Forecasting with Large Datasets

Yoel Avraham Furman

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2014

Acknowledgements

I would like to thank my advisor, Kevin Sheppard, for his help and patience throughout the last three years. I am also grateful to Sophocles Mavroeidis, Neil Shephard, Bent Nielsen, and three anonymous referees for their insightful comments and suggestions. I would also like to thank the Department of Economics and the Oxford-Man Institute of Quantitative Finance for their financial support and elucidating seminars. The computations in the thesis were performed in MATLAB and R, with considerable use of Kevin Sheppard's MFE Toolbox, Gary Koop's materials from his Bayesian Econometrics course, and the glmnet library in R. Lastly, I am thankful for my family, Katie Rudd, and Mowgli, for having faith in me and supporting me throughout my research.

Abstract

Forecasting with Large Datasets

Yoel Avraham Furman

Oriel College

Submitted for the degree of Doctor of Philosophy

Michaelmas Term, 2014

This thesis analyzes estimation methods and testing procedures for handling large data series. The first chapter introduces the use of the adaptive elastic net, and the penalized regression methods nested within it, for estimating sparse vector autoregressions. That chapter shows that under suitable conditions on the data generating process this estimation method satisfies an oracle property. Furthermore, it is shown that the bootstrap can be used to accurately conduct inference on the estimated parameters. These properties are used to show that structural VAR analysis can also be validly conducted, allowing for accurate measures of policy response. The strength of these estimation methods is demonstrated in a numerical study and on U.S. macroeconomic data. The second chapter continues in a similar vein, using the elastic net to estimate sparse vector autoregressions of realized variances to construct volatility forecasts. It is shown that the use of volatility spillovers estimated by the elastic net delivers substantial improvements in forecast ability, and can be used to indicate systemic risk among a group of assets. The model is estimated on realized variances of equities of U.S. financial institutions, where it is shown that the estimated parameters translate into two novel indicators of systemic risk. The third chapter discusses the use of the bootstrap as an alternative to asymptotic Wald-type tests. It is shown that the bootstrap is particularly useful in situations with many restrictions, such as tests of equal conditional predictive ability that make use of many orthogonal variables, or 'test functions'. The testing procedure is analyzed in a Monte Carlo study and is used to test the relevance of real variables in forecasting U.S. inflation.

Contents

Introduction

Recent years have seen a surge in the availability of economic and financial data. This growth in data is evident in both the time-series and cross-sectional dimensions. In the time-series dimension, tick-by-tick data has allowed researchers to analyze financial markets at a far more granular level than before, allowing for careful economic analyses of market microstructure (Hasbrouck, 2013) as well as considerably more accurate estimation of return volatility (Andersen et al., 2003). In the cross-sectional dimension, the addition of previously neglected data series has allowed for improvements in economic modeling and forecasting. A case in point is the use of large datasets by central banks to model the evolution of financial (Kogan and Tian, 2012) and macroeconomic variables (Negro and Otrok, 2008).

The availability of these new data has brought with it a slew of statistical models and testing procedures to handle them. In the time-series dimension, noisy tick-by-tick returns data are used in realized variance estimators of intra-day volatility, which account for the intra-day noise and consistently estimate quadratic variation of general price paths. In the cross-sectional dimension, curse of dimensionality concerns arising out of large cross-sections can be alleviated by penalized regression estimators, which significantly shrink estimation variance at little cost to estimation bias. Along with these large-dataset estimation procedures there has also been an interest in conducting multiple tests of many hypotheses, such as the stepwise testing algorithms in Romano and Wolf (2005) and impulse indicator saturation in Hendry et al. (2006). In the following three papers we explore some of these methods with application to macroeconomic data and U.S. equity volatility.

The first paper discusses the use of the adaptive elastic net to estimate large and highly parameterized reduced-form vector autoregressions (VARs). The adaptive elastic net is a penalized regression estimator which shrinks parameters toward zero to reduce the mean-square-error of parameter estimates. Like the LASSO (Tibshirani, 1996), the adaptive elastic net also performs selection, so that parameter estimates associated with weak predictors are excluded from the model altogether. This estimator additionally borrows from the finite-sample properties of the ridge estimator and keeps grouping effects between correlated regressors, which are frequently encountered in macroeconomic VARs. In this paper, we show that these features of the adaptive elastic net are useful in

estimating medium and large sized VARs. Further, we show that if the true stochastic process driving the data is a sparse VAR where some parameters are exactly zero, then the adaptive elastic net detects the sparsity pattern perfectly as the sample size diverges and the asymptotic distribution of the resulting parameter estimates is the same as the one estimated by OLS on the correct model (the so-called 'oracle property'). Moreover, we also show that structural-VAR analysis can be performed consistently, and impulse-response confidence intervals can be consistently estimated with the bootstrap. The strength of this estimator is demonstrated in a Monte Carlo experiment which mimics typical monetary VARs. The estimator is further compared with competing methods on U.S. macroeconomic data, where it is shown that the estimator forecasts accurately and displays expected dynamic behavior in the form of impulse-responses to a monetary policy shock.

The second paper introduces a similar sparse VAR model to estimate volatility spillovers between a wide pool of financial assets which share common features. The model is a generalization of the Heterogeneous Autoregressive (HAR) model of Corsi (2009), which takes high-frequency realized variance estimators as inputs to a restricted autoregressive forecasting equation that mimics the long-memory features of financial volatility. This generalization allows for spillovers from other assets, which can be informative in forecasting future, inter-day volatility. The model uses the elastic net penalty to focus only on the spillovers that matter most in forecasting, while maintaining all common, factor-like features of the volatility series. This method is applied to a wide group of stocks of U.S. financial institutions, which are likely to spill over volatility to one another due to the systemic nature of the sector. The paper demonstrates that these spillovers (as captured in this model) are significant in out-of-sample predictive performance and enjoy particular success over the 2008-2010 financial crisis years, when systemic risk was an important concern. This motivates us to analyze the Granger causal parameters estimated by the model, and to visualize the directed spillover network implied by these parameters. We see that the resulting volatility spillover network is highly responsive to systemic crisis and buildup, and can be used as an indicator of systemic risk.

The third and final paper shifts the focus to hypothesis testing using large datasets. The paper describes a bootstrap procedure for testing many simple null hypotheses, focusing on the canonical example of the Giacomini and White (2006) test for equal conditional predictive ability with many test functions. We show that the Giacomini and White (2006) test suffers from weaknesses in size and power when the number of tested parameters is large. This problem can arise if the investigator wishes to test the null using many test functions, and thus achieve a more informative result. Moreover, though the test (and other similar Wald-type tests) provides a simple rejection rule, the test does not indicate which set of parameters violates the null in the event of rejection. Our bootstrap alternatives are shown to be much more robust to the number of tested

parameters, and are shown to have good size and power uniformly over the number of tested parameters. Moreover, these bootstrap procedures also give an indication for which parameter restrictions violate the null. We apply these tests to test the predictive ability of a standard Phillips curve (multivariate) forecast relative to an unobserved component stochastic volatility model (univariate) forecast, and conclude that information from real variables does not add significant predictive power in forecasting U.S. inflation.

Chapter 1

VAR Estimation with the Adaptive Elastic Net

Abstract. We propose the adaptive elastic net for estimating reduced-form Vector Autoregressions. Unlike competing methods, this estimator preserves the standard structural-VAR toolkit but at the same time leads to accurate forecasts. We show that this estimator satisfies an oracle property asymptotically, and enjoys a finite-sample grouping effect. We also demonstrate the asymptotic validity of the bootstrap in constructing unconditional-on-model-selection confidence intervals. Bootstrap inference leads to valid impulse response function inference, so that structural VAR and policy analysis is asymptotically justified. The estimator is evaluated in a simulation study and on U.S. quarterly macroeconomic data.
Keywords: Vector Autoregression, Elastic Net, LASSO, Oracle, Bootstrap
JEL: C53, C32, C51, E32, E37

We propose a penalized regression estimation strategy for empirical macroeconomic modeling that generates dependable forecasts and holds structural meaning for causal inference and policy analysis. In the past, the literature was largely unable to reconcile these desires in one modeling framework. The empirical macro literature has generally diverged into two categories: small, simple models that sacrifice size and forecast ability for fairly clear structural interpretations, and large-scale models that rarely have clear structural meaning but produce more accurate forecasts. In particular, Vector Autoregressions (VARs) popularized by Sims (1980) and estimated using OLS fall in the 'small but interpretable' category, while the class of Dynamic Factor Models (DFMs) popularized by Stock and Watson (2002) allows for the use of many variables and tends to outperform VARs in terms of forecast accuracy[1]. In this paper, we propose an alternative VAR estimation strategy which allows for a many-variable VAR structure without sacrificing forecast ability. Like the DFMs, our proposed penalized VAR method exploits wide cross-sections to produce more accurate forecasts. This estimation procedure still results

[1]There are recent exceptions to both of these cases. Large VARs have been estimated by, for example Banbura et al. (2010). Similarly, there have been recent results on estimating structural DFMs in Forni et al. (2009), Forni and Gambetti (2010), and Barigozzi et al. (2014). There is also important work on marrying the approaches in Bernanke et al. (2005).

in a Granger-causal, easily interpretable model, however, in which the standard classical structural VAR analysis can be conducted for policy evaluation. As an additional feature, this method also leads to a sparse VAR, so that Granger-non-causality can be inferred directly in estimation. We show that the standard VAR toolkit is still applicable with adaptive elastic net penalized VARs at no cost to forecast ability.

A closely related, alternative approach is the use of Bayesian methods in VARs (Litterman, 1980). Bayesian methods preserve the structure of the classical VAR and perform well in forecasting exercises. The Bayesian component of these models shrinks VAR parameters to some prior belief about these parameters. The use of Bayesian priors in VARs partially lifts the curse of dimensionality by using the information in the prior as a restriction on the model. Lifting the curse of dimensionality allows the Bayesian approach to capture the best of both worlds by allowing for many variables and forecasting accurately, while maintaining the conveniently interpretable VAR form. In particular, Giannone et al. (2012) and Banbura et al. (2010) suggested a large hierarchical Bayesian VAR with many variables that appears to perform as well as DFMs in forecasting. Our proposed alternative to these Bayesian VAR methods shares their strengths but also improve upon some of their potential drawbacks.

The commonly used Bayesian VARs suffer from possible weaknesses stemming from their priors and from a lack of sparsity. Prior beliefs are in part manifestations of posterior events; that is, the priors are frequently chosen after knowing that certain events have occurred (for example, random-walk type behavior). More generally, the prior depends on user preferences, and it is possible that essentially 'what you get is what you put in': that is, Bayesian VARs only succeed when the prior fits very closely with actual posterior outcomes. This point was argued forcefully by Phillips (1990), who suggested that commonly imposed and seemingly diffuse priors are in fact strongly biased toward favorable outcomes. A second drawback of Bayesian procedures is that inference depends on the tightness of prior beliefs. If the prior is tight, posterior confidence intervals will also be tight, generating a potential bias in inference. Lastly, though Bayesian procedures shrink parameters, the more popular Bayesian procedures generally [2] do not *select* them. That is, all variables are included in the VAR, even when the VAR may have a sparse structure, where many of the true VAR coefficients are exactly zero. This sparsity in the VAR corresponds to the belief that some variables may not Granger cause other variables. Neither the OLS-estimated VAR model nor its more popular Bayesian counterparts address the sparsity concern.

Here, we suggest the use of penalized likelihood models in estimating VARs, and in particular the adaptive elastic net (Zou and Zhang, 2009) for sparse VAR estimation. Like competing Bayesian methods, the adaptive elastic net shrinks coefficients in a VAR model

[2] An important exception is the use of the double-exponential prior, discussed further below. These priors, to the best of our knowledge, have not been used in VAR analysis to date.

in order to improve forecast accuracy. In fact, the adaptive elastic net can be viewed as a Bayesian procedure with a modified double-exponential prior, which does not share the imposed unit-root behavior of traditional Bayesian VARs, and it need not be estimated with a computationally intensive MCMC procedure. This is done in a frequentist way by imposing a penalty parameter on the least-squares problem that leads to potentially large reductions in mean-square error. This reduction comes about by introducing a small bias in the VAR coefficients (coming from the penalty) while dramatically decreasing their variance. Unlike the more popular Bayesian methods, however, this approach admits a sparse representation by excluding variables with little predictive power from the model. Moreover, the selection and shrinkage is conducted adaptively in an entirely data-driven way, so that less relevant variables tend to be shrunk more or excluded altogether. This property allows us to demonstrate that the adaptive elastic-net achieves an oracle property: that is, as the sample size diverges, the adaptive elastic-net finds the correct sparsity pattern and estimates the non-zero VAR coefficients as efficiently as an agent, an 'oracle', who uses OLS having already known the correct sparsity pattern initially. This asymptotic oracle property is complemented by a finite sample grouping effect inherited from the elastic net. The grouping effect shrinks highly correlated regressors toward one another, which limits arbitrary selection issues common with similar penalized likelihood estimators. Unlike many Bayesian methods which require MCMC simulation methods, algorithms exist to solve the adaptive elastic net with the same computational complexity as the least-squares problem. These properties make the adaptive elastic net a highly attractive estimation strategy for VARs that are limited by the curse of dimensionality.

In addition to these features of estimation, we also suggest bootstrap procedures for conducting inference and impulse response analysis on the VAR parameters. Chatterjee and Lahiri (2011) showed that the residual bootstrap is consistent in estimating the sampling distribution of the adaptive LASSO in the i.i.d., fixed regressor setting. We extend their results to the case of VARs and generalize them to the adaptive elastic net. This provides an asymptotic justification for the user to conduct standard residual bootstrap inference on the estimated coefficients. Moreover, this bootstrap method can also be used in generating impulse response function confidence intervals that are important for macroeconomic policy analysis. Importantly, we emphasize that these confidence intervals are not conditional on the correct model being selected *ex ante* and explicitly take into account the error in model selection. Hence, even a non-degenerate sampling distribution of coefficients that are incorrectly estimated to be zero can be estimated using the bootstrap.

We evaluate adaptive elastic net estimation of VARs and its inference in numerical simulations, and we check its ability to forecast against a standard DFM and a BVAR over an actual data set. In the numerical study, we investigate the root-mean-square-forecast-error of the procedure and the ability of the procedure to generate accurate

impulse responses and their associated bootstrap confidence intervals. We show that the adaptive elastic net performs relatively well in these respects. Moreover, we also empirically compare the adaptive elastic net to its main competitors over a dataset of quarterly U.S. macroeconomic indicators from Stock and Watson (2012). We show that our procedure performs strongly in terms of out-of-sample predictive ability. Moreover, we also show that the impulse response functions generated by a large VAR estimated by the adaptive elastic net displays the commonly theorized behaviour expected from a large, closed-economy such as the United States. That is, even in a large VAR estimated by the adaptive elastic net, structural interpretations and commonly held expectations are preserved.

The rest of the paper explores the adaptive elastic net and its use in VAR models. Section 3.1 briefly reviews the VAR model, casts it in the regression framework, and introduces the adaptive elastic net. Section 3.2 is a formal discussion of the statistical theory underlying the adaptive elastic net. This is complemented by a numerical study in section 3.3, which evaluates the theoretical findings in the previous section. Section 3.4 compares the model to existing forecasting procedures on a U.S. quarterly dataset of macroeconomic variables used in Stock and Watson (2012) and in Giannone et al. (2012). Section 3.5 concludes and the appendix contains proofs and additional figures.

1.1 Adaptive Elastic Net Estimation of VARs

In the following we seek to estimate stationary VAR processes. The standard VAR model with possible additional stationary regressors (e.g. an intercept) can be written as:

$$y_t = B_0 h_t + \sum_{i=1}^{p} B_i y_{t-i} + e_t \tag{1.1}$$

where y_t is a stationary $N \times 1$ vector process, h_t is a stationary $M \times 1$ vector process, and e_t is i.i.d. with mean 0 and covariance matrix Σ. The matrices B_i, $1 \leq i \leq p$ are $N \times N$ parameter matrices which govern the relationship between current values of y_t with past values of y_t, and B_0 is a parameter vector governing the relationship between the additional regressors h_t and y_t. A typical parameter in B_i determines the marginal effect of the ith lag of the jth variable on the current value of the kth variable, and thus has a clear Granger non-causal interpretation: if the parameter is exactly zero, the ith lag of the jth variable has no marginal effect on the expectation of the kth variable. The VAR model thus holds a distinct advantage over factor models such as DFMs, where similar marginal interpretation is considerably more difficult and potentially impossible to determine.

The standard method for estimating the B_i parameters in the VAR is via linear regression of y_t on its lagged values and on the additional regressors. In particular,

let $z_t = (h_t, y'_{t-1}, ...y'_{t-p})'$ contain the vector of p-lagged values of all N variables and the stationary regressors. Then we can collect these in a $T \times (NP+M)$ matrix of explanatory variables $Z = (z_T, ..., z_1)'$. Writing $y_i = (y_{i,T}, ..., y_{i,1})'$ as the $T \times 1$ vector of dependent outcomes for variable y_i, $y = (y'_1, ..., y'_N)'$ and its corresponding error term e written analogously, the VAR model (1.1) can be written in multiple regression form as:

$$y = (I_N \otimes Z)\beta + e = X\beta + e \tag{1.2}$$

where β contains the $N(Np + M)$ parameters and e the residuals in the VAR model.

To estimate the full VAR model we can re-write it as a linear model (1.2). We let K denote the number of explanatory variables in the matrix X. It may be known a priori that the first K_1 of these variables (e.g. an intercept) must be included in the post-estimation model, and parameters corresponding to the remaining K_2 variables are sparse, so that some of the K_2 variables will have no true marginal effect on y post-estimation. That is, the regression model (1.2) can be written as:

$$y = X_1\beta_1 + X_2\beta_2 + e \tag{1.3}$$

where β_1 is a non-sparse matrix corresponding to the K_1 variables which are 'forced' into the model and β_2 is a sparse matrix corresponding to the K_2 variables of which a subset may be excluded. To estimate this model allowing for β_2 to be sparse, we first consider the naive adaptive elastic net:

$$\hat{\beta}^{Nnet} = \text{argmin}_\beta\{||y - X\beta||_2^2 + \lambda_2 \sum_{k=1}^{K} \beta_k^2 + \lambda_1 \sum_{k=K_1+1}^{K} \omega_k|\beta_k|\} \tag{1.4}$$

The naive adaptive elastic net is the solution to the minimization of a modified regression problem. Specifically, it minimizes the sum of squared residuals subject to a convex penalty on the magnitude of β. The penalty consists of two parts: an L_2 penalty on the all the parameters β scaled by a multiplier λ_2, and an L_1 penalty on the (possibly sparse) parameters β_2 by a multiplier λ_1 and adaptive weights ω_k. That is, we can think of the naive adaptive elastic net as a constrained minimization problem subject to constraints on the L_1 norm and on the L_2 norm, where the amount of assigned β is limited by the constraint. Variables in X that have little explanatory power, such as variables that do not Granger-cause other variables in the VAR, will have very small (if not zero) coefficients, since the cost of the constraint limits the gain in minimizing the sum of squared residuals. The constraint shrinks the magnitude of β toward zero, thus reducing the prediction error caused by poorly estimated, small β values.

To examine the shrinkage property more closely, consider the special case where $\lambda_1 = 0$ and the naive adaptive elastic net reduces to ridge regression. Ridge regression is similar to OLS except for a penalty on the squared β coefficients. This penalty increases the

bias over that of the OLS estimates at the benefit of decreased variance, and thus leads to potential reduction in mean-square error. Reduced mean-square error in the estimator can hence lead to a reduction in mean-square forecast error in a VAR estimated with this approach.[3] Moreover, as noted in Zou and Hastie (2005), ridge shrinkage dampens correlations between the variables. This decorrelation step encourages a grouping effect, so that coefficients of correlated variables are clustered together. In the macro VAR context, this means that coefficients for highly correlated variables (like different measures of employment) will be close.

To examine the selection property more closely, consider the converse case when $\lambda_2 = 0$ and $\omega_k = 1$ and the naive adaptive elastic-net reduces to the LASSO (Tibshirani, 1996). The LASSO (Least Absolute Shrinkage and Selection Operator) imposes an L_1 penalty on the β_2 coefficients, and like ridge, the penalty increases the bias over that of the OLS estimates at the benefit of decreased variance. While ridge shrinks the coefficients in proportion to λ_2, LASSO translates the coefficients with a truncation at zero. Thus, if a particular variable has poor explanatory power for the dependent variable relative to λ_1, the LASSO solution will set its coefficient to exactly zero, so that the solution for the entire vector of coefficients will be sparse. When a variable predicts the dependent variable strongly relative to λ_1, the LASSO translation shrinks the coefficient away from the OLS solution[4]. Thus, the LASSO shrinks and selects the regression coefficients. When λ_2 is non-zero, the LASSO penalty is combined with the ridge penalty to form the naive elastic net penalty (Zou and Hastie, 2005), which allows for the truncation properties of the LASSO as well as the grouping effect adopted from ridge regression. In the context of VARs where many of the explanatory variables are lagged dependent variables, LASSO and the naive elastic net shrink some of the coefficients of the lagged dependent variables, and exclude other lagged dependent variables altogether. Thus, these L_1 penalizations impose Granger non-causality in the estimation step of the model.

In many applications, a desirable property in a model selection procedure is the ability to identify the correct model asymptotically. An asymptotically consistent model selection and estimation procedure can be characterized by an *oracle property*, which asserts that as the number of observations diverges: (i) the correct subset of coefficients is identified with probability 1, and (ii) the estimated coefficients from the procedure are estimated at the same consistency rate as if the population (sparse) model was estimated using OLS. Zou (2006) and Zhao and Yu (2006) showed that in the i.i.d. setting, the LASSO does not satisfy the oracle property. Instead, Zou (2006) suggested a modification, the adaptive LASSO, which penalizes each coefficient k by a different magnitude,

[3]The ridge solution can also be formulated in the Bayesian VAR context when the coefficients in the VAR have a Gaussian prior centred at zero. Ridge is therefore a simplification of the Minnesota prior for VARs expressed in differences.

[4]As formally discussed in Knight and Fu (2000) and Zou (2006), the bias introduced by this penalization may be non-negligible, even asymptotically.

$\lambda_1\omega_k$. The adaptive LASSO is given by the naive adaptive elastic net when $\lambda_2 = 0$. The adaptive weights are given by $\hat{\omega}_k = |\beta_{I,k}|^{-\gamma}$, where $\beta_{I,k}$ is an initial estimator[5] and γ is an additional tuning parameter. Variables that are excluded from the initial model carry an infinite penalty in the second model, while those that have large initial coefficients carry almost no penalty in the second model. These adaptive weights allow for the correct selection of variables by severely punishing poor explanatory variables. Simultaneously, important explanatory variables are hardly penalized, so that they are estimated almost as if they were estimated by OLS. This 'biased' penalty allows the adaptive LASSO to asymptotically select the true model with oracle efficiency. In the VAR context, this oracle property ensures that asymptotically, the correct DGP is selected so long as it is nested in the estimated VAR. Moreover, the oracle property also ensures that the coefficients of the estimated VAR are estimated at the same rate of efficiency as those estimated by an oracle using OLS on the correct, nested model.

The naive adaptive elastic net penalty is a linear combination of the ridge and adaptive LASSO penalties. The ridge part of the penalty shrinks the estimated coefficients of all the variables X and induces coefficients of correlated variables to be close to one another. The adaptive LASSO part of the penalty shrinks and selects the coefficients of the variables in X_2 with oracle efficiency. The de-correlation step in ridge encourages a grouping effect between the explanatory variables, so that in a group of highly correlated regressors it is unlikely that only a strict subset of them is dropped from the model after the adaptive LASSO penalization. Note that the ridge part of the penalty induces additional unwanted shrinkage beyond that of the adaptive LASSO. To mitigate this additional shrinkage, Zou and Zhang (2009) scale the minimization problem and define the adaptive elastic net[6]:

$$\hat{\beta}^{Anet} = \Big(1 + \frac{\lambda_2}{NT}\Big)\text{argmin}_\beta\{||y - X\beta||_2^2 + \lambda_2 \sum_{k=1}^{K} \beta_k^2 + \lambda_1 \sum_{k=K_1+1}^{K} \omega_k|\beta_k|\} \tag{1.5}$$

The adaptive elastic net benefits from the selection and asymptotic oracle property of the adaptive LASSO, as well as from the finite-sample grouping effect inherited from the ridge penalty. This makes the adaptive elastic net particularly useful for estimating larger VAR models containing several correlated variables. In these VAR models where many of the variables move in tandem, this estimation procedure leaves out irrelevant variables but does not exclude correlated variables that may be relevant as part of a group. For example, when a VAR model is used to estimate the dynamics in a macroeconomic system, the system is likely to have many variables that have no Granger-

[5]In practice, this is frequently the elastic net estimator, since it can handle more variables than observations effectively and is consistent at the root-T rate.

[6]The naive elastic net is adjusted by a multiplicative constant to give the elastic net.

causal effect on other variables. Simultaneously, the system may have several highly-correlated measures of inflation, each transmitting a similar signal. The adaptive elastic net guarantees the exclusion of the correct irrelevant variables asymptotically, and in finite samples it will give similar coefficients to highly correlated variables, such as different measures of inflation. In the numerical study and the empirical example discussed below we will show that this leads to strong predictive performance.

To derive the properties of the adaptive elastic net in the VAR context, we use the representation (1.5). In order to estimate the model in practice, however, we use a re-written version of the adaptive elastic net estimator which allows us to control the tradeoff between ridge and the adaptive LASSO, and hence the degree of grouping. Specifically, (1.5) can be re-written as:

$$\hat{\beta}^{Anet} = \Big(1 + \frac{\lambda\alpha}{NT}\Big)\text{argmin}_{\beta}\{||y - X\beta||_2^2 + \lambda[\alpha\sum_{k=1}^{K}\beta_k^2 + (1-\alpha)\sum_{k=K_1+1}^{K}\omega_k|\beta_k|]\} \quad (1.6)$$

In this estimation, α represents the tradeoff between the ridge and the adaptive LASSO penalty, where $\alpha = 1$ represents ridge and $\alpha = 0$ represents the adaptive LASSO. The total magnitude of the penalty is given by λ, which we can use to gauge the level of penalization in the model. In the applications below, we choose these tuning parameters by minimizing cross-validated mean-square-error in order to minimize prediction error.

1.2 Properties of the Adaptive Elastic Net

In this section, we discuss several statistical properties of adaptive elastic net estimators (1.5) for VAR regressor coefficients in the model (1.1). Let $\mathcal{A}^o$ denote the active set of true, correctly included regressors in the population model, whose parameters are not indexed, and simply given by β^o. This parameter vector includes β_1^o, the (correctly) included population parameters associated with the set of 'forced' regressors. Adaptive elastic net estimates are given by $\hat{\beta}$, the set of included regressors they imply is given by $\hat{\mathcal{A}}$, and their associated bootstrap estimates are given by $\hat{\beta}^*$. We first discuss a useful finite sample property of the adaptive elastic net, the grouping effect, and then we turn our attention to the asymptotic properties of the estimator and its bootstrap distribution.

The grouping effect of the adaptive elastic net is a property conveniently inherited from the elastic net of Zou and Hastie (2005) which groups estimated coefficients of correlated variables together. To simplify the discussion, suppose that the regressors in (1.2) are centred and scaled so that they have sample mean zero and variance one, and that the dependent variable is centred to have mean zero. As noted above, the grouping effect is the property inherited from additional shrinkage from the ridge penalty which squeezes the coefficients of correlated regressors towards one another. This implies that

highly correlated variables will have coefficients of similar magnitude. Formally, we note the following result, a simple modification of Theorem 1 in Zou and Hastie (2005):

Proposition 1.2.1. *Let $\hat{\beta}(\lambda_1, \lambda_2)$ be the naive adaptive elastic net estimate. Suppose that $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$, so that the two estimates have the same sign. Define:*

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{||\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)||_2}{||y||_2}$$

Letting ρ be the sampling correlation between the i'th and the j'th column of X, then:

$$D_{\lambda_1,\lambda_2}(i,j) \leq \frac{\lambda_1||\omega_i - \omega_j||_2}{2\lambda_2||y||_2} + \frac{\sqrt{2(1-\rho)}}{\lambda_2}$$

and, if $\gamma = 1$ and the initial estimators $\beta_{I,i}, \beta_{I,j}$ are the naive elastic net estimators and also have the same sign, then:

$$D_{\lambda_1,\lambda_2}(i,j) \leq \sqrt{2(1-\rho)}\Big(\frac{1}{2\lambda_2} + \frac{\lambda_1}{\lambda_2^2||\beta_{I,i}\beta_{I,j}||_2}\Big)$$

That is, the distance between coefficients of two regressors whose coefficients are of the same sign is bounded by the distance between the initial estimators and a decreasing function of the sampling correlation between the regressors. Moreover, as the shrinkage parameter λ_2 in the model increases, the bound becomes tighter. This means that closely correlated regressors will have close coefficients. For example, if two regressors were almost identical and were highly positively correlated (e.g. first differences in quarterly non-farm payroll and household survey employment measures), then it is likely that their sampling correlation will be close to one, and the distance between their coefficients will be very small. Note that this is a finite sample, exact result, and is independent of the nature of the regression problem provided it falls into the linear model formulation. A simple adaptation of the proof in Zou and Hastie (2005) is provided in the appendix for completeness.

We now turn to the asymptotic properties of these estimators. To derive these, consider the following assumptions:

Assumption 1.2.1. *The VAR innovations hold the following properties:*

(a) $E(e_{it}^4) < \infty$.

(b) $E(e_t) = E((e_{1t}, ..., e_{Nt})') = 0$.

(c) e_t *are i.i.d. with covariance matrix* Σ.

Assumption 1.2.2. $C = E(\frac{1}{T}Z'Z)$ *is positive definite.*

Assumption 1.2.3. *Assume that* $\gamma > 0$ *and the following rates:*

(a) $\frac{\lambda_1}{\sqrt{T}} \to 0$ *and* $\frac{\lambda_2}{\sqrt{T}} \to 0$.

(b) $\frac{\lambda_1}{T^{(1-\gamma)/2}} \to \infty$.

(c) $\sqrt{T}(\hat{\beta}_I - \beta^o) \in O_p(1)$.

Assumption 1.2.1 is a standard assumption on VAR innovations that allows for the use of the law of large numbers and central limit theorem. Importantly, it places limited distributional assumptions on the errors. Assumption 1.2.2 guarantees that the population $X'X$ matrix of the regressors is invertible and positive definite, a key property for solution uniqueness of the regression equation and for a proper covariance matrix of the final estimator. Assumption 1.2.3 ensures that asymptotically the correct sparsity pattern is obtained: that is, the penalty parameters decay to zero sufficiently quickly for the relevant regressors in $\mathcal{A}^o$ to be included, but slowly enough so that irrelevant variables not in $\mathcal{A}^o$ are left out entirely. Similarly to Kock and Callot (2012), these assumptions provide the adaptive elastic net with the ability to determine the correct sparsity pattern asymptotically and to estimate the non-zero coefficients with oracle-rate efficiency.

Proposition 1.2.2. *Suppose assumptions (1.2.1-1.2.3) are true. Further, let* $\hat{\Sigma} = T^{-1}\sum_{t=1}^{T} \hat{e}_t\hat{e}_t'$. *Then* $\hat{\beta}$ *satisfies:*

(a) $\sqrt{T}$*-consistency:*

(i) $||\hat{\beta} - \beta^o||_2 = O_p(1/\sqrt{T})$

(ii) $\hat{\Sigma} \xrightarrow{p} \Sigma$

(b) The oracle property:

(i) $\sqrt{T}(\hat{\beta}_{\mathcal{A}^o} - \beta^o_{\mathcal{A}^o}) \to N(0, (I_N \otimes C)^{-1}_{\mathcal{A}^o}[\Sigma \otimes C]_{\mathcal{A}^o}(I_N \otimes C)^{-1}_{\mathcal{A}^o})$

(ii) $Pr(\hat{\mathcal{A}} = \mathcal{A}^o) \to 1$

The proposition closely mimics that for the adaptive LASSO provided by Kock and Callot (2012) and Zou (2006) and an analogous proof is provided in the appendix for completeness. The first part of the proposition asserts that the adaptive elastic net estimators converges in probability to the true parameter value at the same rate as the ordinary least squares estimator. It then follows that the covariance matrix estimator implied by the adaptive elastic net estimate also converges in probability to its true value. The oracle property proven by the proposition asserts that for the coefficients that are in $\mathcal{A}^o$, the distribution of the non-zero estimates converges in law to the asymptotic distribution of the least-squares estimator estimated by the oracle who knows the true

sparsity pattern prior to estimating the model. Finally, these properties are then used to show that the true sparsity pattern is asymptotically revealed by the adaptive elastic net, in probability.

Note that the asymptotic distributional result in (1.2.2) is non-uniform in β, and in particular, will unlikely hold for local-to-zero coefficients that approach zero at the rate of $\sqrt{T}$, an important concern studied in Leeb and Potscher (2008). In our context however, the coefficients do not vary with the sample size, and this is not an immediate concern. Additionally, works by Potscher and others suggest that in practice the distribution of oracle estimators in finite samples converges slowly and unevenly to the asymptotic distribution. In particular, the asymptotic distribution provides no standard errors or confidence intervals for those parameters estimated to be exactly zero. We note here that although this is an important observation, the asymptotic result continues to hold in the fixed population parameter case, and in finite samples we prefer to rely on bootstrap inference methods.

In many applications, including macroeconomic VARs, standard errors and confidence intervals for the coefficient estimates are needed to gauge the level of uncertainty in the model. As mentioned above, the asymptotic distribution from the oracle result in Proposition 1.2.2 yields unstable inference in finite samples. Moreover, the asymptotic distribution of parameters estimated to be zero is degenerate, providing no standard errors for parameters equal to zero. Because of this, as in Chatterjee and Lahiri (2013), we propose using the standard residual bootstrap of Freedman (1981) with the adaptive elastic net to simulate the distribution of the adaptive elastic net estimates. This procedure was originally used for the adaptive lasso by Chatterjee and Lahiri (2013), who proved the consistency of the bootstrap distribution for the adaptive lasso estimator's sampling distribution in the i.i.d, fixed regressor case. Here, we follow their work analogously and prove their results for the adaptive elastic net in the stationary VAR problem when there are no additional stationary regressors h_t. We conjecture that the stationary bootstrap of Politis and Romano (1994) could be validly used when there are additional stationary regressors, but for simplicity of exposition we consider the simpler case. Specifically, we construct the residual sequence:

$$(\hat{e}_1', ..., \hat{e}_T')' = \hat{e} = y - X\hat{\beta}$$

From the sequence $\{\hat{e}_t\}$, we can construct a sequence of centred residuals:

$$e_t^* = \hat{e}_t - \bar{e}_t = \hat{e}_t - \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t$$

With these centred residuals, we can sample with replacement vectors of the form e_t^* and

construct the bootstrap version of the VAR using $(y^*_{1-p}, ..., y^*_0) = (y_{1-p}, ..., y_0)$:

$$y^*_t = \sum_{i=1}^{p} \hat{B}_i y^*_{t-i} + e^*_t, \quad t = 1, ..., T$$

Then, using the new bootstrap sample of $\{y^*_t\}_{t=1-p}^T$ we estimate the adaptive elastic net on the bootstrapped data:

$$\hat{\beta}^* = (1 + \frac{\lambda_2}{NT})\text{argmin}_\beta\{||y^* - X^*\beta||_2^2 + \lambda_2 \sum_{k=1}^{K} \beta_k^2 + \lambda_1 \sum_{k=K_1+1}^{K} \hat{\omega}_k|\beta_k|\} \tag{1.7}$$

where y^* and X^* are constructed as before, replacing the data with their bootstrapped analogues. We focus here on $Q_T = \sqrt{T}(\hat{\beta} - \beta^o)$, the scaled and centred version of the estimator, whose sampling distribution, $G_T(\cdot)$ we seek to estimate via the bootstrap procedure above. The bootstrap analogue to Q_T is given by $Q^*_T = \sqrt{T}(\beta^* - \hat{\beta})$, and we focus on the conditional distribution of Q^*_T given the observations $y_{1-p}, ..., y_T$:

$$G^*_T(B) = Pr_*(Q^*_T \in B), \quad B \in \mathcal{B}(\mathbb{R}^K)$$

where Pr_* denotes the probability measure conditional on $y_{1-p}, ..., y_T$, and $\mathcal{B}(\mathbb{R}^K)$ denotes the Borel σ-algebra on $\mathbb{R}^K$. For $G^*_T(\cdot)$ to be a useful approximation for $G_T(\cdot)$, we want the distance between these two distributions to be small. Asymptotically, we want the distance between these distribution functions to vanish:

Proposition 1.2.3. *Suppose that assumptions (1.2.1-1.2.3) hold. Then:*

$$d_2(\hat{G}_T(\cdot), G_T(\cdot)) \longrightarrow 0 \ \textit{as } T \to \infty, \ \textit{with probability 1}$$

where $d_2(\cdot, \cdot)$ denotes the Mallow's metric metrizing weak convergence on the set of all probability measures on $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$.

The proposition asserts that asymptotically the two distribution functions converge onto one another. Furthermore, let ∂B denote the boundary of a Borel set B and let $G_\infty(\cdot)$ denote the limiting distribution of Q_T. Then from Proposition 1.2.3, it follows that for any Borel set $B \subset \mathbb{R}^K$ such that $G_\infty(\partial B) = 0$,

$$Pr_*(Q^*_T \in B) - Pr(Q_T \in B) \to 0 \ \text{ as } T \to \infty$$

This implies that the bootstrapped distribution of the adaptive elastic can be justifiably used to estimate the limiting distribution of the estimator, asymptotically. Moreover, we can construct valid estimates of confidence sets for β^o using the bootstrap distribution.

Let $q^*(\alpha)$ denote the α-quantile of the bootstrap distribution of $||\hat{Q}_T^*||_2$, $\alpha \in (0,1)$. Then the set defined by:

$$I_{T,\alpha} = \{q \in \mathbb{R}^K : ||q - \hat{\beta}||_2 \leq T^{-1/2} q^*(\alpha)\}$$

is an approximate confidence set for β^o of level α, as the following corollary shows. For that purpose, let Q_∞ denote the limiting random vector such that $Q_T \to Q_\infty$ in law; that is, Q_∞ has distribution G_∞. Further, let $q(\alpha)$ denote the α-quantile of $||Q_\infty||_2$, $\alpha \in (0,1)$.

Corollary 1.2.4. *If $\{j : \beta_j^o \neq 0\}$ is non-empty, then $Pr(\beta^o \in I_{T,\alpha}) \to \alpha$.*

The corollary asserts that the size of the bootstrapped confidence set converges to its asymptotically correct value in probability, provided that at least one of the regressors is correctly included. This implies that we can consistently use confidence sets and hypothesis tests on the bootstrap distribution of the estimated parameters in the VAR, even when some of those parameters are estimated to be zero.[7]

In addition to the estimation and bootstrap inference of the model parameters, it is also often useful for the VAR model parameters to yield accurate impulse responses with suitable confidence intervals. Consider the model (1.1). This dynamical system's impulse responses Ψ_i to the reduced form innovations e_t are given by the recursive relation $\Psi_i = [\psi_{lk,i}] = \sum_{j=1}^{i} \Psi_{i-j} B_j, i = 0, 1, 2, ...,$ where $\Psi_0 = I_N$ and $B_j = 0$ for $j > p$. That is, the impulse responses are matrix polynomials of the coefficients. These reduced form impulse responses represent *correlated* shocks e_t and their impact on y_{t+i}, rather than shocks caused by one specific variable. We are interested in causal structural form shocks rather than correlated reduced form shocks, so to that end we define P such that $PP' = \Sigma$ and P is lower triangular, and then we obtain the structural innovations $u_t = Pe_t$. Then the structural, orthogonalized impulse responses to the shocks u_t are defined by $\Theta_i = [\theta_{lk,i}] = \Psi_i P, i = 0, 1, 2, ...,$ where these shocks represent the effect of a unit increase in one of the orthogonalized innovations on the entire system i-periods ahead. Letting $b = vec(B_1, B_2, ..., B_p)$ and $\sigma = vech(\Sigma)$, where $vec(\cdot)$ and $vech(\cdot)$ denote the stacking and half-stacking operators respectively, the structural impulse responses $\theta_{lk,i}$ of a shock on variable k on variable l after i periods ahead can be viewed as a continuously differentiable but nonlinear functions of the model parameters, $\theta_{lk,i}(b, \sigma)$.

Estimates $\hat{\theta}_{lk,i}$ of the population impulse responses $\theta_{lk,i}^o$ can be obtained by plugging-in the $\hat{\beta}$ estimate and the associated $\hat{\Sigma}$ parameter into the impulse response function $\theta_{lk,i}(b, \sigma)$. Likewise, the sampling distribution of the impulse responses can be estimated by plugging-in the bootstrap parameters and the estimated covariance matrix

[7] As in corollary 1.2.4, this is accurate only if there is at least one true variable included in the linear restriction for the confidence interval or the test under consideration. In the event of the restriction being conducted only on irrelevant variables not in $\mathcal{A}^o$, the asymptotic distribution is degenerate and the bootstrap confidence interval will be wide.

into $\theta_{lk,i}(b,\sigma)$, $\theta^* = \theta^*_{lk,i}(b^*,\hat{\sigma})$. Moreover, these impulse response estimates and their associated bootstrap confidence intervals converge onto the asymptotic limit of the impulse response estimator and onto its asymptotic distribution, respectively. Suppose that the probability limit of the impulse response estimator $\hat{\theta}_{lk,i}$ is given by $\theta^o_{lk,i}$, then:

Corollary 1.2.5. *Let $\vartheta^*_{lk,i} = \sqrt{T}(\theta^*_{lk,i} - \hat{\theta}_{lk,i})$ with distribution $H^*_T(\cdot)$, and $\hat{\vartheta}_{lk,i} = \sqrt{T}(\hat{\theta}_{lk,i} - \theta^o_{lk,i})$ with distribution $H_\infty(\cdot)$, and let $h^*_{lk,i}(\alpha)$ denote the α-quantile of $H^*_T(\cdot)$. Then:*

(a) $\hat{\theta}_{lk,i} \xrightarrow{p} \theta^o_{lk,i}$.

(b) The distance $d_2(H^*_T(\cdot), H_\infty(\cdot)) \to 0$ *as* $T \to \infty$.

(c) Provided that $\theta_{lk,i}(b^o,\sigma^o)$ is a function of at least one non-zero element of b^o, then intervals of the form:

$$M_{T,\alpha} = \{w \in \mathbb{R} : (w - \hat{\theta}_{lk,i}) \leq T^{-1/2} h^*_{lk,i}(\alpha)\}$$

have asymptotically valid size, $Pr(\theta^o_{lk,i} \in M_{T,\alpha}) \to \alpha$ *as* $T \to \infty$.

This means that we can construct asymptotically valid impulse responses and confidence intervals for these impulse responses, provided that we have correctly included at least one relevant regressor in $\mathcal{A}^o$. This makes standard impulse response analysis available and asymptotically justifiable when estimating VARs with the adaptive elastic net.

In sum, we have shown that the standard structural VAR tool-kit is readily available for use with the adaptive elastic net. Moreover, the end user need not worry (asymptotically) about including potentially irrelevant variables, since the adaptive elastic net will shrink these variables toward zero or exclude them altogether. The advantages of this methodology, however, also extend to its ease of computation. Specifically, we can estimate the entire path of adaptive elastic net coefficients efficiently using at least two different algorithms: LARS (Efron et al., 2004) and coordinate descent (Friedman et al., 2009). Both of these algorithms solve the adaptive elastic net at the same computational complexity rate as solving the OLS problem.

1.3 Numerical Study

In this section we demonstrate the strength of penalized regression models in estimating reduced-form VARs[8] when the true process is a sparse, stationary VAR with behavior similar to a canonical monetary VAR. Specifically, we simulate 1000 Monte Carlo experiments of the trivariate VAR(1) process:

[8] An alternative approach is to estimate structural VARs directly. We do not pursue that here, since we assume (a) sparsity in the reduced form, and (b) as in the Bayesian literature, the user would estimate the reduced form and then conduct structural analysis.

$$y_t = B_1 y_{t-1} + \epsilon_t = B_1 y_{t-1} + P\eta_t \tag{1.8}$$

where:

$$B_1 = \begin{pmatrix} 0.8 & 0 & -0.1 \\ 0 & 0.2 & -0.1 \\ 0 & 0 & 0.1 \end{pmatrix},$$

and $\epsilon_t \sim N(0, \Sigma)$, $\eta_t \sim N(0, I_3)$, with:

$$\Sigma = \begin{pmatrix} 1 & 0 & -0.25 \\ 0 & 1 & -0.25 \\ -0.25 & -0.25 & 1 \end{pmatrix}$$

and P is the Cholesky square-root of Σ. This VAR contains dynamics mimicking those of a macroeconomic dynamical system. The matrix of lag coefficients is upper triangular with a maximum eigenvalue of 0.8, so the system is highly persistent but stationary. This reflects the high degree of persistence and a belief of stationarity in estimated trivariate monetary VARs. As in reduced-form monetary VARs, the error terms are generally correlated, since a structural shock is usually absorbed in multiple variables immediately. Since our primary object of interest is estimation rather than correct identification, we assume that we know the recursive ordering of these shocks. As for the dynamics of the individual variables:

(a) The first variable is very persistent, with slight effects arising from the third variable. The second variable does not Granger-cause the first. This variable is analogous in behavior to inflation, which is highly persistent and is believed to decrease with changes in the interest rate.

(b) The second variable is considerably less persistent, and is also Granger-caused by the third variable but not by the first. This variable is analogous to changes in employment, which is weakly persistent and is also believed to decrease with changes in the interest rate.

(c) The third variable is very mildly persistent, and is only Granger-caused by itself. This corresponds to a belief that shocks to interest rates growth decay rapidly but do not vanish immediately. Moreover, changes in the interest rate are present-and-forward looking only.

Like monetary VARs, the dynamics of the system are rich, with heightened complexity arising from correlated shocks, differing levels of persistence, and a difficult sparsity pattern to uncover.

In order to combat the differing levels of sparsity and persistence for each data series, we estimate each equation in the VAR separately. This allows for each equation to have its own degree of penalization; otherwise, the chosen level of penalization would be a pooled average between all equations, and the models would not perform as well. We employ five-fold cross-validation in choosing our penalty parameter λ from a wide grid of values in order to penalize in a predictively optimal way. To handle models with a grouping effect penalty ($\alpha \neq 0$), we also tune α over a small grid of values, $\alpha \in (0.25, 0.5, 0.75)$. This fixes the degree of grouping in the estimator and allows us to conduct comparisons without worrying about grouping effect models that are de facto adaptive LASSO or ridge models[9]. To avoid tuning on another parameter, we fix $\gamma = 1$. In each estimated model we estimate a VAR(3) to proceed as if we did not know the true lag length. Moreover, for each sample size we also estimate the full model with three lags as well as the oracle model (which only estimates the non-zero population coefficients) by least-squares equation-by-equation, so that we can compare the oracle to the adaptive elastic net directly. We recognize that there are potential efficiency gains to estimating the pooled oracle model[10] (Zellner, 1962), but we conjecture that these are negligible in the present context and we believe that the comparison to the adaptive elastic net is fair without using cross-equation information.

We investigate several evaluation criteria over varying sample sizes. For each method, we estimate the VAR with three different sample sizes: small ($T = 27$), medium ($T = 54$), and large ($T = 270$)[11]. For small systems, penalized regressions should outperform OLS considerably, but perhaps not outperform the oracle which enjoys greater degrees of freedom in addition to knowing the true sparsity pattern. For the medium sized system, the oracle might slightly outperform the penalized likelihood methods, and the oracle property might start to benefit the adaptive methods for the series with a higher signal-to-noise ratio (persistence). At the large sample size, penalized regression methods should all be close, with adaptive methods possibly slightly outperforming due to their oracle property.

Before evaluating the penalized regression methods over the criteria outlined below, we also notice some insightful patterns when looking at the selected cross-validated parameters α and λ. Table 1.1 shows the average penalty parameters selected by cross-validation for each time-series. The left panel shows that the λ parameter is considerably different for different series and varying sample sizes. This suggests that our decision to penalize each equation separately is reasonable from a prediction perspective. Further, we see a clear pattern between the series: the noisier the series, the lower the value of λ required to minimize cross-validated mean-square-error. We hypothesize that this occurs because

[9] If one were to tune at α values near 0 or 1, it would then become more difficult to distinguish elastic net models from adaptive LASSO or ridge.

[10] That is, the full regression imposing the oracle zero-coefficient restrictions.

[11] These correspond to one, two, and ten observations per estimated parameter.

the signal is poorer for noisier series, so to achieve the optimal cross-validated mean-square-error by retrieving this poor signal, we require a lower penalty. A similar pattern emerges when varying sample sizes: the larger the sample size, the lower the value of λ required to minimize cross-validated mean-square-error. In larger sample sizes, parameter estimation variance becomes more negligible and cross-validation chooses a λ parameter that is closer to the OLS solution. In contrast with the results for λ, the average optimal α shown on the right panel is very close to 0.5 for all series and all sample sizes. There is weak evidence suggesting that α is higher for more persistent series, with no clear variation pattern over the sample size. Thus, we find that cross-validation selects the λ penalization parameter in a predictable way, and the selected α parameter may depend weakly on the level of persistence in the series.

T	Y1	Y2	Y3
Small	4.74	3.61	2.94
Medium	3.18	2.50	1.89
Large	1.23	1.14	0.78

(a) λ

T	Y1	Y2	Y3
Small	0.52	0.48	0.48
Medium	0.53	0.50	0.50
Large	0.52	0.50	0.49

(b) α

Table 1.1: Left: mean penalization parameter (λ). Right: mean trade-off between ridge and LASSO parameter (α). Mean values are reported for each series by small (T=27), medium (T=54), and large (T=270) sample sizes.

The first evaluation criterion we consider is the 'distance' between the estimated VAR and the true, population DGP. To check this, we estimate the root mean square forecast error (RMSFE) for the VAR for one period-ahead recursive forecasts. This measure gives us an understanding of how well these models forecast, and how close the model is to the population DGP in an out-of-sample sense. As in Giannone et al. (2012) and others, we note that out-of-sample performance can also be thought of as model validation, since it indicates error in both parameter estimation and model specification.

Table 1.2 reports the fractions (%) of the one-period ahead RMSFEs of the estimated models to the RMSFE under the true parameter estimates averaged across all simulations with an out-of-sample period of 1000 observations. As the sample size grows, all methods improve in their forecast ability. In large samples, estimated fractions hit just over 100%, reflecting the case where parameter estimates are nearly indistinguishable from the true parameter values[12]. The table shows that the RMSFE for the penalized regression methods approaches that of the oracle and of the true model as the sample size grows, reflecting the strength of these methods in prediction. The simple OLS model is nearly uniformly beaten given the large amount of forecast variance inherited from variance in

[12]This is particularly relevant for the noisiest series, $Y3$.

later lag coefficient estimates. For the penalized models, the 'adaptive' part of the penalties improves the forecast ability as the sample size shifts. This is related to the degree of noise in the series: $Y3$ is considerably noisy, so the adaptive part of penalty may underperform as a result of spurious initial estimates, whereas $Y1$ is heavily autoregressive, so initial estimates bias the penalty in an improved direction. All of the penalized regression methods perform comparably for $Y2$, suggesting that the trade-off between poor initial estimates and adaptive penalization is approximately even for autoregressive variables with slight autoregressive tendencies. Note that as the sample size grows, however, the difference between the sparse penalized regression methods vanishes as they all forecast at oracle or near-oracle efficiency. Ridge, the only non-sparse estimator, under performs in larger samples for series with larger persistence. This is probably due to little correlation between the explanatory variables, which also explains why elastic net models tend to under-perform the LASSO[13].

We may also wonder what proportion of the forecast error is due purely to incorrect model selection. Table 1.3 shows the RMSFE fractions of the sparse penalized regression methods using their selected set of variables and population parameters: that is, estimation is performed only to uncover the sparsity pattern and then the RMSFE is computed with the population coefficients conditional on the estimated (possibly incorrect) sparsity pattern. The RMSFE fractions are considerably smaller than those in Table 1.2 across all procedures, suggesting that most of the error in the original estimation methods in Table 1.2 does not arise out of inaccurate selection. Moreover, the fractions in Table 1.3 are very similar across procedures, suggesting that all procedures select similar sparsity patterns. Hence, deviations in the original fractions in Table 1.2 arise out of estimation error in the parameters. We can therefore largely attribute the apparent large-sample out-performance of adaptive procedures in Table 1.2 to their bias-correction ability[14], and not to their ability to uncover the true sparsity pattern.

In addition to its ability to forecast, we are also interested in the ability of the adaptive elastic net to generate accurate structural impulse response functions. In a typical monetary VAR, we frequently consider a shock to interest rate growth. To that end, in Table 1.4 we consider the root mean square error of the impulse response of a one standard deviation shock of $Y3$, the variable whose properties mirror interest rate changes, onto all the variables[15]. Here, we estimate the sample covariance matrix as usual, but we order the variables according to the correct order, which is known. The tables reveal several things. First, as expected, impulse responses tend to have higher RMSEs when

[13]Previously conducted simulations suggest that ridge and elastic net performed better than LASSO in cases with many correlated lagged dependent variables. We will also see this later in the empirical example.

[14]That is, the ability of the magnitude of the penalty to vanish in large samples for the correctly included parameters, thus eliminating bias.

[15]We limit our attention to the adaptive elastic net since it is the primary model we focus on in this study, and computational costs limit our ability to conduct large impulse response analyses.

	Y1	Y2	Y3
Ridge	134.65	108.88	104.14
LASSO	124.94	110.12	104.74
Ad LASSO	119.90	112.85	108.10
Elastic Net	126.78	110.71	105.78
Ad EN	120.22	114.13	109.71
OLS	133.60	133.11	132.62
Oracle	109.34	105.07	102.04

(a) Small Sample RMSFE Ratio

	Y1	Y2	Y3
Ridge	113.31	105.39	101.65
LASSO	108.38	105.35	102.03
Ad LASSO	105.44	105.78	103.14
Elastic Net	109.50	105.35	102.27
Ad EN	106.27	105.91	103.78
OLS	111.31	111.14	111.44
Oracle	103.12	102.12	101.02

(b) Medium Sample RMSFE Ratio

	Y1	Y2	Y3
Ridge	102.48	101.44	100.54
LASSO	101.21	101.03	100.57
Ad LASSO	100.88	100.89	100.64
Elastic Net	101.39	101.05	100.62
Ad EN	100.94	100.90	100.78
OLS	101.74	101.72	101.75
Oracle	100.40	100.39	100.18

(c) Large Sample RMSFE Ratio

Table 1.2: One-period ahead root-mean-square forecast error (RMSFE) fractions in percentage terms relative to the RMSFE estimated with the true parameters for various modelling procedures and three sample sizes: small (T=27), medium (T=54), and large (T=270).

	Y1	Y2	Y3
LASSO	105.02	104.20	100.43
Ad LASSO	105.79	104.12	100.40
Elastic Net	103.28	103.59	100.40
Ad EN	104.01	103.64	100.39

(a) Small Sample

	Y1	Y2	Y3
LASSO	100.28	102.51	100.42
Ad LASSO	100.36	102.53	100.38
Elastic Net	100.20	102.07	100.39
Ad EN	100.30	102.10	100.37

(b) Medium Sample

	Y1	Y2	Y3
LASSO	100.12	100.14	100.28
Ad LASSO	100.32	100.27	100.23
Elastic Net	100.06	100.09	100.24
Ad EN	100.17	100.17	100.20

(c) Large Sample

Table 1.3: The one-period ahead root-mean-square forecast error (RMSFE) fractions of the true parameter model post-penalized selection in percentage terms relative to the RMSFE estimated with the full set of true parameters for various modelling procedures and three sample sizes: small (T=27), medium (T=54), and large (T=270).

the sample size is small. Second, the adaptive elastic net tends to have similarly sized RMSEs as the oracle when the sample size is large. Moreover, for small sample sizes, the adaptive elastic tends to even outperform the oracle, especially at shorter horizons. Lastly, the adaptive elastic net errors less than the oracle for impulse response functions on series with high persistence, as in series $Y1$. This echoes the asymptotic properties in the previous section, as well as the strength of the adaptive elastic net in estimating impulse response functions in small, finite samples.

So far, we have seen that the adaptive elastic net estimates the model with strong forecast accuracy and generates accurate impulse response functions. In order to conduct structural analysis however, accurate inference for impulse responses is also desirable. For this purpose, we use 1000 bootstrap estimates and check the nominal coverage ratio (the percentage of times the true impulse response is in the bootstrap confidence interval) for the 95% symmetric confidence interval of the impulse response function. Since there are no additional stationary exogenous variables here, we use the i.i.d. residual bootstrap. According to the asymptotic theory presented above, the bootstrap confidence interval should contain the true impulse response in 95% of the simulations for both of these estimators. In spite of this result, Killian (1998) notes that even the OLS-estimated bootstrap yields poor confidence intervals in autoregressive models due to finite-sample

IRF RMSE	Y1						Y2						Y3					
	Small		Medium		Large		Small		Medium		Large		Small		Medium		Large	
Horizon	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or
1	1.40	1.87	1.13	1.30	0.69	0.56	1.46	1.96	1.11	1.32	0.69	0.61	1.29	1.76	1.06	1.28	0.70	0.54
2	1.68	1.45	1.26	1.09	0.70	0.51	1.25	0.74	0.92	0.55	0.44	0.25	1.00	0.51	0.78	0.31	0.42	0.11
3	1.71	1.15	1.26	0.87	0.68	0.41	1.24	0.36	0.89	0.24	0.40	0.09	0.90	0.23	0.73	0.11	0.36	0.02
4	1.30	0.92	0.98	0.69	0.55	0.33	0.54	0.20	0.35	0.11	0.16	0.03	0.53	0.11	0.29	0.04	0.09	0.01
5	1.05	0.75	0.76	0.55	0.44	0.26	0.46	0.12	0.26	0.05	0.07	0.01	0.38	0.06	0.21	0.02	0.05	0.00
6	0.89	0.63	0.61	0.45	0.35	0.21	0.42	0.08	0.19	0.03	0.04	0.01	0.42	0.03	0.24	0.01	0.05	0.00
7	0.71	0.54	0.50	0.37	0.28	0.17	0.27	0.06	0.13	0.01	0.03	0.00	0.23	0.02	0.11	0.00	0.02	0.00
8	0.60	0.48	0.40	0.31	0.22	0.14	0.22	0.04	0.10	0.01	0.02	0.00	0.22	0.01	0.09	0.00	0.02	0.00

Table 1.4: Scaled (by 10) root mean square errors (RMSEs) for the impulse response function of $Y3$ onto $Y1$, $Y2$, and $Y3$ over a 1-8 period horizon. As expected, RMSEs are smaller for larger sample sizes.

Coverage	Y1						Y2						Y3					
	Small		Medium		Large		Small		Medium		Large		Small		Medium		Large	
Horizon	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or	AE	Or
1	0.41	1.00	0.56	1.00	0.80	1.00	0.41	1.00	0.45	1.00	0.81	1.00	0.23	1.00	0.42	1.00	0.74	1.00
2	0.40	0.49	0.61	0.79	0.88	1.00	0.36	0.69	0.35	0.59	0.91	0.76	0.48	0.93	0.61	0.95	0.78	0.98
3	0.36	0.61	0.60	0.62	0.92	0.99	0.40	0.96	0.37	0.94	0.93	0.60	0.50	1.00	0.62	1.00	0.78	1.00
4	0.23	0.42	0.34	0.48	0.90	0.92	0.22	0.70	0.24	0.65	0.54	0.68	0.32	0.93	0.34	0.95	0.55	0.98
5	0.31	0.55	0.36	0.43	0.82	0.78	0.37	0.92	0.36	0.92	0.55	0.72	0.41	1.00	0.41	1.00	0.53	1.00
6	0.34	0.43	0.41	0.40	0.77	0.64	0.42	0.72	0.42	0.68	0.68	0.77	0.50	0.93	0.60	0.95	0.71	0.98
7	0.46	0.55	0.49	0.40	0.79	0.53	0.52	0.92	0.50	0.91	0.85	0.79	0.55	1.00	0.65	1.00	0.80	1.00
8	0.31	0.45	0.41	0.40	0.77	0.44	0.29	0.74	0.30	0.70	0.84	0.81	0.39	0.93	0.45	0.95	0.76	0.98

Table 1.5: Bootstrap coverage ratio for the impulse response function of $Y1$ onto $Y1$, $Y2$, and $Y3$ over a 1-8 period horizon. Bootstrap coverage ratios are larger for larger sample sizes.

bias toward zero in the estimated persistence parameters. This bias becomes magnified in highly non-linear impulse response functions, resulting in particularly biased and skewed finite-sample impulse response confidence intervals. In the penalized regression setting, this attenuation bias is exacerbated since parameters are biased toward zero by construction. Hence, though the oracle property of the adaptive elastic net should result in the correct asymptotic coverage ratio, we expect the finite-sample coverage ratio to be worse for this model than for the oracle model.

Table 1.5 shows the coverage ratios for the impulse responses of $Y3$ onto the other variables for the adaptive elastic net and the oracle estimators. The oracle confidence interval is conservative in all sample sizes, possibly reflecting the bias noted in Killian (1998). By contrast, the adaptive elastic net bootstrap interval is too small, particularly in small samples, and does not fully converge to that of the oracle nor to the asymptotic coverage ratio of 95%. Despite this shortcoming, the coverage ratio of the adaptive elastic net's confidence interval considerably improves when the sample size grows, reflecting convergence to the asymptotic distribution. Thus, although coverage appears small for adaptive elastic net regression, the confidence interval moves toward the asymptotic coverage ratio, indicating that poor coverage stems from a bias problem in bootstrap estimated VAR impulse responses that is exacerbated by adaptive elastic net estimation.

To investigate this further, we plot in Figure 1.1 the average impulse responses over all simulations. Specifically, we consider the impulse response of a shock to $Y3$ on the dependent variables $Y1$, $Y2$, and $Y3$ with the medium sample size, $T = 54$. The plots show that the estimated impulse responses of the adaptive elastic net estimates and the oracle estimates are very close, reflecting the identical consistency rate for the two estimators. Moreover, they are not far from the true impulse response, suggesting that the adaptive elastic net is an accurate estimator for the structural behavior of the dynamical system. We can also see from these plots that confidence intervals for the adaptive elastic net estimator are close to those of the oracle. They tend to be too wide and heavily skewed at later horizons, and are too tight over short horizons. This suggests that the bootstrap confidence intervals for the adaptive elastic net are poor in finite samples, and perhaps very large sample sizes are required for them to converge to those of the oracle.

To summarize, the simulations demonstrate that the adaptive elastic net is effective in estimating VARs and their dynamic behavior. Forecasts from all penalized regression methods tend to be accurate, even in small samples. This suggests that VARs estimated with this procedure will produce accurate forecasts. Furthermore, the impulse responses generated by the adaptive elastic net are very close to those of the oracle, and sometimes even outperform the oracle. This suggests that the adaptive elastic net is effective in modeling macroeconomic systems generated by processes similar to stationary VARs, as well as in modeling the policy responses of such systems. In spite of these positive results, the bootstrap confidence intervals of the adaptive elastic net are tighter than

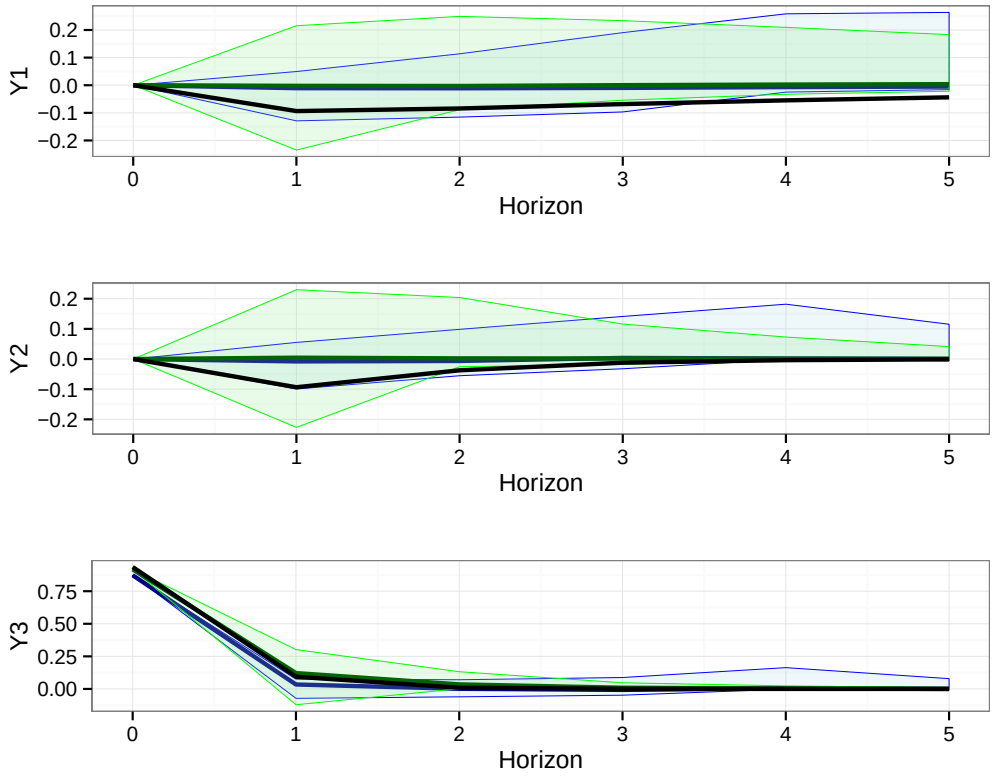


Figure 1.1: Adaptive elastic net estimated impulse response functions and their associated 95% confidence intervals for the effect of a one standard-deviation shock to $Y3$ on $Y1$, $Y2$, and $Y3$ after 0-5 periods for the medium sample ($T = 54$) estimates. The black lines denote the true impulse responses of the DGP, the dark blue lines denote the impulse responses estimated by the adaptive elastic net, and the dark green lines denote the impulse responses estimated by the oracle. Likewise, the filled in areas denote the confidence interval estimated by the oracle (green) and by the adaptive elastic net (blue).

those of the oracle in finite samples, with demonstrated skew. In small samples, the bootstrap confidence interval is limited in its ability, and tends to be too small relative to the nominal coverage ratio. This suggests that when the sample size is large, bootstrap inference may be more dependable, whereas finite sample improvements to the bootstrap method may be needed when facing small samples.

1.4 Empirical Study

In this section we demonstrate the strength of the adaptive elastic net in estimating large VARs empirically. The data used here come from Stock and Watson (2012), who have constructed a quarterly data set on 200 macroeconomic aggregates from 1959-Q3 to 2011-Q3, thus including the last recession and recovery. The series are transformed in similar fashion to Christiano et al. (1998), where filtering is minimal just to the point of near-stationarity; for details, see Table 1.7 in the appendix. In general, real aggre-

gates are taken in log-differences, price indices are taken in log-differences, and financial aggregates are taken in raw differences. Despite some of the literature on the possible non-stationarity of inflation series, we choose to express these as they are to preserve the amount of information that they possess. In double-difference form, much of this information is lost and turns into noise. Moreover, previous work on macroeconomic VARs and impulse response analysis such as (Christiano et al., 1998) tends to consider inflation series in levels, and so we maintain that tradition in order to place our results in the standard empirical macro context.

As in the previous section, estimation is done equation-by-equation, conforming to the notion that different series may be driven by a varying number of explanatory variables, and therefore require different factors of shrinkage and selection. We allow for variation in λ and in α and choose them by five-fold cross-validation across a two-dimensional grid. We add an unpenalized constant to each equation (equivalent of centering), and we scale the regressors to have unit variance for greater computational stability. Lastly, we borrow from the Bayesian literature on the Minnesota prior and we choose not to penalize the first lag of the particular dependent variable under consideration (Doan et al., 1984). The motivation behind this assumption is that own lags tend to matter more than other lags, and that the first lag is particularly informative, while later lags are far less influential. We find that this leads to improved forecasts and smoother impulse responses, and acknowledge that this modification is particularly well-suited to some of these data.

In the spirit of Giannone et al. (2012) we consider three types of VARs: a small trivariate VAR(5), a medium VAR(5) with seven variables, and a large VAR(5) with 25 variables. Notice that the large VAR cannot be estimated by least-squares, so some form of shrinkage is necessary for estimation. With that in mind, we estimate each VAR with OLS (where possible), ridge, LASSO, elastic net, adaptive LASSO, and adaptive elastic net. We estimate each model with an expanding window starting with 1974-Q4 and iterating until 2011-Q3. For each estimation, we construct 1-quarter ahead and 1-year ahead recursive forecasts, so that we can evaluate the forecasts over 146 and 143 quarterly observations, respectively. For comparison, we also estimate (1) a white noise process with non-zero mean, (2) a DFM with six factors and proceed exactly as in Stock and Watson (2002)[16] to construct direct forecasts, and (3) a Bayesian VAR with a Minnesota-type prior and shrinkage parameter varying with the sample size similar to Banbura et al. (2010)[17].

[16]Here, we double difference price indices as in Stock and Watson (2002).

[17]Specifically, we estimate AR(1) processes for each series, and then center the first own-lag coefficient on this estimate. This allows for some series, such as inflation, to be centered near a unit root, while other series can be centered closer to zero. The same filtering is done here as in the penalized regression models, which would otherwise considerably weaken the standard, unit-root centered Minnesota prior.

	Small			Medium			Large		
	GDP	Inflation	FedFunds	GDP	Inflation	FedFunds	GDP	Inflation	FedFunds
White Noise	0.82	0.56	1.03	0.82	0.56	1.03	0.82	0.56	1.03
DFM	0.77	0.25	1.17	0.77	0.25	1.17	0.77	0.25	1.17
BVAR	0.85	0.32	1.25	0.81	0.32	1.20	0.79	0.32	1.24
OLS	0.87	0.30	1.14	0.99	0.38	1.28	NA	NA	NA
Ridge	0.77	0.24	1.14	0.77	0.26	1.08	0.72	0.25	1.03
LASSO	0.80	0.25	1.14	0.77	0.25	1.07	0.73	0.25	1.02
Adaptive LASSO	0.85	0.26	1.09	0.79	0.26	1.12	0.82	0.27	1.13
Elastic Net	0.80	0.25	1.11	0.75	0.26	1.17	0.73	0.27	1.06
Adaptive Elastic Net	0.83	0.27	1.13	0.78	0.28	1.14	0.83	0.32	1.16

	Small			Medium			Large		
	GDP	Inflation	FedFunds	GDP	Inflation	FedFunds	GDP	Inflation	FedFunds
White Noise	2.33	2.17	2.23	2.33	2.17	2.23	2.33	2.17	2.23
DFM	3.76	0.48	2.89	3.76	0.48	2.89	3.76	0.48	2.89
BVAR	2.49	1.48	2.84	2.42	1.35	2.60	2.34	1.29	2.61
OLS	2.31	1.25	2.58	2.36	1.54	2.81	NA	NA	NA
Ridge	2.27	0.99	2.32	2.21	0.99	2.40	2.14	0.96	2.38
LASSO	2.34	1.06	2.35	2.22	0.99	2.31	2.12	1.00	2.47
Adaptive LASSO	2.41	1.16	2.40	2.27	1.01	2.43	2.23	1.12	2.75
Elastic Net	2.28	1.06	2.43	2.21	1.07	2.42	2.09	1.06	2.43
Adaptive Elastic Net	2.35	1.17	2.49	2.23	1.22	2.51	2.30	1.27	2.64

Table 1.6: Comparisons of one quarter ahead (top) and one-year ahead (bottom) root mean square forecast error (scaled by 100) for various models on three dependent variables: GDP, Inflation, and the Fed Funds rate. These also cover three different sizes: a small, medium, and large VAR.

The out-of-sample forecasting results in Table 1.6 suggest that penalized regression methods including the adaptive elastic net perform well in prediction. We first notice that the OLS estimates in the medium VAR yield considerably worse forecasts than those in the small VAR. This occurs because of the curse of dimensionality, which implies that as the number of coefficients increases without an increase in the sample size, the coefficients are estimated with higher variance. This then leads to worse forecasts, as seen in Table 1.6. For the noisier series, GDP growth and the change in the Fed Funds rate, the white noise forecast performs well, whereas for the more persistent series, inflation, this forecast performs quite poorly. The DFM forecasts improve on the white noise forecasts for the less noisy series, output growth and in particular inflation, but are weaker in forecasting the interest rate. This DFM is hard to beat, but penalized regression methods perform fairly well in comparison. They also perform closer to the white noise model for the noisiest series, the interest rate. Moreover, penalized regression methods also tend to perform comparably if not better than the Bayesian VARs with Minnesota-type prior. Within the class of penalized regression models, other than forecasting the interest rate in the medium VAR, the adaptive versions of the LASSO and the elastic net tend to slightly under-perform the non-adaptive versions. One reason why this may have occurred is because of small sample issues as in the simulation, where the initial estimators bias the penalty in the wrong direction. In this case, the LASSO and elastic net may act as robust prediction models but their adaptive versions may exacerbate the noise in these models. We note further that the ridge method performs very well, suggesting that simple shrinkage may be enough for prediction purposes and that grouping effects matter a great deal. In all, we see that penalized regression methods tend to predict accurately, and that this accuracy does not dissolve as the number of variables increases.

In addition to the strength of penalized regression methods in forecasting macroeconomic data, we also see that the adaptive elastic net delivers reasonable impulse responses, and can hence be used in standard structural macroeconomic policy analysis. As in previous work on impulse response analysis (Christiano et al., 1998), we consider the example of a monetary policy shock to the Federal Funds rate. In this structural study, we estimate the adaptive elastic net for the 25-variable VAR(5) over the full sample of observations and estimate the associated covariance matrix as in the theory section above. We then also apply the bootstrap method as above and collect 500 bootstrap estimates of the parameters. We identify the structural shocks using the standard Cholesky decomposition of the estimated covariance matrix, where we assume the following standard recursive ordering (Giannone et al., 2012). We assume that financial variables are the most reactive and can hence react immediately to a shock in the Federal Funds rate. We then assume that price indices cannot react contemporaneously to monetary policy, but can react to shocks in real variables, whereas the reverse is not the case. We note,

however, that in our analysis the difference in ordering of real and nominal variables is largely inconsequential to the shape of the impulse responses.

The effects of a monetary tightening are summarized in Figure 1.2, and greater details can be found in Figures 1.4-1.7 in the appendix. The effects of a monetary tightening in the estimated VAR adhere to the commonly upheld intuitions and theoretical models. That is, real activity declines due to increased incentives to save rather than spend. Prices are slow to react, the so-called 'price puzzle', but when they do, they tend to fall as less money is left in circulation. Interest rates naturally rise at first, but then decay down to zero within one year, since this is a monetary shock rather than a sustained monetary tightening. Among the financial variables, interest rates behave as the federal funds rate, the S&P first falls as is expected, but then returns to regularity (perhaps too quickly), and the exchange rate briefly appreciates reflecting interest rate parity. In sum, the impulse responses estimated by the adaptive elastic net adhere to the commonly held intuitions and theoretical results that are expected after a monetary tightening.

We may also be interested in the degree of penalization and its implication for the VAR structure. On the left panel of Figure 1.3, the plots show the magnitude of penalization λ over the entire estimation horizon for GDP growth (navy), inflation (green), and the Fed Funds rate (red). The penalization level decays in later periods when there are greater degrees of freedom for all three series. This suggests that as the degrees of freedom increase and parameter estimation variance falls, the estimated model requires less penalization to achieve an optimal forecast. We also note that the magnitude of penalization is considerably lower for inflation than for the Fed Funds rate and GDP growth. Recall that the parameter on the first lag is not penalized, so that variation in highly persistent series will be largely explained by the first lag. Thereafter, the remaining variation in the dependent variable will be very noisy. Cross-validation therefore selects a lower level of penalization for inflation since its high persistence is driven largely by its first lag, and very little penalization is required thereafter to predict a very noisy series and achieve an optimal cross-validated mean-square-error. These cross-validation results are closely in line with the simulation.

On the right panel of Figure 1.3, the plots show the tradeoff between ridge and LASSO (α) given by cross-validation. These series are considerably noisier than the left panel, but nonetheless show local stability. Notably, we see that α is below 0.5 on average for inflation, and higher than 0.5 on average for changes in the interest rate. This suggests that grouping effects may be slightly more important in forecasting noisier series.

The empirical results on the adaptive elastic net's ability to forecast and to correctly capture commonly held intuitions and observations about monetary tightening suggests that this model estimation strategy is highly suitable for macroeconomic VAR analysis. The adaptive elastic net is shown to forecast accurately regardless of the dimension of the parameter space. Moreover, even with the richness of a many-variable system, the

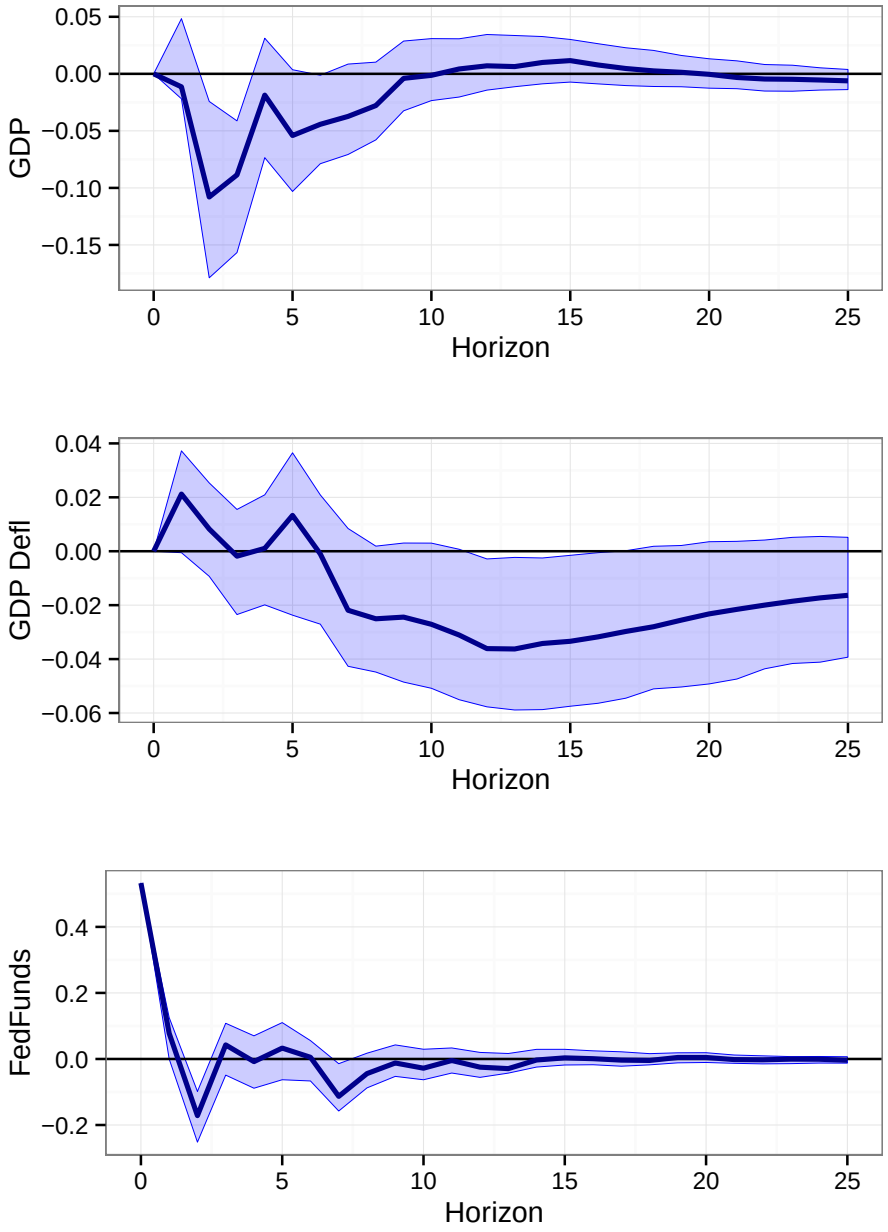


Figure 1.2: Adaptive elastic net estimated impulse response functions and their associated 90% confidence intervals for the effect of a one standard-deviation monetary policy shock to the FedFunds rate on GDP, Inflation, and the FedFunds rate after 0-25 quarters.

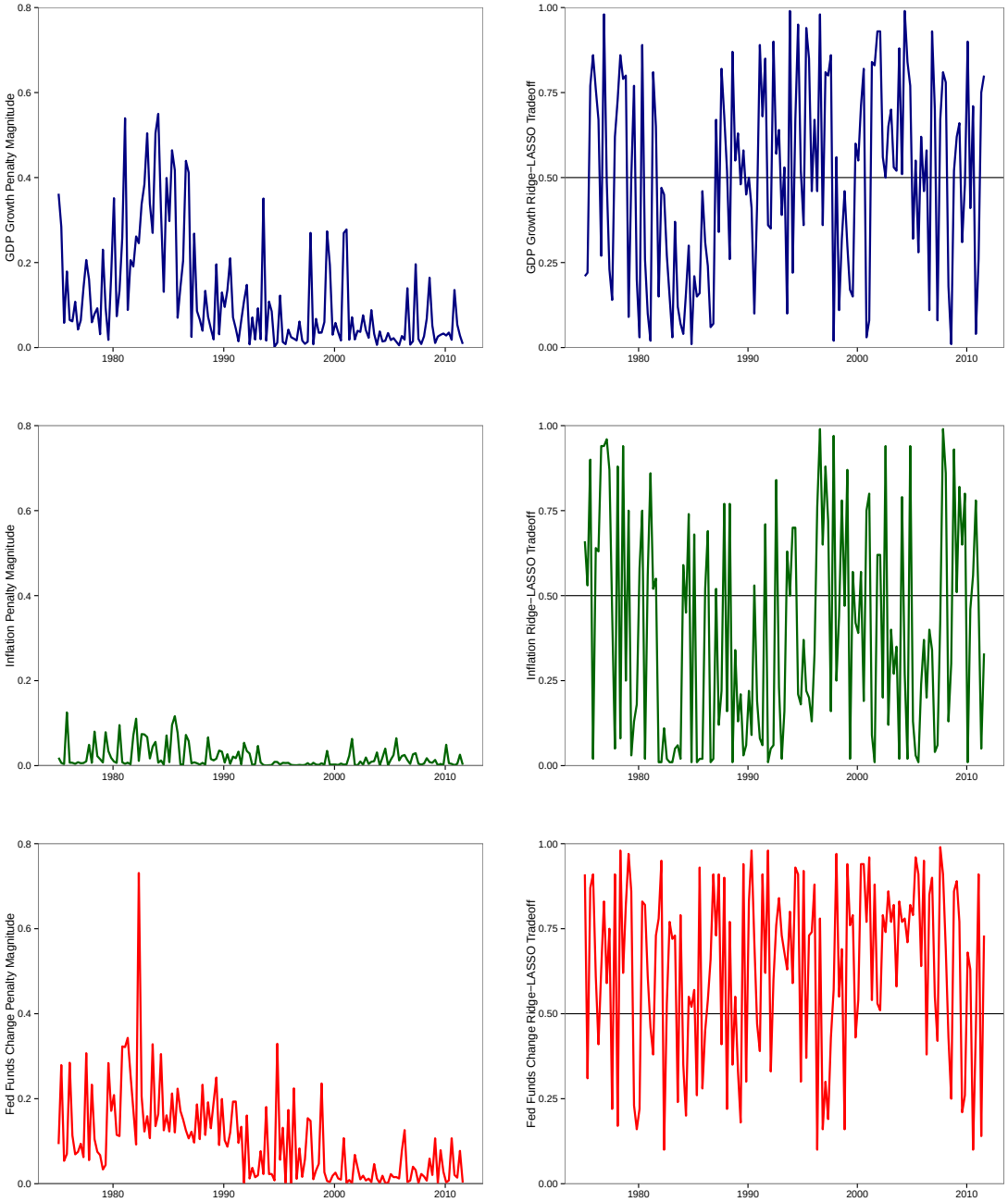


Figure 1.3: λ (left) and α (right) penalty parameters for GDP growth (navy), inflation (green), and the change in the Fed Funds rate (red) for the adaptive elastic net for each estimation in the expanding window sample.

adaptive elastic net can replicate the expected outcomes of a monetary policy shock, making it appealing from a policy-making point of view. Thus, the adaptive elastic net is consistent both with the data and with the commonly upheld macroeconomic beliefs. Moreover, the empirical dynamics of the magnitude of penalization reflect the simulation design, so that penalization is well-understood. This makes the adaptive elastic net highly attractive for medium and large VAR estimation and analysis.

1.5 Extensions and Conclusions

In this paper we proposed the use of the adaptive elastic net to estimate stationary vector autoregressive processes. The adaptive elastic net applies shrinkage and selection, shrinking irrelevant variables and excluding highly irrelevant variables altogether from the model. The adaptive elastic net enjoys the grouping effect of the ridge penalty, so that coefficients of highly correlated variables are close to one another. Furthermore, the 'adaptive' part of the penalty leads to the oracle property asymptotically, which ensures that the correct VAR lag structure is discovered and that non-zero parameters are estimated at the same rate as with OLS including only the non-zero variables. These properties allow not only for consistent estimation of the parameters in high-dimensional VARs, but also allow for accurate impulse response function estimation. Furthermore, we showed that the bootstrap is asymptotically justifiable in constructing confidence intervals for the parameter estimates and for the impulse responses. This makes the standard VAR tool-kit available using the adaptive elastic net, and to a great extent alleviates the curse of dimensionality commonly encountered in macroeconomic data. Moreover, unlike competing Bayesian procedures, this method is easily estimable and it ensures a sparse representation. This representation also lends itself to clear, Granger-causality interpretation, which is convenient in interpreting the VAR coefficients. The adaptive elastic net and associated penalized regression estimators appear to be very attractive for VAR modeling.

Several interesting extensions of this work present themselves for further research. First, we note that there are other penalized likelihood methods available that may be well suited for VARs. For example, there is the smoothly-clipped absolute deviation (SCAD) estimator of Fan and Li (2001), the Dantzig Selector of Candes and Tao (2007), the Autometrics procedure of Doornik (2009), and the interesting work by Belloni and Chernozhukov (2013) on using post-L_1 penalty estimation in a fixed regressor context. The first two papers discuss alternative penalized likelihood schemes that are not nested within the adaptive elastic net, and therefore excluded from the present analysis. The Autometrics algorithms are also excluded from this analysis since they have been heavily investigated in other work, for instance Epprecht et al. (2013), and since they too do not fit into the adaptive elastic net framework. In the last paper, the authors have shown

that estimation of the regression function using least-squares after L_1 selection reduces the risk of the overall model by bias-correcting the non-zero coefficients. This may also be an effective tool in estimation and inference of stationary VARs.

A second extension is to exploit cross-equation information. In this paper, we estimated the model equation by equation, rationalizing this with our conjecture that the variables that we are estimating generally vary in their degrees of persistence and sparsity, and therefore require different levels of shrinkage and selection. In the OLS setting, however, equation-by-equation modeling can usually be improved upon by estimating a pooled multivariate regression, as in Zellner (1962). In that context, information about the covariance matrix of the error terms is useful in minimizing the variance of the least-squares parameter estimates. An important caveat to this, however, is that provided that all of the regressors are included in all of the models, a pooled regression and equation by equation estimation are equivalent. Used explicitly in our setting, this 'seemingly unrelated regressions' concern is unclear since we are leaving variables out of the model while estimating simultaneously. Moreover, it is unlikely that the variance of the ensuing parameter estimates can be reduced by pooling across models, since all regressors are included in all estimations. In investigating post-L_1 penalization estimation, however, it may be useful to check whether taking the union of all regressors in the second, unpenalized estimation stage, can reduce the variance of the parameter estimates. We leave this discussion for further work.

Another extension of this paper involves improvements in impulse response confidence interval estimation, such as in Killian (1998). In his paper, Killian documented that the standard bootstrap, though shown to have an improvement over asymptotic confidence intervals through Edgeworth expansions (Bose, 1988), leads to biased and skewed confidence intervals in finite samples. Killian (1998) suggests correcting for this bias and skew by using bootstrap after bootstrap inference, and shows through simulations that this leads to improved finite sample confidence intervals. This procedure can be used in the context of the adaptive elastic net, and may improve the coverage ratio of the bootstrap confidence intervals presented here.

In this paper we have shown that the adaptive elastic net is a well-suited estimation strategy for VARs that are ordinarily limited by the curse of dimensionality. We have demonstrated that this method produces accurate forecasts without sacrificing the interpretable VAR representation. Structural VAR analysis and policy responses can also be used under this methodology without requiring *ex ante* correct model specification, thereby treating selection error as parameter estimation error. Our results are validated in a simulation analysis and on U.S. macroeconomic data. Finally, we note that the adaptive elastic net can be easily computed on several common software packages.

Appendix 1.A: Proofs

We suppose that the data $y_{1-p}, ..., y_0, y_1, ..., y_T$ are generated by the stationary VAR process (1.1). This process arises from the probability space $(\Omega, \{\mathcal{F}_t\}_{t=-\infty}^{T}, \mathcal{F}, \mathcal{P})$, with the filtration $\{\mathcal{F}_t\}_{t=-\infty}^{T}$, sigma-algebra $\mathcal{F}$, and probability measure $\mathcal{P}$.

Proof (Proposition 1.2.1).

This proof is taken largely from Zou and Hastie (2005). Notice that if $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$ it means that both coefficients are non-zero and have the same sign. Let X_i denote the i'th column of the matrix X. If the estimated coefficients are non-zero, as assumed, they must satisfy the first order conditions:

$$-2X_i'\{y - X\hat{\beta}(\lambda_1, \lambda_2)\} + \lambda_1\omega_i sgn\{\hat{\beta}_i(\lambda_1, \lambda_2)\} + 2\lambda_2\hat{\beta}_i(\lambda_1, \lambda_2) = 0 \tag{1.9}$$

$$-2X_j'\{y - X\hat{\beta}(\lambda_1, \lambda_2)\} + \lambda_1\omega_j sgn\{\hat{\beta}_j(\lambda_1, \lambda_2)\} + 2\lambda_2\hat{\beta}_j(\lambda_1, \lambda_2) = 0 \tag{1.10}$$

Subtracting the equation 1.9 from 1.10 gives:

$$(X_i' - X_j')\{y - X\hat{\beta}(\lambda_1, \lambda_2)\} - \lambda_1(\omega_i - \omega_j)/2 - \lambda_2(\hat{\beta}_i - \hat{\beta}_j) = 0$$

Letting $\hat{r}(\lambda_1, \lambda_2) = y - X\hat{\beta}(\lambda_1, \lambda_2)$:

$$(X_i' - X_j')\hat{r}(\lambda_1, \lambda_2) = (\lambda_1/2)(\omega_i - \omega_j) + \lambda_2(\hat{\beta}_i - \hat{\beta}_j)$$

Then:

$$\hat{\beta}_i - \hat{\beta}_j = \frac{(X_i' - X_j')\hat{r}(\lambda_1, \lambda_2)}{\lambda_2} + \frac{\lambda_1(\omega_j - \omega_i)}{2\lambda_2}$$

so that:

$$\begin{aligned} D_{\lambda_1,\lambda_2}(i,j) &= \frac{||\beta_i - \beta_j||_2}{||y||_2} \\ &\leq \frac{||\hat{r}(\lambda_1, \lambda_2)||_2||X_i' - X_j'||_2}{\lambda_2||y||_2} + \frac{\lambda_1||\omega_i - \omega_j||_2}{2\lambda_2||y||_2} \\ &\leq \frac{\sqrt{2(1-\rho)}}{\lambda_2} + \frac{\lambda_1||\omega_i - \omega_j||_2}{2\lambda_2||y||_2} \end{aligned}$$

where the last line follows from the fact that the X_i's are standardized and that $||\hat{r}(\lambda_1, \lambda_2)||_2^2 + \lambda_1||\omega\hat{\beta}||_1 + \lambda_2||\hat{\beta}||_2^2 \leq ||\hat{r}(\beta = 0)||_2^2$ and therefore $||\hat{r}(\lambda_1, \lambda_2)||_2 \leq ||\hat{r}(\beta = 0)||_2 = ||y||_2$.

Further, notice that if $\gamma = 1$ and the initial estimator is the naive elastic net:

$$\begin{aligned}
\frac{||\omega_i - \omega_j||_2}{||y|_2} &= \Big|\Big|\frac{1}{\beta_{I,i}} - \frac{1}{\beta_{I,j}}\Big|\Big|_2 ||y||_2^{-1} \\
&= \Big|\Big|\frac{\beta_{I,j} - \beta_{I,i}}{\beta_{I,j}\beta_{I,i}}\Big|\Big|_2 ||y||_2^{-1} \\
&\leq D_{I,\lambda_1,\lambda_2}(i,j)||\beta_{I,j}\beta_{I,i}||_2^{-1} \\
&\leq \frac{\sqrt{2(1-\rho)}}{\lambda_2||\beta_{I,j}\beta_{I,i}||_2}
\end{aligned}$$

where $D_{I,\lambda_1,\lambda_2}(i,j)$ denotes the distance between the initial naive elastic net estimators for X_i and X_j, as established in Zou and Hastie (2005). We can then obtain another expression for the second-stage distance:

$$D_{\lambda_1,\lambda_2}(i,j) \leq \frac{\sqrt{2(1-\rho)}}{\lambda_2} + \frac{\lambda_1\sqrt{2(1-\rho)}}{2\lambda_2^2||\beta_{I,j}\beta_{I,i}||_2}$$

□

Proof (Proposition 1.2.2).

This proof is largely an extension of Kock and Callot (2012) and Zou (2006) and is included for completeness. Consider the objective function:

$$L_T(\beta) = ||y - X\beta||_2^2 + \lambda_1 \sum_{k=K_1+1}^{K} \omega_k|\beta_k| + \lambda_2 \sum_{k=1}^{K} \beta_k^2$$

Let $\beta = \beta^o + \frac{u}{\sqrt{T}}$, then we can write the objective function (1.5) as:

$$L_T(u) = ||y - X(\beta^o + \frac{u}{\sqrt{T}})||_2^2 + \lambda_1 \sum_{k=K_1+1}^{K} \omega_k|\beta_k^o + \frac{u}{\sqrt{T}}| + \lambda_2 \sum_{i=k}^{K}(\beta_k^o + \frac{u}{\sqrt{T}})^2$$

note that if $\hat{u} = \text{argmin}L_T(u)$, then $\hat{\beta}_N = \beta^o + \frac{\hat{u}}{\sqrt{T}}$ so that $\hat{u} = \sqrt{T}(\hat{\beta}_N - \beta^o)$, where $\hat{\beta}_N$ is the naive adaptive elastic net estimator. Let's define:

$$
\begin{aligned}
V_T(u) &= L_T(u) - L_T(0) \\
&= ||y - X(\beta^o + \frac{u}{\sqrt{T}})||_2^2 - ||y - X\beta^o||_2^2 \\
&+ \lambda_1 \sum_{k=K_1+1}^{K} \omega_k(|\beta_k^o + \frac{u_k}{\sqrt{T}}| - |\beta_k^o|) + \lambda_2 \sum_{k=1}^{K} (\beta_k^o + \frac{u_k}{\sqrt{T}})^2 - \beta_k^{o2} \\
&= \frac{u'X'Xu}{T} - \frac{2u'X'e}{\sqrt{T}} \\
&+ \lambda_1 \sum_{k=K_1+1}^{K} \omega_k(|\beta_k^o + \frac{u_k}{\sqrt{T}}| - |\beta_k^o|) + \lambda_2 \sum_{k=1}^{K} 2\frac{\beta_k^o u_k}{\sqrt{T}} + \frac{u_k^2}{T}
\end{aligned}
$$

By Theorem 11.2.1 in Brockwell and Davis (2009), we see that $\frac{u'X'Xu}{T} \xrightarrow{p} u'(I_N \otimes C)u$, and from Proposition 7.9 in Hamilton (1994) that $\frac{X'e}{\sqrt{T}} \xrightarrow{L} W \sim N(0, \Sigma \otimes C)$. Hence:

$$
\frac{u'X'Xu}{T} + \frac{u'X'e}{\sqrt{T}} \xrightarrow{L} u'(I_N \otimes C)u - 2u'W
$$

The second penalty term's contribution converges on zero:

$$
\lambda_2 \sum_{k=1}^{K} 2\frac{\beta_k^o u_k}{\sqrt{T}} + \frac{u_k^2}{T} \to 0
$$

since $\frac{\lambda_2}{\sqrt{T}} \to 0$ and $2u_k\beta_k^o + \frac{u_k^2}{\sqrt{T}}$ is finite. Now, consider the β_k^o in $\mathcal{A}^o$, that is $\beta_k^o \neq 0$, then:

$$
\begin{aligned}
\lambda_1 \omega_k(|\beta_k^o + \frac{u_k}{\sqrt{T}}| - |\beta_k^o|) &= \lambda_1 \Big|\frac{1}{\beta_{I,k}}\Big|^\gamma \frac{u_k}{\sqrt{T}} \Big(|\beta_k^o + \frac{u_k}{\sqrt{T}}| - |\beta_k^o|\Big) / \frac{u_k}{\sqrt{T}} \\
&= \frac{\lambda_1}{\sqrt{T}} \Big|\frac{1}{\beta_{I,k}}\Big|^\gamma u_k \Big(|\beta_k^o + \frac{u_k}{\sqrt{T}}| - |\beta_k^o|\Big) / \frac{u_k}{\sqrt{T}} \\
&\xrightarrow{p} 0, \quad \forall u_k \in \mathbb{R}
\end{aligned}
$$

since:

(i) $\frac{\lambda_1}{\sqrt{T}} \to 0$

(ii) $|\frac{1}{\beta_{I,k}}|^\gamma \xrightarrow{p} |\frac{1}{\beta_k^o}|^\gamma$

(iii) $u_k \Big(|\beta_k^o + \frac{u_k}{\sqrt{T}}| - |\beta_k^o|\Big) / \frac{u_k}{\sqrt{T}} \to u_k sgn(\beta_k^o)$

If on the other hand $\beta_k^o = 0$ so $\beta_k^o \notin \mathcal{A}^o$, then:

$$\begin{aligned}
\lambda_1\omega_k\Big(|\beta_k^o + \frac{u_k}{\sqrt{T}}| - |\beta_k^o|\Big) &= \frac{\lambda_1}{\sqrt{T}}\Big|\frac{1}{\beta_{I,k}}\Big|^\gamma |u_k| \\
&= \frac{\lambda_1}{T^{(1-\gamma)/2}}\Big|\frac{1}{\sqrt{T}\beta_{I,k}}\Big|^\gamma |u_k| \\
&\xrightarrow{p} \begin{cases} \infty & \text{if } u_k \neq 0 \\ 0 & \text{if } u_k = 0 \end{cases}
\end{aligned}$$

since $\frac{\lambda_1}{T^{(1-\gamma)/2}} \to \infty$ and $\sqrt{T}\beta_{I,k} \in O_p(1)$. From these results we see that:

$$V_T(u) \xrightarrow{L} V(u) = \begin{cases} u'(I_N \otimes C)u - 2u'W & \text{if } u_k = 0, \forall k \notin \mathcal{A}^o \\ \infty & \text{if } u_k \neq 0 \text{ for some } k \notin \mathcal{A}^o \end{cases}$$

Because $V_T(u)$ is convex and $V(u)$ has a unique minimum, Theorem 2 of Knight and Fu (2000) shows that $\text{argmin}V_T(u) \xrightarrow{L} \text{argmin}V(u)$ and hence:

$$\begin{aligned}
\hat{u}_{\mathcal{A}^C} &\xrightarrow{L} 0 \\
\hat{u}_{\mathcal{A}^o} &\xrightarrow{L} N(0, [I_N \otimes C]^{-1}_{\mathcal{A}^o}[\Sigma \otimes C]_{\mathcal{A}^o}[I_N \otimes C]^{-1}_{\mathcal{A}^o})
\end{aligned}$$

or alternatively:

$$\begin{aligned}
\sqrt{T}(\hat{\beta}_{\mathcal{A}^C,N} - \beta^o_{\mathcal{A}^C}) &\xrightarrow{L} 0 \\
\sqrt{T}(\hat{\beta}_{\mathcal{A},N} - \beta^o_{\mathcal{A}^o}) &\xrightarrow{L} N(0, [I_N \otimes C]^{-1}_{\mathcal{A}^o}[\Sigma \otimes C]_{\mathcal{A}^o}[I_N \otimes C]^{-1}_{\mathcal{A}^o})
\end{aligned}$$

Which yields the consistency at rate-$\sqrt{T}$ and the oracle asymptotic distribution for the naive adaptive elastic net. Using the result that $\frac{\lambda_2}{\sqrt{T}} \to 0$:

$$\begin{aligned}
\sqrt{T}(\hat{\beta}_{\mathcal{A}^C} - \beta^o_{\mathcal{A}^C}) &\xrightarrow{L} 0 \\
\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta^o_{\mathcal{A}^o}) &\xrightarrow{L} N(0, [I_N \otimes C]^{-1}_{\mathcal{A}^o}[\Sigma \otimes C]_{\mathcal{A}^o}[I_N \otimes C]^{-1}_{\mathcal{A}^o})
\end{aligned}$$

That is, the oracle rate is achieved for the adaptive elastic net. Moreover, the convergence rate for the covariance matrix estimator follows directly from the rate for the parameter estimate. It remains for us to prove the asymptotic oracle sparsity pattern.

Let us first consider the naive adaptive elastic net. Note that if $Pr(\hat{\beta}_{\mathcal{A}^{oC}} = 0) \to 1$ then $Pr(\hat{\mathcal{A}} = \mathcal{A}^o) \to 1$ since $\hat{\beta}_{\mathcal{A}^o}$ is $\sqrt{T}$-consistent. To show this, assume $\hat{\beta}_j \neq 0$ for $j \notin \mathcal{A}^o$. Denote, as before, X_j as the j'th column of X. From the first order condition:

$$2X_j'(y - X\hat{\beta}) + \lambda_1\omega_j sgn(\hat{\beta}_j) + 2\lambda_2\hat{\beta}_j = 0$$

which is equivalent to:

$$\frac{2X_j'(y - X\hat{\beta})}{\sqrt{T}} + \frac{\lambda_1 \omega_j sgn(\hat{\beta}_j)}{\sqrt{T}} + \frac{2\lambda_2 \hat{\beta}_j}{\sqrt{T}} = 0 \tag{1.11}$$

The second term in the expression has the following property:

$$\Big|\frac{\lambda_1 \omega_j sgn(\hat{\beta}_j)}{\sqrt{T}}\Big| = \frac{\lambda_1 \omega_j}{\sqrt{T}} = \frac{\lambda_1}{T^{(1-\gamma)/2}}\Big|\frac{1}{\sqrt{T}\hat{\beta}_j}\Big|^{\gamma} \to \infty$$

as before. On the other hand, the first term is:

$$\frac{2X_j'(y - X\hat{\beta})}{\sqrt{T}} = \frac{2X_j'(e - X\hat{\beta} + X\beta^o)}{\sqrt{T}} = \frac{2X_j'e}{\sqrt{T}} - \frac{2X_j'X\sqrt{T}(\hat{\beta} - \beta^o)}{T}$$

Without loss of generality suppose that β_j^o is the population coefficient of the k'th variable in the i'th equation, then:

$$\frac{X_j'e}{\sqrt{T}} \xrightarrow{L} N(0, \Sigma_{ii}C_{kk}), \quad \frac{X_j'X}{T} \xrightarrow{p} (I_N \otimes C)_j$$

the j'th row of $(I_N \otimes C)$, and $\sqrt{T}(\hat{\beta} - \beta^o)$ is $O_p(1)$ as before. That is, the first term is $O_p(1)$. Finally, we consider the third term:

$$\frac{2\lambda_2 \hat{\beta}_j}{\sqrt{T}} \to 0$$

since $\frac{\lambda_2}{\sqrt{T}} \to 0$ and $2\hat{\beta}_j$ is a constant plus a local to zero noise of order $O_p(1/\sqrt{T})$. Hence, $Pr(\hat{\beta}_j \neq 0) = Pr(2X_j'(y - X\hat{\beta}) + \lambda_1 \omega_j sgn(\hat{\beta}_j) + 2\lambda_2 \hat{\beta}_j = 0) \to 0$ as $T \to \infty$ and therefore $Pr(\hat{\mathcal{A}} = \mathcal{A}^o) \to 1$ as $T \to \infty$. □

In order to prove Proposition 1.2.3, we need a weak convergence result for the residual bootstrap in the time-series case. We will need to prove several small lemmas along the way in order to prove the weak convergence lemma given below, which will give us our necessary result for proving the proposition. Note that we will prove weak convergence using the Mallow's $d_2(F, G)$ metric between two distribution functions, defined as:

$$d_2(F, G) = \inf_{X \sim F, Y \sim G}\{E(X - Y)^2\}^{1/2}$$

We note that the distance metric satisfies the triangle inequality and that if this distance converges to zero, it implies that the two distribution functions converge. This is seen in the following lemma proven in Bickel and Freedman (1981):

Lemma 1.5.1. *Let α_T, α be two probability measures. Then $d_2(\alpha_T, \alpha) \to 0$ if and only*

if

$$E_{\alpha_T}(X^2) = \int x^2 \alpha_T(dx) \to E_\alpha(X^2) = \int x^2 \alpha(dx) \tag{1.12}$$

and α_T *converges weakly to* α*.*

We use this distance to show the following weak convergence lemma:

Lemma 1.5.2 (Weak Convergence)**.** *Let* $z_t = (y'_{t-1}, ..., y'_{t-p})'$ *and* $z_t^* = (y_{t-1}^{*'}, ..., y_{t-p}^{*'})'$*. Then:*

$$d_2\Big(\frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T} z_t^* \otimes e_t^*\Big), \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T} z_t \otimes e_t\Big)\Big) \to 0 \text{ as } T \to \infty \tag{1.13}$$

The proof of Lemma 1.5.2 comes in several parts. We denote by $\sigma(Y)$ the sigma-algebra generated by $y_{1-p}, ..., y_0, y_1, ..., y_T$. First, we have convergence between the bootstrap residuals' distribution and the fundamental distribution driving the true errors.

Lemma 1.5.3. *Suppose* $\{e_t^*\}_{t=1}^T$ *are the centred residuals used in the bootstrap and* $\{e_t\}_{t=1}^T$ *are true errors from the underlying DGP. Define* $J = 1, ..., T$ *and let* $Pr(J = k) = \frac{1}{T}$*. Further, let* F *denote the distribution of* e_t*,* F_t *denote the empirical distribution of* e_t*,* $\hat{F}_t$ *denote the empirical distribution of the estimated residuals* $\hat{e}_t$*, and* F_t^* *denote the distribution of the centred* e_t^**. Then:*

$$\begin{aligned}
E(||\hat{e}_J - e_J||_2^2 | \sigma(Y)) &= O_p(1/T) && (1.14)\\
E(||\hat{e}_J^* - \hat{e}_J||_2^2 | \sigma(Y)) &= O_p(1/T) && (1.15)\\
d_2(F_t^*, F) &\le d_2(F_t^*, \hat{F}_t) + d_2(\hat{F}_t, F) \\
&\le d_2(F_t^*, \hat{F}_t) + d_2(\hat{F}_t, F_t) + d_2(F_t, F) \\
&\to 0 \text{ as } T \to \infty && (1.16)\\
E_{F_t^*}(||e_t^*||_2^2 | \sigma(Y)) &\xrightarrow{p} E_F(||e_t||_2^2|) && (1.17)
\end{aligned}$$

Proof (Lemma 1.5.3).

$$
\begin{aligned}
E(||\hat{e}_J - e_J||_2^2|\sigma(Y)) &= \frac{1}{T}\sum_{t=1}^{T}||\hat{e}_t - e_t||_2^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}||\sum_{j=1}^{p}(\hat{B}_j - B_j)y_{t-j}||_2^2 \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}||(\hat{B}_j - B_j)y_{t-j}||_2^2 \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}||(\hat{B}_j - B_j)||_2^2||y_{t-j}||_2^2
\end{aligned}
$$

But recall that $||(\hat{B}_j - B_j)||_2^2 \in O_p(1/T)$, so $E(||\hat{e}_J - e_J||_2^2|\sigma(Y)) = O_p(1/T)$. Further:

$$
\begin{aligned}
E(||\hat{e}_J^* - \hat{e}_J||_2^2|\sigma(Y)) &= ||\frac{1}{T}\sum_{t=1}^{T}\hat{e}_t||_2^2 \\
&\leq \frac{1}{T^2}\sum_{t=1}^{T}||\hat{e}_t||_2^2 \\
&= O_p(1/T)
\end{aligned}
$$

For the next part, from Bickel and Freedman (1981) Lemma 8.4 we see that $d_2(F_t, F) \to 0$. Now for $d_2(\hat{F}_t, F_t)^2$, we need to take the infimum over all joint distributions with marginals $\hat{F}_t$ and F_t, the sampling distributions of the regression residuals and true errors. Note that e_J and $\hat{e}_J$ are sampled from F_t and $\hat{F}_t$, respectively, so that:

$$
\begin{aligned}
d_2(\hat{F}_t, F_t)^2 &= \inf_{\hat{e}_t \sim \hat{F}_t, e_t \sim F_t} E(||\hat{e}_t - e_t||_2^2) \\
&\leq E(||\hat{e}_J - e_J||_2^2) \\
&\leq E(E(||\hat{e}_J - e_J||_2^2|\sigma(Y)))
\end{aligned}
$$

but the inner conditional expectation is $O_p(1/T)$, so $d_2(\hat{F}_t, F_t)^2 \to 0$ in probability. Using the results from the two conditional expectation expressions above, a similar argument also yields $d_2(\hat{F}_t, F_t^*) \to 0$. This then gives us the result that $d_2(F_t^*, F) \to 0$ as $T \to 0$. With this result and using Lemma 1.5.1, we obtain the final equation. □

Recall that y_t has an infinite Vector Moving-Average representation (VMA(∞)), $y_T = \sum_{j=1}^{\infty}\Psi_{j-1}e_{T-j+1}$ and the bootstrap version y_t^* has representation $y_T^* = \sum_{j=1}^{T}\hat{\Psi}_{j-1}e^*_{J=T-j+1} + y_0$. Define:

$$
\begin{aligned}
w_T^* &= \sum_{j=1}^{T} \Psi_{j-1} e^*_{J=T-j+1} + y_0 \\
w_T &= \sum_{j=1}^{T} \Psi_{j-1} e_{J=T-j+1} + y_0
\end{aligned}
$$

where the $e_{J=T-j+1}$ and $e^*_{J=T-j+1}$ terms are the resampled realizations of the true errors and bootstrap errors, respectively.

Lemma 1.5.4. *Then:*

(a) $E(||y_t^* - w_t^*||_2^2 | \sigma(Y)) = O_p(1/T), \quad d_2(y_t^*, w_t^*) \to 0$

(b) $E(||w_t^* - w_t||_2^2 | \sigma(Y)) = O_p(1/T), \quad d_2(w_t^*, w_t) \to 0$

(c) $d_2(w_t, y_t) \to 0$

From the triangle inequality we then also have that $d_2(y_t^*, y_t) \to 0$.

Proof (Lemma 1.5.4).

(a)

$$
\begin{aligned}
E(||y_t^* - w_t^*||_2^2 | \sigma(Y)) &= E(|| \sum_{j=1}^{t} (\hat{\Psi}_{j-1} - \Psi_{j-1}) e_J^* ||_2^2 | \sigma(Y)) \\
&\leq E(\sum_{j=1}^{t} ||\hat{\Psi}_{j-1} - \Psi_{j-1}||_2^2 ||e_J^*||_2^2 | \sigma(Y)) \\
&\leq \sum_{j=1}^{t} ||\hat{\Psi}_{j-1} - \Psi_{j-1}||_2^2 E(||e_J^*||_2^2 | \sigma(Y))
\end{aligned}
$$

Note that Ψ_j is a polynomial in $B_1, ..., B_j$. Hence, by Slutsky's Theorem it is clear that $\hat{\Psi}_j \xrightarrow{p} \Psi_j$, for all j, since $\hat{B}_j \xrightarrow{p} B_j$. Moreover, since $||\hat{B}_j - B_j||_2^2 \in O_p(1/T)$, so by the triangle inequality and the Cauchy-Schwarz inequality, $||\hat{\Psi}_j - \Psi_j||_2^2 \leq ||\hat{B}_j - B_j||_2^2 = O_p(1/T)$. Then $\sum_{j=1}^{T} ||\hat{\Psi}_j - \Psi_j||_2^2 E(||e_J^*||_2^2 | \sigma(Y)) = O_p(1/T)$ using the last part of the previous lemma.

(b)

$$\begin{aligned} E(||w_t^* - w_t||_2^2|\sigma(Y)) &= E(||\sum_{j=1}^{t}(\hat{e}_J - e_J)||_2^2||\Psi_j||_2^2|\sigma(Y)) \\ &= \sum_{j=1}^{t}||\Psi_j||_2^2 E(||\hat{e}_J - e_J||_2^2|\sigma(Y)) \\ &= O_p(1/T) \end{aligned}$$

since $E(||\hat{e}_J - e_J||_2^2|\sigma(Y))$ is $O_p(1/T)$.

(c)

$$\begin{aligned} E(||w_t - y_t||_2^2|\sigma(Y)) &= E(||\sum_{j=1}^{t}\Psi_j(e_J - e_{t-j+1})||_2^2|\sigma(Y)) \\ &= \sum_{j=1}^{t}||\Psi_j||_2^2 E(||(e_J - e_{t-j+1})||_2^2|\sigma(Y)) \end{aligned}$$

Note that $d_2(F_t, F) \to 0$ which then gives $E(||(e_J - e_{t-j+1})||_2^2|\sigma(Y)) = 0$.

□

We can now prove the weak convergence lemma.

Proof (Lemma 1.5.2).

Let $\tilde{z}_t = (w_{t-1}', ..., w_{t-p}')'$ and $\tilde{z}_t^* = (w_{t-1}^{*'}, ..., w_{t-p}^{*'})'$. We prove the following three equations and then by the triangle inequality we obtain Lemma 1.5.2:

$$d_2\Big(\frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}\tilde{z}_t \otimes e_t\Big), \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}\tilde{z}_t^* \otimes e_t^*\Big)\Big) \to 0 \tag{1.18}$$

$$d_2\Big(\frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}z_t \otimes e_t\Big), \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}\tilde{z}_t \otimes e_t\Big)\Big) \to 0 \tag{1.19}$$

$$d_2\Big(\frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}z_t^* \otimes e_t^*\Big), \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}\tilde{z}_t^* \otimes e_t^*\Big)\Big) \to 0 \tag{1.20}$$

Note that:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\tilde{z}_t \otimes e_t) - (\tilde{z}_t^* \otimes e_t^*) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{z}_t \otimes (e_t - e_t^*) + (\tilde{z}_t - \tilde{z}_t^*) \otimes e_t^*$$

From which:

$$\begin{aligned} & E(||\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{z}_t\otimes(e_t-e_t^*) \quad + \quad (\tilde{z}_t-\tilde{z}_t^*)\otimes e_t^*||_2^2|\sigma(Y)) \\ = & \frac{1}{\sqrt{T}}E(||\sum_{t=1}^{T}\tilde{z}_t\otimes(e_t-e_t^*) \quad + \quad (\tilde{z}_t-\tilde{z}_t^*)\otimes e_t^*||_2^2|\sigma(Y)) \\ \leq & \frac{1}{\sqrt{T}}\sum_{t=1}^{T}E(||\tilde{z}_t\otimes(e_t-e_t^*)||_2^2|\sigma(Y)) \quad + \quad E(||(\tilde{z}_t-\tilde{z}_t^*)\otimes e_t^*||_2^2|\sigma(Y)) \\ \leq & \frac{N}{\sqrt{T}}\sum_{t=1}^{T}E(||\tilde{z}_t||_2^2|\sigma(Y))E(||(e_t-e_t^*)||_2^2|\sigma(Y)) \quad + \quad E(||(\tilde{z}_t-\tilde{z}_t^*)||_2^2|\sigma(Y))E(||e_t^*||_2^2|\sigma(Y)) \end{aligned}$$

Now we have $E(||(e_t-e_t^*)||_2^2|\sigma(Y)) = O_p(1/T)$ and $E(||(\tilde{z}_t-\tilde{z}_t^*)||_2^2|\sigma(Y)) = O_p(1/T)$ from Lemma 1.5.3 and Lemma 1.5.4. It then follows that:

$$d_2\Big(\frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}\tilde{z}_t\otimes e_t\Big), \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}\tilde{z}_t^*\otimes e_t^*\Big)\Big)\to 0$$

Now, for the second expression:

$$\begin{aligned} E(||\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(z_t-\tilde{z}_t)\otimes e_t||_2^2|\sigma(Y)) &= \frac{1}{\sqrt{T}}E(||\sum_{t=1}^{T}(z_t-\tilde{z}_t)\otimes e_t||_2^2|\sigma(Y)) \\ &\leq \frac{N}{\sqrt{T}}\sum_{t=1}^{T}E(||(z_t-\tilde{z}_t)||_2^2|\sigma(Y)))E(||e_t||_2^2|\sigma(Y)) \end{aligned}$$

since $E(||(z_t-\tilde{z}_t)||_2^2|\sigma(Y))) = O_p(1/T)$ by Lemma 1.5.4 and $E(||e_t||_2^2)$ is finite by assumption. This implies that:

$$d_2\Big(\frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}z_t\otimes e_t\Big), \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T}\tilde{z}_t\otimes e_t\Big)\Big)\to 0$$

Finally, for the third expression:

$$\begin{aligned} E(||\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\tilde{z}_t^*-z_t^*)\otimes e_t^*||_2^2|\sigma(Y)) &= \frac{1}{\sqrt{T}}E(||\sum_{t=1}^{T}(\tilde{z}_t^*-z_t^*)\otimes e_t^*||_2^2|\sigma(Y)) \\ &\leq \frac{N}{\sqrt{T}}\sum_{t=1}^{T}E(||(\tilde{z}_t^*-z_t^*)||_2^2|\sigma(Y)))E(||e_t^*||_2^2|\sigma(Y)) \end{aligned}$$

since $E(||(\tilde{z}_t^*-z_t^*)||_2^2|\sigma(Y))) = O_p(1/T)$ by Lemma 1.5.4 and $E(||e_t^*||_2^2)$ is finite by assumption and by Lemma 1.5.3. This implies that:

$$d_2\Big(\frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T} z_t^* \otimes e_t^*\Big), \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T} \tilde{z}_t^* \otimes e_t^*\Big)\Big) \to 0$$

□

Note that weak convergence implies that $\frac{1}{\sqrt{T}} X^{*\prime} e^* \xrightarrow{L} \frac{1}{\sqrt{T}} X^{\prime} e$, so that the scaled sample bootstrap moment condition converges in distribution to the distribution of the sample moment condition of the population variables.

Proof (Proposition 1.2.3).

This proof follows the consistency of the bootstrap proof in Chatterjee and Lahiri (2011), with some modifications for time-series estimation and bootstrap, for adaptive elastic net instead of adaptive LASSO, and for multivariate estimation. Let $A_T \in \mathcal{F}$ denote the event that the adaptive elastic net correctly identifies the sparsity pattern. That is, $A_T = \{\omega \in \Omega : \hat{\mathcal{A}}(\omega) = \mathcal{A}^o\}$. Furthermore, consider the difference in minimization problems for the bootstrap naive adaptive elastic net:

$$V_T^*(u) = \frac{u' X^{*\prime} X^* u}{T} - \frac{2u' X^{*\prime} e^*}{\sqrt{T}} + \lambda_1 \sum_{i=1}^{N^2 p} |\hat{\beta}_{I,i}^*|^{-\gamma} (|\hat{\beta}_i + \frac{u}{\sqrt{T}}| - |\hat{\beta}_i|) + \lambda_2 \sum_{i=1}^{N^2 p} 2\frac{\hat{\beta}_i u_i}{\sqrt{T}} + \frac{u_i^2}{T}$$

such that:

$$\sqrt{T}(\hat{\beta}_N^* - \hat{\beta}) = \text{argmin}_{u \in \mathbb{R}^{N^2 p}} V_T^*(u)$$

where $\hat{\beta}_N^*$ is the naive adaptive elastic net estimator under the residual bootstrap. Since $Pr(A_T) \to 1$, there is a subsequence $\{T_k\}$ for which $P(\text{limsup}_{k\to\infty} A_{T_k}^C) = 0$. Now let $\tilde{\omega}_0^C \in \mathcal{F}$ be the union of the set $\text{limsup}_{k\to\infty} A_{T_k}^C$ and the set where Lemma 1.5.2 and Lemma 1.5.4 fail to hold, so that $Pr(\tilde{\omega}_0) = 1$. Fixing $\omega \in \tilde{\omega}_0$, then there exists T_ω so that for all $T \geq T_\omega$, $\hat{\mathcal{A}} = \mathcal{A}^o$. Note that on $\tilde{\omega}_0$:

$$\mathcal{L}(V_{T_k}^*(u)|\sigma(Y)) \to \mathcal{L}(V(u)) \tag{1.21}$$

using Lemma 1.5.4, Lemma 1.5.2, and arguments akin to those in the proof of Proposition 1.2.2. Then on $\tilde{\omega}_0$, weak convergence of $\mathcal{L}(\sqrt{T}(\beta_{T_k}^* - \hat{\beta})|\sigma(Y))$ to $G_\infty(\cdot)$ follows from the argmin theorem of Knight and Fu (2000). □

Proof (Corollary 1.2.4).

The distribution $G_\infty(\cdot)$ has a continuous distribution on $\mathbb{R}$ so long as at least one of the correct variables is included (otherwise the distribution is degenerate), so the corollary follows from Proposition 1.2.3. □

Proof (Corollary 1.2.5).

(a) This follows from the convergence in probability of the adaptive elastic net estimator and the convergence in probability of the covariance matrix estimator.

(b) This follows by applying Proposition 7.4 in Hamilton (1994) to the non-orthoganlized impulse responses, and then applying Slutsky's Theorem to obtain the distributional result for the orthogonalized impulse responses.

(c) As in the above, the limiting distribution is continuous as long as the impulses are functions of non-zero parameters. □

Appendix 1.B: Tables and Figures

Short Desc.	VARs Trans.	DFM Trans.	Small	Medium	Large
GDP	log-diff	log-diff	x	x	x
GDP Defl	log-diff	diff-log-diff	x	x	x
FedFunds	diff	diff	x	x	x
Consumption	log-diff	log-diff		x	x
EmpHrs:nfb	log-diff	log-diff		x	x
Emp:Services	log-diff	log-diff			x
CPH:NFB	log-diff	log-diff		x	x
S&P 500	log-diff	log-diff			x
m2	log-diff	log-diff			x
Ex rate: major	log-diff	log-diff			x
Investment	log-diff	log-diff		x	
FixedInv:NonRes	log-diff	log-diff			x
FixedInv:Res	log-diff	log-diff			x
PCE Def	log-diff	log-diff-diff			x
IP: Total index	log-diff	log-diff-diff			x
GDPI Defl	log-diff	log-diff-diff			x
Emp:Nonfarm	log-diff	log-diff			x
PPI	log-diff	log-diff-diff			x
CPI	log-diff	log-diff-diff			x
TB-1YR	diff	diff			x
TB-10YR	diff	diff			x
Capu Tot	raw	raw			x
Cons. Expectations	raw	raw			x
Emp:Total (HH Survey)	log-diff	log-diff			x
Unemp Rate	raw	raw			x
Price:Oil	log-diff	log-diff-diff			x

Table 1.7: Data series and their transformations. The Short Desc. column denotes the mnemonic for a particular data series. The VARs transformation denotes the transformation used in estimating all of the VARs, including the OLS, Bayesian VAR, and penalized likelihood. The DFM transformation denotes the transformation as in Stock and Watson (2002). All variables are expressed in annualized form. The other columns have 'x's marked where a variable is included in the small, medium, and large VARs.

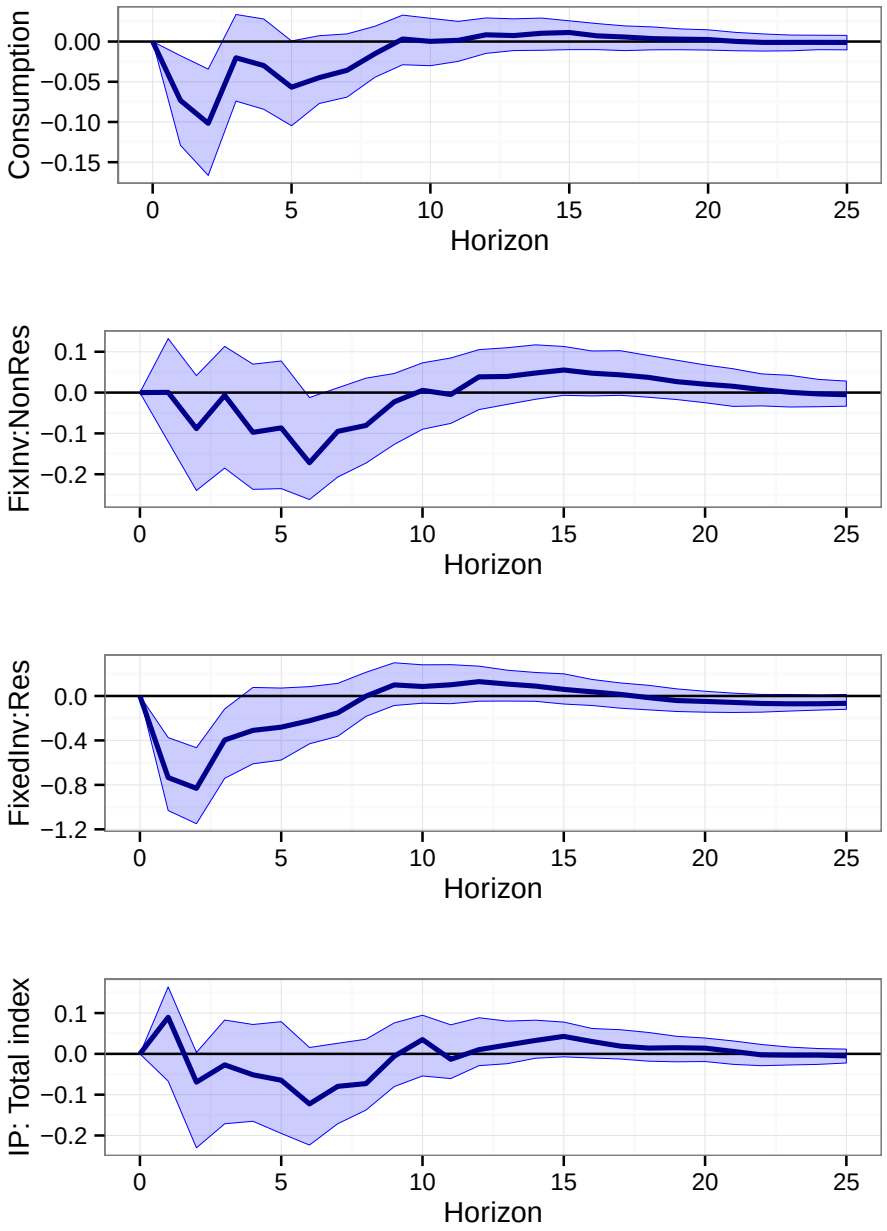


Figure 1.4: Adaptive elastic net estimated impulse response functions and their associated 90% confidence intervals for the effect of a one standard-deviation monetary policy shock to the FedFunds rate on Consumption, Non-Residential Investment, Residential Investment, and Industrial Production after 0-25 quarters. All series appear to decline.

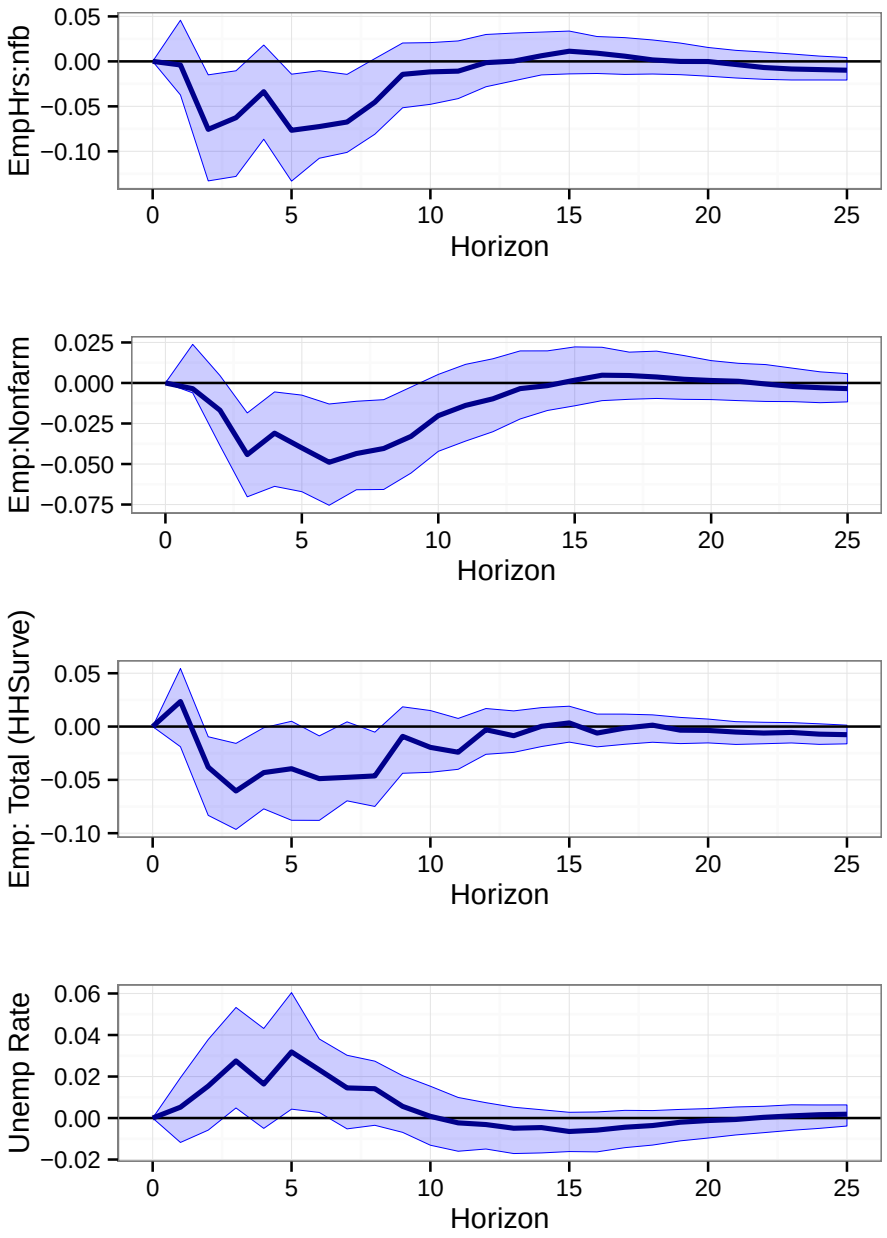


Figure 1.5: Adaptive elastic net estimated impulse response functions and their associated 90% confidence intervals for the effect of a one standard-deviation monetary policy shock to the FedFunds rate on hours worked, employment as measured by nonfarm payrolls, employment as measured by survey data, and the unemployment rate after 0-25 quarters. Employment is seen to decline, unemployment rises, and hours worked are reduced.

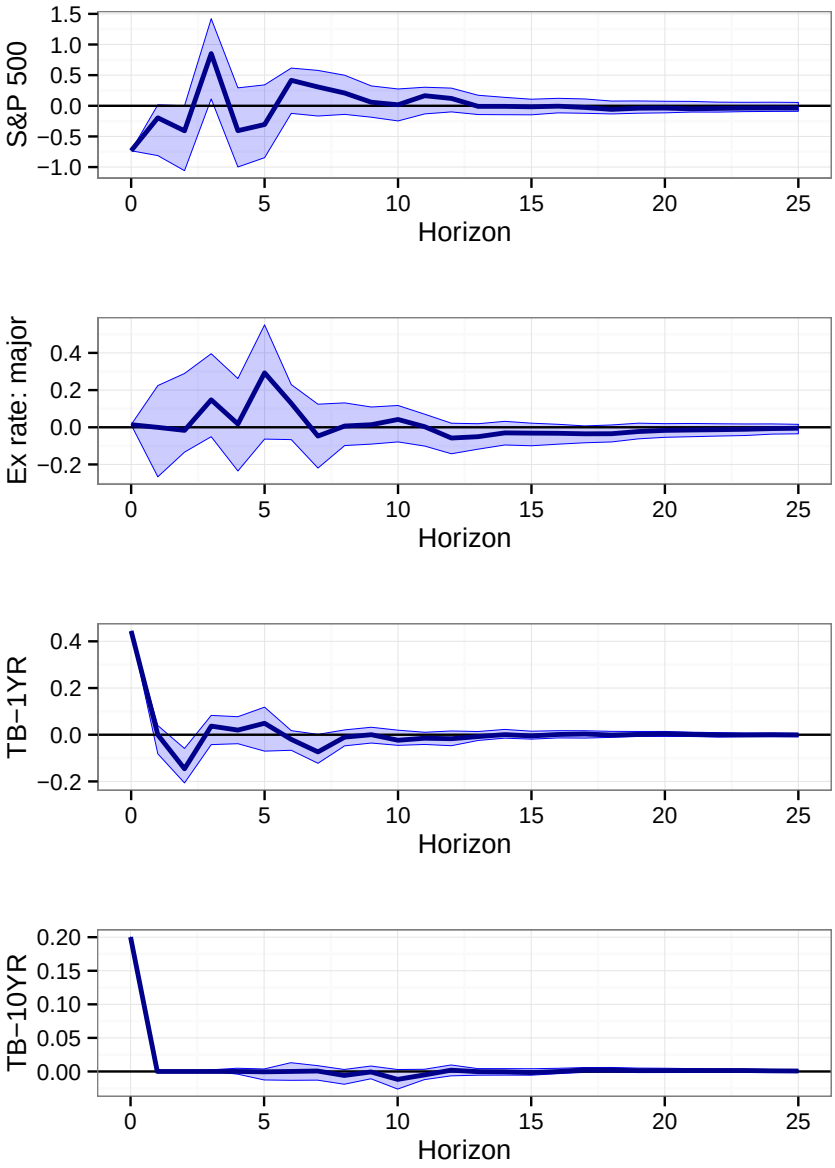


Figure 1.6: Adaptive elastic net estimated impulse response functions and their associated 90% confidence intervals for the effect of a one standard-deviation monetary policy shock to the FedFunds rate on the S&P 500 Index, the exchange rate, and the 1-year and 10-year maturity T-bill rates after 0-25 quarters. Stocks initially drop, then oscillate around zero reflecting reduced demand for stocks and negative economic sentiment. The exchange rate appreciates slightly to keep up with interest rate parity. Shorter maturity bonds react more harshly to monetary shocks, and both bond rates decay after an initial spike.

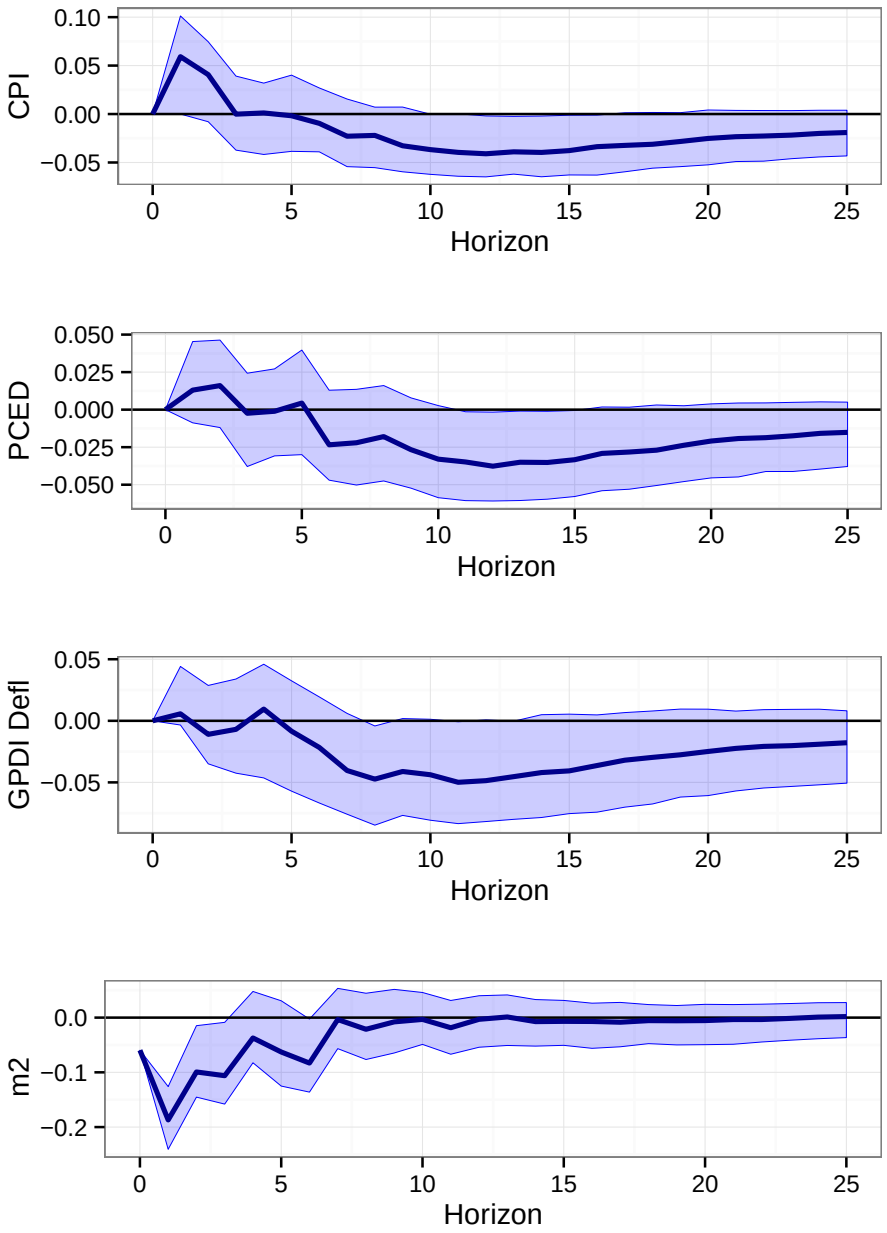


Figure 1.7: Adaptive elastic net estimated impulse response functions and their associated 90% confidence intervals for the effect of a one standard-deviation monetary policy shock to the FedFunds rate on CPI, consumer expenditures, GDPI deflator, and the money supply as measured by M2 after 0-25 quarters. The inflation indices are seen to decline after 5 quarters, as is frequently observed in VAR analysis and termed the 'price puzzle'. The money supply falls as expected.

Chapter 2

Wide Volatility Spillover Networks

Abstract. We introduce a conditional variance model that combines persistent volatility dynamics with spillovers from a wide cross-section of assets. We use elastic net estimation on a large, restricted VAR of realized measures to model these volatility dynamics. We show that despite the many parameters resulting from this wide cross-section, this spillover autoregressive (SPAR) realized variance model forecasts accurately and can be used in estimating large volatility spillover networks. These volatility spillover networks manifest and visually depict Granger-causal contagion of risk between a group of assets. We apply this model to identify systemic risk across U.S. financial institutions over the period 2001-2010, and show that our model identifies systemic risk buildup over the 2007-2008 financial crisis.
Keywords: Elastic Net, Systemic Risk, Realized Variance, Volatility
JEL: C53, C32, C51

We introduce a predictive volatility model that allows for volatility spillovers from a wide cross-section of assets. Estimating and forecasting the volatility of asset returns is paramount in understanding risk, and is crucial for option pricing, portfolio construction, and risk management. Traditionally, time-varying volatility is modeled using daily returns in univariate models, such as the GARCH model of Bollerslev (1986). In part due to the curse of dimensionality, many models of volatility dynamics are largely[1] confined to univariate time-series which either severely limit or altogether prohibit interday spillovers from one security to another. In recent years, the use of ultra-high-frequency returns has allowed for more accurate measures of intraday, open-to-close volatility as well as interday, close-to-close volatility dynamics. In particular, Andersen et al. (2003) thoroughly review the non-parametric realized variance measure of intraday variance and Corsi (2009) introduced the Heterogeneous Autoregressive (HAR) parametric model to describe the interday evolution of volatility over time. Like their predecessors, these more

[1] A notable exception is the family of multivariate GARCH models, reviewed for instance in Bauwens et al. (2006), which focus more on covariance dynamics rather than volatility. Because covariances involve considerably more parameters to estimate, these models are stringently limited by the curse of dimensionality.

recent models of the interday evolution of volatility largely[2] consider univariate or limited multivariate interday volatility dynamics.

We introduce an interday volatility spillover model that exploits a wide, multivariate cross-section of assets as well as high-frequency realized variance measures to model and forecast volatility. This model is formulated as a large vector autoregression (VAR) with linear restrictions similar to those in the HAR model, thus mimicking the long-memory features of volatility discussed in Corsi (2009). The use of VARs to model realized variance dynamics is not novel (Andersen et al., 2003), but in the past the cross-sectional dimension of these VARs has been limited by the curse of dimensionality. As an alternative, Veredas and Luciani (2012) estimate long-memory factor models of cross-sections of volatilities, but their method abandons the Granger-causal VAR structure. By contrast, the spillover model introduced here purposefully exploits a wide cross-section of assets as volatility predictors in a causal VAR. When estimated with elastic net regression (Zou and Hastie, 2005), this *SP*illover *A*uto*R*egressive (SPAR) realized variance model overcomes the curse of dimensionality by penalizing the parameters corresponding to the spillover effects. The elastic net penalty is chosen to optimize predictive ability, and thus curbs the parameter estimation variance of the spillover parameters to increase predictive power. We show that the resulting SPAR realized variance model significantly outperforms the univariate HAR model benchmark in out-of-sample forecast ability. The difference in forecast performance is particularly sharp during turbulent episodes, when an accurate model is needed most.

In addition to its success in out-of-sample forecasting, elastic net estimation in the SPAR model also allows for in-sample visualization of causal volatility spillover. We obtain spillover effect parameter-estimates using the predictive selection property of the elastic net. Of the wide pool of asset volatilities used as explanatory variables, the SPAR model selects parameters for the subset most relevant for prediction. This selection property reveals a directed, weighted network of volatility spillover. Each node in the network corresponds to an asset volatility, and each directed, weighted edge corresponds to the lagged marginal effect of one asset's volatility onto another asset's volatility. When this network is highly connected, a shock to the volatility of one asset permeates throughout the network, and spillover is systemic rather than idiosyncratic. This spillover network implied by the SPAR model hence allows for a visualization of contagion and systemic risk. Moreover, the elastic net encourages grouping effects and factor structures among correlated predictors, allowing for several weak predictors to be 'systemic as a herd'. Thus, elastic net estimation of the SPAR model allows for systemic risk to arise out of weak predictors and strong predictors, without arbitrary exclusions of several small but important systemic risk drivers.

[2]Notable multivariate exceptions are Andersen et al. (2003) and Noureldin et al. (2012). Both of these models are again heavily limited by the curse of dimensionality in the included number of assets.

The out-of-sample success of the SPAR model over turbulent periods suggests that the corresponding spillover networks are particularly important and meaningful over the 2007-2009 financial crisis. We estimate the SPAR model for the volatilities of equities of U.S. financial institutions over this period and analyze their structural features. In particular, the network estimated from the SPAR model allows for a convenient visualization of the spread of risk from one institution to another. We analyze the micro, firm-level spillover patterns as well as the macro, network-level systemic dynamics. This provides us with several perspectives on contagion, all of which identify considerable systemic behavior from early 2007 through 2009. This approach to systemic risk identification explicitly addresses interactions between financial assets, and provides a directed, Granger-causal interpretation of the flow of risk between financial institutions. The volatility-spillover-based systemic risk measure successfully identifies systemic risk as the outcome of network interactions and is shown to be particularly informative in 2007-2009.

The systemic risk indicators introduced here are grounded in an expanding literature on systemic risk identification and measurement. Some of these contributions, like the $CoVaR$ method of Adrian and Brunnermeier (2011), construct a non-structural risk measure, but rely on structural factors for systemic risk forecasting. Similarly, Hautsch et al. (2014) construct a network model that relies on structural macroeconomic and firm-specific factors. These models allow one to understand the sources of a particular financial institution's risk. The difficulty in their structural framework is that it relies on infrequent balance sheet data and a predefined set of risk factors which may not reflect all interactions occurring in the network of institutions. Moreover, these two papers rely on quantile regression estimates, which can result in crossing quantile-curves (Bondell et al., 2010) and instability when using different quantiles in the tails of the distribution. This poses a considerable methodological hurdle for these approaches, since they are precisely focused on these unstable tails.

By contrast, models like the MES of Acharya, Pederson, Phillipon, and Richardson (2010) and its econometric implementation by Brownlees and Engle (2011) rely on frequently observed non-structural factors. Brownlees and Engle (2011) posit that all financial institutions are related to the market return, as in a standard CAPM (Sharpe, 1964) framework, and estimate the expected loss for each financial institution if the market return suffers a major loss. That is, they define a systemic event as a market loss, and then forecast the result of such a loss on each financial institution. The market return is hence a non-structural factor that defines the systemic event.

A third approach most similar to the one pursued here is explored by Diebold and Yilmaz (2011). They do not assume a factor model, but instead build a VAR model of realized variances, representing the evolution of the risk network of financial institutions. A systemic event is not defined through the collapse of a factor: rather, Diebold and Yilmaz (2011) forecast the systemic risk of a particular financial institution by analyzing

its impulse response in the VAR. Financials whose volatility innovations lead to high impulse responses are seen as systemically risky. This model has the major advantage of describing a crisis event as the result of a process and not as its cause[3]. Despite this benefit, the model suffers from curse of dimensionality constraints in the VAR which limit the amount of financial institutions that can be modeled. Moreover, Diebold and Yilmaz (2011) specify a finite-order VAR process which does not account for long memory features frequently attributed to volatility. In this study, we also use volatility spillovers between institutions to arrive at a network-based systemic risk measure, but we account for the long-memory of volatility and we specifically exploit a wide cross-section of financial institutions.

Our discussion is also related to recent research on estimated dependence-networks between stock returns. In Barigozzi and Brownlees (2014), the authors suggest measuring non-directed networks of time-series using graphical LASSO and LASSO-VAR methods. Although Barigozzi and Brownlees (2014) use a very similar estimation method to our own, their method implies a network of non-causal dependence[4], whereas our data represent Granger-causal spillover. Billio et al. (2012) propose a Granger-causal network measure of return spillover more similar to the model pursued in our discussion. Their contribution differs from ours by addressing spillover in returns rather than spillover in volatility[5].

Our paper is divided into four sections which explore the topic of volatility spillover and systemic risk estimation. Section 3.1 introduces the SPAR realized variance model and discusses elastic net estimation of this wide, restricted VAR. Section 3.2 is a brief discussion of the realized kernel measure of realized variance and the general behavior of the inputs in our main model. Section 3.3 analyzes the SPAR model's performance over a decade with a focus on the 2007-2009 financial crisis. This SPAR model implies a Granger-causal interpretation of network connectedness, which turns out to be very useful in identifying periods of systemic risk. Section 3.4 uses the volatility spillover network revealed by the SPAR realized variance model to explore network-based systemic risk identification and build-up.

[3]Brownlees and Engle (2011) define the crisis event through a collapse in the market, and Adrian and Brunnermeier (2011) through a rise in leverage, illiquidity, and other fundamental factors that lead to systemic risk build-up.

[4]They estimate the HAC covariance matrix of serially correlated returns using LASSO and find that approximately 88% of the dependence in their network arises out of contemporaneous correlation rather than spillover. Hence, the dependence they account for is largely ascribed to correlation rather than spillover.

[5]Daily returns are very noisy, and it is therefore difficult to estimate and assess the significance of autoregressive and lagged spillover parameters from them. By contrast, volatilities are highly autocorrelated and predictable, suggesting that lagged parameters are more influential and more likely to be significant in a volatility model.

2.1 The SPAR Realized Variance Model

The introduction of realized variance measures in recent years has led to the use of these measures in forecasting and modeling interday volatility dynamics. A frequently used benchmark model in this category is the HAR realized variance model of Corsi (2009), which posits that variances[6] evolve according to a cascade-like structure, where the realized variance measure depends linearly on its one-day, one-week, and one-month lagged values.:

$$RM_t = \phi_0 + \phi_1 RM_{t-1} + \phi_2 \bar{RM}_{t-1:t-5} + \phi_3 \bar{RM}_{t-1:t-22} + \eta_t \tag{2.1}$$

Here, RM_t denotes the realized variance measure on day t, $\bar{RM}_{t-1:t-p}$ denotes the sample average of the realized variance measure over the last p-lags, and η_t is an idiosyncratic error term. This formulation is equivalent to a restricted autoregressive (AR) model of order 22 on realized variances, with restrictions that impose long-memory-mimicking behavior. This is a standard linear model which is usually estimated using least-squares, and its formulation can be justified both economically and statistically. The economic justification behind this model stems from an asset pricing model where three heterogeneous kinds of active portfolio managers exist: those who update their holdings daily, weekly, and monthly. This portfolio-rebalancing reasoning is used to justify these specific lags in the volatility evolution equation (2.1). From a statistical perspective, the weekly and monthly lags approximate the observed long-memory in volatility, while the one-day lag captures news effects on volatility that are important in short-term forecasts. Regardless of which interpretation one believes, Corsi (2009) shows that this model forecasts realized variance with reasonable accuracy, and hence the HAR is a hard-to-beat, common benchmark.

Though the HAR model captures a rich set of dynamics for one security with a parsimonious and easily estimable representation, it may be that other securities have a lagged, Granger-causal effect. In particular, news effects from one security may spill over to another security over the short-run. As an example, given the natural connectedness between financial institutions, it is reasonable to believe that the volatility of one institution's security may impact the volatility of another institution's security the following day. This may arise out of a delayed news transmission mechanism, such as uncertainty around one bank's abilities to repay its repo loans to its lenders the following day. We depict the difference in these modeling approaches in Figure 2.1.

Figure 2.1 shows two different volatility spillover networks on the left and right panels. The nodes in the graph represent individual securities (1, 2, and 3) and the directed arrows

[6]This model is frequently expressed in logs of realized measures in order to ensure positive variances as the final outcome of the forecast. Moreover, as this model is expressed in logs, the model parameters can be thought of as elasticities.

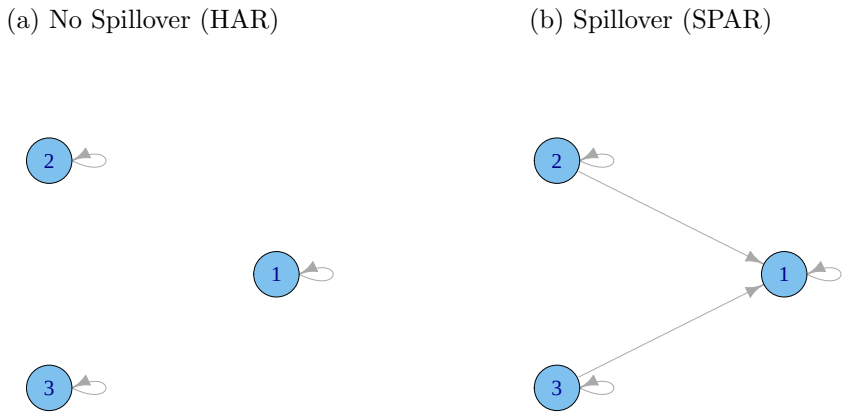


Figure 2.1: Left: model with no spillover such as the HAR model, and each security's volatility only affects itself. Right: spillover model, where there is volatility spillover from the volatility of security 2 and security 3 onto the volatility of security 1.

represent short-run spillovers from one security to another. The left panel represents standard univariate models like the HAR model, which allow for own-lag effects but no spillover effects. Here, arrows only point to the nodes from which they originate, hence excluding any spillover. In the right panel, each security affects itself, but there is additional spillover from securities 2 and 3 onto security 1. That is, volatilities in securities 2 and 3 have an additional marginal effect on the volatility of security 1 the following day, reflecting risk contagion. Such spillover effects may be important in interday volatility dynamics, particularly for securities that share common features.

To capture such spillover effects in an econometric model, we propose the Spillover Autoregressive (SPAR) realized variance model. The SPAR model captures the approximate long-memory and own-news effects of volatility in the HAR model, along with recent information about a related, wide cross-section of assets. Consider a vector of realized measures for K securities, $\mathbf{RM}_t = (RM_{1,t}, ..., RM_{K,t})'$. The SPAR model for the (log) realized measures is given by the restricted VAR(22):

$$\mathbf{RM}_t = \Phi_0 + (\Phi_1 + \beta)\mathbf{RM}_{t-1} + \Phi_2 \mathbf{R\bar{M}}_{t-1:t-5} + \Phi_3 \mathbf{R\bar{M}}_{t-1:t-22} + \eta_t \tag{2.2}$$

The notation in this restricted VAR is analogous to that of the HAR: Φ_0 is a $K \times 1$ vector of constants, whereas Φ_1, Φ_2, and Φ_3 are $K \times K$ diagonal matrices of parameters corresponding to own-lag and long-memory-mimicking effects. The SPAR model deviates from the HAR by introducing the β spillover-parameter matrix, which has free parameters except for zeros down the main diagonal. These spillover parameters allow for lagged volatilities of other assets $k \neq i$ to affect the volatility of security i. As in standard VARs and the HAR model, the SPAR model can be written as a standard linear model for each

security i. That is, we can write the equation-by-equation SPAR model as:

$$RM_{i,t} = \phi_{i,0} + \phi_{i,1} RM_{i,t-1} + \phi_{i,2} \bar{RM}_{i,t-1:t-5} + \phi_{i,3} \bar{RM}_{i,t-1:t-22} + \sum_{k \neq i} \beta_{i,k} RM_{k,t-1} + \eta_{i,t} \quad (2.3)$$

When the model is formulated in terms of logs of the realized measure, the coefficients $\beta_{i,k}$ represent the Granger-causal, marginal effect of a percentage increase in the realized variance of security k in period $t-1$ on the realized variance of security i in period t, in percentage terms.

In the past, formulations like (2.2) and (2.3) were limited because spillovers were assumed to be minor or altogether absent and the dimensionality of β could not be estimated efficiently. In much of the extant literature, the $k \neq i$ securities were assumed to have no marginal effect on security i, and hence their $\beta_{i,k}$ values were implicitly or explicitly assumed to be zero. In models where this was not the case, such as Andersen et al. (2003) and Diebold and Yilmaz (2011), least-squares estimation of the spillovers presented a curse of dimensionality challenge. If there were K assets being modeled, the number of spillover parameters grew at the rate of $O(K^2)$. The more assets K there were being modeled, the higher the variance of the estimated parameters. When K is moderately large and the history of each security is short, this quadratic order of growth can be very limiting. Hence, most models of volatility dynamics relied on univariate models, or low-dimensional multivariate models to predict volatility. In our approach, by contrast, we allow the data to dictate whether such restrictions are valid (that is, which spillovers are non-zero) while simultaneously handling the large dimensionality. We account for Granger non-causality and the curse of dimensionality in estimation, where we assume sparsity in the $\beta_{i,k}$ spillover-effect-motivated coefficients and apply a sparsity-inducing estimator.

To combat the curse of dimensionality and obtain a sparse parameter estimate geared toward optimal prediction, we impose an elastic net penalty on the spillover parameters. The elastic net is a positive and convex penalty on β, which imposes a cost to adding more β to the estimated model. Specifically, the SPAR model for security i estimated by the elastic net minimizes the objective function:

$$L_i = \sum_{t=1}^{T} \eta_{i,t}^2 + \lambda_i \sum_{k \neq i} (\alpha_i |\beta_{i,k}| + (1 - \alpha_i) \beta_{i,k}^2) \quad (2.4)$$

where λ_i is a positive constant and α_i is a constant between zero and one. Compared with the least-squares solution, the elastic net attenuates the parameters $\beta_{i.k}$ toward zero by adding-in a positive penalty term scaled by $\lambda_i > 0$ for each $\beta_{i,k}$. This penalty term is a linear combination of two penalties serving distinct purposes: an L_1 penalty for

shrinkage and sparse selection of $\beta_{i,k}$, and an L_2 penalty to maintain grouping effects between regressors.

To understand this optimization procedure and its advantages, we consider comparative statics on the λ_i and α_i penalty parameters. The parameter λ_i controls the degree of shrinkage and selection in the model, whereas the α_i parameter determines the trade-off between L_1 and L_2 norm penalties. When $\lambda_i = 0$, the penalty vanishes and the optimization problem (2.4) is identical to the least-squares problem.

Suppose then that $\lambda_i > 0$ and consider first the case where $\alpha_i = 1$, resulting in the LASSO (least absolute shrinkage and selection operator) penalty of Tibshirani (1996). Since $\lambda_i > 0$ the minimization function increases in the magnitude of each $\beta_{i,k}$, thus leading to solutions that shrink the magnitude of $\beta_{i,k}$ away from the least-squares solution toward zero. Moreover, LASSO results in a sparse solution arising out of the absolute value penalty. This penalty acts as a truncation threshold for $\beta_{i,k}$ values near zero, since in those cases the gradient of of the penalty may be greater than that of the sum of squared residuals. Specifically, when a spillover-effect variable k does not add enough predictive power in minimizing the sum of squared residuals relative to λ_i, its associated $\beta_{i,k}$ coefficient is set to zero. This is the predictive selection property of the LASSO, which shrinks the coefficients of all spillover-effect variables, and truncates the coefficients of variables with weak explanatory power. In the extreme case where λ_i is greater than some threshold, $\bar{\lambda}_i$, the cost of adding more $\beta_{i,k}$ to the model is greater than the benefit to reducing the sum of squared residuals for all k values of $\beta_{i,k}$, resulting in a no-spillover, univariate HAR model. As λ_i decreases toward zero, the penalty becomes less severe and variables are added incrementally in such a way as to optimally minimize the sum of squared residuals. When $\lambda_i = 0$ the LASSO solution collapses to least-squares. Hence, when varying λ_i, the LASSO part of the elastic net penalty selects a subset of the spillover variables and shrinks their associated coefficients for optimal prediction.

On the other hand, when $\alpha_i = 0$ the optimization problem excludes the L_1 penalty and becomes a ridge problem. Ridge shrinks the least-squares parameters but does not truncate them, resulting in a shrunk but non-sparse solution. This shrinkage in ridge regression is used for its grouping effect, where correlated explanatory variables are assigned coefficients of similar magnitude[7]. This makes the ridge portion of the penalty particularly appealing for the SPAR model, since volatilities of securities tend to rise simultaneously and may have correlated, 'herd-like' behavior, such as factor structures. Indeed, there is a close relationship between the grouping effect of ridge regression, which penalizes in a way that respects correlation structures (and hence common factors) and factor models estimated by principal components, such as the volatility model of Veredas and Luciani (2012). Fitted values from ridge regression are equivalent to those obtained

[7]For details, see Theorem 1 in Zou and Hastie (2005).

by shrunk regression of the dependent variable on the principal components of the regressors, where the regression coefficients are shrunk more aggressively for (less relevant) principal components associated with smaller eigenvalues [8]. Whereas in factor models these less important principal components are excluded entirely, ridge regression uses all factors, but penalizes the less relevant factors more severely than the more relevant factors. Hence, unlike factor models such as the CAPM, ridge regression embeds all factor information in estimation, while assigning the greatest weight to the most common factors, which contain the central group structures. Ridge penalization in the SPAR model thus extracts the more common features (or factors) between the volatilities and exploits them in predicting the dependent volatility series.

When $\alpha \in (0, 1)$ the elastic net borrows from both LASSO and ridge. Specifically, it leads to a sparse model where only spillovers that matter most for prediction are included. At the same time, the ridge component of the elastic net induces correlated spillover variables to have similar coefficients, resulting in less arbitrary selection among these variables and preservation of common features. Moreover, the ridge component of the penalty exploits these common features in similar manner to factor models. This makes the elastic net particularly effective in predictive volatility spillover modeling, where a sparse solution is desirable but many of the data series are highly correlated and share common features.

Since predictive ability is a key concern for our volatility model, we generally choose the tuning parameter λ_i by five-fold cross-validation. This parameter determines the degree of penalization in the model, and hence the degree of shrinkage and sparsity. Choice of parameter dictated by minimization of cross-validated mean-square-error results in a model that is robust to over-fitting and is predictively accurate. In a general setting, tuning the α_i parameter via cross-validation is also feasible, but two-dimensional cross-validation is computationally burdensome and in the sensitivity analyses below it is shown that choice of α_i does not significantly alter the resulting model. Hence, we set $\alpha_i = \alpha = 0.5$ for all of the models estimated below as a democratic compromise between ridge and LASSO.

We motivate this model with contagion modeling and systemic risk identification in the financial network. Here, the primary objective of the volatility model is to construct accurate volatility forecasts using the information contained in the network of financial institutions[9] and to reveal which institutions are spilling over the most risk. The network of financial institutions is changing over time, both in terms of interactions between institutions and in terms of the actual number of institutions. For instance, the characteristics of the individual volatilities and of the financial network are likely to have changed after

[8] For a more detailed treatment, see Hastie et al. (2009).

[9] That is, we exclude other factors that may be considered relevant in other contexts.

the collapse of Lehman Brothers. This type of network-changing event occurs quite frequently: it is not unusual for a new institution to be introduced, for two institutions to merge, or for institutions (particularly smaller ones) to collapse or fall out of the network. That is, the fundamental DGP describing the network of financial institution's volatilities is itself evolving, suggesting that modeling the network as one DGP is a fundamentally wrong exercise.

To resolve this time-varying DGP problem we acknowledge that a large evolving DGP is difficult to estimate and we seek a robust estimate of a local DGP by estimating a family of one-year rolling models.[10] For any rolling model there is a window-length trade-off between adequately long windows to contain sufficient degrees of freedom, and reasonably short windows to capture evolution in the DGP. We consider a one-year horizon window where lagged variables are included only for financial institutions that are contained in the network. One year was chosen because it is long enough to add a considerable amount of degrees of freedom while being sufficiently short to allow for changes in the DGP. In the current setting, the window-length remains fixed in every estimation for computational ease and consistency, though this could be extended for longer-term forecasts. At the end of our analysis, we show that the window-length does not dramatically change the aggregate results.

By design, the shrinkage and selection properties of elastic net estimators are ideal in this forecasting setting where there are many correlated explanatory variables, only a few are significant, and prediction is a primary concern. The major drawback of using the elastic net is that it is engineered for prediction and not for statistical explanations of outcomes: that is, the elastic net does not select or shrink coefficients based on their statistical significance, but on their ability to explain the outcome in a predictive setting. Other model selection approaches, such as the Autometrics algorithm implemented by Doornik (2009), focus on statistical significance of regressors, encompassing in a local DGP, and model congruency. Though the elastic net is not based on tests of statistical significance, it is also not plagued by their frequent low-power and type I error. Moreover, the course of causation is generally clear when dealing with lagged volatilities, so mistakes related to causality which may come up when using the elastic net in other settings are less relevant here[11]. Because prediction and ease of computation are the primary concerns in this volatility spillover framework, the elastic net technique is adopted here.

Finally, due to the convexity of the elastic net minimization problem, estimation of the SPAR model is computationally efficient. In the computations performed next, both the estimation and the cross-validation are performed at the same complexity order as solving the least-squares problem. The sample used here contains approximately 80 volatilities

[10]This is not the only way to deal with this problem: there are alternatives such as locally weighted models with non-zero weights (here we effectively use a locally weighted SPAR with zero weights for observations outside of the rolling window).

[11]Unlike graphical-LASSO applications, for instance.

in every cross-section with approximately 2,000 daily estimations for each series. This translates to about 160,000 elastic net estimations, with cross-validation computed each time. In total, this estimation was completed at almost the same speed as a rolling univariate GARCH model, while handling a much larger multivariate dataset.

2.2 Realized Volatility of U.S. Financial Institutions

Having introduced our volatility spillover model, we now turn to its primary inputs. These are realized variance measures of intraday, open-to-close variation in the price process. In recent years there has been a growing literature on the use of these non-parametric, sample-variance type estimators to estimate the quadratic variation of a price process using high-frequency trading data. Unlike earlier variance models which estimated variance using daily returns, realized variance measures quadratic variation: that is, price variation *within* the day. In estimating quadratic variation of the price process, we need to handle the market microstructure noise arising from bid/ask bounce, liquidity effects, and misreporting of trades. The realized kernel class of estimators introduced by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) estimate the variation in the efficient price process: that is, the price process net of market-microstructure noise. In the following we summarize the methodology used in estimating realized kernels as presented in Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009).

2.2 Realized Kernel

Consider the continuous price process Y_s over a sample interval $[0, T]$, where T is the end of the trading day. This process is assumed to be a Brownian semi-martingale plus jump process:

$$Y_t = \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + J_t \tag{2.5}$$

where $J_t = \sum_{i=1}^{N_t} C_i$ is a jump process having a finite number of jumps in any bounded interval of time, and $N_t < \infty$ counts the number of jumps in the interval $[0, t]$. It is assumed that μ_s is a predictable, locally bounded drift process, W_s is a Brownian motion, and σ_s is a cadlag volatility process.

The desired variance quantity is the quadratic variation of Y_s over $[0, T]$:

$$[Y_T] = \int_0^T \sigma_s^2 ds + \sum_{i=1}^{N_T} C_i^2 \tag{2.6}$$

Unfortunately, instead of observing Y_s directly, one observes a related set of recorded noisy observations:

$$X_{\tau_0}, ..., X_{\tau_n}, \quad 0 = \tau_0 < \tau_1 < ... < \tau_n = T$$

where X_{τ_k} is additively related to Y_{τ_k} through the noise term U_{τ_k}:

$$X_{\tau_k} = Y_{\tau_k} + U_{\tau_k}$$

The noise term is assumed to be mean zero with finite variance ω. The existence[12] of this noise term implies that the use of a naive estimator using X_{τ_k} to estimate quadratic variation results in a biased estimate that includes the unknown ω. For this reason we choose to model quadratic variation using the realized kernel, a positive, HAC estimator:

$$K(X) = \sum_{h=-H}^{H} k\left(\frac{h}{H+1}\right)\gamma_h, \qquad \gamma_h = \sum_{j=H+1}^{n} r_j r_{j-h} \tag{2.7}$$

where each r_k is the recorded return calculated between times τ_{k-1} and τ_k, and $k(x)$ is a kernel-weight function, in this case the non-flat Parzen kernel[13] given by:

$$k(x) = \begin{cases} 1 - 6x^2 + 6x^3 & 0 \leq x \leq 1/2 \\ 2(1-x)^3 & 1/2 \leq x \leq 1 \\ 0 & x > 1 \end{cases}$$

This weight function smoothes the return covariances, giving an estimate of quadratic variation that is robust to micro-structure noise. Specifically, the realized kernel estimator is a consistent estimator for the quadratic variation if $K(U) \xrightarrow{p} 0$ and $K(Y) \xrightarrow{p} [Y]$. For these reasons, the realized kernel is a very attractive estimator of quadratic variation.

Estimation of the realized kernel requires a practical estimator of the bandwidth. To eliminate the bias and variance of the noise (and thus achieve consistency) requires H of at least order $n^{1/2}$. Barndorff-Nielsen et al. (2009) chose the bandwidth given by $H = c^* \xi^{4/5} n^{3/5}$, where c^* and ξ are defined by the Parzen kernel and by subsampled realized variance, respectively. In the empirical analysis used here the same bandwidth is chosen.

Accurate estimation of realized kernels requires careful cleaning of the data. High-frequency trading data contains mis-recorded trades, time delays from exchange to exchange, and other potential compromising sources of data inaccuracy. To obtain an estimator that represents the data optimally, inaccurate observations are given a lower weight, in this case, zero. Hansen and Lunde (2006) show that ignoring a large number

[12] See for instance, Hansen and Lunde (2006).

[13] The Bartlett kernel more frequently encountered with HAC estimators leads to a slower rate of convergence, so the non-flat Parzen kernel is preferred.

of observations can improve the accuracy of realized volatility estimators. For computational simplicity, we use only trade data and employ the cleaning rules suggested by Barndorff-Nielsen et al. (2009).

2.2 Data and Descriptive Statistics

To estimate the realized measures required for the SPAR volatility model, we use high-frequency traded prices. The data comes from the Trade and Quote (TAQ) database, which contains tick-by-tick prices for a large number of assets, including equities of many publicly traded financial institutions in the United States. We focus on prices of these institutions since we will use these later for systemic risk identification. The data cover September 1, 2000 to December 31, 2010, which includes the 2008 financial crisis and the 'dot-com bubble', but excludes earlier market dislocations. As the structure of the financial network is not assumed to be static, the loss of data on earlier market dislocations is not seen as a drawback. The panel contains data on 115 U.S. financial institutions and categorized as in Brownlees and Engle (2011) as broker-dealers, depositories, insurers, or additional financial services firms designated in an 'others' category.

We summarize the realized volatility series for each category over time. Table 2.1 shows descriptive statistics on the mean annualized realized volatility series for each financial category. Broker-Dealers and Others tend to exhibit higher average volatility than the other categories. This is consistent with the perception that during the highest period of volatility in the sample (the 2008 financial crisis), it was broker-dealers and real estate companies who sustained the larger losses.

	Broker-Dealers	Depositories	Insurers	Others
Bottom Quartile	16.93	13.99	15.80	17.27
Median	22.36	19.86	21.77	23.65
Mean	28.80	26.68	28.28	31.28
Top Quartile	32.90	30.71	31.66	35.42

Table 2.1: Mean annualized realized volatility statistics by category.

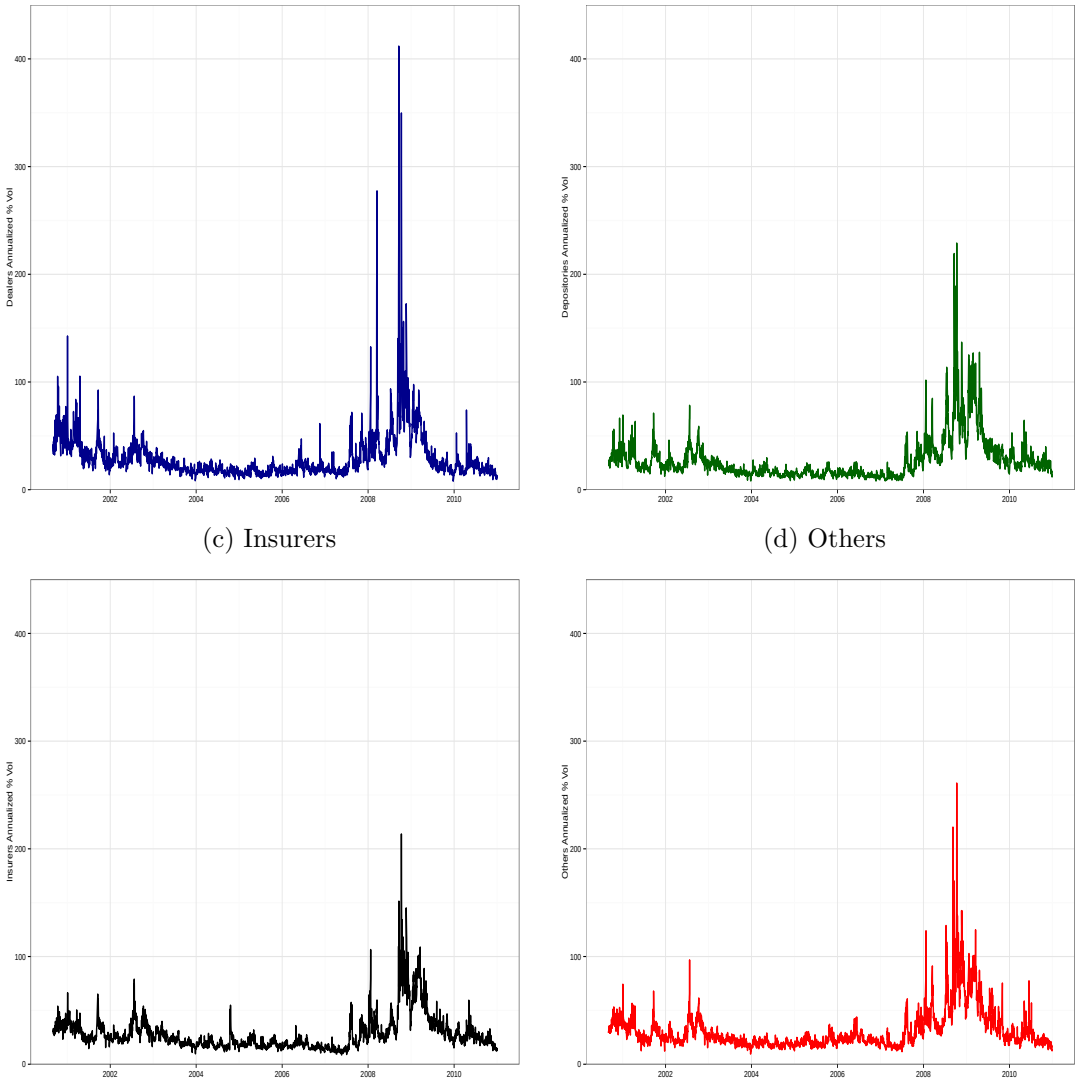
The plots in Figure 2.2 show the cross-sectional mean volatility over time by category[14]. These plots show that volatilities were generally low across the board following the dot-com crash and increased sharply in the third quarter of 2007 and into 2008. Volatilities were high for all groups in the sample in 2000 following the dot-com crash, and began to level off in early 2001. Volatilities remained generally low (with the exception of September 11, 2001) until the collapses of Enron and Worldcom toward the end of 2002. Following these smaller dislocations, volatilities remained low until the 2007-2008 sub-prime mortgage crisis, when volatilities rose and remained exceedingly high for a long

[14]In the remaining sections we keep these colors to distinguish between different types of institutions.

time. Figure 2.2 shows that although all financial institutions' volatilities rose in 2008, broker-dealers rose by one and a half times as much as the rest. Indeed, it was generally broker-dealers who were perceived to be the most systemically risky institutions.

(a) Broker-Dealers

(b) Depositories



(c) Insurers

(d) Others

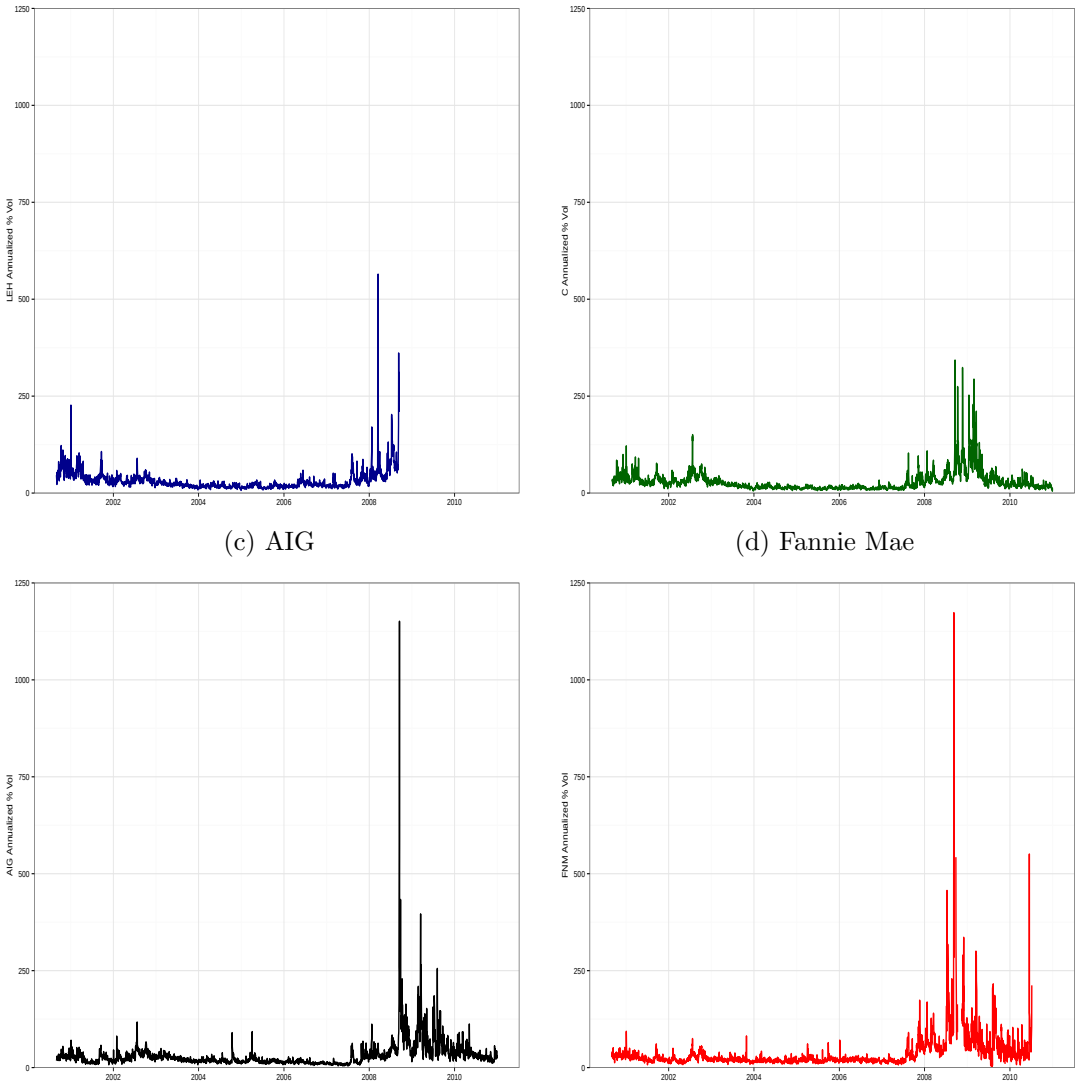
Figure 2.2: Mean annualized realized volatility by financial category over the period 2000-2010. Realized volatility series are grouped by Broker-Dealers, Depositories, Insurers, and Others.

The tremendous buildup in realized volatility seen at the category level is even more stark for particular stressed institutions. Figure 2.3 shows the volatility series for particularly troubled firms in every category: Lehman Brothers, Citigroup, AIG, and Fannie Mae. Realized volatility is generally stable and low for all of these series for the vast majority of the sample, with a buildup of volatility in mid 2007 which spiked at the Bear Stearns and Lehman Brothers failures. Note that these institutions, which were generally considered to be more risky due to large bailouts, lawsuits, failure, restructuring, or a

large degree of solvency concern had significantly higher volatility than their respective category averages. Thus, though volatility was generally high in 2008, high volatilities were higher-than-average for those firms that were particularly deep in trouble. Volatility of financial institutions therefore embodies market dislocations as well as capital losses at the level of the firm. In the next section, we show that we can forecast volatility more accurately for such institutions using a model that builds-in potential systemic behavior by allowing for a wide cross-section of explanatory variables.

(a) Lehman Brothers

(b) Citigroup



(c) AIG

(d) Fannie Mae

Figure 2.3: Annualized realized volatility for an institution in each financial category over the period 2000-2010: Lehman Brothers (Broker-Dealer), Citigroup (Depositories) , AIG (Insurers), and Fannie Mae (Others).

2.3 SPAR Model Results: Out-Of-Sample Performance

To assess the relative importance of spillover effects in the SPAR model, we compare the predictive ability of this model with that of the no-spillover HAR model. To this end, we estimate both models with a 1-year rolling window for the data set of U.S. financial institutions described above. In particular, we are interested in the relative predictive ability of the SPAR over the HAR model over stressful periods, such as the 2007-2008 financial crisis. If the SPAR model outperforms the HAR model, the addition of volatility spillover (as captured by the elastic net penalty) yields significant gains in volatility forecasting on top of the simple addition of high-frequency measures.

We first consider a pseudo-out-of-sample comparison by analyzing the penalization magnitude parameters λ_i across the cross-section and over time. Recall that this parameter reflects the degree of shrinkage and selection as chosen by minimization of cross-validated mean-square-error. If λ_i is low across the cross-section of K securities, cross-validation indicates that spillovers are important for the predictive ability of the model for all securities. Conversely, when λ_i is high across the cross-section, cross-validation indicates that the HAR model is a better fit. When there is considerably variation in the cross-section of λ_i parameters, contagious models fit better for some series and univariate models for others.

Figure 2.4 shows a fan plot of the cross-sectional empirical density of the λ_i penalization parameters at each point in time. There, dark colors (red) show the center of the distribution and light colors (yellow) show the tails. First, we notice a large fall in nearly all tuning parameters over 2002-2003, suggesting that spillovers were important over this time-period. The tuning parameters revert to higher levels in 2004 until mid-2007, when an abrupt change occurs and all λ_i's shift down. Note that over this period the distribution of λ_i has a higher variance and a higher median, suggesting that spillovers are less informative. In mid-2007 the distribution falls and becomes tighter around the median, indicating that spillovers have become more informative across the cross-section of financial institutions. This behavior persists during the financial crisis, with parameters only slowly beginning to rise over 2010. These results indicate that spillovers were particularly useful for volatility prediction over the financial crisis, and this behavior was uniform across the set of financial institutions.

Next, we construct a pure out-of-sample, Diebold and Mariano type (Diebold and Mariano, 1995) of comparison to determine how the SPAR model performs against the HAR model. This approach follows the Sheppard and Shephard (2010), Vuong (1989), and Rivers and Vuong (2002) implementation of ideas presented in Cox (1961). This is done by comparing Quasi-Likelihood (QLIK) loss functions for variance forecasts. The H-period ahead QLIK loss function is given by:

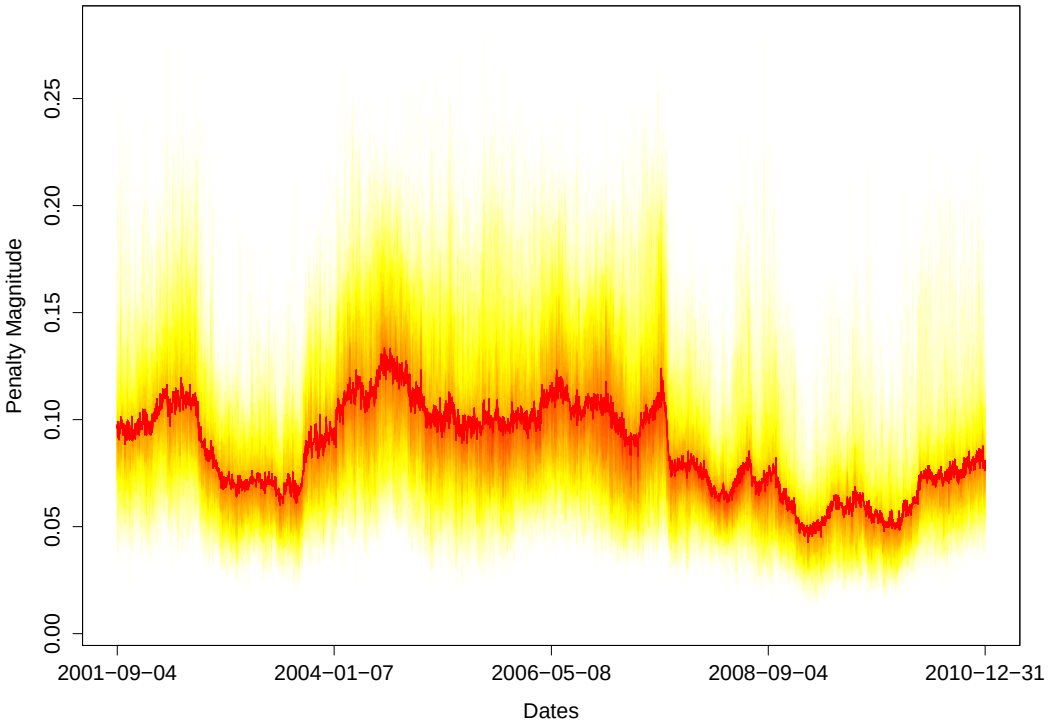


Figure 2.4: The cross-sectional empirical density of the penalty magnitude parameters λ_i across the set of all financial institutions over time as a fan plot. Dark red colors indicate a greater probability mass, whereas lighter yellows indicate less probability mass.

$$Loss(RK_{t+h}, \hat{\sigma}_{t+h|t-1}) = \frac{RK_{t+h}}{\hat{\sigma}^2_{t+h|t-1}} - \log\left(\frac{RK_{t+h}}{\hat{\sigma}^2_{t+h|t-1}}\right) - 1, \quad h = 1, ..., H \tag{2.8}$$

where RK_{t+h} is the realized kernel estimate at period $t+h$ and $\hat{\sigma}_{t+h|t-1}$ is the volatility forecast constructed in period $t-1$. Patton (2011) and Patton and Sheppard (2009) have shown this loss function to be robust to certain types of noise in the variance estimate. To compare the forecasting ability of two models, it is natural to compare their loss-difference:

$$\begin{aligned}
\delta_{t,h} &= Loss(RK_{t+h}, \hat{\sigma}_{t+h|t-1}) - Loss(RK_{t+h}, \hat{\hat{\sigma}}_{t+h|t-1}), \\
&= \left\{\frac{RK_{t+h}}{\hat{\sigma}^2_{t+h|t-1}} + \log(\hat{\sigma}^2_{t+h|t-1})\right\} - \left\{\frac{RK_{t+h}}{\hat{\hat{\sigma}}^2_{t+h|t-1}} + \log(\hat{\hat{\sigma}}^2_{t+h|t-1})\right\} \\
&= -2\log\left\{\frac{f(RK_{t+h}|1, \hat{\sigma}^2_{t+h|t-1})}{f(RK_{t+h}|1, \hat{\hat{\sigma}}^2_{t+h|t-1})}\right\}
\end{aligned}$$

where $\hat{\sigma}_{t+h|t-1}$ denotes the SPAR and $\hat{\hat{\sigma}}_{t+h|t-1}$ the HAR volatility forecasts. The function

$f(x|k,\theta)$ is a Gamma density function with parameters k and θ evaluated at x, so that the loss-difference represents a likelihood-ratio type statistic.

To statistically assess the differences in loss, consider the $h+1$-step ahead average difference in loss:

$$\hat{\delta}_h = \frac{1}{T-h} \sum_{t=h+1}^{T} \delta_{t,h}, \tag{2.9}$$

This is an estimator of the expected difference in loss for the h-step ahead forecasts, $\delta_h = E(\delta_{t,h})$. Note that the latter would be the unconditional average likelihood ratio between the two models if the SPAR forecast were to be cast into a likelihood framework. When the average difference in loss is negative ($\hat{\delta}_h < 0$) it suggests that the true unconditional expected log-likelihood ratio is negative and therefore that the SPAR model would be favored. Conversely, when $\hat{\delta}_h > 0$, it suggests that the alternative model is favored. Further, the hypothesis $\delta_h = 0$ (the two forecasting models yield equal average loss) can be tested formally using the central limit theorem:

$$\sqrt{T}(\hat{\delta}_h - \delta_h) \sim N(0, V)$$

where the variance V can be estimated using a HAC estimator. The tests of interest are:

$$H_0 : \ \delta_h = 0, \ \text{versus } H_1^A : \ \delta_h < 0 \text{ or } H_1^B : \ \delta_h > 0$$

which we can conduct with a usual t-statistic. Recall that our data contains a wide panel of volatility series over 2001-2010. We proceed using only the volatility series that are available over the entire sample (there are thirty-six such series) and for these series, comparing the one-step-ahead forecast from the 1-year HAR and the 1-year SPAR realized variance models. For the SPAR models, we allow the maximal size of the network to change over time, depending on which realized variance series are available at any given time.

The figures below present results on the test-statistics for these financial institutions' volatility series. Figure 2.5 shows the empirical distribution function of thirty-six one-step ahead ($h = 1$) average loss-difference test-statistics for the entire 2001-2010 out-of-sample time-period. The difference in loss is negative for all thirty-six series but one, indicating that the average loss is considerably lower for the SPAR model. For approximately 80% of the series the hypothesis that the losses are equal or greater than zero is rejected at the 95% confidence level, lending strong evidence in favor of the SPAR variance model[15]. Moreover, none of the series significantly favor the HAR model. The strong negative trend

[15]Note that there is some cross-sectional dependence between these loss-difference series, and therefore we cannot consider these tests as independent. Nonetheless, the relative uniformity in rejection yields very strong support in favor of the SPAR model.

in loss-differences suggests that the SPAR model outperforms HAR and is certainly not dominated by it. This evidence in favor of SPAR volatility models suggests that under the elastic net estimation method, volatility spillover effects coupled with realized measures can significantly improve volatility forecasts.

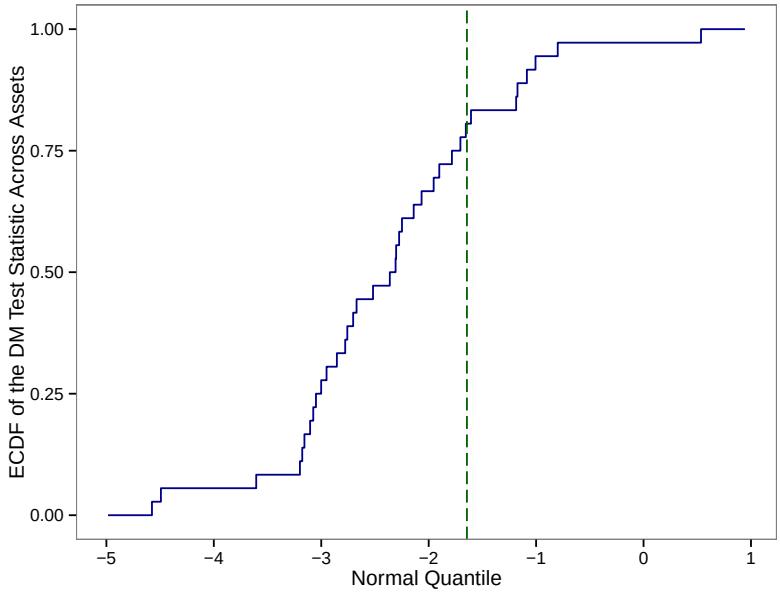


Figure 2.5: Empirical distribution function of Diebold-Marino type test statistics over thirty-six assets, where the loss is between the SPAR and the HAR variance models for the entire time-series sample (approximately 2250 observations).

To home in on particular periods of interest, Figure 2.6 shows the percentage of the thirty-six assets rejecting the Diebold-Mariano type test at the 95% confidence level over each individual year separately. That is, over each year, we estimate the loss-difference test-statistic for each realized variance series, and compute the proportions rejecting the one-sided test at the 95% confidence level in favor of the SPAR and of the HAR models, respectively. If these tests were independent, the proportion of assets that would incorrectly reject under the null would be 5%, as marked by the dashed red line. We plot the proportions for each year and check for temporal trends. We see that in line with the previous figure where the SPAR model seemed to dominate over the entire sample, the SPAR model appears to reject for a greater proportion of the assets throughout time. In particular, the SPAR model performs considerably better in 2007-2009, the years during and immediately after the financial crisis. This suggests that volatility spillovers are particularly prevalent over these years, and that from a forecasting perspective, the SPAR model is adept at estimating their effect. Moreover, well over 5% of the assets reject over this period. Though we cannot make formal inference on this fact since the tests are dependent, this strongly indicates that these rejections are not due to type I error. These

results validate and motivate the use of the SPAR model to estimate volatility spillover networks and to use these networks to identify systemic risk.

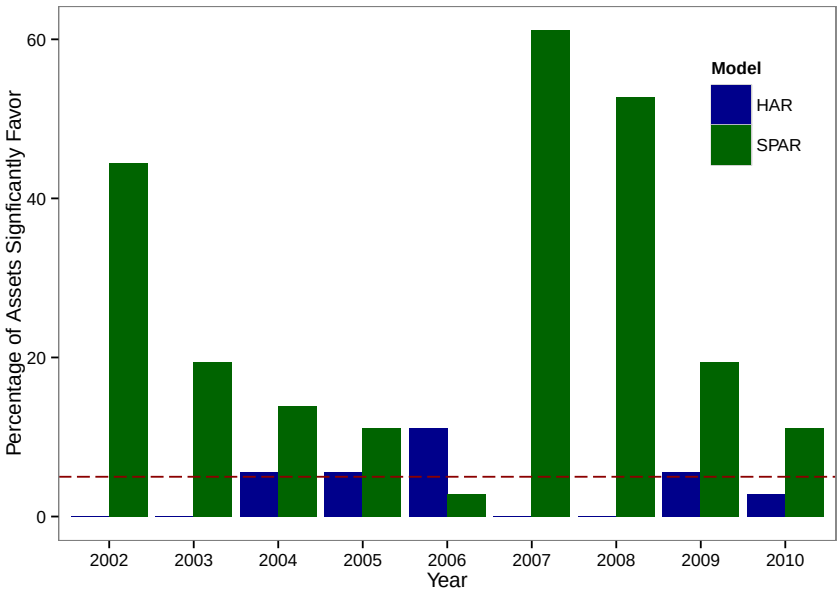


Figure 2.6: Percentage of assets that significantly reject equal predictive ability in favor of one model by year. Rejections are given by Diebold-Mariano type test statistics that are below or above the normal 95% quantile in each year.

2.4 Systemic Risk with Volatility Spillover Networks

Having observed the strong predictive performance of volatility spillover networks via the SPAR model, we now analyze the nature of the selected networks themselves. These networks represent the short-run Granger-causal effect of lagged volatilities of the whole set of assets onto the volatility of the dependent asset. In our setting the assets are stocks of the largest U.S. financial institutions, and these networks represent the effect of a risk spillover from one institution to another. If there is a greater level of risk spillover in aggregate, financial risk is more contagious, and systemic risk is endemic.

The network implied by volatility spillover comes from the selected sparsity pattern in each estimation. Recall that elastic net estimation will set many parameters to zero, so that the respective financial institutions will not spill over any volatility to other institutions. That is, we can define the unweighted connectedness from institution j to i at any given time as:

$$\hat{a}_{i,j} = \begin{cases} 1 & \text{if} \quad \hat{\beta}_{i,j} \neq 0 \\ 0 & \text{if} \quad \hat{\beta}_{i,j} = 0 \end{cases} \tag{2.10}$$

This allows us to define the adjacency matrix for a directed graph at any point in time. Denote $\hat{\mathcal{A}}$ as the matrix with elements $\hat{a}_{i,j}$ representing which institutions spillover into other institutions, and which do not[16]. Note that such a matrix is a vast simplification of the information given by the volatility spillover model, but it nevertheless allows for simple visualization of the network and some illustrative analysis. From a more informative perspective, we can analogously define the weighted connectedness by a similar matrix, $\hat{\mathcal{B}}$, which coincides with the estimated spillover-effect parameter matrix $\hat{\beta}$ in the SPAR model. We use this information along with some simple graph theory to approximate the spread of volatility in our system of financial institutions.

From a network perspective, the estimated adjacency matrix $\hat{\mathcal{A}}$ corresponds to the graph $\mathcal{G}(V, E)$ at each point in time. This graph contains a set of nodes (or vertices) V corresponding to the financial institutions and a set of edges E corresponding to the volatility spillover between them. There is an edge $e_{i,j}$ between i and j if and only if $\hat{a}_{i,j} = 1$, that is, if there is volatility spillover from node (institution) i to node (institution) j. We can use this representation to view and analyze the volatility spillover graph at any given point in time by connecting the nodes (financial institutions) with the edges. This gives us a visual illustration of the estimated spillover network at any given time, as in Figure 2.1.

A financial institution spills over volatility to many other institutions if it has many edges coming out of its corresponding node. Likewise, a node absorbs a lot of external volatility if it receives many edges coming from other nodes. Determination of the impact of the node on the entire graph and hence its contagiousness is called *centrality*, which can be measured in a variety of ways. A simple and intuitive way is to count the number of edges flowing from the node: this is called *degree centrality*. In the case of a weighted graph, the degree centrality of a node corresponds to the sum of the weights of the edges flowing from that node and represents the marginal impact of that node on all the other nodes. That is, the degree centrality is the 'beta contribution' to the system from that financial institution. When this beta contribution is high, that financial institution is spilling-over considerable amounts of risk to other institutions and can thus be considered systemically risky. Since the number of financial institutions changes over time, we compute a 'mean beta contribution' which normalizes the beta contribution by the total number of nodes, thus allowing for comparisons across time. Formally, the mean beta contribution of institution i is given by:

$$\hat{C}_i = \frac{1}{|V|} \sum_{j=1}^{|V|} \hat{\beta}_{j,i} \tag{2.11}$$

[16]Note that the notation here is the same as that usually denoting the active set of regressors $\hat{\mathcal{A}}$ in a penalized regression model. This is deliberate, since the adjacency matrix holds exactly this information.

where $|V|$ is the dimensionality of the network at any point in time and $\hat{\beta}_{j,i}$ are the estimated parameters in the volatility spillover network.

To understand the aggregate change in the graph of financial institutions over time, we can decompose its adjacency matrix into lower-dimensional central features and investigate their evolution over time. One such feature is the mean of all of the mean beta contributions:

$$\hat{R} = \frac{1}{|V|} \sum_{i=1}^{|V|} \hat{C}_i \tag{2.12}$$

The higher this aggregate indicator $\hat{R}$, the more spillover there is in the network as a proportion of the number of nodes, and the larger the degree of systemic risk. Hence, we use this aggregate mean beta contribution indicator as an indicator of systemic risk in the entire financial system. It is also interesting to see whether spillover is coming in a diffuse or concentrated way. That is, whether systemic risk arises out of one group of financial institutions or from several herds. To do this, we decompose the adjacency matrix into its eigenvalue-eigenvector representation, and analyze the behavior of the largest eigenvalue over time. The corresponding eigenvector is the main principal component of the adjacency matrix, which represents the direction of maximal variation in the entries of the adjacency matrix. Likewise, the eigenvector corresponding to the next largest eigenvalue represents the next direction of maximal variation in the adjacency matrix that is orthogonal to the first eigenvector. Continuing in this fashion, the eigenvectors, or principle components, correspond to orthogonal directions of maximal change. Their corresponding eigenvalues represent the magnitude of variation that these eigenvectors explain. Moreover, each eigenvalue divided by the sum of all of the eigenvalues corresponds to the share of variation in the adjacency matrix explained by the corresponding eigenvector. We can hence define a concentration indicator which measures the degree of variation in the weighted adjacency matrix $\hat{\mathcal{B}}$ explained by the direction associated with the largest degree of variation (eigenvalue). This spillover concentration indicator is given by:

$$\hat{\kappa} = \frac{\hat{\lambda}_1}{\sum_{j=1}^{|V|} \hat{\lambda}_j} \tag{2.13}$$

where the estimated eigenvalues $\hat{\lambda}_j$ of the weighted adjacency matrix $\hat{\mathcal{B}}$ are ordered in descending order, $\hat{\lambda}_1 > \hat{\lambda}_2 > ... > \hat{\lambda}_{|V|}$. When this indicator rises over time, the dispersion in the adjacency matrix can be explained more by the first eigenvector. That is, the adjacency matrix and the volatility spillover pattern is more concentrated and less diffuse. When this occurs, risk spillover can be explained in a more systemic and less diffuse fashion.

Over the 2007-2008 financial crisis it was believed that U.S. financial institutions were systemically related, with many institutions spilling risk onto one another. Here,

we see whether this systemic nature is reflected in the volatility spillovers between them. We look at the sparsity pattern implied by the volatility spillovers and at the values of the estimated coefficients. That is, we look at the unweighted and weighted networks implied by the volatility spillovers and we analyze the features of these networks. We examine specific institutions that had undergone default or duress, the magnitude and concentration of the estimated systemic risk network, and the nature of this network before and after the 2008 Lehman Brothers collapse.

2.4 Institution-Level Systemic Buildup

As a starting point, we consider one large institution that did not fail during this period, Goldman Sachs. Figure 2.7 shows the mean beta contribution of Goldman Sachs in 2008, before the Bear Stearns takeover and during the collapse of Lehman Brothers. This beta contribution shows that Goldman Sachs was spilling over considerable risk before and during the Bear Stearns failure in March 2008. After the takeover was announced, Goldman Sachs' contribution fell until September, at which point it jumped with the Lehman Brothers collapse. Finally, this beta contribution descended to low levels by October. This indicates that estimated volatility spillover networks are responsive to rapid changes in the risk climate, with particular sensitivity to systemic events.

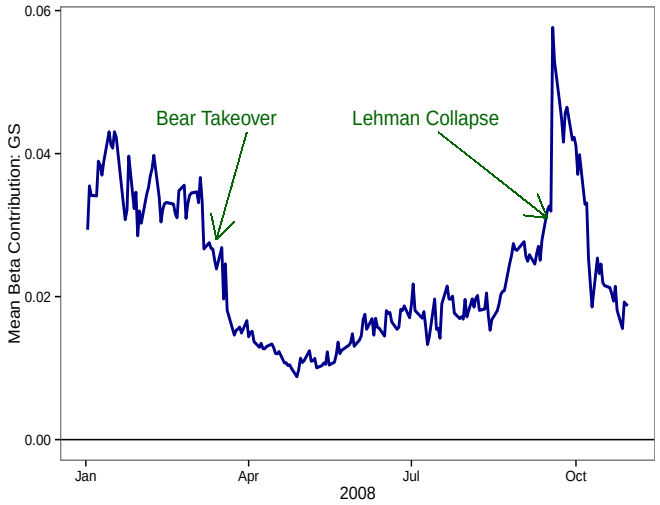


Figure 2.7: Spillover effect of Goldman Sachs over 2008 measured by mean beta contribution.

We now turn to the institutions that we focused on earlier, which were under partial or severe duress in 2008: Lehman Brothers, Citigroup, AIG, and Fannie Mae. Figure 2.8 shows that the mean beta contribution of Lehman Brothers rose consistently through the summer of 2008 until skyrocketing in September. Finally, Lehman Brothers failed and fell out of the network. At this point, other troubled financial institutions like AIG and Citigroup continued to spill over risk and had even higher spillover effects, reflecting

an increase in contagion. Notice that this is in distinction to the beta contribution of Goldman Sachs, which levelled off after the collapse of Lehman Brothers. This again shows that systemically risky institutions frequently have a buildup of beta contribution, and that this contribution is particularly pronounced during systemic events. Moreover, the beta contribution of stable institutions tends to diminish after such events, whereas the contribution of unstable institutions remains high.

Similar patterns emerge with the large government-sponsored-enterprise, Fannie Mae. Figure 2.8 shows the spillover effect of Fannie Mae over the 2007-2009 period. As the market internalized the detrimental impact of mortgage backed securities in 2007, the beta contribution for Fannie Mae rose. In August of 2007 Fannie Mae devalued, and its beta contribution spiked. Spillover again rose in July of 2008, when U.S. Treasury Secretary Paulson announced his intention to seek congressional approval for the backing of Fannie Mae by the U.S. Treasury. We can see here that Fannie Mae's beta contribution rose consistently during these periods of duress and systemic buildup.

2.4 Aggregate Systemic Buildup

Having observed the characteristics of individual systemic institutions, we can also analyze the behavior of the system in aggregate. First, we visually inspect the network before and during the financial crisis. To this end, we use the unweighted and weighted adjacency matrices and analyze their resulting graphs. Figures 2.9 and 2.10 show the entire estimated weighted volatility spillover networks on September 19, 2006 and 2008 respectively. For each figure, the edges are darker when the spillover weights are larger. Likewise, the nodes are darker when their degree centrality is larger; that is, when they spill-over volatility to more nodes. The two figures reveal that there is a considerably greater degree of volatility spillover after the Lehman collapse in 2008, since the number of edges increases are well as their darkness. Furthermore, many of the nodes are darker, reflecting the greater degree of connectedness for their associated financial institutions. Among these are AIG, Washington Mutual, and large broker-dealers such as Goldman Sachs and Morgan Stanley. These institutions were under duress or severe uncertainty at this time, and this fact is reflected in the estimated volatility spillover network.

To focus on the change in the network surrounding the collapse of Lehman Brothers, we also pick a handful of the larger institutions from each category and plot the edges between them[17]. The first graph, Figure 2.11, shows the volatility spillover sub-network on July 10, 2008, prior to the collapse of Lehman Brothers. There are some connections here, but for the most part this network is relatively sparse, reflecting little volatility spillover prior to the systemic event. This contrasts starkly with Figure 2.12, which shows the same sub-graph on September 19, 2008, immediately after the Lehman Brothers

[17]Note that these institutions were picked only for their size and familiarity, not for their connectedness.

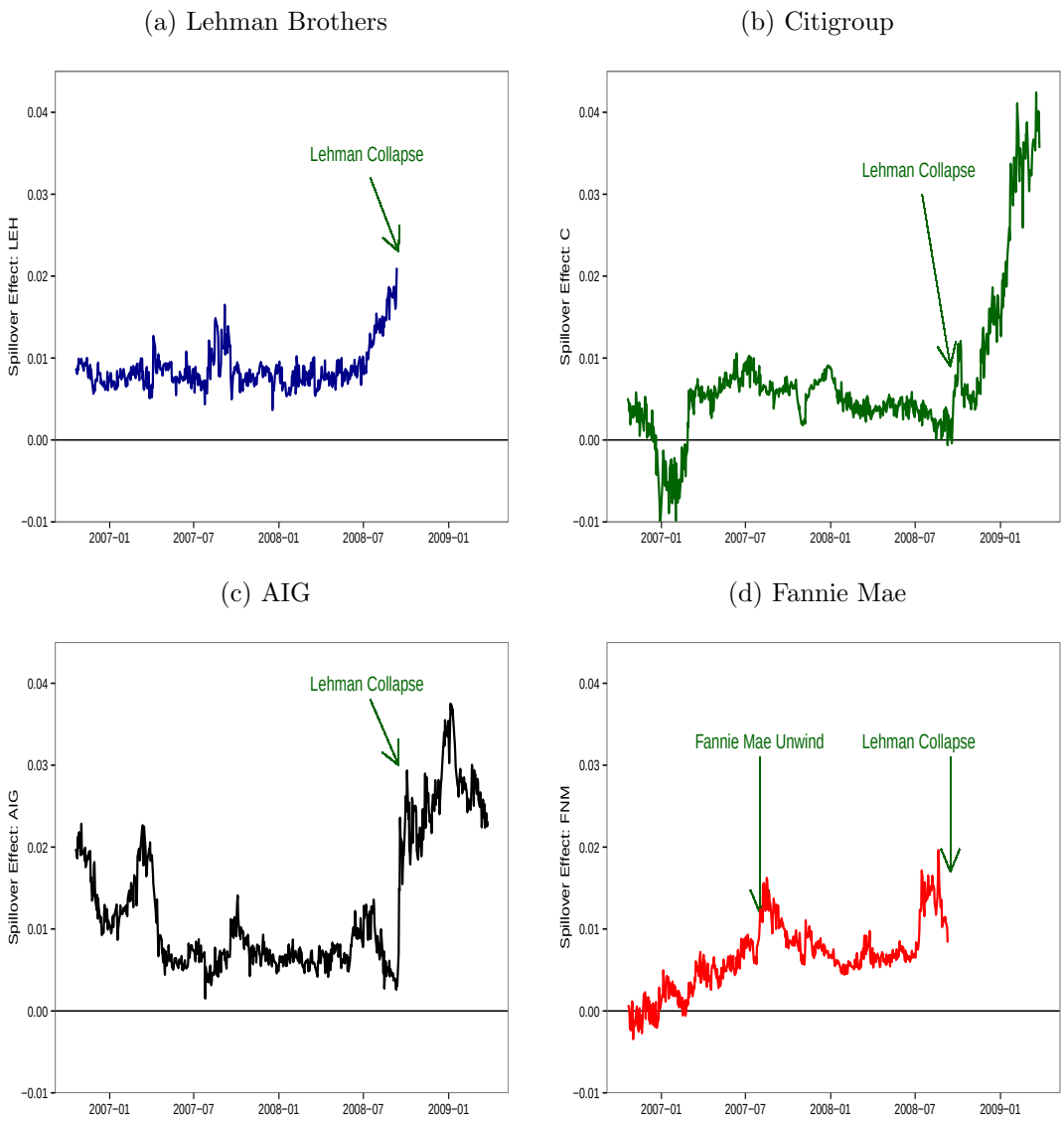


Figure 2.8: Spillover effect as measured by mean beta contribution for institutions in each financial category over the period 2007-2009: Lehman Brothers (Broker-Dealer), Citigroup (Depositories) , AIG (Insurers), and Fannie Mae (Others).

collapse. Here, we clearly see every node engaged with at least one other node. The greater degree of connectedness is easily seen in this sub-graph, reflecting greater volatility spillover throughout all the selected financial institutions.

These 'snapshot' graphs reflect the behavior we see in our systemic risk indicator, the mean of the mean beta contributions. Due to the logarithmic transformation, this indicator can be interpreted as the marginal percentage amount of volatility spillover resulting from a one percent change in the previous lagged volatility. Figure 2.13 shows this aggregate mean beta contribution over time. This indicator begins climbing in 2007, with a particular jump in July-August of 2007, when the sub-prime mortgage problem became particularly prominent. The indicator stays at this level without falling back to

its prior level, suggesting that systemic risk had far from dissolved. In June of 2008 this indicator begins to climb again, with a large bump during the Lehman collapse, and only peaks in early 2009. At this stage, the indicator attains a level more than twice as high as two years prior, suggesting that volatility spillover accounted for twice as much of 1-day ahead volatility as in 2006. This period reflects the buildup of contagion in the network of financial institutions, and the instability seen in the financial system. In February of 2009, as Secretary Geithner announced the Public-Private Investment Program (PPIP) as the last $300 billion bailout in the Troubled Asset Relief Program (TARP), this indicator begins to fall, reflecting a weakening of contagion in the financial network. The ability of this indicator to identify periods of increasing risk buildup motivates us to choose this as our aggregate systemic risk indicator.

We also consider the degree of concentration in the estimated network over time. This may be important to determine whether the aggregate risk indicator's buildup is driven by a small, concentrated number of factors, or by a large, diffuse number of factors. Figure 2.14 shows the degree of concentration as measured by the share of the largest eigenvalue described above. The plot shows that starting in June 2008 and ending in January 2009[18], variation in the spillover network was explained more and more by the first principle component. Moreover, this behavior was also echoed by the second and third principle components. This indicates that variation in volatility spillover was more and more predictable by a small set of systemic factors, and that additional aggregate risk spillover was particularly concentrated.

We note that our results are highly robust to the choice of sample size and α tuning parameter. Figure 2.15 shows the aggregate beta contribution systemic risk indicator for different sizes of rolling windows. The dark blue line reflects the 1-year horizon results seen above, and the dashed lines reflect the 2-year (green), 9-month (purple), and 6-month (red) rolling windows. All of these horizons show very similar behavior, with slightly greater instability occurring at shorter horizons. Similarly, Figure 2.16 shows the sensitivity of the aggregate beta contribution systemic risk indicator to the tradeoff between LASSO and ridge tuning parameter α. We see remarkable stability, suggesting that this tuning parameter has little impact in aggregate but possibly greater impact in modeling individual series.

In sum, volatility spillovers reveal much about the financial network over the 2007-2008 period. This is seen at the institution level, as well as on aggregate. We suggest the aggregate mean beta contribution indicator as the systemic risk indicator, where a large increase in this indicator represents considerable levels of connectedness in the network of financial institutions. We showed that this indicator was able to reflect a systemic shift in

[18]Note also the large jump and fall during the Bear Stearns takeover. Though risk spillover increased in aggregate over this period, and the network was more concentrated as seen here, concentration then levelled off until the summer.

2007 that persisted and magnified throughout 2008. This indicator is also complemented by the largest eigenvalue share indicator, which reflects the degree of concentration in aggregate spillover. When both of these indicators are on the rise, there is a greater degree of connectedness and risk, and this risk emanates from a small group of systemic sources. Such situations result in periods of financial fragility which probably ought to be mitigated or accounted for by the regulatory authority.

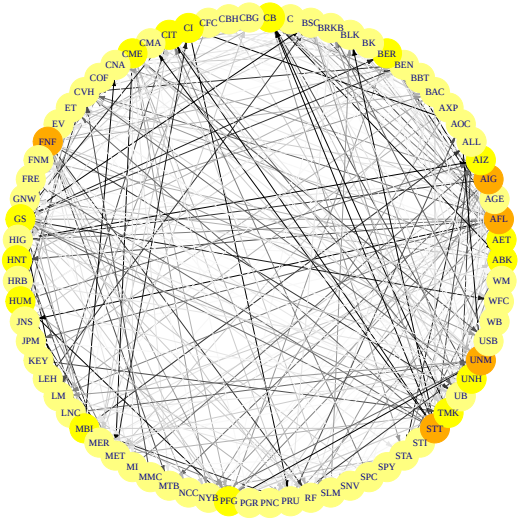


Figure 2.9: Volatility spillover network on September 19, 2006, two years prior to the Lehman collapse. Nodes represent institutions, and directed edges represent directed spillovers. The darker the edge, the larger the value of the coefficient. The darker the color of the node, the higher the degree centrality.

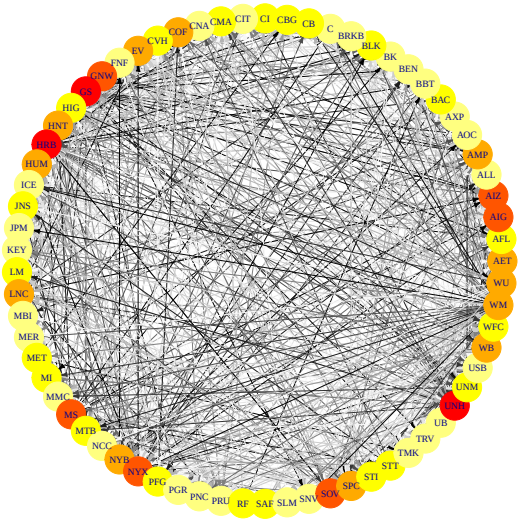


Figure 2.10: Volatility spillover network on September 19, 2008, immediately after the Lehman collapse. Nodes represent institutions, and directed edges represent directed spillovers. The darker the edge, the larger the value of the coefficient. The darker the color of the node, the higher the degree centrality.

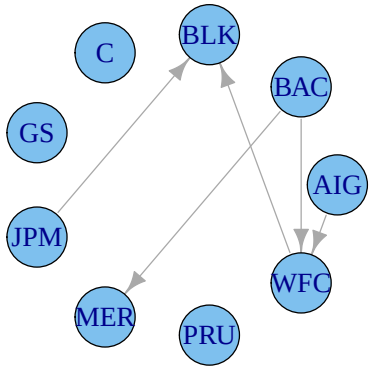


Figure 2.11: Graph showing directed volatility spillover for a sub-graph of the volatility spillover network on July 10, 2008, prior to the Lehman collapse. Nodes represent tickers: Citigroup (C), Blackrock (BLK), Bank of America (BAC), AIG (AIG), Wells Fargo (WFC), Prudential (PRU), Merrill Lynch (MER), J.P. Morgan (JPM), and Goldman Sachs (GS).

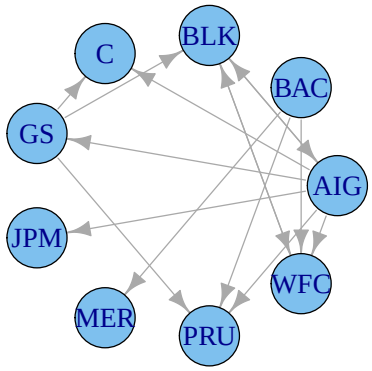


Figure 2.12: Graph showing directed volatility spillover for the same sub-graph as Figure 2.11 of the volatility spillover network on September 19, 2008, following the Lehman collapse.

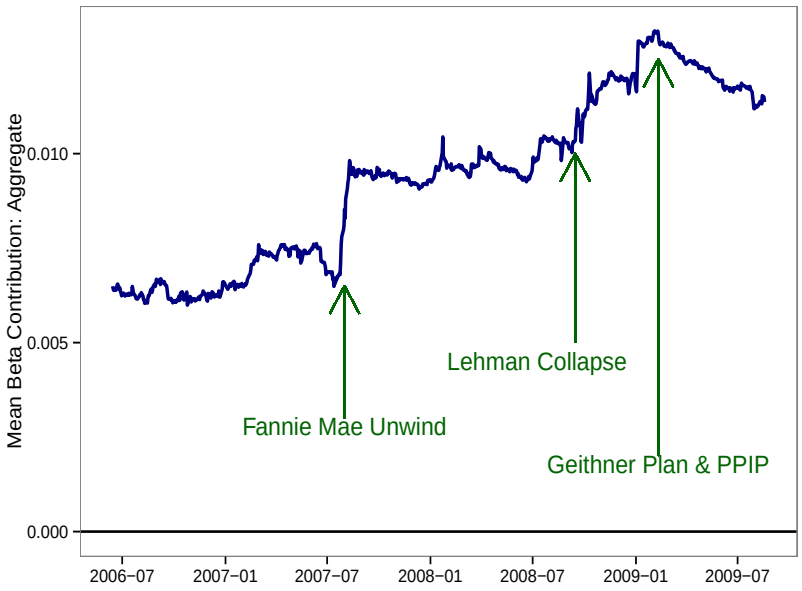


Figure 2.13: Aggregate Mean Beta Contribution (systemic risk indicator) for the weighted volatility spillover network implied by the SPAR volatility model.

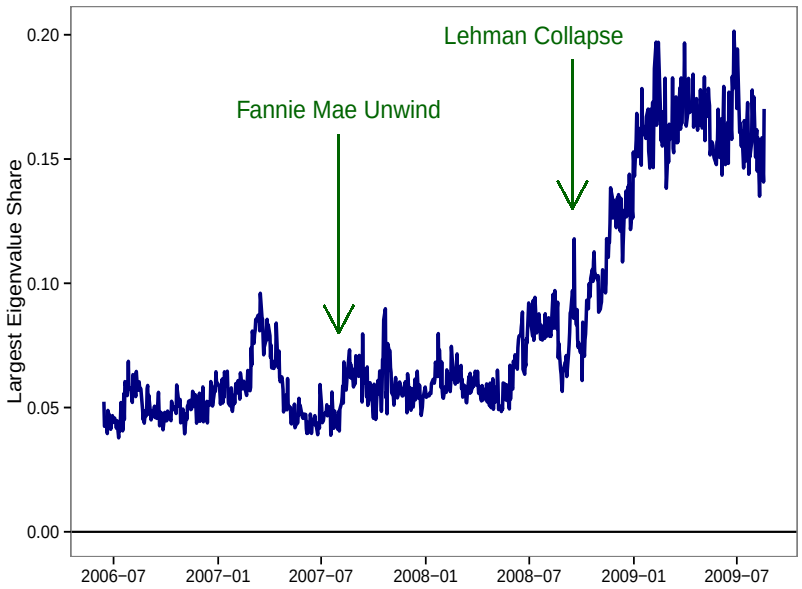


Figure 2.14: Share of the largest eigenvalue for the weighted volatility spillover network implied by the SPAR volatility model.

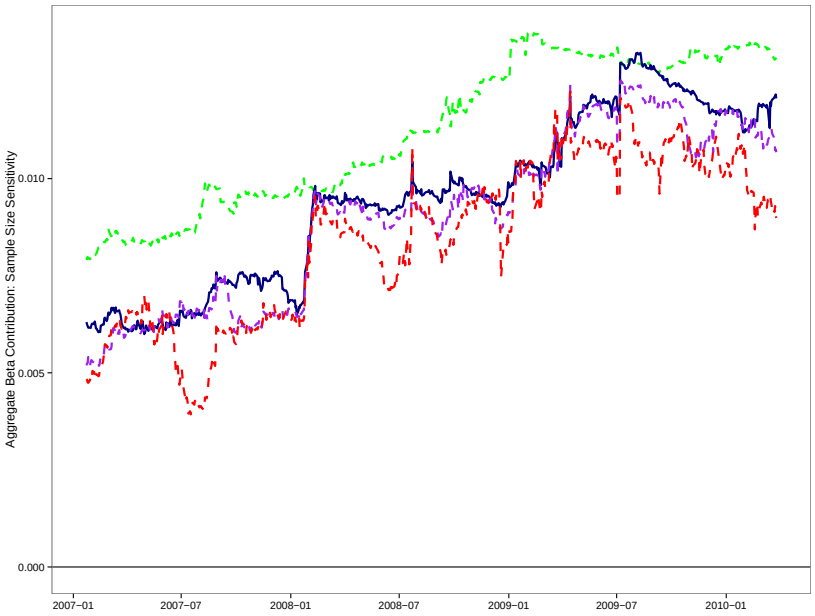


Figure 2.15: Sensitivity of the Mean Beta Contribution to the rolling window horizon: 2-year (dashed green), 1-year (blue), 9-month (dashed purple), and 6-month (dashed red).

Figure 2.16: Sensitivity of the Mean Beta Contribution to the elastic net tuning parameter α, which determines the tradeoff between LASSO and ridge: $\alpha = 0$ in dashed green, $\alpha = 0.25$ in dashed light blue, $\alpha = 0.5$ in blue, $\alpha = 0.75$ in dashed purple, and $\alpha = 1$ in dashed red.

2.5 Extentions and Conclusions

In this paper we introduced an inter-day predictive model for volatility using high-frequency methods that allows for a wide cross-section of explanatory variables. By estimating the model with the elastic net, the resulting spillover autoregressive (SPAR) realized variance model can handle very large cross-sections of explanatory variables that are potentially highly correlated. We showed that this model significantly improves forecasts, and it performs particularly well during periods of risk turmoil: that is, periods when accurate forecasts are needed most. We then turned our attention to the estimation of systemic risk in the network of U.S. financial institutions, and we showed that the sparsity pattern and parameter estimates resulting from the SPAR realized variance model have an elegant, intuitive explanation in terms of networks. Specifically, we saw that institutions that were particularly systemically risky and had undergone some financial duress also exhibited clear patterns in the volatility spillover network. In particular, the spillover network was highly responsive to systemic events, as well as more gradual systemic risk buildup. This analysis culminated in the construction and analysis of the aggregate mean beta contribution indicator, which appears to identify periods of systemic risk buildup successfully, as seen by its persistent and gradual rise from mid 2007 until early 2009. We additionally suggest the use of a concentration indicator, which, when used in combination with the aggregate beta contribution indicator, explains whether spillovers are increasing due to systemic or diffuse behavior. We showed that in the summer of 2008, aggregate systemic risk as measured by the aggregate mean beta contribution indicator increased in a concentrated manner.

This paper is a novel addition to the literature on systemic risk indicators and measures. We note that our systemic risk indicator is not factor-based, and arises purely out of the network of financial institutions. This way, future periods of systemic risk can be identified even if they do not appear in a particular pre-chosen factor. Furthermore, our measure is grounded in the volatility forecasting literature. We account for long-memory as well as short-term news impacts by using long term and short-term horizons in the SPAR realized variance model. Moreover, we use high-frequency variance estimates, which respond to change at a faster rate than slow-moving accounting data or even daily GARCH-type models. Despite these boons to our model, there are several small drawbacks. First, as our model is volatility-based, we do not differentiate between 'good-volatility' and 'bad-volatility': that is, volatility arising from positive price movements versus negative price movements. We do not consider this to be a severe drawback, however, since volatility tends to rise more from negative returns than positive ones, and moreover, volatility is a reflection of uncertainty (and hence risk) regardless of the movement in price. An additional drawback is that our systemic risk indicator only identifies systemic risk, but does not indicate the quantity under risk; that is, this is not a measure

of capital loss. To this end, we would need to extend our results onto return space as in Brownlees and Engle (2011), which we leave as a topic for further research.

We additionally note that the SPAR realized variance model can be applied in other problems. The question of volatility spillover was addressed in the general context of U.S. equities for example in Diebold and Yilmaz (2008), and can be applied to any particular set of assets with sufficient liquidity. Assets with common features and autoregressive behavior are expected to have possible volatility spillover, and the SPAR model can be used to analyze these assets. Moreover, the SPAR model requires fewer degrees of freedom than competing methods, and can hence analyze spillovers over short horizons with wide cross-sections. This application can be useful for exchange rates, for example, which are frequently believed to have regime-dependent behavior (Bollen et al., 2000). A rolling model such as the one we presented above could account for such regime shifts and at the same time allow for a wide cross-section to be used.

Chapter 3

Tests of Equal Conditional Predictive Ability With Many Test Functions

Abstract. We propose bootstrap procedures for two-sided tests of simple nulls. We demonstrate that variants of the Bootstrap Reality Check (White, 2000) procedure can be used in place of more traditional Wald-type tests, such as tests of equal conditional predictive ability. Further, we show that these procedures are particularly adept in testing nulls involving a large number of parameter restrictions, such as nulls of equal conditional predictive ability with many test functions. In a simulation study, we show the strong size and power properties of the bootstrap procedures when compared against the standard asymptotic test of Giacomini and White (2006). Lastly, we apply these bootstrap procedures to determine the predictive value of real variables in U.S. inflation forecasting.

Keywords: Forecast Evaluation, Out-Of-Sample, Hypothesis Test, Bootstrap
JEL: C53, C32, C51, E32, E37

We suggest bootstrap procedures for tests of simple nulls with many restrictions, focusing on the key example of tests of equal conditional predictive ability with many test functions. Tests of predictive ability are essential tools for researchers and practitioners, and are usually implemented with an estimated Wald-type test statistic. These tests have been shown to be more informative and more powerful (Giacomini and White, 2006) when enriched with additional variables, or test functions. The use of these test functions provides power against the null, but this added power comes at a rapidly increasing finite-sample cost stemming from estimation-error in the test statistic. In this paper, we present known bootstrap procedures as estimation-error-robust alternatives to Wald-type test statistics. The theoretical underpinnings of these bootstrap procedures guarantee they convergence to the same limiting distribution as the Wald-type test statistic. At the same time, the bootstrap procedures estimate far fewer parameters, strengthening them against finite-sample estimation error. In the following discussion we provide a simple theoretical explanation of the asymptotic equivalence and the finite-sample advantage of using these bootstrap procedures, and we show these properties in numerical simulations. Lastly,

we revisit the problem of U.S. inflation forecasting using a backward-looking Phillips curve, and examine the importance of lagged real variables on out-of-sample forecasts of inflation.

Though these bootstrap procedures are valid in general settings, we concentrate on tests of predictive accuracy because of the practical importance of predictive testing. Predictive accuracy is a fundamental criterion for economic modeling, forecasting, and policy evaluation. When a model predicts accurately, it captures features of the fundamental process driving the data instead of the noise that is particular to past realizations of that process. Out-of-sample predictive robustness of a model hence serves to validate the model outside of the data on which it was estimated, thus mitigating over-fitting concerns. For this reason researchers have often chosen to evaluate their model or compare their model to an industry benchmark using out-of-sample predictive accuracy tests. Notable tests include Mincer and Zarnowitz (1969), a regression-based test which determines whether forecasts are significant in determining future outcomes, and Diebold and Mariano (1995), which compares the performance of two forecasting methods.

We couch our discussion in the context of the forecast comparison tests introduced in Diebold and Mariano (1995) and further developed in Giacomini and White (2006). More specifically, we focus as in Giacomini and White (2006) on comparisons of two forecasting methods and test whether the difference in their forecast-error-losses is predictable. That is, we test whether one method is predictively equal to another method conditional on all the information available, so that the loss difference between the two methods forms a martingale-difference-sequence (MDS). In practice, testing whether forecast error loss differences form a MDS requires specification of a particular set of (lagged) variables, or test functions, which represent the information available at the time the forecast is made. With the increasing availability of data, a researcher may wish to include many such variables to achieve greater power against the null. In finite samples, however, the test of Giacomini and White (2006) can be severely limited by the number of test functions used. This limitation arises out of the construction of the Wald-type test statistic, which requires estimation and inversion of a covariance matrix which can rapidly become large and unstable as the researcher includes additional test functions.

We introduce testing procedures that differ from that of Giacomini and White (2006) by using bootstrap methods in determining critical values. The bootstrap embeds the underlying dependence structure of the original data into the re-sampled data, and hence circumvents estimation of potentially large covariance matrices. These bootstrap procedures thus deliver potential improvements to power and considerable improvements to size in situations where the dimension of the test functions (the number of variables) is moderately large, and the asymptotic convergence of the Wald-type test statistic in Giacomini and White (2006) is less reliable. As in Giacomini and White (2006), our bootstrap testing procedures are accurate when the number of variables is small relative to

the sample size, but unlike their test, the bootstrap procedures remain accurate when the number of variables is proportional to or greater than the sample size. Furthermore, one of the bootstrap procedures additionally pinpoints the specific variables that represent aberrations from the MDS null, informing the researcher of specific missing information in the forecasts.

The tests presented here are rooted in the work of White (2000), Hansen (2005), and Romano and Wolf (2005) on tests of superior predictive ability. The Bootstrap Reality Check of White (2000) showed that the stationary bootstrap can be used to approximate the asymptotic distribution of the maximum of correlated multivariate normal random variables, a fundamental property that we utilize here. White (2000) considered a composite null of superior predictive ability, which was further developed in Hansen (2005), where it was shown that improvements in power can be gained from studentization and truncation. These tests of White (2000) and Hansen (2005) test whether *any* parameter among a set of parameters is less than or equal to zero. These tests do not consider the question of finding *which* parameters are greater than zero. In Romano and Wolf (2005), the properties of White (2000) and Hansen (2005) are used to find the parameters that result in violations of this null. Specifically, Romano and Wolf (2005) develop a stepwise testing procedure that controls the familywise error rate (FWE), that is, the probability of incorrectly rejecting at least one null. With this stepwise testing procedure, Romano and Wolf (2005) not only establish whether all of the parameters satisfy the null, but also find which parameters belong to the alternative.

In this paper we use the bootstrap procedures in White (2000), Hansen (2005), and Romano and Wolf (2005) to test simple rather than composite nulls, such as nulls of equal conditional predictive ability. We demonstrate that the typical Wald-type test statistic as in Giacomini and White (2006) can be misleading when the dimensionality of the estimated parameters is even moderately large. We frame our discussion in the context of the test of Giacomini and White (2006), but we note that the test is widely applicable to other large-dimensional Wald-tests which rely on asymptotic theory.

The discussion proceeds as follows. In Section 3.1 we present the test of Giacomini and White (2006) for testing equal conditional predictive ability and explain its vulnerability to a large number of restrictions. Section 3.2 briefly explains the methodology and assumptions required to test the null using the bootstrap. Section 3.3 is a numerical study demonstrating the appeal of the bootstrap approach in the context of tests of equal conditional predictive ability, and Section 3.4 applies the bootstrap procedures to U.S. inflation forecasting. In Section 3.5 we discuss extensions of this procedure to other Wald-type tests, and in Section 3.6 we conclude.

3.1 Testing for Equal Conditional Predictive Ability

As in Giacomini and White (2006), suppose that the data under question are driven by a stochastic process $W = \{W_t : \Omega \to \mathbb{R}^{s+1}, s \in \mathbb{N}, t = 1, ..., T\}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We partition $W_t = (Y_t', X_t')'$ and let Y_t be a univariate process of interest and X_t be a vector of predictive variables which may be observed at time $t-1$, so that $\mathcal{F}_t = \sigma(W_1', ..., W_t', X_{t+1}')$. For clarity of exposition, we adopt the notation of Giacomini and White (2006) throughout this section and denote random variables by upper case letters and their realizations in lower case.

We are interested in testing whether one of two univariate forecasting methods f or g results in better τ-step ahead predictive accuracy. That is, we evaluate which method f or g forecasts the realization $Y_{t+\tau}$ better at time t. Each of the forecasting methods is $\mathcal{F}_t$-measurable and formulated at time t, so that only information known at time t can be used in constructing the forecast. As in Giacomini and White (2006) we denote the forecasts by $\hat{f}_{m,t} \equiv f(w_t, w_{t-1}, ..., w_{t-m+1}; \hat{\beta}_{m,t})$ and $\hat{g}_{m,t} \equiv g(w_t, w_{t-1}, ..., w_{t-m+1}; \tilde{\beta}_{m,t})$, where each forecasting method can use an estimated parameter vector $\beta_{t,m}$, and subscripts denote that the forecast is formulated at time t and utilizes only the m most recent observations. Note that though Y_t is univariate, we are not confined to point forecasts - the framework used here will exploit the loss function of the forecast which will be univariate, hence allowing for density, interval, and probability forecasts. Moreover, this framework allows for non-nested models of any type, including parameteric, non-parametric, and semi-parametric models. Thus, like Giacomini and White (2006), the framework here is very flexible with regard to the nature of the forecasting methods, with the exception of nested models.

In order to determine predictive ability, each model is evaluated with a loss function using a pseudo-out-of-sample approach. Let T be the total sample size available. The data from $t = 1, ..., m$ are used for estimation of the first set of parameters, and the first τ-steps ahead forecasts are formed at time m and compared with $y_{m+\tau}$, the realization of $Y_{t+\tau}$. This procedure is rolled forward, so that the data from $t = 2, ..., m+1$ are used for estimation of the second set of parameters, and their resulting forecasts are compared to $y_{m+\tau+1}$. This rolling window method produces an out-of-sample period of length $n = T - \tau - m + 1$, which we use as an input into our loss function. To assess the predictive strength of f requires the specification of a loss function $L_{t+\tau}(\hat{f}_{m,t}, Y_{t+\tau})$ which specifies the degree of loss accumulated by using the forecast rather than the true realization. The loss function can come from economic theory (e.g. utility, cost) or it can be a simple measure of statistical accuracy. In the numerical and empirical sections we focus on mean-square-error loss,

$$L_{t+\tau}(\hat{f}_{m,t}, Y_{t+\tau}) = (\hat{f}_{m,t} - Y_{t+\tau})^2$$

simply because this is the loss most frequently encountered in forecasting applications[1]. Given an n-long sequence of losses for each forecasting method, we can take their difference and investigate whether this loss-difference sequence is a random, mean-zero sequence, or whether it is unusually small, large, or predictable.

The test proposed by Giacomini and White (2006) tests the null that the difference in losses constitutes a martingale difference sequence. That is, each loss difference is mean zero (the losses are on average equal, and f and g are equally accurate) conditional on all of the information available at time t:

$$\begin{aligned} H_0 \quad &: \quad E[L_{t+\tau}(\hat{f}_{m,t}, Y_{t+\tau}) - L_{t+\tau}(\hat{g}_{m,t}, Y_{t+\tau})|\mathcal{F}_t] \\ &\equiv \quad E[\delta_{t+\tau}|\mathcal{F}_t] = 0 \text{ almost surely } t = 1, 2, ... \end{aligned} \tag{3.1}$$

In this context of forecast comparisons, we let $\delta_{t+\tau}$ denote the loss differences, suppressing the index m. We do this to emphasize that this testing procedure is valid for any $\delta_{t+\tau}$ sequence satisfying the assumptions of Giacomini and White (2006) as well as the more stringent ones discussed below, which imply m-invariance.

To test the martingale-difference sequence null (3.1), we note that it can be restated as the moment condition $E[h_t\delta_{t+\tau}] = 0$ for all $\mathcal{F}_t$-measurable functions h_t. In practice, we restrict ourselves to a finite dimensional set of variables, or test functions (Stinchcombe and White, 1998), so that h_t is a $q \times 1$ vector of values known at time t. Given a specific choice of test function h_t, we can test the h-specific null $H_{0,h} : E[h_t\delta_{t+\tau}] = 0$. This null tests whether the test function h_t is correlated with the sequence of loss differences. A rejection of this null suggests that the loss-differences are predictable by the test function, and are therefore not a martingale-difference sequence. An important simple special case is when no time-varying information is used, and $h = 1$. This corresponds to the standard unconditional test of predictive ability in Diebold and Mariano (1995), which investigates whether the loss-differences are biased in one direction at any point in time. When h_t is time-varying, the tested null is more restrictive, allowing for no conditional or unconditional biases.

To test the null (3.1), Giacomini and White (2006) show that a Wald-type statistic of the form[2]:

$$GW_{n,m}^h = n\Big(n^{-1}\sum_{t=m}^{T-\tau} h_t\delta_{t+\tau}\Big)'\hat{\Omega}_n^{-1}\Big(n^{-1}\sum_{t=m}^{T-\tau} h_t\delta_{t+\tau}\Big) = n\bar{Z}'_{m,n}\hat{\Omega}_n^{-1}\bar{Z}_{m,n} \tag{3.2}$$

[1]Other loss functions are commonly used but are less popular. Important cases are asymmetric loss functions and quasi-likelihood based loss functions. The discussion presented here is valid regardless of the from of the loss function. For further details on loss functions, see Lee (2007).

[2]This test statistic is equivalent to the joint test of significance of the regression of $\delta_{t+\tau}$ on h_t using a HAC covariance matrix.

is asymptotically justified. Here, $\bar{Z}_{m,n} = n^{-1}\sum_{t=m}^{T-\tau} Z_{m,t+\tau}$, $Z_{m,t+\tau} = h_t\delta_{t+\tau}$, and $\hat{\Omega}_n = n^{-1}\sum_{t=m}^{T-\tau} Z_{m,t+\tau}Z'_{m,t+\tau} + n^{-1}\sum_{j=1}^{\tau-1}\omega_{n,j} \times \sum_{t=m+j}^{T-\tau}[Z_{m,t+\tau}Z'_{m,t+\tau-j} + Z_{m,t+\tau-j}Z'_{m,t+\tau}]$ is the $q \times q$ covariance matrix estimator, where $\omega_{n,j}$ is a weight function such that $\omega_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, ..., \tau - 1$ (e.g. Newey and West (1987)). Under the mild regularity conditions in Assumption 3.1.1 presented below, Theorems 1 and 3 in Giacomini and White (2006) ensure that this statistic asymptotically approaches the distribution of a χ^2_q random variable under the null (3.1), and diverges to infinity under the alternative:

Assumption 3.1.1. *(i) $\{W_t\}$ and $\{h_t\}$ are mixing with ϕ of size $-r/(r-2)$, $r > 2$, or α of size $-r/(r-2)$, $r > 2$; (ii) $E[Z_{m,t+\tau}, i]^{r+\delta} < \infty$ for some $\delta > 0$, $i = 1, ..., q$ and for all t; (iii) $\Omega_n = n^{-1}\sum_{t=m}^{T-\tau} E[Z_{m,t+\tau}Z'_{m,t+\tau}] + n^{-1}\sum_{j=1}^{\tau-1}\sum_{t=m+j}^{T-\tau}(E[Z_{m,t+\tau}Z'_{m,t+\tau-j}] + E[Z'_{m,t+\tau-j}Z_{m,t+\tau}])$ is uniformly positive definite.*

Assumption 3.1.1 (i) imposes weak dependence restrictions on the underlying data and the test functions, allowing for possible heterogeneity. This is an attractive feature, since heterogeneity is believed to be a common feature of economic and financial data. Assumption 3.1.1 (ii) imposes existence of at least second moments, and (iii) imposes that the long-run variance matrix of $Z_{m,t}$ is well-behaved. If the data satisfy Assumption 3.1.1, a level α test can be constructed which rejects the null hypothesis of equal conditional predictive ability when $GW_{n,m} > \chi^2_{q,1-\alpha}$, where $\chi^2_{q,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2_q distribution.

This test statistic depends on the size of the out-of-sample period n for correct asymptotic coverage. The test largely ignores the dimensionality of the test functions, however, which is assumed to be fixed. This assumption can be very limiting, even when q is moderately large. The limitation occurs because the test statistic requires estimation of the long-run variance matrix Ω_n, which is a $q \times q$ sized matrix. In order to estimate the test statistic we thus require estimates of $O(q^2)$ parameters. The number of parameters required to estimate the test statistic grows at a quadratic rate, while the sample size grows linearly. Hence, as the number of test functions increases, we can expect the accuracy of the long-run variance estimator to worsen and the asymptotic analysis which the test relies on to fail.

3.2 Bootstrap Solutions

When the number of test functions is large the Giacomini and White (2006) test statistic (and Wald-type test statistics in general) can be unreliable due to the dimensionality of the long-run variance matrix estimator. Here, we propose bootstrap alternatives which are also asymptotically valid but do not require the direct estimation of this variance matrix. To ensure asymptotic validity, however, these alternative tests require stronger assumptions on the underlying data process.

First, we drop the m notation and exclude the possibility of nested models. In the tests of Diebold and Mariano (1995) and Giacomini and White (2006), f and g are forecasting methods which can be nested, and the m subscript is emphasized to denote that the test depends on evaluating the forecasting methods in finite samples rather than at their limit. Here, we exclude nested models, so that parameter estimates for f and g never coincide in the limit nor in finite samples. Second, we exclude the possibility of non-stationary heterogeneity in the loss-difference sequence and impose that Z_t is strictly stationary. Third, we assume existence of first moments of Z_t and concentrate on the $1 \times q$ vector of parameters $\theta = E(Z_t)$, writing the null (3.1) as the restriction $H_0 : \theta = 0$. Lastly, we assume the additional assumptions in West (1996) on the data Z_t, and in particular, we require the existence of a long run covariance matrix $\Omega = E(Z_t Z_t')$. Since we are interested in testing restrictions on the vector θ, we focus on $J_n(\mathbb{P})$, the sampling distribution of the centered and scaled means, $\sqrt{n}(\bar{Z}_t - \theta)$. Finally, we add the following assumption from Romano and Wolf (2005) which embeds the assumptions in West (1996) on the data structure and those in White (2000) on the (bootstrap) estimate of the data structure:

Assumption 3.2.1. *(i) $\mathbb{P}$ is the true probability law driving Z_t; (ii) $J_n(\mathbb{P})$ converges in distribution to a limit distribution $J(\mathbb{P})$ which is continuous; (iii) If $\hat{\mathbb{P}}_n$ is an estimate of $\mathbb{P}$ based on the data $\{Z_1, ..., Z_n\}$, $J_n(\hat{\mathbb{P}})$ consistently estimates $J(\mathbb{P})$, so that $\rho(J_n(\hat{\mathbb{P}}), J(\mathbb{P})) \to 0$ in probability for every metric ρ metrizing weak convergence.*

Assumption 3.2.1 (i) imposes a probability structure on the data. The convergence condition (ii) forces the limiting distribution of $J_n(\mathbb{P})$ to be non-degenerate, and in the context of tests of conditional predictive ability this holds so long as the two forecasting models in question are non-nested, as shown in West (1996). When condition (ii) is satisfied, condition (iii) follows when $\mathbb{P}$ is a stationary distribution and $\hat{\mathbb{P}}$ is the bootstrap distribution generated from the stationary bootstrap of Politis and Romano (1994) or related block bootstrap procedures, as proven in White (2000).

When Assumption 3.2.1 is satisfied, the results in West (1996) imply that $J_n(\mathbb{P}) = \sqrt{n}(\bar{Z} - \theta) \xrightarrow{L} N(\theta, \Omega)$. Traditional Wald-type tests consistently estimate Ω and form the GW_α test statistic introduced above, relying on this asymptotic distribution to test the zero-mean null, $H_0 : \theta = 0$. As we noted earlier, however, moderately large q can render estimation of the Ω matrix unreliable. To avoid estimation of the covariance matrix we note that under this null, $\sqrt{n}|\bar{Z}| \xrightarrow{L} |Z|$, where $Z \sim N(0, \Omega)$. Moreover, under this null we also have:

$$\max_{s=1,...,q} \sqrt{n}(|\bar{Z}_s|) \xrightarrow{L} \max_{s=1,...,q} |Z_s|$$

In general, there is no closed form for the distribution of a maximum of the absolute value of a correlated Normal random vector. As noted in White (2000), however, we can consistently estimate this distribution with the bootstrap.

3.2 Bootstrap Algorithm

We use the bootstrap to approximate the sampling distribution of the centered and scaled $\bar{Z}$ and rely on Assumption 3.2.1 condition (iii) to achieve asymptotic validity. Specifically, we resample B sequences of length n with replacement from the estimated sequence $\{Z_1, ..., Z_n\}$. If there is stationary time-series dependence in this sequence, resampling should be done via the stationary bootstrap, or a block bootstrap alternative that satisfies Assumption 3.2.1 condition (iii). For each estimated bootstrap sample $b = 1, ..., B$ we compute the scaled absolute value of the $q \times 1$ vector of mean-differences, $\sqrt{n}|\bar{Z}_b^* - \bar{Z}|$, where we use the $*$ superscript to denote estimates from bootstrapped re-samples. For each such bootstrap vector we find the maximum element, giving a sampling distribution $J(\hat{\mathbb{P}})$ of the scaled maximum of the absolute values. We can then use the percentile method of Hall (1998) to construct a level α test of the null, as detailed in the algorithm below:

Algorithm 3.2.1. *To test $H_0 : \theta = 0$:*

1. *Compute the vector of means, $\bar{Z} = n^{-1} \sum_{t=1}^{n} Z_t$ and the maximum of the absolute values of $\bar{Z}$, $z \equiv max_{s=1,...,q}|\bar{Z}|$*

2. *For each $b = 1, ..., B$, generate bootstrap samples of length n, $\{Z_{b,1}^*, ..., Z_{b,n}^*\}$, using the stationary bootstrap. For each such bootstrap sample, estimate the mean of the bootstrap sample $\bar{Z}_b^* = n^{-1} \sum_{t=1}^{n} Z_{b,t}^*$, center by the estimated mean of the original sample, and take absolute values to obtain the sequence $\{V_b^*\}_{b=1}^B = \{|\bar{Z}_b^* - \bar{Z}|\}_{b=1}^B$.*

3. *Find the maximum across all q elements of each $1 \times q$ vector in each bootstrap sample element $v_b^* = max_{s=1,...,q} V_b^*$. The collection of maxima across the B bootstrap samples gives the empirical distribution function $\hat{F}(v)$ estimate of the desired distribution.*

4. *Find the quantile $v_{1-\alpha} = min\{v : \hat{F}(v) \geq 1 - \alpha\}$, and reject the null if $z > v_{1-\alpha}$.*

By Corollary 2.7 and Proposition 2.8 in White (2000), the test in Algorithm 3.2.1 has asymptotic size α under the null and under the alternative the probability of its rejection approaches 1 as the sample size n diverges. Note that this test is asymptotically valid, but does not require estimation of the long-run covariance matrix Ω. Hence, it is considerably less susceptible to estimation error arising out of a large value of q than the Giacomini and White (2006) test.

3.2 Studentized Bootstrap Algorithm

The results of Hansen (2005) and Romano and Wolf (2005) suggest that studentization in Algorithm 3.2.1 is likely to generate improvements in power and may also improve size. Hall (1992) and Gotze and Kunsch (1996) show that the bootstrap applied to studentized statistics can provide asymptotic refinements in size. This arises out of improvements to bootstrap approximations when using pivotal quantities, which do not depend on the asymptotic distribution's parameters. In the algorithm presented below, studentization does not create exact pivotal quantities (due to covariance terms), but might still allow finite sample improvements from the univariate variances. As we see in the numerical section, improvements to size because of asymptotic refinement are possible but may be of a questionable order of magnitude. On the other hand, studentization can lead to large improvements in power[3]. This arises because studentization rescales each series to the same unit of standard deviation, thus allowing for 'apples-to-apples' comparisons between series. For instance, if one series has power against the null and another does not, and the first series has much smaller standard deviation, then the first series will be masked by the second and seldom appear as the maximum. This can lead to acceptance of the null under the alternative. Hence, we add the following assumption and modify the algorithm by studentization:

Assumption 3.2.2. *Let* $\hat{\sigma}_s = \sqrt{n^{-1}\sum_{t=1}^{n}(Z_{t,s} - \bar{Z}_s)^2}$ *and* $\sigma^*_{b,s} = \sqrt{n^{-1}\sum_{t=1}^{n}(Z^*_{b,t,s} - \bar{Z}_s)^2}$. *Assume that for each* $s = 1, ..., q$, *both* $\sqrt{n}\hat{\sigma}_s$ *and* $\sqrt{n}\sigma^*_{b,s}$ *converge to* σ_s *in probability.*

Algorithm 3.2.2. *To test* $H_0 : \theta = 0$*:*

1. *Compute the vector of studentized means for each* $s = 1, ..., q$, $\zeta_s = \frac{\bar{Z}_s}{\hat{\sigma}_s}$ *and the maximum of the absolute values of* ζ_s, $\zeta \equiv max_{s=1,...,q}|\zeta_s|$

2. *For each* $b = 1, ..., B$, *generate bootstrap samples of length* n, $\{Z^*_{b,1}, ..., Z^*_{b,n}\}$, *using the stationary bootstrap. For each such bootstrap sample, estimate the mean of the bootstrap sample* $\bar{Z}^*_b = n^{-1}\sum_{t=1}^{n} Z^*_{b,t}$ *and the bootstrap standard deviation* $\sigma^*_{b,s}$ *for each series* $s = 1, ..., q$. *Center by the estimated mean of the original sample and scale by the bootstrap standard deviation, and take absolute values to obtain the sequence* $\{U^*_{b,s}\}_{b=1}^{B} = \{|(\bar{Z}^*_{b,s} - \bar{Z}_s)/\sigma^*_{b,s}|\}_{b=1}^{B}$.

3. *Find the maximum across all* q *elements of each* $1 \times q$ *vector in each bootstrap sample element* $u^*_b = max_{s=1,...,q}U^*_{b,s}$. *The collection of maxima across the* B *bootstrap samples gives the empirical distribution function* $\hat{F}(u)$ *estimate of the desired distribution.*

4. *Find the quantile* $u_{1-\alpha} = min\{u : \hat{F}(u) \geq 1 - \alpha\}$, *and reject the null if* $\zeta > u_{1-\alpha}$.

[3]For details, see Appendix C in Romano and Wolf (2005) or Example 4 in Hansen (2005).

Assumptions 3.2.2 and 3.2.1 ensure that the studentized bootstrap test provides correct size and power behavior asymptotically, and may improve the size and power of the regular bootstrap test in finite samples when the series Z_t are heteroskedastic.

3.2 Stepwise Algorithm

The tests proposed in Algorithms 3.2.1 and 3.2.2 are viable bootstrap alternatives to the Wald-type test, but they do not indicate which of the tested parameters are greater than zero under the alternative hypothesis[4]. When these tests are rejected, they indicate that one or more of the elements in the parameter vector θ is nonzero. It may be that several elements of this parameter vector θ are nonzero and it is of interest to the researcher to find these elements. For instance, in the context of rejection of the test of equal conditional predictive ability, the researcher may be interested in finding the test functions that are correlated with the loss differentials. A test function that is correlated with the loss differentials reveals that one model is weaker than the other because the first model is missing useful predictive information embedded in the test function. Determining which test function results in rejection of the null is hence very useful information to the researcher, as that variable should probably be included in the candidate predictive model.

To find the parameters that are greater than zero, we use the results of Romano and Wolf (2005) on stepwise testing procedures that control the familywise error rate (FWE) asymptotically. In this setup we test the null sequentially, element-by-element, for each parameter θ_s in θ, $s = 1, ..., q$:

$$H_{0,s} : \theta_s = 0 \quad \text{vs.} \quad H_{A,s} : \theta_s \neq 0$$

In this stepwise testing procedure tests are conducted sequentially, controlling each test for the probability of false rejection of its null. As in Romano and Wolf (2005), we accomplish this control by controlling the FWE. This FWE is the probability of at least one test being incorrectly rejected under the joint null probability distribution. Formally, we denote by $I_0 = I_0(\mathbb{P}) \subset \{1, ..., q\}$ the indices of the set of true hypotheses under the probability measure $\mathbb{P}$, so that the FWE is then defined to be:

$$\text{FWE}_\alpha \equiv \text{Prob}\{\text{Reject at least one } H_s : s \in I_0(\mathbb{P})\}^5 \tag{3.3}$$

Traditionally, the FWE has been controlled using the Bonferroni and Holm methods (Harvey and Liu, 2014), which can be extremely conservative. In the context of testing under a composite null, Romano and Wolf (2005) propose an improved stepwise testing

[4]The bootstrap tests indicate that the maximum of θ is greater than zero, but do not make assertions about the other, smaller elements of θ.

[5]In the event that I_0 is empty, the FWE is equal to zero.

algorithm based on the results of White (2000) that asymptotically bounds the FWE below some pre-specified size α. Their results hold for any measure $\mathbb{P}$ satisfying the assumptions above. Specifically, Assumption 3.2.1 ensures that their results are applicable in the context of simple, large-dimensional nulls such as tests of equal conditional predictive ability with many test functions.

We modify Algorithm 3.1 in Romano and Wolf (2005) to test large-dimensional simple nulls such as those in tests of equal conditional predictive ability. We begin by reordering the indices $1, ..., q$ in the order $r1, ..., rq$, so that the absolute value of the mean statistics are in descending order $|\bar{Z}_{r1}| \geq |\bar{Z}_{r2}| \geq ... \geq |\bar{Z}_{rq}|$. We make individual decisions in a stepwise manner in which we maintain a rectangular joint confidence region for the vector θ with nominal joint coverage probability $1 - \alpha$. In the first step, the confidence region takes the form:

$$[|\bar{Z}_{r1}| - c_1, \infty) \times ... \times [|\bar{Z}_{rq}| - c_1, \infty) \tag{3.4}$$

where the critical value c_1 is chosen such that the coverage region has the proper joint asymptotic coverage probability under the null hypothesis $H_{0,r1}$. That is, we choose c_1 such that the *joint* coverage probability is $1 - \alpha$, and the dependence across series is explicit. As in Algorithms 3.2.1 and 3.2.2, we use the bootstrap to find c_1 and we check whether the first confidence interval $[|\bar{Z}_{r1}| - c_1, \infty)$ contains zero. If not, we reject the corresponding null for the maximum value $\theta_{r1} \equiv \max \theta_s = 0$. The test we present here extends these earlier tests and specifies values for subsequent critical values c_j. We use the bootstrap distribution to find these c_j coefficients and test the restrictions on the other $\theta_{rj} \leq \max \theta_s$ parameters sequentially. We reject the nulls if the associated confidence region suggests that the associated parameter is unlikely to be zero. The goal of our algorithm is to determine values for the c_j coefficients $j = 1, ..., q$:

$$[|\bar{Z}_{rj}| - c_j, \infty) \times ... \times [|\bar{Z}_{rq}| - c_j, \infty) \tag{3.5}$$

such that this joint confidence region has asymptotic probability $1 - \alpha$ under the null $H_{0,rj}$. As in the $j = 1$ case, if a particular individual confidence interval $|\bar{Z}_{rk}| - c_j$ contains zero, the respective null $H_{0,rk} : \theta_{rk} = 0$ is rejected for $k \geq j$. Our algorithm computes the sequence c_j of critical values with the bootstrap and conducts the sequential test while controlling FWE.

To compute the c_j coefficients we use the same bootstrap approach used earlier, but we also obtain an estimate of the sampling distribution for subsequent order statistics beyond the maximum. That is, we find all the parameters $\theta_{rj} > 0$ using the following algorithm:

Algorithm 3.2.3. *To test each $H_{0,j} : \theta_{rj} = 0$, $j = 1, ..., q$:*

1. *Set* $j = 1$ *and* $R_0 = 0$

2. *For* $R_{j-1} + 1 \leq k \leq q$*:*

   (a) *Compute the critical value* c_j *using the data* $\{Z_{rk,t}, ..., Z_{rq,t}\}_{t=1}^{n}$ *with Algorithm 3.2.1 and confidence level equal to the FWE.*

   (b) *If* $0 \notin [|\bar{Z}_{rk}| - c_j, \infty)$*, reject the null hypothesis* $H_{0,j}$ *and set* $i = k$*. For* $i > k$*:*

      i. *Reject the null hypothesis* $H_{0,i}$ *if* $0 \notin [|\bar{Z}_{ri}| - c_j, \infty)$*.*

      ii. *Otherwise, let* R_j *be the total number of null hypotheses rejected so far. Then set* $j = j + 1$*,* $k = i$*, break the inner loop and return to step 2.*

   (c) *Else, stop and break loop. No more hypotheses can be rejected.*

Algorithm 3.2.3 finds all the values of θ that are greater than zero such that the probability that any element $\theta_s = 0$ when the algorithm indicated otherwise is smaller than the FWE, asymptotically as n diverges. Despite the complicated multiple testing problem, this procedure allows for careful control of 'size' (via FWE) at little additional cost to the user. Moreover, by using Algorithm 3.2.2 and studentized test statistics, Algorithm 3.2.3 can be easily extended to potentially improve size and power properties. Note that the first step of this algorithm is the same as Algorithm 3.2.1, while the potential next steps of the algorithm reveal additional information beyond acceptance or rejection of the multi-dimensional null in Algorithms 3.2.1 and 3.2.2. Hence, this multiple testing algorithm provides a 'free-lunch' relative to Algorithms 3.2.1 and 3.2.2, giving the researcher additional feedback on which null restrictions are unlikely to hold true.

To summarize, we have suggested bootstrap procedures that can be used in place of traditional Wald-type tests for testing large-dimensional simple nulls, such as those frequently encountered in tests of equal conditional predictive ability. The traditional tests suffer from the need to estimate a large dimensional variance matrix which grows at a faster rate than the growth rate of data. The bootstrap tests presented here embed the covariance matrix into the resampled data and forego its estimation altogether. This property renders bootstrap tests more rubust to estimation error than the typical test statistic for testing large-dimensional simple nulls. Furthermore, the sequential procedure outlined in Algorithm 3.2.3 finds the particular elements in the tested parameter vector which represent the largest deviations from the null, thereby pointing the researcher towards the reasons for a rejection of the null. To assess these procedures more closely, we now turn to a numerical study which compares these procedures to traditional Wald-type tests like the Giacomini and White (2006) test for equal conditional predictive ability.

3.3 Numerical Study

In this section we investigate the resilience of the Giacomini and White (2006) test and the bootstrap procedures described above to the inclusion of many test functions. Specifically, we examine the size and power properties of the original Wald-type test statistic, and the studentized and non-studentized versions of the test in Algorithm 3.2.3, which we call the 'non-studentized' and 'studentized' tests respectively. We recall that rejection and non-rejection of the null under that bootstrap procedures is determined by the first step in Algorithm 3.2.3, which is simply Algorithm 3.2.1 for the non-studentized case and Algorithm 3.2.2 for the studentized case. If the first step is rejected, Algorithm 3.2.3 then proceeds and finds all other order statistics that are unusually large given their sampling distribution. Hence, the size and power of the test are determined purely by the first step, and the following steps provide additional information on the sources of deviation from the null.

For comparative convenience, we follow a similar setup to the Monte Carlo analysis of Giacomini and White (2003)[6]. As in Giacomini and White (2003), we limit this numerical discussion to mean-square-error losses, since they are often encountered in practice. We largely focus on stationary simulation scenarios where all of the tests above are valid, but in the end of our discussion we conduct sensitivity analyses to changes in parameters as well as to heterogeneity in the data structure.

3.3 Size Properties

We first investigate the properties of the Giacomini and White (2006) test and the alternative studentized and non-studentized bootstrap tests under the null hypothesis. Instead of applying the tests directly to a simple process satisfying the null (e.g. white noise), we use the simulation procedure of Giacomini and White (2003). This simulation more closely parallels the forecast evaluation process of practical researchers and it will allows us to conduct a comparable sensitivity check of our tests to heterogeneity in the data (see sensitivity check below). To that end, we note the following result from Giacomini and White (2003):

Proposition 3.3.1. *$E[(Y_{t+1} - \hat{f}_t)^2 - (Y_{t+1} - \hat{g}_t)^2|\mathcal{F}_t] = E[\Delta L_{t+1}|\mathcal{F}_t] = 0$ if and only if either $\hat{f}_t = \hat{g}_t$ almost surely, or $E[Y_{t+1}|\mathcal{F}_t] = (\hat{f}_t + \hat{g}_t)/2$.*

Importantly, the proposition asserts that the mean-square-error loss differences between *any* two forecasts is a MDS if the conditional expectation of the predicted process is

[6]We use the working paper of Giacomini and White (2006), which differs from the published version by comparing non-nested models instead of nested ones. Since the asymptotic justification for our bootstrap procedures relies on non-nested models, we feel the working paper numerical study setup is more appropriate.

an average between the two forecasts. To generate data under the MDS null hypothesis $H_0 : E[\Delta L_{t+1}|\mathcal{F}_t] = 0$, we hence construct Y_{t+1} sequences using this approach. Our setup then differs slightly from that of Giacomini and White (2003), who consider heterogeneous data for construction of the forecast sequences. Here, we exclude the possibility of heterogeneity and focus on a strictly stationary underlying data series. We do this by first constructing a sample of data to be used for the forecasts, $W_t = N(0, 1)$, where $t = 1, ..., 1120$. Using a rolling look-back window of 120 observations ($m = 120$ in the language of Giacomini and White (2006)), we then construct $n = 1000$ forecast sequences $\{\hat{f}_t, \hat{g}_t\}_{t=121}^{1120}$ of $\{W_t\}_{t=121}^{1120}$. As in Giacomini and White (2003), the forecast f is computed as a rolling mean using the previous 120 observations, and the forecast g is an AR(1) model-implied forecast:

$$\hat{f}_{m,t} = (W_{t-m+1} + ... + W_t)/m \tag{3.6}$$

$$\hat{g}_{m,t} = \hat{\alpha}_{m,t} + \hat{\beta}_{m,t} W_t \tag{3.7}$$

With these forecasts, we then define $Y_{t+1} = (\hat{f}_t + \hat{g}_t)/2 + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \sim N(0, \sigma^2)$, and $\sigma = 1$. By Proposition 3.3.1, we hence obtain a 1000-observations-long martingale-difference-sequence of forecast-loss-differences. Furthermore, if we simulate K sequences of noise, $\{\varepsilon_{t+1}\}_{t=121}^{1120}$, we obtain K such Monte Carlo replications of the martingale-difference-sequence.

Recall that we are primarily interested in the resilience of each of the tests to the dimension of the test function. We hence choose $h_t \sim N(0, I_q)$, which is independent of $\{\hat{f}_t, \hat{g}_t, \varepsilon_{t+1}\}$, and we vary $q \in \{5, 10, 20, 30, 100, 300\}$. Since the test function is independent of the loss-differences, the MDS null is preserved. Since the test of Giacomini and White (2006) assumes q is fixed, the number of test functions should not affect the p-values of the test. But, as the discussion in the previous section suggested, we expect the Giacomini and White (2006) test statistic to have considerably worse size as q increases. Moreover, we expect the bootstrap procedures to be considerably more robust to the size of q than the Giacomini and White (2006) test.

To determine the level of each test, we conduct $K = 1000$ simulated replications for each configuration of q and estimate each test on each one of these replicated sequences. For the bootstrap procedures, we use 1000 bootstrap replications[7] to conduct the test. We collect the 1000 p-values for all the tests and present them as empirical CDFs in Figure 3.1 below.

[7]In preliminary analyses we did not find that a greater number of bootstrap samples was more informative.

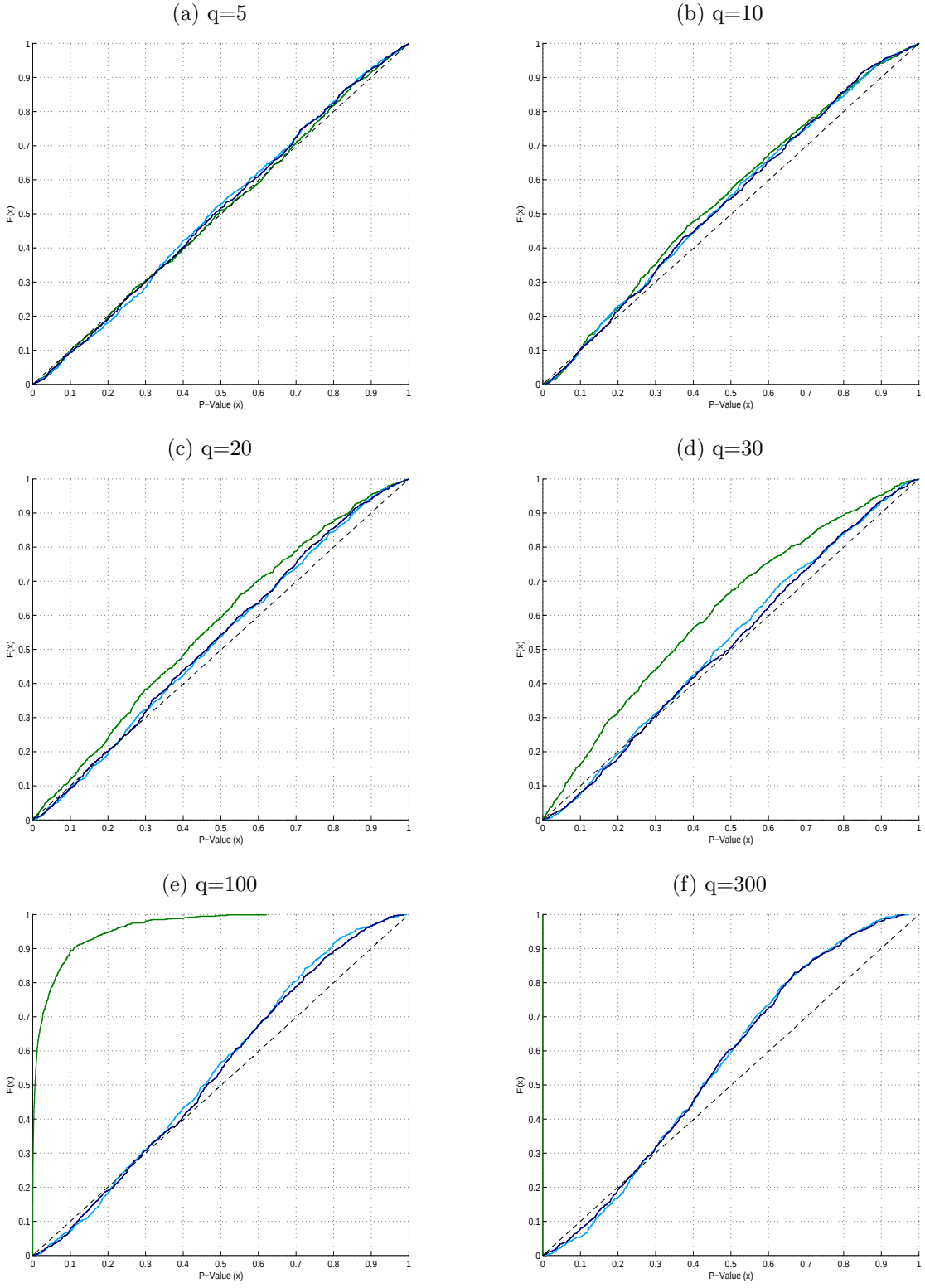


Figure 3.1: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the dimension of the test function q is varied under the null. The test function is homoskedastic and the out-of-sample size is 1000.

Figure 3.1 shows the empirical CDFs of the p-values of each test as the dimension of the test function, q, is varied. Accurate size of any test is reflected by an empirical CDF of p-values that is close to the 45% line (dashed black line), so that the p-values are distributed uniformly on $(0, 1)$. When q is small, we observe that the empirical CDFs of all three tests are close to this 45% line. As q increases, we note a tendency of the Giacomini and White (2006) test to over-reject. When q is large (100 or 300), the Giacomini and White (2006) test nearly always rejects. We note that the bootstrap tests also have slightly distorted size when q increases, but the distortion is considerably smaller. Moreover, the bootstrap tests tend to be conservative for common levels of tests ($\alpha \leq 0.10$), so type I error is limited below the level of the test. Hence, the bootstrap tests are considerably more robust to the number of test functions, and their few failings are generally a 'lesser of two evils' to the researcher.

The results of Figure 3.1 suggest that the studentized and non-studentized bootstrap tests perform comparably, and the need to studentize seems unnecessary. Note that in this case all the test functions have the same variance, and studentization does not act as a pivot. To determine the effects of studentization, we conduct the same analysis with the test function $h_t \sim N(0, \Sigma)$, where $\Sigma = diag(1, .., q)$. As before, the test function is independent of the loss-differences and the martingale-difference-sequence null is preserved, albeit with differing variances across the series.

We document the effect of heteroskedasticity between the data series in Figure 3.2. The results of Figure 3.2 show that studentization has very little effect on size. We conjecture that the potential asymptotic refinements arising out of partially-pivotal quantities are minor in this simulation. Heuristically, we note that under the null the scale of the series is irrelevant, since the maximum of all series is centered around zero. If one series had belonged to the alternative, its effect would have been masked if that series had low variance. In the next section, we observe that the main advantages to studentization are hence mostly motivated by power under the alternative rather than size under the null.

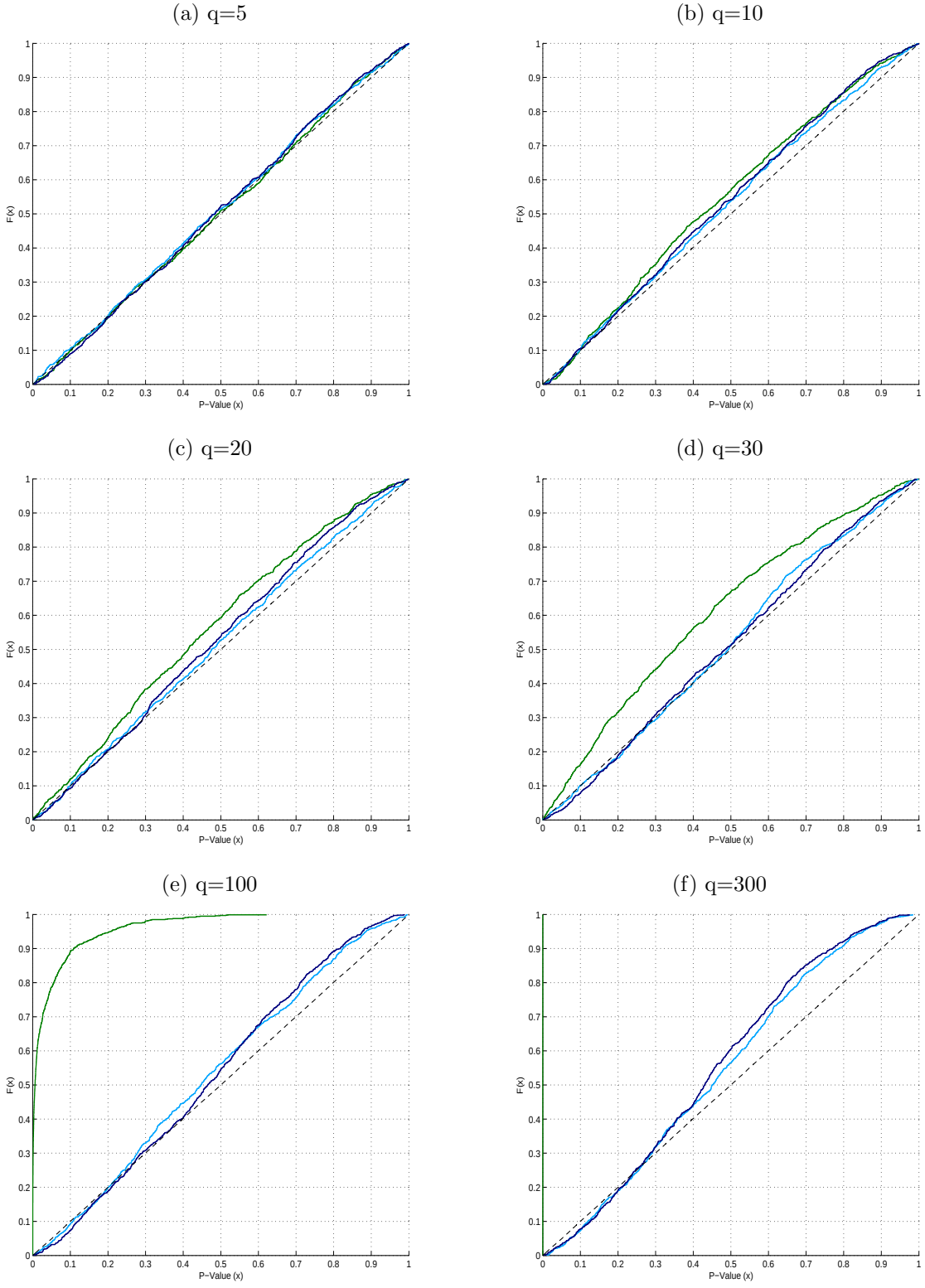


Figure 3.2: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the dimension of the test function q is varied under the null. The test function is heteroskedastic and the out-of-sample size is 1000.

3.3 Power Properties

In keeping with the Monte Carlo analysis of Giacomini and White (2003), we analyze the power properties of our tests against serially correlated alternatives. As in their analysis, we consider the alternative:

$$H_{a,\rho} : E[\Delta L_{t+1}|\mathcal{F}_t] = \rho \Delta L_t \tag{3.8}$$

where ρ is an autocorrelation parameter and where we use q lags of the loss differences as test functions. We simulate data under this alternative by setting $Y_{t+1} = (\hat{f}_t + \hat{g}_t)/2 - \rho \Delta L_t/(2(\hat{f}_t - \hat{g}_t)) + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \sim N(0, \sigma^2)$, and $\sigma = 1$.

To gain a clear understanding of the probability of rejection under different alternatives, we allow for varying parametrizations under the alternative. We vary the degree of autocorrelation ρ in the loss-difference-sequence from 0.05 to 0.5 in increments of 0.05. We expect all of the tests to reject more often when serial correlation is higher, since higher autocorrelation leads to an alternative that is further from the null. The forecast values $\hat{f}_t$ and $\hat{g}_t$ are generated as before, and the initial values of ΔL_t are drawn from $N(0, 1)$. Again, we generate 1000 observations ($n = 1000$), 1000 bootstrap samples for the tests, and 1000 Monte Carlo replications for each parametrization (ρ, q).

Recall that we are primarily interested in the effects of q on rejection of the null under the alternative (3.8). From the results of Figure 3.1, we learned that when there are many test functions the Giacomini and White (2006) test tends to over-reject whereas bootstrap tests do not. Hence, we expect that under the alternative, the Giacomini and White (2006) test (with many test functions) may have artificially slightly higher rejection probability due to (favorable) estimation error in the test statistic. We investigate the power of all three tests in Figure 3.3, which shows the empirical CDFs of the p-values of all tests for various values of ρ and q.

Figure 3.3 shows that the power of the bootstrap procedures is largely unhindered by the number of test functions. For more distant alternatives from the null, that is, larger values of ρ, the bootstrap procedures appear to outperform the Giacomini and White (2006) test. This is particularly more acute when q is large, and many weak test functions are added. The exception to this success is when the alternative is close to the null, the number of test functions is large, and the studentized version of the bootstrap has weaker power than the Giacomini and White (2006) test. We note that the variances of all series are the same, so we expect the power of the studentized bootstrap to be weaker than the non-studentized bootstrap, since all series are in the same 'units of standard deviation'[8]. The non-studentized bootstrap and the Giacomini and White (2006) test perform comparably when they outperform the studentized bootstrap, so we suspect that this outperformance is purely due to unnecessary studentization. To analyze situations

[8]For more details, see Hansen (2005).

in which studentization may be beneficial, we also consider the more appropriate case where the series are heteroskedastic.

To examine the heteroskedastic case, we scale the test functions (lagged loss-differences) by a constant corresponding to the lag number. That is, we scale the j'th lagged loss-differences by j, so that higher-order lags (with less predictive ability than lower order lags) have higher variance. As explained above, when studentization is not performed, alternatives that are closer to the null (those associated with larger lags) with higher variances may lead to incorrect acceptance of the null (type II error). Under this simulation scenario, we expect that the non-studentized bootstrap will perform poorly and that studentization will improve power considerably.

In Figure 3.4 we see the benefits of studentization to power when the data series are heteroskedastic. The non-studentized bootstrap test preforms very poorly under this scenario, since the least informative test functions are also the noisiest ones, and the noise is not studentized. These weak test functions lead to parameters that are very close to the null, and the non-studentized bootstrap estimate of the sampling distribution of the maximum is hence heavily perturbed by them. By contrast, the studentized bootstrap test preforms very well, and mostly even better than the Giacomini and White (2006) test.

We may also be interested in the success of the stepwise testing procedure at identifying all the parameters θ that are greater than zero. We note that $\theta_1 > \theta_2 > ... > \theta_q > 0$ under the serial correlation alternative, so we expect the first-lag series to be rejected most often, then the second-lag series most often, and so on. We plot the frequency of rejections at the 10% level across simulations and parametrizations for each lag in Figure 3.5 for the homoskedastic case, and in Figure 3.6 for the heteroskedastic case. Figure 3.5 shows that the frequency of rejections for both studentized and non-studentized bootstrap tests decays gradually with the lag length, much like an autocorrelation function. When the autocorrelation parameter ρ is larger, the frequency of rejection is higher across the board, as expected. Further, as in Figure 3.3, Figure 3.5 shows that studentization can lead to fewer rejections than non-studentization under the homoskedastic alternatives. Finally, under the heteroskedastic alternatives the non-studentized bootstrap procedure performs very poorly, hardly ever rejecting. On the other hand, the studentized procedure performs very well and rejects frequently. In all, the stepwise procedure rejects frequently for alternative hypotheses far away from the null, and infrequently for alternative hypotheses that are near to the null.

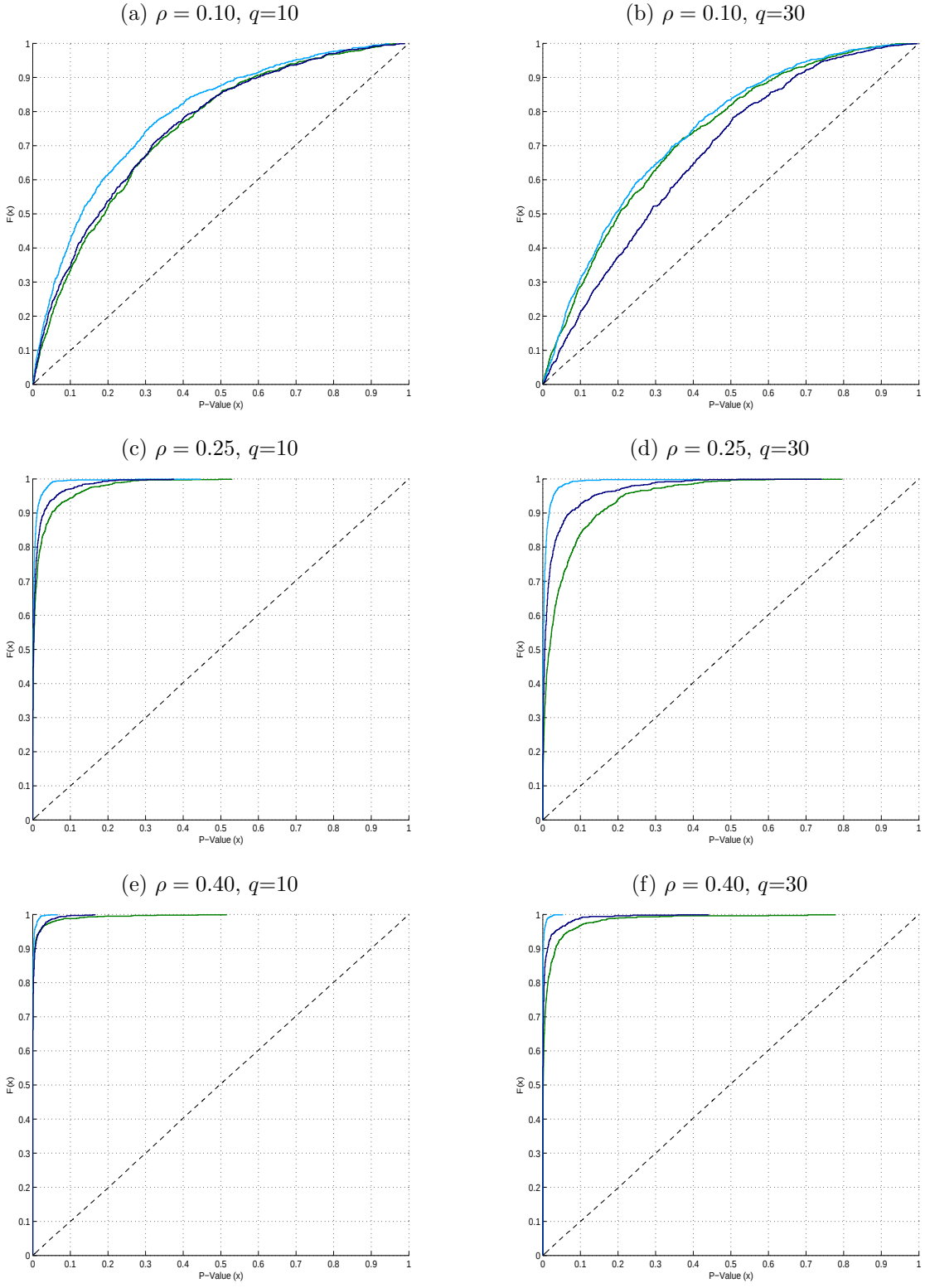


Figure 3.3: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the dimension of the test function q is varied under three alternatives: $\rho = \{0.10, 0.25, 0.40\}$. The test function is homoskedastic and the out-of-sample size is 1000.

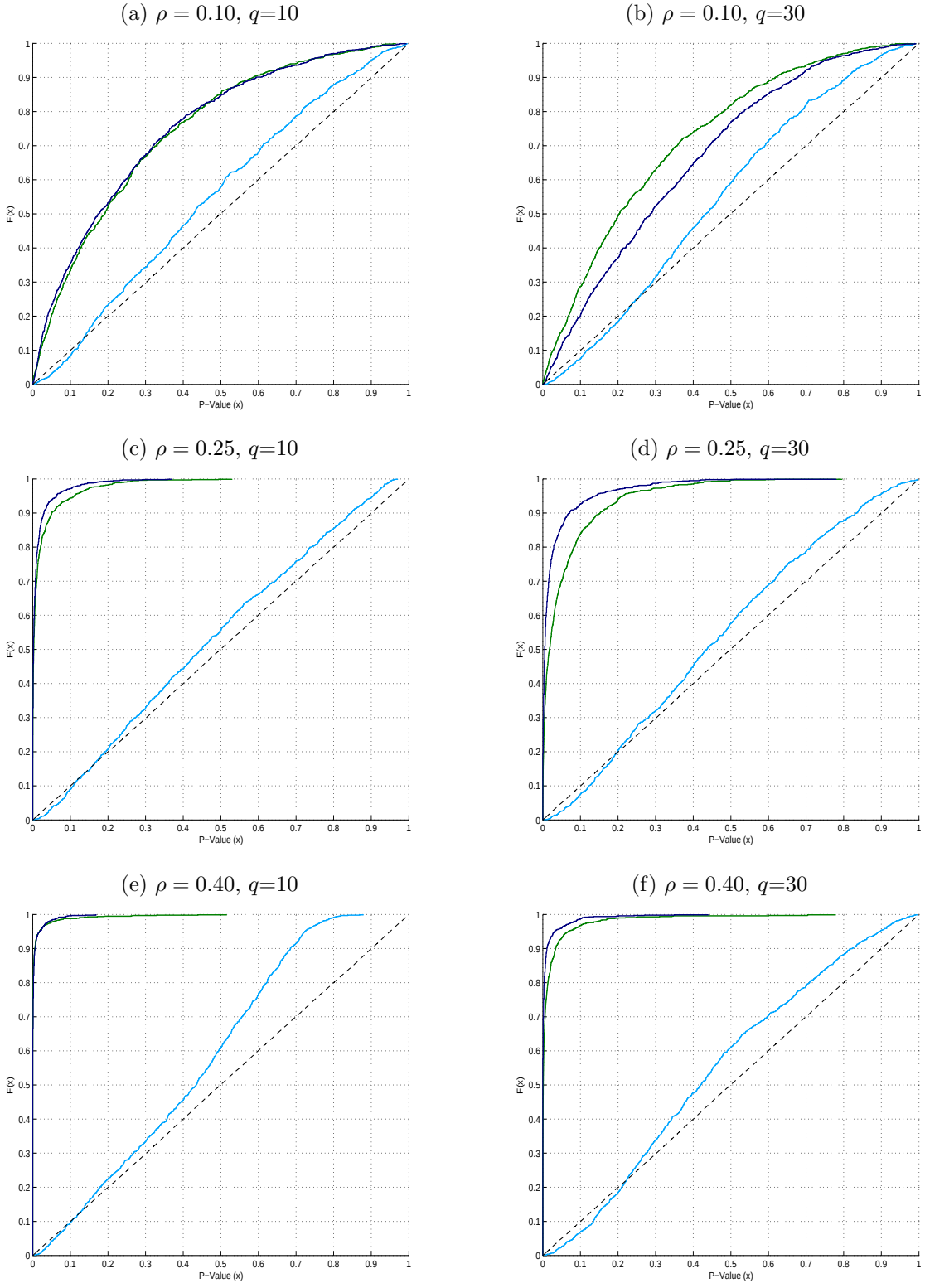


Figure 3.4: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the dimension of the test function q is varied under three alternatives: $\rho = \{0.10, 0.25, 0.40\}$. The test function is heteroskedastic and the out-of-sample size is 1000.

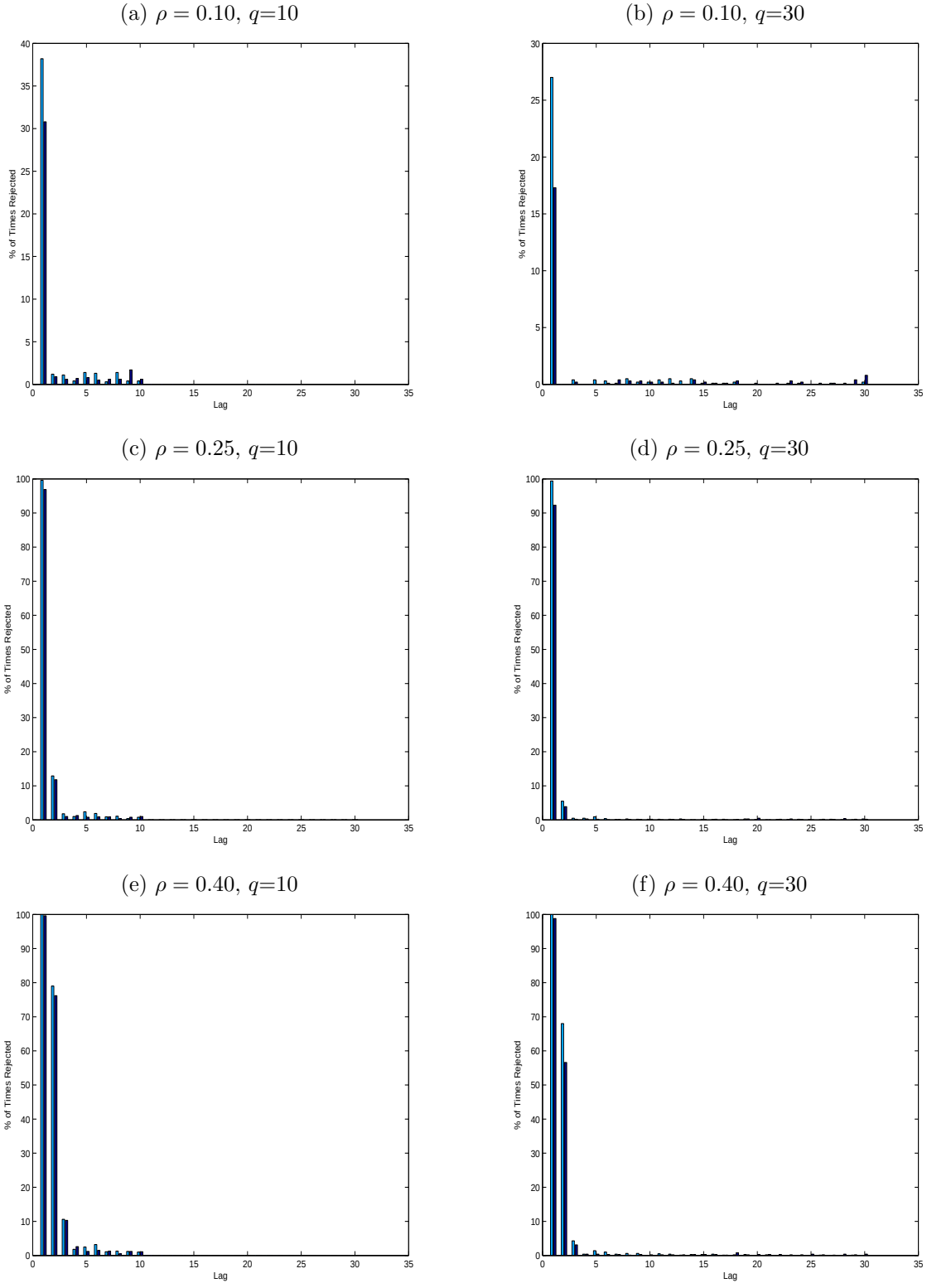


Figure 3.5: Histogram of the frequency of rejection of each lag across the Monte Carlo replications of the non-studentized bootstrap test (light blue) and the studentized bootstrap test (dark blue) at the 10% level as the dimension of the test function q is varied under three alternatives: $\rho = \{0.10, 0.25, 0.40\}$. The test function is homoskedastic and the out-of-sample size is 1000.

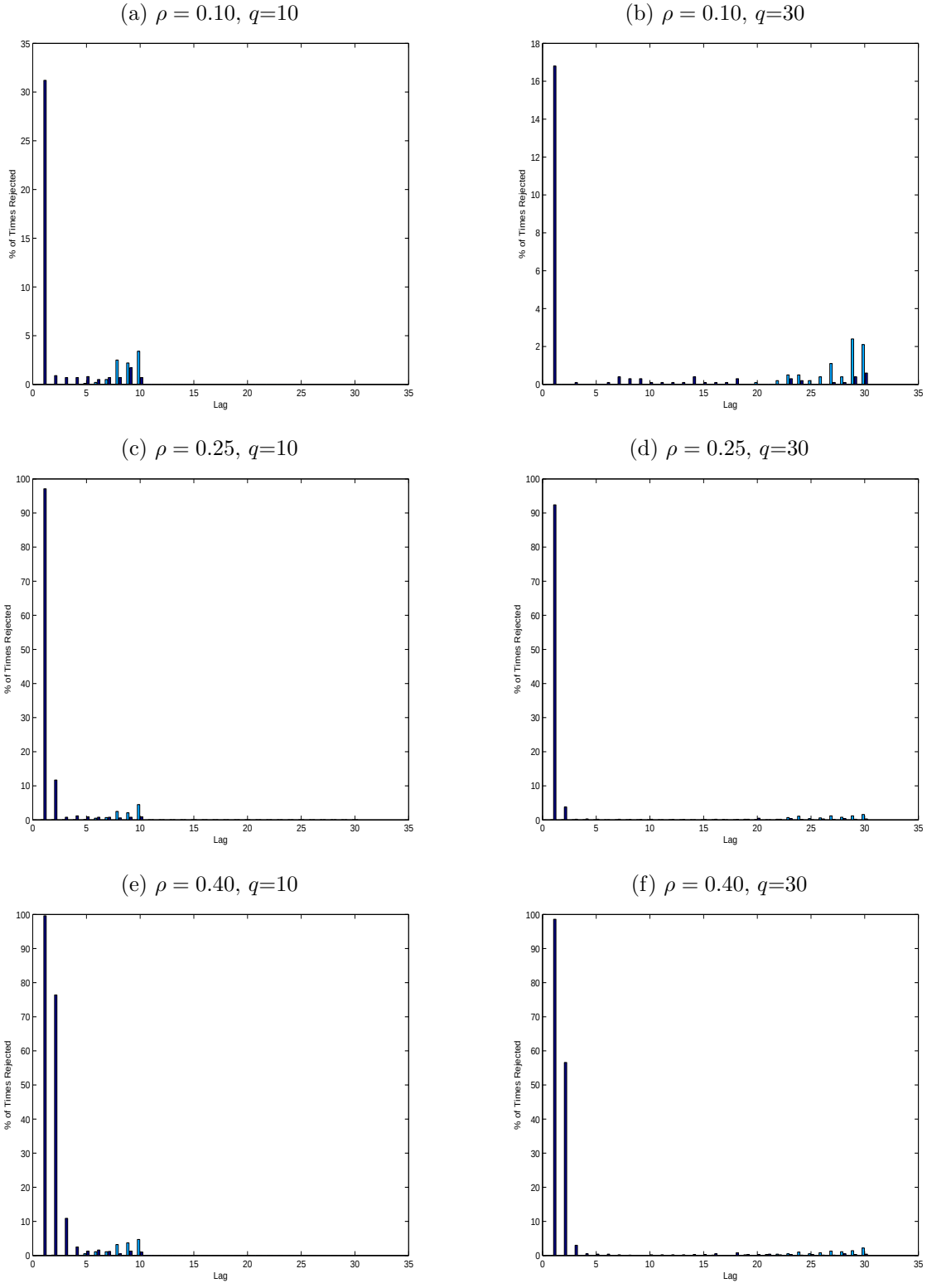


Figure 3.6: Histogram of the frequency of rejection of each lag across the Monte Carlo replications of the non-studentized bootstrap test (light blue) and the studentized bootstrap test (dark blue) at the 10% level as the dimension of the test function q is varied under three alternatives: $\rho = \{0.10, 0.25, 0.40\}$. The test function is heteroskedastic and the out-of-sample is 1000.

3.3 Sensitivity to Parameters and Data Heterogeneity

Practical research frequently involves testing under weaker assumptions than those above. To determine the efficacy of the bootstrap procedures above under different situations, we analyze the sensitivity of the bootstrap procedures to other parameter values and heterogeneity in the data. All figures in this section can be found in Appendices A and B.

First, we compute the tests under different look-back horizons and forecast error volatilities. We allow for variation in the variance of the noise, $\sigma^2 \in \{0.10, 1, 3\}$, and variation in the look-back horizon, $m \in \{36, 60, 120\}$. For simplicity, we fix the number of test functions at $q = 30$ and analyze the effect of changing the parameter values individually. Changes in m and σ largely reflect the same effect: an increase in the variance of the forecast error. In Figure 3.7, we see that changes to this variance due to smaller look-back horizon or larger error variance have little effect on the efficacy of these tests on size. As σ is increased, there seems to be little effect on the size of the test. As m increases, however, we see possible slight improvements to size, suggesting that the bootstrap approximation may improve as noise is reduced.

In terms of power, Figure 3.8 shows the empirical CDFs of the p-values of the tests under the alternative $\rho = 0.4$ and $q = 30$ as the variance of the error terms σ and the look-back window m are varied. The variance of the error term has little (if any) effect on power, but the look-back window m appears to improve power as it increases. This is particularly true of the Giacomini and White (2006) test, which has lower power than the bootstrap tests universally, and particularly when the look-back estimation window is low. Hence, we conclude that all tests are robust to error variance. Power is improved, however, for the Giacomini and White (2006) test when the look-back horizon increases, and is largely unaffected for the bootstrap tests.

Next, we examine the robustness of the tests to heterogeneity in the data structure. We repeat the exercise before, but introduce heterogeneity in the forecast errors as in Giacomini and White (2003). We do this by changing the underlying data series W_t to U.S. growth rate in CPI inflation (second differences of log monthly CPI) available in the dataset of Stock and Watson (2012). This data series has 631 monthly observations spanning 1959 to 2011, thus including very different periods in U.S. inflation history, such as the high inflation of the 1970's, the Great Moderation, the low inflation of the 2000's, and the recent 2007-2009 Great Recession. As before, we examine the rolling forecast-error-loss-differences using f and g from above and a look-back of $m = 120$, leaving $n = 511$ out-of-sample observations for the tests[9].

[9]Implicitly we have shortened the out-of-sample window from 1000 to 511 observations. This in itself deteriorates size and power, but this deterioration is even across all tests.

Figure 3.9 shows the size of all the tests under the null with heterogeneous data. We immediately notice that the bootstrap procedures have considerably worse size due to heterogeneity. As before, the size of the Giacomini and White (2006) test is severely deteriorated by the number of test functions used, while the bootstrap test is considerably more robust. This difference is particularly acute at high values of q, where the bootstrap is dominant.

Figure 3.10 shows the power of all the tests under the alternatives $\rho \in \{0.10, 0.25, 0.4\}$ with heterogeneous data. We note the poorer performance of all tests, and particularly the bootstrap tests, which are far too conservative at low-level tests ($\alpha \leq 0.10$). We ascribe the apparent stronger power of the Giacomini and White (2006) test for larger q to be misleading, since once again, larger q results in greater probability of rejection for that test under both the null and the alternative. Thus, we conclude that data heterogeneity deteriorates power considerably for all tests.

The deterioration of size and power for all tests under data heterogeneity suggests that heterogeneity matters a great deal for the performance of these tests. We note that though there is no asymptotic justification for the correct convergence of the bootstrap tests under heterogeneity, size under the null is still more stable for the bootstrap tests than for the Giacomini and White (2006) test, particularly when q is large. Though poor power is a problem for the bootstrap tests, correct size of these tests ensures that type I error is unlikely and a rejected null under the bootstrap tests is more likely to be correct than for the Giacomini and White (2006) test.

To summarize, the numerical study has shown that using many test functions in the Giacomini and White (2006) test can drastically inflate type I error and that bootstrap procedures are robust to this problem under both stationary and heterogeneous data. Furthermore, power is generally slightly improved with the bootstrap procedures when the data source is stationary. Moreover, when the data series are heteroskedastic, studentization can drastically improve the performance of the bootstrap. Finally, the stepwise testing procedure is shown to be adept at finding several of the large deviations from the null when there are multiple such deviations. Hence, the numerical study shows that the procedures above are viable alternatives to the Giacomini and White (2006) test, can lead to large improvements in size and power, and can detect multiple deviations from the null under the alternative.

In the next section, we apply the tests above to U.S. inflation forecasting. With the surfeit of available test functions, we re-examine this important application with the bootstrap procedures since they will likely add confidence to the result with less fear of type I error.

3.4 Modeling U.S. Inflation

In this section, we re-examine the strength of a univariate time-series model relative to the multivariate Phillips curve model in forecasting U.S. inflation. Inflation modeling and forecasting is an important problem in macroeconometric research and policy making. U.S. inflation has evolved considerably over time, and with it, the most commonly used estimation and forecasting models. As noted in Stock and Watson (2006), inflation was highly volatile in the 1970's and early 1980's, and this volatility sharply declined during the Great Moderation, post-1984 and throughout the 1990's. Forecasting inflation has hence become 'easier' in some sense, because root-mean-square-errors have become considerably lower as a result of this reduced volatility. In another sense, however, inflation has become more difficult to forecast. Specifically, it has been shown in Atkeson and Ohanian (2001) that simple univariate models improve annual inflation forecasts over multivariate models, and in particular, the previously popular backward-looking Phillips curve models. A backward-looking Phillips curve model posits that inflation is related to lagged unemployment, or some other measure of real economic activity. In the past, it was thought that real activity variables help account for the variation in inflation (Gordon, 1997). However, the studies of Atkeson and Ohanian (2001) and Stock and Watson (2006) have shown that these real indicators have little predictive accuracy for forecasting U.S. inflation after 1984, and are beaten by univariate models throughout the 1990's. These papers attribute the subsequent failure of Phillips curve models to changes in the economy, and these changes do not suggest a stable relationship between the unemployment rate and inflation. Moreover, this empirical finding is supported by theoretical motivations. The theoretical analyses in Friedman (1968), Lucas (1972), Phelps (1969), Fischer (1977), and Taylor (1980) do not suggest a compelling reason for a stable relationship between unemployment and inflation. They instead demonstrate that the relationship between inflation and unemployment should shift with changes in agents' expectations, which in turn fluctuate with the state of the economy. Furthermore, the results of Stock and Watson (2006) and Stock and Watson (2010) suggest that a more sophisticated univariate model, the unobserved-component-stochastic-volatility (UCSV) model, can forecast inflation better than Phillips curve forecasts before and after 1984. The success of this UCSV model is attributable to the model's ability to detect shifts in the state of the economy.

Here, we use the bootstrap procedures described above to conduct tests of equal conditional predictive ability for U.S. CPI and PCE inflation forecasts under a common bookward-looking Phillips curve and the UCSV model. Unlike previous studies, we use one-month ahead forecasts from monthly data in Stock and Watson (2012), which is more granular than the quarterly data in Stock and Watson (2010) and which additionally includes the zero-lower-bound, post-Great Recession period. Moreover, this dataset

contains information on other important macroeconomic variables that can be used as test functions. Thus, our analysis provides a novel comparison in terms of both granularity of data and the use of many test functions to determine a controversial and important macroeconomic relationship.

Following the strong results of Atkeson and Ohanian (2001), univariate inflation forecasting models such as the UCSV model have seen considerable success. The UCSV model assumes inflation is a noisy series around a trend which evolves according to a random walk, where the noise volatility in both the trend and in the actual series evolves according to a stochastic volatility process. The UCSV model can be written in state-space form as:

$$\begin{aligned} \pi_t &= \tau_t + \eta_t, \quad \eta_t = \sigma_{\eta,t}\xi_{\eta,t} && (3.9)\\ \tau_t &= \tau_{t-1} + \varepsilon_t, \quad \varepsilon_t = \sigma_{\varepsilon,t}\xi_{\varepsilon,t} && (3.10)\\ \log(\sigma^2_{\eta,t}) &= \log(\sigma^2_{\eta,t-1}) + \nu_{\eta,t} && (3.11)\\ \log(\sigma^2_{\varepsilon,t}) &= \log(\sigma^2_{\varepsilon,t-1}) + \nu_{\varepsilon,t} && (3.12) \end{aligned}$$

where $\xi_t = (\xi_{\eta,t}, \xi_{\varepsilon,t})'$ is distributed $N(0, I_2)$, $\nu_t = (\nu_{\eta,t}, \nu_{\varepsilon,t})$ is i.i.d. $N(0, \gamma I_2)$ and ξ_t and ν_t are independent, with $\gamma = 0.013$[10]. This model is computed by Markov Chain Monte Carlo (MCMC) using a diffuse, inverse-Wishart prior for the initial condition. In Figure 2 of Stock and Watson (2006), the authors note that the volatility of the trend component fluctuates considerably over time while the transitory volatility remains stable. This change in permanent volatility accounts for the strong performance of this model over both high inflation volatility periods (pre-1984) and low inflation volatility periods (post-1984).

By contrast, the backward-looking Phillips curve model allows for spillover from lagged real variables onto inflation. In particular, unemployment was suggested as an indicator of future inflation based on an observed empirical negative relationship between the two variables in Phillips (1958) and Samuelson and Solow (1960). In inflation forecasting applications such as Stock and Watson (2010), the relationship is written as an autoregressive-distributed-lag (ADL) model:

$$\Delta\pi_t = \mu + \sum_{j=1}^{p} \alpha_j \Delta\pi_{t-j} + \sum_{i=1}^{r} \beta_i u_{t-i} + \nu_t \tag{3.13}$$

where u_t is the unemployment rate and ν_t is a mean-zero error term. The model (3.13) is generally estimated using least-squares and direct forecasts are used.

The Phillips curve and UCSV models fundamentally differ. The Phillips curve model uses lagged information about inflation, but unlike the UCSV model, it also incorporates

[10]The specific choice of this scale parameter has little impact on the trend and on the forecast.

multivariate information into the prediction. Moreover, the Phillips curve model does not nest the UCSV model, and the univariate features of inflation are captured in a different way under the UCSV model. Since these two models are non-nested, we can use the bootstrap procedures above to compare their forecasts reliably.

We compare the two forecasting schemes using mean-square-prediction-error losses of annualized inflation as measured by CPI and PCE price-level indices. To construct forecasts, we use a rolling forecasting scheme for both models, with $m = 130$ as the look-back horizon[11]. To estimate the lag-order of the Phillips curve forecasts, we use the AIC as in Stock and Watson (2006), choosing lags for unemployment (r) and change in inflation (p) separately with lags between 1 and 12 months. To obtain stationary test functions for use in the tests, we first transform the monthly data in Stock and Watson (2012) according to the transformations the authors lay out. In particular, we use log differences for real series, log double-differences for price-levels, and no transformation for series that are expressed as annualized rates of change. Then we lag these series by one month to obtain the total set of 105 test functions. Instruments include real variables like Industrial Production, inflation-change variables like PPI, and financial variables like the S&P500 Index return. As in Stock and Watson (2006), we examine the entire out-of-sample window from 1970 to 2011 as one, as well as the windows 1970-1990 and 1990-2008 separately, so that we account for possible breaks in the 1980's and late 2000's. The original series and forecasts are plotted in Appendix 3.C.

The relative predictive ability of the two models is compared in Tables 3.1 and 3.2. Table 3.1 shows the Diebold and Mariano (1995) test statistics for the hypothesis that the expected loss under the Phillips curve model is less than the expected loss under the UCSV model over the three out-of-sample periods. The results indicate that the expected loss under the UCSV model is smaller than under the Phillips curve model. This is particularly significant for PCE, which has higher test statistics in all three out-of-sample windows. The results are echoed in Table 3.2, which shows the p-values of the Giacomini and White (2006) test (GW), the bootstrap test (BT), and the studentized bootstrap test (BTS). These tests all reject the null hypothesis that the forecast-error-losses form a martingale-difference-sequence at the 99% confidence level. The unanimous rejection of the null across the board suggests the UCSV is much more accurate at forecasting inflation at the one month horizon, and the rejection of the bootstrap tests suggests that type I error is highly unlikely.

These results confirm the findings of Stock and Watson (2006) and Atkeson and Ohanian (2001) that univariate inflation forecasts tend to outperform Phillips curve forecasts. This casts doubt on Phillips curve models for inflation, and suggests that accounting for the economic environment in a univariate framework (with stochastic volatility, for

[11] We use this look-back so that our first forecast is for 1970, as in Stock and Watson (2006).

	CPI	PCE
Full Sample	4.85**	6.05**
1970-1990	3.61**	4.80**
1990-2008	3.42**	4.62**

Table 3.1: Diebold-Mariano test statistics for tests of equal predictive ability of (MSE) loss differences between the UCSV and the Phillips curve for CPI and PCE inflation forecasts. Positive test statistics indicate that UCSV model has lower mean loss than the Phillips curve model.** indicates significance at the 99% confidence level.

	GW	BT	BTS
Full Sample	0.000**	0.000 **	0.002**
1970-1990	0.000**	0.004 **	0.007 **
1990-2011	0.000**	0.000 **	0.022 **

(a) CPI

	GW	BT	BTS
Full Sample	0.000**	0.000 **	0.001**
1970-1990	0.000**	0.000 **	0.001 **
1990-2011	0.000**	0.000 **	0.003 **

(b) PCE

Table 3.2: P-values for GW, bootstrap (BT), and studentized boostrap (BTS) tests of equal conditional predictive ability for CPI and PCE inflation forecasts under the UCSV and Phillips curve models. ** indicate significance at the 99% confidence level.

example) may be a better description of inflation. Though in this paper we consider unemployment as the real economic indicator, other papers like Stock and Watson (2010) have confirmed similar findings with gap measures[12] and other real indicators for one year forecasts. Because of the strong rejections of the null over all windows and the comparable performance of Phillips curves to one another in Stock and Watson (2006), we conjecture that the results here extend to these alternative Phillips curve specifications and other forecast horizons. The results suggest that the Phillips curve models largely arose as over-fitted models to the inflation environment pre-1970, and that an accurate univariate model that captures changes in the economy is an improved representation of inflation dynamics. We leave the task of a full comparison across specifications and horizons as an area for further research.

3.5 Extensions

The bootstrap procedures in the algorithms above are generally applicable in other settings where Wald tests are used and the number of restrictions on the parameters is large. As long as the underlying data satisfy the assumptions above, the bootstrap procedures will have asymptotically correct coverage. This allows us to test general, two-sided simple nulls.

A common example where this applies is in testing restrictions on a linear regression

[12]That is, instead of using the raw unemployment rate, the authors also consider deviations from that rate and a trend, such as an estimate of the NAIRU (non-accelerating inflation rate of unemployment).

of a dependent variable on a (possibly large) set of independent variables[13]. The population coefficients on the independent variables, β, can be estimated by OLS, giving coefficients $\hat{\beta}$. Researchers are frequently interested in knowing whether the parameters β are significantly different from β_0, some benchmark values. To test this restriction, $H_0 : \beta = \beta_0$, a Wald test is most commonly used, and a covariance matrix estimate between all of the estimated regression coefficients is required. Just as in the Giacomini and White (2006) test, if the number of regressors is moderately large, this covariance matrix can be unstable and the test may be incorrectly rejected. To remedy this problem we can use the bootstrap approach in a similar spirit to our previous methodology.

Specifically, we let Z_t denote the matrix of all data, where the first column is the dependent variable and subsequent columns are independent variables. After estimating the parameter vector $\hat{\beta}$ by regressing the first column on all the rest, we seek to estimate the sampling distribution of $\sqrt{T}(\hat{\beta} - \beta_0)$. As before, using an appropriate bootstrap method we resample new matrices Z_t^*, obtain new parameter estimates $\hat{\beta}^*$, and center them by $\hat{\beta}$. We then obtain a joint confidence region for the parameters, and reject the null appropriately using this confidence region. That is, if β_0 is outside the joint confidence region, we reject the null hypothesis.

This application to the multiple regression framework is particularly useful in the context of inference with many weak instruments in simultaneous equations models. Chernozhukov and Hansen (2008) introduce a Wald test for testing the significance of parameters corresponding to endogenous variables that are (weakly) identified by a set of instruments. In particular, they consider testing restrictions of the form $H_0 : \beta \neq \beta_0$ in the simultaneous equations model (given in structural form):

$$Y = X\beta + \varepsilon \tag{3.14}$$

$$X = Z\Pi + V \tag{3.15}$$

When the instruments Z are highly correlated with the dependent variables X, this model can be reliably estimated by 2SLS. In that case, testing the null $H_0 : \beta = \beta_0$ can by done using the 2SLS estimator and its asymptotic distribution. When the instruments are weakly correlated with X, however, the 2SLS asymptotic distribution is a poor approximation to the desired sampling distribution of the estimator. In that event, Chernozhukov and Hansen (2008) suggest an alternative test which takes advantage of the reduced form equation under the null. In particular, they note that the equation (3.15) can be written as:

$$Y - X\beta_0 = Z\alpha + \varepsilon \tag{3.16}$$

[13]This example is briefly described in Romano and Wolf (2005) as a possible extension of their test.

Under the null $H_0 : \beta = \beta_0$, the validity of the instruments guarantees that $\alpha = 0$. Hence, an alternative formulation of the null $H_0 : \beta = \beta_0$ is the null $H_0' : \alpha = 0$ in the reduced form model (3.16). This new null hypothesis relies on a Wald test statistic, which, as seen before, can be plagued by type I error if the dimension of the instruments is large. In the common case of many weak instruments, this test will frequently be rejected under the null.

As in the simple multiple regression case above, we can remedy the size of the test when the number of instruments is large using the bootstrap procedures explained above. Moreover, the stepwise bootstrap procedure can be used to identify which instruments represent particularly strong departures from the null. Armed with that knowledge, applied researchers can decide whether the instruments are invalid or whether $\beta = \beta_0$. If few restrictions are rejected, it may be that those instruments are invalid and the null is correct. Alternatively, if many restrictions are rejected, this may be a sign that the overall restriction $\beta = \beta_0$ is improbable.

3.6 Conclusion

In this paper we investigated the resilience of Wald-type test statistics, and the Giacomini and White (2006) statistic in particular, to the inclusion of many restrictions. We explained intuitively that these test statistics can be misleading due to the dimensionality of the restrictions, and that they are prone to type I error under the null. Our numerical study corroborated this explanation, and showed that the problem is severe even when the number of restrictions (or test functions) is moderately large.

As an alternative, we suggested bootstrap approximations to the distributions of these statistics, which were inspired by the Bootstrap Reality Check of White (2000) and the stepwise tests of Romano and Wolf (2005). We motivated the rationale for the robustness of bootstrap procedures to the dimensionality of the restrictions under the null. Furthermore, we showed in the numerical study that the bootstrap tests have strong size and power properties irrespective of the number of restrictions. This robustness was evident even under heterogeneity in the data source and different parameter choices for the DGP. Lastly, we demonstrated that studentization costs little in performance under homoskedastic DGPs, but considerably improves the bootstrap tests under heteroskedastic DGPs.

To complement the theoretical investigation of these tests, we also investigated the practical use of the bootstrap tests in an empirical example and suggested useful extensions to other empirical work. We showed that univariate inflation models which allow for changes in the economic environment and the UCSV model, in particular, outperform Phillips curve models in inflation forecasting. This finding was robust to the inflation

series used and (using the bootstrap tests) to inclusion of many test functions, suggesting that the UCSV model is a significant improvement to the Phillips curve. The use of the bootstrap procedure can hence robustly confirm previous results such as this important macroeconomic finding, and can also be used in alternative setting such as tests in the presence of many weak instruments in microeconometrics. In sum, the presented bootstrap procedures form an important new tool for applied economic research.

Appendix 3.A: Sensitivity to Parameters

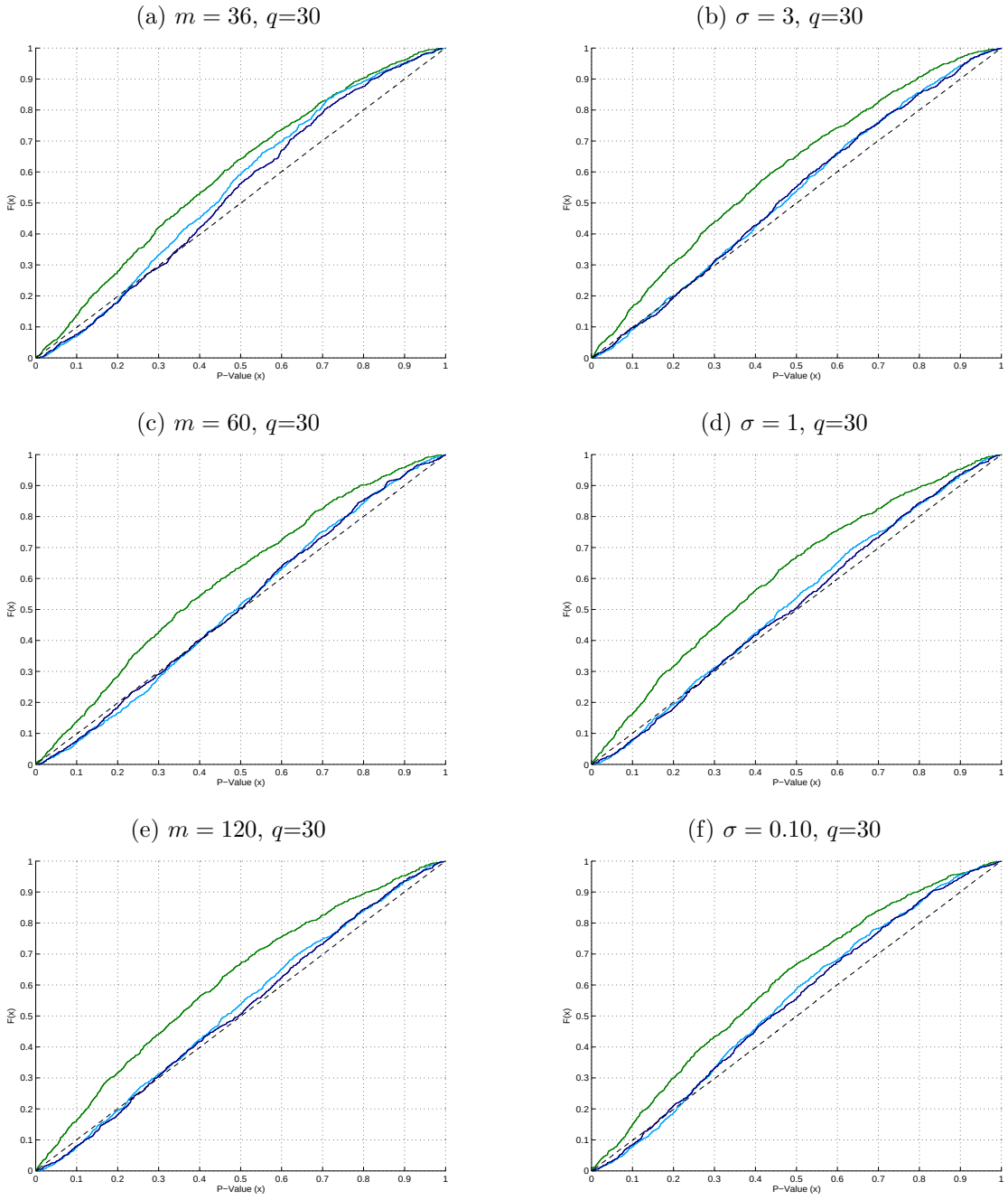


Figure 3.7: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the variance of the innovation σ and the forecast estimation window m are varied under the null. The test functions are homoskedastic and the sample size is 1000.

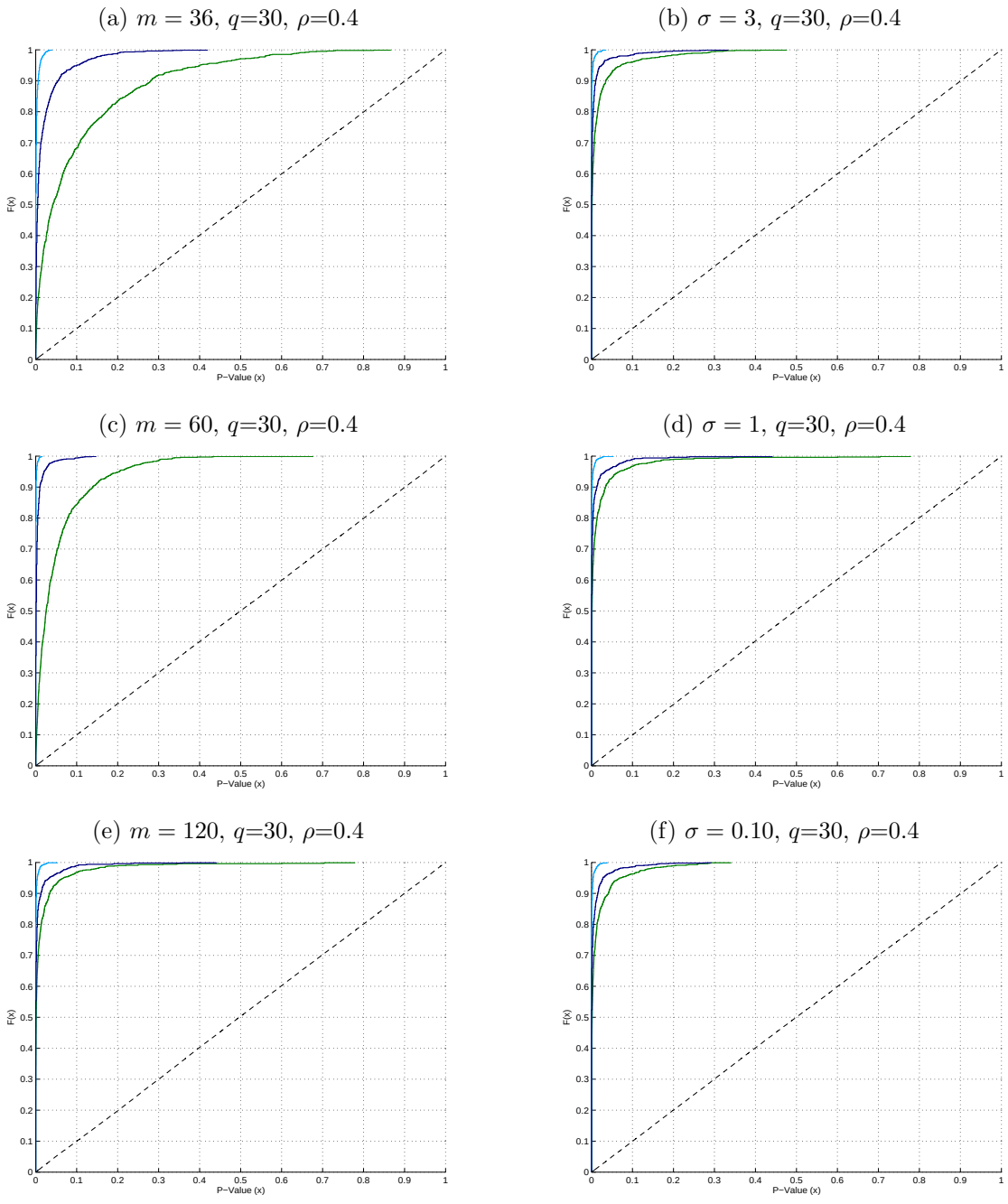


Figure 3.8: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the variance of the innovation σ and the forecast estimation window m are varied under the alternative $\rho = 0.4$. The test functions are homoskedastic and the sample size is 1000.

Appendix 3.B: Sensitivity to Data Heterogeneity

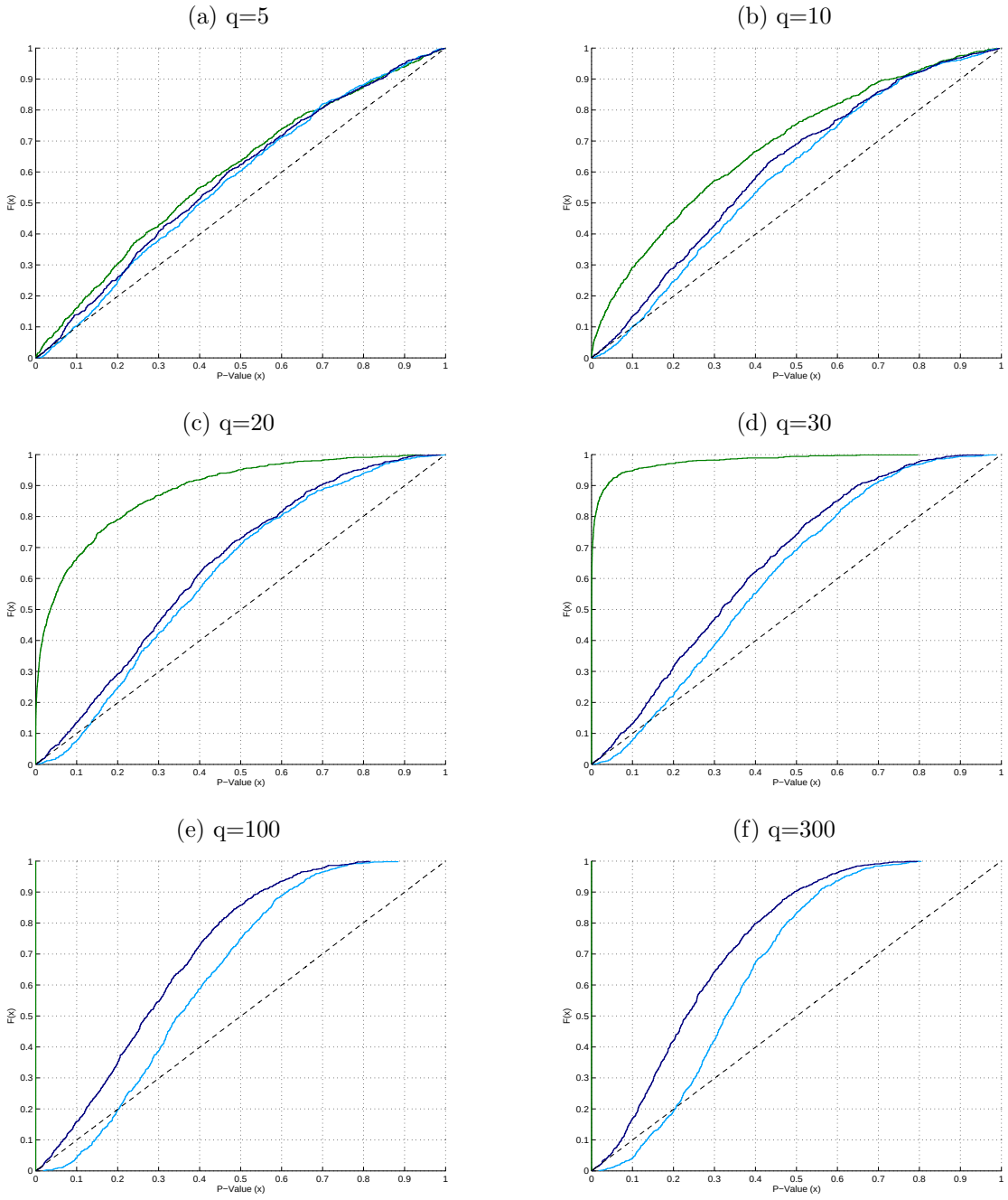


Figure 3.9: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the dimension of the test function q is varied under the null. The test functions are homoskedastic, the sample size is 511, and the underlying data are CPI inflation growth rates.

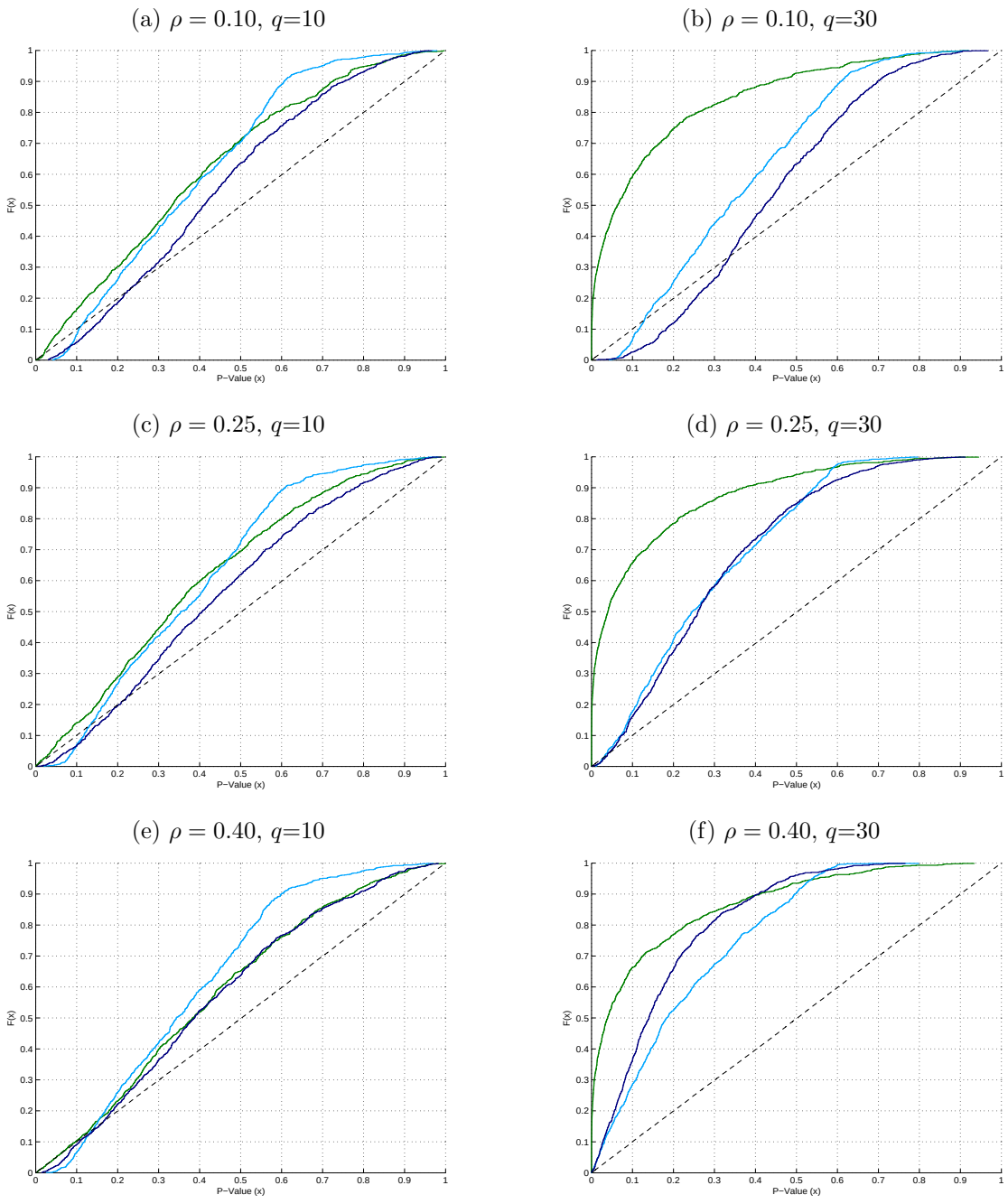
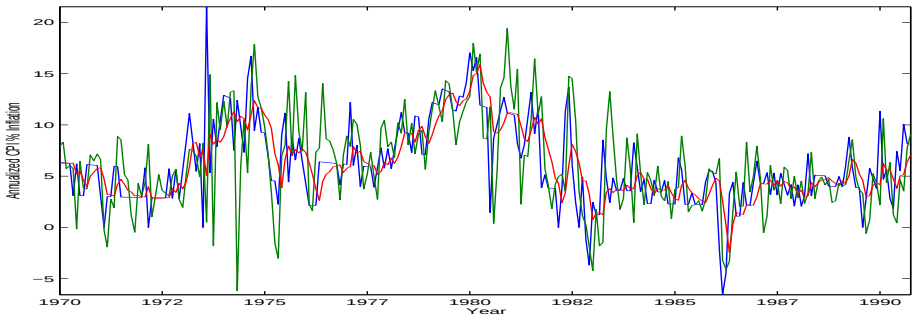


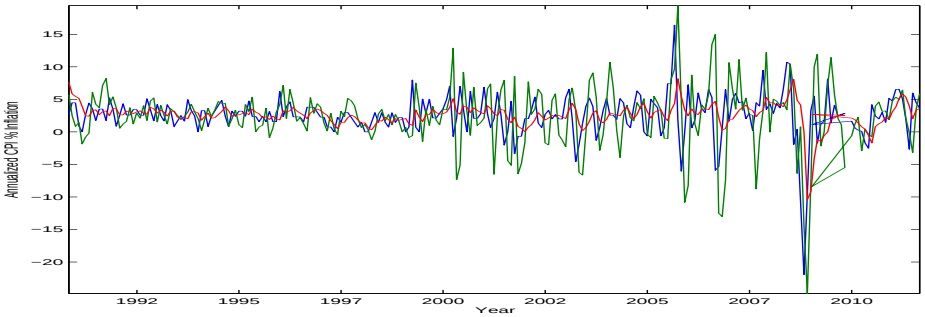
Figure 3.10: Empirical CDFs for the 1000 p-values across the Monte Carlo replications of the Giacomini and White (2006) test (green), non-studentized bootstrap test (light blue), and the studentized bootstrap test (dark blue) as the dimension of the test function q is varied under three alternatives: $\rho = \{0.10, 0.25, 0.40\}$. The test functions are homoskedastic, the sample size is 511, and the underlying data are CPI inflation growth rates.

Appendix 3.C: Inflation Series and Forecasts

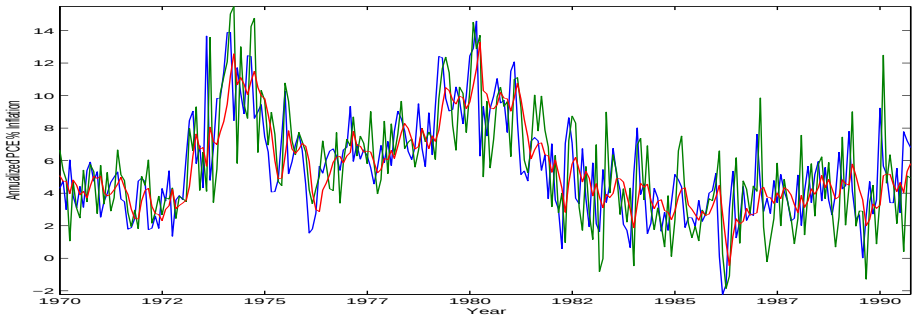
(a) CPI Inflation and Forecasts: 1970-1990



(b) CPI Inflation and Forecasts: 1990-2011



(c) PCE Inflation and Forecasts: 1970-1990



(d) PCE Inflation and Forecasts: 1990-2011

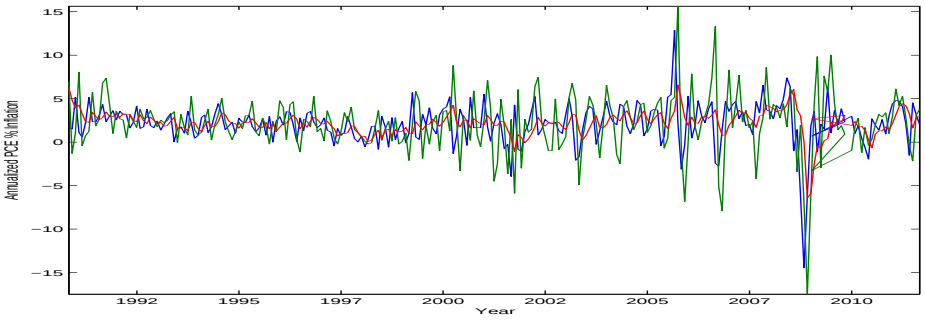


Figure 3.11: Inflation series and forecasts for CPI and PCE annualized % inflation. Original series are in blue, Phillips curve forecasts are in red, and UCSV forecasts are in red.

References

Acharya, V., L. Pederson, T. Phillipon, and M. Richardson (2010). Measuring systemic risk. NYU Working Paper.

Adrian, T. and M. K. Brunnermeier (2011, September). CoVaR. Technical Report, FRB of New York: Staff report No. 348.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*, 529–626.

Atkeson, A. and L. E. Ohanian (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review 25*, 2–11.

Banbura, M., D. Giannone, and L. Reichlin (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics 25*, 71–92.

Barigozzi, M. and C. Brownlees (2014). NETS: Network Estimation for Time Series. Technical report, Barcelona GSE.

Barigozzi, M., A. M. Conti, and M. Luciani (2014). Do euro area countries respond asymmetrically to the common monetary policy? *Oxford Bulletin of Economics and Statistics 76*, 693–714.

Barndorff-Nielsen, O., P. R. Hansen, A. Lunde, and N. Shephard (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*.

Barndorff-Nielsen, O., P. R. Hansen, A. Lunde, and N. Shephard (2009). Realised kernels in practice: Trades and quotes. *Econometrics Journal 04*, 1–32.

Bauwens, L., S. Laurent, and J. V. K. Rombouts (2006, February). Multivariate GARCH models: a survey. *Journal of Applied Econometrics 21*, 79–109.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*, 521–547.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: A factor augmented vector autoregressive (favar) approach. *Quarterly Journal of Economics 1*, 387–422.

Bickel, P. J. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics 9*(6).

Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics 104*, 535–559.

Bollen, N. P., S. F. Grau, and R. E. Whaley (2000). Regime switching in foreign exchange rates: Evidence from currency option prices. *Journal of Econometrics 94*, 239–276.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

Bondell, H. D., B. J. Reich, and H. Wang (2010). Non-crossing quantile regression curve estimation. *Biometrika*, 1–13.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *The Annals of Statistics 16*(4).

Brockwell, P. J. and R. A. Davis (2009). *Time Series: Theory and Methods*. Springer Series in Statistics.

Brownlees, C. T. and R. F. Engle (2011, June). Volatility, correlation, and tails for systemic risk measurement. Working Paper.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6).

Chatterjee, A. and S. Lahiri (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics 41*, 1232–1259.

Chatterjee, A. and S. N. Lahiri (2011). Bootstraping lasso estimators. *Journal of the American Statistical Association 106*(494), 608–625.

Chernozhukov, V. and C. Hansen (2008). The reduced form: A simple approach to inference with weak instruments. *Economics Letters 100*, 68–71.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1998, August). Monetary policy shocks: What have we learned and to what end?

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7*, 174–196.

Cox, D. R. (1961). Tests of seperate families of hypothesis. In *Proceedings of the Berkeley Symposium*, Volume 4, pp. 105–123.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*, 253–265.

Diebold, F. X. and K. Yilmaz (2008). Measuring financial asset return and volatility spillovers, with application to global equity markets. *NBER Working Paper No. 13811*.

Diebold, F. X. and K. Yilmaz (2011, October). On the network topology of variance decompositions: Measuring the connectedness of financial firms. NBER Working Paper No. 17490.

Doan, T., R. B. Litterman, and C. A. Sims (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews 3*, 1–100.

Doornik, J. (2009). *Autometrics in The Methodology and Practice of Econometrics*, pp. 88–121. Oxford University Press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(3), 407–499.

Epprecht, C., D. Guegan, and A. Veiga (2013). Comparing variable selection techniques for linear regression: Lasso and autometrics.

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.

Fischer, S. (1977). Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy 85*, 191–205.

Forni, M. and L. Gambetti (2010, March). The dynamic effects of monetary policy: A structural factor model approach. *Journal of Monetary Economics 57*, 203–216.

Forni, M., D. Giannone, M. Lippi, and L. Reichlin (2009, October). Opening the black box: Structural factor models with large cross sections. *Econometric Theory 25*, 1319–1347.

Freedman, D. (1981). Bootstrapping regression models. *The Annals of Statistics 9*, 1218–1228.

Friedman, J., T. Hastie, and R. Tibshirani (2009, April). Regularization paths for generalized linear models via coordinate descent.

Friedman, M. (1968). The role of monetary policy. *American Economic Review 58*, 1–17.

Giacomini, R. and H. White (2003, April). Tests of conditional predictive ability. Working Paper.

Giacomini, R. and H. White (2006, November). Tests of conditional predictive ability. *Econometrica 74*, 1545–1578.

Giannone, D., M. Lenza, and G. E. Primiceri (2012, November). Prior selection for vector autoregressions. *ECB Working Paper Series* (1494).

Gordon, R. J. (1997). The time-varying NAIRU and its implications for economic policy. *Journal of Economic Perspectives 11*, 11–32.

Gotze, F. and H. R. Kunsch (1996). Second order correctness of the blockwise bootstrap for stationary observations. *The Annals of Statistics 24*(5), 1914–1933.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hall, P. (1998). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics 16*, 927–953.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics 23*(4), 365–380.

Hansen, P. R. and A. Lunde (2006). Realised variance and market microstructure noise. *Journal of Business and Economic Statistics*.

Harvey, C. R. and Y. Liu (2014). Backtesting. *Working Paper*.

Hasbrouck, J. (2013). High frequency quoting: Short-term volatility in bids and offers. *Working Paper*.

Hastie, T., R. Tibshirani, and J. Friedman (2009, February). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

Hautsch, N., J. Schaumburg, and M. Schienle (2014). Forecasting systemic impact in financial networks. *International Journal of Forecasting 30*, 781–794.

Hendry, D. F., S. Johansen, and C. Santos (2006). Selecting a regression saturated by indicators. *Working Paper*.

Killian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*.

Knight, K. and W. Fu (2000, October). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356–1378.

Kock, A. B. and L. A. Callot (2012). Oracle efficient estimation and forecasting with the adpative lasso and the adaptive group lasso in vector autoregressions. *Creates Research Paper 38*.

Kogan, L. and M. Tian (2012). Firm characteristics and empirical factor models: A data-mining experiment. Technical report, Board of Governors of the Federal Reserve System.

Lee, T.-H. (2007, March). Loss functions in time series forecasting.

Leeb, H. and B. M. Potscher (2008). Can one estimate the unconditional distribution of post-model selection estimators? *Econometric Theory 24*, 338–376.

Litterman, R. B. (1980). A Bayesian procedure for forecasting with vector autoregression. *Working Paper, Massachusetts Institute of Technology, Department of Economics*.

Lucas, R. E. J. (1972). Expectations and the neutrality of money. *Journal of Economic Theory 4*, 103–24.

Mincer, J. and V. Zarnowitz (1969). *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, Chapter "The Evaluation of Economic Forecasts", pp. 1–46. National Bureau of Economic Research.

Negro, M. D. and C. Otrok (2008). Dynamic factor models with time-varying parameters: Measuring changes in international business cycles. Technical report, Federal Reserve Bank of New York.

Newey, W. K. and K. D. West (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–708.

Noureldin, D., N. Shephard, and K. Sheppard (2012, October). Multivariate high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics 27*.

Patton, A. J. (2011, January). Volatility forecast evaluation and comparison using imperfect volatility proxies. *Journal of Econometrics 160*, 246–256.

Patton, A. J. and K. K. Sheppard (2009). *Evaluating volatility forecasts in Handbook of Financial Time Series*. Springer-Verlag.

Phelps, E. S. (1969). The new microeconommics in inflation and employment theory. *American Economic Review 59*, 147–60.

Phillips, A. W. (1958, November). The relation between unemployment and the rate of change of money wage rates in the united kingdom. *Economica 25*, 283–299.

Phillips, P. C. (1990). To criticize the critics: An objective Bayesian analysis of stochastic trends. *Cowles Foundation Discussion Paper No. 950*.

Politis, D. N. and J. P. Romano (1994, December). The stationary bootstrap. *Journal of the American Statistical Association 89*(428).

Rivers, D. and Q. H. Vuong (2002). Model selection for nonlinear dynamic models. *The Econometrics Journal 5*, 1–39.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Samuelson, P. A. and R. M. Solow (1960, May). Analytical aspects of anti-inflation policy. *American Economic Review 50*, 177–194.

Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance 19*(3), 425–442.

Sheppard, K. K. and N. Shephard (2010). Realising the future: forecasting with high frequency based volatility (HEAVY) models. *Journal of Applied Econometrics 25*, 197–231.

Sims, C. A. (1980, January). Macroeconomics and reality. *Econometrica 48*(1), 1–48.

Stinchcombe, M. and H. White (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory 14*, 295–325.

Stock, J. H. and M. W. Watson (2002, April). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147–162.

Stock, J. H. and M. W. Watson (2006). Why has U.S. inflation become harder to forecast? *NBER Working Paper No. 12324*.

Stock, J. H. and M. W. Watson (2010). Modeling inflation after the crisis. *NBER Working Paper No. 16488*.

Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics 30*(4), 481–493.

Taylor, J. B. (1980). Aggregate dynamics and staggered contracts. *Journal of Political Economy 88*, 1–23.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Royal Statistics Soc. B. 58*(1), 267–288.

Veredas, D. and M. Luciani (2012). Estimating and forecasting large panels of volatilities with approximate dynamic factor models. Technical report, ECARES.

Vuong, Q. H. (1989). Likelihood ratio test for model selection and non-nested hypotheses. *Econometrica 57*, 307–333.

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

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association 57*, 348–368.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society 67*, 301–320.

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics 37*(4), 1733–1751.