ORACLE EFFICIENT ESTIMATION AND FORECASTING WITH THE ADAPTIVE LASSO AND THE ADAPTIVE GROUP LASSO IN VECTOR AUTOREGRESSIONS

LAURENT A.F. CALLOT AND ANDERS BREDAHL KOCK

VU AMSTERDAM, AARHUS UNIVERSITY AND CREATES

Abstract. We show that the adaptive Lasso (aLasso) and the adaptive group Lasso (agLasso) are oracle efficient in stationary vector autoregressions where the number of parameters per equation is smaller than the number of observations. In particular, this means that the parameters are estimated consistently at a $\sqrt{T}$-rate, that the truly zero parameters are classified as such asymptotically and that the non-zero parameters are estimated as efficiently as if only the relevant variables had been included in the model from the outset. The group adaptive Lasso differs from the adaptive Lasso by dividing the covariates into groups whose members are all relevant or all irrelevant. Both estimators have the property that they perform variable selection and estimation in one step.

We evaluate the forecasting accuracy of these estimators for a large set of macroeconomic variables. The plain Lasso is found to be the most precise procedure overall. The adaptive and the adaptive group Lasso are less stable but mostly perform at par with common factor models.

Key words: Vector autoregression, VAR, adaptive Lasso, Group Lasso, Forecasting, Factor models, LSTAR.

JEL classifications: C32, C53, E17.

1. Introduction

In recent years large data sets have become increasingly available and as a result techniques to handle these have been the object of considerable research. When building a

Date: February 20, 2013.

The second author would like to thank Timo Teräsvirta for guiding him through his Ph.D and for showing him the fascinating world of forecasting. However, the shortcomings of this paper are not Timo's responsibility. The authors also wish to thank the Centre for Research in the Econometric Analysis of Time Series (CREATES), funded by the Danish National Research Foundation, for financial support.

model to explain the behavior of a variable it is not uncommon that the set of potential explanatory variables can be very large. Traditional techniques for model selection rely on a sequence of tests or the application of information criteria. However, neither of these is very useful when the number of potential explanatory variables is large since the number of tests or information criteria to be calculated increases exponentially in the cardinality of the set of covariates. Hence, alternative routes have been explored and in particular regularized estimators have received a lot of attention in the statistics literature. The most prominent member of this class is the least absolute shrinkage and selection operator (Lasso) of Tibshirani (1996). Since its inception, the statistical properties of Lasso-type estimators have been studied intensively with particular focus on the *oracle property.* An estimator is said to possess the oracle property if i) it selects the correct sparsity pattern with probability tending to one (i.e leaves out all irrelevant variables and retains all relevant variables) and ii) estimates the non-zero coefficients with the same rate and asymptotic distribution as if only the relevant variables had been included in the model from the outset. Put differently, the oracle property guarantees that the estimator performs as well as if the true model had been revealed to the researcher in advance by an oracle.

A lot of research has been carried out investigating the oracle property of various shrinkage type estimators: bridge-type estimators were investigated by Knight and Fu (2000), the SCAD by Fan and Li (2001), the adaptive Lasso by Zou (2006), the Bridge and Marginal Bridge by Huang et al. (2008) and Sure independence screening by Fan and Lv (2008). The working assumption in the literature is that even though the set of potential explanatory variables may be large (sometimes even considerably larger than the sample size) only a small subset of these variables are relevant for the task of explaining the left hand side variable, i.e. the model is sparse. Most focus has been on the cross sectional setting with either fixed or independently identically distributed covariates while much less attention has been paid to the case of dependent data. Some exceptions are Wang et al. (2007), Caner and Knight (2011), Kock (2012), Medeiros and Mendes (2012), Song and Bickel (2011), Kock and Callot (2012) and Liao and Phillips (2012). Specifically, the first three references consider univariate models (Caner and Knight (2011) and Kock (2012) consider stationary as well as non-stationary models). The three latter references deal with multivariate models. More precisely Song and Bickel (2011), Kock and Callot (2012) deal with high-dimensional models while Liao and Phillips (2012) deal with the low-dimensional but non-stationary case. Finally, Caner (2011) uses bridge estimators to select factors in approximate factor models. In this paper we further fill this gap by considering stationary vector autoregressive models of the type

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

where y_t is $N \times 1$ and e_t is i.i.d. with mean 0 and covariance matrix Σ. $B_i,\ 1 \leq i \leq p$ are the $N \times N$ parameter matrices. The properties of the model will be made precise in the next section.

It is likely that many entries in the B_i matrices are equal to zero, i.e. the B_i are sparse. This could be because of p being larger than the true number of lags or because there are gaps in the lag structure (e.g. $B_1 \neq 0,\ B_2 = B_3 = 0$ and $B_4 \neq 0$ for quarterly data). Another reason could be that lags of a subset of the variables are irrelevant for the task of explaining another subset of variables which manifests itself by zero restrictions on certain entries of the B_i. Granger non-causality is an extreme case of this latter example. In the first part of this paper we show that the adaptive Lasso of Zou (2006) possesses the oracle property when applied to stationary vector autoregressions. Hence, it selects the correct sparsity pattern asymptotically and the non-zero parameters are estimated as precisely as if the true model had been known in advance and only the relevant variables had been included and estimated by least squares.

In equation (1) it is likely that zero parameters occur in groups. For example all lags of a specific length may be irrelevant resulting in $B_i = 0$ for some $1 \leq i \leq N$. Alternatively, all lags of a certain variable may be irrelevant in explaining another variable. Utilizing this group structure may lead to improved (finite sample) performance of the (adaptive) Lasso. Hence, inspired by Wang and Leng (2008) we combine the group Lasso of Yuan and Lin (2006) with the adaptive Lasso to make use of this grouping structure. We show that the adaptive group Lasso possesses a variant of the oracle property if one correctly groups (a subset) of the potential explanatory variables.

The above results motivate using the adaptive (group) Lasso in impulse response analysis of e.g. a macroeconomic shock. To do so one simply uses the estimated model and carries out the impulse response analysis as usual in a VAR. In this sense, the adaptive (group) Lasso constitutes an alternative way of handling high-dimensionality in impulse response analysis than the factor augmented VAR (FAVAR) of Bernanke et al. (2005)[1].

Since vector autoregressions have been used extensively for forecasting, an obvious question is how well the VAR performs in this respect when estimated by the Lasso, the adaptive Lasso or the adaptive group Lasso. In particular, we investigate the performances of these estimators for forecasting in large macroeconomic datasets. The benchmark models for this type of forecasting exercise are common factor models. The common factor approach is supported by a long tradition in macroeconomic theory of assuming that a small set of underlying variables drives the business cycle and are responsible for the bulk of the variation of macroeconomic time series. Stock and Watson (2002); Ludvigson and Ng (2009) *inter alia* document the strong forecasting power of these types of models for large US macroeconomic datasets. Motivated by this we shall compare the forecast accuracy of the Lasso type estimators to the one of factor models. A comparison to a simple linear autoregression of order one and the Bayesian VAR is also made. The potential

[1] We would like to thank an anonymous referee for pointing us towards this motivation.

gains in forecast accuracy from exploiting non-linearities in the data are investigated by also including the logistic smooth transition autoregression (LSTAR) of Teräsvirta (1994) into the comparison. Interestingly, it is found that the Lasso on average forecasts most precisely. The factor models show a very stable performance, while the forecast errors from the adaptive Lasso and the adaptive group Lasso are much more erratic.

In the next section we introduce the VAR model and some notation. Section 3 introduces the adaptive lasso and section 4 the adaptive group Lasso. Section 5 discusses the forecasting experiment and presents the results. All proofs are relegated to the appendix.

2. Model and notation

As mentioned in the introduction we are concerned with stationary VARs, meaning that all roots of $|I_N - \sum_{j=1}^{p} B_j z^j|$ lie outside the unit circle[2].

It is convenient to write the model in (1) as a standard regression model. To do so let $Z_t = (y_{t-1}', ..., y_{t-p}')'$ be the $Np \times 1$ vector of explanatory variables at time t in each equation $i = 1, ..., N$ and $Z = (Z_T, ..., Z_1)'$ the $T \times Np$ matrix of covariates. Set $X = I_N \otimes Z$ where $\otimes$ denotes the Kronecker product. Let $y_i = (y_{T,i}, ..., y_{1,i})'$ be the $T \times 1$ vector of observations on the ith variable ($i = 1, ..., N$) and ϵ_i the corresponding vector of error terms for variable i. Defining $y = (y_1', ..., y_N')'$ and $\epsilon = (\epsilon_1', ..., \epsilon_N')'$, we may write (1) as

$$y = X\beta^* + \epsilon \tag{2}$$

where β^* contains N^2p parameters. It is this model we will estimate by the adaptive and the adaptive group Lasso. We assume that N and p are fixed and independent of the sample size. In particular, we assume that the number of parameters per equation, Np, is less than the sample size T. For the setting where these quantities are allowed to diverge

[2] Here z is complex and for any square matrix A, $|A|$ denotes its determinant.

with the sample size we refer to Kock and Callot (2012) who however don't consider the adaptive group Lasso.

While β^* contains N^2p parameters, only a subset of those might be relevant to model the dynamics of the vector y. The adaptive Lasso discussed in section 3 is able to discard the zero parameters and estimate the non-zero ones with an oracle efficient asymptotic distribution.

2.1. **Further notation.** Let $\mathcal{A} = \{i : \beta_i^* \neq 0\}$ index the set of nonzero β_i^*s (the active set) and let $|\mathcal{A}|$ be its cardinality. For any vector $x \in \mathbb{R}^n$ $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$ denotes its euclidean norm. Furthermore, for any $A \subseteq \{1, ..., n\}$, x_A denotes the vector consisting only of the elements indexed by A. Most often $n = N^2p$ in this paper. If M is a square matrix, M_A denotes the submatrix of M consisting of the rows and columns indexed by A. We let $\to_d$ and $\to_p$ denote convergence in distribution and probability, respectively. For any set $A \subseteq \{1, ..., n\}$, A^c denotes its complement in $\{1, ..., n\}$.

Finally, let $C = E(\frac{1}{T}Z'Z)$ which is time independent by the stationarity assumption.

3. The adaptive Lasso

As noted by Zhao and Yu (2007) the Lasso is only model selection consistent under rather restrictive assumptions which rule out highly dependent covariates as may be encountered in VAR models. Hence, we shall apply the adaptive Lasso, which was proposed by Zou (2006) as a solution to the lack of model selection consistency of the Lasso, to estimate the parameters in (2). The adaptive Lasso estimates β^* by minimizing the following objective function.

$$L_T(\beta) = \|y - X\beta\|^2 + \lambda_T \sum_{i=1}^{N^2p} \hat{w}_i |\beta_i| \tag{3}$$

where $\hat{w}_i$ is a set of weights such that $\hat{w}_i = |\hat{\beta}_{I,i}|^{-\gamma}$, $\gamma > 0$ with $\hat{\beta}_I$ a $\sqrt{T}$-consistent (initial) estimator of β^*. We shall use the least squares estimator[3]. The most common choice of γ is $\gamma = 1$. λ_T is a sequence whose properties determine the asymptotic properties of the adaptive Lasso. Note that the standard Lasso corresponds to the case of $\hat{w}_i = 1$, i.e. all parameters receive an equal penalty. In other words the difference between the Lasso and its adaptive version is that the latter chooses its penalty terms more intelligently (adaptively): If $\beta_i^* = 0$ for some $i = 1, ..., N^2p$ the initial least squares estimator is likely to be close to zero and so $\hat{w}_i$ tends to be large resulting in a large penalty of β_i. Hence, the adaptive Lasso is more likely to correctly classify β_i^* as zero. By a similar logic, the penalty on β_i is relatively small when $\beta_i^* \neq 0$. As we shall see in the theorems to follow, these more intelligent weights result in an improved asymptotic performance of the adaptive Lasso compared to the regular Lasso.

The objective function (3) reveals the computational advantage of the (adaptive) lasso compared to e.g. information criteria since (3) is a convex optimization problem for which many efficient optimization procedures exist. Information criteria generally penalize model complexity by an ℓ_0-penalty instead of the ℓ_1-penalty used by lasso type estimators. It is exactly the switch from ℓ_0 to ℓ_1-penalty which yields the computational advantage enabling us to consider high dimensional problems which would be impossible or very hard to approach by means of ℓ_0-penalization. As we will see next, the convex program (3) is not only fast to solve but its solution, the adaptive Lasso estimator, which we shall denote by $\hat{\beta}$, also possesses the oracle property.

Assumptions

[3] As already noted by Zou (2006) the initial estimator need not be $\sqrt{T}$-consistent. The assumptions made below can be altered such that theorems 1 and 2 still apply in the case where the initial estimator converges at a slower rate. However, we will not pursue this avenue any further here since we *do* have access to a $\sqrt{T}$-consistent consistent initial estimator.

1: $\epsilon_{i,1}$ has finite fourth moments for $i = 1, ..., N$. Recall as well that $e_t = (\epsilon_{1,t}, ..., \epsilon_{N,t})'$ are mean zero iid vectors with covariance matrix Σ.

2: $C = E(\frac{1}{T}Z'Z)$ is positive definite.

3: All roots of $|I_N - \sum_{j=1}^{p} B_j z^j|$ lie outside the unit circle.

4: $\hat{w}_i = |\hat{\beta}_{I,i}|^{-\gamma}$ with $\hat{\beta}_{I,j}$ being the least squares estimator of β^*_j.

Assumption 1 is relatively standard and used to ensure that $\frac{1}{\sqrt{T}}X'\epsilon$ converges in distribution to a gaussian random variable. But any assumption yielding this convergence will suffice for our purpose. Assumption 2 is reasonable since it simply rules out perfect collinearity because if C would not be positive definite there would exist a nonzero $Np \times 1$ vector v such that

$$0 = v'Cv = \frac{1}{T}E(v'Z'Zv) = \frac{1}{T}\sum_{t=1}^{T} E(v'Z_t)^2$$

implying that $v'Z_t = 0$ almost surely for $t = 1, ..., T$ and hence that the covariates are linearly dependent. No procedure can be expected to distinguish between such variables, and assumption 2 rules out this situation.

Assumption 3 is a stationarity assumption while assumption 4 repeats our choice of initial estimator to construct the weights. See footnote 3 for other choices of the initial estimator.

We are now in a position to state our first theorem.

Theorem 1. *Let assumptions 1-4 be satisfied and suppose that* $\frac{\lambda_T}{\sqrt{T}} \to 0$ *and* $\frac{\lambda_T}{T^{1/2-\gamma/2}} \to \infty$. *Then* $\hat{\beta}$ *satisfies the following:*

1. $\sqrt{T}$*-consistency:* $\left\| \sqrt{T}\left(\hat{\beta} - \beta^*\right) \right\| \in O_p(1)$
2. *Oracle (i):* $P(\hat{\beta}_{\mathcal{A}^c} = 0) \to 1$
3. *Oracle (ii):* $\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta^*_{\mathcal{A}}) \to_d N\left(0, [(I_N \otimes C)_{\mathcal{A}}]^{-1}[\Sigma \otimes C]_{\mathcal{A}}[(I_N \otimes C)_{\mathcal{A}}]^{-1}\right)$

The assumption $\frac{\lambda_T}{T^{1/2-\gamma/2}} \to \infty$ is needed for the adaptive Lasso to shrink truly zero parameters to zero. It requires the penalty sequence λ_T to increase sufficiently fast[4]. On the other hand, $\frac{\lambda_T}{\sqrt{T}} \to 0$ prevents λ_T from increasing too fast. This is needed to prevent the adaptive Lasso from classifying non-zero parameters as zero.

Part 1 of Theorem 1 states that the adaptive Lasso converges at the usual $\sqrt{T}$-rate. This means that no $\hat{\beta}_j,\ j \in \mathcal{A}$ will be set equal to 0 since for all $j \in \mathcal{A},\ \hat{\beta}_j$ converges in probability to $\beta_j^* \neq 0$. Part 2 is the first part of the oracle property: all truly zero parameters are set exactly equal to zero asymptotically. This is a strengthening of the consistency result in part 1 since this only ensures convergence in probability to 0 of $\hat{\beta}_{\mathcal{A}^c}$. Part 1 and 2 together imply that $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ where $\hat{\mathcal{A}}$ denotes the estimated active set. Part 3 states that the non-zero coefficients have the same asymptotic distribution as if the system in (2) had been estimated by least squares *only including the relevant variables* – i.e. only including the variables in the active set $\mathcal{A}$. In conclusion, the adaptive Lasso performs variable selection and estimation simultaneously and possesses the oracle property in the sense that it performs as well as if an oracle had revealed the true model prior to estimation.

Compared to Kock and Callot (2012) Theorem 1 considers the low-dimensional setting where the number of covariates is fixed. The gain from this additional assumption is a faster rate of convergence ($\sqrt{T}$ as opposed to almost $\sqrt{T}$ [5]). Due to space limitations we refer the interested reader to Kock and Callot (2012) for more details as well as simulation results illustrating Theorem 1.

4. ADAPTIVE GROUP LASSO

If certain groups of variables are either jointly zero or non-zero it may be useful to utilize this information to get more efficient (finite sample) estimates. For this reason Yuan

[4] Strictly speaking λ_T is only required to be increasing if $0 < \gamma \leq 1$ but since $\gamma = 1$ is the most common choice we shall use the word increasing without risk of confusion.

[5] By almost $\sqrt{T}$ we mean that the rate of convergence only differs from $\sqrt{T}$ by a logarithmic factor.

and Lin (2006) introduced the group Lasso which penalizes different groups of variables differently. Later, Wang and Leng (2008) combined the ideas of the group Lasso and the adaptive Lasso into the adaptive group Lasso. We shall now show that the latter possesses a variant of the oracle property when used to estimate vector autoregressive models. Assume that the $N^2p \times 1$ parameter vector has been partitioned into M disjoint groups, i.e. $\cup_{i=1}^{M} G_i = \{1, ..., N^2p\}$ and $G_i \cap G_j = \emptyset$ for $i \neq j$. A group G_i is said to be active if at least one of the entries of $\beta^*_{G_i}$ is non-zero. Without any confusion with the previously introduced notation we shall denote the set of active groups by $\mathcal{A} \subseteq \{1, ..., M\}$. $\mathcal{G} = \cup_{i \in \mathcal{A}} G_i \subseteq \{1, ..., N^2p\}$ denotes the union of the active groups.

The adaptive group Lasso estimates the parameters by minimizing the following objective function

$$\tilde{L}_T(\beta) = \|y - X\beta\|^2 + \lambda_T \sum_{j=1}^{M} \tilde{w}_j \|\beta_{G_j}\| \tag{4}$$

where $\tilde{w}_j$ is a set of weights such that $\tilde{w}_j = \left\|\hat{\beta}_{I,G_j}\right\|^{-\gamma}$, $\gamma > 0$ with $\hat{\beta}_{I,G_j}$ a $\sqrt{T}$-consistent estimator of $\beta^*_{G_j}$. As was the case the for the adaptive Lasso we will use the least squares estimator as initial estimator. Denote the adaptive group Lasso estimator by $\tilde{\beta}$. Note the difference with the objective function of the adaptive Lasso in (3): now the penalty is applied group-wise as opposed to being applied to each parameter individually. The economic motivation for this is that one might conjecture that either all variables in a specific group are relevant or none of them are. Imposing this (correct) restriction may increase efficiency. We shall investigate the empirical performance in terms of forecasting accuracy in the next section. But first we state the adaptive group Lasso equivalent of Theorem 1.

Theorem 2. *Let assumptions 1-4 be satisfied and suppose that* $\frac{\lambda_T}{\sqrt{T}} \to 0$ *and* $\frac{\lambda_T}{T^{1/2-\gamma/2}} \to \infty$. *Then* $\tilde{\beta}$ *satisfies the following:*

1. *$\sqrt{T}$-consistency:* $\left\|\sqrt{T}\left(\hat{\beta}-\beta^*\right)\right\| \in O_p(1)$
2. *Oracle (i):* $P(\hat{\beta}_{\mathcal{G}^c}=0) \to 1$
3. *Oracle (ii):* $\sqrt{T}(\hat{\beta}_{\mathcal{G}}-\beta^*_{\mathcal{G}}) \to_d N\left(0, [(I_N \otimes C)_{\mathcal{G}}]^{-1}[\Sigma \otimes C]_{\mathcal{G}}[(I_N \otimes C)_{\mathcal{G}}]^{-1}\right)$

The assumptions underlying Theorem 2 are identical to the ones made to establish Theorem 1 and the intuition on the rate of increase of λ_T is also the same: it must be large enough to shrink all inactive groups of parameters to zero while being small enough to avoid doing so for any active group of parameters.

Part 1 of Theorem 2 states the $\sqrt{T}$-consistency of the adaptive group Lasso. Hence, no relevant variables will be excluded asymptotically since $\tilde{\beta}_i \to_p \beta^*_i \neq 0$ for $i \in \mathcal{G}$. Part 2 yields that all inactive groups are also classified to be inactive asymptotically. So all groups consisting *only* of parameters whose true value is zero will also be set exactly equal to zero with probability tending to one. However, note that this claim is not made about those parameters whose true value is zero but are (mistakenly) located in an active group. Their behavior is described in part 3 of the theorem: all parameters belonging to an active group are estimated with the same asymptotic distribution as if least squares had been applied to (2) only including variables belonging to $\mathcal{G}$. On the downside this means that the adaptive group Lasso only performs better than least squares including all variables if one is able to identify a group consisting only of zeros. On the other hand, the asymptotic distribution is equivalent to the one of least squares including all variables if one fails to do so and hence there is no efficiency loss. The empirical performance of the adaptive group Lasso estimator is investigated in the forecasting section. As we shall see there, many groups are found to be inactive in practice.

4.1. **Some limitations.** As it stands, the oracle property sounds almost too good to be true – and in some sense it is. In a series of papers, Leeb and Pötscher (2005, 2008); Pötscher and Leeb (2009) shed critical light on consistent model selection procedures and shrinkage type estimators in particular. They point out that most results, including the ones in this paper, are for pointwise asymptotics (sometimes also referred to as fixed parameter asymptotics). The adaptive Lasso performs well in such a setting, but if uniform asymptotics are considered it may not be able to distinguish certain non-zero parameters from zero ones. In particular, the problematic regions are disks with radius proportional to $1/\sqrt{T}$. Furthermore, even though the asymptotic distribution of the truly non-zero parameters is the same as if least squares had been applied only including the relevant variables one may find that the finite sample distributions can be highly bimodal – with mass at zero and in an interval around the true parameter value. Finally, using the mean square estimation error as loss function, the uniform (uniform over the parameter space) loss of *any* consistent model selection technique of the standard linear regression model may be shown to be infinite while the one of the least squares estimator can be shown to be finite.

5. Forecasting

In this section we investigate the empirical performance of the Lasso, the adaptive Lasso and the adaptive group Lasso in terms of forecasting macroeconomic variables with a large number of predictors. Vector autoregressive models have been used extensively for forecasting since their inception and are still a popular tool for this purpose in macroeconometrics. Hence, it is of interest to investigate whether novel estimation methods of vector autoregressions can lead to more precise forecasts in data rich settings.

5.1. **The data.** We use the data from Ludvigson and Ng (2009), which is an updated version of the data used in Stock and Watson (2002). The data set contains 131 U.S. monthly macroeconomic indicators, from January 1964 to December 2007. Detailed description of the series as well as the transformations required to make the series I(0) can be found in appendix A of Ludvigson and Ng (2009). The series fall in 8 broad economic categories:

(1) Output and Income (17 series)

(2) Labor market (32 series)

(3) Housing (10 series)

(4) Consumption, Orders and Inventory (14 series)

(5) Money and Credit (11 series)

(6) Bonds and Exchange rates (22 series)

(7) Prices (21 series)

(8) Stock market (4 series)

All variables are forecasted $h = 1, 3,$ and 12 months ahead. The initial training sample uses data between 1964:3[6] and 1999:12 which amounts to 430 observations. We allow for a maximum of 2 lags per equation, which together with an intercept requires the estimation of 263 parameters per equation. All the parameters are estimated on the initial sample, then forecasts of y_t at t=1999:12+h, $h = 1, 3, 12$ are computed. Parameters for all models are then re-estimated on data from 1964:3 to 2000:1 and forecasts computed at horizons $h = 1, 3, 12$. This expanding window scheme is repeated until the final out of sample forecast is computed for $2007 : 12$. At the one month horizon 96 forecasts are made and correspondingly less at longer horizons.

The categories mentioned above serve as natural groups for the adaptive group Lasso and we shall indeed use these as candidate groups for this estimator. Different groups are constructed for different lags, such that there are 16 groups for a VAR(2).

For each of the 131 series the mean square forecast errors relative to the mean square forecast error of a VAR(1) estimated by least squares are calculated[7]. Then the average of the relative mean square forecast errors is calculated within each group resulting in one measure of forecast accuracy for each of the eight groups mentioned above.

[6] Two initial observations are lost during the transformation of the variables to I(0)

[7] More precisely the lag length of the unrestricted VAR was chosen by BIC and it was always found to be one.

All forecasts are direct forecasts. The argument for direct forecasts is that they are tailored to the specific forecast horizon of interest. Furthermore, the absence of any sort of recursion makes direct forecasts relatively robust at longer forecast horizons. To forecast h periods ahead with the VAR, the following model is estimated.

$$y_{t+h} = \sum_{l=1}^{p} B_l^h y_{t-l+1} + \epsilon_{t+h}^h \ , \quad t \in [p, ..., T-h]$$

Here the superscript h highlights the fact that a separate model is estimated for each horizon. Our main interest is now to investigate how the estimators discussed fare when it comes to forecasting with this model.

5.2. **Implementation.** The Lasso and the adaptive Lasso are estimated using the `glmnet` package for `R 2.15`, which implements the algorithm by Friedman et al. (2010). The optimization problem is solved on an equally spaced grid of 100 values of the penalty parameter such that the largest λ_{max} on this grid is the smallest value of λ for which all variables are discarded from the model. The lower bound of the grid is computed as $\lambda_{min} = 10^{-5}\lambda_{max}$, which is close to zero. The value of λ_T is then selected by the Bayesian Information Criterion (BIC). γ is fixed at one and it is our experience that no gains can be achieved in terms of more precise forecasts by also searching over a grid of γs. The risk of overfitting in sample seems to be too high to justify such a search.

The adaptive group Lasso is estimated using the `grplasso` package, implementing the algorithm in Meier et al. (2008). Again λ_T is selected by BIC while γ is set to one. All the packages required for the computation of the results in this paper are publicly available at `CRAN`[8], and the code is available upon request.

[8] `www.cran.r-project.org`

5.3. **Competing models.** The forecasts of the above mentioned procedures are compared to forecasts from common factor models, simple linear autoregressions, smooth transition models, and the Bayesian VAR.

For the common factor model we follow the approach by Stock and Watson (2002). The forecasting equation for a given variable y_i is given by:

$$y_{t+h,i} = \alpha_i^h + \sum_{j=1}^{m} \hat{F}'_{t-j+1}\beta_{i,j}^h + \sum_{j=1}^{p} \delta_{i,j}^h y_{t-j+1,i} + \epsilon_{t+h,i}^h \ , \quad t \in [\max(m,p), ..., T-h]$$

The vector of common factors $\hat{F}_t$ and the parameters are estimated using a two step procedure. First the common factors $\hat{F}_t$ are estimated using principal component analysis on the training sample containing all 131 series. The number of principal components to retain for the second step is then selected and the parameters α_i^h, $\beta_{i,j}^h$, and $\delta_{i,j}^h$ are estimated by least squares on the training sample.

We use $m = p = 1$ throughout but vary the number of factors. Specifically, models with 1,3 and 5 common factors are estimated. The results for these models are denoted $CF1, CF3$ and $CF5$ in the tables below. In addition, results for a common factor model where the number of factors is chosen by BIC are reported. The number of common factors searched over is one to five. The corresponding results are denoted $CF\ BIC$ in the tables below. We also experimented using lags of the common factors, but this didn't bring substantial nor consistent improvement to the forecasting accuracy of the model.

The two univariate models considered are an $AR(1)$ and a Logistic Smooth Transition AutoRegressive (LSTAR) model. The forecasts from the $AR(1)$ model for $y_{t,i}$ are generated by

$$y_{t+h,i} = \alpha_i^h + \beta_i^h y_{t,i} + \epsilon_{t+h,i}^h \ , \quad t \in [1, ..., T-h]$$

where the parameters are estimated by least squares. The forecasts of the LSTAR (see Teräsvirta (1994)) are created by the following model for variable $y_{t,i}$

$$\begin{aligned} y_{t+h,i} &= \left(\alpha_{1,i}^h + \beta_{1,i}^h y_{t,i}\right)\left(1 - G\left(y_{t,i}, \gamma_i, \tau_i\right)\right) \\ &\quad + \left(\alpha_{2,i}^h + \beta_{2,i}^h y_{t,i}\right)\left(G\left(y_{t,i}, \gamma_i, \tau_i\right)\right) + \epsilon_{t+h,i}^h \end{aligned}$$

where $G(y_{t,i}, \gamma_i, \tau_i) = \left(1 + \exp\left[-\gamma_i(y_{t,i} - \tau_i)\right]\right)^{-1}$ is the logistic function. For the LSTAR we use y_t as the threshold variable. τ_i indicates the location of the transition and γ_i measures the speed of transition.

Forecasts based on a Bayesian VAR (BVAR) are also computed. The model is estimated as in Bańbura et al. (2009), imposing a Minnesota prior on the transformed series, which are all assumed to be stationary. The VAR can be estimated by ordinary least squares. The prior is imposed by augmenting the sample with matrices of dummy observations, the construction of which is described in Bańbura et al. (2009). The hyperparameter λ governing the tightness of the prior distribution is selected using a binary search algorithm seeking to minimize the BIC. We estimate and forecast with a BVAR containing a single lag, since it turned out to deliver more precise forecasts than a BVAR using two lags.

5.4. **Insane forecasts.** It is well known that in particular non-linear models may sometimes provide forecasts that are clearly unreasonable. Swanson and White (1995) suggests to weed out unreasonable forecasts by means of an insanity filter. We shall follow this suggestion by replacing a forecast by the most recent observation of the estimation window if it does not lie in the interval given by the most recent observation in the estimation window plus/minus three times the standard deviation of the data in the estimation window. As noted in Kock and Teräsvirta (2012) the particular choice of insanity filter is not overly important – what matters is that the insane forecasts are weeded out. To treat all forecast procedures on an equal footing the insanity filter is implemented for all procedures.

5.5. **Results.** Table 5.1 contains the relative mean square forecast errors (relative MSE) for each group of variables as well as for all variables when averaged over all horizons $h = 1, 3$, and 12. The lowest relative MSE is in bold face. From this table it is seen that the plain Lasso gives the most precise forecasts for most categories, and that it is the overall best estimator. The Lasso actually has a relative mean square forecast error below one for all groups of variables indicating its stability. Note also that except for the output and income group and the stock market group the most precise procedure is always the Lasso.

It is striking that the plain Lasso almost always outperforms its adaptive versions, sometimes by a wide margin. The BVAR also gives quite precise forecasts. The poor performance of the BVAR and both adaptive estimators for the housing group as well as to a lesser degree for the other groups has a common cause. All three estimators use a VAR estimated by OLS as the initial estimator (the BVAR uses a VAR(1) and the adaptive (group) Lasso uses a VAR(2)), providing weights for the adaptive Lasso penalties and estimates of the variance of the residuals used to construct the prior for the BVAR. The VAR(2) performs quite poorly relative to the VAR(1) in terms of MSE, and even more so at longer horizons, resulting in bad initial estimates for the second step estimators[9]. Given this very poor initial estimator it is not too surprising that the second step estimators forecast (relatively) inaccurately. Despite this impediment the adaptive procedures provide forecasts that are in general close to, or better than, common factor forecasts.

As can be expected from a non-linear procedure like the LSTAR, it performs very well for some series and quite poorly for others. This is in line with the commonly made observation that non-linear procedures are somewhat "risky" – a fact which can make

[9] We also experimented with using the VAR(2) as initial estimator for the BVAR but this yielded very imprecise forecasts in particular at the 12 month horizon. In this sense the BVAR actually has an unfair advantage compared to the adaptive (group) Lasso since it is based on a more precise initial estimator.

	Output and Income	Labor Market	Housing	Consumption Orders Inventories	Money and Credit	Bonds and Exchange Rates	Prices	Stock Market	Total
BVAR	0.561	0.600	1.262	0.645	0.640	0.772	0.714	0.588	0.723
Lasso	0.446	**0.464**	**0.466**	**0.490**	**0.585**	0.359	**0.616**	0.455	**0.485**
aLasso	0.494	0.647	3.961	0.680	0.609	**0.321**	0.732	0.479	0.990
agLasso	0.524	0.698	6.537	0.795	0.613	1.468	0.743	**0.454**	1.479
				Factor model forecasts					
CF 1	0.429	0.495	0.778	0.569	0.629	0.416	0.686	0.495	0.562
CF 3	**0.417**	0.512	0.806	0.522	0.630	0.399	0.663	0.512	0.558
CF 5	0.421	0.515	0.772	0.524	0.630	0.403	0.669	0.510	0.555
CF BIC	0.436	0.508	0.829	0.578	0.643	0.443	0.683	0.499	0.577
				Univariate forecasts					
lstar	0.683	0.672	0.752	0.703	1.146	0.483	11.066	0.871	2.047
ar(1)	0.959	0.862	0.764	0.932	0.774	0.687	1.185	1.216	0.922

Table 5.1. Mean square errors relative to the VAR(1). Average across all horizons. The last column, Total, contains the average across all series.

them very useful in forecast combination schemes. To highlight this riskiness note that the LSTAR outperforms the plain AR(1) for six out of eight series while it still has a considerably larger relative mean square forecast error than common factor models and Lasso-type estimators due to its occasionally very imprecise forecasts.

Factor models deliver reasonably precise forecasts for all types of variables resulting in relative mean square forecast errors below one for all groups. On the other hand, no gains seem to be made from applying BIC to select the number of factors as opposed to simply fixing the number of these.

	Output and Income	Labor Market	Housing	Consumption Orders Inventories	Money and Credit	Bonds and Exchange Rates	Prices	Stock Market
BVAR	0.947	0.905	0.865	0.914	0.871	0.859	0.926	0.892
Lasso	0.639	**0.557**	**0.591**	0.567	0.667	**0.362**	0.707	**0.521**
aLasso	0.763	0.883	6.257	1.066	0.721	0.367	1.013	0.590
agLasso	0.854	1.076	10.609	1.314	0.732	2.855	1.030	0.517
				Factor model forecasts				
CF 1	0.600	0.683	0.825	0.753	0.817	0.543	0.874	0.585
CF 3	**0.555**	0.717	0.825	0.649	0.817	0.518	0.824	0.589
CF 5	0.569	0.733	0.831	0.674	0.818	0.539	0.826	0.585
CF BIC	0.601	0.683	0.824	0.753	0.816	0.543	0.874	0.585
				Univariate forecasts				
lstar	0.581	0.580	0.615	0.526	1.128	0.363	0.616	0.597
ar(1)	0.580	0.578	0.618	**0.511**	**0.423**	0.373	**0.613**	0.562

Table 5.2. Mean square error relative to the VAR(1). 96 one step ahead forecasts.

	Output	Labor	Housing	Consumption	Money	Bonds and	Prices	Stock
	and	Market		Orders	and	Exchange		Market
	Income			Inventories	Credit	Rates		
Lasso								
Output	0.025	0.040	0.012	**0.059**	0.016	0.021	0.001	0.004
Labor	0.024	**0.094**	0.025	0.048	0.010	0.018	0.007	0.030
Housing	0.009	0.024	**0.218**	0.008	0.052	0.051	0.002	0.064
Consumption	0.022	0.030	0.045	**0.107**	0.017	0.018	0.006	0.019
Money	0.012	0.018	0.009	0.020	**0.137**	0.037	0.016	0.026
Bonds	0.007	0.028	0.021	0.035	0.007	0.079	0.023	**0.081**
Prices	0.016	0.003	0.002	0.033	0.033	0.012	**0.077**	0.000
Stock	0.000	0.007	0.000	0.039	0.024	0.073	0.010	**0.138**
adaptive Lasso								
Output	0.011	0.013	0.000	**0.104**	0.065	0.002	0.020	0.001
Labor	0.012	0.025	0.002	**0.100**	0.070	0.002	0.020	0.001
Housing	0.036	0.051	0.059	**0.233**	0.095	0.014	0.039	0.006
Consumption	0.009	0.012	0.001	**0.078**	0.048	0.002	0.017	0.001
Money	0.001	0.002	0.001	0.009	**0.022**	0.001	0.003	0.000
Bonds	0.018	0.019	0.001	**0.108**	0.059	0.007	0.023	0.004
Prices	0.001	0.001	0.000	**0.012**	0.010	0.000	0.005	0.000
Stock	0.000	0.003	0.000	0.024	**0.083**	0.000	0.021	0.000
adaptive group Lasso								
Output	0.030	0.032	0.012	**0.052**	0.016	0.020	0.001	0.004
Labor	0.038	**0.078**	0.023	0.036	0.010	0.016	0.007	0.024
Housing	0.006	0.017	**0.188**	0.002	0.033	0.044	0.001	0.031
Consumption	0.018	0.024	0.029	**0.087**	0.015	0.016	0.001	0.014
Money	0.012	0.013	0.008	0.031	**0.111**	0.030	0.014	0.036
Bonds	0.016	0.015	0.019	0.023	0.001	0.068	0.011	**0.078**
Prices	0.020	0.003	0.003	0.014	0.025	0.010	**0.067**	0.017
Stock	0.000	0.006	0.000	0.044	0.043	**0.071**	0.004	0.037

TABLE 5.3. Selection frequency for each group of variables in each equation. Average over 96 forecasts at horizon 1. Largest share of selected variables in bold for each equation.

Table 5.2 contains the relative mean square forecast errors for each group of variables at the one month horizon. Table 5.2 shows that common factor models as well as the Lasso and the adaptive Lasso deliver forecasts that are in general more accurate than those obtained by a VAR estimated by least squares. The Lasso outperforms common factor models for most groups. The adaptive group Lasso performs quite poorly, faring worse than the benchmark VAR in 5 of the groups while being the best model for the Stock Market series.

The Bayesian VAR performs better than the benchmark VAR but by a smaller margin than the Lasso.

The two univariate forecasting models have very similar forecasting performances in most instances, being close to the best models. The AR(1) forecasts are the most precise

	Output	Labor	Housing	Consumption	Money	Bonds and	Prices	Stock	Total
	and	Market		Orders	and	Exchange		Market	
	Income			Inventories	Credit	Rates			
				Lasso					
Output	0.833	**2.591**	0.238	1.664	0.354	0.919	0.037	0.030	6.665
Labor	0.804	**6.034**	0.506	1.340	0.223	0.779	0.275	0.236	10.196
Housing	0.301	1.557	**4.356**	0.227	1.151	2.227	0.086	0.512	10.419
Consumption	0.754	1.911	0.900	**3.006**	0.369	0.790	0.256	0.155	8.141
Money	0.403	1.149	0.176	0.565	**3.023**	1.606	0.655	0.209	7.787
Bonds	0.244	1.773	0.419	0.974	0.145	**3.473**	0.973	0.647	8.647
Prices	0.542	0.211	0.036	0.936	0.728	0.537	**3.224**	0.001	6.216
Stock	0.003	0.427	0.000	1.081	0.526	**3.208**	0.419	1.107	6.771
				adaptive Lasso					
Output	0.385	0.857	0.009	**2.920**	1.420	0.072	0.857	0.011	6.531
Labor	0.422	1.586	0.033	**2.807**	1.546	0.093	0.860	0.007	7.353
Housing	1.207	3.275	1.172	**6.511**	2.092	0.636	1.641	0.050	16.584
Consumption	0.295	0.768	0.018	**2.187**	1.066	0.068	0.708	0.009	5.119
Money	0.023	0.100	0.027	0.243	**0.486**	0.038	0.117	0.004	1.039
Bonds	0.597	1.230	0.029	**3.031**	1.298	0.330	0.950	0.031	7.497
Prices	0.018	0.056	0.000	**0.333**	0.230	0.003	0.189	0.001	0.831
Stock	0.010	0.174	0.000	0.682	**1.826**	0.005	0.865	0.000	3.562
				adaptive group Lasso					
Output	1.478	0.000	0.037	**4.904**	0.310	0.000	0.000	0.039	6.768
Labor	0.044	3.556	1.405	**2.873**	0.007	0.000	0.000	0.186	8.072
Housing	0.000	0.000	**0.188**	0.000	0.000	0.000	0.000	0.000	0.188
Consumption	0.000	0.000	0.000	**5.874**	0.000	0.000	0.000	0.112	5.986
Money	0.000	0.000	0.114	0.000	**5.884**	0.000	0.000	0.045	6.043
Bonds	0.000	0.000	0.758	0.318	0.000	**5.057**	0.000	0.717	6.850
Prices	0.000	0.000	0.000	0.000	0.436	0.000	**11.476**	0.004	11.916
Stock	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**1.646**	1.646

Table 5.4. Average number of variables per equation. Average over 96 forecasts at horizon 1. Largest number of selected variables in bold for each equation. The rightmost column, Total, gives the average total number of variables included in each equation.

for three groups at this horizon. The LSTAR model is less stable than the AR(1), being almost the best model for Bonds and Exchange Rates and the worst for Money and Credit.

To shed further light on these findings Table 5.3 reports the share of variables from a given group (in columns) retained in the equations for variables from another given group (in rows). The two leftmost entries of the first row in Table 5.3 should be read as: in the equations where the left-hand side variable belongs to the Output and Income group, 2.5% of the candidate explanatory variables from the Output and Income group were retained and 4% of the candidate explanatory variables belonging to the Labor Market group were retained on average. The boldfaced number is the largest share for a given row. Some striking differences between the behavior of the adaptive Lasso and the other two regularization estimators appear. The adaptive Lasso selects a large share of variables belonging to the Consumption, Orders, Inventories group for most equations. The largest share selected by the other two estimators is often on the diagonal: lags of variables belonging to the same group as the left hand side variable are most often used as predictors. Another feature is that most of these shares are quite small, indicating the selected models are very sparse. This is confirmed by Table 5.4 which reports the average number of variables selected per group and for the whole equation. The Lasso selects between 6 and 11 out of the 262 candidate variables in each equations. The adaptive group Lasso often selects no variables at all in the housing equations, with an average of 0.188 variables per equation. Interestingly, this is also the group where this estimator performs worst. As mentioned previously this seems to be due to the imprecise initial estimates.

	Output	Labor	Housing	Consumption	Money	Bonds and	Prices	Stock
	and	Market		Orders	and	Exchange		Market
	Income			Inventories	Credit	Rates		
BVAR	0.436	0.580	1.270	0.711	0.552	1.100	0.666	0.445
Lasso	**0.372**	**0.456**	**0.450**	**0.507**	**0.549**	0.391	**0.586**	**0.425**
aLasso	0.408	0.626	4.342	0.627	0.559	**0.336**	0.635	0.429
agLasso	0.417	0.668	7.304	0.746	0.567	1.202	0.650	0.427
				Factor model forecasts				
CF 1	0.389	0.480	0.834	0.626	0.557	0.427	0.642	0.448
CF 3	0.381	0.480	0.877	0.574	0.560	0.411	0.623	0.473
CF 5	0.388	0.486	0.870	0.567	0.559	0.409	0.632	0.469
CF BIC	0.375	0.475	0.841	0.611	0.557	0.465	0.636	0.450
				Univariate forecasts				
lstar	0.7334	0.7178	0.8210	0.7918	1.1548	0.5436	16.2912	1.0083
ar(1)	1.1481	1.0042	0.8365	1.1427	0.9493	0.8433	1.4706	1.5428

TABLE 5.5. Mean square error relative to the VAR(1). 94 three step ahead forecasts.

Table 5.5 reports the results for the 3-months ahead forecasts. The Lasso consistently forecasts more precisely than every other procedure except for the Bonds and Exchange Rates group where it is not far behind the adaptive Lasso. The relative MSEs are slightly smaller than those obtained at the 1-month horizon. The two adaptive estimators still perform very poorly for the Housing group. At this horizon the BVAR is less precise than the Lasso and for two series it is even less precise than the benchmark. Just as for the one month horizon, common factor models provide very reliable forecasts that are almost as accurate as the Lasso forecasts for most series. The LSTAR performs quite well for a few groups, and in general outperforms the AR(1). However, it fails badly for the Prices group. Similar observations can be made for 6-month ahead forecasts (not reported).

At the one year horizon (Table 5.6) the relative mean square errors of most procedures are even lower than at shorter horizons. The adaptive Lasso delivers the most accurate forecasts for the Bonds and Exchange Rates group as well as for the Stock market groups. As previously it is close to the best procedure for most other groups. The Lasso delivers forecasts that are very close to being the most precise for every groups, as do to a lesser extent the common factor models. The BVAR performs very well at this horizon, being the best forecast method for half of the groups, despite faring poorly for the housing group. However, this poor performance on the housing group can be found for any procedure based on the VAR(2) initial estimates.

	Output	Labor	Housing	Consumption	Money	Bonds and	Prices	Stock
	and	Market		Orders	and	Exchange		Market
	Income			Inventories	Credit	Rates		
BVAR	**0.299**	**0.316**	1.650	**0.311**	**0.497**	0.357	0.551	0.426
Lasso	0.328	0.379	**0.357**	0.397	0.538	0.325	0.554	**0.418**
aLasso	0.310	0.431	1.285	0.346	0.548	**0.260**	0.548	**0.418**
agLasso	0.302	0.350	1.699	0.324	0.539	0.348	0.549	**0.418**
				Factor model forecasts				
CF 1	**0.299**	0.322	0.674	0.328	0.512	0.279	0.542	0.451
CF 3	0.314	0.340	0.716	0.343	0.513	0.267	0.542	0.473
CF 5	0.306	0.325	0.614	0.331	0.514	0.261	0.550	0.475
CF BIC	0.331	0.365	0.822	0.370	0.557	0.320	**0.540**	0.462
				Univariate forecasts				
lstar	0.7334	0.7178	0.8210	0.7918	1.1548	0.5436	16.2912	1.0083
ar(1)	1.1481	1.0042	0.8365	1.1427	0.9493	0.8433	1.4706	1.5428

Table 5.6. Mean square error relative to the recursive VAR(1). 85 twelve step ahead forecasts.

6. Conclusion

In this paper we have studied the properties of the adaptive Lasso and the adaptive group Lasso when applied to stationary vector autoregressions of a fixed dimension. The adaptive Lasso was shown to possess the oracle property in the sense that all truly zero parameters will be classified as such asymptotically, while the estimators of the non-zero parameters have the same asymptotic distribution as if least squares had been used to estimate the model *only* including the relevant variables.

Since many variables are naturally classified into groups of similar variables (like in the large macroeconomic dataset used in this paper) one may naturally ask the question whether certain *group* of variables are relevant for the task of explaining another variable. For this reason the asymptotic properties of the adaptive group Lasso were investigated and it was shown that it possesses a version of the oracle property.

The performance of these two estimators in terms of forecast precision was investigated by comparing different forecasting procedures using the data by Ludvigson and Ng (2009). The plain Lasso was found to give the most precise forecasts on average while its adaptive variants had problems forecasting, in particular the housing series, due to imprecise initial least squares estimates of the VAR(2) model. The forecasts from the common factor models were relatively precise for all series while the non-linear LSTAR was much more volatile.

Even though this work has been concerned with stationary autoregressions we believe that the results may be generalized to integrated series by using the limit results in Sims et al. (1990). This would alter the rates of convergences obtained in this paper since the rates would now depend on the order of integration.

7. Proof (appendix)

Proof of Theorem 1: The proof is inspired by the proof of Theorem 2 in Zou (2006) and the proof of Theorem 2 in Kock (2012). Letting $\beta = \beta^* + \frac{u}{\sqrt{T}}$ the objective function (3) may also be written as

$$L_T(u) = \left\| y - X\left(\beta^* + \frac{u}{\sqrt{T}}\right)\right\|^2 + \lambda_T \sum_{i=1}^{N^2 p} \hat{w}_i \left|\beta_i^* + \frac{u_i}{\sqrt{T}}\right| \tag{5}$$

Let $\hat{u} = \arg\min L_T(u)$. It follows that $\hat{\beta} = \beta^* + \frac{\hat{u}}{\sqrt{T}}$ and so $\hat{u} = \sqrt{T}\left(\hat{\beta} - \beta^*\right)$. Next, define

$$\begin{aligned} V_T(u) &= L_T(u) - L_T(0) \\ &= \left\| y - X\left(\beta^* + \frac{u}{\sqrt{T}}\right)\right\|^2 - \left\| y - X\beta^*\right\|^2 + \lambda_T \sum_{i=1}^{N^2 p} \hat{w}_i \left(\left| \beta_i^* + \frac{u_i}{\sqrt{T}} \right| - |\beta_i^*| \right) \\ &= u' \frac{X'X}{T} u - 2\frac{u'X'\epsilon}{\sqrt{T}} + \lambda_T \sum_{i=1}^{N^2 p} \hat{w}_i \left(\left| \beta_i^* + \frac{u_i}{\sqrt{T}} \right| - |\beta_i^*| \right) \end{aligned} \tag{6}$$

By Theorem 11.2.1 in Brockwell and Davis (2009) it follows that $\frac{u'X'Xu}{T} \to u'(I_N \otimes C)u$ in probability for any $u \in \mathbb{R}^{N^2 p}$. Furthermore, it follows from Proposition 7.9 in Hamilton (1994) (see also expression 11.A.3 page 341 in Hamilton (1994)) that $\frac{X'\epsilon}{\sqrt{T}} \to_d W \sim \mathcal{N}(0, \Sigma \otimes C)$. Hence,

$$u' \frac{X'X}{T} u - 2\frac{u'X'\epsilon}{\sqrt{T}} \to_d u'(I_N \otimes C)u - 2u'W \tag{7}$$

In addition, if $\beta_i^* \neq 0$

$$\begin{aligned} \lambda_T \hat{w}_i \left(\left| \beta_i^* + \frac{u_i}{\sqrt{T}} \right| - \left| \beta_i^* \right| \right) &= \lambda_T \left| \frac{1}{\hat{\beta}_{I,i}} \right|^{\gamma} \frac{u_i}{\sqrt{T}} \left(\left| \beta_i^* + \frac{u_i}{\sqrt{T}} \right| - \left| \beta_i^* \right| \right) / \left(\frac{u_i}{\sqrt{T}} \right) \\ &= \frac{\lambda_T}{T^{1/2}} \left| \frac{1}{\hat{\beta}_{I,i}} \right|^{\gamma} u_i \left(\left| \beta_i^* + \frac{u_i}{\sqrt{T}} \right| - \left| \beta_i^* \right| \right) / \left(\frac{u_i}{\sqrt{T}} \right) \\ &\to 0 \text{ in probability} \end{aligned} \tag{8}$$

for every $u_i \in \mathbb{R}$ since (i): $\lambda_T / T^{1/2} \to 0$, (ii): $\left|1/\hat{\beta}_{I,i}\right|^{\gamma} \to \left|1/\beta_i^*\right|^{\gamma} < \infty$ in probability and (iii): $u_i \left(\left| \beta_i^* + \frac{u_i}{\sqrt{T}} \right| - \left| \beta_i^* \right| \right) / \left(\frac{u_i}{\sqrt{T}} \right) \to u_i \text{sign}(\beta_i^*)$.

If, on the other hand, $\beta_i^* = 0$

$$\lambda_T \hat{w}_i \left(\left| \beta_i^* + \frac{u_i}{\sqrt{T}} \right| - |\beta_i^*| \right) = \frac{\lambda_T}{T^{1/2}} \left| \frac{1}{\hat{\beta}_{I,i}} \right|^{\gamma} |u_i| = \frac{\lambda_T}{T^{1/2-\gamma/2}} \left| \frac{1}{\sqrt{T}\hat{\beta}_{I,i}} \right|^{\gamma} |u_i|$$

$$\to \begin{cases} \infty \text{ in probability if } u_i \neq 0 \\ 0 \text{ in probability if } u_i = 0 \end{cases} \tag{9}$$

since (i): $\frac{\lambda_T}{T^{1/2-\gamma/2}} \to \infty$ and (ii): $\sqrt{T}\hat{\beta}_{I,i}$ is tight. Using the convergence results (7)-(9) in (6) yields

$$V_T(u) \to_d V(u) = \begin{cases} u'(I_N \otimes C)u - 2u'W \text{ if } u_i = 0 \text{ for all } i \in \mathcal{A}^c \\ \infty \ \text{ if } u_i \neq 0 \text{ for some } i \in \mathcal{A}^c \end{cases}$$

Since $V_T(u)$ is convex and $V(u)$ has a unique minimum it follows from Knight (1999) (or alternatively Knight and Fu (2000)) that $\arg\min V_T(u) \to_d \arg\min V(u)$. Hence,

$$\hat{u}_{\mathcal{A}^c} \to_d \delta_0^{|\mathcal{A}^c|} \tag{10}$$

$$\hat{u}_{\mathcal{A}} \to_d N\left(0, [(I_N \otimes C)_{\mathcal{A}}]^{-1} [\Sigma \otimes C]_{\mathcal{A}} [(I_N \otimes C)_{\mathcal{A}}]^{-1}\right) \tag{11}$$

where δ_0 is the Dirac measure at 0 and $|\mathcal{A}^c|$ is the cardinality of $\mathcal{A}^c$ (hence, $\delta_0^{|\mathcal{A}^c|}$ is the $|\mathcal{A}^c|$-dimensional Dirac measure at 0). Notice that (10) implies that $\hat{u}_{\mathcal{A}^c} \to 0$ in probability. An equivalent formulation of (10)-(11) is

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}^c} - \beta^*_{\mathcal{A}^c}) \to_d \delta_0^{|\mathcal{A}^c|} \tag{12}$$

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta^*_{\mathcal{A}}) \to_d N\left(0, [(I_N \otimes C)_{\mathcal{A}}]^{-1} [\Sigma \otimes C]_{\mathcal{A}} [(I_N \otimes C)_{\mathcal{A}}]^{-1}\right) \tag{13}$$

(12)-(13) yield the consistency part of the theorem at the rate of $\sqrt{T}$ for $\hat{\beta}$. (13) also yields the oracle efficient asymptotic distribution for $\hat{\beta}_{\mathcal{A}}$, i.e. part (3) of the theorem. It remains to show part (2) of the theorem; $P(\hat{\beta}_{\mathcal{A}^c} = 0) \to 1$.

Assume $\hat{\beta}_j \neq 0$ for $j \in \mathcal{A}^c$. Then, letting x_j denote the jth column of X, it follows from the first order conditions

$$2x_j'(y - X\hat{\beta}) + \lambda_T \hat{w}_j \text{sign}(\hat{\beta}_j) = 0$$

or equivalently,

$$\frac{2x_j'\left(y - X\hat{\beta}\right)}{T^{1/2}} + \frac{\lambda_T \hat{w}_j \text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0 \tag{14}$$

First, consider the second term in (14)

$$\left|\frac{\lambda_T \hat{w}_j \text{sign}(\hat{\beta}_j)}{T^{1/2}}\right| = \frac{\lambda_T \hat{w}_j}{T^{1/2}} = \frac{\lambda_T}{T^{1/2-\gamma/2}\left|T^{1/2}\hat{\beta}_{I,j}\right|^{\gamma}} \to \infty$$

since $\sqrt{T}\hat{\beta}_{I,j}$ is tight. Regarding the first term in (14),

$$\begin{aligned}\frac{2x_j'\left(y - X\hat{\beta}\right)}{T^{1/2}} &= \frac{2x_j'\left(\epsilon - X[\hat{\beta} - \beta^*]\right)}{T^{1/2}} \\ &= \frac{2x_j'\epsilon}{T^{1/2}} - \frac{2x_j'X}{T}\sqrt{T}[\hat{\beta} - \beta^*]\end{aligned}$$

Assuming β_j is the coefficient to the kth variable in the ith equation (so the jth column of X is the kth variable in the ith equation) it follows from the same arguments as those preceding (6) that $\frac{x_j'\epsilon}{T^{1/2}} \to_d N(0, \Sigma_{ii}C_{kk})$. Furthermore, $\frac{x_j'X}{T} \to_p (I_N \otimes C)_j$ where $(I_N \otimes C)_j$ is the jth row of $(I_N \otimes C)$. Hence, $\frac{x_j'\epsilon}{T^{1/2}}$ and $\frac{x_j'X}{T}$ are tight. The same is the case for $\sqrt{T}[\hat{\beta} - \beta^*]$ since it converges weakly by (12)-(13). Taken together, $\frac{2x_j'\left(y - X\hat{\beta}\right)}{T^{1/2}}$ is tight and so

$$P(\hat{\beta}_j \neq 0) \leq P\left(\frac{2x_j'\left(y - X\hat{\beta}\right)}{T^{1/2}} + \frac{\lambda_T \hat{w}_j \text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0\right) \to 0$$

□

Proof of Theorem 2. The idea of the proof is similar to the one of Theorem 1. Letting $\beta = \beta^* + \frac{u}{\sqrt{T}}$ the objective function (4) may also be written as

$$\tilde{L}_T(u) = \left\| y - X\left(\beta^* + \frac{u}{\sqrt{T}}\right)\right\|^2 + \lambda_T \sum_{i=1}^{M} \tilde{w}_i \left\|\beta^*_{G_i} + \frac{u_{G_i}}{\sqrt{T}}\right\| \tag{15}$$

Let $\tilde{u} = \arg\min \tilde{L}_T(u)$. It follows that $\tilde{\beta} = \beta^* + \frac{\tilde{u}}{\sqrt{T}}$ and so $\tilde{u} = \sqrt{T}\left(\tilde{\beta} - \beta^*\right)$. Next, define

$$\begin{aligned}
\tilde{V}_T(u) &= \tilde{L}_T(u) - \tilde{L}_T(0)\\
&= \left\| y - X\left(\beta^* + \frac{u}{\sqrt{T}}\right)\right\|^2 - \left\|y - X\beta^*\right\|^2 + \lambda_T \sum_{i=1}^{M} \tilde{w}_i \left(\left\|\beta^*_{G_i} + \frac{u_{G_i}}{\sqrt{T}}\right\| - \left\|\beta^*_{G_i}\right\|\right)\\
&= u'\frac{X'X}{T}u - 2\frac{u'X'\epsilon}{\sqrt{T}} + \lambda_T \sum_{i=1}^{M} \tilde{w}_i \left(\left\|\beta^*_{G_i} + \frac{u_{G_i}}{\sqrt{T}}\right\| - \left\|\beta^*_{G_i}\right\|\right) \qquad (16)
\end{aligned}$$

By Theorem 11.2.1 in Brockwell and Davis (2009) it follows that $\frac{u'X'Xu}{T} \to u'(I_N \otimes C)u$ in probability for any $u \in \mathbb{R}^{N^2p}$. Furthermore, it follows from Proposition 7.9 in Hamilton (1994) (see also expression 11.A.3 page 341 in Hamilton (1994)) that $\frac{X'\epsilon}{\sqrt{T}} \to_d W$ where $W \sim \mathcal{N}(0, \Sigma \otimes C)$. Hence,

$$u'\frac{X'X}{T}u - 2\frac{u'X'\epsilon}{\sqrt{T}} \to_d u'(I_N \otimes C)u - 2u'W \tag{17}$$

In addition, if $\beta^*_{G_i} \neq 0$, it follows by continuity of the norm that

$$\left|\lambda_T \tilde{w}_i \left(\left\|\beta^*_{G_i} + \frac{u_{G_i}}{\sqrt{T}}\right\| - \left\|\beta^*_{G_i}\right\|\right)\right| \leq \lambda_T \tilde{w}_i \left\|\frac{u_{G_i}}{\sqrt{T}}\right\| = \frac{\lambda_T}{\sqrt{T}} \frac{\left\|u_{G_i}\right\|}{\left\|\hat{\beta}_{I,G_i}\right\|^{\gamma}} \tag{18}$$

$$\to 0 \text{ in probability}$$

since (i): $\lambda_T / T^{1/2} \to 0$ and (ii): $\frac{1}{\left\|\hat{\beta}_{I,G_i}\right\|^{\gamma}} \to \frac{1}{\left\|\beta^*_{G_i}\right\|^{\gamma}} < \infty$ in probability. If, on the other hand, $\beta^*_{G_i} = 0$

$$\lambda_T \tilde{w}_i \left(\left\|\beta^*_{G_i} + \frac{u_{G_i}}{\sqrt{T}}\right\| - \left\|\beta^*_{G_i}\right\|\right) = \lambda_T \tilde{w}_i \left\|\frac{u_{G_i}}{\sqrt{T}}\right\| = \frac{\lambda_T}{T^{1/2-\gamma/2}} \frac{\left\|u_{G_i}\right\|}{\left\|\sqrt{T}\hat{\beta}_{I,G_i}\right\|^{\gamma}}$$

$$\to \begin{cases} \infty \text{ in probability if } u_{G_i} \neq 0 \\ 0 \text{ in probability if } u_{G_i} = 0 \end{cases} \tag{19}$$

since (i): $\frac{\lambda_T}{T^{1/2-\gamma/2}} \to \infty$ and (ii): $\sqrt{T}\hat{\beta}_{I,G_i}$ is tight. Using the convergence results (17)-(19) in (16)

$$\tilde{V}_T(u) \to_d \tilde{V}(u) = \begin{cases} u'(I_N \otimes C)u - 2u'W \text{ if } u_{G_i} = 0 \text{ for all } i \in \mathcal{A}^c \\ \infty \ \text{ if } u_{G_i} \neq 0 \text{ for some } i \in \mathcal{A}^c \end{cases}$$

Since $\tilde{V}_T(u)$ is convex and $\tilde{V}(u)$ has a unique minimum it follows from Knight (1999) (or alternatively Knight and Fu (2000)) that $\arg\min \tilde{V}_T(u) \to_d \arg\min \tilde{V}(u)$. Hence,

$$\tilde{u}_{\mathcal{G}^c} \to_d \delta_0^{|\mathcal{G}^c|} \tag{20}$$

$$\tilde{u}_{\mathcal{G}} \to_d N\left(0, [(I_N \otimes C)_{\mathcal{G}}]^{-1}[\Sigma \otimes C]_{\mathcal{G}}[(I_N \otimes C)_{\mathcal{G}}]^{-1}\right) \tag{21}$$

where δ_0 is the Dirac measure at 0 and $|\mathcal{G}^c|$ is the cardinality of $\mathcal{G}^c$. Notice that (20) implies that $\tilde{u}_{\mathcal{G}^c} \to 0$ in probability. An equivalent formulation of (20)-(21) is

$$\sqrt{T}(\tilde{\beta}_{\mathcal{G}^c} - \beta^*_{\mathcal{G}^c}) \to_d \delta_0^{|\mathcal{G}^c|} \tag{22}$$

$$\sqrt{T}(\tilde{\beta}_{\mathcal{G}} - \beta^*_{\mathcal{G}}) \to_d N\left(0, [(I_N \otimes C)_{\mathcal{G}}]^{-1}[\Sigma \otimes C]_{\mathcal{G}}[(I_N \otimes C)_{\mathcal{G}}]^{-1}\right) \tag{23}$$

(22)-(23) yield the consistency part of the theorem at the rate of $\sqrt{T}$ for $\tilde{\beta}$. (23) also yields the asymptotic distribution for $\tilde{\beta}_{\mathcal{G}}$, i.e. part 3 of the theorem. It remains to show part 2 of the theorem; $P(\tilde{\beta}_{\mathcal{G}^c} = 0) \to 1$.

Assume $\tilde{\beta}_{G_i} \neq 0$ for $i \in \mathcal{A}^c$. Then all entries $\tilde{\beta}_{G_i,j},\ 1 \leq j \leq |G_i|$ satisfy the first order condition

$$2x_j'(y - X\tilde{\beta}) + \lambda_T \tilde{w}_i \left\|\tilde{\beta}_{G_i}\right\|^{-1} \tilde{\beta}_{G_i,j} = 0$$

or equivalently,

$$\frac{2x_j'(y - X\tilde{\beta})}{T^{1/2}} + \frac{\lambda_T \tilde{w}_i \left\|\tilde{\beta}_{G_i}\right\|^{-1} \tilde{\beta}_{G_i,j}}{T^{1/2}} = 0$$

This also implies

$$\max_{1 \leq j \leq |G_i|} \left| \frac{2x_j'(y - X\tilde{\beta})}{T^{1/2}} \right| = \max_{1 \leq j \leq |G_i|} \left| \frac{\lambda_T \tilde{w}_i \left\|\tilde{\beta}_{G_i}\right\|^{-1} \tilde{\beta}_{G_i,j}}{T^{1/2}} \right| \tag{24}$$

First, consider the right hand side of (24). To this end note that

$$\frac{\max_{1 \leq j \leq |G_i|} |\tilde{\beta}_{G_i,j}|}{\left\|\tilde{\beta}_{G_i}\right\|} \geq \frac{\max_{1 \leq j \leq |G_i|} |\tilde{\beta}_{G_i,j}|}{\sum_{j=1}^{|G_i|} |\tilde{\beta}_{G_i,j}|} \geq \frac{\max_{1 \leq j \leq |G_i|} |\tilde{\beta}_{G_i,j}|}{|G_i| \max_{1 \leq j \leq |G_i|} |\tilde{\beta}_{G_i,j}|} = \frac{1}{|G_i|}$$

This implies

$$\max_{1\leq j\leq |G_i|}\left|\frac{\lambda_T \tilde{w}_i \left\|\tilde{\beta}_{G_i}\right\|^{-1}\tilde{\beta}_{G_i,j}}{T^{1/2}}\right| = \frac{\lambda_T \tilde{w}_i}{T^{1/2}}\frac{\max_{1\leq j\leq |G_i|}|\tilde{\beta}_{G_i,j}|}{\left\|\tilde{\beta}_{G_j}\right\|}$$
$$\geq \frac{\lambda_T}{T^{1/2-\gamma/2}}\frac{1}{\left\|T^{1/2}\hat{\beta}_{I,G_i}\right\|^{\gamma}}\frac{1}{|G_i|} \to_p \infty$$

since $\sqrt{T}\hat{\beta}_{I,G_i}$ is tight. Regarding the left hand side in (24),

$$\frac{2x_j'\left(y - X\tilde{\beta}\right)}{T^{1/2}} = \frac{2x_j'\left(\epsilon - X[\tilde{\beta}-\beta^*]\right)}{T^{1/2}}$$
$$= \frac{2x_j'\epsilon}{T^{1/2}} - \frac{2x_j'X}{T}\sqrt{T}[\tilde{\beta}-\beta^*]$$

Assuming β_j is a coefficient to the kth variable in the ith equation it follows from the same arguments as those preceding (16) that $\frac{x_j'\epsilon}{T^{1/2}} \to_d N(0,\Sigma_{ii}^2 C_{kk})$. Furthermore, $\frac{x_j'X}{T} \to_p (I_N \otimes C)_j$ where $(I_N\otimes C)_j$ is the ith row of $(I_N \otimes C)$. Hence, $\frac{x_j'\epsilon}{T^{1/2}}$ and $\frac{x_j'X}{T}$ are tight. The same is the case for $\sqrt{T}[\tilde{\beta}-\beta^*]$ since it converges weakly by (22)-(23). Taken together, $\frac{2x_j'\left(y-X\tilde{\beta}\right)}{T^{1/2}}$ is tight for all $j = 1, ..., N^2p$. Furthermore,

$$P\left(\max_{1\leq j\leq |G_i|}\left|\frac{2x_j'\left(y - X\tilde{\beta}\right)}{T^{1/2}}\right| > K\right) \leq |G_i| \max_{1\leq j\leq |G_i|} P\left(\left|\frac{2x_j'\left(y - X\tilde{\beta}\right)}{T^{1/2}}\right| > K\right)$$
$$\leq |G_i| \max_{1\leq j\leq |G_i|}\sup_{T\geq 1} P\left(\left|\frac{2x_j'\left(y - X\tilde{\beta}\right)}{T^{1/2}}\right| > K\right)$$

implies

$$\sup_{T\geq 1} P\left(\max_{1\leq j\leq |G_i|}\left|\frac{2x_j'\left(y - X\tilde{\beta}\right)}{T^{1/2}}\right| > K\right) \leq |G_i| \max_{1\leq j\leq |G_i|}\sup_{T\geq 1} P\left(\left|\frac{2x_j'\left(y - X\tilde{\beta}\right)}{T^{1/2}}\right| > K\right)$$

And so, for any $\delta > 0$ by choosing K sufficiently large it follows from the tightness of $\frac{2x_j'\left(y-X\tilde{\beta}\right)}{T^{1/2}}, j \in G_i$ that

$$\inf_{T\geq 1} P\left(\max_{1\leq j\leq |G_i|}\left|\frac{2x_j'\left(y - X\tilde{\beta}\right)}{T^{1/2}}\right| \leq K\right) \geq 1-\delta$$

Since the right hand side in (24) will be larger than K from a certain step and onwards it follows that $P(\tilde{\beta}_{G_i} = 0) \to 1$. □

References

Bańbura, M., D. Giannone, and L. Reichlin (2009): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Bernanke, B., J. Boivin, and P. Eliasz (2005): "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach," *The Quarterly Journal of Economics*, 120, 387–422.

Brockwell, P. and R. Davis (2009): *Time series: theory and methods*, Springer Verlag.

Caner, M. (2011): "Selecting the correct number of factors in approximate factor models: The large panel case with bridge estimators," Tech. rep., Mimeo. North Carolina State University, Raleigh, NC.

Caner, M. and K. Knight (2011): "An Alternative to Unit Root Tests: Bridge Estimators Differentiate between Nonstationary versus Stationary Models and Select Optimal Lag," Working paper, Michigan State University.

Fan, J. and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and J. Lv (2008): "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, 33, 1.

Hamilton, J. (1994): *Time series analysis*, vol. 2, Cambridge Univ Press.

Huang, J., J. Horowitz, and S. Ma (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *The Annals of Statistics*, 36, 587–613.

Knight, K. (1999): "Epi-convergence in distribution and stochastic equi-semicontinuity," *Unpublished manuscript.*

Knight, K. and W. Fu (2000): "Asymptotics for lasso-type estimators," *Annals of Statistics*, 1356–1378.

Kock, A. and L. Callot (2012): "Oracle Inequalities for High Dimensional Vector Autoregressions," *CREATES working paper 2012-05.*

Kock, A. and T. Teräsvirta (2012): "Forecasting Macroeconomic Variables using Neural Network Models and Three Automated Model Selection Techniques," *CREATES Research Papers.*

Kock, A. B. (2012): "Consistent and conservative model selection in stationary and non-stationary autoregressions," *Submitted.*

Leeb, H. and B. Pötscher (2005): "Model selection and inference: Facts and fiction," *Econometric Theory*, 21, 21–59.

——— (2008): "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, 142, 201–211.

Liao, Z. and P. Phillips (2012): "Automated Estimation of Vector Error Correction Models," .

Ludvigson, S. and S. Ng (2009): "Macro factors in bond risk premia," *Review of Financial Studies*, 22, 5027–5067.

Medeiros, M. and E. Mendes (2012): "Estimating High-Dimensional Time Series Models," Tech. rep.

Meier, L., S. Van De Geer, and P. Bühlmann (2008): "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 53–71.

Pötscher, B. and H. Leeb (2009): "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding," *Journal of Multivariate Analysis*, 100, 2065–2082.

Sims, C., J. Stock, and M. Watson (1990): "Inference in linear time series models with some unit roots," *Econometrica: Journal of the Econometric Society*, 113–144.

Song, S. and P. Bickel (2011): "Large vector auto regressions," *arXiv preprint arXiv:1106.3915*.

Stock, J. and M. Watson (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

Swanson, N. and H. White (1995): "A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks," *Journal of Business & Economic Statistics*, 265–275.

Teräsvirta, T. (1994): "Specification, estimation, and evaluation of smooth transition autoregressive models," *Journal of the American Statistical Association*, 208–218.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Wang, H. and C. Leng (2008): "A note on adaptive group lasso," *Computational Statistics & Data Analysis*, 52, 5277–5286.

Wang, H., G. Li, and C. L. Tsai (2007): "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 63–78.

Yuan, M. and Y. Lin (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhao, P. and B. Yu (2007): "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541.

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