SFB 649 Discussion Paper 2017-023

Penalized Adaptive Method in Forecasting with Large Information Set and Structure Change

Xinjue Li *
Lenka Zbonakova *[2]
Wolfgang Karl Härdle *[2]

* Xiamen University, P. R. China
*[2] Humboldt-Universität zu Berlin, Germany

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

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

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

Penalized Adaptive Method in Forecasting with Large Information Set and Structure Change[*]

Xinjue Li[†], Lenka Zboňáková[‡] and Wolfgang Karl Härdle [§]

September 04, 2017

Abstract

In the present paper we propose a new method, the Penalized Adaptive Method (PAM), for a data driven detection of structure changes in sparse linear models. The method is able to allocate the longest homogeneous intervals over the data sample and simultaneously choose the most proper variables with help of penalized regression models. The method is simple yet flexible and can be safely applied in high-dimensional cases with different sources of parameter changes. Comparing with the adaptive method in linear models, its combination with dimension reduction yields a method which selects proper significant variables and detects structure breaks while steadily reduces the forecast error in high-dimensional data. When applying PAM to bond risk premia modelling, the locally selected variables and their estimated coefficient loadings identified in the longest stable subsamples over time align with the true structure changes observed throughout the market.

JEL classification: C13, C20, E37

Keywords: SCAD penalty, propagation-separation, adaptive window choice, multiplier bootstrap, bond risk premia

[*] Financial support from the Deutsche Forschungsgemeinschaft via CRC "Economic Risk" and IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged.

[†] W.I.S.E. - Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian, China (*e-mail: cabinofyunnan@163.com*)

[‡] Corresponding author. C.A.S.E. - Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany (*e-mail: zbonakle@hu-berlin.de*)

[§] C.A.S.E. - Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany; Singapore Management University, 50 Stamford Road, 178899 Singapore, Singapore

1 Introduction

Parameter instability is widely recognized as a crucial issue in forecasting. This instability is caused not only by time-variation of coefficients associated with predictors, but also by varying significance of the predictors themselves. Variable selection is particularly important when the true underlying model has a sparse representation. Ensuring high prediction accuracy requires high quality of discovering the relevant variables and ability of adjusting for time-varying coefficient loadings. To handle such instability it is common to use only the most recent rather than all available observations to estimate the coefficients and identify significant predictors at each point of time.

In out-of-sample forecasting, model parameters are generally estimated using either a recursive or rolling window estimation method. These methods are widespread in many areas, especially in macroeconomics and finance, because structure changes are often encountered. However, none of them answers the question of how to select the proper intervals in which the coefficient loadings can be considered to be stationary. Chen and Niu (2014), Chen and Spokoiny (2015) and Niu et al. (2017), among others, addressed this issue by applying a data driven adaptive window choice (Polzehl and Spokoiny (2005), Polzehl and Spokoiny (2006)) to detect the longest homogeneous intervals over the financial and macroeconomic data samples. The method enables us to detect structural shifts and select large subsamples of constant coefficient loadings for predictors, but switches to smaller sample sizes if a structure change is detected. The procedure is fully data driven and parameters are tuned following a propagation-separation approach.

As pointed out by Chen and Niu (2014) the short memory view is quite realistic and easily understood in the context of business cycle dynamics, policy changes and structural breaks. However, we are going to face another question, where we consider the stability of the coefficient loadings and their significance.

Considering the variable selection problem, the traditional criteria such as AIC and BIC become infeasible due to expensive computation in high-dimensional data (Zou and Li (2008)). Fan and Li (2001) advocate the use of penalty functions satisfying certain conditions so the resulting penalized likelihood estimator possesses the properties of sparsity, continuity and unbiasedness while introducing the Smoothly Clipped Absolute Deviation (SCAD) penalty. Moreover, Fan and Li (2001) gave a comprehensive overview of feature selection and proposed a unified penalized likelihood framework to approach the problem of variable selection. Alternatively, the recent advances of variable selection enable us to construct efficient estimation methods. Zou and Li (2008) developed the one-step SCAD algorithm to solve the estimation procedures based on nonconcave penalized likelihood problems. For the SCAD penalty it has been shown that for the appropriate choice of the regularization parameter the nonconcave penalized likelihood estimates perform as well as the oracle procedure in terms of selecting the correct subset of covariates and consistent estimation of the true nonzero coefficients.

Although both the adaptive method and penalized regression models enjoying oracle properties increase prediction accuracy compared with traditional least squares or maximum likelihood methods, neither of them can provide a complete solution when dealing with parameter instability. On one hand, the adaptive algorithm associates nonzero coefficients to all of the predictors which may result in a too large model. On the other hand, treating the whole sample size as a stationary data and performing variable selection and coefficient shrinkage to fit the model also contradicts the economic background, since it is known there are structural breaks and regime switches observable throughout history. Thus the whole sample size data should not be considered as homogeneous.

It seems unwise to directly use some of the penalized regression methods to deal with the macroeconomic problems. It is because predictors can be important during particular periods of time and insignificant in others when the economic situation changes. Therefore we propose to do the breaking points detection simultaneously with the variable selection in a fully data driven way.

In this paper we derive a new method - the Penalized Adaptive Method (PAM) - which can handle all of the previously described challenges. It provides a new way to perform variable selection and structure breaks detection at the same time, i.e. a way to capture parameter instability. With the use of PAM one can detect the longest homogeneous intervals observable throughout the data sample and simultaneously identify the relevant predictors which improves the performance of the out-of-sample forecasting. In the derived approach we assume that the local model with homogeneous parameters will hold with high probability for the forecast horizon and can be automatically identified.

The rest of the paper is organized as follows. In Section 2 we shortly describe the propagation-separation approach introduced by Polzehl and Spokoiny (2006) and the penalized regression method SCAD (Fan and Li (2001)) with its one-step algorithm developed by Zou and Li (2008). Further into the section we then combine those two methods into so-called PAM. In Section 3 we perform the simulation study. Section 4 deals with the application of PAM to a real dataset consisting of excess bond returns and macrovariables observed on the market. Section 5 concludes.

Both simulation study and real data application were performed with help of R software (R Core Team (2014)) and the codes are available on quantlet.de.

2 Penalized Adaptive Method

As mentioned previously, there are several approaches on how to model time-variation in coefficient loadings. One can simply use rolling windows as it was done for example in Härdle et al. (2016), where the authors modelled time variation observable on the financial market. However, this approach has a drawback of selecting the window size prior to model fitting. Of course, this can be done using some external information about the behaviour of the data, e.g observable business cycles or seasonality.

2.1 Propagation-separation approach

In the proposed framework we want to circumvent the use of *a priori* knowledge about the data by selecting the window in a fully data driven way. We will do so by implementing the propagation-separation approach of Polzehl and Spokoiny (2005) and Polzehl and Spokoiny (2006). In the context of model fitting, propagation condition means that the local model can be extended to a longer interval under the assumption of homogeneity. To the opposite, separation means that the extension is restricted to the homogeneous interval. Let us introduce the notation we are going to use throughout this paper, in order to denote the propagation-separation approach from mathematical point of view.

Assume a linear model with a vector of responses $Y = (Y_1, Y_2, \dots, Y_n)^\top$, a vector of parameters $\beta = (\beta_1, \dots, \beta_p)^\top$, an $(n \times p)$ design matrix X and a vector of independent errors ε_i with zero mean and variance σ^2. In this work we assume that the number of parameters p and that the parameter vector β is sparse, i.e. only some number q, $q < n$, of the true coefficients are nonzero.

Now divide the sample of n observations into M nested subintervals. Then for each time point t we have

$$I_t^{(1)} \subset I_t^{(2)} \subset I_t^{(3)} \subset \dots \subset I_t^{(M)},$$

with $n_t^{(m)}$ observations in each subinterval $I_t^{(m)}$, for $m = 1, \dots, M$. Number of subintervals M is arbitrary, however should be reasonably small, so the computation and model fitting is feasible. Increments of observations between two adjacent intervals do not have to be constant.

The algorithm starts with fitting a local model with maximum likelihood (ML) method for the shortest interval $I_t^{(1)}$

$$\tilde{\beta}_t^{(1)} = \arg \max_{\beta} L(\beta, I_t^{(1)}), \tag{1}$$

where $L(\cdot)$ stands for the joint log-likelihood function. The interval $I_t^{(1)}$ is homogeneous by assumption, therefore should be short enough so this assumption holds with high probability. In order to explain the adaptive algorithm, we closely follow Chen and Niu (2014). Let us denote the so called adaptive estimator of the m-th interval by $\widehat{\beta}_t^{(m)}$. The adaptive estimator of the first subinterval $I_t^{(1)}$ is equal to the ML estimator $\tilde{\beta}_t^{(1)}$, which holds because of the previously stated assumption of local homogeneity throughout the interval $I_t^{(1)}$.

Then the propagation-separation approach means that we are testing for significant changes across the neighbouring subsamples with the use of the following test statistic

$$T_t^{(m)} = |2L(\tilde{\beta}_t^{(m)}, I_t^{(m)}) - 2L(\widehat{\beta}_t^{(m-1)}, I_t^{(m)})|^{1/2}, \quad m = 2, \dots, M. \tag{2}$$

Here $\tilde{\beta}_t^{(m)}$ stands for the ML estimator of the vector of parameters using the subinterval $I_t^{(m)}$ and $\widehat{\beta}_t^{(m-1)} = \tilde{\beta}_t^{(m-1)}$ is the previously accepted adaptive estimator from the subinterval $I_t^{(m-1)}$. Both log-likelihood functions in (2) are evaluated over the subinterval $I_t^{(m)}$ so the test statistic $T_t^{(m)}$ measures the difference between the adaptive estimator from the previous subsample and the current ML estimator. Correctly calibrated set of critical values $\zeta_1, \dots, \zeta_M$ is crucial in quantifying the significance level. We refer to Chen and Niu (2014) or Niu et al. (2017) for a calibration relevant for an unpenalized linear model.

Having the set of critical values $\zeta_1, \dots, \zeta_M$, the algorithm proceeds as follows

Adaptive Algorithm

```
1. Initialization: $\widehat{\beta}_t^{(1)} = \tilde{\beta}_t^{(1)}$
2. $m = 2$
3. While $T_t^{(m)} \leq \zeta_m$ and $m \leq M$
       $\widehat{\beta}_t^{(m)} = \tilde{\beta}_t^{(m)}$
       $m = m + 1$
4. Final estimate $\widehat{\beta}_t^{(l)} = \widehat{\beta}_t^{(m-1)}$, for $l \geq m$
```

According to Chen and Niu (2014), after detecting a structure change in the dataset using step 3 from the algorithm, the final estimate from step 4 is the ML estimate from the longest identified homogeneous interval and it is used as a valid estimate also for longer subsamples. Since we want to correctly identify all of the possible change points in our data, after detecting the change we initiate the algorithm from the beginning with a smaller data sample.

2.2 SCAD penalty

So far we were dealing with a linear model, where the number of parameters is predefined or chosen by one of the variable selection methods available. As mentioned previously, variable selection with use of BIC or AIC criteria might not be computationally feasible when dealing with high-dimensional data. Therefore our aim is to combine the foregoing adaptive algorithm with penalized regression methods, which serves the objective of simultaneous dimension reduction and nonstationarity detection.

For this purpose we are using the smoothly clipped absolute deviation (SCAD) method introduced by Fan and Li (2001). The reason why we choose nonconcave SCAD penalty over the least absolute shrinkage and selection operator (LASSO) developed by Tibshirani (1996) is that the SCAD penalty yields oracle estimator under some conditions on a shrinkage parameter λ. LASSO selects the true model consistently (almost) only if the irrepresentable condition is satisfied (Zhao and Yu (2006)). As this would impose constraints on the design matrix X, we are omitting the application of LASSO as a variable selection method.

Moreover, SCAD estimator enjoys three important properties desirable in penalized regression model fitting, which are sparsity, continuity and unbiasedness. All of them play a crucial role in PAM when it comes to calibration of critical values and, finally, longest homogeneous subinterval identification.

However, a drawback of SCAD penalty is its nonconcavity. Fan and Li (2001) proposed an algorithm with local quadratic approximation (LQA) of SCAD penalty to be able to perform the shrinkage and selection as a minimization problem. Zou and Li (2008) revisited the task of finding the solution to penalized likelihood problem and developed an algorithm with local linear approximation (LLA) of the broad class of penalty functions with SCAD among others. In their work they showed the proposed method outperforms the LQA approach, in a sense that it automatically adapts a sparse solution. What is more, the computational cost is significantly reduced by using only one iteration step as the efficiency of the algorithm is the same as for the fully iterative method. This holds under the assumption that the initial estimators are reasonably chosen.

In this paper we perform the penalized likelihood estimation of the vector of parameters, i.e. we maximize the objective function

$$Q(\beta) = \sum_{i=1}^{n} l_i(\beta) - n \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{3}$$

with $l_i(\cdot)$ a non-penalized log-likelihood function for an observed couple (Y_i, X_i) and $p_\lambda(\cdot)$ a penalty function with parameter $\lambda > 0$. The SCAD penalty is in Fan and Li (2001) defined as a continuous differentiable function

$$p'_\lambda(|\beta_j|) = \lambda \left\{ I(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda) \right\}, \tag{4}$$

for some $a > 2$ ($a = 3.7$ was suggested as a generally good choice) and $\lambda > 0$. By $I(\cdot)$ we denote an indicator function and $(\cdot)_+ = \max(0, \cdot)$.

Following the LLA approach by Zou and Li (2008), the general penalty function $p_\lambda(|\beta_j|)$ can be locally approximated by

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + p'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|), \tag{5}$$

for some $\beta_j \approx \beta_j^{(0)}$. Then the k-th iteration step estimator of their proposed procedure is defined as follows

$$\beta^{(k+1)} = \arg\max_\beta \left\{ \sum_{i=1}^n l_i(\beta) - n \sum_{j=1}^p p'_\lambda(|\beta_j^{(k)}|)|\beta_j| \right\} \tag{6}$$

for $k = 1, 2, \ldots$, and $\beta^{(0)}$ being a non-penalized maximum likelihood estimator. The iteration process stops if the sequence $\{\beta^{(k)}\}$ converges. We refer to Zou and Li (2008) for the proof of convergence and oracle properties of the one-step SCAD estimator under condition that the penalty parameter λ satisfies

$$\sqrt{n}\lambda_n \to \infty \quad \text{and} \quad \lambda_n \to 0. \tag{7}$$

Here we use a subscript n to denote the dependency of λ_n on number of observations n in the model.

2.3 Penalized Adaptive Method

As discussed, both propagation-separation approach and penalized regression with SCAD penalty function have their advantages in capturing non-stationarity and dimension reduction, respectively. To combine the properties of these two methods, we propose an algorithm called Penalized Adaptive Method (PAM). In PAM we are building a procedure, which deals with non-stationary and high-dimensional data simultaneously in a data driven way. Adaptive way of choosing window size helps us in determining the longest homogeneous subsample of a given dataset and penalized regression reduces the dimension so the interpretability of the model is improved.

One of the differences from previously introduced propagation-separation approach lies in using a penalized likelihood function $Q(\beta)$ rather than its non-penalized counterpart, i.e. the test statistic takes the form

$$T_t^{(m)} = |2Q(\tilde{\beta}_t^{(m)}, I_t^{(m)}) - 2Q(\widehat{\beta}_t^{(m-1)}, I_t^{(m)})|^{1/2}, \quad m = 2, \ldots, M, \tag{8}$$

where $Q(\beta, \cdot)$ is defined as previously in equation (3) with the second argument denoting the interval over which is the function evaluated. $\tilde{\beta}_t^{(m)}$ is a SCAD estimator over the subinterval $I_t^{(m)}$, i.e. $\tilde{\beta}_t^{(m)} = \arg\max Q(\beta, I_t^{(m)})$ and $\widehat{\beta}_t^{(m-1)} = \arg\max Q(\beta, I_t^{(m-1)})$ is a SCAD estimator from previously accepted homogeneous subinterval $I_t^{(m-1)}$. Nevertheless, a major difference comes into play when one focuses on calibration of the critical values as a crucial part of the adaptive method itself. Distribution of the test statistic is unknown for the non-penalized case in (2) and for the penalized case (8) one important question arises; how do we compute confidence sets for sparse estimators of β? Fan and Li (2001) derived a formula for variance approximation of the non-zero components of the SCAD estimator of β. However, as pointed out in their work, estimated standard deviation for zero components of the estimator is 0 and therefore one is unable to do any inference related to those elements of vector β. This was also a problem in Tibshirani (1996) and it was reconsidered by Chatterjee and Lahiri (2011). In their work, Chatterjee and Lahiri (2011) developed a modified residual bootstrap which, under some mild conditions, consistently estimates variance of all of the parameter values, both zero and nonzero. Moreover, they show that for the adaptive LASSO (Zou (2006)) the residual bootstrap yields consistent variance estimators even without use of any modification.

Despite these results, residual bootstrap should not be used when the design matrix is random. Instead, one should perform bootstrap on the observed couples (Y_i, X_i), $i = 1, \ldots, n$, where $X_i \in \mathbb{R}^p$ is the i-th row of the design matrix X. Since we are interested in evaluating likelihood based confidence sets, one of the possibilities at hand is so-called wild or multiplier bootstrap. For the case of non-penalized likelihood Spokoiny and Zhilova (2015) developed useful theoretical results valid even for small or moderate sample sizes with possible model misspecification. In this section we are going to relate their results with the method used for critical values calibration in PAM.

2.3.1 Multiplier Bootstrap for Penalized Likelihood

In order to describe the multiplier bootstrap procedure for likelihood based functions, we closely follow Spokoiny and Zhilova (2015) with extension of the notation for the penalized likelihood case. Let us use the notation from previous chapters for the non-penalized log-likelihood function $L(\beta) = \sum_{i=1}^n l_i(\beta)$, i.e. $l_i(\beta)$ denotes the parametric logarithmic density of the i-th observation in a given sample. Assume a set of i.i.d. scalar random variables u_i, $i = 1, \ldots, n$, which are independent of Y and X, if X is considered random. Further assumptions about the so-called multipliers are that $\mathrm{E}(u_i) = 1$, $\mathrm{var}(u_i) = 1$ and $\mathrm{E}(\exp(u_i)) < \infty$. Multiplying the elements of $L(\beta)$ by the defined random variables u_i we get the bootstrap penalized log-likelihood function as follows

$$Q^\circ(\beta) = \sum_{i=1}^n l_i(\beta) u_i - n \sum_{j=1}^p p_\lambda(|\beta_j|). \tag{9}$$

Denoting $\mathrm{E}^{\circ}(\cdot) = \mathrm{E}(\cdot|Y)$ we then can write

$$\arg\max_{\beta} \mathrm{E}^{\circ}\, Q^{\circ}(\beta) = \arg\max_{\beta} Q(\beta) = \tilde{\beta}. \tag{10}$$

Thus, as pointed out by Spokoiny and Zhilova (2015), the target parameter in the bootstrap world coincides with the penalized MLE of the real world. The penalized MLE of the bootstrap world is then defined as

$$\tilde{\beta}^{\circ} = \arg\max_{\beta} Q^{\circ}(\beta). \tag{11}$$

It is important to note, that the parameter λ of the SCAD method is the same for $Q(\beta)$ and $Q^{\circ}(\beta)$. Then one circumvents the problem of penalizing elements of vector β by a different amount in the real and the bootstrap case, which could lead to unstable results.

If one wishes to approximate the distribution of the test statistic $\{2Q(\tilde{\beta}) - 2Q(\beta^{*})\}^{1/2}$, where β^{*} denotes the real unknown parameter vector, it can be done (up to some approximation error) by using the ratio $\{2Q^{\circ}(\tilde{\beta}^{\circ}) - Q^{\circ}(\tilde{\beta})\}^{1/2}$, where all of the elements are known. Then, similarly, one can use this approximation for finding critical values for the aforementioned test statistic. Specifically, let $1 - \alpha \in (0, 1)$ be a determined confidence level of a testing procedure. It is then straightforward to follow, that the approximation of a desired quantile of the distribution of the likelihood ratio

$$\zeta_{\alpha}^{*} = \inf\{z \geq 0 : \mathrm{P}\{Q(\tilde{\beta}) - Q(\beta^{*})\} > z^{2}/2 \leq \alpha\} \tag{12}$$

can be evaluated as

$$\zeta_{\alpha}^{\circ} = \inf\{z \geq 0 : \mathrm{P}^{\circ}\{Q^{\circ}(\tilde{\beta}^{\circ}) - Q^{\circ}(\tilde{\beta})\} > z^{2}/2 \leq \alpha\}, \tag{13}$$

where P° denotes conditional probability given vector of observation Y. For further theory on embedding multiplier bootstrap into non-penalized likelihood function setting, we refer the reader to Spokoiny and Zhilova (2015).

2.3.2 Critical Values Calibration

Using results from the previous section, we propose to calibrate corresponding critical values for PAM by using the multiplier bootstrap method based on penalized likelihood function from (3). Speaking in terms of the adaptive window choice the procedure is as follows. Assume sequence of nested intervals $I_t^{(1)} \subset I_t^{(2)} \subset \ldots \subset I_t^{(M)}$ for a given time point t. Suppose the interval $I_t^{(m-1)}$ is homogeneous. We have the SCAD estimator

$$\tilde{\beta}_t^{(m-1)} = \arg\max_{\beta} Q(\beta, I_t^{(m-1)}), \tag{14}$$

which we also call an adaptive estimator for the given subinterval and denote it by $\widehat{\beta}_t^{(m-1)}$. Next step is to analogously find $\tilde{\beta}_t^{(m)}$ and corresponding penalty parameter

$\lambda_t^{(m)}$. Remember, the second parameter of the SCAD penalty function, a, is kept constant and equal to 3.7 as suggested by Fan and Li (2001).

We implement the multiplier bootstrap into assessing quantiles of the test statistic from (8) by simulating a large number B of i.i.d. multipliers u_i, $i = 1, \ldots, n_t^{(m)}$. Computing

$$|2Q^{\circ b}(\tilde{\beta}_t^{\circ b(m)}, I_t^{(m)}) - 2Q^{\circ b}(\tilde{\beta}_t^{(m)}, I_t^{(m)})|^{1/2}, \quad b = 1, \ldots, B, \tag{15}$$

we get an approximate distribution of $|2Q(\tilde{\beta}_t^{(m)}, I_t^{(m)}) - 2Q(\beta_t^{*(m)}, I_t^{(m)})|^{1/2}$, a transformation of the real likelihood ratio if one denotes the real unknown parameter over the supposedly homogeneous interval $I_t^{(m)}$ by $\beta_t^{*(m)}$. Next, for a given confidence level $1 - \alpha \in (0, 1)$ the critical value is defined as

$$\zeta_{t\alpha}^{(m)} = \inf\{z \geq 0 : \mathrm{P}^{\circ}\{Q^{\circ}(\tilde{\beta}_t^{\circ(m)}, I_t^{(m)}) - Q^{\circ}(\tilde{\beta}_t^{(m)}, I_t^{(m)})\} > z^2/2 \leq \alpha\}. \tag{16}$$

Comparing the test statistic from (8) to the above defined critical value we either reject the homogeneity hypothesis, if $T_t^{(m)} > \zeta_{t\alpha}^{(m)}$, for the given confidence level, or move to the next step in PAM algorithm and prolong the subsample regarded as homogeneous.

3 Simulation Study

In order to justify the use of multiplier bootstrap in critical values calibration we include a simulation study concerning its performance. Using the LLA algorithm of Zou and Li (2008), we need multipliers u_i, $i = 1, \ldots, n$ to be non-negative. Therefore we propose to use either $u_i \sim \mathrm{Exp}(1)$ or u_i having a bounded distribution on interval $(0, 4)$ with a pdf

$$f(u_i) = \begin{cases} \dfrac{3}{14} & \text{if} \quad 0 \leq u_i \leq 1; \\ \\ \dfrac{1}{12} & \quad 1 < u_i \leq 4. \end{cases} \tag{17}$$

In the simulation study we consider a linear model $Y = X\beta + \varepsilon$ with number of observations n and number of parameters p from which only $q \leq p$ are nonzero. Design matrix X is taken from p-dimensional normal distribution as follows

$$\{X_i\}_{i=1}^n \sim \mathrm{N}_p(0, \Sigma), \tag{18}$$

with elements $\{\sigma_{ij}\}_{i,j=1}^p$ of the covariance matrix Σ satisfying $\sigma_{ij} = 0.5^{|i-j|}$. Error terms ε_i are simulated as i.i.d. from $\mathrm{N}(0, 1)$. We consider $n = 50, 100, 200$ to assess performance for small to medium sized samples. Number of parameters p is set to be $p = 10$ and for each n we define $q = 3, 5$ as number of real nonzero parameters, i.e. $\beta = (1, 1, 1, 0, \ldots, 0)^\top$ or $\beta = (1, 1, 1, 1, 1, 0 \ldots, 0)^\top$. For each of the studied settings we simulated 1 000 scenarios and for each scenario we simulated 10 000 u_i's both exponentially and bounded distributed in order to obtain an approximation of the penalized likelihood distribution.

Summary of the simulation is given in Table 1. We set $\alpha = 0.1, 0.05, 0.025$ to compute the upper quantiles of the bootstrap penalized likelihood ratio distribution. In Table 1 one can see percentage of occurrences of the event, when the real likelihood ratio $Q(\tilde{\beta}) - Q(\beta^*)$ was smaller or equal than the respective quantile of its approximated distribution.

		Bound	Exp	Bound	Exp	Bound	Exp
n	q	90 %		95 %		97.5 %	
50	3	85.7	85.9	90.6	91.2	94.3	95.2
	5	80.3	80.1	87.8	89.6	94.0	95.4
100	3	94.0	93.9	96.8	97.0	98.0	98.4
	5	89.4	89.4	94.2	94.8	96.8	97.5
200	3	95.4	95.5	97.7	98.3	99.3	99.4
	5	89.7	89.5	94.0	94.4	96.3	97.0

Table 1: Multiplier bootstrap performance

As can be seen from Table 1, for the case of $n = 50$ approximated quantiles of the penalized likelihood ratio statistic are smaller than expected. This is due to use of the penalized regression method, which creates bigger estimation error for small sample sizes and this error is not mirrored in the bootstrap world. Also the performance of bootstrap quantiles increases with increasing ratio q/p, the ratio of nonzero to all parameters in the model. For the cases of $n = 100, 200$ and $q = 5$ we see, that the bootstrap quantiles approximate the real quantiles reasonably well. This holds for both bounded and exponentially distributed multipliers u_i, $i = 1, \ldots, n$. Explanation lies in the use of SCAD penalty and its oracle properties, which with increasing n guarantee that the model parameters are consistently fitted. Moreover, with penalty parameter λ_n approaching zero as number of observations n increases one can see that the penalized likelihood $Q(\beta)$ approaches its non-penalized form $L(\beta)$ where the number of parameters of the model is equal to q and inference from Spokoiny and Zhilova (2015) can be applied.

3.1 Change Point Detection

In the following we perform a simulation study regarding the use of bootstrap critical values in a change point detection, i.e. in the propagation-separation approach to adaptive window choice. We again assume a linear model $Y = X\beta + \varepsilon$, with the same design matrix X and the error term ε as before. For this study we use number of subintervals $M = 20$ and we keep increments between successive subintervals constant, $n_t^{(m+1)} - n_t^{(m)} = 50$, for $m = 1, \dots, M-1$. Then we define the true parameter vector $\beta_i^* \in \mathbb{R}^p$, $p = 10$, $i = 1, \dots, n_t^{(M)}$ as

$$\beta_i^* = \begin{cases} (1,1,1,1,1,0,\dots,0) & \text{if} \quad i < i_{cp}; \\ (2,2,2,2,2,0,\dots,0) & \qquad i \geq i_{cp}, \end{cases} \tag{19}$$

where i_{cp} denotes an observation with a change point. Further, for comparison, we use multipliers u_i with exponential and bounded distribution, where the latter is defined by (17). We simulated 500 scenarios for 18 different i_{cp}, i.e. for each scenario there was only one change point occurring throughout the set of all observations $n_t^{(M)}$. Results of the multiplier bootstrap performance are summarized in Table 2 and Table 3, respectively.

i_{cp}	50	100	150	200	250	300
Corr	100	100	100	100	100	100
1stCorr	100	79.6	78.2	78.2	78.2	78.2
i_{cp}	350	400	450	500	550	600
Corr	100	100	100	99.8	99.6	99.8
1stCorr	78.2	78.2	78.2	78.2	78.0	78.0
i_{cp}	650	700	750	800	850	900
Corr	100	99.4	99.6	98.6	99.4	98.2
1stCorr	78.2	77.6	77.8	76.6	77.8	76.8

Table 2: Percentage of correctly identified change points with use of $u_i \overset{iid}{\sim} \text{Exp}(1)$.

i_{cp}	50	100	150	200	250	300
Corr	100	100	100	100	100	100
1stCorr	100	79.6	78.2	78.2	78.2	78.2
i_{cp}	350	400	450	500	550	600
Corr	100	100	100	99.8	99.6	99.8
1stCorr	78.2	78.2	78.2	78.2	78.0	78.0
i_{cp}	650	700	750	800	850	900
Corr	100	99.4	99.6	98.4	99.4	98.2
1stCorr	78.2	77.6	77.8	76.6	77.8	76.8

Table 3: Percentage of correctly identified change points with use of $u_i \overset{iid}{\sim}$ bounded from (17).

In the aforementioned tables we denote a percentage of correctly identified change points by "Corr" and "1stCorr" stands for a percentage of correctly identified change points, which were identified as the first ones occurring. One can see that the latter went abruptly down when the change point occurred later than in the second subinterval $I_t^{(2)}$. This is the issue we addressed previously noticing that the critical values are generally smaller than the real quantiles of the test statistic (8) if the sample size is small. Therefore they incorrectly reject the homogeneity hypothesis of the second interval and this error mirrors in the lower number of correctly identified occurrences of the first change points in the given set of observations. After this abrupt decline, the value of "1stCorr" stays approximately constant for both distributions of u_i's, meaning that the homogeneity hypothesis of the interval $I_t^{(2)}$ is falsely rejected in approximately 20 % of the scenarios for all of the 18 cases. Otherwise the method performs reasonably well with high values of "Corr" in all of the cases.

With reference to the simulation results one would suggest the use of moderate number of observations for model fitting or a correction of calibrated critical values corresponding to smaller sample sizes.

4 Excess Bond Premia Modelling

In this section we use the previous results and apply PAM to the excess bond premia modelling problem. Motivation for this application comes mainly from Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009), where they used linear model with macro factors in order to forecast bond risk premium, which was regarded, by the expectation hypothesis, as unforecastable in the past. Cochrane and Piazzesi (2005) reconsidered model of Fama and Bliss (1987), who proved that the expectation hypothesis does not hold and compared it to their newly proposed factor model which was shown to outperform the preceding one.

However, all of the previous authors considered the coefficient loadings in their models to be homogeneous throughout the whole sample size and if not, they assumed the factor models compensate for the non-stationarity (Ludvigson and Ng (2009)). Our aim is to introduce possible time-varying coefficient loadings into the modelling and also propose a different dimension reduction which will not come from factor models, but rather from a penalized regression. The advantage of the latter lies in direct association of the modelled bond risk premia with actual macroeconomic variables, which simplifies model interpretation. To the best of our knowledge such an approach has not yet been implemented in the case of macroeconomic modelling.

As for the notation, we closely follow Cochrane and Piazzesi (2005) throughout the chapter. Let us denote the log bond prices by $p_t^{(k)}$ = log price of k-year discount bond at time t. Then the log yield is determined by

$$y_t^{(k)} = -\frac{1}{k} p_t^{(k)}.$$

Further, log forward rate for loans between time $t + k - 1$ and $t + k$ specified at time t is

$$f_t^{(k)} = p_t^{(k-1)} - p_t^{(k)}$$

and the log holding period return from buying a k-year bond at time t and selling it at time $t + 1$ as a $(k - 1)$-year bond is denoted by

$$r_{t+1}^{(k)} = p_{t-1}^{(k-1)} - p_t^{(k)}.$$

Finally, for the excess log returns we write

$$rx_{t+1}^{(k)} = r_{t+1}^{(k)} - y_t^{(1)}, \quad \text{for } k = 2, 3, 4, 5.$$

In their work, Cochrane and Piazzesi (2005) started with considering linear regressions with excess log returns for all maturities as dependent variables and all of the related forward rates as predictors, i.e.

$$rx_{t+1}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} y_t^{(1)} + \beta_2^{(k)} f_t^{(2)} + \ldots + \beta_5^{(k)} f_t^{(5)} + \varepsilon_{t+1}^{(k)}, \tag{20}$$

for $k = 2, 3, 4, 5$. Further they specified a single factor for modelling expected excess returns for all k. Since we are interested in time-variation and dimension reduction, we omit the description of their further model. Nevertheless, let us take the model from (20) as a baseline model with which we later compare performance of PAM.

In what follows we deviate from work of Cochrane and Piazzesi (2005) in the sense that we adopt results of Ludvigson and Ng (2009) who showed that inclusion of macro variables, or factors based on macro variables to be more specific, improves the model fit and its forecasting performance. This serves our purpose, since with PAM we can include large number of covariates and reduce the dimension of the model afterwards.

We use monthly US data obtained from CRSP and Federal Reserve Board. The time span over which the data sample was taken is from January 1995 to December 2014. In addition to the forward rates as defined previously we consider their 1-year lagged values and other macroeconomic variables which are specified in Table 4. Altogether the predictors yield a dimension of $p = 20$.

As discussed previously, when using multiplier bootstrap in combination with penalized likelihood function, the critical values are small and falsely detect change points in the model if the number of observations in the sample n is small. Therefore we use a cubic spline interpolation of given covariates to obtain “weekly” observations and therein increase our information set.

Let us now specify the proposed model. For each $k = 2, 3, 4, 5$ we assume

$$rx^{(k)}_{t+1} = \beta_0 t^{(k)} + \beta^{(k)\top}_{1t} f_t + \beta^{(k)\top}_{2t} f_{t-1} + \beta^{(k)\top}_{2t} M_t + \varepsilon^{(k)}_{t+1}, \tag{21}$$

where $f_t = (y^{(1)}_t, f^{(2)}_t, \dots, f^{(5)}_t)^\top$ and f_{t-1} is defined analogously. Vector M_t then defines all of the macro variables from Table 4. Apart from adding more predictors into the model, there is another difference between our approach and that of Cochrane and Piazzesi (2005), we allow for time-variation which is denoted by the subscript t.

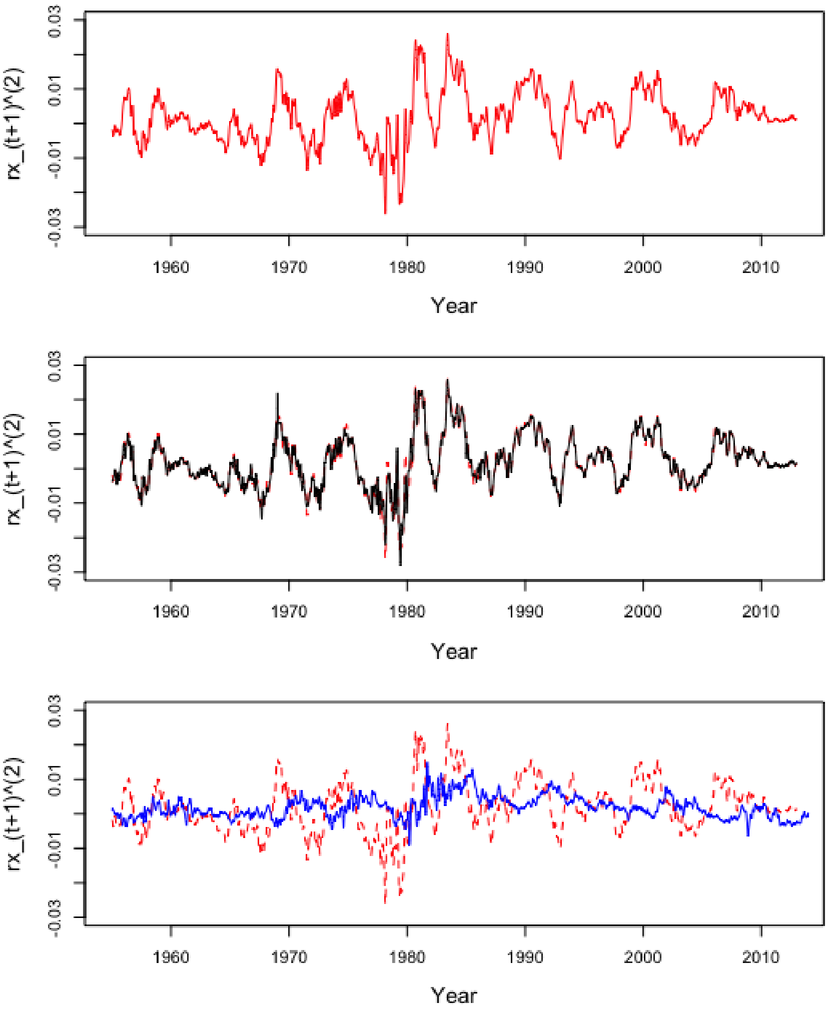

Figure 1: Comparison of a fitted baseline forward rates model (blue) and PAM (black) performance where the true excess bond risk premia for maturity 2 years is depicted as a red curve.

5 Concluding Remarks

In the presented paper we introduced a novel approach for capturing nonstationarity and reducing dimension of the model by Penalized Adaptive Method. We argued its

1.	Federal fund returns
2.	10-year treasury bond returns
3.	7-year treasury bond returns
4.	5-year treasury bond returns
5.	2-year treasury bond returns
6.	1-year treasury bond returns
7.	Inflation rate (year-on-year price changing rate)
8.	Industrial production (year-on-year growth of industrial production)
9.	Returns on S&P 500 index
10.	Value-weighted return excluding dividend

Table 4: List of macroeconomic variables

advantages and applied the procedure to excess bond risk premia modelling, where we compared our model to the baseline of Cochrane and Piazzesi (2005). Further extension of PAM which we hope could be of interest lies in implementing the case when $p \geq n$ and/or modelling quantiles.

References

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso Estimators, Journal of the American Statistical Association **106**: 608–625.

Chen, Y. and Niu, L. (2014). Adaptive Dynamic NelsonSiegel Term Structure Model with Applications, Journal of Econometrics **180**: 98–115.

Chen, Y. and Spokoiny, V. (2015). Modeling Nonstationary and Leptokurtic Financial Time Series, Econometric Theory **31**: 703–728.

Cochrane, J. H. and Piazzesi, M. (2005). Bond Risk Premia, American Economic Review **95**: 138–160.

Fama, E. F. and Bliss, R. R. (1987). The Information in Long-Maturity Forward Rates, American Economic Review **77**: 680-692.

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

Härdle, W. K, Wang, W. and Yu, L. (2016). TENET: Tail-Event driven NETwork risk, Journal of Econometrics **192**: 499–513.

Ludvigson, S. C. and Ng, S. (2009). Macro Factors in Bond Risk Premia, The Review of Financial Studies **22**: 5027–5067.

Niu, L., Xu, X. and Chen, Y. (2017). An Adaptive Approach to Forecasting Three Key Macroeconomic Variables for Transitional China, Economic Modelling in press.

Polzehl, J. and Spokoiny, V. (2005). Spatially Adaptive Regression Estimation: Propagation-Separation Approach, WIAS Preprint No. 218.

Polzehl, J. and Spokoiny, V. (2006). Propagation-Separation Approach for Local Likelihood Estimation, Probablity Theory and Related Fields **135**: 335–362.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. `http://www.R-project.org/` (Accessed: 15th April 2015).

Spokoiny, V. and Zhilova, M. (2015). Bootstrap Confidence Sets Under Model Misspecification, The Annals of Statistics **43**: 2653–2675.

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

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

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

Zou, H. and Li, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models, The Annals of Statistics **36**: 1509–1533.

SFB 649 Discussion Paper Series 2017

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

001 "Fake Alpha" by Marcel Müller, Tobias Rosenberger and Marliese Uhrig-Homburg, January 2017.

002 "Estimating location values of agricultural land" by Georg Helbing, Zhiwei Shen, Martin Odening and Matthias Ritter, January 2017.

003 "FRM: a Financial Risk Meter based on penalizing tail events occurrence" by Lining Yu, Wolfgang Karl Härdle, Lukas Borke and Thijs Benschop, January 2017.

004 "Tail event driven networks of SIFIs" by Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle and Yarema Okhrin, January 2017.

005 "Dynamic Valuation of Weather Derivatives under Default Risk" by Wolfgang Karl Härdle and Maria Osipenko, February 2017.

006 "RiskAnalytics: an R package for real time processing of Nasdaq and Yahoo finance data and parallelized quantile lasso regression methods" by Lukas Borke, February 2017.

007 "Testing Missing at Random using Instrumental Variables" by Christoph Breunig, February 2017.

008 "GitHub API based QuantNet Mining infrastructure in R" by Lukas Borke and Wolfgang K. Härdle, February 2017.

009 "The Economics of German Unification after Twenty-five Years: Lessons for Korea" by Michael C. Burda and Mark Weder, April 2017.

010 "Data Science & Digital Society" by Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, May 2017.

011 "The impact of news on US household inflation expectations" by Shih-Kang Chao, Wolfgang Karl Härdle, Jeffrey Sheen, Stefan Trück and Ben Zhe Wang, May 2017.

012 "Industry Interdependency Dynamics in a Network Context" by Ya Qian, Wolfgang Karl Härdle and Cathy Yi-Hsuan Chen, May 2017.

013 "Adaptive weights clustering of research papers" by Larisa Adamyan, Kirill Efimov, Cathy Yi-Hsuan Chen, Wolfgang K. Härdle, July 2017.

014 "Investing with cryptocurrencies - A liquidity constrained investment approach" by Simon Trimborn, Mingyang Li and Wolfgang Karl Härdle, July 2017.

015 "(Un)expected Monetary Policy Shocks and Term Premia" by Martin Kliem and Alexander Meyer-Gohde, July 2017.

016 " Conditional moment restrictions and the role of density information in estimated structural models" by Andreas Tryphonides, July 2017.

017 "Generalized Entropy and Model Uncertainty" by Alexander Meyer-Gohde, August 2017.

018 "Social Security Contributions and the Business Cycle" by Anna Almosova, Michael C. Burda and Simon Voigts, August 2017.

019 "Racial/Ethnic Differences In Non-Work At Work" by Daniel S. Hamermesh, Katie R. Genadek and Michael C. Burda, August 2017.

020 "Pricing Green Financial Products" by Awdesch Melzer, Wolfgang K. Härdle and Brenda López Cabrera, August 2017.

021 "The systemic risk of central SIFIs" by Cathy Yi-Hsuan Chen and Sergey Nasekin, August 2017.

022 "Das deutsche Arbeitsmarktwunder: Eine Bilanz" by Michael C. Burda and Stefanie Seele, August 2017.

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

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

SFB 649 Discussion Paper Series 2017

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

023 "Penalized Adaptive Method in Forecasting with Large Information Set and Structure Change" by Xinjue Li, Lenka Zbonakova and Wolfgang Karl Härdle, September 2017.

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

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