In-Sample and Out-of-Sample Fit: Their Joint Distribution and its Implications for Model Selection and Model Averaging
(Criteria-Based Shrinkage for Forecasting)

Peter Reinhard Hansen*
Stanford University
Department of Economics
Stanford, CA 94305
Email: peter.hansen@stanford.edu

Preliminary version: November 25, 2007
Prepared for the
5th ECB Workshop on Forecasting Techniques

Abstract

We consider the case where a parameter, θ, is estimated by maximizing a criterion function, $Q(\mathcal{X}, \theta)$. The estimate is then used to evaluate the criterion function with the same data, $\mathcal{X}$, as well as with an independent data set, $\mathcal{Y}$. The *in-sample fit* and *out-of-sample fit* relative to that of θ_0, the "true" parameter, are given by $T_{x,x} = Q(\mathcal{X}, \hat{\theta}_x) - Q(\mathcal{X}, \theta_0)$ and $T_{y,x} = Q(\mathcal{Y}, \hat{\theta}_x) - Q(\mathcal{Y}, \theta_0)$. We derive the limit distribution of $(T_{x,x}, T_{y,x})$ for a large class of criterion functions and show that $T_{x,x}$ and $T_{y,x}$ are strongly negatively related. The implication is that good in-sample fit translates directly into poor out-of-sample fit. This result forms the basis for a unified framework for discussing aspect of model selection, model averaging, and the effects of data mining. The limit distribution can also be used to motivate a particular form of shrinkage, called *qrinkage*, where in-sample parameter estimates are modified to off-set the overfit of the criterion function, hence the name. This form of shrinkage is particularly simple in the context of regression models, such as the factor-based forecasting models.

Keywords: Qrinkage, Out-of-Sample Likelihood, Model Selection, Model Averaging, Data Mining, Forecasting.

*I thank Jan Magnus, Mark Watson, Kenneth West, and participants at the 2006 Stanford Institute for Theoretical Economics workshop on *Economic Forecasting under Uncertainty*, for valuable comments. The author is also affiliated with CREATES at the University of Aarhus, a research center funded by Danish National Research Foundation.

1 Introduction

Much of applied econometrics is motivated by some form of out-of-sample use. An obvious example is the forecasting problem, where a model is estimated with in-sample data, while the objective is to construct a good out-of-sample forecast. The out-of-sample motivation is intrinsic to many other problems. For example, when a sample is analyzed in order to make inference about aspects of a general population, the objective is to get a good model for the general population, not a model that necessarily explains all the variation in the sample. In this case one may view the general population (less the sample used in the empirical analysis) as the "out-of-sample".

The main contribution of this paper is the result established in Theorem 1, which reveals a strong connection between the in-sample fit and the out-of-sample fit of a model, in a general framework. The result has important implications for model selection by information criteria, because these are shown to have some rather unfortunate and paradoxical properties. The result also provides important insight about model averaging and shrinkage methods. Furthermore, the result provides a theoretical foundation for the use of out-of-sample analysis.

It is well known that as more complexity is added to a model the better will the model fit the data in-sample, while the contrary tends to be true out-of-sample. See, e.g. Chatfield (1995). This aspect is evident from the following example, which serves to illustrated some of the results in this paper.

Consider the regression model, $y_t = x_t^{\mathrm{T}}\beta_0 + \varepsilon_t$, where $\varepsilon_t \sim iid\,\mathrm{N}(0,1)$. The sample $\{y_t,x_t\}_{t=1}^n$ is available for inference about $\beta_0 \in \mathbb{R}^k$ while our true objective concerns $\{y_t,x_t\}_{t=n+1}^{2n}$. We shall refer to the two periods as the *in-sample* and *out-of-sample* periods, respectively, and we use the notation $\mathcal{X} = \{y_t,x_t\}_{t=1}^n$ and $\mathcal{Y} = \{y_t,x_t\}_{t=n+1}^{2n}$. To make the in-sample and out-of-sample regressors comparable, we assume that $\sum_{t=1}^n x_t x_t^{\mathrm{T}} = \sum_{t=n+1}^{2n} x_t x_t^{\mathrm{T}}$.

Suppose that our objective is to minimize the *out-of-sample* expected mean-squared error, or equivalently maximize

$$Q(\beta) = \mathrm{E}\left\{Q(\mathcal{Y},\beta)\right\} = \mathrm{E}\left\{-\sum_{t=n+1}^{2n}(y_t - x_t^{\mathrm{T}}\beta)^2\right\}.$$

It can be verified that β_0 is the solution to this problem. Since β_0 is unknown to us, we must pick a value for β based on the available information. One possibility is to choose the

β that maximizes

$$Q(\mathcal{X}, \beta) = -\sum_{t=1}^{n} (y_t - \beta^{\mathrm{T}} x_t)^2.$$

The solution is the well known least squares estimator, $\hat{\beta}_x = \sum_{t=1}^{n} y_t x_t^{\mathrm{T}} / \sum_{t=1}^{n} x_t x_t^{\mathrm{T}}$, which is also the maximum likelihood estimator in this setting.

In the present situation it is well known that $T_{x,x} = Q(\mathcal{X}, \hat{\beta}_x) - Q(\mathcal{X}, \beta_0) \sim \chi^2_{(k)}$. The fact that $Q(\mathcal{X}, \hat{\beta}_x) > Q(\mathcal{X}, \beta_0)$ (almost surely) is called overfitting, and the expected overfit is here $\mathrm{E}(T_{x,x}) = k$. The converse is true out-of-sample, because $T_{y,x} = Q(\mathcal{Y}, \hat{\beta}_x) - Q(\mathcal{Y}, \beta_0)$ has a negative expected value, specifically $\mathrm{E}(T_{y,x}) = -k$. This merely confirms the well known result that overparameterized models tend to do poorly out-of-sample, despite good in-sample fit. This can motivate the use of information criteria, such as AIC and BIC that explicitly make a trade-off between the complexity of a model and how well the model fits the data.

Our theoretical result provides additional insight and reveals a stronger connection between the in-sample fit and out-of-sample fit. One implication of our analysis is that

$$\mathrm{E}\left[Q(\mathcal{Y}, \hat{\beta}_x) - Q(\mathcal{Y}, \beta_0) | \mathcal{X}\right] = -\left[Q(\mathcal{X}, \hat{\beta}_x) - Q(\mathcal{X}, \beta_0)\right],$$

which shows that more (in-sample) overfitting results in a lower expected fit out-of-sample. This observation is important for model selection and model averaging.

In this paper we derive the (joint) limit distribution of $(T_{x,x}, T_{x,y})$ for a general class of criteria functions, which includes loss functions that are commonly used for the evaluation of forecasts. The limit distribution for the out-of-sample quantity, $T_{y,x}$ has features that are similar to those seen in quasi maximum likelihood analysis, see e.g. White (1994). The limit distribution is particularly simple when an information-matrix style equality holds. This inequality holds when the criterion function is a correctly specified likelihood function. In this case we have that $(T_{x,x}, T_{x,y}) \xrightarrow{d} (Z_1^{\mathrm{T}} Z_1, -Z_1^{\mathrm{T}} Z_1 + 2 Z_1^{\mathrm{T}} Z_2)$, where Z_1 and Z_2 are independent Gaussian distributed random variables, $Z_1, Z_2 \sim \mathrm{N}_k(0, I_k)$. Thus the out-of-sample quantity, $T_{y,x}$, does not have a limit distribution that is simply (minus one times) a $\chi^2_{(k)}$. The additional term appears because $\hat{\beta}_x$ does not maximize $Q(\mathcal{Y}, \beta)$.

Comments out theoretical results:

- An interesting special case is that where the criterion function is the log-likelihood function. Our result provide the limit distribution of the out-of-sample likelihood

ratio statistic, $\mathrm{LR}_{y,x} = 2\left\{\log L(\mathcal{Y}, \hat{\theta}_x) - \log L(\mathcal{Y}, \theta_0)\right\}$. In fact we establish the joint distribution of $(\mathrm{LR}_{y,x}, \mathrm{LR}_{x,x})$, where $\mathrm{LR}_{x,x}$ is the conventional (in-sample) likelihood ratio statistic, $\mathrm{LR}_{x,x} = 2\left\{\log L(\mathcal{X}, \hat{\theta}_x) - \log L(\mathcal{X}, \theta_0)\right\}$.

- An implication of our result is that one is less likely to produce spurious results out-of-sample than in-sample. The reason is that an over-parameterized model tends to do worse than a parsimonious (but correct) model out-of-sample. It will take a lot of luck for an overparameterized model to offset its disadvantage in an out-of-sample comparison with the simpler model. Thus when a complex model outperforms a simpler model out-of-sample it is stronger evidence in favor of the larger model, than had the outperformance been found in-sample (other things being equal).

- A useful decomposition for discussing model selection and model averaging.

- Model Selection: Finding the best model is obscured by sampling and estimation error, as the noise conceals the true ranking of models. Based on our theoretical result we will argue that standard model selection criteria are poorly suited for the problem of selecting a model with a good out-of-sample fit, this is particularly the case in model-rich environments. Shrinkage methods or model averaging are more promising avenues for dealing with this issue.

- Model Averaging: We shall discuss model averaging based on our theoretical results.

- Our theoretical result provides a deep understanding of the observations made in Clark and West (2007). They consider the situation with two regression models – one being nested in the other – where the parameters are estimated by least squares and the mean squared (prediction) error is used as criterion function. The observation made in Clark and West (2007) is that MSPE is *expected* to be smaller for parsimonious models. This motivates a correction of a particular test. Our results reveals that source of the smaller expected MSPE, is the close connection between estimation error and out-of-sample MSPE. Furthermore, we show that this aspect of estimation and out-of-sample prediction holds in a rather general framework.

- Estimating the expected overfit by subsampling, bootstrapping, or the jackknife. The latter has been used in this context by Hansen and Racine (2007).

- This result motivates a particular form of shrinkage, called qrinkage. Qrinkage is particularly simple to apply in regression models, and shrinkage arguments may explain some of the empirical success of the principal component-based forecasts. Several forms of shrinkage have been proposed in the literature, see Hastie, Tibshirani, and Friedman (2001) for an introduction to a large number of shrinkage methods.

While parameter instability is an important issue for forecasting, this is not the focus of this paper. Though we shall comment on this issue where appropriate. Forecasting in an environment with non-constant parameters is an active field of research, see e.g. Hendry and Clements (2002), Pesaran and Timmermann (2005), and Rossi and Giacomini (2006).

Much caution is warranted when asserting the merits of a particular model, based on an out-of-sample comparison. Estimation error may entirely explain the out-of-sample outcome. This is particular relevant if one suspects that parameters are poorly estimated. Thus critiquing a model could backfire by directing attention to the econometrician having estimated the parameters poorly, e.g. by using a relatively short estimation period, or an estimation method that does not maximize the appropriate criterion function. These aspects are worth having in mind, when more sophisticated models are compared to a simple parsimonious benchmark model, as is the case in Meese and Rogoff (1983) and Atkeson and Ohanian (2001).

2 Theoretical Results

We consider a situation where the criterion function and estimation problem can be expressed within the framework of extremum estimators/M-estimators, see Huber (1981). In our exposition we will adopt the framework of Amemiya (1985).

The objective is given in terms of a non-stochastic criterion function $Q(\theta)$, which attains a unique global maximum, $\theta_0 = \arg\max_{\theta\in\Theta} Q(\theta)$. We will refer to θ_0 as the *true* parameter value. The empirical version of the problem is based on a random criterion $Q(\mathcal{X}, \theta)$, where $\mathcal{X} = (X_1, \ldots, X_n)$ is the sample used for the estimation.

To take an example, the criterion function may be the mean squared error, $Q(\mu) = -\mathrm{E}(X_t - \mu)^2$ with the empirical criterion function given by $Q(\mathcal{X}, \mu) = -\sum_{t=1}^{n}(X_t - \mu)^2$.

The extremum estimator is defined by

$$\hat{\theta}_x = \arg\max_{\theta\in\Theta} Q(\mathcal{X}, \theta).$$

We adopt the following standard assumptions from the theory on extremum estimators, see e.g. Amemiya (1985).

Assumption 1 *$\bar{Q}(\mathcal{X},\theta) = n^{-1}Q(\mathcal{X},\theta) \xrightarrow{p} Q(\theta)$ uniformly in θ on a open neighborhood of θ_0, as $n \to \infty$.*

$Q''(\mathcal{X},\theta) = \partial^2 Q(\mathcal{X},\theta)/\partial\theta\partial\theta^{\mathrm{T}}$ exists and is continuous in an open neighborhood of θ_0,

$\bar{Q}''(\mathcal{X},\theta) \xrightarrow{p} -\mathcal{I}(\theta)$ uniformly in θ in an open neighborhood of θ_0,

$\mathcal{I}(\theta)$ is continuous in a neighborhood of θ_0 and $\mathcal{I}_0 = \mathcal{I}(\theta_0) \in \mathbb{R}^{k\times k}$ is negative definite.

$n^{-1/2}Q'(\mathcal{X},\theta_0) \xrightarrow{d} \mathrm{N}\{0,\mathcal{J}_0\}$, where $\mathcal{J}_0 = \lim_{n\to\infty} \mathrm{E}\left\{n^{-1}Q'(\mathcal{X},\theta_0)Q'(\mathcal{X},\theta_0)^{\mathrm{T}}\right\}$.

Assumption 1 guarantees that $\hat{\theta}_x$ (eventually) will be given by the first order condition $Q'(\mathcal{X},\hat{\theta}_x) = 0$. In what follows, we assume that n is sufficiently large that this is indeed the case.[1] The assumptions are stronger than necessary. The differentiability (both first and second) can be dispensed with and replaced with weaker assumptions, e.g. by adopting the setup in Hong and Preston (2006).

We have in mind a situation where the estimate, $\hat{\theta}_x$, is to be computed from n observations, $\mathcal{X} = (X_1,\ldots,X_n)$, however the object of interest is tied to $Q(\mathcal{Y},\hat{\theta}_x)$, where $\mathcal{Y} = (Y_1,\ldots,Y_m)$ denotes m observations that are drawn from the same distribution as that of X. In the context of forecasting, $\mathcal{Y}$ will represent the data from the out-of-sample period, say the last m observations as illustrated below.

$$\underbrace{X_1,\ldots,X_n}_{=\mathcal{X}},\underbrace{X_{n+1},\ldots,X_{n+m}}_{=\mathcal{Y}}.$$

We are particularly interested in the two quantities

$$T_{x,x} = Q(\mathcal{X},\hat{\theta}_x) - Q(\mathcal{X},\theta_0), \quad \text{and} \quad T_{y,x} = Q(\mathcal{Y},\hat{\theta}_x) - Q(\mathcal{Y},\theta_0).$$

The first quantity, $T_{x,x}$, is a measure of in-sample "fit". We have $Q(\mathcal{X},\hat{\theta}_x) \geq Q(\mathcal{X},\theta_0)$, because $\hat{\theta}_x$ maximizes $Q(\mathcal{X},\theta)$. In this sense, $Q(\mathcal{X},\hat{\theta}_x)$ will reflect a value that is too good relative to that of the true parameter $Q(\mathcal{X},\theta_0)$, hence the notion of overfitting. The second quantity, $T_{y,x}$, is a measure of out-of-sample fit. Unlike the in-sample statistic, there is no guarantee that $T_{y,x}$ is non-negative. In fact, because θ_0 is the best ex-ante value for θ, the out-of-sample measure, $T_{y,x}$, will tend to be negative.

[1] When there are multiple solutions to the FOC, one can simply choose the one that yields the largest value of the criterion function, that is $\hat{\theta}_x = \arg\max_{\theta\in\{\theta:Q'(\mathcal{X},\theta)=0\}} Q(\mathcal{X},\theta)$.

Note that we consider the natural situation where θ is estimated by maximizing the criterion function in-sample, $Q(\mathcal{X}, \cdot)$, and the very same criterion function is the one used for the out-of-sample evaluation, $Q(\mathcal{Y}, \cdot)$.

We have the following result concerning the limit distribution of $(T_{x,x}, T_{y,x})$.

Theorem 1 *Given Assumption 1. $\theta \in \mathbb{R}^k$. Suppose $\frac{m}{n} \to \pi$. Then*

$$2 \begin{pmatrix} T_{x,x} \\ T_{x,y} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \zeta_1 \\ 2\sqrt{\pi}\zeta_2 - \pi\zeta_1 \end{pmatrix}, \qquad \text{as } n \to \infty,$$

where $\zeta_1 = Z_1^{\mathrm{T}} \Lambda Z_1$, $\zeta_2 = Z_1^{\mathrm{T}} \Lambda Z_2$ and Z_1 and Z_2 are independent Gaussian random variables $Z_i \sim \mathrm{N}_k(0, I_k)$, and $\Lambda = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_k\right)$, $\lambda_1, \ldots, \lambda_k$ being the eigenvalues of $[\mathcal{I}_0^{-1}\mathcal{J}_0]$.

Remark. Too good in-sample fit (overfit), $T_{x,x,} \gg 0$, translates into mediocre out-of-sample fit. This aspect is particularly important when multiple models are compared in-sample for the purpose of selecting a model to be used out-of-sample, because

$$Q(\mathcal{X}, \hat{\theta}_x^{(j)}) = Q(\mathcal{X}, \theta_0^{(j)}) + Q(\mathcal{X}, \hat{\theta}_x^{(j)}) - Q(\mathcal{X}, \theta_0^{(j)}),$$

and the more models that are being compared with approximately the same $Q(\mathcal{X}, \theta_0^{(j)})$, the more likely it is that the best in-sample performance, as defined by $\max_j Q(\mathcal{X}, \hat{\theta}_x^{(j)})$, is attained by a model with a large $T_{x,x}^{(j)}$, hence a poor out-of-sample fit.

[Selecting the model with the best in-sample fit for the purpose of out-of-sample forecasting, is an *act of hubris*... the (large) value of $T_{x,x}^{(j)}$ its nemesis.]

Remark. The result offers insight about the merits of model averaging, as we shall discuss in the next section.

The theoretical result formulated in Theorem 1 relates the estimated model to that of the model using population values for the parameters. The implications for comparing two arbitrary models, nested or non-nested, is straight forward as will be evident from our analysis in the next Section.

Next we consider the special case where the criterion function is a correctly specified log-likelihood function.

2.1 Out-Of-Sample Likelihood Analysis

A special case is that where the criterion function is given in the form of the likelihood function.

When k parameters are estimated and evaluated using the same data, it is well known that the log-likelihood function, $\ell(\mathcal{X}, \hat{\theta}_x)$ is expected to be about $k/2$ units better than the log-likelihood function evaluated at the true parameters, $\ell(\mathcal{X}, \theta_0)$. In this setting we used $\hat{\theta}_x = \hat{\theta}(\mathcal{X})$ to denote the maximum likelihood estimator. The $k/2$ follows from the fact that the likelihood ratio statistic, $\mathrm{LR}_{x,x} = 2\{\ell(\mathcal{X}, \hat{\theta}_x) - \ell(\mathcal{X}, \theta_0)\}$ is asymptotically distributed as a χ^2 with k degrees of freedom (in regular problems).

It is less known that the converse is true when the log-likelihood function is evaluate out-of-sample. In fact, the asymptotic distribution of $\mathrm{LR}_{y,x} = 2\{\ell(\mathcal{Y}, \hat{\theta}_x) - \ell(\mathcal{Y}, \theta_0)\}$ has expected value $-k$, if $\mathcal{X}$ and $\mathcal{Y}$ are independent and identically distributed. Again we see how expected in-sample overfit translates into expected out-of-sample underfit. The out-of-sample log-likelihood function, $\ell(\mathcal{Y}, \hat{\theta}_x)$, is related to the predictive likelihood introduced by Lauritzen (1974). We could call $\ell(\mathcal{Y}, \hat{\theta}_x)$ the *plug-in predictive likelihood*. Due to overfitting, the plug-in predictive likelihood need not produce an accurate estimate of the distribution of $\mathcal{Y}$, which is typically the objective in the literature on predictive likelihood, see Bjørnstad (1990) for a review.

As we have seen in the general formulation of this problem, $\mathrm{LR}_{x,x}$ and $\mathrm{LR}_{y,x}$ are closely related, and more so than having opposite expected values. Not surprisingly, will we see that $\mathrm{LR}_{x,x} = Z_1^{\mathrm{T}} Z_1 + o_p(1)$ while $\mathrm{LR}_{y,x} = -Z_1^{\mathrm{T}} Z_1 + 2Z_1^{\mathrm{T}} Z_2 + o_p(1)$, where Z_1 and Z_2 are two independent random variables, $Z_i \sim \mathrm{N}_k(0, I_k)$, $i = 1, 2$. So the (random) in-sample overfit, $Z_1^{\mathrm{T}} Z_1$, translates directly into an out-of-sample underfit, $-Z_1^{\mathrm{T}} Z_1$.

To make this result precise. Let $\{X_i\}$, be a sequence of iid random variables in $\mathbb{R}^p$ with density $g(x)$, and suppose that

$$g(x) = f_{\theta_0}(x), \quad \text{for some } \theta_0 \in \Theta \subset \mathbb{R}^k, \tag{1}$$

so that the model is correctly specified model. The in-sample and out-of-sample log-likelihood functions are given by

$$\ell(\mathcal{X}, \theta) \equiv \sum_{i=1}^{n} \log f(X_i; \theta), \quad \text{and} \quad \ell(\mathcal{Y}, \theta) \equiv \sum_{i=n+1}^{n+m} \log f(X_i; \theta).$$

The in-sample maximum likelihood estimator, $\hat{\theta}_x = \arg\max_\theta \ell(\mathcal{X}, \theta)$, is given by $\ell'(\mathcal{X}, \hat{\theta}_x) = 0$.

Theorem 2 *Assume that $\ell(\mathcal{X}, \theta)$ satisfies Assumption 1, and that $\ell(\mathcal{X}, \cdot)$ is correctly specified as formulated in (1). Then the information matrix equality holds, $\mathcal{I}_0 = \mathcal{J}_0$, and the*

in-sample and out-of-sample likelihood ratio statistics,

$$\mathrm{LR}_{x,x} \equiv 2\{\ell(\mathcal{X}, \hat{\theta}_x) - \ell(\mathcal{X}, \theta_0)\} \quad \textit{and} \quad \mathrm{LR}_{y,x} \equiv 2\{\ell(\mathcal{Y}, \hat{\theta}_x) - \ell(\mathcal{Y}, \theta_0)\},$$

are such that (with $\pi = \lim_{n\to\infty} \frac{m}{n}$*)*

$$\begin{pmatrix} \mathrm{LR}_{x,x} \\ \mathrm{LR}_{y,x} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \zeta_1 \\ 2\sqrt{\pi}\zeta_2 - \pi\zeta_1 \end{pmatrix}, \qquad \textit{as } n \to \infty,$$

where $\zeta_1 = Z_1^{\mathrm{T}} Z_1$*,* $\zeta_2 = Z_1^{\mathrm{T}} Z_2$ *and* Z_1 *and* Z_2 *are independent Gaussian random variables* $Z_i \sim \mathrm{N}_k(0, I_k)$.

When $n = m$ we see that the limit distribution of (two times) the in-sample log-likelihood and the out-of-sample log-likelihood, $2\{\ell(\mathcal{X}, \hat{\theta}_x) - \ell(\mathcal{Y}, \hat{\theta}_x)\} = \mathrm{LR}_{x,x} - \mathrm{LR}_{y,x}$, has the expected value,

$$\mathrm{E}\{\zeta_1 - (2\zeta_2 - \zeta_1)\} = \mathrm{E}\{2\zeta_1\} = 2k.$$

This expectation can be used to motivate the Akaike's information criterion (AIC), see Akaike (1974). AIC assumes that the likelihood function is correctly specified. The proper penalty to use for misspecified models was derived by Takeuchi (1976) (QMLE results).

The additional insight provided by Theorem 2, is that whenever a model fits the in-sample data abnormally well, this will result in a meager value of the out-of-sample log-likelihood, due to the term, ζ_1, with opposite signs in the limit distribution. This offers a theoretical explanation for the *AIC paradox* in a very general setting. Shimizu (1978) analyzed the problem of selecting the order of an autoregressive process, and found that in-sample fit was strongly negatively related to out-of-sample fit (here expressed in our terminology).

The classical result, $\mathrm{LR}_{x,x} \xrightarrow{d} \chi^2(k)$, is a special case of Theorem 2, so the interesting part of the Theorem is the result for the out-of-sample likelihood ratio. Given the our results in Theorem 1, we are not surprised to find that $\mathrm{LR}_{y,x}$ has a negative expected value and is closely tied to the usual in-sample log-likelihood ratio, $\mathrm{LR}_{x,x}$, as ζ_1 appears in both expressions.

Corollary 3 *When the in-sample and out-of-sample size is the same,* $m = n$*, we have*

$$\begin{aligned} \mathrm{E}(\zeta_1) &= +k, & \mathrm{var}(\zeta_1) &= k^2 + 2k, \\ \mathrm{E}(2\zeta_2 - \zeta_1) &= -k, & \mathrm{var}(2\zeta_2 - \zeta_1) &= k^2 + 6k. \end{aligned}$$

Next, we look at the results of Theorem 2 in the context of a linear regression model.

Example 1 *Consider the linear regression model,*

$$Y = X\beta + u.$$

To avoid notational confusion, we will use subscripts, 1 *and* 2*, to represent the in-sample and out-of-sample periods, respectively. In sample we have* $Y_1, u_1 \in \mathbb{R}^n$*,* $X_1 \in \mathbb{R}^{n\times k}$*, and* $u_1|X_1 \sim iid\, \mathrm{N}_n(0, \sigma^2 I_n)$*, and the well known result for the the sum-of-squared residuals,*

$$\begin{aligned}\hat{u}_1^{\mathrm{T}}\hat{u}_1 &= Y_1^{\mathrm{T}}Y_1 - \hat{\beta}_1^{\mathrm{T}}X_1^{\mathrm{T}}Y_1 - Y_1^{\mathrm{T}}X_1\hat{\beta}_1 + \hat{\beta}_1^{\mathrm{T}}X_1^{\mathrm{T}}X_1\hat{\beta}_1 \\ &= Y_1^{\mathrm{T}}(I - P_{X_1})Y_1 = u_1^{\mathrm{T}}(I - P_{X_1})u_1,\end{aligned}$$

where we have introduced the notation $P_{X_1} = X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}$*, and we find*

$$2\left\{\ell_1(\hat{\beta}_1) - \ell_1(\beta_0)\right\} = -\hat{u}_1^{\mathrm{T}}\hat{u}_1/\sigma^2 + u_1^{\mathrm{T}}u_1/\sigma^2 = u_1^{\mathrm{T}}P_{X_1}u_1/\sigma^2 \sim \chi^2_{(k)}.$$

Similarly, out-of-sample we have

$$\begin{aligned}\hat{u}_2^{\mathrm{T}}\hat{u}_2 &= Y_2^{\mathrm{T}}Y_2 - 2\hat{\beta}_1^{\mathrm{T}}X_2^{\mathrm{T}}Y_2 + \hat{\beta}_1^{\mathrm{T}}X_2^{\mathrm{T}}X_2\hat{\beta}_1 \\ &= Y_2^{\mathrm{T}}Y_2 - 2Y_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}Y_2 + Y_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}X_2(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}Y_1 \\ &= u_2^{\mathrm{T}}u_2 - 2u_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}u_2 + u_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}X_2(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}})u_1 \\ &\quad +\beta_0^{\mathrm{T}}X_2^{\mathrm{T}}X_2\beta_0 - 2\beta_0^{\mathrm{T}}X_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}X_2\beta_0 + \beta_0^{\mathrm{T}}X_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}X_2(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}X_1\beta_0 \\ &\quad +u_1^{\mathrm{T}}(-2X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}X_2 + 2X_1(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}X_2)\beta_0 + u_2^{\mathrm{T}}(2X_2 - 2X_2X_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1})\beta_0,\end{aligned}$$

where the last two terms are both zero. If we define $W = \frac{n}{m}(X_1^{\mathrm{T}}X_1)^{-1}X_2^{\mathrm{T}}X_2 \xrightarrow{p} I$*, we find*

$$\begin{aligned}2\sigma^2\left\{\ell_2(\hat{\beta}_2) - \ell_2(\beta_0)\right\} &= u_2^{\mathrm{T}}u_2 - \hat{u}_2^{\mathrm{T}}\hat{u}_2 \\ &= 2u_1^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1/2}\sqrt{\frac{m}{n}}W^{1/2}(X_2^{\mathrm{T}}X_2)^{-1/2}X_2^{\mathrm{T}}u_2 + u_1^{\mathrm{T}}X_1\frac{m}{n}W(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}})u_1 \\ &= \sigma^2\left\{\sqrt{\tfrac{m}{n}}2Z_1^{\mathrm{T}}Z_2 - \tfrac{m}{n}Z_1^{\mathrm{T}}Z_1\right\} + o_p(1)\end{aligned}$$

where we defined $Z_1 = \sigma^{-1}(X_1^{\mathrm{T}}X_1)^{-1/2}X_1^{\mathrm{T}}u_1$ *and* $Z_2 = \sigma^{-1}(X_2^{\mathrm{T}}X_2)^{-1/2}X_2^{\mathrm{T}}u_2$ *so that* $u_1^{\mathrm{T}}P_{X_1}u_1\sigma^2 Z_1^{\mathrm{T}}Z_1$*, since* Z_1 *and* Z_2 *are independent and both distributed as* $\mathrm{N}_k(0, I)$*, and the structure of Theorems 1 and 2 emerges.*

2.2 Extensions

Out-of-sample forecast evaluation has been analyzed with different estimation schemes, known as the *fixed*, *rolling*, and *recursive* schemes[REF: McCracken...]. Under the fixed scheme the parameters are estimated once and the same point estimate is used for the entire out-of-sample period. In the rolling and recursive schemes the parameter is reestimated every time a forecast is made. The recursive scheme use all past observations for the estimation, whereas the rolling scheme only use a limited number of the most recent observations. The number of observations used for the estimation with the rolling scheme is typically constant, but one can also use a random number of observations, defined by some stationary data dependent process, see e.g. Giacomini and White (2006).

The results presented in Theorem 1 are based on the fixed scheme, but can be adapted to forecast comparisons using the rolling and recursive schemes. Still, Theorem 1 speaks to the general situation where a forecast is based on estimated parameters, and have implications for model selection and model averaging as we discuss in the next section.

For example under the recursive schemes, the expected out-of-sample underfit for a correctly specified model is approximate

$$\begin{aligned} k\sum_{i=1}^{m}\frac{1}{n+i} &= k\frac{1}{m+n}\sum_{s=n+1}^{m+n}\frac{m+n}{s} \\ &\approx k\int_{\frac{1}{1+\pi}}^{1}\frac{1}{u}\mathrm{d}u \to k\int_{\frac{1}{1+\pi}}^{1}\frac{1}{u}\mathrm{d}u = k\log(1+\pi) < k, \end{aligned}$$

where $\pi = \lim \frac{m}{n}$, which is consistent with McCracken (200x), who established this result in the context of regression models.

[ADD ADDITIONAL DETAILS ON ROLLING/RECURSIVE]

3 Implications

We now turn to a situation where we estimate more that a single model.

Consider M different specifications (models) that each have their own "true" parameter value, denoted by $\theta_0^{(j)}$. It is useful to think of the different models as restricted version or a larger nesting model, $\theta \in \Theta$. The jth model is now characterized by $\theta \in \Theta^{(j)} \subset \Theta$, and its true value is $\theta_0^{(j)} = \arg\max_{\theta\in\Theta^{(j)}} Q(\theta)$. We shall assume that Assumption 1 applies to all models, so that $\hat{\theta}_x^{(j)} \xrightarrow{p} \theta_0^{(j)}$, where $\hat{\theta}_x^{(j)} = \arg\max_{\theta\in\Theta^{(j)}} Q(\mathcal{X},\theta)$. So $\theta^{(j)}$ reflects the best

possible ex-ante value for θ. The nesting model need not be interesting as a model per se. In many situation this model will be so heavily parameterized that it would make little sense to estimate it directly.

When we evaluate the in-sample fit of a model, a relevant question is whether a large value of $Q(\mathcal{X}, \hat{\theta}_x^{(j)})$ reflects genuine superior performance or is due to sampling variation. The following decomposition shows that the sampling variation comes in two flavors, one of them being particularly nasty. The in-sample fit can be decomposition as follows:

$$Q(\mathcal{X}, \hat{\theta}_x^{(j)}) = \underbrace{Q(\theta_0^{(j)})}_{\text{Genuine}} + \underbrace{Q(\mathcal{X}, \theta_0^{(j)}) - Q(\theta_0^{(j)})}_{\text{White noise}} + \underbrace{Q(\mathcal{X}, \hat{\theta}_x^{(j)}) - Q(\mathcal{X}, \theta_0^{(j)})}_{\text{Deceptive noise}}. \tag{2}$$

We have labelled the two random terms as *white noise* and *deceptive noise*, respectively. The first component reflects the best possible value for this model, that would be realized if one knew the true value, $\theta_0^{(j)}$. The second term is pure sampling error that is unaffected by our choice for $\hat{\theta}$, so this term simply induces a layer of noise that makes it harder to infer $Q(\theta_0^{(j)})$ from $Q(\mathcal{X}, \hat{\theta}_x^{(j)})$. The last term is the culprit. From Theorem 1 we have that $Q(\mathcal{X}, \hat{\theta}_x^{(j)}) - Q(\mathcal{X}, \theta_0^{(j)})$ is strongly negatively related to $Q(\mathcal{Y}, \hat{\theta}_x^{(j)}) - Q(\mathcal{Y}, \theta_0^{(j)})$. So the larger this term is in-sample, the worse a fit can we expect to see out-of-sample. So this term is deceiving, because increases the observed criterion function, $Q(\mathcal{X}, \hat{\theta}_x^{(j)})$, which decreasing the expected value of $Q(\mathcal{Y}, \hat{\theta}_x^{(j)})$.

When comparing two arbitrary models, nested or nonnested, the identity (2) show how the results of the previous Section carry over to this situation. We have

$$\begin{aligned} Q(\mathcal{X}, \hat{\theta}_x^{(1)}) - Q(\mathcal{X}, \hat{\theta}_x^{(2)}) \;\; = \;\; & Q(\theta_0^{(1)}) - Q(\theta_0^{(2)}) \\ & + \{Q(\mathcal{X}, \theta_0^{(1)}) - Q(\theta_0^{(1)})\} - \{Q(\mathcal{X}, \theta_0^{(2)}) - Q(\theta_0^{(2)})\} \\ & + \{Q(\mathcal{X}, \hat{\theta}_x^{(1)}) - Q(\mathcal{X}, \theta_0^{(1)})\} - \{Q(\mathcal{X}, \hat{\theta}_x^{(2)}) - Q(\mathcal{X}, \theta_0^{(2)})\}, \end{aligned}$$

and the similar decomposition of the out-of-sample criterion, shows that overfitting can strongly influence the out-of-sample ranking of models. The first term in the expression above vanishes when both models nest the true model. For example if the two models are nested, and the smaller model nests the true model.

3.1 Data Mining

Theorem 1 provides a theoretical justification for the dogma that *out-of-sample analysis is less likely to produce spurious results than is in-sample analysis.*[2] In other words one is less likely to encounter a spuriously large value of $Q(\mathcal{Y}, \hat{\theta}_x)$ than is the case for $Q(\mathcal{X}, \hat{\theta}_x)$. An implication is that a good empirical result found out-of-sample is far more impressive than had it been found in-sample. When a larger model outperforms a smaller nested model in an out-of-sample comparison, this is evidence that the larger model is the better of the two.

Thus when confronted with an out-of-sample empirical result in which the conventional model has been outperformed by a more sophisticated model, it deserves attention. In fact, the excess performance may be impressive, even if the better performing model was found after a search over a moderate set of alternative specifications (data mining).

In practice it is typically impossible to determine the "aggregate mining" that led to the discovery of a particular empirical result. Besides the data exploration undertaken by the researcher who found the result, the same data may have been analyzed by many other researcher. Furthermore, the study that led to the result in question may have been influenced by previous studies of the same data.[3] This issue is particularly relevant for the analysis of time-series. If one is unable to assess the extent to which the data has been mined, then out-of-sample results would be more credible than in-sample results. In-sample, the excess performance of a complex model has to be substantially better than that of the simpler benchmark before the result deserves much attention (when data mining has occurred).

Suppose that we are to compare a large number of alternatives to a benchmark model, which is characterized by the belief that θ_B is the true value for θ. We shall quantify how likely a search over alternative models is to produce a "spurious" result, in-sample as well as out-of-sample. By spurious result, we mean a situation where the best performing model outperforms the benchmark by more than would be expected had just a single model been

[2] West (1996) acknowledged that a formal statistical justification for the use of out-of-sample analysis did not exit, but conjectured a source that is consistent with our findings. West wrote: "out-of-sample comparisons sometimes bring surprising and important insights (e.g. Nelson (1972) and Meese and Rogoff (1983)), *perhaps because inadvertent over-fitting that results from repeated profession wide use of a limited body of data.*" (Our italic).

[3] Possible impact of studies using different data can also be problematic, unless the two sets of data are independent.

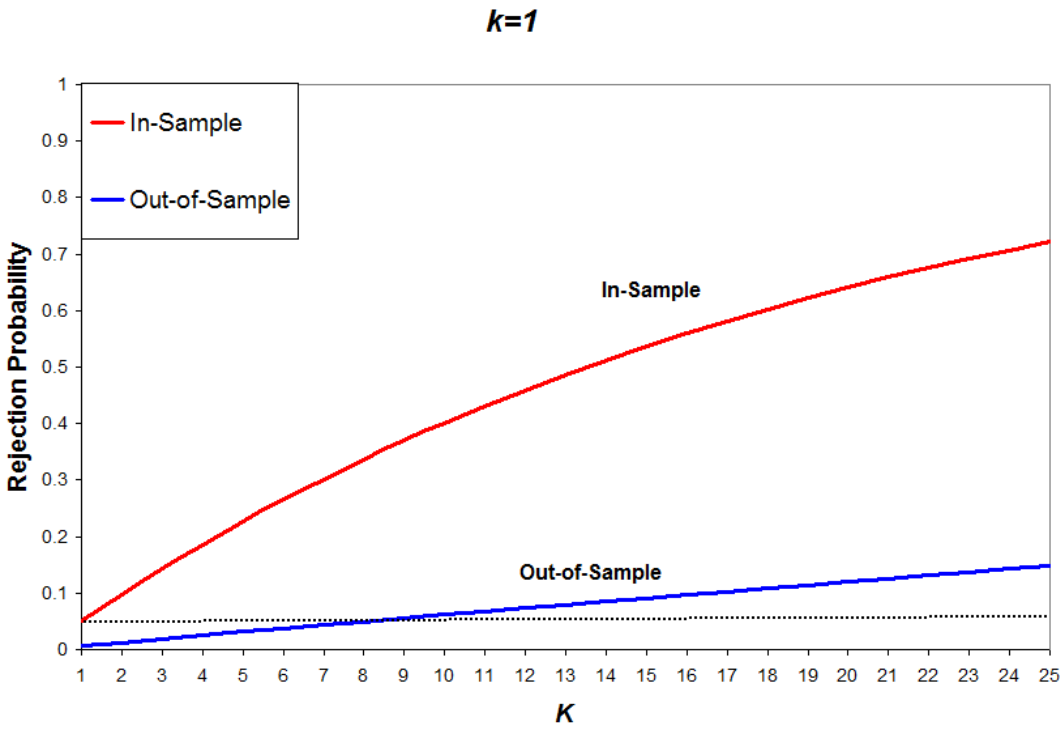

Figure 1: Regression models with one regressor are estimated and $\max_{k=1,\ldots,K} \mathrm{LR}_{x,x}$ and $\max_{k=k,\ldots,K} \mathrm{LR}_{y,x}$ are computed. The figure shows the frequency by which these statistics exceed the 5%-critical value of a χ^2-distribution with one degree of freedom. As K increases we see that both frequencies increase, but the damage done by "data mining" is far more severe in-sample than out-of-sample.

compared to the benchmark.

Suppose that c_α is the critical value associated with the test statistic, $Q(\mathcal{X}, \hat{\theta}_x) - Q(\mathcal{X}, \theta_B)$, under the null hypothesis that $\theta = \theta_B$. We report the frequencies by which

$$\sup_{j=1,...,M} Q(\mathcal{X}, \hat{\theta}_x) \geq Q(\mathcal{X}, \theta_B) + c_\alpha,$$

and

$$\sup_{j=1,...,M} Q(\mathcal{Y}, \hat{\theta}_x) \geq Q(\mathcal{Y}, \theta_B) + c_\alpha,$$

where M is the number of models being compared to the benchmark. Naturally, using c_α will not control the size of this test because it does not account for the search over specifications. Nor does it account for the estimation error in the out-of-sample comparison. Figures 1 and 2 illustrate one such situation using a simple regression design. The (true) benchmark model is $y_i = \varepsilon_i$, where ε_i are iid N$(0, 1)$, whereas the pool of alternative specifications, all have the same number of regressors ($k = 1$ or $k = 3$), that are selected from a set of K orthogonal regressors. Figure 1 displays the results for the case where all models have a single regressor ($k = 1$), and Figure 2 displays the results for $k = 3$. We have $n = m = 50$ in both designs.

Not surprisingly, do we see that a search over many model exacerbate the best empirical fit. This is true in-sample as well as out-of-sample, but much less so out-of-sample. In fact, when three regressors are used, it takes a substantial degree of data mining before the true benchmark is substantively out-performed in the out-of-sample comparison.

This finding contradicts the conclusion made in Inoue and Kilian (2004). They argue that in-sample comparisons are superior to out-of-sample tests. Specifically they write: "we question the notion that in-sample tests of predictability are more susceptible to size distortions than out-of-sample tests"; and "We conclude that results of in-sample tests of predictability will typically be more credible than results of out-of-sample tests".

The overfitting problem can be more severe in an environment with parameter instability. In this setting, the in-sample pseudo-true parameter value likely differs from the out-of-sample pseudo-true parameter value, creating an even larger gab between in-sample fit and out-of-sample fit.

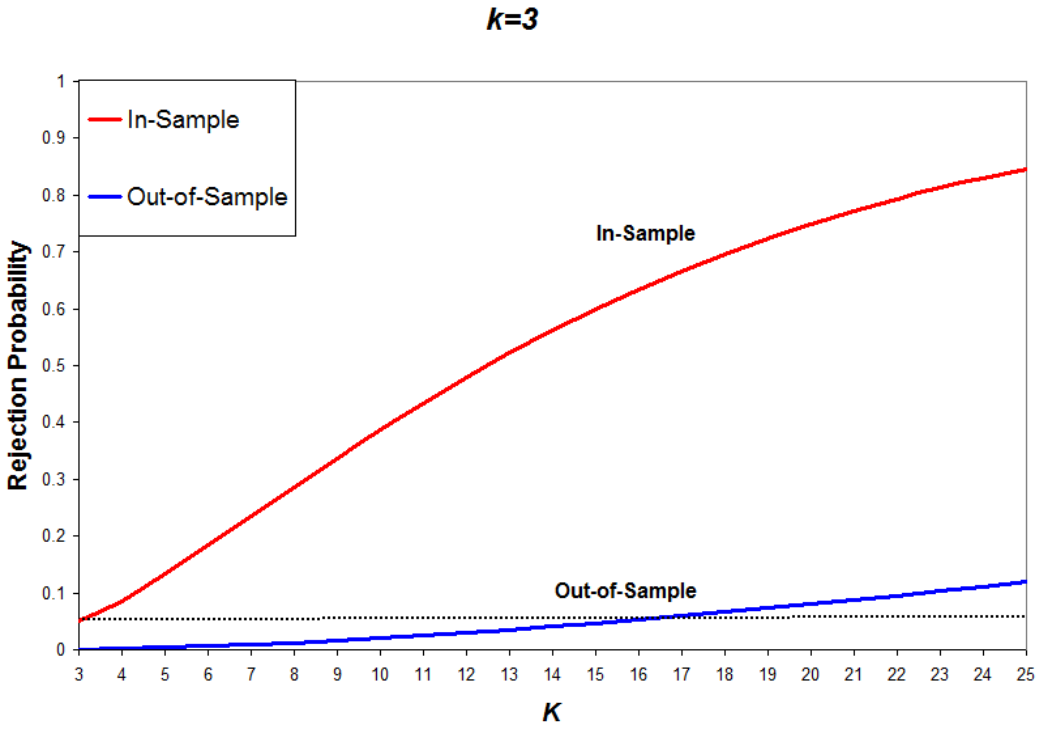

Figure 2: Regression models with exactly three regressors are estimated where the regressors are selected from a pool consisting of K regressors. The largest in-sample and out-of-sample statistics, $\mathrm{LR}_{x,x}$ and $\mathrm{LR}_{y,x}$ are computed. The figure shows the frequency by which these statistics exceed the 5%-critical value of a χ^2-distribution with three degrees of freedom. Naturally, as K increases we see that the rejection rates increase. However, the damage done by "data mining" is far more severe in-sample.

3.2 Model Selection: An Act of Hubris?

An important implication of (2) arises in this situation where multiple models are being compared. We have seen that sampling variation comes in two forms, the relative innocuous type, $Q(\mathcal{X}, \theta_0^{(j)}) - Q(\theta_0^{(j)})$, and the vicious type $Q(\mathcal{X}, \hat{\theta}_x^{(j)}) - Q(\mathcal{X}, \theta_0^{(j)})$. The latter the overfit that translate into an underfit, out-of-sample, and the implication of this term is that we do *not* want to select the model with the largest value of $Q(\theta_0^{(j)})$. Instead, the best choice is the solution to:

$$\arg\max_j \left[Q(\theta_0^{(j)}) - \{ Q(\mathcal{X}, \hat{\theta}_x^{(j)}) - Q(\mathcal{X}, \theta_0^{(j)}) \} \right].$$

It may seem paradoxical that we would prefer a model that does not (necessarily) explain the in-sample data well, but it is the logical consequence of the fact that in-sample overfitting translates into out-of-sample underfit.

In a model-rich environment, this is a knockout blow to standard model selection criteria such as AIC. The larger the pool of candidate models, the more likely is it that one of these models has a larger value of $Q(\theta_0^{(j)})$. But the downside of expanding a search to include additional models is that it adds (potentially much) noise to the problem. If the models being added to the comparison is no better than the best model, then standard model selection criteria, such as AIC or BIC will tend to select a model with an increasingly worse expected out-of-sample performance, i.e. a small $Q(\mathcal{Y}, \hat{\theta}_x^{(j)})$. Even if slightly better models are added to the set of candidate models, the improved performance, may not offset the additional noise that is added to the selection problem. If the model with the best in-sample performance, $j^* = \arg\max_j Q(\mathcal{X}, \hat{\theta}_x^{(j)})$, is indeed the best model in the sense of have the largest value of $Q(\theta^{(j)})$, then this does not guarantee a good out-of-sample performance. The reason is that the model with the best in-sample performance (possibly adjusted for degrees of freedom) is rather likely to have a large in-sample overfit, $Q(\mathcal{X}, \hat{\theta}_x^{(j)}) - Q(\mathcal{X}, \theta^{(j)})$. Since this reduces the expected out-of-sample performance, $Q(\mathcal{Y}, \hat{\theta}_x^{(j)})$, it is not obvious that selecting the model with the best (adjusted) in-sample fit is the right thing to do.

This phenomenon is often seen in practice. For example, flexible non-linear specifications tend to do better than a parsimonious model in terms of fitting the data in-sample, but substantially worse out-of-sample. This does not reflect that the true underlying model is necessarily linear, only that the gain from the nonlinearity is not large enough to offset the burden of estimating the additional parameters. See e.g. Diebold and Nason (1990).

The terminology "predictable" and "forecastable" is used in the literature to distinguish between these two sides of the forecasting problems, see Hendry and Hubrich (2006) for a recent example and discussion.

Suppose that a large number of models are being compared and suppose for simplicity that all models have the same number of parameters, so that no adjustment for the degrees of freedom is needed. We imagine a situation where all models are equally good in terms of $Q(\theta_0^{(j)})$. When the observed in-sample criterion function, $Q(\mathcal{X}, \hat{\theta}_x^{(j)})$, is larger for model A than model B, this would suggest that model A may be better than B. However, if we were to select the model with the best in-sample performance,

$$j^* = \arg\max_j Q(\mathcal{X}, \hat{\theta}_x^{(j)}),$$

we could very well be selecting the model with the largest sampling error $Q(\mathcal{X}, \hat{\theta}_x^{(j)}) - Q(\mathcal{X}, \theta_0^{(j)})$. When all models are equally good, *one may be selecting the model with the worst expected out-of-sample performance by choosing the one with the best in-sample performance.*

[ADD EXAMPLE]

It is rather paradoxical that AIC will tend to favor the model with the worst expected out-of-sample performance in this environment, and that the worst possible configuration for AIC is the one where all models in the comparison are as good as the best model. This is a direct consequence of the AIC paradox, mentioned earlier. This is not a criticism of AIC *per se*, rather it is a drawback of choosing a single model from a large pool of equally good models. Note that one would be better of by selecting a model at random in this situation.

Rather than selecting a single model, a more promising avenue to good out-of-sample performance is to aggregate the information across models, in some parsimonious way, such as model averaging.

There may be situations where the selection of a single model potentially can be useful. For example, in on unstable environment one model may be more robust to parameter changes than others. See Rossi and Giacomini (2006) for model selection in this environment. Forecasting the level or increment of a variable is effectively the same problem. But the distinction could be important for the robustness of the estimated model, as pointed out by David Hendry. Hendry argues that a model for differences is less sensitive to structural changes in the mean that a model for the level, so the former may be the best choice for

forecasting if the underlying process has time-varying parameters.

3.3 Model Averaging

The idea of combining forecast goes back to Bates and Granger (1969), see also Granger and Newbold (1977), Diebold (1988), Granger (1989), and Diebold and Lopez (1996). Forecast averaging has been used extensively in applied econometrics, and is often found to produce one of the best forecasts, see e.g. Hansen (2005). Choosing the optimal linear combination of forecasts empirically has proven difficult (this is also related to Theorem 1). Successful methods include the *Akaike weights*, see Burnham and Anderson (2002), and Bayesian model averaging, see e.g. Wright (2003). Weights that are deduced from a generalized Mallow's criterion (MMA) has recently been developed by Hansen (2006, 2007), and these are shown to be optimal in and asymptotic mean square error sense. Clark and McCracken (2006) use a very appealing framework with weakly nested models. In their local-asymptotic framework, the larger model is strictly speaking the correct model, however it is only slightly different from the nested model, and Clark and McCracken (2006) shows the advantages of model averaging in this context.

To gain some intuition, consider the average criterion function,

$$M^{-1}\sum_{j=1}^{M} Q(\mathcal{X},\hat{\theta}_x^{(j)}) = M^{-1}\sum_{j=1}^{M} Q(\mathcal{X},\theta_0^{(j)}) + M^{-1}\sum_{j=1}^{M}\{Q(\mathcal{X},\hat{\theta}_x^{(j)}) - Q(\mathcal{X},\theta_0^{(j)})\}. \quad (3)$$

Suppose that model averaging simply amounts to take the average criterion function (it does not). The last term in (3) is trivially smaller than the largest deceptive term, $\max_j\{Q(\mathcal{X},\hat{\theta}_x^{(j)}) - Q(\mathcal{X},\theta_0^{(j)})\}$. Therefore, if the models are similar in terms of $Q(\mathcal{X},\theta_0^{(j)})$, then averaging can eliminate much of the bias caused by the deceptive noise, without being too costly in terms of reducing the genuine value. Naturally, averaging over models does not in general lead to a performance that is simply the average performance. Thus for a deeper understanding we need to look at this aspect in a more detailed manner. The decomposition (2) is useful for this problem.

Define

$$\mu(\theta) = Q(\theta),$$

and

$$\eta_j(\theta) = Q(\mathcal{X},\theta) - Q(\mathcal{X},\theta_0)$$

Write (2) as $\xi_i(\tilde{\theta}) = \mu_i + \varepsilon_i + \nu_i(\tilde{\theta})$. Then our problem is to

[to be added]

4 Estimation

[This is a preliminary draft: Methods discussed in this section are mostly based on unproven conjectures.]

For the purpose of estimation we will assume that the empirical criterion function is additive, $Q(\mathcal{X}, \theta) = \sum_{t=1}^{n} q_t(x_t, \theta)$, and is such that $\{q_t(x_t, \theta)\}_{t=1}^{n}$ is stationary and

$$s_t(x_t, \theta) = \tfrac{\partial}{\partial \theta} q_t(x_t, \theta),$$

evaluated at the true parameter value, $s_t(x_t, \theta_0)$, is a martingale difference sequence. In addition to X_t, the variable, x_t, may also include lagged values of X_t. For example, if the criterion function is the log-likelihood for an autoregressive model of order one, then $x_t = (X_t, X_{t-1})^{\mathrm{T}}$ and $q_t(x_t, \theta) = -\frac{1}{2}\{\log \sigma^2 + (X_t - \varphi X_{t-1})^2/\sigma^2\}$

Recall the decomposition (2),

$$Q(\mathcal{X}, \hat{\theta}_x) = Q(\theta_0) + Q(\mathcal{X}, \theta_0) - Q(\theta_0) + Q(\mathcal{X}, \hat{\theta}_x) - Q(\mathcal{X}, \theta_0).$$

The properties of the last term, may be estimated by splitting the sample into two halves, $\mathcal{X}_1$ and $\mathcal{X}_2$, say. We estimate θ using $\mathcal{X}_1$ and leaving $\mathcal{X}_2$ for the "out-of-sample" evaluation. Hence we compute $\hat{\theta}_{x_1} = \hat{\theta}(\mathcal{X}_1)$ and the relative fit,

$$\eta = Q(\mathcal{X}_1, \hat{\theta}_{x_1}) - Q(\mathcal{X}_2, \hat{\theta}_{x_1}).$$

We may split the sample in S different ways, and index the quantities for each split by $s = 1, \ldots, S$. Taking the average

$$\frac{1}{S} \sum_{s} \eta_s,$$

will produce an estimate of $2\mathrm{E}\left\{Q(\mathcal{X}, \hat{\theta}_x) - Q(\mathcal{X}, \theta_0)\right\}$, thereby give us an estimate of the expected difference between the in-sample fit and the out-of-sample fit. (This would also produce and estimate of the proper penalty term to be used in AIC).

More generally we could consider a different sample split $n = n_1 + n_2$, and study $\eta = Q(\mathcal{X}_1, \hat{\theta}_{x_1}) - \frac{n_1}{n_2} Q(\mathcal{X}_2, \hat{\theta}_{x_1})$.

Bootstrap resampling, will also enable us to compute

$$\varepsilon_b = Q(\mathcal{X}_b^*, \hat{\theta}_x) - Q(\mathcal{X}, \hat{\theta}_x),$$

which may used to estimate aspects of the quantity, $Q(\mathcal{X}, \theta_0) - Q(\theta_0^{(j)})$.

Related references... Shibata (1997), Kitamura (1999), Hansen and Racine (2007)

[Aspects of multiperiod ahead forecasts to be discussed...]

5 Qrinkage

Shrinkage is another way to mitigate the problems induced by in-sample estimation error. Hastie, Tibshirani, and Friedman (2001) is a recent book that describes many of these methods. The factor model approach by SW is a popular way do deal with the overfitting problem in macroeconomic forecasting. Stock and Watson (2005a) consider several shrinkage methods and compare their risk functions, including the *bagging* method by Breiman (1996), see Kitamura (1999) and Inoue and Kilian (2007) for the use of bagging in an econometric setting. Risk function have previously been used to compare shrinkage methods by Magnus and Durbin (1999) and Magnus (2002). Pre-testing, where in-significant parameters are dropped from the model before a forecast is produced, is commonly used in this context. There are several aspects of pretesting that are problematic for inference, see e.g. Judge and Bock (1978), Leeb and Pötscher (2003), and Danilov and Magnus (2004). Nevertheless, its simplicity is appealing, and for the purpose of forecasting it is certainly better than estimating a large model that accumulates much of estimation error. We can in this sense view pretesting as a particular form of shrinkage.

Some shrinkage methods tend to select sparse models, i.e. models with relatively few non-zero parameters. Miller (2002) emphasize the virtues of selecting a simple and interpretable model. This aspect is also an integral part of some shrinkage methods such as *nonnegative garrote* by Breiman (1995) and the *lasso* by Tibshirani (1996).

One possible way to adjust the estimated model, prior to using it out-of-sample, is to change the parameter estimate away from $\hat{\theta}_x$, until the criterion function is reduced by a desired amount, ψ say. A natural choice for ψ is $\psi_0 = \mathrm{E}\{Q(\mathcal{X}, \hat{\theta}_x) - Q(\mathcal{X}, \theta_0)\}$, since this would offsets the expected bias of the criterion function. An obstacle to this approach is that the solution to

$$\theta : Q(\mathcal{X}, \theta) = Q(\mathcal{X}, \hat{\theta}_x) - \psi_0,$$

will not be unique in most situations. So which of the many solutions should we choose? This issues can be resolved by introducing a *gravity model*.

A gravity model is characterized by a parameter value, θ^*. Qrinkage amounts to shrinking the unrestricted estimate, $\hat{\theta}_x$, towards the gravity point, θ^*, until the criterion function is reduced by a prespecified amount. A natural choice is ψ_0, because it offset the in-sample bias in the value of the criterion function. However, in some cases, one may want to shrink by more or less than ψ_0. If a selection over different shrunken models is the way that the final model is chosen, then more shrinkage is typically needed to offset the bias induced by the selection.

Manganelli (2006) has independently proposed a very similar form of shrinkage. The starting point in Manganelli (2006) is a judgemental forecast. The judgemental forecast is adopted unless there is statistical evidence to suggest this forecast is inferior. When the judgemental forecasts is at odds with the empirical evidence, Manganelli suggests to adjust the judgemental forecast until it no longer is significantly at odds with the data, using some prespecified significance level. The judgemental forecast is similar to the gravity model in our framework, and the significance level is used to control the extend of shrinkage.

The shrinkage towards the gravity model, can be done in various ways. In the regression context we can adopt a nonnegative garrote style shrinkage. For example, if $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are the (unrestricted) point estimates, then we consider the solution to the constrained optimization problem,

$$\min_{c_1,\ldots,c_k} \sum_{i=1}^{n} (Y_i - c_1\hat{\beta}_1 X_{1,i} - \cdots - c_k\hat{\beta}_k X_{k,i})^2, \qquad \text{s.t.} \quad c_j \geq 0 \quad \text{and} \quad \sum_{j=1}^{k} c_j \leq s.$$

The extent of shrinkage is controlled by s. So we can "tighten" the estimates by shrinking s towards zero, until the criterion function is reduced by the desired amount, ψ say. Let $c_1(\psi), \ldots, c_k(\psi)$ be the resulting shrinkage factors, then the final shrinkage estimates are given by

$$\tilde{\beta}_j = c_j(\psi)\hat{\beta}_j, \qquad \text{for} \qquad j = 1, \ldots, k.$$

5.1 Qrinkage in Regression Models

Consider the simple linear regression model,

$$Y = X\beta + \varepsilon,$$

where $\varepsilon|X \sim N(0, \sigma_\varepsilon^2 I)$.

It is well known that minus two times the log-likelihood is given by

$$-2\ell(\sigma_\varepsilon^2, \beta) = \frac{n}{\sigma_\varepsilon^2}(S_{yy} - \beta^{\mathrm{T}} S_{xy} - S_{yx}\beta + \beta^{\mathrm{T}} S_{xx}\beta),$$

where we have used the definitions $S_{yy} = Y^{\mathrm{T}}Y/n$, $S_{xy} = X^{\mathrm{T}}Y/n$, $S_{yx} = S_{xy}^{\mathrm{T}}$, and $S_{xx} = X^{\mathrm{T}}X/n$.

We shall estimate the parameters by least squares and shrink the estimator of β towards the gravity point. Here we can take $\beta^* = 0$ without loss of generality. (If $\beta^* = b \neq 0$ we can reparameterize the model $\tilde{Y} = Y - Xb = X\tilde{\beta} + \varepsilon$, where $\tilde{\beta} = (\beta - b)$.)

Let $S_{xx} = V^{\mathrm{T}}\Lambda V$ be the diagonalization of S_{xx}, and define $\gamma = V\beta$.

$$\begin{aligned} Y &= XV^{\mathrm{T}}V\beta + \varepsilon, \\ &= W\gamma + \varepsilon \\ &= (W_1\gamma_1 + \cdots + W_k\gamma_k) + \varepsilon, \end{aligned}$$

where $W^{\mathrm{T}}W = VX^{\mathrm{T}}XV^{\mathrm{T}} = VV^{\mathrm{T}}\Lambda VV^{\mathrm{T}} = \Lambda$, such that the regressors are orthogonal.

If we define $\delta \equiv VS_{yx}$ we have that

$$\hat{\gamma}_i = \frac{S_{yx}q_i}{q_i^{\mathrm{T}}S_{xx}q_i} = \frac{\delta_i}{\lambda_i}.$$

If we hold σ_ε^2 fixed we have that

$$\begin{aligned} -2\ell(\gamma) &= \frac{n}{\sigma_\varepsilon^2}(S_{yy} - 2\delta^{\mathrm{T}}\gamma + \gamma^{\mathrm{T}}\Lambda\gamma) \\ &= \frac{n}{\sigma_\varepsilon^2}(S_{yy} - 2\sum_{i=1}^{k}\delta_i\gamma_i + \lambda_i\gamma_i^2), \end{aligned}$$

and the idea is to shrink γ_i towards zero such that $-2\ell(\cdot)$ is reduced by one unit, as this would be the bias of two times the log-likelihood when the model is correctly specified. Thus, for each i we seek the solution to

$$\frac{n}{\sigma_\varepsilon^2}\left\{\lambda_i(\kappa_i\gamma_i)^2 - 2\delta_i(\kappa_i\gamma_i)\right\} = \frac{n}{\sigma_\varepsilon^2}(\lambda_i\gamma_i^2 - 2\delta_i\gamma_i) + 1,$$

where $\kappa_i \in [0, 1]$ (if equality cannot be achieved, we set $\kappa_i = 0$).

$$\begin{aligned} 0 &= \lambda_i\gamma_i^2\kappa_i^2 + (-2\delta_i\gamma_i)\kappa_i + (2\delta_i\gamma_i - \lambda_i\gamma_i^2 - \frac{\sigma_\varepsilon^2}{n}) \\ &= \lambda_i\frac{\delta_i^2}{\lambda_i^2}\kappa_i^2 + (-2\delta_i\frac{\delta_i}{\lambda_i})\kappa_i + (2\delta_i\frac{\delta_i}{\lambda_i} - \lambda_i\frac{\delta_i^2}{\lambda_i^2} - \frac{\sigma_\varepsilon^2}{n}) \end{aligned}$$

$$= \frac{\delta_i^2}{\lambda_i}\kappa_i^2 + \frac{-2\delta_i^2}{\lambda_i}\kappa_i + (\frac{\delta_i^2}{\lambda_i} - \frac{\sigma_\varepsilon^2}{n}),$$

or equivalently

$$\kappa_i^2 - 2\kappa_i + (1 - \frac{\lambda_i}{\delta_i^2}\frac{\sigma_\varepsilon^2}{n}) = 0,$$

which has the two roots given by

$$\frac{2 \pm \sqrt{4 - 4(1 - \frac{\lambda_i}{\delta_i^2}\frac{\sigma_\varepsilon^2}{n})}}{2} = 1 \pm \sqrt{\frac{\lambda_i}{\delta_i^2}\frac{\sigma_\varepsilon^2}{n}}.$$

Since the gravity point is the origin, the relevant root is the smaller of the two, so that

$$\kappa_i^* = \max(0, 1 - \sqrt{\frac{\lambda_i}{\delta_i^2}\frac{\sigma_\varepsilon^2}{n}}) = \max(0, 1 - \frac{1}{\sqrt{n}}\frac{\sigma_\varepsilon}{\sigma_y}\frac{1}{|\rho_{y,w_i}|}). \tag{4}$$

There are several interesting observations to be made from the shrinkage formula (4).

1. The more observations we have the less shrinkage.

2. The more noise-to-signal ($\sigma_\varepsilon/\sigma_y$) the more shrinkage

3. The larger is the (absolute) correlation between Y and W_i, the less shrinkage.

The qrinkage estimator, $\tilde{\gamma}_i = \hat{\gamma}_i\kappa_i^*$, can be rewritten as

$$\tilde{\gamma}_i = \begin{cases} \hat{\gamma}_i + \Delta_{i,n} & \text{if } \hat{\gamma}_i < -\Delta_{i,n} \\ 0 & \text{if } -\Delta_i \leq \hat{\gamma}_i \leq \Delta_i \\ \hat{\gamma}_i - \Delta_{i,n} & \text{if } \hat{\gamma}_i > \Delta_i \end{cases} \qquad \Delta_{i,n} = \frac{1}{\sqrt{n}}\sqrt{\sigma_\varepsilon^2/\lambda_i}.$$

So in the regression context, the qrinkage estimator is known as the Burr estimator, see Magnus (2002).[4] Put differently, the likelihood-based shrinkage can motivate the Burr estimator.

Shrinking a parameter all the way to zero, may not reduce the criterion function by the desired amount. In such case, one may want to shrink other parameters further, such that the aggregate reduction of the criterion function is the desired amount.

Figure 3 illustrates Qrinkage in a 6-month ahead forecasting exercise for personal Income, using a simple autoregressive model.

[4] I thank Jan Magnus for pointing this out to me.

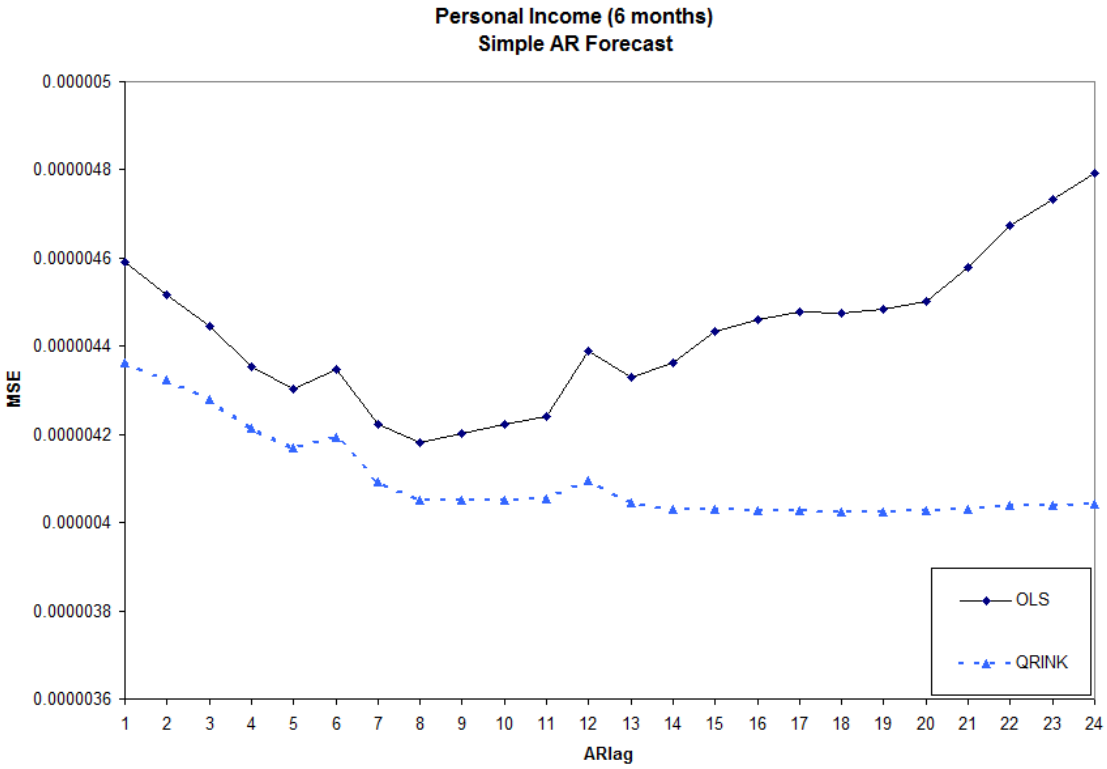

Figure 3: MSE of six-month ahead forecast of Personal Income using the OLS estimates from an autoregressive model and the corresponding Qrinkage estimates. The forecast of the unrestricted OLS estimator initially gets better as more lags are included in the model, but then deteriorates rapidly. The Qrinkage estimate is much less sensitive to including a large number of lags (because qrinkage sets them to zero).

5.1.1 Ordered Qrinkage in Regression Models

Rewrite

$$y_t = \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t,$$

as

$$y_t = \alpha_1 \tilde{x}_{1,t} + \cdots + \alpha_k \tilde{x}_{k,t} + \varepsilon_t,$$

where $\tilde{x}_{1,t} = x_{1,t}$, $\tilde{x}_2 = (I - P_1)x_2$, $\tilde{x}_3 = (I - P_{1:2})x_3$, , $\tilde{x}_k = (I - P_{1:k-1})x_k$. Note that

$$\tilde{x}_i^{\mathrm{T}} \tilde{x}_j = x_i^{\mathrm{T}} (I - P_{1:i-1})(I - P_{1:j-1}) x_j = 0$$

because

$$(I - P_{1:i-1})(I - P_{1:j-1}) = I - P_{1:i-1} - P_{1:j-1} + P_{1:i-1}P_{1:j-1} = I - P_{1:j-1},$$

and

$$x_i^{\mathrm{T}} (I - P_{1:j-1}) = x_i^{\mathrm{T}} - x_i^{\mathrm{T}} = 0.$$

The resulting model can be viewed as a "soft" alternative to conventional model selection methods that are based on information criteria. The qrinkage "selection" does not have the discontinuity of standard information criteria, such as AIC and BIC, where a sharp threshold determines whether a parameter is set to zero, or kept in the model at its unrestricted point estimate.

[Show how to recover estimates of β from those of α]

5.1.2 Unit Roots

Shrink towards a unit root.. extra useful because the maximum likelihood estimator tends to be biased away from the unit root. Part of the explanation for the empirical success of the Minnesota prior, introduced by Doan, Litterman, and Sims (1984).

5.1.3 Partial Qrinkage in Regression Models

Consider now

$$Y = X\beta + Z\varphi + \varepsilon,$$

and suppose that we are only interested in shrinking the parameters associated with X, while leaving the coefficients associated with Z unrestricted. (Naturally, our shrinking of $\hat{\beta}$

will cause the least squares estimate of φ to change, so these are not entirely unaffected by the qrinkage).

Here we define $R_0 = \{I - Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}\}Y$ and $R_1 = \{I - Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}\}X$, and consider the concentrated regression equation

$$R_0 = R_1\beta + \tilde{\varepsilon}$$

and define $S_{00} = R_0^{\mathrm{T}}R_0/n$, $S_{10} = R_1^{\mathrm{T}}R_0$, $S_{01} = S_{10}^{\mathrm{T}}$, and $S_{11} = R_1^{\mathrm{T}}R_1/n$, and decompose $S_{11} = V^{\mathrm{T}}\Lambda V$ and define $\delta = V^{\mathrm{T}}S_{10}$. The estimator of β is given by

$$\hat{\beta} = S_{11}^{-1}S_{10}$$

and $\hat{\gamma} = V\hat{\beta}$, and shrink according to (4).

5.1.4 Qrinkage Interpretation of Diffusion Indexes

The expression (4) is also interesting, as it shows that the first principal component (those with a large λ_i) should be shrunk less than the last PC (those with a small λ_i). This may explain the empirical success of forecasting using diffusion indices by Stock and Watson, who keeps the regression coefficients associated with the first few principal components in the final models, whereas all other coefficient are set to zero. In principle there is no reason to expect that the t-statistics associated with the first principal components should be larger that those associated with the last principal components. Since this is often found empirically, this suggest that the first principal components are capturing real economic features that are useful of forecasting a great variety of variables.

Here revisiting the data analyzed in Stock and Watson (2005b)/Stock and Watson (2005a). [hof.xls – thank Mark Watson for data].

Qrinkage offers an alternative to selecting a fixed number of factors. The standard approach has been: regression coefficients associated with the principal components are either kept at their unrestricted point estimate, or forced to be zero. An approach that is similar to model selection methods and pre-testing. See e.g. Stock and Watson (2002a, 2002b) and Bai and Ng (2002). Qrinkage provides a smooth transition between these binary choices.

The choice of factors is often made without consideration to the variable being forecasted. This is an odd aspect of this approach, but it does have the advantage of reducing

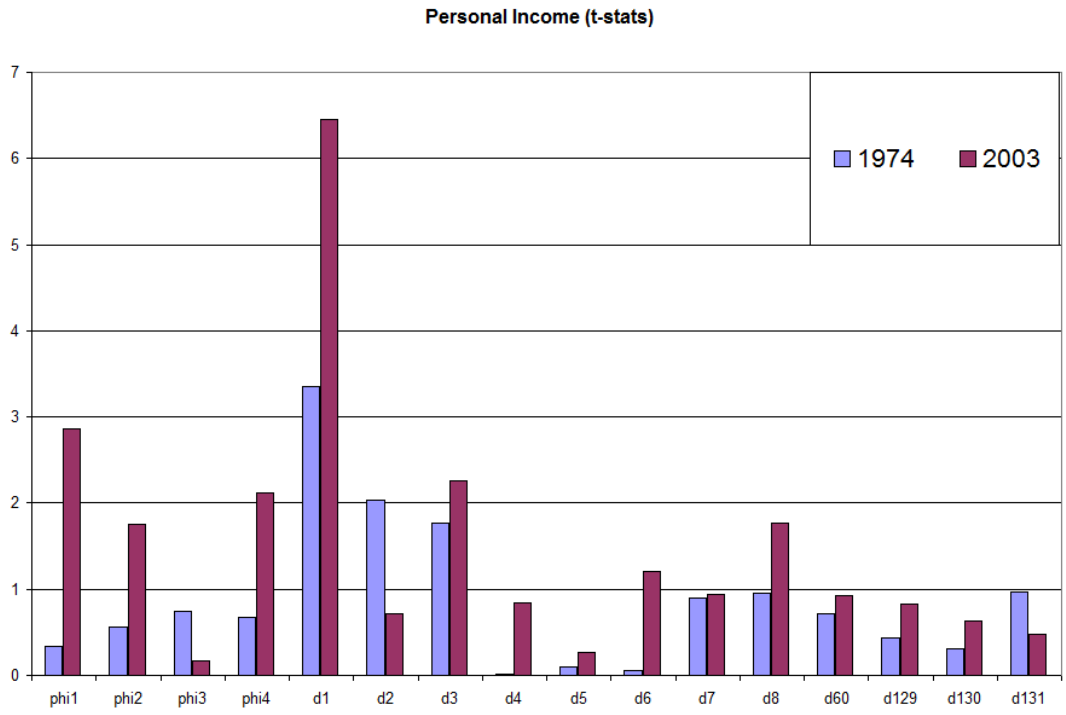

Figure 4: The absolute value of several t-statistics are reported. Left columns refer to the statistics obtained using data up until 1974 and right columns give the t-statistics using the larger sample up until 2003. The dependent variable is personal income, the regression model includes four lags, and a set of possible principal component. di refers to the ith principal component.

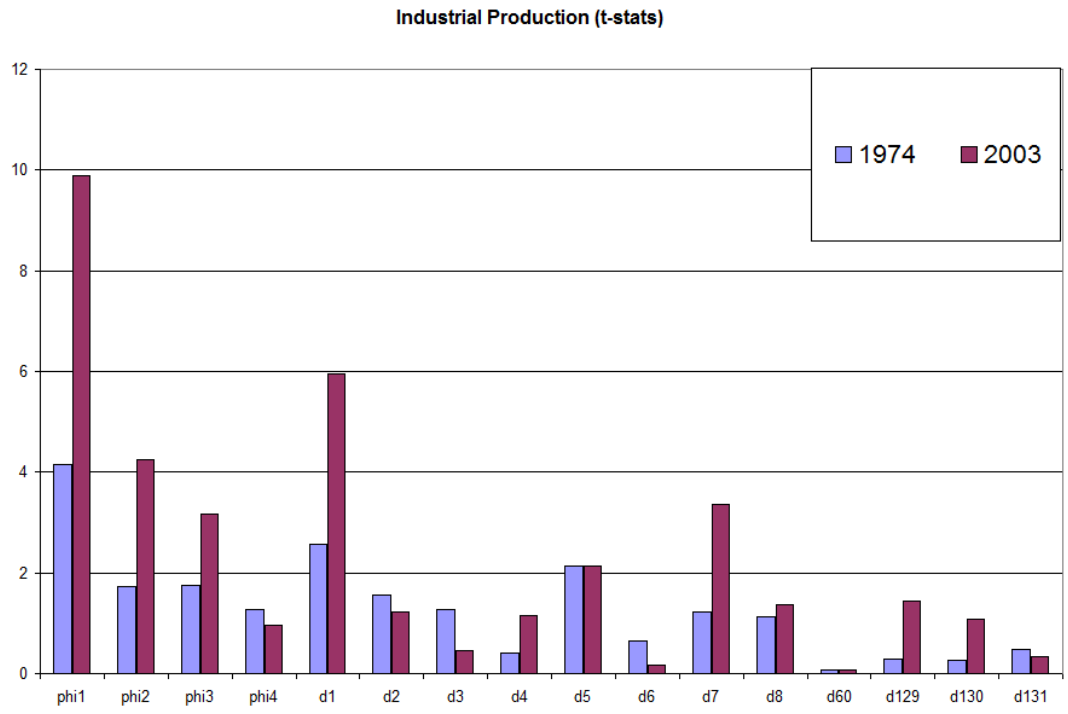

Figure 5: The absolute value of several t-statistics are reported. Left columns refer to the statistics obtained using data up until 1974 and right columns give the t-statistics using the larger sample up until 2003. The dependent variable is industrial production, the regression model includes four lags, and a set of possible principal component. di refers to the ith principal component

the overfitting problem. Qrinkage lets the data speak as to which factors are relevant, while keeping the overfitting in check. Another way to let the "data speak for themselves" is to use other forms of shrinkage, such as that proposed by Bai and Ng (2007), who use the terminology of "target predictors".

The method of *sliced inverse regression* is an approach that shares some of the features of the principle components, without disregarding the relation between predictors and the variable to be forecasted when making the data reduction. The sliced inverse regression was introduced by Li (1991), and has not been used much in econometrics, see Chen and Smith (2007) for a recent exception. For a good description of the relation between SIR and related methods, see Naik, Haferty, and Tsai (2000).

In practice one often finds the most "significant" regressors to be those associated with the first principal components. A good example of this situation is illustrated in Figure 4, where Personal Income is the dependent variable. At times the data seems to ask for a other factors than the first few, as seen in Figure 5. This figure displays the results for the case where Industrial Production is the dependent variable, and we see that the 7th principal component is rather significant according to its t-statistic.

5.2 Combining Forecasts

In the context of point forecasting, there is an natural alternative to taking a (weighted) average of the parameters in the competing models. Instead we can take a linear combination of the individual point forecasts. Let Y_{t+1} be the variable to be forecasted, and let $\hat{Y}_t^{(j)}$, $j = 1, \ldots, M$ be the competing forecasts. We can stack the forecast into the vector $\hat{Y}_t = (\hat{Y}_t^{(1)}, \ldots, \hat{Y}_t^{(M)})^{\mathrm{T}}$. Given the empirical success of principal components in the context of Stock and Watson, it would be natural to consider the principal components of the vector of forecasts, $\hat{Y}_t$. We may decompose the individual forecast into

$$\hat{Y}_t^{(j)} = \mathrm{E}(Y_{t+1}|\mathcal{F}_t) + [\text{bias}]^{(j)} + [\text{error}]_t^{(j)}.$$

If the target variable is a persistent variables, such as inflation, an interest rates, or the GDP growth-rate, then $\mathrm{E}(Y_{t+1}|\mathcal{F}_t)$ may define (or be closely related to) the first principal component of $\hat{Y}_t$. Thus the first principal component (suitably scaled) will in this situation be quite similar to the equal-weighted combination of the individual forecasts. It would be very interesting to study this aspect empirically, because this could link the empirical

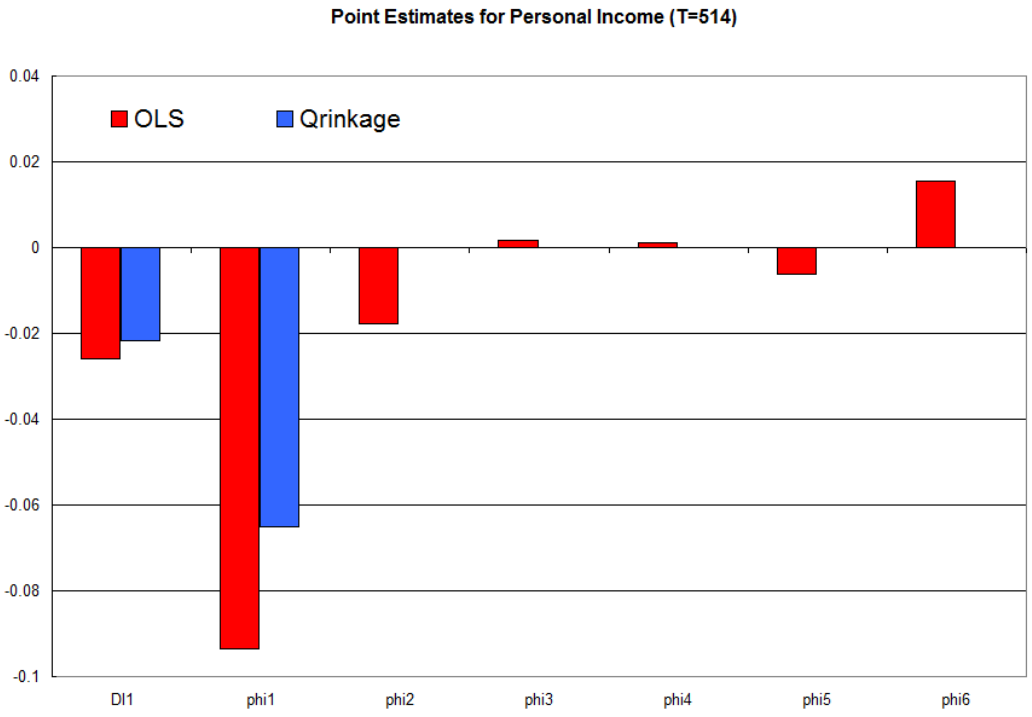

Figure 6: Least squares point estimates and the corresponding qrinkage estimates. Here we see that all but the first two regression parameters are shrunken all the way to zero by the qrinkage procedure.

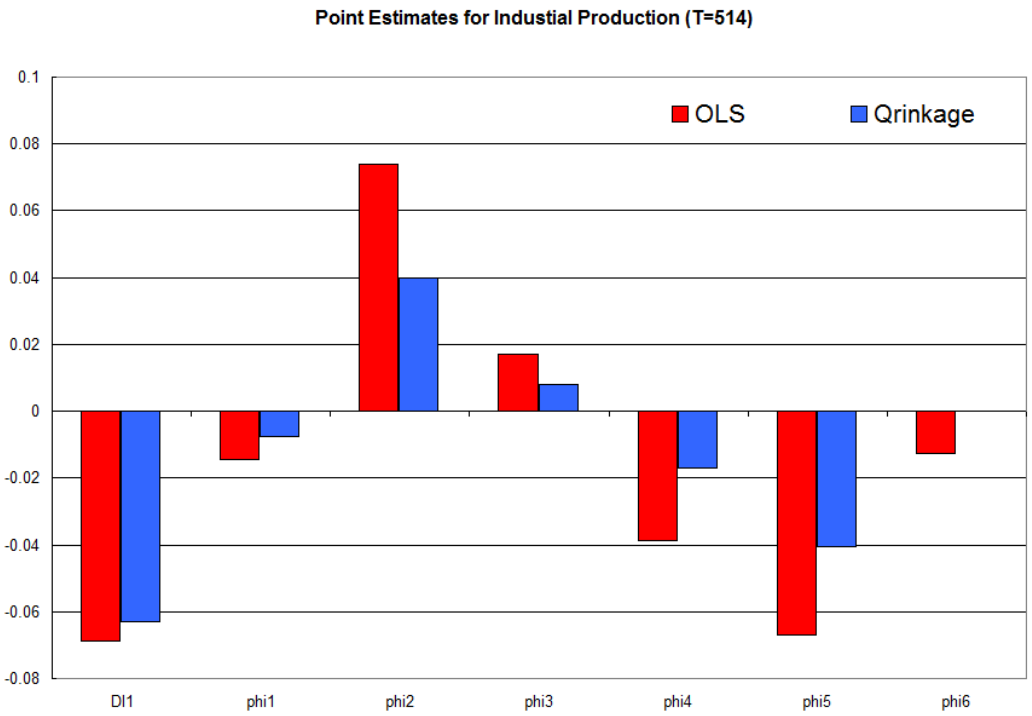

Figure 7: Least squares point estimates and the corresponding qrinkage estimates.

success of the equal-weighted forecasts, to that found in the context of SW. Furthermore, it would suggest ways to improve upon the equal-weighted forecasts, as the first principal component may not be exactly proportional to ι (the vector of ones), and it may also useful to incorporated more than the first principal component in the construction of a combined forecast.

5.3 Qrinkage of Weak/Many Instruments

Instrumental variables

$$\begin{aligned} Y_i &= X_i^{\mathrm{T}}\beta + u_i \\ X_i &= Z_i^{\mathrm{T}}\pi + v_i. \end{aligned}$$

The TSLS estimator is given by $\hat{\beta}_{\mathrm{IV}} = (\hat{X}^{\mathrm{T}}\hat{X})^{-1}\hat{X}^{\mathrm{T}}Y$ where $\hat{X} = P_Z X$ and $P_Z = Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}$. A key problem with the TSLS is that $(\hat{X}^{\mathrm{T}}\hat{X})$ is too large, particularly when the instruments, Z, are and/or used in large numbers. The problem is that the first-stage regression will explain more variation in X, than had the true value of π been used. It would be straight forward to "qrink" $\hat{\pi} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X$, such that the second-stage regression would involve less variable regressors. It would be interesting to compare the resulting estimator to k-class estimators: $\hat{\beta}_k = [X^{\mathrm{T}}\{I - k(I - P_Z)\}X]^{-1}X^{\mathrm{T}}\{I - k(I - P_Z)\}Y$.

6 Concluding Remarks

[To be added]

We have seen that model selection by information criteria, such as AIC, is an act of hubris in a model-rich environment.

Qrinkage has been applied in Chun (2007) and his results are rather promising for the use of qrinkage in practise.

References

Akaike, H. (1974): "A New Look at the Statistical Model Identification," *IEEE transactions on automatic control*, 19, 716–723.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, MA.

Atkeson, A., and L. E. Ohanian (2001): "Are Phillips Curves Useful for Forecasting Inflation?," *Federal Reserve Bank of Minneapolis Quarterly Review*, 25.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2007): "Forecasting Time Series Using Targeted Predictors," working paper.

Bates, J. M., and C. W. J. Granger (1969): "The Combination of Forecasts," *Operational Research Quarterly*, 20, 451–468.

Bjørnstad, J. F. (1990): "Predictive Likelihood: A Review," *Statistical Science*, 5, 242–265.

Breiman, L. (1995): "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

——— (1996): "Bagging Predictors," *machine learning*, 26, 123–140.

Burnham, K. P., and D. R. Anderson (2002): *Model Selection and MultiModel Inference*. Springer, New York, 2nd edn.

Chatfield, C. (1995): "Model Uncertainty, Data Mining and Statistical Inference," *Journal of the Royal Statistical Society, Series A*, 158, 419–466.

Chen, P., and A. Smith (2007): "Dimension Reduction Using Inverse Regression and Nonparametric Factors," working paper.

Chun, A. L. (2007): "Forecasting Interest Rates and the Macroeconomy: Blue Chip Clairvoyants, Econometrics or Qrinkage?," Working paper.

Clark, T. E., and M. W. McCracken (2006): "Combining Forecasts from Nested Models," Manuscript. Federal Reserve Bank of Kansas City.

Clark, T. E., and K. D. West (2007): "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 127, 291–311.

Danilov, D. L., and J. R. Magnus (2004): "On the Harm That Ignoring Pretesting Can Cause," *Journal of Econometrics*, 122, 27–46.

Diebold, F. X. (1988): "Serial Correlation and the Combination of Forecasts," *Journal of Business and Economic Statistics*, 6, 105–111.

Diebold, F. X., and J. A. Lopez (1996): "Forecast Evaluation and Combination," in *Handbook of Statistics*, ed. by G. S. Maddala, and C. R. Rao, vol. 14, pp. 241–268. North-Holland, Amsterdam.

Diebold, F. X., and J. A. Nason (1990): "Nonparametric Exchange Rate Prediction?," *Journal of International Economics*, 28, 315–332.

Doan, T., R. Litterman, and C. Sims (1984): "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometrics Reviews*, 3, 1–100.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

Granger, C. W. J. (1989): "Combining Forecasts – Tweenty Years Later," *Journal of Forecasting*, 8, 167–173.

Granger, C. W. J., and P. Newbold (1977): *Forecasting Economic Time Series*. Academic Press, Orlando.

Hansen, B. E. (2006): "Least Squares Forecast Averaging," working paper.

——— (2007): "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189.

Hansen, B. E., and J. S. Racine (2007): "Jackknife Model Averaging," working paper.

Hansen, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23, 365–380.

Hastie, T., R. Tibshirani, and J. Friedman (2001): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hendry, D. F., and M. P. Clements (2002): "Pooling of Forecasts," *Econometrics Journal*, 5, 1–26.

Hendry, D. F., and K. Hubrich (2006): "Forecasting Economic Aggregates by Disaggregates," ECB working paper.

Hong, H., and B. Preston (2006): "Nonnested Model Selection Criteria," working paper.

Huber, P. (1981): *Robust Statistics*. Wiley, New York.

Inoue, A., and L. Kilian (2004): "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," *Econometrics Reviews*, 23, 371–402.

——— (2007): "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPIInflation," *Journal of the American Statistical Association*, forthcoming.

Judge, G. G., and M. E. Bock (1978): *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. North-Holland, Amsterdam.

Kitamura, Y. (1999): "Predictive Inference and the Bootstrap," working paper.

Lauritzen, S. L. (1974): "Sufficiency, Prediction and Extreme Models," *Scandinavian Journal of Statistics*, 1, 128–134.

Leeb, H., and B. Pötscher (2003): "The Finite-Sample Distribution of Post-Model-Selection Estimators, and Uniform Versus Non-Uniform Approximations," *Econometric Theory*, 19, 100–142.

Li, K.-C. (1991): "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342.

Magnus, J. R. (2002): "Estimation of the Mean of a Univariate Normal Distribution with Known Variance," *Econometrics Journal*, 5, 225–236.

Magnus, J. R., and J. Durbin (1999): "Estimation of Regression Coefficients of Interest When Other Regression Coefficients are of No Interest," *Econometrica*, 67, 639–643.

Manganelli, S. (2006): "A New Theory of Forecasting," working paper.

Meese, R., and K. Rogoff (1983): "Exchange Rate Models of the Seventies. Do They Fit Out of Sample?," *Journal of International Economics*, 14, 3–24.

Miller, A. (2002): *Subset Selection in Regression*. Chapman and Hall/CRC, Boca Raton, 2nd edn.

Naik, P. A., M. R. Haferty, and C.-L. Tsai (2000): "A New Dimension Reduction Approach for Data-Rich Marketing Environments: Sliced Inverse Regression," *Journal of Marketing Research*, 37, 88–101.

Nelson, C. R. (1972): "The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy," *American Economic Review*, 62, 902–917.

Pesaran, H., and A. Timmermann (2005): "Small Sample Properties of Forecasts from Autoregressive Models under Structural Breaks," *Journal of Econometrics*, 129, 183–217.

Rossi, B., and R. Giacomini (2006): "Non-Nested Model Selection in Unstable Environments," working paper.

Shibata, R. (1997): "Bootstrap Estimate of Kullback-Leibler Information for Model Selection," *Statistica Sinica*, 7, 375–394.

Shimizu, R. (1978): "Entropy Maximization Principle and Selecting of the Order of an Autoregressive Gaussian Process," *Annals of the Institute of Statistical Mathematics*, 30, 263–270.

Stock, J. H., and M. W. Watson (2002a): "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147–162.

——— (2005a): "An Empirical Comparison of Methods for Forecasting Using Many Predictors," working paper.

——— (2005b): "Implications of Dynamic Factor Models for VAR Analysis," working paper.

Takeuchi, K. (1976): "Distribution of Informational Statistics and a Criterion of Model Fitting," *Suri-Kagaku (Mathematical Sciences)*, 153, 12–18, (In Japanese).

Tibshirani, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

West, K. D. (1996): "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

White, H. (1994): *Estimation, Inference and Specification Analysis*. Cambridge University Press, Cambridge.

Wright, J. H. (2003): "Bayesian Model Averaging and Exchange Rate Forecasts," working paper.

A Appendix of Proofs

[some details to be added].

Proof of Theorem 1. To simplify notation we write $Q_x(\cdot)$ as short for $Q(\mathcal{X}, \cdot)$. Since $\hat{\theta}_x$ is given by $Q'_x(\hat{\theta}_x) = 0$, we have

$$0 = Q'_x(\hat{\theta}_x) = Q'_x(\theta_0) + Q''_x(\tilde{\theta})(\hat{\theta}_x - \theta), \quad \text{where } \tilde{\theta} \in [\theta_0, \hat{\theta}_1]$$

so that

$$(\hat{\theta}_x - \theta_0) = \left[-Q''_x(\tilde{\theta})\right]^{-1} Q'_x(\theta_0),$$

and we have

$$\begin{aligned} Q_x(\hat{\theta}_x) - Q_x(\theta_0) &= \frac{1}{2}(\hat{\theta}_x - \theta_0)^{\mathrm{T}} \left[-Q''_x(\tilde{\theta})\right] (\hat{\theta}_x - \theta_0) \\ &= Q'_x(\theta_0)^{\mathrm{T}} \left[-Q''_x(\theta_0)\right]^{-1} Q'_x(\theta_0) + o_p(1) \end{aligned}$$

For the out-of-sample period we have

$$\begin{aligned} Q_y(\hat{\theta}_x) - Q_y(\theta_0) &= Q'_y(\theta_0)^{\mathrm{T}}(\hat{\theta}_x - \theta) + \frac{1}{2}(\hat{\theta}_x - \theta_0)^{\mathrm{T}} Q''_y(\tilde{\theta})(\hat{\theta}_x - \theta_0) + o_p(1) \\ &= Q'_y(\theta_0)^{\mathrm{T}} \left[-Q''_x(\theta_0)\right]^{-1} Q'_x(\theta_0) \\ &\quad + \frac{1}{2} Q'_x(\theta_0)^{\mathrm{T}} \left[-Q''_x(\theta_0)\right]^{-1} Q''_y(\tilde{\theta}) \left[-Q''_x(\theta_0)\right]^{-1} Q'_x(\theta_0) + o_p(1). \end{aligned}$$

Since $m^{-1} Q''_y(\tilde{\theta}_n) \xrightarrow{p} \mathcal{I}_0$ and $n^{-1} Q''_x(\tilde{\theta}_n) \xrightarrow{p} \mathcal{I}_0$, whenever $\tilde{\theta}_n \xrightarrow{p} \theta_0$, we have

$$\begin{aligned} Q_y(\hat{\theta}_x) - Q_y(\theta_0) &= \sqrt{\frac{m}{n}} Q'_y(\theta_0)^{\mathrm{T}} \{\mathcal{I}_0\}^{-1} Q'_x(\theta_0) \\ &\quad + \frac{1}{2}\frac{m}{n} Q'_x(\theta_0)^{\mathrm{T}} \{\mathcal{I}_0\}^{-1} Q'_x(\theta_0) + o_p(1). \end{aligned}$$

■

The in-sample and out-of-sample log-likelihood functions are given by

$$\ell_x(\theta) \equiv \sum_{i=1}^{n} \log f(X_i; \theta), \quad \text{and} \quad \ell_y(\theta) \equiv \sum_{i=n+1}^{n+m} \log f(X_i; \theta).$$

We assume that the likelihood functions satisfy Assumption 1. The in-sample and out-of-sample maximum-likelihood estimators are given by

$$\hat{\theta}_x = \arg\max_{\theta} \ell_x(\theta) \qquad \text{and} \qquad \hat{\theta}_y = \arg\max_{\theta} \ell_y(\theta).$$

Assumption 1 ensures that $S_x(\hat{\theta}_x) = S_y(\hat{\theta}_y) = 0$, where the scores are defined by

$$S_x(\theta) \equiv \frac{\partial}{\partial \theta} \ell_x(\theta) \qquad \text{and} \qquad S_y(\theta) \equiv \frac{\partial}{\partial \theta} \ell_y(\theta).$$

The Hessians,

$$H_x(\theta) \equiv \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \ell_x(\theta) \qquad \text{and} \qquad H_y(\theta) \equiv \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \ell_y(\theta),$$

are such that $\hat{\theta} \xrightarrow{p} \theta_0 \Rightarrow H_z(\hat{\theta})[H_z(\theta_0)]^{-1} \xrightarrow{p} I_k$ for both $z = x$ and $z = y$.

We can factorize the log-likelihood function and express the scores and the Hessians as

$$S_x(\theta) = \sum_{i=1}^{n} s_i(\theta) \quad \text{and} \quad S_y(\theta) = \sum_{i=n+1}^{n+m} s_i(\theta),$$

and

$$H_x(\theta) = \sum_{i=1}^{n} h_i(\theta) \quad \text{and} \quad H_y(\theta) = \sum_{i=n+1}^{n+m} h_i(\theta),$$

where $s_i(\theta) \equiv \frac{\partial}{\partial \theta} \log f(X_i; \theta)$ and $h_i(\theta) = \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \log f(X_i; \theta)$.

Proof of Theorem 2. In this standard likelihood framework we have that

$$0 = S_{1,\hat{\theta}_1} = S_{1,\theta_0} + H_{1,\tilde{\theta}}(\hat{\theta}_1 - \theta_0),$$

where $\tilde{\theta}$ lies between $\hat{\theta}_1$ and θ_0, such that

$$(\hat{\theta}_1 - \theta_0) = [-H_{1,\tilde{\theta}}]^{-1} S_{1,\theta_0}.$$

Correct specification ensures that

$$\Sigma_s \equiv E[s_{i,\theta_0} s_{i,\theta_0}^{\mathrm{T}}] = -E[h_{i,\theta_0}],$$

also known as the information matrix equality, and regularity conditions ensure that

$$-H_{1,\tilde{\theta}} = -\frac{1}{n} \sum_{i=1}^{n} h_{i,\tilde{\theta}} \xrightarrow{p} E[h_{i,\theta_0}] = \Sigma_s.$$

Thus if we define

$$Z_{1,n} = \Sigma_s^{-1/2} \tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{i,\theta_0} \quad \text{and} \quad Z_{2,m} = \Sigma_s^{-1/2} \tfrac{1}{\sqrt{m}} \sum_{i=n+1}^{n+m} s_{i,\theta_0},$$

it follows that $Z_{1,n} \xrightarrow{d} Z_1$, as $n \to \infty$, and $Z_{2,m} \xrightarrow{d} Z_2$ where $(Z_1^{\mathrm{T}}, Z_2^{\mathrm{T}})^{\mathrm{T}} \sim N_{2k}(0, I_{2k})$.

A Taylor expansion of the in-sample log likelihood function yields

$$\ell_1(\theta_0) = \ell_1(\hat{\theta}_1) + S^{\mathrm{T}}_{1,\hat{\theta}_1}(\hat{\theta}_1 - \theta_0) + \frac{1}{2}(\hat{\theta}_1 - \theta_0)^{\mathrm{T}} H_{1,\breve{\theta}}(\hat{\theta}_1 - \theta_0), \tag{5}$$

for some $\breve{\theta}_1$, that lies between θ_0 and $\hat{\theta}_1$, such that

$$\ell_1(\hat{\theta}_1) - \ell_1(\theta_0) = \frac{1}{2} S^{\mathrm{T}}_{1,\theta_0}[-H_{1,\theta_0}]^{-1} S_{1,\theta_0} + o_p(1) \xrightarrow{d} \frac{1}{2} Z_1^{\mathrm{T}} Z_1.$$

The out-of-sample score is given by

$$S_{2,\hat{\theta}_1} = S_{2,\theta_0} + H_{2,\breve{\theta}}(\hat{\theta}_1 - \theta_0) = S_{2,\theta_0} + H_{2,\breve{\theta}}[-H_{1,\tilde{\theta}}]^{-1} S_{1,\theta_0}$$

where $\breve{\theta}$ lies between $\hat{\theta}_1$ and θ_0, and we note that $S_{2,\hat{\theta}_1} \neq 0$ almost surely, (unlike the in-sample score $S_{1,\hat{\theta}_1}) = 0$).

Consider now a Taylor expansion of the out-of-sample likelihood

$$\ell_2(\hat{\theta}_1) = \ell_2(\theta_0) + S^{\mathrm{T}}_{2,\theta_0}(\theta_0 - \hat{\theta}_1) + \frac{1}{2}(\hat{\theta}_1 - \theta_0)^{\mathrm{T}} H_{2,\tilde{\theta}}(\hat{\theta}_1 - \theta_0),$$

such that

$$\begin{aligned}
\ell_2(\hat{\theta}_1) - \ell_2(\theta_0) &= S^{\mathrm{T}}_{2,\theta_0}[-H_{1,\tilde{\theta}}]^{-1} S_{1,\theta_0} + \tfrac{1}{2} S^{\mathrm{T}}_{1,\theta_0}[H_{1,\tilde{\theta}}]^{-1}[H_{2,\tilde{\theta}}][H_{1,\tilde{\theta}}]^{-1} S_{1,\theta_0} \\
&= \sqrt{\tfrac{m}{n}} Z^{\mathrm{T}}_{2,m} Z_{1,m} - \tfrac{1}{2}\tfrac{m}{n} Z^{\mathrm{T}}_{1,m} Z_{1,m} + o_p(1). \\
&= \sqrt{\tfrac{m}{n}} Z_2^{\mathrm{T}} Z_1 - \tfrac{1}{2}\tfrac{m}{n} Z_1^{\mathrm{T}} Z_1 + o_p(1).
\end{aligned}$$

■