Variable Selection in Predictive Regressions

Serena Ng*

May 2012

Abstract

This chapter reviews methods for selecting empirically relevant predictors from a set of N potentially relevant ones for the purpose of forecasting a scalar time series. I first discuss criterion based procedures in the conventional case when N is small relative to the sample size, T. I then turn to the large N case. Regularization and dimension reduction methods are then discussed. Irrespective of the model size, there is an unavoidable tension between prediction accuracy and consistent model determination. Simulations are used to compare selected methods from the perspective of relative risk in one period ahead forecasts.

Keywords: Principal components, Factor models, Regularization, Information Criteria.

JEL Classification: C1, C2, C3, C5

*Department of Economics, Columbia University. Email: serena.ng@columbia.edu
Correspondence: Columbia University, 420 W. 118 St., MC 3308, New York, NY 10027. This paper is prepared for the Handbook of Forecasting. I thank Graham Elliott, Bruce Hansen, Chu-An Liu, Alexei Onatski and Allan Timmermann for many helpful comments. Financial support from from the NSF (SES-0962431) is gratefully acknowledged.

1 Introduction

This chapter considers linear models for explaining a scalar variable when a researcher is given T historical observations on N potentially relevant predictors but that the population regression function is well approximated by a set of empirically relevant predictors whose composition is unknown. The problem is to determine the identity of these predictors. I consider the variable selection problem both when the number of potentially relevant predictors is small and when it is large. I distinguish models with few relevant predictors from those with many relevant predictors that may possibly have a factor structure. The common factors in the predictor set are distinguished from those in the variable of interest. I also distinguish between discretionary and 'must have' regressors to accommodate variables (such as lags) that practitioners for one reason or another choose to keep.

Three types of variable (model) selection procedures are distinguished:- criterion based methods, regularization, and dimension reduction procedures. Section 2 begins with a discussion of information criteria and sequential testing procedures in the classical setting when N is small relative to T. I then turn to the data-rich case when N is large. Regularization methods are discussed in section 3 with special focus on L_1 type penalties. Section 4 concerns constructing components to reduce the dimension of the predictor set. The relation between factor analysis, principal components, and partial least squares is reviewed. Section 5 discusses some unresolved issues, in particular, whether to target components/factors to the variable of interest, and whether constructed predictors should be treated like the observed ones. The analysis wraps up with a discussion of the tension between optimal prediction and consistent model selection. These issues are illustrated by means of monte-carlo simulations.

The discussion on a variety of methods reflects my view that which procedure is best will likely depend on the true data structure which we unfortunately do not know. Regularization seems to better suit situations when all but a few observed predictors have non-zero effects on the regression function while dimension reduction methods seem more appropriate when the predictors are highly collinear and possibly have a factor structure. The best model may not be identified if the set of candidate models is narrowed by the method used to select predictors. Nonetheless, in spite of considering a broad array of methods, the review remains incomplete and far from exhaustive. The discussion is presented at a general level leaving the readers to references for technical details and assumptions. Cross-validation, bayesian methods, model averaging and forecast combinations as well as many issues related to the general-to-specific modeling strategy outlined in Campos, Ericsson, and Hendry (1994) are omitted. I also do not provide empirical or monte-carlo forecast comparisons; such results can be found in Stock and Watson (2006, 2010), Kim and Swanson (2010),

as well as Pesaran, Pick, and Timmermann (2011). These papers also contain useful references to applications of methods being reviewed.

The following notation will be adopted. For an arbitrary $m \times n$ matrix A, let $\mathbf{A}_j$ be the j-th column of A. The submatrix formed from the first r columns of A is denoted $A_{1:r}$. For a $N \times 1$ vector $z \in \mathbb{R}^n$, the L_2 norm is $\|z\|_2^2 = \sum_{i=1}^N z_i^2$, the L_1 norm is $\|z\|_1 = \sum_{i=1}^N |z_i|$, and the L_0 norm is $\|z\|_0 = \sum_{j=1}^N I_{z_j \neq 0}$. The singular value decomposition of a $T \times N$ matrix X when $T > N$ is $X = U_X D_X V_X'$ where D_X is a diagonal matrix of singular values with $d_{X,1} \geq d_{X,2} \ldots \geq d_{X,N}$ along the diagonal, U_X and V_X are $T \times N$ and $N \times N$ orthogonal matrices spanning the column and row space of X respectively, with $(V_X')^{-1} = V_X$, $U_X' U_X = I_N$ and $V_X' V_X = V_X V_X' = I_N$. Also let $x_+ = \max(x, 0)$, $x_- = \min(-x, 0)$, $\text{sgn}\,(x) = 1$ if $x > 0$, $\text{sgn}\,(x) = -1$ if $x < 0$, $\text{sgn}\,(x) = 0$ if $x = 0$. To conserve on notation, I use ϵ_t to generically denote the error of the predictive regression irrespective of the predictors and $\mathcal{E}$ is its vector analog.

In the statistics and machine learning literature, the exercise of using inputs (Z) to learn about an outcome (y) is known as supervised learning. This is to be contrasted with unsupervised learning which concerns how an outcome is organized or clustered without reference to observed inputs. The exercise of model based economic forecasting is a form of supervised learning in which the object of interest is the value of y at some time $T + h$ and for which historical data on $(y_1, \ldots, y_T)'$ and other inputs are available. Denote by $W_t = (w_{1t}, \ldots, w_{Mt})'$ a set of M 'must have' predictors that typically include lags of y_t and deterministic terms such as dummy variables that control for irregular events in the sample. Often, researchers also have at their disposal a set of N potentially relevant predictors $X_t = \begin{pmatrix} x_{1t} & x_{2t} & \ldots & x_{NT} \end{pmatrix}'$. These regressors are predetermined and chosen with the forecast horizon h in mind. To simplify notation, reference of the predictors and y_t to h will be suppressed. Let $Z_t = (W_t' \ X_t')'$. Throughout, each y_t is assumed to be mean zero, the regressors are demeaned and scaled so that for each $i = 1, \ldots, M + N$, $\sum_{t=1}^T z_{it} = 0$ and $\sum_{t=1}^T z_{it}^2 = 1$.

A predictive regression that includes all available predictors is

$$y_t = W_t' \alpha + X_t' \beta + \epsilon_t, \tag{1}$$

where for $t = 1, \ldots T$, ϵ_t is white noise with variance σ^2. Let $Y = (y_1, \ldots, y_T)'$ and $\delta = (\alpha' \ \beta')'$. The predictive regression in matrix form is

$$Y \ = \ W\alpha + X\beta + \mathcal{E} = Z\delta + \mathcal{E}.$$

The best linear unbiased h period forecast given information up to period T is given by the linear projection:

$$y_{T+h|T} = W_{T+h|T}' \alpha + X_{T+h|T}' \beta.$$

Equation (1) is of interest in a variety of applications. For example, an out-of-sample forecast of inflation with $h > 0$ can be obtained with W_t being lags of inflation and X_t being indicators of slackness in the goods and labor markets. Many econometric exercises involve the in-sample prediction with $h = 0$. In instrumental variable estimation, y_t would be one of the many endogenous variable in the system, W_t would be exogenous variables, and X_t would be the potentially valid instruments of the endogenous regressor y_t. In risk-return analysis, y_t could be the excess return or volatility for holding an asset over h periods. Given information W_t and X_t available to econometricians, predictive regressions can be used to construct the conditional mean and volatility of asset returns. A central question in these applications is the robustness of these estimates to the choice of predictors. Predictive regressions are also useful for testing hypothesis such as rational expectations and/or market efficiency. For example, if theory suggests that bond risk premia reflects real macroeconomic risk, a finding that financial variables appearing as X_t in (1) are significant would be at odds with theory. As discussed in Ludvigson and Ng (2011), whether one accepts or rejects the hypothesis often rests on the choice of predictor set X_t.

The best linear prediction is clearly infeasible because $\delta = (\alpha' \ \beta')'$ is unknown. Assuming that Z is full column rank, δ can be replaced by the least squares estimates:

$$\widehat{\delta}_{LS} = \text{argmin}_\delta \|Y - Z\delta\|_2^2 = \ (Z'Z)^{-1}Z'Y.$$

Since $Z'Z = V_Z D_Z^2 V_Z'$, it follows that

$$\widehat{\delta}_{LS} = V_Z D_Z^{-1} U_Z' Y = \sum_{i=1}^{N+M} \frac{\mathbf{U}_{Z,i}' Y}{d_{Z,i}} \mathbf{V}_{Z,i}.$$

The in-sample least squares fit is

$$\widehat{Y}_{LS} = Z\widehat{\delta}_{LS} = U_Z U_Z' Y \tag{2}$$

and assuming that $W_{T+h|T}$ and $X_{T+h|T}$ are available, the feasible h-period ahead prediction is

$$\widehat{y}_{T+h|T} \ = \ W_{T+h|T}'\widehat{\alpha} + X_{T+h|T}'\widehat{\beta} \ = Z_{T+h|T}'\widehat{\delta}_{LS}.$$

Although the least squares estimate $\widehat{\delta}_{LS}$ is $\sqrt{T}$ consistent for δ, the mean-square forecast error is increasing in $\dim(\beta)$ for given $\dim(\alpha)$, and not every potentially important predictor is actually relevant. Retaining the weak predictors can introduce unwarranted sampling variability to the prediction. The objective of the exercise is to form an accurate forecast using the available information. I focus on quadratic loss and hence accuracy is measure is defined in terms of mean-squared forecast error.

Let $\mathcal{A}$ be an index set containing the positions of the variables deemed empirically relevant. Henceforth, $X_{\mathcal{A}}$ will be referred to as the 'empirically relevant' or 'active set' of predictors. Let $\widehat{\beta}_A$

be a $N \times 1$ vector of estimates whose j-th element is zero if the corresponding regressor's index is not in $\mathcal{A}$, and equal to the least square estimates otherwise. Two ways of forecasting can be envisioned. In the first case, only a small subset of X has predictive power. In the second case, the best forecast is achieved by using information in a large number of predictors, however small the contribution of each series is in explaining Y. Belloni and Chernozhukov (2011) refer to sparsity as the condition when the number of non-zero entries in the population coefficient vector β is much smaller than the dimension of β. Following these authors, the predictor set in first situation is said to be sparse. It is then fitting to characterize the predictor set in the second situation as dense. The difference between the two comes down to the dimension of $X_{\mathcal{A}}$ relative to the sample size T.

2 Criterion Based Methods when $N < T$

Mallows (1973) is amongst the first to determine $X_{\mathcal{A}}$ on the basis of prediction accuracy. His criterion is the scaled sum of squared errors:

$$\frac{SSR_p}{\sigma^2} = \frac{1}{\sigma^2}\left(\delta - \widehat{\delta}_A\right)' Z'Z\left(\delta - \widehat{\delta}_A\right)$$

where SSR_p is the sum of squared residuals in a regression of Y on W and $X_{\mathcal{A}}$. The subscript p refers to the number of regressors included in the regression. In the framework given by (1), $p = \dim(\widehat{\alpha}) + \dim(\mathcal{A})$ is less than T. Assuming that the regressors Z are non-random and that the errors are homoskedastic, Mallows (1973) shows that a useful estimate of $E(\frac{SSR_p}{\sigma^2})$ is

$$CP_p = \frac{1}{\widehat{\sigma}^2} SSR_p - T + 2p,$$

where $\widehat{\sigma}^2$ is an accurate estimate of σ^2. He also proposes two multivariate generalization of CP: one that replaces $\frac{SSR_p}{\sigma^2}$ by a weighted sum of squared errors, and another that uses an estimate δ_A that is not least squares based.

The CP criterion defines $X_{\mathcal{A}}$ as the subset of explanatory variables that corresponds to the lowest point in the plot of CP against p. Mallows (1973) does not recommend to blindly follow this practice because the rule will not be reliable when a large number of subsets are close competitors to the minimizer of CP. Li (1987) considers the squared difference between the true and the estimated conditional mean $L_T(p) = \frac{1}{T}\|y_{T+h|T} - \widehat{y}_{T+h|T}\|^2$ as the criterion for prediction accuracy. He relates the CP to cross-validation methods and shows that it is optimal when the regression errors are homoskedastic in the sense that $\frac{L_T(\widehat{p})}{\inf_{p\in\mathcal{P}} L_T(p)} \xrightarrow{p} 1$, where $\mathcal{P} = (1, 2, \ldots, N+M)$ is an index set. These results are extended to allow for heteroskedastic errors in Andrews (1991).

The CP criterion is related to a large class of information criteria that determines the size of a model as follows:

$$p_{IC} = \text{argmin}_{\,p=1,\ldots p_{\max}} IC_p, \quad IC_p = \left[\log \widehat{\sigma}_p^2 + p\frac{C_T}{T}\right],$$

where $p_{\max}$ is the maximum number of variables considered. The criterion function has three components. The first is $\widehat{\sigma}_p^2$, which measures the fit of a model with p parameters. The second is p, which defines the complexity of the model. The third is $\frac{C_T}{T}$, a term that penalizes model complexity in favor of parsimony. The factor of T in the penalty term is appropriate whenever the variance of $\widehat{\delta}$ tends to zero at rate T. The choice of C_T is crucial and will be discussed below.

Model selection procedures are probably most analyzed in the context of autoregressions in which case $Z_t = X_t = (y_{t-1}, \ldots, y_{t-p})'$, W_t is empty, and p is small relative to the sample size T. Because the predictors in an autoregression have a natural (time) ordering, the varaible selection problem is computationally simple. A p-th order autoregression uses p lags and the model selection problem reduces to the determination of the lag length, p. Akaike (1969, 1970) propose to measure adequacy by the final prediction error $E\left(y_{T+h} - \widehat{y}_{T+h}\right)^2$ which can be viewed as a weighted version of Mallows' criterion with all weight given to the final observation. Assuming that a constant is included in the autoregression and that the true order of the autoregression p is known, Akaike suggests the large sample approximation:

$$E\left(y_{T+h} - \widehat{y}_{T+h}\right)^2 \approx \left(1 + \frac{p+1}{T}\right)\sigma^2.$$

To make the criterion operational, Akaike first replaces σ^2 in the above expression by $\frac{1}{T-p-1}SSR_p$ and then chooses p to minimize the statistic:

$$FPE_p = \left(1 + \frac{p+1}{T}\right)\frac{SSR_p}{T-p-1} \equiv \frac{T+p+1}{T-p-1}\widehat{\sigma}_p^2,$$

where $\widehat{\sigma}_p^2 = \frac{1}{T}SSR_p$. Note that as $T \to \infty$, such a strategy is equivalent to choosing p by minimizing $\log FPE_p = \log \widehat{\sigma}_p^2 + \frac{2p}{T}$. Assuming that the true p increases with T, Shibata (1981) shows that the FPE and CP are asymptotically equivalent.

Phillips (1979) and others note that minimizing the conditional mean squared forecast error CMSFE= $E\left[\left(y_{T+h} - \widehat{y}_{T+h}\right)^2 | y_1, \ldots, y_T\right]$ may be more relevant in practice as researchers only observe one draw of the data. Ing and Yu (2003) approximate the CMSFE by

$$V_p = (1 + \frac{p}{T})\widehat{\sigma}_p^2 + \frac{1}{T}\mathbf{X}'_{1:p}S_{XX}^{-1}\mathbf{X}_{1:p}\widehat{\sigma}_p^2$$

where $S_{XX} = \frac{1}{T}\mathbf{X}'_{1:p}\mathbf{X}_{1:p}$, and $\mathbf{X}_{1:p}$ is a matrix consisting of p lags of the dependent variable. The authors show that V_p has a stronger correlation with CMSFE than the FPE.

Taking advantage of the ordered nature of time series data, many theoretical results are also available for selection of parametric time series models. Hannan and Deistler (1988) show that the p_{IC} chosen for autoregressions is asymptotically proportional to $\log T$ when the observed data are stationary ARMA processes. This logarithmic rate of increase extends to ARMAX and multivariate

models. Practical issues in using information criteria are discussed in Ng and Perron (2005). In particular, all autoregressions of order p must be estimated using T-$p_{\max}$ observations even if $p < p_{\max}$. This is necessary for the goodness of fit component of information criteria to not depend on the complexity component of the criteria.

Sequential testing procedures can also be used to select models. It is generally used when the number of candidate models to be considered is small, as is the case of autoregressions. A general-to-specific (top-down) method starts from the largest model which in the case of autoregression would be the $p_{\max}$ lags of the dependent variable. One checks if the coefficient on the last (ie. $p_{\max}$-th) lag is zero at some prescribed significance level. If it is not significant, the model with $p_{max} - 1$ lags is estimated and the last lag in this regression (ie. $p_{\max} - 1$) is tested. If it is not, a model with $p_{\max} - 2$ lags is estimated, and so on. The test on the last lag is repeated until the estimated coefficient on the last lag is found significant. General to specific procedures are detailed in Hendry and Doornik (2001). It is also possible to consider a specific-to-general (bottom-up) approach that starts with the smallest model possible. However, Hall (1994) finds that such a specific-to- general approach is generally not valid for pure AR models and its finite sample properties are inferior to general-to-specific approaches.

Sequential t tests and information criteria are *stepwise, data dependent,* rules that start by setting all coefficients equal to zero, and then build a sequence of models that include one additional variable at a time. Top down (bottom up) sequential testing is a form of backward (forward) stepwise regression. Stepwise methods share two common features. First, the coefficients of the variables already included in the regression are adjusted when a new variable is added or deleted. Stepwise algorithms are 'greedy' because the locally optimal choices made at each stage may not be globally optimal. Second, they perform what is known as 'hard thresholding': a variable is either in or out of the predictor set. An undesirable feature of this is that a regressor set selected from N available predictors may disagree with the one chosen when N is increased or decreased slightly. In other words, hard thresholding is sensitive to small changes in the data because of discreteness of the decision rule (also known as the bouncing beta problem). Furthermore, as discussed in Fan and Li (2001), a good understanding of stepwise methods requires an analysis of the stochastic errors in the various stages of the selection problem, which is not a trivial task.

The crucial parameter in a sequential testing procedure is the size of the test. If the size is too small, the critical value will be large and few variables will be selected. But information criteria can also be seen from a stepwise testing perspective. The AIC and BIC choose a test size that corresponds to critical values of $\sqrt{2}$ and $\sqrt{\log T}$, respectively. Now seen from the perspective of information criteria, a two tailed five percent t test corresponds to a C_T of 1.96. The variable selection problem boils down to the choice of C_T with large values favoring parsimonious models.

Different values for C_T have been proposed but the most widely used ones are probably $\log T$ and 2. The BIC (Bayesian Information Criterion) of Schwarz (1978) assigns a non-zero prior probability to a model of small dimension. Maximizing an approximation to the posterior probability of the model is equivalent to minimizing the IC with $C_T = \log T$. In addition to the FPE, a C_T of two can also be motivated from perspective of the Kullback-Leibler (KL) distance. Following Cavanaugh (1997),the KL distance between the candidate model parameterized by δ_p and the true model with density g is

$$D(\delta_p) = E_g(-2\log L(y|\delta_p))$$

where E_g denotes expectation taken with respect to the true density, $L(\delta_p|y)$ is the likelihood of the candidate model. While δ_p can be estimated from the data, the KL still cannot be used to evaluate models without knowledge of g. Akaike (1974) considers the expectation of KL when the candidate models nest the true model parameterized by δ_0.

$$\begin{aligned} E_0[D(\widehat{\delta}_p)] &= E_0(-2\log L(\widehat{\delta}_p|y)) \\ &\quad + \Big[E_0(-2\log L(\delta_0|y)) - E_0(-2\log L(\widehat{\delta}_p|y))\Big] \\ &\quad + \Big[(E_0(D(\widehat{\delta}_p)) - E_0(-2\log L(\delta_0|y))\Big]. \end{aligned} \tag{3}$$

The second order expansion of each of the last two terms is the likelihood ratio statistic which can be approximated by p since the expected value of a χ^2 random variable with p degrees of freedom is p. The expected KL suggests to select the best model minimizing

$$-2\log L_T(\widehat{\delta}_p|y) + 2p.$$

In the least squares case this further simplifies to

$$T\log(\frac{SSR_p}{T}) + 2p$$

Minimizing this criterion function is equivalent to minimizing the IC with $C_T = 2$. As noted earlier, the FPE and CP select the same model as the AIC. Hurvich and Tsai (1989) propose a small sample correction that replaces $2k$ by $\frac{1+p/T}{1-(p+2)/T}$ which amounts to adding a non-stochastic term of $\frac{2(p+1)(p+2)}{T(T-p-2)}$ to the AIC.

When the true model is not in the set of candidate models considered and possibly infinite dimensional, Takeuchi (1976) suggests to approximate each of the last two terms of (3) by

$$\mathrm{tr}(J(\delta_0)I(\theta_0)^{-1}) \tag{4}$$

where $J(\delta_0) = E_g[(\frac{\partial}{\partial\delta}\log L(\delta|y))(\frac{\partial}{\partial\delta}\log L(\delta|y)')]_{\delta=\delta_0}$ and $I(\delta_0) = E_g[-\frac{\partial^2\log L(\delta|y)}{\partial\delta_i\partial\delta_j}]_{\delta=\delta_0}$. The TAIC penalty is twice the quantity in (4). If δ is close to δ_0, $J(\delta_0)$ will be close to $I(\delta_0)$. The trace term

is approximately p and the TIC reduces to the AIC. The TIC has the flavor of determining the best model taking into account the sampling error of the quasi-maximum likelihood estimates.

To make the TIC operational without knowing g, observed Fisher information and the outer product of the scores evaluated at $\widehat{\delta}_p$ are used in place of $J(\delta_0)$ and $I(\delta_0)$, respectively. The TIC is computationally more demanding but it could be useful when the ARMA parameters are not well identified in general, in view of the MAIC proposed in Ng and Perron (2001). The criterion adjusts the AIC by a data dependent term so that it is robust to near cancellation of the unit roots in both the autoregressive and moving average polynomials. This is precisely the situation when $I(\theta_0)$ is far from $J(\theta_0)$.

Other selection procedures have been proposed. The PIC of Phillips and Ploberger (1996) can be seen as a generalization of the BIC. Like the TIC, it also uses a data dependent term in place of k as a measure of model complexity. But most have been shown to be related to the AIC or the BIC. For example, Rissanen (1986b) suggests using a predictive principle that minimizes the accumulative squares of prediction errors. Wei (1992) shows that the resulting model selection rule is asymptotically equivalent to the BIC for ergodic models. Rissanen (1986a) uses coding theory to choose a model with the minimum description length (MDL). The MDL of a fitted model has a component that depends on complexity, and another that depends on the fit. As discussed in Stine (2004), the MDL behaves like the AIC for some choice of coding parameters and the BIC for special choice of the prior.

Let m^0 be the true model, $\widehat{m}_T$ be the model selected using a procedure, and m_T^{opt} be the model that minimizes the squared loss, $L_T(m)$. A model selection procedure is said to be consistent if the probability of selecting the true model approaches one as the sample size increases, ie. $P(\widehat{m}_T = m^0) \to 1$. A concept related to consistency is asymptotic loss efficiency, defined in Shao (1997) as $L_T(\widehat{m}_T)/L_T(m_T^{opt}) \xrightarrow{p} 1$. Both notions are to be distinguished from consistent estimation of the regression function or of prediction. Consistent model selection can, however, conflict with the objective of mean-squared prediction accuracy because while the parameter estimates may be biased when the selected model is too small, the parameter estimates will not be efficient if the model is too large.

Establishing optimal values of C_T has generated much research interest, but the assumptions vary across studies. Shibata (1980) considers selecting the lag order of infinite order Gaussian autoregressions. He assumes that the data used for estimation are independent of those used in forecasting. Using the criterion $E_y(\widehat{y}_{t+h} - y_{t+h})^2 = \|\widehat{\alpha} - \alpha\|^2 + \widehat{\sigma}_p^2$, he shows that the (finite) p selected by the AIC is efficient in the sense that no other selection criterion achieves a smaller conditional mean squared prediction error asymptotically. Lee and Karagrigoriou (2001) obtain similar results for non-Gaussian autoregressions. However, Ing and Wei (2003) extend the analysis

to allow the sample used for prediction to overlap with that used in estimation. The issue is that while $C_T = 2$ will find the best model amongst the incorrect ones, the dimension of the selected model tends to be unnecessarily large. Kunitomo and Yamamoto (1985) show that under-specifying the order of the finite autoregression may actually be beneficial for prediction.

More generally, AIC is understood to fall short when it comes to consistent model selection. Shibata (1976) shows that the AIC (and thus the FPE and CP) has a non-zero probability of over-parameterizing finite order autoregressions. Shibata (1984) considers a generalized final prediction error that replaces $C_T = 2$ in the FPE with some other value, say, κ. His theoretical analysis suggests that κ needs to exceed one for prediction efficiency, and simulations suggest that approximate efficiency is still low when κ is set to two. Atkinson (1980) points out that a C_T of two might still be too small if the prediction problem is ill-conditioned. The observation that $C_T = 2$ will not lead to consistent selection of finite dimensional models is subsequently proved using various arguments.

When it comes to consistent model selection, results tend to favor a C_T that increases with T. Geweke and Meese (1981) show in a stochastic regressors setup that this condition is necessary for consistent model selection. Speed and Yu (1993) also show that the BIC with $C_T = \log T$ is desirable for prediction. Asymptotic efficiency of the BIC is also shown in Shao (1997). While it appears that $C_T = \log T$ is both consistent and optimal for prediction of finite dimensional (parametric) models with observed regressors. However, a finite dimensional model is not always the accepted framework for analysis. The apparent lack of a rule that delivers both consistent model selection and optimal prediction will be discussed again in Section 6.

3 Regularization Methods

One problem with information criteria when there is a large set of predictors with no natural ordering is that enumeration of 2^N predictive regressions is necessary. If $N = 10$, the number of candidate models is 1024, and when $N = 20$, the number increases to 1048576. Even with very fast computers, evaluating 2^N models and interpreting all the results would be impractical. Furthermore, a prediction rule that works well in the estimation sample need not perform well in the prediction sample. This problem is more serious when there are many predictors since the large number predictors span a high dimensional space that is likely to capture most of the variation in the dependent variable. In the extreme case when $N = T$, a perfect fit can be found but only because the model is explaining random noise. Regularization goes some ways in resolving these two problems.

In statistics and machine learning, overfitting occurs when making a model fit better in-sample has the consequence of poor out-of-sample fit. It usually occurs when a model has too many

variables relative to the number of observations. Any method that prevents overfitting the data is a form of regularization. Information criterion and sequential testing perform L_0 regularization and can be written as

$$p_{IC} = \text{argmin}_{\ p=1,\ldots,pmax} \log \widehat{\sigma}_p^2 + \frac{C_T \|\beta\|_0}{T}.$$

since $\|\beta\|_0$ is the number of non-zero components of β. But information criteria were developed under the assumption that the regressor matrix Z has full column rank. The parameter estimates will be sensitive to small changes in the data when the eigenvalues of Z are nearly zero, which is a source of the bouncing beta problem. One way to alleviate the problem is to down-weigh the less important predictors, a method known as shrinkage. Stock and Watson (2009) use shrinkage as the unifying framework to discuss various forecast methods. For variable selection, a general shrinkage framework is bridge regressions:

$$\widehat{\delta}_B = \text{argmin}_\beta \|Y - Z\delta\|_2^2 + \gamma \sum_{j=1}^{M+N} |\delta_j|^\eta, \qquad \eta > 0.$$

The ridge estimator (also known as Tikhonov regularization) due to Hoerl and Kennard (1970) is a special case with $\eta = 2$. It is also a Bayesian estimator with Gaussian prior. The ridge estimates are defined as

$$\begin{aligned} \widehat{\delta}_R &= (Z'Z + \gamma I_{M+N})^{-1} Z'Y \\ &= \sum_{i=1}^{M+N} a_{Z,i} \frac{\mathbf{U}'_{Z,i} Y}{\widehat{d}_{Z,i}} \mathbf{V}_{Z,i} \end{aligned} \tag{5}$$

where for $i = 1, \ldots, M+N$, $a_{Z,i} = \frac{d_{Z,i}^2}{d_{Z,i}^2 + \gamma} \leq 1$. The ridge estimator thus shrinks the i-th least squares estimate by an amount that depends on the i-th eigenvalue of $Z'Z$. If all $M+N$ predictors are identical, each coefficient is $1/(M+N)$ of the size of the coefficient in a single regression. The ridge estimator can be cast as a least squares problem using the augmented data

$$Z_\gamma = \begin{pmatrix} Z \\ \sqrt{\gamma} I_N \end{pmatrix}, \quad Y_\gamma = \begin{pmatrix} Y \\ 0 \end{pmatrix}.$$

As written, the L_2 penalty treats all predictors equally and cannot distinguish must have predictors from discretionary ones though this can be easily modified to penalize only the N parameters β and not the M parameters α. While the ridge estimator will alleviate the problem of highly collinear regressors, most coefficient estimates will remain non-zero. The reason is that a convex penalty with $\eta > 1$ will not yield a sparse model and efficiency of the estimator decreases with p. The more serious limitation of the L_2 penalty is that least squares estimation is infeasible when $p > T$ even when Z has full column rank.

3.1 LASSO

A method that has received a great deal of attention in the statistics literature is the least absolute shrinkage selection operator (LASSO) of Tibshirani (1996). In the simple case without the must have regressors W (ie. $Z = X$ and $\delta = \beta$), LASSO solves the quadratic programming problem:

$$\widehat{\beta}_{LASSO} = \text{argmin}_\beta \frac{1}{2}\|Y - X\beta\|_2^2 \quad \text{subject to} \sum_{j=1}^{N} |\beta_j| < s$$

for some $s > 0$. The Lagrange formulation is

$$\widehat{\beta}_{LASSO} = \text{argmin}_\delta \frac{1}{2} \|Y - X\beta\|_2^2 + \gamma\|\beta\|_1.$$

Obviously, LASSO is a bridge estimator with $\eta = 1$. It is also a Bayesian estimator with a Laplace (or double exponential) prior.

The main difference between a LASSO and a ridge regression is the use of a L_1 instead of an L_2 penalty. This difference turns out to important because an L_2 penalty only shrinks coefficients to zero but never sets them to zero exactly. In contrast, an L_1 penalty can set an estimate to zero, thereby excluding the corresponding variable from the active set. LASSO thus performs shrinkage and variable selection simultaneously, a property known as soft-thresholding. Because of the sparseness of the final active set, the LASSO estimates tend to be much less variable than the ridge estimates.

A second difference is that the ridge coefficients of correlated predictors are shrunk towards each other, while LASSO tends to pick one and ignore the rest of the correlated predictors. This latter property is a consequence of the fact that the LASSO penalty is convex but not strictly convex. In regularization problems, a strictly convex penalty has the effect that predictors with similar properties will have similar coefficients. A strictly convex penalty can be obtained by taking a convex combination of a L_1 and a L_2 penalty. The result is the 'elastic net' (EN) estimator

$$\widehat{\beta}_{EN} = \text{argmin}_\delta \frac{1}{2} \|Y - X\beta\|_2^2 + \gamma_1 \sum_{j=1}^{N} |\beta_j| + \gamma_2 \sum_{j=1}^{N} \beta_j^2.$$

The penalty function is strictly convex when $\frac{\gamma_2}{\gamma_1+\gamma_2} > 0$. An appeal of the EN estimator is that strongly correlated variables are chosen as a group. By defining the augmented data

$$X^+ = (1+\gamma_2)^{-1/2} \begin{pmatrix} X \\ \sqrt{\gamma_2} I_N \end{pmatrix}, \qquad Y^+ = \begin{pmatrix} Y \\ 0_N, \end{pmatrix},$$

the elastic net estimator can be formulated as a LASSO problem with regularization parameter $\gamma_{EN} = \frac{\gamma_1}{\sqrt{1+\gamma_2}}$. The EN problem can be treated as though it is LASSO problem.[1]

[1] A review of LASSO and related methods is provided by Belloni and Chernozhukov (2011). Technical details can be found in Hesterberg, Choi, Meier, and Fraley (2008); Fan and Lv (2010); Belloni and Chernozhukov (2011).

There are many ways to write the LASSO problem and each yields different insight. For example, using

$$\|\beta\|_1 = \sum_{j=1}^{N} |\beta_j| = \text{sgn } (\beta)'\beta,$$

the LASSO penalty can be written as sgn $(\beta)'\beta$, while the ridge penalty $\beta'\beta$. Must have predictors W can be incorporated by considering the problem

$$\min_{\alpha,\beta} \frac{1}{2}\|Y - W\alpha - X\beta\|_2^2 + \gamma \text{ sgn } (\beta)'\beta.$$

Note that the L_1 penalty is only applied to β. Let M_W be the idempotent matrix that projects onto the space orthogonal to W. The first order conditions hold that for any $j \in \mathcal{A}$,

$$\mathbf{X}_j' M_W (Y - X\beta) = \gamma \text{sgn } (\beta_j), \tag{6}$$

implying that $|\mathbf{X}_k' M_W (Y - X\beta)| \leq \gamma$ for $k \notin \mathcal{A}$. This makes clear that LASSO regressions with W can be analyzed as if data $\widetilde{X} = M_W X$ and $\widetilde{Y} = M_W Y$ were given. To simplify the discussion, the rest of this section assumes $Z = X$ and without considering the must have predictors W.

An implication of the L_1 penalty is that the LASSO objective function is not differentiable. Indeed, the first order conditions involve 2^N inequality constraints to reflect the 2^N possibilities for the sign of β. As a consequence, the estimator has no closed form solution except when $N = 1$. In that case, the estimator can be expressed as:

$$\widehat{\beta}_{LASSO} = (\widehat{\beta}_{LS,1} - \gamma)_+ \text{sgn } (\widehat{\beta}_{LS,1}). \tag{7}$$

However, Fu (1998) shows that this result for $N = 1$ can be exploited even when $N > 1$. The idea is to find the solution to

$$\frac{1}{2}\|Y - \sum_{k \neq j} \mathbf{X}_k' \beta_k - \mathbf{X}_j' \beta_j\|_2^2 + \gamma \sum_{k \neq j}^{N} \text{sgn } (\beta_k)\beta_k + \gamma \text{sgn } (\beta_j)\beta_j$$

for each $j = 1, \ldots N$ while holding $k \neq j$ fixed and iterative until the estimates converge. In this coordinate-wise descent algorithm, the partial residual $Y - \sum_{k \neq j} \mathbf{X}_k' \beta_k$ is treated as the dependent variable, and $\mathbf{X}_j$ is the single regressor whose coefficient estimate $\widetilde{\beta}_{LS,j}$ is defined by (7). The LASSO path traces out $\beta(\gamma)$ as the regularization parameter γ changes. Rosset and Zhu (2007) show that the optimal path $\widehat{\beta}(\gamma)$ is piecewise linear in γ. This is an attractive property because the solution path can be computed at the same cost as a least squares calculation. A more efficient solution can be obtained by using the homotopy algorithm of Osborne, Presnell, and Turlach (2000), which is related to forward stagewise regressions.

3.2 Forward Stagewise and Least Angle Regression

To motivate LASSO as a forward stagaewise regression, consider the effect of increasing $\widehat{\beta}_{LS,j}$ by $\Delta > 0$ for some $j \in [1, n]$ with $\mathbf{X}_j'\mathbf{X}_j = 1$. Let $\widetilde{\beta}_{LS} = \widehat{\beta}_{LS} + \Delta \cdot 1_j$ where 1_j is zero except in the j-position. By direct calculations,

$$\begin{aligned}
\mathcal{L}(\widetilde{\beta}_{LS}; j) - \mathcal{L}(\widehat{\beta}_{LS}) &\equiv \sum_{t=1}^{T}(y_t - X_t'(\widehat{\beta}_{LS} + \Delta 1_j))^2 - \sum_{t=1}^{T}(y_t - X_t'\widehat{\beta})^2 \\
&= \sum_{t=1}^{T}(\widehat{\epsilon}_t - X_{tj}\Delta)^2 - \sum_{t=1}^{T}\widehat{\epsilon}_t^2 \\
&= \sum_{t=1}^{T} -2\Delta\widehat{\epsilon}_t X_{tj} + \Delta^2 X_{tj}^2.
\end{aligned}$$

The above implies that the change in sum of squared residuals as a result of perturbing the j-th potential regressor is determined by its correlation with the least squares residuals. For given Δ, the predictor that generates the largest decrease in sum of squared residuals is the one most correlated with the fitted residuals at each step. This idea of 'gradient descent' has long been used in optimization problems. What is new is that gradient descent can be adapted to model fitting if it is considered in function space where in regression analysis, the function of interest is the conditional mean. This insight, due to Friedman (2001), is the principle behind forward stagewise regressions which can generically be described as follows:

Forward Stagewise Regression initialize $r = Y$ and $\beta = 0_N$. Let ν be some small number. Repeat (1) and (2) until r is uncorrelated with all predictors:

1. find j such that $\mathbf{X}_j$ is most correlated with the current residuals, r.
2. update $\beta_j = \beta_j + \nu \cdot \text{sgn}\left(\text{corr}\left(\mathbf{X}_j, r\right)\right)$ and $r = r - \nu \cdot \text{sgn}\left(\text{corr}\left(\mathbf{X}_j, r\right)\right)\mathbf{X}_j$.

A forward stagewise regression creates a coefficient path that includes one variable at a time and sequentially updates the fit. At each stage, the variable most correlated with the current residuals is chosen, and each predictor is always moved in the direction of corr $(\mathbf{X}_j, r)$. The active set $X_{\mathcal{A}}$ is then determined by a stopping rule that would terminate the algorithm. In principle, the variables can move as a group. As discussed in Hastie, Tibshirani, and Friedman (2001), an incremental forward stagewise regression that moves one variable at a time can be easily devised.

An important development in regularized regressions is the least angle regression (LAR) due to Efron, Hastie, Johnstone, and Tibshirani (2004). LAR sequentially builds up the regression fit by increasing the coefficient of the predictor until it is no longer the one most correlated with

the residual, at which point, the competing predictor joins the active set. In other words, the predictors in the active set are pushed in the joint least squares direction until some other regressor matches their correlation with the current residuals. Under LAR, all predictors in the active set have common correlation c with the current residual r:

$$\mathbf{X}_j' r = c \cdot \text{sgn}\,(\mathbf{X}_j' r) \tag{8}$$

while $\mathbf{X}_k' r \leq c$ for $k \notin \mathcal{A}$. Theorem 3 of Efron, Hastie, Johnstone, and Tibshirani (2004) indicates that the degree of freedom after m steps of LAR is approximately m. This suggests to stop after m steps by minimizing the statistic $CP = (1/\widehat{\sigma}^2)SSR_m - T + 2m$, where SSR_m is the sum of squared residuals at the m-th step.

LAR is important because it provides a unifying view of LASSO and seemingly related statistical procedures. The LAR moment condition defined by (8) is evidently similar to that of LASSO given in (6) because both update the fit based on the relation between the predictors and current residuals. While LAR puts no sign restrictions, $\widehat{\beta}_{j,LASSO}$ agrees in sign with sgn $(\text{corr}(\mathbf{X}_j, r))$. Hence as shown in Efron, Hastie, Johnstone, and Tibshirani (2004), the LAR-LASSO algorithm requires that the coefficient be removed from the active set and joint least squares recomputed when a non-zero coefficient hits zero.

While it is clear that LASSO performs shrinkage via the L_1 penalty, less obvious is that methods that do not directly impose an L_1 penalty implicitly mimic features of the L_1 loss and hence can be implemented using LAR. For example, the L_2 boosting of Buhlmann and Yu (2003) restricts successive revisions in $\widehat{\beta}_j$ to agree in sign with sgn $(\text{corr}(\mathbf{X}_j, r))$. Also related is forward stagewise regression which computes the best direction at each stage. If the direction of predictor j does not agree with the sign of corr $(r, \mathbf{X}_j)$, the direction is projected onto the positive cone spanned by the signed predictors. Thus a forward stagewise regression uses only the non-negative least squares directions while LAR use also the negative directions in the active set of variables. In this sense, LAR is a democratic forward stagewise regression.

As seen earlier, information criteria is a form of L_0 regularization. Statistical theory does not favor L_1 penalty over L_0 per se. Heavy shrinkage approximates L_1 regularization which may improve mean-squared prediction accuracy if the bias-variance trade-off is favorable. Ideally, one would like a procedure to have the oracle property of selecting the correct subset model and has an estimation/prediction error rate that is as good as if the true underlying model were known. However, LASSO is not an oracle procedure because any regularization yields biased estimates that may lead to suboptimal estimation risk.

The crucial parameter in L_1 regularization problems is obviously γ. Donoho, Johnstone, Kerkyacharian, and Picard (1995) show that with suitable choice of γ, the LASSO estimates can be

near-minimax optimal with the sparsity property that the zero components of the true parameter vector will be estimated to be zero with probability approaching one as the sample size increases. But how should γ be chosen? As shown in Buena (2008), consistent subset variable selection using LASSO when $N > T$ requires a carefully chosen penalty parameter. Fan and Li (2001) recommend to use penalties such that the resulting estimators have three properties: (i) sparsity, such that small estimated coefficients are automatically set to zero; (ii) near unbiasedness especially when the true coefficients are large; and (iii) continuity in the data to reduce instability in model prediction. They find that if data-driven rules are used to select γ, LASSO tends to have many false positive variables in the selected model. Fan and Lv (2010) note that stringent conditions must hold for LASSO to consistently select the true model. Zou (2006) suggests to re-weight the penalty function in order for LASSO to have the oracle property. This leads to the adaptive LASSO estimator

$$\widehat{\beta}_{LASSO} = \text{argmin}_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \gamma \sum_{j=1}^{N} \psi_j |\beta_j|,$$

where ψ_j are weights that can be set to some initial estimator of β_j such as $\widehat{\beta}_{LS,j}$. Belloni and Chernozhukov (2012) propose a data dependent rule for γ and analyze the two roles (shrinkage/estimation and model selection) played by LASSO. They show that applying least squares to a model selected by LASSO (known as OLS post LASSO) performs at least as well as LASSO in terms of the rate of convergence and has smaller bias. The reason is that LASSO only omits components with relative small coefficients relative to the oracle, and the OLS post LASSO estimator removes some of the shrinkage bias from LASSO. The estimator can be superior to LASSO and has a better convergence rate than LASSO if the selected model includes all components of the pseudo-true model.

Meinshausen and Buhlmann (2006) consider graphical models for estimating entries of the inverse covariance matrix of N Gaussian series observed over T periods. They fit a LASSO model to each variable using all other variables as predictors. They show that LASSO will consistently estimate the non-zero entries of the inverse covariance matrix when N increases faster than T, but only if a neighborhood stability condition is satisfied. This is related to the 'irrepresentable' condition, which is shown in Zang and Yu (2006) to be almost necessary and sufficient for 'sign' consistency of LASSO under more general conditions. That is to say, the probability that the sign of $\widehat{\beta}_{LASSO}$ agrees with that of β tends to one as the sample size increases. This ensures that $\widehat{\beta}$ has the same support as the true regression coefficients with probability one asymptotically. Loosely speaking, the condition requires that the correlation between the relevant and the irrelevant predictors not to be too large. This condition is, however, quite restrictive when N is large.

Meinshausen and Buhlmann (2006) find that the regularization parameter γ in LASSO that is optimal for prediction is not always optimal for variable selection. More precisely, the γ that yields the smallest mean-squared prediction error tends to suggest more predictors than are present in the true model, even though the true model is contained in the selected model with high probability. Using an orthogonal design, Leng, Lin, and Wahba (2006) show that if the criterion of prediction accuracy is used to choose γ, the set of variables selected by LASSO are not consistent for the true set of important predictions.

While L_1 type regularization solves many problems inherent in ridge regressions, it apparently does not eliminate the conflict between consistent model selection and prediction. Fan and Lv (2010) noted that the problem of collinearity amongst predictors is especially challenging in high dimensional model selection because spurious collinearity can give rise to overfitting. An alternative that has received increased attention when the regressors are highly correlated is to combine information from the observables.

4 Dimension Reduction Methods

While regularization picks out the empirical relevant variables from amongst the potentially relevant ones, a different approach is to use all data available intelligently. For example, one can use a subset of the regressors at a time and then combine the forecasts produced by the different subset of regressors. This is the method of model averaging pioneered by Bates and Granger (1969), reviewed in Timmermann (2006), and further developed in Hansen (2008); Hansen and Racine (2012). Here, I focus on methods that simultaneously consider all predictors.

4.1 Principal Components and Factor Augmented Regressions

A popular technique that combines the potentially relevant predictors X_t into new predictors is principal components. By definition, the $T \times N$ principal components of X are defined as

$$X_{PC} = XV_X = U_X D_X.$$

The j-th principal component $\mathbf{X}_{PC,j}$ is the linear combination of X that captures the j-th largest variation in X. The left singular vectors of X multiplied by the eigenvalues are also known as the factor scores. A principal component regression replaces the $T \times N$ predictor matrix X with a $T \times r_X$ sub-matrix of principal components. Let $X_{PC,1:r_X}$ be the first r_X columns of X_{PC} that corresponds to the r_X largest eigenvalues of X. To fix ideas, suppose that there are no must have

predictors W. The estimator using the first r principal components as regressors is

$$\begin{aligned}\widehat{\beta}_{PC} &= (X'_{PC,1:r_X}X_{PC,1:r_X})^{-1}X'_{PC,1:r_X}Y \\ &= V_{X,1:r_X}D^{-1}_{X,1:r_X}U'_{X,1:r_X}Y \\ &= \sum_{i=1}^{r_X}\frac{1}{d_{X,i}}\mathbf{U}'_{X,i}Y\mathbf{V}_{X,i}. \end{aligned} \tag{9}$$

The in-sample fit is

$$\widehat{Y}_{PC} = X_{PC,1:r_X}\widehat{\beta}_{PC} = U_{X,1:r_X}U'_{X,1:r_X}Y.$$

Notice that compared to the least squares estimator, the sum only involves $r_X \leq N$ components. In other words, $\widehat{\beta}_{PC}$ puts a unit weight on the first r_X components and ignores the remaining ones. Thus r_X controls the degree of shrinkage from $\widehat{\beta}_{LS}$ towards zero. This contrasts with the ridge estimator in which all singular values $\widehat{d}_{X,i}$ are shrunk towards zero.

Principal component analysis is often seen as a numerical tool that reduces the dimension of the data but has weak statistical foundations because no probability model is specified. It is thus an unsupervised dimension reduction technique. In contrast, factor analysis assumes that the data have a specific structure. However, Tipping and Bishop (1999) show using a small T large N setup that a principal components regression model can be seen as a Gaussian latent variable model that is closely related to factor analysis. The distinction between principal components and factor analysis may not be as sharp as once thought.

While a factor interpretation is not necessary to motivate the use of principal components as regressors, more analytical results are available when a factor structure is imposed. Suppose that y_t can be well approximated by the infeasible regression

$$y_t = W'_t\alpha + F'_t\beta_F(L) + \epsilon_t \tag{10}$$

where F_t is a $r_Y \times 1$ vector of unobserved common factors, $\beta_F(L)$ is a polynomial in the lag operator of order p_F. A factor augmented regression is obtained when $\widehat{F}_t$ is used in place of F_t in (10), as though F_t were observed. Stone and Brooks (1990) calls $\widehat{F}_t$ the constructed predictors while Stock and Watson (2002a,b) refer to $\widehat{F}_t$ as diffusion indices. A h period ahead diffusion index forecast is

$$\widehat{y}_{T+h|T} = W'_{T|h|T}\widehat{\alpha} + \widehat{F}'_{T+h|T}\widehat{\beta}_F(L).$$

The key to factor augmented regressions is that the latent factors can be estimated precisely from a large number of the observed predictors x_{it} that can be represented by the factor model

$$x_{it} = \lambda'_i\mathbb{F}_t + e_{it} \tag{11}$$

where $\mathbb{F}_t$ is a $r_X \times 1$ vector of latent common factors, λ_i are the loadings, and e_{it} are the idiosyncratic errors. As the factors relevant for forecasting need not be the same as the set of pervasive factors in X_t, F_t (of dimension r_Y) is kept distinct from $\mathbb{F}_t$(of dimension r_X).

Factor analysis is attributed to Spearman (1904) who suggests that intelligence is composed of a factor common to all attributes such as mathematics, language, music, etc., as well as factors that are specific to each attribute. Associated with a factor model is the population covariance structure $\Sigma_X = \Lambda\Sigma_{\mathbb{F}}\Lambda' + \Sigma_\epsilon$. In classical factor analysis, Σ_ϵ is typically a diagonal matrix, meaning that the errors e_{it} are uncorrelated over i and t. Chamberlain and Rothschild (1983) allow e_{it} to be weakly correlated both serially and cross-sectionally and call factor models with these properties 'approximate factor models'. For $X_t = (x_{1t}, \ldots, x_{NT})'$ to have r_X strong pervasive factors in an approximate factor model, the r_X largest eigenvalues of the $N \times N$ population covariance matrix of X_t must diverge to infinity as N increases. There are thus r_X 'factor eigenvalues' and $N - r_X$ 'idiosyncratic eigenvalues'. A factor structure is said to be strong if the factor eigenvalues and well separated largest idiosyncratic eigenvalue and $\Lambda'\Lambda/N \to \Phi$ for some Φ that is non-degenerate. Connor and Korajczyk (1993) were the first to use the method of principal components to estimate approximate factor models. The idea is that when N is large, the variation of ϵ_{it} will then be dominated by that of the common component $\lambda_i'\mathbb{F}_t$. The eigenvalue decomposition of Σ_X will be asymptotically equivalent to that of $\Sigma_X - \Sigma_\epsilon$ when N tends to infinity.

When y_t also belongs to X_t, r_Y can be set to r_X, making $\widehat{F}_t$ the r_X static principal components of X.[2] Thus one may write $\widehat{F} = \widehat{\mathbb{F}}_{1:r_X} = \sqrt{T}U_{X,1:r_X} = \sqrt{T}D_X^{-1}\mathbf{X}_{PC}$. The relation between principal components regression and factor augmented regression is easy to see when $p_F = 0$ and W_t is empty. Then $\widehat{\beta}_F(L) = \widehat{\beta}_F$,

$$\widehat{\beta}_F = \frac{1}{T}\widehat{\mathbf{F}}'Y = \frac{1}{\sqrt{T}}U_{X,1:r_X}'Y = \frac{1}{\sqrt{T}}D_X^{-1}\widehat{\beta}_{PC,i}. \tag{12}$$

The diffusion index forecast is

$$\widehat{Y}_F = \sum_{j=1}^{r_X} U_{X,j}U_{X,j}'Y = \widehat{Y}_{PC}.. \tag{13}$$

A review of factor based forecasts is given in Stock and Watson (2006). Of note from (13) is that $\widehat{Y}_{PC}$ and $\widehat{Y}_F$ are numerically equivalent. This suggests to use the principal components as regressors in factor augmented regression. This is useful because compared to maximum likelihood estimation, principal components are easy to construct. Furthermore, using the probability structure of a

[2]Static principal components are distinguished by dynamic principal components, developed in Brillinger (1981) for large T fixed N, and extended in Forni, Hallin, Lippi, and Reichlin (2000) to large panels. Boivin and Ng (2005) finds that with appropriate choice of the tuning parameters, dynamic and static factors yield similar forecasts. However, estimation of static factors is computationally simpler. The relation between static and dynamic factors can be found in Forni, Hallin, Lippi, and Reichlin (2005), Bai and Ng (2008b), Stock and Watson (2005).

model with strong factors, statistical statements about principal component estimates can be made. Connor and Korajczyk (1993) show that $\widehat{F}_{1:r_X}$ consistently estimates the space spanned by the common factors as $N \to \infty$ with T fixed. Assuming N and T are both large, Stock and Watson (2002a) show uniform convergence of $\widehat{\mathbb{F}}_t$ to the space spanned by $\mathbb{F}_t$. But to validate use of $\widehat{\mathbb{F}}_t$ as regressors, weaker results suffice. Bai and Ng (2002) show that if $\widehat{F}_t$ is a $k > 1$ vector of factor estimates, there is a matrix H of rank $\min(k, r_X)$ such that $C_{NT}^2 \left(\frac{1}{T} \sum_{t=1}^{T} \|\widehat{\mathbb{F}}_t(k) - H\mathbb{F}_t\|^2 \right) = O_p(1)$, where $C_{NT} = \min[\sqrt{N}, \sqrt{T}]$.

In practice, p_Y lags of y_t are usually included in the regression, thereby defining the must have predictors $W_t = (1, y_t, y_{t-1}, \ldots, y_{t-p_Y})'$. To accommodate W_t, consider the generalized factor representation of X:

$$X = W_X \Psi + \mathbb{F}\Lambda + e$$

where W_X could overlap with W in the prediction equation. The presence of W_X necessitates a different way to estimate the principal components. To proceed, note that if Ψ were observed, then $\widetilde{X} = X - W_X\Psi = \mathbb{F}\Lambda + e$ has a factor structure. Furthermore, if F were observed, then Ψ can be estimated by a least squares regression of $XM_{\mathbb{F}}$ on $W_X M_{\mathbb{F}}$ where $M_{\mathbb{F}} = I - \mathbb{F}(\mathbb{F}'\mathbb{F})^{-1}\mathbb{F}'$. Stock and Watson (2005) suggest an iterative principal components estimator whose properties are formally analyzed in Bai (2009):

Algorithm: Iterative Principal Components:

1 Estimation of $\mathbb{F}$: Initialize $\widetilde{X}_W = X$.

   i Let $\widehat{\mathbb{F}}$ be $\sqrt{T}$ times the eigenvectors corresponding to the r_X largest eigenvalues of $\widetilde{X}\widetilde{X}'$. Let $\widetilde{\Lambda}$ be obtained by least squares regression of X on $\widetilde{\mathbb{F}}$.

   ii Estimate Ψ by regressing $XM_{\widehat{\mathbb{F}}}$ on $W_X M_{\widehat{\mathbb{F}}}$ where $M_{\widehat{\mathbb{F}}} = I - \widehat{\mathbb{F}}(\widehat{\mathbb{F}}'\widehat{\mathbb{F}})^{-1}\widehat{\mathbb{F}}'$. Let $\widetilde{X}_W = X - W_X\widehat{\Psi}$. Return to step (i) until $\widehat{\Psi}$ converges.

2 Regress Y on W and $\widehat{F}$ to obtain $(\widehat{\alpha}, \widehat{\beta}_F)$, where $\widehat{F} \subset \widehat{\mathbb{F}}$.

The principal components estimates can always be obtained by iterative estimation whether or not W_t is present. In psychometrics, there is a long tradition in estimating factor models by the method of alternating least squares (also referred to as PRINCIPALS). These matrix decomposition methods do not require specification of a probability model, see, eg, Young, Takane, and de Leeuw (1978). The econometrics literature specifies a probability model and shows that iterative principal components can consistently estimate the space spanned by the factors even in the presence of W.

A criticism of factor augmented regressions is that the factors are estimated without taking into account that the objective is to forecast Y. Factors that have good explanatory power for X may

not be good predictors for Y even if $y_t \subset X_t$. More precisely, a factor augmented regression first estimates $\mathbb{F}$ by maximizing $R^2_X = 1 - \|X - \mathbb{F}\Lambda\|^2/\|X\|^2$ where $\Lambda = (\mathbb{F}'\mathbb{F})^{-1}\mathbb{F}'X$. Given $\widehat{\mathbb{F}} = XV_{X,1:r_X}$, estimates of α and β are then obtained by maximizing $R^2_Y = 1 - \|Y - W\alpha - \widehat{\mathbb{F}}\beta(L)\|^2/\|Y\|^2$. While we can select $\widehat{F}_t$ from $\widehat{\mathbb{F}}_t$, a problem that will be discussed in the next section, the $\widehat{\mathbb{F}}$ are constructed the same way irrespective of Y. The next section discuss selected methods that address this problem.

4.2 Reduced Rank and Partial Least Squares Regressions

Rao (1964) suggests reduced rank regressions that find F with the fit of Y taken into account. The objective is to maximize $R^2_Y = 1 - \|Y - F\beta_F\|^2/\|Y\|^2$ with respect to β and $F = XV_R$. Taking $\widehat{\beta}_F$ to be $(F'F)^{-1}F'Y$, the concentrated objective function

$$\|Y - F(F'F)^{-1}F'Y\|^2$$

is minimized subject to the constraint that $F'F = I$ and $F = XV_R$. Since the problem reduces to maimizing $\text{tr}(Y'FF'Y)$, the solution is to take $\widehat{F}$ to be the first r_R unit eigenvectors of $P_XYY'P_X$. Since $P_X = X(X'X)^{-1}X'$ is the projector on the subspace spanned by the columns of X, $\widehat{F}$ is in the subspace of X. From $V_R = (X'X)^{-1}X'\widehat{F}$, the implicit estimates from a reduced rank regression of Y on X is $\widehat{\beta}_X = V_R\widehat{\beta}_{\widehat{F}}$.

Two other methods that target the components to Y are canonical correlation analysis (CCA) and partial least squares (PLS). Both allow Y to be multivariate. CCA is due to Hotelling (1936). For one component, CCA maximizes the correlation coefficient

$$\rho = \frac{w_x'XY'w_y}{\sqrt{(w_x'XX'w_x)(w_y'YY'w_y)}}$$

by solving for projection vectors w_x and w_y. For multiple components, CCA maximizes $\text{tr}(W_x'XY'W_y)$ subject to $W_X'XX'W_x = I$ and $W_y'YY'W_y = I$. The projection matrix W_x is given by the r_C eigenvectors of the generalized eigenvalue problem $XY'(YY')^{-1}YX'w_x = \mu XX'w_x$ where μ is the eigenvalue.

The method of partial least squares, developed in Wold (1969), is especially popular with chemical engineers. Sun, Ji, Yu, and Ye (2009) show that CCA differs from PLS in that the latter maximizes covariance instead of correlation between Y and X. Statistical aspects of PLS are discussed in Dijkstra (1983). Wold's NIPALS algorithm when Y is a column vector is as follows (Kramer (2007)):

Algorithm PLS: Demean Y and also standardize X. Let $X^1 = X$. For $m = 1, \ldots r_P$:

i Set $w^m = X^{m\prime}Y$;

ii Define $\widehat{F}^m = X^m w^m$;

iii update $X^{m+1} = M^m X^m$ and $Y^{m+1} = M^m Y^m$ where $M^m = I - P^m$ and $P^m = \widehat{F}^m(\widehat{F}^{m\prime}\widehat{F}^m)^- \widehat{F}^{m\prime}$.

The PLS prediction is $\widehat{Y}_{PLS} = \sum_{j=1}^{r_P} P^m Y$. It can be shown that $\widehat{F}^m = X^m V^m$ where V^m is the eigenvector corresponding to the m-th eigenvalue of $X^{m\prime}Y^m Y^{m\prime}X^m$. The algorithm can also be understood as first regressing Y on X^1 to get least squares coefficients $\widehat{\beta}^1_{PLS}$ that is, up to a factor, the weight vector w^1. Since $\widehat{F}^1$ is a weighted average of Y using the covariance between X^1 and Y as weights, PLS forms the $\widehat{F}^1$ with information about Y taken into account. Subsequent components are formed by choosing w^{m+1} to maximize $\text{cov}\,(X^{m+1}w^{m+1}, Y)$ subject to the constraint that $\|w^{m+1}\| = 1$ and orthogonal to $\widehat{F}_1, \ldots \widehat{F}^m$, noting that X^m has the effect of $\widehat{F}^m$ partialled out from X. The acronym PLS has also been taken to mean 'projection to latent structure' since it chooses the subspaces of the column space of X sequentially and project Y onto these subspaces. Notably, PLS also indirectly optimizes on the explained variance of X.

The least squares estimator obtains when $N = r_P$, making r_P the regularization parameter of a PLS regression. Lingjaerde and Christophersen (2000) show that

$$\begin{aligned} \widehat{\beta}_{PLS} &= \sum_{i=1}^{r_P} \frac{b_{X,i}}{d_{X,i}} \mathbf{U}_i' Y \mathbf{V}_{X,i} \\ b_{X,i} &= 1 - \prod_{j=1}^{r_P} \left(1 - \frac{d^2_{X,j}}{\theta_j}\right) \end{aligned} \tag{14}$$

where θ_j are the eigenvalues of a matrix with columns that form the orthogonal basis of $\mathcal{K} = \{X'Y, (X'X)^{-1}X'Y, \ldots (X'X)^{r_P-1}X'Y\}$, the Krylov space of $X'X$ and $X'Y$. Obviously, θ_j depends on Y and $\widehat{\beta}_{PLS}$ is non-linear function of Y. The PLS shrinkage factor is stochastic because of the dependence on Y and has the peculiar feature that it can exceed one. An alternative to PLS is latent root regressions of Webster, Grant, and Mason (1974) which forms the principal components of the augmented data $[Y|X]$.

Stone and Brooks (1990) show that PCA, PLS and OLS can all be analyzed from the perspective of generalized canonical correlations. Reduced rank regressions and PLS can be in principle be generalized to include must have predictors by working with the residuals from projecting Y and X on W. There is on-going work that constructs components adapted to Y. See, for example, Li (1991) for sliced inverse regressions and the model based approach of Cook and Forzani (2008).

5 Three Practical Problems

The methods discussed in the previous two sections are all biased regression techniques. They seek to shrink the OLS coefficient vector away from directions in the predictor space that have

low variance. Ridge regressions reweigh $\widehat{\beta}_{LS}$ using the eigenvalues of X. LASSO uses rectangular weights to truncate the small coefficients to zero. Principal component regressions use rectangular weights to truncate small eigenvalues to zero. Partial least squares re-weighs the least squares estimates according to the eigenvalues of X and $X'Y$ and additionally truncates small eigenvalues of X to zero. Note that the active regressor set $X_{\mathcal{A}}$ associated with all these methods usually coincides with X, in contrast to LASSO and information criteria type procedures. Even though all methods perform some form of shrinkage, they produce different models. Which one is best depends on the objective of the exercise and the data structure on hand.

This section discusses three problems that are still being debated or warrant further work. The first is whether or not to construct components with the variable of interest in mind. The second concerns variable selection when the predictors are themselves estimated. The third is the robustness of model selection rules over the parameter space.

5.1 To Target or Not to Target

As the principal components of X do not depend on Y, linearity of $\widehat{Y}_{PC}$ in Y ensures that the shrinkage produced by principal components decrease as r_X increases. While PLS is designed to shrink away from the predictor space in the low variance directions, Frank and Friedman (1993) find that PLS routinely inflates the high variance directions. The consequence in finite samples is to increase both the bias and the variance of the coefficient estimates. This suggests that the PLS shrinkage may not decrease with r_P. There is an apparent trade-off between the information content of the components, and ease in controlling the degree of shrinkage. At least for PLS, targeting the components to Y does not necessarily give better finite sample properties. It is however unclear whether this non-monotonicity of the shrinkage factor documented for PLS is generic of methods that target the components to Y.

Helland and Almoy (1994) assume normality and derive asymptotic criteria for comparing principal component regressions and PLS. Simulations in Almoy (1996) suggest that these methods generally have similar properties for the data generating processes considered. Kiers and Smilde (2007) find that PLS work well when the coefficients of the population regression lie in the subspace spanning the first few principle components of the predictor variables.

There has always been a disagreement as to whether one should reduce the dimension of X on the basis of the marginal distribution of X, or the conditional distribution of Y given X. As Cook (2007) points out, Fisher (1924) recognizes the need for dimension reduction in regression analysis but cautions that predictors might be spuriously chosen if reference is made to the dependent variable. On the other hand, Cox (1968, p.272) among others see no strong reason why Y should not be closely related to the least important principal component. Kiers and Smilde (2007) take

the view that aiming to explain both the predictors and the endogenous variable will be better able to yield models that predict well both in and out of samples. Li (2007) conjectures that the first principal component of an arbitrary covariance matrix of X will have a tendency to be more correlated with Y than other principal components of X. Nonetheless, he concludes in favor of dimension reduction of X with reference to Y especially when N is large. However, the T and N considered in these simulations are much smaller than typical configurations of macroeconomic data.

Bai and Ng (2006b) call variables selected for the purpose of predicting Y the 'targeted predictors'. They evaluate the usefulness of forming targeted predictors from 132 potentially relevant predictors by soft and hard thresholding for the purpose of forecasting inflation. They find that targeted predictors generally yield better forecasts but the composition of the predictors changes with the forecast horizon. This leads to the point raised by Hansen (2010) that in multi-period forecast, the final prediction error is approximately the expected sample sum of squared residuals plus a penalty term that is a function of the long-run variance rather than the short-run variance appropriate for one-step ahead forecasts. This imples that criteria developed for one-period ahead prediction are biased for the final prediction error of multi-step forecasts. This suggests that targeting is necessary at least with respect to the forecast horizon.

5.2 Determining the Number of Generated Predictors

It may sometimes be necessary to replace latent predictors by estimated ones. As is known from Pagan (1984), the variance of the second-step estimates are inflated by the sampling error in the first stage estimation. This has implications for variable selection. Consider first the small N setup. Suppose that one of the potential predictors F_t is latent but that a small number of observables X_t are available to form an estimate $\widehat{F}_t$ using a first step regression. The feasible prediction model is $y_{t+h} = W_t'\alpha + \widehat{F}_t'\gamma_F + \epsilon_{t+h}$. To see which of the available predictors $(W_t', \widehat{F}_t')'$ are relevant for predicting Y, Bai and Ng (2008a) suggest a modified FPE:

$$\widehat{FPE}_p = \log \widehat{\sigma}_p^2 + \frac{2p}{T-p} + \frac{\widehat{c}_n}{T-p}$$

where $\widehat{c}_n = \widehat{\gamma_F}' \widehat{Avar}(\widehat{F}_T)\widehat{\gamma_F}/\widehat{\sigma}_p^2$, and $\widehat{Avar}(\widehat{F}_T)$ is the asymptotic variance that arises from having to estimate F_T. The additional penalty $\frac{\widehat{c}_n}{T-p}$ accounts for the sampling variability due to regressors generated by coefficients that are $\sqrt{T}$ consistent. Notably, the adjustment factor is asymptotically negligible as $T \to \infty$ for fixed p. Adjustment terms of this nature can be expected for other model selection procedures.

When there are N possibly larger than T predictors that contain information about F_t (possibly a vector), the columns of $\widehat{F}_t$ are no longer estimated from first step regressions but are now the

principal components of X_t. The feasible factor augmented regression is

$$y_{t+h} = W_t'\alpha + \widehat{F}_t'\beta_F(L) + \epsilon_{t+h}$$

where $W_t = (1, y_t, y_{t-1}, \ldots, y_{t-p_Y})'$ and $\widehat{F}_t \subset \mathbb{F}_t$ is of dimension r_Y, while $\widehat{\mathbb{F}}_t$ is of dimension r_X. As noted in Eickmeier and Ziegler (2008), there is much heterogeneity in empirical work about the choice of both parameters. Some simply fix r_X and r_Y a priori. Others use data dependent methods such as the PCP and ICP criteria of Bai and Ng (2002) to optimally determine r_X. These are generalizations of the CP and IC to a panel context. Instead of a penalty of C_T as discussed in Section 2, the penalty term of $\min(N,T)$ is now a function of both N and T.

The PCP and ICP take as given that the objective is consistent estimation of r_X. As pointed out earlier, consistent selection of the model size does not usually lead to a model that yields minimum forecast errors. Onatski (2011) studies the problem of factor selection from the point of view of optimal prediction of *all* series in the panel so that $r_X = r_Y$. He extends Mallows's CP criterion to a factor augmented regression without W. Aassuming that $N/T \to c \in (0, +\infty)$ as $N, T \to \infty$, he suggests a new penalty term to reflect the bias in the forecasts when r_X is incorrectly specified. The results are, however, specific to the unusual objective of forecasting all series in a panel.

In the more usual case when interest is in forecasting only one series that happens to be one of the series in X, then r_Y can arguably be taken to be r_X in the factor augmented regression. Assuming that r_X does not increase with N or T, Bai and Ng (2006a) show under strong factor asymptotics that $\widehat{F} = \widehat{\mathbb{F}}_{1:r_X}$ can be treated in factor augmented regressions as though they were the latent F_t provided $\sqrt{T}/N \to 0$. In other words, there is no need to adjust the standard errors for the fact that $\widehat{F}_t$ are estimated from a preliminary step. This is unlike the generated regressors problem considered in Pagan (1984). In those problems, there is an $O_p(1)$ term that reflects sampling variability in the $\sqrt{T}$ consistent estimates of a first step regression. This term is of order $O_p(\frac{\sqrt{T}}{\min[N,T]})$ when the first step estimates are the principal components of a large panel. However, while this term tends to zero if $\sqrt{T}/N \to 0$, Ludvigson and Ng (2011) show that when $\sqrt{T}/N$ is not negligible, generated regressors in the form of estimated factors will induce an asymptotic bias in $\widehat{\beta}_F$. This effect on bias contrasts with the effect of inflated variance in the small N setup. The nature of this asymptotic bias is further analyzed in Goncalves and Perron (2011) in the context of bootstrapping. The implications for the determination of r_Y remain to be studied.

The assumption that $r_Y = r_X$ is somewhat strong as the factors that are pervasive in $x_{1t}, \ldots x_{NT}$ need not be the most important predictors for the series y_t. If $\widehat{F}_t$ was not estimated, we would simply determine r_Y by the methods discussed in information criteria or regularization. But in factor augmented regressions, $\widehat{F}_t$ are the principal component estimates Bai and Ng (2008a) suggest a modified stopping rule for boosting to account for the fact that $\widehat{F}_t$ are the principal components

estimates. They suggest to add another penalty term to information criteria:

$$ICP = \log(\widehat{\sigma}_p^2) + \frac{pC_T}{T} + \frac{r_Y C_N}{N}$$

where r_Y is the number of estimated predictors in the regression and $p = M + r_Y$ is the total number of predictors in the model being considered. An estimated predictor is penalized more heavily than an observed one. The overall penalty of an additional predictor would then vanish at rate rate of $min(N, T)$. Stock and Watson (2002a) suggest to use a modified information criteria to select r_Y for forecasting y_{t+h}:

$$ICP = \log(\widehat{\sigma}_p^2) + p \cdot g(T)$$

where $\widehat{\sigma}_p^2$ is SSR_p/T, SSR_p is the sum of squared residuals from estimating the diffusion index equation with p factors. Under the assumption that $\frac{\log N}{\log T} \to \rho > 2$, they show that prob $(\widehat{r}_Y = r_Y) \to 1$ if (i) $g(T) \to 0$ and (ii) $T^b g(T) \to \infty$ where $b < \min(.5\rho - 1, 1)$. Stock and Watson (1998) suggest to use $g(T) = \omega \log(T)/\delta_{NT}$ where $\delta_{NT} = \min(N^{1/2}/T^{1+\epsilon}, T^{1-\epsilon})$, ϵ is a small and positive number, and ω is a positive constant. Notably, both modifications require consideration of both N and T even though the prediction equation is estimated from a sample of size T.

5.3 Consistent Model Selection or Efficient Prediction?

This chapter is about methods that determine the composition of the best predictor set. Whether the predictors are observed or constructed from a preliminary step, the problem in practice comes down to the choosing a parameter that will determine how parsimonious a regression model one desires. The vast literature seems to converge towards two types of regularization parameters. One increases with the sample size (such as the BIC), and one is a constant (such as the AIC).

At least in the classical $T > N$, it is generally thought that the BIC is good if the true model is finite dimensional; otherwise the AIC finds the smallest possible model for prediction, cf. Yang (2007). Nonetheless, this view of the relative merits of AIC and BIC has not gone unchallenged. Kabaila (2002) cautions that result in Shibata (1981) that favors the AIC over the BIC is based on first fixing the data generating process, and then providing a pointwise analysis of β as T increases. This efficiency result apparently breaks down when the comparison is based on varying the data generating mechanism with T fixed to possibly some large value. Stone (1979) also notes that the comparison between the AIC and BIC is sensitive to the type of asymptotic analysis used, and there can be situations when the AIC is consistent but the BIC is not.

In the statistics literature, the tension between model consistency and optimal prediction is referred to as the AIC-BIC dilemma. The question of whether the strengths of the AIC and BIC can be combined to yield a better procedure is analyzed in Yang (2005, 2007). The main finding is

that model selection procedures cannot be both consistent and minimax rate optimal and in this sense, the strengths of the AIC and BIC cannot be shared.[3] Yang (2007) simulates iid data using the model $y_t = f(x_t) + \epsilon_t$; under Model (0), $f_0(x) = \alpha$, and under Model 1, $f_1(x) = \alpha + \beta x_t$. He shows that while the BIC is pointwise risk adaptive,[4] the AIC is minimax-rate adaptive. Yang (2007) favors combining models when different selection methods do not come to a consensus. LASSO was not in Yang's analytical or numerical analysis.

To see if the findings of Yang (2007) prevail in more general settings, I conduct a monte carlo exercise with data generated from the following models.

DGP	β	Predictors
1: $y_t = .5y_{t-1} + \beta y_{t-2} + e_t$	[-.5,.5]	(b) $1, y_{t-1}$
		(c) $1, y_{t-1}, y_{t-2}$
		(d) $1, y_{t-1}, y_{t-2}, y_{t-3}$
2: $y_t = .8x_t + \beta x_{t-1} + e_t + .5e_{t-1}$	[-.5,.5]	(b) $1, y_{t-1}$
		(c) $1, y_{t-1}, x_t$
		(d) $1, y_{t-1}, x_t, y_{t-2}$
		(e) $1, y_{t-1}, x_t, y_{t-2}, x_{t-1}$
		(f) $1, y_{t-1}, x_t, y_{t-2}, x_{t-1}, y_{t-2}$
		(g) $1, y_{t-1}, x_t, y_{t-2}, x_{t-1}, y_{t-3}, x_{t-2}$
3: $y_t = .8x_t + .5x_{t-1} + e_t + \beta e_{t-1}$	[-.5,.5]	(b) $1, y_{t-1}$
		(c) $1, y_{t-1}, x_t$
		(d) $1, y_{t-1}, x_t, y_{t-2}$
		(e) $1, y_{t-1}, x_t, y_{t-2}, x_{t-1}$
		(f) $1, y_{t-1}, x_t, y_{t-2}, x_{t-1}, y_{t-3}$
		(g) $1, y_{t-1}, x_t, y_{t-2}, x_{t-1}, y_{t-3}, x_{t-2}$
4: $y_{1t} = .4y_{t-1} + \lambda_1 \widehat{F}_t + e_{1t} + \beta e_{t-1}$	[-.5,.5]	(b) $1, y_{t-1}$
		(c) $1, y_{t-1}, \widehat{F}_t = \widehat{\mathbb{F}}_{1t}$
		(d) $1, y_{t-1}, \widehat{F}_t = (\widehat{\mathbb{F}}_{1t}\widehat{\mathbb{F}}_{2t})'$
		(e) $1, y_{t-1}, \widehat{F}_t = (\widehat{\mathbb{F}}_{1t}\widehat{\mathbb{F}}_{2t}, y_{t-2})'$
		(f) $1, y_{t-1}, \widehat{F}_t = (\widehat{\mathbb{F}}_{1t}\widehat{\mathbb{F}}_{2t}, y_{t-2}, \widehat{\mathbb{F}}_{1t-1})'$
		(g) $1, y_{t-1}, \widehat{F}_t = (\widehat{\mathbb{F}}_{1t}\widehat{\mathbb{F}}_{2t}, y_{t-2}, \widehat{\mathbb{F}}_{1t-1}\widehat{\mathbb{F}}_{2t-1})'$

where $x_t = .5x_{t-1} + u_t$, $u_t \sim N(0,1)$ and $e_t \sim N(0,.5)$, e_t and u_t are mutually uncorrelated. For each DGP, prediction model (a) has an intercept but no covariate. Results are based on $S = 2,000$ replications for $T = 100, 200$ and 500.[5] MATLAB 2012a is used to conduct the simulations. The LASSO results are based on cross-validation as implemented in MATLAB.

Let $\widehat{y}^m_{T+1|T}$ be the prediction when the estimates are based on model m as determined by the either AIC, BIC or LASSO. Relative risk is computed as the ratio of the risk associated $\widehat{y}^m_{T+1|T}$

[3] For data generated by $y_t = f(x_t) + e_t$ and risk $R_T(\widehat{f}, f) = E\|\widehat{f} - f\|_2^2$, minimax prediction risk is $\inf_{\widehat{f}} \sup_f R_T(\widehat{f}, f)$.

[4] A selection procedure is said to be pointwise risk adaptive if the estimator of $f(x_0)$ based on the selection procedure is as good as the better of $\widehat{f}_0(x_0)$ and $\widehat{f}_1(x_0)$.

[5] Results for the AIC and BIC using 20,000 replications are available.

relative to the lowest risk amongst models considered:

$$RR_m = \frac{\frac{1}{S}\sum_{s=1}^{S}(y_{s,T+1|T} - \widehat{y}^m_{s,T+1|T})^2}{\min_m \frac{1}{S}\sum_{s=1}^{S}(y_{s,T+1|T} - \widehat{y}^m_{s,T+1|T})^2}, \quad m = AIC, BIC, LASSO.$$

A relative risk above one indicates that the procedure does not produce the best possible prediction. While AIC and BIC only consider predictors ordered as listed above, the predictors selected by LASSO can be unordered. For example, in model 3, LASSO could select x_t and y_{t-3}, a configuration that would not be considered by AIC or BIC. Thus RR_{BIC} may not equal RR_{LASSO} even if both procedures select two predictors. For each of the four models, relative risk and the average model size (including the intercept) are graphed. In all the figures, the dark solid line is BIC, the broken line with a dot is AIC, and the dash line is LASSO.

In Model 1, the data are generated from an AR(2) model in which the true β is varied between -.5 and .4. The sum of the autoregressive parameters is thus between 0 and 0.9. The left panel of Figure 1 shows that the relative risk function for all three procedures are non-linear in β. The three methods have similar risk when $|\beta| = .1$. The AIC and LASSO have higher relative risks than the BIC when $|\beta| < .1$. However, the BIC pays a high price for parsimony in this parameter range. When $.1 \leq |\beta| \leq .25$, the BIC can have a higher risk than both LASSO and the AIC. The right panel shows that BIC chooses smaller models than AIC as expected. However, LASSO chooses a model that is even more parsimonious than the BIC when $\beta > .1$ and yet has lower relative risks. One explanation is that LASSO has the added flexibility to choose the lagged regressors in an unordered manner while the AIC/BIC only consider ordered sets of lags. For $T = 500$, the AIC has the highest risk when $|\beta| > .25$ because it selects the largest model. For this parameter space, the results accord with the folk wisdom that the AIC is not desirable when the true model is finite dimensional. The results (not reported) are fairly similar when the DGP includes an exogenous regressor ($y_t = .8x_t + .5y_{t-1} + \beta y_{t-2} + e_t$) or if y_t is generated by a distributed lag of x_t so that the regressors are lags of x_t instead of y_t.

While the correct model size in the first example is finite, the next two examples consider infinite dimensional models. In Example 2, y_t is a distributed lag of x_t with a moving average error. Least squares regression of y_t on x_t is not efficient in this case. An equivalent representation of y_t is an autoregressive distributed lag model of infinite order. This is approximated by a finite number of lags of y_t and x_t in the regression. Figure 2 shows that the risk functions are not symmetric around $\beta = 0$. Risk is much higher when β is positive than when it is negative. The BIC has the highest relative risk especially when β is large and positive. The right panel shows that this corresponds to situations when the BIC selects model sizes that are smallest. Interesting, larger models do not necessary translate into lower relative risks. The AIC tends to select noticeably larger models than LASSO, but LASSO tends to have slightly lower risks.

The third model considered is similar to Example 2, except that the free parameter is now the moving-average coefficient which is varied from -.5 to .5. When $\beta = 0$, the true model size is two. For all other values of β, the true model size is infinite though the empirically relevant predictor set is expected to be small. The size of the largest approximate model considered is seven. Figure 3 shows that the relative risk functions become more symmetric around zero as T increases. The BIC risks tend to increase with β. Of note is that the lack of a systematic relation between risk and model size. LASSO tends to have the lowest risk even though it does not always select the smallest model.

For example four, $N = 100$ potentially relevant predictors are generated as $x_{it} = \rho_i x_{it-1} + \epsilon_{it} + \lambda_i \mathbb{F}_t$. Each x_{it} is a stable AR(1) process with a factor structure in the errors and where $\rho_i \sim U[0, .8]$. The single factor is an AR(1) process with unit innovation variance while the idiosyncratic error ϵ_{it} is N(0,1). The variable of interest, y_t, is taken to be x_{1t} and thus $\beta = \lambda_1$. The true predictor set is the one-dimensional $\mathbb{F}_t$ but the empirically relevant predictor set is large. Two factors are formed from the principal components of one lag of X_t, ie. $X_{t-1} = (x_{1t-1}, \ldots, x_{Nt-1})'$. When $\beta = 0$, both $\widehat{F}_{1t}$ and $\widehat{F}_{2t}$ are irrelevant; when $\beta \neq 0$, $\widehat{F}_{1t}$ is relevant but $\widehat{F}_{2t}$ is not. Figure 4 shows that while diffusion index forecasts are effective when $\beta \neq 0$, relative risk can be high when $\beta = 0$ and $\widehat{F}_t$ are used as predictors. The BIC selects the most parsimonious models especially when β is small or zero, yet its risk properties are indistinguishable from LASSO.

The examples show that in finite samples, neither the BIC nor AIC dominate one another. Forecasts based on small models need not have lower risks even if the true number of predictors is finite. Pointwise arguments that favor a selection procedure may not be useful guides in practice. Large and small values of regularization parameters can both be justified depending on the optimality principle. The BIC has the lowest risk in example 4 but has the highest risk in example 2. The relative risk of the BIC is most sensitive to the true parameter value, a feature that is especially clear in model 1. In our simulations, LASSO has rather stable risk functions; it systematically dominates the AIC and often has lower relative risks than the BIC. This is true whether the variables to be selected are observed or being constructed. It could be specific to the design of the predictor sets since AIC and BIC only consider the ordered subsets but not all possible combinations of variables available as in LASSO. But this then underscores an advantage of LASSO, namely, that the predictors do not need to be ordered. Clearly, there is ample room for further investigation into these issues.

6 Conclusion

This paper has considered variable selection using information criteria, regularization, and dimension reduction from the perspective of prediction. But a predictive regression serves many purposes and its usefulness goes beyond prediction. For example, Ng and Perron (2001) show that the correct lag length need not yield a unit root test with the best size and/or power. Potscher (1991) is concerned with the adverse effects of pretesting for inference. Leeb and Potscher (2005, 2008) show that the distributions of estimators depend on the outcome of model-selection and cannot be uniformly estimated. As discussed in Hansen (2005) in the context of selecting observed predictors, what is best depends on the objective on hand. Still, practitioners need to be wary of these caveats, and this paper attemps to highlight some of these issues.

A message that is emphasized in this paper is the tension between the objective of consistent model selection and accurate prediction. This is true for large or small available predictor sets, and whether or not predictors need to be constructed. This point is transpired in the simulations presented here. The discussion has placed emphasis on the large N case (possibly larger than T) because the situation is only recently empirically relevant and problem is not as well understood.

The variable selection problem is by no means solved. While the problem is being actively studied by statisticians, there are also issues specific to economic data that need to be better understood. Case in point is generated predictors. Intuition suggests that model selection rules should be more conservative when the predictors are themselves estimated. As well, economic data are often not iid but are weakly dependent and often cross-sectionally correlated. More work is needed to understand the theory and practice of selecting constructed predictors in data rich environments.

References

Akaike, H. (1969): "Fitting Autoregressions for Predictions," *Annals of the Institute of Statistical Mathematics*, 21, 243–247.

——— (1970): "Statistical Predictor Identification," *Annals of Institute of Statistical Mathematics*, 22, 203–217.

——— (1974): "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Almoy, T. (1996): "A Simulation Study on Comparison of Prediction Methods when Only a Few Components are Relevant," *Computatonal Statistics and Data Anaylsis*, 21, 87–107.

Andrews, D. (1991): "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generaized Cross-Valdiation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

Atkinson, A. (1980): "A Note on the Generalized Information Criterion for Choice of a Model," *Biometrika*, 67, 413–418.

Bai, J. (2009): "Panel Data Models with Interative Fixed Effects," *Econometrica*, 77:4, 1229–1280.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70:1, 191–221.

——— (2006a): "Confidence Intervals for Diffusion Index Forecasts and Inference with Factor-Augmented Regressions," *Econometrica*, 74:4, 1133–1150.

——— (2006b): "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, forthcoming.

——— (2008a): "Boosting Diffusion Indices," *Journal of Applied Econometrics*, forthcoming.

——— (2008b): "Large Dimensional Factor Analysis," *Foundations and Trends in Econometrics*, 3:2, 89–163.

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

Belloni, A., and V. Chernozhukov (2011): "High Dimensional Sparse Econometric Models: An Introduction," *Lecture Notes in Statistics*, 203, 121–156.

——— (2012): "Least Squares After Model Selection in High Dimensional Sparse Models," *Bernoulli*, forthcoming.

Boivin, J., and S. Ng (2005): "Undertanding and Comparing Factor Based Forecasts," *International Journal of Central Banking*, 1:3, 117–152.

Brillinger, D. (1981): *Time Series: Data Analysis and Theory*. Wiley, San Francisco.

Buena, F. (2008): "Consistent Selection via the LASSO for High Dimensional Approximating Regression Models," *Institute of Mathematical Statistics Collections*, pp. 122–137.

Buhlmann, P., and B. Yu (2003): "Boosting with the L_2 Loss: Regression and Classification," *Journal of the American Statistical Assocation*, 98, 324–339.

Campos, J., N. Ericsson, and D. F. Hendry (1994): "Cointegration tests in the presence of structural breaks," *Journal of Econometrics*, forthcoming.

Cavanaugh, J. (1997): "Unifying the Derivations of the Akaike and Corrected Akaike Inforation Criteria," *Statistics and Probability Letters*, 33, 201–208.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets," *Econometrica*, 51, 1281–2304.

Connor, G., and R. Korajczyk (1993): "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, 48:4, 1263–1291.

Cook, D. (2007): "Fisher Lecture: Dimension Reduction in Regression," *Statistical Science*, 22:1, 1–26.

Cook, D., and L. Forzani (2008): "Principal Fitted Components for Dimension Reduction in Regression," *Statistical Science*, 23(4), 485–501.

Cox, D. (1968): "Notes on Some Aspects of Regression Analysis," *Journal of Royal Statsitical Society Series A*, 131, 265–279.

Dijkstra, T. (1983): "Some Comments on Maximum Likelihood and Partial Least Squares Methods," *Journal of Econometrics*, 22, 67–90.

Donoho, D., I. Johnstone, G. Kerkyacharian, and D. Picard (1995): "Wavelet Shrinkage Asymptopia ?," *Journal of the Royal Statistical Society Series B*, 57, 301–337.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004): "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

Eickmeier, S., and C. Ziegler (2008): "How Successful are Dynamic Factor Models at Forecasting Output and Inflation," *Journal of Forecasting*, 27:3, 237–265.

Fan, J., and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., and J. Lv (2010): "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148.

Fisher, R. (1924): "The Influence of Rainfall on the Yield of Wheat at Rothamsted," *Philosophy Transactions of the Royal Society Series B*, 213, 89–142.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The Generalized Dynamic Factor Model: Identification and Estimation," *Review of Economics and Statistics*, 82:4, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005): "The Generalized Dynamic Factor Model, One Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830–840.

Frank, I., and J. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35:2, 109–135.

Friedman, J. (2001): "Greedy Function Approxmiation: a Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232.

Fu, W. (1998): "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7:3, 397–416.

Geweke, J., and R. Meese (1981): "Estimating Regression Models of Finite but Unknown Order," *International Economic Review*, 23:1, 55–70.

Goncalves, S., and B. Perron (2011): "Bootstrapping Factor-Augmented Regression Models," mimeo, University of Montreal.

Hall, A. (1994): "Testing for a Unit Root in Time Series with Pretest Data Based Model Selection," *Journal of Business and Economics Statistics*, 12, 461–470.

Hannan, E. J., and M. Deistler (1988): *The Statistical Theory of Linear Systems*. John Wiley, New York.

Hansen, B. (2005): "Challenges for Econometric Model Selection," *Econometric Theory*, 21, 60–68.

——— (2008): "Least Squares Forecast Averaging," *Journal of Econometrics*, 146, 342–350.

——— (2010): "Multi-Step Forecast Model Selection," mimeo, University of Wisconsin.

Hansen, B., and J. Racine (2012): "Jacknife Model Averaging," *Journal of Econometrics*, 167, 28–46.

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

Helland, I., and T. Almoy (1994): "Comparison of Prediction Methods When Only a Few Components are Relevant," *Journal of the American Statistical Association*, 89, 583–591.

Hendry, D., and J. Doornik (2001): "Automatic Econometric Model Selection," .

Hesterberg, T., N. Choi, L. Meier, and C. Fraley (2008): "Least Angle and L_1 Penalized Regression: A Review," *Statistics Surveys*, 2, 61–92.

Hoerl, A., and R. Kennard (1970): "Ridge Regression: Biased Estiation for Nonorthogonal Problems," *Technometrics*, 8, 27–51.

Hotelling, H. (1936): "Relation Between Two Sets of Variables," *Biometrika*, 28, 312–377.

Hurvich, M., and C. Tsai (1989): "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 78, 297–307.

Ing, C., and C. Wei (2003): "On Same-realization Prediction in an Infinite-Order Autoregressive Process," *Journal of Multivariate Analysis*, 85, 130–155.

Ing, C., and S. Yu (2003): "On Estimating Conditional Mean-Squard Prediction Error in Autoregressive Models," *Journal of Time Series Analysis*, 24:4, 401–422.

Kabaila, P. (2002): "On Variation Selection in Linear Regression," *Econometric Theory*, 18, 913–925.

Kiers, H., and A. Smilde (2007): "A Comparison of Various Methods for Multivariate Regression with Highly Collinear Variables," *Statistical Methods and Applications*, 16(2), 193–228.

Kim, H., and N. Swanson (2010): "Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence," mimeo, Rutgers University.

Kramer, N. (2007): "An Overview on the Shrinkage Properties of Partial Least Squares Regression," *Computational Statistics*, 22, 249–273.

KUNITOMO, N., AND T. YAMAMOTO (1985): "Properties of Predictors in Misspecified Autoregressive Time Series," *Journal of the American Statistical Association*, 80:392, 941–950.

LEE, S., AND A. KARAGRIGORIOU (2001): "An Asymptotically Optimal Selection of the Order of a Linear Process," *Sanklya, Series A*, 63, 93–106.

LEEB, H., AND B. POTSCHER (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 29–59.

——— (2008): "Can One Estimate the Unconditional Distribution of Post-Model-Secltion Estimators," *Econometric Theory*, 24:2, 338–376.

LENG, C., Y. LIN, AND G. WAHBA (2006): "A Note on the Lasso and Related Procedures in Model Selection," *Statistical Sinica*, 16, 1273–1284.

LI, B. (2007): "Comment: Fisher Lecture: Dimension Reduction in Regression," *Statistical Science*, 22:1, 32–35.

LI, K. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Valdiation and Generalized Cross-Valdiation: Discrete Index Set," *The Annals of Statistics*, 15, 985–975.

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

LINGJAERDE, O., AND N. CHRISTOPHERSEN (2000): "Shrinage Structure of Partial Least Squares," *Scandinavian Journal of Statistics*, 27, 459–473.

LUDVIGSON, S., AND S. NG (2011): "A Factor Analysis of Bond Risk Premia," in *Handbook of Empirical Economics and Finance*, ed. by D. Gilles, and A. Ullah, pp. 313–372. Chapman and Hall.

MALLOWS, C. L. (1973): "Come Comments on C_p," *Technometrics*, 15, 661–675.

MEINSHAUSEN, N., AND P. BUHLMANN (2006): "High Dimensional Graphs and Variable Selection with Lasso," *Annals of Statistics*, 34:3, 1436–1462.

NG, S., AND P. PERRON (2001): "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power," *Econometrica*, 69:6, 1519–1554.

——— (2005): "A Note on the Selection of Time Series Models," *Oxford Bulletin of Economics and Statistics*, 67:1, 115–134.

ONATSKI, A. (2011): "Factor Augmented Regressions When the Number of Factors May be Misspecified Factor Models," Cambridge University.

OSBORNE, M. A., B. PRESNELL, AND B. TURLACH (2000): "A New Appropach to Variable Selection in Least Squares Problem," *IMA Journal of Numerical Analysis*, 20:3, 389–403.

PAGAN, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221–247.

PESARAN, H., A. PICK, AND A. TIMMERMANN (2011): "Variable Selection, Estimation and Inference for Multi-Period Forecasting Problems," *Journal of Economics*, forthcoming.

PHILLIPS, P., AND W. PLOBERGER (1996): "An Asymptotic Theory for Bayesian Inference for Time Series," *Econometrica*, 64(2), 381.

Phillips, P. C. B. (1979): "The Sampling Distribution of Forecasts from a First-Order Autoregression," *Journal of Econometrics*, 9:3, 241–261.

Potscher, B. (1991): "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163–185.

Rao, C. (1964): "The Use and Interpretation of Principal Components in Applied Research," *Sankhya*, 26, 329–358.

Rissanen, J. (1986a): "Modeling the Shortest Data Description," *Automatica*, 14, 465–471.

——— (1986b): "A Predictive Least Squares Principle," *IMA Journal of Mathematics Control Information*, 3, 211–222.

Rosset, S., and J. Zhu (2007): "Piecewise Linear Regularized Solution Paths," *Annals of Statistics*, 35:3, 1012–1030.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1997): "An Asymptotic Theory for Linear Model Selection," *Statistical Sinica*, 7, 221–242.

Shibata, R. (1976): "Selection of the Order of an Autoregressive Model by Akaike's Information Criteria," *Biometriak*, 63, 117–126.

——— (1980): "Asymptotic efficient selection of the order of the model for estimating parameters of a linear process," *Annals of Statistics*, 8, 147–164.

——— (1981): "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54.

——— (1984): "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, 71, 43–49.

Spearman, C. (1904): "General Intelligence, Objectively Determined and Measured," *American Journal of Psychology*, 15, 201–293.

Speed, T., and B. Yu (1993): "Model Selection and Prediction: Normal Regression," *Annals of Institute of Statistical Mathematics*, 45:1, 35=54.

Stine, R. (2004): "Model Selection Using Information Theory and the MDL Principle," *Sociological Methods and Research*, 33:2, 230–260.

Stock, J., and M. Watson (2009): "Generalized Shrinkage Methods for Forecasting Using Many Predictors," Princeton University.

——— (2010): "Dynamic Factor Models," in *Oxford Handbook of Economic Forecasting*, Oxford. Oxford University Press.

Stock, J. H., and M. W. Watson (1998): "Diffusion Indexes," NBER Working Paper 6702.

——— (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*, 20:2, 147–162.

——— (2005): "Implications of Dynamic Factor Models for VAR analysis," NBER WP 11467.

——— (2006): "Forecasting with Many Predictors," in *Handbook of Forecasting*. North Holland.

Stone, M. (1979): "Comments on Model Selection Criteria of Akaike and Schwarz," *Journal of Royal Statistical Society Series B*, 41, 276–278.

Stone, M., and R. Brooks (1990): "Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regressions," *Journal of the Royal Statistical Society B*, 52(2), 237–269.

Sun, L., S. Ji, S. Yu, and J. Ye (2009): "On the Equivalence Between Canonical Correlation Analysis and Orthonormalized Partial Least Squares," in *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pp. 1230–1235, San Francisco, CA. Morgan Kaufmann Publishers INC.

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

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statistical Society Series B*, 58:1, 267–288.

Timmermann, A. (2006): "Forecast Combinations," in *Handbook of Forecasting*, vol. 1, pp. 135–196, Amsterdam. Elseiver.

Tipping, M., and C. Bishop (1999): "Probabilistic Principal Component Analysis," *Journal of Royal Statistical Society Series B*, 61:3, 611–622.

Webster, J., R. Grant, and R. Mason (1974): "Latent Root Regression Analysis," *Technometrics*, 16, 513–532.

Wei, C. (1992): "On Predictive Least Squares Principle," *Annals of Statistics*, 20:1, 1–42.

Wold, H. (1969): "Nonlinear Estimation by Iteraltive Least Squares," in *Festschrift for J. Neymann*, pp. 411–444, New York. Wiley.

Yang, Y. (2005): "Can the Strengths of AIC and BIC be Shared? A Conflict Between Model Identification and Regression Estimation," *Biometrika*, 92, 937–950.

——— (2007): "Prediction/Estimation with Simple Linear Models: Is it Really That Simple?," *Econometric Theory*, 23, 1–36.

Young, F., Y. Takane, and J. de Leeuw (1978): "Principal Components of Mixed Measurement Level Multivariate Data: An Alternating Least Squares Method with Optimal Scaling Features," *Psychometrika*, 43, 279–281.

Zang, P., and B. Yu (2006): "On Model Selection Consistency of Lasso," *Journal of Machine Learning*, 7, 2541–2563.

Zou, H. (2006): "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Figure 1

$$y_t = .5\, y_{t-1} + \beta\, y_{t-2} + e_t$$

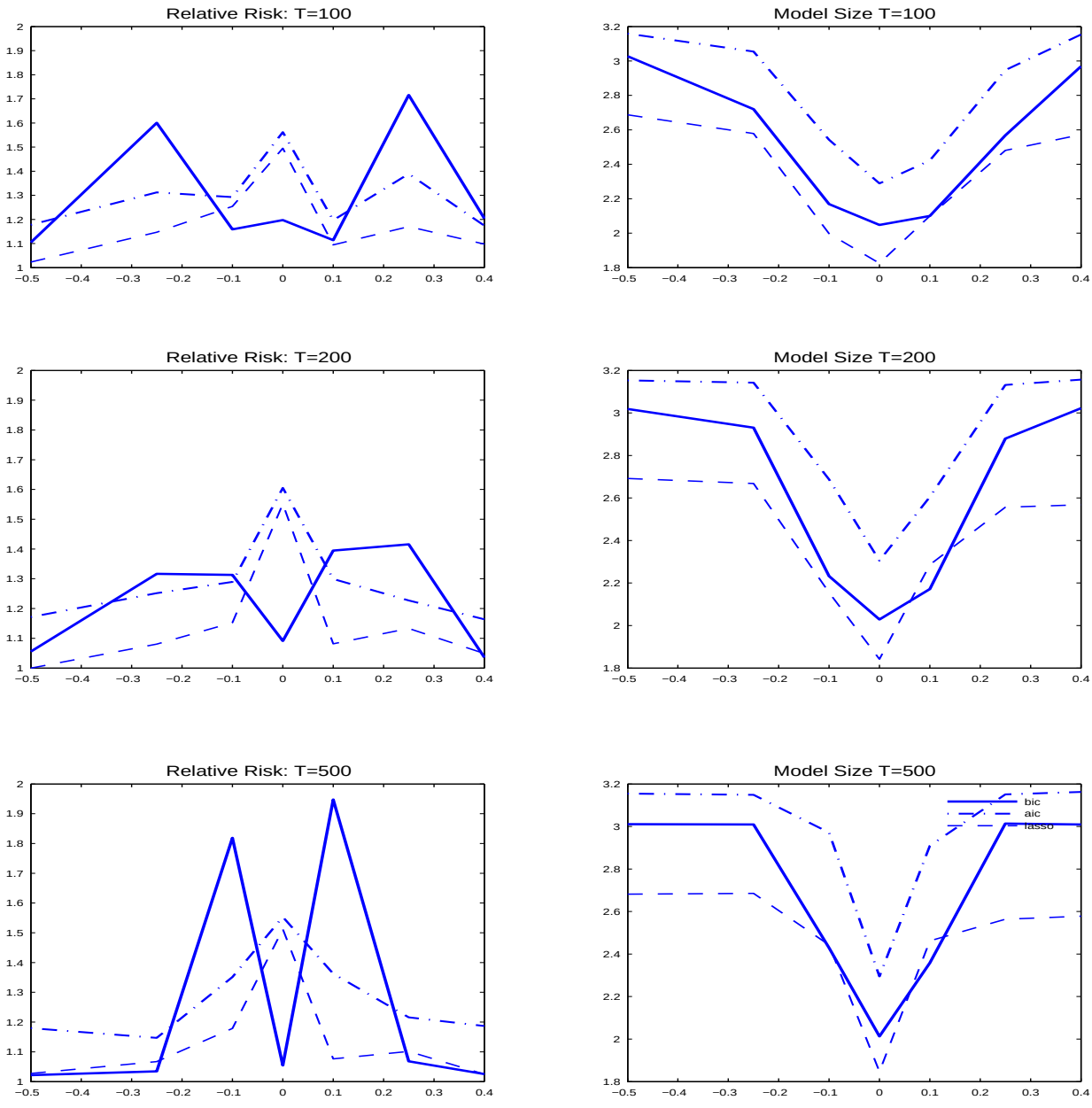
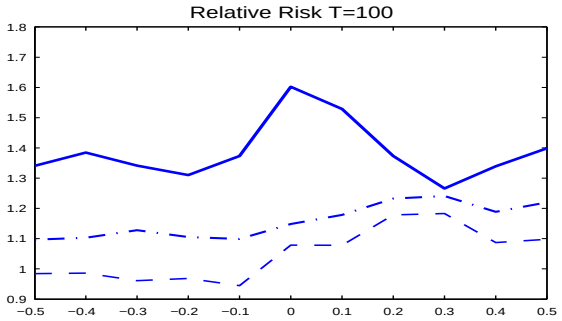
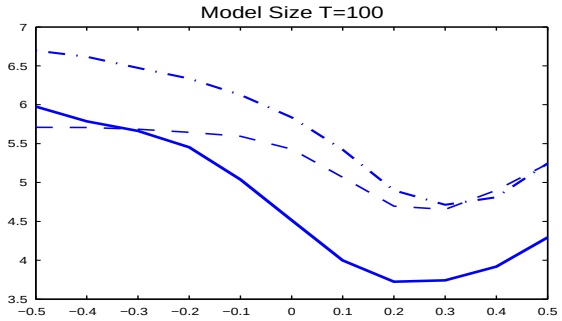
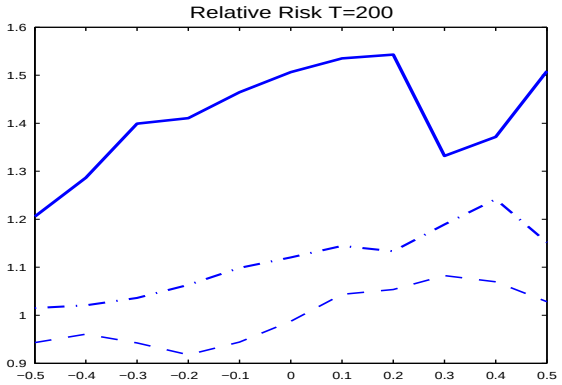


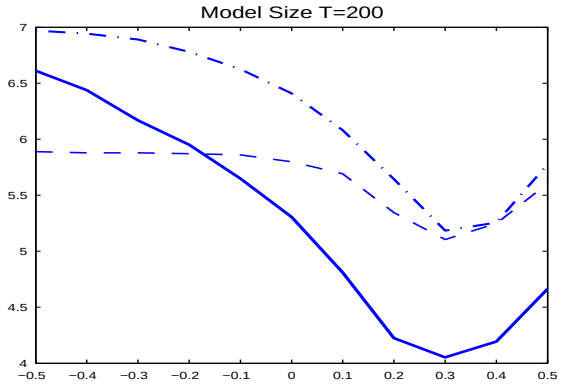
Figure 2

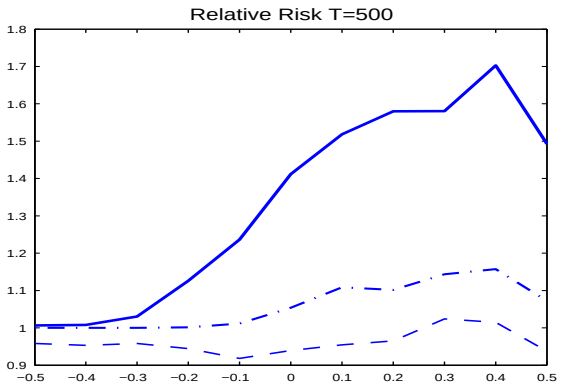
$$y_t = .8\, x_t + \beta\, x_{t-1} + e_t + .5\, e_{t-1}$$











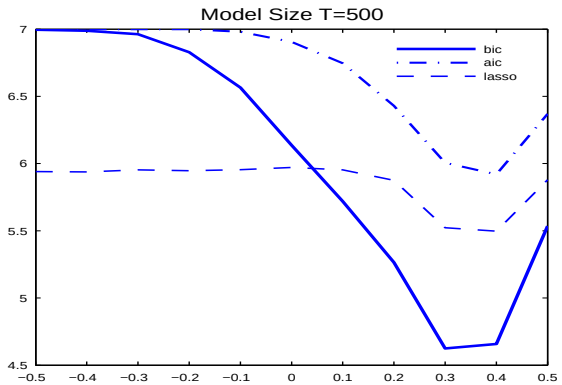
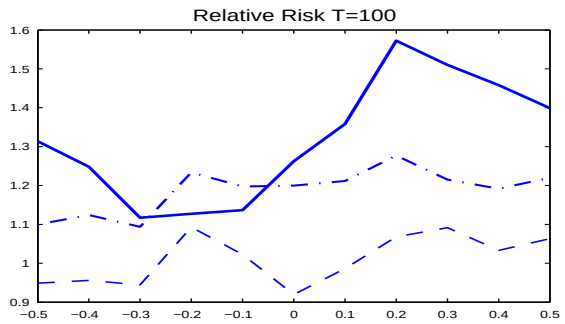
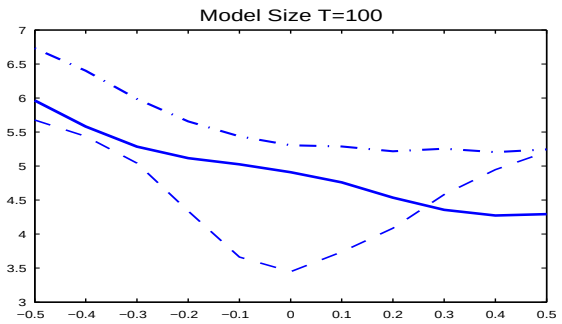
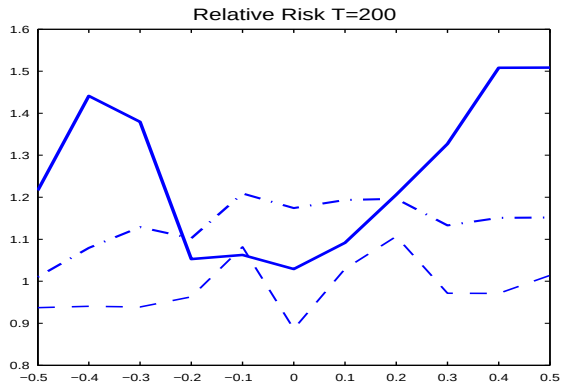


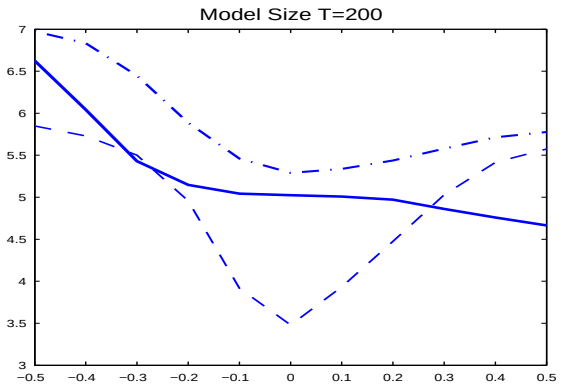
Figure 3

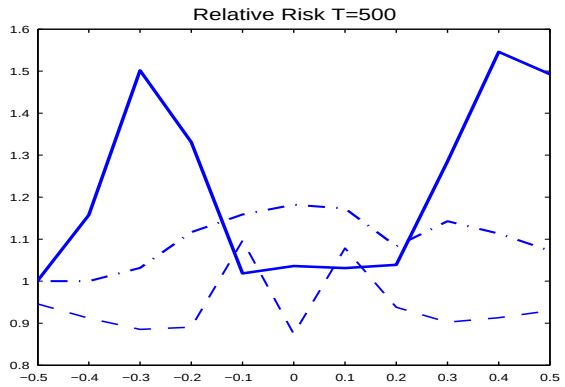
$$y_t = .8\, x_t + .5\, x_{t-1} + e_t + \beta\, e_{t-1}$$











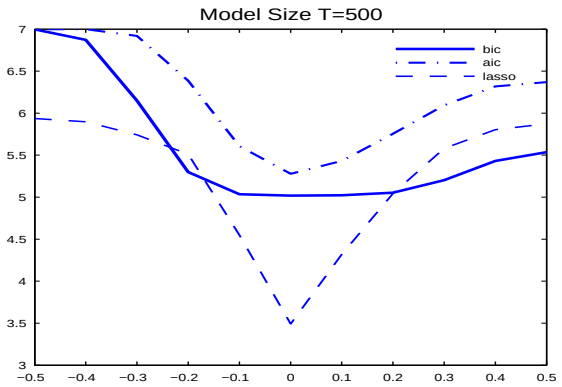


Figure 4

$$y_t = .5\ y_{t-1} + e_t + \beta F_t$$