Q}_{zx}^{-1} - Q_{zx}^{-1}\right\|$ are $O_p\left(m/\sqrt{R}\right)$,

(b) $\left\|\hat{Q}_{zx}^{-1}\right\| < c^{-1}$ with probability approaching 1,

(c) $\left\|\xi_{zu}\right\|$ is $O_p\left(\sqrt{m}\right)$.

**Proof.** (a) Observe that $E\left[\left\|\hat{Q}_{zx} - Q_{zx}\right\|_F^2\right]$ is bounded by

$$
\leq \quad \frac{1}{R}E\left[\sum_{i=1}^m \sum_{j=1}^m (z_{t,i}x_{t,j} - E\left[z_{t,i}x_{t,j}\right])^2\right]
$$

$$
+ \frac{2}{R^2}\sum_{1\leq t<s\leq R} E\left[\sum_{i=1}^m \sum_{j=1}^m |(z_{t,i}x_{t,j} - E\left[z_{t,i}x_{t,j}\right])(z_{s,i}x_{s,j} - E\left[z_{s,i}x_{s,j}\right])|\right]
$$

$$
\leq \quad \frac{m^2}{R}\left(\max_{1\leq i\leq m} E\left[z_{t,i}^4\right]\max_{1\leq j\leq m} E\left[x_{t,j}^4\right] + \left(\max_{1\leq i\leq m} E\left[z_{t,i}^2\right]\max_{1\leq j\leq m} E\left[x_{t,j}^2\right]\right)^2\right)
$$

$$
+ \frac{16m^2}{R^2}\left(\left(\max_{1\leq i\leq m} E\left[z_{t,i}^{4\nu}\right]\right)^{1/2\nu}\left(\max_{1\leq j\leq m} E\left[x_{t,j}^{4\nu}\right]\right)^{1/2\nu} + \max_{1\leq i\leq m} E\left[z_{t,i}^2\right]\max_{1\leq j\leq m} E\left[x_{t,j}^2\right]\right)^2
$$

$$
\times \sum_{1\leq t<s\leq R} \alpha_{s-t}^{1-1/\nu}
$$

$$
\leq \quad O\left(\frac{m^2}{R}\right),
$$

using Lemma 3, Assumptions 1(i,ii) and the Minkowski and triangular inequalities. Because $\|B\| \leq \|B\|_F$, we conclude that $\left\|\hat{Q}_{zx} - Q_{zx}\right\|$ is $O_p\left(m/\sqrt{R}\right)$. Now, $\left\|\hat{Q}_{zx}^{-1} - Q_{zx}^{-1}\right\| \leq \left\|\hat{Q}_{zx}^{-1}\right\|\left\|\hat{Q}_{zx} - Q_{zx}\right\|\|Q_{zx}^{-1}\| \leq O_p\left(m/\sqrt{R}\right)$ using also the result in (b).

(b) Because $\hat{Q}_{zx}^{-1} = Q_{zx}^{-1} - \hat{Q}_{zx}^{-1}\left(\hat{Q}_{zx} - Q_{zx}\right)Q_{zx}^{-1}$, we have

$$
\left\|\hat{Q}_{zx}^{-1}\right\| \leq \|Q_{zx}^{-1}\| + \left\|\hat{Q}_{zx}^{-1}\right\|\left\|\hat{Q}_{zx} - Q_{zx}\right\|\|Q_{zx}^{-1}\|
$$

and hence

$$
\left\|\hat{Q}_{zx}^{-1}\right\| \leq \left(1 - \left\|\hat{Q}_{zx} - Q_{zx}\right\|\|Q_{zx}^{-1}\|\right)^{-1}\|Q_{zx}^{-1}\| \leq \left(1 - o_p(1)c^{-1}\right)^{-1}c^{-1}
$$

from Lemma 2(a) and the first result in (a). The result follows.

(c) Observe that

$$
E\left[\|\xi_{zu}\|^2\right] \quad = \quad E\left[\frac{1}{R}\sum_{i=1}^m \left(\sum_{t=1}^R z_{t,i}u_t\right)^2\right] = \frac{1}{R}\sum_{i=1}^m E\left[\sum_{t=1}^R\sum_{s=1}^R z_{t,i}z_{s,i}u_t u_s\right]
$$

$$
= \quad \text{tr}\left(V_{zu}\right) \leq m\bar{\lambda}\left(V_{zu}\right) = O_p\left(m\right),
$$

using 2(ii) and because the trace of a square matrix equals a sum of its eigenvalues (MN, thm.17, p.19), and the conclusion follows. ∎

31

# B  Appendix: proofs related to case $Q_{\partial f} = 0$

Denote

$$
\begin{aligned}
\Sigma_{\partial^2 f} &= Q_{zx}'^{-1} Q_{\partial^2 f} Q_{zx}^{-1}, \\
\xi_{\partial f} &= \frac{1}{\sqrt{P}} \sum_{t=R+1}^{R+P} \frac{\partial f_t}{\partial \beta}, \\
\hat{Q}_{\partial^2 f}^* &= \frac{1}{P} \sum_{t=R+1}^{R+P} \frac{\partial^2 f_t^*}{\partial \beta \partial \beta'}.
\end{aligned}
$$

**Lemma 5** *Under assumptions 2 and 4,*

*(a)* $\|\Sigma_{\partial^2 f}\| < C/c^2$ *and* $\left\|\Sigma_{\partial^2 f}^{1/2}\right\| < \sqrt{C}/c,$

*(b)* $\left\|\xi_{\partial f}\right\|$ *is* $O_p\left(\sqrt{m}\right).$

**Proof.** (a) Because $Q_{\partial^2 f}$ is symmetric, using Lemma 1(a), $\left|\bar{\lambda}\left(Q_{\partial^2 f}\right)\right| = \|Q_{\partial^2 f}\| < C.$ (b) Note that $\Sigma_{\partial^2 f}$ is symmetric, so by Lemmas 2(a) and 1(a) $\|\Sigma_{\partial^2 f}\| \leq \left|\bar{\lambda}\left(Q_{\partial^2 f}\right)\right| \|Q_{zx}^{-1}\|^2 < C/c^2$ using (a). Next, $\left\|\Sigma_{\partial^2 f}^{1/2}\right\| = \|\Sigma_{\partial^2 f}\|^{1/2}$ as in Lemma 2(b). (b) Observe that

$$
\begin{aligned}
E\left[\|\xi_{\partial f}\|^2\right] &= E\left[\frac{1}{P} \sum_{i=1}^{m} \left(\sum_{t=R+1}^{R+P} \frac{\partial f_t}{\partial \beta}\right)^2\right] = \frac{1}{P} \sum_{i=1}^{m} E\left[\sum_{t=R+1}^{R+P} \sum_{s=R+1}^{R+P} \frac{\partial f_t}{\partial \beta_i} \frac{\partial f_s}{\partial \beta_i}\right] \\
&= \sum_{i=1}^{m} E\left[\left(\frac{\partial f_t}{\partial \beta_i}\right)^2\right] + \frac{2}{P} \sum_{R+1 \leq t < s \leq R+P} \sum_{i=1}^{m} E\left[\frac{\partial f_t}{\partial \beta_i} \frac{\partial f_s}{\partial \beta_i}\right] \\
&\leq \operatorname{tr}\left(V_{\partial f}\right) + \frac{16}{P} \sum_{R+1 \leq t < s \leq R+P} \alpha_{s-t}^{1-1/\nu} \sum_{i=1}^{m} \left(E\left[\left|\frac{\partial f_s}{\partial \beta_i}\right|^{2\nu}\right]\right)^{1/\nu} \\
&\leq m\bar{\lambda}\left(V_{\partial f}\right) + O\left(m\right) = O\left(m\right),
\end{aligned}
$$

using 4(iv, v) and because the trace of a square matrix equals a sum of its eigenvalues (MN, thm.17, p.19). The conclusion follows. ■

**Lemma 6** *Under the asymptotics of assumption 5 and conditions of assumptions 2, 3 and 4,*

*(a)*

$$
\frac{\xi_{zu}' Q_{zx}'^{-1} Q_{\partial^2 f} Q_{zx}^{-1} \xi_{zu}}{m} = \psi_1 + o_p\left(1\right),
$$

*(b)*
$$\left| \frac{\xi'_{zu}\hat{Q}'^{-1}_{zx}\hat{Q}_{\partial^2 f}\hat{Q}^{-1}_{zx}\xi_{zu}}{m} - \frac{\xi'_{zu}Q'^{-1}_{zx}Q_{\partial^2 f}Q^{-1}_{zx}\xi_{zu}}{m} \right| \leq o_p(1),$$

*(c)*
$$\left| \frac{\xi'_{zu}\hat{Q}'^{-1}_{zx}\hat{Q}^*_{\partial^2 f}\hat{Q}^{-1}_{zx}\xi_{zu}}{m} - \frac{\xi'_{zu}\hat{Q}'^{-1}_{zx}\hat{Q}_{\partial^2 f}\hat{Q}^{-1}_{zx}\xi_{zu}}{m} \right| \leq o_p(1).$$

**Proof.** (a) Note that

$$\frac{\xi'_{zu}\Sigma_{\partial^2 f}\xi_{zu}}{m} = \frac{1}{mR}\sum_{i=1}^{m}\sum_{j=1}^{m}(\Sigma_{\partial^2 f})_{ij}\left(\sum_{t=1}^{R}z_{i,t}u_t\right)\left(\sum_{s=1}^{R}z_{j,s}u_s\right)$$

$$= \frac{1}{R}\sum_{t=1}^{R}\frac{(z_t u_t)'\Sigma_{\partial^2 f}(z_t u_t)}{m} + \frac{2}{R}\sum_{1\leq t<s\leq R}\frac{(z_t u_t)'\Sigma_{\partial^2 f}(z_s u_s)}{m}$$

$$= A_1 + 2A_2,$$

say. Consider the first term

$$A_1 = \frac{1}{R}\sum_{t=1}^{R}\frac{(z_t u_t)'\Sigma_{\partial^2 f}(z_t u_t)}{m}.$$

Its expectation is

$$E[A_1] = E\left[\frac{(z_t u_t)'\Sigma_{\partial^2 f}(z_t u_t)}{m}\right] = E\left[\frac{\text{tr}(\Sigma_{\partial^2 f}(z_t z'_t u_t^2))}{m}\right]$$

$$= \frac{\text{tr}(Q'^{-1}_{zx}Q_{\partial^2 f}Q^{-1}_{zx}V_{zu})}{m} = \frac{\text{tr}(Q_{\partial^2 f}Q^{-1}_{zx}V_{zu}Q'^{-1}_{zx})}{m} \to \psi_1,$$

and the variance is

$$V[A_1] = \frac{1}{m^2 R^2}\sum_{t=1}^{R}\sum_{s=1}^{R}E\left[\text{tr}\left(\Sigma_{\partial^2 f}\left(z_t z'_t u_t^2 - V_{zu}\right)\right)\text{tr}\left(\Sigma_{\partial^2 f}\left(z_s z'_s u_s^2 - V_{zu}\right)\right)\right]$$

$$= \frac{1}{m^2 R}E\left[\text{tr}\left(\Sigma_{\partial^2 f}\left(z_t z'_t u_t^2 - V_{zu}\right)\right)^2\right]$$

$$+ \frac{2}{m^2 R^2}\sum_{1\leq t<s\leq R}E\left[\text{tr}\left(\Sigma_{\partial^2 f}\left(z_t z'_t u_t^2 - V_{zu}\right)\right)\text{tr}\left(\Sigma_{\partial^2 f}\left(z_s z'_s u_s^2 - V_{zu}\right)\right)\right]$$

$$= V_1 + 2V_2,$$

say. Now,

$$
\begin{aligned}
V_1 &= \frac{1}{m^2 R} E\left[\left(\sum_{i=1}^m \sum_{j=1}^m [\Sigma_{\partial^2 f}]_{ij}\left(z_{t,j}z_{t,i}u_t^2 - [V_{zu}]_{ji}\right)\right)^2\right] \\
&\leq \frac{1}{m^2 R}\sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m \sum_{l=1}^m \left|[\Sigma_{\partial^2 f}]_{ij}[\Sigma_{\partial^2 f}]_{kl}\right|\left(E\left[\left|z_{t,i}z_{t,j}z_{t,k}z_{t,l}u_t^4\right|\right] + \left|[V_{zu}]_{ij}[V_{zu}]_{kl}\right|\right) \\
&\leq \frac{1}{m^2 R}\left(\max_{1\leq i\leq m} E\left[z_{i,t}^8\right]\right)^{1/2} E\left[u_t^8\right]^{1/2}\left(\sum_{i=1}^m \sum_{j=1}^m \left|[\Sigma_{\partial^2 f}]_{ij}\right|\right)^2 + \frac{1}{m^2 R}\mathrm{tr}\left(\Sigma_{\partial^2 f}V_{zu}\right)^2 \\
&\leq \frac{1}{m^2 R}C\left(m^{3/2}\|\Sigma_{\partial^2 f}\|\right)^2 + \frac{1}{m^2 R}\left(m\psi_1 + o(m)\right)^2 \\
&\leq O\left(\frac{m}{R}\right) = o(1),
\end{aligned}
$$

using Lemma 5(a) and Assumptions 1(i,ii), 2 and 6(i). The other term in $V[A_1]$ satisfies

$$
\begin{aligned}
|V_2| &\leq \frac{1}{m^2 R^2}\sum_{1\leq t<s\leq R}\left(E\left[\mathrm{tr}\left(\Sigma_{\partial^2 f}\left(z_t z_t' u_t^2 - V_{zu}\right)\right)\mathrm{tr}\left(\Sigma_{\partial^2 f}\left(z_s z_s' u_s^2 - V_{zu}\right)\right)\right]\right) \\
&\leq \frac{8}{m^2 R^2}E\left[\left|\mathrm{tr}\left(\Sigma_{\partial^2 f}\left(z_t z_t' u_t^2 - V_{zu}\right)\right)\right|^{2\nu}\right]^{1/\nu}\sum_{1\leq t<s\leq R}\alpha_{s-t}^{1-1/\nu} \\
&\leq \frac{8}{m^2 R^2}E\left[\left|\sum_{i=1}^m \sum_{j=1}^m [\Sigma_{\partial^2 f}]_{ij}\left(z_{t,j}z_{t,i}u_t^2 - [V_{zu}]_{ji}\right)\right|^{2\nu}\right]^{1/\nu}\left(R\sum_{k=1}^\infty \alpha_k^{1-1/\nu}\right) \\
&\leq \frac{8}{m^2 R}\left(\left(m^{3/2}\|\Sigma_{\partial^2 f}\|\left(\max_{1\leq i\leq m} E\left[z_{i,t}^{8\nu}\right]\right)^{1/4\nu} E\left[u_t^{8\nu}\right]^{1/4\nu}\right)^{2\nu} +?\left|\mathrm{tr}\left(\Sigma_{\partial^2 f}V_{zu}\right)\right|^{2\nu}\right)^{1/\nu} \\
&\qquad\qquad \times \sum_{k=1}^\infty \alpha_k^{1-1/\nu} \\
&\leq O\left(\frac{m}{R}\right) = o(1),
\end{aligned}
$$

using in addition Lemmas 3 and 5(a) and the Minkowski and triangular inequalities.

The second term in $\xi_{zu}' \Sigma_{\partial^2 f}\xi_{zu}/m$ equals twice

$$
A_2 = \frac{1}{R}\sum_{1\leq t<s\leq R}(z_s u_s)'\frac{\Sigma_{\partial^2 f}}{m}(z_t u_t).
$$

Let us look at the MSE of this expression:

$$
\begin{aligned}
MSE &= E\left[\frac{1}{R^2}\sum_{1\leq t_1<s_1\leq R}\sum_{1\leq t_2<s_2\leq R}(z_{s_1}u_{s_1})'\frac{\Sigma_{\partial^2 f}}{m}(z_{t_1}u_{t_1})(z_{s_2}u_{s_2})'\frac{\Sigma_{\partial^2 f}}{m}(z_{t_2}u_{t_2})\right] \\
&= \frac{1}{R^2}\sum_{1\leq t_1<R}\sum_{t_1<s_1\leq R}\sum_{1\leq t_2<s_1}E\left[(z_{s_1}u_{s_1})'\frac{\Sigma_{\partial^2 f}}{m}(z_{t_1}u_{t_1})(z_{s_1}u_{s_1})'\frac{\Sigma_{\partial^2 f}}{m}(z_{t_2}u_{t_2})\right],
\end{aligned}
$$

34

because all terms with $s_1 < s_2$ or $s_1 > s_2$ are zero because of the MDS structure of $u_t$.

Next, the foregoing expression is a sum of two, the first of which is

$$MSE_1 = \frac{1}{R^2} \sum_{1 \leq t < s \leq R} E\left[ \left( (z_s u_s)' \frac{\Sigma_{\partial^2 f}}{m} (z_t u_t) \right)^2 \right].$$

Denote $\zeta_t = \Sigma_{\partial^2 f}^{1/2} (z_t u_t)$, then

$$
\begin{aligned}
MSE_1 &= \frac{1}{m^2 R^2} \sum_{1 \leq t < s \leq R} E\left[ (\zeta_t' \zeta_s)^2 \right] = \frac{1}{m^2 R^2} \sum_{1 \leq t < s \leq R} E\left[ \sum_{i=1}^{m} \sum_{j=1}^{m} \zeta_{t,i} \zeta_{t,j} \zeta_{s,i} \zeta_{s,j} \right] \\
&= \frac{1}{m^2 R^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{1 \leq t < s \leq R} E\left[ \zeta_{t,i} \zeta_{t,j} \right]^2 + \frac{1}{m^2 R^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{1 \leq t < s \leq R} cov\left[ \zeta_{t,i} \zeta_{t,j}, \zeta_{s,i} \zeta_{s,j} \right] \\
&= MSE_{1,1} + MSE_{1,2},
\end{aligned}
$$

say. Now,

$$
\begin{aligned}
MSE_{1,1} &= \frac{1}{m^2 R^2} \frac{R(R-1)}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} E\left[ \zeta_{t,i} \zeta_{t,j} \right]^2 \leq \frac{1}{2m^2} \left\| E\left[ \zeta_t \zeta_t' \right] \right\|_F^2 \\
&\leq \frac{1}{2m^2} \left( \sqrt{m} \left\| E\left[ (z_s u_s)' \Sigma_{\partial^2 f} (z_t u_t) \right] \right\| \right)^2 \\
&\leq \frac{1}{2m} \left\| \Sigma_{\partial^2 f} V_{zu} \right\|^2 \leq \frac{1}{2m} \left\| \Sigma_{\partial^2 f} \right\|^2 \bar{\lambda} (V_{zu})^2 = o(1),
\end{aligned}
$$

using Assumptions 2(ii) and 4(i) and Lemma 1(d), and

$$
\begin{aligned}
|MSE_{1,2}| &\leq \frac{1}{m^2 R^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{1 \leq t < s \leq R} \left| cov\left[ \zeta_{t,i} \zeta_{t,j}, \zeta_{s,i} \zeta_{s,j} \right] \right| \\
&\leq \frac{8}{m^2 R^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{1 \leq t < s \leq R} \alpha_{s-t}^{1-1/\nu} E\left[ \left| \zeta_{t,i} \zeta_{t,j} \right|^{2\nu} \right]^{1/\nu} \\
&\leq \frac{8}{m^2 R^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{1 \leq t < s \leq R} \alpha_{s-t}^{1-1/\nu} \left( \sum_{k=1}^{m} \sum_{l=1}^{m} \left| \left( \Sigma_{\partial^2 f}^{1/2} \right)_{ik} \left( \Sigma_{\partial^2 f}^{1/2} \right)_{jl} \right| E\left[ u_t^{4\nu} |z_{t,k} z_{t,l}|^{2\nu} \right] \right)^2 \\
&\leq \frac{8}{m^2 R^2} E\left[ u_t^{8\nu} \right]^{1/2\nu} \left( \max_{1 \leq i \leq m} E\left[ z_{i,t}^{8\nu} \right] \right)^{1/2\nu} \left( \sum_{i=1}^{m} \sum_{k=1}^{m} \left| \left( \Sigma_{\partial^2 f}^{1/2} \right)_{ik} \right| \right)^2 \sum_{1 \leq t < s \leq R} \alpha_{s-t}^{1-1/\nu} \\
&\leq \frac{8}{m^2 R^2} E\left[ u_t^{8\nu} \right]^{1/2\nu} \left( \max_{1 \leq i \leq m} E\left[ z_{i,t}^{8\nu} \right] \right)^{1/2\nu} \left( m^{3/2} \left\| \Sigma_{\partial^2 f}^{1/2} \right\| \right)^2 \left( R \sum_{i=1}^{\infty} \alpha_i^{1-1/\nu} \right) \\
&\leq O\left( \frac{m}{R} \right) = o(1),
\end{aligned}
$$

using in addition Assumption 1(i,ii), Lemmas 3 and 5(a), and the Minkowski and triangular inequalities.

The second ingredient of $MSE$ is, using Assumption 1(i,ii) and Lemmas 3, 1(c) and 5(a),

$$
\begin{aligned}
MSE_2 &= \frac{2}{R^2} \sum_{1 \le t_1 < R} \sum_{t_1 < s_1 \le R} \sum_{t_1 < t_2 < s_1} E\left[ (z_{s_1} u_{s_1})' \frac{\Sigma_{\partial^2 f}}{m} (z_{t_1} u_{t_1}) (z_{s_1} u_{s_1})' \frac{\Sigma_{\partial^2 f}}{m} (z_{t_2} u_{t_2}) \right] \\
&\le \frac{2}{m^2 R^2} \sum_{1 \le t_1 < R} \sum_{t_1 < s \le R} \sum_{t_1 < t_2 < s} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} \left| (\Sigma_{\partial^2 f})_{ij} \right| \left| (\Sigma_{\partial^2 f})_{kl} \right| E\left[ \left| z_{i,s} z_{j,t_1} z_{k,s} z_{l,t_2} u_s^2 u_{t_1} u_{t_2} \right| \right] \\
&\le \frac{16}{m^2 R^2} \left( \max_{1 \le i \le m} E\left[ |z_{i,t}|^{8\nu} \right] E\left[ |u_t|^{8\nu} \right] \right)^{1/2\nu} \left( \sum_{i=1}^{m} \sum_{j=1}^{m} \left| (\Sigma_{\partial^2 f})_{ij} \right| \right)^2 \\
&\quad \times \sum_{1 \le t_1 < R} \sum_{t_1 < s \le R} \sum_{t_1 < t_2 < s} \alpha_{\max(t_2 - t_1, s - t_2)}^{1 - 1/\nu} \\
&< \frac{16}{m^2 R^2} C^{1/\nu} \left( m^{3/2} \left\| \Sigma_{\partial^2 f} \right\| \right)^2 \left( 2R \sum_{i=1}^{\infty} i \alpha_i^{1-1/\nu} \right) \\
&\le O\left( \frac{m}{R} \right) = o(1).
\end{aligned}
$$

Summarizing,

$$
\frac{\xi_{zu}' \Sigma_{\partial^2 f} \xi_{zu}}{m} \xrightarrow{p} \psi_1.
$$

(b) Consider the difference

$$
\begin{aligned}
&\left| \frac{\xi_{zu}' \hat{Q}_{zx}'^{-1} \hat{Q}_{\partial^2 f} \hat{Q}_{zx}^{-1} \xi_{zu}}{m} - \frac{\xi_{zu}' Q_{zx}'^{-1} Q_{\partial^2 f} Q_{zx}^{-1} \xi_{zu}}{m} \right| \\
&\le \frac{1}{m} \left\| \hat{Q}_{zx}'^{-1} \hat{Q}_{\partial^2 f} \hat{Q}_{zx}^{-1} - Q_{zx}'^{-1} Q_{\partial^2 f} Q_{zx}^{-1} \right\| \left\| \xi_{zu} \right\|^2 \\
&\le \frac{1}{m} O_p\left( \frac{m}{\sqrt{R}} \right) \cdot O_p(m) \\
&= o_p(1),
\end{aligned}
$$

using Lemma 4(c) and because

$$
\begin{aligned}
&\left\| \hat{Q}_{zx}'^{-1} \hat{Q}_{\partial^2 f} \hat{Q}_{zx}^{-1} - Q_{zx}'^{-1} Q_{\partial^2 f} Q_{zx}^{-1} \right\| \\
&\le \left\| \left( \hat{Q}_{zx}^{-1} - Q_{zx}^{-1} \right)' \hat{Q}_{\partial^2 f} \left( \hat{Q}_{zx}^{-1} + Q_{zx}^{-1} \right) \right\| + \left\| Q_{zx}'^{-1} \left( \hat{Q}_{\partial^2 f} - Q_{\partial^2 f} \right) Q_{zx}^{-1} \right\| \\
&\le \left\| \hat{Q}_{zx}^{-1} - Q_{zx}^{-1} \right\| \left( \left\| \hat{Q}_{\partial^2 f} - Q_{\partial^2 f} \right\| + \left\| Q_{\partial^2 f} \right\| \right) \left( \left\| \hat{Q}_{zx}^{-1} - Q_{zx}^{-1} \right\| + 2 \left\| Q_{zx}^{-1} \right\| \right) \\
&\quad + \left\| Q_{zx}^{-1} \right\|^2 \left\| \hat{Q}_{\partial^2 f} - Q_{\partial^2 f} \right\| \\
&\le O_p\left( m/\sqrt{R} \right) \left( O_p\left( m/\sqrt{P} \right) + C \right) \left( O_p\left( m/\sqrt{R} \right) + 2c^{-1} \right) + c^{-2} O_p\left( m/\sqrt{P} \right) \\
&= O_p\left( \frac{m}{\sqrt{R}} \right).
\end{aligned}
$$

using Lemma 4(a,b).

(c) Using assumption 4(iii),

$$\left\| \hat{Q}^*_{\partial^2 f} - \hat{Q}_{\partial^2 f} \right\| \le \left( \frac{1}{P} \sum_{t=R}^{R+P} d_t \right) \|\beta^* - \beta\| \le \left( \frac{1}{P} \sum_{t=R}^{R+P} d_t \right) O_p\left( \frac{m}{\sqrt{R}} \right) = o_p(1),$$

because

$$\|\beta^* - \beta\| \le \left\| \hat{\beta} - \beta \right\| \le \sqrt{\frac{1}{R}} \left\| \hat{Q}^{-1}_{zx} \right\| \|\xi_{zu}\| \le O_p\left( \sqrt{\frac{m}{R}} \right)$$

by Lemma 4(b,c), hence

$$\left| \frac{\xi'_{zu} \hat{Q}'^{-1}_{zx} \hat{Q}^*_{\partial^2 f} \hat{Q}^{-1}_{zx} \xi_{zu}}{m} - \frac{\xi'_{zu} \hat{Q}'^{-1}_{zx} \hat{Q}_{\partial^2 f} \hat{Q}^{-1}_{zx} \xi_{zu}}{m} \right| \le \frac{1}{m} \|\xi_{zu}\|^2 \left\| \hat{Q}^{-1}_{zx} \right\|^2 \left\| \hat{Q}^*_{\partial^2 f} - \hat{Q}_{\partial^2 f} \right\|$$

$$\le \frac{1}{m} O_p(m) C^2 o_p(1) = o_p(1).$$

∎

**Proof of Theorem 1.** Consider the Taylor expansion up to second order:

$$\sqrt{P}(\bar{\hat{f}} - Ef) = \sqrt{P}(\bar{f} - Ef) + \sqrt{P}\frac{1}{P} \sum_{t=R+1}^{R+P} \frac{\partial f_t}{\partial \beta'}\left( \hat{\beta} - \beta \right) + \frac{\sqrt{P}}{2}\left( \hat{\beta} - \beta \right)' \hat{Q}^*_{\partial^2 f}\left( \hat{\beta} - \beta \right)$$

$$= \sqrt{P}(\bar{f} - Ef) + \sqrt{\frac{1}{R}} \xi'_{\partial f} \hat{Q}^{-1}_{zx} \xi_{zu} + \frac{1}{2}\sqrt{\frac{Pm^2}{R^2}} \frac{\xi'_{zu} \hat{Q}'^{-1}_{zx} \hat{Q}^*_{\partial^2 f} \hat{Q}^{-1}_{zx} \xi_{zu}}{m}.$$

Note first that, using the Cauchy–Schwartz inequality and Lemmas 4(b,c) and 5(b),

$$\left| \xi'_{\partial f} \hat{Q}^{-1}_{zx} \xi_{zu} \right| = \left| \xi'_{\partial f} \hat{Q}^{-1}_{zx} \xi_{zu} \right| \le \left\| \hat{Q}^{-1}_{zx} \right\| \|\xi_{\partial f}\| \|\xi_{zu}\|$$

$$< c^{-1} \|\xi_{\partial f}\| \|\xi_{zu}\|$$

$$\le O_p(m).$$

Now using Lemma 6(a,b,c), rearranging and summarizing,

$$\sqrt{P}(\bar{\hat{f}} - Ef) = \sqrt{P}(\bar{f} - Ef) + \frac{O_p(m)}{\sqrt{R}} + \frac{1}{2}\sqrt{\frac{Pm^2}{R^2}}(\psi_1 + o_p(1)) + \frac{1}{6}\sqrt{\frac{Pm^3}{R^3}} O_p\left( m^{3/2} \right).$$

Provided that $\psi_1 \ne 0$, if $Pm^2/R^2 \to \infty$, the noise dominates the first signal term. If $Pm^2/R^2 \to 0$, all noise terms asymptotically vanishes. If $Pm^2/R^2 \to \mu_1 > 0$,

$$\sqrt{P}(\bar{\hat{f}} - Ef) = \sqrt{P}(\bar{f} - Ef) + \frac{O_p(m)}{\sqrt{R}} + \frac{\sqrt{\mu_1} + o(1)}{2}(\psi_1 + o_p(1))$$

$$+ \frac{\sqrt{\mu_2} + o(1)}{6}\sqrt{\frac{m}{R}} O_p(1)$$

$$= \zeta_f + \frac{\sqrt{\mu_1}}{2}\psi_1 + o_p(1) \xrightarrow{d} \mathcal{N}\left( \frac{\sqrt{\mu_1}}{2}\psi_1, V_f \right).$$

∎

**Proof of Theorem 2.** From the proof of Theorem 1,

$$
\begin{aligned}
t_1 &= \frac{1}{\sqrt{V_f + o_p(1)}} \left( \zeta_f + \frac{\sqrt{\mu_1}}{2} \psi_1 + o_p(1) - \frac{\sqrt{P}m}{2R} \psi_1 \right) \\
&= \frac{\zeta_f}{\sqrt{V_f}} + \frac{\sqrt{\mu_1}}{2\sqrt{V_f}} \psi_1 + o_p(1) - \frac{1}{2} \frac{\sqrt{\mu_1} + o(1)}{\sqrt{V_f + o_p(1)}} \left( \psi_1 + o_p(1) \right) \\
&= \frac{\zeta_f}{\sqrt{V_f}} + o_p(1) \xrightarrow{p} \mathcal{N}(0, 1).
\end{aligned}
$$

# C  Appendix: proofs related to case $Q_{\partial f} \neq 0$

Denote

$$
\tilde{a}' = Q_{\partial f}' Q_{zx}^{-1}.
$$

**Lemma 7** *Under assumptions 2 and 6,*

*(a)* $\|\tilde{a}\| < \sqrt{\mathring{m}} C^2$ *and* $\|\tilde{a}\|_1 < \sqrt{m\mathring{m}} C^2$,

*(b)* $\sum_{i=1}^{m} \sum_{j=1}^{m} \left| \tilde{a}_i \tilde{a}_j \left[ V_{zu} \right]_{ij} \right| < C \|\tilde{a}\|_1^2$.

**Proof.** (a) First, $\|\tilde{a}\| \leq \|Q_{zx}^{-1}\| \|Q_{\partial f}\| < \sqrt{\mathring{m}} C^2$. Second, $\|\tilde{a}\|_1 \leq \sqrt{m} \|\tilde{a}\| < \sqrt{m\mathring{m}} C^2$, where the first inequality is GV, eqn. 2.3.12. (b) $\sum_{i=1}^{m} \sum_{j=1}^{m} \left| \tilde{a}_i \tilde{a}_j \left[ V_{zu} \right]_{ij} \right| \leq \|\tilde{a}\|_1^2 \max_{i,j} \left[ V_{zu} \right]_{ij} \leq \|\tilde{a}\|_1^2 \|V_{zu}\|$, where the second inequality is GV, eqn. 2.3.8. ∎

**Lemma 8** *Under the asymptotics of assumption 7 and conditions of assumptions 2, 3 and 6,*

*(a)*
$$
\frac{Q_{\partial f}' Q_{zx}^{-1} \xi_{zu}}{\sqrt{\mathring{m}}} \xrightarrow{d} \mathcal{N}(0, \psi_2),
$$

*(b)*
$$
\left| \frac{\hat{Q}_{\partial f}' \hat{Q}_{zx}^{-1} \xi_{zu}}{\sqrt{\mathring{m}}} - \frac{Q_{\partial f}' Q_{zx}^{-1} \xi_{zu}}{\sqrt{\mathring{m}}} \right| \leq o_p(1),
$$

*(c)*

$$\left| \hat{Q}_{\partial f}^{*\prime} \hat{Q}_{zx}^{-1} \xi_{zu} - \hat{Q}_{\partial f}^{\prime} \hat{Q}_{zx}^{-1} \xi_{zu} \right| \leq o_p(1).$$

**Proof.** (a) Denote $a' = \tilde{a}'/\sqrt{\mathring{m}R}$ and $x_{Rt} = a' z_t u_t$. Note that $x_{Rt}$ is a martingale difference array (MDS) with variance $\sigma_{Rt}^2 = E[x_{Rt}^2] = Q_{\partial f}' \Sigma_\beta Q_{\partial f}/(\mathring{m}R)$. Consider

$$s_R^2 = \sum_{t=1}^{R} \sigma_{Rt}^2 = \frac{Q_{\partial f}' \Sigma_\beta Q_{\partial f}}{\mathring{m}} \to \psi_2.$$

We will now show that $s_R^{-1} \sum_{t=1}^{R} (x_{Rt}^2 - \sigma_{Rt}^2) \xrightarrow{p} 0$. Indeed,

$$
\begin{aligned}
\sum_{t=1}^{R} (x_{Rt}^2 - \sigma_{Rt}^2) &= \sum_{t=1}^{R} \left( \frac{(\tilde{a}' z_t u_t)^2}{\mathring{m}R} - \frac{Q_{\partial f}' \Sigma_\beta Q_{\partial f}}{\mathring{m}R} \right) + o(1) \\
&= \frac{1}{\mathring{m}R} \tilde{a}' \sum_{t=1}^{R} (z_t z_t' u_t^2 - V_{zu}) \tilde{a} + o(1).
\end{aligned}
$$

The leading term has zero expectation, and the mean squared error

$$
\begin{aligned}
MSE &= \frac{1}{\mathring{m}^2 R^2} E\left[ \sum_{t=1}^{R} \sum_{s=1}^{R} \tilde{a}' (z_t z_t' u_t^2 - V_{zu}) \tilde{a}\tilde{a}' (z_s z_s' u_s^2 - V_{zu}) \tilde{a} \right] \\
&= \frac{1}{\mathring{m}^2 R^2} \sum_{t=1}^{R} E\left[ \left( \tilde{a}' (z_t z_t' u_t^2 - V_{zu}) \tilde{a} \right)^2 \right] \\
&\quad + \frac{2}{\mathring{m}^2 R^2} \sum_{1 \leq t < s \leq R} E\left[ \tilde{a}' (z_t z_t' u_t^2 - V_{zu}) \tilde{a}\tilde{a}' (z_s z_s' u_s^2 - V_{zu}) \tilde{a} \right] \\
&= MSE_1 + 2MSE_2,
\end{aligned}
$$

say. Now,

$$
\begin{aligned}
MSE_1 &= \frac{1}{\mathring{m}^2 R} E\left[ \left( \tilde{a}' (z_t z_t' u_t^2 - V_{zu}) \tilde{a} \right)^2 \right] \\
&= \frac{1}{\mathring{m}^2 R} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} \tilde{a}_i \tilde{a}_j \tilde{a}_k \tilde{a}_l E\left[ \left( z_{t,i} z_{t,j} u_t^2 - [V_{zu}]_{ij} \right) \left( z_{t,k} z_{t,l} u_t^2 - [V_{zu}]_{kl} \right) \right] \\
&\leq \frac{1}{\mathring{m}^2 R} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} |\tilde{a}_i \tilde{a}_j \tilde{a}_k \tilde{a}_l| \left( E\left[ |z_{t,i} z_{t,j} z_{t,k} z_{t,l} u_t^4| \right] + |[V_{zu}]_{ij} [V_{zu}]_{kl}| \right) \\
&\leq \frac{1}{\mathring{m}^2 R} \left( \left( \max_{1 \leq i \leq m} E\left[ |z_{i,t}|^8 \right] \right)^{1/2} E[u_t^8]^{1/2} + \left( \max_{i,j} |[V_{zu}]_{ij}| \right)^2 \right) \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} |\tilde{a}_i \tilde{a}_j \tilde{a}_k \tilde{a}_l| \\
&\leq \frac{1}{\mathring{m}^2 R} \left( \left( \max_{1 \leq i \leq m} E\left[ |z_{i,t}|^8 \right] \right)^{1/2} E[u_t^8]^{1/2} + \|V_{zu}\|^2 \right) \|\tilde{a}\|_1^4 \\
&\leq O\left( \frac{m^2}{R} \right) = o(1),
\end{aligned}
$$

39

using Lemma 7(a) and Assumptions 1(i,ii), 2 and 6(i). The other term in $MSE$ satisfies

$$
\begin{aligned}
|MSE_2| \;&\leq\; \frac{1}{\mathring{m}^2 R^2} \sum_{1 \leq t < s \leq R} \left| E\left[ \tilde{a}' \left( z_t z_t' u_t^2 - V_{zu} \right) \tilde{a} \tilde{a}' \left( z_s z_s' u_s^2 - V_{zu} \right) \tilde{a} \right] \right| \\
&\leq\; \frac{8}{\mathring{m}^2 R^2} E\left[ \left| \tilde{a}' \left( z_t z_t' u_t^2 - V_{zu} \right) \tilde{a} \right|^{2\nu} \right]^{1/\nu} \sum_{1 \leq t < s \leq R} \alpha_{s-t}^{1-1/\nu} \\
&\leq\; \frac{8}{\mathring{m}^2 R^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \tilde{a}_i \tilde{a}_j E\left[ \left| z_{t,i} z_{t,j} u_t^2 - [V_{zu}]_{ij} \right|^{2\nu} \right]^{1/\nu} \left( R \sum_{k=1}^{\infty} \alpha_k^{1-1/\nu} \right) \\
&\leq\; \frac{8}{\mathring{m}^2 R} \left( \sum_{i=1}^{m} \sum_{j=1}^{m} |\tilde{a}_i \tilde{a}_j| \left( E\left[ u_t^{8\nu} \right]^{1/4\nu} \left( \max_{1 \leq i \leq m} E\left[ |z_{i,t}|^{8\nu} \right] \right)^{1/4\nu} + \left| [V_{zu}]_{ij} \right| \right) \right)^2 \sum_{k=1}^{\infty} \alpha_k^{1-1/\nu} \\
&\leq\; \frac{8}{\mathring{m}^2 R} \left( C^{1/2\nu} \|\tilde{a}\|_1^2 + C \|\tilde{a}\|_1^2 \right)^2 \sum_{k=1}^{\infty} \alpha_k^{1-1/\nu} \\
&\leq\; O\left( \frac{m^2}{R} \right) = o(1),
\end{aligned}
$$

using in addition Lemmas 3 and 7(b), and the Minkowski and triangular inequalities.

To summarize, we have proved that $s_R^{-1} \sum_{t=1}^{R} \left( x_{Rt}^2 - \sigma_{Rt}^2 \right) \xrightarrow{p} 0$, and hence the condition (a) of Theorem 12.4.1 of Davidson (2000) for MDS arrays is satisfied. The condition (b)(i) is satisfied by the stationarity assumption 1(i). Indeed, the array $x_{Rt}$ is strictly stationary with respect to $t$ for fixed $R$. In conclusion,

$$
\Xi_{zu} = \frac{Q'_{\partial f} Q_{zx}^{-1} \xi_{zu}}{\sqrt{\mathring{m}}} = s_R \cdot \frac{1}{s_R} \sum_{t=1}^{R} x_{Rt} \xrightarrow{d} \lim_{m \to \infty} s_R \cdot \mathcal{N}(0,1) = \mathcal{N}(0, \psi_2).
$$

(b) Consider the difference

$$
\begin{aligned}
\left| \frac{\hat{Q}'_{\partial f} \hat{Q}_{zx}^{-1} \xi_{zu}}{\sqrt{\mathring{m}}} - \frac{Q'_{\partial f} Q_{zx}^{-1} \xi_{zu}}{\sqrt{\mathring{m}}} \right| \;&\leq\; \frac{1}{\sqrt{\mathring{m}}} \left| \left( \hat{Q}'_{\partial f} \hat{Q}_{zx}^{-1} - Q'_{\partial f} Q_{zx}^{-1} \right) \xi_{zu} \right| \\
&\leq\; \frac{1}{\sqrt{\mathring{m}}} \left\| \hat{Q}'_{\partial f} \hat{Q}_{zx}^{-1} - Q'_{\partial f} Q_{zx}^{-1} \right\| \|\xi_{zu}\| \\
&\leq\; \frac{1}{\sqrt{\mathring{m}}} O_p\left( \sqrt{\frac{m^2 \mathring{m}}{R}} \right) O_p(\sqrt{m}) \\
&=\; o_p(1),
\end{aligned}
$$

40

because

$$\left\|\hat{Q}'_{\partial f}\hat{Q}_{zx}^{-1} - Q'_{\partial f}Q_{zx}^{-1}\right\|$$

$$\leq \left\|\hat{Q}'_{\partial f}\hat{Q}_{zx}^{-1} - \hat{Q}'_{\partial f}Q_{zx}^{-1}\right\| + \left\|\hat{Q}'_{\partial f}Q_{zx}^{-1} - Q'_{\partial f}Q_{zx}^{-1}\right\|$$

$$\leq \left\|\hat{Q}_{zx}^{-1} - Q_{zx}^{-1}\right\|\left(\left\|\hat{Q}_{\partial f} - Q_{\partial f}\right\| + \|Q_{\partial f}\|\right) + \left\|\hat{Q}_{\partial f} - Q_{\partial f}\right\|\|Q_{zx}^{-1}\|$$

$$\leq O_p\left(m/\sqrt{R}\right)\left(O_p\left(m/\sqrt{P}\right) + O\left(\sqrt{\mathring{m}}\right)\right) + O_p\left(\mathring{m}/\sqrt{P}\right)c^{-1}$$

$$= O_p\left(\frac{m\sqrt{\mathring{m}}}{\sqrt{R}} + \frac{\mathring{m}}{\sqrt{P}}\right) = O_p\left(\sqrt{\frac{m^2\mathring{m}}{R}}\right).$$

(c) Using assumption 6(ii),

$$\left\|\hat{Q}^*_{\partial f} - \hat{Q}_{\partial f}\right\| \leq \left(\frac{1}{P}\sum_{t=R}^{R+P}d_t\right)\|\beta^* - \beta\| \leq O_p\left(\frac{m}{\sqrt{R}}\right),$$

because

$$\|\beta^* - \beta\| \leq \left\|\hat{\beta} - \beta\right\| \leq \sqrt{\frac{1}{R}}\left\|\hat{Q}_{zx}^{-1}\right\|\|\xi_{zu}\| \leq O_p\left(\sqrt{\frac{m}{R}}\right)$$

by Lemma 4(b,c), hence

$$\left|\hat{Q}^{*\prime}_{\partial f}\hat{Q}_{zx}^{-1}\xi_{zu} - \hat{Q}'_{\partial f}\hat{Q}_{zx}^{-1}\xi_{zu}\right| \leq \left\|\hat{Q}^*_{\partial f} - \hat{Q}_{\partial f}\right\|\left\|\hat{Q}_{zx}^{-1}\right\|\|\xi_{zu}\|$$

$$\leq O_p\left(\frac{m}{\sqrt{R}}\right)c^{-1}O_p\left(\sqrt{m}\right)$$

$$\leq o_p\left(1\right).$$

∎

**Proof of Theorem 3.** Consider the Taylor expansion up to first order:

$$\left|\hat{Q}^{*\prime}_{\partial f}\hat{Q}_{zx}^{-1}\xi_{zu} - \hat{Q}'_{\partial f}\hat{Q}_{zx}^{-1}\xi_{zu}\right| \leq o_p\left(1\right).$$

$$\sqrt{P}(\bar{\hat{f}} - Ef) = \sqrt{P}\left(\bar{f} - Ef\right) + \sqrt{P}\frac{1}{P}\sum_{t=R}^{R+P}\frac{\partial f_t^*}{\partial\beta'}\left(\hat{\beta} - \beta\right)$$

$$= \sqrt{P}\left(\bar{f} - Ef\right) + \sqrt{P}\frac{1}{P}\sum_{t=R}^{R+P}\frac{\partial f_t}{\partial\beta'}\sqrt{\frac{1}{R}}\hat{Q}_{zx}^{-1}\xi_{zu} + \sqrt{\frac{P}{R}}\left(\hat{Q}^{*\prime}_{\partial f}\hat{Q}_{zx}^{-1}\xi_{zu} - \hat{Q}'_{\partial f}\hat{Q}_{zx}^{-1}\xi_{zu}\right).$$

Using Lemma 8(b,c), rearranging and summarizing,

$$\sqrt{P}(\bar{\hat{f}} - Ef) = \sqrt{P}\left(\bar{f} - Ef\right) + \sqrt{\frac{P\mathring{m}}{R}}\left(\frac{Q'_{\partial f}Q_{zx}^{-1}\xi_{zu}}{\sqrt{\mathring{m}}} + o_p(1)\right) + \sqrt{\frac{P}{R}}o_p(1).$$

41

Provided that $\psi_2 \neq 0$ and $m^3/R \to 0$, if $P\mathring{m}/R \to \infty$, the noise dominates the first signal term. If $P\mathring{m}/R \to 0$, all noise terms asymptotically vanish. If $P\mathring{m}/R \to \mu_2 > 0$,

$$
\begin{aligned}
\sqrt{P}(\bar{\hat{f}} - Ef) &= \sqrt{P}\left(\bar{f} - Ef\right) + \left(\sqrt{\mu_2} + o(1)\right)\left(\frac{Q'_{\partial f}Q_{zx}^{-1}\xi_{zu}}{\sqrt{\mathring{m}}} + o_p(1)\right) + o_p(1) \\
&= \zeta_f + \sqrt{\mu_2}\frac{Q'_{\partial f}Q_{zx}^{-1}\zeta_{zu}}{\sqrt{\mathring{m}}} + o_p(1) \\
&\xrightarrow{d} \mathcal{N}\left(0, V_f + \mu_2\psi_2\right).
\end{aligned}
$$

■