

IRTG 1792 Discussion Paper 2019-030



Combining Penalization and Adaption in High Dimension with Application in Bond Risk Premia Forecasting

Xinjue Li ^{*}
Lenka Zboňáková ^{*2}
Weining Wang ^{*2}
Wolfgang Karl Härdle ^{* *2 *3 *4}



- * Xiamen University, China
- *2 Humboldt-Universität zu Berlin, Germany
- *3 Singapore Management University, Singapore
- *4 Charles University, Czech Republic

This research was supported by the Deutsche
Forschungsgesellschaft through the
International Research Training Group 1792
"High Dimensional Nonstationary Time Series".

<http://irtg1792.hu-berlin.de>
ISSN 2568-5619

International Research Training Group 1792

Combining Penalization & Adaption in High Dimension with Application in Bond Risk Premia Forecasting

Xinjue Li^{a,*}, Lenka Zboňáková^b, Weining Wang^b

Wolfgang Karl Härdle ^{a,b,c,d}

Abstract

The predictability of a high-dimensional time series model in forecasting with large information sets depends not only on the stability of parameters but also depends heavily on the active covariates in the model. Since the true empirical environment can change as time goes by, the variables that function well at the present may become useless in the future. Combined with the instable parameters, finding the most active covariates in the parameter time-varying situations becomes difficult. In this paper, we aim to propose a new method, the Penalized Adaptive Method (PAM), which can adaptively detect the parameter homogeneous intervals and simultaneously select the active variables in sparse models. The newly developed method is able to identify the parameters stability at one hand and meanwhile, at the other hand, can manage of selecting the active forecasting covariates at every different time point. Comparing with the classical models, the method can be applied to high-dimensional cases with different sources of parameter changes while it steadily reduces the forecast error in high-dimensional data. In the out-of-sample bond risk premia forecasting, the Penalized Adaptive Method can reduce the forecasting error(RMSPE and MAPE) around 24% to 50% comparing with the other forecasting methods.

JEL classification: C4, C5, E4, G1

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

^aW.I.S.E. - Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, 361005, Fujian, China

^bC.A.S.E. - Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Spandauer Str.1,10178 Berlin, Germany

^cSingapore Management University, 50 Stamford Road, 178899 Singapore, Singapore

^dDepartment of Mathematics and Physics, Charles University Prague, Ke Karlovu 2027/3, 12116 Praha 2, Czech

*Correspondence to: Rm N310, Economic Building, Xiamen University, Xiamen, 361005, Fujian, China, *E-mail: XinjueLi@outlook.com*

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 an 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 parameter variations 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 stable. 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 identify parameter homogeneity and select large subsamples of constant coefficient loadings for predictors, but switches to smaller sample sizes if parameter inhomogeneity 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 other exogenous economic shocks. However, in this work we 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). One of the possibilities at hand for dealing with large dimensions is the LASSO introduced by Tibshirani (1996) and recently applied to a system of high-dimensional regression equations by Chernozhukov et al. (2018). Further, Fan and Li (2001) advocate the use of other 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 variable 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 sta-

tionary data and performing variable selection and coefficient shrinkage to fit the model also contradicts the economic background, since it is known that there are exogenous economic shocks and regime switches observable throughout history. Thus, the whole sample size should not be considered as homogeneous.

It seems unwise to directly use some of the penalized regression methods to deal with a parameter-varying macroeconomic problem. 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 break point 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 homogeneous interval 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 advantages of PAM are documented by applying the method to the excess bond risk premia modelling problem. Comparison of the in-sample and out-of-sample fit of our proposed method with the baseline models from Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009) shows significant improvement in terms of various model accuracy measures when applying the former.

The rest of the paper is organized as follows. In Section 2 we first describe the SCAD penalized regression method with its one-step algorithm developed by Zou and Li (2008) and the adaptive propagation-separation approach. Further into

the section we then combine those two methods into the 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. The theoretical results are shown in Section 2.5 and Section 5 concludes.

2 Penalized Adaptive Method

Although many economies and financial markets have been experiencing policy shocks and business cycles such as the global recession (2008), the European debt crisis(2010), the Brexit(2016) and the current trade war(2018), mostly used econometric models are based on the assumption of time homogeneity. But, the market and institutional changes have long been assumed to cause structural instability and uncertainty in economic time series. Ignoring these instabilities(both from the structural side and the variable side) can adversely affect the modelling, estimation and forecasting. As mentioned previously, there are several approaches on how to model time-variation in coefficient loadings. By using the external information about the business cycles, 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 the drawback of selecting the window size and also the variables prior to model fitting and in general it stays an unsolved issue affecting the interpretability of the statistical results.

As discussed before, both the adaptive approach and penalized SCAD regression have their respective advantages in capturing the homogeneity in stationary or in non-stationary time series and in dimension reduction. To combine the properties of this two methods, we propose a new method which can detect the homogeneous

subintervals and meanwhile automatically select the variables in high dimensional situations.

In the following part, we develop an alternative and more robust parametric approach to the local stability analysis that relies on a finite-sample theory of testing a growing sequence of historical time intervals on homogeneity. We first discuss the penalization method applied as a tool to estimate and select the variables in a given interval I . Then we discuss the test statistics employed to test the homogeneity of the interval I . Later, we rigorously describe the adaptive estimation procedure, the implementations and the selection of the parameters entering the adaptive procedure.

2.1 SCAD Penalty

We consider a linear model

$$y_{t+1} = \beta_t^\top x_t + \varepsilon_{t+1} \quad (1)$$

with the sample size of n and where y_{t+1} is the response. $x_t = (1, x_{t,1}, \dots, x_{t,p})^\top$, $\beta_t = (\beta_{t,0}, \beta_{t,1}, \dots, \beta_{t,p})^\top$. In order to simplify the economic forecasting model, we follow Chen et al. (2010) and Chen and Niu (2014) to set the errors $\{\varepsilon_{t+1}\}_{t=1}^n$ a set of *i.i.d* random variables with zero mean and variance σ_{t+1}^2 . In this paper we assume that the parameter vector β_t is sparse, i.e. only some number q_t , $1 \leq q_t < p$, of the true coefficients are nonzero.

We are now dealing with a linear sparse model, where the active covariates are unknown and will shift as the time goes by. At every point of time, the active covariates have to be chosen by one of the available variable selection methods. For this purpose we are using the smoothly clipped absolute deviation (SCAD) method introduced by Fan and Li (2001). The proposed nonconcave

SCAD penalty yields an oracle estimator under some conditions on a shrinkage parameter λ and this property plays a crucial role in the 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 general, we can simply consider a given interval I with sample size n since the whole search and estimation can be repeated at different time points. Considering of variable selection, we perform the penalized (quasi) likelihood estimation of the parameter β and maximize the objective function on the given interval I ,

$$L_{Q,I}(\beta) = L_I(\beta) - n \sum_{j=1}^p P_\lambda(|\beta_j|), \quad (2)$$

where $L_I(\beta) = \sum_{i=1}^n \ell_i(\beta)$ is the non-penalized log-likelihood function with $\ell_i(\cdot)$ to be the non-penalized log-likelihood function for an observed $(p+1)$ -tuple (y_{i+1}, x_i) and $P_\lambda(\cdot)$ a penalty function with parameter $\lambda > 0$. The SCAD penalty is defined as a continuous differentiable function with a derivative

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

for some $a > 2$ ($a = 3.7$ was suggested as a generally good choice) and $\lambda > 0$, where by $\mathbf{I}(\cdot)$ 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

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\tilde{\beta}_j^{(0)}|) + P'_\lambda(|\tilde{\beta}_j^{(0)}|)(|\beta_j| - |\tilde{\beta}_j^{(0)}|)$$

for some $\beta_j \approx \tilde{\beta}_j^{(0)}$ and $\tilde{\beta}^{(0)}$ is a non-penalized maximum likelihood estimator.

Then the estimator of the proposed procedure is defined as follows

$$\tilde{\beta} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \ell_i(\beta) - n \sum_{j=1}^p P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\beta_j| \right\}$$

Based on Zou and Li (2008), we set our objective function $Q_I(\beta)$ on the given interval I to be

$$Q_I(\beta) = \sum_{i=1}^n \ell_i(\beta) - n \sum_{j=1}^p P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\beta_j|. \quad (3)$$

We refer to the proof of convergence and oracle properties of the one-step SCAD estimator under condition that the penalty parameter λ satisfies

$$\sqrt{n}\lambda \rightarrow \infty \quad \text{and} \quad \lambda \rightarrow 0. \quad (4)$$

with n to be the sample size.

The choice of the parameter λ over a grid of values satisfying conditions (4) is performed with the use of BIC modified for the penalized regression case as follows

$$\text{BIC}_\lambda = \log(\hat{\sigma}_\lambda^2) + q \frac{\log(n)}{n} C_n, \quad (5)$$

where $\hat{\sigma}_\lambda^2 = n^{-1} \text{SSE}_\lambda = n^{-1} \sum_{i=1}^n (y_{i+1} - \hat{\beta}(\lambda)^\top x_i)^2$ and C_n is some positive constant. Here we denote $\hat{\beta}(\lambda)$ explicitly as a function of λ in order to indicate its dependency on the choice of the penalization parameter. Consistency of (5) in selecting a true model was proved by Wang and Leng (2007), where they discussed diverging number of parameters and therefore proposed $C_n = \log\{\log(p)\}$.

Chand (2012) discussed the choice of the constant C_n in a greater detail. For moderate to large sample sizes with a fixed parameter dimension p he showed the BIC performs best with $C_n = \sqrt{n}/p$.

2.2 Test of homogeneity

The adaptive estimation procedure crucially relies on the test of local time-homogeneity of an interval $I = [t_0, T]$. The null hypothesis for I means that the observations follow the parametric model (1) with a fixed parameter β , leading to the quasi-MLE estimate $\tilde{\beta}_I$ and the corresponding fitted log-likelihood $Q_I(\tilde{\beta}_I)$. If the observations follow the parametric model (1) with a parameter β_J for the subinterval $J = [t_0, s]$ and with a different parameter β_{J^c} for the counterpart $J^c = [s + 1, T]$, the test of homogeneity can be performed using the test statistic $T_{I,s}$:

$$\begin{aligned} T_{I,s} &= \max_{\beta_J, \beta_{J^c} \in \Theta} \{Q_J(\beta_J) + Q_{J^c}(\beta_{J^c})\} - \max_{\beta \in \Theta} Q_I(\beta) \\ &= Q_J(\tilde{\beta}_J) + Q_{J^c}(\tilde{\beta}_{J^c}) - Q_I(\tilde{\beta}_I). \end{aligned} \quad (6)$$

Suppose that we have a growing set $I_{t,1} \subset I_{t,2} \subset I_{t,3} \subset \dots \subset I_{t,K}$ of historical interval candidates which satisfy that for each k , the interval $I_{t,k} = [t - m_k + 1, t]$, $m_k = k \cdot c$, $c > 0$, with the right-end point t is fixed. If $I_{t,k}$ is considered as a homogeneous interval, then the observations follow the parametric model (1) should have a fixed parameter on $I_{t,k}$. The quasi-MLE estimate $\tilde{\beta}_{I_{t,k}}$ should have no significant difference with the quasi-MLE estimate $\tilde{\beta}_{I_{t,m}}$ of interval $I_{t,m}$, $m = 1, \dots, k - 1$. Similar to (6), the procedure which checks every interval $I_{t,k}$ on homogeneity relies on the test statistic $T_t^{(k,m)}$, where $m = 1, \dots, k - 1$,

$$T_t^{(k,m)} = Q_{I_{t,m}}(\tilde{\beta}_{I_{t,m}}) + Q_{I_{t,k} \setminus I_{t,m}}(\tilde{\beta}_{I_{t,k} \setminus I_{t,m}}) - Q_{I_{t,k}}(\tilde{\beta}_{I_{t,k}}). \quad (7)$$

The considered problem of testing homogeneity of interval $I_{t,k}$ can be stated as

$$\begin{aligned} H_{0,t} : & \quad Y_t \sim \mathbb{P}_1, \quad \text{for } t \in I_{t,k} \\ H_{1,t} : & \quad \begin{cases} Y_t \sim \mathbb{P}_1, & \text{for } t \in I_{t,k-1} \\ Y_t \sim \mathbb{P}_2, & \text{for } t \in I_{t,k} \setminus I_{t,k-1}, \end{cases} \end{aligned} \quad (8)$$

for $k = 2, \dots, K$ and where $\mathbb{P}_1, \mathbb{P}_2$ are measures defined on a parametric family $\mathbb{P}(\beta)$, i.e. $\mathbb{P}_1, \mathbb{P}_2 \in \{\mathbb{P}(\beta), \beta \in \Theta \subseteq \mathbb{R}^p\}$.

2.3 Adaptive search for the longest homogeneous interval

The algorithm starts with fitting a local model with the quasi-MLE for the interval $I_{t,k}, k = 1, 2, \dots, K$,

$$\tilde{\beta}_{I_{t,k}} = \arg \max_{\beta} Q_{I_{t,k}}(\beta).$$

In default, the smallest interval $I_{t,1}$ is accepted automatically as homogeneous. Then the adaptive procedure checks every larger interval $I_{t,k}, k = 2, \dots, K$ on homogeneity using the test statistics (7). If $I_{t,k}$ has been selected as the homogeneous interval, then the quasi-MLE estimate $\tilde{\beta}_{I_{t,m}}$ on $I_{t,m}, m = 1, 2, \dots, k-1$ should have no significant difference with the quasi-MLE estimate $\tilde{\beta}_{I_{t,k}}$ on $I_{t,k}$. The selected interval \hat{I}_t corresponds to the largest accepted interval $I_{t,\hat{k}}$ with index \hat{k} should satisfy the following conditions:

$$\forall m \in \{1, 2, \dots, \hat{k} - 1\}, T_t^{(\hat{k}, m)} \leq \varsigma_{t,m}; \quad (9)$$

$$\exists m_0 \in \{1, 2, \dots, \hat{k}\}, T_t^{(\hat{k}+1, m_0)} \geq \varsigma_{t,m_0}, \quad (10)$$

where the critical values $\varsigma_{t,m}, m = 1, \dots, K-1$ are discussed in the next section. This procedure then leads to the adaptive estimate $\hat{\beta}_t = \hat{\beta}_{I_{t,\hat{k}}} = \tilde{\beta}_{I_{t,\hat{k}}}$ corresponding to the selected interval $\hat{I}_t = I_{t,\hat{k}}$. The complete description of the procedure includes two steps. (A) Fixing the set-up and the parameters of the procedure. (B) Search for the longest interval of homogeneity.

(A) Set-up parameters

- A1. Select the growing set $I_{t,1}, \dots, I_{t,K}$ of historical interval-candidates where $I_{t,k} = [t - m_k + 1, t]$ and each m_k , which represents the length of the interval $I_{t,k}$, independent of time t .
- A2. Select the critical value $\varsigma_{t,1}, \dots, \varsigma_{t,K-1}$ in (9),(10)

(B) Adaptive search and estimation

- B1. $\hat{\beta}_{I_{t,1}} = \tilde{\beta}_{I_{t,1}}$
- B2. Test the $H_{0,t}$ null hypothesis of homogeneity for the interval $I_{t,k}$ according to the test procedures of (9),(10) and the critical values $\varsigma_{t,m}$ obtained in (A2). If $H_{0,t}$ is rejected, go to (B4). Otherwise proceed with (B3).
- B3. Set $\hat{\beta}_t = \hat{\beta}_{I_{t,k}} = \tilde{\beta}_{I_{t,k}}$. Further, set $k := k + 1$. If $k \leq K$, repeat (B3); otherwise go to (B4).
- B4. $I_{t,k-1}$ is the longest homogeneous interval, and $\hat{I}_t = I_{t,k-1}$, $\hat{\beta}_t = \tilde{\beta}_{I_t} = \tilde{\beta}_{I_{t,k-1}}$. Moreover, if $k \leq K$, $\hat{\beta}_{I_k} = \dots = \hat{\beta}_{I_K} = \hat{I}_t$.

The step (B) performs the search for the longest time-dependent homogeneous interval. Initially, as assumed, $I_{t,1}$ is the shortest homogeneous interval. If the null hypothesis has been accepted and I_{k-1} is accepted as homogeneous, one should continue with $I_{t,k}$ by employing test (9),(10) in step (B2). If the $H_{0,t}$ has been accepted, then $I_{t,k}$ is accepted as homogeneous in step (B3), otherwise the procedure terminates in step (B4) and $I_{t,k-1}$ is the longest homogeneous interval at time t . The longest interval accepted as homogeneous at time t is used for estimation in step (B4). Suppose $I_{t,\hat{k}}$ is the selected homogeneous interval and note that the selected \hat{k} is corresponding with the shorter homogeneous intervals $I_{t,m}$, $m \leq \hat{k}$, the interval selection procedure has selected out of $I_{t,1}, \dots, I_{t,\hat{k}}$ as homogeneous intervals.

2.4 Choice of critical values

The presented method of choosing the interval of homogeneity \hat{I} can be viewed as multiple testing procedure. The critical values for this procedure are selected using the general approach of testing theory: to provide a prescribed performance of the procedure under the null hypothesis in the pure parametric situation. This means that the procedure is trained on the data generated from the pure parametric time-homogeneous model from step (A1). The correct choice in this situation is to set the largest considered interval $I_{t,K}$ to be a homogeneous interval and a choice $I_{t,\hat{k}}$ with $\hat{k} < K$ can be interpreted as a false alarm. We select the minimal critical values ensuring a small probability of such a false alarm. Suppose $I_{t,K}$ is the homogenous interval, then $I_{t,k}$, $k = 1, 2, \dots, K-1$ are also homogenous intervals. Based on (B2), for $T_t^{(l,k)} = Q_{I_{t,k}}(\tilde{\beta}_{I_{t,k}}) + Q_{I_{t,l} \setminus I_{t,k}}(\tilde{\beta}_{I_{t,l} \setminus I_{t,k}}) - Q_{I_{t,l}}(\tilde{\beta}_{I_{t,l}})$, $l = k+1, \dots, K$, there should be a set of critical values $\varsigma_{t,l,k}$, $l = k+1, \dots, K$ satisfy $T_t^{(l,k)} \leq \varsigma_{t,l,k}$, $l = k+1, \dots, K$. Since we are focusing on parameter estimation rather than on hypothesis testing, for simplicity we can set

$$\varsigma_{t,k} = \max\{\varsigma_{t,k+1,k}, \dots, \varsigma_{t,K,k}\}, \quad (11)$$

to make sure that the null hypothesis $H_{0,t}$ for $I_{t,l}$, $l = k+1, \dots, K$ are all accepted.

One way to select the critical values is using the multiplier bootstrap technic which follows from Klochkov et al. (2019). Recall the notations from previous sections that the non-penalized log-likelihood function of a given interval I with sample size n is represented as $L(\beta) = \sum_{i=1}^n \ell_i(\beta)$, i.e. $\ell_i(\beta)$ denotes the parametric logarithmic density of the i -th observation. $\tilde{\beta}^{(0)}$ stands for the non-penalized quasi-MLE estimate. Assume a set of *i.i.d.* scalar random variables u_i , $i = 1, \dots, n$, which satisfy $\mathbb{E}(u_i) = 1$, $\text{Var}(u_i) = 1$ and $\mathbb{E}\{\exp(u_i)\} < \infty$. The

bootstrapped penalized log-likelihood function as follows

$$Q_I^\circ(\beta) = \sum_{i=1}^n u_i \left\{ \ell_i(\beta) - n \sum_{j=1}^p P'_\lambda(|\tilde{\beta}_j^{(0)}|) |\beta_j| \right\}. \quad (12)$$

Denoting $E^\circ(\cdot) = E(\cdot|Y, X, \lambda)$, where $Y = (y_1, \dots, y_n)^\top$, $X = (x_1, \dots, x_n)^\top$, we then can have $E^\circ Q_I^\circ(\beta) = E Q_I(\beta)$ and

$$\arg \max_{\beta} E^\circ Q_I^\circ(\beta) = \arg \max_{\beta} Q_I(\beta) = \tilde{\beta},$$

where $Q_I^\circ(\beta)$ is denoted to be the bootstrapped penalized log-likelihood function on a specific interval I and the corresponding penalized quasi-MLE of the bootstrap world is $\tilde{\beta}_I^\circ = \arg \max_{\beta} Q_I^\circ(\beta)$.

In order to circumvent the problem of penalizing elements of vector β by a different amount in the real and the bootstrap case in a finite sample size situation, we set the parameter λ of the SCAD method to be the same for $Q(\beta)$ and $Q^\circ(\beta)$. Asymptotically, the parameter λ approaches zero, see (4), as needed for the oracle properties of the SCAD estimator, and therefore the condition of equal λ 's is no longer required.

If one wishes to approximate the distribution of the test statistic from (7), it can be done (up to some approximation error in finite samples) by using the bootstrapped penalized likelihood ratio, $l = k + 1, \dots, K$,

$$\begin{aligned} T_t^{\circ(l,k)} &= Q_{I_{t,k}}^\circ(\tilde{\beta}_{I_{t,k}}^\circ) + Q_{I_{t,l} \setminus I_{t,k}}^\circ(\tilde{\beta}_{I_{t,l} \setminus I_{t,k}}^\circ) \\ &\quad - \sup_{\beta \in \Theta} \{Q_{I_{t,k}}^\circ(\beta) + Q_{I_{t,l} \setminus I_{t,k}}^\circ(\beta + \tilde{\beta}_{I_{t,l} \setminus I_{t,k}} - \tilde{\beta}_{I_{t,k}})\}. \end{aligned} \quad (13)$$

Here the term $\tilde{\beta}_{I_{t,l} \setminus I_{t,k}} - \tilde{\beta}_{I_{t,k}}$ is devoted to compensate the biases of the estimators $\tilde{\beta}_{I_{t,k}}^\circ$ and $\tilde{\beta}_{I_{t,l} \setminus I_{t,k}}^\circ$ in the bootstrap world. According to Klochkov et al. (2019), the distribution of $T_t^{\circ(l,k)}$ conditioned on the data mimics the distribution of the original test $T_t^{\circ(l,k)}$ with high probability.

Specifically, let $1 - \alpha \in (0, 1)$ be a determined confidence level of a testing procedure. Based on (11), the approximation of a desired quantile of the distribution of the original test statistic from (7)

$$S_{t,k,\alpha} = \max_{l \in \{k+1, \dots, K\}} \inf\{t \geq 0 : \mathbf{P}(T_t^{(l,k)} > t) \leq \frac{k}{K}\alpha\}$$

can be evaluated as

$$S_{t,k,\alpha}^\circ = \max_{l \in \{k+1, \dots, K\}} \inf\{t \geq 0 : \mathbf{P}^\circ(T_t^{\circ(l,k)} > t) \leq \frac{k}{K}\alpha\}, \quad (14)$$

where \mathbf{P}° denotes the conditional probability given observations of $\{y_{t+1}\}_{t \in I_{t,k}}$, $\{x_t\}_{t \in I_{t,k}}$ and values of λ .

At each point of time t , we implement the multiplier bootstrap into determining quantiles of the test statistic from (7) by simulating n_b sets of *i.i.d.* multipliers u_i , $i = 1, \dots, |I_k|$. For each set of multipliers, we can denote as $u_{b,i}$, $b = 1, \dots, n_b, i = 1, \dots, |I_k|$, $l = k + 1, \dots, K$, computing

$$\begin{aligned} T_t^{b,\circ(l,k)} &= Q_{I_{t,k}}^{b,\circ}(\tilde{\beta}_{I_{t,k}}^{b,\circ}) + Q_{I_{t,l} \setminus I_{t,k}}^{b,\circ}(\tilde{\beta}_{I_{t,l} \setminus I_{t,k}}^{b,\circ}) \\ &\quad - \sup_{\beta \in \Theta} \{Q_{I_{t,k}}^{b,\circ}(\beta) + Q_{I_{t,l} \setminus I_{t,k}}^{b,\circ}(\beta + \tilde{\beta}_{I_{t,l} \setminus I_{t,k}} - \tilde{\beta}_{I_{t,k}})\}. \end{aligned}$$

Therefore, we can get an approximate distribution of $T_t^{(l,k)}$ under the homogeneous situation and can evaluate the respective $(1 - \alpha)\%$ quantile as in (14). Comparing the test statistic from (7) to the defined critical value we either reject the homogeneity hypothesis $H_{0,t}$, if (9) is satisfied, for the given confidence level, or move to the next step and prolong the subsample regarded as homogeneous.

2.5 The small modeling bias and the stability

In this section, we collect basic results describing the quality of the proposed estimation procedure. First, we define the concept of small modeling bias and

discuss although the parametric assumption may not be precisely fulfilled but the PAM process can also be used. Then we discuss certain stability properties of the proposed method.

Without loss of generality, we discuss the quality of estimating the underlying parameter vector β^\diamond by $\tilde{\beta}_{I_k}, k = 1, \dots, K$ based on a certain given interval set $\{I_k\}_{k=1}^K$ instead of $\{I_{t,k}\}_{k=1}^K$ for all t since the discussion for every different interval set which corresponds to a different time point will still be hold. We denote the parameter dimension by p and the sample size by n . We also need to introduce the stochastic part of the likelihood process: $\zeta_{Q_{I_k}}(\beta) = Q_{I_k}(\beta) - \mathbf{E}(Q_{I_k}(\beta))$. Set $D_0^2 = -\nabla_\beta^2 \mathbf{E}(Q_{I_k}(\beta^\diamond))$ and the required assumptions are the basic conditions \mathcal{ED}_0 and \mathcal{L}_r required in Spokoiny and Zhilova (2015).

Assumption 1, Spokoiny and Zhilova (2015), condition \mathcal{ED}_0 . For $1 \leq k \leq K$, there exist a positive-definite symmetric matrix V_0^2 and constants $g > 0, \nu_0 > 1$ such that $\text{Var}\{\nabla_\beta \zeta_{Q_{I_k}}(\beta^\diamond)\} \leq V_0^2, |\varrho| \leq g$ and

$$\sup_{\gamma \in \mathbb{R}^p} \log \mathbf{E}(\exp\{\varrho \frac{\gamma^\top \nabla_\beta \zeta_{Q_{I_k}}(\beta^\diamond)}{\|V_0 \gamma\|}\}) \leq \nu_0^2 \varrho^2 / 2.$$

Assumption 2, Spokoiny and Zhilova (2015), condition \mathcal{L}_r . For $1 \leq k \leq K$, and for each $r \geq r_0$ there exists a value $b(r) > 0$ s.t. $rb(r) \rightarrow \infty$ for $r \rightarrow \infty$ and $\forall \beta : \|D_0(\beta - \beta^\diamond)\| = r$, it holds

$$-2\{\mathbf{E}(Q_{I_k}(\beta)) - \mathbf{E}(Q_{I_k}(\beta^\diamond))\} \geq r^2 b(r),$$

where $D_0^2 = -\nabla_\beta^2 \mathbf{E}(Q_{I_k}(\beta^\diamond))$.

Based on definition of $\beta^\diamond = \arg \max_\beta \mathbf{E}(Q_{I_k}(\beta))$, apparently $\nabla_\beta \mathbf{E}(Q_{I_k}(\beta^\diamond)) = 0$.

The Taylor expansion of $\mathbf{E}(Q_{I_k}(\beta))$ around β^\diamond indicate that,

$$\begin{aligned} -2(\mathbf{E}(Q_{I_k}(\beta)) - \mathbf{E}(Q_{I_k}(\beta^\diamond))) &= (\beta - \beta^\diamond)^\top \mathbf{E}\left(\frac{\partial^2 Q_{I_k}(\bar{\beta})}{\partial \beta \partial \beta^\top}\right)(\beta - \beta^\diamond) \\ &= \|D(\bar{\beta})(\beta - \beta^\diamond)\|^2 \end{aligned}$$

where $\bar{\beta}$ lies between β and β^\diamond , $D(\bar{\beta})^2 = \mathbf{E}\left(\frac{\partial^2 Q_{I_k}(\bar{\beta})}{\partial \beta \partial \beta^\top}\right)$. Since $\tilde{\beta} \xrightarrow{P} \beta^\diamond$, then $-2(\mathbf{E}(Q_{I_k}(\tilde{\beta})) - \mathbf{E}(Q_{I_k}(\beta^\diamond))) \xrightarrow{P} r^2$. Define the loss function $Q(m, k, \beta)$ as,

if $m \leq k$,

$$Q(m, k, \beta) = Q_{I_m}(\tilde{\beta}_{I_m}) + Q_{I_k \setminus I_m}(\tilde{\beta}_{I_k \setminus I_m}) - Q_{I_k}(\beta),$$

if $k < m$, the loss function is,

$$Q(m, k, \beta) = Q_{I_k}(\tilde{\beta}_{I_k}) + Q_{I_m \setminus I_k}(\tilde{\beta}_{I_m \setminus I_k}) - Q_{I_m}(\beta).$$

where $\tilde{\beta}_I$ denotes the quasi-MLE on the interval I . We also define the function $G(m, k)$ as

$$G(m, k) = \begin{cases} Q(m, k, \tilde{\beta}_{I_k}); & m \leq k, \\ Q(m, k, \tilde{\beta}_{I_m}); & m > k. \end{cases}$$

By definition, the value of $G(m, k)$ is non-negative and represents the homogeneity deviation of the maximum log penalized likelihood process from the shorter interval to the longer interval. More generally speaking, $G(m, k)$ can measure the homogeneity difference between the interval I_m and the interval I_k . Based on $G(m, k)$, we have the following results.

Theorem 1. *In the case of SCAD penalty with the penalty parameter λ satisfies (4) and holding the assumptions 1 and 2, for $1 \leq m \leq K$, $1 \leq k \leq K$,*

$$\mathbf{E}_{\beta^\diamond} |Q(m, k, \beta^\diamond)| \leq \mathcal{R}, \quad (15)$$

where $\mathcal{R} > 0$ is a constant.

This result gives a non-asymptotic and fixed upper bound for the risk of quasi-MLE estimate that applies to an arbitrary sample size $|I|$.

2.5.1 Small modeling bias condition

Now we extend our discussion to the situation when the parametric assumption is not precisely fulfilled but the deviation from the true model is small in a modeling bias. To measure the distance of a parametric model from the nonparametric process, we should introduce for every interval I_k and every parametric $\beta \in \Theta$ a random quantity. Therefore, we define the Kullback Leibler divergence between the nonparametric measure and the parametric measure as,

$$\Delta_{I_k}(\beta) = \sum_{i \in I_k} \mathcal{K}\{P_{(y_{i+1}, x_i)}, P_\beta\}$$

where y_{i+1} is the response variable and $x_i = (1, x_{i,1}, \dots, x_{i,p})^\top$ is the covariates vector. We assume that the error terms are independent and identically distributed and denote $P_{(y_{i+1}, x_i)}, P_\beta$ corresponding to the marginal distribution of ε_i with respect to (y_{i+1}, x_i) and $\beta \in \Theta$. To characterize the parametric feature, we define a small modeling bias (SMB) condition

$$\Delta_{I_k}(\beta) \leq \Delta \tag{16}$$

which simply means that, for some $\beta \in \Theta$, $\Delta_{I_k}(\beta)$ is bounded by a small constant with high probability. This implies that the model can be well approximated on the interval I_k by the parametric parameter β .

Theorem 2. *Let the (16) hold for some interval I_k and $\beta \in \Theta$. Then, in the case of SCAD penalty with the penalty parameter λ satisfies (4), we have*

$$\mathbb{E} \log(1 + Q(m, k, \beta)/\mathcal{R}) \leq 1 + \Delta, 1 \leq m \leq K,$$

where $\mathcal{R} > 0$ is the parametric risk bound.

This result shows that the estimation loss $Q(m, k, \beta)$ normalized by the parametric risk \mathcal{R} is stochastically bounded by a constant proportional to $\exp(\Delta)$. If Δ is not large, $\exp(\Delta)$ shows the risk bound of using the parametric modeling for approximation under the SMB condition.

2.5.2 Stability

Informally, the best parametric fit to the underlying model (1) on I_k can be defined by minimizing the value $\mathbb{E}(\Delta_{I_k}(\beta))$ over $\beta \in \Theta$. The oracle index k^* can be estimated as $k^* = \arg \max_k \{\Delta_k(\beta) \leq \Delta\}$ and $\tilde{\beta}_{I_{k^*}}$ can be viewed as the best estimate.

For a fixed $\Delta > 0$, (16) does not hold for $k > k^*$ and unfortunately, the underlying $\Delta_{I_{k^*}}$ is unknown. Therefore, we need to make sure that when our final adaptive estimated \hat{k} overshoots the oracle k^* ($\hat{k} > k^*$), the estimate does not vary too much.

According to the construction, the penalized adaptive procedure of testing a growing sequence of historical time intervals on homogeneity described above provides a stable performance by following the parametric model (1).

Theorem 3. *In the case of overshooting $\hat{k} > k^*$, the estimate is accurate enough in the sense that*

$$\begin{aligned} Q(I_{k^*}, I_{\hat{k} \setminus k^*}, \tilde{\beta}_{I_{\hat{k}}}) &= Q_{I_{k^*}}(\tilde{\beta}_{I_{k^*}}) + Q_{I_{\hat{k}} \setminus I_{k^*}}(\tilde{\beta}_{I_{\hat{k}} \setminus I_{k^*}}) - Q_{I_{\hat{k}}}(\tilde{\beta}_{I_{\hat{k}}}) \\ &\leq \varsigma_{k^*}. \end{aligned}$$

This result provides the prescribed performance that when overshooting $\hat{k} > k^*$, the final estimate will not destroy the risk bond. Moreover, Theorem 4 also implies similar performance below.

Theorem 4. *Let the (16) and hold the assumptions 1 and 2, $1 \leq m \leq K$, then in the case of SCAD penalty with the penalty parameter λ satisfies (4),*

$$\mathbb{E} \log\left(1 + \frac{Q(m, k^*, \beta)}{\mathcal{R}}\right) \leq \Delta + 1 \quad (17)$$

$$\mathbb{E} \log\left(1 + \frac{G(k^*, \hat{k})}{\mathcal{R}}\right) \leq \Delta + 3 + \log(1 + \varsigma_{k^*}/\mathcal{R}). \quad (18)$$

Due to Theorem 4, even for the further steps of the penalized adaptive algorithm with $\hat{k} > k^*$ the homogeneity difference $G(k^*, \hat{k})$ between $I_{\hat{k}}$ and I_{k^*} which is unknown can not be too large. The penalized adaptive estimate $\hat{\beta}$ belongs with a high probability to the confidence set of the oracle estimate $\tilde{\beta}_{I_{k^*}}$.

Next, we move to discuss on which condition, when $\hat{k} < k^*$, the proposed penalized adaptive estimate can mimic the oracle estimate or the final adaptive estimate provides the same (in order) accuracy as the oracle estimate on the basis of available data using the sequential test of homogeneity. Let us propose the last assumption in the paper.

Assumption 3. *Suppose the final adaptive procedures stop at \hat{k} , then there is a constant $\rho > 0$, satisfy*

$$\mathbb{E}_{\beta^\circ}(d(\hat{k}, k)) \leq \rho\mathcal{R}, \quad (19)$$

where

$$d(\hat{k}, k) = \begin{cases} Q(\hat{k}, k, \hat{\beta}_{I_k}); & \hat{k} \leq k, \\ Q(\hat{k}, k, \hat{\beta}_{I_{\hat{k}}}); & \hat{k} > k. \end{cases}$$

Similar to $G(\hat{k}, k)$, $d(\hat{k}, k)$ is also a measure which represents the homogeneity difference between the interval I_k and the final adaptively selected interval $I_{\hat{k}}$. According to the penalized adaptive estimation procedure, we can always select a set of critical values $\varsigma_1, \dots, \varsigma_K$ which can satisfy Assumption 3.

It follows, if $\hat{k} \leq k$, then

$$\begin{aligned}
& \mathbf{E}_{\beta^\circ}(Q(\hat{k}, k, \hat{\beta}_{I_k})) \\
&= \mathbf{E}_{\beta^\circ}(Q_{I_k}(\tilde{\beta}_{I_k}) + Q_{I_k \setminus I_{\hat{k}}}(\tilde{\beta}_{I_k \setminus I_{\hat{k}}}) - Q_{I_k}(\hat{\beta}_{I_k})) \\
&= \mathbf{E}_{\beta^\circ}(Q_{I_k}(\tilde{\beta}_{I_k}) + Q_{I_k \setminus I_{\hat{k}}}(\tilde{\beta}_{I_k \setminus I_{\hat{k}}}) - Q_{I_k}(\tilde{\beta}_{I_k}) + Q_{I_k}(\tilde{\beta}_{I_k}) - Q_{I_k}(\hat{\beta}_{I_k})) \\
&\leq 3\mathcal{R} + \mathbf{E}_{\beta^\circ}(Q_{I_k}(\tilde{\beta}_{I_k}) - Q_{I_k}(\hat{\beta}_{I_k}))
\end{aligned}$$

and when $\varsigma_k \rightarrow \infty, k = 1, \dots, K, \hat{\beta}_{I_k} \rightarrow \tilde{\beta}_{I_k}$, therefore $E_{\beta^\circ}(Q(\hat{k}, k, \hat{\beta}_{I_k}))$ is bounded corresponding to \mathcal{R} .

If $\hat{k} > k$, then

$$\begin{aligned}
& \mathbf{E}_{\beta^\circ}(Q(\hat{k}, k, \hat{\beta}_{I_{\hat{k}}})) \\
&= \mathbf{E}_{\beta^\circ}(Q_{I_k}(\tilde{\beta}_{I_k}) + Q_{I_{\hat{k}} \setminus I_k}(\tilde{\beta}_{I_{\hat{k}} \setminus I_k}) - Q_{I_{\hat{k}}}(\hat{\beta}_{I_k})) \\
&= \mathbf{E}_{\beta^\circ}(Q_{I_k}(\tilde{\beta}_{I_k}) + Q_{I_{\hat{k}} \setminus I_k}(\tilde{\beta}_{I_{\hat{k}} \setminus I_k}) - Q_{I_{\hat{k}}}(\tilde{\beta}_{I_{\hat{k}}})) \\
&\leq 3\mathcal{R}.
\end{aligned}$$

Combine the discussions for both situations, apparently, there exist $\varsigma_1, \dots, \varsigma_K$ which can satisfy (19).

Theorem 5. *Let $\beta \in \Theta$ and $\Delta > 0$ to be such that $\mathbf{E}(\Delta_{I_{k^*}}(\beta)) \leq \Delta$ for some $k^* \leq K$. If the Assumption 3 is hold, then in the case of SCAD penalty with the penalty parameter λ satisfies (4),*

$$\mathbf{E} \log\left(1 + \frac{d(\hat{k}, k^*)}{\mathcal{R}}\right) \leq \Delta + \rho. \tag{20}$$

Under the SMB condition $\mathbf{E}(\Delta_{I_{k^*}}(\beta)) \leq \Delta$ and Assumption 3, Theorem 5 documents that the penalized adaptive estimate does not induce larger (in order) errors into estimation than the oracle estimate. As prescribed, based on assumption 3, the final adaptive estimate claims the same accuracy with the oracle estimate.

3 Simulation Study

In the following part, we perform two simulation studies regarding the use of multiplier bootstrap in critical value selection. In this two simulations, we set three different homogeneous scenarios and in each scenario the active variables are different from the other scenarios. The main point of this simulation study is to show that the newly developed method in this paper can successfully detect the longest homogeneous interval and at the same time can manage of selecting the correct variables.

For the multiplier bootstrap procedure, we propose to use either $u_i \sim \text{Exp}(1)$, $u_i \sim \text{Pois}(1)$ or u_i having a bounded distribution on interval $[0, 4]$ with a pdf

$$f(u_i) = \begin{cases} \frac{3}{14} & \text{if } 0 \leq u_i \leq 1; \\ \frac{1}{12} & 1 < u_i \leq 4. \end{cases} \quad (21)$$

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

$$\{X_i\}_{i=1}^n \sim N_p(0, \Sigma),$$

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 $N(0, 1)$. We consider $n = 500$ to assess performance for $k = 10$ with $m_k = 50 \cdot k$. Number of parameters p is set to be $p = 10$.

In the first simulation, there are tow different parameter homogeneity shifting points which locate at $t = 51$, $t = 101$ and indicate that there are three different homogenous scenarios.

The first set of parameter β is list as

	1	2	3
Scenarios	$1 \leq t < 50$	$51 \leq t < 100$	$101 \leq t \leq 500$
β	(1,1,1,1,1,0,0,0,0,0)	(1,1,1,0,0,0,0,0,0,0)	(1,1,1,1,1,0,0,0,0,0)

Table 1: The homogeneity shifting points locate at $t=51$ and $t=101$.

In the second simulation, there are also three different homogenous scenarios with tow different parameter homogeneity shifting points which locate at $t = 101$ and $t = 151$. The second set of parameter β is listed as

	1	2	3
Scenarios	$1 \leq t < 100$	$101 \leq t < 150$	$151 \leq t < 500$
β	(1,1,1,1,1,0,0,0,0,0)	(1,1,1,0,0,0,0,0,0,0)	(1,1,1,1,1,0,0,0,0,0)

Table 2: The homogeneity shifting points locate at $t=101$ and $t=151$.

For each of the setting we simulated 1000 times and for each time we simulated u_i from the three aforementioned distributions in order to obtain an approximation of the distribution of the penalized likelihood ratio.

For the choice of the penalization parameter λ we defined BIC as in (5) with $C_n = \max(1, \sqrt{n}/p)$. This was specified according to suggestions from Chand (2012).

Results of the first simulation are given in Table 3 and Table 4. In Table 3 one can see the percentage of correctly identifying the homogeneity of a interval which corresponds to a certain parameter homogeneity shifting point. In the simulation, there are two parameter homogeneity shifting points, the homogeneity shifting point a and the homogeneity shifting point b . The first homogeneity shifting point a corresponds to $t = 51$. With the high probability of correctly identifying the homogeneity shifting point a , the PAM method can, with high probability, correctly identify the homogeneity of the interval $1 \leq t < 50$. The last homogeneity shifting point b corresponds to $t = 101$. With the high probability of correctly identifying the last homogeneity shifting point, the PAM method can

also identify the homogeneity of the intervals: $51 \leq t < 100$ and $101 \leq t < 500$ with high probability.

In Table 4, one can see the percentage of correctly selecting the active variables in a certain scenario. In the simulation, there are three scenarios, since there is no parameter inhomogeneity in the first scenario, PAM performs similar in variable selection compared to the normal SCAD method. Therefore, we are focusing only on the second and the third scenarios in which there are not only shift in parameters, but also there are changes in active variables. Based on Table 4, PAM method have high probability of correctly selecting the active variables when the situation come across with changes both in active variables and in parameters.

Results of the second simulation are given in Table 5 and Table 6. In simulation two, the first parameter homogeneity shifting point located at $t = 101$ rather than at $t = 51$ and the second parameter homogeneity shifting point located at $t = 151$ rather than at $t = 101$. In this situation, the PAM method can also, with high probability, correctly identify the homogeneity of the intervals: $1 \leq t < 100$, $101 \leq t < 150$ and $151 \leq t < 500$.

In Table 6, similar to simulation one, we are also showing the percentage of correctly selecting the active variables in a certain scenario by using PAM. Based on Table 6, we can conclude that the PAM method can correctly select the active variables in the situation of parameter shifting with high probability.

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 the 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.

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

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

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

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

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

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

Finally, for the excess log returns we write

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

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.

$$rx_{t+1}^{(m)} = \beta_0^{(m)} + \beta_1^{(m)} y_t^{(1)} + \beta_2^{(m)} f_t^{(2)} + \dots + \beta_5^{(m)} f_t^{(5)} + \varepsilon_{t+1}^{(m)}, \quad (22)$$

for $m = 2, 3, 4, 5$. Further they specified a single factor for modelling expected excess returns for all k as follows

$$rx_{t+1}^{(m)} = b_m(\gamma_0 + \gamma_1 y_t^{(1)} + \gamma_2 f_t^{(2)} + \dots + \gamma_5 f_t^{(5)}) + \varepsilon_{t+1}^{(m)}, \quad (23)$$

where vector $\gamma = (\gamma_0, \dots, \gamma_5)^\top$ is the same for all $m = 2, 3, 4, 5$ and b_m satisfies $\frac{1}{4} \sum_{m=2}^5 b_k = 1$ in order to allow for a separate identification of the given set of parameters.

In what follows we deviate from the cited work in the sense that we consider inclusion of macro variables, what was shown to improve the model fit and its forecasting performance, see Ludvigson and Ng (2009). This serves our purpose, since with PAM we can include a large number of covariates and reduce the dimension of the model afterwards.

The factor model of Ludvigson and Ng (2009) is defined by the following

$$rx_{t+1}^{(m)} = \alpha^\top F_t + \beta^\top Z_t + \varepsilon_{t+1}, \quad (24)$$

where F_t is an $(r \times 1)$ vector of latent common factors, α a corresponding vector of factor loadings, Z_t is a $(s \times 1)$ vector of directly observable covariates and β_t its associated parameter vector. For their empirical study, they chose the number of

estimated factors $r = 8$ and considered two models, one with the single forward factor of Cochrane and Piazzesi (2005) included and one without. According to a minimized BIC criterion the subset of either five, for the first case, or six, for the latter case, common factors was selected. The description of their estimation method is omitted here and can be found in the original work of Ludvigson and Ng (2009). Later in the section we take all of the models (22), (23) and (24), both with five and six factors, as baselines with which we compare the forecasting performance of PAM.

For our proposed model we use the raw data of Jurado et al. (2015), where we select a subset of collected macro variables and for the sake of comparison with the models of Ludvigson and Ng (2009) we follow their transformation suggestions and apply them to the raw dataset. The selected predictors can be classified into three groups, which capture the situation on the bond market, the stock market or describe the macroeconomic environment. Complete list of the used macro variables and their transformations can be found in Table 7. In addition to the macroeconomic variables, we also use log yield and log forward rates defined previously as explanatory variables. Altogether the predictors yield a dimension of $p = 36$. The time span over which the sample of covariates was taken is January 1960 to December 2010 and the observations of bond risk premia as dependent variables were taken from January 1961 to December 2011.

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

$$r x_{t+1}^{(m)} = \beta_{0t}^{(m)} + \beta_{1t}^{(m)\top} f_t + \beta_{2t}^{(m)\top} M_t + \varepsilon_{t+1}^{(m)},$$

where $f_t = (y_t^{(1)}, f_t^{(2)}, \dots, f_t^{(5)})^\top$ and vector M_t defines all of the macro variables from Table 7. Please note that in our model we allow for time-variation of the vector of parameters β_t .

For our empirical study, we take $\{I_k\}_{k=1}^5$ and consider the increments between two adjacent subintervals to be 4 years, i.e. $m_k = 48 \cdot k, k = 1, \dots, 5$, for monthly observations. This comes from the fact, that business cycles as defined by The National Bureau of Economic Research (NBER) last on average around 5.5 years, therefore reducing this span and assuming it as homogeneous sample is regarded as a reasonable choice. Moreover, from the ADNS model of Chen and Niu (2014), where they focused on the short term explanation of the macroeconomic situation, one can see that the average length of the stable subsample is around 2.5-3.5 years. The specified length of the subintervals and their increments should also yield better coverage probabilities of the multiplier bootstrap based confidence regions for the estimated parameters.

As mentioned previously, in our study we compare the performance of PAM with formerly described models of Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009). Authors of both works considered the time span ranging from January 1964 to December 2003, which might have influenced their results. Replicating the forward factor from Cochrane and Piazzesi (2005) and the reasoning behind using it, we come to a conclusion that time-variation of coefficients in this type of real data cannot be omitted. The ‘tent-shape’ characteristic of the parameters corresponding to yields and forward rates no longer holds if one considers a longer time span, as can be seen in Figure 1. Moreover, the line shapes differ across the maturities of considered bonds. Therefore, in order to thoroughly analyse and compare the performance of the stated baseline models and PAM, we use both lengths of the data, January 1964 to December 2003 and January 1961 to December 2011.

Firstly, we compare the fitting performance of the used methods. As measures of the model accuracy we compute the root mean squared error (RMSE), the mean absolute error (MAE), R^2 and R_{adj}^2 for 1-year excess log returns of 2-, 3-, 4-

and 5-year bonds as dependent variables. For calculation of adjusted R^2 we use the number of covariates or factors as number of parameters in case of baseline models and average number of nonzero coefficients over the whole time range in case of PAM model. For the calibration of critical values, we use 1 000 multipliers with the Pois(1) distribution, since in the simulation section they yielded the best coverage probability results in the small sample case. For the homogeneity testing the confidence level of 95 % was applied.

The fitting procedure summary can be found in Table 8, where we use abbreviations CP, CP1F, LN5F and LN6F for models (22), (23) and (24), respectively, with five or six factors used in the latter case. Here we omit the single factor representation of five and six factor models of Ludvigson and Ng (2009) since, as shown by the authors, they yield very similar results to those where each factor is considered as a separate covariate. Graphical comparison for the case of 2-year bond excess returns is presented in Figures 2 and 3.

As can be seen from Table 8, the PAM method performs the best in terms of used fitting performance measures. On average it reduces the RMSE and MSE to one fourth of the RMSE and MSE of the models used by Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009). The coefficient of determination R^2 and its adjusted value R^2_{adj} attain values as high as 98 %, what greatly outperforms the baseline models. This performance largely owes to the possibility of time variation in coefficients throughout the whole time span of the data and use of many covariates without grouping them into common factors.

For the shorter time span (from January 1964 to December 2003) the average length of homogeneous time intervals is 4.4, 6.7, 6.7, 5.7 for the 2-, 3-, 4- and 5-year bond excess returns, respectively. This is in agreement with the findings of Chen and Niu (2014), where a short memory view of the yield curve modelling

has been promoted.

For the 2-year bond excess returns, the homogeneous intervals were shortest, i.e. the change point was found between all of the time intervals apart from the time spans between the years of 1980-83 and 1984-87. The average number of selected covariates was 11.5, with minimum 3 and maximum 19. In all of the sub-samples, the 2-year forward rate $f_t^{(2)}$ and spread between Moody's Baa corporate bond yield and Federal Funds interest rate were chosen as explanatory variables. From the rest of the possible covariates, the ones with acronyms *sfygt1*, *sfygt5*, and *sfyaaac* were chosen in more than 80 % of the sub-samples, and thus, modelling the development of 2-year bond excess returns mainly by spread between Moody's corporate bond yield or US Treasury Bills interest rates and Federal Funds interest rate.

The model for 3-year bond excess returns yields average significant parameter dimension of 13.8 with a minimum of 9 and maximum of 21. Number of change points detected is 5 and the covariates selected in more than 80 % of cases are $f_t^{(2)}$, $f_t^{(3)}$, *fygt5*, *sfygm6*, *sfygt1*, *sfygt5*, *sfygt10* and *sfyaaac*. Hence, the discussed model chooses similar covariates to those for the 2-year bond excess returns with a use of different maturities, which can be understood as the effect of longer time to maturity of the dependent variable.

The results for 4- and 5-year bond excess returns are very similar to the previous ones with an average number of chosen covariates 14.5 in both of the cases. The set of chosen macro variables in most of the sub-samples was very similar to the models above. However, the pattern of chosen forward rates broke down in case of the 5-year bond excess returns, where the yield $y_t^{(1)}$ together with the forward rates $f_t^{(2)}$, $f_t^{(3)}$ were chosen in more than 80 % of the sub-samples. In case of 4-year bond excess return these were the forward rates $f_t^{(2)}$ and $f_t^{(4)}$.

Investigation of the longer time period spanning between January 1961 and December 2011 yields very similar results to those reported above and thus we omit its lengthy description.

Comparison of our model fitted by the PAM method to the baseline models (22), (23) and (24) can be summarized in a few highlights. First of all, our findings align with the assertion of Cochrane and Piazzesi (2005) by selecting forward rates as the significant explanatory variables in most of the sub-samples and hence proving their power in modelling the development of bond risk excess premia. However, we can see, that the most significant are the forward rates over the periods which are included in the maturity of the specified bond, in contrast to the single factor including all of the forward rates. Second, the conclusions of Ludvigson and Ng (2009) are also present in our model, since the specific macro variables are almost always included in the homogeneous models providing us with a better fit compared to the single forward factor model of Cochrane and Piazzesi (2005). Last, but not least, allowing the coefficient loadings to vary over time we capture the unstable situation over the markets, where the stationarity assumption is violated.

As the target of our interest lies rather in forecasting than in in-sample fitting performance of PAM, we move our focus on prediction over a one-year horizon ahead. We use the data sample over a period from January 1961 to December 2011 and we make an out-of-sample forecast with a starting point December 2000. For the model fitting we use all of the observed data prior to January 2001 and predict excess bond returns over a one-year horizon, i.e. we predict the values corresponding to December 2001. Then we recursively adjust the fitted models to the sample including January 2001 and predict over next year (January 2002), etc. For the evaluation of forecasting accuracy we use root mean squared prediction error (RMSPE) and mean absolute prediction error (MAPE)

as suitable measures. For the calibration of PAM, we again use 1 000 multipliers generated from the Pois(1) distribution and choose 99 % as a confidence level for the homogeneity testing. Table 9 collects all of the results for the three compared methods. Graphical output can be seen in Figure 4.

From Table 9 it is visible that the PAM method outperforms all of the models (22), (23) and (24) when one deals with forecasting of excess bond returns over a 1-year period ahead. It achieves the best forecasting performance in terms of RMSPE and MAPE, reducing it by 24 - 50 % depending on the baseline model chosen. This effect owes to the possibility of time variation of coefficient loadings, which can capture the instability over the financial markets. Particularly in the forecasting period used in this section, where the global financial crisis of the years 2008 - 2009 is included. In Figure 4 the abrupt rise of the observed values of excess bond premia for all of the investigated maturities related to the period of the early 2000s after the Dotcom Bubble and the years of the global financial crisis is detectable.

This is a natural behaviour of the market since the investors have to be compensated for the risk with higher bond risk premia. Looking at the Figure 4 one can see that whereas the model of Cochrane and Piazzesi (2005) fails to capture the parameter inhomogeneity completely, the six-factor model of Ludvigson and Ng (2009) and PAM react to the development of the curve.

According to the Federal Reserve announcements, the Federal Reserve started buying billions of mortgage-backed securities in late 2008, and by June 2010, the amount of bank debt, mortgage-backed securities, and Treasury notes reached its peak of 2.1 trillion USD. This kind of stimulation pushed the economy to grow and shifted the expectations of the market, the bond risk premia stopped increasing and had a decreasing trend at the early stage of 2009. We can see

that PAM manages to forecast this period more promptly than the investigated alternative methods.

Concluding from Figure 4 we can say that PAM captures the upward and downward turns of the excess bond returns more efficiently than the alternatives used for comparison, since its core assumption is the non-stationary of the modelled data. Indeed, the average lengths of the homogeneous intervals used for the 1-year ahead prediction are 4.8, 5.0, 5.4 and 5.3 years for the 2-, 3-, 4- and 5-year bond excess returns, respectively, which is in a large contrast to the whole sample size of the Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009) methods. The covariates, which are mostly used for the 1-year ahead prediction of the respective excess bond returns are the ones, which were used for the in-sample fit, what is a natural result.

With the foregoing summary of the PAM performance at hand, we conclude that our proposed method provides a useful tool for modelling time variation of the coefficient loadings especially when dealing with forecasting of non-stationary and possibly high-dimensional models.

5 Concluding Remarks

In the present paper we proposed a novel approach for dealing with a challenging statistical inference arising with the occurrence of big data. The introduced Penalized Adaptive Method (PAM) can capture the non-stationarity and conduct effective model reduction simultaneously.

The performance of PAM was argued theoretically as well as practically, where simulation methods were implemented. For the real data application we chose the problem of excess bond risk premium modelling and its forecastability, where

we compared PAM with a several baseline models based on the work of Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009). These authors developed a technique, which is useful from the practitioner's point of view because of its simplicity and good interpretability. However, these models omit two important aspects of modeling the variations which include the variable-variation and parameter-variation.

It is well known that the expectations in the market together with the government policies can shift the whole economic trend. Therefore, a new method which is not only capable of providing higher forecasting accuracy but also able to identify the macro-covariates useful in determining the bond excess returns will certainly have strong economic implications. Our proposed Penalized Adaptive Method fits perfectly in the gap between methods dealing with nonstationarity and methods of variable selection.

It is intuitive that the expectations and the government policies are changing in different periods of economic cycles and hence cause the time-variation of the economic fundamentals. By using PAM, which is designed to identify significant variables and detect homogeneous intervals simultaneously, the simplicity and interpretability of the model is preserved whereas its fit and forecasting ability can be largely outperformed as seen from its in-sample and out-of-sample performance. Mainly, it reduces the root mean squared prediction error and mean absolute prediction error by up to 50 % of the models using whole data sample for the model fitting. This improvement comes at a cost of a more computationally intensive method, but its gains should be of interest for any type of users.

The proposed PAM method is fully data-driven and therefore can be applied to variety of problems which include the high dimensional economic situations occurring in the real world.

Appendix

Since $Q_I(\beta_I) = L_I(\beta_I) - |I| \sum_{j=1}^p P_\lambda(|\tilde{\beta}_{I,j}^{(0)}|) |\beta_{I,j}|$, if we want to discuss the moment bounds for $Q(m, k, \beta^\diamond)$, where β^\diamond is the true underlying parameter, we can firstly refer to the moment bounds for the likelihood ratio process obtained by Lemma 2.7 in Spokoiny et al. (2013).

Lemma 1. (Spokoiny et al. (2013), Lemma 2.7) *Holding the assumptions 3 and 7 and define the positive loss function $|L_I(\tilde{\beta}_I, \beta^\diamond)|^r, r > 0$, then,*

$$\mathbf{E}_{\beta^\diamond} |L_I(\tilde{\beta}_I, \beta^\diamond)|^r < \mathcal{R}_r,$$

where $L_I(\cdot)$ is the no-penalized likelihood function of the given interval I , $L(\tilde{\beta}, \beta^\diamond) = L(\tilde{\beta}) - L(\beta^\diamond)$.

Theorem 1. *In the case of SCAD penalty with the penalty parameter λ satisfies (4) and holding the assumptions 1 and 2, for $1 \leq m \leq K$,*

$$\mathbf{E}_{\beta^\diamond} |Q(m, k, \beta^\diamond)| \leq \mathcal{R}, \quad (25)$$

where $\mathcal{R} > 0$ is a constant.

Proof of Theorem 1.

$$\begin{aligned} & \mathbf{E}_{\beta^\diamond} |Q_{I_{k-1}}(\tilde{\beta}_{I_{k-1}}) + Q_{I_k \setminus I_{k-1}}(\tilde{\beta}_{I_k \setminus I_{k-1}}) - Q_{I_k}(\beta^\diamond)| \\ & \leq \mathbf{E}_{\beta^\diamond} |Q_{I_{k-1}}(\tilde{\beta}_{I_{k-1}}) - Q_{I_{k-1}}(\beta^\diamond)| + \mathbf{E}_{\beta^\diamond} |Q_{I_k \setminus I_{k-1}}(\tilde{\beta}_{I_k \setminus I_{k-1}}) - Q_{I_k \setminus I_{k-1}}(\beta^\diamond)| \\ & = H_1 + H_2, \end{aligned}$$

where,

$$\begin{aligned} H_1 &= \mathbf{E}_{\beta^\diamond} |Q_{I_{k-1}}(\tilde{\beta}_{I_{k-1}}) - Q_{I_{k-1}}(\beta^\diamond)| \\ H_2 &= \mathbf{E}_{\beta^\diamond} |Q_{I_k \setminus I_{k-1}}(\tilde{\beta}_{I_k \setminus I_{k-1}}) - Q_{I_k \setminus I_{k-1}}(\beta^\diamond)| \end{aligned}$$

The term of H_1 can be expanded into the following,

$$\begin{aligned} H_1 &\leq \mathbf{E}_{\beta^\circ} |L_{I_{k-1}}(\tilde{\beta}_{I_{k-1}}) - L_{I_{k-1}}(\beta^\circ)| \\ &\quad + \mathbf{E}_{\beta^\circ} ||I_{k-1}| \sum_{j=1}^p (P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\tilde{\beta}_{I_{k-1},j}| - P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\beta_j^\circ|)|. \end{aligned}$$

According to SCAD penalty,

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

we can see that, if $\tilde{\beta}_j^{(0)} \geq a\lambda$ then $P'_\lambda(|\tilde{\beta}_j^{(0)}|) = 0$. Because, $\tilde{\beta}_j^{(0)}$ is fixed and when $\tilde{\beta}_j^{(0)} \neq 0$, $\lambda \rightarrow 0$, there exist $N > 0$, satisfy that $|I_{k-1}| > N$, $P'_\lambda(|\tilde{\beta}_j^{(0)}|) = 0$.

If $|I_{k-1}| > N$, $\mathbf{E}_{\beta^\circ} ||I_{k-1}| \sum_{j=1}^p (P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\tilde{\beta}_{I_{k-1},j}| - P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\beta_j^\circ|) = 0$, therefore, $H_1 = \mathbf{E}_{\beta^\circ} |Q_{I_{k-1}}(\tilde{\beta}_{I_{k-1}}) - Q_{I_{k-1}}(\beta^\circ)|$ is bounded according to Lemma 1.

If $|I_{k-1}| \leq N$, then apparently, $\mathbf{E}_{\beta^\circ} ||I_{k-1}| \sum_{j=1}^p (P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\tilde{\beta}_{I_{k-1},j}| - P'_\lambda(|\tilde{\beta}_j^{(0)}|)|\beta_j^\circ|)$ is bounded.

Therefore, H_1 is bounded. For the same reason, H_2 is also bounded. \square

Lemma 2. Let P, P_0 , be two measures s.t. $\mathbf{E} \log(dP/dP_0) \leq \Delta < \infty$, and for any random variable \mathfrak{z} , with $\mathbf{E}(\mathfrak{z}) < \infty$, we have $\mathbf{E} \log(1 + \mathfrak{z}) \leq \Delta + \mathbf{E}_0(\mathfrak{z})$.

Proof of Lemma 2. $f(x) = xy - x \log(x) + x$ attains maximum at the point $x = e^y$, thus $f(x) \leq f(e^y)$, and we have $xy \leq x \log(x) - x + e^y$. Let $x = dP/dP_0$ and $y = \log(1 + \mathfrak{z})$,

$$\begin{aligned} \mathbf{E}_0 dP/dP_0 \log(1 + \mathfrak{z}) &= \mathbf{E}(\log(1 + \mathfrak{z})) \\ &\leq \mathbf{E}_0(dP/dP_0 \log(dP/dP_0) - dP/dP_0 + 1 + \mathfrak{z}) \\ &\leq \Delta + \mathbf{E}_0(\mathfrak{z}) \end{aligned}$$

\square

Theorem 2. Let the (16) hold for some interval I_k and $\beta \in \Theta$. Then, in the case of SCAD penalty with the penalty parameter λ satisfies (4), we have

$$\mathbf{E} \log(1 + Q(m, k, \beta)/\mathcal{R}) \leq 1 + \Delta, 1 \leq m \leq K,$$

where $\mathcal{R} > 0$ is the parametric risk bound.

Proof of Theorem 2. Based on Theorem 1 and Lemma 2,

$$\begin{aligned} \mathbb{E} \log\left(1 + \frac{Q(m, k, \beta)}{\mathcal{R}}\right) &\leq \Delta + \mathbb{E}_\beta\left(\frac{Q(m, k, \beta)}{\mathcal{R}}\right) \\ &\leq \Delta + 1. \end{aligned}$$

□

Theorem 3. *In the case of overshooting $\hat{k} > k^*$, the estimate is accurate enough in the sense that,*

$$Q(I_{k^*}, I_{\hat{k} \setminus k^*}, \tilde{\beta}_{I_{\hat{k}}}) \leq \varsigma_{k^*}.$$

Proof of Theorem 3. The result directly follows the adaptive procedure that,

$$\begin{aligned} Q(I_{k^*}, I_{\hat{k} \setminus k^*}, \tilde{\beta}_{I_{\hat{k}}}) &= Q_{I_{k^*}}(\tilde{\beta}_{I_{k^*}}) + Q_{I_{\hat{k}} \setminus I_{k^*}}(\tilde{\beta}_{I_{\hat{k}} \setminus I_{k^*}}) - Q_{I_{\hat{k}}}(\tilde{\beta}_{I_{\hat{k}}}) \\ &\leq \varsigma_{k^*}. \end{aligned}$$

□

Theorem 4. *Let the (16) and holding the assumptions 1 and 2, then in the case of SCAD penalty with the penalty parameter λ satisfies (4),*

$$\begin{aligned} \mathbb{E} \log\left(1 + \frac{Q(m, k^*, \beta)}{\mathcal{R}}\right) &\leq \Delta + 1 \\ \mathbb{E} \log\left(1 + \frac{G(k^*, \hat{k})}{\mathcal{R}}\right) &\leq \Delta + 3 + \log(1 + \varsigma_{k^*}/\mathcal{R}). \end{aligned}$$

Proof of Theorem 4. According to Theorem 1 and Lemma 2,

$$\begin{aligned} \mathbb{E} \log\left(1 + \frac{Q(m, k^*, \beta)}{\mathcal{R}}\right) &\leq \Delta + \mathbb{E}_\beta\left(\frac{Q(m, k^*, \beta)}{\mathcal{R}}\right) \\ &\leq \Delta + 1. \end{aligned}$$

Recall that

$$G(k^*, \hat{k}) = \begin{cases} Q(k^*, \hat{k}, \tilde{\beta}_{I_{k^*}}) & \hat{k} \leq k^* \\ Q(k^*, \hat{k}, \tilde{\beta}_{I_{\hat{k}}}) & k^* < \hat{k} \end{cases}$$

Based on Lemma 2, we have,

$$\begin{aligned}
\mathbb{E} \log\left(1 + \frac{G(k^*, \hat{k})}{\mathcal{R}}\right) &= \mathbb{E} \log\left(1 + \frac{G(k^*, \hat{k})}{\mathcal{R}}\right) \mathbf{1}(\hat{k} \leq k^*) \\
&\quad + \mathbb{E} \log\left(1 + \frac{G(k^*, \hat{k})}{\mathcal{R}}\right) \mathbf{1}(k^* < \hat{k}) \\
&\leq \Delta + \mathbb{E}_{\beta^\circ} \left(\frac{|Q_{I_{\hat{k}}}(\tilde{\beta}_{I_{\hat{k}}}) - Q_{I_{\hat{k}}}(\beta^\circ)| + |Q_{I_{k^*} \setminus I_{\hat{k}}}(\tilde{\beta}_{I_{k^*} \setminus I_{\hat{k}}}) - Q_{I_{k^*} \setminus I_{\hat{k}}}(\beta^\circ)|}{\mathcal{R}} \right) \\
&\quad + \mathbb{E}_{\beta^\circ} \left(\frac{|Q_{I_{k^*}}(\tilde{\beta}_{I_{k^*}}) - Q_{I_{k^*}}(\beta^\circ)|}{\mathcal{R}} \right) + \log(1 + \varsigma_{k^*}/\mathcal{R}) \\
&\leq \Delta + 3 + \log(1 + \varsigma_{k^*}/\mathcal{R})
\end{aligned}$$

□

Theorem 5. *Let $\beta \in \Theta$ and $\Delta > 0$ to be such that $\mathbb{E}(\Delta_{I_{k^*}}(\beta)) \leq \Delta$ for some $k^* \leq K$. If the Assumption 3 is hold, then in the case of SCAD penalty with the penalty parameter λ satisfies (4),*

$$\mathbb{E} \log\left(1 + \frac{d(\hat{k}, k^*)}{\mathcal{R}}\right) \leq \Delta + \rho. \tag{26}$$

Proof of Theorem 5. Based on Lemma 2,

$$\begin{aligned}
\mathbb{E} \log\left(1 + \frac{d(\hat{k}, k^*)}{\mathcal{R}}\right) &\leq \Delta + \mathbb{E}_{\beta^\circ} \left(\frac{d(\hat{k}, k^*)}{\mathcal{R}} \right) \\
&\leq \Delta + \rho.
\end{aligned}$$

□

Acknowledgements

Financial supports from the Deutsche Forschungsgemeinschaft via CRC ‘‘Economic Risk’’ and IRTG 1792 ‘‘High Dimensional Non Stationary Time Series’’, Humboldt-Universität zu Berlin, from the National Natural Science Foundation of China (71528008) ‘‘Adaptive Methods for real-time forecasting and monitoring of macroeconomic and financial markets indicators’’ and from the China Scholarship Council(201806310176) are gratefully acknowledged.

References

- Chand, S. (2012). On Tuning Parameter Selection of Lasso-Type Methods - A Monte Carlo Study, Proceedings of 9th International Bhurban Conference on Applied Sciences & Technology: 120–129.
- Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso Estimators, Journal of the American Statistical Association **106**(494): 608–625.
- Chatterjee, A. and Lahiri, S. N. (2013). Rates of Convergence of the Adaptive Lasso Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap, The Annals of Statistics **41**(3): 1232–1259.
- Chen, Y., Härdle, W., Pigorsch, U. (2010). Localized Realized Volatility Modelling, Journal of the American Statistical Association **105**: 1376–1393.
- Chen, Y. and Niu, L. (2014). Adaptive Dynamic Nelson-Siegel Term Structure Model with Applications, Journal of Econometrics **180**(1): 98–115.
- Chen, Y. and Spokoiny, V. (2015). Modeling Nonstationary and Leptokurtic Financial Time Series, Econometric Theory **31**(4): 703–728.
- Chernozhukov, V., Härdle, W. K., Huang, C. and Wang, W. (2018). LASSO-Driven Inference in Time and Space, Annals of Statistics, Revise & Resubmit, *arXiv preprint arXiv:1806.05081*.
- Cochrane, J. H. and Piazzesi, M. (2005). Bond Risk Premia, American Economic Review **95**(1): 138–160.
- Fama, E. F. and Bliss, R. R. (1987). The Information in Long-Maturity Forward Rates, American Economic Review **77**(4): 680-692.
- Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likeli-

- hood and its Oracle Properties, *Journal of the American Statistical Association* **96**(456): 1348–1360.
- Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters, *The Annals of Statistics* **32**(3): 928–961.
- Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software* **33**(1): 1–22.
- Härdle, W. K. and Mammen, E. (1993). Comparing Nonparametric versus Parametric Regression Fits, *Annals of Statistics* **21**(4): 1926–1947.
- Härdle, W. K., Wang, W. and Yu, L. (2016). TENET: Tail-Event driven NETWORK risk, *Journal of Econometrics* **192**(2): 499–513.
- Jurado, K., Ludvigson, S. C. and Ng, S. (2015). Measuring Uncertainty, *American Economic Review* **105**(3): 1177–1216.
- Klochkov, Y., Härdle, W. K. and Xu, X. (2019). Localizing Multivariate CAViaR, IRTG 1792 discussion paper **2019-007**.
- Kim, Y., Choi, H. and Oh, H. S. (2008). Smoothly Clipped Absolute Deviation in High Dimensions, *Journal of the American Statistical Association* **103**(484): 1665–1673.
- Kwon, S. and Kim, Y. (2012). Large Sample Properties of the SCAD-Penalized Maximum Likelihood Estimation on High Dimensions, *Statistica Sinica* **22**(2): 629–653.
- Ludvigson, S. C. and Ng, S. (2009). Macro Factors in Bond Risk Premia, *The Review of Financial Studies* **22**(12): 5027–5067.

- Niu, L., Xu, X. and Chen, Y. (2017). An Adaptive Approach to Forecasting Three Key Macroeconomic Variables for Transitional China, *Economic Modelling* **66**: 201–213.
- 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, *Probability Theory and Related Fields* **135**(3): 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/>.
- Spokoiny, V. (2017). Penalized Maximum Likelihood Estimation and Effective Dimension, *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* **53**(1): 389–429.
- Spokoiny, V. and Zhilova, M. (2015). Bootstrap Confidence Sets Under Model Misspecification, *The Annals of Statistics* **43**(6): 2653–2675.
- Spokoiny, V., Wang, W and Härdle, W.K. (2013). Local Quatile Regression, *Journal of Statistical Planning and Inference* **143**(7):1109–1129.
- Suvorikova, A., Spokoiny, V. and Buzun, N. (2015). Multiscale Parametric Approach for Change Point Detection, *Information Technology and Systems 2015*: 979–996.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society: Series B* **58**(1): 267–288.
- Wang, H. and Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation, *Journal of the American Statistical Association* **102**(479): 1039–1048.

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**(476): 1418–1429.

Zou, H. and Li, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models, *The Annals of Statistics* **36**(4): 1509–1533.

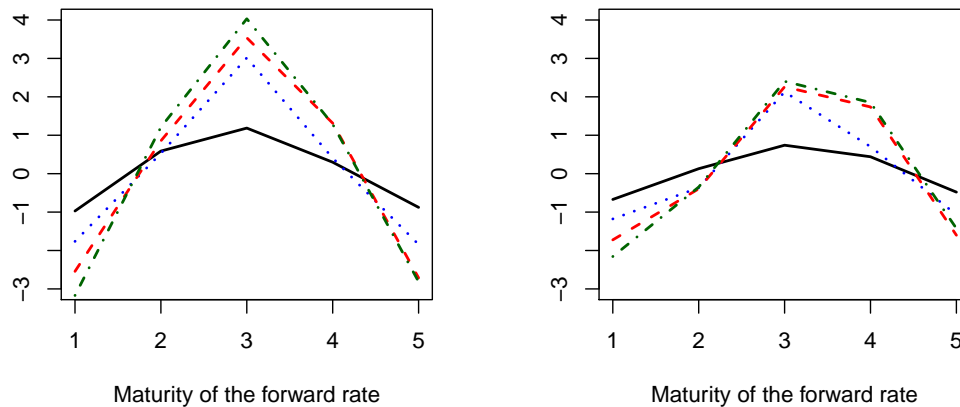
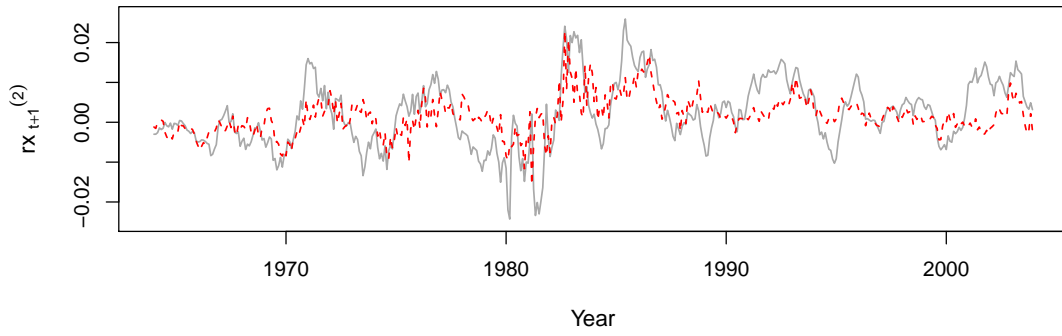
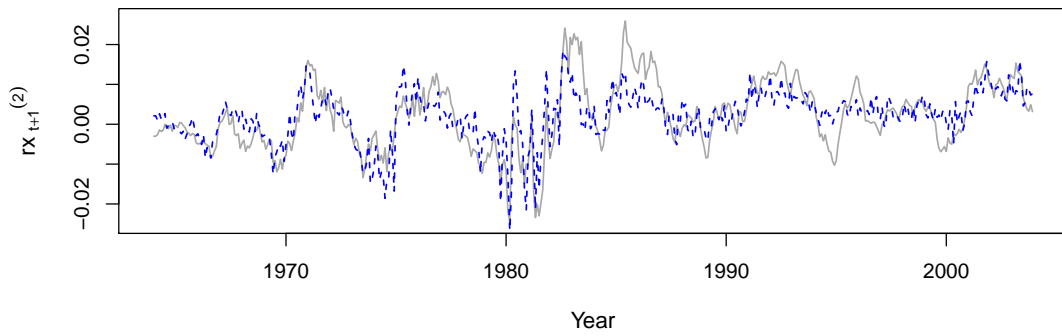


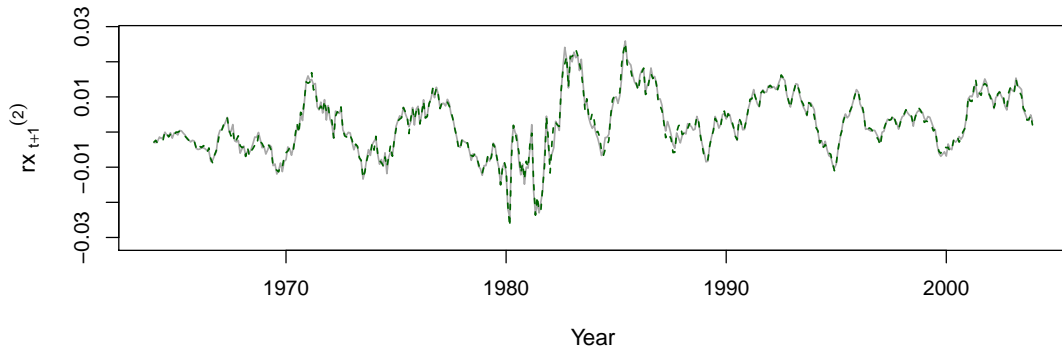
Figure 1: Regression coefficients of 1-year excess log returns on forward rates for 011964-122003 (left) and for 011961-122011 (right). Solid, dotted, dashed and dot-dashed lines denote 2-, 3-, 4- and 5-year maturity of the bond, respectively.



(a) Forward factor model



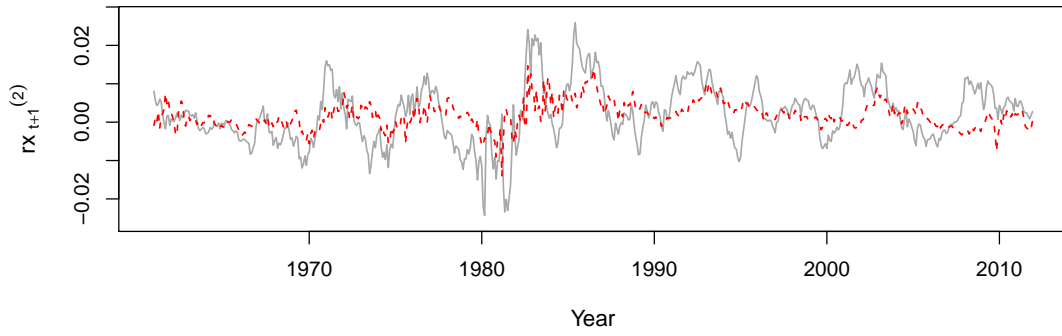
(b) Model with six macro factors



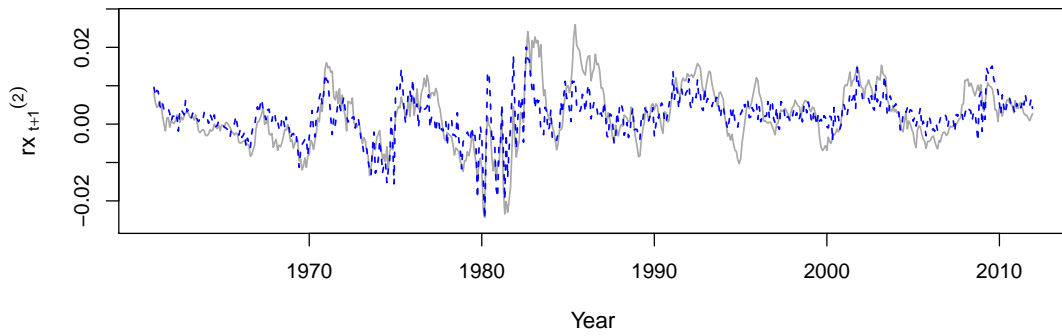
(c) PAM

Figure 2: Fitted CP1F, LN6F and PAM models (dashed) with observed values of 2-year bond excess log returns (solid) for the time period 011964-122003.

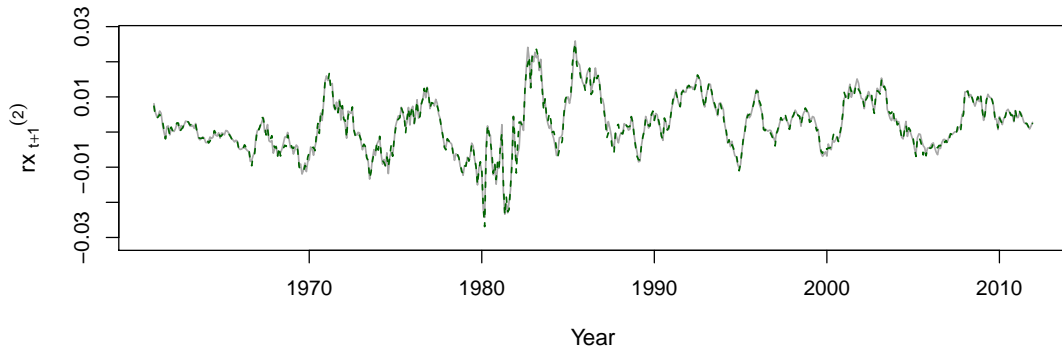





(a) Forward factor model



(b) Model with six macro factors



(c) PAM

Figure 3: Fitted CP1F, LN6F and PAM models (dashed) with observed values of 2-year bond excess log returns (solid) for the time period 011961-122011.  CPAinsample

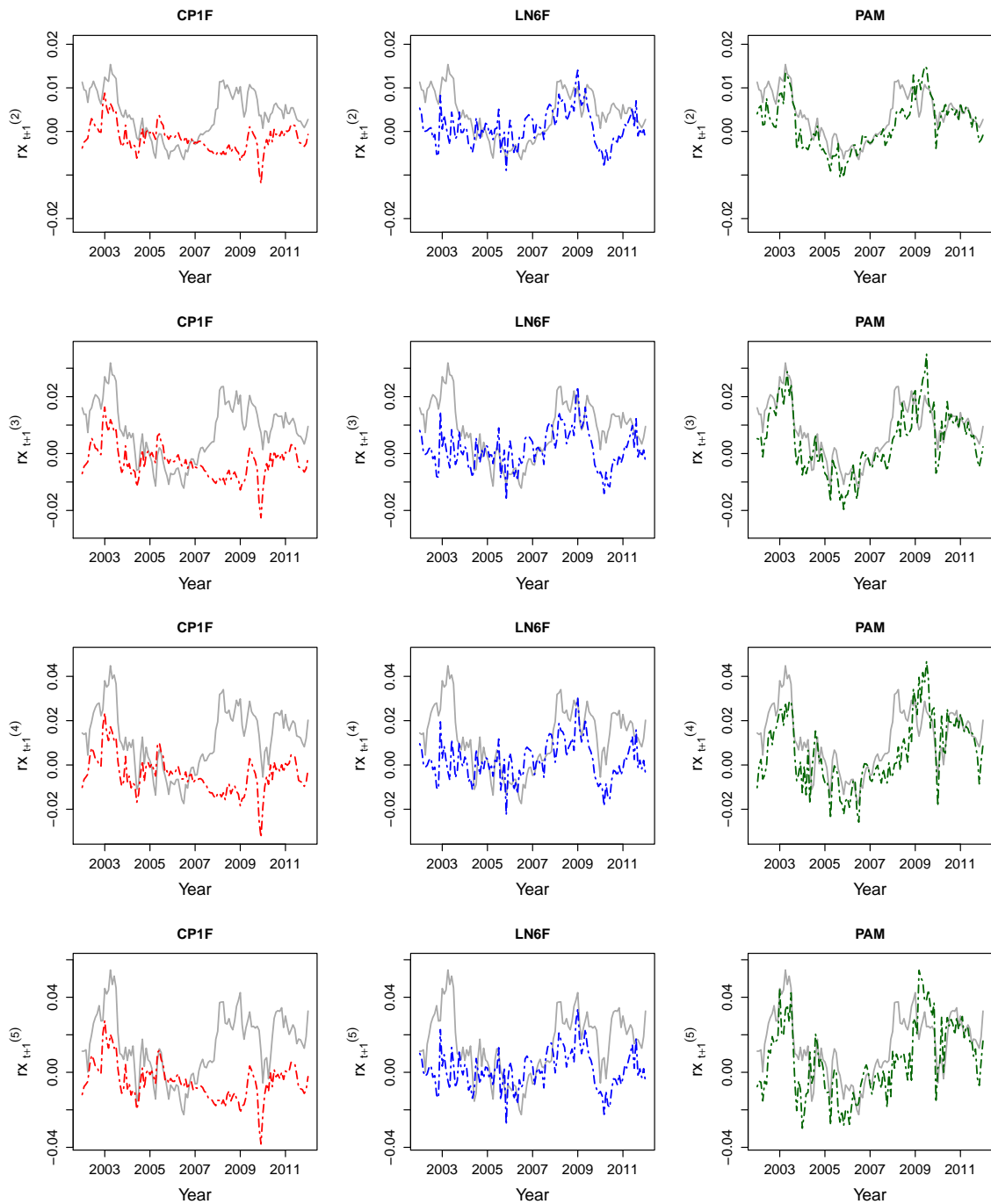



Figure 4: Predicted values of CP1F (red), LN6F (blue) and PAM (green) models (dashed) with observed values of k -year bond excess log returns, $k = 2, 3, 4, 5$, (solid) for the time period 122001-122011.  CPAoutsample

u_i	Point a	Point b
	t=51	t=101
Exp(1)	0.9360	0.8761
Pois(1)	0.9340	0.8724
Bounded	0.9460	0.8949

Table 3: Percentage of correctly identifying the parameter homogeneity shifting points (point a and point b) with $\alpha = 0.05$, $u_i \stackrel{iid}{\sim}$ bounded from (21), $u_i \stackrel{iid}{\sim}$ Exp(1) and $u_i \stackrel{iid}{\sim}$ Pois(1), $m_k = 50$, $M = 10$.

u_i	Scenario 2	Scenario 3
	$51 \leq t < 100$	$101 \leq t \leq 500$
Exp(1)	0.9161	0.9122
Pois(1)	0.9150	0.9104
Bounded	0.9266	0.9266

Table 4: Percentage of correctly fitting in variable selection with $\alpha = 0.05$, $u_i \stackrel{iid}{\sim}$ bounded from (21), $u_i \stackrel{iid}{\sim}$ Exp(1) and $u_i \stackrel{iid}{\sim}$ Pois(1), $m_k = 50$, $M = 10$.

u_i	Point a	Point b
	t=101	t=151
Exp(1)	0.9860	0.9223
Pois(1)	0.9860	0.9209
Bounded	0.9840	0.9309

Table 5: Percentage of correctly identifying the parameter homogeneity shifting points (point a and point b) with $\alpha = 0.05$, $u_i \stackrel{iid}{\sim}$ bounded from (21), $u_i \stackrel{iid}{\sim}$ Exp(1) and $u_i \stackrel{iid}{\sim}$ Pois(1), $m_k = 50$, $M = 10$.


u_i	Scenario 2	Scenario 3
	$101 \leq t < 150$	$151 \leq t \leq 500$
Exp(1)	0.9674	0.9666
Pois(1)	0.9672	0.9666
Bounded	0.9654	0.9650

Table 6: Percentage of correctly fitting in variable selection with $\alpha = 0.05$, $u_i \stackrel{iid}{\sim}$ bounded from (21), $u_i \stackrel{iid}{\sim}$ Exp(1) and $u_i \stackrel{iid}{\sim}$ Pois(1), $m_k = 50$, $M = 10$.


Number	Description	Notation	Transform
1.	Personal Income	a0m52	$\Delta \log$
2.	Real Consumption	a0m224_r	$\Delta \log$
3.	Industrial Production Index (Total)	ips10	$\Delta \log$
4.	NAPM Production Index (Percent)	pmp	–
5.	Civilian Labor Force: Employed, Total	lhemp	$\Delta \log$
6.	Unemployment Rate: All workers, 16 years & over (Percent)	lhur	Δ
7.	NAPM Employment Index (Percent)	pnemp	–
8.	Money Stock M1	fm1	$\Delta^2 \log$
9.	Money Stock M2	fm2	$\Delta^2 \log$
10.	Money Stock M3	fm3	$\Delta^2 \log$
11.	S&P500 Common Stock Price Index: Composite	fspcom	$\Delta \log$
12.	Interest Rate: Federal Funds (% p.a.)	fyff	Δ
13.	Commercial Paper Rate	cp90	Δ
14.	Interest Rate: US Treasury Bill, Sec Mkt, 3-m (% p.a.)	fygm3	Δ
15.	Interest Rate: US Treasury Bill, Sec Mkt, 3-m (% p.a.)	fygm6	Δ
16.	Interest Rate: US Treasury Const Maturities, 1-y (% p.a.)	fygt1	Δ
17.	Interest Rate: US Treasury Const Maturities, 5-y (% p.a.)	fygt5	Δ
18.	Interest Rate: US Treasury Const Maturities, 10-y (% p.a.)	fygt10	Δ
19.	Bond Yield: Moody's Aaa Corporate (% p.a.)	fyaaac	Δ
20.	Bond Yield: Moody's Baa Corporate (% p.a.)	fybaac	Δ
21.	cp90 - fyff Spread	scp90	–
22.	fygm3 - fyff Spread	sfygm3	–
23.	fygm6 - fyff Spread	sfygm6	–
24.	fygt1 - fyff Spread	sfygt1	–
25.	fygt5 - fyff Spread	sfygt5	–
26.	fygt10 - fyff Spread	sfygt10	–
27.	fyaaac - fyff Spread	sfyaaac	–
28.	fybaac - fyff Spread	sfybaac	–
29.	Spot Market Price Index: all commodities	psccom	$\Delta^2 \log$
30.	NAPM Commodity Prices Index (Percent)	pmcp	–
31.	CPI-U: All items	punew	$\Delta^2 \log$

Table 7: List of macroeconomic variables from Ludvigson and Ng (2009), with the same notation and transformations. Note that Δ denotes the first difference of the series and $\Delta \log$ and $\Delta^2 \log$ denote the first and second differences of the logarithm of the series, respectively.

		Jan 1964 - Dec 2003				Jan1961 - Dec 2011			
		RMSE	MAE	R^2	R^2_{adj}	RMSE	MAE	R^2	R^2_{adj}
$rx_{t+1}^{(2)}$	CP	0.007	0.005	0.322	0.315	0.007	0.005	0.215	0.208
	CP1F	0.007	0.005	0.318	0.316	0.007	0.005	0.204	0.203
	LN5F	0.007	0.005	0.365	0.357	0.006	0.004	0.377	0.371
	LN6F	0.005	0.004	0.579	0.574	0.005	0.004	0.501	0.496
	PAM	0.001	0.001	0.980	0.979	0.001	0.001	0.979	0.979
$rx_{t+1}^{(3)}$	CP	0.012	0.010	0.340	0.333	0.012	0.010	0.224	0.217
	CP1F	0.012	0.010	0.338	0.336	0.012	0.010	0.220	0.219
	LN5F	0.012	0.009	0.385	0.377	0.011	0.008	0.383	0.377
	LN6F	0.010	0.008	0.532	0.526	0.010	0.008	0.463	0.458
	PAM	0.003	0.002	0.970	0.970	0.002	0.002	0.970	0.970
$rx_{t+1}^{(4)}$	CP	0.017	0.013	0.370	0.363	0.017	0.013	0.253	0.247
	CP1F	0.017	0.013	0.369	0.368	0.017	0.013	0.251	0.250
	LN5F	0.016	0.013	0.414	0.407	0.015	0.012	0.401	0.395
	LN6F	0.015	0.012	0.486	0.479	0.015	0.011	0.420	0.414
	PAM	0.004	0.003	0.968	0.967	0.003	0.003	0.967	0.966
$rx_{t+1}^{(5)}$	CP	0.021	0.016	0.344	0.337	0.021	0.016	0.231	0.225
	CP1F	0.021	0.016	0.344	0.343	0.021	0.016	0.229	0.228
	LN5F	0.020	0.016	0.386	0.378	0.019	0.015	0.368	0.362
	LN6F	0.019	0.015	0.461	0.454	0.018	0.014	0.398	0.392
	PAM	0.005	0.003	0.965	0.964	0.005	0.003	0.962	0.961

Table 8: RMSE and MAE of fitted PAM, Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009) models. Model with the smallest values of RMSE and MAE and greatest values of R^2 and R^2_{adj} is marked in bold.  CPAoutsample

		RMSPE	MAPE	$\frac{\text{RMSPE}_{\text{PAM}}}{\text{RMSPE}}$	$\frac{\text{MAPE}_{\text{PAM}}}{\text{MAPE}}$
$rx_{t+1}^{(2)}$	CP	0.008	0.007	0.50	0.43
	CP1F	0.008	0.006	0.50	0.50
	LN5F	0.008	0.006	0.50	0.50
	LN6F	0.006	0.005	0.67	0.60
	PAM	0.004	0.003	–	–
$rx_{t+1}^{(3)}$	CP	0.015	0.013	0.47	0.46
	CP1F	0.015	0.013	0.47	0.46
	LN5F	0.015	0.013	0.47	0.46
	LN6F	0.012	0.010	0.58	0.60
	PAM	0.007	0.006	–	–
$rx_{t+1}^{(4)}$	CP	0.021	0.017	0.57	0.59
	CP1F	0.021	0.018	0.57	0.56
	LN5F	0.021	0.018	0.57	0.56
	LN6F	0.017	0.013	0.71	0.77
	PAM	0.012	0.010	–	–
$rx_{t+1}^{(5)}$	CP	0.025	0.021	0.64	0.62
	CP1F	0.026	0.021	0.62	0.62
	LN5F	0.026	0.022	0.62	0.59
	LN6F	0.021	0.017	0.76	0.76
	PAM	0.016	0.013	–	–

Table 9: Forecasting performance of PAM, Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009) models. Model with the smallest values of RMSPE and MAPE is marked in bold.  CPAoutsample

IRTG 1792 Discussion Paper Series 2019



For a complete list of Discussion Papers published, please visit
<http://irtg1792.hu-berlin.de>.

- 001 "Cooling Measures and Housing Wealth: Evidence from Singapore" by Wolfgang Karl Härdle, Rainer Schulz, Taojun Xie, January 2019.
- 002 "Information Arrival, News Sentiment, Volatilities and Jumps of Intraday Returns" by Ya Qian, Jun Tu, Wolfgang Karl Härdle, January 2019.
- 003 "Estimating low sampling frequency risk measure by high-frequency data" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.
- 004 "Constrained Kelly portfolios under alpha-stable laws" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.
- 005 "Usage Continuance in Software-as-a-Service" by Elias Baumann, Jana Kern, Stefan Lessmann, February 2019.
- 006 "Adaptive Nonparametric Community Detection" by Larisa Adamyan, Kirill Efimov, Vladimir Spokoiny, February 2019.
- 007 "Localizing Multivariate CAViaR" by Yegor Klochkov, Wolfgang K. Härdle, Xiu Xu, March 2019.
- 008 "Forex Exchange Rate Forecasting Using Deep Recurrent Neural Networks" by Alexander J. Dautel, Wolfgang K. Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2019.
- 009 "Dynamic Network Perspective of Cryptocurrencies" by Li Guo, Yubo Tao, Wolfgang K. Härdle, April 2019.
- 010 "Understanding the Role of Housing in Inequality and Social Mobility" by Yang Tang, Xinwen Ni, April 2019.
- 011 "The role of medical expenses in the saving decision of elderly: a life cycle model" by Xinwen Ni, April 2019.
- 012 "Voting for Health Insurance Policy: the U.S. versus Europe" by Xinwen Ni, April 2019.
- 013 "Inference of Break-Points in High-Dimensional Time Series" by Likai Chen, Weining Wang, Wei Biao Wu, May 2019.
- 014 "Forecasting in Blockchain-based Local Energy Markets" by Michael Kostmann, Wolfgang K. Härdle, June 2019.
- 015 "Media-expressed tone, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang K. Härdle, Yanchu Liu, June 2019.
- 016 "What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble" by Cathy Yi-Hsuan Chen, Roméo Després, Li Guo, Thomas Renault, June 2019.
- 017 "Portmanteau Test and Simultaneous Inference for Serial Covariances" by Han Xiao, Wei Biao Wu, July 2019.
- 018 "Phenotypic convergence of cryptocurrencies" by Daniel Traian Pele, Niels Wesselhöfft, Wolfgang K. Härdle, Michalis Kolossiatis, Yannis Yatracos, July 2019.
- 019 "Modelling Systemic Risk Using Neural Network Quantile Regression" by Georg Keilbar, Weining Wang, July 2019.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
<http://irtg1792.hu-berlin.de>

This research was supported by the Deutsche
Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2019



For a complete list of Discussion Papers published, please visit
<http://irtg1792.hu-berlin.de>.

- 020 "Rise of the Machines? Intraday High-Frequency Trading Patterns of Cryptocurrencies" by Alla A. Petukhina, Raphael C. G. Reule, Wolfgang Karl Härdle, July 2019.
- 021 "FRM Financial Risk Meter" by Andrija Mihoci, Michael Althof, Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, July 2019.
- 022 "A Machine Learning Approach Towards Startup Success Prediction" by Cemre Ünal, Ioana Ceasu, September 2019.
- 023 "Can Deep Learning Predict Risky Retail Investors? A Case Study in Financial Risk Behavior Forecasting" by A. Kolesnikova, Y. Yang, S. Lessmann, T. Ma, M.-C. Sung, J.E.V. Johnson, September 2019.
- 024 "Risk of Bitcoin Market: Volatility, Jumps, and Forecasts" by Junjie Hu, Weiyu Kuo, Wolfgang Karl Härdle, October 2019.
- 025 "SONIC: SOcial Network with Influencers and Communities" by Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, Yegor Klochkov, October 2019.
- 026 "Affordable Uplift: Supervised Randomization in Controlled Experiments" by Johannes Haupt, Daniel Jacob, Robin M. Gubela, Stefan Lessmann, October 2019.
- 027 "VCRIX - a volatility index for crypto-currencies" by Alisa Kim, Simon Trimborn, Wolfgang Karl Härdle, November 2019.
- 028 "Group Average Treatment Effects for Observational Studies" by Daniel Jacob, Wolfgang Karl Härdle, Stefan Lessmann, November 2019.
- 029 "Antisocial Online Behavior Detection Using Deep Learning" by Elizaveta Zinovyeva, Wolfgang Karl Härdle, Stefan Lessmann, November 2019.
- 030 "Combining Penalization and Adaption in High Dimension with Application in Bond Risk Premia Forecasting" by Xinjue Li, Lenka Zboňáková, Weining Wang, Wolfgang Karl Härdle, December 2019.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
<http://irtg1792.hu-berlin.de>

This research was supported by the Deutsche
Forschungsgemeinschaft through the IRTG 1792.