A Winner's Curse for Econometric Models: On the Joint Distribution of In-Sample Fit and Out-of-Sample Fit and its Implications for Model Selection

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

Preliminary version: September 28, 2010

Abstract

We consider the case where a parameter, θ, is estimated by maximizing a criterion function, $Q(\mathcal{X}, \theta)$. The estimate, $\hat{\theta} = \hat{\theta}(\mathcal{X})$, 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 the true, or quasi-true, parameter, θ^*, are defined by $\eta = Q(\mathcal{X}, \hat{\theta}) - Q(\mathcal{X}, \theta^*)$ and $\tilde{\eta} = Q(\mathcal{Y}, \hat{\theta}) - Q(\mathcal{Y}, \theta^*)$, respectively. We derive the joint limit distribution of $(\eta, \tilde{\eta})$ for a broad class of criterion functions and the joint distribution reveals that η and $\tilde{\eta}$ are strongly negatively related. The implication is that good in-sample fit translates into poor out-of-sample fit, one-to-one.

The result exposes a winner's curse problem when multiple models are compared in terms of their in-sample fit. The winner's curse has important implications for model selection by standard information criteria such as AIC and BIC.

Keywords: Out-of-Sample Fit, Model Selection, AIC, BIC, TIC, Forecasting.

*I thank Serena Ng, Joe Romano, Mark Watson, Kenneth West, and participants at the 2006 SITE 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 of the estimated model. Perhaps the most 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 necessarily one that explains all the variation in the sample. In this case one may view the general population (less the sample used for 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 relation between the in-sample fit and the out-of-sample fit of a model, in a general framework. This exposes a winner's curse that has important implications for model selection by information criteria, because these are shown to have some rather unfortunate and paradoxical properties. Theorem 1 also provides important insight about model averaging and shrinkage methods.

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). For the purpose of model selection, this has motivated the use of information criteria that involve a penalty term for the complexity. The following example serves to illustrate some of the results in this paper.

Example 1 *Let $\mathcal{X} = (X_1, \ldots, X_n)$ and $\mathcal{Y} = (Y_1, \ldots, Y_n)$ represent the in-sample and out-of-sample, respectively. Suppose that $X_i, Y_i \sim iid\, \mathrm{N}(\theta^*, 1)$, $i = 1, \ldots, n$, so that $Z_1 = n^{-1/2} \sum_{i=1}^n (X_i - \theta^*)$ and $Z_2 = n^{-1/2} \sum_{i=1}^n (Y_i - \theta^*)$ are independent standard normal random variables. Using the log-likelihood function, or equivalently the criterion function, $Q(\mathcal{X}, \theta) = -\sum_{i=1}^n (X_i - \theta)^2$, we find that $\hat{\theta} = \hat{\theta}(\mathcal{X}) = \bar{X} = n^{-1} \sum_{i=1}^n X_i$, solves $\max_\theta Q(\mathcal{X}, \theta)$. The in-sample fit at $\hat{\theta}$ relative to that at the true parameter θ^* is*

$$\eta = Q(\mathcal{X}, \hat{\theta}) - Q(\mathcal{X}, \theta^*) = \left\{ n^{-1/2} \sum_{i=1}^n (X_i - \theta^*) \right\}^2 = Z_1^2,$$

which is distributed as a $\chi^2_{(1)}$. The fact that $Q(\mathcal{X}, \hat{\theta}) > Q(\mathcal{X}, \theta^*)$ (almost surely) is called overfitting, and the expected overfit is $E(\eta) = 1$. The out-of-sample criterion function is more interesting. We have

$$\tilde{\eta} \;=\; Q(\mathcal{Y}, \hat{\theta}) - Q(\mathcal{Y}, \theta^*) = \sum_{i=1}^n (Y_i - \theta^*)^2 - (Y_i - \hat{\theta})^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}(Y_i-\theta^*)^2-(Y_i-\theta^*+\theta^*-\hat{\theta})^2 \\
&= \sum_{i=1}^{n}-(\theta^*-\hat{\theta})^2+2(Y_i-\theta^*)(\hat{\theta}-\theta^*) \\
&= -\left\{n^{-1/2}\sum_{i=1}^{n}(X_i-\theta^*)\right\}^2+2\sum_{i=1}^{n}(Y_i-\theta^*)n^{-1}\sum_{i=1}^{n}(X_i-\theta^*) \\
&= -Z_1^2+2Z_2Z_1.
\end{aligned}
$$

So the out-of-sample relative fit, $\tilde{\eta}$, has a non-standard distribution that involves a product of two independent Gaussian variables minus a χ^2 distributed random variable. We note that the expected in-sample overfit is positive, $\mathrm{E}(\eta)=+1$, *and the converse is true out-of-sample since* $\mathrm{E}(\tilde{\eta})=-1$. *Thus* $\mathrm{E}(\eta-\tilde{\eta})=+2$ *and this difference has motivate Akaike's information criterion (and related criteria) that explicitly make a trade-off between the complexity of a model and how well the model fits the data.*

Our theoretical result sheds additional light on the connection between in-sample overfit and out-of-sample underfit. In the example above, we note that Z_1^2 appears in both expressions with opposite signs. This turns out to be a feature of the limit distribution of $(\eta,\tilde{\eta})$ in a general framework. The connection between η and $\tilde{\eta}$ is therefore far stronger than one of expectations. For instance, in Example 1 we note that the conditional distribution of $\tilde{\eta}$ given $\mathcal{X}$ is $\mathrm{N}(-\eta,4\eta)$, so that

$$\mathrm{E}\left(\tilde{\eta}|\mathcal{X}\right)=-\eta.$$

This shows that in-sample overfitting results in a lower out-of-sample fit – not only in expectation – but one-to-one.

In this paper we derive the joint limit distribution of $(\eta,\tilde{\eta})$ for a general class of criteria, which includes loss functions that are commonly used for the evaluation of forecasts. The limit distribution for the out-of-sample quantity, $\tilde{\eta}$, has features that are similar to those seen in quasi maximum likelihood analysis, see White (1994) for a comprehensive treatment. The limit distribution is particularly simple when an information-matrix type equality holds. This equality holds when the criterion function is a correctly specified likelihood function. When Q is a correctly specified log-likelihood function and $\theta\in\Theta\subset\mathbb{R}^k$ we have an asymptotic multivariate version of the result we found in Example 1, specifically

$$(\eta,\tilde{\eta})\xrightarrow{d}(Z_1'Z_1,-Z_1'Z_1+2Z_1'Z_2),$$

where Z_1 and Z_2 are independent Gaussian distributed random variables, $Z_1,Z_2\sim\mathrm{N}_k(0,I_k)$.

The fact that in-sample overfit translates into out-of-sample underfit has important

implications for model selection. Model selection by standard information criteria, such as AIC and BIC, tend to favor models that have a large η in the sample used for estimation. We shall refer to this as the *winner's curse* of model selection. The winner's curse is particularly relevant in model-rich environments where many models may have a similar expected fit when evaluated at their respective population parameters. So we will argue that standard information criteria are poorly suited for the selecting a model with a good out-of-sample fit in model-rich environments. In the context of forecasting this can explain the empirical success of shrinkage methods and combining models, such as model averaging.

Another implication of the theoretical 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 more parsimonious model out-of-sample. In an out-of-sample comparison, it will take a great deal of luck for an overparameterized model to offset its disadvantage relative to a simpler model, in particular when both models nests the true model. Therefore, when a complex model is found to outperform a simpler model out-of-sample, it is stronger evidence in favor of the larger model, than had the same result been found in-sample (other things being equal).

Parameter instability is an important issue for forecasting, because it may result in major forecast failures, see e.g. Hendry and Clements (2002), Pesaran and Timmermann (2005), and Rossi and Giacomini (2006), and references therein. Interestingly, we will show that a major discrepancy between the empirical in-sample fit and out-of-sample fit can be induced by model selection, even if all parameters are constant. This phenomenon is particularly likely to occur in model rich environments where a model is selected by a conventional model selection method such as AIC or BIC.

2 The Joint Distribution of In-Sample Fit and Out-of-Sample Fit

We consider a situation where the criterion function and estimation problem can be expressed within the framework of extremum estimators/M-estimators, see e.g. 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, $\theta^* = \arg\max_{\theta\in\Theta} Q(\theta)$. We will refer to θ^* as the *true* parameter value. The empirical version of the problem is based on a random criterion function $Q(\mathcal{X}, \theta)$, where $\mathcal{X} = (X_1, \ldots, X_n)$ is the sample used for the estimation. In Example 1 we have, $Q(\theta) = -\mathrm{E}(X_1 - \theta)^2$, whereas the empirical criterion function is $Q(\mathcal{X}, \theta) = -\sum_{t=1}^{n}(X_t - \theta)^2$, so that $\bar{Q}(\mathcal{X}, \theta) = n^{-1}Q(\mathcal{X}, \theta) \xrightarrow{p} Q(\theta)$.

The extremum estimator is defined by

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

and we define $S(\mathcal{X},\theta) = \partial Q(\mathcal{X},\theta)/\partial\theta$ and $H(\mathcal{X},\theta) = \partial^2 Q(\mathcal{X},\theta)/\partial\theta\partial\theta'$. Throughout this paper we let k denote the dimension of θ, so that $\theta \in \Theta \subset \mathbb{R}^k$. We shall 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^*, as $n \to \infty$.*

(i) $H(\mathcal{X},\theta)$ exists and is continuous in an open neighborhood of θ^*,

(ii) $-n^{-1}H(\mathcal{X},\theta) \xrightarrow{p} \mathcal{I}(\theta)$ uniformly in θ in an open neighborhood of θ^*,

(iii) $\mathcal{I}(\theta)$ is continuous in a neighborhood of θ^* and $\mathcal{I}_0 = \mathcal{I}(\theta^*) \in \mathbb{R}^{k\times k}$ is positive definite.

(iv) $n^{-1/2}S(\mathcal{X},\theta^*) \xrightarrow{d} \mathrm{N}(0,\mathcal{J}_0)$, where $\mathcal{J}_0 = \lim_{n\to\infty} \mathrm{E}\left\{n^{-1}S(\mathcal{X},\theta^*)S(\mathcal{X},\theta^*)'\right\}$.

Assumption 1 guarantees that $\hat{\theta}$ (eventually) will be given by the first order condition $S(\mathcal{X},\hat{\theta}) = 0$. 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 (2008).

We have in mind a situation where the estimate, $\hat{\theta}$, is to be computed from n observations, $\mathcal{X} = (X_1,\ldots,X_n)$. The object of interest is $Q(\mathcal{Y},\hat{\theta})$, 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 consider the situation where θ is estimated by maximizing the criterion function in-sample, $Q(\mathcal{X},\cdot)$, and the very same criterion function is used for the out-of-sample evaluation, $Q(\mathcal{Y},\cdot)$. We are particularly interested in the following two quantities

$$\eta = Q(\mathcal{X},\hat{\theta}) - Q(\mathcal{X},\theta^*), \quad \text{and} \quad \tilde{\eta} = Q(\mathcal{Y},\hat{\theta}) - Q(\mathcal{Y},\theta^*).$$

The first quantity, η, is a measure of *in-sample fit* (or in-sample overfit). We have $Q(\mathcal{X},\hat{\theta}) \geq Q(\mathcal{X},\theta^*)$, because $\hat{\theta}$ maximizes $Q(\mathcal{X},\theta)$. In this sense, $Q(\mathcal{X},\hat{\theta})$ will reflect a value that is

[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 $\hat{\theta} = \arg\max_{\theta\in\{\theta:S(\mathcal{X},\theta)=0\}} Q(\mathcal{X},\theta)$.

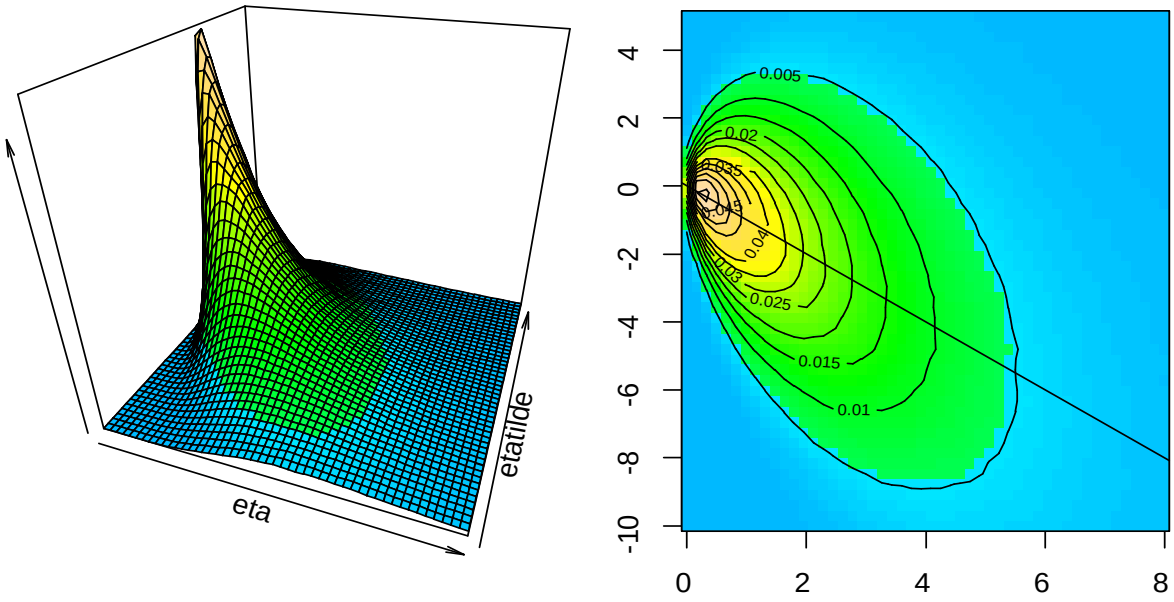

Figure 1: The joint density of $(\eta, \tilde{\eta})$ for the case with $k = 3$ and $\Lambda = I$.

too good relative to that of the true parameter $Q(\mathcal{X}, \theta^*)$, hence the label "overfit". The second quantity, $\tilde{\eta}$, is a measure of *out-of-sample fit*. Unlike the in-sample statistic, there is no guarantee that $\tilde{\eta}$ is non-negative. In fact, $\tilde{\eta}$, will tend to be negative because θ^* is the best ex-ante value for θ. We have the following result concerning the limit distribution of $(\eta, \tilde{\eta})$.

Theorem 1 *Given Assumption 1 and* $\frac{m}{n} \to \pi$, *we have*

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

where $\zeta_1 = Z_1'\Lambda Z_1$, $\zeta_2 = Z_1'\Lambda Z_2$ *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* $\mathcal{I}_0^{-1}\mathcal{J}_0$.

The joint distribution for the case with $k = 3$, $\Lambda = I$, and $\pi = 1$ is plotted in Figure 1. The left panel has the joint density and the right panel is the corresponding contour plot. The plots illustrates the joint distribution of η and $\tilde{\eta}$ and the negative correlation between η and $\tilde{\eta}$ is evident in the contour plot. The downwards sloping line in the contour plot shows the conditional mean, $\mathrm{E}(\tilde{\eta}|\eta) = -\eta$.

Remark. Too good in-sample fit (overfit), $\eta \gg 0$, 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. The reason is that the observed fit can be written as,

$$Q(\mathcal{X}, \hat{\theta}_j) = Q(\mathcal{X}, \theta_j^*) + Q(\mathcal{X}, \hat{\theta}_j) - Q(\mathcal{X}, \theta_j^*) = Q(\mathcal{X}, \theta_j^*) + \eta_j.$$

If several models are approximately equally good, and have roughly the same value of $Q(\mathcal{X}, \theta_j^*)$, then is it quite likely that the best in-sample performance, as defined by $\max_j Q(\mathcal{X}, \hat{\theta}_j)$, is attained by a model with a large η_j, which translated directly into poor out-of-sample fit.

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 and we address this issue 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

In this section we study the case where the criterion function is a correctly specified likelihood function. We denote the log-likelihood function by $\ell(\mathcal{X}, \theta)$, and suppose that $Q(\mathcal{X}, \theta) = 2\ell(\mathcal{X}, \theta)$ where $\theta \in \Theta \subset \mathbb{R}^k$. In this case $\hat{\theta} = \hat{\theta}(\mathcal{X})$ is the maximum likelihood estimator, and in regular problems with a correctly specified likelihood function, it is well known that the likelihood ratio statistic,

$$\text{LR} = \eta = 2\{\ell(\mathcal{X}, \hat{\theta}) - \ell(\mathcal{X}, \theta^*)\},$$

is asymptotically distributed as a χ^2 with k degrees of freedom. So on average, $\ell(\mathcal{X}, \hat{\theta})$ is about $k/2$ larger than the log-likelihood function evaluated at the true parameters, $\ell(\mathcal{X}, \theta^*)$.

It is less known that the converse is true when the log-likelihood function is evaluated out-of-sample. In fact, the asymptotic distribution of the out-of-sample statistic,

$$\widetilde{\text{LR}} = \tilde{\eta} = 2\{\ell(\mathcal{Y}, \hat{\theta}) - \ell(\mathcal{Y}, \theta^*)\},$$

has an expected value that is $-k$, when $\mathcal{X}$ and $\mathcal{Y}$ are independent and identically distributed. Again we see that expected in-sample overfit translates into expected out-of-sample underfit. The out-of-sample log-likelihood function, $\ell(\mathcal{Y}, \hat{\theta})$, is related to the predictive likelihood introduced by Lauritzen (1974). We could call $\ell(\mathcal{Y}, \hat{\theta})$ 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.

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

$$g(x) = f(x;\theta^*), \quad \text{almost everywhere for some } \theta^* \in \Theta \subset \mathbb{R}^k, \tag{1}$$

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, $\hat{\theta} = \arg\max_\theta \ell(\mathcal{X},\theta)$, is given by $\frac{\partial}{\partial\theta}\ell(\mathcal{X},\hat{\theta}) = 0$.

Corollary 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 with* $\eta =$

$$\begin{pmatrix} \mathrm{LR} \\ \widetilde{\mathrm{LR}} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_1'Z_1 \\ 2\sqrt{\pi}Z_1'Z_2 - \pi Z_1'Z_1 \end{pmatrix}, \qquad \text{as } n \to \infty \text{ and } \tfrac{m}{n} \to \pi,$$

where Z_1 and Z_2 are independent with $Z_i \sim \mathrm{N}_k(0, I_k)$, for $i = 1, 2$.

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

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

This expectation motivated the Akaike's information criterion (AIC), see Akaike (1974). The AIC penalty, $2k$, is derived under the assumption that the likelihood function is correctly specified. The proper penalty to use for misspecified models was derived by Takeuchi (1976), who derived this results within the quasi maximum likelihood framework.

Corollary 3 *When $m = n$ the limit distribution of $(\eta, \tilde{\eta})' = (\mathrm{LR}, \widetilde{\mathrm{LR}})'$ has mean $(+k, -k)'$, and variance-covariance matrix,*

$$2\begin{pmatrix} k & -k \\ -k & 3k \end{pmatrix},$$

and the conditional distribution of $\tilde{\eta}$ given η is, in the limit, $N(-\eta, 4\eta)$.

The conditional density of $\tilde{\eta}$ given η is plotted in Figure 2, for various values of η.

An implication is that the unconditional limit distribution of $\tilde{\eta}$ is mixed Gaussian, $\tilde{\eta} \sim N(-\eta, 4\eta)$, with a χ^2-distributed mixing parameter.

The negative correlation between LR and $\widetilde{\text{LR}}$ that we formulated in Corollary 2, offers a theoretical explanation for the so-called *AIC paradox* in a very general setting. Shimizu (1978) analyzed the problem of selecting the order of an autoregressive process, and noted that AIC tends to select too large an order when it is most unfortunate to do so.

2.2 Related Results and Some Extensions

The expected value of $\tilde{\eta}$, as computed from the limit distribution in Theorem 1, is related to results 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 the *expected* MSPE is smaller for the parsimonious model.[2] In our notation, Clark and West are concerned with $\text{E}(\tilde{\eta})$ which increases with the number of regressors in the model. Clark and West (2007) use this finding to motivate a correction of a particular test. The joint distribution of $(\eta, \tilde{\eta})$ reveals some interesting aspects of this problem, and shows that the results in Clark and West (2007) hold in a general framework, beyond the regression models and the MSPE criterion.

Out-of-sample forecast evaluation is often analyzed with different estimation schemes, known as the *fixed*, *rolling*, and *recursive* schemes, see e.g. McCracken (2007). Under the fixed scheme the parameters are estimated once and this point estimate is used throughout the out-of-sample period. In the rolling and recursive schemes the parameter is reestimated every time a forecast is made. The recursive scheme uses all past observations for the estimation, whereas the rolling scheme only uses 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 approximately

$$k \sum_{i=1}^{m} \frac{1}{n+i} \quad = \quad k \frac{1}{m+n} \sum_{s=n+1}^{m+n} \frac{m+n}{s}$$

[2] This feature is also used to motivate and derive Akaike's information criterion.

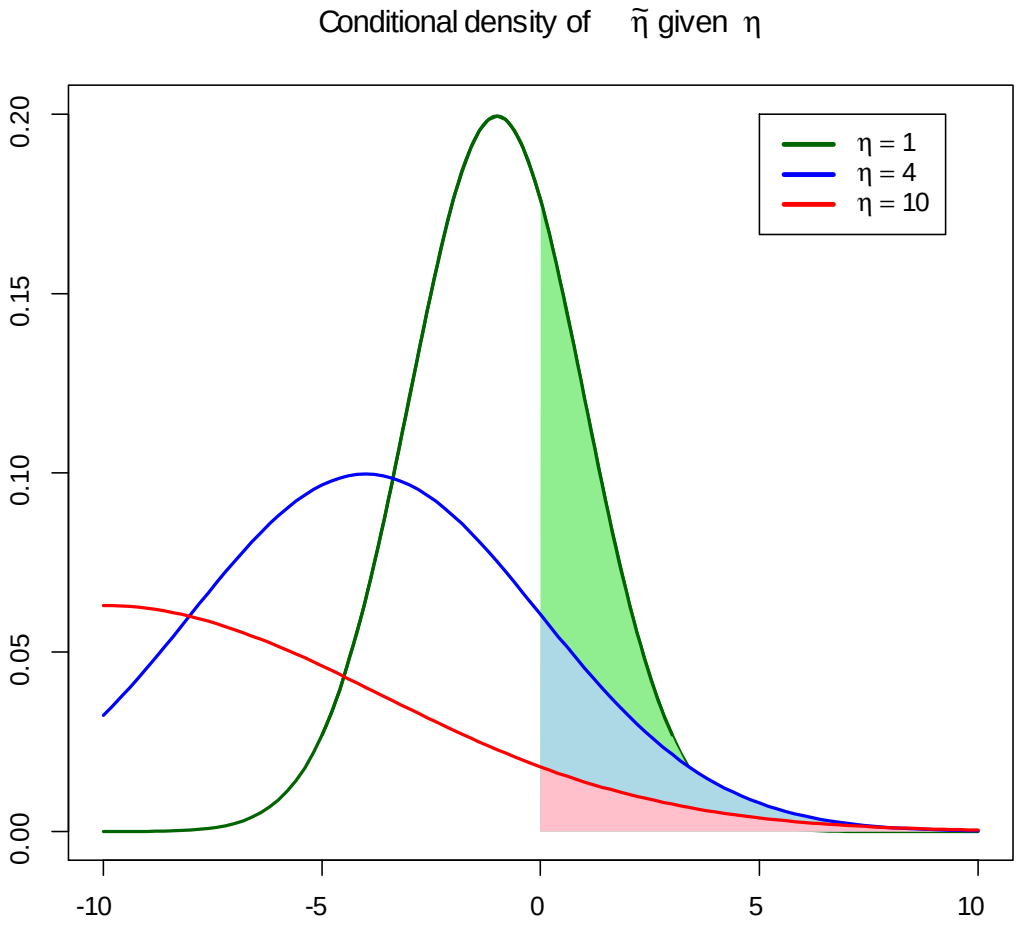

Figure 2: The conditional distribution of $\tilde{\eta}$ given η is, in the limit, $\mathrm{N}(-\eta, 4\eta)$, when Q is a correctly specified log-likelihood function. Here we have plotted conditional density for three values of η. In this case η is the usual in-sample likelihood ratio statistic and $\tilde{\eta}$ can be interpreted as an out-of-sample likelihood ratio statistic.

$$\approx \quad k\int_{\frac{1}{1+\pi}}^{1}\frac{1}{u}\mathrm{d}u \to k\int_{\frac{1}{1+\pi}}^{1}\frac{1}{u}\mathrm{d}u = k\log(1+\pi),$$

where $\pi = \lim_{n\to\infty}\frac{m}{n}$. This is consistent with McCracken (2007) who, in the context of regression models, derived the asymptotic distribution of what can be labelled as an aggregate out–of-sample fit. Given our previous results it is evident that the aggregate out-of-sample fit will be negatively correlated with the aggregate in-sample overfit, yet the joint dependence is more complicated than that of Theorem 1.

3 Implications of Theorem 1

We now turn to a situation where we estimate more than a single model. The relation between models is important in this context. For example the joint distribution of $(\eta_1,\ldots,\eta_m)$, where η_j is the in-sample overfit of the j-th model is important for model selection.

Consider M different models that each have their own "true" parameter value, denoted by θ_j^*. It is useful to think of the different models as restricted version of 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 $\theta_j^* = \arg\max_{\theta\in\Theta_j} Q(\theta)$. We shall assume that Assumption 1 applies to all models, so that $\hat{\theta}_j \xrightarrow{p} \theta_j^*$, where $\hat{\theta}_j = \arg\max_{\theta\in\Theta_j} Q(\mathcal{X},\theta)$. So θ_j^* reflects the best possible ex-ante value for $\theta\in\Theta_j$. The nesting model need not be interesting as a model per se. In many situations 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 small value of $Q(\mathcal{X},\hat{\theta}_j)$ 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:

$$Q(\mathcal{X},\hat{\theta}_j) = \underbrace{Q(\theta_j^*)}_{\text{Genuine}} + \underbrace{Q(\mathcal{X},\theta_j^*) - Q(\theta_j^*)}_{\text{Ordinary noise}} + \underbrace{Q(\mathcal{X},\hat{\theta}_j) - Q(\mathcal{X},\theta_j^*)}_{\text{Deceptive noise}}. \tag{2}$$

We have labelled the two random terms as *ordinary 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, θ_j^*. The second term is pure sampling error that does not depend on $\hat{\theta}$, so this term simply induces a layer of noise that makes it harder to infer $Q(\theta_j^*)$ from $Q(\mathcal{X},\hat{\theta}_j)$. The last term is the culprit. From Theorem 1 we have that $\eta_j = Q(\mathcal{X},\hat{\theta}_j) - Q(\mathcal{X},\theta_j^*)$ is strongly negatively related to $\tilde{\eta}_j = Q(\mathcal{Y},\hat{\theta}_j) - Q(\mathcal{Y},\theta_j^*)$. So η_j is deceiving as it increases the observed criterion function, $Q(\mathcal{X},\hat{\theta}_j)$, while decreasing the expected value of $Q(\mathcal{Y},\hat{\theta}_j)$.

3.1 Model Selection by In-Sample Information Criteria

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, $Q(\mathcal{X}, \theta_j^*) - Q(\theta_j^*)$, and the vicious type $Q(\mathcal{X}, \hat{\theta}_j) - Q(\mathcal{X}, \theta_j^*)$. The latter is the overfit that translate into an out-of-sample underfit, and the implication of this term is that we may *not* want to select the model with the largest value of $Q(\theta_j^*)$. Instead, the best choice is the solution to:

$$\arg\max_j \left\{ Q(\theta_j^*) - \eta_j \right\}.$$

It may seem paradoxical that we would prefer a model that does not (necessarily) explain the in-sample data as well as alternative models, but it is the logical consequence of Theorem 1, specifically the fact that in-sample overfitting translates into out-of-sample underfit.

In a model-rich environment we view this to be 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 better value of $Q(\theta_j^*)$. 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 are 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 $Q(\mathcal{Y}, \hat{\theta}_j)$. 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, $j^* = \arg\max_j Q(\mathcal{X}, \hat{\theta}_j)$, 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, $\eta_j \gg 0$. Since this reduces the expected out-of-sample performance, $Q(\mathcal{Y}, \hat{\theta}_j)$, 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 will often fit the data better than a parsimonious model 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 $Q(\theta_j^*)$. When the observed in-sample criterion function, $Q(\mathcal{X}, \hat{\theta}_j)$, 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,

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

we could very well be selecting the model with the largest sampling error $Q(\mathcal{X}, \hat{\theta}_j) - Q(\mathcal{X}, \theta_j^*)$. 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.* This point is illustrated in the following example.

Example 2 *Suppose we estimate K regression models,*

$$y_i = \beta_j x_{j,i} + \varepsilon_{j,i},$$

by least squares, so that $\hat{\beta}_j = \sum_{i=1}^n x_{j,i} y_i / \sum_{i=1}^n x_{j,i}^2$, $j = 1, \ldots, K$. Here $\beta_j = \mathrm{E}(y_i x_{j,i})/\mathrm{E}(x_{j,i}^2)$ and we let $\theta = (\beta_1, \ldots, \beta_K)'$ and consider the least squares criterion, $Q(\mathcal{X}, \theta) = -\sum_{i=1}^n (y_i - \theta' X_i)^2$. In this setting, θ_j, which is associated with the j-th regression model, is an K-dimensional vector with all but the j-th element being equal to zero.

We have

$$\begin{aligned} -Q(\mathcal{X}, \hat{\theta}_j) &= \sum_{i=1}^n (y_i - \hat{\beta}_j x_{j,i})^2 = \sum_{i=1}^n \varepsilon_{j,i}^2 + (\hat{\beta}_j - \beta_j)^2 x_{j,i}^2 - 2(\hat{\beta}_j - \beta_j) x_{j,i} \varepsilon_{j,i} \\ &= \sum_{i=1}^n \varepsilon_{j,i}^2 - \frac{\left(\sum_{i=1}^n x_{j,i} \varepsilon_{j,i}\right)^2}{\sum_{i=1}^n x_{j,i}^2}, \end{aligned}$$

so that

$$\eta_j = Q(\mathcal{X}, \hat{\theta}_j) - Q(\mathcal{X}, \theta_j^*) = \frac{\left(\sum_{i=1}^n x_{j,i} \varepsilon_{j,i}\right)^2}{\sum_{i=1}^n x_{j,i}^2}.$$

Suppose that $(\varepsilon_i, x_{j,i})$, $i = 1, \ldots, n$, $j = 1, \ldots, K$ are mutually independent, all having a standard normal distribution, and the true model be $y_i = \varepsilon_i$, so that $\varepsilon_{j,i} = \varepsilon_i$ for all j. It follows that $Q(\mathcal{X}, \theta_j^*) = -\sum_{i=1}^n \varepsilon_i^2$ for all j, and we have

$$\begin{pmatrix} \frac{n^{-1/2} \sum_{i=1}^n x_{1,i} \varepsilon_i}{\sqrt{n^{-1} \sum_{i=1}^n x_{1,i}^2}} \\ \vdots \\ \frac{n^{-1/2} \sum_{i=1}^n x_{K,i} \varepsilon_i}{\sqrt{n^{-1} \sum_{i=1}^n x_{K,i}^2}} \end{pmatrix} \xrightarrow{d} N_K(0, I_K),$$

so that the limit distribution of $(\eta_1, \ldots, \eta_K)'$ is a vector of independent $\chi_{(1)}^2$-distributed

random variables.

In our previous notation we have

$$\begin{aligned}
\mu_j &= -\sum_{i=1}^{n} \mathrm{var}(\varepsilon_i) = -n, \\
\nu_j &= \sum_{i=1}^{n} (1 - \varepsilon_i^2), \\
\eta_j &= \sum_{i=1}^{n} (\varepsilon_i^2 - \hat{\varepsilon}_{j,i}^2), \qquad \textit{with} \quad \hat{\varepsilon}_{j,i} = y_i - \hat{\beta}_j x_{j,i}.
\end{aligned}$$

With $m = n$, the out-of-sample criterion is

$$\begin{aligned}
-Q(\mathcal{Y}, \hat{\theta}_j) &= \sum_{i=n+1}^{2n} \varepsilon_i^2 + \hat{\beta}_j^2 x_{j,i}^2 - 2\hat{\beta}_j x_{j,i}\varepsilon_i \\
&= \sum_{i=n+1}^{2n} \varepsilon_i^2 + \frac{\left(\sum_{i=1}^{n} x_{j,i}\varepsilon_i\right)^2}{\sum_{i=1}^{n} x_{j,i}^2} \frac{\sum_{i=n+1}^{2n} x_{j,i}^2}{\sum_{i=1}^{n} x_{j,i}^2} - 2\frac{\sum_{i=1}^{n} x_{j,i}\varepsilon_i \sum_{i=n+1}^{2n} x_{j,i}\varepsilon_i}{\sum_{i=1}^{n} x_{j,i}^2}
\end{aligned}$$

and it follows that

$$\mathrm{AIC}_j = -\sum_{i=1}^{n} \varepsilon_i^2 + \frac{\left(\sum_{i=1}^{n} x_{j,i}\varepsilon_i\right)^2}{\sum_{i=1}^{n} x_{j,i}^2} - 2,$$

is such that $\mathrm{E}(\mathrm{AIC}_j) - \mathrm{E}Q(\mathcal{Y}, \hat{\theta}_j) \to 0$ *as* $n \to \infty$. *However, the* AIC *of the selected model,* $\mathrm{AIC}_{j^*} = \max_j \mathrm{AIC}_j$, *is not an unbiased estimate of its out-of-sample performance* $\mathrm{E}Q(\mathcal{Y}, \hat{\theta}_{j^*})$.

In Example 2 we have the paradoxical outcome that AIC_j picks the model with the worst expected out-of-sample fit, and the model with the best expected out-of-sample fit is the one that minimizes AIC. Table 1 contains the expected value of AIC_{j^*} for $K = 1, \ldots, 20$, the average value of $Q(\mathcal{Y}, \hat{\theta}_{j^*})$, their difference. The average value of the smallest $\mathrm{AIC}_{j^\dagger}$ and its corresponding average value for $Q(\mathcal{Y}, \hat{\theta}_{j^\dagger})$.

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

	Maximum AIC			Minimum AIC	
K	AIC_{max}	$Q(Y, \hat{\theta}_{j^*})$	Bias	AIC_{min}	$Q(Y, \hat{\theta}_{j^\dagger})$
1	-101.00	-101.01	0.01	-101.00	-101.01
2	-100.36	-101.66	1.30	-101.63	-100.37
3	-99.90	-102.13	2.23	-101.80	-100.19
4	-99.54	-102.50	2.97	-101.88	-100.12
5	-99.24	-102.81	3.57	-101.91	-100.08
6	-98.99	-103.07	4.08	-101.94	-100.06
7	-98.77	-103.30	4.53	-101.95	-100.04
8	-98.57	-103.49	4.92	-101.96	-100.03
9	-98.40	-103.67	5.28	-101.97	-100.02
10	-98.24	-103.84	5.60	-101.97	-100.02
11	-98.09	-103.98	5.89	-101.98	-100.01
12	-97.96	-104.12	6.17	-101.98	-100.01
13	-97.83	-104.25	6.42	-101.98	-100.01
14	-97.72	-104.36	6.65	-101.99	-100.01
15	-97.61	-104.48	6.87	-101.99	-100.00
16	-97.51	-104.58	7.07	-101.99	-100.00
17	-97.41	-104.68	7.27	-101.99	-100.00
18	-97.32	-104.77	7.45	-101.99	-100.00
19	-97.23	-104.86	7.63	-101.99	-100.00
20	-97.15	-104.94	7.79	-101.99	-100.00

Table 1: The expected values of the largest and smallest AIC are compute as a function of the number of models, K, along with the corresponding out-of-sample criterion values. In this setup, AIC selects the worst model, whereas the model with the smallest AIC is indeed the best model.

out by Clements and Hendry (1998), see also Hendry (2004). They argue that a model for differences is less sensitive to structural changes in the mean than a model for the level, so the former may be the best choice for forecasting if the underlying process has time-varying parameters.

The literature on model selection: Inoue and Kilian (2006)... Ng and Perron (2005).

3.2 Local Model Asymptotics

[To be completed].

3.3 Resolution to Winner's Curse

Shrinkage and model combination are methods that implicitly dodge the winner's curse problem. Thus methods are helpful in reducing η, which in turn improved the out-of-sample performance. A particular for of shrinkage amounts to adding restrictions on θ, such as $\theta = \theta(\psi)$ where ψ is of lower dimension, and this will tend to reduce η. A drawback is that shrinkage and model combination can reduce μ. For instance, shrinkage of the type above will be useful if there exists a ψ^*, so that $\theta^* = \theta(\psi^*)$. However, if no such ψ^* exits, the value of shrinkage becomes a trade-off between the positive effect it has on η and loss associates with, $Q(\theta^*) - \sup_\psi Q(\theta(\psi)) > 0$.

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 an 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,

$$M^{-1}\sum_{j=1}^{M} Q(\mathcal{X}, \hat{\theta}_j) = M^{-1}\sum_{j=1}^{M} Q(\mathcal{X}, \theta_j^*) + M^{-1}\sum_{j=1}^{M}\{Q(\mathcal{X}, \hat{\theta}_j) - Q(\mathcal{X}, \theta_j^*)\}. \tag{3}$$

Suppose that model averaging simply amounts to take the average criterion function (it does

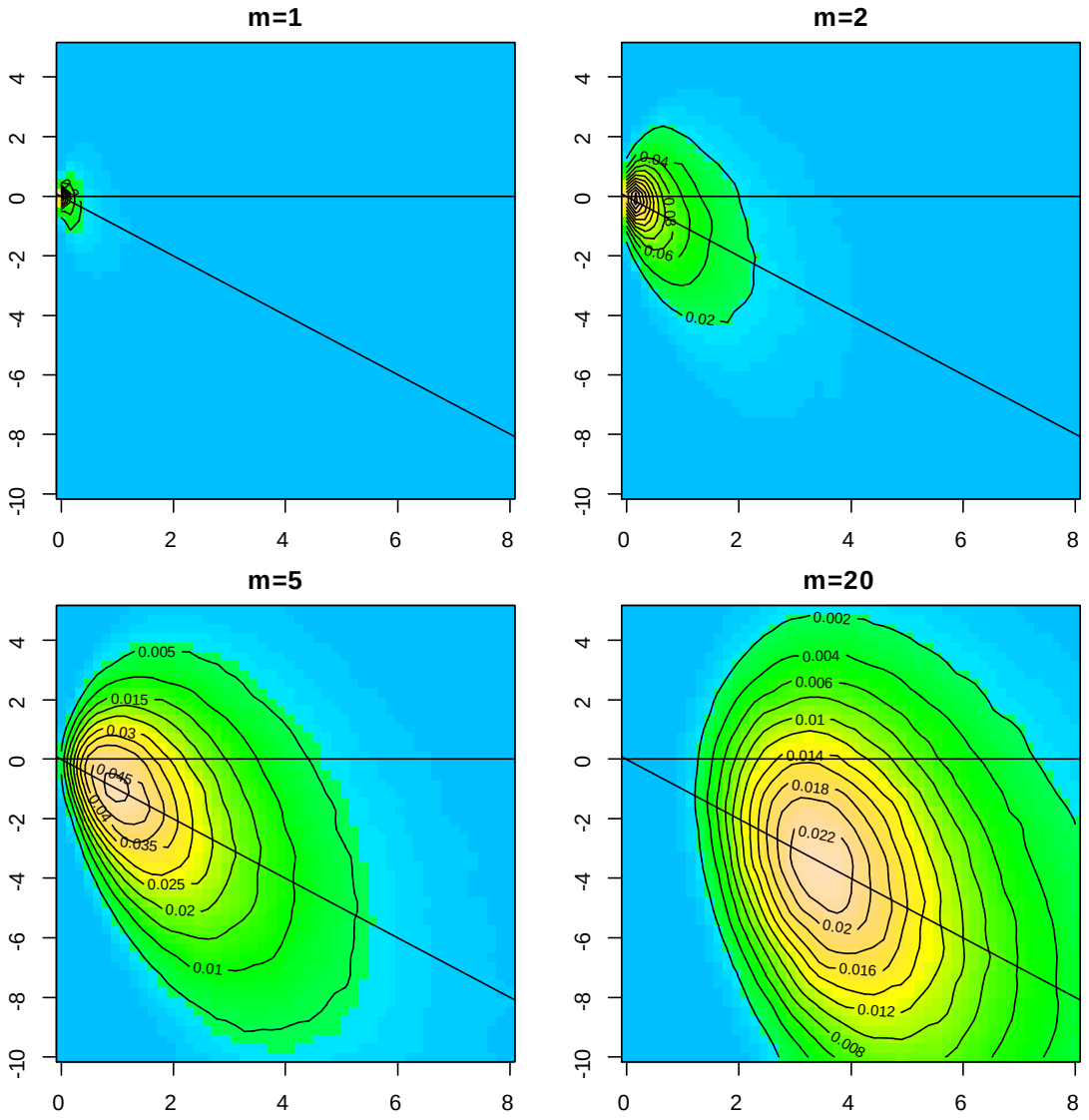

Figure 3: Winner's curse of model selection illustrated by contour plots for the joint distribution of $(\eta_{j^*}, \tilde{\eta}_{j^*})$, where $j^* = \arg\max_{j=1,\ldots,m} \eta_j$.

not). The last term in (3) is trivially smaller than the largest deceptive term, $\min_j\{Q(\mathcal{X},\hat{\theta}_j)-Q(\mathcal{X},\theta_j^*)\}$. Therefore, if the models are similar in terms of $Q(\mathcal{X},\theta_j^*)$, 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.

4 Empirical Application

We present empirical results for three problems. The first application studies the term structure of interest rates, and will illustrate the connection between η and $\tilde{\eta}$. The second considers the forecasting problem using the Stock and Watson data that consists of 131 macro economic variables, see Stock and Watson (2005). This application will demonstrate the severity of the winner's curse. The third application studies a portfolio selection problem. Simulating time series of returns, using a design based on empirical estimates from Jobson and Korkie (1980), we seek the portfolio weights that maximizes certainty equivalent returns. This application will illustrate that shrinkage can substantially improve the out-of-sample performance, because it reduces the overfitting problem.

4.1 An Empirical Illustration: VAR for the US Term Structure

Let X_t denote a vector of interest rates with five different maturities, 3, 6, 12, 60, 120 months. The monthly time series of interest rates were downloaded from the Federal Reserve Economic Data (FRED). (TB3MS, TB6MS, GS1, GS5, and GS10). The time-series span the period 1959:01–2008:05. We estimate the cointegration vector autoregressive (VAR) model,

$$\Delta X_t = \alpha\beta' X_{t-1} + \sum_{j=1}^{p-1}\Gamma_j \Delta X_{t-j} + \mu + \varepsilon_t,$$

using different laglength, $p = 1,\ldots,12$, and different cointegration rank $r = 0,\ldots,5$. The VARs are estimated by least squares, which is equivalent to maximum likelihood when a Gaussian specification is used, see Johansen (1991).

Rather than estimating the parameters with the full sample we divided the sample into odd months, $\mathcal{T}_{\text{odd}}$, and even months, $\mathcal{T}_{\text{even}}$, and estimate the parameters, $\theta = (\alpha,\beta,\Gamma_1,\ldots,\Gamma_{p-1},\mu)$, by maximizing, either

$$Q_{\text{odd}}(\cdot,\theta) = -\frac{T_{\text{odd}}}{2}\log\left|\frac{1}{T_{\text{odd}}}\sum_{t\in\mathcal{T}_{\text{odd}}}\varepsilon_t\varepsilon_t'\right|,$$

or

$$Q_{\text{even}}(\cdot,\theta) = -\frac{T_{\text{even}}}{2}\log\left|\frac{1}{T_{\text{even}}}\sum_{t\in\mathcal{T}_{\text{even}}}\varepsilon_t\varepsilon_t'\right|,$$

where $\varepsilon_t = \Delta X_t - \alpha\beta' X_{t-1} - \sum_{j=1}^{p-1}\Gamma_j\Delta X_{t-j} - \mu$, and with T_{odd} and T_{even} being the cardinality of $\mathcal{T}_{\text{odd}}$ and $\mathcal{T}_{\text{even}}$, respectively. We only include observations from 1960:01 and onwards in $\mathcal{T}_{\text{odd}}$ and $\mathcal{T}_{\text{even}}$, such that we always have a sufficient number of initial observations for $p = 1,\ldots,12$. This is done such that it makes sense to compare the log-likelihoods for different values of p.

Let $\hat{\theta}_{\text{odd}}$ and $\hat{\theta}_{\text{even}}$ denote the two sets of parameter estimates. The in-sample fits, $Q_{\text{odd}}(\cdot,\hat{\theta}_{\text{odd}})$ and $Q_{\text{even}}(\cdot,\hat{\theta}_{\text{even}})$, are reported in the upper panel of Table 2, and the corresponding out-of-sample fits, $Q_{\text{odd}}(\cdot,\hat{\theta}_{\text{even}})$ and $Q_{\text{even}}(\cdot,\hat{\theta}_{\text{odd}})$, are reported in the lower panel of Table 2. Interestingly, the best out-of-sample fit is provided by $(p,r) = (2,5)$ in both cases. For comparison, AIC and BIC selects (p,r) to be $(10,2)$ and $(2,0)$ respectively, for the odd sample and $(10,4)$ and $(1,3)$ respectively, for the odd sample. The AIC and BIC statistics are reported in Table 7. The AIC and BIC statistics in Table 7 are (compared with the conventional way of computing these statistics) scaled by minus a half to make them directly comparable with out-of-sample criterion.

The (column-wise) increments in $Q(\cdot,\cdot)$ as the laglength, p, is increased in steps of one, are reported in Table 8. Theorem 1 predict a linear relationship between the in-sample and out-of-sample increments. Figure 4 provides a scatter plot of these increments, for using increments where the smaller model is always $p \geq 3$.

4.2 Forecasting macroeconomic variables: The winners curse

In this section we analyze the 131 macro economic time series from Stock and Watson (2005). We estimate a relatively simple benchmark model, and compare the out-of-sample performance of this model to a model that adds an additional regressor. The regressor being added is the one that improves the in-sample fit the most.

From $X_{i,t}$, $i = 1,\ldots,131$ macro economic variables, we first compute the principal components, $\text{PC}_{i,t}$ using data for the period 1960:01-1994:12.

The benchmark prediction model for each of the variables is given by

$$\hat{X}_{i,t+h} = \alpha + \beta X_{i,t} + \gamma \text{PC}_{1,t},$$

with $h = 12$, such that we consider the problem of one-year-ahead prediction. The parameters, α, φ and γ are estimated by least squares over the in-sample period, 1960:01-1994:12.

In-sample log-likelihood criterion

	Odd months						Even months					
	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
$p=1$	2832.53	2850.34	2863.63	2871.16	2874.11	2874.76	2886.80	2942.69	2966.24	2980.96	2987.08	2987.19
$p=2$	2909.84	2925.03	2938.58	2948.22	2951.38	2952.70	2952.67	3000.73	3029.78	3039.33	3048.10	3048.32
$p=3$	2957.21	2970.95	2982.23	2989.19	2993.22	2993.63	2994.89	3041.74	3065.98	3073.42	3079.72	3079.91
$p=4$	2993.48	3006.75	3017.18	3024.61	3028.73	3029.46	3041.23	3071.17	3095.28	3102.98	3109.33	3109.58
$p=5$	3029.50	3040.88	3048.26	3054.65	3057.46	3058.19	3057.51	3085.59	3106.87	3114.13	3120.43	3120.59
$p=6$	3059.57	3071.17	3080.75	3086.68	3089.42	3090.15	3082.60	3107.34	3122.93	3130.52	3135.51	3135.66
$p=7$	3083.91	3097.40	3104.39	3109.40	3113.22	3113.75	3133.18	3156.76	3171.43	3178.24	3182.07	3182.20
$p=8$	3114.24	3129.37	3138.62	3142.61	3145.66	3146.02	3176.28	3194.68	3210.81	3218.15	3222.01	3222.15
$p=9$	3147.07	3161.07	3174.01	3178.79	3179.84	3179.90	3205.72	3227.06	3241.97	3250.05	3254.92	3254.97
$p=10$	3172.11	3188.65	3199.05	3203.53	3204.56	3204.72	3247.27	3269.36	3287.38	3296.67	3299.91	3299.93
$p=11$	3192.76	3211.83	3222.42	3226.99	3227.48	3227.56	3269.46	3291.57	3307.26	3316.55	3320.13	3320.16
$p=12$	3217.52	3238.75	3248.78	3252.82	3253.10	3253.10	3294.72	3314.26	3329.86	3338.01	3342.79	3342.79

Out-of-sample log-likelihood criterion

	Odd months						Even months					
	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
$p=1$	2830.33	2830.81	2832.88	2843.65	2848.35	2848.98	2884.78	2897.34	2923.88	2941.21	2950.63	2951.93
$p=2$	2851.67	2853.65	2859.11	2866.41	2873.09	2874.13	2888.99	2901.24	2924.67	2937.93	2951.44	2952.66
$p=3$	2848.93	2845.34	2846.07	2852.75	2856.10	2856.68	2888.64	2888.19	2902.29	2913.50	2918.82	2921.47
$p=4$	2804.19	2816.66	2825.13	2833.62	2834.61	2835.36	2840.13	2847.97	2866.98	2876.32	2876.13	2879.63
$p=5$	2810.55	2815.53	2819.40	2822.51	2824.16	2824.78	2832.48	2834.77	2838.72	2849.55	2849.50	2853.05
$p=6$	2801.23	2802.30	2813.23	2815.58	2816.87	2817.64	2821.25	2819.77	2830.57	2837.17	2838.54	2841.02
$p=7$	2778.48	2779.66	2791.36	2794.20	2797.97	2798.59	2834.37	2826.16	2834.93	2839.22	2840.04	2842.28
$p=8$	2754.54	2755.37	2763.80	2768.54	2773.15	2773.98	2797.69	2791.81	2801.41	2799.91	2804.68	2805.98
$p=9$	2742.00	2742.41	2751.40	2755.05	2759.44	2760.06	2759.62	2765.29	2765.31	2763.76	2767.40	2768.22
$p=10$	2705.72	2706.90	2712.77	2715.22	2719.79	2720.18	2727.39	2733.59	2733.77	2733.72	2737.91	2738.88
$p=11$	2711.91	2710.92	2713.45	2716.26	2722.15	2722.60	2711.75	2714.78	2716.05	2716.76	2719.00	2720.85
$p=12$	2694.56	2692.79	2691.94	2695.15	2699.25	2699.16	2669.73	2671.07	2674.09	2676.57	2678.54	2678.88

Table 2: This table reports values of the in-sample criterion function (upper panels) and out-of-sample criterion function (lower panels) for different values of (p, r). The criterion function was maximized for the "odd" observation in the left panels and for the "even" observations in the right panels. AIC selects (p, r) to be (10,2) for the "odd" sample and (10,4) for the "even" sample, whereas BIC selects (2,0) and (1,3).

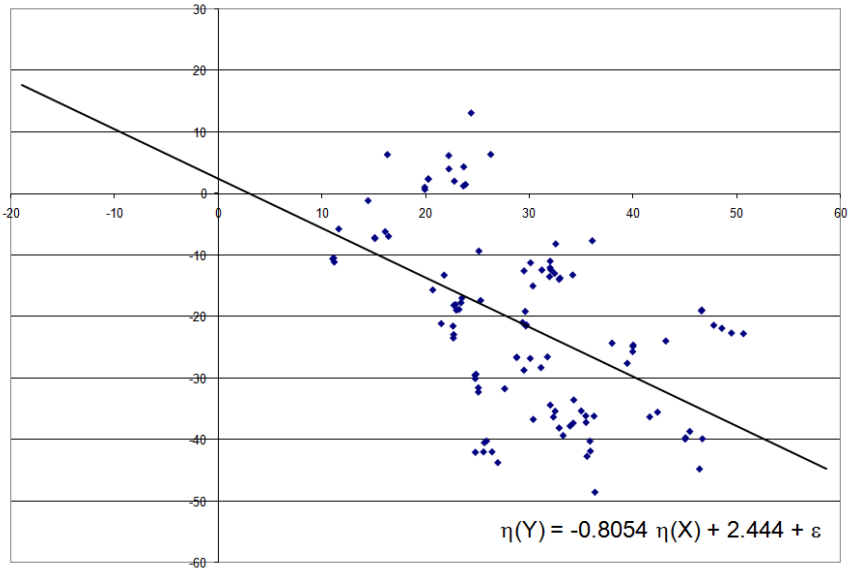

Figure 4: Changes in the out-of-sample fit plotted against the corresponding change in in-sample fit, that results from adding one lag to the VAR, starting with $p = 3$. We have nine observations for each of the two subsamples and each of the six possible values for r.

The larger model includes an additional regressor,

$$\hat{X}_{i,t+h} = \alpha + \beta X_{i,t} + \gamma \mathrm{PC}_{1,t} + \psi Z_t,$$

where Z_{t-1} is chosen from the pool of 260 regressors, that consists of the other 130 macro variables and the other 130 principal components, i.e. , $Z_{t-1} = X_{j,t-1}$ with $j \neq i$, or $Z_{t-1} = \mathrm{PC}_{j,t-1}$, $j \geq 2$. The parameters of this model are also estimated by least squares.

We evaluate the in-sample and out-of-sample residual sum of square

$$\hat{\sigma}^2_{\mathcal{X}} = n^{-1} \sum_{t=1}^{n} \hat{\varepsilon}_t^2 \qquad \text{and} \qquad \hat{\sigma}^2_{\mathcal{Y}} = m^{-1} \sum_{t=n+1}^{n+m} \hat{\varepsilon}_t^2.$$

Stock and Watson (2005) focus on the nine series in Table 3: PI, IP, UR, EMP, TBILL, TBOND, PPI, CPI, PCED.

We note the winners curse in Figure that is a scatter plot of $\Delta Q_{\mathcal{Y}}$ against $\Delta Q_{\mathcal{X}}$.

Figure 5 presents the result for all 131 variables. This figure is a scatter plot of the percentage change in out-of-sample fit relative to the percentage change of in-sample fit. We note the strong negative relation, as illustrated by the estimated regression line.

	$\hat{\sigma}^2_X$	$\hat{\sigma}^2_Y$	$\hat{\sigma}^2_{*,X}$	ΔQ_X	$\hat{\sigma}^2_{*,Y}$	ΔQ_Y
PI	3.61	2.95	2.75	27.21%	4.19	-34.98%
IP	21.02	10.38	11.96	56.36%	12.09	-15.22%
UR	1.02	0.26	0.55	62.44%	0.56	-76.75%
EMP	46.67	25.05	36.06	25.78%	34.39	-31.70%
TBILL	2.75	1.54	2.41	13.28%	2.41	-45.04%
TBOND	1.21	0.62	0.95	24.53%	0.44	35.61%
PPI	26.57	24.13	23.48	12.35%	24.61	-1.94%
PCI	10.79	14.76	10.27	4.92%	14.71	0.35%
PCED	7.52	3.17	6.96	7.82%	3.32	-4.57%

Table 3: The average residual sum of squares for the benchmark model and extended model. The extended model substantially improves the in-sample fit, whereas the out-of-sample fit tends to be substantially worse than that of the benchmark. Among the nine variables, the largest percentage in-sample improvement is found for the unemployment rate, UR, +62.44%. This is also the variable where the out-of-sample fit deteriorates the most, −76.75%

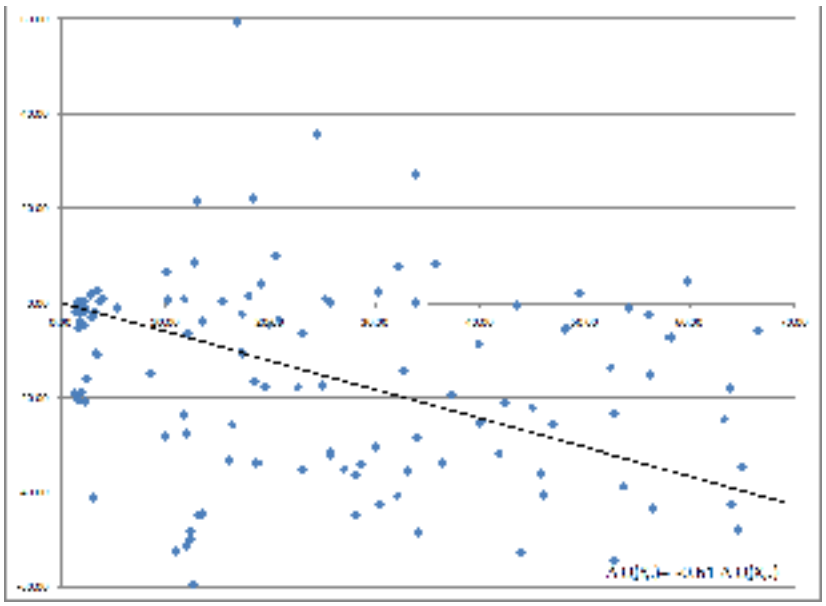

Figure 5: A scatter plot of the percentage reduction in the out-of-sample MSE plotted against the percentage reduction of the in-sample MSE.

4.3 Portfolio Choice

In this section we consider a standard portfolio choice problem where the overfitting is known to be very problematic. This problem will illustrate three issues. First it will show that the overfitting problem can be worse in small samples. The basic reason is that our asymptotic result in Theorem 1 relies on certain quantities having converged to their probability limit, which is the same in-sample and out-of-sample. However, in finite samples the may be a sizable difference between the relevant in-sample and out-of-sample quantities. Second, we will use the portfolio choice problem to illustrate the means by which shrinkage is beneficial as well as the drawbacks associated with shrinkage. Third, adding constraints to the optimization problem is a way to reduce the overfitting problem, and because overfitting is very problematic in this setting, almost any form of restriction will tend to improve the out-of-sample fit. Thus, the observation that a particular constraint is helpful need not be evidence that the imposed structure has a deeper meaning. The main point here is that in empirical applications where the overfitting problem is large, one might be prone to think that a given structure has a deeper explanation, because it is found to be very useful out-of-sample. However, such conclusions may be spuriously driven by the overfitting problem.

Let X_t be an N-dimensional vector of returns in period t, and consider the case where $X_t \sim \text{iid}\, \mathrm{N}_N(\mu, \Sigma)$, for $t = 1, \ldots, T$. Suppose that the criterion is to maximize certainty equivalent returns. Formally, the problem is

$$\max_{w \in \mathbb{R}^N} w'\mu - \tfrac{\gamma}{2} w'\Sigma w, \qquad \text{subject to } \iota' w = 1,$$

in the absence of a risk-free asset, while in the presence of a risk-free asset the problem is given by

$$\max_{w \in \mathbb{R}^N} w_0\mu_0 + w'\mu - \tfrac{\gamma}{2} w'\Sigma w, \qquad \text{subject to } \iota' w = 1 - w_0.$$

The solutions to these two problems are well known and given by

$$w^* = \Sigma^{-1}\left(\mu\tfrac{1}{\gamma} + \iota\tfrac{1-\iota\Sigma^{-1}\mu/\gamma}{\iota'\Sigma^{-1}\iota}\right) \quad \text{and} \quad w^* = \gamma^{-1}\Sigma^{-1}(\mu - \mu_0\iota),$$

respectively.

The empirical criterion function is given by

$$Q(\mathcal{X}, w) = \sum_{t=1}^{T} w'X_t - \tfrac{\gamma}{2} w' \sum_{t=1}^{T} (X_t - \bar{X})(X_t - \bar{X})' w.$$

Here T plays the role of n, and the average in-sample certainty equivalent return may be

defined by $\bar{Q}(\mathcal{X}, w) = \frac{1}{T}Q(\mathcal{X}, w)$.

Using empirical estimates taken from Jobson and Korkie (1980). They estimated the mean and variance-covariance matrix for 20 randomly selected assets using monthly for the sample period: December, 1949 to December 1975. We use their empirical estimates as the population parameters in our simulations. Our results are based on 100,000 simulations, and we set $\gamma = 2/30$ that results in reasonably values of the CER.

First we consider the case with five assets. Table 4 presents results for this case using various sample sizes. Table 5 presents the corresponding results for the case with 20 assets, where the overfitting problem is more severe. It takes a ridiculously large sample for the empirically chosen portfolio, $\hat{w}$, to produce better CER out-of-sample than the equi-weighted portfolio.

Overfitting can be reduced by shrinkage methods. We shrink the unrestricted estimator by imposing the constraint

$$\frac{\|\hat{w}_c - e\|_2}{\|\hat{w} - e\|_2} \leq c, \qquad \text{with } \|x\|_2 = \sqrt{x'x} \quad \text{and } c \geq 0,$$

where e denotes the equi-weighted portfolio, i.e. $e_i = \frac{1}{N}$ for all $i = 1, \ldots, N$. The solution to the constrained optimization problem is simply $\hat{w}_c = c\hat{w} + (1-c)e$. Imposing constraints affects the value of the population parameter. In this case, the population parameter under c-shrinkage is given by $w_c^* = cw^* + (1 - \ c)e$, for $c \leq 1$ and $w_c^* = w^*$ for $c > 1$. Naturally, we have $\bar{Q}(w_c^*) \leq \bar{Q}(w^*)$ and this reduction of the criterion function at the population parameters is the drawback of shrinkage. The advantages of shrinkage is that it reduces the overfit. The smaller is c, the more concentrated is the distribution of η_c near zero. This in turn reduced the out-of-sample underfit, and the question is whether the gains in $\tilde{\eta}_c = Q(\mathcal{Y}, \hat{w}_c) - Q(\mathcal{Y}, w_c^*)$ are sufficiently large to offset the reduction in the population criterion function.

For simplicity we focus on the case without a risk-free asset. The average in-sample CER, $\bar{Q}(\mathcal{X}, \hat{w}_c)$, and out-of-sample CER, $\bar{Q}(\mathcal{Y}, \hat{w}_c)$, are presented in Figure 6, along with the average in-sample overfit in CER, defined by η_c/T.

5 Estimation

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

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

Without a risk-free asset ($N = 5$)

T	η	$\tilde{\eta}$	η/T	$\tilde{\eta}/T$	$\bar{Q}(\mathcal{X}, w^*)$	$\bar{Q}(\mathcal{X}, \hat{w})$	$\bar{Q}(\mathcal{Y}, \hat{w})$	$\bar{Q}(\mathcal{Y}, w_e)$
60	36.67	-43.73	0.61	-0.73	0.34	0.95	-0.38	0.07
120	34.92	-37.89	0.29	-0.32	0.34	0.63	0.02	0.06
180	34.24	-36.23	0.19	-0.20	0.34	0.53	0.14	0.06
240	34.06	-35.42	0.14	-0.15	0.33	0.48	0.19	0.05
360	33.68	-34.45	0.09	-0.10	0.33	0.43	0.24	0.05
480	33.69	-34.32	0.07	-0.07	0.33	0.40	0.26	0.05
600	33.49	-34.12	0.06	-0.06	0.33	0.39	0.27	0.05
1200	33.54	-33.54	0.03	-0.03	0.33	0.36	0.30	0.05
6000	33.38	-33.48	0.01	-0.01	0.33	0.34	0.33	0.05

With a risk-free asset ($N = 5$)

T	η	$\tilde{\eta}$	η/T	$\tilde{\eta}/T$	$\bar{Q}(\mathcal{X}, w^*)$	$\bar{Q}(\mathcal{X}, \hat{w})$	$\bar{Q}(\mathcal{Y}, \hat{w})$	$\bar{Q}(\mathcal{Y}, \hat{w}_e)$
60	45.29	-56.71	0.75	-0.95	0.42	1.17	-0.53	0.17
120	42.53	-47.32	0.35	-0.39	0.41	0.77	0.02	0.25
180	41.49	-44.53	0.23	-0.25	0.41	0.64	0.17	0.27
240	41.18	-43.39	0.17	-0.18	0.41	0.58	0.23	0.28
360	40.62	-41.86	0.11	-0.12	0.41	0.52	0.29	0.29
480	40.56	-41.60	0.08	-0.09	0.41	0.49	0.32	0.30
600	40.31	-41.25	0.07	-0.07	0.41	0.48	0.34	0.30
1200	40.38	-40.61	0.03	-0.03	0.41	0.44	0.38	0.31
6000	40.08	-40.28	0.01	-0.01	0.41	0.42	0.40	0.31

Table 4: Certainty equivalent return (CER) using different portfolio choices with $N = 5$ assets and different sample sizes that are listed in the first column. The average in-sample overfit and out-of-sample underfit in Q are reported in columns two and three. These translate into overfit and underfit in CER are η/T and $\tilde{\eta}/T$, respectively. So η/T measures how much overfitting inflates the in-sample CER. The last four columns report CER for the (infeasible) optimal portfolio weights, w^*, the empirical weights, $\hat{w}$, and equal weights, w_e.

Without a risk-free asset ($N = 20$)

T	η	$\tilde{\eta}$	η/T	$\tilde{\eta}/T$	$\bar{Q}(\mathcal{X}, w^*)$	$\bar{Q}(\mathcal{X}, \hat{w})$	$\bar{Q}(\mathcal{Y}, \hat{w})$	$\bar{Q}(\mathcal{Y}, w_e)$
60	234.42	-540.35	3.91	-9.01	0.86	4.76	-8.15	0.43
120	186.06	-268.06	1.55	-2.23	0.85	2.40	-1.38	0.42
180	174.23	-220.70	0.97	-1.23	0.85	1.82	-0.38	0.42
240	169.06	-201.31	0.70	-0.84	0.85	1.55	0.01	0.42
360	164.08	-184.04	0.46	-0.51	0.85	1.30	0.34	0.42
480	161.89	-176.24	0.34	-0.37	0.85	1.19	0.48	0.42
600	160.30	-171.30	0.27	-0.29	0.85	1.11	0.56	0.42
1200	157.32	-162.72	0.13	-0.14	0.84	0.98	0.71	0.42
6000	155.23	-156.37	0.03	-0.03	0.85	0.87	0.82	0.42

With a risk-free asset ($N = 20$)

T	η	$\tilde{\eta}$	η/T	$\tilde{\eta}/T$	$\bar{Q}(\mathcal{X}, w^*)$	$\bar{Q}(\mathcal{X}, \hat{w})$	$\bar{Q}(\mathcal{Y}, \hat{w})$	$\bar{Q}(\mathcal{Y}, \hat{w}_e)$
60	266.52	-667.30	4.44	-11.12	0.89	5.33	-10.23	0.30
120	206.31	-309.14	1.72	-2.58	0.88	2.60	-1.69	0.37
180	191.91	-249.55	1.07	-1.39	0.88	1.94	-0.51	0.40
240	185.53	-225.31	0.77	-0.94	0.88	1.65	-0.06	0.41
360	179.54	-203.95	0.50	-0.57	0.88	1.37	0.31	0.42
480	176.83	-194.38	0.37	-0.40	0.88	1.25	0.47	0.43
600	174.96	-188.47	0.29	-0.31	0.88	1.17	0.56	0.43
1200	171.35	-177.90	0.14	-0.15	0.87	1.02	0.73	0.44
6000	168.80	-170.36	0.03	-0.03	0.87	0.90	0.85	0.44

Table 5: Certainty equivalent return (CER) using different portfolio choices with $N = 20$ assets and different sample sizes that are listed in the first column. The average in-sample overfit and out-of-sample underfit in Q are reported in columns two and three. These translate into overfit and underfit in CER are η/T and $\tilde{\eta}/T$, respectively. So η/T measures how much overfitting inflates the in-sample CER. The last four columns report CER for the (infeasible) optimal portfolio weights, w^*, the empirical weights, $\hat{w}$, and equal weights, w_e. For the case with a risk-free asset, the ratio of wealth invested in the risk-free asset is chosen empirically.

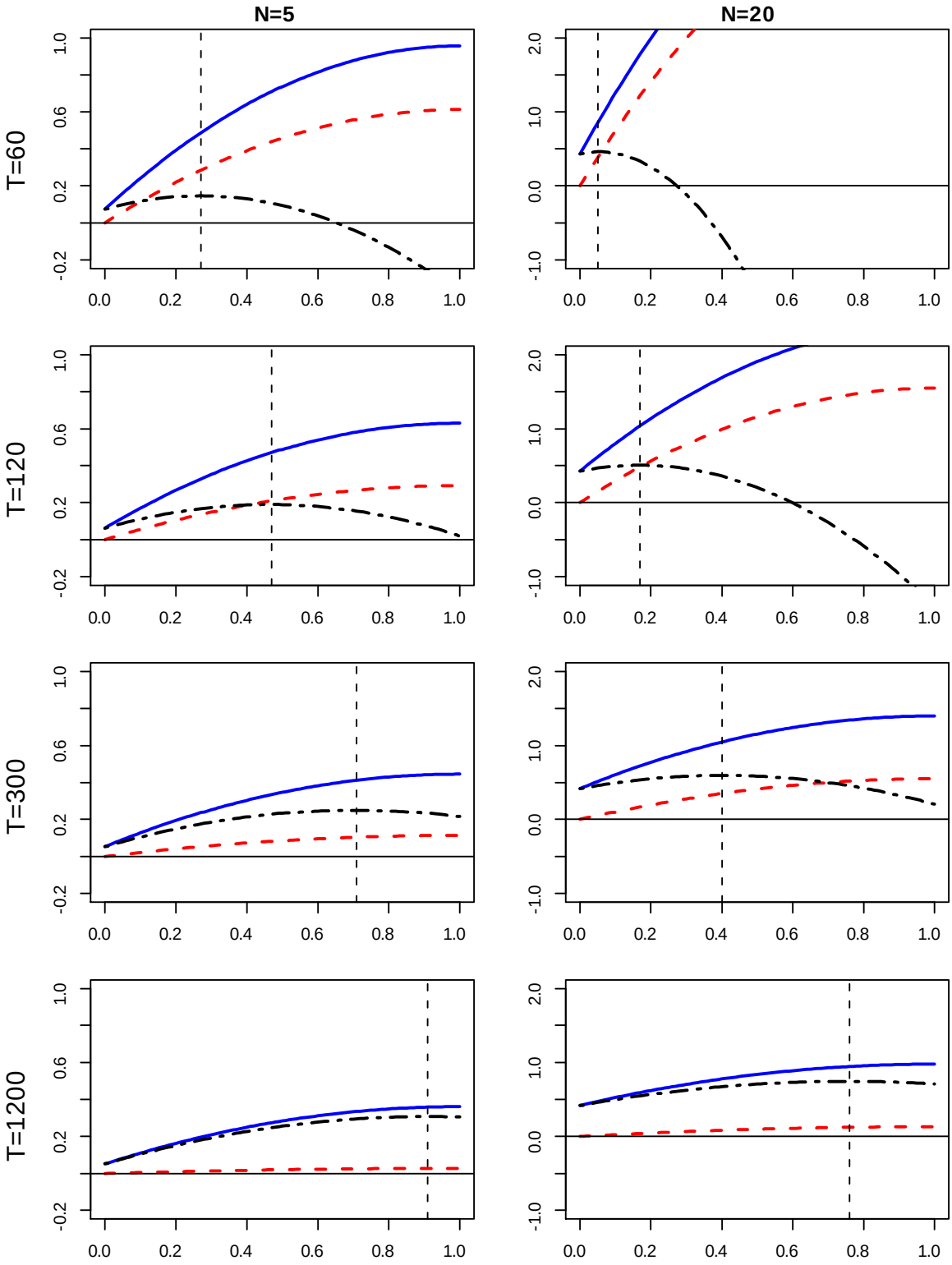

Figure 6: Average certainty equivalent returns obtained in-sample and out-of-sample with $N = 5$ and $N = 20$ and four different sample sized. The value of the skrinkage parameter, c, is given by the x-axis. The solid line is the in-sample CER, $\bar{Q}(\mathcal{X}, \hat{\theta}_c)$, the dashed line is the average in-sample overfit η_c, and the dash-dotted line is the out-of-sample CER, $\bar{Q}(\mathcal{Y}, \hat{\theta}_c)$. The vertical lines identifies the value of c that maximizes the out-of-sample CER.

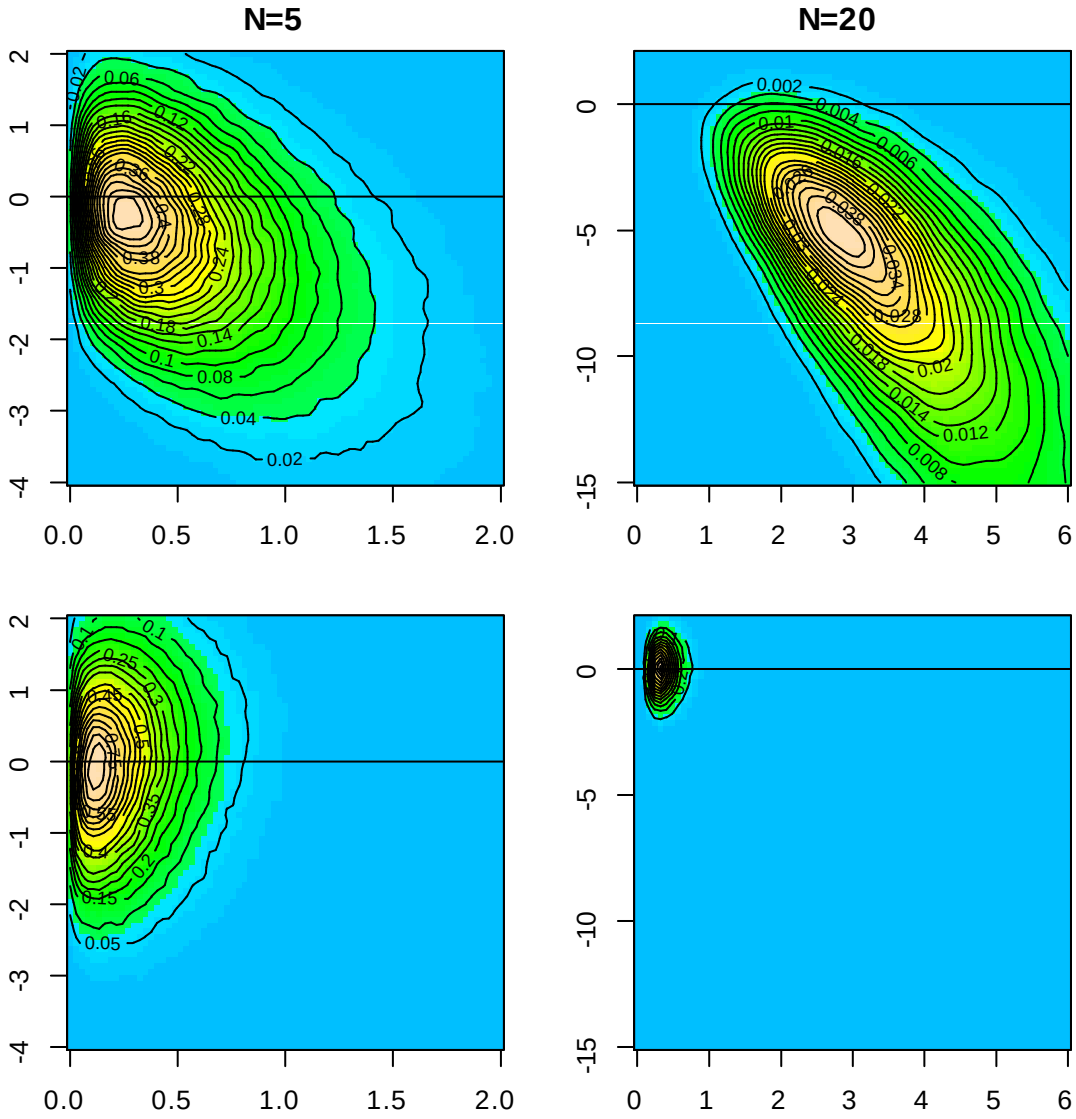

Figure 7: Sample size $T = 60$. The joint distribution of $(\eta/T, \tilde{\eta}/T)$ for unrestricted portfolio weights are given in the two upper panels. The lower panels illustrates the joint distribution of $(\eta_c/T, \tilde{\eta}_c/T)$ where the portfolio weights are the solution to a constrained optimization problem, which essentially shrinks the unrestricted weights towards an equi-weighted portfolio.

evaluated at the true parameter value, $s_t(x_t, \theta^*)$, 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 $x_t = (X_t, X_{t-1})'$ 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),

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

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 $\hat{\theta}(\mathcal{X}_1)$ and the relative fit,

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

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

$$\frac{1}{S}\sum_s \psi_s,$$

will produce an estimate of $2\mathrm{E}\left\{Q(\mathcal{X}, \theta^*) - Q(\mathcal{X}, \hat{\theta})\right\}$, thereby give us an estimate of the expected difference between the in-sample fit and the out-of-sample fit. (This approach would also produce an 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 $\psi = Q(\mathcal{X}_1, \hat{\theta}(\mathcal{X}_1)) - \frac{n_1}{n_2} Q(\mathcal{X}_2, \hat{\theta}(\mathcal{X}_1))$.

Bootstrap resampling, will also enable us to compute

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

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

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

Estimation by the jackknife, as in Hansen and Racine (2007) is also a possibility.

6 Concluding Remarks

[To be completed]

An implication of the "Winner's Curse Problem" is that a parsimonious model may not possess the traits of a parsimonious model, when the model is selected from a larger family of parsimonious models.

Selecting the true model, or the (in population) best approximating model should not

be the dominant criterion when the purpose is to select a model with good out-of-sample properties. The reason is that the true model need not be the best choice, because it may have a larger overfit than another model, and the overfit can more that offset the degree to which the true model dominates the other model in population.

Under the out-of-sample paradigm the relevant question for model selection is "how good is the selected model, relative to other models" rather than "how frequently is the true model selected". For instance, it may be the case that the true model is only selected with its overfit is large.

A tightly parameterized model that is selected after an extensive search may not be parsimonious due to the winner's curse.

Cross-validation IC better than in-sample ICs such as AIC and BIC.

This result forms the basis for a unified framework for discussing aspect of model selection, model averaging, and the effects of data mining.

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 by estimating the parameters with one criterion but evaluating the models with a different criterion. 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).

In empirical problems where overfitting is very problematic, such as portfolio choice over a large number of assets, almost any type of constraint on the optimization problem will improve out-of-sample performance. So to conclude that a particular structure has a deeper meaning (beyond reducing the overfitting problem) would require additional arguments beyond the fact that it improves the out-of-sample fit.

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.

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.

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.

Clark, T. E., and M. W. McCracken (2006): "Combining Forecasts from Nested Models," .

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

Clements, M. P., and D. F. Hendry (1998): *Forecasting Economic Time Series*. Cambridge University Press, Cambridge.

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.

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.

Hendry, D. F. (2004): "Robustifying Forecasts from Equilibrium-Correction Models," *Working Paper*.

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 (2008): "Bayesian Averaging, Prediction and Nonnested Model Selection," NBER Working Paper No. W14284.

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

Inoue, A., and L. Kilian (2006): "On the Selection of Forecasting Models," *Journal of Econometrics*, 130, 273–306.

Jobson, J. D., and B. Korkie (1980): "Estimation for Markowitz Efficient Portfolios," *Journal of the American Statistical Association*, 75, 544–554.

Johansen, S. (1991): "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 59, 1551–1580.

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.

McCracken, M. W. (2007): "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics*, 140, 719–752.

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.

Ng, S., and P. Perron (2005): "A Note on the Selection of Time Series Models," *Oxford Bulletin of Economics and Statistics*, 67, 115–134.

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

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

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

Proof of Theorem 1. To simplify notation we write $Q_x(\cdot)$ as short for $Q(\mathcal{X}, \cdot)$, and with a similar simplification for $S_x(\cdot)$ and $H_x(\cdot)$. Assumption 1, it is well known that $\hat{\theta} \xrightarrow{p} \theta^*$, that $\hat{\theta}$ is characterized by $S_x(\hat{\theta}) = 0$, and that $n^{-1/2} S_x(\theta^*) \xrightarrow{d} N(0, \mathcal{J}_0)$. Thus,

$$0 = S_x(\hat{\theta}) = S_x(\theta^*) + H_x(\tilde{\theta})(\hat{\theta} - \theta), \quad \text{where } \tilde{\theta} \in [\theta^*, \hat{\theta}]$$

so that $(\hat{\theta}-\theta^*) = \left[-H_x(\tilde{\theta})\right]^{-1} S_x(\theta^*)$. A second order Taylor expansion of $Q_x(\theta^*)$, about $\hat{\theta}$ yields,

$$\begin{aligned} Q_x(\hat{\theta}) - Q_x(\theta^*) &= \frac{1}{2}(\hat{\theta}-\theta^*)' \left[-H_x(\check{\theta})\right](\hat{\theta}-\theta^*) \\ &= \frac{1}{2} S_x(\theta^*)' \left[-H_x(\theta^*)\right]^{-1} S_x(\theta^*) + o_p(n^0), \end{aligned}$$

with $\check{\theta} \in [\theta^*, \hat{\theta}]$. Here we used that $H_x(\theta_n) - H_x(\theta) = o_p(n)$, whenever $\theta_n \xrightarrow{p} \theta^*$, and that $S_x(\theta^*) = O_p(n^{1/2})$. Out-of-sample, a Taylor expansion of $Q_y(\hat{\theta})$ about θ^* yields

$$\begin{aligned} Q_y(\hat{\theta}) - Q_y(\theta^*) &= S_y(\theta^*)'(\hat{\theta}-\theta) + \frac{1}{2}(\hat{\theta}-\theta^*)' H_y(\check{\theta})(\hat{\theta}-\theta^*) \\ &= S_y(\theta^*)' \left[-H_x(\tilde{\theta})\right]^{-1} S_x(\theta^*) \\ &\quad -\frac{1}{2} S_x(\theta^*)' \left[H_x(\tilde{\theta})\right]^{-1} \left[-H_y(\check{\theta})\right] \left[H_x(\tilde{\theta})\right]^{-1} S_x(\theta^*), \end{aligned}$$

with $\check{\theta} \in [\theta^*, \hat{\theta}]$.

Now define $V_{1,n} = n^{-1/2} \mathcal{J}_0^{-1/2} S_x(\theta^*)$ and $V_{2,n} = m^{-1/2} \mathcal{J}_0^{-1/2} S_y(\theta^*)$. Since $-n^{-1} H_x(\tilde{\theta}) \xrightarrow{p} \mathcal{I}_0$ and $-m^{-1} H_y(\check{\theta}) \xrightarrow{p} \mathcal{I}_0$, it follows that

$$\begin{aligned} Q_y(\hat{\theta}) - Q_y(\theta^*) &= \sqrt{\frac{m}{n}} V_{2,n}' \mathcal{J}_0^{1/2} \mathcal{I}_0^{-1} \mathcal{J}_0^{1/2} V_{1,n} \\ &\quad + \frac{1}{2}\frac{m}{n} V_{1,n}' \mathcal{J}_0^{1/2} \mathcal{I}_0^{-1} \mathcal{J}_0^{1/2} V_{1,n} + o_p(1). \end{aligned}$$

D Then by Assumption 1 and independence between $\mathcal{X}$ and $\mathcal{Y}$, it follows that $(V_{1,n}', V_{2,n}')' \xrightarrow{d} N_{2k}(0, I_{2k})$, so that

$$2(\eta, \tilde{\eta}) \xrightarrow{d} (V_1' A V_1, 2V_1' A V_2 - V_1' A V_1),$$

where $A = \mathcal{J}_0^{1/2} \mathcal{I}_0^{-1} \mathcal{J}_0^{1/2}$. Now write $Q'\Lambda Q = A$ where $Q'Q = I$ and Λ being a diagonal matrix with the eigenvalues of $A = \mathcal{J}_0^{1/2} \mathcal{I}_0^{-1} \mathcal{J}_0^{1/2}$, and define $Z_1 = QV_1$ and $Z_2 = QV_2$. Since $Ax = \lambda x$ for $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^k$ implies that $\mathcal{I}_0^{-1} \mathcal{J}_0 y = \lambda y$ with $y = \mathcal{J}_0^{-1/2} x$, it follows that the eigenvalues of $\mathcal{J}_0^{1/2} \mathcal{I}_0^{-1} \mathcal{J}_0^{1/2}$ coincide with those of $\mathcal{I}_0^{-1} \mathcal{J}_0$. This completes the proof. ■

B Special Cases and Additional Empirical Results

B.1 Log Likelihood for Regression Model

Here we look at the results of Corollary 2 in the context of a linear regression model.

Example 3 *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,*

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

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

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

Similarly, out-of-sample we have

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

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

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

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

	$\bar{k}$	AIC	$-\mu+\eta_{j^*}$	$-\mu$	$-\nu$	$-\eta_{j^*}$	$-\mu_{ave}$
Equal	3.2610	95.70	122.68	108.50	5.16	14.18	101.21
Linear	3.3180	95.55	126.99	112.75	9.59	14.25	106.85
Quadratic	3.3880	95.44	130.40	116.02	12.98	14.38	111.70
Cubic	3.4360	95.38	132.16	117.67	14.68	14.49	114.67

Table 6: The first column identifies the design in the simulation experiment. The average number of regressors, AIC, etc, are reported. The last column states the genuine quality of the 'model' that is a simple average across all estimated models.

B.2 Simulation

Example 4 *Consider the family of regression models,*

$$Y_t = \beta'_{(j)} Z_{(j),t} + \varepsilon_{(j),t}, \qquad t = 1, \ldots, n,$$

where $Z_{(j),t}$, $j = 1, \ldots, M$, *is a subset of a pool of explanatory variables,* $Z_{1,t}, \ldots, Z_{K,t}$.

Suppose that

$$Z_{i,t} = X_t + V_{i,t}, \qquad i = 1, \ldots, K,$$

where $X_t \sim$ iid $\mathrm{N}(0,1)$ *and* $V_t \sim$ iid $\mathrm{N}_K(0, \gamma^2 I_K)$, *while the dependent variable is given by*

$$Y_t = \alpha(X_t + w'V_t) + U_t, \qquad U_t \sim \text{iid } \mathrm{N}(0,1), \qquad w'w = 1. \tag{4}$$

The family of regression models will consist of all subset regressions with k regressions, with $k = 1, \ldots, k_{\max} \leq K$.

For a given value of $\rho \in (0,1)$, *we set* $\alpha = \frac{\rho}{\sqrt{(1-\rho^2)(1+\gamma^2)}}$ *so that* ρ^2 *is the population* R^2 *in (4).*

We choose the vector of "weight", w, in four different ways. Equal: $w_i = 1/\sqrt{K}$, *Linear:* $w_i \propto i$, *Quadratic:*$w_i = i^2$, *and Cubic:* $w_i \propto i^3$.

Taking average over simulations: $\bar{k}$ *is the number of regressors in the selected model. AIC is the AIC value of the selected model,* $-\mu_j = \mathrm{E}(U'_{(j)}U_{(j)})$, $\nu_j = \mathrm{E}(U'_{(j)}U_{(j)}) - U'_{(j)}U_{(j)}$, *and* $\eta_j = U'_{(j)}U_{(j)} - \hat{U}'_{(j)}\hat{U}_{(j)}$.

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 models.

B.3 Additional empirical results for the US Term Structure of Interest Rates

B.4 Another Application: ARMA Estimation for Realized Kernel Estimator

Realized Kernel estimator applied to SPY x_t =logRK$_t$

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2},$$

with $\varepsilon_t \sim$

$$Q_A = -\frac{1}{2} \sum_{t \text{ odd}} \left(\log \sigma^2 + \frac{\hat{\varepsilon}_t^2}{\sigma^2} \right), \qquad \text{with } \hat{\varepsilon}_t =$$

	IMA(1,1)		ARMA(1,1)		ARMA(1,2)		ARMA(2,1)	
	A	B	A	B	A	B	A	B
ϕ_1	1.00	1.00	0.90	0.81	0.87	0.88	0.57	1.26
ϕ_2	–	–	–	–	–	–	0.23	-0.31
θ_1	0.62	0.55	0.53	0.32	0.52	0.40	0.23	0.78
θ_2	–	–	–	–	-0.06	0.11	–	–
μ	0.00	0.00	-0.12	-0.24	-0.15	-0.15	-0.23	-0.06
σ^2	0.19	0.18	0.18	0.17	0.18	0.17	0.18	0.17
$\max \ell_A$	142.57	*140.31*	150.82	*143.79*	151.45	*141.02*	152.10	*140.32*
$\max \ell_B$	*152.18*	153.70	*165.50*	170.12	*162.90*	171.52	*159.14*	172.06

B.5 Details concerning Portfolio Choice

Simulation design based on the estimates from Jobson and Korkie (1980) who randomly selected 20 stocks. The mean vector and covariance matrix was estimated with monthly returns for the sample December, 1949 to December 1975.

$$\hat{\mu} =$$

$$\begin{pmatrix} 0.50 & 0.90 & 1.10 & 1.74 & 1.82 & 1.11 & 0.91 & 1.18 & 1.35 & 1.07 & 1.16 & 1.23 & 0.81 & 1.18 & 0.88 & 1.20 & 0.72 & 1.16 & 0.92 & 1.25 \end{pmatrix}'$$

$$\hat{\Sigma} =$$

Akaike's Information Criterion (AIC)

	Odd months						Even months					
	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
$p=1$	2827.53	2836.34	2842.63	2845.16	2845.11	2844.76	2881.80	2928.69	2945.24	2954.96	2958.08	2957.19
$p=2$	2879.84	2886.03	2892.58	2897.22	2897.38	2897.70	2922.67	2961.73	2983.78	2988.33	2994.10	2993.32
$p=3$	2902.21	2906.95	2911.23	2913.19	2914.22	2913.63	2939.89	2977.74	2994.98	2997.42	3000.72	2999.91
$p=4$	2913.48	2917.75	2921.18	2923.61	2924.73	2924.46	2961.23	2982.17	2999.28	3001.98	3005.33	3004.58
$p=5$	2924.50	2926.88	2927.26	2928.65	2928.46	2928.19	2952.51	2971.59	2985.87	2988.13	2991.43	2990.59
$p=6$	2929.57	2932.17	2934.75	2935.68	2935.42	2935.15	2952.60	2968.34	2976.93	2979.52	2981.51	2980.66
$p=7$	2928.91	2933.40	2933.39	2933.40	2934.22	2933.75	2978.18	2992.76	3000.43	3002.24	3003.07	3002.20
$p=8$	2934.24	2940.37	2942.62	2941.61	2941.66	2941.02	2996.28	3005.68	3014.81	3017.15	3018.01	3017.15
$p=9$	2942.07	2947.07	2953.01	2952.79	2950.84	2949.90	3000.72	3013.06	3020.97	3024.05	3025.92	3024.97
$p=10$	2942.11	2949.65	2953.05	2952.53	2950.56	2949.72	3017.27	3030.36	3041.38	3045.67	3045.91	3044.93
$p=11$	2937.76	2947.83	2951.42	2950.99	2948.48	2947.56	3014.46	3027.57	3036.26	3040.55	3041.13	3040.16
$p=12$	2937.52	2949.75	2952.78	2951.82	2949.10	2948.10	3014.72	3025.26	3033.86	3037.01	3038.79	3037.79

Bayesian Information Criterion (BIC)

	Odd months						Even months					
	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
$p=1$	2818.35	2810.63	2804.06	2797.41	2791.85	2789.66	2872.62	2903.00	2906.70	2907.25	2904.86	2902.14
$p=2$	2824.74	2814.40	2808.10	2803.55	2798.20	2796.68	2867.63	2890.17	2899.37	2894.75	2895.01	2892.39
$p=3$	2801.19	2789.41	2780.83	2773.61	2769.12	2766.69	2838.96	2860.30	2864.70	2857.97	2855.76	2853.11
$p=4$	2766.55	2754.28	2744.86	2738.10	2733.71	2731.61	2814.44	2818.86	2823.12	2816.65	2814.50	2811.91
$p=5$	2731.65	2717.50	2705.02	2697.23	2691.53	2689.43	2759.84	2762.41	2763.84	2756.93	2754.73	2752.05
$p=6$	2690.80	2676.88	2666.60	2658.34	2652.58	2650.46	2714.06	2713.28	2709.03	2702.44	2698.93	2696.24
$p=7$	2644.23	2632.19	2619.32	2610.15	2605.46	2603.15	2693.77	2691.83	2686.66	2679.29	2674.62	2671.91
$p=8$	2603.64	2593.24	2582.64	2572.44	2566.98	2564.50	2665.99	2658.88	2655.16	2648.32	2643.68	2640.99
$p=9$	2565.55	2554.03	2547.11	2537.70	2530.24	2527.47	2624.56	2620.38	2615.45	2609.35	2605.72	2602.93
$p=10$	2519.68	2510.69	2501.23	2491.53	2484.04	2481.37	2595.23	2591.81	2589.98	2585.10	2579.84	2577.02
$p=11$	2469.42	2462.95	2453.69	2444.07	2436.05	2433.30	2546.55	2543.14	2538.99	2534.10	2529.18	2526.37
$p=12$	2423.25	2418.96	2409.13	2398.98	2390.75	2387.92	2500.94	2494.96	2490.72	2484.70	2480.97	2478.14

Table 7: AIC and BIC (multiplied by minus a half to make them directly comparible with out-of-sample criterion).

$Q(p,r) - Q(p-1,r)$. In-sample: "Odd" observations

	Odd months (in-sample): η_1						Even months: η_2					
	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
$p=1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p=2$	77.31	74.69	74.96	77.06	77.27	77.93	4.21	3.90	0.79	-3.28	0.80	0.73
$p=3$	47.37	45.92	43.65	40.97	41.84	40.93	-0.35	-13.06	-22.38	-24.43	-32.62	-31.19
$p=4$	36.27	35.79	34.95	35.41	35.50	35.83	-48.52	-40.21	-35.31	-37.18	-42.69	-41.84
$p=5$	36.02	34.13	31.08	30.04	28.74	28.74	-7.65	-13.21	-28.26	-26.77	-26.63	-26.57
$p=6$	30.07	30.29	32.50	32.03	31.96	31.95	-11.23	-14.99	-8.15	-12.38	-10.96	-12.04
$p=7$	24.34	26.23	23.63	22.73	23.80	23.60	13.12	6.39	4.36	2.04	1.50	1.26
$p=8$	30.33	31.97	34.23	33.21	32.44	32.27	-36.68	-34.35	-33.52	-39.31	-35.35	-36.30
$p=9$	32.82	31.70	35.39	36.18	34.18	33.88	-38.07	-26.52	-36.11	-36.15	-37.28	-37.76
$p=10$	25.05	27.58	25.04	24.74	24.72	24.82	-32.23	-31.70	-31.54	-30.04	-29.49	-29.34
$p=11$	20.65	23.18	23.37	23.46	22.92	22.84	-15.63	-18.81	-17.71	-16.96	-18.91	-18.03
$p=12$	24.75	26.92	26.36	25.82	25.62	25.54	-42.03	-43.71	-41.96	-40.19	-40.45	-41.97

$Q(p,r) - Q(p-1,r)$. In-sample: "Even" observations

	Odd months: η_2						Even months (in-sample): η_1					
	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
$p=1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p=2$	21.35	22.84	26.23	22.76	24.74	25.15	65.87	58.05	63.54	58.37	61.02	61.13
$p=3$	-2.75	-8.32	-13.04	-13.66	-16.99	-17.45	42.21	41.00	36.20	34.10	31.62	31.59
$p=4$	-44.73	-28.68	-20.95	-19.14	-21.50	-21.32	46.34	29.44	29.30	29.56	29.61	29.68
$p=5$	6.36	-1.13	-5.73	-11.11	-10.45	-10.58	16.28	14.42	11.59	11.16	11.11	11.01
$p=6$	-9.32	-13.23	-6.17	-6.93	-7.28	-7.14	25.09	21.75	16.07	16.39	15.08	15.06
$p=7$	-22.75	-22.64	-21.87	-21.38	-18.90	-19.05	50.58	49.42	48.50	47.72	46.56	46.54
$p=8$	-23.94	-24.29	-27.56	-25.66	-24.82	-24.61	43.10	37.92	39.38	39.91	39.93	39.96
$p=9$	-12.54	-12.96	-12.40	-13.49	-13.71	-13.92	29.44	32.38	31.16	31.90	32.91	32.82
$p=10$	-36.28	-35.51	-38.64	-39.84	-39.66	-39.88	41.55	42.30	45.41	46.63	44.99	44.96
$p=11$	6.20	4.02	0.68	1.05	2.36	2.42	22.19	22.21	19.88	19.87	20.22	20.22
$p=12$	-17.36	-18.13	-21.51	-21.12	-22.89	-23.44	25.26	22.69	22.60	21.46	22.66	22.64

Table 8: Columnwise increments in $Q(\cdot, \hat{\theta})$. In-sample increments in upper-left and lower-right panels. Out-of-sample in upper-right and lower-left panels.

$$\begin{pmatrix}
53.6 & 6.6 & 19.8 & 34.1 & 6.3 & 5.8 & 16.9 & 15.3 & 9.6 & 10.1 & 10.3 & 18.9 & 8.5 & 14.6 & 14.6 & 14.3 & 27.9 & 25.1 & 11.8 & 16.9 \\
6.6 & 29.8 & 16.7 & 20.7 & 6.5 & 11.8 & 8.4 & 9.3 & 10.8 & 11.2 & 9.6 & 8.8 & 13.5 & 14.1 & 16.5 & 8.8 & 14.9 & 16.7 & 22.8 & 10.3 \\
19.8 & 16.7 & 82.9 & 48.0 & 18.7 & 21.0 & 22.2 & 16.2 & 16.3 & 18.9 & 21.6 & 27.0 & 8.8 & 23.3 & 17.4 & 22.3 & 36.7 & 41.4 & 21.4 & 27.7 \\
34.1 & 20.7 & 48.0 & 178.1 & 27.5 & 19.3 & 32.4 & 26.9 & 18.3 & 22.4 & 21.8 & 41.7 & 17.3 & 42.9 & 26.3 & 30.3 & 66.0 & 47.6 & 20.7 & 43.9 \\
6.3 & 6.5 & 18.7 & 27.5 & 118.1 & 26.3 & 23.9 & 12.4 & 14.2 & 23.1 & 31.0 & 13.1 & 5.4 & 20.4 & 9.9 & 14.3 & 17.1 & 20.2 & 13.5 & 18.5 \\
5.8 & 11.8 & 21.0 & 19.3 & 26.3 & 57.1 & 20.2 & 11.7 & 15.2 & 16.3 & 13.7 & 19.3 & 7.8 & 21.5 & 11.3 & 13.2 & 13.5 & 12.3 & 16.8 & 18.1 \\
16.9 & 8.4 & 22.2 & 32.4 & 23.9 & 20.2 & 52.1 & 15.3 & 12.1 & 17.7 & 18.0 & 21.4 & 9.6 & 26.4 & 16.2 & 15.2 & 25.6 & 24.8 & 15.5 & 25.3 \\
15.3 & 9.3 & 16.2 & 26.9 & 12.4 & 11.7 & 15.3 & 48.3 & 9.7 & 9.4 & 8.6 & 14.4 & 9.9 & 11.3 & 13.3 & 17.0 & 32.1 & 21.7 & 14.3 & 15.8 \\
9.6 & 10.8 & 16.3 & 18.3 & 14.2 & 15.2 & 12.1 & 9.7 & 29.8 & 11.2 & 13.1 & 13.8 & 7.3 & 16.7 & 11.4 & 8.2 & 15.7 & 20.6 & 14.8 & 10.7 \\
10.1 & 11.2 & 18.9 & 22.4 & 23.1 & 16.3 & 17.7 & 9.4 & 11.2 & 35.1 & 22.6 & 13.0 & 7.9 & 17.6 & 10.7 & 12.6 & 16.2 & 21.5 & 14.2 & 14.7 \\
10.3 & 9.6 & 21.6 & 21.8 & 31.0 & 13.7 & 18.0 & 8.6 & 13.1 & 22.6 & 47.6 & 16.6 & 6.0 & 19.8 & 9.3 & 13.5 & 20.5 & 18.8 & 13.3 & 17.7 \\
18.9 & 8.8 & 27.0 & 41.7 & 13.1 & 19.3 & 21.4 & 14.4 & 13.8 & 13.0 & 16.6 & 65.6 & 7.9 & 23.1 & 11.6 & 25.8 & 35.8 & 26.4 & 17.0 & 23.7 \\
8.5 & 13.5 & 8.8 & 17.3 & 5.4 & 7.8 & 9.6 & 9.9 & 7.3 & 7.9 & 6.0 & 7.9 & 23.5 & 12.0 & 14.3 & 8.5 & 15.2 & 14.2 & 15.8 & 9.7 \\
14.6 & 14.1 & 23.3 & 42.9 & 20.4 & 21.5 & 26.4 & 11.3 & 16.7 & 17.6 & 19.8 & 23.1 & 12.0 & 51.2 & 16.4 & 14.7 & 26.2 & 25.6 & 20.4 & 20.9 \\
14.6 & 16.5 & 17.4 & 26.3 & 9.9 & 11.3 & 16.2 & 13.3 & 11.4 & 10.7 & 9.3 & 11.6 & 14.3 & 16.4 & 28.7 & 12.2 & 19.9 & 24.3 & 22.4 & 13.8 \\
14.3 & 8.8 & 22.3 & 30.3 & 14.3 & 13.2 & 15.2 & 17.0 & 8.2 & 12.6 & 13.5 & 25.8 & 8.5 & 14.7 & 12.2 & 56.0 & 32.3 & 24.5 & 13.1 & 14.5 \\
27.9 & 14.9 & 36.7 & 66.0 & 17.1 & 13.5 & 25.6 & 32.1 & 15.7 & 16.2 & 20.5 & 35.8 & 15.2 & 26.2 & 19.9 & 32.3 & 109.5 & 50.8 & 18.6 & 32.3 \\
25.1 & 16.7 & 41.4 & 47.6 & 20.2 & 12.3 & 24.8 & 21.7 & 20.6 & 21.5 & 18.8 & 26.4 & 14.2 & 25.6 & 24.3 & 24.5 & 50.8 & 131.8 & 27.0 & 29.2 \\
11.8 & 22.8 & 21.4 & 20.7 & 13.5 & 16.8 & 15.5 & 14.3 & 14.8 & 14.2 & 13.3 & 17.0 & 15.8 & 20.4 & 22.4 & 13.1 & 18.6 & 27.0 & 44.7 & 16.1 \\
16.9 & 10.3 & 27.7 & 43.9 & 18.5 & 18.1 & 25.3 & 15.8 & 10.7 & 14.7 & 17.7 & 23.7 & 9.7 & 20.9 & 13.8 & 14.5 & 32.3 & 29.2 & 16.1 & 58.7
\end{pmatrix}$$