Master Thesis

High-dimensional VAR analysis of regional house prices in United States

Author: Adam Krčál

Supervisor: Ing. Ondřej Čížek PhD.

Department of Econometrics
Faculty of Informatics and Statistics
University of Economics, Prague

Abstract

In this thesis the heterogeneity of regional real estate prices in United States is investigated. A high dimensional VAR model with additional exogenous predictors, originally introduced by Fan, Lv, et al. (2011), is adopted. In this framework, the common factor in regional house prices dynamics is explained by exogenous predictors and the spatial dependencies are captured by lagged house prices in other regions. For the purpose of estimation and variable selection under high-dimensional setting the concept of Penalized Least Squares (PLS) with different penalty functions (e.g. LASSO penalty) is studied in detail and implemented. Moreover, clustering methods are employed to identify subsets of statistical regions with similar house prices dynamics. It is demonstrated that these clusters are well geographically defined and contribute to a better interpretation of the VAR model. Next, we make use of the LASSO variable selection property in order to construct the impulse response functions and to simulate the prices behavior when a shock occurs. And last but not least, one-period-ahead forecasts from VAR model are compared to those from the Diffusion Index Factor Model by Stock and Watson (2002), a commonly used model for forecasts.

Keywords: regional house prices, penalized least squares, LASSO, VAR model, hierarchical clustering, impulse response analysis

Abstrakt

V této diplomové práci jsou prozkoumány závislosti mezi regionálními cenami nemovitostí ve Spojených státech amerických. K tomuto účelu je implementován VAR (Vector Autoregressive) model navržený Fanem a kol. (2011). V tomto konceptu jsou ceny v daných regionech modelovány pomocí zpožděných cen v ostatních regionech. Protože model obsahuje velké množství vysvětlujících proměnných, nelze použít tradiční metody odhadu (např. MNČ). Odhad a zároveň výběr relevantních proměnných je tedy proveden pomocí metody penalizovaných nejmenších čtverců (PLS) s penalizační funkcí LASSO. V teoretické části je představen koncept PLS a jeho varianty, v praktické části je proveden odhad a interpretace VAR modelu a odhad DIF modelu (Stock a Watson (2002)), který je jedním ze zástupců faktorových modelů používaných pro předpovědi. Pro lepší uchopení výsledků odhadu jsou pomocí hierarchického shlukování identifikovány shluky regionů, kde se ceny pohybují podobným způsobem. Výsledné shluky lze velmi dobře interpretovat z geografického hlediska. Protože PLS s penalizační funkcí LASSO pokládá nevýznamné proměnné rovny nule, jsou implementovány i funkce odezvy ke sledování pohybu potenciálního šoku systémem. Nakonec je provedeno srovnání předpovědí z obou modelů a vyhodnocena jejich přesnost.

Klíčová slova: regionální ceny nemovitostí, penalizované nejmenší čtverce, LASSO, VAR model, shluková analýza, funkce odezvy

Declaration of Independence

I hereby declare that I have written this thesis without any help from others and without the use of documents and aids other than those stated above. I have mentioned all used sources and cited them correctly according to established academic citation rules.

January 6, 2016 in Prague .....

Adam Krčál

Acknowledgements

I would like to thank my supervisor, Ing. Ondřej Čížek PhD. from University of Economics in Prague and dr. Laurent Callot from the VU University in Amsterdam for their patient guidance and valuable advice.

Contents

1 Introduction

Real estate market is widely recognized as a very important one due to its size and impact on the state of economy. According to US Census, in 2011 the equity in real estates composed more than 28.3% of all assets owned by an ordinary US household. Moreover, mortgages represent an important component of the financial intermediaries portfolios (Tsatsaronis and Zhu (2004)). The overall outstanding mortgage debt in US according to Board of Governors of the Federal Reserve System, was enormous 13.4 *trillion* dollars at the beginning of 2015, which more than anything else documents the size of the real estate and mortgage market.

It is a well known fact that the recent global economic crisis evolved from the subprime mortgage crisis in the US house market. New financial instruments, such as mortgage-backed securities (MBS), experienced a great boom in the years preceding the crisis. They were sophisticated but highly non-transparent and the credit risk connected to the mortgage collateral was severely underestimated. Consequently the US market was flooded by low-quality (subprime) mortgages and the house prices began to grow above fundamentals. The house price bubble bursted in the mid-2006, which led to massive defaults and eventually to a global financial distress.

This example demonstrates that even major events in the financial world can be closely connected to the house market. Understanding the behaviour and dynamics of house prices and their role in the globalized financial world then seems to be a key challenge the academic researchers should cope with. In recent years there has been a lot of research in this field, see section 2.

This paper focuses on the spatial aspect of the house price dynamics in the United States. In almost 400 statistical regions the house price dynamics is highly heterogeneous. In some states or metropolitan areas the peak before the bubble burst is extremely high and other remain stable for the entire observation period. To gain an insight into the regional characteristics, we make use of hierarchical clustering algorithms to form clusters of statistical areas with similar house prices dynamics. See section 5.3 for figures.

Besides the heterogeneity we would expect a high spatial dependence among regions. The capitol of Michigan and a major industrial center, Detroit, recently became infamous for being the biggest city in US to ever experienced a bankruptcy. The failure of public services resulted in a massive departure of inhabitants and a substantial drop of house prices. In the mid 2013 Detroit was auctioning old houses for few hundreds dollars (Hackman (2013)). It is very rational to expect that such an anomaly in the house market in particular region is likely to affect (either positively or negatively) house prices in neighbouring regions or regions with a strong economic connection to this region.

In order to examine the spatial characteristics, the Vector Autoregressive Model (VAR)

with additional exogenous variables, as proposed by Fan, Lv, et al. (2011), is employed. Lagged house prices in other regions serve as explanatory variables whereas exogenous variables capture the common factor. The problem is set into the high-dimensional framework because statistical indicators of house prices growth are being assembled for several hundreds of regions while we can only hope for 100 observations (in case of quarterly data). For the purpose of estimation and variable selection we adopt the concept of Penalized Least Squares (PLS). The main task of this paper is to study the results thoroughly and to uncover possibly interesting patterns.

In terms of predictions in the high-dimensional framework, various types of factor models are commonly used, e.g. the diffusion index factor model by Stock and Watson (2002). Our final task is thus to compare forecasts from our VAR model with those from the factor model and determine whether they could be useful to some extent.

2 House prices literature review

The Global Financial Crisis 2007 – 2008 also triggered a boom in the academic field. A huge amount of papers is produced every year to study different aspects of the crisis. Due to its unpredictability and complexity it became a true phenomenon. Thus, not surprisingly, a significant fraction of published papers is dedicated to house prices issues. Some study the subprime mortgage crisis, others attempt to identify the house price determinants. There has been a lot of research in the years preceding the crisis, though.

For instance, Leamer (2007) collects a powerful evidence that house prices are strongly connected to the business cycle. He further stresses that a weakness in housing sector and in residential investments is very likely to contribute to recessions. Poterba et al. (1991) study the role of the demographic (age) structure of US population on house price dynamics and find out this link does not hold across regions. Iacoviello and Neri (2008) employ a dynamic stochastic equilibrium model (DSGE) to study the housing market. Their aim is to study the shocks that hit the residential investments and the house prices. Afterwards they examine their impact on the wider economy. To determine what drives the house prices, Tsatsaronis and Zhu (2004) use a structural VAR model on macroeconomic variables and mortgage finance indicators such as GDP, income and interest rates. Gallin (2006) tries to verify the assumption that there exist a long-run relationship between house prices and fundamentals such as income, population and user cost. He founds only a little evidence of cointegration. Goetzmann et al. (2012) react to the recent crisis and argue that expectations based on econometric models tend to underestimate the probability of a rapid price decrease and could have contribute to the asset prices bubble.

In the paper by Y. Li and Leatham (2010) the Large-scale Bayesian Vector Autore-

gressive (LBVAR) and Dynamic Factor Model (DFM) are applied to obtain regional US house prices forecasts. Fan, Lv, et al. (2011) demonstrated that when the spatial dependencies between regional house prices are taken into account, predictive accuracy of DFM or VAR models can be improved significantly. Their VAR model is set into a high dimensional framework estimated by modern variable selection techniques. These and a few other papers provided a strong incentive for us to study the house prices, in particular in US where the spatial heterogeneity is present.

3 Models

The basic OLS regression model

$$y = X\beta + \epsilon, \tag{1}$$

is based on such generating process that for fixed predictors matrix X, fixed vector of true parameters β and stochastic error term ϵ different values of response variable y are generated. Given that X and β is fixed, y has the same covariance structure as ϵ. Imposing Gauss-Markov assumptions (see for example Davidson and MacKinnon (1999)) on the error term ensures that the OLS estimator is BLUE (Best Linear Unbiased Estimator) and the observations of y have constant and finite variance and are not mutually correlated. However, due to the spatial nature of some data (for instance biological or geographical data) causes significant correlation in realizations of a spatially distributed random variable. Anselin and Bera (1998) define the spatial autocorrelation as a coincidence of value similarity with locational similarity. In other words, high or low values for a random variable tend to cluster in space. We can immediately see the problem: if we draw a sample of locations from a spatially autocorrelated random process and do not have panel data, then effectively we have a sample size of *one* for each location.

3.1 Spatial econometrics approach

In order to model the spatial dependence, a wide range of methods has been developed. The Spatial Autoregressive Model (SAR), studied in detail for instance by Anselin (1980), is frequently used in econometric analyses. Several possible representations of a general SAR model are described by LeSage (1999) or Kissling and Carl (2008). A mixed regressive-spatial autoregressive model which implies that the levels of response variable y depend on y in the neighbouring regions, has the following form:

$$y = \rho W y + X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{2}$$

where y is a $n \times 1$ vector of cross-sectional dependent variables. X represents a $n \times m$ matrix of m original predictors. Error term ϵ usually comes from a normal distribution with variance σ^2. ρ and β are standard 1×1 and $m \times 1$ vectors of (auto)regressive coefficients. SAR model representation in (2) assumes that the autoregressive process occurs only in the response variable y. For further discussion on different models, e.g. the model with an autoregressive process in the error term, see Kissling and Carl (2008) or LeSage (1999).

Another important element, $n \times n$ weight matrix W, captures the spatial structure of the cross-sectionally dependent variable y. Depending on the context, W can be defined in various ways. Generally, element w_{ij} represents a measure of the distance between locations i and j. The neighbourhood structure can be identified by an adjacency grid or Euclidean distance. In the geographical context, it might be convenient to take the Earth surface curvature into account. For this purpose the *haversine* formula can be used. Naturally, matrix W has zeros on the main diagonal. The main purpose of ρWY term is to capture the spatial dependence in the observations of the response variable.

Apart from the cross-sectional dimension, economic data usually vary in time. Therefore our problem can be set into the *panel data* framework. According to Viton (2010), most such models assume balanced panels and adopt a time-invariant unobserved component. Thus for the SAR model we have:

$$y_{it} = \rho \sum_{j=1}^{n} w_{ij} y_{jt} + \alpha_i + \beta_i x_t + u_{it}, \quad i = 1, \ldots, n, \tag{3}$$

where α_i is time-invariant intercept component, x_{it} is an element of $T \times m$ matrix X of predictors which does not contain constant term, w_{ij} is an element of weight matrix W, ρ and β_i represent regression coefficients and u_{it} is an idiosyncratic error term. This model is known as a *fixed effects* model.

3.2 Vector autoregressive approach

Fan, Lv, et al. (2011) employ a different approach to obtain more precise predictions of house prices in US. They propose the following simple benchmark model:

$$y_t = \beta x_t + u_t \tag{4}$$

where $y_t = (y_{1,t}, \ldots, y_{n,t})'$ is the n-dimensional response variable of house prices, β represents $n \times m$ coefficient matrix for corresponding fixed $m \times 1$ vector of exogenous predictors x_t (a t-th column of $T \times m$ matrix X of exogenous predictors) and finally $u_t = (u_{1,t}, \ldots, u_{n,t})'$ is an n-dimensional white noise innovation process. To account for

spatial dependencies they employ a high-dimensional VAR(1) with additional exogenous predictors. Thus (4) becomes a VAR(p) that has the following form:

$$y_{i,t} = \sum_{j=1}^{n} a_1^{ij} y_{j,t-1} + \cdots + \sum_{j=1}^{n} a_p^{ij} y_{j,t-p} + \beta_i x_t + u_{i,t}, \quad i = 1, \ldots, n.$$

Rewriting in the matrix form (using notation from Lütkepohl (2005) and (4)) we have

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \beta x_t + u_t. \tag{5}$$

where terms $A_k(a_k^{ij}), k = 1, \ldots, p$ are fixed structural $n \times n$ coefficient matrices and other terms are as above. Macroeconomic variables with different lags are aggregated in only one matrix of predictors to distinguish exogenous factors from those that are generated by the model.

Unlike the spatial econometrics approach, no weight matrix is incorporated and the term ρW is replaced by coefficient matrix A_1 (in the VAR(1) case). Thus the 'weights' are obtained directly by estimation and do not necessary reflect the spatial structure represented by W that we assume the observations have. In (3), ρ is a single parameter assigned to n explanatory variables that represent house prices in other regions. Through parameter ρ weighted by W the response variable (house price measure in i-th region) is generated. However, matrices $A_k(a_k^{ij}), k = 1, \ldots, p$ then contain pn^2 parameters to estimate. Since data for hundreds of US statistical areas are available and the dimensionality of a VAR model increases quadratically, the issue of high dimensionality may arise very easily. The length of economic time series is usually highly limited and thus the number of parameters to be estimated can easily exceed the number of observations, i.e. $n + m > T$ for every singe equation in (5). It can been shown that the simple OLS estimator cannot be used since the analytical solution does not exist: In $\hat{\beta} = (X^T X)^{-1} X^T y$ the $X^T X$ term becomes a $n + m \times n + m$ singular matrix of rank at most equal to T and hence the inverse $(X^T X)^{-1}$ does not exist. Fortunately there has been a major progress in the field of high-dimensionality in recent years a many relevant techniques to estimate and perform the variable selection simultaneously were introduced.

Forecasting house prices locally is important because price dynamics over regions with different economic or demographic profile behaves quite differently (see section 5.1). Although significant predicting power of many key macroeconomic variables such as income or GDP has already been proven, (lagged) house prices in regions that are close either in economic or geographic sense, may play a great role in evaluating levels of house prices in particular region . For instance national level of disposable income may capture a common trend in house prices growth in New York but at the same time it makes a good sense to test whether recent steep growth of real estate prices in Boston, which is in financial sense strongly connected to New York, is likely to affect the local house market.

Here we would like to emphasize that predictors, such as aforementioned GDP or income are assumed to be *exogenous* which is rather a strong assumption. To declare that the GDP or even interest rates should not be generated by a macroeconomic model along with house prices would be audacious and in contradiction with the discussion above. Nevertheless we are primarily interested in the spatial aspect and forecasts. For this purpose, it is convenient to filter the common factor out. Otherwise some kind of DSGE model could be employed.

Thus, unlike (3), VAR(p) model described by Fan, Lv, et al. (2011) models the cross-sectional correlation *explicitly* and does not impose an assumption that the correlation structure is represented by the distance matrix. This feature allows us to interpret results or to identify potentially interesting patters. For instance we may expect that suburb areas are sensitive to price changes in metropolitan areas but not the other way round.

In the previous paragraphs we expressed some reasons why we stick to the VAR(p) model. We make use of modern techniques of estimation from the Penalized Least Squares (PLS) family, namely the *least absolute shrinkage and selection operator (LASSO)* proposed by Tibshirani (1996). See section 4 for details.

3.3 Factor models

Least but not least, a wide variety of *factor models* was developed exclusively for the purpose of forecasting.

In the developed economies, thousands of macro-economic time series are accessible. However, models that are currently used in economical forecasting cannot contain hundreds of explanatory variables. One possibility is to perform a variable selection, as described in the previous chapter, but then the out-of-sample performance rests ultimately on the small subset of selected variables (Stock and Watson (2002)). Nevertheless, macrovariables are usually strongly correlated and can be replaced by a small number of factors that explain almost all variability within the predictors. Thus, if we are interested in forecasting, factor models should be considered as a reasonable choice. These factors can be obtained in various ways. In the paper by Jungbacker and Koopman (2008), the factors are treated as unobservable. Thus, the resulting model has a state space representation and the signal extraction and likelihood evaluation are provided by the Kalman filter. This model is commonly referred to as the Dynamic Factor Model (DFM). Alternatively, Stock and Watson (2002) propose the Diffusion Index Factor Model (DIF). In this framework, a two step estimation procedure is employed. First, the factors are obtained via the principal component analysis. Second, their loadings are estimated by regressing the response variable on estimated factors and response variable lags. The diffusion index factor model set into our panel data framework has the following form:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \psi_1 f_{t-1} + \cdots + \psi_q f_{t-q} + u_t. \tag{6}$$

Fixed $n \times n$ matrices $\phi_1, \ldots, \phi_p$ are assumed to be diagonal. The $n \times k$ matrices $\psi_1, \ldots, \psi_q$, where k stands for the number of factors, contain factor loadings. And finally, y_t and f_t represent the $n \times 1$ vector of response variables (house price measures for n regions) and $k \times 1$ vector of factors respectively. Thus every equation contains only $p + qk$ parameters. The estimation can be easily carried out by OLS. Forecasts based on this model are used as a benchmark to evaluate forecasting performance of (5).

4 Estimation within the high-dimensional setting

High dimensionality poses many challenges for theoretical research as well as for applications. It arises not exclusively in economics and finance but also in sciences such as biology or ecology. High dimensional modelling refers to models with $p >> n$, i.e. where the number of parameters (p) significantly exceeds the number of observations (n)[1]. Theoretical aspects, like different types of asymptotics cannot be neglected[2] (see for example Bühlmann and Van De Geer (2011) or Belloni et al. (2011)).

A large amount of explanatory variables is usually taken into account in the initial stage of modelling. Econometricians are interested in the variable selection – to point out the variables with the strongest explanatory power and exclude the rest from the model. A stepwise procedure, either backward stepwise elimination or forward selection, seems to be the most natural way; the variable with the lowest absolute t-value is excluded from the model in each step or included in the model respectively. This approach, however, suffers from lack of objectivity and exhibits somewhat not 'nice' theoretical properties (Fan and R. Li (1999)). Moreover, backward elimination is not applicable when $n < p$. The best subset selection has such advantage that it considers every subset of variables and simply picks the best one in terms of some criteria. However, the number of models to be estimated grows non-polynomially as more variables are taken into account. Thus in many cases the complete set description is computationally infeasible. Next, modern techniques of variable selection such the Penalized Least Squares (PLS) and others were developed. In this paper, we focus exclusively on the PLS family.

4.1 Penalized Least Squares

If we consider the canonical regression model (1), the PLS optimization problem is defined as follows:

[1] Note the change in notation compared to the previous section

[2] Relatively high dimensionality refers to an asymptotic framework, where the growth of p is of a smaller order of the sample size n (i.e. $p = o(n)$). If p grows polynomially with n (i.e. $p = O(n^\alpha)$ for some $\alpha > 0$) we refer to a ultra high dimensionality.

$$\min_{\beta \in R^p} \left(\frac{1}{2n} ||y - X\beta||_2^2 + \sum_{i=1}^{p} p_\lambda(|\beta_i|) \right) \tag{7}$$

where $||\cdot||_2$ denotes the L_2 norm, i.e. $||y - X\beta||_2^2 = \sum_{i=1}^{n}(y_i - X_i\beta)^2$ is the OLS term and p_λ is a penalty function identified by the regularization parameter $\lambda \geq 0$. The purpose of the non-decreasing penalty function p_λ is to assign a certain penalty to non-zero parameters. Thus, every increase in the value of a particular estimate β_j leads to an increase in value of the penalty function. An algorithm designed to solve a PLS minimization problem must seek balance between the goodness of fit and the size of the penalty. An additional variable included in the model may provide a better solution since the goodness of fit improvement (OLS term $||y - X\hat{\beta}||_2^2$) outweighs the increase of the penalty. Similarly, a different explanatory variable that explains only a little variance of the response variable, may not be included in the model since it is simply not worth it. Thus the estimation of parameters and variable selection is carried out simultaneously. One must bear in mind that in general PLS does not set non-relevant variables to zero. However penalty function can be specified in such way that this property holds (see the PLS properties description in this chapter). For the case of orthogonal design matrix X[3] we have $X^T X = nI_p$ and the ordinary least squares estimator reduces to $\hat{\beta} = n^{-1}X^T y$. Fan and R. Li (2005) argue that imposing this restriction leads to:

$$\min_{\beta \in R^p} \left(\frac{1}{2n} ||y - X\hat{\beta}||_2^2 + \frac{1}{2} ||\hat{\beta} - \beta||_2^2 + \sum_{i=1}^{p} p_\lambda(|\beta_i|) \right). \tag{8}$$

Clearly we can drop the first term since it does not contain β and (8) can be reduced to the following PLS minimization problem:

$$\min_{\beta \in R^p} \left(\frac{1}{2} ||z - \beta||_2^2 + p_\lambda(|\beta|) \right) \tag{9}$$

where $z = (X^T X)^{-1} X^T y = n^{-1} X^T y$ is the OLS estimator. Fan and R. Li (1999) further argue that (9) is equivalent to the following univariate componentwise optimization problem.

$$\min_{\beta_j \in R} \left(\frac{1}{2}(z_j - \beta_j)^2 + p_{\lambda_j}(|\beta_j|) \right), \quad \text{for } j = 1, \ldots, p.$$

where β_j and z_j is the j-th component of β and z respectively. This form is very convenient because one can look at each β_j separately. We suppress the subscript j and let

$$Q(\beta) = \frac{1}{2}(z - \beta)^2 + p_\lambda(|\beta|).$$

[3]Orthogonal matrix is a square matrix whose columns and rows are orthogonal unit vectors. Orthogonal vectors are perpendicular to each other.

The first derivative of $Q(\beta)$ (see for example Fan and R. Li (2005)) is

$$Q'(\beta) = \beta - z + p'_\lambda(|\beta|)\text{sgn}(\beta) = \text{sgn}(\beta)(|\beta| + p'_\lambda(|\beta|)) - z, \tag{10}$$

where $sgn()$ represents the *signum* function and p'_λ is the first derivative of p_λ. According to Antoniadis and Fan (2011), the PLS estimator in (9) may yield the following properties (defined in terms of the derivative (10)):

1. *sparsity* if $\min_{\beta\neq 0}[|\beta| + p'_\lambda(|\beta|)] > 0$. In this case estimated coefficients in absolute value smaller than a certain threshold are set to zero. This is a key feature in the high dimensional framework because complexity of the original model must be reduced. If $|z| < \min_{\beta\neq 0}[|\beta| + p'_\lambda(|\beta|)]$, the derivative (10) is positive for all positive βs and negative for all negative βs (see figure 1). Consequently the PLS estimator $\hat{\beta} = 0$ because $\text{argmin}_\beta\ \text{sgn}(\beta)(|\beta| + p'_\lambda(|\beta|)) - z = 0$. If $|z| > \min_{\beta\neq 0}[|\beta| + p'_\lambda(|\beta|)]$, two crossings (solutions) exist and the larger one represents the desired non-zero PLS estimator.

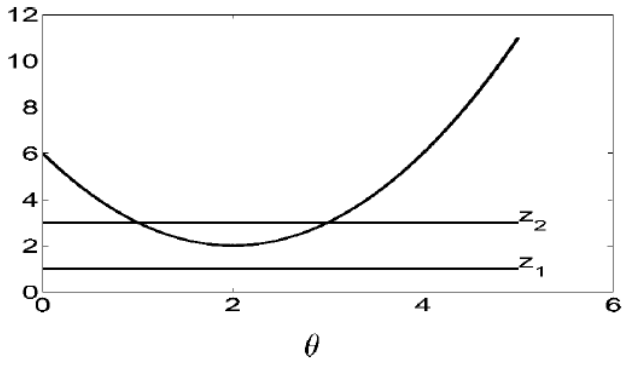


Figure 1: A plot of $\beta + p'_\lambda(|\beta|)$ against $\beta \geq 0$. Source: Fan and R. Li (2001).

2. *approximate unbiasedness* if $p'_\lambda(|\beta|) = 0$ for large $|\beta|$, in which case the resulting estimator is nearly unbiased, especially when the true coefficient β is large, to reduce model bias (see for example Zhang and Huang (2008) or Fan and R. Li (2001)). The above mentioned condition effectively means that the penalty assigned to large βs is directly proportional to $|\beta|$. Consequently, $\hat{\beta} = z$ and the estimator is approximately unbiased.

3. *continuity* if and only if $\text{argmin}_\beta[|\beta| + p'_\lambda(|\beta|)] = 0$, i.e. the penalty function must be continuous in data. This property helps to maintain prediction stability of the model. The above mentioned sparsity condition implies the continuity property.

4. *oracle property* when the estimator asymptotically identifies the true subset of variables. Let $A = \{j : \beta_j \neq 0\}$ be the true subset of predictors and assume that

$|A| = p_0 < p$. According to Zou (2006) and Huang and Xie (2007), a procedure of subset selection, that produces $\hat{\beta}$, is called an *oracle* procedure if it asymptotically satisfies the following conditions:

- Identifies the right subset model, i.e. $\{j : \hat{\beta}_j \neq 0\} = A$
- Has the optimal estimation rate, i.e. $\sqrt{n}(\hat{\beta}_A - \beta_A) \xrightarrow{d} N(0, \Sigma)$, where 0 is a null vector and Σ is the true covariance matrix knowing the true subset model[4].

In order to determine whether (9) satisfies the above mentioned conditions, the form of the penalty function must be explicitly defined. In practice, L_q penalties are commonly used. L_q norm of a vector x is defined as $||x||_q = \sum_{i=1}(x_i^q)^{\frac{1}{q}}$. PLS with the L_2 penalty is equivalent to the Tikhonov regularization (commonly known as the *ridge regression*) proposed by Tikhonov (1963):

$$p_\lambda(|\beta|) = \lambda||\beta||_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}.$$

The L_q penalty with $0 < q \leq 2$, where q is a subject to optimization, was proposed by Frank and J. H. Friedman (1993) and is called bridge regression. In the case of L_0 penalty ($p_\lambda = \frac{\lambda^2}{2} I(|z| > 0)$, the same penalization is given to all non-zero coefficients and the minimization problem 9 then results in a combinatorial search through all possible subsets of variables and thus might consume a huge amount of computational time.

The *hard-thresholding* rule $\hat{\beta} = zI(|z| > \lambda)$ is represented for instance by a penalty function of the following form:

$$p_\lambda(|\beta|) = |\beta|I(|\beta| \leq \lambda) + \lambda/2 I(|\beta| > \lambda), \tag{11}$$

where λ is the regularization parameter. Note that-hard thresholding rule is not continuous (see figure 3). Fan and R. Li (1999) show that the hard thresholding rule is equivalent to the backward stepwise elimination where in each step the variable with the highest t-value is removed. In particular, for the orthogonal design matrix X, simply the variable with smallest $|\hat{\beta}|$ is eliminated. Suppose now that the elimination is carried out k times. The remaining variables are those with the highest $p - k$ values of $|\hat{\beta}|$. This is equivalent to using the hard-thresholding rule with thresholding parameter $\gamma \in (|\hat{\beta}|^{(k)}, |\hat{\beta}|^{(k+1)})$.

Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO) which is nothing else than L_1 penalty:

$$p_\lambda(\beta) = \lambda||\beta||_1 = \sum_{i=1}^{p} |\beta_i|$$

[4]Meaning that $\sqrt{n}(\hat{\beta}_A - \beta_A)$ converges *in distribution* to a multivariate normal distribution with particular characteristics.

and the optimization problem is defined as follows:

$$\min_{\beta \in R^p} \left(\frac{1}{2n} ||y - X\beta||_2^2 + \lambda \sum_{i=1}^{p} |\beta_i) \right).$$

LASSO yields some interesting properties. First, unlike the ridge regression (a together with the hard thresholding rule (11), clipped penalty (12) and SCAD penalty) the coefficients can be set exactly to zero, i.e. the sparsity condition is fulfilled. Clear explanation is provided for instance in Bühlmann and Van De Geer (2011). Moreover, along with SCAD, LASSO is continuous (see figure 3 which plots OLS estimator against PLS estimates). However, LASSO suffers from bias which was studied in detail for instance by Zhang and Huang (2008). This issue is well documented in figure (2). The dotted line is straight thus the condition for *approximate unbiasedness*, $p'_\lambda(|\beta|) = 0$, is not fulfilled for any $|\beta|$. LASSO leads to the following solution:

$$\hat{\beta} = \text{sgn}(z)(|z| - \lambda)_+ = \begin{cases} 0, & \text{for} \quad |z| < \lambda; \\ \text{sgn}(z)(|z| - \lambda), & \text{for} \quad |z| \geq \lambda. \end{cases}$$

This is a *soft-thresholding* rule which is much finer than the hard-thresholding one.

To address the bias problem, an extension called adaptive LASSO was introduced by Zou (2006). In this framework certain weights are defined and assigned to the penalty function. Next the author shows that with a proper choice of regularization parameter λ the resulting estimator has the oracle property (unlike LASSO). Thus we have:

$$\min_{\beta \in R^p} \left(\frac{1}{2n} ||y - X\beta||_2^2 + \lambda \sum_{i=1}^{p} \hat{w}_i |\beta_i| \right).$$

Zou (2006) further suggests that $\hat{w}_i = 1/|\hat{\beta}_i^*|^\gamma$ where $\hat{\beta}_i^*$ is an estimator obtained for instance by OLS but preferably, when OLS is not available, by ridge regression. Since ridge regression sets no parameters to zero, $\hat{w}_i$ is always positive for $i = 1, \ldots, n$. See figure (3) for the adaptive LASSO with two different values of threshold parameter γ plotted against the OLS estimate. Clearly, the bias is eliminated.

Antoniadis and Fan (2011) introduced a clipped L_1 function $p_\lambda(|\beta|) = \lambda \min(|\beta|, \lambda)$ and showed that the solution is a mixture of soft and hard thresholding rule:

$$\hat{\beta} = \text{sgn}(z)(|z| - \lambda)_+ I(|z| \leq 1.5\lambda) + zI(|z| > 1.5\lambda). \tag{12}$$

So far the most sophisticated penalty function which is based on (12), was proposed by Fan and R. Li (1999) and is called the smoothly clipped absolute deviation penalty (SCAD):

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{for} \quad |\beta| \leq \lambda; \\ -(\beta^2 - 2a\lambda|\beta| + \lambda^2)/[2(a-1)], & \text{for} \quad \lambda < |\beta| \leq a\lambda; \\ (a+1)\lambda^2/2, & \text{for} \quad |\beta| > a\lambda. \end{cases}$$

The first derivative of SCAD is defined as follows (Huang and Xie (2007)):

$$p'_\lambda(|\beta|) = \begin{cases} \text{sgn}(\beta)\lambda, & \text{for} \quad |\beta| \leq \lambda; \\ \text{sgn}(\beta)(a\lambda - |\beta|)/(a-1), & \text{for} \quad \lambda < |\beta| \leq a\lambda; \\ 0, & \text{for} \quad |\beta| > a\lambda. \end{cases}$$

and has the following solution:

$$\hat{\beta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{for} \quad |z| \leq 2\lambda; \\ [(a-1)z - \text{sgn}(z)a\lambda]/(a-2), & \text{for} \quad 2\lambda < |z| \leq a\lambda; \\ z, & \text{for} \quad |z| > a\lambda. \end{cases}$$

The penalty function has two unknown parameters, λ and a and is continuously differentiable outside 0. Its derivative vanishes outside $[-a\lambda, a\lambda]$. Parameter a is usually set to 3.7. In figure 2 the three 'regimes' of SCAD penalty are clearly observable. For $|\beta| \in [0, \lambda)$ the value of the function grows linearly with $|\beta|$ and for $|\beta| > a\lambda$ the penalization assigned to β grows proportionally as $|\beta|$ increases. Thus the continuity property holds. Moreover, Huang and Xie (2007) showed that SCAD penalty has oracle property.

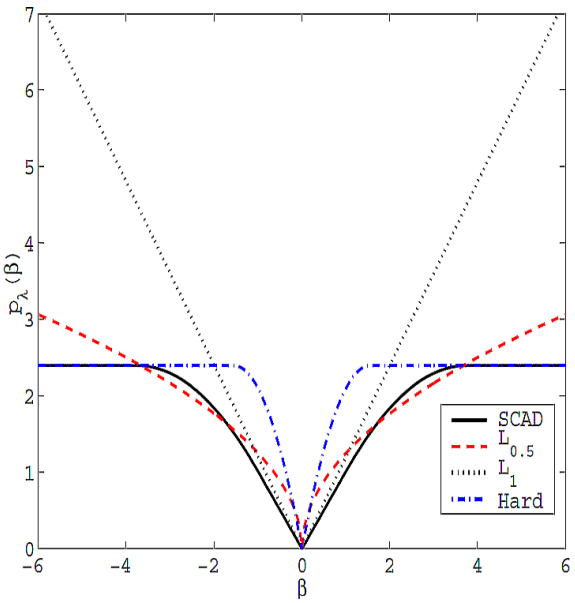


Figure 2: Different penalty functions of β – SCAD, Bridge $L_{0.5}$, LASSO L_1 and the hard thresholding rule. Source: Fan and R. Li (2005).

Apart from the well established theoretical framework, we must discuss possible computational issues that may arise while implementing PLS. PLS with LASSO or adaptive LASSO leads in fact to a convex minimization problem and effective algorithms based on the Newton-Raphson method may be used (Bühlmann and Van De Geer (2011)). In particular, the *glmnet* package that we make use of, employs the *coordinate descent* method (see section (4.3)). Osborne et al. (2000) utilizes the fact that LASSO can be specified as a quadratic program with convex objective function and a linear constraint. Efron et al. (2004) established the Least Angle Regression (LARS) which is an another fast and effective algorithm. Non-convex optimization problems, that arise when SCAD and similar folded concave penalty functions are employed, are problematic. For instance Zhang (2010) introduced and extended LARS algorithm called PLUS, which can be used when the penalty function is a quadratic spline such as the SCAD. In this paper we avoid usage of SCAD entirely.

The final question is how to determine the tuning parameter λ. For the diverging number of parameters the traditional model selection instruments such as the information criteria might not identify the true model consistently. In the PLS framework this issue is crucial because otherwise some interesting properties of the penalty functions do not exist. Therefore Wang et al. (2009) proposed the modified Bayesian Information Criterion (BIC*) a demonstrated that it selects the true model consistently regardless the choice of penalty function. The modified BIC has the following form:

$$BIC_\lambda = \log(\hat{\sigma}_\lambda^2) + |S_\lambda| \frac{\log n}{n} C_n, \tag{13}$$

where $\hat{\sigma}_\lambda^2 = n^{-1}||y - X\beta||^2$, S_λ represents the subset of predictors that was chosen by $\hat{\beta}_\lambda$ and n is number of observations. For $C_n = 1$ modified BIC reduces to the traditional BIC. Wang et al. (2009) further argue that theoretically C_n is only required to go to infinity as $d \to \infty$ (d represents the number of non-zero parameters) and that a function with arbitrary slow rate of convergence can be used, but in numerical experiments they use $C_n = \log\{\log(d)\}$. The optimal tuning parameter is given by $\hat{\lambda} = \arg\min_\lambda(BIC_\lambda)$.

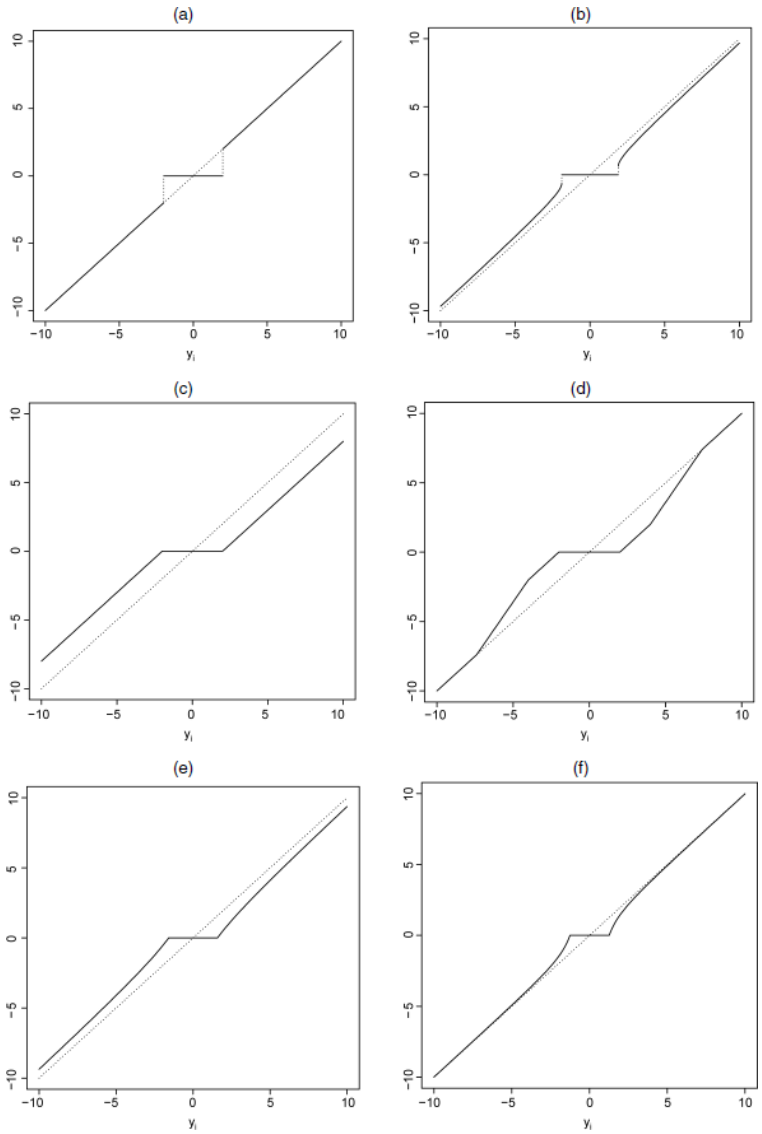


Figure 3: PLS estimates with various penalty functions ($\lambda = 2$) (y-axis) against OLS estimates (x-axis). (a) the Hard; (b) Bridge $L_{0.5}$; (c) the Lasso; (d) the SCAD; (e) the Adaptive Lasso $\gamma = 0.5$; and (f) the Adaptive Lasso, $\gamma = 2$. The dotted line represents the OLS estimate plotted against OLS estimate. Source: Zou (2006).

4.2 PLS for the VAR model

If we consider the VAR(1) with exogenous variables defined in the previous section (5) and adopt the notation from (7), the PLS minimization problem with LASSO penalty has the following form:

$$\min_{\beta_i, A_1^i \in R^n,} \left(\frac{1}{2T} ||y_i - X\beta_i - A_1^i y_{t-1}||_2^2 + \lambda_i \sum_{j=1}^{n} |a_1^{ij}| \right),$$

where $A_1^i (a_1^{ij})$ is the i-th row of a fixed $n \times n$ matrix of autoregressive coefficients. β_i is the i-th column of $n \times m$ matrix of regressive coefficients assigned to X ($T \times m$ matrix of exogenous predictors). Variable $y_t = (y_{1,t}, \ldots, y_{n,t})'$ is an $n \times 1$ vector that represents house price measures in regions $1, \ldots, n$. And finally the response variable y_i represents a T-dimensional vector of house prices in region i. Clearly, only parameters contained in A_1^i are subject to penalization. Variables in X are always included in the model. Generalization to VAR(p) is straightforward.

To obtain all estimates, the PLS optimization procedure must be applied to each VAR(p) equation (for $i = 1, \ldots, n$). Thus we end up with n PLS problems with different regularization parameters λ_i but in case of LASSO, adaptive LASSO and ridge regression this is not an issue since fast and effective algorithms exist.

4.3 The *glmnet* package for Matlab

The *glmnet* package was originally developed for *R-project* users by J. Friedman et al. (2010). The port to Matlab environment is carried out by Qian et al. (2013). The optimization procedure is a subroutine written in Fortran. It contains extremely effective procedures for fitting the entire LASSO or elastic-net path for generalized linear regression models (GLM), including logistic and multinomial regression, Poisson regression or the Cox model using the *cyclical coordinate descent* optimization algorithm (CCD). Suppose we have a multivariate function $f(x)$. CCD iteratively optimizes $f(x)$ along one direction, i.e. solving a univariate minimization problem in each step of the loop while other variables remain fixed. According to the official documentation by Hastie and Qian (2012), the general objective function for the *gaussian* family has the following form:

$$\min_{\beta_0,\beta\in R^p} \frac{1}{2n}||y - X\beta||_2^2 + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1] \tag{14}$$

where the response y is n-dimensional, X is a $n \times p$ matrix of predictors and β represents a $p \times 1$ vector of parameters. λ is the regularization parameter and $\alpha \in [0; 1]$ represents a compromise between LASSO ($\alpha = 1$) and the ridge ($\alpha = 0$) penalty. Thus the penalty function used here is a linear combination of LASSO and ridge called *elastic net* which was introduced by Zou and Hastie (2005). Since (14) is a continuously differentiable function, the CCD algorithm for our objective function can be summarized as follows:

1. Choose an initial parameter vector $\hat{\beta}^{(0)}$
2. Do until the convergence is reached or a termination condition is satisfied. At each step s:
   - Denote current estimate as $\hat{\beta}^{(s)}$
   - Choose an index j from $1, \ldots, p$
   - Using the gradient at $\beta_j^{(s)} = \hat{\beta}_j^{(s)}$ compute the update as:

     $$\hat{\beta}_j^{(s+1)} = \frac{S(\frac{1}{n}\sum_{i=1}^n x_{ij}(y_i - y_i^{(j)}), \lambda\alpha)}{1 + \lambda(1-\alpha)},$$

     where $y_i^{(j)} = \sum_{k\neq j} x_{ik}\hat{\beta}_k$ and $S(z, \gamma)$ is a soft-thresholding operator $sign(z)(|z| < \gamma)_+$. This formula applies when the variables in X are standardized, i.e. have zero mean and unit variance. For the *gaussian* family X is standardized by default.

Index j in the second step can be chosen in different ways. Except the fact that *glmnet* uses warm starts and active set of iterations, authors do not provide additional details.

For more details on effective optimization algorithms see Wu and Lange (2008). The *glmnet* function has the following syntax:

$$fit = glmnet(x, y, family, options)$$

- **x** is a $n \times p$ matrix of input variables
- **y** is the response variable, which is quantitative for the *gaussian* family, *binary* for the binomial family, etc.
- **family** specifies the type of GLM (the full list is mentioned above)
- **options** is a structure set by the *glmnetSet* function

According to the Qian et al. (2013), the *glmnetSet* structure contains the following options (only those relevant for the *gaussian* family are reported):

- **options.alpha:** The mixing parameter, with $0 < \alpha <= 1$ such that $\alpha = 1$ is the LASSO and $\alpha = 0$ is the ridge penalty.
- **options.nlambda:** The number of lambda values, default is 100.
- **options.lambda:** A user supplied lambda sequence.
- **options.standardize:** Logical for x variable standardization, prior to fitting the model sequence. The coefficients are always returned on the original scale.
- **options.weights:** Observation weights.
- **options.intr:** Should intercept be fitted (default = true) or set to zero (false).
- **options.lambda_min:** Smallest value for λ, as a fraction of λ_{max}, the (data derived) entry value (i.e., the smallest value for which all coefficients are zero)?
- **options.thresh:** Convergence threshold for coordinate descent.
- **options.dfmax:** Limit the maximum number of variables in the model.
- **options.pmax:** Limit the maximum number of variables ever (in each iteration) to be nonzero.
- **options.exclude:** Indices of variables to be excluded from the model.
- **options.penalty_factor:** Separate penalty factors can be applied to each coefficient. Can be 0 for some variables, which implies that these variables are always included in the model. Default is 1 for all variables.

- **options.cl:** Two-row matrix with the first row being the lower limits for each coefficient and the second the upper limits.
- **options.gtype:** Two algorithm types are supported (only) for family = 'gaussian'. The default when $p < 500$ is options.gtype = 'covariance'. This can be much faster than options.gtype='naive' which can be more efficient for $p >> n$ situations, or when $p > 500$.
- **options.ltype:** If 'Newton' then the exact hessian is used (default), while 'modified.Newton' uses an upper-bound on the hessian, and can be faster.

And the most important output arguments are:

- **fit:** A structure.
- **fit.a0:** Intercept sequence of length *length(fit.lambda)*.
- **fit.beta:** $p \times length(fit.lambda)$ matrix of coefficients.
- **fit.lambda:** The actual sequence of lambda values used.
- **fit.dev:** The fraction of deviance explained (for "gaussian" family, this is the R-square).
- **fit.df:** The number of nonzero coefficients for each value of lambda.
- **fit.dim:** Dimension of coefficient matrix (ices).
- **fit.call:** A cell including the names of all the input variables in the parent environment.

Effectively, *glmnet* computes a monotonously increasing sequence of λ_j values in such way that $\max_j(\lambda_j)$ is the smallest value of λ_j for which no penalized variables are included in the model, $\min_j(\lambda_j)$ is given by **options.lambda_min** and $j = 1, \ldots,$ **options.nlambda**. Then, for every λ_j a vector of estimates is computed and the selection must be carried out separately (see section 4.1).

5 An application to US house prices

5.1 Data description

Without a solid data background no analysis would be possible. Our dataset can be divided into two subsets: house prices data and exogenous predictors data. To measure the house prices growth, the house price index (HPI) is commonly used. In United States, in particular, three main indices based on different methodologies are regularly assembled:

- *Case-Shiller index* by Standard & Poor's is based on the weighted, repeat-sales (WRS) methodology proposed by Case and Shiller (1989). Weighted, repeat-sales means that it measures average price changes in repeat sales or refinancings on the *same* properties, which are assumed to undergo no significant changes (see Nagaraja et al. (2014))[5]. Full methodology is described in Calhoun (1996). It is published monthly for 20 and 10 most important metropolitan areas (MSAs) in US.
- *House price index (HPI) by Federal Housing Finance Agency (FHFA)* is a weighted, repeat-sales index for single family detached properties using data on conventional conforming mortgage transactions obtained from the Federal Home Loan Mortgage Corporation (Freddie Mac) and the Federal National Mortgage Association (Fannie Mae) (Calhoun (1996)). It is based on a modified WRS methodology and is published quarterly for 384 metropolitan areas and divisions. A wide range of composite indices is constructed as well.
- *Residential price index (RPI) by FNC Inc.* is based both on public records of sales transactions and proprietary appraisal data collected by FNC (FNC (2010)). Individual indices for 30 major metropolitan areas as well as composite indices for 10, 20, 30, 100 metropolitan statistical areas (MSA) are published monthly.

According to Calhoun (1996) there are several differences between the HPI and Case-Shiller index. Unlike the Case-Shiller, the all-transaction variant of HPI also takes into account the mortgage refinance appraisals, rather than purchase prices merely. The price trends of the most expensive properties have a greater influence on Case-Shiller index since it is value weighted. There is no such issue in case of HPI (see figure 4). FHFA publishes composite indices for states and census divisions as well as local indices on the MSA level. The residential price index is largely based on Case-Shiller but takes the rising quality of the houses into account. In this paper, we make use of FHFA house

[5] Another commonly used method of constructing a house price index is the *hedonic regression*. In this framework the value of a particular estate is decomposed into constituent characteristics that are believed to contribute to the resulting value. The repeat sales methodology poses a nonnegligible advantage: it addresses the problem that the hedonic regression does not capture all characteristics (see Nagaraja et al. (2014)).

price indices exclusively used since they are constructed for the entire set of MSAs. In particular three datasets are assembled:

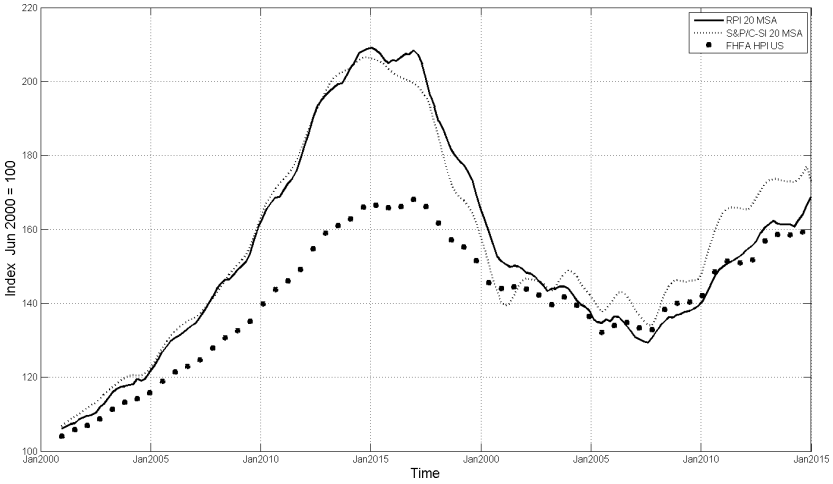


Figure 4: Comparison of composite residential house price index (FNC) for top 20 metropolitan statistical areas (MSA), composite S&P/Case-Shiller house price index for top 20 MSA and national FHFA house price index (HPI) (; 2000 – 2015, June 2000 = 100, seasonally unadjusted, HPI is quarterly.

- 384 metropolitan statistical areas and divisions, 1994 Q1 – 2013 Q1, seasonally unadjusted (*regions384*)

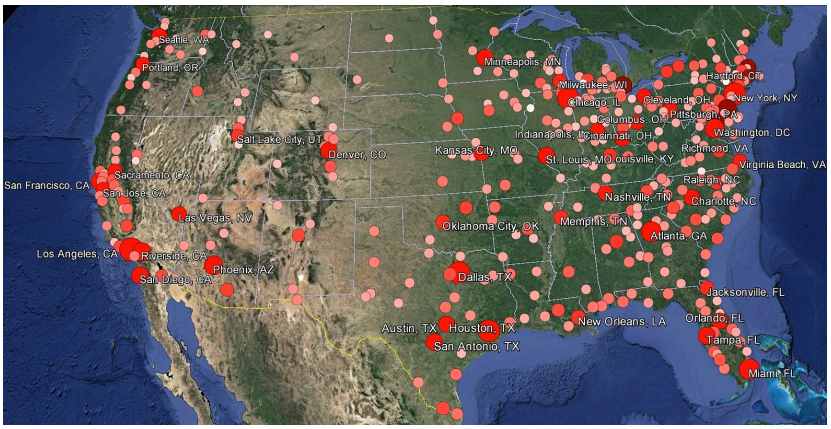


Figure 5: 384 metropolitan statistical areas sorted by population (darker color and bigger dot means bigger population), graphs by Gephi and Google Earth.

- 100 largest metropolitan areas, 1991 Q1 – 2014 Q4, seasonally adjusted (*metro100*)
- 51 US states, 1986 Q1 – 2014 Q4, seasonally unadjusted (*states51*)

Dataset *regions384* is key to our analysis. This decision is driven by the fact that it contains enough entries to observe heterogeneity and spatial dependencies. We also in-

spect the *metro100* dataset since population in larger MSAs is more likely produce enough house purchases and mortgage refinances to make the resulting price representative (Fan, Lv, et al. (2011)).

As mentioned in section 4, apart from house prices we also make use of several national level macroeconomic variables. Our choice is based on the past experience with house prices forecasting; we incorporate variables whose influence on house prices was proven in scientific studies. These variables are mostly of macroeconomic nature:

- *Real gross domestic product (GDP).* Low GDP means an overall lack of demand in the market, which drives the prices down. The real GDP in 2009 dollars is used.
- *Industrial production index (IPI)* is another measure of economic performance.
- *Consumer price index (CPI)* captures the inflation. We are interested in house price changes driven by change in fundamentals, not by inflation. The aggregate CPI for all urban consumers (all items) is incorporated.
- *Interest rate* is directly connected to the house prices. Lower interest rates make the mortgages affordable for a wider range of households. The demand and consequently the prices increase. According to Tsatsaronis and Zhu (2004), house prices are more sensitive to the short term interest rates in markets where the mortgage contracts include floating rate. In United States the mortgage interest rate is fixed. Despite this fact we make use both of the dollar based 3-month London Interbank Offered Rate (LIBOR) and 30-year fixed average mortgage rate for United States.
- *Disposable income* is another macroindicator that may positively influence the house prices. Intuitively, the higher income households have, the higher demand on the house market they comprise. However, according to Tsatsaronis and Zhu (2004) and Gallin (2006) income has a surprisingly small explanatory power. Despite these findings, the real disposable income in 2009 dollars is used.

5.2 Data processing

From all datasets, non-continental states (Alaska and Hawaii) and their statistical areas are excluded to make the visualisations of the results well arranged. The time span is as long as possible with respect to the available length of individual time series. MSAs that would shift the whole dataset due to an extremely short time series were excluded. We ended up with 377 regions for *regions384* dataset, 99 metropolitan areas for *metro100* dataset and 49 states for *states51* dataset.

Since all house prices data we collected are in the form of base indices, we adjust the macrovariables in the same manner. Furthermore we performed the seasonal adjustments

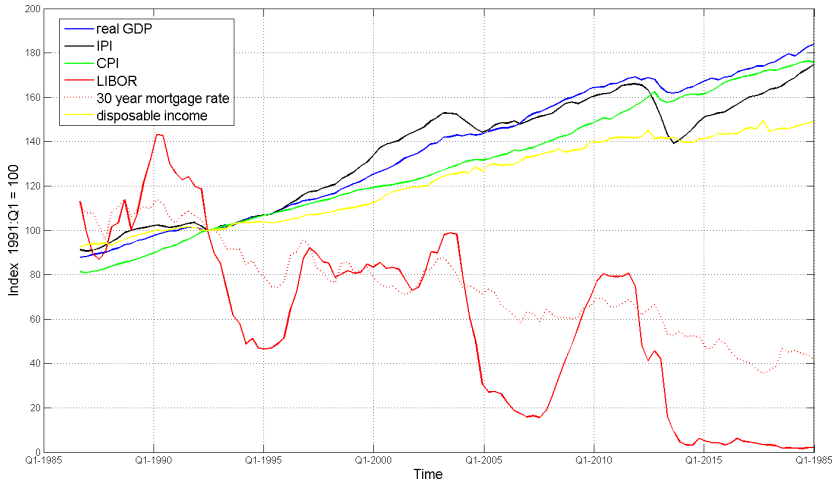


Figure 6: Quarterly time series of indices, 1986 Q1 – 2014 Q4, 1991 Q1 = 100, real GDP in 2009 dollars, industrial production index (IPI), consumer price index (CPI), London Interbank Offered Rate (LIBOR), 30-year fixed average mortgage rate and real disposable income in 2009 dollars.

using the *X13arima* procedure. To avoid spurious regression, we tested all series for stationarity by ADF test with intercept (C), trend-intercept (CT) and no intercept (NC) specification:

$$\text{(NC):} \qquad \Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-k} + \epsilon_t$$

$$\text{(C):} \qquad \Delta y_t = \alpha + \gamma y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-k} + \epsilon_t$$

$$\text{(CT):} \qquad \Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-k} + \epsilon_t$$

The number of lags of response variable to include in the ADF regression was determined by BIC. We test the null hypothesis $H_0 : \gamma = 0$ against the one-sided alternative $H_1 : \gamma < 0$. The test statistics ($t-$ratio of $\hat{\gamma}$) has tabulated critical values.

Only a fraction of time series in all datasets on 0.05 level is stationary (see table 1) and hence the VAR(p) and other models are estimated on first differences. Thus we must bear in mind that such model can capture only the short-term relations among regions.

	states51	*metro100*	*regions384*	ex. predictors
ADF_{NC}	0	0	0	1
ADF_{C}	0	1	2	0
ADF_{CT}	13	17	29	1
N of series	49	99	377	6

Table 1: Number of stationary time series according to particular specification of Augmented Dickey-Fuller regression.

5.3 Heterogeneity in regional house prices

As mentioned in introduction, there is a high level of spatial heterogeneity in house prices. Some states and metropolitan areas are characterized by a steep growth of house prices, others experience a moderate growth during the entire observation period. Different series of house prices may clearly exhibit different dynamics. Thus the first step of our analysis is, not surprisingly, an attempt to assemble several groups of metropolitan areas or states with similar dynamics. We further show that the resulting arrangement is not random but follows a certain geographical pattern. Although conclusions from this chapter are not fundamental for the VAR(p) model estimation itself, they help us comprehend and understand the main results.

For this purpose we perform the *cluster analysis*, which is a statistical technique to divide a set of n objects into $p << n$ subsets. From a wide range of algorithms we, after some tests, selected the hierarchical clustering algorithm. In this framework the objects are iteratively connected to form the user-specified amount of clusters. Whether certain object is merged with an existing one depends on which cluster linkage method the algorithm makes use of. For instance, the between-groups linkage method iteratively forms clusters with respect to the overall longest 'distance' among the clusters. Similarly, within-groups linkage method pursues the shortest total distance among the objects in the same clusters. The distance between two objects can be defined in various ways - we discussed some in the section (3). However the most natural way to measure the resemblance between two time series of quantitative data is the simple (Pearson) correlation coefficient. The analysis is performed in PASW Statistics (SPSS). For a detailed description of hierarchical clustering algorithms see the documentation of SPSS[6]. HPI plots were generated by Matlab and maps with network graphs were created using Gephi, a free software for network visualizations (Bastian et al. (2009)) and via the ExportToEarth plugin exported to the Google Earth environment (GoogleEarth (2013)).

For *metro100* and *states51* dataset three clusters are assembled. We discovered that smaller clusters do not reflect different kinds of dynamics well and are difficult to interpret. In figures 7 and 9 we observe the HPI for states and metropolitan areas respectively.

[6]In particular, we make use of PASW Statistics 18.0.0.

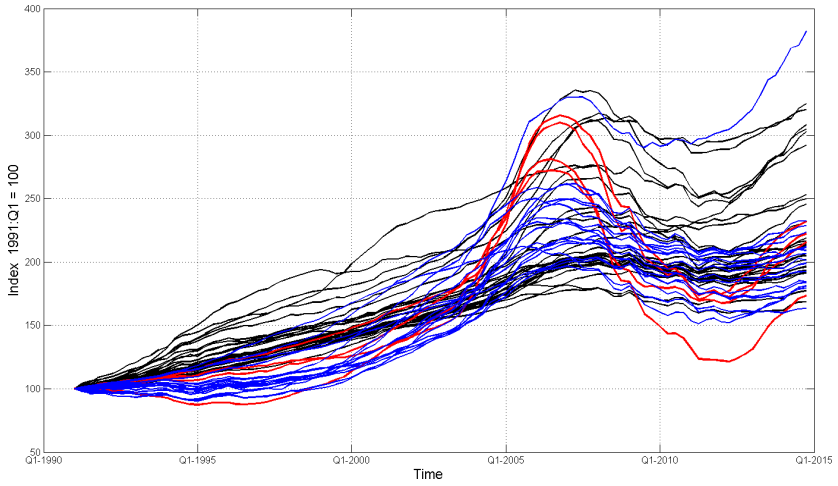


Figure 7: Results from the cluster analysis, *states51* dataset: the extreme peak states are red, moderate peak states are blue and non-peak states are black. Data are HPI with 1991 Q1 = 100.

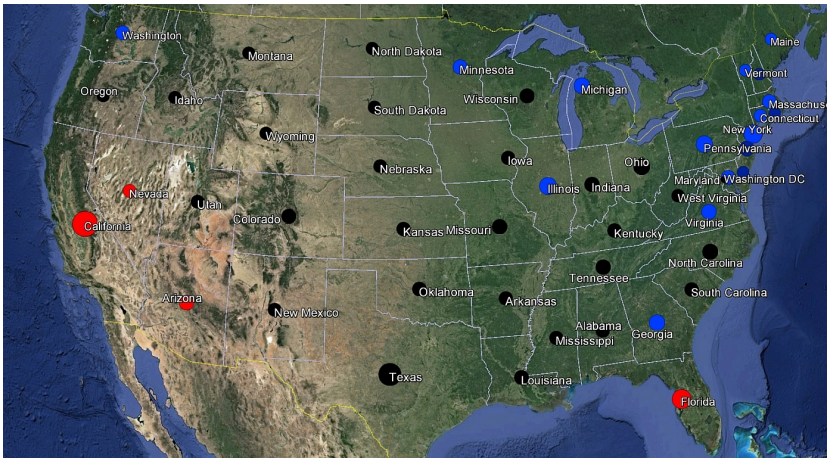


Figure 8: Results from the cluster analysis, *states51* dataset: the high peak states are red, moderate peak east coast states are blue and moderate-peak continental states are black.

According to hierarchical cluster analysis we can divide them into three groups. The red group consists of states/MSAs where the peak in the mid 2006 was extremely high and the drop that followed was immensely deep – namely states as California, Arizona, Nevada and Florida. Taking into account their location, an explanation arises naturally. Affordable mortgages convinced many Americans to search for new homes in highly attractive subtropical locations and seaside resorts and the prices were pushed even more above their fundamentals. This idea is supported by the fact that Nevada experienced 35 % , Arizona 24 %, Florida 18 % and California 10 % growth of population during 2000 – 2010 period. By the end of 2010, house prices were back on the pre-crisis level. As for 100

largest metropolitan areas (*metro100* dataset), the resulting clusters are not distributed well in the geographic sense, which may indicate that lesser MSAs should not be omitted in the further analyses.

The moderate peak group, marked by the blue color, consists mainly of states that are located on the east coast and is driven by the Boston-New York-New Jersey-Philadelphia-Washington, DC agglomeration. East coast was hit not so hard by the real estate market collapse.

And finally, the third group, marked by the black color, exhibits only moderate or even non-peak dynamics. As evident from the figures 8 and 10, these states/MSAs are located in the mid-west territory and are characterized by a lower level of urbanization, smaller population and agriculture of a great importance.

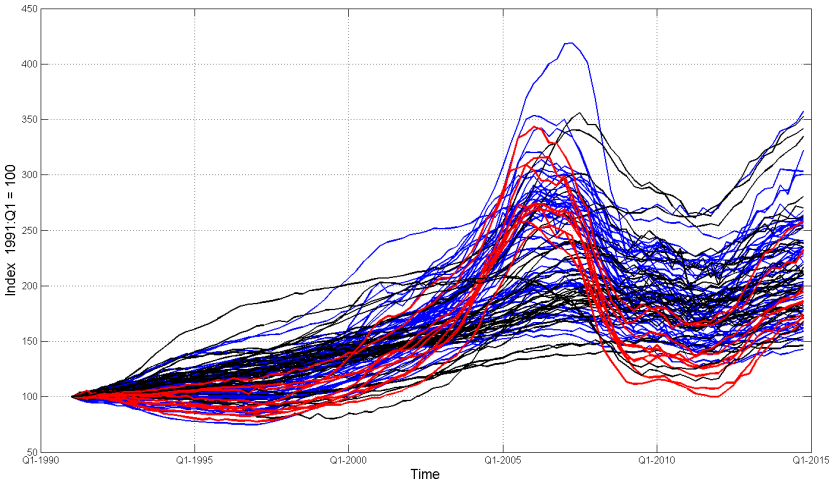


Figure 9: Results from the cluster analysis, *metro100* dataset, HPI with 1991 Q1 = 100.

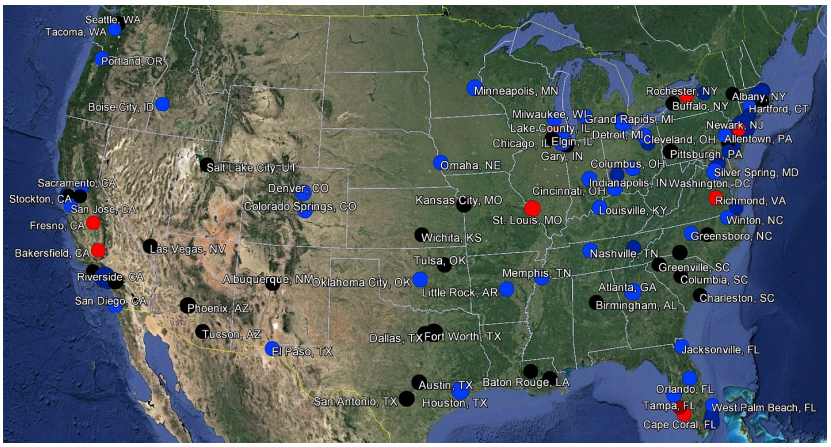


Figure 10: Results from the cluster analysis, *metro100* dataset.

The most interesting results were obtained by applying hierarchical cluster analysis (using within-groups linkage method) on the *regions384* dataset (figures 11 and 12). Every group from the total of four is very clearly defined in the geographic sense. First cluster (yellow) is concentrated around Detroit agglomeration in Michigan. House prices in these MSAs experience their peak in the late 2005 which is considerably earlier than in other bubble MSAs. They also grow almost linearly in the years preceding the burst of the bubble and decrease slowly in the subsequent period.

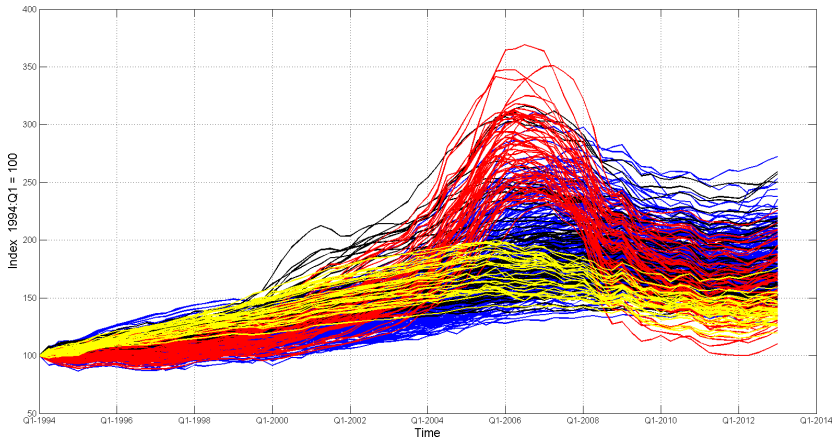


Figure 11: Results from the cluster analysis, *regions384* dataset, HPI with 1994 Q1 = 100.

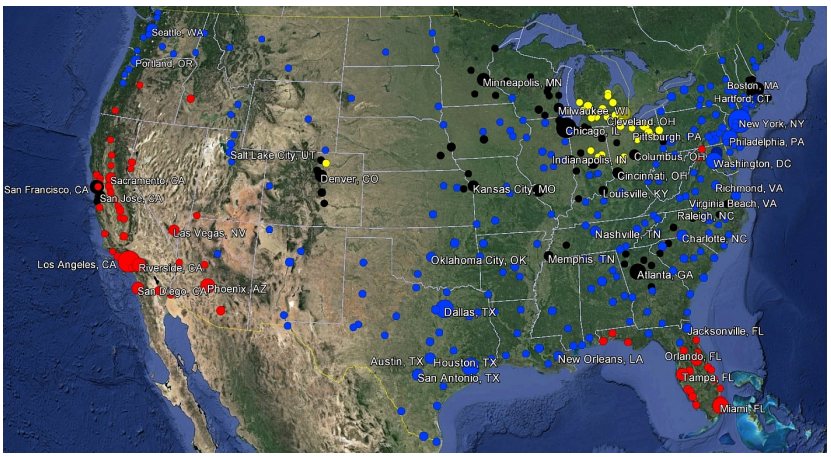


Figure 12: Results from the cluster analysis, *regions384* dataset.

Red, high peak MSAs, that were already characterized in the previous paragraphs, are located on the south-western coast and on the Florida peninsula. The blue group has rather diverse dynamics – it contains both MSAs that experienced the price boom (Boston-New York-New Jersey-Philadelphia-Washington, DC agglomeration on the east coast and Seattle on the on the west) and those that did not (MSAs in Rocky Mountains

or in Texas). And finally the black group, which shows a moderate peak behaviour, is located mainly in the Great Plains in the midwestern US but includes Boston and Atlanta agglomerations and affiliated MSAs as well. To sum up, we have:

1. *Blue group*: the largest one, rather heterogeneous (in terms of house prices dynamics), located on the east coast around New York agglomeration, on the north-western coast and in the south (Dallas and New Orleans neighbourhood).
2. *Yellow group*: geographically homogeneous group with a moderate peak dynamics located around Detroit and Cleveland.
3. *Black group*: heterogeneous, moderate peak MSAs located in the mid-west, Chicago, Atlanta and Boston neighbourhoods.
4. *Red group*: homogeneous group with bubble-like dynamics, seaside resorts such as Florida and California.

In the figure 13 we present a network graph generated by the Force Atlas algorithm. Nodes represent individual MSAs and edges a strong positive correlation (threshold is set to 0.98). Being a force-directed algorithm, Force Atlas positions the nodes so that the edges have more or less equal length and that there are as few crossing edges as possible. Since there are relatively few edges between particular groups, i.e. the mutual within group correlation dominates the between group correlation, the groups have a 'nice' spatial distribution. This picture also gives an intuition how 'close' the clusters are to each other in terms of mutual correlation.

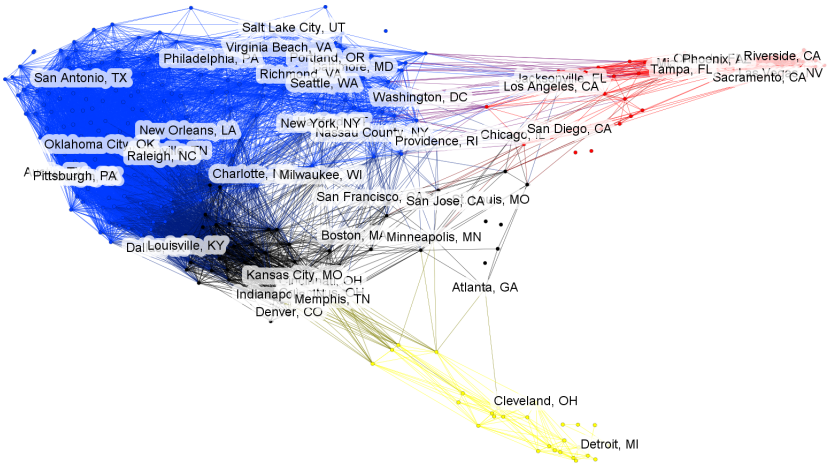


Figure 13: Network graph by Gephi software. The edge between two nodes exists if correlation between their house prices series exceeds 0.98. Arrangement into groups is given by cluster analysis. The layout is generated by Force Atlas algorithm.

To give an intuition about the volatility in house prices during the crisis we compare the maximum value of HPI in particular series with its standard deviation. In table 2 we

can observe that series from the yellow cluster (2.) have the highest spikes compared to their standard deviation. On the contrary, the magnitude of spikes in red cluster series (4.) does not deviate from their volatility to such extent. Almost 100 % of HPI series have their maximum greater than four standard deviations which may indicate that of more or less pronounced bubble is present in the entire set of MSAs. In the figure 14 the ranking of individual MSAs based on their maximum/standard deviation ratio is depicted.

	4 × std	5 × std	6 × std	7 × std	8 × std
1.	99.6%	87.9%	59.6%	29.1%	9.4%
2.	100.0%	100.0%	100.0%	96.3%	55.6%
3.	100.0%	84.1%	73.9%	52.2%	26.1%
4.	100.0%	53.4%	3.4%	0.0%	0.0%

Table 2: Fraction of series in *regions384* dataset from particular cluster that have its maximum greater than $x\times$ standard deviation.

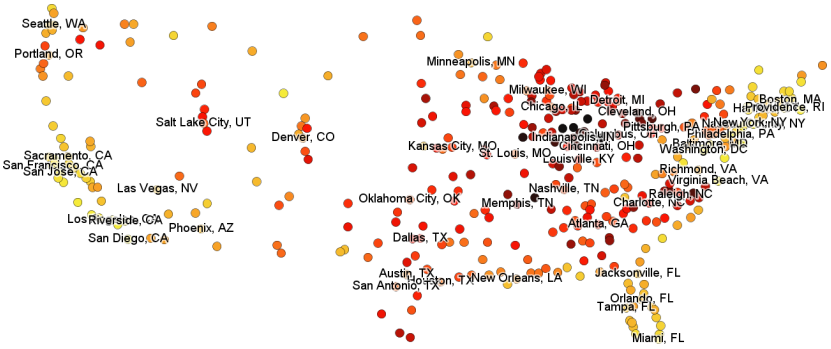


Figure 14: Visualisation of maximum to standard deviation ratio for the regional HPI series. (*regions384* dataset). Yellow MSAs have the smallest ratio, dark red and black MSAs have the highest.

5.4 Common factor modelling

We have (5):

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \beta x_t + u_t. \tag{15}$$

The coefficient matrix β in (15) is left unpenalized. Thus the exogenous predictors collected in X are always included in the model. We hope to capture the common factor of house prices dynamics while the unexplained spatial patterns are modelled as a VAR process (terms $A_1 y_{t-1} + \cdots + A_p y_{t-p}$) via the variable selection. We stick to the simple VAR(1) because further lags contribute with little explaining power and even bring more autocorrelation to the residuals, i.e. higher fraction of sub-models represented by individual VAR(p) equations exhibits a significant autocorrelation according to the Ljung-Box test. Moreover, the interpretation of higher order VAR in the multidimensional setting would be problematic.

In the next step we try select the best proxy for the common factor. Effectively we estimate (15) using PLS for different specification and lags of common factor x_t. One possibility is to take the simple arithmetic mean of HPI in all regions or the composite HPI for United States. Next we consider to include the macrovariables described in the section 5.1.

In fact, to select the best subset for all 6 variables (GDP, IPI, CPI, short term rate, mortgage rate and disposable income) and more than 2 lags is computationally impossible. Suppose that we observed that one LASSO optimization procedure for *metro100* takes 0.63 seconds. Thus the best subset selection for 6 variables $\times$ 4 lags would take 2936 hours[7]. However, one may argue that the selection can simply be carried out by PLS, as the coefficients in A_1 matrix (15). In fact, after some experiments, we observed that these variables are rarely selected, which contradicts our idea to separate the common factor and the spatial dependencies.

To overcome selection issues and to account for all variables, we computed the first principal component of all variables and used it as a proxy. Since the macrovariables may not be capable to capture the common factor satisfactorily, we added the national level HPI, which is constructed as a composite indicator of house prices in the largest metropolitan areas. According to the Bartlett's test of sphericity[8], on $< 1\%$ level we can not reject the null hypothesis that the variables are linearly independent a thus the factor/ principal component analysis is applicable. The first and the only principal component explains satisfying 90,7 % of the total variance. However, in practice the first principal

[7] Altogether we have 24 predictors. Thus the number of possible subsets of all cardinalities is given by $\sum_{i=1}^{24} \binom{24}{n}$.

[8] Bartlett's test of sphericity tests whether the observed correlation matrix is equal to identity matrix, i.e. $H_0 : \hat{\Sigma} = I$ against $H_1 : \hat{\Sigma} \neq I$. or $H_0 : |\hat{\Sigma}| = 1$ against $H_1 : |\hat{\Sigma}| = 0$

component does not perform very well.

variables	AIC	BIC	modBIC	HQ
CPI(3), USHPI (3)	597.947	632.141	574.287	611.612
RGDP (3), USHPI (3)	607.420	640.786	582.824	620.754
RGDP (2), USHPI (4)	608.839	642.757	584.969	622.394
CPI(3), USHPI (4)	608.388	643.257	585.568	622.323
RGDP(3), USHPI (4)	611.113	644.848	586.952	624.595
All Variables (3)	614.259	705.526	659.597	650.734
All Variables (1)	656.052	742.565	694.703	690.626
1st PC (3)	643.016	681.872	639.132	658.545
1st PC (1)	657.388	697.655	656.905	673.481

Table 3: Best subsets according to modified BIC containing up to 2 variables and information criteria for the principal component and for full set of variables. Estimated by LASSO using *regions384* dataset. Numbers in brackets stand for lag order.

For illustration, in the table 3 the best sub-models according to modified BIC containing up to 2 variables estimated by LASSO for *regions384* dataset are reported. Results for *metro100* dataset are to be found in the appendix (table 14). Modified BIC and other information criteria are computed for each equation separately and then counted up.

The national level HPI and its lags up to 4th order appear to have a non-negligible explanatory power. In terms of specification issues the Ljung-Box test of residual autocorrelation and ARCH test of heteroskedasticity are employed. If we consider the *regions384* dataset, table 4 documents the number of autocorrelated residual series for selected specifications of common factor. When the common factor is modelled by 3 lags of all macrovariables and national level HPI, the number of VAR equations with significant residual correlation up to 4th order and heteroskedasticity is minimized [9]. We further observe similar results for *metro100* dataset (table 15 in appendix) and for *states50* dataset (table 16).

Due to the fact that for each common factor specification different penalized variables are selected, we do not put an excessive emphasis on these results. We also take into account the *interpretability* of patterns that arise in the estimates of A_1 coefficients for different specifications of the common factor. To construct a VAR model with exogenous predictors for house prices in 377 metropolitan areas (*regions384* dataset) we eventually decided to include 3 lags of real GDP, income, LIBOR, mortgage rate, CPI, IPI and national level HPI. Similarly, we include two lags of aforementioned variables in the complementary models (*metro100* and *states50* dataset respectively).

[9] It is minimized with respect to all subsets that consist of up to two variables and subsets reported in table 4

LASSO	Ljung-Box test				ARCH	sparsity	
regions384	1.	2.	3.	4.	test	pen.	total
1st PC (1)	21.5%	19.9%	36.9%	34.2%	30.0%	3.3%	3.6%
1st PC (1–2)	18.6%	19.1%	30.5%	30.0%	31.6%	3.5%	4.1%
1st PC (1–3)	22.3%	18.8%	27.9%	26.3%	26.8%	3.2%	4.0%
1st PC (1–4)	20.7%	20.4%	30.5%	27.9%	26.8%	3.3%	4.4%
All variables (1)	21.2%	24.4%	32.4%	29.2%	34.2%	3.2%	5.0%
All variables (1–2)	22.3%	18.3%	28.1%	21.0%	28.9%	3.1%	6.7%
All variables (1–3)	25.7%	22.5%	22.5%	20.2%	18.0%	2.8%	8.1%
All variables (1–4)	23.1%	24.7%	22.5%	21.0%	21.0%	2.6%	9.5%
HPIUS (1)	27.1%	32.1%	46.2%	45.4%	33.4%	3.1%	3.4%
HPIUS (1–2)	27.6%	28.6%	40.3%	41.9%	33.2%	3.3%	3.9%
HPIUS (1–3)	29.2%	24.7%	27.6%	27.3%	25.7%	3.3%	4.1%
HPIUS (1–4)	29.2%	23.3%	24.9%	24.4%	22.0%	3.3%	4.3%

Table 4: Results for different specifications of matrix X using LASSO estimator and *regions384* dataset. The table contains: Ljung-Box test of autoccorelation in residuals of 1. – 4. order (percentage of total residual series that are correlated on $\alpha = 0.05$), ARCH test of heteroskedasticity (percentage of total residual series that exhibit heteroskedasticity on $\alpha = 0.05$) and percentage of non-zero coefficients (total and penalized). PC stands for principal component and HPIUS for the national level HPI.

5.5 VAR model estimation results

The VAR model for different datasets was estimated by LASSO using the *glmnet* package for Matlab (Qian et al. (2013), see section 4.3). We also tested the adaptive LASSO but we obtained much less sparse matrices of coefficient estimates. This is also the reason why it provides better in-sample fit according to modified BIC and other information criteria. In addition, the weighting scheme of adaptive LASSO deteriorates the pattern that exists within the LASSO estimates. Thus we stick to simple LASSO in the entire section.

In figure 15 the matrix of estimates (A_1) from (15) is visualised. If $\alpha_1^{ij} > 0$, a blue dot is placed at (i, j) and an increase of house prices in region j (x-axis) at time $t-1$ causes (in Granger sense) an increase of contemporaneous house prices in region i (y-axis). And similarly, a red dot is placed, when $\alpha_1^{ij} < 0$. The rows and columns of A_1 matrix are sorted on the basis of the cluster analysis results (see figure 12) such that their group membership is respected. Other coefficients are set to zero. The complete allocation of all 377 MSAs is given in the section C in appendix. Dashed lines represent borders between individual groups.

Group 1 (blue) (as listed in 5.3) exhibits no clear pattern, which corresponds to heterogeneity in terms of dynamics (figure 11). Nevertheless, a potential change in house prices of MSAs in group 4 (red) seem not to affect prices in group 1 (and groups 2 and 3 as well) at all. Since no other group is so distant from the rest in terms of crisis depth

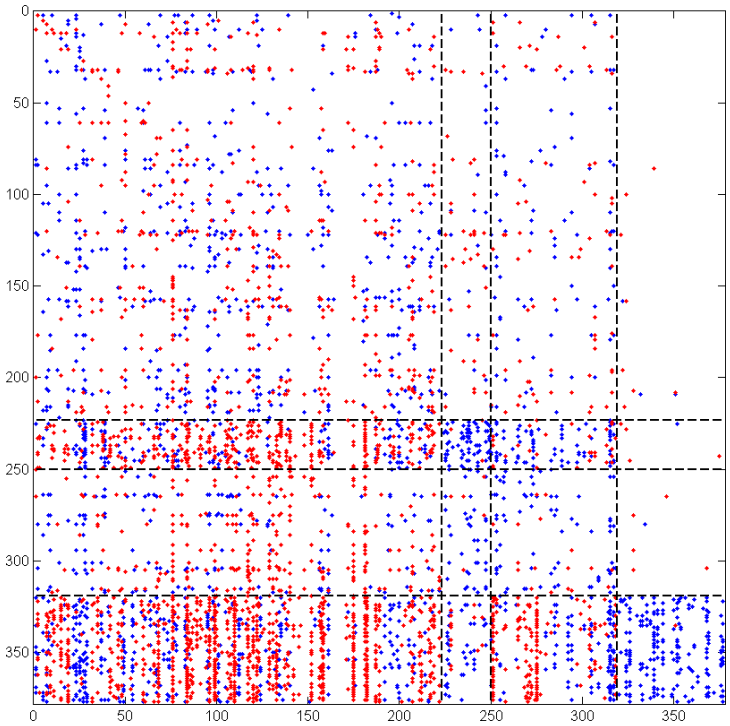


Figure 15: Visual representation of the sorted matrix of estimated coefficients (*regions384* dataset). Estimates for individual equations are in rows. Blue dots represent positive coefficients, red negative. Horizontal and vertical lines represent borders between the clusters.

group	R^2_{adj}	AIC	BIC	BIC*	HQ	Ljung-Box test, order				ARCH	n
						1.	2.	3.	4.	test	
1.	0.542	0.917	1.725	2.015	1.240	30.5%	24.2%	26.0%	22.9%	16.1%	223
2.	0.784	0.053	1.462	2.375	0.616	11.1%	18.5%	7.4%	7.4%	11.1%	27
3.	0.580	0.718	1.603	1.969	1.072	24.6%	24.6%	29.0%	26.1%	23.2%	69
4.	0.916	1.086	2.675	3.793	1.721	15.5%	15.5%	8.6%	8.6%	22.4%	58

Table 5: Statistics for individual groups. Model with dataset *regions384* as the input. According to 5.3, in figure 12 the first group is marked by blue, second by yellow, third by black and fourth by red color.

that followed after the bubble burst, these results could be anticipated. MSAs in Group 2 (yellow) are closely connected to each other. A positive shock is likely to spread quickly and cause an increase in house prices in the entire group. Majority of these MSAs is also affected negatively by the MSAs in the first group and positively by the third group. Group 3 (black) is in many ways similar to group 1. On the contrary, the submatrix containing coefficients that explain house price dynamics of the fourth group is much less sparse. MSAs from groups 1 and 3 have mostly negative influence. Furthermore, the level of positive interconnection is as high as within the group 2.

Table 5 contains average information criteria and fractions of correlated and het-

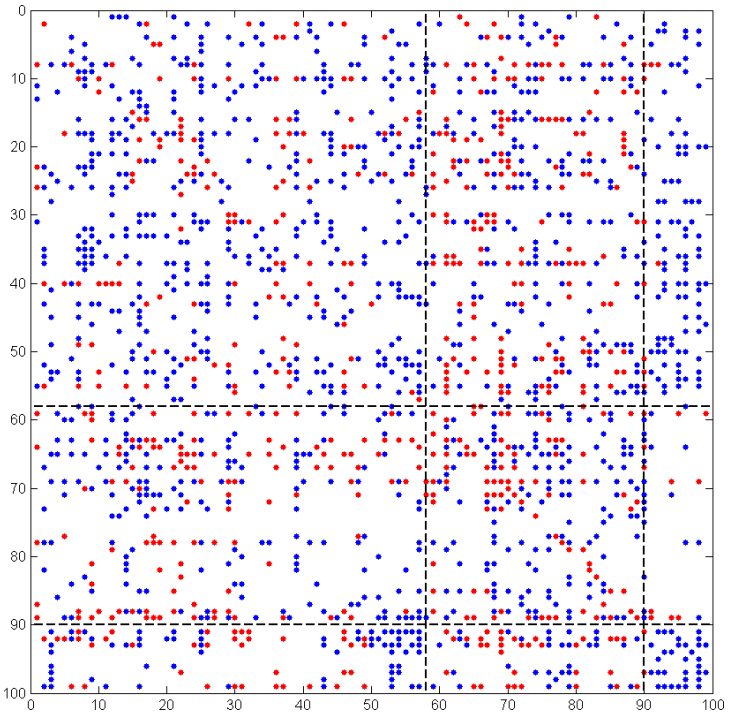


Figure 16: Visual representation of the sorted matrix of estimated coefficients (*metro100* dataset). Estimates for individual equations are in rows. Blue dots represent positive coefficients, red negative. Horizontal and vertical lines represent borders between the clusters.

eroscedastic residual series for individual groups. For MSA with plenty of explanatory variables (groups 2 and 4) the modified BIC, which penalizes the number of coefficients most, is the highest but in exchange the fraction of autocorrelated residual series is the lowest and the adjusted R^2 highest. In other words, the fact that the equation is overfitted is compensated by better residual characteristics.

Sparse matrix of estimates for metro100 dataset (figure 16) can, unlike the previous model, hardly be interpreted. The non-zero coefficients are randomly distributed, which documents the importance of lesser MSAs when modelling spatial dependencies. According to our calculations, the average modified BIC for 100 largest MSAs decreases by 35 % when lesser MSAs, that are usually directly connected to the most important and largest metropolitan areas, are included in the model. In further analysis we omit this dataset since the latter two appear to be more suitable.

And finally, in figure 17 the sparse matrix of estimates for VAR(1) model of *states51* dataset is depicted. We may observe that states in the second (red) group (California, Nevada, Arizona and Florida) as well as states from the third group (black) are positively interconnected. Similarly as in the *regions384* case, the high peak cluster has the best fit but in exchange for a high number of parameters. This finding can be related to an

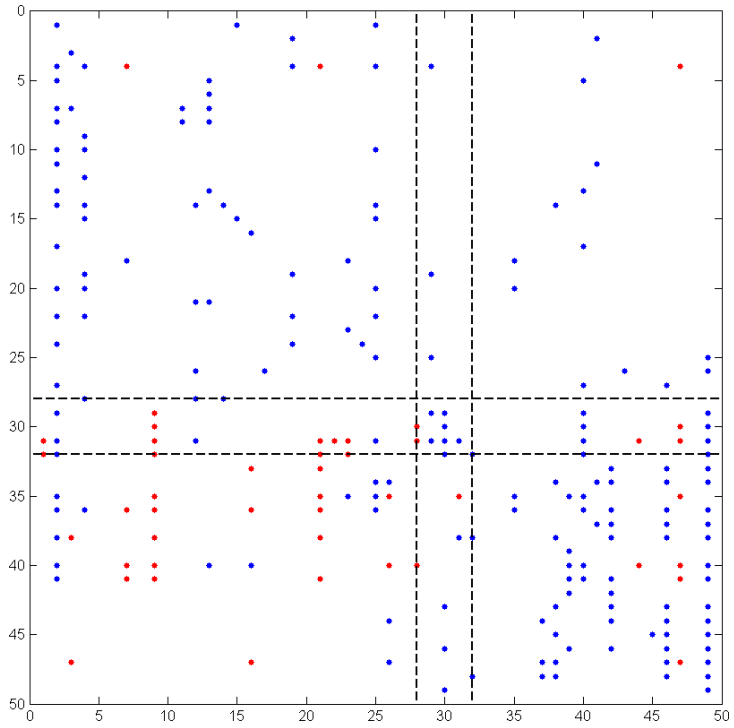


Figure 17: Visual representation of sorted estimates matrix (*states51* dataset). Blue dots represent positive coefficients, red negative. Horizontal and vertical lines represent borders between the clusters.

increased vulnerability of 'bubble' MSAs/states to price changes in other regions.

group	R^2_{adj}	AIC	BIC	BIC*	HQ	Ljung-Box test, order				ARCH	n
						1.	2.	3.	4.	test	
1.	0.470	0.608	1.061	1.093	0.791	53.6%	50,0%	60.7%	60.7%	46.4%	28
2.	0.861	1.800	2.425	2.591	2.052	0%	0%	0%	0%	100%	4
3.	0.708	0.911	1.456	1.559	1.131	23.5%	17.6%	58.8%	47.1%	76.5%	17

Table 6: Statistics for individual groups. Model with dataset *states51* as the input. In figure 8 the first group is marked by blue, second by red and third by black color.

In all applications above the heteroskedasticity in residuals is a major issue. In general, the larger the spike is, the more likely the variance of residual series is a function of time.

At this place we would like to stress that these results should be interpreted carefully. Due to the fact that no standard errors[10] are computed and no inference is carried out, we recommend to keep the sense of perspective and inspect these results as a big picture, possibly with the help of cluster analysis results, rather than in terms of individual MSAs

[10]Tibshirani (1996) argues that for a non-linear and non-differentiable function (PLS LASSO estimator) it is difficult to obtain accurate estimates of standard errors. He proposes a closed form formula based on a transformation of the penalty function or bootstrap but both of these techniques are only approximative.

or states. To gain even more insight into the spatial pattern of house prices, we further apply the following approach. We merge the metropolitan areas into 20 clusters according to their geo-coordinates (figure 18). Using the sorted matrix of estimates (figure 19) we can point out some interesting results:

- Clusters 3–7 (mid-west and Florida) are mostly likely to experience a decrease when prices on the east coast (clusters 1 & 2) grew in the previous quarter.
- Clusters 3–7 form a mutually interconnected block that decrease in prices when regions located in the Great Plains (11–15) experience an increase.
- This relation appears to hold for the north western coast (1) as well.
- On the contrary, sub-matrix of estimated coefficients for MSAs located in the Great Plains, Rocky Mountains and on the north-western coast (13–19) is much more sparse. Texas (12) is more likely to share features with clusters 3–7.
- And finally, MSAs on the south-western coast are divided into two clusters, 17 and 20, that behave as a one homogeneous cluster – are positively interconnected and mostly negatively connected to the clusters located in the midwest and on the central eastern coast.

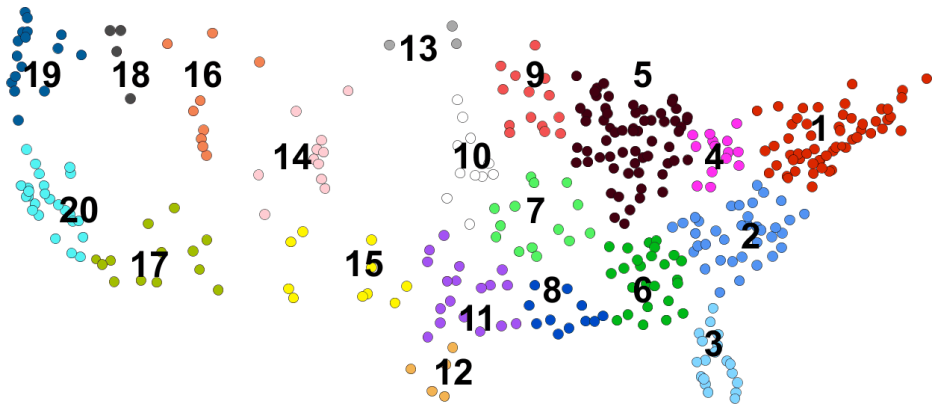


Figure 18: Twenty MSA clusters, average within-group linkage method.

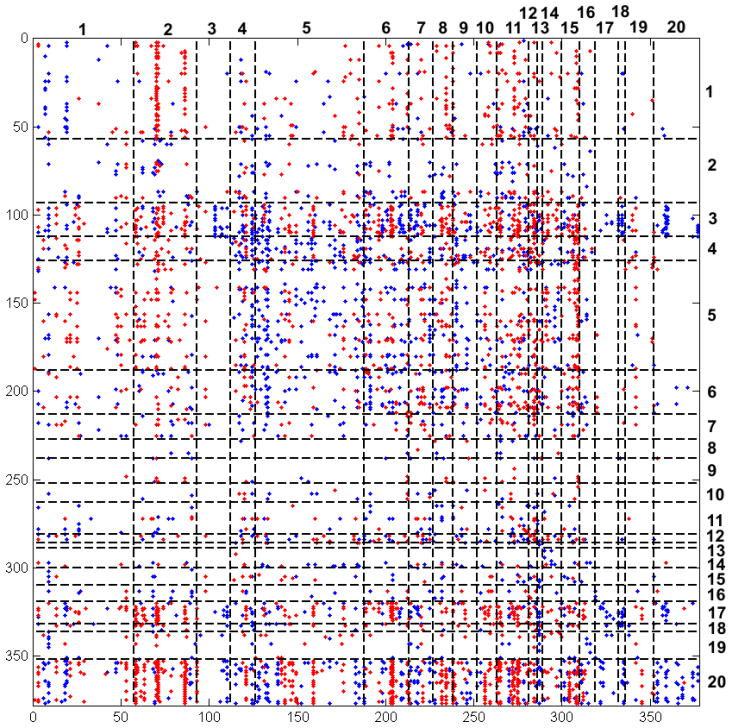


Figure 19: Visualisation of matrix of estimates (*regions384* dataset). Blue dots represent positive coefficients, red negative. Horizontal and vertical lines represent the borders between groups according to figure 18. Estimates for individual equations are in rows.

5.6 DIF model estimation results

As for the DIF model (6), several model specifications were tested (tables 7 and 8). For both datasets, only three factors (principal components) with eigenvalue greater than 1 were computed. These factors explain 98.6 % and 97.9 % respectively of the total variance in the datasets (see scree plots in figure 26). According to BIC, the best models contain only first lag of 3-dimensional factor f_t and up to six lags of response variable y_t. As benchmark models for the purpose of forecasting we select DIF(3,1) and DIF(5,1) since they give mostly uncorrelated residual series and moreover provide the most accurate forecasts (see section 5.8). Slightly less than one third of them is heteroskedastic but other specifications suffer from this issue as well.

lags (y_t)	lags (f_t)	Ljung-Box, order 1.	2.	3.	4.	ARCH test	AIC	BIC	BIC*	HQ	R^2_{adj}
3	1	1	7	18	19	134	472.95	543.92	501.23	488.51	51.17%
1	1	19	111	151	147	180	506.42	553.02	525.03	494.57	47.07%
2	1	7	91	132	128	162	504.61	563.30	528.02	505.83	48.09%
4	1	0	4	8	13	132	476.58	560.03	509.80	507.63	51.41%
5	1	0	0	2	2	114	470.81	566.92	509.03	518.41	52.61%
3	2	1	11	27	33	124	472.74	579.20	515.17	536.77	52.75%
6	1	1	0	0	0	112	472.25	581.24	515.54	537.41	53.20%

Table 7: Statistics for diffusion index factor model (DIF) on *regions384* dataset with various lag specifications. The table contains: Ljung-Box test of autoccorelation in residuals of 1. – 4. order (# of residual series that are correlated on $\alpha = 0.05$), ARCH test of heteroskedasticity (# of residual series that exhibit heteroskedasticity on $\alpha = 0.05$), information criteria and adjusted R^2. BIC* stands for modified BIC.

lags (y_t)	lags (f_t)	Ljung-Box, order 1.	2.	3.	4.	ARCH test	AIC	BIC	BIC*	HQ	R^2_{adj}
3	1	0	0	6	4	32	40.04	48.09	43.29	42.07	63.15%
4	1	0	0	1	2	34	40.29	49.75	44.11	44.07	63.65%
3	2	0	1	5	6	28	37.72	49.81	42.60	45.20	65.50%
5	1	0	0	0	0	29	40.06	50.94	44.45	45.69	64.45%
1	1	5	37	41	41	39	47.26	52.56	49.40	46.19	57.89%
6	1	0	0	0	1	26	40.24	52.57	45.21	47.84	65.06%
4	2	0	1	2	2	29	38.35	51.87	43.81	47.84	65.71%

Table 8: Statistics for diffusion index factor model (DIF) on *states51* dataset with various lag specifications. The table contains: Ljung-Box test of autoccorelation in residuals of 1. – 4. order (# of residual series that are correlated on $\alpha = 0.05$), ARCH test of heteroskedasticity (# of residual series that exhibit heteroskedasticity on $\alpha = 0.05$), information criteria and adjusted R^2. BIC* stands for modified BIC.

5.7 The contagion

In a high-dimensional system of equations where various lags off all considered variables explain the current levels, the impulse-response analysis can be useful. This technique allows us to trace out the effect of an exogenous shock, that occurs in one of the variables, through the entire system. It also makes sense in our spatial framework. For instance, we may consider a rapid and unexpected increase of house prices in a major metropolitan area, which may be caused by a new law or tax policy, and monitor what happens in neighbouring or distant MSAs and how the 'contagion' spreads.

We assume that the mean of y_t variable for $t < 0$ is equal to zero vector and the unit exogenous shock occurs only in i-th MSA, i.e. $u_{i,0} = 1$. Thus we have $y_0 = u_0 = (0, \ldots, 1, \ldots, 0)'$. In addition, we require the further shocks to be equal to zero ($u_t = (0, \ldots, 0)'$ for $t > 0$) to distinguish system changes caused by the initial shock from the

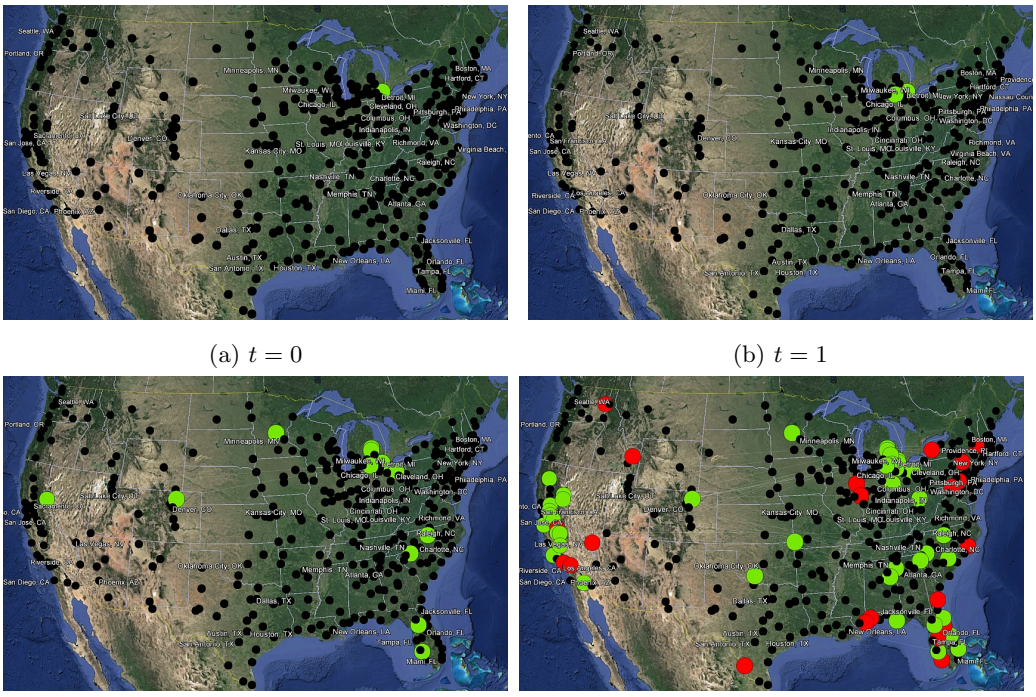
noise. We consider the VAR(1) (15) model without the common factor in x_t because it is not generated by the model. Thus the following results reflect only deviations from the common trend. We have:

$$y_t = A_1 y_{t-1} + u_t$$

and hence

$$y_0 = u_0, \quad y_1 = A_1 y_0, \quad \cdots, \quad y_t = A_1^t y_0$$

We suppose that a positive unit exogenous shock occurs in Detroit, Michigan. In figure 20 response vectors y_t for $t = 0, 1, \ldots, 7$ are visualised. Black dots represent MSAs that are not influenced by the shock at time t. Green and red MSAs experience an increase and decrease respectively. The shock spreads rather quickly through the yellow group region around Michigan and Ohio; house prices mostly increase.

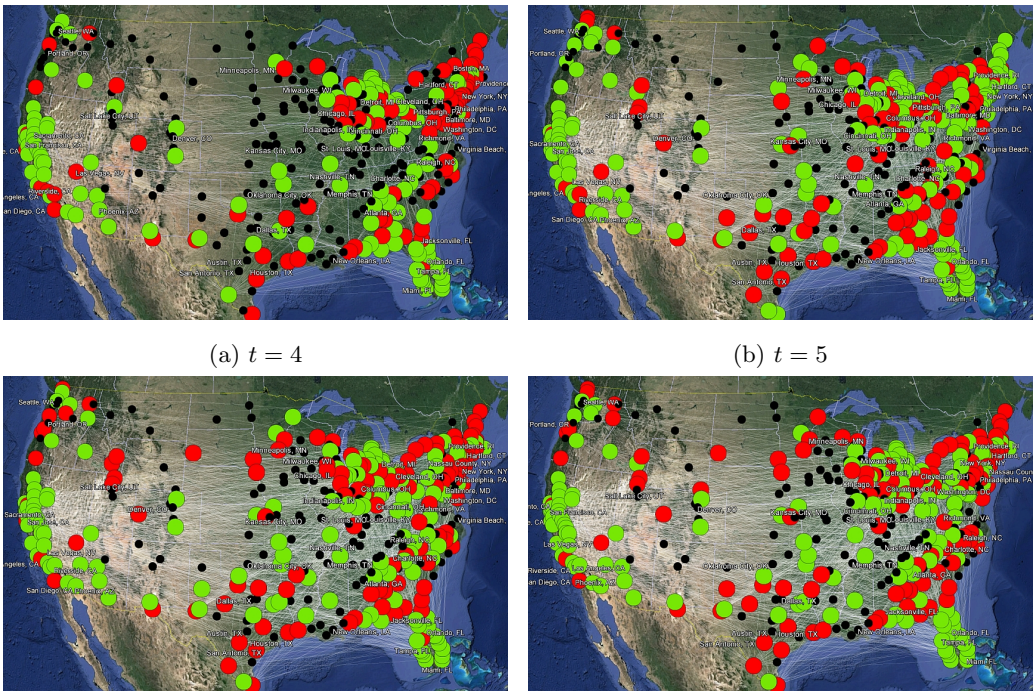


(a) $t = 0$ (b) $t = 1$

(c) $t = 2$ (d) $t = 3$

Figure 20: Impulse-response function visualisation, unit positive shock in Detroit, time $t = 0, \ldots, 3$.

In three or four quarters the shock is transmitted to the south-western coast and to the Florida peninsula, causing growth of house prices as well. Negative relation between the Detroit region and the densely populated north-eastern coast is evident in the figure 21, (c) and (d). After seven quarters the shock is completely transmitted to all relevant metropolitan areas.

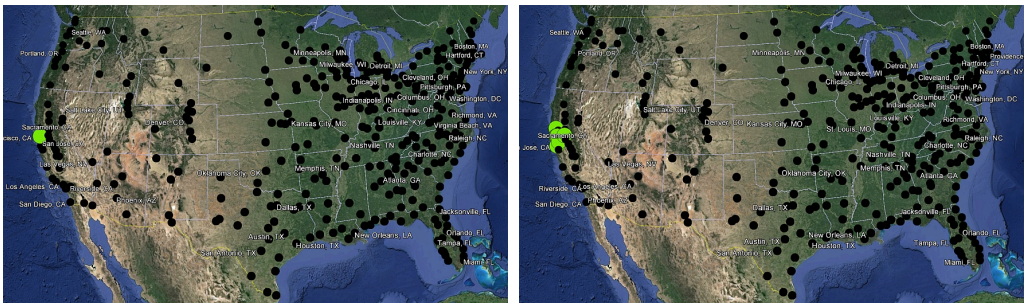


(a) $t = 4$ (b) $t = 5$

(c) $t = 6$ (d) $t = 7$

Figure 21: Impulse-response function visualisation, unit positive shock in Detroit, time $t = 4, \ldots, 7$.

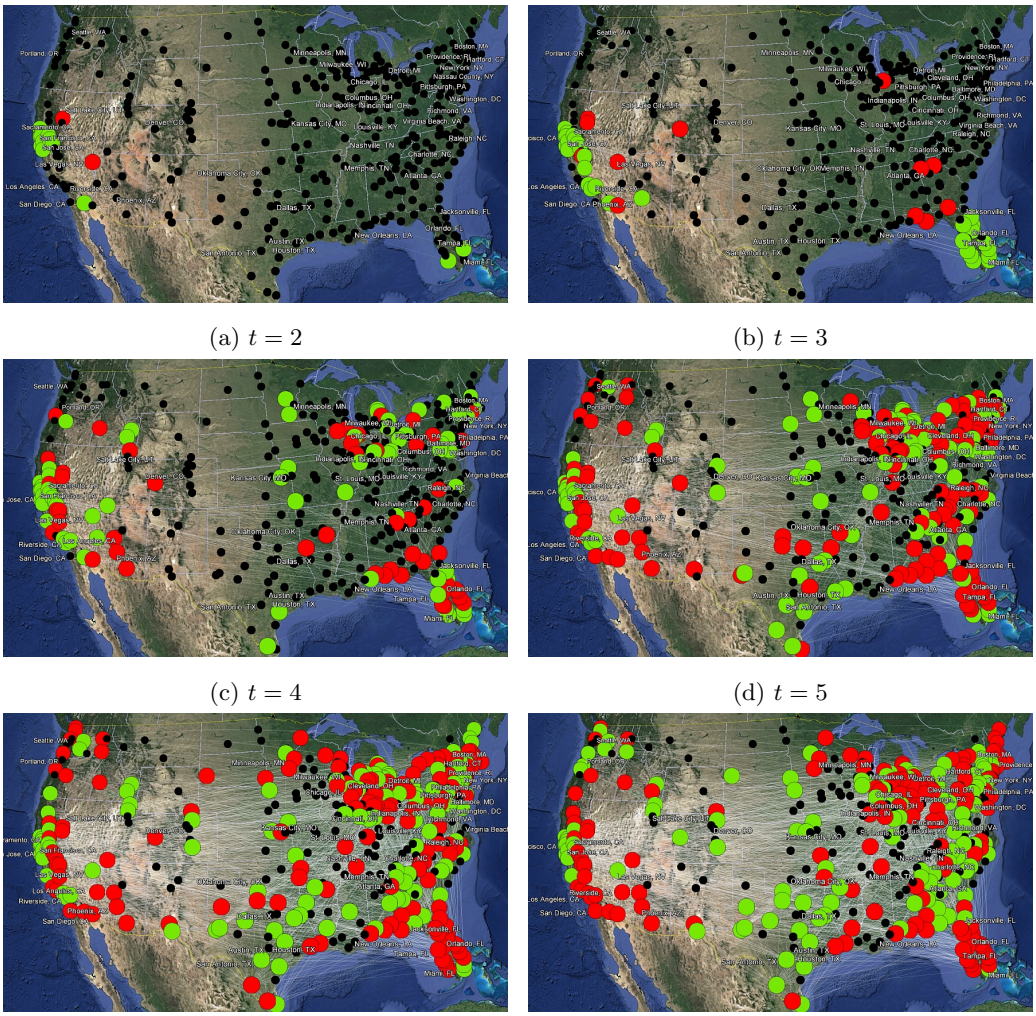
Now suppose that a unit positive shock occurs on the western coast, say in San Francisco (figure 22). Within the first two quarters the shock is transmitted to other MSAs in California, Arizona and Nevada. At the third quarter Florida and the mid-west is hit.



(a) $t = 0$ (b) $t = 1$

Figure 22: Impulse-response function visualisation, unit positive shock in San Francicso, time $t = 0, \ldots, 1$.

In the next quarters, the positive shock is transformed into a drop of house prices in Florida and eventually in California as well. By our VAR model only the short-term relationship is modelled and thus it makes no sense to look further.



(a) $t = 2$ (b) $t = 3$

(c) $t = 4$ (d) $t = 5$

(e) $t = 6$ (f) $t = 7$

Figure 23: Impulse-response function visualisation, unit positive shock in San Francicso, time $t = 0, \ldots, 7$.

5.8 Forecasts

In this section we examine the out-of-sample performance of our model. We compare the forecast accuracy of VAR(1) model (5) estimated by PLS using various penalty functions with the simple benchmark model (4) and the diffusion index factor model (6). We compute the Mean Square Error (MSE) and Mean Absolute Error (MAE) and since our models suffers from heteroskedasticity in residuals, we also report the Mean Square Error

(HMSE) and Mean Absolute Error (HMAE) corrected for heteroskedasticity:

$$MSE = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{h}(\hat{y}_{i,j} - y_{i,j})^2,$$

$$MAE = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{h}|\hat{y}_{i,j} - y_{i,j}|,$$

$$HMSE = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{h}\left(1 - \frac{\hat{y}_{i,j}}{y_{i,j}}\right)^2,$$

$$HMAE = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{h}\left|1 - \frac{\hat{y}_{i,j}}{y_{i,j}}\right|,$$

where $\hat{y}_{i,j}$ is a forecast of house price index at time $T+j$ in region i and similarly, $y_{i,j}$ is the true value of house price index at time $T+j$ in region i. h is the forecast horizon and n stands for number of regions in particular dataset. Apart from the average measures over all equations in VAR model, we also compute the medians over n since the average is undesirably deviated by a couple of extreme outliers.

First, we divide our datasets into out-of-sample and in-sample subset of size $h = 16$ and $T - 16$ respectively and construct the one-quarter ahead forecasts based on rolling regression with fixed sample window $T-16$. For both datasets (*regions384* and *states51*) the out-of-sample period spans from 2009 Q2 to 2013 Q1. The choice of the forecast window size is motivated by the fact that we do not expect overly accurate forecasts during the ambivalent period from 2009 to 2013. In the table 9 the out-of-sample performance of selected models is reported.

According to MSE, VAR(1) estimated by PLS (LASSO and ridge penalty) clearly outperforms the simple benchmark model. All average measures are biased upward by the outliers, especially HMSE. Therefore we consider the median measures to be more suitable for interpretation. Naturally, the DIF model gives the best results, even better than the ridge regression. At this point we have to stress that neither VAR nor DIF model give satisfactory accurate forecasts. The mean absolute error ranges from 2.03 to 3 which indicates that the house price index forecast at time $T+1$ given the information set up to time T is likely to deviate up to 2 units (percentage points) on average from the true value. For comparison, average quarterly change of the house price index in the *regions384* dataset is 0.95.

Measures of predictive accuracy of individual clusters (according to LASSO) are reported in the table 10. The mean HMSE is negatively affected by a single outlier (Lafayette, LA). The fourth group (red) appears to provide the worst forecasts, but when the median measure corrected for heteroskedasticity (HMSE and HMAE) are taken, fourth cluster is suddenly the best. Fitted values usually underestimate the large spikes that are present in the red cluster time series, which results in heteroskedastic residuals.

dataset: *regions384*	mean				median			
2009:Q2 – 2013:Q1	MSE	MAE	HMSE	HMAE	MSE	MAE	HMSE	HMAE
VAR (1,3), LASSO	14.58	2.71	37406.8	7.55	9.41	2.42	19.53	2.54
VAR (1,3), ad. LASSO	17.56	3.00	16765.7	8.02	10.78	2.69	30.84	3.19
VAR (1,3), ridge	13.63	2.62	37050.0	7.63	9.26	2.35	18.14	2.51
DIF (3,1)	7.73	2.03	10815.8	4.73	5.32	1.83	7.01	1.68
DIF (5,1)	8.88	2.14	11361.4	5.30	5.80	1.91	9.52	1.91
DIF (3,3)	8.86	2.19	8965.5	4.58	6.35	2.01	10.48	1.99
benchmark (3), no const.	19.18	2.94	42499.5	8.53	9.81	2.45	20.83	2.69
benchmark (3), const.	20.43	3.03	6668.6	6.54	11.55	2.58	24.45	2.77

Table 9: Forecasting performance of selected models, *regions384* dataset, mean and median of measures over regions. VAR model (5): numbers in brackets stand for VAR order and lag order of exogenous variables. DIF model (6): autoregressive lag order and factor lag order. Benchmark model (4): lag order of predictors.

	mean				median				
group	MSE	MAE	HMSE	HMAE	MSE	MAE	HMSE	HMAE	n
1.	13.17	2.65	92706.9	10.37	10.58	2.53	17.07	2.49	223
2.	17.29	3.22	229.3	3.63	15.05	3.07	26.28	3.02	27
3.	11.89	2.53	158.6	2.93	8.77	2.42	12.59	2.08	69
4.	54.82	5.63	3279.4	6.22	47.53	5.33	11.91	2.07	58

Table 10: Forecasting performance of LASSO, individual clusters, *regions384* dataset. According to 5.3, in figure 12 the first group is marked by blue, second by yellow, third by black and fourth by red color.

As for the *states51* dataset, table 11 contains the results. In general, HPI in states for the 2009 Q2 – 2013 Q1 period is forecasted more accurate than in metropolitan areas. We can observe the same pattern as above: VAR(1) outperforms the benchmark model but DIF model gives the best results.

dataset: *states51*	mean				median			
2009:Q2 – 2013:Q1	MSE	MAE	HMSE	HMAE	MSE	MAE	HMSE	HMAE
VAR (1,3), LASSO	9.09	2.09	299.0	3.45	5.60	1.85	9.21	1.91
VAR (1,3), ad. LASSO	13.28	2.58	9047.2	8.96	9.03	2.36	34.67	3.17
VAR (1,3), ridge	8.49	2.07	210.8	3.12	5.48	1.79	9.25	1.70
DIF (3,1)	6.65	1.87	721.7	3.87	3.90	1.57	10.20	1.83
DIF (5,1)	7.47	1.95	563.1	3.92	4.26	1.64	16.76	2.09
DIF (3,3)	7.22	1.94	3288.0	5.92	4.16	1.64	11.20	1.97
benchmark (2), no const.	11.12	2.27	367.3	3.51	5.81	1.87	9.80	1.92
benchmark (2), const.	13.89	2.49	171.3	3.26	8.92	2.08	12.08	2.15

Table 11: Forecasting performance of selected models, *states51* dataset, mean and median of measures over regions. VAR model (5): numbers in brackets stand for VAR order and lag order of exogenous variables. DIF model (6): autoregressive lag order and factor lag order. Benchmark model (4): lag order of predictors.

The LASSO provides the worst forecast for the states from the most volatile (red) group. They do not have the highest average HMSE but have the highest median HMSE (see table 12).

	mean				median				
group	MSE	MAE	HMSE	HMAE	MSE	MAE	HMSE	HMAE	n
1.	3.95	1.44	19.40	1.70	2.14	1.22	5.31	1.42	28
2.	12.12	2.68	115.90	3.14	13.47	2.87	40.42	2.75	4
3.	5.10	1.70	289.26	3.82	4.38	1.59	12.41	2.19	17

Table 12: Forecasting performance of LASSO, individual clusters, *regions384* dataset. According to 5.3, in figure 12 the first group is marked by blue, second by yellow, third by black and fourth by red color.

We also computed forecasts for the peak period from 2005 Q1 to 2008 Q4 and found out that the models performance ranking remains the same and the forecasts are generally less accurate. In most cases, especially when the peak is high, no model is able to capture the sudden drop in house prices even though the the forecast window is only one quarter wide. Plots of forecasts along with true values and fitted values for selected MSAs and states are reported in the section E of appendix (VAR(1) with LASSO penalty and DIF(3,1) models).

5.9 Model prediction stability

In each step of the rolling regression via PLS, new variables are selected. We may ask whether during the crisis, when majority of HPI series experience more or less pronounced spikes, the outcome of the selection remains stable. In other words, we can test the robustness of our estimation techniques.

	2009:Q2–2013:Q1		2005:Q1–2008:Q4	
	LASSO	adaptive LASSO	LASSO	adaptive LASSO
non-zero → non-zero	8274	17982	7936	13130
non-zero → 0	733	12510	363	9600
0 → non-zero	3612	14205	546	11887
0 → 0	137427	105349	141201	115429
# parameters at $t = T$	9007	30492	8299	22730
% of stable non-zero coefficients	91.9%	59.0%	95.6%	57.8%

Table 13: Stability of the rolling PLS regression coefficients.

Table 13 measures changes in coefficients for two different forecast periods with the first being the standard period defined above and the latter including the point where most of the series reached their peak. In the second quarter of 2009 there were 9007 non-zero LASSO coefficients from which 8274 (91.9 %) remained non-zero also at the end

of the forecast period. On the contrary, a shift of the in-sample subset by h quarters appears to have a huge impact on the weights calculated for adaptive LASSO because almost 59 % non-zero coefficients is at the end of the forecast period set to zero. As for the peak forecast period (2005 Q1 – 2008 Q4), the number of non-zero coefficients remains even more stable for LASSO and less stable for adaptive LASSO. Graphs 24 and 25 demonstrate that even in volatile periods LASSO estimator is stable. Taking this evidence into account, we can claim that the relations among house prices in US regions in our VAR model are rather strong.

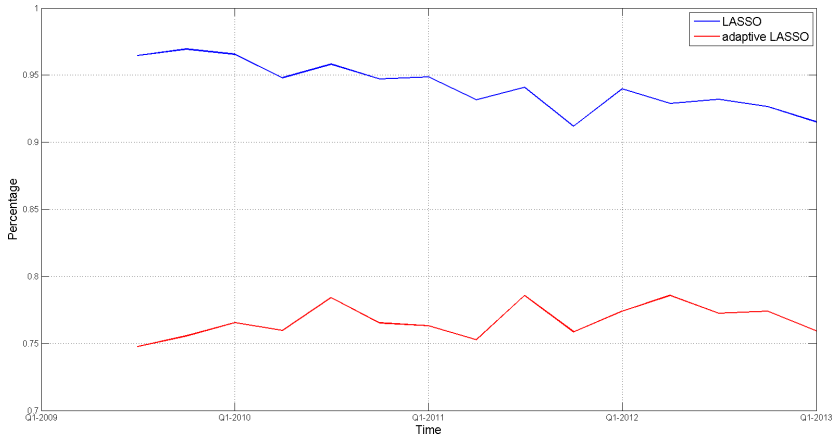


Figure 24: Percentage of coefficients that remain non-zero in each step of the rolling regression estimated by LASSO and adaptive LASSO. Dataset *regions384* and forecast period 2009:Q2–2013:Q1.

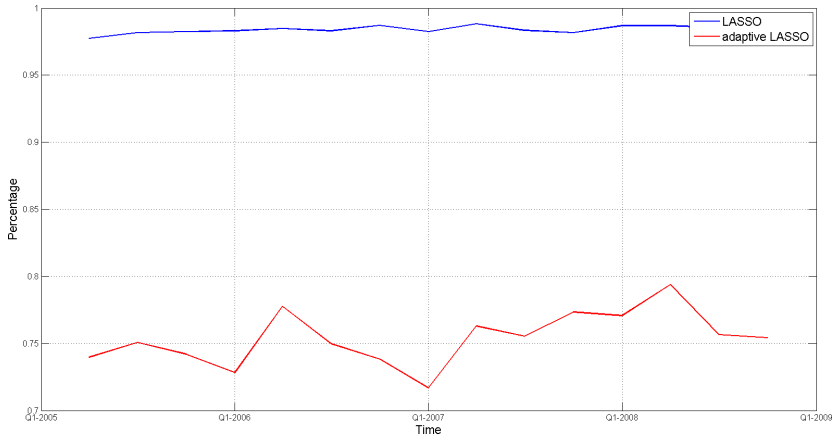


Figure 25: Percentage of coefficients that remain non-zero in each step of the rolling regression estimated by LASSO and adaptive LASSO. Dataset *regions384* and forecast period 2005:Q1–2008:Q4.

6 Conclusion

In this paper we try to gain an insight into the spatial dependencies among the regional house prices in United States. We identify several consistent clusters of MSAs and states with similar dynamics and study their behaviour in the VAR model proposed by Fan, Lv, et al. (2011). We discover that house prices in the traditional bubble regions such as the south western coast (California) and Florida are likely to decrease when prices on the north eastern coast increase. The opposite relation is substantially weaker. Our VAR model is able to fit the house prices in 'bubble' regions very well but gives worse forecasts compared to the diffusion index factor model (DIF). However, none of the models we implemented is able to forecast a rapid drop in prices that overwhelming majority of bubble MSAs experienced at the end of 2007. To be honest, we would not expect any model to do so. The DIF model provides the best forecasts in all respects and the reader may ask why to bother with the high-dimensional VAR model when one can obtain more accurate forecast by OLS. Its true contribution lies in the explicitly modelled dynamic spatial dependencies. Even though the results are somewhat vague since no statistical inference is carried out, they give a good intuition about relations that exist in the system.

Findings from this paper might be useful for spatial economists as they may provide foundations for a serious research. Obviously, we leave many suggestions for the future research. For instance, different techniques of variable selection can be used. Also, we made use of LASSO as an computationally undemanding procedure but we disregarded the SCAD penalty and many others. Next, a higher order VAR model could be considered since some relations may require further lags to be uncovered. It would also be interesting to inspect the accuracy of approximative standard errors. And finally, the last idea that comes into mind is cointegration. Long-term relationships might exist but their identification in high-dimensional setting poses a great challenge for the theoretical framework, that has not been developed yet.

References

Anselin, L. (1980). "Estimation methods for spatial autoregressive structures." In: *Regional Science Dissertation & Monograph Series, Program in Urban and Regional Studies, Cornell University*(8).

Anselin, L. and K. A. Bera (1998). "Spatial dependence in linear regression models with an introduction to spatial econometrics". In: *Statistics textbooks and monographs* 155, pp. 237–290.

Antoniadis, A. and J. Fan (2011). "Regularization of wavelet approximations". In: *Journal of the American Statistical Association*.

Bastian, M., S. Heymann, and M. Jacomy (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: URL: `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

Belloni, A., V. Chernozhukov, and C. Hansen (2011). "Inference for high-dimensional sparse econometric models". In: *arXiv preprint arXiv:1201.0220*.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Calhoun, C. A. (1996). "OFHEO house price indexes: HPI technical description". In: *Office of Federal Housing Enterprise Oversight*.

Case, K. E. and R. J. Shiller (1989). "The Efficiency of the Market for Single-Family Homes". In: *The American Economic Review*, pp. 125–137.

Davidson, R. and J. G. MacKinnon (1999). *Econometric Theory and Methods*. Oxford University Press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). "Least angle regression". In: *The Annals of statistics* 32(2), pp. 407–499.

Fan, J. and R. Li (1999). "Variable Selection via Penalized Likelihood". In: *Department of Statistics Papers*.

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

Fan, J. and R. Li (2005). "Statistical Challenges with High Dimensionality". In: *Korean Statistical Society Journal* 2006, pp. 1–16.

Fan, J., J. Lv, and L. Qi (2011). "Sparse high dimensional models in economics". In: *Annual review of economics* 3, p. 291.

FNC (2010). "FNC Residential Price Index: Research Report". In: *FNC Inc.* Pp. 125–137.

Frank, L. E. and J. H. Friedman (1993). "A statistical view of some chemometrics regression tools". In: *Technometrics* 35(2), pp. 109–135.

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

Gallin, J. (2006). "The long-run relationship between house prices and income: evidence from local housing markets". In: *Real Estate Economics* 34(3), pp. 417–438.

Goetzmann, W. N., L. Peng, and J. Yen (2012). "The subprime crisis and house price appreciation". In: *The Journal of Real Estate Finance and Economics* 44(1-2), pp. 36–66.

GoogleEarth (2013). *United States of America, 39° 38'40.55" N and 96° 04'18.47" W.*

Hackman, R. (2013). "Where can young people buy a house for 500 dollars? Detroit". In: *The Guardian*. http://www.theguardian.com/money/2013/nov/05/young-people-detroit-buy-500-house.

Hastie, T. and J. Qian (2012). "Glmnet Vignette". In:

Huang, J. and H. Xie (2007). "Asymptotic oracle properties of SCAD-penalized least squares estimators". In: *Lecture Notes-Monograph Series*, pp. 149–166.

Iacoviello, M. M. and S. Neri (2008). "Housing market spillovers: evidence from an estimated DSGE model". In: *National Bank of Belgium Working Paper*(145).

Jungbacker, B. and S. J. Koopman (2008). "Likelihood-based analysis for dynamic factor models". In: *Tinbergen Institute Discussion Paper*.

Kissling, W. D. and G. Carl (2008). "Spatial autocorrelation and the selection of simultaneous autoregressive models". In: *Global Ecology and Biogeography* 17(1), pp. 59–71.

Krčál, Adam (2015). "Spatial analysis of regional house prices in United States". MA thesis. Vrije Universiteit Amsterdam.

Leamer, E. E. (2007). "Housing is the business cycle". In: *National Bureau of Economic Research*.

LeSage, J. P. (1999). "The theory and practice of spatial econometrics". In: *University of Toledo. Toledo, Ohio* 28, p. 33.

Li, Y. and D. J. Leatham (2010). "Forecasting Housing Prices: Dynamic Factor Model versus LBVAR Model". In: (103667).

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Nagaraja, C., L. Brown, and S. Wachter (2014). "Repeat sales house price index methodology". In: *Journal of Real Estate Literature* 22(1), pp. 23–46.

Osborne, M., B. Presnell, and A. Turlach (2000). "On the lasso and its dual". In: *Journal of Computational and Graphical statistics* 9(2), pp. 319–337.

Poterba, J. M., D. N. Weil, and R. Shiller (1991). "House price dynamics: The role of tax policy and demography". In: *Brookings Papers on Economic Activity*, pp. 143–203.

Qian, J., T. Hastie, J Friedman, R. Tibshirani, and N. Simon (2013). "Glmnet for Matlab". In: URL: `http://www.stanford.edu/~hastie/glmnet_matlab/`.

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

Stock, J. H. and M. W. Watson (2011). "Dynamic factor models". In: *Oxford Handbook of Economic Forecasting* 1, pp. 35–59.

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

Tikhonov, A. (1963). "Solution of incorrectly formulated problems and the regularization method". In: *Soviet Math. Dokl.* Vol. 5, pp. 1035–1038.

Tsatsaronis, Kevin and Huang Zhu (2004). "What drives housing price dynamics: cross-country evidence". In: *BIS Quarterly Review, March*.

Viton, P. A. (2010). "Notes on spatial econometric models". In: *City and regional planning* 870(03), pp. 9–10.

Wang, H., B. Li, and C. Leng (2009). "Shrinkage tuning parameter selection with a diverging number of parameters". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(3), pp. 671–683.

Wu, T. T. and K. Lange (2008). "Coordinate descent algorithms for lasso penalized regression". In: *The Annals of Applied Statistics*, pp. 224–244.

Zhang, C. (2010). "Nearly unbiased variable selection under minimax concave penalty". In: *The Annals of Statistics*, pp. 894–942.

Zhang, C. and J. Huang (2008). "The sparsity and bias of the lasso selection in high-dimensional linear regression". In: *The Annals of Statistics*, pp. 1567–1594.

Zou, H. (2006). "The adaptive lasso and its oracle properties". In: *Journal of the American statistical association* 101(476), pp. 1418–1429.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), pp. 301–320.

A PLS model selection

Variables	AIC	BIC	modBIC	HQ
USHPI (3)	178.062	196.691	188.512	185.589
USHPI (2)	181.309	198.003	189.290	188.055
Income (4)	181.309	198.003	189.290	188.055
Income (3)	176.357	198.267	189.986	185.210
USHPI (3), Income (4)	174.214	197.656	190.272	183.687
All variables (1)	184.212	215.987	211.691	197.051
All variables (4)	183.795	216.269	212.235	196.917
1st PC (4)	177.177	198.710	191.265	185.878
1st PC (1)	179.938	199.643	191.675	187.901

Table 14: Best subsets according to modified BIC containing up to 2 variables and information criteria for the principal component and for full set of variables. Estimated by LASSO using *metro100* dataset. Numbers in brackets stand for lag order.

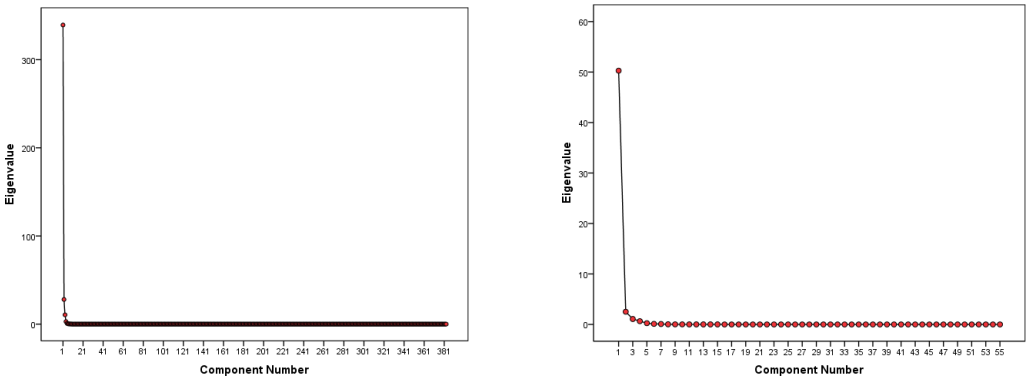
LASSO		Ljung-Box test			ARCH	sparsity	
metro100	1.	2.	3.	4.	test	pen.	total
1st PC (1)	23.2%	23.2%	33.3%	35.4%	34.3%	18.1%	19.1%
1st PC (1–2)	24.2%	27.3%	31.3%	33.3%	38.4%	16.9%	18.9%
1st PC (1–3)	21.2%	25.3%	29.3%	35.4%	37.4%	17.3%	20.2%
1st PC (1–4)	24.2%	25.3%	32.3%	35.4%	40.4%	17.5%	21.4%
All variables (1)	14.1%	18.2%	25.3%	29.3%	30.3%	16.5%	23.1%
All variables (1–2)	24.2%	21.2%	25.3%	24.2%	28.3%	15.3%	27.7%
All variables (1–3)	21.2%	25.3%	24.2%	25.3%	24.2%	13.7%	31.2%
All variables (1–4)	25.3%	26.3%	27.3%	26.3%	16.2%	11.9%	33.9%
HPIUS (1)	23.2%	18.2%	35.4%	37.4%	37.4%	16.4%	17.4%
HPIUS (1–2)	22.2%	22.2%	32.3%	37.4%	28.3%	17.0%	19.0%
HPIUS (1–3)	19.2%	25.3%	33.3%	37.4%	30.3%	16.2%	19.1%
HPIUS (1–4)	18.2%	25.3%	30.3%	39.4%	31.3%	16.5%	20.4%

Table 15: Results for different specifications of matrix X using LASSO estimator and *metro100* dataset. The table contains: Ljung-Box test of autoccorelation in residuals of 1. – 4. order (percentage of total residual series that are correlated on $\alpha = 0.05$), ARCH test of heteroskedasticity (percentage of total residual series that exhibit heteroskedasticity on $\alpha = 0.05$) and percentage of non-zero coefficients (total and penalized). PC stands for principal component and HPIUS for the national level HPI.

LASSO		Ljung-Box test			ARCH	sparsity	
states50	1.	2.	3.	4.	test	pen.	total
1st PC (1)	14.3%	34.7%	71.4%	77.6%	81.6%	11.2%	13.2%
1st PC (1–2)	20.4%	49.0%	73.5%	73.5%	81.6%	10.4%	14.3%
1st PC (1–3)	26.5%	34.7%	49.0%	46.9%	67.3%	9.4%	15.1%
1st PC (1–4)	30.6%	36.7%	49.0%	42.9%	71.4%	8.5%	16.1%
All variables (1)	22.4%	28.6%	63.3%	67.3%	69.4%	8.8%	21.3%
All variables (1–2)	38.8%	34.7%	55.1%	51.0%	61.2%	7.3%	29.5%
All variables (1–3)	59.2%	61.2%	75.5%	71.4%	59.2%	5.3%	35.3%
All variables (1–4)	57.1%	59.2%	79.6%	73.5%	61.2%	5.4%	41.8%
HPIUS (1)	16.3%	51.0%	73.5%	77.6%	79.6%	10.6%	12.6%
HPIUS (1–2)	12.2%	44.9%	69.4%	75.5%	75.5%	10.3%	14.2%
HPIUS (1–3)	14.3%	40.8%	71.4%	73.5%	75.5%	11.1%	16.8%
HPIUS (1–4)	28.6%	49.0%	77.6%	79.6%	75.5%	10.2%	17.8%

Table 16: Results for different specifications of matrix X using LASSO estimator and *states50* dataset. The table contains: Ljung-Box test of autoccorelation in residuals of 1. – 4. order (percentage of total residual series that are correlated on $\alpha = 0.05$), ARCH test of heteroskedasticity (percentage of total residual series that exhibit heteroskedasticity on $\alpha = 0.05$) and percentage of non-zero coefficients (total and penalized). PC stands for principal component and HPIUS for the national level HPI.

B Scree plots



(a) *regions384* (b) *states51*

Figure 26: Scree plots, principal component analysis.

C Cluster allocation

group 1 (blue)																group 2 red)		group 3 (black)				group 4 (yellow)			
1	Abilene, TX	36	Camden, NJ	71	Fargo, ND	106	Jefferson City, MO	141	Morristown, TN	176	Rochester, NY	211	Waco, TX	224	Akron, OH	251	Anderson, SC	286	Lima, OH	320	Bakersfield, CA	355	Prescott, AZ
2	Albany, GA	37	Cape Girardeau, MO	72	Farmington, NM	107	Johnson City, TN	142	Mount Vernon, WA	177	Rockford, IL	212	Warner Robins, GA	225	Anderson, IN	252	Appleton, WI	287	Lincoln, NE	321	Bend, OR	356	Punta Gorda, FL
3	Albany, NY	38	Casper, WY	73	Fayetteville, NC	108	Jonesboro, AR	143	Myrtle Beach, SC	178	Salem, OR	213	Washington, DC	226	Ann Arbor, MI	253	Athens, GA	288	Louisville, KY	322	Boise City, ID	357	Redding, CA
4	Albuquerque, NM	39	Cedar Rapids, IA	74	Fayetteville, AR	109	Joplin, MO	144	Nashville, TN	179	Salisbury, MD	214	Waterloo, IA	227	Battle Creek, MI	254	Atlanta, GA	289	Macon, GA	323	Cape Coral, FL	358	Reno, NV
5	Alexandria, LA	40	Champaign, IL	75	Flagstaff, AZ	110	Kankakee, IL	145	Nassau County, NY	180	Salt Lake City, UT	215	Wenatchee, WA	228	Bay City, MI	255	Barnstable, MA	290	Madison, WI	324	Carson City, NV	359	Riverside, CA
6	Allentown, PA	41	Charleston, WV	76	Florence, SC	111	Kennewick, WA	146	Newark, NJ	181	San Angelo, TX	216	Wheeling, WV	229	Canton, OH	256	Boston, MA	291	Manchester, NH	325	Chico, CA	360	Sacramento, CA
7	Altoona, PA	42	Charleston, SC	77	Florence, AL	112	Killeen, TX	147	New Haven, CT	182	San Antonio, TX	217	Wichita, KS	230	Cleveland, OH	257	Boulder, CO	292	Mankato, MN	326	Crestview, FL	361	St George, UT
8	Amarillo, TX	43	Charlotte, NC	78	Fort Smith, AR	113	Kingsport, TN	148	New Orleans, LA	183	Santa Fe, NM	218	Wichita Falls, TX	231	Dayton, OH	258	Burlington, NC	293	Memphis, TN	327	Deltona, FL	362	Salinas, CA
9	Ames, IA	44	Charlottesville, VA	79	Fort Worth, TX	114	Kingston, NY	149	New York, NY	184	Savannah, GA	219	Williamsport, PA	232	Detroit, MI	259	Cambridge, MA	294	Michigan City, IN	328	El Centro, CA	363	San Diego, CA
10	Anniston, AL	45	Chattanooga, TN	80	Gadsden, AL	115	Knoxville, TN	150	Norwich, CT	185	Scranton, PA	220	Wilmington, DE	233	Flint, MI	260	Chicago, IL	295	Milwaukee, WI	329	Fort Lauderdale, FL	364	San Luis Obispo, CA
11	Asheville, NC	46	Cheyenne, WY	81	Gainesville, FL	116	La Crosse, WI	151	Ocean City, NJ	186	Seattle, WA	221	Wilmington, NC	234	Grand Rapids, MI	261	Cincinnati, OH	296	Minneapolis, MN	330	Fresno, CA	365	Santa Ana, CA
12	Atlantic City, NJ	47	Clarksville, TN	82	Gary, IN	117	Lafayette, LA	152	Odessa, TX	187	Sherman, TX	222	Yakima, WA	235	Greeley, CO	262	Colorado Springs, CO	297	Niles, MI	331	Hagerstown, MD	366	Santa Barbara, CA
13	Auburn, AL	48	Cleveland, TN	83	Glens Falls, NY	118	Lake Charles, LA	153	Ogden, UT	188	Shreveport, LA	223	York, PA	236	Holland, MI	263	Columbus, OH	298	Omaha, NE	332	Hanford, CA	367	Santa Rosa, CA
14	Augusta, GA	49	Coeur dAlene, ID	84	Goldsboro, NC	119	Lancaster, PA	154	Oklahoma City, OK	189	Sioux City, IA			237	Jackson, MI	264	Dalton, GA	299	Oshkosh, WI	333	Lake Havasu City, AZ	368	Sebastian, FL
15	Austin, TX	50	College Station, TX	85	Grand Forks, ND	120	Laredo, TX	155	Olympia, WA	190	Sioux Falls, SD			238	Kalamazoo, MI	265	Danville, IL	300	Peabody, MA	334	Lakeland, FL	369	Stockton, CA
16	Baltimore, MD	51	Columbia, MO	86	Grand Junction, CO	121	Las Cruces, NM	156	Owensboro, KY	191	Spokane, WA			239	Kokomo, IN	266	Denver, CO	301	Pueblo, CO	335	Las Vegas, NV	370	Tampa, FL
17	Bangor, ME	52	Columbia, SC	87	Great Falls, MT	122	Lawton, OK	157	Parkersburg, WV	192	Springfield, IL			240	Lansing, MI	267	Des Moines, IA	302	Rochester, MN	336	Los Angeles, CA	371	Tucson, AZ
18	Baton Rouge, LA	53	Columbus, GA	88	Greenville, NC	123	Lebanon, PA	158	Pascagoula, MS	193	Springfield, MA			241	Mansfield, OH	268	Duluth, MN	303	Rockingham County, NH	337	Madera, CA	372	Vallejo, CA
19	Beaumont, TX	54	Columbus, IN	89	Greenville, SC	124	Lewiston, ID	159	Peoria, IL	194	Springfield, MO			242	Monroe, MI	269	Eau Claire, WI	304	Rocky Mount, NC	338	Medford, OR	373	Visalia, CA
20	Bellingham, WA	55	Corpus Christi, TX	90	Gulfport, MS	125	Lewiston, ME	160	Philadelphia, PA	195	State College, PA			243	Muncie, IN	270	Elkhart, IN	305	Rome, GA	339	Merced, CA	374	West Palm Beach, FL
21	Bethesda, MD	56	Corvallis, OR	91	Harrisburg, PA	126	Lexington, KY	161	Pine Bluff, AR	196	Sumter, SC			244	Muskegon, MI	271	Evansville, IN	306	St Cloud, MN	340	Miami, FL	375	Winchester, VA
22	Billings, MT	57	Dallas, TX	92	Harrisonburg, VA	127	Little Rock, AR	162	Pittsburgh, PA	197	Syracuse, NY			245	Saginaw, MI	272	Fond du Lac, WI	307	St Joseph, MO	341	Modesto, CA	376	Yuba City, CA
23	Binghamton, NY	58	Danville, VA	93	Hartford, CT	128	Logan, UT	163	Pittsfield, MA	198	Tacoma, WA			246	Sandusky, OH	273	Fort Collins, CO	308	St Louis, MO	342	Napa, CA	377	Yuma, AZ
24	Birmingham, AL	59	Davenport, IA	94	Hattiesburg, MS	129	Longview, TX	164	Pocatello, ID	199	Tallahassee, FL			247	Springfield, OH	274	Fort Wayne, IN	309	San Francisco, CA	343	Naples, FL		
25	Bismarck, ND	60	Decatur, IL	95	Hot Springs, AR	130	Longview, WA	165	Portland, ME	200	Texarkana, TX			248	Toledo, OH	275	Gainesville, GA	310	San Jose, CA	344	North Port, FL		
26	Blacksburg, VA	61	Dothan, AL	96	Houma, LA	131	Lubbock, TX	166	Portland, OR	201	Topeka, KS			249	Warren, MI	276	Green Bay, WI	311	Santa Cruz, CA	345	Oakland, CA		
27	Bloomington, IN	62	Dover, DE	97	Houston, TX	132	Lynchburg, VA	167	Poughkeepsie, NY	202	Trenton, NJ			250	Youngstown, OH	277	Greensboro, NC	312	Sheboygan, WI	346	Ocala, FL		
28	Bloomington, IL	63	Dubuque, IA	98	Huntington, WV	133	Manhattan, KS	168	Providence, RI	203	Tulsa, OK					278	Hickory, NC	313	South Bend, IN	347	Orlando, FL		
29	Bowling Green, KY	64	Durham, NC	99	Huntsville, AL	134	McAllen, TX	169	Provo, UT	204	Tuscaloosa, AL					279	Indianapolis, IN	314	Spartanburg, SC	348	Oxnard, CA		
30	Bremerton, WA	65	Edison, NJ	100	Idaho Falls, ID	135	Midland, TX	170	Racine, WI	205	Tyler, TX					280	Jackson, TN	315	Weirton, WV	349	Palm Bay, FL		
31	Bridgeport, CT	66	Elizabethtown, KY	101	Iowa City, IA	136	Missoula, MT	171	Raleigh, NC	206	Utica, NY					281	Janesville, WI	316	Terre Haute, IN	350	Palm Coast, FL		
32	Brownsville, TX	67	Elmira, NY	102	Ithaca, NY	137	Mobile, AL	172	Rapid City, SD	207	Valdosta, GA					282	Kansas City, MO	317	Wausau, WI	351	Panama City, FL		
33	Brunswick, GA	68	El Paso, TX	103	Jackson, MS	138	Monroe, LA	173	Reading, PA	208	Victoria, TX					283	Lafayette, IN	318	Winton, NC	352	Pensacola, FL		
34	Buffalo, NY	69	Erie, PA	104	Jacksonville, FL	139	Montgomery, AL	174	Richmond, VA	209	Vineland, NJ					284	Lake County, IL	319	Worcester, MA	353	Phoenix, AZ		
35	Burlington, VT	70	Eugene, OR	105	Jacksonville, NC	140	Morgantown, WV	175	Roanoke, VA	210	Virginia Beach, VA					285	Lawrence, KS			354	Port St Lucie, FL		

Table 17: List of regions and their group arrangement based on cluster analysis.

D Network graphs

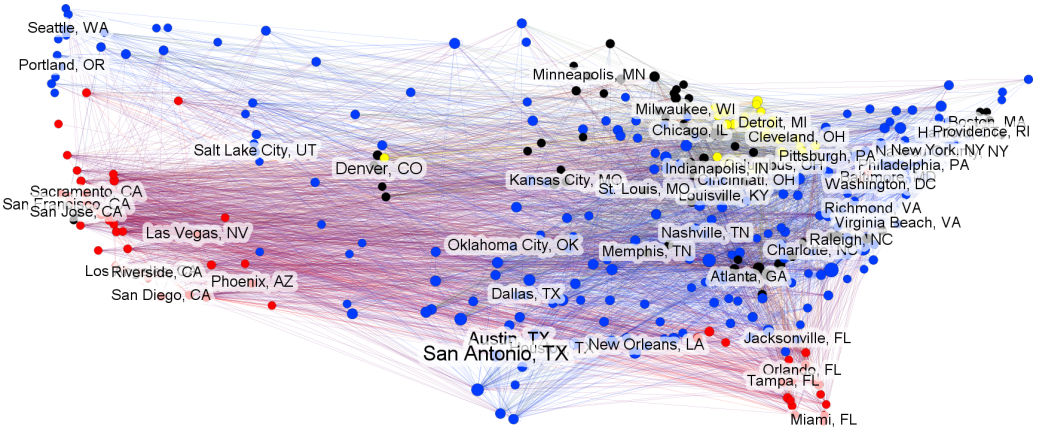


Figure 27: Network graph of VAR(1) model by Gephi software, *regions384* dataset. Layout of the particular node corresponds to its geo-coordinates.

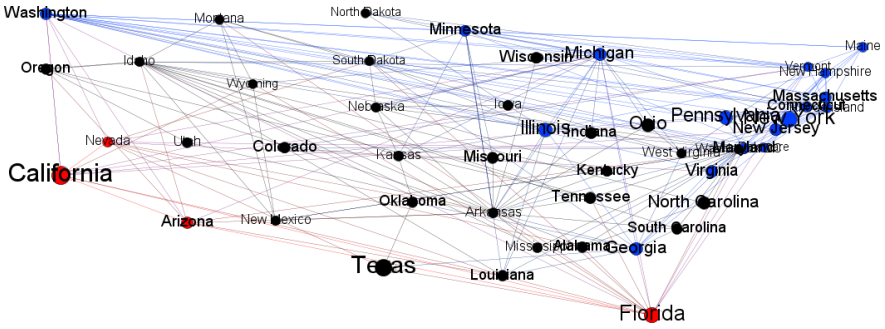


Figure 28: Network graph of VAR(1) model by Gephi software, *states51* dataset. Layout of the particular node corresponds to its geo-coordinates.

E Fitted values and forecasts for selected MSAs

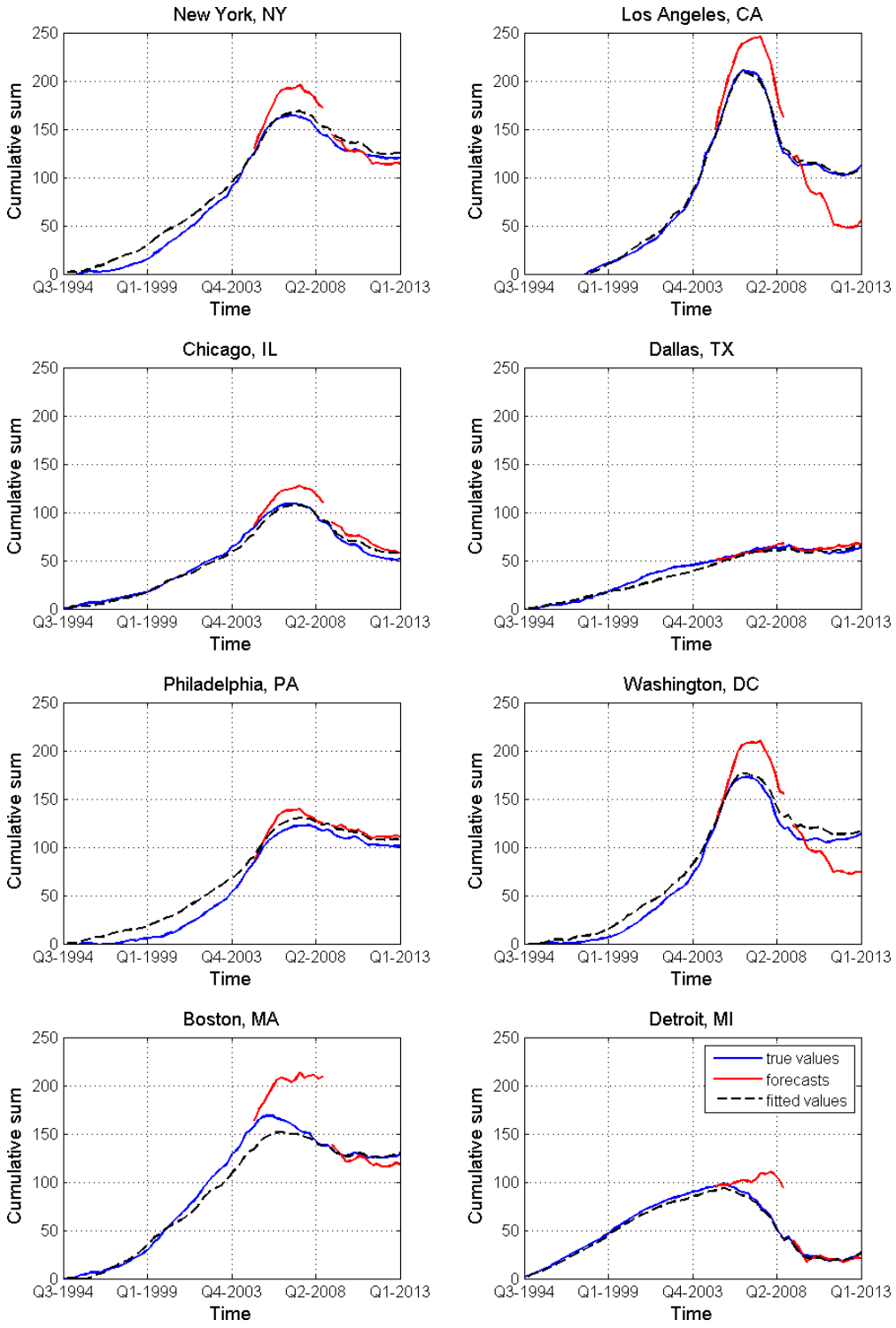


Figure 29: Fitted values and forecasts for selected MSAs, VAR(1) estimated by LASSO.

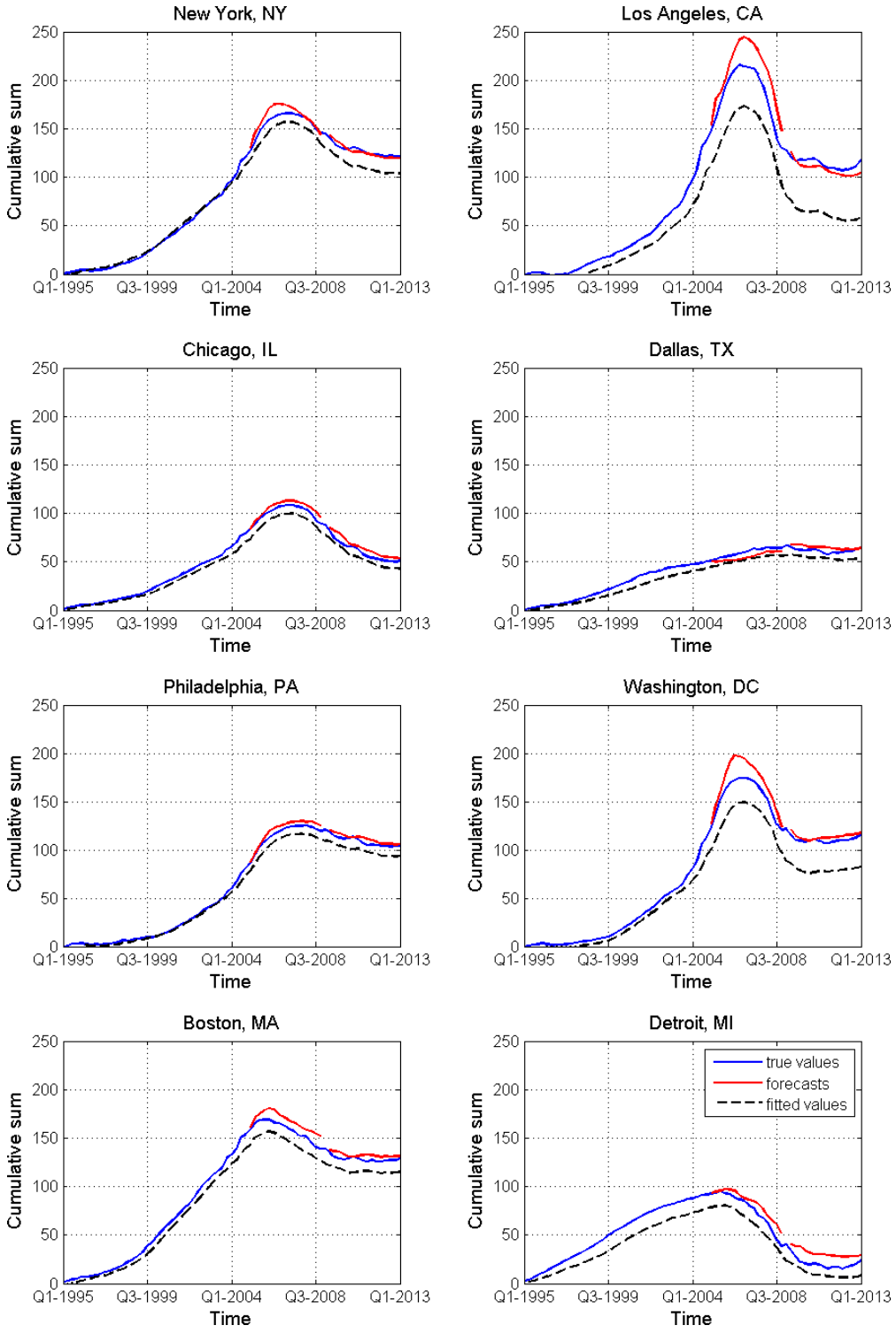


Figure 30: Fitted values and forecasts for selected MSAs, DIF(3,1) model.

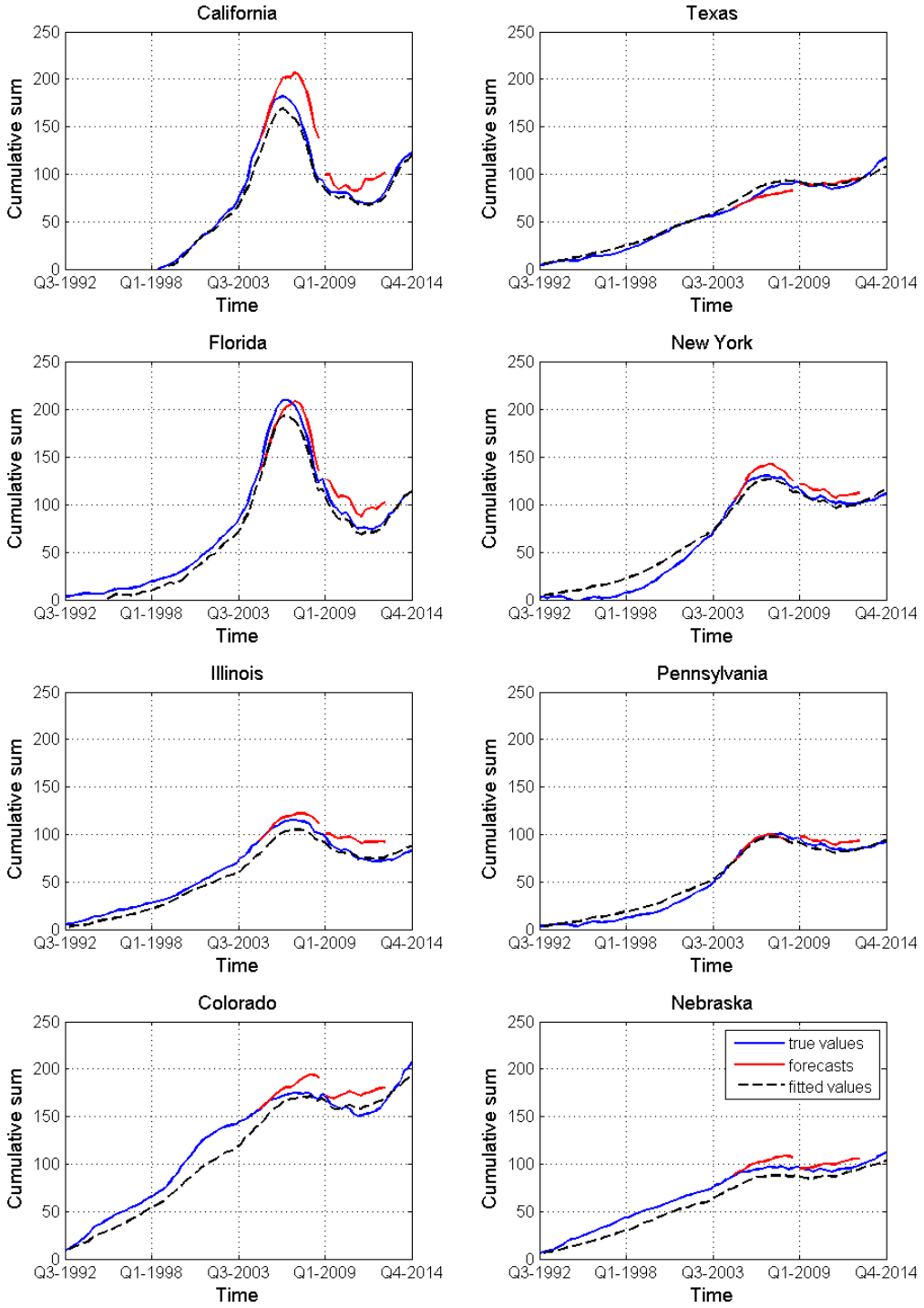


Figure 31: Fitted values and forecasts for selected states, VAR(1) estimated by LASSO.

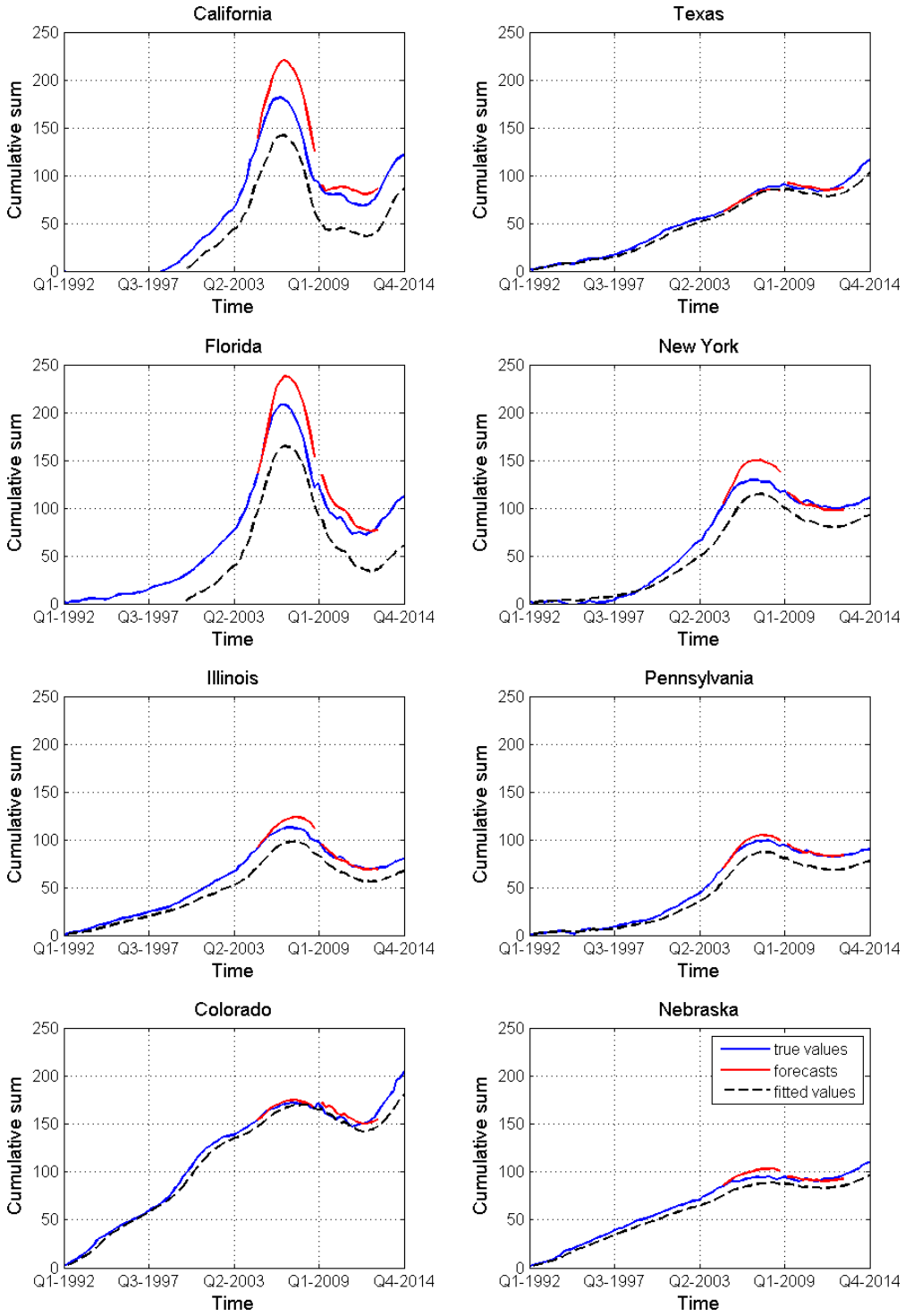


Figure 32: Fitted values and forecasts for selected states, DIF(3,1) model.

F Matlab codes

In the following script (**pls_lasso2.m**) the PLS estimation and some supplementary procedures are carried out. As an input dataset, VAR order, PLS type, lag order of exogenous predictors need to be specified. Then the estimation itself is performed, goodness-of-fit criteria are computed, various heteroskedasticity and autocorrelation tests are performed both for individual clusters and overall. Data are loaded automatically from a user-defined path.

```
%% dataset
addpath('\codes\glmnet_matlab')
addpath('\codes')
dataset = 3;  % 1 for metropolitan areas, 2 for states, 3 for regions ...
    base indices
switch dataset
    case 1
        load('data_metro.mat');
    case 2
        load('data_states.mat');
    case 3
        load('data_regions.mat');
end

%% parameters
predictorsLag = 3;
p = 1; %VAR order

% pred = component;
pred = predictors;

%% differenced data/exogenous predictors
nPredictors = size(pred,2)*predictorsLag;

if dataset > 2
   beg = 33;
else
   beg = 21;
end
dataDif = diff(dataAdjusted,1,1);
predictorsDif = diff(pred,1,1);

%% constructing X
[n, ¬] = size(dataDif);
m1 = size(region,1);
```

```
X = [];
variables = [];

% p lags of HPI in other regions
for i = 1:p
    X = [X dataDif(p-i+1:end-i,:) ];
    variables = [variables; region];
end
%  lags of exogenous predictors
for i = 1:predictorsLag
    X = [X predictorsDif(beg+p-i:end-i,:) ];
    variables = [variables; transpose(names)];
end

[¬, m] = size(X);
results = cell(250,m1);
resid = zeros(n-p,m1);
fitted = zeros(n-p,m1);
significance = zeros(2,m1);
dependence = zeros(1,m1);
significanceRot = zeros(2,m1);
yStored = zeros(n-p,m1);
betaStored = zeros(m1,m);
penalty = zeros(m1,m);
%% equal weighted PLS lasso
options = glmnetSet();
options.intr = false;
options.standardize = true;
options.thresh = 1e-8;
options.nlambda = 100;
type = 1;     %1    'LASSO'
              %2    'adaptiveLASSO'
              %3    'ridge'
switch type
    case 1 % 'LASSO'
    options.alpha = 1; % 1 for lasso, 0 for ridge
    case 2 % 'adaptiveLASSO'
    options.alpha = 1; % 1 for lasso, 0 for ridge
    case 3 % 'ridge'
    options.alpha = 0; % 1 for lasso, 0 for ridge
end

% penalty factor - exogenous variables are not penalized, penalty ...
    (weight) = 0
% penalty factor - HPI in other regions is penalized, penalty (weight) = 1
options.penalty_factor = ones(1,p*m1);
if nPredictors > 0;
   options.penalty_factor = [options.penalty_factor zeros(1,nPredictors)];
end
tic
```

```
% PLS for each region*******************************************
for i = 1:m1
    i
    y = dataDif(1+p:end,i);

    % weights for adaptive lasso************************
    % computed from ridge estimator as stated in the paper
    if type == 2 % 'adaptiveLASSO'
        options.alpha = 0;
        ridgeFit = glmnet(X,y,[],options);
        [¬,minBICindex,¬] = BICselection(ridgeFit,ridgeFit.beta,X,y); ...
            %best model wrt lambda according to modified
                                                                    BIC
        pen = transpose(1./abs(ridgeFit.beta(1:p*m1,minBICindex)));
        if nPredictors > 0;
                pen = [pen zeros(1,nPredictors)];
        end
        options.penalty_factor = pen;

        options.alpha = 1;
    end
    % ************************************************
    penalty(i,:) = options.penalty_factor;
    fit = glmnet(X,y,[],options);  % an instance of glmnet object is ...
        created - a structure

    % BIC for tuning parameter selection - from the sequence of results ...
        (for each lambda) the one with smallest
      modified BIC is chosen
    [¬,minBICindex,¬] = BICselection(fit,fit.beta,X,y);
    beta = fit.beta(:,minBICindex);
    k = find(beta);
    for r = 1:size(k,1)
        results(r,i) = variables(k(r),1);
    end

    resid(:,i) = y - X*beta;
    fitted(:,i) = X*beta;
    betaStored(i,:) = transpose(beta);
    yStored(:,i) = y;

end
sparsity = [size(find(betaStored(:,1:m1)),1) size(find(betaStored),1)];

%% information criteria etc.
criteria = zeros(1,4);
BICs = zeros(m1,1);
modBICs = zeros(m1,1);
AICs = zeros(m1,1);
HQs = zeros(m1,1);
```

```
SSE = zeros(m1,1);
ds = zeros(m1,1);
SST = zeros(m1,1);
R2 = zeros(m1,1);
R2adj = zeros(m1,1);

for i = 1:m1
        d = size(find(betaStored(i,:)),2);
        C = log(log(d));
        if d ≤ 1
           C = 0;
        end
        SSE(i) = sum(resid(:,i).^2); % sum of squares
        SST(i) = sum((yStored(:,i) - mean(yStored(:,i))).^2);
        R2(i) = 1 - SSE(i)/SST(i);
        R2adj(i) = R2(i) - (1 - R2(i))*(d/(n-d-1));
        BICs(i) = log(SSE(i)/n) + d*(log(n)/n);
        modBICs(i) = log(SSE(i)/n) + d*(log(n)/n)*C;
        AICs(i) = log(SSE(i)/n) + d*(2/n);
        HQs(i) = log(SSE(i)/n) + d*(2*log(log(n))/n);
        ds(i,1) = d;
end

criteria(1,1) = sum(AICs);
criteria(1,2) = sum(BICs);
criteria(1,3) = sum(modBICs);
criteria(1,4) = sum(HQs);
toc

%% adjacency matrices to construct network graphs
A = adjacencyMatrix(betaStored);

figure (3)
gplot(A(:,:,1),[longitude latitude]);

%% autocorrelation + heteroskedasticity tests
% h = 1 indicates rejection of the no residual autocorrelation null
% hypothesis in favor of the alternative. (AUTOCORRELATION PRESENT)
% h = 0 indicates failure to reject the no residual autocorrelation null
% hypothesis. (AUTOCORRELATION NOT PRESENT)
lags = 4;

correl = zeros(m1,lags+1);
heterosked = zeros(m1,1);
for i = 1:m1
[correl(i,1:lags),¬,¬,¬] = lbqtest(resid(:,i),'Lags',1:lags);
heterosked(i) = archtest(resid(:,i));
end
correl(:,lags+1) = sum(correl(:,1:lags),2);
LBtestResNW = sum(correl,1);
```

```
heteroskedResults  = sum(heterosked,1);

% heterosked = TestHet(residNW(:,1), X, '-BPK');
SPEC = [LBtestResNW(:,1:lags) heteroskedResults sparsity];


%% matrix rotation - according the cluster membership of every region (row)
rowSort = [3 1];

adjPlus = adjacencyMatrixPlus(betaStored);
adjMinus = adjacencyMatrixMinus(betaStored);
groupFrequency = tabulate(group);

adjPlusRot = [transpose(1:1:m1) population  group adjPlus(:,1:m1)];
adjPlusRot = sortrows(adjPlusRot,rowSort);
index = adjPlusRot(:,1:3);
adjPlusRot(:,1:3) = [];
adjPlusRot = [(1:1:m1); transpose([population group]); adjPlusRot];
adjPlusRot = transpose(adjPlusRot);
adjPlusRot = sortrows(adjPlusRot,rowSort);
adjPlusRot(:,1:3) = [];
adjPlusRot = transpose(adjPlusRot);

adjMinusRot = [transpose(1:1:m1) population  group adjMinus(:,1:m1)];
adjMinusRot = sortrows(adjMinusRot,rowSort);
adjMinusRot(:,1:3) = [];
adjMinusRot = [(1:1:m1); transpose([population group]); adjMinusRot];
adjMinusRot = transpose(adjMinusRot);
adjMinusRot = sortrows(adjMinusRot,rowSort);
adjMinusRot(:,1:3) = [];
adjMinusRot = transpose(adjMinusRot);

spy(adjPlusRot,'-.b',8);
hold on
spy(adjMinusRot,'-.r',8);
hold off

% lines of boundaries between pair of clusters added
color = [54/255 54/255 54/255];
cumul = 0;
for i= 1:size(groupFrequency,1)-1
cumul = cumul + groupFrequency(i,2);
line('XData', [cumul cumul], 'YData', [0 m1+1], 'LineStyle', ...
    '--','LineWidth', 1.5, 'Color','black');

line('XData', [0 m1+1], 'YData', [cumul cumul], 'LineStyle', ...
    '--','LineWidth', 1.5, 'Color','black');
end

%% correlation matrix
```

```
229 R = corrcoef(dataAdjusted);
230 % rotation
231 Rnew = [transpose(1:1:m1) group R];
232 Rnew = sortrows(Rnew,[2 1]);
233 Rnew(:,1:2) = [];
234 Rnew = [(1:1:m1); transpose(group); Rnew];
235 Rnew = transpose(Rnew);
236 Rnew = sortrows(Rnew,[2 1]);
237 Rnew(:,1:2) = [];
238 Rnew = transpose(Rnew);
239
240 % full network - creates a network assuming every region is connected to ...
        all the remaining
241 A = ones(m1,m1);
242 A = A - eye(m1);
243 netwFull = createNetwork(A,region);
244 netWeights = reshape(R,m1*m1,1);
245 for i = 1:size(netWeights,1)
246     if netWeights(i,1) == 1
247        netWeights(i,1) = NaN;
248     end
249 end
250
251 %% network according to model
252
253 networkModel = createNetwork(adjacencyMatrix(betaStored(:,1:m1)),region);
254 networkModelPlus = createNetwork(adjPlus,region);
255 networkModelMinus = createNetwork(adjMinus,region);
256
257 %% group characteristics
258 charsUnsorted = [num2cell(transpose(1:1:m1)) num2cell(group) ...
        num2cell(population) region num2cell(AICs) num2cell(BICs) ...
        num2cell(modBICs) num2cell(HQs) num2cell(correl(:,1:lags)) ...
        num2cell(heterosked) num2cell(R2) num2cell(R2adj)];
259 charsSorted = sortrows(charsUnsorted,[2 1]);
260 charsSorted(:,1:4) = [];
261
262 groupStat = zeros(size(groupFrequency,1),size(charsSorted,2));
263 cumul = 0;
264 for i = 1:size(groupFrequency,1)
265
266     groupStat(i,:) = sum(cell2mat(charsSorted(cumul+1:cumul + ...
            groupFrequency(i,2),:)),1);
267     cumul = cumul + groupFrequency(i,2);
268
269 end
270
271 groupStat = [groupStat groupFrequency(:,2)];
```

In the next script (**IRS.m**) the impulse response functions are computed and consequently transfomed into networks that monitor where the shock came from at time t, who was affected by the shock and whether the shock was positive or negative. Several *.csv* files are generated as an input to Gephi that creates a network graph (see section (5.7)).

```
%% impulse response analysis for p = 1
% pls_lasso2.m must be executed before

h = 20;
yt = zeros(m1,h+1);
% shock in which region/variable
shockVarIndex = 308;
yt(shockVarIndex,1) = 1; %yo


% IR function for horizon h and p = 1
   for i = 1:h
       yt(:,i+1) = betaStored(:,1:m1)*yt(:,i);

   end
yCumul = cumsum(yt,2);

% adjacency matrix
betaT = zeros(m1,m1,h);
aMatrix = zeros(m1,m1,h);
netw = cell(2000,2*h);

nod = cell(m1+1,9,h);

nod(1,7,:) = cellstr('state');
nod(1,8,:) = cellstr('color');
nod(1,9,:) = cellstr('size');

for j = 1:h
nod(:,1:6,j) = regionStat;
netw(1,2*j-1) = cellstr('source');
netw(1,2*j) = cellstr('target');

k = find(yt(:,j));
    for l = 1:size(k)
        betaT(:,k(l),j) = betaStored(:,k(l),1);
    end

    aMatrix(:,:,j) = adjacencyMatrix(betaT(:,:,j));
    q = sum(sum(aMatrix(:,:,j)));
    netw(2:q+1,2*j-1:2*j) = createNetwork(aMatrix(:,:,j), region);

    for i = 1:m1
```

```
44          if yCumul(i,j) < 0
45               gr = 'negative';
46               cl = '#ff0000';
47               st = 1;
48          elseif  yCumul(i,j) == 0
49               gr = 'not affected';
50               cl = '#000000';
51               st = 0;
52          elseif  yCumul(i,j) > 0
53               gr = 'positive';
54               cl = '#77e805';
55               st = 1;
56          else
57               gr = '-';
58          end
59          nod(i+1,7,j) = cellstr(gr);
60          nod(i+1,8,j) = cellstr(cl);
61          nod(i+1,9,j) = num2cell(st);
62
63      end
64      % saving results into csv file (using non-standard cell2csv function)
65      switch dataset
66       case 1
67           cell2csv(strcat('C:\Adam\VU\Thesis\codes\gephi_data\IRS\metro\net',
68           num2str(j+1),'.csv'), ...
                 netw(1:1+sum(sum(aMatrix(:,:,j))),2*j-1:2*j), ';', 2013, '.');
69           cell2csv(strcat('C:\Adam\VU\Thesis\codes\gephi_data\IRS\metro\nod',
70           num2str(j),'.csv'), nod(:,:,j), ';', 2013, '.');
71       case 2
72           cell2csv(strcat('C:\Adam\VU\Thesis\codes\gephi_data\IRS\states\net',
73           num2str(j+1),'.csv'), ...
                 netw(1:1+sum(sum(aMatrix(:,:,j))),2*j-1:2*j), ';', 2013, '.');
74           cell2csv(strcat('C:\Adam\VU\Thesis\codes\gephi_data\IRS\states\nod',
75           num2str(j),'.csv'), nod(:,:,j), ';', 2013, '.');
76       case 3
77           cell2csv(strcat('C:\Adam\VU\Thesis\codes\gephi_data\IRS\regions\net',
78           num2str(j+1),'.csv'), ...
                 netw(1:1+sum(sum(aMatrix(:,:,j))),2*j-1:2*j), ';', 2013, '.');
79           cell2csv(strcat('C:\Adam\VU\Thesis\codes\gephi_data\IRS\regions\nod',
80           num2str(j),'.csv'), nod(:,:,j), ';', 2013, '.');
81      end
82
83  end
84
85  initial = cell(1+m1,2);
86  initial(1,1) = cellstr('source');
87  initial(1,2) = cellstr('target');
88
89  switch dataset
90      case 1
```

```
91          cell2csv('C:\Adam\VU\Thesis\codes\gephi_data\IRS\metro\net1.csv', ...
                initial, ';', 2013, '.');
92      case 2
93          cell2csv('C:\Adam\VU\Thesis\codes\gephi_data\IRS\states\net1.csv', ...
                initial, ';', 2013, '.');
94      case 3
95          cell2csv('C:\Adam\VU\Thesis\codes\gephi_data\IRS\regions\net1.csv', ...
                initial, ';', 2013, '.');
96  end
```

Last but not least, forecasts are carried out by the **predictions.m** script.

```
1   %% predictions
2   addpath('C:\Adam\VU\Thesis\codes\glmnet_matlab')
3
4   model = 1;  % 1 for PLS;                run pls_lasso2.m first!
5               % 2 for DFM;                run dfm2.m first!
6               % 3 for benchmark model     run olsX2.m first!
7
8   h = 16; % lenght of the out of sample period
9   % 1 means 1991Q1 for metropolitan areas dataset
10  % 1 means 1994Q1 for regions dataset
11
12  [n, m] = size(X);
13  ind = 0;
14  s = n - h - ind;
15
16  forecast = zeros(h,m1);
17  %% forecasts
18  betaPredict = zeros(m1,m,h-1);
19  sparsity = zeros(1,h);
20
21  switch model
22      case 1 %PLS ------------------------------------------------------------
23          % glmnet options
24          options = glmnetSet();
25          options.intr = false;
26          options.standardize = true;
27          options.thresh = 1e-8;
28          options.nlambda = 100;
29
30          % penalty factor
31          options.penalty_factor = ones(1,p*m1);
32          options.penalty_factor = [options.penalty_factor ...
                zeros(1,nPredictors)];
33          switch type
34              case 1 % 'LASSO'
35              options.alpha = 1; % 1 for lasso, 0 for ridge
36
```

```
37              case 2 % 'adaptiveLASSO'
38              options.alpha = 1; % 1 for lasso, 0 for ridge
39
40              case 3 % 'ridge'
41              options.alpha = 0; % 1 for lasso, 0 for ridge
42          end
43
44          for i = 1:h
45             i
46             XPr = X(i:s+i-1,:);
47             for j = 1:m1
48                 yPr = yStored(i:s+i-1,j);
49                     % weights for adaptive lasso************************
50                      if type == 2
51                          options.alpha = 0;
52                          ridgeFit = glmnet(XPr,yPr,[],options);
53                          [¬,minBICindex,¬] = ...
                                BICselection(ridgeFit,ridgeFit.beta,XPr,yPr);
54                          pen = transpose(1./abs(ridgeFit.beta(1:p*m1,
55                          minBICindex)));
56                          if nPredictors > 0;
57                                  pen = [pen zeros(1,nPredictors)];
58                          end
59                          options.penalty_factor = pen;
60                          options.alpha = 1;
61                      end
62                   % ************************************************
63                 fit = glmnet(XPr,yPr,[],options);
64
65                 % BIC selection
66                 [¬,minBICindex] = BICselection(fit,fit.beta,XPr,yPr);
67                 beta = fit.beta(:,minBICindex);
68                 betaPredict(j,:,i) = transpose(beta);
69             end
70
71             forecast(i,:) = X(s+i,:)*betaPredict(:,:,i)';
72             sparsity(i) = size(find(betaPredict(:,:,i)),1);
73          end
74
75      case 2 % DFM ...
            ------------------------------------------------------------
76          for i = 1:h
77             i
78             for j = 1:m1
79                 yPr = yStored(i:s+i-1,j);
80                 XPr = XStored(i:s+i-1,:,j);
81
82                 fit = fitlm(XPr,yPr,'Intercept',false);
83
84                 beta = transpose(fit.Coefficients.Estimate);
```

```
85                  betaPredict(j,:,i) = transpose(beta);
86                  forecast(i,j) = XStored(s+i,:,j)*betaPredict(j,:,i)';
87              end
88
89          end
90      case 3 % benchmark ...
            ------------------------------------------------------
91          for i = 1:h
92              i
93              XPr = X(i:s+i-1,:);
94              for j = 1:m1
95                  yPr = yStored(i:s+i-1,j);
96
97                  fit = fitlm(XPr,yPr,'Intercept',false);
98
99                  beta = transpose(fit.Coefficients.Estimate);
100                 betaPredict(j,:,i) = transpose(beta);
101             end
102             forecast(i,:) = X(s+i,:)*betaPredict(:,:,i)';
103         end
104
105 end
```

The list of codes presented above is not complete at all, many user defined functions are stored in separated files. In codes themselves, several rows were omitted as well. The full set of codes and datasets can be found in the multimedia attachment.