SFB 649 Discussion Paper 2011-048

Large Vector Auto Regressions

Song Song *
Peter J. Bickel *

* University of California, USA

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

SFB 649 ECONOMIC RISK BERLIN

Large Vector Auto Regressions

Song Song[*], Peter J. Bickel [†]

July 31, 2011

Abstract

One popular approach for nonstructural economic and financial forecasting is to include a large number of economic and financial variables, which has been shown to lead to significant improvements for forecasting, for example, by the dynamic factor models. A challenging issue is to determine which variables and (their) lags are relevant, especially when there is a mixture of serial correlation (temporal dynamics), high dimensional (spatial) dependence structure and moderate sample size (relative to dimensionality and lags). To this end, an *integrated* solution that addresses these three challenges simultaneously is appealing. We study the large vector auto regressions here with three types of estimates. We treat each variable's own lags different from other variables' lags, distinguish various lags over time, and is able to select the variables and lags simultaneously. We first show the consequences of using Lasso type estimate directly for time series without considering the temporal dependence. In contrast, our proposed method can still produce an estimate as efficient as an *oracle* under such scenarios. The tuning parameters are chosen via a data driven "rolling scheme" method to optimize the forecasting performance. A macroeconomic and financial forecasting problem is considered to illustrate its superiority over existing estimators.

Keywords: Time Series, Vector Auto Regression, Regularization, Lasso, Group Lasso, Oracle estimator

JEL classification: C13, C14, C32, E30, E40, G10

1 Introduction

Macroeconomic forecasting is one of the central tasks in Economics. Broadly speaking, there are two approaches, structural and nonstructural forecasting. Structural forecasting, which aligns itself with

[*]University of California, Berkeley. Email: songsong@stat.berkeley.edu

[†]University of California, Berkeley. Email: bickel@stat.berkeley.edu

economic theory, and hence rises and falls with that, recedes following the decline of Keynesian theory. In recent years, new dynamic stochastic general equilibrium theory has been developed, and structural macroeconomic forecasting is poised for resurgence. Nonstructural forecasting, in contrast, attempts to exploit the reduced-form correlations in observed macroeconomic time series, has little reliance on economic theory, has always been working well and continues to be improved. Various univariate and multivariate time series analyzing techniques have been proposed, e.g. the auto regression (AR), moving average (MA), autoregressive moving average (ARMA), generalized autoregressive conditional heteroskedasticity (GARCH), vector auto regression (VAR) models among many others. A very challenging issue for this nonstructural approach is to determine which variables and (their) lags are relevant. If we omit some "important" variables by mistake, it potentially creates an omitted variable bias with adverse consequences for both structural analysis and forecasting. For example, Christiano et al. (1999) points out that the positive reaction of prices in response to a monetary tightening, the so-called *price puzzle*, is an artefact resulting from the omission of forward-looking variables, such as the commodity price index. Recently, Bańbura et al. (2010) shows that, when using the cross-sectional dimension related shrinkage, the forecasting performance of small monetary vector auto regression can be improved by adding additional macroeconomic variables and sectoral information. To illustrate this, we consider an example of interest rate forecasting. Nowadays people primarily use univariate or multivariate time series models, e.g. the Vasicek, CIR, Jump-Diffusion, Regime-Switching, and time-varying coefficients models, all of which are mostly based on the information from the interest rate time series itself. However, in practice, the central bank (Fed) bases their decisions of interest rate adjustment (as a monetary policy instrument) heavily on the national macroeconomic situation by taking many macro and financial measures into account. Bringing in this additional spatial (over the space of variables instead of from a geographic point of view; also used in future for convenience) information will therefore help improve its forecasting performance. Another example about the interactions between macroeconomics and finance comes from modeling credit defaults by also using macroeconomic information, since variation in aggregate default rates over time presumably reflects changes in general economic conditions also. Figlewski et al. (2006) find credit events are significantly affected by macroeconomic factors. Not only macroeconomics could affect finance, finance could also affect macroeconomics. For example, the economic crisis typically starts from the stock market crash. All of these call for an *integrated* analysis of macroeconomics and finance. Thus recently there has been a growing trend of using large panel macroeconomic and financial time series for forecasting, impulse response study and structural analysis, Forni et al. (2000), Stock and Watson (2002a), Stock and Watson (2002b), also seen at Forni et al. (2005), Stock and Watson (2005b), Giannone et al.

(2005), and Bańbura et al. (2010) for latest advancements.

Besides its presence in empirical macroeconomics, high dimensional data, where information often scatters through a large number of interrelated time series, is also attracting increasing attention in many other fields of economics and finance. In neuro-economics and behavioral finance, one uses high dimensional functional magnetic resonance imaging data (fMRI) to analyze the brain's response to certain risk related stimuli as well as identifying its activation area, Worsley et al. (2002) and Myšičková et al. (2011). In quantitative finance, one studies the dynamics of the implied volatility surface for risk management, calibration and pricing purposes, Fengler et al. (2007). Other examples and research fields for very large dimensional time series include mortality analysis, Lee and Carter (1992); bond portfolio risk management or derivative pricing, Nelson and Siegel (1987) and Diebold and Li (2006); international economy (many countries); industrial economy (many firms); quantitative finance (many assets) analysis among many others.

On the methodology side, if people still use either low dimensional (multivariate) time series techniques on a few subjectively (or from some background knowledge) selected variables or high dimensional "static" methods which are initially designed for independent data, they might either disregard potentially relevant information (temporal dynamics and spatial dependence) to produce suboptimal forecasts, or bring in additional risk. Examples include the already mentioned prize puzzle and interest rate forecasting problems. The more scattered and dynamic the information is, the severer this loss becomes. This modeling becomes more challenging under the situation that macroeconomic data we typically deal with has only low frequencies, e.g. monthly or yearly. For example, the popularly used dataset introduced by Stock and Watson (2005a) contains 131 monthly macro indicators covering a broad range of categories including income, industrial production, capacity, employment and unemployment, consumer prices, producer prices, wages, housing starts, inventories and orders, stock prices, interest rates for different maturities, exchange rates and money aggregates and so on. The time span is from January 1959 to December 2003 (so $T = 540$). In summary, we can see that the challenge of modeling high dimensional time series, especially the macroeconomic ones, comes from a mixture of serial correlation (temporal dynamics), high dimensional (spatial) dependence structure and moderate sample size (relative to dimensionality and lags). To this end, an *integrated* solution addressing these three challenges simultaneously is appealing.

To circumvent this problem, dynamic factor models have been considered to be quite successful recently in the analysis of large panels of time series data, Forni et al. (2000), Stock and Watson (2002a), Stock and Watson (2002b), also seen at Forni et al. (2005), Giannone et al. (2005), Park et al. (2009) and Song et al. (2010) (nonstationary case). They rely on the assumption that the bulk

of dynamics interrelations within a large dataset can be explained and represented by a few common factors (low dimensional time series). Less general models in the literature include *static* factor models proposed by Stock and Watson (2002a), Stock and Watson (2002b) and *exact* factor model suggested by Sargent and Sims (1977) and Geweke (1977).

Compared with the well studied dynamic factor models through the use of dynamic principal component analysis, the vector auto regressive (VAR) models have several natural advantages. For example, compared with the dynamic factor models' typical 2-step estimation procedure: dimension reduction first and low dimensional time series modeling, the VAR approach is able to model the high dimensional time series in one step, which may lead to greater efficiency. It also allows variable-to-variable relationship (impulse response) analysis and facilitates corresponding interpretation, which is not feasible in the factor modeling setup since the variables are "represented" by the corresponding factors. Historically, the VAR models are not appropriate for analyzing high dimensional time series because they involve the estimation of too many (J^2P, where J is the dimensionality and P is the number of lags) parameters. Thus they are primarily implemented on relatively low dimensional situations, e.g. the Baysian VARs (BVAR) by Doan et al. (1984) or still through the idea of factor modeling, e.g. the factor-augmented VAR (FAVAR) by Bernanke et al. (2005). However, based on recent advances in variable selection, shrinkage and regularization theory from Tibshirani (1996), Zou (2006) and Yuan and Lin (2006), large unrestricted vector auto regression becomes an alternative for the analysis of large dynamic systems. Therefore, the VAR framework can also be applied to empirical problems that require the analysis of more than a handful of time series. Mol et al. (2008) and Bańbura et al. (2010) proceed that from the Bayesian point of view. Chudik and Pesaran (2007) consider the case that $P = 1$ and both J and T are large through some "neighboring" procedure, which can be viewed as a special case of the "segmentized grouping" as we study here (details in Subsection 2.4). In the univariate case ($J = 1$), Wang et al. (2007) studies the regression coefficient and autoregressive order shrinkage and selection via the lasso when P is large and T is small (relative to P).

In this article, we will study the large vector auto regressions when $J, P \to \infty$ and T is moderate (relative to JP). Comparing to prior works in (large) vector auto regressions, the novelty of this article lies in the following perspectives. *First*, from the variable selection and regularization point of view, the theoretical properties of many existing methods have been established under the independent scenario, which is rarely met in practice and contradicts the original time series setup (if used directly). Disregarding the serial correlation in variable selection and regularization can be dangerous in the sense that various risk bounds in fact depend on the degree of time dependence, as we will

illustrate later. We propose a new methodology to address this serial correlation (time dependence) issue together with high dimensionality and moderate sample size, which enables us to obtain the consistency of variable selection even under the dependent scenario, i.e. to reveal the equilibrium among them. *Second*, our method is able to do variable selection and lag selection simultaneously. In previous literature, variable selection is usually carried out first, and then the corresponding estimate's performances w.r.t. different number of lags are compared through some information criteria to select the "optimal" number of lags. By doing so, we neglect the "interaction" between variable selection and lag selection. Additionally, when the number of the lag's candidates to be searched over is large, it is also computationally inefficient, due to the cost of the repeated variable selection procedures. *Third*, we differentiate the variable of interest's own lags (abbreviated as *own lags* afterwards) from the ones of other variables (abbreviated as *others' lags* afterwards). Their relative weights are also allowed to be varied when predicting different variables, while in other literature, they are assumed to stay the same. This is due to the fact that the dynamic of some variables is driven by itself, while for a different variable, it might be driven by the dynamics of others. When we include a vast number of macroeconomic and financial time series, assuming the same weight seems to be too restrictive. *Fourth*, our method is based on a more computationally efficient approach, which mostly uses the existing packages, e.g. the LARS (least angle regression) package, developed by Efron et al. (2004), while most other works in the literature go through the Bayesian approach that requires the choice of priors.

The rest of the article is organized as follows. In the next section, we present the main ingredients of the large vector autoregressive model (Large VAR) together with several corresponding estimation procedures and comparisons among them. The estimates' properties are presented in Section 3. In Section 4, the method is applied to the motivating macroeconomic forecasting problem, which shows that it outperforms some major existing method. Section 5 contains concluding remarks with discussions about relaxing some assumptions. All technical proofs are sketched in the appendix.

2 The Large VAR Model and Its Estimation

In this section, we introduce the model with three different estimates first, then discuss the data driven choice of hyperparameters to optimize the forecasting performance, provide a numerical algorithm, and finally summarize comparisons among these three estimates.

2.1 The Model

Assume that the high dimensional time series $\{Y_{tj}\}_{t=1,j=1}^{T\ J}$ is generated from

$$Y_t^\top = Y_{t-1}^\top B_1 + \ldots + Y_{t-P}^\top B_P + U_t^\top \tag{1}$$

$$\underbrace{\begin{pmatrix} Y_T^\top \\ Y_{T-1}^\top \\ \ldots \end{pmatrix}}_{T\times J} = \underbrace{\begin{pmatrix} Y_{T-1}^\top & Y_{T-2}^\top & \ldots & Y_{T-P}^\top \\ Y_{T-2}^\top & Y_{T-3}^\top & \ldots & Y_{T-1-P}^\top \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}}_{T\times JP} \underbrace{\begin{pmatrix} B_1 \\ B_2 \\ \ldots \end{pmatrix}}_{JP\times J} + \underbrace{\begin{pmatrix} U_T^\top \\ U_{T-1}^\top \\ \ldots \end{pmatrix}}_{T\times J}$$

$$Y = XB + U, \qquad \text{(compact form)} \tag{2}$$

where

- $Y = (Y_T^\top, \ldots, Y_1^\top)^\top$ with $Y_t^\top = (Y_{t1}, \ldots, Y_{tJ})$;
- $X = (X_T^\top, \ldots, X_1^\top)^\top$ (the lags of Y) with $X_t = (Y_{t-1}^\top, \ldots, Y_{t-P}^\top)^\top$;
- $B_1, \ldots, B_P$ are $J \times J$ autoregressive matrices, where P is the number of lags, $B = (B_1, \ldots, B_P)^\top$ is the $JP \times J$ matrix containing all coefficients $\{B_{pij}\}_{p=1,i=1,j=1}^{P\ \ J\ \ J}$, and $B_{\cdot\cdot j}, B_{p\cdot j}, B_{\cdot i\cdot}, B_{pi\cdot}$ is the jth column of B and B_p, ith row of B and B_p respectively;
- $U = (U_T^\top, \ldots, U_1^\top)^\top$, where U_t is a J-dimensional noise and independent of X_t.

All Y_t, X_t and U_t are assumed to have mean zero. The $J \times J$ covariance matrix of U_t, $\mathsf{Cov}(U_t)$, is assumed to be independent of t. Here we assume $\mathsf{Cov}(U_t)$ to be diagonal, say I_{J*J}. In our case, it is justified by the fact that the variables in the panel we will consider for estimation are standardized and demeaned. Similar assumption is also carried out in Mol et al. (2008). The relaxation allowing nonzero off-diagonal entries is discussed in Section 5.

We can see that given large J and P, we have to estimate a total of J^2P parameters, which is much larger than the moderate number of observations JT, i.e. $JP \gg T$. Consequently, ordinary least squares estimation is not feasible. Additionally, due to the structural change points in the macro and financial data (although not explored in this paper), the effective number of observations used for estimation could be much smaller than the original T. Thus we can see that on one hand, we do not want to impose any restrictions on the parameters and attain some general representations; on the other hand, it is known that making the model unnecessarily complex can degrade the efficiency of the resulting parameter estimate and yield less accurate predictions, as well as making interpretation and variable selection difficult. Hence, to avoid over fitting, *regularization* and *variable selection* are necessary. In the following, we are going to discuss the estimation procedure with different kinds of

regularization (illustrated in Figure 1). Before moving on, we incorporate the following very mild belief, as also considered in Bańbura et al. (2010): *the more recent lags should provide more reliable information than the more distant ones*, which tries to strike a balance between attaining model simplicity and keeping the historic information.

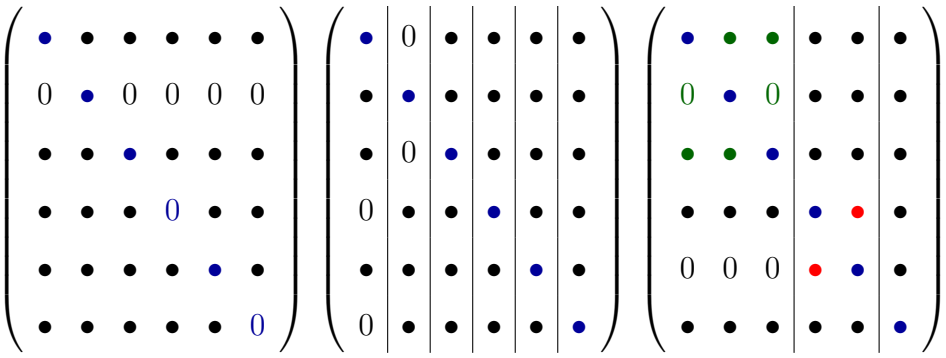

Figure 1: Illustration of three different types of estimates.

2.2 Universal Grouping

Without loss of generality, we start from considering one coefficient matrix, say B_p with entries $\{B_{pij}, 1 \leqslant i, j \leqslant J\}$. Inspired by Bańbura et al. (2010), we note the fact that the dynamic of some variable is driven by itself, while for a different variable, it might be driven by the dynamics of others. Consequently, we treat the variables' own lags (diagonal terms of B_p) different from others' lags (off-diagonal terms of B_p) and impose different regularizations for them. We assume that the off-diagonal coefficients of B_p are not only sparse, but also have the same sparsity pattern across different columns, which we call *group sparsity*. Thus we base our selection solution on group Lasso techniques (Yuan and Lin (2006)) for the off-diagonal terms and Lasso techniques (Tibshirani (1996)) for the diagonal terms here. We use B_{pj-j} to denote the vector composed of $\{B_{pji}\}_{i \neq j}$ and W_{-j} to denote the $(J-1) \times (J-1)$ diagonal matrix $\text{diag}[w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_J]$ where w_i is the positive real-valued weight associated with the ith variable for $1 \leqslant i \leqslant J$. It is included here primarily for practical implementation since if w_i is chosen as the $\text{Std}(Y_i)$, it is equivalent (subsection 2.6 for details) to standardize the predictors so that they all have zero mean and unit variance, Tibshirani (1996), which is also preferable for comparisons to prior works.

Specifically, given the above notations, we use the group Lasso type penalty $\sum_{j=1}^{J} \|B_{pj-j} W_{-j}\|_2$ and Lasso type penalty $\mu \sum_{j=1}^{J} w_j |B_{pjj}|$ to impose regularizations on other regressors' lags and predicted variables' own lags respectively and have the following penalty for the B_p matrix:

$$\sum_{j=1}^{J} \|B_{pj-j} W_{-j}\|_2 + \mu \sum_{j=1}^{J} w_j |B_{pjj}| \leqslant C p^{-\alpha}, \tag{3}$$

with some generic constant C.

- The hyperparameter μ controls the extent to which others' lags are less (more) "important" than the own lags. When μ is large, the penalty assigned to own lags is larger than to others' lags. As a result, it is more likely that the off-diagonal entries are shrunk to 0 instead of the diagonal ones, which corresponds to the case that the variable's dynamic is driven by itself, and vice versa when μ is small.

- The item $p^{-\alpha}$ reflects different regularization for different lags (over time). It becomes smaller when p gets larger. This is consistent with the previous belief: *the more recent lags should provide more reliable information than the more distant ones.* Thus as a result, large amounts of shrinkage are towards the more distant lags, whereas small amounts of shrinkage are towards the more recent ones. The hyperparameter α governs the relative importance of distant lags w.r.t. the more recent ones. Other decreasing functions of p, e.g. $f(p) = \log(p)^{-\alpha}, f(p) = \exp(p)^{-\alpha}$ could also be used. However, we do not consider a general representation (and use a data driven way to estimate $f(1), \ldots, f(p), \ldots, f(P)$ correspondingly) to avoid too many tuning parameters, especially when $P \to \infty$.

Since we have P coefficient matrices $B_1, \ldots, B_P$, summing (3) up over p (after multiplying p^α on both sides) yields $\sum_{p=1}^{P}\sum_{j=1}^{J} p^\alpha \|B_{pj-j}W_{-j}\|_2 + \mu\sum_{p=1}^{P}\sum_{j=1}^{J} p^\alpha w_j |B_{pjj}| \leqslant CP$. If we couple this to the quadratic loss $\{2J(T-P)\}^{-1}\sum_{t=P+1}^{T}\|Y_t^\top - X_t^\top B\|_2^2$ through Lagrange multipliers, we have equation (4):

$$\begin{aligned} & \min_B \{J(T-P)\}^{-1}\sum_{t=P+1}^{T}\|Y_t^\top - X_t^\top B\|_2^2 + \lambda\Big(\sum_{p=1}^{P}\sum_{j=1}^{J} p^\alpha \|B_{pj-j}W_{-j}\|_2 + \mu\sum_{p=1}^{P}\sum_{j=1}^{J} p^\alpha w_j|B_{pjj}|\Big) \\ \overset{\gamma=\lambda\mu}{=} & \min_B \{J(T-P)\}^{-1}\sum_{t=P+1}^{T}\|Y_t^\top - X_t^\top B\|_2^2 + \lambda\sum_{p=1}^{P}\sum_{j=1}^{J} p^\alpha \|B_{pj-j}W_{-j}\|_2 + \gamma\sum_{p=1}^{P}\sum_{j=1}^{J} p^\alpha w_j|B_{pjj}| \quad (4) \end{aligned}$$

with hyperparameters λ, γ and α. We call this estimate $\hat{B}$ the ***universal grouping*** estimate. As the number of variables J increases, the autocoefficients should be shrunk more in order to avoid over-fitting, as already discussed by Mol et al. (2008).

Using the group Lasso type regularization for the off-diagonal terms actually poses some strong assumptions on the underlying structure, which is not realistic from an economic point of view. **Remark 2.2.1** *First*, we just have one hyperparameter μ ($\mu = \gamma/\lambda$) to control the relative weights between own lags and others' lags. This means that the weights between own's lags and others' lags are the same across different dimensions which is hardly met in practice. Correspondingly, when we select the "optimal" μ to optimize the forecasting performance, we are actually optimizing the

averaged forecasting performance for all J variables instead of the variable of particular interest. This might produce suboptimal forecasts. Bańbura et al. (2010) considers a special case that own lags are always more "important" than others' lags, which might be less general than ours. **Remark 2.2.2** *Second*, using the L_2 norm $\|B_{pj-j}W_{-j}\|_2$ might shrink all off-diagonal terms in the same row $(\{B_{pj\cdot}\}_{\cdot\neq j})$ to zero simultaneously, which implicitly means that, for the jth corresponding variable, we assume it is either significant for all the other $J-1$ variables or not for any other $J-1$ variables at all. This is, again, too strong from an economic point of view.

2.3 No Grouping

To amend the deficiencies of the universal grouping estimate, we estimate the autocoefficient matrix B column by column instead of all at once. Without loss of generality, we consider the jth column $B_{\cdot\cdot j}$ here. Since $B_{\cdot\cdot j}$ is a vector, we can use the Lasso type penalties for both own lags and others' lags. By following similar ideas and abbreviations in subsection 2.2, we have equation (5) to get $\hat{B}_{\cdot\cdot j}$:

$$\begin{aligned} & \min_{B_{\cdot\cdot j}}(T-P)^{-1}\sum_{t=P+1}^{T}(Y_{tj}-X_t^\top B_{\cdot\cdot j})^2+\lambda_j\Big(\sum_{p=1}^{P}\sum_{i\neq j}p^\alpha w_i|B_{pij}|+u_j\sum_{p=1}^{P}p^\alpha w_j|B_{pjj}|\Big) \\ \overset{\gamma_j=\lambda_j\mu_j}{=} & \min_{B_{\cdot\cdot j}}(T-P)^{-1}\sum_{t=P+1}^{T}(Y_{tj}-X_t^\top B_{\cdot\cdot j})^2+\lambda_j\sum_{p=1}^{P}\sum_{i\neq j}p^\alpha w_i|B_{pij}|+\gamma_j\sum_{p=1}^{P}p^\alpha w_j|B_{pjj}| \end{aligned} \tag{5}$$

with hyperparameters λ_j, γ_j and α. The subindex j is added to λ_j and γ_j to emphasize that they could vary when estimating different $B_{\cdot\cdot j}$'s, $1\leqslant j\leqslant J$. We call this estimate $\hat{B}=(\hat{B}_{\cdot\cdot 1},\ldots,\hat{B}_{\cdot\cdot J})$ the ***no grouping*** estimate.

Remark 2.3.1 Because of different μ_j's ($\mu_j=\gamma_j/\lambda_j$) for different columns' estimates $\hat{B}_{\cdot\cdot j}$, we allow individualized weights between own lags and others' lags and could tune λ_j's and γ_j's to produce optimal forecasting performance for each variable of interest, say the jth. **Remark 2.3.2** Also for the same reason, we could get rid of the disadvantage that all off-diagonal terms in one row might be shrunk to 0 simultaneously.

For simplicity of notation, we drop the common subindex j and write $Y_{tj}=y_t$, $B_{\cdot\cdot j}=\beta$, $B_{pij}=c_i, i\neq j$, $B_{pjj}=d_p$, $\lambda_j=\lambda$, $\gamma_j=\gamma$, and (5) becomes:

$$\begin{aligned} \min_\beta Q_T(\beta) &= \min_\beta(T-P)^{-1}\sum_{t=P+1}^{T}(y_t-X_t^\top\beta)^2+\lambda\sum_{p=1}^{P}\sum_{i\neq j}p^\alpha w_i|c_i|+\gamma\sum_{p=1}^{P}p^\alpha w_p|d_p| \\ &= \min_\beta(T-P)^{-1}\sum_{t=P+1}^{T}(y_t-X_t^\top\beta)^2+\lambda_i\sum_{i=1}^{P(J-1)}|c_i|+\gamma_p\sum_{p=1}^{P}|d_p| \end{aligned} \tag{6}$$

with $\lambda_i=\lambda p^\alpha w_i$ and $\gamma_p=\gamma p^\alpha w_p$.

2.4 Segmentized Grouping

Since the large panel of macroeconomic and financial data sets usually have some natural "segment" structure, e.g. multiple interest rate time series w.r.t. different maturities, different number of employees w.r.t. different industrial sectors, different price indices w.r.t. different goods etc, if we take this information into account instead of estimating B either all at once or column by column, we could also do it segment by segment. Without loss of generality, we consider the ith segment $B_{\cdot\cdot\mathcal{N}_i}$. $\mathcal{N}_i$ is the index set for the ith segment and $N_i = |\mathcal{N}_i|$ denotes the cardinality of the set $\mathcal{N}_i$. $Y_{t\mathcal{N}_i}^\top$ is the corresponding part of $Y_t^\top$. We also use $W_{\mathcal{N}_i}$ to denote the $N_i \times N_i$ diagonal matrix with diagonal entries $\{w_i\}_{i\in\mathcal{N}_i}$ and $W_{\mathcal{N}_i - j}$ to denote the $(N_i - 1) \times (N_i - 1)$ diagonal matrix with diagonal entries $\{w_i\}_{i\in\mathcal{N}_i, i\neq j}$.

Under this situation, we have: own lags, others' (in the same segment) lags and others' (outside the segment) lags for the estimation of the ith segment's corresponding autoregressive coefficients $B_{\cdot\cdot\mathcal{N}_i}$. Also following similar ideas and abbreviations in subsection 2.2, we have the following estimation equation:

$$\begin{aligned}
&\min_{B_{\cdot\cdot N_i}} \{N_i(T-P)\}^{-1} \textstyle\sum_{t=P+1}^{T} \|Y_{t\mathcal{N}_i}^\top - X_t^\top B_{\cdot\cdot\mathcal{N}_i}\|_2^2 \\
+\lambda_{\mathcal{N}_i}\Big(&\textstyle\sum_{p=1}^{P}\sum_{j\notin N_i} p^\alpha \|B_{pj\cdot} W_{\mathcal{N}_i}\|_2 + \mu_{1\mathcal{N}_i}\sum_{p=1}^{P} p^\alpha w_j |B_{pjj}| + \mu_{2\mathcal{N}_i}\sum_{p=1}^{P}\sum_{j\in N_i} p^\alpha \|B_{pj-j} W_{\mathcal{N}_i - j}\|_2\Big) \\
&\min_{B_{\cdot\cdot N_i}} \{N_i(T-P)\}^{-1} \textstyle\sum_{t=P+1}^{T} \|Y_{t\mathcal{N}_i}^\top - X_t^\top B_{\cdot\cdot\mathcal{N}_i}\|_2^2 \\
+\lambda_{\mathcal{N}_i} &\textstyle\sum_{p=1}^{P}\sum_{j\notin\mathcal{N}_i} p^\alpha \|B_{pj\cdot} W_{\mathcal{N}_i}\|_2 + \gamma_{\mathcal{N}_i}\sum_{p=1}^{P} p^\alpha w_j |B_{pjj}| + \eta_{\mathcal{N}_i}\sum_{p=1}^{P}\sum_{j\in N_i} p^\alpha \|B_{pj-j} W_{\mathcal{N}_i - j}\|_2 \quad (7)
\end{aligned}$$

with hyperparameters $\lambda_{\mathcal{N}_i}, \gamma_{\mathcal{N}_i}, \eta_{\mathcal{N}_i}, \alpha$, $\gamma_{\mathcal{N}_i} = \lambda_{\mathcal{N}_i}\mu_{1\mathcal{N}_i}$ and $\eta_{\mathcal{N}_i} = \lambda_{\mathcal{N}_i}\mu_{2\mathcal{N}_i}$, $i = 1, \ldots, I$ where I is the overall number of segments. We call this estimate $\hat{B} = (\hat{B}_{\cdot\cdot\mathcal{N}_1}, \ldots, \hat{B}_{\cdot\cdot\mathcal{N}_I})$ the ***segmentized grouping*** estimate.

Chudik and Pesaran (2007) consider the case $P = 1$ and T is large (relative to J) through some "neighboring" procedure, which can be viewed as a special case of the "segmentized grouping" we studied here.

2.5 Forecast Evaluation and Choice of Parameters

The three penalization methods discussed above critically depend on penalty parameter selection for their performance in model selection, parameter estimation and prediction accuracy. Here we have hyper-parameters λ, γ (universal grouping), $\lambda_j, \gamma_j, 1 \leqslant j \leqslant J$ (universal grouping), $\lambda_i, \gamma_i, \eta_i, 1 \leqslant i \leqslant I$ and α, and choose them via a data driven "rolling scheme". To simulate real-time forecasting, we conduct an out-of-sample experiment. Let T_0 and T_1 denote the beginning and the end of the

Figure 2: Illustration of the Rolling Scheme

evaluation sample respectively. The point estimate of the jth variable's forecast is denoted by $\widehat{y}_{j,t|\sigma(t)}^{(\lambda,\gamma,\alpha)}$ based on $\sigma(t)$, the information up to time t. The point estimate of the one-step-ahead forecast is computed as in equation (5), and the h-step-ahead forecasts are computed in similar spirit. Out-of-sample forecast accuracy is measured in terms of mean squared forecast error (MSFE):

$$MSFE_{j,h}^{(\lambda,\gamma,\alpha)} = \frac{1}{T_1 - T_0 - h + 1} \sum_{t=T_0}^{T_1-h} (\widehat{y}_{j,t+h|\sigma(t)}^{(\lambda,\gamma,\alpha)} - y_{j,t+h|\sigma(t)})^2.$$

We report results for MSFE relative to the benchmark (random walk with drift) model's (abbreviated as $MSFE_{j,h}^{(0)}$), as also considered by Bańbura et al. (2010), i.e.

$$RMSFE_{j,h}^{(\lambda,\gamma,\alpha)} = \frac{MSFE_{j,h}^{(\lambda,\gamma,\alpha)}}{MSFE_{j,h}^{(0)}}.$$

The parameters are estimated using the observations from the most recent 10 years (rolling scheme) as illustrated in Figure 2. The parameters are set to yield a desired fit for the variable(s) of interest from T_0 to T_1. In other words, to obtain the desired magnitude of fit, the search is performed over a grid of λ, γ and α to minimize $\sum_{j=1}^{J} RMSFE_{j,h}^{(\lambda,\gamma,\alpha)}$ (universal grouping); λ_j, γ_j and α to minimize $RMSFE_{j,h}^{(\lambda,\gamma,\alpha)}$ (no grouping); $\lambda_i, \gamma_i, \eta_i$ and α to minimize $\sum_{j\in\mathcal{N}_i} RMSFE_{j,h}^{(\lambda,\gamma,\eta,\alpha)}$ (segmentized grouping) respectively. Due to computational cost, we prefix α to be 1 or 2 first, and then do the search of λ's and γ's over loose grids. For the nice performing λ's and γ's, we search over denser grids around them afterwards. The *parfor* command in Matlab is used to facilitate parallel computations to fasten this process. Also, using the least angle regression package provided at www-stat.stanford.edu/∼tibs/glmnet-matlab, makes the computation time together with the parameter selection very moderate in our experience.

2.6 Algorithm

Motivated by the adaptive lasso procedure, Zou (2006), if we define

- $\mathcal{P} = \text{diag}[1^\alpha, 2^\alpha, \ldots, P^\alpha] \otimes I_{J\times J}$, where $\text{diag}[1^\alpha, 2^\alpha, \ldots, P^\alpha]$ is the diagonal matrix with diagonal entries $\{1^{-\alpha}, 2^{-\alpha}, \ldots, P^{-\alpha}\}$, $\otimes$ is the Kronecker product and $I_{J\times J}$ is the $J \times J$ identity matrix;
- $\mathcal{W} = I_{P\times P} \otimes \text{diag}[w_1, w_2, \ldots, w_J]$;
- $\tilde{X}^\top = X^\top \mathcal{W}^{-1}\mathcal{P}^{-1}$ and $\tilde{B} = \mathcal{P}\mathcal{W}B$

and note the fact that $X^\top B$ in (2) is the same as $X^\top \mathcal{W}^{-1}\mathcal{P}^{-1}\mathcal{P}\mathcal{W}B = \tilde{X}^\top \tilde{B}$, we have the following estimation procedure (the proof is very simple and hence is omitted):

(1) Generate $\tilde{X}^\top = X^\top \mathcal{W}^{-1}\mathcal{P}^{-1}$;

(2) Corresponding to the three different estimates (4), (5) and (7), solve:

$$\min_{\tilde{B}} \{J(T-P)\}^{-1} \sum_{t=P+1}^{T} \|Y_t^\top - \tilde{X}_t^\top \tilde{B}\|_2^2 + \lambda \sum_{p=1}^{P}\sum_{j=1}^{J} \|\tilde{B}_{pj-j}\|_2 + \gamma \sum_{p=1}^{P}\sum_{j=1}^{J} |\tilde{B}_{pjj}|, \tag{8}$$

$$\min_{\tilde{B}_{\cdot\cdot j}} (T-P)^{-1} \sum_{t=P+1}^{T} (Y_{tj} - \tilde{X}_t^\top \tilde{B}_{\cdot\cdot j})^2 + \lambda_j \sum_{p=1}^{P}\sum_{i\neq j} |\tilde{B}_{pij}| + \gamma_j \sum_{p=1}^{P} |\tilde{B}_{pjj}|, \tag{9}$$

$$\min_{B_{\cdot\cdot\mathcal{N}_i}} \{N_i(T-P)\}^{-1} \sum_{t=P+1}^{T} \|Y_{t\mathcal{N}_i}^\top - X_t^\top \tilde{B}_{\cdot\cdot\mathcal{N}_i}\|_2^2 + \lambda_{\mathcal{N}_i} \sum_{p=1}^{P}\sum_{j\notin N_i} \|\tilde{B}_{pj\cdot}\|_2 + \gamma_{\mathcal{N}_i} \sum_{p=1}^{P} |\tilde{B}_{pjj}|$$
$$+ \eta_{\mathcal{N}_i} \sum_{p=1}^{P}\sum_{j\in N_i} \|\tilde{B}_{pj-j}\|_2;; \tag{10}$$

(3) Output $\hat{B} = \mathcal{W}^{-1}\mathcal{P}^{-1}\hat{\tilde{B}}$ with $\hat{\tilde{B}}$ minimizing (8) (universal grouping), $\hat{\tilde{B}} = (\hat{\tilde{B}}_{\cdot\cdot 1}, \ldots, \hat{\tilde{B}}_{\cdot\cdot J})$, $\hat{\tilde{B}}_{\cdot\cdot j}$ minimizing (9) (no grouping); $\hat{\tilde{B}} = (\hat{\tilde{B}}_{\cdot\cdot\mathcal{N}_1}, \ldots, \hat{\tilde{B}}_{\cdot\cdot\mathcal{N}_I})$, $\hat{\tilde{B}}_{\cdot\cdot\mathcal{N}_i}$ minimizing (10) (segmentized grouping).

At Step (2), motivated by Wang et al. (2007), as we have more than one penalty terms (mixed Lasso and group Lasso), we could iterate between penalties to solve it as the standard (group) Lasso problem.

For the "no grouping" estimate, by noting that $\gamma_j = \mu_j\lambda_j$ and $\gamma_j \sum_{p=1}^{P} |\tilde{B}_{pjj}| = \lambda_j \sum_{p=1}^{P} |\mu_j \tilde{B}_{pjj}|$, the estimation procedure above is equivalent to:

(1) Generate $\tilde{X}^\top = X^\top \mathcal{W}'^{-1}\mathcal{P}^{-1}$ with $\mathcal{W}' = I_{P\times P} \otimes \text{diag}[w_1, w_2, w_{j-1}, u_j w_j, w_{j+1}, \ldots, w_J]$ for estimating $B_{\cdot\cdot j}, 1 \leqslant j \leqslant J$;

(2) Corresponding to (9), solve:

$$\min_{\tilde{B}_{\cdot\cdot j}} (T-P)^{-1} \sum_{t=P+1}^{T} (Y_{tj} - \tilde{X}_t^\top \tilde{B}_{\cdot\cdot j})^2 + \lambda_j \sum_{p=1}^{P}\sum_{i=1}^{J} |\tilde{B}_{pij}|; \tag{11}$$

(3) Output $\hat{B} = \mathcal{W}'^{-1}\mathcal{P}^{-1}\hat{\tilde{B}}$ with $\hat{\tilde{B}} = (\hat{\tilde{B}}_{\cdot\cdot 1}, \ldots, \hat{\tilde{B}}_{\cdot\cdot J})$, $\hat{\tilde{B}}_{\cdot\cdot j}$ minimizing (9) (no grouping).

At Step (2), we could avoid iterating between multiple penalties and just solve it as the standard Lasso, e.g. by using the least angle regression package provided at www-stat.stanford.edu/∼tibs/glmnet-matlab.

In "large J, small T" paradigms, to get parsimonious models, shrinkage with penalization in model selection can shrink insignificant regression coefficients towards zero exactly, but at the same time, significant coefficients are shrunk as well though they are retained in selected working models, Wainwright (2009) and Huang et al. (2008). To this end, we only use our method for the variable (and lag) selection, but not for estimation, Chernozhukov et al. (2011). Thus, we implement the ordinary least squares estimation for the selected variables (and lags) from (8), (9) and (10) w.r.t. three different estimates.

2.7 Comparison

Now it is a matter of what kind of regularization techniques among these three choices to use in practice. *First*, as already discussed in Remark 2.2.1 and 2.3.1, from the allowing individualized (for the variable of particular interest) weights between own lags and others' lags and individualized forecasting performance optimization point of view, the "no grouping" approach is the best, the "universal grouping" one is the worst, and the "segmentized grouping" one is in between. *Second*, as in Remark 2.2.2 and 2.3.2, from whether all off-diagonal autocoefficients in one row are shrunk to zero point of view, the "no grouping" one is still favored. *Third*, as in subsection 2.5, the tuning parameters w.r.t. the universal grouping, no grouping or segmentized grouping estimates are selected to optimize the averaged forecasting performance for all variables, the specific variable's forecasting performance or the averaged forecasting performance for the variables in the same segment respectively. When different variables' time series have very distinct patterns, this individualized optimization is preferred. *Fourth*, for the estimation of large coefficient matrices, due to the strong *group-sparse* assumption on the underlying structure as mentioned in subsection 2.2, the group lasso type estimator actually has a sharper theoretical risk bound, Huang and Zhang (2009) for more details. In particular, they show that group Lasso is more robust to noise due to the stability associated with group structure and thus requires a smaller sample size to satisfy the sparse eigenvalue condition required in modern sparsity analysis. And the universal grouping estimate is also more computationally efficient since the whole autocoefficient matrix is estimated at once. However, note that the statistical error is a combination of modeling error and estimation error. Even though the group Lasso type estimate might have smaller

estimation error, due to the strong assumption to the underlying structure, the overall risk might not be smaller, as we discussed in subsection 2.2. Moreover, the typical macroeconomic data has low frequency, i.e. monthly. Thus the computational cost is not a severe problem since we only need to update the model once per month at most. Due to all these, we suggest the no grouping estimate for practical implementation as a compromise between flexibility and realization of assumptions. For this reason and technical simplicities, we mainly study the theoretical properties of the no grouping estimate as defined in (6) afterwards.

3 Estimates' Properties

In this section, we first show that, under the time series setup, if we just use the classic Lasso estimator, the risk bound will depend on the time dependence level as in Theorem 3.1. To circumvent this problem, through reweighting over time, our estimate in (6) can still produce an estimator, which is shown in Theorem 3.2 and 3.3, to be equivalent to an appropriate oracle. The techniques of the proofs are closely built upon those in Lounici et al. (2009), Bickel et al. (2009), Lounici (2008) and Wang et al. (2007).

3.1 Dependence Matters?

Now we will illustrate how the temporal dependence level affects the risk bounds of the Lasso type estimator. For technical simplicities, we consider the univariate AR(P) (or MA(P)) model with $P \to \infty$, i.e. $J = 1$ for equations (1) and (2):

$$e_t = x_{t1}\theta_1 + \ldots, x_{tP}\theta_P + \epsilon_t = x_t^\top \theta + \epsilon_t, \tag{12}$$

with the regressors $(x_{t1}, \ldots, x_{tP}) = x_t^T$, the coefficients $(\theta_1, \ldots, \theta_P) = \theta^\top$ and the error term ϵ_t. We also define $\mathbf{x}$ as a $T \times P$ matrix with the t, pth entry as x_{tp} and $e = (e_1, \ldots, e_T)^\top$. $x_{tp} = e_{t-p}$ (or ϵ_{t-p}) corresponds to the AR(P) (or MA(P)) model. In this situation, since there are no "others lags" ($J = 1$) and θ is a vector, the standard Lasso estimator $\hat{\theta}$ is defined through:

$$\min_\theta (T-P)^{-1} \sum_{t=P+1}^{T} (e_t - x_t^\top \theta)^2 + 2\lambda \|\theta\|_1. \tag{13}$$

We assume there is a *true* coefficient θ^* for (12) and define $M(\theta^*) = \sum_{p=1}^{p} \mathbf{1}(\theta_p^* \neq 0)$ and $M(\hat{\theta}) = \sum_{p=1}^{p} \mathbf{1}(\hat{\theta}_p \neq 0)$. Before moving on, we recall the fractional cover theory based definition first, which was introduced by Janson (2004) and can be viewed as a generalization of m-dependency. Given a set $\mathcal{T}$ and random variables V_t, $t \in \mathcal{T}$, we say:

- A subset $\mathcal{T}'$ of $\mathcal{T}$ is *independent* if the corresponding random variables $\{V_t\}_{t\in\mathcal{T}'}$ are independent.
- A family $\{\mathcal{T}_j\}_j$ of subsets of $\mathcal{T}$ is a *cover* of $\mathcal{T}$ if $\bigcup_j \mathcal{T}_j = \mathcal{T}$.
- A family $\{(\mathcal{T}_j, \mathbf{w_j})\}_\mathbf{j}$ of pairs $(\mathcal{T}_j, \mathbf{w_j})$, where $\mathcal{T}_j \subseteq \mathcal{T}$ and $\mathbf{w_j} \in [\mathbf{0}, \mathbf{1}]$ is a *fractional cover* of $\mathcal{T}$ if $\sum_j \mathbf{w_j} \mathbf{1}_{\mathcal{T}_\mathbf{j}} \geqslant \mathbf{1}_\mathcal{T}$, i.e. $\sum_{j:t\in\mathcal{T}_j} \mathbf{w_j} \geqslant \mathbf{1}$ for each $t \in \mathcal{T}$.
- A (fractional) cover is *proper* if each set $\mathcal{T}_j$ in it is independent.
- $\mathcal{X}(\mathcal{T})$ is the size of the smallest proper cover of $\mathcal{T}$, i.e. the smallest m such that $\mathcal{T}$ is the union of m independent subsets.
- $\mathcal{X}^*(\mathcal{T})$ is the minimum of $\sum_j \mathbf{w_j}$ over all proper fractional covers $\{(\mathcal{T}_j, \mathbf{w_j})\}_\mathbf{j}$.

Notice that, in spirit of these notations, $\mathcal{X}(\mathcal{T})$ and $\mathcal{X}^*(\mathcal{T})$ depend not only on $\mathcal{T}$ but also on the family $\{V_t\}_{t\in\mathcal{T}}$. Further note that $\mathcal{X}^*(\mathcal{T}) \geqslant 1$ (unless $\mathcal{T} = \varnothing$) and that $\mathcal{X}^*(\mathcal{T}) = 1$ if and only if the variables $V_t, t \in \mathcal{T}$ are independent, i.e. $\mathcal{X}^*(\mathcal{T})$ is a measure of the dependence structure of $\{V_t\}_{t\in\mathcal{T}}$. For example, if V_t only depends on $V_{t-1}, \ldots, V_{t-k}$ but is independent of all $\{V_s\}_{s<t-k}$, we will have $k+1$ independent sets:

$$\begin{aligned}
\mathcal{T}_1 &= \{V_1, V_{(k+1)+1}, V_{2(k+1)+1}, \ldots\},\\
\mathcal{T}_2 &= \{V_2, V_{(k+1)+2}, V_{2(k+1)+2}, \ldots\},\\
&\cdots\\
\mathcal{T}_{k+1} &= \{V_{k+1}, V_{(k+1)+(k+1)}, V_{2(k+1)+(k+1)}, \ldots\},
\end{aligned}$$

s.t. $\bigcup_{j=1}^{k+1} \mathcal{T}_j = \mathcal{T}$. So $\mathcal{X}^*(\mathcal{T}) = k+1$ (if $k+1 < T$).

Before stating the first main result of this section, we make the following two assumptions.

ASSUMPTION 3.1 *With a high probability q, $\forall$ p, the random variables x_{tp} and ϵ_t satisfy*

$$|\epsilon_t x_{tp}| \leqslant b_t \text{ and } \quad T^{-1}\sum_{t=1}^{T} b_t^2 \leqslant C'$$

for some constants $b_t, C' > 0, t = 1, \ldots, T$.

ASSUMPTION 3.2 *There exists a positive number $\kappa = \kappa(s)$ such that*

$$\min\left\{\frac{|x^\top\Delta|_2}{\sqrt{T}|\Delta_\mathcal{R}|_2} : |\mathcal{R}| \leqslant s, \Delta \in \mathbb{R}^P\backslash\{0\}, \| \Delta_{\mathcal{R}^c} \|_1 \leqslant 3 \| \Delta_\mathcal{R} \|_1 \right\} \geqslant \kappa,$$

where $\mathcal{R}^c$ denotes the complement of the set of indices $\mathcal{R}$, $\Delta_\mathcal{R}$ denotes the vector formed by the coordinates of the vector Δ w.r.t. the index set $\mathcal{R}$.

Assumption 3.2 is the *restricted eigenvalue* assumption from Bickel et al. (2009), which is essentially a restriction on the eigenvalues of the Gram matrix $\Psi_T = x^\top x/T$ as a function of sparsity s. To see this, recall the definitions of *restricted eigenvalues* and *restricted correlations* in Bickel et al. (2009):

$$\psi_{\min}(u) = \min_{z\in\mathbb{R}^P:1\leqslant\mathcal{M}(z)\leqslant u} \frac{z^\top \Psi_T z}{|z|_2^2}, \quad 1 \leqslant z \leqslant P,$$

$$\psi_{\max}(u) = \max_{z\in\mathbb{R}^P:1\leqslant\mathcal{M}(z)\leqslant u} \frac{z^\top \Psi_T z}{|z|_2^2}, \quad 1 \leqslant z \leqslant P,$$

$$\psi_{m_1,m_2} = \max\Big\{\frac{f_1^\top x_{I_1}^\top x_{I_2} f_2}{T|f_1|_2|f_2|_2} : I_1 \bigcap I_2 = \varnothing, |I_i| \leqslant m_i, f_i \in \mathbb{R}^{I_i}\backslash\{0\}, i = 1,2\Big\},$$

where $|I_i|$ denotes the cardinality of I_i and x_{I_i} is the $T \times |I_i|$ submatrix of x obtained by removing from x the columns that do not correspond to the indices in I_i. Lemma 4.1 in Bickel et al. (2009) shows that if the *restricted eigenvalue* of the Gram matrix Ψ_T satisfies $\psi_{\min}(2s) > 3\psi_{s,2s}$ for some integer $1 \leqslant s \leqslant P/2$, Assumption 3.2 holds.

We can now state our first main result.

THEOREM 3.1 *Consider the model* (12) *for* $P \geqslant 3$, $T \geqslant 1$ *and random variables* $V_t = \epsilon_t x_{tp}, t \in \mathcal{T}$. *Let the random variables* x_{tp} *and* ϵ_t *satisfy Assumption 3.1 for any* p, *all diagonal elements of the matrix* $x^\top x/T$ *euqal to* 1, *and* $M(\theta^*) \leqslant s$. *Furthermore, let* κ *be defined as in Assumption 3.2, and* ϕ_{max} *be the maximum eigenvalue of the matrix* $X^\top X/T$. *Let* $\lambda = \sqrt{\mathcal{X}^*(\mathcal{T})(\log P)^{1+\delta'}C'/T}, \quad \delta' > 0$. *Then with probability at least* $q(1 - P^{-\delta'})$, *for any solution* $\hat{\theta}$ *of (13), we have:*

$$T^{-1}\| x(\hat{\theta} - \theta^*) \|^2 \leqslant 16s\mathcal{X}^*(\mathcal{T})(\log P)^{1+\delta'}C'/T\kappa^2, \tag{14}$$

$$\| \hat{\theta} - \theta^* \|_1 \leqslant 16 \leqslant 16s\sqrt{\mathcal{X}^*(\mathcal{T})(\log P)^{1+\delta'}C'/T}/\kappa^2, \tag{15}$$

$$M(\hat{\theta}) \leqslant 64\phi_{max}^2 s/\kappa^2. \tag{16}$$

Before explaining the results, we would like to discuss some related results first. Suppose x in (12) has full rank P and ϵ_t is $\mathrm{N}(0, \sigma^2)$. Consider the least squares estimate ($P \leqslant T$) $\hat{\theta}_{OLS} = (xx^\top)^{-1}xe$. Then from standard least squares theory, we know that the prediction error $\|x^T(\hat{\theta}_{OLS} - \theta^*)\|_2^2/\sigma^2$ is χ_p^2-distributed, i.e.

$$\mathsf{E}\,\frac{\|x^T(\hat{\theta}_{OLS} - \theta^*)\|_2^2}{T} = \frac{\sigma^2}{T}P. \tag{17}$$

In the sparse situation if ϵ_t is $\mathrm{N}(0, \sigma^2)$ (different from our case), Corollary 6.2 of Bühlmann and van de Geer (2011) shows that the Lasso estimate obeys the following *oracle inequality*:

$$\frac{\|x^T(\hat{\theta}_{Lasso} - \theta^*)\|_2^2}{T} \leqslant C_0\frac{\sigma^2 \log P}{T}M(\theta^*) \tag{18}$$

with a large probability and some constant C_0. The additional $\log P$ factor here could be seen as the price to pay for not knowing the set $\{\theta_p^*, \theta_p^* \neq 0\}$, Donoho and Johnstone (1994).

Similar to the i.i.d. Gaussian situation discussed above, the term $s(\log P)^{1+\delta'}$ in (20) could be interpreted as the price to pay for not knowing the set $\{\theta_p^*, \theta_p^* \neq 0\}$. Here we have $(\log P)^{1+\delta'}$ instead of $\log P$ because we deviate from the typical i.i.d. Gaussian situation and establish the result under the more general Assumption 3.1, which could be thought as the "finite second moment" condition. And the δ' term is the price to pay for this deviation.

For the case of $x_{tp} = \epsilon_{t-p}$ (MA(P) model), by the definition of $\mathcal{X}^*(\mathcal{T})$ and V_t, if $k^* = \max\{p, \text{s.t.}\theta_p^* \neq 0, \theta_{p+1}^* = \theta_{p+2}^* = \ldots = \theta_P^* = 0\})$, we have $\mathcal{X}^*(\mathcal{T}) = k^* + 1$ (if $k^* + 1 < T$). The RHS of (14) becomes $16s(k^*+1)(\log P)^{1+\delta'}/T\kappa^2$. For the case of $x_{tp} = e_{t-p}$ (AR(P) model), which is equivalent to MA(∞) and $\mathcal{X}^*(\mathcal{T}) \leqslant T$, the RHS of (14) becomes $16s(\log P)^{1+\delta'}/\kappa^2$. Thus $\mathcal{X}^*(\mathcal{T})$ could be interpreted as a measure on how many past lags V_t depends on. Additionally, when the time dependence level increases, by the definition of the Gram matrix, κ will decrease since it characterizes how strong $x_{t1}, \ldots, x_{tP}$ depend on each other. Still using the MA(k^*) example considered above, κ could be thought as a measure on how strong e_t depends on $e_{t-1}, \ldots, e_{t-k^*}$, which is a complement of the measure of $\mathcal{X}^*(\mathcal{T})$ on the time dependence level. In both cases (MA(P) or AR(P)), Theorem 3.1 states that if we use the standard Lasso estimate directly for the time series, the bounds get larger when the dependence level ($\mathcal{X}^*(\mathcal{T})$) increases and κ decreases. In other words, the bound is minimized when $\mathcal{X}^*(\mathcal{T}) = 1$, which corresponds to the independent situation in the literature. When $\mathcal{X}^*(\mathcal{T})$ reaches T, it will be offset by the T in the denominator. Thus the risk bound does not decrease when T increases. The intuition behind is clear: if the dependence level is strong, then the additional information brought by a "new" observation will be effectively less, i.e. the overall information from $\{V_t\}_{t=1}^T$ will be less correspondingly, which will result in increasing estimates' risk bounds. Consequently, we expect the selection not to be stable and to be very sensitive to minor perturbation of the data. In this sense, we do not expect variable selection to provide results that lead to clearer economic interpretation than principal components or Ridge regression.

3.2 Consistency of Selection

To study the oracle properties of the estimator in (6), we assume that there is a correct model with the regression and autoregression coefficients $\beta^* = (c^{*\top}, d^{*\top})\top = (c_1^*, \ldots, c_{P(J-1)}^*, d_1^*, \ldots, d_P^*)'$. Furthermore, we assume that there are a total of $p_0 \leqslant P(J-1)$ non-zero other-lag coefficients and $q_0 \leqslant P$ non-zero own-lag coefficients. For convenience, we define $S_1 = \{1 \leqslant i \leqslant P(J-1), c_i^* \neq 0\}$, $\hat{S}_1 = \{1 \leqslant i \leqslant P(J-1), \hat{c}_i \neq 0\}$, $S_2 = \{1 \leqslant p \leqslant P, d_p^* \neq 0\}$ and $\hat{S}_2 = \{1 \leqslant p \leqslant P, \hat{d}_p \neq 0\}$. Then, the sets S_1 and S_2 contain the indices of the significant others-lag and own-lag coefficients respectively, and their complements S_1^c and S_2^c contain the indices of the insignificant coefficients. Next, let

$c^*_{S_1}$ denote the $p_0 \times 1$ significant other-lag coefficient vector with $\hat{c}_{S_1}$ being its associated estimator. Moreover, other related parameters and their corresponding estimators are analogously defined (e.g. $c^*_{S_1^c}, \hat{c}_{S_1^c}, d^*_{S_2}, \hat{d}_{S_2}, d^*_{S_2^c}, \hat{d}_{S_2^c}$). Finally, let $\beta^*_1 = (c^{*\prime}_{S_1}, d^{*\prime}_{S_2})'$ and $\beta^*_2 = (c^{*\prime}_{S_1^c}, d^{*\prime}_{S_2^c})'$ with corresponding estimates $\hat{\beta}_1, \hat{\beta}_2$. To facilitate the study, we also introduce the notations

$$\begin{aligned} a_T &\overset{\text{def}}{=} \max(\lambda_i, \gamma_p, i \in S_1, p \in S_2), \\ b_T &\overset{\text{def}}{=} \min(\lambda_i, \gamma_p, i \in S_1^c, p \in S_2^c), \end{aligned}$$

where λ_i and γ_p are functions of T. To investigate the theoretical properties of $\hat{\beta}$, we introduce the following conditions:

A1 The sequence $\{X_t\}$ is independent of ε_t ($\varepsilon_t \overset{\text{def}}{=} U_{tj}$);

A2 All roots of polynomial $1 - \sum_{p=1}^{P} d^*_p z_p$ are outside the unit circle;

A3 ε_t has finite fourth-order moment, i.e. $\mathsf{E}(\varepsilon_t^4) < \infty$;

A4 $\forall j$, Y_{tj} (as components of the covariate X_t) is strictly stationary and ergodic with finte second-order moment (i.e. $\mathsf{E}\,\|Y_{tj}\|^2 < \infty$).

The technical conditions above are typically used to assure the $\sqrt{T}$-consistency and asymptotic normality of the unpenalized least squares estimator.

We also rewrite equation (6) as

$$\min_{\beta} Q_T(\beta) = \min_{\beta} \sum_{t=P+1}^{T} (y_t - X_t^\top \beta)^2 + T \sum_{i=1}^{P(J-1)} \lambda_i |c_i| + T \sum_{p=1}^{P} \gamma_p |d_p| \tag{19}$$

by multiplying $2(T-P)$ and writing $T-P$ as T without confusion. Define $\sum_{t=P+1}^{T}(y_t - X_t^\top \beta)^2$ as $L_T(\beta)$, $x_t^T = (X_{j,t}, X_{2j,t}, \ldots, X_{Jj,t})$, the own lags corresponding to the coefficients c, and $z_t^\top = X_t^\top \backslash x_t^\top$, others' lags corresponding to the coefficients d respectively. Then $X_t^\top \beta = x_t^\top c + z_t^\top d$.

We first investigate the consistency of the estimator of (6).

LEMMA 3.1 *Assume that $a_T = o(1)$ as $T \to \infty$. Then, under conditions (A1-A4), $\exists$ a local minimizer $\widehat{\beta}$ of (6) s.t.*

$$\|\widehat{\beta} - \beta^*\| = \mathcal{O}_p(T^{-1/2} + a_T).$$

The proof is given in the appendix. Lemma 3.1 implies that, if the tuning parameters associated with the significant regressors converge to 0 at a speed faster than $T^{-1/2}$, then there is a local minimizer of (6), which is $\sqrt{T}$-consistent. Next, we show that, if the tuning parameters associated with the insignificant regressors shrink to 0 slower than $T^{-1/2}$, then their coefficients can be estimated exactly as 0 with probability tending to 1.

THEOREM 3.2 (Consistency of Selection) *Assume that* $b_T\sqrt{T} \to \infty$ *and* $\|\widehat{\beta}-\beta^*\| = \mathcal{O}(T^{-1/2})$, *then*

$$\mathrm{P}(\widehat{\beta}_2 = 0) \to 1.$$

Theorem 3.2 shows that our method can produce a sparse solution for insignificant coefficients consistently. Furthermore, this theorem, together with Lemma 3.1, indicates that the $\sqrt{T}$-consistent estimator must satisfy $\mathrm{P}(\widehat{\beta}_2 = 0) \to 1$ when the tuning parameters fulfill the appropriate conditions. Finally, we obtain the asymptotic distribution of this estimator.

THEOREM 3.3 *Assume that* $a_T\sqrt{T} \to 0$ *and* $b_T\sqrt{T} \to \infty$. *Then, under conditions (A1-A4), the "nonzero" components* $\widehat{\beta}_1$ *of the local minimizer* $\widehat{\beta}$ *in Lemma 3.1 satisfies*

$$(\widehat{\beta}_1 - \beta_1^*)\sqrt{T} \xrightarrow{d} N(0, \Sigma_0^{-1}),$$

$$\mathrm{P}(\hat{S}_1 = S_1) \to 1, \quad \mathrm{P}(\hat{S}_2 = S_2) \to 1,$$

where Σ_0 *is the submatrix of* Σ *corresponding to* β_1^*, $\Sigma = \mathrm{diag}(B, C)$, $B = \mathsf{E}(x_t x_t^\top)$ *and* $C = \mathsf{E}(z_t z_t^\top)$.

Theorem 3.3 implies that, if the tuning parameters satisfy the conditions $a_T\sqrt{T} \to 0$ and $b_T\sqrt{T} \to \infty$, then, asymptotically, the resulting estimator can be as efficient as the oracle estimator. And our method can produce a sparse solution for significant coefficients consistently.

Since consistency of selection is established here, if we use the ordinary least squares estimation for the selected variables, we can avoid the log term on (18) and (14).

4 Application

We use the dataset of Stock and Watson (2005a) for illustration. This dataset contains 131 monthly macro indicators covering a broad range of categories including income, industrial production, capacity, employment and unemployment, consumer prices, producer prices, wages, housing starts, inventories and orders, stock prices, interest rates for different maturities, exchange rates, money aggregates and so on. The time span is from January 1959 to December 2003. We apply logarithms to most of the series except those already expressed in rates. The variables of special interest include a measure of real economic activity, a measure of prices and a monetary policy instrument. As in Christiano et al. (1999), we use employment as an indicator of real economic activity measured by the number of employees on non-farm payrolls (EMPL). The level of prices is measured by the consumer price index (CPI) and the monetary policy instrument is the Federal Funds Rate (FFR). All 131

		$P=1$	$P=4$	$P=7$	$P=13$	$P=25$	BVAR
	EMPL	0.3333	0.3336	0.3338	0.3341	0.3335	0.46
$h=1$	CPI	0.3623	0.3618	0.3613	0.3621	0.3623	0.50
	FFR	0.4279	0.4281	0.4281	0.4284	0.4287	0.75
	EMPL	0.5191	0.5188	0.5192	0.5191	0.5189	0.38
$h=3$	CPI	0.4990	0.4992	0.4986	0.4995	0.4996	0.40
	FFR	0.4615	0.4614	0.4619	0.4617	0.4628	0.94
	EMPL	0.4730	0.4730	0.4735	0.4729	0.4736	0.50
$h=6$	CPI	0.4880	0.4874	0.4884	0.4885	0.4891	0.40
	FFR	0.5237	0.5242	0.5243	0.5243	0.5250	1.29
	EMPL	0.4997	0.4991	0.4992	0.4997	0.5002	0.78
$h=12$	CPI	0.4689	0.4687	0.4689	0.4694	0.4686	0.44
	FFR	0.4201	0.4199	0.4201	0.4200	0.4216	1.93

Table 1: RMSFE w.r.t. different choices of h and P.

variables' lags are used as regressors. As discussed earlier, because of the stationary requirement of our method, the series are transformed to obtain stationarity so that many of the series are (2nd order) differences of the raw data series (or logarithm of the raw series).

We evaluate the forecast performance over the period from T_0 = January 70 to T_1 = December 03 and for forecast horizons up to one year ($h = 1, 3, 6, 12$). The order of the VAR is set to be $P = 1, 4, 7, 13, 25$. The resulting performance is summarized in Table 1 with comparisions to the ones of Bańbura et al. (2010) listed under the "BVAR" column. As we can see, unlike the information criteria based on lag selection techniques, the RMSFE is very robust to the initial choice of P, which primarily benefits from the "re-weighting over lags" technique ($p^{-\alpha}$) we used before. For this specific data set, $P = 1$ seems enough. But in general, since we never know the true value of lags, we can include a large enough P at the beginning to allow flexibility without worrying about over fitting. Moreover, for the one-step-ahead forecast, our method outperforms for EMPL, CPI and FFR, while when $h \geqslant 3$, it outperforms mainly for EMPL and FFR, especially for the latter one. This results from the fact that different time series might have quite different behaviors, so if we just have the "universal" penalty parameter for all of them as in Bańbura et al. (2010), the corresponding forecasting performance might not be optimized. For reference purpose, we also provide the factor-augmented vector autoregressive results of Bernanke et al. (2005) in Figure 3.

		FAVAR 1 factor			FAVAR 3 factors			*LARGE*
		$p = 13$	$p =$ BIC	BVAR	$p = 13$	$p =$ BIC	BVAR	BVAR
$h = 1$	EMPL	1.36	0.54	0.70	3.02	0.52	0.65	0.46
	CPI	1.10	0.57	0.65	2.39	0.52	0.58	0.50
	FFR	1.86	0.98	0.89	2.40	0.97	0.85	0.75
$h = 3$	EMPL	1.13	0.55	0.68	2.11	0.50	0.61	0.38
	CPI	0.80	0.49	0.55	1.44	0.44	0.49	0.40
	FFR	1.62	1.12	1.03	3.08	1.16	0.99	0.94
$h = 6$	EMPL	1.33	0.73	0.87	2.52	0.63	0.77	0.50
	CPI	0.74	0.52	0.55	1.18	0.46	0.50	0.40
	FFR	2.07	1.31	1.40	3.28	1.45	1.27	1.29
$h = 12$	EMPL	1.15	0.98	0.92	3.16	0.84	0.83	0.78
	CPI	0.95	0.58	0.70	1.98	0.54	0.64	0.44
	FFR	2.69	1.43	1.93	7.09	1.46	1.69	1.93

Figure 3: Results of Bernanke et al. (2005) and Bańbura et al. (2010)

5 Concluding Remarks and Discussions

To summarize, in this article, we first show that under the time series setup, if we still use the classic Lasso type estimator, the risk bound will increase when the time dependence level increases, however, our method could still achieve the consistency of variable selection under such scenario; second, our method is able to do variable selection and lag selection simultaneously, and is rather robust to the initial choice of lags; third, we allow individualized weights between own and others' lags. All these have been confirmed by the real forecasting performance in the previous section and come at a low computational cost.

Some issues we do not explore here include *nonstationarity*, *rank test, cointegration* and *causal test*. For a typical macroeconomic data set, the nonstationarity comes from seasonality, business cycle and economic developments. In spirit of Song et al. (2010), this motivates us to add a nonstationary component $U\Gamma$ to equation (2) as below

$$Y_t = Z_t\Gamma + X_tB + U_t,$$

where $Z_t = (Z_1(t), \ldots, Z_R(t))^\top$ contains R basis functions of time consisting of Fourier series with different frequencies and segment by segment ortho-normal polynomials with corresponding $R \times J$ coefficient matrix Γ, and X_t, B are the same as in equation (1). Studying this extended model deserves further investigation and will be presented in a separate paper. If we want to consider the rank test, cointegration and causal test, what we need for this high dimensional time series is not the ones in the univariate case, but the high dimensional *simultaneous* tests, which might be much more difficult.

Heteroscedasticity with Cross-section Correlations.

We consider $\mathsf{Cov}(U_t) = \Sigma$ with nonzero off-diagonal entries in Σ. Assume that we have a consistent estimate $\hat{\Sigma}$ for Σ (which is another challenging task since Σ is a $J \times J$ matrix) with Cholesky decomposition $\hat{\Sigma} = C^\top C$, where C is an upper triangular matrix with inverse D (which is also an

upper triangular matrix). Without loss of generality, assume all diagonal entries of $\hat{\Sigma}, C$ and D are equal to 1. Transform the original X_t by D to generate $\tilde{X}_t$ ($\tilde{X}_t = X_t D$) s.t. $\mathsf{Cov}(U_t D) = I$. Under this situation, we are no longer selecting the original variables, but linear transformations of them. Thus we must show that this does not affect the inference. We have

$$\begin{aligned}
&\tilde{\beta}_1\tilde{x}_{t1} + \tilde{\beta}_2\tilde{x}_{t2} + \ldots + \tilde{\beta}_J\tilde{x}_{tJ} \\
=&\tilde{\beta}_1 x_{t1} + \tilde{\beta}_2(d_{12}x_{t1} + x_{t2}) + \ldots + \tilde{\beta}_J(\sum_{j=1}^{J-1} d_{jJ}x_{tj} + x_{tJ}) \\
=&(\tilde{\beta}_1 + \sum_{j=2}^{J}\tilde{\beta}_j d_{1j})x_{t1} + (\tilde{\beta}_2 + \sum_{j=3}^{J}\tilde{\beta}_j d_{2j})x_{t2} + \ldots + \tilde{\beta}_J x_{tJ}.
\end{aligned}$$

If the off-diagonal entries of D, $\{d_{ij}\}_{i<j}$, are much smaller than the diagonal entries 1, it is likely that the selected nonzero sets of $\tilde{\beta}$'s (or $\tilde{c}, \tilde{d}$) are the same as the selected nonzero sets $\hat{S}_1$ and $\hat{S}_2$ of β's, which have been shown to be the same as the *oracle* ones, Theorem 3.2 and 3.3. By the definition of D, this means that the off-diagonal entries of $C, \hat{\Sigma}$ and Σ should also be much smaller than their diagonal entries 1, e.g. the cross-section correlations must be weak enough, which aligns with the case for dynamic factor models, Forni et al. (2000).

Acknowledgement The main results of this paper were first presented at the annual Winter Meeting of the Econometric Society, Denver, Jan, 2011. We are grateful for seminar participants' many interesting comments on several versions of the paper. In particular, I would like to thank Prof Peter Bickel for sponsoring my stay at the University of California, Berkeley. This work is also partially supported by Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin from May, 2011 to Oct, 2011, which is greatly acknowledged. We would also like to thank Prof Marta Bańbura, Prof Domenico Giannone and Prof Lucrezia Reichlin for sharing the codes of BVAR.

6 Appendix

Proof of Theorem 3.1 The proof of this theorem is based on the ones of Lemma 3.1 and Theorem 3.1 in Lounici et al. (2009) up to a modification of the bound on $\mathrm{P}(\mathcal{A}^c)$ with random event $\mathcal{A} = \left\{ \max_{1\leqslant p\leqslant P} \sum_{t=1}^{T} \varepsilon_t X_{tp} \leqslant \lambda T \right\}$, where n, M and T there are equivalent to T, P and 1 here respectively.

The intermediate results in the proof of Theorem 3.1 in Lounici et al. (2009) show that

$$T^{-1}\| X(\widehat{B} - B^*) \|^2 \leqslant 16s\lambda^2/\kappa^2, \tag{20}$$

$$\| \widehat{B} - B^* \| \leqslant 16 \leqslant 16s\lambda/\kappa^2, \tag{21}$$

$$M(\widehat{B}) \leqslant 64\phi_{max}^2 s/\kappa^2. \tag{22}$$

We have:

$$\mathrm{P}(\mathcal{A}^c) = \mathrm{P}\Big(\max_{1\leqslant p\leqslant P}\sum_{t=1}^{T}\varepsilon_t X_{tp} > \lambda T\Big) \leqslant P\,\mathrm{P}\Big(\sum_{t=1}^{T}\varepsilon_t X_{tp} > \lambda T\Big) \stackrel{\text{def}}{=} P\,\mathrm{P}\Big(\sum_{t=1}^{T} V_t > \lambda T\Big).$$

Then, by the (extended) Mcdiarmid inequality, see Theorem 2.1 of Janson (2004), with random vectors $\{V_t\}_{t=1}^T$, we have

$$\mathrm{P}(\mathcal{A}^c) \leqslant P\,\mathrm{P}\Big(\sum_{t=1}^{T} V_t > \lambda T\Big) \leqslant P\exp\left\{-\frac{\lambda^2 T}{\mathcal{X}^*(\mathcal{T})\sum_t b_t^2/T}\right\} \leqslant P^{-\delta'}$$

with $\lambda = \sqrt{\mathcal{X}^*(\mathcal{T})(\log P)^{1+\delta'}C'/T}$, $\quad \delta' > 0$, which, together with (20), (21) and (22), leads to (14), (15) and (16). □

Proof of Lemma 3.1 The proofs are closely built upon those of Wang et al. (2007). Let $\delta = (u^\top, v^\top)^\top$, $u = (u_1, \ldots, u_{P(J-1)})^\top$, $v = (v_1, \ldots, v_P)^\top$, $\alpha_T = T^{-1/2} + a_n$ and $\{\beta^* + \alpha_T\delta : \|\delta\| \leqslant e\}$ be the ball around β^*. Then, for $\|\delta\| = e$, we have

$$\begin{aligned}
D_T(\delta) &= Q_T(\beta^* + \alpha_T\delta) - Q_T(\beta^*) \\
&= L_T(\beta^* + \alpha_T\delta) - L_T(\beta^*) + T\sum_{i\in S_1}\lambda_i(|c_i^* + \alpha_T u_i| - \|c_i^*\|) + T\sum_{j\in S_2}\gamma_j(|d_j^* + \alpha_T v_j| - \|d_j^*\|) \\
&= L_T(\beta^* + \alpha_T\delta) - L_T(\beta^*) - T\alpha_T\sum_{i\in S_1}\lambda_i|u_i| - T\alpha_T\sum_{j\in S_2}\gamma_j|v_j| \\
&= L_T(\beta^* + \alpha_T\delta) - L_T(\beta^*) - T\alpha_T^2 p_0 e - T\alpha_T^2 q_0 e \\
&= L_T(\beta^* + \alpha_T\delta) - L_T(\beta^*) - T\alpha_T^2(p_0 + q_0)e.
\end{aligned} \tag{23}$$

Furthermore,

$$\begin{aligned}
L_T(\beta^* + \alpha_T\delta) - L_T(\beta^*) &= \sum_t\{u_t - a_T z_t^\top v - a_T u^\top x_t\}^2 - \sum_t u_t^2 \\
&= a_T^2\sum_t\{(z_t^\top v)^2 + u^\top x_t x_t^\top u\} && (24) \\
&\quad -2a_T\sum_t u_t z_t^\top v && (25) \\
&\quad +2a_T^2\sum_t z_t^\top v u^\top x_t. && (26)
\end{aligned}$$

By employing the martingale central limit theorem and the ergodic theorem, we can show that (24) $= Ta_T^2\{\delta^\top\Sigma\delta + o_p(1)\}$, (25) $= \delta^\top\mathcal{O}_p(Ta_T^2)$ and (26) $= Ta_T^2 o_p(1) = o_p(Ta_T^2)$. Because (24) dominates the terms (25), (26) and $T\alpha_T^2(p_0+q_0)e$ in equation (23), for any given $\epsilon > 0$, there is a large constant e such that

$$\mathrm{P}[\inf_{\|\delta\|=e} Q_T(\beta^* + \alpha_T\delta) > Q_T(\beta^*)] \geqslant 1-\epsilon.$$

This implies that, with probability at least $1-\epsilon$, there is a local minimizer in the ball $\{\beta^* + \alpha_T\delta : \|\delta\| \leqslant e\}$, Bickel et al. (1998) and Fan and Li (2001). Consequently, there is a local minimizer of $Q_T(\beta)$ such that $\|\hat{\beta} - \beta^*\| = \mathcal{O}_p(\alpha_T)$. This completes the proof. □

Proof of Theorem 3.2 The proof is essentially the same as those of Theorem 2 of Zou (2006) and Wang et al. (2007). For $i \in S_1^c$, assume that there is a local minimizer $\hat{\beta}$ with $\hat{c}_i \neq 0$. By the KKT optimality condition, we have

$$\begin{aligned} 0 &= \frac{\partial L_T(\hat{\beta})}{c_i} + T\lambda_i sgn(\hat{c}_i) \\ &= \frac{\partial L_T(\beta^*)}{c_i} + T\Sigma_i(\hat{\beta} - \beta^*)\{1 + o_p(1)\} + T\lambda_i sgn(\hat{c}_i), \end{aligned} \tag{27}$$

where Σ_i denotes the ith row of Σ and $i \in S_1^c$. By employing the central limit theorem, the first term in equation (27) is of order $\mathcal{O}_p(T^{1/2})$. Furthermore, the condition in Theorem 3.2 implies that its second term is also of order $\mathcal{O}_p(T^{1/2})$. Both are dominated by $T\lambda_i$ since $b_T\sqrt{T} \to \infty$. Therefore, the sign of equation (27) is dominated by the sign of $\hat{c}_i$. Thus (27) can not be equal to 0. Consequently, we must have $\hat{c}_i = 0$ in probability. Analogously, we can show that $\mathrm{P}(\hat{d}_{S_2^c} = 0) \to 1$. This completes the proof. □

Proof of Theorem 3.3 Applying Lemma 3.1 and Theorem 3.2, we have $\mathrm{P}(\widehat{\beta}_2 = 0) \to 1$. Hence, the minimizer of $Q_T(\beta)$ is the same as that of $Q_T(\beta_1)$ with probability tending to 1. This implies that the estimator $\hat{\beta}_1$ satisfies the equation

$$\frac{\partial Q_T(\beta_1)}{\beta_1}\Big|_{\beta_1=\hat{\beta}_1} = 0. \tag{28}$$

According to Lemma 3.1, $\hat{\beta}_1$ is $\sqrt{T}$-consistent. Thus, the Taylor series expansion of equation (28) yields

$$\begin{aligned} 0 &= \frac{1}{\sqrt{T}}\frac{\partial L_T(\hat{\beta}_1)}{\beta_1} + g(\hat{\beta}_1)\sqrt{T} \\ &= \frac{1}{\sqrt{T}}\frac{\partial L_T(\beta_1^*)}{\beta_1} + g(\beta_1^*)\sqrt{T} + \Sigma_0\sqrt{T}(\hat{\beta}_1 - \beta_1^*) + o_p(1), \end{aligned}$$

where g is the first-order derivative of the penalty function

$$\sum_{i\in S_1}\lambda_i|c_i| + \sum_{j\in S_2}\gamma_j|d_j|,$$

and $g(\hat{\beta}_1) = g(\beta_1^*)$ when T is sufficiently large. Furthermore, it can be easily shown that $g(\beta_1^*)\sqrt{T} = \mathcal{O}_p(1)$, which implies that

$$\begin{aligned}(\hat{\beta}_1 - \beta_1^*)\sqrt{T} &= \frac{\Sigma_0^{-1}}{\sqrt{T}} \frac{\partial L_T(\beta_1^*)}{\partial \beta_1} + \mathcal{O}_p(1) \\ &\xrightarrow{d} \mathrm{N}(0, \Sigma_0^{-1}).\end{aligned}$$

The next step is to show $\mathrm{P}(\hat{S}_1 = S_1) \to 1$ and $\mathrm{P}(\hat{S}_2 = S_2) \to 1$. $\forall i \in S_1$ and $p \in S_2$, the asymptotic normality result indicates that $\hat{c}_i \xrightarrow{p} c_i^*$ and $\hat{d}_p \xrightarrow{p} d_p^*$, where $\xrightarrow{p}$ stands for convergence in probability. Thus $\mathrm{P}(i \in \hat{S}_1) \to 1$ and $\mathrm{P}(p \in \hat{S}_2) \to 1$. It suffices to show that $\forall i' \notin S_1$ and $p' \notin S_2$, $\mathrm{P}(i' \in \hat{S}_1) \to 0$ and $\mathrm{P}(p' \in \hat{S}_2) \to 0$, which have been shown by Theorem 3.2. This completes the proof. □

References

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large bayesian vector auto regressions. Journal of Applied Econometrics, 25(1):71–92.

Bernanke, B., Boivin, J., and Eliasz, P. S. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. The Quarterly Journal of Economics, 120(1):387–422.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. (1998). Efficient and adaptive estimation for semiparametric models, 2nd edition. Springer Verlag.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. Annals of Statists, 37(4):1705–1732.

Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Heidelberg: Springer Verlag.

Chernozhukov, V., Belloni, A., and Hasen, C. (2011). Estimation and inference methods for high-dimensional sparse econometric models. Submitted.

Christiano, L., Eichenbaum, M., and Evans, C. (1999). Monetary policy shocks: What have we learned and to what end? Handbook of Macroeconomics, 1(1):65–148.

Chudik, A. and Pesaran, M. (2007). Infinite dimensional vars and factor models. Cambridge Working Papers in Economics 0757, Faculty of Economics, University of Cambridge.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. Journal of Econometrics, 130:337–364.

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

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3):pp. 425–455.

Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32:407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360.

Fengler, M. R., Härdle, W., and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. Journal of Financial Econometrics, 5(2):189–218.

Figlewski, S., Frydman, H., and Liang, W. (2006). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. NYU Stern Finance Working Paper No. FIN-06-007. Available at SSRN: http://ssrn.com/abstract=934438.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. The Review of Economics and Statistics, 82(4):540–554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. Journal of the American Statistical Association, 100:830–840.

Geweke, J. (1977). The dynamic factor analysis of economic time series. Latent Variables in Socio-Economic Models, eds. D. J. Aigner and A. S. Goldberg, Amsterdam: North-Holland, pages 365–383.

Giannone, D., Reichlin, L., and Sala, L. (2005). Monetary policy in real time. In NBER Macroeconomics Annual 2004, Volume 19, NBER Chapters, pages 161–224. National Bureau of Economic Research, Inc.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive lasso for sparse highdimensional regression. Statistica Sinica, 18:1603–1618.

Huang, J. and Zhang, T. (2009). The Benefit of Group Sparsity. ArXiv e-prints.

Janson, S. (2004). Large deviations for sums of partly dependent random variables. Random Structures Algorithms, 24(3):234–248.

Lee, R. D. and Carter, L. (1992). Modeling and forecasting the time series of u.s. mortality. Journal of the American Statistical Association, 87(419):659–671.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. Electronic Journal of Statistics, 2:90–102.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. Proceedings of Conference on Learning Theory (COLT) 2009.

Mol, C. D., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? Journal of Econometrics, 146(2):318 – 328.

Myšičková, A., Song, S., Mohr, P. N., Heekeren, H. R., and Härdle, W. K. (2011). Risk patterns and correlated brain activities. Working paper.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modeling of yield curves. Journal of Business, 60:473–489.

Park, B. U., Mammen, E., Härdle, W., and Borak, S. (2009). Time series modelling with semiparametric factor dynamics. Journal of the American Statistical Association, 104(485):284–298.

Sargent, T. J. and Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Papers 55, Federal Reserve Bank of Minneapolis.

Song, S., Härdle, W., and Ritov, Y. (2010). Dynamic factor models for high dimensional nonstationary time series. Under revision.

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

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics, 20(2):147–62.

Stock, J. H. and Watson, M. W. (2005a). An empirical comparison of methods for forecasting using many predictors. Manuscript, Princeton University.

Stock, J. H. and Watson, M. W. (2005b). Implications of dynamic factor models for VAR analysis. NBER Working Papers 11467, National Bureau of Economic Research, Inc. available at http://ideas.repec.org/p/nbr/nberwo/11467.html.

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

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). IEEE Trans. Inf. Theor., 55:2183–2202.

Wang, H., Li, G., and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. Journal Of The Royal Statistical Society Series B, 69(1):63–78.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fmri data. NeuroImange, 15:1–15.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, 68(1):49–67.

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

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Localising temperature risk" by Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin and Weining Wang, January 2011.
002 "A Confidence Corridor for Sparse Longitudinal Data Curves" by Shuzhuan Zheng, Lijian Yang and Wolfgang Karl Härdle, January 2011.
003 "Mean Volatility Regressions" by Lu Lin, Feng Li, Lixing Zhu and Wolfgang Karl Härdle, January 2011.
004 "A Confidence Corridor for Expectile Functions" by Esra Akdeniz Duran, Mengmeng Guo and Wolfgang Karl Härdle, January 2011.
005 "Local Quantile Regression" by Wolfgang Karl Härdle, Vladimir Spokoiny and Weining Wang, January 2011.
006 "Sticky Information and Determinacy" by Alexander Meyer-Gohde, January 2011.
007 "Mean-Variance Cointegration and the Expectations Hypothesis" by Till Strohsal and Enzo Weber, February 2011.
008 "Monetary Policy, Trend Inflation and Inflation Persistence" by Fang Yao, February 2011.
009 "Exclusion in the All-Pay Auction: An Experimental Investigation" by Dietmar Fehr and Julia Schmid, February 2011.
010 "Unwillingness to Pay for Privacy: A Field Experiment" by Alastair R. Beresford, Dorothea Kübler and Sören Preibusch, February 2011.
011 "Human Capital Formation on Skill-Specific Labor Markets" by Runli Xie, February 2011.
012 "A strategic mediator who is biased into the same direction as the expert can improve information transmission" by Lydia Mechtenberg and Johannes Münster, March 2011.
013 "Spatial Risk Premium on Weather Derivatives and Hedging Weather Exposure in Electricity" by Wolfgang Karl Härdle and Maria Osipenko, March 2011.
014 "Difference based Ridge and Liu type Estimators in Semiparametric Regression Models" by Esra Akdeniz Duran, Wolfgang Karl Härdle and Maria Osipenko, March 2011.
015 "Short-Term Herding of Institutional Traders: New Evidence from the German Stock Market" by Stephanie Kremer and Dieter Nautz, March 2011.
016 "Oracally Efficient Two-Step Estimation of Generalized Additive Model" by Rong Liu, Lijian Yang and Wolfgang Karl Härdle, March 2011.
017 "The Law of Attraction: Bilateral Search and Horizontal Heterogeneity" by Dirk Hofmann and Salmai Qari, March 2011.
018 "Can crop yield risk be globally diversified?" by Xiaoliang Liu, Wei Xu and Martin Odening, March 2011.
019 "What Drives the Relationship Between Inflation and Price Dispersion? Market Power vs. Price Rigidity" by Sascha Becker, March 2011.
020 "How Computational Statistics Became the Backbone of Modern Data Science" by James E. Gentle, Wolfgang Härdle and Yuichi Mori, May 2011.
021 "Customer Reactions in Out-of-Stock Situations – Do promotion-induced phantom positions alleviate the similarity substitution hypothesis?" by Jana Luisa Diels and Nicole Wiebach, May 2011.

SFB 649, Spandauer Str. 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Extreme value models in a conditional duration intensity framework" by Rodrigo Herrera and Bernhard Schipp, May 2011.
023 "Forecasting Corporate Distress in the Asian and Pacific Region" by Russ Moro, Wolfgang Härdle, Saeideh Aliakbari and Linda Hoffmann, May 2011.
024 "Identifying the Effect of Temporal Work Flexibility on Parental Time with Children" by Juliane Scheffel, May 2011.
025 "How do Unusual Working Schedules Affect Social Life?" by Juliane Scheffel, May 2011.
026 "Compensation of Unusual Working Schedules" by Juliane Scheffel, May 2011.
027 "Estimation of the characteristics of a Lévy process observed at arbitrary frequency" by Johanna Kappus and Markus Reiß, May 2011.
028 "Asymptotic equivalence and sufficiency for volatility estimation under microstructure noise" by Markus Reiß, May 2011.
029 "Pointwise adaptive estimation for quantile regression" by Markus Reiß, Yves Rozenholc and Charles A. Cuenod, May 2011.
030 "Developing web-based tools for the teaching of statistics: Our Wikis and the German Wikipedia" by Sigbert Klinke, May 2011.
031 "What Explains the German Labor Market Miracle in the Great Recession?" by Michael C. Burda and Jennifer Hunt, June 2011.
032 "The information content of central bank interest rate projections: Evidence from New Zealand" by Gunda-Alexandra Detmers and Dieter Nautz, June 2011.
033 "Asymptotics of Asynchronicity" by Markus Bibinger, June 2011.
034 "An estimator for the quadratic covariation of asynchronously observed Itô processes with noise: Asymptotic distribution theory" by Markus Bibinger, June 2011.
035 "The economics of TARGET2 balances" by Ulrich Bindseil and Philipp Johann König, June 2011.
036 "An Indicator for National Systems of Innovation - Methodology and Application to 17 Industrialized Countries" by Heike Belitz, Marius Clemens, Christian von Hirschhausen, Jens Schmidt-Ehmcke, Axel Werwatz and Petra Zloczysti, June 2011.
037 "Neurobiology of value integration: When value impacts valuation" by Soyoung Q. Park, Thorsten Kahnt, Jörg Rieskamp and Hauke R. Heekeren, June 2011.
038 "The Neural Basis of Following Advice" by Guido Biele, Jörg Rieskamp, Lea K. Krugel and Hauke R. Heekeren, June 2011.
039 "The Persistence of "Bad" Precedents and the Need for Communication: A Coordination Experiment" by Dietmar Fehr, June 2011.
040 "News-driven Business Cycles in SVARs" by Patrick Bunk, July 2011.
041 "The Basel III framework for liquidity standards and monetary policy implementation" by Ulrich Bindseil and Jeroen Lamoot, July 2011.
042 "Pollution permits, Strategic Trading and Dynamic Technology Adoption" by Santiago Moreno-Bromberg and Luca Taschini, July 2011.
043 "CRRA Utility Maximization under Risk Constraints" by Santiago Moreno-Bromberg, Traian A. Pirvu and Anthony Réveillac, July 2011.

SFB 649, Spandauer Str. 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

HUMBOLDT-UNIVERSITÄT ZU BERLIN

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

044 "Predicting Bid-Ask Spreads Using Long Memory Autoregressive Conditional Poisson Models" by Axel Groß-Klußmann and Nikolaus Hautsch, July 2011.
045 "Bayesian Networks and Sex-related Homicides" by Stephan Stahlschmidt, Helmut Tausendteufel and Wolfgang K. Härdle, July 2011.
046 “The Regulation of Interdependent Markets“, by Raffaele Fiocco and Carlo Scarpa, July 2011.
047 “Bargaining and Collusion in a Regulatory Model“, by Raffaele Fiocco and Mario Gilli, July 2011.
048 “Large Vector Auto Regressions“, by Song Song and Peter J. Bickel, August 2011.

SFB 649, Spandauer Str. 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".