IRTG 1792 Discussion Paper 2018-039

Penalized Adaptive Forecasting with Large Information Sets and Structural Changes

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

* Humboldt-Universität zu Berlin, Germany
*[2] Xiamen University, PR China

This research was supported by the Deutsche Forschungsgemeinschaft 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

Penalized Adaptive Forecasting with Large Information Sets and Structural Changes*

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

August 22, 2018

Abstract

In the present paper we propose a new method, the Penalized Adaptive Method (PAM), for a data driven detection of structural 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 the 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 properly selects significant variables and detects structural breaks while steadily reduces the forecast error in high-dimensional data.

JEL classification: C12, C13, C50, E47, G12

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

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

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

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

§ 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; W.I.S.E. - Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian, China

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 structural 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 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 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, 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 selec-

tion 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 that there are structural breaks 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 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 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 structural 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 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 shortly describe the propagation-separation approach and the penalized regression method SCAD with its one-step algorithm developed by Zou and Li (2008). 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 5 and Section 6 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 the drawback of selecting the window size prior to model fitting. Although it may be done in some cases using external information about the behaviour of the data, e.g observable business cycles or seasonality, in general it stays an unsolved issue affecting the interpretability of the statistical results.

2.1 Propagation-Separation Approach

In the proposed framework we aim to circumvent the use of *a priori* assumptions about the data behaviour by selecting the windows 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) and its extension by Suvorikova et al. (2015). In the context of model fitting, the propagation condition means that the local model can be extended to a longer interval under an 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 the mathematical point of view.

Assume a linear model with a vector of responses $Y = (Y_1, Y_2, \ldots, Y_n)^\top$, a vector of parameters $\beta = (\beta_1, \ldots, \beta_p)^\top$, 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 parameter vector β is sparse, i.e. only some number $q < p$ 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

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

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

The considered problem of testing homogeneity can be stated in terms of hypothesis testing as follows

$$\begin{aligned} H_0 : &\quad Y_t \sim \mathbb{P}_1, \quad \text{for } t \in I_t^{(m)} \\ H_1 : &\begin{cases} Y_t \sim \mathbb{P}_1, & \text{for } t \in I_t^{(m-1)} \\ Y_t \sim \mathbb{P}_2, & \text{for } t \in I_t^{(m)} \setminus I_t^{(m-1)}, \end{cases} \end{aligned} \tag{1}$$

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

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

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

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. 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 $I_t^{(1)}$.

The propagation-separation approach is then applied, meaning that we are testing for significant changes across the neighbouring subsamples with the use of the following generalized likelihood ratio test statistic adapted from Suvorikova et al. (2015)

$$T_{Lt}^{(m)} = \max_{\beta} L(\beta, I_t^{(m-1)}) + \max_{\beta} L(\beta, I_t^{(m)} \setminus I_t^{(m-1)}) - \max_{\beta} L(\beta, I_t^{(m)}) \tag{2}$$

A correctly calibrated set of critical values $\zeta_1, \ldots, \zeta_M$ is crucial in quantifying the significance level of the given test. We refer to Chen and Niu (2014) or Niu et al. (2017)

for a calibration relevant for an unpenalized linear model without use of the generalized likelihood ratio principle or to Suvorikova et al. (2015) for a calibration using multiplier bootstrap method (Spokoiny and Zhilova, 2015) which will be described later.

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

Adaptive Algorithm

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

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 the 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 the nonconcave SCAD penalty is that this penalized method yields an oracle estimator under some conditions on a shrinkage parameter λ. Moreover, SCAD estimators enjoy 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.

When one deals with a model where the number of parameters p is larger than the number of observations n, LASSO serves as a good initial step of the iterative algorithm and under the irrepresentable condition (Zhao and Yu, 2006) maintains oracle properties. Kim et al. (2008) developed an efficient algorithm similar to LLA for the case of high-dimensional data which always converges to a local minimum. Moreover, they showed oracle properties of SCAD for this case.

In the present paper we perform the penalized (quasi) 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 $(p+1)$-tuple (Y_i, 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

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

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

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

for $k = 0, 1, \ldots$, and $\widehat{\beta}^{(0)}$ being a non-penalized maximum likelihood estimator. The iteration process stops if the sequence $\{\widehat{\beta}^{(k)}\}$ 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

$$\sqrt{n}\lambda_n \to \infty \quad \text{and} \quad \lambda_n \to 0. \tag{4}$$

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

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

$$\mathrm{BIC}_\lambda = \log(\widehat{\sigma}^2_\lambda) + q\frac{\log(n)}{n}C_n, \tag{5}$$

where $\widehat{\sigma}^2_\lambda = n^{-1}\mathrm{SSE}_\lambda = n^{-1}\|Y - X\widehat{\beta}(\lambda)\|_2^2$ and C_n is some positive constant. Here we denote $\widehat{\beta}(\lambda)$ 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 $C_n = \sqrt{n}/p$.

2.3 Penalized Adaptive Method

As discussed before, both the propagation-separation approach and penalized SCAD regression have their advantages in capturing non-stationarity and dimension reduction, respectively. To combine the properties of these two methods, we propose PAM. In PAM we are building a procedure, which deals with non-stationary and high-dimensional data simultaneously. The adaptive way of choosing a window size helps us in determining a homogeneous subsample and the penalized regression reduces the dimension.

One of the differences from the 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

$$\begin{aligned} T_t^{(m)} &= \max_\beta Q(\beta, I_t^{(m-1)}) + \max_\beta Q(\beta, I_t^{(m)} \setminus I_t^{(m-1)}) \\ &\quad -\max_\beta Q(\beta, I_t^{(m)}), \quad m = 2, \ldots, M, \end{aligned} \tag{6}$$

where $Q(\beta, \cdot)$ is defined as previously in equation (3) with the second argument denoting the interval over which the function is evaluated. Here we also adapt the notation for the penalized case; $\tilde{\beta}_t^{(m)}$ now denotes 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 the 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. Non-asymptotic distribution of the test statistic is unknown for the non-penalized case in (2) and for the penalized case (6) 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 deviations for zero components of the estimator are 0 and therefore one is unable to do any inference related to those elements of vector β. This problem was also mentioned in Tibshirani (1996), p. 273. Chatterjee and Lahiri (2011) proposed a remedy developed a modified residual bootstrap which, under some mild conditions, consistently estimates the mean squared error (MSE) of all of the parameter values, both zero and nonzero, of the LASSO method. With this result one is then able to quantify the uncertainty associated with the estimation procedure and construct confidence regions for all of the values of vector β. Moreover, they showed that for the adaptive LASSO (Zou, 2006) the residual bootstrap yields consistent MSE estimate even without use of any modification. Since the one-step SCAD procedure can be easily related to the adaptive LASSO method by introducing a different set of weights for the elements of β, those results apply also to this case. Stating the theoretical results is, however, beyond the scope of this paper.

Despite the appealing conclusions about the applicability of the residual bootstrap, we do not pursue the method in this work. Instead, we perform bootstrap on the observed $(p+1)$-tuples (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 a so-called wild or multiplier bootstrap (Härdle and Mammen, 1993). 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. Suvorikova et al. (2015) then extended the foregoing work into change point detection problem. In this section we are going to relate those 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) slightly extending their notation for the penalized likelihood case. Let us use the notation from previous sections 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 $\mathsf{E}(u_i) = 1$, $\mathsf{Var}(u_i) = 1$ and $\mathsf{E}\{\exp(u_i)\} < \infty$. Multiplying the elements of $L(\beta)$ by the defined random variables u_i we get the bootstrap penalized log-likelihood function as follows

$$Q^{\circ}(\beta) = \sum_{i=1}^{n} u_i \left\{ l_i(\beta) - \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}. \tag{7}$$

Denoting $\mathsf{E}^{\circ}(\cdot) = \mathsf{E}(\cdot|Y, \lambda)$ we then can write that $\mathsf{E}^{\circ}\, Q^{\circ}(\beta) = \mathsf{E}\, Q(\beta)$ and

$$\arg\max_{\beta} \; \mathsf{E}^{\circ}\, Q^{\circ}(\beta) = \arg\max_{\beta} Q(\beta) = \tilde{\beta},$$

which follows from the properties of the SCAD estimator and the LLA algorithm. Please note that 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

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

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 in a finite sample size situation. 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.

2.3.2 Critical Values Calibration

If one wishes to approximate the distribution of the test statistic from (6), it can be done (up to some approximation error in finite samples) by using the bootstrapped penalized likelihood ratio

$$\begin{aligned} T_t^{\circ(m)} \;=\; & \max_{\beta} Q^{\circ}(\beta, I_t^{(m-1)}) + \max_{\beta} Q^{\circ}(\beta, I_t^{(m)} \setminus I_t^{(m-1)}) \\ & - \max_{\beta} Q^{\circ}(\beta_{ts}, I_t^{(m)}), \end{aligned} \tag{8}$$

where the maximization of the bootstrapped penalized likelihood function of the whole interval $I_t^{(m)}$ is taken over values of β_{ts} satisfying

$$\beta_{ts} = \begin{cases} \beta & \text{for} \quad I_t^{(m-1)}; \\ \beta + \tilde{\beta}_{t12} & \text{for} \quad I_t^{(m)} \setminus I_t^{(m-1)}. \end{cases}$$

Here the term

$$\tilde{\beta}_{t12} = \underset{\beta}{\operatorname{argmax}}\, Q(\beta, I_t^{(m)} \setminus I_t^{(m-1)}) - \underset{\beta}{\operatorname{argmax}}\, Q(\beta, I_t^{(m-1)}) \tag{9}$$

corrects a bias of the bootstrap calibration. One can then use this approximation for finding critical values for the aforementioned test statistic under H_0. 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 generalized penalized likelihood ratio test statistic from (6)

$$\zeta_{t\alpha}^{*(m)} = \inf\{z \geq 0 : \mathsf{P}(T_t^{(m)} > z) \leq \alpha\}$$

can be evaluated as

$$\zeta_{t\alpha}^{\circ(m)} = \inf\{z \geq 0 : \mathsf{P}^{\circ}(T_t^{\circ(m)} > z) \leq \alpha\}, \tag{10}$$

where P° denotes the conditional probability given observations of Y and values of λ. In Section 5 we justify the embedding of the multiplier bootstrap into the approximation of the distribution of our test statistic $T_t^{(m)}$ from (6). For the theory of the non-penalized likelihood setting in finite samples we refer the reader to Suvorikova et al. (2015).

As discussed previously, SCAD penalty yields oracle properties only under some conditions the parameter λ has to satisfy, mainly its dependence on the number of observations n in the sample. In order to assess the properties of the bootstrapped penalized likelihood ratio from (8) which would mirror a homogeneous situation, the third right-hand side term of (8) is evaluated with the use of λ from the first right-hand side term adjusted for the longer sample size. Remember, the second parameter of the SCAD penalty function, a, is kept constant and set to 3.7.

We implement the multiplier bootstrap into determining quantiles of the test statistic from (6) by simulating a large number n_b of i.i.d. multipliers u_i, $i = 1, \ldots, n_t^{(M)}$. Computing

$$\begin{aligned} T_t^{\circ b(m)} \quad = \quad & \max_{\beta} Q^{\circ b}(\beta, I_t^{(m-1)}) + \max_{\beta} Q^{\circ b}(\beta, I_t^{(m)} \setminus I_t^{(m-1)}) \\ & - \max_{\beta} Q^{\circ b}(\beta_{ts}, I_t^{(m)}), \end{aligned}$$

for each $b = 1, \ldots, n_b$ we get an approximate distribution of $T_t^{(m)}$ under the homogeneous situation and can evaluate the respective $(1-\alpha)\%$ quantile as in (10). Comparing the test statistic from (6) to the defined critical value we either reject the homogeneity hypothesis H_0, if $T_t^{(m)} > \zeta_{t\alpha}^{\circ(m)}$, for the given confidence level, or move to the next step in PAM and prolong the subsample regarded as homogeneous.

3 Simulation Study

In order to justify the use of multiplier bootstrap in critical values calibration for the penalized likelihood ratio test we present a simulation study. Using the LLA algorithm of Zou and Li (2008) combined with `glmnet` by Friedman et al. (2010), we need multipliers u_i, $i = 1, \ldots, n$ to be non-negative. Therefore 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

$$f(u_i) = \begin{cases} \dfrac{3}{14} & \text{if} \quad 0 \leq u_i \leq 1; \\ \dfrac{1}{12} & \qquad 1 < u_i \leq 4. \end{cases} \tag{11}$$

In the simulation study 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. Design matrix X is taken from a p-dimensional normal distribution as follows

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

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 $\text{N}(0, 1)$. We consider $n = 100, 200, 400$ to assess performance for small to medium sized samples and for convenience we use $M = 2$ which splits the samples into two equally sized parts. 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 1 000 u_i's with 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) and our own simulations, which are not reported here.

Summary of the simulations is given in Table 1. We set $\alpha = 0.1, 0.05, 0.025, 0.01$ to compute the upper quantiles of the bootstrap penalized likelihood ratio distribution.

In Table 1 one can see the percentage of occurrences of the event, when the real likelihood ratio test statistic T was smaller than or equal to the respective quantile of its approximated distribution.

				Confidence level			
n	p	q	$\mathcal{L}(u_i)$	90 %	95 %	97.5 %	99 %
			Bounded	73.2	81.6	86.2	90.8
100	10	3	Exp(1)	72.6	80.7	86.7	91.2
			Pois(1)	83.7	89.3	93.5	96.7
			Bounded	65.7	74.8	82.1	88.9
100	10	5	Exp(1)	64.8	74.3	81.9	89.3
			Pois(1)	77.7	84.9	91.5	96.6
			Bounded	90.6	94.5	97.3	98.6
200	10	3	Exp(1)	89.9	94.9	97.0	98.7
			Pois(1)	93.2	96.2	98.4	99.2
			Bounded	86.9	92.8	96.4	98.2
200	10	5	Exp(1)	86.0	92.7	96.3	98.2
			Pois(1)	90.8	95.7	97.8	99.3
			Bounded	96.9	98.1	99.4	99.8
400	10	3	Exp(1)	96.7	98.2	99.2	99.8
			Pois(1)	97.1	98.4	99.4	99.9
			Bounded	94.1	97.2	98.5	99.0
400	10	5	Exp(1)	93.4	97.2	98.5	99.2
			Pois(1)	94.9	97.8	98.6	99.3
			Bounded	96.7	98.6	99.6	99.8
400	20	3	Exp(1)	96.7	98.5	99.5	99.7
			Pois(1)	97.8	99.3	99.6	99.9
			Bounded	94.5	97.3	98.7	99.4
400	20	5	Exp(1)	93.9	97.0	98.7	99.5
			Pois(1)	96.0	98.1	99.0	99.6

Table 1: Empirical coverage probabilities

As can be seen from Table 1 the performance of the quantiles obtained by multiplier bootstrap method largely depends on the number of observations n in the respective penalized likelihood functions and on the number of active parameters q as well. Bounded and exponentially distributed multipliers lead to very similar results, which are both outperformed by the multipliers generated from Pois(1) distribution. This difference is especially pronounced in cases of $n = 100$ and $n = 200$.

For small samples ($n = 100$), the SCAD method tends to overfit the true model and therefore there is a larger variance of estimator of the vector of parameters β, both in the real and the bootstrapped penalized likelihood case, which leads to underestimation of the real quantiles of the penalized likelihood ratio statistic by the bootstrapped ones.

With a growing sample size, the performance of the multiplier bootstrap improves and finally leads to more conservative bootstrapped quantiles. This is a result of combination of SCAD method and a bootstrap. As discussed before, when dealing with sparse models, it is important to be able to perform statistical inference not only for the nonzero parameters, but for the zero components of the parameter vector as well. Here the multiplier bootstrapped quantiles, even if more conservative, are applicable, as they cover the true confidence regions of the vector of parameters of the studied model.

The issue of conservative quantiles can be seen as the case from Spokoiny and Zhilova (2015) which was regarded as a misspecified model. Misspecification introduces bias into the model of interest and so does SCAD. As discussed in previous sections and proved in Zou and Li (2008), SCAD attains oracle properties only with a growing sample size n, which is not always available in the real world data.

3.1 Change Point Detection

In the following we perform a simulation study regarding the use of bootstrapped 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 a number of subintervals either $M = 10$ or $M = 5$ with $n_t^{(1)} = 50$ and $n_t^{(1)} = 100$, respectively. The size of the increments between successive subinterval is an arbitrary choice as PAM is in this matter a very flexible method. For simplicity we keep the increments constant, i.e. $n_t^{(m+1)} - n_t^{(m)} = 50$ or 100, for $m = 1, \ldots, M-1$. Then we define the true parameter vector $\beta_i^* \in \mathbb{R}^p$, $p = 10$, $i = 1, \ldots, n_t^{(M)}$ as

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

where i_{cp} denotes an observation with a change point. Further, for comparison, we use multipliers u_i with Exp(1), Pois(1) and bounded distributions, where the latter is defined by (11). We generated 1 000 scenarios with four different i_{cp}'s in case of $M = 10$ and for three i_{cp}'s in case of $M = 5$. For each scenario there was only one change point occurring throughout the set of all observations $n_t^{(M)} = 500$ and the confidence level for the hypothesis testing was set to $(1-\alpha) = 95\%$. Results of the multiplier bootstrap performance for the described settings are summarized in Table 2 and Table 3.

In the aforementioned tables we denote a percentage of correctly identified change

	i_{cp}	50	100	200	400
Bounded	Corr	99.9	100.0	100.0	99.9
	1stCorr	99.9	80.0	57.9	33.8
Exp(1)	Corr	99.9	100.0	100.0	100.0
	1stCorr	99.9	78.9	57.1	33.6
Pois(1)	Corr	99.9	100.0	99.9	100.0
	1stCorr	99.9	88.8	74.5	53.0

Table 2: Percentage of correctly identified change points in number of active parameters at 95 % confidence level with use of $u_i \overset{iid}{\sim}$ bounded from (11), $u_i \overset{iid}{\sim}$ Exp(1) and $u_i \overset{iid}{\sim}$ Pois(1), $n_t^{(m-1)} - n_t^{(m)} = 50$, $M = 10$.

PAMsimCP

	i_{cp}	100	200	400
Bounded	Corr	100.0	99.8	100.0
	1stCorr	100.0	93.3	84.9
Exp(1)	Corr	100.0	99.8	100.0
	1stCorr	100.0	93.2	84.1
Pois(1)	Corr	100.0	99.9	100.0
	1stCorr	100.0	95.1	88.8

Table 3: Percentage of correctly identified change points in number of active parameters at 95 % confidence level with use of $u_i \overset{iid}{\sim}$ bounded from (11), $u_i \overset{iid}{\sim}$ Exp(1) and $u_i \overset{iid}{\sim}$ Pois(1), $n_t^{(m-1)} - n_t^{(m)} = 100$, $M = 5$.

PAMsimCP

points by "Corr" and "1stCorr" stands for a percentage of correctly identified change points, which were identified as the first ones occurring.

From Table 2 we can see that our proposed PAM identified the true change point in almost all of the generated scenarios. However in the rows of "1stCorr" one can see the consequences of the underestimated quantiles of the real penalized likelihood ratio. In every prolongation of the subintervals, roughly 10 % (in case of Poisson distributed multipliers) or 20 % (in case of other selected distributions) of scenarios are falsely rejected to be homogeneous, which results in a worse performance of the PAM if the true change point occurs in later sections of the data sample. This effect can be partially overcome by using larger increments between the successive subintervals, as can be seen from Table 3. The false rate in this case is significantly smaller and true change points are again correctly identified in almost every generated scenario.

In the previous change point detection simulation, the parameters of the model were

cut off abruptly and set to zero, which resulted in a change of the number of active parameters. In a real world scenario, the case of smaller changes in parameters might be more common. Therefore in the next simulation we investigate performance of the multiplier bootstrap in a change point detection, where the L_1-norm of the vector of parameters β was the subject of change, while the number of active parameters stayed constant. We kept all of the scenario settings the same and simulated the change point in β as follows

$$\beta_i^* = \begin{cases} (1,1,1,1,1,0,\ldots,0) & \text{if} \quad i < i_{cp}; \\ (1,0.8,0.6,0.4,0.2,0,\ldots,0) & \quad\quad i \geq i_{cp}. \end{cases}$$

Results of the simulations are stated in Table 4 and Table 5. Percentage of correctly identified change points and their occurrence are very similar to those from the previously generated scenarios in Table 2 and Table 3 and thus confirm our inference from above.

	i_{cp}	50	100	200	400
Bounded	Corr	99.8	99.9	100.0	99.9
	1stCorr	99.8	79.4	58.4	33.5
Exp(1)	Corr	99.9	99.8	99.9	99.8
	1stCorr	99.9	79.5	57.1	33.8
Pois(1)	Corr	99.5	99.9	99.8	99.9
	1stCorr	99.5	88.7	74.5	53.8

Table 4: Percentage of correctly identified change points in L_1-norm of parameters at 95 % confidence level with use of $u_i \overset{iid}{\sim}$ bounded from (11), $u_i \overset{iid}{\sim}$ Exp(1) and $u_i \overset{iid}{\sim}$ Pois(1), $n_t^{(m-1)} - n_t^{(m)} = 50$, $M = 10$.

PAMsimCP

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.

	i_{cp}	100	200	400
Bounded	Corr	100.0	99.9	100.0
	1stCorr	100.0	93.2	84.5
Exp(1)	Corr	100.0	99.9	100.0
	1stCorr	100.0	93.1	84.3
Pois(1)	Corr	100.0	99.9	100.0
	1stCorr	100.0	95.2	88.8

Table 5: Percentage of correctly identified change points L_1-norm of parameters at 95 % confidence level with use of $u_i \overset{iid}{\sim}$ bounded from (11), $u_i \overset{iid}{\sim}$ Exp(1) and $u_i \overset{iid}{\sim}$ Pois(1), $n_t^{(m-1)} - n_t^{(m)} = 100$, $M = 5$.

PAMsimCP

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

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

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.

$$rx_{t+1}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} y_t^{(1)} + \beta_2^{(k)} f_t^{(2)} + \ldots + \beta_5^{(k)} f_t^{(5)} + \varepsilon_{t+1}^{(k)}, \tag{12}$$

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

$$rx_{t+1}^{(k)} = b_k(\gamma_0 + \gamma_1 y_t^{(1)} + \gamma_2 f_t^{(2)} + \ldots + \gamma_5 f_t^{(5)}) + \varepsilon_{t+1}^{(k)}, \tag{13}$$

where vector $\gamma = (\gamma_0, \ldots, \gamma_5)^\top$ is the same for all $k = 2, 3, 4, 5$ and b_k satisfies $\frac{1}{4}\sum_{k=2}^{5} b_k = 1$ 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 allow 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 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

$$rx_{t+1}^{(k)} = \alpha^\top F_t + \beta^\top Z_t + \varepsilon_{t+1}, \tag{14}$$

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 (12), (13) and (14), 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 6. 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 $k = 2, 3, 4, 5$ we assume

$$rx_{t+1}^{(k)} = \beta_0 t^{(k)} + \beta_{1t}^{(k)\top} f_t + \beta_{2t}^{(k)\top} M_t + \varepsilon_{t+1}^{(k)},$$

where $f_t = (y_t^{(1)}, f_t^{(2)}, \ldots, f_t^{(5)})^\top$. Vector M_t then defines all of the macro variables from Table 6. Apart from adding more predictors into the model, please note that we allow for time-variation of the vector of parameters β_t.

For our empirical study we consider increments between adjacent subintervals to be 4 years, i.e. $n_t^m - n_t^{m-1} = 48$ 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^2_{adj} 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.

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	lhem	$\Delta \log$
6.	Unemployment Rate: All workers, 16 years & over (Percent)	lhur	Δ
7.	NAPM Employment Index (Percent)	pmemp	–
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 6: 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.

The fitting procedure summary can be found in Table 7, where we use abbreviations

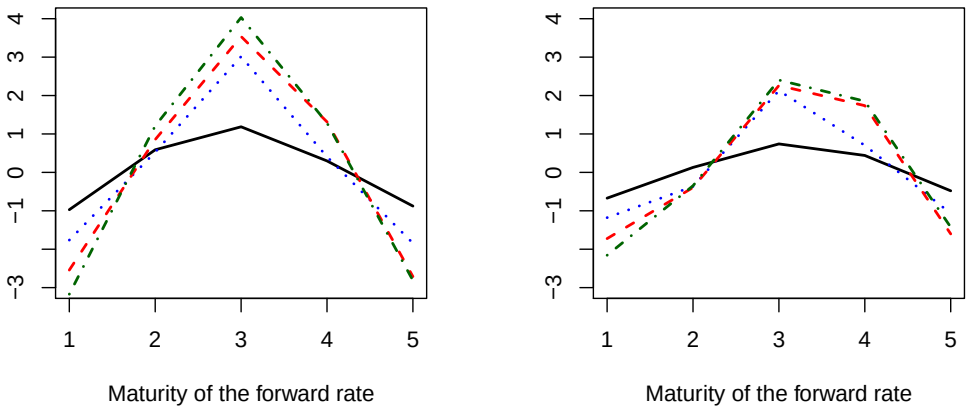


Figure 1: Regression coefficients of 1-year excess log returns on forward rates for 011964-122003 (left) and for 011961-122011 (right). Solid, dotted, dashed and dot-dashed lines denote 2-, 3-, 4- and 5-year maturity of the bond, respectively.

PAMCocPia

CP, CP1F, LN5F and LN6F for models (12), (13) and (14), 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 7, 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

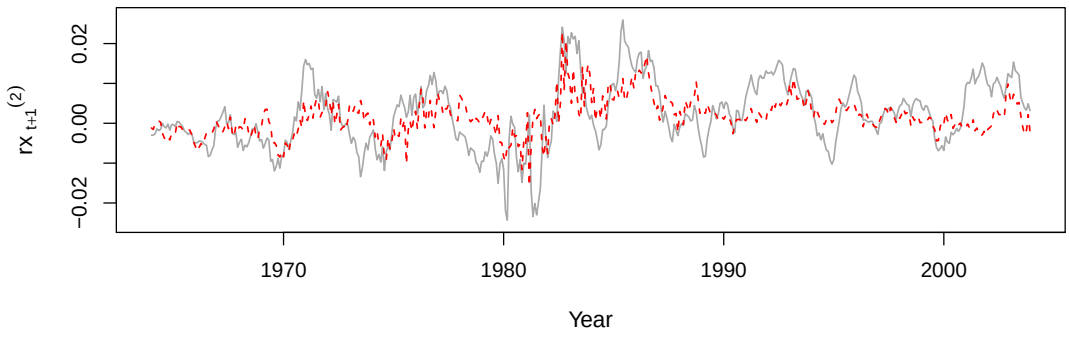
		Jan 1964 - Dec 2003				Jan1961 - Dec 2011			
		RMSE	MAE	R^2	R^2_{adj}	RMSE	MAE	R^2	R^2_{adj}
$rx^{(2)}_{t+1}$	CP	0.007	0.005	0.322	0.315	0.007	0.005	0.215	0.208
	CP1F	0.007	0.005	0.318	0.316	0.007	0.005	0.204	0.203
	LN5F	0.007	0.005	0.365	0.357	0.006	0.004	0.377	0.371
	LN6F	0.005	0.004	0.579	0.574	0.005	0.004	0.501	0.496
	PAM	**0.001**	**0.001**	**0.980**	**0.979**	**0.001**	**0.001**	**0.979**	**0.979**
$rx^{(3)}_{t+1}$	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**
$rx^{(4)}_{t+1}$	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**
$rx^{(5)}_{t+1}$	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 7: 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.

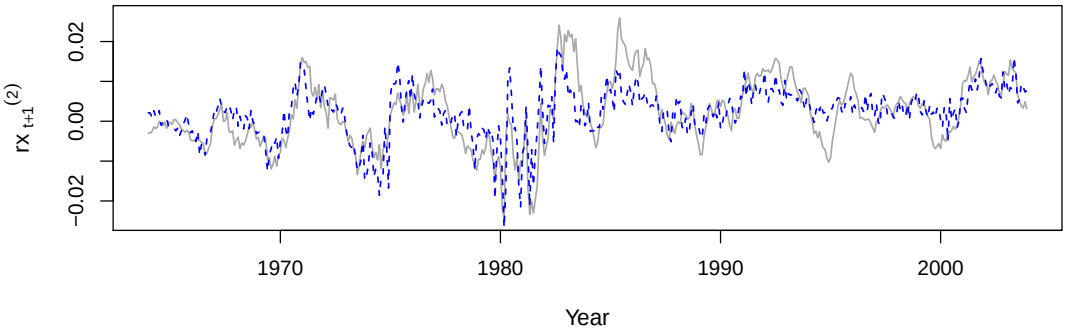
PAMinsam

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 $f^{(2)}_t$ 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.

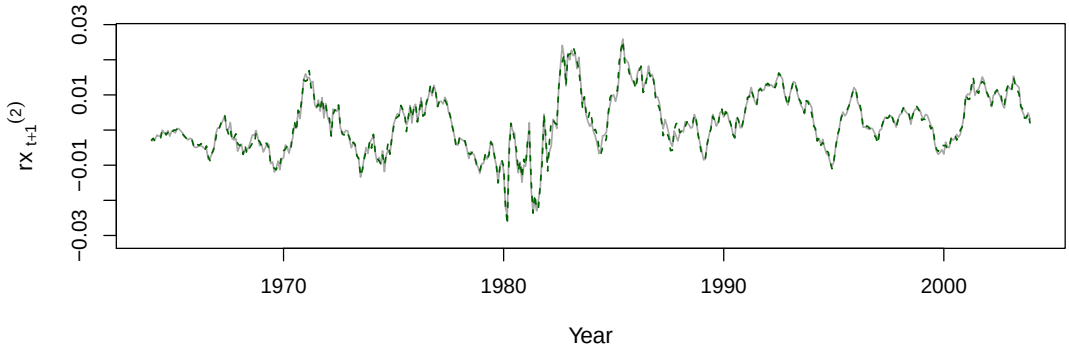
The model for 3-year bond excess returns yields average significant parameter dimension of 13.8 with a minimum of 9 and maximum of 21. Number of change points detected is 5 and the covariates selected in more than 80 % of cases are $f^{(2)}_t$, $f^{(3)}_t$, *fygt5*, *sfygm6*, *sfygt1*, *sfygt5*, *sfygt10* 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.



(a) Forward factor model



(b) Model with six macro factors

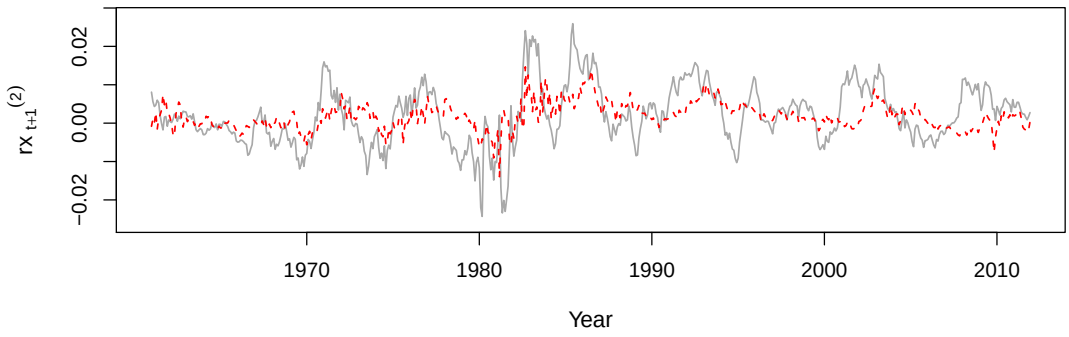


(c) PAM

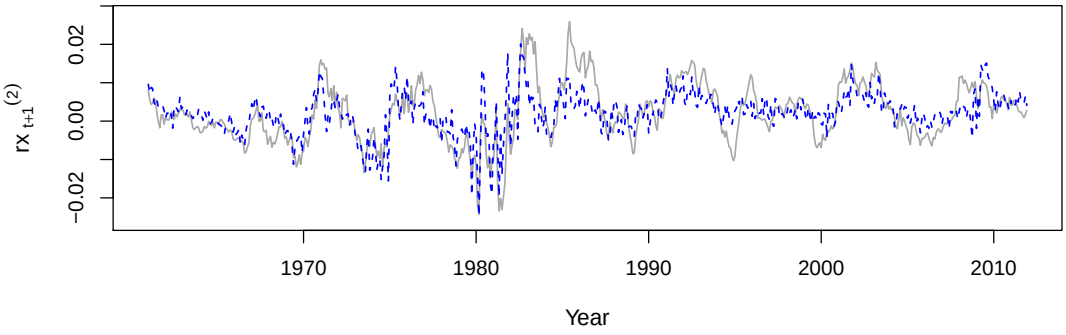
Figure 2: Fitted PAM, CP1F and LN6F models (dashed) with observed values of 2-year bond excess log returns (solid) for the time period 011964-122003.

PAMinsam

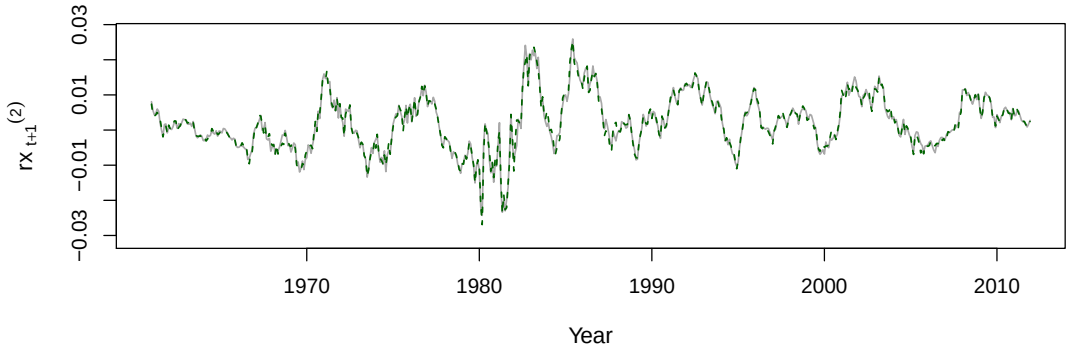
The results for 4- and 5-year bond excess returns are very similar to the previous ones with an average number of chosen covariates 14.5 in both of the cases. The set of chosen macro variables in most of the sub-samples was very similar to the models above. However, the pattern of chosen forward rates 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



(a) Forward factor model



(b) Model with six macro factors



(c) PAM

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

PAMinsam

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 (12), (13) and (14) 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 8 collects all of the results for the three compared methods. Graphical output can be seen in Figure 4.

From Table 8 it is visible that the PAM method outperforms all of the models (12), (13) and (14) 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

		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 8: 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.

PAMoutsam

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 structure change 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

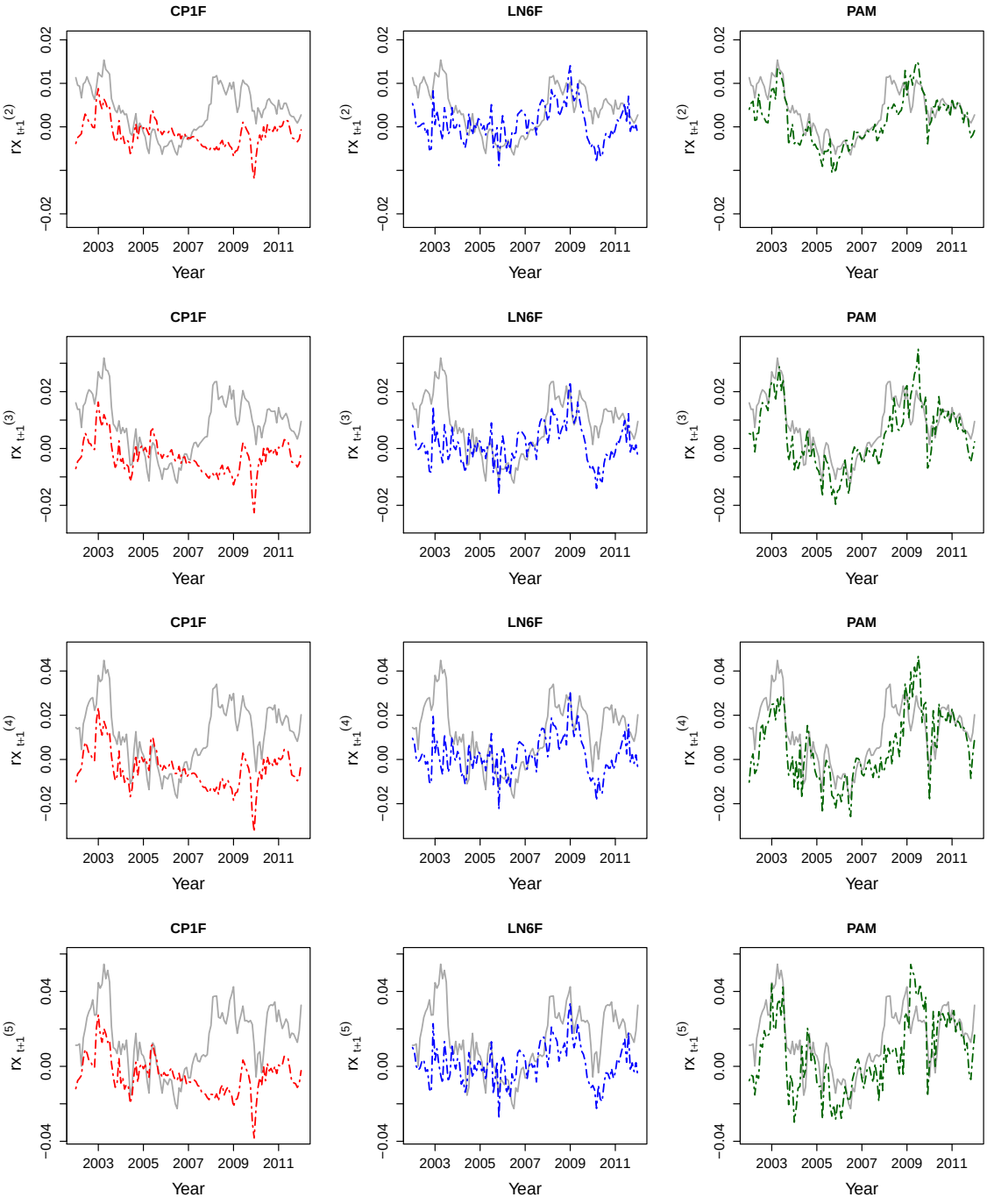


Figure 4: Predicted values of PAM (green), CP1F (red) and LN6F (blue) models (dashed) with observed values of k-year bond excess log returns, $k = 2, 3, 4, 5$, (solid) for the time period 122001-122011.

PAMoutsam

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 length 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 Theoretical Results

In this section we aim to justify the use of multiplier bootstrap approach in case of the SCAD penalized likelihood ratio. As we want to show the results in a quite general setting, we adopt the approach of Kwon and Kim (2012) who showed that SCAD penalized likelihood estimator has indeed oracle properties even for the case of growing dimensions p_n and q_n. Here we take the notation from before, where we denoted the parameter dimension by p and the number of true nonzero coefficients of the model by q and indicate their dependence on the number of observations, n, by the subscript.

Kwon and Kim (2012) showed for the diverging $p_n = \mathcal{O}(n^k)$, where $k \geq 1$, and q_n that the local maximizer of the SCAD penalized likelihood is asymptotically equal to the oracle MLE, $\widehat{\beta}_n^{\text{MLE}}$. They define the latter as the maximizer of the likelihood function subject to the condition $\beta_{nj} = 0$, for $q_n < j \leq p_n$, which satisfies

$$\|\widehat{\beta}_n^{\text{MLE}} - \beta_n^*\| = \mathcal{O}_p\left(\sqrt{\frac{q_n}{n}}\right),$$

where β_n^* is the true parameter vector. Moreover, their results can be strengthened for the global maximizer of the SCAD penalized likelihood in case of $p_n \leq n$ and strictly concave log-likelihood functions.

As we intend to show applicability of the multiplier bootstrap method for the penalized likelihood, we will consider the latter case, i.e. asymptotic results for the global SCAD penalized likelihood maximizer and thereby restrict ourselves to the case of $p_n \leq n$. In what follows, we adopt and adjust the conditions of Kwon and Kim (2012) so we can state our results.

For each n consider Y_{ni}, $i \leq n$, to be i.i.d. random variables with a density $f_n(Y_{n1}, \beta_n^*)$

with $\beta_n^* \in \Theta_n \subseteq \mathbb{R}^{p_n}$. Elements of the true parameter β_n^* can be, without loss of generality, rearranged so that the first q_n of them are nonzero and the rest is equal to zero. Now, let us focus on the regularity conditions assuring oracle properties of the SCAD penalized estimator. Adopting the notation from Kwon and Kim (2012), we denote generic positive constants by $M_1, \ldots, M_7$.

Condition (A1). For some constants c_1 and c_2 which satisfy $0 < 6c_1 < c_2 \leq 1$, it holds

$$q_n = \mathcal{O}(n^{c_1}), \qquad \min_{1 \leq j \leq q_n} n^{(1-c_2)/2} |\beta_{nj}^*| \geq M_1.$$

Here we adjusted the condition for the smaller values of c_1 as is needed for the results for the penalized likelihood ratio from below.

Condition (A2). The first and second derivatives of the log-likelihood $\log f_n(Y_{n1}, \beta_n)$ satisfy

$$\begin{aligned}
\mathsf{E}_{\beta_n^*} \left\{ \frac{\partial \log f_n(Y_{n1}, \beta_n^*)}{\partial \beta_{nj}} \right\} &= 0, \\
\mathsf{E}_{\beta_n^*} \left\{ \frac{\partial^2 \log f_n(Y_{n1}, \beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl}} \right\} &= - \mathsf{E}_{\beta_n^*} \left\{ \frac{\partial \log f_n(Y_{n1}, \beta_n^*)}{\partial \beta_{nj}} \frac{\partial \log f_n(Y_{n1}, \beta_n^*)}{\partial \beta_{nl}} \right\},
\end{aligned}$$

for all $1 \leq j, l \leq p_n$ and $n \geq 1$.

Condition (A3). The first $q_n \times q_n$ submatrix, $I_{n1}(\beta_n^*)$, of the Fisher information matrix

$$I_n(\beta_n^*) = \mathsf{E}_{\beta_n^*} \left[\left\{ \frac{\partial \log f_n(Y_{n1}, \beta_n^*)}{\partial \beta_n} \right\} \left\{ \frac{\partial \log f_n(Y_{n1}, \beta_n^*)}{\partial \beta_n} \right\}^\top \right]$$

is positive definite and it holds

$$0 < M_2 < \gamma_{min}\{I_{n1}(\beta_n^*)\} \leq \gamma_{max}\{I_{n1}(\beta_n^*)\} < M_3 < \infty,$$

for all $n \geq 1$, with $\gamma_{min}(\cdot)$ and $\gamma_{max}(\cdot)$ denoting the smallest and largest eigenvalues of the considered matrix.

Condition (A4). There exists a sufficiently large open subset $B_n \subset \Theta_n$ which contains the true parameter β_n^* such that for almost all Y_{ni} their density function is three times differentiable for all $\beta_n \in B_n$. Moreover, there exist functions $U_{njkl}(\cdot)$ satisfying

$$\left| \frac{\partial^3 \log f_n(Y_{n1}, \beta_n)}{\partial \beta_{nj} \partial \beta_{nk} \partial \beta_{nl}} \right| < U_{njkl}(Y_{ni}),$$

for any $\beta_n \in B_n$ and for all $1 \leq j, k, l \leq p_n$ and $n \geq 1$.

Condition (A5). For the second and fourth moments of the log-likelihood it holds

$$\mathsf{E}_{\beta_n^*}\left\{\frac{\partial \log f_n(Y_{n1},\beta_n^*)}{\partial \beta_{nj}}\right\}^2 < M_4, \qquad \mathsf{E}_{\beta_n^*}\left\{\frac{\partial^2 \log f_n(Y_{n1},\beta_n^*)}{\partial \beta_{nj}\partial \beta_{nl}}\right\}^2 < M_5,$$
$$\mathsf{E}\left\{U_{njkl}(Y_{ni})\right\}^2 < M_6,$$

for all $1 \le j,k,l \le p_n$ and $n \ge 1$.

Condition (A6). There exists a positive constant M_7 and a convex open subset $\Omega_n \subset \Theta_n$ such that both $\widehat{\beta}_n^{\mathrm{MLE}}$ and β_n^* belong to Ω_n and

$$\min_{\beta_n \in \Omega_n} \gamma_{min}(\beta_n) > M_7,$$

for all sufficiently large n. Here $\gamma_{min}(\beta_n)$ denotes the smallest eigenvalue of the matrix of the second derivatives of the negative log-likelihood

$$-\frac{1}{2n}\sum_{i=1}^{n}\frac{\partial^2 \log f_n(Y_{ni},\beta_n)}{\partial \beta_n^2}$$

at β_n.

In their work (Theorem 2), Kwon and Kim (2012) show that $\mathsf{P}(\tilde{\beta}_n = \widehat{\beta}_n^{\mathrm{MLE}}) \to 1$ with n tending to infinity, where $\tilde{\beta}_n$ denotes the global maximizer of the SCAD penalized likelihood on the set Ω_n. This holds under the conditions *(A1)* to *(A6)* and if $\lambda_n = o(n^{-(1-c_2+c_1)/2})$ and $p_n/(\sqrt{n}\lambda_n)^2 \to 0$ as $n \to \infty$.

Their result is very important for the theory of the multiplier bootstrap combined with the SCAD penalized likelihood and it shows that the SCAD estimator satisfies $\|\tilde{\beta}_n - \beta_n^*\| = \mathcal{O}_p(\sqrt{q_n/n})$, where $\tilde{\beta}_{nj} = 0$ for $q_n < j \le p_n$ if n is large enough.

In the following we consider Fisher and Wilks type of expansions of the penalized likelihood ratio similar to Spokoiny (2017), where the author dealt with a quadratic penalization in a finite sample case.

For the purposes of this section we denote a vector of derivatives with respect to β_n by ∇, where we omit the subscript β_n for simplicity. A gradient vector of the penalty term is denoted by $\nabla P_{\lambda_n}(\beta_n) = (p'_{\lambda_n}(|\beta_{n1}|)\mathrm{sgn}(\beta_{n1}),\ldots,p'_{\lambda_n}(|\beta_{np_n}|)\mathrm{sgn}(\beta_{np_n}))^\top$ and the diagonal matrix of its second derivatives by $\nabla^2 P_{\lambda_n}(\beta_n) = \mathrm{diag}\left\{p''_{\lambda_n}(|\beta_{n1}|),\ldots,p''_{\lambda_n}(|\beta_{np_n}|)\right\}$. Further, we define the penalized information matrix as

$$D_Q^2 = -\mathsf{E}\,\nabla^2 Q(\beta_n^*) = -\mathsf{E}\,\nabla^2 L(\beta_n^*) + n\nabla^2 P_{\lambda_n}(\beta_n^*).$$

For the ease of notation the subscript "1" will be used in the following to denote the first q_n elements or the first $q_n \times q_n$ submatrix of the vector or matrix, respectively. For

example, the first $q_n \times q_n$ submatrix of D_Q^2 will be denoted by D_{Q1}^2 and analogously its inverse D_{Q1}^{-2} which is defined for all n given the condition *(A3)*. The proofs of theorems are relegated to the Appendix.

Firstly, we come to the Wilks approximation type of result for the SCAD penalized likelihood ratio.

Theorem 1. *Assume a model with a dimension such that* $p_n/(\sqrt{n}\lambda_n)^2 \to 0$ *as* $n \to 1$ *and that conditions (A1) to (A6) hold. Then*

$$Q(\tilde{\beta}_n) - Q(\beta_n^*) = \frac{1}{2}\|D_{Q1}^{-1}\nabla_1 Q(\beta_n^*)\|^2 + o_p(1), \tag{15}$$

if the SCAD penalty parameter satisfies $\lambda_n = o(n^{-(1-c_2+c_1)/2})$.

In order to be able to obtain a similar approximation for the bootstrapped penalized likelihood ratio, we need show the $\sqrt{n/q_n}$-consistency and sparsity property of the global maximizer of (7).

Note that the choice of the multipliers can affect the concavity of the log-likelihood function, however, this is not the case of the present work, since the multipliers are non-negative and for sufficiently large n there is almost surely at least one positive u_i for all of the discussed distributions. Thus, condition *(A6)* is valid also for the bootstrapped log-likelihood function, where we shall replace the positive constant M_7 by M_{10} for the sake of correctness.

Theorem 2. *Under the conditions of Theorem 1 the following holds for the global maximizer of* (7), $\tilde{\beta}_n^\circ$,

$$\|\tilde{\beta}_n^\circ - \beta_n^*\| = \mathcal{O}_p\left(\sqrt{\frac{q_n}{n}}\right) \text{ as well as } \|\tilde{\beta}_n^\circ - \tilde{\beta}_n\| = \mathcal{O}_p\left(\sqrt{\frac{q_n}{n}}\right).$$

Moreover, if n *is sufficiently large,* $\tilde{\beta}_{nj}^\circ = 0$ *for* $q_n < j \leq p_n$.

Now we can state the desired results.

Theorem 3. *If the conditions from Theorem 1 are satisfied, we can write*

$$Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) = \frac{1}{2}\|D_{Q1}^{-1}\left\{\nabla_1 Q^\circ(\beta_n^*) - \nabla_1 Q(\beta_n^*)\right\}\|^2 + o_p(1), \tag{16}$$

for the multiplier bootstrapped SCAD penalized likelihood ratio.

With the assertions of the previous theorems at hand we are able to show the asymptotic validity of the bootstrap approximation of the real world. Let us denote by $G_n(\tilde{\beta}_n, \beta_n^*)$ the cumulative distribution function of the SCAD penalized likelihood $Q(\tilde{\beta}_n) - Q(\beta_n^*)$ and by $G_n^\circ(\tilde{\beta}_n^\circ, \tilde{\beta}_n)$ its bootstrapped counterpart conditioned on the data and the value of the parameter λ_n. Then we can summarize the previous theorems into the following.

Theorem 4. *Under the conditions of Theorem 1 it holds*

$$\rho(G_n, G_n^\circ) \xrightarrow{P} 0, \quad \text{as } n \to \infty,$$

with $\rho(\cdot, \cdot)$ denoting the Prokhorov metric on a set of all probability measures on $(\mathbb{R}^{p_n}, \mathcal{B}(\mathbb{R}^{p_n}))$.

Moreover, if $q_n \geq 1$, then the statement of the last theorem can be applied to justify the approximation of the quantiles of the penalized likelihood by their bootstrapped variants, cf. Chatterjee and Lahiri (2011) and their conclusions about the bootstrap confidence intervals. In their more recent work, Chatterjee and Lahiri (2013), showed, that the residual bootstrap approximation of the distribution of their considered test statistic is more efficient, in terms of convergence rates, than the usual asymptotic inference based on the knowledge of its asymptotic distribution. Similar results concerning multiplier bootstrap and the SCAD method might be of great interest, however are beyond the scope of this paper.

Let us now come back to the case of the generalized penalized likelihood ratio from Section 2. Consider the null hypothesis from (1) given as $H_0 : \tilde{\beta}_t^{(m-1)} = \tilde{\beta}_t^{(m)}$. Then, under H_0, the foregoing theorems and their statements can be analogously applied to the test statistic $T_t^{(m)}$ and its bootstrapped version $T_t^{\circ(m)}$ from (6) and (8), respectively. Furthermore, as already pointed out, the term $\tilde{\beta}_{12}$ from (9) corrects the bias of the bootstrapped version of the test statistics in case there is a structural change in the considered subintervals. Owing to this fact, one gets the approximation of the distribution of the SCAD penalized likelihood ratio as if the H_0 was true.

6 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 forecastability of excess bond risk premia modelling, 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, exactly those two aspects of their models omit an important characteristic of financial data and that is its time-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 occurring in the real world. Especially, its extensions for the case of modelling time series, quantiles or both is of our interest in the future work.

Appendix

All of the proofs from section 5 together with the additional results of interest are collected here.

For the proof of Theorem 1 we need the following Lemma.

Lemma 1. *Under the conditions of Theorem 1 we can write*

$$Q(\tilde{\beta}_n) - Q(\beta_n^*) = \frac{1}{2}\|D_{Q1}(\tilde{\beta}_{n1} - \beta_{n1}^*)\|^2 + o_p(1). \tag{17}$$

Proof of Lemma 1. By Taylor's expansion of $Q(\beta_n^*)$ around $\tilde{\beta}_n$ we get

$$Q(\tilde{\beta}_n) - Q(\beta_n^*) = \frac{1}{2}(\beta_n^* - \tilde{\beta}_n)^\top \left\{-\nabla^2 Q(\beta_n^+)\right\}(\beta_n^* - \tilde{\beta}_n), \tag{18}$$

where the vector β_n^+ lies between β_n^* and $\tilde{\beta}_n$.

For sufficiently large n, as demonstrated by Kwon and Kim (2012), the vector $(\beta_n^* - \tilde{\beta}_n)$ has the last $p_n - q_n$ elements equal to zero with probability tending to 1. Therefore, assuming that the SCAD penalized estimator $\tilde{\beta}_n$ successfully recovers zero components of the true parameter vector β_n^*, it suffices to show the expansion from (18) only for the nonzero part of the vector and the corresponding $q_n \times q_n$ submatrix of $\nabla^2 Q(\beta_n^+)$, i.e.

$$Q(\tilde{\beta}_n) - Q(\beta_n^*) = \frac{1}{2}(\beta_{n1}^* - \tilde{\beta}_{n1})^\top \left\{-\nabla_1^2 Q(\beta_n^+)\right\}(\beta_{n1}^* - \tilde{\beta}_{n1}).$$

Further, we can write

$$\begin{aligned}
\nabla_1^2 Q(\beta_n^+) &= \nabla_1^2 L(\beta_n^+) - n\nabla_1^2 P_{\lambda_n}(\beta_n^+) - \nabla_1^2 L(\beta_n^*) + \nabla_1^2 L(\beta_n^*) - \mathsf{E}\,\nabla_1^2 Q(\beta_n^*) + \mathsf{E}\,\nabla_1^2 Q(\beta_n^*) \\
&= \{\nabla_1^2 L(\beta_n^+) - \nabla_1^2 L(\beta_n^*)\} + \{\nabla_1^2 L(\beta_n^*) - \mathsf{E}\,\nabla_1^2 L(\beta_n^*)\} \\
&\quad + \{n\nabla_1^2 P_{\lambda_n}(\beta_n^*) - n\nabla_1^2 P_{\lambda_n}(\beta_n^+)\} + \mathsf{E}\,\nabla_1^2 Q(\beta_n^*) \\
&\stackrel{\text{def}}{=} I_1 + I_2 + I_3 - D_{Q1}^2.
\end{aligned}$$

For the term I_1 and some β_n^{++} lying between β_n^* and β_n^+ we have

$$\begin{aligned}
\|I_1\|^2 = \|(\beta_{n1}^+ - \beta_{n1}^*)^\top \nabla_1\{\nabla_1^2 L(\beta_n^{++})\}\|^2 &\le \sum_{j,k,l=1}^{q_n} \left\{\frac{\partial^3 L(\beta_n^{++})}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}}\right\}^2 \|\beta_n^+ - \beta_n^*\|^2 \\
&\le \sum_{j,k,l=1}^{q_n} \left\{\sum_{i=1}^{n} U_{njlk}(Y_{ni})\right\}^2 \mathcal{O}_p\left(\frac{q_n}{n}\right) = \mathcal{O}_p(q_n^3 n^2)\mathcal{O}_p\left(\frac{q_n}{n}\right) = \mathcal{O}_p(q_n^4 n),
\end{aligned}$$

with the functions $U_{njkl}(Y_{ni})$ taken from Condition *(A4)*. From this we get

$$\frac{1}{2}(\beta_{n1}^* - \tilde{\beta}_{n1})^\top I_1(\beta_{n1}^* - \tilde{\beta}_{n1}) \le \|I_1\|\|\beta_n^* - \tilde{\beta}_n\|^2 = \mathcal{O}_p(q_n^2\sqrt{n})\mathcal{O}_p\left(\frac{q_n}{n}\right) = o_p(1), \tag{19}$$

where the last equation invokes *(A1)*. The term I_2 can be bounded by the Chebyshev's inequality as follows

$$\begin{aligned}
\mathsf{P}\left\{\left\|\frac{1}{n}\nabla_1^2 L(\beta_n^*) - \frac{1}{n}\,\mathsf{E}\,\nabla_1^2 L(\beta_n^*)\right\| \ge \frac{\varepsilon}{q_n^2}\right\} &\le \frac{q_n^4}{n^2\varepsilon^2}\,\mathsf{E}\sum_{j,k=1}^{q_n}\left\{\frac{\partial^2 L(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nk}} - \mathsf{E}\,\frac{\partial^2 L(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nk}}\right\}^2 \\
&= \frac{q_n^4}{n^2\varepsilon^2}\mathcal{O}(q_n^2 n) = o(1).
\end{aligned}$$

Hence,

$$\left\| \frac{1}{n} \nabla_1^2 L(\beta_n^*) - \frac{1}{n} \mathsf{E}\, \nabla_1^2 L(\beta_n^*) \right\| = o_p\left(\frac{1}{q_n^2}\right) \tag{20}$$

and we have

$$\frac{1}{2}(\beta_{n1}^* - \tilde{\beta}_{n1})^\top I_2 (\beta_{n1}^* - \tilde{\beta}_{n1}) \leq \|I_2\| \|\beta_n^* - \tilde{\beta}_n\|^2 = o_p\left(\frac{n}{q_n^2}\right) \mathcal{O}_p\left(\frac{q_n}{n}\right) = o_p\left(\frac{1}{q_n}\right) = o_p(1). \tag{21}$$

Having $p''_{\lambda_n}(\beta_{nj}) = 0$ for all $q_n + 1 \leq j \leq p_n$, the norm of $n^{-1} I_3$ satisfies

$$\|\nabla_1^2 P_{\lambda_n}(\beta_n^*) - \nabla_1^2 P_{\lambda_n}(\beta_n^+)\| = \left[\sum_{j=1}^{q_n} \left\{ p''_{\lambda_n}(\beta_{nj}^*) - p''_{\lambda_n}(\beta_{nj}^+) \right\}^2 \right]^{1/2}.$$

The term on the right hand side can be bounded due to the smoothness condition *(D)* from Fan and Peng (2004), which is satisfied for SCAD if *(A1)* holds. It assumes that there are positive constants M_8 and M_9 such that, if $\beta_1, \beta_2 > M_8 \lambda_n$, then $|p''_{\lambda_n}(\beta_1) - p''_{\lambda_n}(\beta_2)| \leq M_9 |\beta_1 - \beta_2|$. Thus, for n large enough, all of the nonzero coefficients are larger than $M_8 \lambda_n$ for some generic constant M_8 and we can write

$$\begin{aligned} \left[\sum_{j=1}^{q_n} \left\{ p''_{\lambda_n}(\beta_{nj}^*) - p''_{\lambda_n}(\beta_{nj}^+) \right\}^2 \right]^{1/2} &\leq \left[\sum_{j=1}^{q_n} \left\{ M_9 |\beta_{nj}^* - \beta_{nj}^+| \right\}^2 \right]^{1/2} = M_9 \|\beta_n^* - \beta_n^+\| \\ &= \mathcal{O}_p\left(\sqrt{\frac{q_n}{n}}\right) = o_p\left(\frac{1}{q_n^{5/2}}\right). \end{aligned}$$

Further,

$$\frac{1}{2}(\beta_{n1}^* - \tilde{\beta}_{n1})^\top I_3 (\beta_{n1}^* - \tilde{\beta}_{n1}) \leq \frac{1}{2} \|I_3\| \|\beta_n^* - \tilde{\beta}_n\|^2 = o_p\left(\frac{n}{q_n^{5/2}}\right) \mathcal{O}_p\left(\sqrt{\frac{q_n}{n}}\right) = o_p(1). \tag{22}$$

Combining (19), (21) and (22) yields

$$\frac{1}{2}(\beta_{n1}^* - \tilde{\beta}_{n1})^\top \{-\nabla_1^2 Q(\beta_n^+)\} (\beta_{n1}^* - \tilde{\beta}_{n1}) = \frac{1}{2}(\beta_{n1}^* - \tilde{\beta}_{n1})^\top \{-\mathsf{E}\, \nabla_1^2 Q(\beta_n^*)\} (\beta_{n1}^* - \tilde{\beta}_{n1}) + o_p(1)$$

and with this the proof of (17) is complete. □

With the result of Lemma 1 we can now show a type of Fisher expansion for the penalized likelihood ratio.

Theorem 5. *Under the conditions of Theorem 1 it holds*

$$D_{Q1}(\tilde{\beta}_{n1} - \beta_{n1}^*) = D_{Q1}^{-1} \nabla_1 Q(\beta_n^*) + o_p\left(\frac{1}{\sqrt{q_n}}\right).$$

Proof of Theorem 5. Using the fact that the term $\nabla_1 Q(\tilde{\beta}_n) = \nabla_1 L(\tilde{\beta}_n) - n\nabla_1 P_{\lambda_n}(\tilde{\beta}_n) = 0$ and its Taylor's expansion around β_n^* we get

$$\begin{aligned} 0 = \nabla_1 L(\beta_n^*) + (\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \nabla_1^2 L(\beta_n^*) + \frac{1}{2}(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \nabla_1^2 \{\nabla_1 L(\beta_n^+)\}(\tilde{\beta}_{n1} - \beta_{n1}^*) \\ - n\nabla_1 P_{\lambda_n}(\beta_n^*) - n(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \nabla_1^2 P_{\lambda_n}(\beta_n^{++}), \end{aligned}$$

where β_n^+ and β_n^{++} lie between $\tilde{\beta}_n$ and β_n^*. The equation can be rewritten into

$$\begin{aligned} (\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \{\nabla_1^2 L(\beta_n^*) - \mathsf{E}\,\nabla_1^2 L(\beta_n^*) + n\nabla_1^2 P_{\lambda_n}(\beta_n^*) - n\nabla_1^2 P_{\lambda_n}(\beta_n^{++}) + \mathsf{E}\,\nabla_1^2 Q(\beta_n^*)\} \\ = -\nabla_1 Q(\beta_n^*) - \frac{1}{2}(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \nabla_1^2 \{\nabla_1 L(\beta_n^+)\}(\tilde{\beta}_{n1} - \beta_{n1}^*) \end{aligned} \tag{23}$$

which is

$$(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \{I_2 + I_3 - D_{Q1}^2\} = -\nabla_1 Q(\beta_n^*) - I_4.$$

For the terms I_2 and I_3 it holds

$$(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top I_2 \leq \|\tilde{\beta}_n - \beta_n^*\| \|I_2\| = \mathcal{O}_p\left(\sqrt{\frac{q_n}{n}}\right) o_p\left(\frac{n}{q_n^2}\right) = o_p\left(\sqrt{\frac{n}{q_n^3}}\right) \tag{24}$$

and

$$(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top I_3 \leq \|\tilde{\beta}_n - \beta_n^*\| \|I_3\| = \mathcal{O}_p\left(\sqrt{\frac{q_n}{n}}\right) o_p\left(\frac{n}{q_n^{5/2}}\right) = o_p\left(\sqrt{\frac{n}{q_n^4}}\right). \tag{25}$$

Using the assumptions of the theorem we can write

$$\begin{aligned} \left\| \frac{1}{n} I_4 \right\|^2 &\leq \frac{1}{n^2} \sum_{j,k,l=1}^{q_n} \left\{ \frac{\partial^3 L(\beta_n^+)}{\partial \beta_{nj} \partial \beta_{nk} \partial \beta_{nl}} \right\}^2 \|\tilde{\beta}_n - \beta_n^*\|^4 \leq \frac{1}{n^2} \sum_{i=1}^{n} n \sum_{j,k,l=1}^{q_n} U_{njkl}^2(Y_{ni}) \|\tilde{\beta}_n - \beta_n^*\|^4 \\ &= \mathcal{O}_p(q_n^3) \mathcal{O}_p\left(\frac{q_n^2}{n^2}\right) = o_p\left(\frac{1}{nq_n}\right) \end{aligned}$$

and thus

$$\|I_4\| = o_p\left(\sqrt{\frac{n}{q_n}}\right). \tag{26}$$

Putting (23), (24), (25) and (26) together, we obtain

$$\{-\mathsf{E}\,\nabla_1^2 Q(\beta_n^*)\}(\tilde{\beta}_{n1} - \beta_{n1}^*) = D_{Q1}^2(\tilde{\beta}_{n1} - \beta_{n1}^*) = \nabla_1 Q(\beta_n^*) + o_p\left(\sqrt{\frac{n}{q_n}}\right).$$

From *(A3)* it now follows

$$D_{Q1}(\tilde{\beta}_{n1} - \beta_{n1}^*) = D_{Q1}^{-1} \nabla_1 Q(\beta_n^*) + o_p\left(\frac{1}{\sqrt{q_n}}\right),$$

which completes the proof. □

Now we can move to the proof of the Wilks type of approximation for the penalized likelihood ratio.

Proof of Theorem 1. Here we use the Taylor's expansion of $Q(\tilde{\beta}_n)$ around β_n^* which is

$$\begin{aligned} Q(\tilde{\beta}_n) &= Q(\beta_n^*) + (\tilde{\beta}_n - \beta_n^*)^\top \nabla Q(\beta_n^*) + \frac{1}{2}(\tilde{\beta}_n - \beta_n^*)^\top \nabla^2 Q(\beta_n^*)(\tilde{\beta}_n - \beta_n^*) \\ &\quad + \frac{1}{6}\nabla^\top\{(\tilde{\beta}_n - \beta_n^*)^\top \nabla^2 Q(\beta_n^+)(\tilde{\beta}_n - \beta_n^*)\}(\tilde{\beta}_n - \beta_n^*) \\ &\stackrel{\text{def}}{=} Q(\beta_n^*) + (\tilde{\beta}_n - \beta_n^*)^\top \nabla Q(\beta_n^*) + I_5 + I_6, \end{aligned}$$

for some vector β_n^+ lying between $\tilde{\beta}_n$ and β_n^*.

In order to prove the equation (15), we need to expand and bound the terms I_5 and I_6. For I_5 we get

$$\nabla^2 Q(\beta_n^*) - \mathsf{E}\,\nabla^2 Q(\beta_n^*) + \mathsf{E}\,\nabla^2 Q(\beta_n^*) = \nabla^2 L(\beta_n^*) - \mathsf{E}\,\nabla^2 L(\beta_n^*) + \mathsf{E}\,\nabla^2 Q(\beta_n^*)$$

and hence

$$\begin{aligned} \frac{1}{2}(\tilde{\beta}_n - \beta_n^*)^\top \nabla^2 Q(\beta_n^*)(\tilde{\beta}_n - \beta_n^*) &= \frac{1}{2}(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \nabla_1^2 Q(\beta_n^*)(\tilde{\beta}_{n1} - \beta_{n1}^*) \\ &\le \frac{1}{2}\|\nabla_1^2 L(\beta_n^*) - \nabla_1^2\,\mathsf{E}\,L(\beta_n^*)\|\|\tilde{\beta}_n - \beta_n^*\|^2 \\ &\quad + \frac{1}{2}(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top\,\mathsf{E}\,\nabla_1^2 Q(\beta_n^*)(\tilde{\beta}_{n1} - \beta_{n1}^*) \\ &\le \frac{1}{2}(\tilde{\beta}_{n1} - \beta_{n1}^*)^\top\,\mathsf{E}\,\nabla_1^2 Q(\beta_n^*)(\tilde{\beta}_{n1} - \beta_{n1}^*) + o_p(1) \\ &= -\frac{1}{2}\|D_{Q1}(\tilde{\beta_{n1}} - \beta_{n1}^*)\|^2 + o_p(1), \end{aligned} \tag{27}$$

where the second inequality is based on (21).

Using Cauchy-Schwarz inequality and the fact that the third derivatives of the SCAD penalty are zero for all of the values of the function's argument, we can bound I_6 by the following

$$\begin{aligned} |I_6| &\le \left|\frac{1}{6}\nabla^\top\{(\tilde{\beta}_n - \beta_n^*)^\top \nabla^2 L(\beta_n^+)(\tilde{\beta}_n - \beta_n^*)\}(\tilde{\beta}_n - \beta_n^*)\right| \\ &= \left|\frac{1}{6}\sum_{j,k,l=1}^{q_n} \frac{\partial^3 L(\beta_n^+)}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}}(\tilde{\beta}_{nj} - \beta_{nj}^*)(\tilde{\beta}_{nk} - \beta_{nk}^*)(\tilde{\beta}_{nl} - \beta_{nl}^*)\right| \\ &\le \frac{1}{6}\left[\sum_{j,k,l=1}^{q_n}\left\{\frac{\partial^3 L(\beta_n^+)}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{ml}}\right\}^2\right]^{1/2}\|\tilde{\beta}_n - \beta_n^*\|^3 \\ &\le \frac{1}{6}\left[\sum_{i=1}^{n} n\sum_{j,k,l=1}^{q_n} U_{njkl}^2(Y_{ni})\right]^{1/2}\|\tilde{\beta}_n - \beta_n^*\|^3 = \mathcal{O}_p(nq_n^{3/2})\mathcal{O}_p\left(\sqrt{\frac{q_n^3}{n^3}}\right) \\ &= \mathcal{O}_p\left(\sqrt{\frac{q_n^6}{n}}\right) = o_p(1). \end{aligned} \tag{28}$$

Using (27) and (28) and the sparsity property of $\tilde{\beta}_n$ and β_n^* we get

$$\begin{aligned}
\mathcal{O}_p(1) &= Q(\tilde{\beta}_n) - Q(\beta_n^*) - (\tilde{\beta}_{n1} - \beta_{n1}^*)^\top \nabla_1 Q(\beta_n^*) + \frac{1}{2}\|D_{Q1}(\tilde{\beta}_{n1} - \beta_{n1}^*)\|^2 \\
&= Q(\tilde{\beta}_n) - Q(\beta_n^*) - \frac{1}{2}\|D_{Q1}^{-1}\nabla_1 Q(\beta_n^*)\|^2 + \frac{1}{2}\|D_{Q1}^{-1}\nabla_1 Q(\beta_n^*)\|^2 \\
&\quad - (\tilde{\beta}_{n1} - \beta_{n1}^*)^\top D_{Q1} D_{Q1}^{-1}\nabla_1 Q(\beta_n^*) + \frac{1}{2}\|D_{Q1}(\tilde{\beta}_{n1} - \beta_{n1}^*)\|^2 \\
&\geq Q(\tilde{\beta}_n) - Q(\beta_n^*) - \frac{1}{2}\|D_{Q1}^{-1}\nabla_1 Q(\beta_n^*)\|^2 + \frac{1}{2}\|D_{Q1}(\tilde{\beta}_{n1} - \beta_{n1}^*) - D_{Q1}^{-1}\nabla_1 Q(\beta_n^*)\|^2,
\end{aligned} \tag{29}$$

where the last term on the right hand side of (29) is of order $\mathcal{O}_p(1)$ as a consequence of Theorem 5. Thus the identity (15) holds. □

For the validity of the bootstrap method we firstly need to prove the assertion of Theorem 2.

Proof of Theorem 2. In this proof we closely follow the technique of Fan and Peng (2004) and Kwon and Kim (2012) to show the existence of a $\sqrt{n/q_n}$-consistent local (global) maximizer of $L^\circ(\beta_n)$ subject to $\beta_{nj} = 0$ for all $q_n < j \leq p_n$ denoted by $\widehat{\beta}_n^\circ$ and its asymptotic equivalence with the maximizer of $Q^\circ(\beta_n)$ denoted by $\tilde{\beta}_n^\circ$.

Let us define a sequence $\alpha_n = \sqrt{q_n/n}$ and a vector $\omega \in \mathbb{R}^{p_n}$ such that $\omega_j = 0$ for $q_n < j \leq p_n$. Then, set $\|\omega\| = C$, with C being a large enough constant. If we can show that for any given ε there exists such a constant C that it holds

$$\mathsf{P}\left\{\sup_{\|\omega\|=C} L^\circ(\beta_n^* + \alpha_n\omega) < L^\circ(\beta_n^*)\right\} \geq 1 - \varepsilon, \tag{30}$$

for n large enough, then with a probability tending to 1 there is a local maximizer, $\widehat{\beta}_n^\circ$, satisfying $\|\widehat{\beta}_n^\circ - \beta_n^*\| = \mathcal{O}_p(\sqrt{q_n/n})$ in the ball $\{\beta_n^* + \alpha_n\omega, \|\omega\| \leq C\}$. For strictly concave likelihood functions with respect to β_{nj}, for $j \leq q_n$, this result is naturally extended to a global maximizer of $L^\circ(\beta_n)$.

Define

$$V_n(\omega) = L^\circ(\beta_n^* + \alpha_n\omega) - L^\circ(\beta_n^*)$$

and use the Taylor's expansion to rewrite $V_n(\omega)$ as

$$\begin{aligned}
V_n(\omega) &= \alpha_n \nabla^\top L^\circ(\beta_n^*)\omega + \frac{1}{2}\omega^\top \nabla^2 L^\circ(\beta_n^*)\omega\alpha_n^2 + \frac{1}{6}\nabla^\top\{\omega^\top \nabla^2 L^\circ(\beta_n^+)\omega\}\omega\alpha_n^3 \\
&\stackrel{\text{def}}{=} I_1^\circ + I_2^\circ + I_3^\circ,
\end{aligned}$$

where β_n^+ lies between β_n^* and $\beta_n^* + \alpha_n\omega$.

Knowing that $\partial L(\beta_n^*)/\partial\beta_{nj} = \mathcal{O}_p(\sqrt{n})$ we can show the same result for $\partial L^\circ(\beta_n^*)/\partial\beta_{nj}$, where one can use the fact that u_i's are independent of the data. This further yields that

$$|I_1^\circ| = |\nabla^\top L^\circ(\beta_n^*)\omega| \leq \alpha_n \|\nabla_1^\top L^\circ(\beta_n^*)\| \|\omega\| = \mathcal{O}_p(\alpha_n\sqrt{nq_n})\|\omega\| = \mathcal{O}_p(\alpha_n^2 n)\|\omega\|.$$

The term I_2° can be expanded into the following

$$\begin{aligned}\frac{1}{n}I_2^\circ = \frac{1}{2}\omega^\top &\left\{\frac{1}{n}\nabla^2 L^\circ(\beta_n^*) - \frac{1}{n}\,\mathsf{E}^\circ\,\nabla^2 L^\circ(\beta_n^*)\right\}\omega\alpha_n^2 \\ &+ \frac{1}{2}\omega^\top\left\{\frac{1}{n}\,\mathsf{E}^\circ\,\nabla^2 L^\circ(\beta_n^*) - \frac{1}{n}\,\mathsf{E}\,\nabla^2 L(\beta_n^*)\right\}\omega\alpha_n^2 \\ &+ \frac{1}{2}\omega^\top\left\{\frac{1}{n}\,\mathsf{E}\,\nabla^2 L(\beta_n^*)\right\}\omega\alpha_n^2. \end{aligned} \tag{31}$$

In addition we have

$$\begin{aligned}&\mathsf{P}\left\{\left\|\frac{1}{n}\nabla_1^2 L^\circ(\beta_n^*) - \frac{1}{n}\,\mathsf{E}^\circ\,\nabla_1^2 L^\circ(\beta_n^*)\right\| \geq \frac{\varepsilon}{q_n^2}\right\} \\ &\qquad \leq \frac{q_n^4}{\varepsilon^2}\sum_{j,k=1}^{q_n}\mathsf{Var}^\circ\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_n(Y_{ni},\beta_n^*)u_i}{\partial\beta_{nj}\partial\beta_{nk}}\right\} \\ &\qquad = \frac{q_n^4}{\varepsilon^2 n^2}\mathcal{O}(q_n^2 n) = o(1).\end{aligned}$$

In other words

$$\left\|\frac{1}{n}\nabla_1^2 L^\circ(\beta_n^*) - \frac{1}{n}\,\mathsf{E}^\circ\,\nabla_1^2 L^\circ(\beta_n^*)\right\| = o_p\left(\frac{1}{q_n^2}\right). \tag{32}$$

The same holds for the second part of (31) as was shown in (20). Thus,

$$I_2^\circ = \frac{1}{2}n\alpha_n^2 o_p\left(\frac{1}{q_n^2}\right)\|\omega\|^2 + \frac{1}{2}\alpha_n^2\omega^\top\{\mathsf{E}\,\nabla_1^2 L(\beta_n^*)\}\omega = -\frac{1}{2}n\alpha_n^2\omega^\top I_n(\beta_n^*)\omega + o_p(1)n\alpha_n^2\|\omega\|^2.$$

Subsequently, the term I_3° can be bounded as

$$\begin{aligned}|I_3^\circ| &= \left|\frac{1}{6}\nabla^\top\left\{\omega^\top\nabla^2 L^\circ(\beta_n^*)\omega\right\}\omega\alpha_n^3\right| = \left|\frac{1}{6}\sum_{j,k,l=1}^{q_n}\frac{\partial^3 L^\circ(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}}\omega_j\omega_k\omega_l\alpha_n^3\right| \\ &\leq \frac{1}{6}\sum_{i=1}^{n}\left[\sum_{j,k,l=1}^{q_n}\left\{\frac{\partial^3 \log f_n(Y_{ni},\beta_n^*)u_i}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}}\right\}^2\right]^{1/2}\|\omega\|^3\alpha_n^3 \\ &\leq \frac{1}{6}\sum_{i=1}^{n}\left\{\sum_{j,k,l=1}^{q_n}U_{njkl}^2(Y_{ni})u_i^2\right\}^{1/2}\|\omega\|^3\alpha_n^3 = \mathcal{O}_p(q_n^{3/2}\alpha_n)n\alpha_n^2\|\omega\|^3 \\ &= o_p(n\alpha_n^2)\|\omega\|^3.\end{aligned} \tag{33}$$

Hence, the results are analogous to the non-bootstrapped case and, for $\|\omega\|$ large enough, the negative term I_2° dominates the rest of $V_n(\omega)$, which proves (30) and the existence of $\widehat{\beta}_n^\circ$ such that $\|\widehat{\beta}_n^\circ - \beta_n^*\| = \mathcal{O}_p(\sqrt{q_n/n})$ as well as $\|\widehat{\beta}_n^\circ - \tilde{\beta}_n\| = \mathcal{O}_p(\sqrt{q_n/n})$.

Following the approach of Kwon and Kim (2012) and showing that

$$\mathsf{P}\left\{\max_{\beta_n\in\Omega_n} Q^\circ(\beta_n) \le Q^\circ(\widehat{\beta}_n^\circ)\right\} \to 1, \qquad \text{as } n\to\infty, \tag{34}$$

we prove the asymptotic equivalence of $\widehat{\beta}_n^\circ$ and $\tilde{\beta}_n^\circ$ for the strictly concave log-likelihood functions and thereby the $\sqrt{n/q_n}$-consistency and sparsity of the latter.

Firstly, we need to show that for $\widehat{\beta}_n^\circ$ it holds

$$\mathsf{P}\left\{\max_{q_n<j\le p_n}\left|\frac{\partial L^\circ(\widehat{\beta}_n^\circ)}{\partial\beta_{nj}}\right| \le n\lambda_n\right\} \to 1, \qquad \text{as } n\to\infty. \tag{35}$$

We use Taylor's expansion around β_n^* for all $q_n < j \le p_n$

$$\begin{aligned}\frac{\partial L^\circ(\widehat{\beta}_n^\circ)}{\partial\beta_{nj}} &= \frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}} + (\widehat{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \nabla_1 \frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}} \\ &\quad + \frac{1}{2}(\widehat{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \nabla_1^2 \frac{\partial L^\circ(\beta_n^+)}{\partial\beta_{nj}} (\widehat{\beta}_{n1}^\circ - \beta_{n1}^*),\end{aligned}$$

where β_n^+ lies between β_n^* and $\widehat{\beta}_n^\circ$. Expanding the right-hand side and using the properties of β_n^* and $\widehat{\beta}_n^\circ$ and Cauchy-Schwarz inequality, we can write

$$\begin{aligned}&\mathsf{P}\left\{\max_{q_n<j\le p_n}\left|\frac{\partial L^\circ(\widehat{\beta}_n^\circ)}{\partial\beta_{nj}}\right| > n\lambda_n\right\} \\ &\le \mathsf{P}\left\{\max_{q_n<j\le p_n}\left|\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}}\right| > \frac{n\lambda_n}{5}\right\} \\ &\quad + \mathsf{P}\left\{\max_{q_n<j\le p_n}\left\|\nabla_1\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}} - \mathsf{E}^\circ\,\nabla_1\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}}\right\| \|\widehat{\beta}_n^\circ - \beta_n^*\| > \frac{n\lambda_n}{5}\right\} \\ &\quad + \mathsf{P}\left\{\max_{q_n<j\le p_n}\left\|\nabla_1\frac{\partial L(\beta_n^*)}{\partial\beta_{nj}} - \mathsf{E}\,\nabla_1\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}}\right\| \|\widehat{\beta}_n^\circ - \beta_n^*\| > \frac{n\lambda_n}{5}\right\} \\ &\quad + \mathsf{P}\left\{\max_{q_n<j\le p_n}\left\|\mathsf{E}\,\nabla_1\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}}\right\| \|\widehat{\beta}_n^\circ - \beta_n^*\| > \frac{n\lambda_n}{5}\right\} \\ &\quad + \mathsf{P}\left\{\max_{q_n<j\le p_n}\left\|\nabla_1^2\frac{\partial L^\circ(\beta_n^+)}{\partial\beta_{nj}}\right\| \|\widehat{\beta}_n^\circ - \beta_n^*\|^2 > \frac{2n\lambda_n}{5}\right\} \\ &\overset{\text{def}}{=} P_1 + P_2 + P_3 + P_4 + P_5,\end{aligned}$$

Following Lemma A.1 from Kwon and Kim (2012), we can bound the terms P_1 - P_5.

For P_1 we have, due to conditions *(A2)*, *(A5)* and Markov's inequality, for any constant κ that

$$\mathsf{P}\left\{\left|\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}}\right| > \sqrt{n}\kappa\right\} \le (\sqrt{n}\kappa)^{-2}\,\mathsf{E}\left(\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}}\right)^2 = \mathcal{O}(\kappa^{-2}),$$

for all $j \le p_n$. Thus,

$$P_1 \le \sum_{j=q_n+1}^{p_n} \mathsf{P}\left\{\left|\frac{\partial L^\circ(\beta_n^*)}{\partial\beta_{nj}}\right| > \frac{n\lambda_n}{5}\right\} = \mathcal{O}\left\{\frac{p_n}{(\sqrt{n}\lambda_n)^2}\right\} \to 0,$$

as $n \to \infty$. For the second term we need the following result

$$\mathsf{P}\left\{\left\|\nabla_1 \frac{\partial L^\circ(\beta_n^*)}{\partial \beta_{nj}} - \mathsf{E}^\circ \nabla_1 \frac{\partial L^\circ(\beta_n^*)}{\partial \beta_{nj}}\right\| > \sqrt{nq_n}\kappa\right\} = \mathcal{O}(\kappa^{-2}),$$

for all $j \leq p_n$. Using Chebyshev's inequality and condition *(A5)* we get

$$\begin{aligned}
&\mathsf{P}\left\{\left\|\nabla_1 \frac{\partial L^\circ(\beta_n^*)}{\partial \beta_{nj}} - \mathsf{E}^\circ \nabla_1 \frac{\partial L^\circ(\beta_n^*)}{\partial \beta_{nj}}\right\| > \sqrt{nq_n}\kappa\right\} \\
&\qquad \leq (\sqrt{nq_n}\kappa)^{-2}\, \mathsf{Var}^\circ \sum_{k=1}^{q_n}\left\{\sum_{i=1}^{n} \frac{\partial^2 \log f_n(Y_{ni,\beta_n^*})u_i}{\partial \beta_{nk}\partial \beta_{nk}}\right\} \\
&\qquad = \mathcal{O}(\kappa^{-2}),
\end{aligned}$$

for any positive constant κ. Then we can write

$$\begin{aligned}
P_2 \leq\ & \mathsf{P}\left(\|\widehat{\beta}_n^\circ - \beta_n^*\| > \frac{q_n}{\sqrt{n}}\right) \\
&+ \mathsf{P}\left\{\max_{q_n < j \leq p_n}\left\|\nabla_1 \frac{\partial L^\circ(\beta_n^*)}{\partial \beta_{nj}} - \mathsf{E}^\circ \nabla_1 \frac{\partial L^\circ(\beta_n^*)}{\partial \beta_{nj}}\right\| > \frac{n\sqrt{n}\lambda_n}{5q_n}\right\} \\
=\ & o(1) + \mathcal{O}\left[\frac{p_n}{\left\{n\lambda_n/(q_n\sqrt{q_n})\right\}^2}\right] \to 0,
\end{aligned}$$

for $n \to \infty$.

As already shown in Kwon and Kim (2012) in the proof of Theorem 1, the terms P_3 and P_4 both go to zero with n tending to infinity. Hence, the last term we need to bound is P_5 which can be done by showing that

$$\mathsf{P}\left\{\left\|\nabla_1^2 \frac{\partial L^\circ(\beta_n^+)}{\partial \beta_{nj}}\right\| > nq_n\kappa\right\} = \mathcal{O}(\kappa^{-2}),$$

for κ as before. Using conditions *(A4)*, *(A5)* and Markov's inequality we get

$$\begin{aligned}
&\mathsf{P}\left\{\left\|\nabla_1^2 \frac{\partial L^\circ(\beta_n^+)}{\partial \beta_{nj}}\right\| > nq_n\kappa\right\} \\
&\qquad \leq (nq_n\kappa)^{-2}\, \mathsf{E}\left\{\left\|\nabla_1^2 \frac{\partial L^\circ(\beta_n^+)}{\partial \beta_{nj}}\right\|\right\}^2 \\
&\qquad = (nq_n\kappa)^{-2} \sum_{k,l=1}^{q_n} \mathsf{E}\left\{\sum_{i=1}^{n} \frac{\partial^3 \log f_n(Y_{ni}, \beta_n^+)u_i}{\partial \beta_{nj}\partial \beta_{nk}\partial \beta_{nl}}\right\}^2 \\
&\qquad = \mathcal{O}(\kappa^{-2}).
\end{aligned}$$

Now it follows that

$$\begin{aligned}
P_5 \leq\ & \mathsf{P}\left(\|\widehat{\beta}_n^\circ - \beta_n^*\|^2 > \frac{q_n\sqrt{q_n}}{n}\right) + \mathsf{P}\left\{\max_{q_n < j \leq p_n}\left\|\nabla_1^2 \frac{\partial L^\circ(\beta_n^+)}{\partial \beta_{nj}}\right\| > \frac{2n^2\lambda_n}{5q_n\sqrt{q_n}}\right\} \\
=\ & o(1) + \mathcal{O}\left[\frac{p_n}{\left\{n\lambda_n/(q_n^2\sqrt{q_n})\right\}^2}\right] \to 0,
\end{aligned}$$

as $n \to \infty$, what proves (35).

Hence, we can move on to prove (34), where we firstly use Taylor's expansion as follows

$$L^\circ(\beta_n) - L^\circ(\widehat{\beta}_n^\circ) = (\beta_n - \widehat{\beta}_n^\circ)^\top \nabla L^\circ(\widehat{\beta}_n^\circ) + \frac{1}{2}(\beta_n - \widehat{\beta}_n^\circ)^\top \nabla^2 L^\circ(\beta_n^+)(\beta_n - \widehat{\beta}_n^\circ),$$

for some β_n^+ lying between β_n^* and $\widehat{\beta}_n^\circ$. Using (35) and the definition of $\widehat{\beta}_n^\circ$ we get

$$(\beta_n - \widehat{\beta}_n^\circ)^\top \nabla L^\circ(\widehat{\beta}_n^\circ) = \sum_{j=1}^{p_n} \frac{\partial L^\circ(\widehat{\beta}_n^\circ)}{\partial \beta_{nj}}(\beta_{nj} - \widehat{\beta}_{nj}^\circ) \leq \sum_{j=q_n+1}^{p_n} o_p(n\lambda_n)|\beta_{nj}|.$$

Further, from the bootstrapped equivalent of *(A6)* with the positive constant M_{10} and the Cauchy-Schwarz inequality it follows

$$\frac{1}{2}(\beta_n - \widehat{\beta}_n^\circ)^\top \nabla^2 L^\circ(\beta_n^+)(\beta_n - \widehat{\beta}_n^\circ) \leq -nM_{10}\|\beta_n - \widehat{\beta}_n^\circ\|^2,$$

which implies that

$$Q^\circ(\beta_n) - Q^\circ(\widehat{\beta}_n^\circ) \leq \sum_{j=1}^{p_n} n w_{nj},$$

where

$$\begin{aligned} w_{nj} = {} & o_p(\lambda_n)|\beta_{nj}|\,\mathbf{I}(j > q_n) - M_{10}(\beta_{nj} - \widehat{\beta}_{nj}^\circ)^2 \\ & + n^{-1}\sum_{i=1}^{n} u_i \sum_{j=1}^{p_n} p_{\lambda_n}(|\widehat{\beta}_{nj}^\circ|) - n^{-1}\sum_{i=1}^{n} u_i \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|). \end{aligned}$$

Using the fact that $n^{-1}\sum_{i=1}^{n} u_i = \mathcal{O}_p(1)$ and the rest of the arguments from the proof of Theorem 2 in Kwon and Kim (2012), we can conclude that for all large enough n it holds $\sum_{j=1}^{p_n} w_{nj} \leq 0$, which proves (34) and subsequently the assertions of the theorem.

□

Let us now show the Fisher type of expansion for the bootstrapped penalized likelihood function and its maximizer.

Theorem 6. *Under the conditions of Theorem 1 it holds*

$$D_{Q1}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*) = D_{Q1}^{-1}\nabla_1 Q^\circ(\beta_n^*) + o_p\left(\frac{1}{\sqrt{q_n}}\right).$$

Proof of Theorem 6. By Taylor's expansion of $\nabla_1 Q^\circ(\tilde{\beta}_n^\circ)$, which is equal to zero by

definition, around β_n^* we have

$$\begin{aligned}
0 = \nabla_1 Q^\circ(\tilde{\beta}_n^\circ) &= \nabla_1 L^\circ(\tilde{\beta}_n^\circ) - \sum_{i=1}^n u_i \nabla_1 P_{\lambda_n}(\tilde{\beta}_n^\circ) \\
&= \nabla_1 L^\circ(\beta_n^*) + (\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \nabla_1^2 L^\circ(\beta_n^*) + \frac{1}{2}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \nabla_1^2 \{\nabla_1 L^\circ(\beta_n^+)\}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*) \\
&\quad - \sum_{i=1}^n u_i \nabla_1 P_{\lambda_n}(\beta_n^*) - (\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \sum_{i=1}^n u_i \nabla_1^2 P_{\lambda_n}(\beta_n^{++}),
\end{aligned}$$

where β_n^+ and β_n^{++} lie between $\tilde{\beta}_n^\circ$ and β_n^*. We can rewrite this into

$$\begin{aligned}
&(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \left\{ \nabla_1^2 L^\circ(\beta_n^*) - \sum_{i=1}^n u_i \nabla_1^2 P_{\lambda_n}(\beta_n^{++}) \right\} \\
&\qquad = -\nabla_1 Q^\circ(\beta_n^*) - \frac{1}{2}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \nabla_1^2 \{\nabla_1 L^\circ(\beta_n^+)\}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*). \qquad (36)
\end{aligned}$$

Subsequently, the term on the left-hand side can be expanded as

$$\begin{aligned}
\nabla_1^2 L^\circ(\beta_n^*) - \sum_{i=1}^n u_i \nabla_1^2 P_{\lambda_n}(\beta_n^{++}) &= \nabla_1^2 L^\circ(\beta_n^*) - \mathsf{E}^\circ \nabla_1^2 L^\circ(\beta_n^*) \\
&\quad + \nabla_1^2 L(\beta_n^*) - \mathsf{E} \nabla_1^2 L(\beta_n^*) \\
&\quad + n \nabla_1^2 P_{\lambda_n}(\beta_n^*) - \sum_{i=1}^n u_i \nabla_1^2 P_{\lambda_n}(\beta_n^{++}) \\
&\quad + \mathsf{E} \nabla_1^2 Q(\beta_n^*) \\
&\stackrel{\text{def}}{=} I_4^\circ + I_2 + I_5^\circ - D_{Q1}^2,
\end{aligned}$$

where $\|I_4^\circ\| = \mathcal{o}_p(n/q_n^2)$ and $\|I_2\| = \mathcal{o}_p(n/q_n^2)$ as shown in (32) and (20), respectively. Subsequently, both terms of I_5° tend to zero in probability, due to the properties of β_n^*, β_n^{++} and the SCAD penalty with parameter λ, i.e. we can write $\|I_5^\circ\| = \mathcal{o}_p(1)$. Thus,

$$(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \left\{ \nabla_1^2 L^\circ(\beta_n^*) - \sum_{i=1}^n u_i \nabla_1^2 P_{\lambda_n}(\beta_n^{++}) \right\} = (\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top D_{Q1}^2 + \mathcal{o}_p\left(\sqrt{\frac{n}{q_n^3}}\right).$$

Next we have, for sufficiently large n,

$$\begin{aligned}
&\left\| \frac{1}{2n}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*)^\top \nabla_1^2 \{\nabla_1 L^\circ(\beta_n^+)\}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*) \right\|^2 \\
&\qquad \le \frac{1}{n^2} \sum_{j,k,l=1}^{q_n} \left\{ \frac{\partial^3 L^\circ(\beta_n^+)}{\partial \beta_{nj} \partial \beta_{nk} \partial \beta_{nl}} \right\}^2 \|\tilde{\beta}_n^\circ - \beta_n^*\|^4 \\
&\qquad \le \frac{1}{n^2} \sum_{i=1}^n n \sum_{j,k,l=1}^{q_n} \left\{ U_{njkl}^2(Y_{ni}) u_i^2 \right\} \|\tilde{\beta}_n^\circ - \beta_n^*\|^4 \\
&\qquad = \mathcal{O}_p(q_n^3) \mathcal{O}_p\left(\frac{q_n^2}{n^2}\right) = \mathcal{o}_p\left(\frac{1}{nq_n}\right).
\end{aligned}$$

Hence, it follows that the norm of the last term on the right-hand side of (36) is of order $\mathcal{O}_p(\sqrt{n/q_n})$ and we can rewrite the equation into

$$D_{Q1}^2(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*) = \nabla_1 Q^\circ(\beta_n^*) + \mathcal{O}_p\left(\sqrt{\frac{n}{q_n}}\right),$$

from which it follows

$$D_{Q1}(\tilde{\beta}_{n1}^\circ - \beta_{n1}^*) = D_{Q1}^{-1}\nabla_1 Q^\circ(\beta_n^*) + \mathcal{O}_p\left(\sqrt{\frac{1}{q_n}}\right),$$

what completes the proof. □

Combining theorems 5 and 6 we get the following Fisher type of expansion for the SCAD estimator $\tilde{\beta}_n$ and its bootstrapped counterpart $\tilde{\beta}_n^\circ$

$$D_{Q1}(\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1}) = D_{Q1}^{-1}\left\{\nabla_1 Q^\circ(\beta_n^*) - \nabla_1 Q(\beta_n^*)\right\} + \mathcal{O}_p\left(\frac{1}{\sqrt{q_n}}\right). \tag{37}$$

Collected results together with the following Lemma lead to the proofs of theorems 3 and 4.

Lemma 2. *Under the conditions of Theorem 1 we can write*

$$Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) = \frac{1}{2}\|D_{Q1}(\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1})\|^2 + \mathcal{O}_p(1). \tag{38}$$

Proof of Lemma 2. By Taylor's expansion of $Q^\circ(\tilde{\beta}_n)$ around $\tilde{\beta}_n^\circ$ we get

$$Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) = \frac{1}{2}(\tilde{\beta}_n - \tilde{\beta}_n^\circ)^\top\left\{-\nabla^2 Q^\circ(\beta_n^+)\right\}(\tilde{\beta}_n - \tilde{\beta}_n^\circ), \tag{39}$$

for some β_n^+ lying between $\tilde{\beta}_n$ and $\tilde{\beta}_n^\circ$. Assuming that n is large enough we can, similarly as in the proof of Lemma 1, use the sparsity property of $\tilde{\beta}_n$ and $\tilde{\beta}_n^\circ$ and write

$$Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) = \frac{1}{2}(\tilde{\beta}_{n1} - \tilde{\beta}_{n1}^\circ)^\top\left\{-\nabla_1^2 Q^\circ(\beta_n^+)\right\}(\tilde{\beta}_{n1} - \tilde{\beta}_{n1}^\circ).$$

Then we have

$$\begin{aligned}
\nabla_1^2 Q^\circ(\beta_n^+) &= \nabla_1^2 L^\circ(\beta_n^+) - \nabla_1^2 L^\circ(\beta_n^*) + \sum_{i=1}^n u_i \nabla_1^2 P_{\lambda_n}(\beta_n^*) - \sum_{i=1}^n u_i \nabla_1^2 P_{\lambda_n}(\beta_n^+) \\
&\quad + \nabla_1^2 L^\circ(\beta_n^*) - \mathsf{E}^\circ\,\nabla_1^2 L^\circ(\beta_n^*) + \sum_{i=1}^n (1-u_i)\nabla_1^2 P_{\lambda_n}(\beta_n^*) \\
&\quad + \nabla_1^2 L(\beta_n^*) - \mathsf{E}\,\nabla_1^2 L(\beta_n^*) + \mathsf{E}\,\nabla_1^2 L(\beta_n^*) \\
&\stackrel{\text{def}}{=} I_6^\circ + n^{-1}\sum_{i=1}^n u_i I_3 + I_4^\circ + I_7^\circ + I_2 - D_{Q1}^2.
\end{aligned}$$

From (20), (22) and (32), it follows that

$$\frac{1}{2}(\tilde{\beta}_{n1}-\tilde{\beta}^{\circ}_{n1})^{\top}\left(n^{-1}\sum_{i=1}^{n}u_iI_3+I_4^{\circ}+I_2\right)(\tilde{\beta}_{n1}-\tilde{\beta}^{\circ}_{n1})=o_p(1) \tag{40}$$

and we need to show the same for I_6° and I_7°.

Let us consider the term I_6° first. Its norm can be bounded for some β_n^{++} lying between β_n^{+} and β_n^{*} accordingly

$$\|I_6^{\circ}\|^2=\|(\beta^{+}_{n1}-\beta^{*}_{n1})^{\top}\nabla_1^{\top}\left\{\nabla_1^2L^{\circ}(\beta_n^{++})\right\}\|^2\leq\sum_{j,k,l=1}^{q_n}\left\{\frac{\partial^3L^{\circ}(\beta_n^{++})}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}}\right\}^2\|\beta_n^{+}-\beta_n^{*}\|^2$$
$$\leq\mathcal{O}_p\left(\frac{q_n}{n}\right)\sum_{j,k,l=1}^{q_n}n\sum_{i=1}^{n}U^2_{njkl}(Y_{ni})u_i^2=\mathcal{O}_p(q_n^4n).$$

Then

$$\frac{1}{2}(\tilde{\beta}_{n1}-\tilde{\beta}^{\circ}_{n1})^{\top}I_6^{\circ}(\tilde{\beta}_{n1}-\tilde{\beta}^{\circ}_{n1})\leq\frac{1}{2}\|I_6^{\circ}\|\|\tilde{\beta}_n-\tilde{\beta}_n^{\circ}\|^2=\mathcal{O}_p(q_n^2\sqrt{n})\mathcal{O}_p\left(\frac{q_n}{n}\right)=o_p(1), \tag{41}$$

due to conditions *(A1)*, *(A1)* and assertion of Theorem 2.

For the term I_7° it holds

$$\frac{1}{2}(\tilde{\beta}_{n1}-\tilde{\beta}^{\circ}_{n1})^{\top}I_7^{\circ}(\tilde{\beta}_{n1}-\tilde{\beta}^{\circ}_{n1})\xrightarrow{\mathsf{P}}0,\qquad\text{as } n\to\infty, \tag{42}$$

because of the properties of the SCAD penalty and our assumptions on λ and non-zero coefficients of the parameter vector β_n^{*}.

Finally, by the same arguments as in Lemma 1, the combination of (40), (41) and (42) completes the proof of (38). □

Proof of Theorem 3. In order to prove the Wilks type of expansion for the bootstrapped penalized likelihood ratio, we use the Taylor's expansion of $Q^{\circ}(\tilde{\beta}_n^{\circ})$ around $\tilde{\beta}_n$

$$Q^{\circ}(\tilde{\beta}_n^{\circ})=Q^{\circ}(\tilde{\beta}_n)+(\tilde{\beta}_n^{\circ}-\tilde{\beta})^{\top}\nabla Q^{\circ}(\tilde{\beta}_n)+\frac{1}{2}(\tilde{\beta}_n^{\circ}-\tilde{\beta}_n)^{\top}\nabla^2Q^{\circ}(\tilde{\beta}_n)(\tilde{\beta}_n^{\circ}-\tilde{\beta}_n)$$
$$+\frac{1}{6}\nabla^{\top}\left\{(\tilde{\beta}_n^{\circ}-\tilde{\beta}_n)^{\top}\nabla^2Q^{\circ}(\beta_n^{+})(\tilde{\beta}_n^{\circ}-\tilde{\beta}_n)\right\}(\tilde{\beta}_n^{\circ}-\tilde{\beta}_n), \tag{43}$$

with β_n^{+} lying between $\tilde{\beta}_n$ and $\tilde{\beta}_n^{\circ}$.

The second last term of the right-hand side can be expanded and bounded in a similar fashion as the right-hand side term of the equation (39) and thus can be set to be equal to $-\frac{1}{2}\|D_{Q1}(\tilde{\beta}^{\circ}_{n1}-\tilde{\beta}_{n1})\|^2+o_p(1)$.

Knowing that the third derivatives of the SCAD penalty function are zero for all of the values of the coefficients β_{nj}, we can relate the last term from the expansion to the

term I_3° from (33), where we set $\alpha_n = 1$ and $\omega = \tilde{\beta}_n^\circ - \tilde{\beta}_n$. This means that the last term from (43) is of order $o_p(1)$.

Thus, invoking the sparsity property, we have

$$\begin{aligned}
o_p(1) &= Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) - (\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1})^\top \nabla_1 Q^\circ(\tilde{\beta}_n) + \frac{1}{2}\left\| D_{Q1}(\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1}) \right\|^2 \\
&= Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) - (\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1})^\top \left\{ \nabla_1 Q^\circ(\tilde{\beta}_n) - \nabla_1 Q(\tilde{\beta}_n) \right\} \\
&\quad + \frac{1}{2}\left\| D_{Q1}(\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1}) \right\|^2 \\
&= Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) - (\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1})^\top D_{Q1} D_{Q1}^{-1} \left\{ \nabla_1 Q^\circ(\tilde{\beta}_n) - \nabla_1 Q(\tilde{\beta}_n) \right\} \\
&\quad - \frac{1}{2}\left\| D_{Q1}^{-1} \left\{ \nabla_1 Q^\circ(\tilde{\beta}_n) - \nabla_1 Q(\tilde{\beta}_n) \right\} \right\|^2 + \frac{1}{2}\left\| D_{Q1}^{-1} \left\{ \nabla_1 Q^\circ(\tilde{\beta}_n) - \nabla_1 Q(\tilde{\beta}_n) \right\} \right\|^2 \\
&\quad + \frac{1}{2}\left\| D_{Q1}(\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1}) \right\|^2 \\
&\geq Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) - \frac{1}{2}\left\| D_{Q1}^{-1} \left\{ \nabla_1 Q^\circ(\tilde{\beta}_n) - \nabla_1 Q(\tilde{\beta}_n) \right\} \right\|^2 \\
&\quad + \frac{1}{2}\left\| D_{Q1}(\tilde{\beta}_{n1}^\circ - \tilde{\beta}_{n1}) - D_{Q1}^{-1} \left\{ \nabla_1 Q^\circ(\tilde{\beta}_n) - \nabla_1 Q(\tilde{\beta}_n) \right\} \right\|^2 + o_p(1),
\end{aligned}$$

where the second last term from the inequality is of order $o_p(1)$ as shown in (37). Hence, (16) holds. □

Proof of Theorem 4. If we can show that $G_n(\cdot)$ and $G_n^\circ(\cdot)$ have the same limiting distribution Ω_n, the assertion of Theorem 4 will follow.

By the conclusion of Theorem 1 we have that $Q(\tilde{\beta}_n) - Q(\beta_n^*) = \frac{1}{2}\|D_{Q1}^{-1}\nabla_1 Q(\beta_n^*)\|^2 + o_p(1)$. Using the notation $D_L^2 = -\,\mathsf{E}\,\nabla^2 L(\beta_n^*)$ and further D_{L1}^2 and D_{L1}^{-2} for its first $q_n \times q_n$ submatrix and the corresponding inverse, we can rewrite the formula into

$$Q(\tilde{\beta}_n) - Q(\beta_n^*) = \frac{1}{2}\|D_{L1}^{-1}\nabla_1 L(\beta_n^*)\|^2 + o_p(1).$$

This holds because of the properties of the SCAD penalty function and our assumptions on the true coefficients β_n^*, see condition *(A1)*, and the penalty parameter λ_n. By the same arguments, for the bootstrapped penalized likelihood ratio it follows that

$$Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n) = \frac{1}{2}\|D_{L1}^{-1}\left\{\nabla_1 L^\circ(\beta_n^*) - \nabla_1 L(\beta_n^*)\right\}\|^2 + o_p(1).$$

Similarly as in Fan and Peng (2004) one can show that

$$2\left\{Q(\tilde{\beta}_n) - Q(\beta_n^*)\right\} \xrightarrow{\mathcal{L}} \chi^2_{q_n},$$

which follows from the fact that $D_{L1}^{-1}\nabla_1 L(\beta_n^*)$ is asymptotically normally distributed, i.e.

$$D_{L1}^{-1}\nabla_1 L(\beta_n^*) \xrightarrow{\mathcal{L}} \mathrm{N}_{q_n}(0, I_{q_n}).$$

For the conditional distribution of the bootstrapped penalized likelihood ratio conditioned on the data and the penalty parameter λ_n, the procedure of showing the asymptotic distribution is analogous.

For $D_{L1}^{-1}\{\nabla_1 L^\circ(\beta_n^*) - \nabla_1 L(\beta_n^*)\}$ we have

$$\mathsf{E}^\circ\left[D_{L1}^{-1}\{\nabla_1 L^\circ(\beta_n^*) - \nabla_1 L(\beta_n^*)\}\right] = \mathsf{E}^\circ \sum_{i=1}^{n} D_{L1}^{-1}\nabla_1 \log f_n(Y_{ni}, \beta_n^*)(u_i - 1) = 0.$$

In order to show that the variance of $G^\circ(\cdot)$ tends to I_{q_n} in probability, we firstly need to show that

$$\left\| \frac{1}{n}\nabla_1 L(\beta_n^*)\nabla_1^\top L(\beta_n^*) + \frac{1}{n}\,\mathsf{E}\,\nabla_1^2 L(\beta_n^*) \right\| = o_p\left(\frac{1}{q_n^2}\right).$$

This can be proved analogously to the bound in (20).

Then it follows

$$\begin{aligned}
&\mathsf{Var}^\circ\left[D_{L1}^{-1}\{\nabla_1 L^\circ(\beta_n^*) - \nabla_1 L(\beta_n^*)\}\right] \\
&\quad= \mathsf{Var}^\circ\left[\sum_{i=1}^{n} D_{L1}^{-1}\nabla_1 \log f_n(Y_{ni}, \beta_n^*)(u_i - 1)\right] \\
&\quad= \sum_{i=1}^{n} D_{L1}^{-1}\nabla_1 \log f_n(Y_{ni}, \beta_n^*)\nabla_1^\top \log f_n(Y_{ni}, \beta_n^*) D_{L1}^{-1} \\
&\quad= D_{L1}^{-1}\left\{\sum_{i=1}^{n} \nabla_1 \log f_n(Y_{ni}, \beta_n^*)\nabla_1^\top \log f_n(Y_{ni}, \beta_n^*) - D_{L1} + D_{L1}\right\} D_{L1}^{-1} \\
&\quad\leq \gamma_{max}\{I_{n1}(\beta_n^*)\}\left\| \frac{1}{n}\nabla_1 L(\beta_n^*)\nabla_1^\top L(\beta_n^*) + \frac{1}{n}\,\mathsf{E}\,\nabla_1^2 L(\beta_n^*)\right\| + I_{q_n} \\
&\quad= I_{q_n} + o_p(1).
\end{aligned}$$

Now it suffices to show that the random variable $D_{L1}^{-1}\nabla_1 \log f_n(Y_{ni}, \beta_n^*)(u_i - 1)$ satisfies Lindeberg's condition, which is done analogously as in Fan and Peng (2004) for the non-bootstrapped case, and by the central limit theorem it then holds that

$$D_{L1}^{-1}\{\nabla_1 L^\circ(\beta_n^*) - \nabla_1 L(\beta_n^*)\} \xrightarrow{\mathcal{L}} \mathrm{N}_{q_n}(0, I_{q_n}).$$

Thus, we can conclude that

$$2\left\{Q^\circ(\tilde{\beta}_n^\circ) - Q^\circ(\tilde{\beta}_n)\right\} \xrightarrow{\mathcal{L}} \chi_{q_n}^2,$$

which we needed to prove. □

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**: 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**: 1232–1259.

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.

Chernozhukov, V., Härdle, W. K., Huang, C. and Wang, W. (2018). LASSO-Driven Inference in Time and Space, *arXiv preprint arXiv:1806.05081.*

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.

Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters, The Annals of Statistics **32**: 928–961.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software **33**: 1–22.

Härdle, W. K. and Mammen, E. (1993). Comparing Nonparametric versus Parametric Regression Fits, Annals of Statistics **21**: 1926–1947.

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

Jurado, K., Ludvigson, S. C. and Ng, S. (2015). Measuring Uncertainty, American Economic Review **105**: 1177–1216.

Kim, Y., Choi, H. and Oh, H. S. (2008). Smoothly Clipped Absolute Deviation in High Dimensions, Journal of the American Statistical Association **103**: 1665–1673.

Kwon, S. and Kim, Y. (2012). Large Sample Properties of the SCAD-Penalized Maximum Likelihood Estimation on High Dimensions, Statistica Sinica **22**: 629–653.

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 **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**: 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**: 389–429.

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

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**: 267–288.

Wang, H. and Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation, Journal of the American Statistical Association **102**: 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**: 1418–1429.

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

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

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 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.

021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.

022 " Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.

023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.

024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.

025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.

026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.

027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.

028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.

030 "Gaussian Process Forecast with multidimensional distributional entries" by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018.

031 "Instrumental variables regression" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

032 "Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective" by Li Guo, Yubo Tao and Wolfgang Karl Härdle, July 2018.

033 "Optimal contracts under competition when uncertainty from adverse selection and moral hazard are present" by Natalie Packham, August 2018.

034 "A factor-model approach for correlation scenarios and correlation stress-testing" by Natalie Packham and Fabian Woebbeking, August 2018.

035 "Correlation Under Stress In Normal Variance Mixture Models" by Michael Kalkbrener and Natalie Packham, August 2018.

036 "Model risk of contingent claims" by Nils Detering and Natalie Packham, August 2018.

037 "Default probabilities and default correlations under stress" by Natalie Packham, Michael Kalkbrener and Ludger Overbeck, August 2018.

038 "Tail-Risk Protection Trading Strategies" by Natalie Packham, Jochen Papenbrock, Peter Schwendner and Fabian Woebbeking, August 2018.

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 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

039 "Penalized Adaptive Forecasting with Large Information Sets and Structural Changes" by Lenka Zbonakova, Xinjue Li and Wolfgang Karl Härdle, August 2018.

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.