DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

The Prior Adaptive Group Lasso and the Factor Zoo

Kristoffer Pons Bertelsen

CREATES Research Paper 2022-05

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 9352 1936

The Prior Adaptive Group Lasso and the Factor Zoo*

Kristoffer Pons Bertelsen†

Abstract

This paper develops and presents the prior adaptive group lasso (pag-lasso) for generalized linear models. The pag-lasso is an extension of the prior lasso, which allows for the use of existing information in the lasso estimation. We show that the estimator exhibits properties similar to the adaptive group lasso. The performance of the pag-lasso estimator is illustrated in a Monte Carlo study. The estimator is used to select the set of relevant risk factors in asset pricing models while requiring that the chosen factors must be able to price the test assets as well as the unselected factors. The study shows that the pag-lasso yields a set of factors that explain the time variation in the returns while delivering estimated pricing errors close to zero. We find that canonical low-dimensional factor models from the asset pricing literature are insufficient to price the cross section of the test assets together with the remaining traded factors. The required number of pricing factors to include at any given time is closer to 20.

Keywords: Asset Pricing, Factor Selection, Factor Zoo, High-Dimensional Modeling, Prior Information, Variable Selection

JEL: C13, C33, C38, C51, C55, C58, G12

This version: January 20, 2022

*The author thanks Tom Engsted, Anders Bredahl Kock, Sophocles Mavroeidis, Jeremy Large, and Yunus Emre Ergemen for useful comments and discussions. The author thanks seminar participants at Aarhus University and University of Oxford. This work was supported by CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78) funded by the Danish National Research Foundation, and the Center for Scientific Computing, Aarhus (CSCAA).

†CREATES, Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, DK-8210 Aarhus V, Denmark. E-mail: kristoffer@econ.au.dk.

1 Introduction

This paper develops and presents the prior adaptive group lasso (pag-lasso), which combines the prior lasso of Jiang, He, and Zhang (2016) with the adaptive group lasso (aglasso) of Wang and Leng (2008), allowing the researcher to perform simultaneous variable selection and parameter estimation on sets of variables with a natural grouping while taking previously obtained information into account. In line with Jiang et al. (2016) developing the prior lasso, we refer to this previously obtained information on the model specification as the "prior information".[1] We demonstrate a way to utilize this prior information that is also flexible enough to not throw away potentially relevant information in the current data set. We prove that the pag-lasso features the same consistency rates as the adaptive group lasso in a generalized linear model as seen in Wang and Tian (2019). The performance of the pag-lasso is illustrated in a Monte Carlo study and, depending on the quality of the prior information, we show that there are substantial gains to be made by taking this information into account for the variable selection. The estimator even copes when the prior information only includes a few of the relevant variables. The simulation study illustrates the sensitivity of the pag-lasso to the weight placed on the prior information relative to the current set of observations. Unsurprisingly, placing more weight on the prior information is beneficial when the prior information is accurate.

The lasso (Tibshirani, 1996) was first introduced as a special case of the more general bridge estimator (Frank and Friedman, 1993). The lasso had some appealing properties in its ability to simultaneously select variables and estimate coefficients. The lasso is, however, consistent only under rather restrictive conditions (Knight and Fu, 2000; Meinshausen and Bühlmann, 2006). Many adaptations have been developed to improve on the properties of the lasso, including the smoothly clipped absolute deviation penalty (SCAD) (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), and the adaptive lasso (Zou, 2006). Subsequent studies have established the statistical properties of the various estimators in different settings. Zou (2006) proves the oracle property for the adaptive lasso with a fixed number of parameters, and Huang, Ma, and Zhang

[1] Although one might be able to link this prior information to a "prior" as understood in a Bayesian setting, the analysis in the present paper is distinctly non-Bayesian.

(2008) expand this to a parameter space that grows with the sample size and they are able to accommodate more variables than observations under some conditions. Nardi and Rinaldo (2008) show selection and estimation consistency for the group lasso in similar low and high-dimensional settings. Wang, You, and Lian (2015) expand convergence and selection consistency to generalized linear models for models with more parameters than observations. For the adaptive group lasso, Zhang and Xiang (2016) are able to show consistency and asymptotic normality when using OLS as the initial estimator, although this restricts the number of parameters to be less than the number of observations. Wei and Huang (2010) allow for more parameters than observations by using the group lasso as the initial estimator and are able to demonstrate selection consistency for the adaptive group lasso. Wang and Tian (2019) show selection and estimation consistency for generalized linear models when the number of groups diverges with the sample size, as well as selection consistency when the number of parameters exceeds the number of observations.

The pag-lasso is applied to the selection of the set of relevant risk factors from the vast set of risk factors and anomalies proposed in the asset pricing literature since the development of the capital asset pricing model (Sharpe, 1964; Lintner, 1965) and the arbitrage pricing theory (Ross, 1976). For a good overview of the so called "Factor Zoo", see Harvey and Liu (2019). There have been numerous attempts at choosing between these factors as an asset pricing model containing several hundred risk factors can be impractical. Ahmed, Bu, and Tsvetanov (2019) compares some of the most prominent and mainstream factors models and find that the four-factor model by Stambaugh and Yuan (2016) outperforms the q-factor model by Hou, Xue, and Zhang (2015), the Fama-French five-factor model (Fama and French, 2015), and the six-factor model by Barillas and Shanken (2018). Bryzgalova (2019) takes a brute force approach and computes all possible combinations of a universe of 50 factors, which is indeed effective in covering all factor models that have been or could have been suggested, but hardly efficient and certainly not feasible considering the actual universe of potential factors easily exceeding 500. Harvey, Liu, and Zhu (2016) propose that researchers use a larger threshold when testing the significance of the abnormal returns derived from new risk factors and anomalies to correct for multiple testing problems, and Feng, Giglio, and Xiu (2020) propose a lasso based procedure that can be used to evaluate new factors that should

control for the explanatory power already present in the existing set of suggested risk factors. One can also consider the factor zoo as a set of noisy approximations of the true set of underlying latent risk factors as done by, e.g., Kozak, Nagel, and Santosh (2020) and Lettau and Pelger (2020a), which use principal component analysis to recover the latent risk factors and vastly shrink the factor space.

This paper deploys the pag-lasso to select a subset of the proposed factors and anomalies, by requiring that the factors not only help explain the test assets but also the remaining tradeable risk factors that are not chosen by the estimator. In particular, we combine three different approaches from the literature on machine learning and asset pricing. First, our approach builds on a lasso-type estimator (cf. Freyberger, Neuhierl, and Weber, 2020; Feng et al., 2020) because of its ability to perform variable selection when faced with many possible explanatory variables. Second, we require that the asset pricing factors chosen by our approach should price not only the set of test portfolios but also the set of factors left out by the estimator (cf. Barillas and Shanken, 2018). Finally, we use the fact that pricing errors should be close to zero when we are using the true set of factors (cf. Lettau and Pelger, 2020b). We demonstrate that the pag-lasso is able to identify a set of relevant factors that price the cross section of returns of both the test assets as well as the remaining factors and anomalies not included in the relevant set. This stands in stark contrast to the aglasso, which is only able to identify a sufficient set of factors in a few of the samples. We are also able to show the evolution of the set of relevant risk factors over time which is stable for many of the included factors indicating some degree of robustness of the pag-lasso procedure. The study is repeated using the mainstream factor models by Carhart (1997), Fama and French (2015), and Hou et al. (2015) in the prior set for the pag-lasso, and we show that they are unable to provide a set of factors sufficient for pricing the test assets as well as the excluded factors.

Section 2 presents the pag-lasso estimator in a generalized linear model framework. Section 3 develops the statistical properties of the pag-lasso. Section 4 provides the Monte Carlo study comparing various formulations of the pag-lasso and the aglasso. Section 5 applies the pag-lasso to the selection of risk factors from the asset pricing literature. Section 6 concludes the paper.

2 The Method

Consider the generalized linear model of Nelder and Wedderburn (1972). We assume that there exists a real matrix of covariates X and a real vector of responses, Y, where the density of Y is assumed to have an exponential form

$$f\left(Y|X\right) = \exp\left(\xi(X) - \phi\left(\xi(X)\right) + \psi(Y)\right),$$

given known functions $\xi(\cdot)$, $\phi(\cdot)$, and $\psi(\cdot)$. The expected value of the response variable given the data is then given by $\mathbb{E}\left[Y|X\right] = \phi'\left(\xi(X)\right)$, where $\phi'\left(\beta\right)$ is the first derivative of $\phi(\cdot)$ and is assumed to exist. In order to parametrise the model we use the link function $g(\cdot)$ to model the link between the expectation, $\mathbb{E}\left[Y|X\right]$, and the linear combination, $X\beta$, such that $g\left(\mathbb{E}\left[Y|X\right]\right) = X\beta$, and thus

$$\mathbb{E}\left[Y|X\right] = \phi'\left(\xi(X)\right) = g^{-1}\left(X\beta\right),$$

where $g(\cdot)$ is assumed to be invertible. We will consider the canonical link where $g^{-1}\left(X\beta\right) = \phi'\left(X\beta\right)$, and, consequently, $\xi(X) = X\beta$. The vector of parameters, β, is split into p_n groups indexed by $j = 1, \ldots, p_n$ and with individual length d_j. The total length of β is then $q_n = \sum_{j=1}^{p_n} d_j$. Consequently, X will be a $N \times q_n$ dimensional matrix, where N is the number of observations, and we denote row $i = 1, \ldots, N$ of X as X_i, referring to the ith observation of X. Similarly Y is a vector of length N, and Y_i is the ith observation of Y. We will continue with the following log-likelihood

$$\ell_n\left(\beta; X, Y\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i X_i^T\beta - \phi\left(X_i^T\beta\right)\right).$$

The adaptive group lasso allows for the selection of variables with a natural grouping. Inspired by the improvements made by the adaptive lasso over the original lasso, the adaptive group lasso utilizes an initial estimator to calculate adaptive weights for the penalization term that differs across the various groups of variables. The adaptive group lasso maximizes the following objective

function with respect to β

$$\begin{aligned} Q_n(\beta; X, Y) &= \ell_n(\beta; X, Y) - \lambda_{n,j} \sum_{j=1}^{p_n} \|\beta_j\|_2 \\ &= \ell_n(\beta; X, Y) - \lambda_n \sum_{j=1}^{p_n} \|\tilde{\beta}_j\|_2^{-1} \|\beta_j\|_2 \end{aligned}$$

where λ_n is a tuning parameter for the penalty term, $\tilde{\beta}$ is the initial estimator used to create adaptive weights, and $\|\cdot\|_2$ is the Euclidean norm. The value of the objective function decreases for non-zero parameter estimates. The tuning parameter λ_n determines the degree of regularisation of the parameters. Small values of λ_n leads to less regularisation, with the limiting case, $\lambda_n = 0$, gives in no regularisation at all and results in the standard OLS estimator. For $\lambda_n \to \infty$ the penalty dominates the likelihood in the objective function and sets all parameters to zero. The optimal value of λ_n is often found through minimisation of some information criterion or cross validation. $\lambda_{n,j} = \lambda_n \|\tilde{\beta}_j\|_2^{-1}$ illustrates the adaptive nature of the penalization of non-zero estimates. The initial estimator is typically chosen as some consistent estimator like the group lasso[2] such that truly non-zero parameters have estimates larger in magnitude than those for irrelevant variables, in general. Hence, we expect that non-zero parameters are penalized less than the irrelevant parameters yielding a more accurate variable selection and less bias in the parameter estimation.

The pag-lasso augments the aglasso by allowing for pre-existing knowledge to guide the variable selection. We assume that the pre-existing knowledge can be summarized in Y^p. The information collected in Y^p can be constructed in various ways, but one possible scenario would be to have some knowledge regarding the value or relevance of the parameters prior to estimation. This can be the case in very high dimensional settings with small sample sizes, as seen in genome-wide association studies (see Jiang et al., 2016). In these studies, it is very costly to generate more observations, and the possibility of utilising findings from previous studies can be of great value.

The intuition is that the estimator is penalized for making predictions that deviate from Y^p, in addition to minimizing the error relative to the observations in Y. The importance placed in Y^p

[2]OLS can be used in low-dimensional settings. Using the lasso and the group lasso as the initial estimator is theoretically equivalent as they are both consistent estimators. However, we find that using the lasso as the initial estimator yields more variables to be included in the second step of the estimation compared to the group lasso.

relative to Y is given by the weight $\eta > 0$. A high value of η indicates a strong belief in the model summarized in Y^p[3]. The objective function of the pag-lasso to be maximised is then given by

$$\begin{aligned} Q_n\left(\beta; X, Y, Y^p\right) &= \ell_n\left(\beta; X, Y\right) + \eta \ell_n\left(\beta; X, Y^p\right) - \lambda_n \sum_{j=1}^{p_n} \left\|\tilde{\beta}_j\right\|_2^{-1} \left\|\beta_j\right\|_2 \\ &= S_n\left(\beta; X, Y, Y^p\right) + \lambda_n \sum_{j=1}^{p_n} \left\|\tilde{\beta}_j\right\|_2^{-1} \left\|\beta_j\right\|_2, \end{aligned} \tag{1}$$

where $S_n = \frac{1}{n}\sum_{i=1}^n \left(Y_i X_i^T \beta - \phi\left(X_i^T\beta\right)\right) + \frac{\eta}{n}\sum_{i=1}^n \left(Y_i^p X_i^T \beta - \phi\left(X_i^T\beta\right)\right)$, $\tilde{\beta}$ is some consistent estimator used to construct the adaptive weights, and Y^p is constructed using some prior information. The tuning parameter η determines the weight placed on the prior information. For $\eta = 0$, no weight is placed on the prior information and the estimator collapses to the adaptive group lasso. For $\eta \to \infty$ only the prior information is taken into account and the "current" data set is completely disregarded. In the simple setting of linear regression, the pag-lasso estimator maximises the objective function

$$\begin{aligned} Q_n\left(\beta; X, Y, Y^p\right) &= \frac{1}{n}\left\|Y - X\beta\right\|_2^2 + \frac{\eta}{n}\left\|Y^p - X\beta\right\|_2^2 + \lambda_n \sum_{j=1}^{p_n} \left\|\tilde{\beta}_j\right\|_2^{-1} \left\|\beta_j\right\|_2 \\ &= S_n\left(\beta; X, Y, Y^p\right) + \lambda_n \sum_{j=1}^{p_n} \left\|\tilde{\beta}_j\right\|_2^{-1} \left\|\beta_j\right\|_2, \end{aligned}$$

where $S_n\left(\beta; X, Y, Y^p\right) = \frac{1}{n}\left\|Y - X\beta\right\|_2^2 + \frac{\eta}{n}\left\|Y^p - X\beta\right\|_2^2$.

The estimator minimizing the objective function in (1) can be written as (cf. Jiang et al. (2016))

$$\hat{\beta} = \arg\max_{\beta} \left(\ell_n\left(\beta; X, \tilde{Y}\right) - \lambda_n \sum_{j=1}^{p_n} \left\|\tilde{\beta}_j\right\|_2^{-1} \left\|\beta_j\right\|_2 \right), \tag{2}$$

with $\tilde{Y} = \left(Y + \eta Y^p\right)/(1+\eta)$, which is of the same form as the adaptive group lasso. Hence the pag-lasso can be solved using the same algorithms as the adaptive group lasso. For some given

[3]If the information summarized in Y^p is derived from a large number of existing studies from credible sources, then that could warrant a larger value of η. In this setting, the new data might be considered to be noisier than some average of the pre-existing studies, and the researcher wishes to extract any additional information from this new sample.

prior information on the values of β denoted by β^p we can calculate Y^p as

$$Y^p = \phi'\left(X\beta^p\right).$$

One way to quantify the prior information is by considering a subset of the available variables which are believed to be relevant with high certainty, denoted by S^p. β^p is then obtained as the group lasso estimate

$$\beta^p = \arg\max_{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i X_i^T\beta - \phi\left(X_i^T\beta\right)\right) - \lambda_n\sum_{j=1}^{p_n} w_j\left\|\beta_j\right\|_2\right),$$

where

$$w_j = \begin{cases} 0, & j \in S^p \\ \sqrt{d_j}, & j \notin S^p. \end{cases}$$

This ensures that all variables in S^p are estimated as non-zero while at the same time allowing additional estimates to be non-zero if the corresponding variables contain sufficient information.

First, let (Y_i, X_i) with $i = 1, \ldots, n$ be iid samples from some population $(\mathcal{Y}, \mathcal{X})$. The use of the pag-lasso assumes sparsity in the set of parameters. Here, sparsity refers to the fact that the parameters for all variables in most of the groups are exactly zero. We denote the true vector of parameters as $\beta_0 = \left(\beta_{01}^T, \ldots, \beta_{0p_n}^T\right)^T$, and define $\beta_0 = \left(\beta_0^{(1)T}, \beta_0^{(2)T}\right)^T$ with $\beta_0^{(1)} = \left(\beta_{0j}^T, j = 1, \ldots, k_n\right)^T$, and $\beta_0^{(2)} = \left(\beta_{0j}^T, j = k_n + 1, \ldots, p_n\right)^T$. Without loss of generality, we assume that $\left\|\beta_{0j}\right\|_2 \neq 0$ for $j = 1, \ldots, k_n$, and $\left\|\beta_{0j}\right\|_2 = 0$ for $j = k_n + 1, \ldots, p_n$. As a result, we also define $X_i = \left(X_i^{(1)T}, X_i^{(2)T}\right)$, for $i = 1, \ldots, n$, and $X = (X_1, \ldots, X_n)^T \triangleq \left(X^{(1)}, X^{(2)}\right)$ with $X^{(1)} = \left(x^1, \ldots, x^{k_n}\right)$ and $X^{(2)} = \left(x^{k_n+1}, \ldots, x^{p_n}\right)$. We denote the total number of parameters and total number of non-zero parameters as $q_n = \sum_{j=1}^{p_n} d_j$ and $q_0 = \sum_{j=1}^{k_n} d_j$, respectively.

Furthermore, we define the parameter space as $\Omega_n \subseteq \mathbb{R}^{q_n}$. For any $\beta \in \Omega_n$, let $\phi\left(X\beta\right) = \left(\phi\left(X_1^T\beta\right), \ldots, \phi\left(X_n^T\beta\right)\right)^T$, $\phi'\left(X\beta\right) = \left(\phi'\left(X_1^T\beta\right), \ldots, \phi'\left(X_n^T\beta\right)\right)^T$, and $\Sigma\left(\beta\right) = \text{diag}\left(\phi''\left(X_1^T\beta\right), \ldots, \phi''\left(X_n^T\beta\right)\right)$. From the generalized linear model it follows that $\mathbb{E}\left[Y\right] = \phi'\left(X\beta_0\right)$ and $\text{cov}\left(Y\right) =$

$\Sigma(\beta_0)$. Define $\Sigma = \frac{1}{n} X^T \Sigma(\beta_0) X$ and $\Sigma_{(1)} = \frac{1}{n} X^{(1)T} \Sigma(\beta_0) X^{(1)}$. Let $\tau_{\min}(A)$ and $\tau_{\max}(A)$ denote the smallest and largest eigenvalues of any given symmetric matrix, A. Let $\theta_1 = \min_{j=1,\ldots,k_n} \left\|\beta_{0j}\right\|_2$. For simplicity we use the constant $M > 0$ in various settings throughout the paper. It is allowed to take different values in different contexts.

3 Theoretical Results

3.1 Generalized Linear Models

This section is an adaptation of the theorems in Wang and Tian (2019) for the aglasso in generalized linear models. The following results are divided into two categories. First, we consider the statistical properties of the pag-lasso in a low dimensional setting where the number of parameters to be estimated is smaller than the number of observations. In the second case, the number of parameters is allowed to be larger than the number of observations.

3.1.1 The case of $q_n < n$

Before presenting the theoretical results, we provide the following assumptions.

(A1) $q_n = O(p_n)$ and $q_0 = O(k_n)$.

(A2) There exists an initial estimator $\tilde{\beta}$, such that $\left\|\tilde{\beta} - \beta_0\right\|_2 = O_P\left((p_n/n)^{1/2}\right)$.

(A3) The eigenvalues of Σ are bounded away from zero and infinity.

(A4) There exists some constant $M > 0$, such that $\max_{1 \le i \le n} \max_{1 \le j \le p_n} \left\|x_{ij}\right\| \le M$.

(A5) There exists a large constant $M > 0$ and a large open subset, $\mathbb{B}_n \subset \Omega_n$ that contains β_0, such that for almost all x_i, we have $\left|\phi^{(3)}\left(x_i^T \beta\right)\right| \le M$.

(A6) There exist constants $0 < 3c_1 < c_2 \le 1$ and $M > 0$, such that $p_n = O(n^{c_1})$ and $n^{(1-c_2)/2}\theta_1 \ge M$.

(A7) There exists some constant $M > 0$, such that $\mathbb{E}\left[\left(\varepsilon_i + \eta\varepsilon_i^p\right)^2\right] \le M$, and for $i, l \in \{1, \ldots, N\}$ we have that $\operatorname{cov}\left[x_i, \varepsilon_l + \eta\varepsilon_l^p\right] = 0$.

Assumptions (A1)-(A6) are identical to those presented in Wang and Tian (2019) and assumption (A7) is required for the following theorems to hold for the pag-lasso. Assumption (A1) gives some bound on the number of variables in each group but still allows for the number of groups and variables in each group to grow with the sample size, n. By (A1), we have that (A2) is consistent with an estimator with a diverging number of parameters in Fan and Peng (2004). Assumptions (A3)-(A5) are regularity conditions also used for the derivations in Fan and Peng (2004). Assumption (A6) allows for the number of groups to diverge to infinity and bounds the size of the non-zero parameters in $\left\{\left\|\beta_{0j}\right\|_2 : j = 1, \ldots, k_n\right\}$ away from zero. It is similar to condition (8) in Zhao and Yu (2006). Assumption (A7) states that the explanatory variables are uncorrelated with the errors across observations and that the variance of the errors is finite. The following theorems exhibit properties similar to those in Wang and Tian (2019), and assumption (A7) is the critical addition used to arrive at these results.[4] Proofs can be found in the appendix.

Theorem 3.1. *Under conditions (A1) - (A7), we have*

$$\left\|\hat{\beta} - \beta_0\right\|_2 = O_p\left((p_n/n)^{1/2}\right),$$

if $\lambda_n n^{(2-c_2+c_1)/2} \to 0$.

This theorem shows the convergence rate of the pag-lasso. The rate is identical to that found in Theorem 3.2 of Portnoy (1984) and Theorem 1 of Fan and Peng (2004) and is conditional on the convergence rate of the initial estimator, $\tilde{\beta}$.

Theorem 3.2. *Let* $\hat{\beta}_* = \left(\hat{\beta}_*^{(1)T}, 0^T\right)^T$, *and define*

$$\hat{\beta}_*^{(1)} = \arg\max_{\beta}\left\{S_n(\beta) - \lambda \sum_{j=1}^{k_n} \left\|\tilde{\beta}_j\right\|_2^{-1} \left|\beta_j\right|_2, \left\|\beta_j\right\| = 0 \text{ for } j = k_n + 1, \ldots, p_n\right\}.$$

Suppose that conditions (A1) - (A7) hold. If $\lambda_n n^{1-c_1} \to \infty$, *then with probability tending to one,* $\hat{\beta}_*$ *is the solution of* (2).

[4]Failing to specify the prior information correctly will make it less likely for assumption (A7) to hold.

This theorem shows that if the penalty parameter, λ_n, increases sufficiently quickly with n, then the pag-lasso estimator is able to set the coefficients relating to irrelevant variables to zero with probability tending to one for large n.

Theorem 3.3. *(Oracle property) Suppose that* $\lambda_n n^{(2-c_2+c_1)/2} \to 0$, $\lambda_n n^{1-c_1} \to \infty$, *and furthermore that* $\mathbb{E}\left[Y_1 - \phi'\left(X_1^T \beta_0\right)\right]^4 < \infty$. *Under conditions (A1) - (A7), the pag-lasso estimator* $\hat{\beta} = \left(\hat{\beta}^{(1)T}, 0^T\right)^T$ *satisfies*

1. *Sparsity:* $\mathbb{P}\left(\left\{j : \left\|\hat{\beta}_j\right\|_2 \neq 0\right\} = \{1, \ldots, k_n\}\right) \to 1$.

2. *Asymptotic normality:*

$$n^{1/2} \alpha_n^T \Sigma_{(1)}^{1/2} \left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) \xrightarrow{d} \mathcal{N}(0, 1),$$

where α_n *is a* $\left(\sum_{j=1}^{k_n} d_j\right)$*-dimensional unit vector.*

This theorem shows the oracle property for the pag-lasso estimator. Required that the penalty parameter is large enough to set the coefficients for irrelevant variables equal to zero (as ensured by $\lambda_n n^{1-c_1} \to \infty$ and Theorem 3.2) and small enough to keep coefficients of relevant parameters as non-zero ($\lambda_n n^{(2-c_2+c_1)/2} \to 0$), the pag-lasso is able to recover the true sparsity pattern and yield root-n consistent estimates of the non-zero coefficients.

3.1.2 The case of $q_n > n$

We will now present the theoretical properties of the pag-lasso in the high-dimensional setting where $p_n > n$. The following definition from Wei and Huang (2010) is used to prove selection consistency.

Definition 1. An estimator $\tilde{\beta}$ is consistent at zero with rate r_n if

$$r_n \max_{j=k_n+1,\ldots,p_n} \left\|\tilde{\beta}\right\|_2 = O_p(1),$$

where $r_n \to \infty$ as $n \to \infty$, and there exists a constant $\xi_0 > 0$ such that for any $\epsilon > 0$,

$$P\left(\min_{j=1,\ldots,k_n} \left\|\tilde{\beta}\right\|_2 > \xi_0 \theta_1\right) > 1 - \epsilon,$$

for n sufficiently large.

For the following results we make the assumptions:

(B1) The eigenvalues of $\Sigma_{(1)}$ are bounded away from zero and infinity.

(B2) Conditions (A1), (A4), (A5), and (A7) hold.

(B3) The initial estimator $\tilde{\beta}$ is consistent at zero with rate $r_n \to \infty$.

(B4) There exist constants $0 < 3c_3 < c_4 \leq 1$ and $M > 0$, such that $k_n = O\left(n^{c_3}\right)$ and such that $n^{(1-c_4)/2}\theta_1 \geq M$.

(B5) There exist constants $M > 0$ and $R > 0$, such that $\mathbb{E}\left[\left|\varepsilon_i + \eta\varepsilon_i^p\right|^r\right] \leq \frac{1}{2} r! M^{r-2} R$ for any $r \geq 2$.

(B6)

$$\frac{k_n}{n r_n^2 \lambda_n^2} \to 0, \qquad \frac{\log\left(p_n\right)}{n r_n^2 \lambda_n^2} \to 0.$$

Assumptions (B1)-(B5) are identical to those presented in Wang and Tian (2019) and assumption (B6) is required for the following theorems to hold for the pag-lasso. Assumption (B1) ensures that $\Sigma_{(1)}$ is positive definite, which is reasonable as long as the number of relevant variables, q_0, is much smaller than the number of observations, n. Assumption (B3) requires a zero-consistent estimator as the initial estimator. By theorem 3 in Wang et al. (2015), the initial estimator can be chosen as the group lasso estimator, which is consistent at zero with rate $\left(n/\left(k_n \log\left(p_n\right)\right)\right)^{1/2}$. Assumption (B5) bounds the moments of the noise and requires the tail distribution of the noise to have an exponential decay. Assumption (B6) restricts the relation between the number of groups and the penalty parameter. When using the group lasso as the initial estimator (B6) can be written as

(B6)'

$$\frac{k_n^2 \log(p_n)}{n^2 \lambda_n^2} \to 0, \qquad \frac{\log(p_n)^2 k_n}{n^2 \lambda_n^2}.$$

Choosing the penalty parameter $\lambda_n = n^{(c_4 - c_3 - 1)/2 - \delta}$ with any small $\delta > 0$, we can have as many as $\exp\left(n^{-\delta + c_4/2 - c_3 + 1/2}\right)$ groups. The proofs for the following theorems can be found in the appendix.

Theorem 3.4. *$\hat{\beta}_* = \left(\hat{\beta}_*^{(1)T}, 0^T\right)^T$ is defined as in Theorem 3.2. Suppose that conditions (B1) - (B4) hold. If $\lambda_n n^{(1 - c_4 + c_3)/2} \to 0$, then*

$$\left\|\hat{\beta}_* - \beta_0\right\|_2 = O_p\left((k_n/n)^{1/2}\right).$$

This theorem shows the convergence rate for the pag-lasso. The proof is similar to that of Theorem 3.1 and is omitted.

Theorem 3.5. *Under conditions (B1) - (B6). If $\lambda_n n^{(1 - c_4 + c_3)/2} \to 0$, then the pag-lasso estimator $\hat{\beta}_* = \left(\hat{\beta}_*^{(1)T}, 0^T\right)^T$ is the solution of (2) with probability tending to one.*

This theorem shows that the pag-lasso sets the coefficients of irrelevant parameters equal to zero.

Theorem 3.6. *Under condition (B4), we have*

$$\mathbb{P}\left(\left\{j : \left\|\hat{\beta}_j\right\|_2 \neq 0\right\} = \{1, \ldots, k_n\}\right) \to 1.$$

This theorem ensures that truly non-zero parameters are also estimated as non-zero parameters by the pag-lasso. Together with Theorem 3.5, this shows that the pag-lasso is able to recover the true sparsity pattern. This is as close to the oracle property as one can get in a high-dimensional setting where it is impossible to arrive at asymptotic normality for the non-zero estimates.

4 Simulation Study

This section compares the pag-lasso with the aglasso (Wang and Leng, 2008). For both estimators, we use the lasso and the group lasso as initial estimators to calculate the adaptive weights. We compare the models across several dimensions.

nVAR The total number of non-zero coefficients in the estimation.

CNZ The number of coefficients that are correctly estimated to be non-zero.

INZ The number of coefficients that are incorrectly estimated to be non-zero.

Contains The share of simulations where the correct model is contained in the estimated model.

Sparsity The share of simulations where all the relevant coefficients are estimated as non-zero and all irrelevant variables are excluded.

Bias Calculated as the L1 norm bias for the non-zero coefficients.

ME The model error is calculated as

$$ME = \frac{1}{N}\left(\hat{\beta} - \beta_0\right)^T X^T X \left(\hat{\beta} - \beta_0\right)$$

MSE The mean squared error quantifies the in-sample fit of the model. For given estimates $\hat{\beta}$ it is calculated as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(X\hat{\beta} - Y\right)^2$$

We consider six different examples consisting of three formulations of β_0 and two different types of simulated explanatory variables. The three different formulations of β_0 are

$$\beta_{01} = \left(-0.5, -2, 0.5, 2, -1.5, 1, 2, -1.5, 2, -2, 1, 1.5, -2, 1, 1.5, \mathbf{0}^T, \ldots, \mathbf{0}^T\right)^T \tag{3}$$

$$\beta_{02} = \left(-0.5, -2, 0.5, \mathbf{0}^T, 2, -1.5, 1, \mathbf{0}^T, 2, -1.5, 2, \mathbf{0}^T, -2, 1, 1.5, \mathbf{0}^T, -2, 1, 1.5, \mathbf{0}^T, \ldots, \mathbf{0}^T\right)^T \tag{4}$$

$$\beta_{03} = \left(-0.5, -2, 0.5, 2, -1.5, 1, \mathbf{0}^T, \ldots, \mathbf{0}^T, 2, -1.5, 2, -2, 1, 1.5, -2, 1, 1.5\right)^T, \tag{5}$$

where $\mathbf{0}$ is a three-dimensional vector of zeros.

The simulated data is created in two ways, where we consider the case with p groups and 3 variables in each group. First, we generate p latent variables, $Z_1, \ldots, Z_p$, from a multivariate normal distribution with zero mean and with the covariance between Z_i and Z_j being $0.5^{|i-j|}$. In the first setting, we trichotomise the data such that for each Z_j we generate three interrelated variables

$$\begin{aligned}
X_{j1i} &= \begin{cases} 1, & \text{if } Z_{ji} < \Phi^{-1}(1/3), \\ 0, & \text{if } Z_{ji} \geq \Phi^{-1}(1/3), \end{cases} \\
X_{j2i} &= \begin{cases} 1, & \text{if } \Phi^{-1}(1/3) \leq Z_{ji} < \Phi^{-1}(2/3), \\ 0, & \text{if } Z_{ji} < \Phi^{-1}(1/3) \text{ or } Z_{ji} \geq \Phi^{-1}(2/3), \end{cases} \\
X_{j3i} &= \begin{cases} 1, & \text{if } Z_{ji} \geq \Phi^{-1}(2/3), \\ 0, & \text{if } Z_{ji} < \Phi^{-1}(2/3). \end{cases}
\end{aligned} \tag{6}$$

Thus yielding p different groups with 3 related variables. Alternatively, we generate the three variables as

$$\begin{aligned}
X_{j1i} &= Z_{ji} \\
X_{j2i} &= Z_{ji}^2 \\
X_{j3i} &= Z_{ji}^3.
\end{aligned} \tag{7}$$

For given β_0 and X, we generate $Y = X\beta_0 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$. This yields six different examples split into (1) Y generated using trichotomised data and β_{01}, (2) Y generated using trichotomised data and β_{02}, (3) Y generated using trichotomised data and β_{03}, (4) Y generated using non-trichotomised data and β_{01}, (5) Y generated using non-trichotomised data and β_{02}, and (6) Y generated using non-trichotomised data and β_{03}. The variables are not standardised prior

to estimation. While standardisation prior to estimation can be useful it is mostly of relevance when the explanatory variables vary significantly in magnitude. The magnitude of the variables generated in (6) is identical, and we don't consider it to be a significant problem for the variables in (7) either. The number of groups is set to $p = 100$, the standard deviation of ϵ is $\sigma_\epsilon = 2$, the number of observations is $N = 100$, the number of relevant groups is $k = 5$, and the number of variables in each group is $d = 3$. To give an idea of the sensitivity to the weight placed on the prior information through η we set $\eta = \{1, 10\}$.

We estimate the pag-lasso in a number of different specifications and compare it to the aglasso. Both are estimated using the lasso and the group lasso as initial estimators. The tuning parameter λ is chosen using ten-fold cross-validation. The different specifications of the pag-lasso cover the various degrees to which the prior contains the true set of relevant variables. They are divided into three categories with a total of six groups.

G1: The prior only includes relevant groups.

- **S1:** The prior includes all five relevant groups.
- **S2:** The prior includes three of the five relevant groups.

G2: The prior includes a mix of relevant and irrelevant groups.

- **S3:** The prior includes all five relevant groups as well as two irrelevant groups.
- **S4:** The prior includes three of the five relevant groups as well as ten irrelevant groups.

G3: The prior only includes irrelevant groups.

- **S5:** The prior includes two irrelevant groups.
- **S6:** The prior includes ten irrelevant groups.

Tables 1, 2 , and 3 show the performance of the pag-lasso using the variable groups presented above and $\eta = 1$ compared to the aglasso using the data obtained from examples 1, 2, and 3. Tables 13, 15, and 17 depict the simulations using the same data generating process but with $\eta = 10$ and are referred to in the appendix.

Table 1: Simulations for example 1 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	11.376	14.041	12.074	14.042	12.417	11.571	12.927
	(0.285)	(0.163)	(0.257)	(0.164)	(0.284)	(0.308)	(0.395)
CNZ	11.368	14.041	12.072	14.041	12.029	11.153	11.161
	(0.284)	(0.163)	(0.257)	(0.164)	(0.258)	(0.289)	(0.288)
INZ	0.009	0.000	0.002	0.001	0.388	0.418	1.767
	(0.016)	(0.000)	(0.008)	(0.004)	(0.110)	(0.110)	(0.247)
Contains	0.250	0.716	0.327	0.718	0.320	0.224	0.225
	(0.043)	(0.045)	(0.047)	(0.045)	(0.047)	(0.042)	(0.042)
Sparsity	0.250	0.716	0.327	0.718	0.280	0.196	0.122
	(0.043)	(0.045)	(0.047)	(0.045)	(0.045)	(0.040)	(0.033)
Bias	14.479	4.261	7.863	4.259	7.948	8.807	8.767
	(0.154)	(0.256)	(0.401)	(0.256)	(0.406)	(0.406)	(0.401)
MSE	26.225	4.926	10.552	4.923	10.591	12.645	12.190
	(9.144)	(3.226)	(6.354)	(3.221)	(6.539)	(6.965)	(6.662)
ME	22.842	2.150	7.690	2.149	7.778	9.757	9.465
	(0.914)	(0.322)	(0.633)	(0.322)	(0.649)	(0.689)	(0.659)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{01} from (3) and X from (7). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Tables 4, 5, and 6 show the performance of the pag-lasso with $\eta = 1$ compared to the aglasso using the data obtained from examples 1, 2, 3 with trichotomised X. Tables 19, 21, and 23 show the simulations using the trichotomised data for the data generating process with $\eta = 10$ and are deferred to the appendix.

The first column of the tables contains the results from the adaptive group lasso estimates, and the remaining six columns are specific to the pag-lasso. The six columns refer to the six different specifications of the prior set of relevant variables presented above. There are some general tendencies from the various simulations which indicate that there are indeed gains to be

Table 2: Simulations for example 2 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	11.802	14.208	12.327	14.196	13.275	13.368	15.573
	(0.276)	(0.147)	(0.254)	(0.147)	(0.312)	(0.345)	(0.446)
CNZ	11.772	14.208	12.321	14.190	12.219	11.466	11.403
	(0.274)	(0.147)	(0.253)	(0.147)	(0.251)	(0.280)	(0.283)
INZ	0.030	0.000	0.006	0.006	1.056	1.902	4.170
	(0.030)	(0.000)	(0.013)	(0.013)	(0.170)	(0.201)	(0.351)
Contains	0.302	0.758	0.367	0.751	0.344	0.257	0.255
	(0.046)	(0.043)	(0.048)	(0.043)	(0.048)	(0.044)	(0.044)
Sparsity	0.298	0.758	0.366	0.750	0.225	0.125	0.065
	(0.046)	(0.043)	(0.048)	(0.043)	(0.042)	(0.033)	(0.025)
Bias	6.486	2.305	2.994	2.325	3.040	4.900	4.802
	(0.105)	(0.149)	(0.162)	(0.150)	(0.163)	(0.208)	(0.206)
MSE	24.564	4.667	10.255	4.696	10.118	11.633	10.922
	(8.595)	(2.688)	(6.424)	(2.687)	(6.327)	(6.665)	(6.073)
ME	21.265	1.904	7.426	1.934	7.407	8.941	8.472
	(0.865)	(0.271)	(0.644)	(0.271)	(0.634)	(0.665)	(0.606)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{02} from (4) and X from (7). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

made when utilising additional information if that information resembles the true data generating process sufficiently close.

Generally, using the lasso as the initial estimator results in more variables being included by the final estimator. There are 15 relevant variables (five relevant groups and three variables in each group), and the aglasso tends to select around 20 variables with the trichotomised data when using the group lasso as the initial estimator. This jumps to 25 variables when using the lasso as the initial estimator. The pag-lasso also experiences a jump in the number of included variables, but it is not as large as seen with the aglasso. Using the group lasso as the initial

Table 3: Simulations for example 3 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	12.135	14.250	12.693	14.256	13.014	12.618	14.430
	(0.274)	(0.146)	(0.244)	(0.144)	(0.273)	(0.313)	(0.426)
CNZ	12.117	14.250	12.687	14.253	12.669	11.853	11.862
	(0.272)	(0.146)	(0.244)	(0.144)	(0.243)	(0.284)	(0.281)
INZ	0.018	0.000	0.006	0.003	0.345	0.765	2.568
	(0.023)	(0.000)	(0.013)	(0.009)	(0.111)	(0.141)	(0.295)
Contains	0.362	0.772	0.437	0.771	0.431	0.332	0.326
	(0.048)	(0.042)	(0.050)	(0.042)	(0.050)	(0.047)	(0.047)
Sparsity	0.359	0.772	0.436	0.770	0.379	0.252	0.131
	(0.048)	(0.042)	(0.050)	(0.042)	(0.049)	(0.043)	(0.034)
Bias	4.801	1.700	1.990	1.693	2.010	3.321	3.431
	(0.091)	(0.132)	(0.132)	(0.133)	(0.130)	(0.171)	(0.171)
MSE	22.702	4.531	8.482	4.517	8.561	10.390	9.933
	(7.532)	(2.715)	(5.412)	(2.575)	(5.459)	(5.858)	(5.286)
ME	19.386	1.752	5.631	1.731	5.758	7.578	7.336
	(0.750)	(0.270)	(0.536)	(0.256)	(0.535)	(0.581)	(0.523)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{03} from (5) and X from (7). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

estimator improves the performance of the estimator because the additional variables included by the lasso are predominantly irrelevant variables. This property is not as pronounced for the non-trichotomised data, however.

Turning to the *Contains* metric, both estimators include all, or mostly all, of the relevant variables with the trichotomised data, thus limiting the possible gains from including more information with the pag-lasso. There are, however, gains from incorporating relevant information, as can be seen from the slight increase in the share of simulations that contains all the relevant variables, which increases from around 80% for the adaptive group lasso to upwards of 99% for

Table 4: Simulations for example 1 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	19.403	15.781	16.375	17.322	27.481	20.797	33.077
	(0.587)	(0.207)	(0.291)	(0.282)	(0.629)	(0.529)	(0.722)
CNZ	14.424	14.972	14.890	14.968	14.848	14.299	14.207
	(0.122)	(0.029)	(0.057)	(0.031)	(0.067)	(0.133)	(0.141)
INZ	4.979	0.808	1.485	2.354	12.632	6.498	18.871
	(0.564)	(0.205)	(0.283)	(0.280)	(0.624)	(0.504)	(0.702)
Contains	0.814	0.991	0.964	0.989	0.951	0.775	0.748
	(0.039)	(0.010)	(0.019)	(0.010)	(0.022)	(0.042)	(0.043)
Sparsity	0.252	0.809	0.664	0.454	0.017	0.088	0.002
	(0.043)	(0.039)	(0.047)	(0.050)	(0.013)	(0.028)	(0.004)
Bias	7.568	7.308	7.469	7.506	8.368	8.180	8.959
	(0.198)	(0.188)	(0.212)	(0.191)	(0.228)	(0.228)	(0.243)
MSE	3.419	3.262	3.191	3.147	2.544	2.890	2.316
	(0.636)	(0.557)	(0.616)	(0.549)	(0.559)	(0.677)	(0.598)
ME	0.901	0.734	0.844	0.851	1.493	1.310	1.866
	(0.044)	(0.042)	(0.054)	(0.044)	(0.060)	(0.065)	(0.065)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{01} from (3) and X from (6). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

the pag-lasso. These gains can be found in the first four specifications of the prior set of relevant variables ($S1, S2, S3$, and $S4$). While $S1$ is the only specification that is completely correct, they are all able to increase the share of simulations that select at least all the relevant variables. If the prior information is completely wrong and does not contain any of the relevant variables, we observe a reduction in the share of simulations that contain the entire set of relevant variables. We see a similar pattern for the non-trichotomised data. However, the two settings with complete misspecification of the prior set of relevant variables, the performance is still close to that of the aglasso.

Table 5: Simulations for example 2 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	21.324	16.062	16.890	17.964	29.688	22.530	35.523
	(0.697)	(0.253)	(0.353)	(0.350)	(0.674)	(0.633)	(0.755)
CNZ	14.517	14.994	14.904	14.988	14.841	14.385	14.232
	(0.115)	(0.013)	(0.055)	(0.019)	(0.070)	(0.129)	(0.143)
INZ	6.807	1.068	1.986	2.976	14.847	8.145	21.291
	(0.682)	(0.252)	(0.347)	(0.349)	(0.668)	(0.612)	(0.739)
Contains	0.845	0.998	0.969	0.996	0.949	0.806	0.761
	(0.036)	(0.004)	(0.017)	(0.006)	(0.022)	(0.040)	(0.043)
Sparsity	0.185	0.769	0.599	0.385	0.012	0.075	0.000
	(0.039)	(0.042)	(0.049)	(0.049)	(0.011)	(0.026)	(0.000)
Bias	4.684	4.249	4.328	5.833	6.808	6.903	7.778
	(0.155)	(0.142)	(0.150)	(0.226)	(0.246)	(0.242)	(0.270)
MSE	3.287	3.230	3.141	3.092	2.447	2.747	2.213
	(0.663)	(0.587)	(0.653)	(0.577)	(0.592)	(0.699)	(0.614)
ME	0.920	0.773	0.903	0.912	1.600	1.431	1.970
	(0.048)	(0.046)	(0.060)	(0.048)	(0.062)	(0.071)	(0.068)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{02} from (4) and X from (6). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

The share of simulations that correctly identifies the sparsity pattern in the coefficients must by definition be lower than the share of simulations in which simply the non-zero coefficients are detected. The difference in the two metrics, *Contains* and *Sparsity*, is found in the estimator's ability to exclude irrelevant variables. The aglasso includes around five irrelevant variables across the different trichotomised examples, and the pag-lasso includes anything from less than one to more than 20 irrelevant variables. This is driven by the accuracy of the prior information. Unsurprisingly, the number of irrelevant selected variables increases with the number of irrelevant variables included in the prior set. However, the number of irrelevant variables that are selected by the pag-lasso also seems to increase when fewer relevant variables are included, as can be seen in

Table 6: Simulations for example 3 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	19.554	15.777	16.512	17.412	27.609	21.024	33.318
	(0.588)	(0.203)	(0.293)	(0.293)	(0.620)	(0.529)	(0.723)
CNZ	14.445	14.976	14.928	14.976	14.910	14.358	14.262
	(0.120)	(0.027)	(0.046)	(0.027)	(0.051)	(0.129)	(0.134)
INZ	5.109	0.801	1.584	2.436	12.699	6.666	19.056
	(0.567)	(0.201)	(0.292)	(0.292)	(0.617)	(0.506)	(0.709)
Contains	0.819	0.992	0.976	0.992	0.970	0.793	0.761
	(0.039)	(0.009)	(0.015)	(0.009)	(0.017)	(0.041)	(0.043)
Sparsity	0.254	0.807	0.674	0.461	0.018	0.088	0.001
	(0.044)	(0.039)	(0.047)	(0.050)	(0.013)	(0.028)	(0.003)
Bias	3.425	2.938	3.013	4.250	5.817	5.243	6.975
	(0.118)	(0.114)	(0.120)	(0.199)	(0.260)	(0.219)	(0.279)
MSE	3.404	3.267	3.171	3.142	2.522	2.865	2.300
	(0.633)	(0.594)	(0.628)	(0.561)	(0.577)	(0.685)	(0.603)
ME	0.904	0.739	0.856	0.864	1.507	1.325	1.873
	(0.044)	(0.045)	(0.055)	(0.046)	(0.061)	(0.069)	(0.068)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{03} from (5) and X from (6). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

the difference between groups $S1$ and $S2$. The aglasso excludes almost all of the irrelevant variables with the non-trichotomised data, thus making the potential gains from using the pag-lasso very small. On the other hand, the number of irrelevant variables that are selected with prior sets $S5$ and $S6$ is also very small compared to what is seen with the trichotomised data, thus limiting the reduction in performance with poorly specified prior sets as well.

Looking at model fit and estimation accuracy, as summarised by the metrics *Bias*, *MSE*, and *ME*, we see that for the trichotomised data, the pag-lasso only leads to minor gains when correctly specified prior sets. However, the reduction in performance is also limited when the prior set is misspecified. The non-trichotomised paint a different picture with vast improvements in *Bias*,

MSE, and *ME* for the pag-lasso. Across all the metrics studied in the simulations, we find that increasing the weight given to the prior information, η, leads to more extreme findings. If the pag-lasso outperforms the aglasso, then increasing η will lead to even stronger outperformance, and, conversely, if the pag-lasso underperforms the aglasso, decreasing η leads to even stronger underperformance.

5 Empirical Study

We apply the pag-lasso to a large set of pricing factors from the asset pricing literature. Since the development of the capital asset pricing model (Sharpe, 1964; Lintner, 1965) and the arbitrage pricing theory (Ross, 1976), the asset pricing literature has contributed with the discovery of a large number of factors that are suggested to be useful in explaining the cross section of returns of financial assets. As the number of proposed factors has continued to increase, the asset pricing literature has sought to sort and select among the candidates in the "Factor Zoo". Harvey et al. (2016) suggest that a larger critical value should be used when testing for new factors in order to adjust for multiple testing. Pukthuanthong, Roll, and Subrahmanyam (2019) set up criteria that must be fulfilled before a factor can be categorised as "genuine", Bryzgalova (2019) uses a brute force approach and estimates all possible combinations of 51 factors to identify the best model in a Bayesian setup[5]. Feng et al. (2020) present a double lasso approach to test for new factors while correctly controlling for previously identified factors, Ahmed et al. (2019) compare a few of the most prominent factor models[6], and Kozak et al. (2020) estimate a low dimensional set of latent factors using principal components.

[5]This approach quickly becomes unfeasible due to the combinatorial increase in computation time. The data used in the present study contains 150 risk factors, which is already much larger than the one used in Bryzgalova (2019). Using this method to sort through the more than 500 factors summarized by Harvey and Liu (2019) will be completely infeasible using modern technology.

[6]We present the performance of the factor models by Carhart (1997), Fama and French (2015), and Hou et al. (2015) in section 5.4 and compare them to our implementation of the prior adaptive group lasso.

5.1 Data

We use the collection of factors and test portfolios that are used in Feng et al. (2020)[7]. The data set contains 150 factors from the asset pricing literature. While this does not come close to covering the entire factor zoo, it is among the largest collections of factors that have been assembled to date. The set of factors consists of all the U.S. equity market factors from Kenneth French's data library, the liquidity factor from Pastor and Stambaugh (2003), the q-factors from Hou et al. (2015), the intermediary asset pricing factors from He, Kelly, and Manela (2017), factors from the AQR data library, and 135 long-short value-weighted proxy portfolios based on characteristics described in Hou, Xue, and Zhang (2020) and Green, Hand, and Zhang (2017). See Feng et al. (2020) for more details on the construction of the factors. A brief overview of the 150 factors can be found in table 25. The factors are observed at a monthly frequency from July 1976 to December 2017.

We use the 202 portfolios used in Giglio and Xiu (2019), which consist of 25 portfolios constructed using a two-way sort by size and book-to-market ratio, 25 portfolios two-way sorted by operating profitability and investment, 25 portfolios two-way sorted by size and variance, 35 portfolios two-way sorted by size and net issuance, 25 portfolios two-way sorted by size and accruals, 25 portfolios two-way sorted by size and momentum, 25 portfolios two-way sorted by size and beta, and 17 industry portfolios.

5.2 Setting up the model

We divide the sample into rolling windows of five years of length. This should provide enough data points to identify a limited number of relevant factors while being short enough to keep most of the fundamental dynamics constant. Then we estimate the relevant set of factors for each window such that we can describe the evolution of the set over time by comparing the different windows.

From the arbitrage pricing theory, we have that given traded factors, f_t, the asset excess returns, r_{it} follow a linear factor model

$$r_{it} = 0 + f_t^T \beta_i, \qquad i = 1, \dots, N,\ t = 1, \dots, T,$$

[7]We thank the authors for kindly providing the data.

where f_t and β_i are $K \times 1$ vectors with factor excess returns and factor loadings, respectively. For a given set of factors and test assets, this relation can be evaluated by testing $\alpha_i = 0$ in the regression

$$r_{it} = \alpha_i + f_t^T \beta_i + \epsilon_{it}, \tag{8}$$

for fixed i, where ϵ_{it} is an iid error term. If $\alpha_i = 0$, then the factors in f_t are said to "price" r_{it}. Like Barillas and Shanken (2018), we want to require the selected factors to be able to price unselected factors in addition to the test assets. The selected factors, of course, trivially price themselves, but we will also require the relation in (8) to hold with each of the K factors acting as test assets.

This system can be stacked into a single equation, similarly to what was done in Hwang and Rubesam (2020). Let $r_i = \left(r_{i1}^T, \ldots, r_{iT}^T\right)^T$ be the $T \times 1$ vector of test asset excess returns for each $i = 1, \ldots, N$. Let $f_j = \left(f_{j1}^T, \ldots, f_{jT}^T\right)^T$ be the $T \times 1$ vector of factor excess returns for each $j = 1, \ldots, K$. Then collect all r_i and f_j such that $r = \left(r_1^T, \ldots, r_N^T, f_1^T, \ldots, f_K^T\right)^T$ is the $(N + K)\, T \times 1$ vector of excess returns for all assets that are required to be priced by the selected factors. Collect the traded factors in the $T \times K$ matrix $f = (f_1, \ldots, f_K)$ and construct the $(N + K)\, T \times (N + K)\, K$ block diagonal matrix of factor excess returns $F = I_{N+K} \otimes f$, where I_{N+K} is the $N + K$ dimensional identity matrix. Then the test regressions in (8) for all test assets and factors can be written as

$$r = \alpha \otimes \iota_T + F\beta + \epsilon,$$

where α is the $N + K \times 1$ vector of intercepts, $\beta = \left(\beta_1^T, \ldots, \beta_{N+K}^T\right)^T$ is the $(N + K)\, K \times 1$ vector of factor loadings, $\epsilon = \left(\epsilon_1^T, \ldots, \epsilon_{N+K}^T\right)^T$ is the $(N + K)\, T \times 1$ vector of residuals, and ι_T is a $T \times 1$ vector of ones. Of course this can also be cast as a standard linear regression by defining $B = \left(\alpha^T, \beta^T\right)^T$ and $X = (\iota_T \otimes I_{N+K}, F)$ such that

$$r = XB + \epsilon.$$

The procedure can also handle non-traded factors by defining g_{lt} as the return of non-traded

factor $l = 1, \ldots, L$ at time t. In a fashion similar to the construction of F above, we can then construct G as the $(N + K)\,T \times (N + K)\,L$ matrix containing the non-traded factors. Since they are non-traded, they will not be included on the left-hand side in r. Then, the combination of F and G will price the set of test assets and traded factors in the regression

$$r = \alpha \otimes \iota_T + F\beta + G\gamma + \epsilon,$$

The factors in F can be divided such that all columns relating to the same factor are grouped together. This allows the group lasso to require that a factor is either included or excluded for all assets. This prevents factor j from being selected as good at pricing asset i but not asset l. If a factor is deemed to be relevant it must be included for all assets, which, in turn, increases the penalty in the lasso objective function. The prior adaptive group lasso then selects the set of factors that best balance the trade-off between pricing the test assets while keeping the set of relevant factors small. To illustrate the grouping structure consider

$$F = \begin{pmatrix} f & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & f & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & f \end{pmatrix} = \begin{pmatrix} f_1 & \cdots & f_K & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & f_1 & \cdots & f_K & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & f_1 & \cdots & f_K \end{pmatrix},$$

Such that the red variables in F are grouped together and the green variables are grouped together. Similarly, the grouping can be illustrated in the vector factor loadings

$$\beta = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1K} & \cdots & \beta_{i1} & \cdots & \beta_{iK} & \cdots & \beta_{N1} & \cdots & \beta_{NK} \end{pmatrix}^T.$$

Let β_j be the subvector of β containing the elements relating to factor j.

For a given prior r^p, the prior adaptive group lasso estimates, $\hat{\alpha}$, and $\hat{\beta}$, are obtained as

$$\begin{aligned}\left[\hat{\alpha}, \hat{\beta}\right] =& \underset{\alpha,\beta}{\arg\min} \left(\frac{1}{T} \left\| r - \alpha \otimes \iota_T - F\beta \right\|_2^2 + \frac{\eta}{T} \left\| r^p - \alpha \otimes \iota_T - F\beta \right\|_2^2 + \lambda_n \sum_{j=1}^{K} w_j \left\| \beta_j \right\|_2 \right) \\ =& \underset{\alpha,\beta}{\arg\min} \left(\frac{1}{T} \left\| \tilde{r} - \alpha \otimes \iota_T - F\beta \right\|_2^2 + \frac{\lambda_n}{1+\eta} \sum_{j=1}^{K} w_j \left\| \beta_j \right\|_2 \right),\end{aligned}$$

where $\tilde{r} = \left(r + r^p\right)/(1+\eta)$. The estimates are obtained using ten-fold cross validation.

There are several ways to construct the prior information summarized in r^P. We will take advantage of the fact that when all relevant factors are included, the factors will price the test assets as well as the remaining factors. Using this property of the asset pricing models as a guiding principle for the estimation has also been utilised in Barillas and Shanken (2018) and Lettau and Pelger (2020a). Hence, we estimate the following group lasso that restricts the intercepts to be zero.

$$\left[\beta^P\right] = \underset{\beta}{\arg\min} \left(\frac{1}{T} \left\| r - F\beta \right\|_2^2 + \lambda_n \sum_{j=1}^{K} \left\| \beta_j \right\|_2 \right),$$

from which the prior information can be computed as $r^p = X\beta^P$.

We estimate the set of relevant risk factors using the pag-lasso and aglasso. We then evaluate these sets of potentially relevant factors using the test by Gibbons, Ross, and Shanken (1989), in the following referred to as GRS. Let the set of factors for evaluation be given by $\mathcal{S} = \left\{ j : \left\| \hat{\beta}_j \right\|_2 \neq 0 \right\}$ with cardinality S. Let $F_{\mathcal{S}}$ and $\beta_{\mathcal{S}}$ be the subsets of F and β, respectively, only containing the elements that are relevant to the factors indexed by S, and let $r_{-\mathcal{S}}$ and $\alpha_{-\mathcal{S}}$ be the subsets of r and α not related to the factors contained in $\mathcal{S}$. Then the model is re-estimated with OLS

$$r_{-\mathcal{S}} = \alpha_{-\mathcal{S}} \otimes \iota_T + F_S \beta_S + \epsilon.$$

The GRS test evaluates the null hypothesis that all the intercepts are equal to zero against the alternative that at least one of the intercepts are different from zero. The test statistic is computed

as

$$\xi = \frac{\hat{\alpha}^T \hat{\Sigma}_\alpha \hat{\alpha}}{1 + \hat{\mu}_F^T \hat{\Sigma}_F \hat{\mu}_F} \sim \chi^2 \left(N + K - S\right),$$

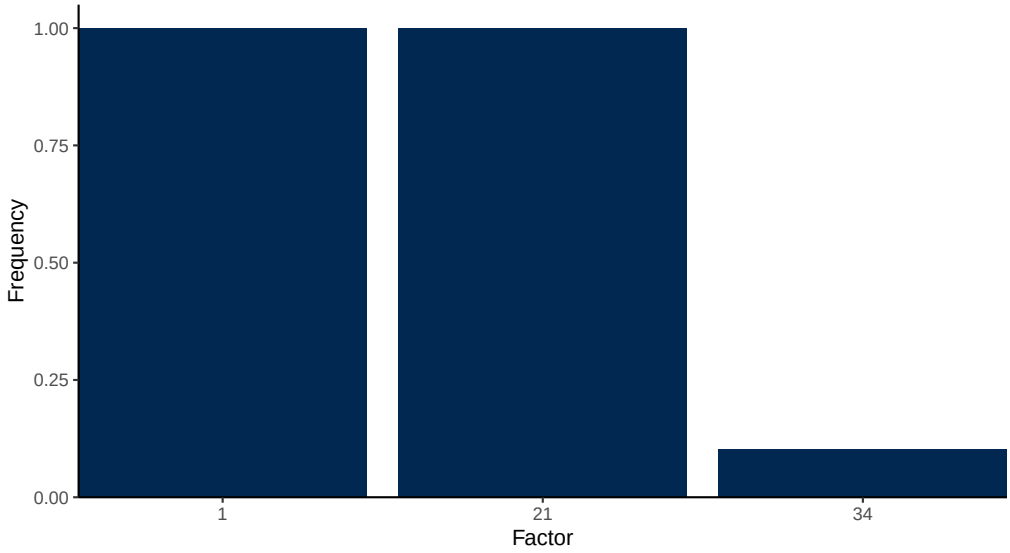
where $\hat{\Sigma}_\alpha$ is the estimated covariance matrix of the intercepts, and $\hat{\mu}_F$ and $\hat{\Sigma}_F$ are the estimated mean and covariances of the factor returns. The degrees of freedom are given by the number of test assets, and the number of potential factors subtracted the number of factors estimated to be relevant.

A number of different studies have used machine learning techniques to estimate the set of relevant risk factors. Freyberger et al. (2020) use the group lasso to model excess returns non-parametrically, Feng et al. (2020) use a double pass lasso to form a test for new factors, Lettau and Pelger (2020b) penalizes pricing errors in principal component framework, and Hwang and Rubesam (2020) models the time series and the cross section using Bayesian methods and all possible (relevant) combinations of the risk factors. These studies are different but related ways to describe the set of relevant asset pricing factors in some form. The present study differs from previous studies in its ability to explicitly name the relevant factors over time while taking the pricing error into account in addition to explaining the time series and cross section of returns. In spirit, our approach combines the penalty for pricing errors in Lettau and Pelger (2020b) and the requirement of the ability of the selected factors to price the excluded factors in Barillas and Shanken (2018). Although, we employ different methods from the lasso literature.

5.3 The relevant factors over time

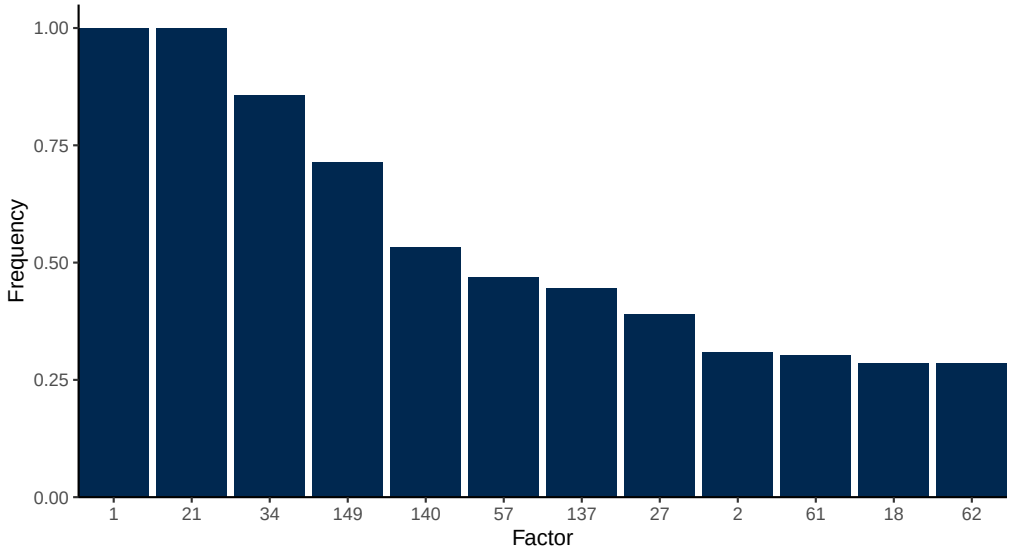
This section provides an overview of the ability of the aglasso and the pag-lasso to produce sets of factors that are able to price the test assets, as well as the factors not selected by the estimator. Figures 1, 2, and 3 show the frequency with which the most important factors are included by the aglasso as well as the pag-lasso with either up to 10 or 20 factors in the prior set. Figures 4, 5, and 6 show how the inclusion of the various factors evolves over time. Figures 7, 8, and 9 show the p-value of the GRS test corresponding to the estimated set of relevant factors at any given period.

Figure 1: Factors chosen by the aglasso



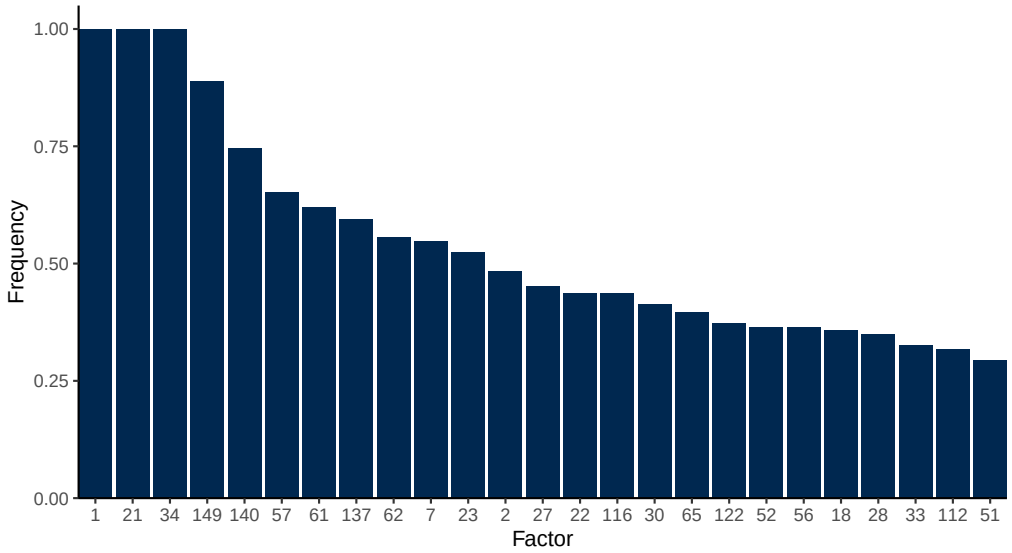
This figure shows the factors chosen by the aglasso. The vertical axis contains the frequency with which the factors are chosen across all sample windows. The horisontal axis contains the factor id referencing to the list in table 25. The factors are sorted by their frequency of inclusion in the estimated set in descending order.

Figure 2: Factors chosen by the pag-lasso with ten prior factors



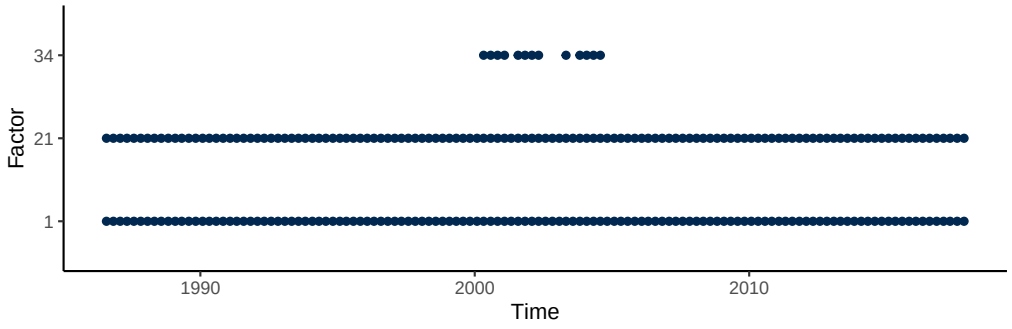
This figure shows the factors chosen by the pag-lasso with up to 10 factors in the prior set. The vertical axis contains the frequency with which the factors are chosen across all sample windows. The horisontal axis contains the factor id referencing to the list in table 25. The factors are sorted by their frequency of inclusion in the estimated set in descending order, and factors chosen with a frequency below 25% are omitted.

Figure 3: Factors chosen by the pag-lasso with 20 prior factors



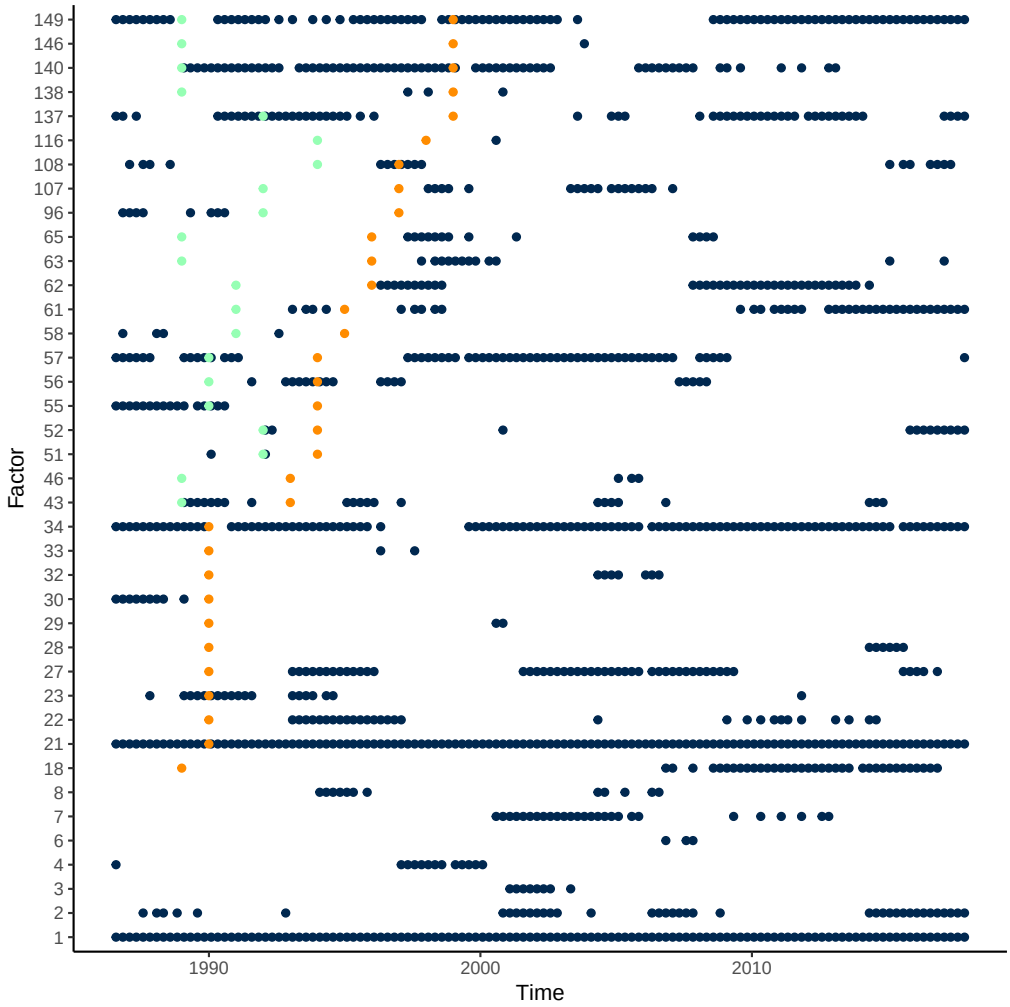
This figure shows the factors chosen by the pag-lasso with up to 20 factors in the prior set. The vertical axis contains the frequency with which the factors are chosen across all sample windows. The horisontal axis contains the factor id referencing to the list in table 25. The factors are sorted by their frequency of inclusion in the estimated set in descending order, and factors chosen with a frequency below 25% are omitted.

Figure 4: Timeline of factors chosen by the aglasso



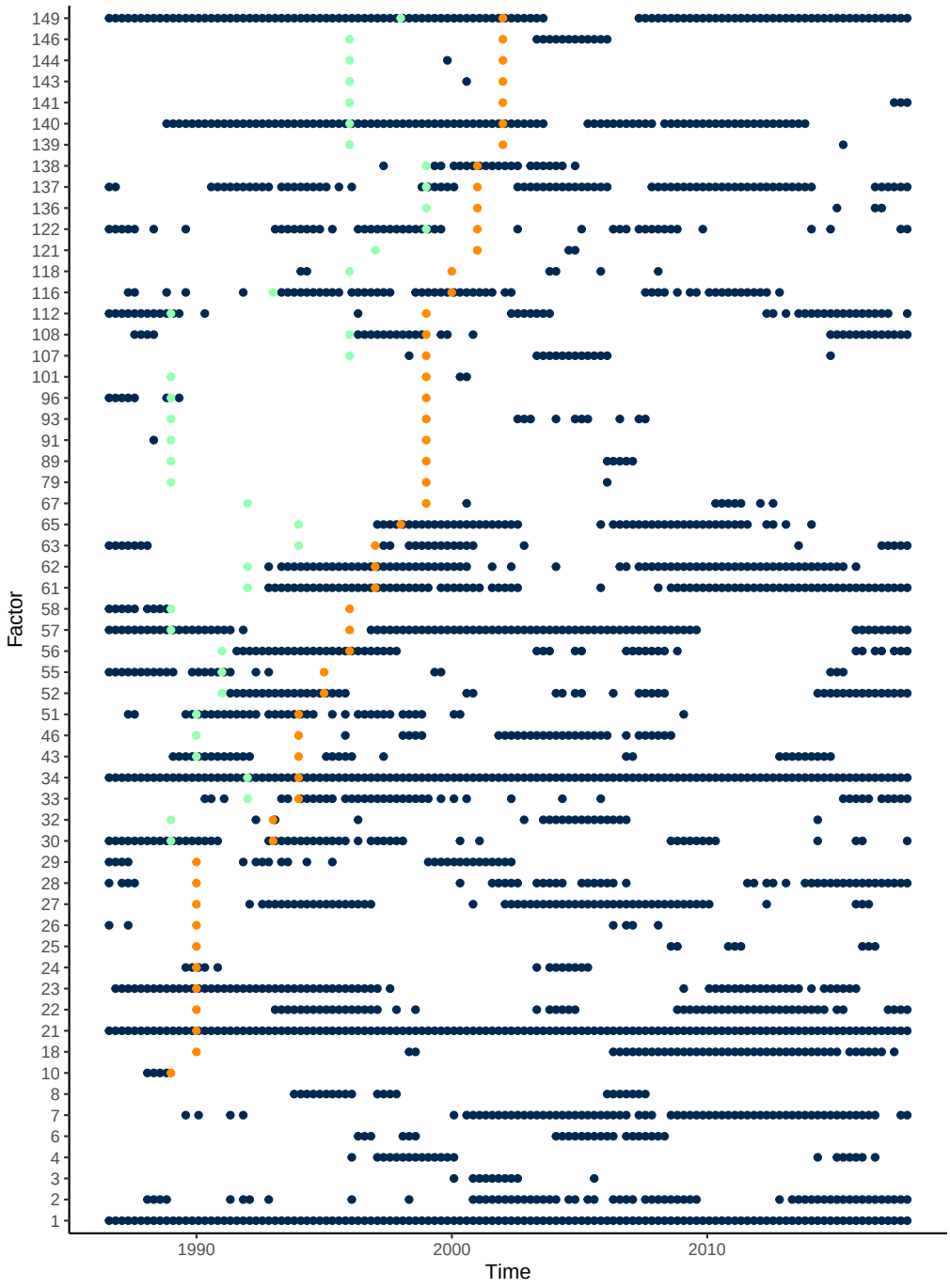
This figure shows a timeline of the periods during which the corresponding factors are included in the estimated set of relevant factors by the aglasso. The vertical axis contains the factor id referencing to the list in table 25.

Figure 5: Timeline of factors chosen by the pag-lasso with ten prior factors



This figure shows a timeline of the periods during which the corresponding factors are included in the estimated set of relevant factors by the pag-lasso with up to ten factors in the prior set. The vertical axis contains the factor id referencing to the list in table 25. The publication date of the factor is shown in orange, and the last date of the sample used in the original study is shown in green.

Figure 6: Timeline of factors chosen by the pag-lasso with 20 prior factors

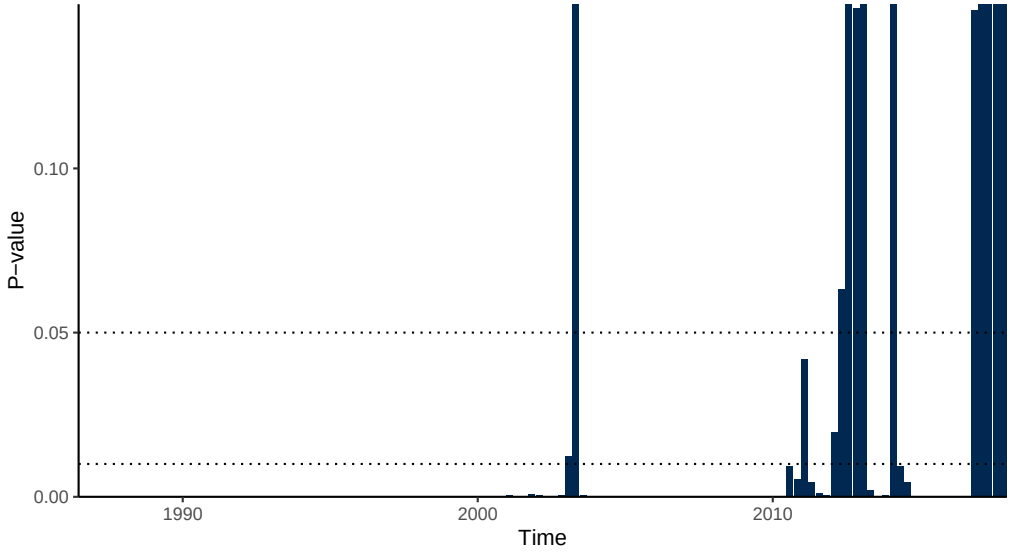


This figure shows a timeline of the periods during which the corresponding factors are included in the estimated set of relevant factors by the pag-lasso with up to 20 factors in the prior set. The vertical axis contains the factor id referencing to the list in table 25. The publication date of the factor is shown in orange, and the last date of the sample used in the original study is shown in green.

We see that the aglasso only estimates three factors to be relevant, namely the excess market return (1) (Black, Jensen, and Scholes, 1972), small minus big (21) (Fama and French, 1993), and momentum (34) (Carhart, 1997). Notably, the excess market return and small minus big are included at all times, whereas the momentum factor is only included during a smaller window from 2000-2005, as can be seen in figure 4. We also see that performing a GRS test on this very restricted set of factors rejects the null hypothesis of no pricing errors at almost all times except a few cases at the end of the sample.

The pag-lasso includes a much larger number of risk factors than the aglasso. This stems from the number of chosen parameters in the prior step of the pag-lasso is much larger due to the restriction of zero intercepts. They are, however, not all included at all times. The factors with the highest selection frequency are the excess market return (1), small minus big (21), momentum (34), the intermediary investment factor (149) (He et al., 2017), and betting against beta (140) (Frazzini and Pedersen, 2014). From figure 8, we see that there are more periods in which the GRS test cannot reject the null that the factors selected by the pag-lasso are able to jointly price the test assets as well as the unselected factors. There are, however, numerous periods in which the GRS test still leads to rejections of zero pricing errors. This changes when allowing as many as 20 factors to be included in the prior set. In addition to the factors included by the pag-lasso with ten factors in the prior set, the factors with the highest inclusion frequency are R&D to sales (57) (Chan et al., 2001), illiquidity (61) (Amihud, 2002), HML devil (137) (Asness and Frazzini, 2013), liquidity (62) (Pastor and Stambaugh, 2003), long-term reversal (7) (De Bondt and Thaler, 1985), short-term reversal (23) (Jegadeesh and Titman, 1993), and market beta (2) (Fama and MacBeth, 1973). We see in figure 9 that factors selected by the prior adaptive group lasso are able to price the test assets as well as the remaining factors in almost all periods. We also see that while the aglasso only includes the momentum factor (34) (Carhart, 1997) in a fraction of the windows, both formulations of the pag-lasso include momentum in many more windows. Across reasonable values of η the chosen set of factors only showed minor differences. The simulations in section 4 showed that the weight placed on the prior information through η plays a significant role. However, the significant overlap between the factors in the prior set and the factors that are

Figure 7: Testing the null of no pricing errors for the aglasso

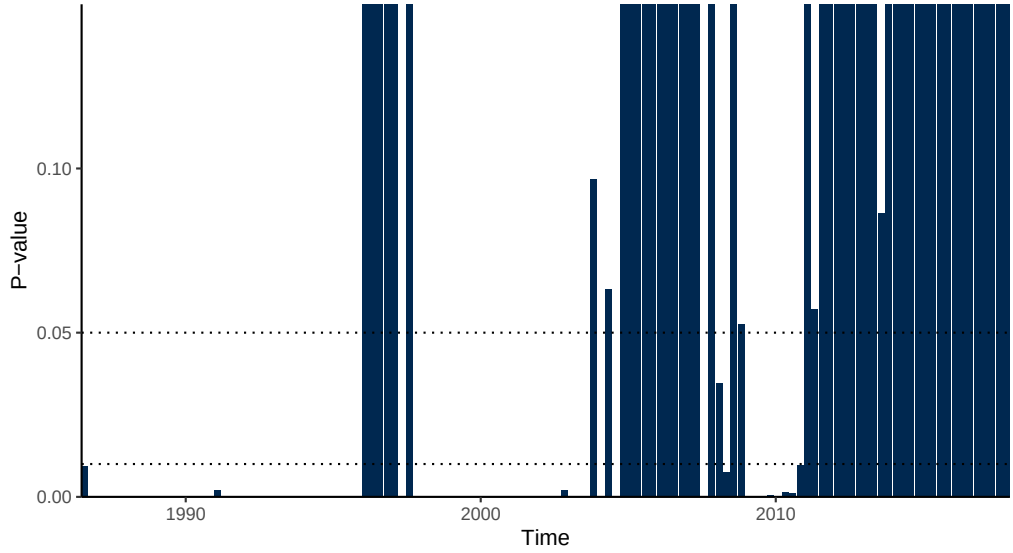


This figure shows the p-values for the GRS test, testing the null of "no pricing errors" when using the factors selected by the aglasso. The test is performed for each sample window across time as shown on the horisontal axis. The vertical axis is truncated from above at 15%. There are horisontal lines showing the 1% and 5% significance levels.

eventually selected could suggest that the explanatory power of the prior factors is so strong that it trumps the remaining factors.

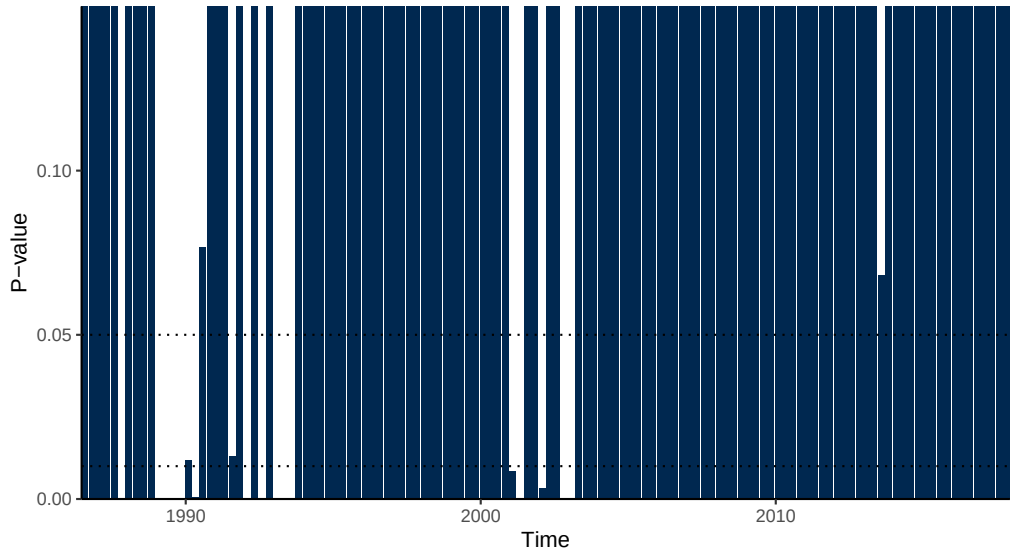
While many of the factors selected by the pag-lasso and the aglasso are well-known in the asset pricing literature, we will give a brief overview. Of the three factors selected by the aglasso we start with the well-known excess market return (Black et al., 1972), which in theory contains all assets in the investor universe. By most practical implementations, this is crudely approximated by the S&P500 index. Second, the aglasso chooses the small minus big factor (Fama and French, 1993), which illustrates that small firms tend to have higher expected returns than large firms. Finally, the aglasso selects the one year momentum factor from Carhart (1997) and Jegadeesh and Titman (1993), which takes advantage of the fact that stocks with high returns for the past year (past winners) tend to keep delivering high returns. In addition to these, the pag-lasso selects the intermediary investment factor (He et al., 2017), which considers the value-weighted equity return of primary dealers. They argue that unsophisticated households are unlikely to be the primary driver of market returns for complex assets as proposed by the consumption CAPM model. Instead,

Figure 8: Testing the null of no pricing errors for the pag-lasso with ten prior factors



This figure shows the p-values for the GRS test, testing the null of "no pricing errors" when using the factors selected by the pag-lasso with up to ten factors in the prior set. The test is performed for each sample window across time as shown on the horisontal axis. The vertical axis is truncated from above at 15%. There are horisontal lines showing the 1% and 5% significance levels.

Figure 9: Testing the null of no pricing errors for the pag-lasso with 20 prior factors



This figure shows the p-values for the GRS test, testing the null of "no pricing errors" when using the factors selected by the pag-lasso with up to 20 factors in the prior set. The test is performed for each sample window across time as shown on the horisontal axis. The vertical axis is truncated from above at 15%. There are horisontal lines showing the 1% and 5% significance levels.

they proxy the marginal investor using primary dealers since they are present in a significant part of all transactions made in the market, and over-the-counter markets in particular. The betting against beta factor (Frazzini and Pedersen, 2014) is based on a strategy that buys low beta stocks and sells high beta stocks, which should yield positive expected returns contrary to the predictions made by the original CAPM model. The R&D to sales factor (Chan et al., 2001) is based on the strategy of buying stocks of firms with a large ratio of R&D expenditures to sales and selling the stocks of those with low ratios. The illiquidity factor (Amihud, 2002) is an illustration of the fact that relatively illiquid stocks have higher expected returns than liquid stocks. The HML devil factor (Asness and Frazzini, 2013) is an alternative to the value factor (Fama and French, 1993) but using a more timely measure of the book to market ratio. The liquidity factor (Pastor and Stambaugh, 2003) uses that stocks with high sensitivity to overall market liquidity have higher expected returns than low sensitivity stocks. The long-term reversal factor (De Bondt and Thaler, 1985) is based on the assumption of overreaction of investors to dramatic news events, which is corrected over the following year(s). The short-term reversal factor (Jegadeesh and Titman, 1993) is a short-term counterpart to the long-term reversal factor by De Bondt and Thaler (1985), considering the reversal experienced in the first month after portfolio construction. The market beta factor (Fama and MacBeth, 1973) is constructed based on the sensitivity of individual stock returns to the market return.

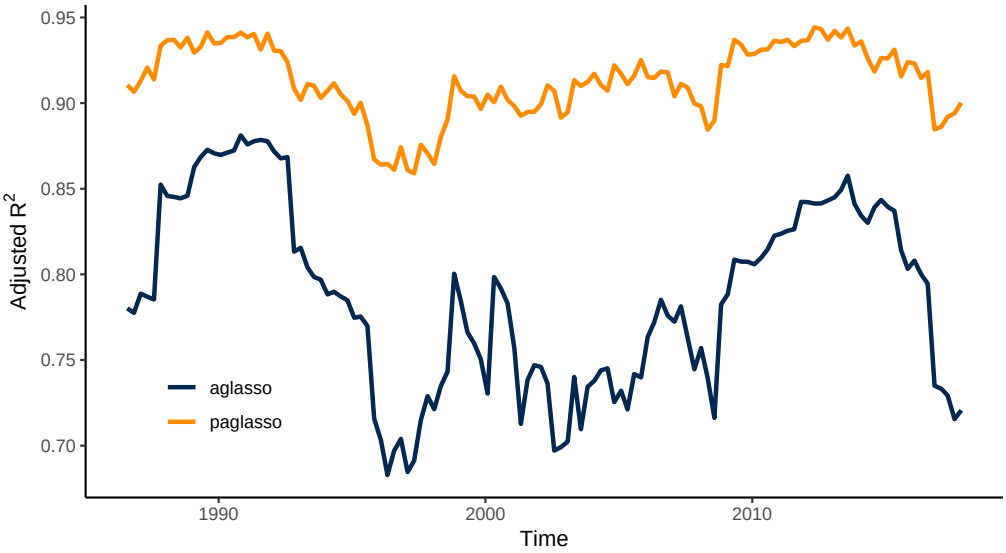
The p-values in figures 7, 8, and 9 are the result of several χ^2 tests. Hence, simply due to multiple testing, we would expect to see a positive number of rejections of the null for given positive significance levels even under the null. We see that for all the presented models, there are indeed times at which the null of zero pricing errors is rejected. Since we have not explicitly corrected for multiple testing, we cannot say for certain that the sets selected by the pag-lasso can or cannot be rejected overall. However, we can comment on the relative performance of the different approaches and it is highly unlikely that the rejections depicted in figure 7 are all due to multiple testing.

Looking at the distribution of the selected factors over time in figures 4, 5, and 6, we see that they are, for the most part, included in consecutive windows. This speaks to some degree of stability of the estimator as well as the underlying factor structure of the returns, such that if a

Figure 10: Adjusted R^2 with ten prior factors

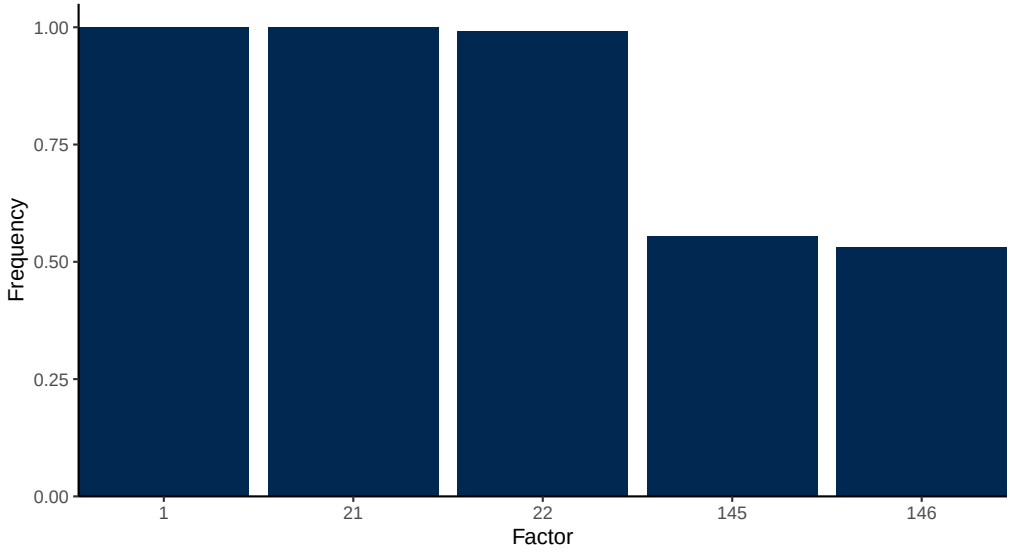
This figure shows the adjusted R^2 across all sample windows for the aglasso and the pag-lasso with up to ten factors in the prior set.

Figure 11: Adjusted R^2 with 20 prior factors



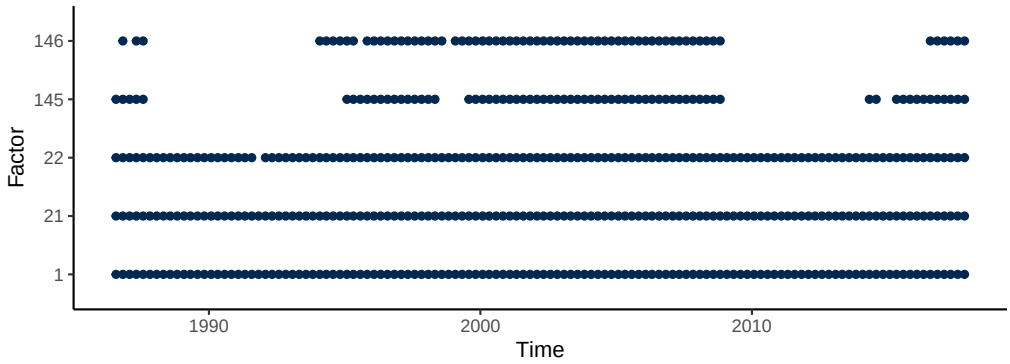
This figure shows the adjusted R^2 across all sample windows for the aglasso and the pag-lasso with up to 20 factors in the prior set.

Figure 12: Factors chosen with the Fama and French (2015) model as prior



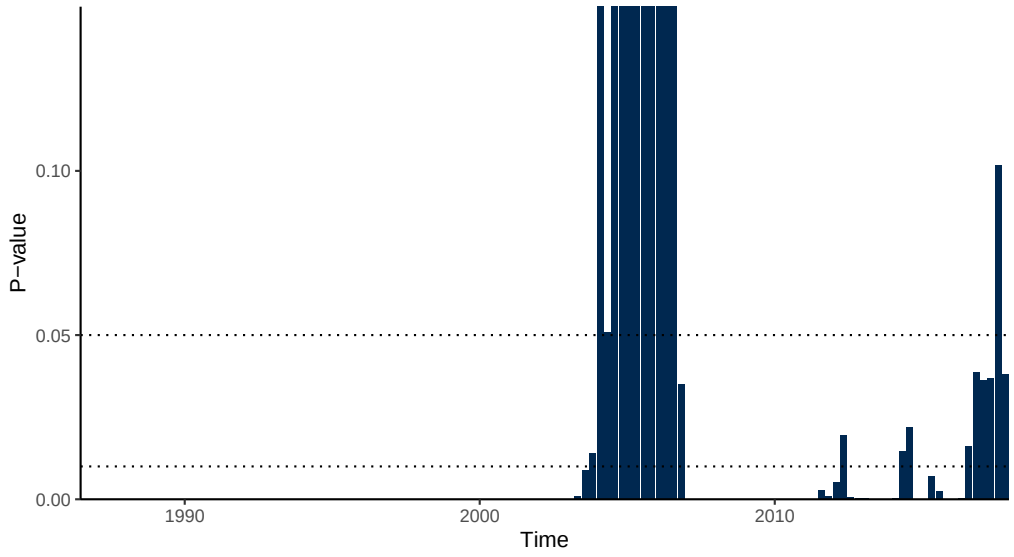
This figure shows the factors chosen by the pag-lasso with the factors from the Fama and French (2015) model in the prior set. The vertical axis contains the factor id referencing to the list in table 25. The horisontal axis contains the factor id referencing to the list in table 25. The factors are sorted by their frequency of inclusion in the estimated set in descending order.

Figure 13: Timeline of factors chosen with the Fama and French (2015) model as prior



This figure shows a timeline of the periods during which the corresponding factors are included in the estimated set of relevant factors by the pag-lasso with the factors from the Fama and French (2015) model in the prior set. The vertical axis contains the factor id referencing to the list in table 25.

Figure 14: Testing the null of no pricing errors with the Fama and French (2015) model as prior



This figure shows the p-values for the GRS test, testing the null of "no pricing errors" when using the factors selected by the pag-lasso with the factors from the Fama and French (2015) model in the prior set. The test is performed for each sample window across time as shown on the horisontal axis. The vertical axis is truncated from above at 15%. There are horisontal lines showing the 1% and 5% significance levels.

factor is included at some point in time it is likely to be included in the future as well as having been relevant in the past. The publication date and end-of-sample date for the discovery of the included factors are indicated by orange and green dots, respectively, in the figures. From this, we do not see any particular evidence that the chosen factors are only relevant before their publication date. This may serve as an indication that the selected factors were not due to data snooping at their discovery.

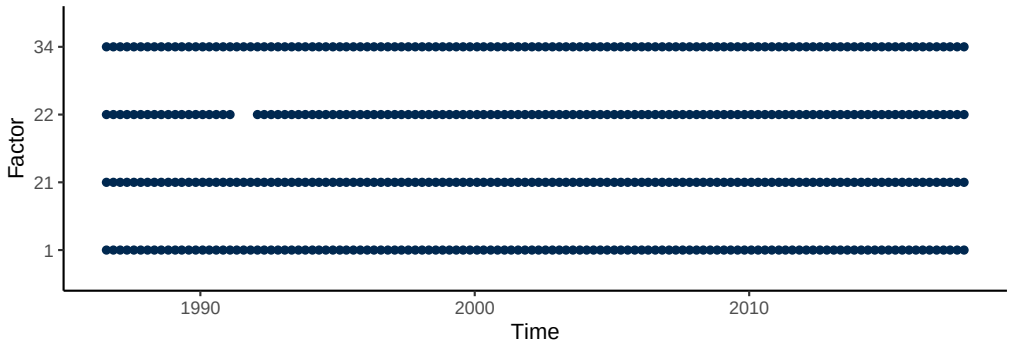
Figures 10 and 11 show the changes in adjusted R^2 for the pag-lasso with either 10 or 20 factors in the prior set compared to the aglasso, and figures 21 and 22 in the appendix show the changes in R^2 for the pag-lasso with either 10 or 20 factors in the prior set compared to the aglasso. The inclusion of additional factors by the pag-lasso compared to the aglasso yields an unsurprising increase in R^2. Perhaps more interesting, we also see an increase in the adjusted R^2, thus penalizing the inclusion of additional variables for the pag-lasso compared to the aglasso. Our results indicate that we need more than the usual (around) five factors used as controls in the literature since we need 10-20 factors in the relevant set in order to price the entire cross-section. Given that the set of relevant factors is not constant over time, the total number of historically relevant pricing factors is much larger.

5.4 Mainstream factor models as the prior set

We can also implement the pag-lasso using the mainstream factor models used in the asset pricing literature. Figures 12, 13, and 14 show the selected set when using the pag-lasso with the five factors of (Fama and French, 2015). The pag-lasso does not add any factors beyond the five-factor model, and the robust minus weak and conservative minus aggressive factors are not selected at all times. Similar to the aglasso, using the Fama-French five-factor model is not enough to achieve pricing in the cross section of the test assets as well as the remaining excluded assets.

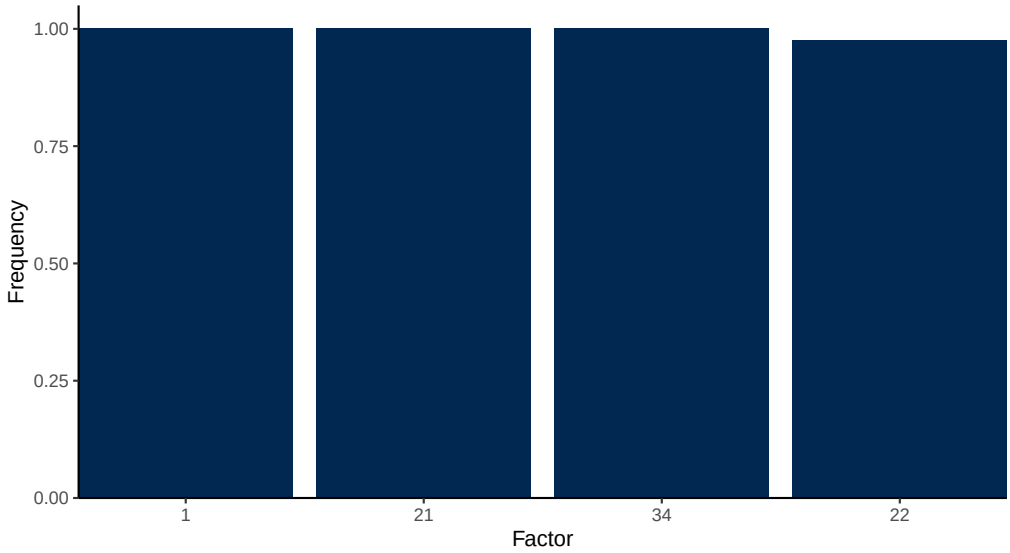
Figures 16, 15, and 17 show the selected sets of factors when the pag-lasso uses the Carhart (1997) four-factor model as the prior set. The final set of relevant factors estimated by the model contains the four factors at almost all times. We find that the Carhart (1997) model is not sufficient to price the cross section of test assets together with the remaining factors.

Figure 15: Timeline of factors chosen with the Carhart (1997) model as prior



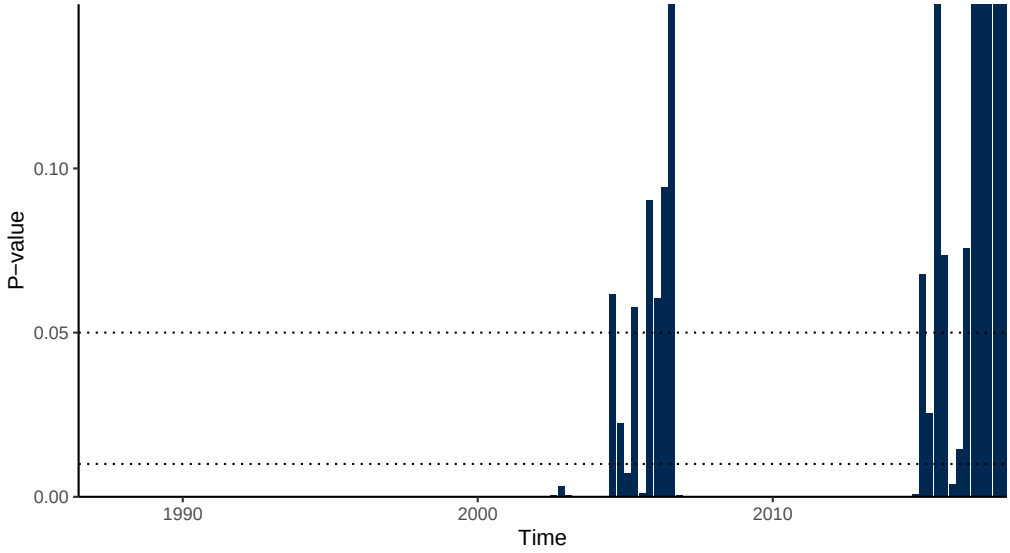
This figure shows a timeline of the periods during which the corresponding factors are included in the estimated set of relevant factors by the pag-lasso with the factors from the Carhart (1997) model in the prior set. The vertical axis contains the factor id referencing to the list in table 25.

Figure 16: Factors chosen with the Carhart (1997) model as prior



This figure shows the factors chosen by the pag-lasso with the factors from the Carhart (1997) model in the prior set. The vertical axis contains the factor id referencing to the list in table 25. The horisontal axis contains the factor id referencing to the list in table 25. The factors are sorted by their frequency of inclusion in the estimated set in descending order.

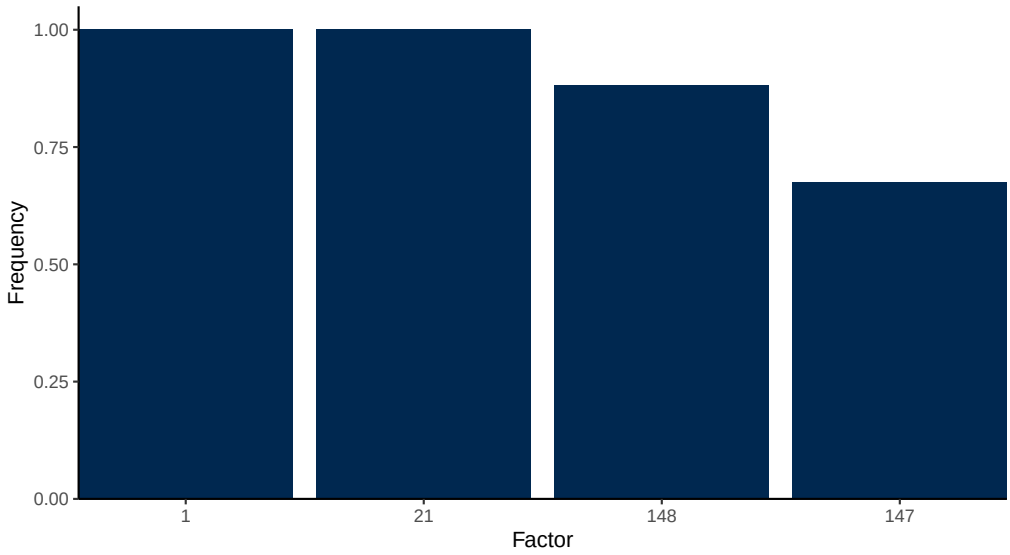
Figure 17: Testing the null of no pricing errors with the Carhart (1997) model as prior



This figure shows the p-values for the GRS test, testing the null of "no pricing errors" when using the factors selected by the pag-lasso with the factors from the Carhart (1997) model in the prior set. The test is performed for each sample window across time as shown on the horisontal axis. The vertical axis is truncated from above at 15%. There are horisontal lines showing the 1% and 5% significance levels.

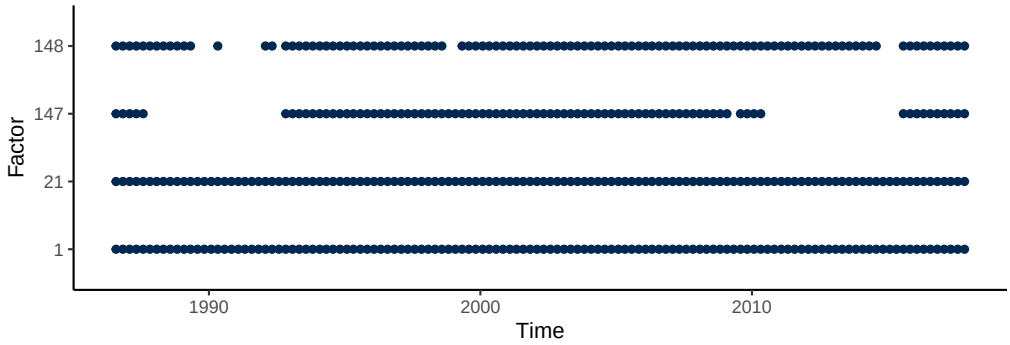
Figures 18, 19, and 20 show the selected sets of factors when the prior adaptive group lasso uses the Hou et al. (2015) q-factor model as the prior set. Similarly to the results above, we do find that using the well-established factors improves pricing slightly. However, they are nowhere near enough to price all the assets at all times.

Figure 18: Factors chosen with the Hou et al. (2015) model as prior



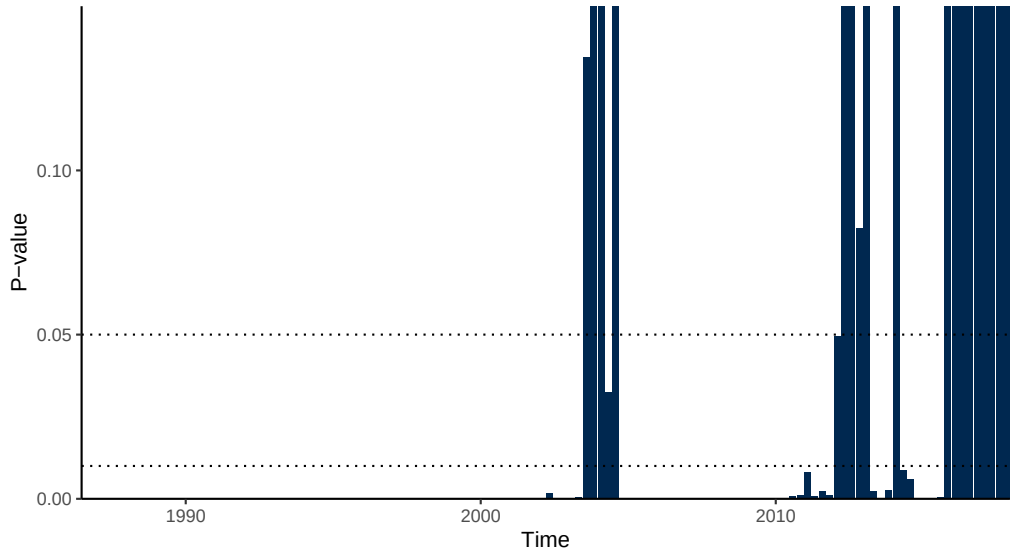
This figure shows the factors chosen by the pag-lasso with the factors from the Hou et al. (2015) model in the prior set. The vertical axis contains the factor id referencing to the list in table 25. The horisontal axis contains the factor id referencing to the list in table 25. The factors are sorted by their frequency of inclusion in the estimated set in descending order.

Figure 19: Timeline of factors chosen with the Hou et al. (2015) model as prior



This figure shows a timeline of the periods during which the corresponding factors are included in the estimated set of relevant factors by the pag-lasso with the factors from the Hou et al. (2015) model in the prior set. The vertical axis contains the factor id referencing to the list in table 25.

Figure 20: Testing the null of no pricing errors with the Hou et al. (2015) model as prior



This figure shows the p-values for the GRS test, testing the null of "no pricing errors" when using the factors selected by the pag-lasso with the factors from the Hou et al. (2015) model in the prior set. The test is performed for each sample window across time as shown on the horisontal axis. The vertical axis is truncated from above at 15%. There are horisontal lines showing the 1% and 5% significance levels.

6 Conclusion

We develop the prior adaptive group lasso (pag-lasso) that simultaneously selects variables from a high-dimensional model and estimates the coefficient values with a natural grouping of the variables while allowing for the use of previously obtained information. We show the selection and estimation consistency analytically in a low-dimensional setting with few variables relative to the number of observations as well as under the assumption of high-dimensionality with many variables relative to the number of observations. If the information used to guide the variable selection is of sufficiently high quality,[8] we establish that the pag-lasso features properties similar to those for the adaptive group lasso (aglasso) derived in Wang and Tian (2019). The analytical results are supported by a Monte Carlo study where we show that the pag-lasso does indeed achieve superior performance to the aglasso when the prior information is sufficiently accurate in finite samples. The simulations also illustrate the effect of the weight placed on the prior information. If

[8]As described by assumption (A7)

the prior information is of sufficient quality, a higher weight leads to improvements in performance. However, if the prior information is inaccurate, a higher weight transmits the inaccuracies into the pag-lasso estimator, thus hurting performance. In general we find that the pag-lasso is robust to omissions of relevant variables in the prior set, as it is still able to include them in the final set of variables. For future research, it would be interesting to derive similar econometric properties as those described in section 3 with a looser restriction on the errors of the prior information, such that the quality of the prior information and η could show up in the convergence rates.

In the empirical application, we use the fact that the pricing error in the regressions should be statistically indistinguishable from zero when including the relevant pricing factors. Using this as the guiding principle when estimating the set of relevant factors with the prior adaptive group lasso, we select more factors than the adaptive group lasso and we are able to price the test assets as well as the remaining traded factors not chosen by the pag-lasso in almost all of the considered periods. The selected set of relevant pricing factors is relatively stable over time, as many of the factors are included for several consecutive periods as opposed to popping in and out at random. The relatively restricted and considerably smaller set selected by the aglasso is, on the other hand, not able to price the test assets and the remaining factors in the majority of the sample windows. We also compare the factors selected by the pag-lasso with the factors used in the mainstream asset pricing factor models by Fama and French (2015), Carhart (1997), and Hou et al. (2015). Again, we find that the mainstream factor models are insufficient in pricing the cross section of test assets together with the factors that are not included in the factor model in question. The factors that are included most often by the pag-lasso include well-known factors like the excess market return (Black et al., 1972), small minus big (Fama and French, 1993), momentum (Carhart, 1997), the intermediary investment factor (He et al., 2017), betting against beta (Frazzini and Pedersen, 2014), R&D to sales (Chan et al., 2001), illiquidity (Amihud, 2002), HML devil (Asness and Frazzini, 2013), liquidity (Pastor and Stambaugh, 2003), long-term reversal (De Bondt and Thaler, 1985), short-term reversal (Jegadeesh and Titman, 1993), and the market beta (Fama and MacBeth, 1973). From Figure 6, showing the evolution of the set of relevant factors over time as estimated by the pag-lasso, we also find that the included factors remain relevant after their publication

dates, serving as evidence against any accusations of data snooping by the original authors of the selected factors. For future research, it would be interesting to consider different avenues to further smooth the estimated set of relevant factors and test the performance against that of the pag-lasso as described in this present paper.

References

Abarbanell, J. S. and B. J. Bushee (1998): "Abnormal returns to a fundamental analysisstrategy," The Accounting Review, 73, 19–45.

Ahmed, S., Z. Bu, and D. Tsvetanov (2019): "Best of the Best: A Comparison of Factor Models," Journal of Financial and Quantitative Analysis, 54, 1713–1758.

Ali, A., L. S. Hwang, and M. A. Trombley (2003): "Arbitrage risk and the book-to-market anomaly," Journal of Financial Economics, 69, 355–373.

Almeida, H. and M. Campello (2007): "Financial Constraints, Asset Tangibility, and Corporate Investment," Review of Financial Studies, 20, 1429–1460.

Amihud, Y. (2002): "Illiquidity and stock returns: Cross-section and time-series effects," Journal of Financial Markets, 5, 31–56.

Amihud, Y. and H. Mendelson (1989): "The Effects of Beta, Bid-Ask Spread, Residual Risk, and Size on Stock Returns," The Journal of Finance, 44, 479–486.

Anderson, C. W. and L. Garcia-Feijóo (2006): "Empirical evidence on capital investment, growth options, and security returns," The Journal of Finance, 61, 171–194.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006): "The cross-section of volatility and expected returns," The Journal of Finance, 61, 259–299.

Asness, C. S. and A. Frazzini (2013): "The devil in HML's details," Journal of Portfolio Management, 39, 49–68.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2019): "Quality minus junk," Review of Accounting Studies, 24, 34–112.

Asness, C. S., R. B. Porter, and R. L. Stevens (2000): "Predicting stock returns usingindustry-relative firm characteristics," Technical Report, AQR Capital Investment.

Balakrishnan, K., E. Bartov, and L. Faurel (2010): "Post loss/profit announcement drift," Journal of Accounting and Economics, 50, 20–41.

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011): "Maxing out: Stocks as lotteries and the cross-section of expected returns," Journal of Financial Economics, 99, 427–446.

Bandyopadhyay, S. P., A. G. Huang, and T. S. Wirhjanto (2010): "The accrual volatility anomaly," Technical Report, University of Waterloo.

Barbee, W. C., S. Mukherji, and G. A. Raines (1996): "Do sales-price and debt-equity explain stock returns better than book-market and firm size?" Financial Analysts Journal, 52, 56–60.

Barillas, F. and J. Shanken (2018): "Comparing Asset Pricing Models," The Journal of Finance, 73, 715–754.

Barth, M. E., J. A. Elliott, and M. W. Finn (1999): "Market Rewards Associated with Patterns of Increasing Earnings," Journal of Accounting Research, 37, 387–413.

Basu, S. (1977): "Investment Performance of Common Stocks in Relation to their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis," The Journal of Finance, 32, 663–682.

Belo, F. and X. Lin (2012): "The Inventory Growth Spread," Review of Financial Studies, 25, 278–313.

Belo, F., X. Lin, and S. Bazdresch (2014): "Labor hiring, investment, and stock return predictability in the cross section," Journal of Political Economy, 122, 129–177.

Bhandari, L. C. (1988): "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence," The Journal of Finance, 43, 507–528.

Black, F., M. C. Jensen, and M. Scholes (1972): "The Capital Asset Pricing Model: Some Empirical Tests," in Studies in the Theory of Capital Markets, Praeger, New York, 79–121.

Boudoukh, J., R. Michaely, M. Richardson, and M. R. Roberts (2007): "On the importance of measuring payout yield: Implications for empirical asset pricing," The Journal of Finance, 62, 877–915.

Bradshaw, M. T., S. A. Richardson, and R. G. Sloan (2006): "The relation between corporate financing activities, analysts' forecasts and stock returns," Journal of Accounting and Economics, 42, 53–85.

Brandt, M. W., R. Kishore, P. Santa-Clara, and M. Venkatachalam (2008): "Earnings announcements are full of surprises," Technical Report, Duke University.

Brown, D. P. and B. Rowe (2007): "The productivity premium in equity returns," Technical Report, University of Wisconsin-Madison.

Bryzgalova, S. (2019): "Bayesian Solutions for the Factor Zoo : We Just Ran Two Quadrillion Models," Available at SSRN: http://dx.doi.org/10.2139/ssrn.3481736, 1–46.

Carhart, M. M. (1997): "On persistence in mutual fund performance," The Journal of Finance, 52, 57–82.

Chan, L. K., J. Lakonishok, and T. Sougiannis (2001): "The stock market valuation of research and development expenditures," The Journal of Finance, 56, 2431–2456.

Chandrashekar, S. and R. K. S. Rao (2009): "The productivity of corporate cash holdingsand the cross-section of expected stock returns," Technical Report, Univerisity of Texas at Austin.

Chen, L. and L. Zhang (2010): "A better three-factor model that explains more anomalies," The Journal of Finance, 65, 563–595.

Chordia, T., A. Subrahmanyam, and V. R. Anshuman (2001): "Trading activity and expected stock returns," Journal of Financial Economics, 59, 3–32.

Daniel, K. and S. Titman (2006): "Market reactions to tangible and intangible information," The Journal of Finance, 61, 1605–1643.

Datar, V. T., N. Y. Naik, and R. Radcliffe (1998): "Liquidity and stock returns: An alternative test," Journal of Financial Markets, 1, 203–219.

De Bondt, W. F. M. and R. Thaler (1985): "Does the Stock Market Overreact?" The Journal of Finance, 40, 793–805.

Desai, H., S. Rajgopal, and M. Venkatachalam (2004): "Value-glamour and accruals mispricing: One anomaly or two?" Accounting Review, 79, 355–385.

Dichev, I. D. (1998): "Is the risk of bankruptcy a systematic risk?" The Journal of Finance, 53, 1131–1147.

Eberhart, A. C., W. F. Maxwell, and A. R. Siddique (2004): "An Examination of Long-Term Abnormal Stock Returns and Operating Performance Following R&d Increases," The Journal of Finance, 59, 623–650.

Eisfeldt, A. L. and D. Papanikolaou (2013): "Organization capital and the cross-section of expected returns," The Journal of Finance, 68, 1365–1406.

Fairfield, P. M., S. Whisenant, and T. L. Yohn (2003): "The differential persistence of accruals and cash flows for future operating income versus future profitability," Review of Accounting Studies, 8, 221–243.

Fama, E. F. and K. R. French (1993): "Common risk factors in the returns on stocks and bonds," Journal of Financial Economics, 33, 3–56.

——— (2015): "A Five-Factor Asset Pricing Model," Journal of Financial Economics, 116, 1–22.

Fama, E. F. and J. D. MacBeth (1973): "Risk, return, and equilibrium: Empirical tests," Journal of Political Economy, 81, 607–636.

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

Fan, J. and H. Peng (2004): "Nonconcave penalized likelihood with a diverging number of parameters," Annals of Statistics, 32, 481–499.

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," The Journal of Finance, 75, 1327–1370.

Francis, J., R. LaFond, P. M. Olsson, and K. Schipper (2004): "Costs of equity and earnings attributes," The Accounting Review, 79, 967–1010.

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

Frazzini, A. and L. H. Pedersen (2014): "Betting against beta," Journal of Financial Economics, 111, 1–25.

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting Characteristics Nonparametrically," The Review of Financial Studies, 33, 2326–2377.

Gettleman, E. and J. M. Marks (2006): "Acceleration strategies," Technical Report, Bentley University.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989): "A Test of the Efficiency of a Given Portfolio," Econometrica, 57, 1121–1152.

Giglio, S. and D. Xiu (2019): "Asset Pricing with Omitted Factors," Chicago Booth Research Paper No.16-21, 1–28.

Green, J., J. R. Hand, and X. F. Zhang (2017): "The characteristics that provide independent information about average u.s. monthly stock returns," Review of Financial Studies, 30, 4389–4436.

Hafzalla, N., R. Lundholm, and E. M. Van Winkle (2011): "Percent accruals," Accounting Review, 86, 209–236.

Harvey, C. R. and Y. Liu (2019): "A Census of the Factor Zoo," SSRN Electronic Journal, 1–7.

Harvey, C. R., Y. Liu, and H. Zhu (2016): ". . . and the Cross-Section of Expected Returns," Review of Financial Studies, 29, 5–68.

Haugen, R. A. and N. L. Baker (1996): "Commonality in the determinants of expected stock returns," Journal of Financial Economics, 41, 401–439.

He, Z., B. Kelly, and A. Manela (2017): "Intermediary asset pricing: New evidence from many asset classes," Journal of Financial Economics, 126, 1–35.

Heston, S. L. and R. Sadka (2008): "Seasonality in the cross-section of stock returns," Journal of Financial Economics, 87, 418–445.

Hirshleifer, D., K. Hou, S. H. Teoh, and Y. Zhang (2004): "Do investors overvalue firms with bloated balance sheets?" Journal of Accounting and Economics, 38, 297–331.

Holthausen, R. W. and D. F. Larcker (1992): "The prediction of stock returns using financial statement information," Journal of Accounting and Economics, 15, 373–411.

Hong, H. and M. Kacperczyk (2009): "The price of sin: The effects of social norms on markets," Journal of Financial Economics, 93, 15–36.

Hou, K. and T. J. Moskowitz (2005): "Market Frictions, Price Delay, and the Cross-Section of Expected Returns," Review of Financial Studies, 18, 981–1020.

Hou, K. and D. T. Robinson (2006): "Industry concentration and average stock returns," The Journal of Finance, 61, 1927–1956.

Hou, K., C. Xue, and L. Zhang (2015): "Digesting anomalies: An investment approach," Review of Financial Studies, 28, 650–705.

——— (2020): "Replicating Anomalies," Review of Financial Studies, 33, 2019–2133.

Huang, A. G. (2009): "The cross section of cashflow volatility and expected stock returns," Journal of Empirical Finance, 16, 409–429.

Huang, J., S. Ma, and C.-H. Zhang (2008): "Adaptive Lasso for Sparse High-Dimensional Regression Models," Statistica Sinica, 18, 1603–1618.

Hwang, S. and A. Rubesam (2020): "Bayesian Selection of Asset Pricing Factors Using Individual Stocks," Journal of Financial Econometrics, nbaa045, 1–46.

Jegadeesh, N. and J. Livnat (2006): "Revenue surprises and stock returns," Journal of Accounting and Economics, 41, 147–171.

Jegadeesh, N. and S. Titman (1993): "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," The Journal of Finance, 48, 65–91.

Jiang, G., C. M. Lee, and Y. Zhang (2005): "Information uncertainty and expected returns," Review of Accounting Studies, 10, 185–221.

Jiang, Y., Y. He, and H. Zhang (2016): "Variable Selection With Prior Information for Generalized Linear Models via the Prior LASSO Method," Journal of the American Statistical Association, 111, 355–376.

Kama, I. (2009): "On the Market Reaction to Revenue and Earnings Surprises," Journal of Business Finance & Accounting, 36, 31–50.

Knight, K. and W. Fu (2000): "Asymptotics for Lasso-type estimators," Annals of Statistics, 28, 1356–1378.

Kozak, S., S. Nagel, and S. Santosh (2020): "Shrinking the cross-section," Journal of Financial Economics, 135, 271–292.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994): "Contrarian Investment, Extrapolation, and Risk," The Journal of Finance, 49, 1541–1578.

Lamont, O., C. Polk, and J. Saaá-Requejo (2001): "Financial Constraints and Stock Returns," Review of Financial Studies, 14, 529–554.

Lerman, A., J. Livnat, and R. R. Mendenhall (2008): "The high-volume return pre-mium and post-earnings announcement drift," Technical Report, Yale University.

Lettau, M. and M. Pelger (2020a): "Estimating Latent Asset-Pricing Factors," Journal of Econometrics, 1–31.

——— (2020b): "Factors That Fit the Time Series and Cross-Section of Stock Returns," Review of Financial Studies, 33, 2274–2325.

Lev, B. and D. Nissim (2004): "Taxable income, future earnings, and equity values," .

Lintner, J. (1965): "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," The Review of Economics and Statistics, 47, 13–37.

Litzenberger, R. H. and K. Ramaswamy (1979): "The effect of personal taxes and dividends on capital asset prices. Theory and empirical evidence," Journal of Financial Economics, 7, 163–195.

Liu, W. (2006): "A liquidity-augmented capital asset pricing model," Journal of Financial Economics, 82, 631–671.

Lou, D. (2014): "Attracting Investor Attention through Advertising," Review of Financial Studies, 27, 1797–1829.

Loughran, T. and J. R. Ritter (1995): "The New Issues Puzzle," The Journal of Finance, 50, 23–51.

Loughran, T. and J. W. Wellman (2011): "New evidence on the relation between the enterprise multiple and average stock returns," Journal of Financial and Quantitative Analysis, 46, 1629–1650.

Lyandres, E., L. Sun, and L. Zhang (2008): "The New Issues Puzzle: Testing the Investment-Based Explanation," Review of Financial Studies, 21, 2825–2855.

Meinshausen, N. and P. Bühlmann (2006): "High-dimensional graphs and variable selection with the Lasso," Annals of Statistics, 34, 1436–1462.

Michaely, R., R. H. Thaler, and K. L. Womack (1995): "Price Reactions to Dividend Initiations and Omissions: Overreaction or Drift?" The Journal of Finance, 50, 573–608.

Miller, M. H. and M. S. Scholes (1982): "Dividends and Taxes: Some Empirical Evidence," Journal of Political Economy, 90, 1118–1141.

Mohanram, P. S. (2005): "Separating winners from losers among low book-to-market stocks using financial statement analysis," Review of Accounting Studies, 10, 133–170.

Moskowitz, T. J. and M. Grinblatt (1999): "Do industries explain momentum?" The Journal of Finance, 54, 1249–1290.

Nardi, Y. and A. Rinaldo (2008): "On the asymptotic properties of the group lasso estimator for linear models," Electronic Journal of Statistics, 2, 605–633.

Nelder, J. and R. Wedderburn (1972): "Generalized Linear Models," Journal of the Royal Statistical Society. Series A (General), 135, 370–384.

Novy-Marx, R. (2011): "Operating Leverage," Review of Finance, 15, 103–134.

——— (2013): "The other side of value: The gross profitability premium," Journal of Financial Economics, 108, 1–28.

Ortiz-Molina, H. and G. M. Phillips (2014): "Real asset illiquidity and the cost of capital," Journal of Financial and Quantitative Analysis, 49, 1–32.

Ou, J. A. and S. H. Penman (1989): "Financial statement analysis and the prediction of stock returns," Journal of Accounting and Economics, 11, 295–329.

Palazzo, B. (2012): "Cash holdings, risk, and expected returns," Journal of Financial Economics, 104, 162–185.

Pastor, L. and R. F. Stambaugh (2003): "Liquidity risk and expected stock returns," Journal of Political Economy, 111, 642–685.

Penman, S. H., S. A. Richardson, and I. Tuna (2007): "The book-to-price effect in stock returns: Accounting for leverage," Journal of Accounting Research, 45, 427–467.

Piotroski, J. D. (2000): "Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers," Journal of Accounting Research, 38, 1–41.

Pontiff, J. and A. Woodgate (2008): "Share issuance and cross-sectional returns," The Journal of Finance, 63, 921–945.

Portnoy, S. (1984): "Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. I. Consistency," Annals of Statistics, 12, 1298–1309.

Pukthuanthong, K., R. Roll, and A. Subrahmanyam (2019): "A protocol for factor identification," Review of Financial Studies, 32, 1573–1607.

Rajgopal, S., T. Shevlin, and M. Venkatachalam (2003): "Does the stock market fully appreciate the implications of leading indicators for future earnings? Evidence from order backlog," Review of Accounting Studies, 8, 461–492.

Rendleman, R. J., C. P. Jones, and H. A. Latané (1982): "Empirical anomalies based on unexpected earnings and the importance of risk adjustments," Journal of Financial Economics, 10, 269–287.

Richardson, S. A., R. G. Sloan, M. T. Soliman, and I. Tuna (2005): "Accrual reliability, earnings persistence and stock prices," Journal of Accounting and Economics, 39, 437–485.

Ross, S. A. (1976): "The arbitrage theory of capital asset pricing," Journal of Economic Theory, 13, 341–360.

Sharpe, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," The Journal of Finance, 19, 425.

Sloan, R. G. (1996): "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings? on JSTOR," The Accounting Review, 71, 289–315.

Soliman, M. T. (2008): "The use of DuPont analysis by market participants," Accounting Review, 83, 823–853.

Stambaugh, R. F. and Y. Yuan (2016): "Mispricing Factors," Review of Financial Studies, 30, 1270–1315.

Thomas, J. K. and H. Zhang (2002): "Inventory changes and future returns," Review of Accounting Studies, 7, 163–187.

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

Titman, S., K. C. Wei, and F. Xie (2004): "Capital investments and stock returns," Journal of Financial and Quantitative Analysis, 39, 677–700.

Tuzel, S. (2010): "Corporate Real Estate Holdings and the Cross-Section of Stock Returns," Review of Financial Studies, 23, 2268–2302.

Valta, P. (2016): "Strategic Default, Debt Structure, and Stock Returns," Journal of Financial and Quantitative Analysis, 51, 197–229.

van der Vaart, A. W. and J. A. Wellner (1997): "Weak Convergence and Empirical Processes: With Applications to Statistics," Journal of the Royal Statistical Society. Series A (General), 160, 596–608.

Wang, H. and C. Leng (2008): "A note on adaptive group lasso," Computational Statistics and Data Analysis, 52, 5277–5286.

Wang, L., Y. You, and H. Lian (2015): "Convergence and sparsity of Lasso and group Lasso in high-dimensional generalized linear models," Statistical Papers, 56, 819–828.

Wang, M. and G. L. Tian (2019): "Adaptive group Lasso for high-dimensional generalized linear models," Statistical Papers, 60, 1469–1486.

Wei, F. and J. Huang (2010): "Consistent group selection in high-dimensional linear regression," Bernoulli, 16, 1369–1384.

Whited, T. M. and G. Wu (2006): "Financial Constraints Risk," Review of Financial Studies, 19, 531–559.

Xing, Y. (2008): "Interpreting the Value Effect Through the Q-Theory: An Empirical Investigation," Review of Financial Studies, 21, 1767–1795.

Zhang, C. and Y. Xiang (2016): "On the oracle property of adaptive group Lasso in high-dimensional linear models," Statistical Papers, 57, 249–265.

Zhao, P. and B. Yu (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 T. Hastie (2005): "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society. Series B (Methodological), 67, 301–320.

A Tables

Table 7: Simulations for example 1 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	12.615	14.026	12.670	14.027	12.919	12.502	13.333
	(0.249)	(0.164)	(0.245)	(0.165)	(0.271)	(0.274)	(0.326)
CNZ	12.605	14.026	12.666	14.026	12.604	12.235	12.297
	(0.248)	(0.164)	(0.245)	(0.165)	(0.247)	(0.259)	(0.256)
INZ	0.010	0.000	0.004	0.001	0.316	0.267	1.036
	(0.018)	(0.003)	(0.011)	(0.006)	(0.101)	(0.089)	(0.189)
Contains	0.427	0.711	0.435	0.713	0.425	0.364	0.373
	(0.049)	(0.045)	(0.050)	(0.045)	(0.049)	(0.048)	(0.048)
Sparsity	0.425	0.711	0.435	0.713	0.380	0.333	0.264
	(0.049)	(0.045)	(0.050)	(0.045)	(0.049)	(0.047)	(0.044)
Bias	13.885	4.283	6.713	4.280	6.837	7.211	7.108
	(0.154)	(0.256)	(0.401)	(0.256)	(0.406)	(0.406)	(0.401)
MSE	21.693	4.956	8.195	4.952	8.345	9.027	8.681
	(9.144)	(3.226)	(6.354)	(3.221)	(6.539)	(6.965)	(6.662)
ME	18.358	2.179	5.397	2.177	5.581	6.221	5.969
	(0.914)	(0.322)	(0.633)	(0.322)	(0.649)	(0.689)	(0.659)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{01} from (3) and X from (7). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 8: Simulations for example 2 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	12.957	14.199	12.966	14.196	13.704	13.680	15.156
	(0.239)	(0.148)	(0.236)	(0.150)	(0.291)	(0.309)	(0.383)
CNZ	12.921	14.193	12.951	14.181	12.855	12.567	12.585
	(0.235)	(0.147)	(0.235)	(0.148)	(0.236)	(0.245)	(0.245)
INZ	0.036	0.006	0.015	0.015	0.849	1.113	2.571
	(0.035)	(0.013)	(0.025)	(0.021)	(0.157)	(0.173)	(0.290)
Contains	0.481	0.753	0.485	0.749	0.464	0.406	0.413
	(0.050)	(0.043)	(0.050)	(0.043)	(0.050)	(0.049)	(0.049)
Sparsity	0.473	0.751	0.483	0.745	0.337	0.268	0.173
	(0.050)	(0.043)	(0.050)	(0.044)	(0.047)	(0.044)	(0.038)
Bias	6.022	2.322	2.720	2.334	2.768	3.594	3.522
	(0.105)	(0.149)	(0.162)	(0.150)	(0.163)	(0.208)	(0.206)
MSE	20.657	4.700	7.831	4.718	7.907	8.392	7.920
	(8.595)	(2.688)	(6.424)	(2.687)	(6.327)	(6.665)	(6.073)
ME	17.384	1.935	5.062	1.955	5.226	5.726	5.415
	(0.865)	(0.271)	(0.644)	(0.271)	(0.634)	(0.665)	(0.606)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{02} from (4) and X from (7). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 9: Simulations for example 3 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	13.191	14.259	13.131	14.256	13.398	13.233	14.460
	(0.226)	(0.145)	(0.232)	(0.144)	(0.253)	(0.270)	(0.347)
CNZ	13.161	14.256	13.116	14.250	13.098	12.789	12.882
	(0.225)	(0.145)	(0.231)	(0.143)	(0.230)	(0.247)	(0.241)
INZ	0.030	0.003	0.015	0.006	0.300	0.444	1.578
	(0.030)	(0.009)	(0.021)	(0.013)	(0.102)	(0.115)	(0.235)
Contains	0.534	0.773	0.526	0.769	0.519	0.471	0.486
	(0.050)	(0.042)	(0.050)	(0.042)	(0.050)	(0.050)	(0.050)
Sparsity	0.530	0.772	0.523	0.767	0.467	0.407	0.281
	(0.050)	(0.042)	(0.050)	(0.042)	(0.050)	(0.049)	(0.045)
Bias	4.275	1.696	1.861	1.699	1.874	2.395	2.421
	(0.091)	(0.132)	(0.132)	(0.133)	(0.130)	(0.171)	(0.171)
MSE	18.637	4.519	6.995	4.520	6.998	7.548	7.102
	(7.532)	(2.715)	(5.412)	(2.575)	(5.459)	(5.858)	(5.286)
ME	15.368	1.742	4.197	1.733	4.231	4.762	4.453
	(0.750)	(0.270)	(0.536)	(0.256)	(0.535)	(0.581)	(0.523)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{03} from (5) and X from (7). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 10: Simulations for example 1 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	23.474	16.328	17.374	17.921	28.540	23.501	35.495
	(0.763)	(0.269)	(0.388)	(0.334)	(0.695)	(0.663)	(0.825)
CNZ	14.353	14.972	14.819	14.968	14.766	14.230	14.129
	(0.133)	(0.029)	(0.074)	(0.031)	(0.083)	(0.142)	(0.151)
INZ	9.121	1.356	2.555	2.954	13.774	9.271	21.366
	(0.743)	(0.267)	(0.379)	(0.333)	(0.690)	(0.642)	(0.809)
Contains	0.797	0.991	0.941	0.989	0.924	0.760	0.732
	(0.040)	(0.010)	(0.024)	(0.010)	(0.026)	(0.043)	(0.044)
Sparsity	0.092	0.701	0.498	0.393	0.014	0.048	0.001
	(0.029)	(0.046)	(0.050)	(0.049)	(0.012)	(0.021)	(0.003)
Bias	7.929	7.387	7.712	7.581	8.583	8.589	9.319
	(0.198)	(0.188)	(0.212)	(0.191)	(0.228)	(0.228)	(0.243)
MSE	3.258	3.198	3.098	3.079	2.461	2.682	2.151
	(0.636)	(0.557)	(0.616)	(0.549)	(0.559)	(0.677)	(0.598)
ME	1.002	0.798	0.974	0.918	1.609	1.536	2.049
	(0.044)	(0.042)	(0.054)	(0.044)	(0.060)	(0.065)	(0.065)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{01} from (3) and X from (6). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 11: Simulations for example 2 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	26.118	16.854	18.321	18.822	30.936	25.962	38.436
	(0.910)	(0.321)	(0.476)	(0.407)	(0.732)	(0.784)	(0.865)
CNZ	14.361	14.991	14.835	14.988	14.754	14.262	14.079
	(0.140)	(0.016)	(0.075)	(0.019)	(0.087)	(0.146)	(0.164)
INZ	11.757	1.863	3.486	3.834	16.182	11.700	24.357
	(0.894)	(0.321)	(0.468)	(0.405)	(0.726)	(0.771)	(0.859)
Contains	0.808	0.997	0.949	0.996	0.922	0.776	0.731
	(0.039)	(0.005)	(0.022)	(0.006)	(0.027)	(0.042)	(0.044)
Sparsity	0.047	0.634	0.426	0.310	0.006	0.025	0.000
	(0.021)	(0.048)	(0.049)	(0.046)	(0.008)	(0.016)	(0.000)
Bias	5.008	4.338	4.518	5.846	6.903	7.214	8.069
	(0.155)	(0.142)	(0.150)	(0.226)	(0.246)	(0.242)	(0.270)
MSE	3.098	3.150	3.020	3.011	2.364	2.514	2.029
	(0.663)	(0.587)	(0.653)	(0.577)	(0.592)	(0.699)	(0.614)
ME	1.053	0.854	1.061	0.993	1.720	1.695	2.185
	(0.048)	(0.046)	(0.060)	(0.048)	(0.062)	(0.071)	(0.068)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{02} from (4) and X from (6). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 12: Simulations for example 3 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	24.192	16.569	17.601	18.192	28.923	24.162	36.123
	(0.793)	(0.316)	(0.408)	(0.361)	(0.714)	(0.704)	(0.841)
CNZ	14.430	14.973	14.853	14.976	14.832	14.307	14.226
	(0.125)	(0.028)	(0.065)	(0.027)	(0.069)	(0.134)	(0.141)
INZ	9.762	1.596	2.748	3.216	14.091	9.855	21.897
	(0.780)	(0.316)	(0.400)	(0.360)	(0.709)	(0.686)	(0.827)
Contains	0.820	0.991	0.951	0.992	0.944	0.780	0.756
	(0.038)	(0.009)	(0.022)	(0.009)	(0.023)	(0.041)	(0.043)
Sparsity	0.097	0.685	0.504	0.388	0.016	0.053	0.000
	(0.030)	(0.046)	(0.050)	(0.049)	(0.013)	(0.022)	(0.000)
Bias	3.556	2.985	3.077	4.241	5.824	5.421	7.086
	(0.118)	(0.114)	(0.120)	(0.199)	(0.260)	(0.219)	(0.279)
MSE	3.214	3.174	3.068	3.051	2.412	2.625	2.092
	(0.633)	(0.594)	(0.628)	(0.561)	(0.577)	(0.685)	(0.603)
ME	1.009	0.833	0.997	0.954	1.647	1.575	2.086
	(0.044)	(0.045)	(0.055)	(0.046)	(0.061)	(0.069)	(0.068)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{03} from (5) and X from (6). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 13: Simulations for example 1 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	11.376	14.041	12.014	14.042	12.379	11.498	13.095
	(0.285)	(0.163)	(0.257)	(0.164)	(0.287)	(0.314)	(0.414)
CNZ	11.368	14.041	12.011	14.041	11.969	10.992	11.035
	(0.284)	(0.163)	(0.257)	(0.164)	(0.258)	(0.292)	(0.290)
INZ	0.009	0.000	0.002	0.001	0.409	0.506	2.060
	(0.016)	(0.000)	(0.008)	(0.004)	(0.113)	(0.120)	(0.270)
Contains	0.250	0.716	0.316	0.718	0.310	0.207	0.211
	(0.043)	(0.045)	(0.046)	(0.045)	(0.046)	(0.041)	(0.041)
Sparsity	0.250	0.716	0.315	0.718	0.268	0.175	0.102
	(0.043)	(0.045)	(0.046)	(0.045)	(0.044)	(0.038)	(0.030)
Bias	14.479	4.261	7.982	4.259	8.063	9.065	8.963
	(0.154)	(0.256)	(0.406)	(0.256)	(0.409)	(0.419)	(0.411)
MSE	26.225	4.926	10.786	4.923	10.808	13.291	12.623
	(9.144)	(3.226)	(6.547)	(3.221)	(6.675)	(7.899)	(7.275)
ME	22.842	2.150	7.924	2.149	7.993	10.403	9.917
	(0.914)	(0.322)	(0.651)	(0.322)	(0.662)	(0.782)	(0.720)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{01} from (3), X from (7) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 14: Simulations for example 1 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	12.615	14.026	12.579	14.027	12.873	12.299	13.354
	(0.249)	(0.164)	(0.247)	(0.165)	(0.273)	(0.287)	(0.348)
CNZ	12.605	14.026	12.576	14.026	12.538	11.957	12.080
	(0.248)	(0.164)	(0.247)	(0.165)	(0.249)	(0.269)	(0.263)
INZ	0.010	0.000	0.004	0.001	0.336	0.342	1.273
	(0.018)	(0.003)	(0.010)	(0.005)	(0.103)	(0.100)	(0.211)
Contains	0.427	0.711	0.418	0.713	0.412	0.325	0.341
	(0.049)	(0.045)	(0.049)	(0.045)	(0.049)	(0.047)	(0.047)
Sparsity	0.425	0.711	0.417	0.713	0.366	0.288	0.223
	(0.049)	(0.045)	(0.049)	(0.045)	(0.048)	(0.045)	(0.042)
Bias	13.885	4.283	6.885	4.281	6.959	7.643	7.436
	(0.154)	(0.256)	(0.406)	(0.256)	(0.409)	(0.419)	(0.411)
MSE	21.693	4.956	8.480	4.952	8.550	9.917	9.289
	(9.144)	(3.226)	(6.547)	(3.221)	(6.675)	(7.899)	(7.275)
ME	18.358	2.179	5.676	2.177	5.786	7.097	6.585
	(0.914)	(0.322)	(0.651)	(0.322)	(0.662)	(0.782)	(0.720)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{01} from (3), X from (7) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 15: Simulations for example 2 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	11.802	14.208	12.267	14.196	13.290	13.542	15.936
	(0.276)	(0.147)	(0.254)	(0.147)	(0.317)	(0.348)	(0.465)
CNZ	11.772	14.208	12.261	14.190	12.174	11.286	11.304
	(0.274)	(0.147)	(0.253)	(0.147)	(0.250)	(0.282)	(0.287)
INZ	0.030	0.000	0.006	0.006	1.116	2.256	4.632
	(0.030)	(0.000)	(0.013)	(0.013)	(0.177)	(0.209)	(0.370)
Contains	0.302	0.758	0.354	0.751	0.332	0.234	0.245
	(0.046)	(0.043)	(0.048)	(0.043)	(0.047)	(0.042)	(0.043)
Sparsity	0.298	0.758	0.353	0.750	0.215	0.101	0.051
	(0.046)	(0.043)	(0.048)	(0.043)	(0.041)	(0.030)	(0.022)
Bias	6.486	2.305	3.026	2.325	3.058	5.230	4.976
	(0.105)	(0.149)	(0.162)	(0.150)	(0.163)	(0.226)	(0.215)
MSE	24.564	4.667	10.494	4.696	10.246	12.143	11.157
	(8.595)	(2.688)	(6.599)	(2.687)	(6.415)	(7.015)	(6.501)
ME	21.265	1.904	7.661	1.934	7.533	9.464	8.744
	(0.865)	(0.271)	(0.661)	(0.271)	(0.644)	(0.697)	(0.645)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{02} from (4), X from (7) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 16: Simulations for example 2 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	12.957	14.199	12.885	14.193	13.692	13.827	15.435
	(0.239)	(0.148)	(0.240)	(0.149)	(0.296)	(0.326)	(0.410)
CNZ	12.921	14.193	12.870	14.181	12.798	12.324	12.420
	(0.235)	(0.147)	(0.238)	(0.148)	(0.238)	(0.254)	(0.252)
INZ	0.036	0.006	0.015	0.012	0.894	1.503	3.015
	(0.035)	(0.013)	(0.025)	(0.019)	(0.161)	(0.193)	(0.317)
Contains	0.481	0.753	0.470	0.749	0.453	0.368	0.386
	(0.050)	(0.043)	(0.050)	(0.043)	(0.050)	(0.048)	(0.049)
Sparsity	0.473	0.751	0.468	0.746	0.319	0.214	0.136
	(0.050)	(0.043)	(0.050)	(0.044)	(0.047)	(0.041)	(0.034)
Bias	6.022	2.322	2.746	2.334	2.793	3.938	3.707
	(0.105)	(0.149)	(0.162)	(0.150)	(0.163)	(0.226)	(0.215)
MSE	20.657	4.700	8.080	4.719	8.064	8.960	8.302
	(8.595)	(2.688)	(6.599)	(2.687)	(6.415)	(7.015)	(6.501)
ME	17.384	1.935	5.306	1.954	5.400	6.300	5.832
	(0.865)	(0.271)	(0.661)	(0.271)	(0.644)	(0.697)	(0.645)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{02} from (4), X from (7) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 17: Simulations for example 3 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	12.135	14.250	12.624	14.256	13.020	12.645	14.742
	(0.274)	(0.146)	(0.245)	(0.144)	(0.275)	(0.321)	(0.449)
CNZ	12.117	14.250	12.618	14.253	12.651	11.682	11.730
	(0.272)	(0.146)	(0.245)	(0.144)	(0.244)	(0.291)	(0.287)
INZ	0.018	0.000	0.006	0.003	0.369	0.963	3.012
	(0.023)	(0.000)	(0.013)	(0.009)	(0.115)	(0.153)	(0.317)
Contains	0.362	0.772	0.422	0.771	0.427	0.309	0.312
	(0.048)	(0.042)	(0.049)	(0.042)	(0.049)	(0.046)	(0.046)
Sparsity	0.359	0.772	0.421	0.770	0.371	0.222	0.110
	(0.048)	(0.042)	(0.049)	(0.042)	(0.048)	(0.042)	(0.031)
Bias	4.801	1.700	2.008	1.693	2.018	3.551	3.601
	(0.091)	(0.132)	(0.133)	(0.133)	(0.130)	(0.191)	(0.184)
MSE	22.702	4.531	8.750	4.517	8.613	11.001	10.363
	(7.532)	(2.715)	(5.485)	(2.575)	(5.491)	(6.888)	(6.084)
ME	19.386	1.752	5.884	1.731	5.806	8.212	7.795
	(0.750)	(0.270)	(0.543)	(0.256)	(0.539)	(0.682)	(0.600)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{03} from (5), X from (7) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 18: Simulations for example 3 using X from (7)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	13.191	14.259	13.065	14.256	13.362	13.122	14.637
	(0.226)	(0.145)	(0.235)	(0.144)	(0.257)	(0.294)	(0.381)
CNZ	13.161	14.256	13.050	14.250	13.047	12.492	12.648
	(0.225)	(0.145)	(0.233)	(0.143)	(0.232)	(0.263)	(0.255)
INZ	0.030	0.003	0.015	0.006	0.315	0.630	1.989
	(0.030)	(0.009)	(0.021)	(0.013)	(0.105)	(0.131)	(0.263)
Contains	0.534	0.773	0.512	0.769	0.510	0.425	0.448
	(0.050)	(0.042)	(0.050)	(0.042)	(0.050)	(0.049)	(0.050)
Sparsity	0.530	0.772	0.509	0.767	0.455	0.342	0.229
	(0.050)	(0.042)	(0.050)	(0.042)	(0.050)	(0.047)	(0.042)
Bias	4.275	1.696	1.878	1.699	1.888	2.670	2.638
	(0.091)	(0.132)	(0.133)	(0.133)	(0.130)	(0.191)	(0.184)
MSE	18.637	4.519	7.155	4.520	7.121	8.360	7.667
	(7.532)	(2.715)	(5.485)	(2.575)	(5.491)	(6.888)	(6.084)
ME	15.368	1.742	4.353	1.733	4.353	5.576	5.047
	(0.750)	(0.270)	(0.543)	(0.256)	(0.539)	(0.682)	(0.600)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{03} from (5), X from (7) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 19: Simulations for example 1 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	19.403	15.304	15.636	16.899	27.668	20.064	33.604
	(0.587)	(0.139)	(0.223)	(0.237)	(0.604)	(0.475)	(0.672)
CNZ	14.424	14.972	14.833	14.969	14.775	14.128	14.015
	(0.122)	(0.029)	(0.070)	(0.031)	(0.082)	(0.146)	(0.156)
INZ	4.979	0.332	0.803	1.930	12.893	5.936	19.589
	(0.564)	(0.136)	(0.209)	(0.234)	(0.599)	(0.442)	(0.655)
Contains	0.814	0.991	0.945	0.990	0.927	0.724	0.695
	(0.039)	(0.010)	(0.023)	(0.010)	(0.026)	(0.045)	(0.046)
Sparsity	0.252	0.915	0.772	0.500	0.013	0.074	0.001
	(0.043)	(0.028)	(0.042)	(0.050)	(0.011)	(0.026)	(0.002)
Bias	7.568	7.259	7.469	7.464	8.460	8.253	9.113
	(0.198)	(0.186)	(0.220)	(0.188)	(0.237)	(0.232)	(0.248)
MSE	3.419	3.335	3.296	3.212	2.584	3.015	2.373
	(0.636)	(0.543)	(0.619)	(0.533)	(0.552)	(0.690)	(0.608)
ME	0.901	0.661	0.771	0.785	1.482	1.258	1.870
	(0.044)	(0.034)	(0.050)	(0.036)	(0.058)	(0.063)	(0.063)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{01} from (3), X from (6) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 20: Simulations for example 1 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	23.474	15.475	16.039	17.076	28.267	21.841	35.388
	(0.763)	(0.176)	(0.295)	(0.261)	(0.648)	(0.583)	(0.760)
CNZ	14.353	14.973	14.746	14.969	14.678	14.067	13.929
	(0.133)	(0.028)	(0.086)	(0.031)	(0.097)	(0.154)	(0.167)
INZ	9.121	0.502	1.292	2.107	13.590	7.774	21.459
	(0.743)	(0.174)	(0.280)	(0.260)	(0.641)	(0.554)	(0.742)
Contains	0.797	0.991	0.918	0.990	0.896	0.712	0.680
	(0.040)	(0.009)	(0.027)	(0.010)	(0.030)	(0.045)	(0.047)
Sparsity	0.092	0.883	0.670	0.481	0.011	0.054	0.001
	(0.029)	(0.032)	(0.047)	(0.050)	(0.010)	(0.023)	(0.003)
Bias	7.929	7.280	7.667	7.484	8.659	8.541	9.410
	(0.198)	(0.186)	(0.220)	(0.188)	(0.237)	(0.232)	(0.248)
MSE	3.258	3.313	3.264	3.190	2.543	2.873	2.249
	(0.636)	(0.543)	(0.619)	(0.533)	(0.552)	(0.690)	(0.608)
ME	1.002	0.684	0.856	0.808	1.567	1.418	2.017
	(0.044)	(0.034)	(0.050)	(0.036)	(0.058)	(0.063)	(0.063)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{01} from (3), X from (6) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 21: Simulations for example 2 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	21.324	15.444	16.023	17.373	29.913	21.534	35.787
	(0.697)	(0.166)	(0.290)	(0.286)	(0.616)	(0.567)	(0.688)
CNZ	14.517	14.994	14.841	14.988	14.718	14.190	13.950
	(0.115)	(0.013)	(0.070)	(0.019)	(0.095)	(0.144)	(0.169)
INZ	6.807	0.450	1.182	2.385	15.195	7.344	21.837
	(0.682)	(0.165)	(0.277)	(0.285)	(0.610)	(0.541)	(0.675)
Contains	0.845	0.998	0.949	0.996	0.914	0.746	0.690
	(0.036)	(0.004)	(0.022)	(0.006)	(0.028)	(0.044)	(0.046)
Sparsity	0.185	0.900	0.718	0.441	0.004	0.074	0.000
	(0.039)	(0.030)	(0.045)	(0.050)	(0.006)	(0.026)	(0.000)
Bias	4.684	4.202	4.290	5.789	6.950	7.145	8.121
	(0.155)	(0.136)	(0.142)	(0.225)	(0.248)	(0.244)	(0.282)
MSE	3.287	3.320	3.255	3.176	2.507	2.898	2.297
	(0.663)	(0.577)	(0.647)	(0.561)	(0.578)	(0.725)	(0.653)
ME	0.920	0.683	0.828	0.828	1.594	1.364	1.975
	(0.048)	(0.037)	(0.055)	(0.040)	(0.061)	(0.068)	(0.067)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{02} from (4), X from (6) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 22: Simulations for example 2 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	26.118	15.693	16.608	17.628	30.537	23.841	38.037
	(0.910)	(0.216)	(0.355)	(0.314)	(0.650)	(0.695)	(0.783)
CNZ	14.361	14.994	14.787	14.988	14.598	14.088	13.767
	(0.140)	(0.013)	(0.082)	(0.019)	(0.111)	(0.162)	(0.192)
INZ	11.757	0.699	1.821	2.640	15.939	9.753	24.270
	(0.894)	(0.215)	(0.343)	(0.313)	(0.644)	(0.671)	(0.772)
Contains	0.808	0.998	0.933	0.996	0.876	0.730	0.661
	(0.039)	(0.004)	(0.025)	(0.006)	(0.033)	(0.044)	(0.047)
Sparsity	0.047	0.856	0.616	0.418	0.002	0.044	0.000
	(0.021)	(0.035)	(0.049)	(0.049)	(0.004)	(0.021)	(0.000)
Bias	5.008	4.210	4.347	5.814	6.989	7.359	8.447
	(0.155)	(0.136)	(0.142)	(0.225)	(0.248)	(0.244)	(0.282)
MSE	3.098	3.292	3.206	3.151	2.469	2.731	2.159
	(0.663)	(0.577)	(0.647)	(0.561)	(0.578)	(0.725)	(0.653)
ME	1.053	0.711	0.913	0.853	1.682	1.556	2.159
	(0.048)	(0.037)	(0.055)	(0.040)	(0.061)	(0.068)	(0.067)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{02} from (4), X from (6) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 23: Simulations for example 3 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	19.554	15.327	15.720	17.004	27.879	20.301	33.864
	(0.588)	(0.147)	(0.224)	(0.250)	(0.608)	(0.493)	(0.656)
CNZ	14.445	14.976	14.865	14.976	14.850	14.196	14.109
	(0.120)	(0.027)	(0.062)	(0.027)	(0.067)	(0.140)	(0.145)
INZ	5.109	0.351	0.855	2.028	13.029	6.105	19.755
	(0.567)	(0.145)	(0.214)	(0.248)	(0.605)	(0.458)	(0.644)
Contains	0.819	0.992	0.955	0.992	0.951	0.741	0.715
	(0.039)	(0.009)	(0.021)	(0.009)	(0.022)	(0.044)	(0.045)
Sparsity	0.254	0.917	0.773	0.501	0.014	0.075	0.000
	(0.044)	(0.028)	(0.042)	(0.050)	(0.012)	(0.026)	(0.000)
Bias	3.425	2.896	2.957	4.231	6.024	5.510	7.338
	(0.118)	(0.108)	(0.114)	(0.201)	(0.267)	(0.219)	(0.288)
MSE	3.404	3.334	3.285	3.203	2.551	2.993	2.354
	(0.633)	(0.575)	(0.630)	(0.548)	(0.556)	(0.710)	(0.622)
ME	0.904	0.673	0.780	0.802	1.498	1.271	1.872
	(0.044)	(0.036)	(0.050)	(0.038)	(0.057)	(0.066)	(0.064)

This table shows the performance of the pag-lasso relative to the aglasso using the group lasso as the initial estimator. The data is generated using β_{03} from (5), X from (6) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 24: Simulations for example 3 using X from (6)

	aglasso	pag-lasso					
		g1-s1	g1-s2	g2-s3	g2-s4	g3-s5	g3-s6
nVAR	24.192	15.552	16.149	17.220	28.572	22.239	35.784
	(0.793)	(0.220)	(0.307)	(0.284)	(0.656)	(0.619)	(0.756)
CNZ	14.430	14.976	14.799	14.976	14.772	14.127	14.004
	(0.125)	(0.027)	(0.075)	(0.027)	(0.081)	(0.147)	(0.162)
INZ	9.762	0.576	1.350	2.244	13.800	8.112	21.780
	(0.780)	(0.220)	(0.296)	(0.282)	(0.652)	(0.592)	(0.742)
Contains	0.820	0.992	0.933	0.992	0.925	0.725	0.701
	(0.038)	(0.009)	(0.025)	(0.009)	(0.026)	(0.045)	(0.046)
Sparsity	0.097	0.887	0.676	0.480	0.013	0.057	0.000
	(0.030)	(0.032)	(0.047)	(0.050)	(0.011)	(0.023)	(0.000)
Bias	3.556	2.907	3.001	4.232	6.009	5.639	7.442
	(0.118)	(0.108)	(0.114)	(0.201)	(0.267)	(0.219)	(0.288)
MSE	3.214	3.309	3.244	3.176	2.499	2.836	2.209
	(0.633)	(0.575)	(0.630)	(0.548)	(0.556)	(0.710)	(0.622)
ME	1.009	0.697	0.856	0.830	1.585	1.445	2.048
	(0.044)	(0.036)	(0.050)	(0.038)	(0.057)	(0.066)	(0.064)

This table shows the performance of the pag-lasso relative to the aglasso using the lasso as the initial estimator. The data is generated using β_{03} from (5), X from (6) and $\eta = 10$ in equation (1). The group names for the pag-lasso columns refers to the prior variable sets given on page 15. The metrics used to evaluate the performance of the models are listed in the first column and explained on page 13. The numbers show the average score by the models for each corresponding metric, and the numbers in parenthesis are the standard error of a sample average.

Table 25: Factor Zoo

ID	Description	Year.pub	Year.end	Avg.Ret.	S.R.	Reference
1	Excess Market Return	1972	1965	0.64%	50.6%	(Black et al., 1972)
2	Market Beta	1973	1968	-0.08%	-5.4%	(Fama and MacBeth, 1973)
3	Earnings to price	1977	1971	0.28%	29.7%	(Basu, 1977)
4	Dividend to price	1979	1977	0.01%	0.6%	(Litzenberger and Ramaswamy, 1979)
5	Unexpected quarterly earnings	1982	1980	0.12%	26.3%	(Rendleman et al., 1982)
6	Share price	1982	1978	0.02%	2.2%	(Miller and Scholes, 1982)
7	Long-Term Reversal	1985	1982	0.34%	36.3%	(De Bondt and Thaler, 1985)
8	Leverage	1988	1981	0.21%	24.3%	(Bhandari, 1988)
9	Cash flow to debt	1989	1984	-0.09%	-17.0%	(Ou and Penman, 1989)
10	Current ratio	1989	1984	0.06%	7.7%	(Ou and Penman, 1989)
11	% change in current ratio	1989	1984	0.00%	0.5%	(Ou and Penman, 1989)
12	% change in quick ratio	1989	1984	-0.04%	-11.9%	(Ou and Penman, 1989)
13	% change sales-to-inventory	1989	1984	0.17%	46.2%	(Ou and Penman, 1989)
14	Quick ratio	1989	1984	-0.02%	-2.9%	(Ou and Penman, 1989)
15	Sales to cash	1989	1984	0.01%	1.5%	(Ou and Penman, 1989)
16	Sales to inventory	1989	1984	0.09%	16.1%	(Ou and Penman, 1989)
17	Sales to receivables	1989	1984	0.14%	22.8%	(Ou and Penman, 1989)
18	Bid-ask spread	1989	1979	-0.04%	-3.3%	(Amihud and Mendelson, 1989)
19	Depreciation / PP&E	1992	1988	0.11%	12.1%	(Holthausen and Larcker, 1992)
20	% change in depreciation	1992	1988	0.08%	23.1%	(Holthausen and Larcker, 1992)
21	Small Minus Big	1993	1991	0.21%	24.5%	(Fama and French, 1993)
22	High Minus Low	1993	1991	0.28%	34.3%	(Fama and French, 1993)
23	Short-Term Reversal	1993	1989	0.15%	21.7%	(Jegadeesh and Titman, 1993)
24	6-month momentum	1993	1989	0.21%	27.8%	(Jegadeesh and Titman, 1993)
25	36-month momentum	1993	1989	0.09%	13.4%	(Jegadeesh and Titman, 1993)
26	Sales growth	1994	1990	0.04%	5.8%	(Lakonishok et al., 1994)
27	Cash flow-to-price	1994	1990	0.31%	32.5%	(Lakonishok et al., 1994)
28	New equity issue	1995	1990	0.10%	8.7%	(Loughran and Ritter, 1995)
29	Dividend initiation	1995	1988	-0.03%	-3.4%	(Michaely et al., 1995)
30	Dividend omission	1995	1988	-0.18%	-18.0%	(Michaely et al., 1995)
31	Working capital accruals	1996	1991	0.22%	46.0%	(Sloan, 1996)

Continued on next page

Table 25 – continued from previous page

ID	Description	Year.pub	Year.end	Avg.Ret.	S.R.	Reference
32	Sales to price	1996	1991	0.35%	41.8%	(Barbee et al., 1996)
33	Capital turnover	1996	1993	-0.11%	-16.6%	(Haugen and Baker, 1996)
34	Momentum	1997	1993	0.63%	50.2%	(Carhart, 1997)
35	Share turnover	1998	1991	-0.02%	-2.1%	(Datar et al., 1998)
36	% change in gross margin - % change in sales	1998	1988	-0.05%	-12.4%	(Abarbanell and Bushee, 1998)
37	% change in sales - % change in inventory	1998	1988	0.14%	42.1%	(Abarbanell and Bushee, 1998)
38	% change in sales - % change in A/R	1998	1988	0.14%	43.5%	(Abarbanell and Bushee, 1998)
39	% change in sales - % change in SG&A	1998	1988	0.09%	19.6%	(Abarbanell and Bushee, 1998)
40	Effective Tax Rate	1998	1988	-0.04%	-9.1%	(Abarbanell and Bushee, 1998)
41	Labor Force Efficiency	1998	1988	-0.03%	-8.5%	(Abarbanell and Bushee, 1998)
42	Ohlson's O-score	1998	1995	0.05%	9.3%	(Dichev, 1998)
43	Altman's Z-score	1998	1995	0.20%	22.1%	(Dichev, 1998)
44	Industry adjusted % change in CAPEX	1998	1988	0.10%	20.5%	(Abarbanell and Bushee, 1998)
45	Number of earnings increases	1999	1992	0.01%	2.8%	(Barth, Elliott, and Finn, 1999)
46	Industry momentum	1999	1995	0.01%	1.4%	(Moskowitz and Grinblatt, 1999)
47	Financial statements score	2000	1996	0.08%	18.4%	(Piotroski, 2000)
48	Industry-adjusted book to market	2000	1998	0.22%	38.0%	(Asness et al., 2000)
49	Industry-adjusted cash flow to price ratio	2000	1998	0.26%	52.1%	(Asness et al., 2000)
50	Industry-adjusted change in employees	2000	1998	-0.01%	-1.5%	(Asness et al., 2000)
51	Industry-adjusted size	2000	1998	0.36%	36.3%	(Asness et al., 2000)
52	Dollar trading volume	2001	1995	0.38%	35.8%	(Chordia et al., 2001)
53	Volatility of liquidity (dollar trading volume)	2001	1995	0.20%	38.8%	(Chordia et al., 2001)
54	Volatility of liquidity (share turnover)	2001	1995	0.02%	2.1%	(Chordia et al., 2001)
55	Advertising Expense-to-market	2001	1995	-0.13%	-15.6%	(Chan et al., 2001)
56	R&D Expense-to-market	2001	1995	0.34%	36.2%	(Chan et al., 2001)
57	R&D-to-sales	2001	1995	0.06%	5.5%	(Chan et al., 2001)
58	Kaplan-Zingales Index	2001	1997	0.22%	25.3%	(Lamont et al., 2001)
59	Change in inventory	2002	1997	0.18%	40.7%	(Thomas and Zhang, 2002)
60	Change in tax expense	2002	1997	0.09%	18.0%	(Thomas and Zhang, 2002)
61	Illiquidity	2002	1997	0.34%	28.6%	(Amihud, 2002)
62	Liquidity	2003	2000	0.38%	38.6%	(Pastor and Stambaugh, 2003)
63	Idiosyncratic return volatility	2003	1997	0.07%	5.1%	(Ali et al., 2003)
64	Growth in long term net operating assets	2003	1993	0.22%	51.8%	(Fairfield et al., 2003)

Continued on next page

Table 25 – continued from previous page

ID	Description	Year.pub	Year.end	Avg.Ret.	S.R.	Reference
65	Order backlog	2003	1999	0.05%	5.7%	(Rajgopal et al., 2003)
66	Changes in Long-term Net Operating Assets	2003	1993	0.24%	56.0%	(Fairfield et al., 2003)
67	Cash flow to price ratio	2004	1997	0.27%	31.7%	(Desai et al., 2004)
68	R&D increase	2004	2001	0.06%	11.1%	(Eberhart et al., 2004)
69	Corporate investment	2004	1995	0.13%	36.4%	(Titman, Wei, and Xie, 2004)
70	Earnings volatility	2004	2001	0.10%	10.7%	(Francis et al., 2004)
71	Abnormal Corporate Investment	2004	1995	0.13%	31.2%	(Titman et al., 2004)
72	Net Operating Assets	2004	2002	0.31%	66.6%	(Hirshleifer et al., 2004)
73	Changes in Net Operating Assets	2004	2002	0.14%	41.6%	(Hirshleifer et al., 2004)
74	Tax income to book income	2004	2000	0.14%	28.3%	(Lev and Nissim, 2004)
75	Price delay	2005	2001	0.07%	16.8%	(Hou and Moskowitz, 2005)
76	# Years since first Compustat coverage	2005	2001	0.01%	1.1%	(Jiang, Lee, and Zhang, 2005)
77	Growth in common shareholder equity	2005	2001	0.15%	27.6%	(Richardson et al., 2005)
78	Growth in long-term debt	2005	2001	0.06%	13.3%	(Richardson et al., 2005)
79	Change in Current Operating Assets	2005	2001	0.19%	34.6%	(Richardson et al., 2005)
80	Change in Current Operating Liabilities	2005	2001	0.03%	6.3%	(Richardson et al., 2005)
81	Changes in Net Non-cash Working Capital	2005	2001	0.11%	25.2%	(Richardson et al., 2005)
82	Change in Non-current Operating Assets	2005	2001	0.21%	44.5%	(Richardson et al., 2005)
83	Change in Non-current Operating Liabilities	2005	2001	0.04%	9.6%	(Richardson et al., 2005)
84	Change in Net Non-current Operating Assets	2005	2001	0.23%	35.4%	(Richardson et al., 2005)
85	Change in Net Financial Assets	2005	2001	0.23%	59.0%	(Richardson et al., 2005)
86	Total accruals	2005	2001	0.19%	44.8%	(Richardson et al., 2005)
87	Change in Short- term Investments	2005	2001	-0.03%	-8.3%	(Richardson et al., 2005)
88	Change in Financial Liabilities	2005	2001	0.18%	56.1%	(Richardson et al., 2005)
89	Change in Book Equity	2005	2001	0.17%	30.0%	(Richardson et al., 2005)
90	Financial statements performance	2005	2001	0.17%	37.1%	(Mohanram, 2005)
91	Change in 6-month momentum	2006	2006	0.21%	29.8%	(Gettleman and Marks, 2006)
92	Growth in capital expenditures	2006	1999	0.14%	30.4%	(Anderson and Garcia-Feijóo, 2006)
93	Return volatility	2006	2000	-0.02%	-1.7%	(Ang et al., 2006)
94	Zero trading days	2006	2003	-0.05%	-4.4%	(Liu, 2006)
95	Three-year Investment Growth	2006	1999	0.11%	23.6%	(Anderson and Garcia-Feijóo, 2006)
96	Composite Equity Issuance	2006	2003	-0.01%	-2.2%	(Daniel and Titman, 2006)
97	Net equity finance	2006	2000	0.08%	9.7%	(Bradshaw et al., 2006)

Continued on next page

Table 25 – continued from previous page

ID	Description	Year.pub	Year.end	Avg.Ret.	S.R.	Reference
98	Net debt finance	2006	2000	0.17%	48.3%	(Bradshaw et al., 2006)
99	Net external finance	2006	2000	0.22%	38.6%	(Bradshaw et al., 2006)
100	Revenue Surprises	2006	2003	0.05%	9.0%	(Jegadeesh and Livnat, 2006)
101	Industry Concentration	2006	2001	0.03%	3.8%	(Hou and Robinson, 2006)
102	Whited-Wu Index	2006	2001	-0.02%	-2.6%	(Whited and Wu, 2006)
103	Return on invested capital	2007	2005	0.18%	29.3%	(Brown and Rowe, 2007)
104	Debt capacity/firm tangibility	2007	2000	0.05%	7.1%	(Almeida and Campello, 2007)
105	Payout yield	2007	2003	0.16%	17.5%	(Boudoukh et al., 2007)
106	Net payout yield	2007	2003	0.16%	17.2%	(Boudoukh et al., 2007)
107	Net debt-to-price	2007	1950	0.02%	2.5%	(Penman et al., 2007)
108	Enterprise book-to-price	2007	2001	0.14%	14.7%	(Penman et al., 2007)
109	Change in shares outstanding	2008	1969	0.24%	36.1%	(Pontiff and Woodgate, 2008)
110	Abnormal earnings announcement volume	2008	2006	-0.08%	-17.0%	(Lerman et al., 2008)
111	Earnings announcement return	2008	2004	0.02%	6.8%	(Brandt et al., 2008)
112	Seasonality	2008	2002	0.16%	17.3%	(Heston and Sadka, 2008)
113	Changes in PPE and Inventory-to-assets	2008	2005	0.19%	42.0%	(Lyandres, Sun, and Zhang, 2008)
114	Investment Growth	2008	2003	0.17%	39.5%	(Xing, 2008)
115	Composite Debt Issuance	2008	2005	0.08%	21.6%	(Lyandres et al., 2008)
116	Return on net operating assets	2008	2002	0.09%	8.6%	(Soliman, 2008)
117	Profit margin	2008	2002	0.02%	4.4%	(Soliman, 2008)
118	Asset turnover	2008	2002	0.06%	6.7%	(Soliman, 2008)
119	Industry-adjusted change in asset turnover	2008	2002	0.14%	41.1%	(Soliman, 2008)
120	Industry-adjusted change in profit margin	2008	2002	-0.01%	-3.2%	(Soliman, 2008)
121	Cash productivity	2009	2009	0.27%	37.6%	(Chandrashekar and Rao, 2009)
122	Sin stocks	2009	2006	0.44%	41.6%	(Hong and Kacperczyk, 2009)
123	Revenue surprise	2009	2005	0.12%	19.3%	(Kama, 2009)
124	Cash flow volatility	2009	2008	0.20%	26.6%	(Huang, 2009)
125	Absolute accruals	2010	2008	-0.05%	-8.6%	(Bandyopadhyay et al., 2010)
126	Capital expenditures and inventory	2010	2006	0.19%	42.8%	(Chen and Zhang, 2010)
127	Return on assets	2010	2005	-0.09%	-13.9%	(Balakrishnan et al., 2010)
128	Accrual volatility	2010	2008	0.19%	26.6%	(Bandyopadhyay et al., 2010)
129	Industry-adjusted Real Estate Ratio	2010	2005	0.11%	17.3%	(Tuzel, 2010)
130	Percent accruals	2011	2008	0.16%	35.0%	(Hafzalla et al., 2011)

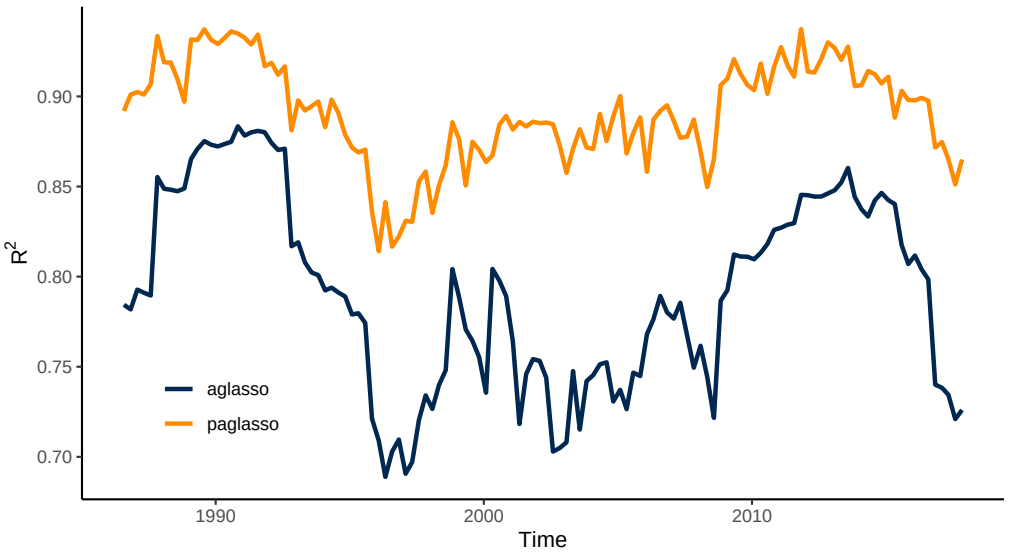
Continued on next page

Table 25 – continued from previous page

ID	Description	Year.pub	Year.end	Avg.Ret.	S.R.	Reference
131	Maximum daily return	2011	2005	0.00%	-0.3%	(Bali, Cakici, and Whitelaw, 2011)
132	Operating Leverage	2011	2008	0.20%	32.8%	(Novy-Marx, 2011)
133	Inventory Growth	2011	2009	0.13%	30.1%	(Belo and Lin, 2012)
134	Percent Operating Accruals	2011	2008	0.15%	28.9%	(Hafzalla et al., 2011)
135	Enterprise multiple	2011	2009	0.11%	17.6%	(Loughran and Wellman, 2011)
136	Cash holdings	2012	2009	0.13%	15.3%	(Palazzo, 2012)
137	HML Devil	2013	2011	0.23%	22.6%	(Asness and Frazzini, 2013)
138	Gross profitability	2013	2010	0.15%	22.5%	(Novy-Marx, 2013)
139	Organizational Capital	2013	2008	0.21%	31.9%	(Eisfeldt and Papanikolaou, 2013)
140	Betting Against Beta	2014	2012	0.91%	92.8%	(Frazzini and Pedersen, 2014)
141	Quality Minus Junk	2014	2012	0.43%	60.1%	(Asness et al., 2019)
142	Employee growth rate	2014	2010	0.08%	12.9%	(Belo et al., 2014)
143	Growth in advertising expense	2014	2010	0.07%	13.0%	(Lou, 2014)
144	Book Asset Liquidity	2014	2006	0.09%	12.3%	(Ortiz-Molina and Phillips, 2014)
145	Robust Minus Weak	2015	2013	0.34%	49.8%	(Fama and French, 2015)
146	Conservative Minus Aggressive	2015	2013	0.26%	46.8%	(Fama and French, 2015)
147	HXZ Investment	2015	2012	0.34%	64.7%	(Hou et al., 2015)
148	HXZ Profitability	2015	2012	0.57%	77.5%	(Hou et al., 2015)
149	Intermediary Investment	2016	2012			(He et al., 2017)
150	Convertible debt indicator	2016	2012	0.11%	26.4%	(Valta, 2016)

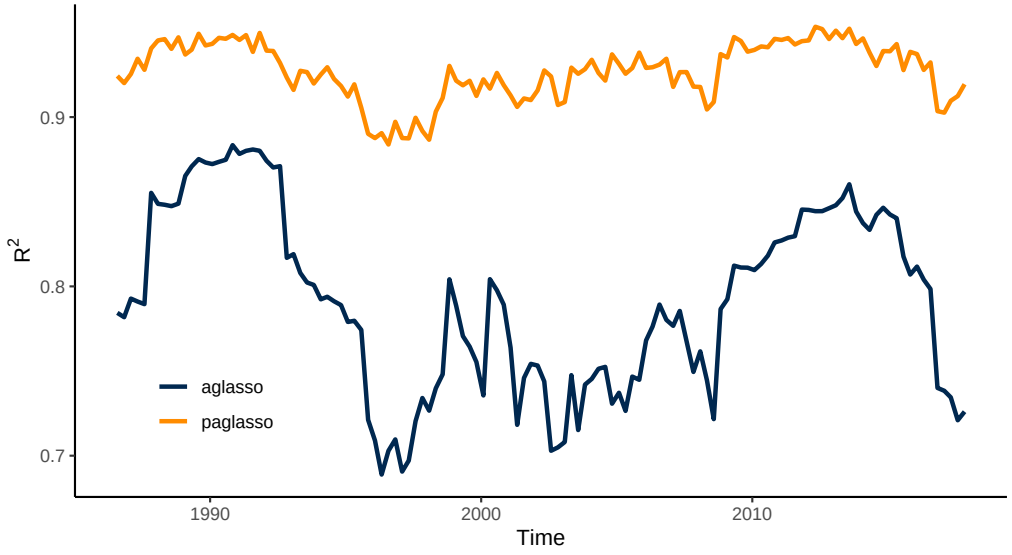
B Figures

Figure 21: R^2 with 10 factors in the prior set



This figure shows the R^2 across all sample windows for the aglasso and the pag-lasso with up to 10 factors in the prior set.

Figure 22: R^2 with 20 factors in the prior set



This figure shows the R^2 across all sample windows for the aglasso and the pag-lasso with up to 20 factors in the prior set.

C Proof of theorem 3.1

Proof. We need to show that there exists a global maximizer, $\hat{\beta}$, for $Q_n\left(\beta; X, Y, Y^p\right)$ in (1). That is, for any given $\epsilon > 0$, there exists a large constant C, such that

$$\mathbb{P}\left[\sup_{\|u\|=C} Q_n\left(\beta_0 + \alpha_n u; X, Y, Y^p\right) < Q_n\left(\beta_0; X, Y, Y^p\right)\right] \geq 1 - \epsilon,$$

Where $\alpha_n \triangleq \sqrt{p_n/n}$. This implies that there exist a local maximum in the ball $\{\beta_0 + \alpha_n u : \|u\| \leq C\}$, with $u = \left(u_1^T, \ldots, u_{p_n}^T\right)^T$ being a $\left(\sum_{j=1}^{p_n} d_j\right) \times 1$ vector, with probability at least $1 - \epsilon$. Then there exist a local maximizer, $\hat{\beta}$, where $\left\|\hat{\beta} - \beta_0\right\| = O_P\left(\alpha_n\right)$. Then by the concavity of $Q_n\left(\beta; X, Y, Y^p\right)$, $\hat{\beta}$ is also the global maximizer.

We have that

$$\begin{aligned}
&Q_n\left(\beta_0 + \alpha_n u; X, Y, Y^p\right) - Q_n\left(\beta_0; X, Y, Y^p\right) \\
&\quad = S_n\left(\beta_0 + \alpha_n u\right) - S_n\left(\beta_0\right) + \lambda_n \sum_{j=1}^{p_n} \left\|\tilde{\beta}_j\right\|^{-1} \left(\left\|\beta_{0,j} + \alpha_n u_j\right\| - \left\|\beta_{0,j}\right\|\right) \\
&\quad \triangleq T_{n1} + T_{n2},
\end{aligned}$$

where $T_{n1} = S_n\left(\beta_0 + \alpha_n u\right) - S_n\left(\beta_0\right)$, $T_{n2} = \lambda_n \sum_{j=1}^{p_n} \left\|\tilde{\beta}_j\right\|^{-1} \left(\left\|\beta_{0,j} + \alpha_n u_j\right\| - \left\|\beta_{0,j}\right\|\right)$, and $S_n\left(\beta\right) = -\frac{1}{2}\left\|Y - X\beta\right\|_2^2 + \frac{\eta}{2}\left\|Y^p - X\beta\right\|_2^2$.

Then we perform a third order Taylor expansion of T_{n1} around $u = 0$ which yields,

$$\begin{aligned}
T_{n1} &= \alpha_n \left(\frac{\partial S_n\left(\beta_0\right)}{\partial \beta}\right)^T u + \frac{1}{2}\alpha_n^2 u^T \frac{\partial^2 S_n\left(\beta_0\right)}{\partial \beta \partial \beta^T} u + \frac{1}{6}\alpha_n^3 u^T \nabla^2 \left(\left(\frac{\partial S_n\left(\beta^*\right)}{\partial \beta}\right)^T u\right) u \\
&\triangleq J_{n1} + J_{n2} + J_{n3}.
\end{aligned}$$

where β^* is between β_0 and $\beta_0 + \alpha_n u$, thus encompassing the approximation error.

Then following Wang and Tian (2019)

$$
\begin{aligned}
J_{n1} &= \alpha_n \left(\frac{1}{n} \sum_{i=1}^{n} x_i^T \left(y_i - \phi' \left(x_i^T \beta_0 \right) \right) + \frac{\eta}{n} \sum_{i=1}^{n} x_i^T \left(y_i^P - \phi' \left(x_i^T \beta_0 \right) \right) \right) u \\
&= \alpha_n \left(\frac{1}{n} \sum_{i=1}^{n} x_i^T \left(y_i - \phi' \left(x_i^T \beta_0 \right) \right) \right) u + \alpha_n \left(\frac{\eta}{n} \sum_{i=1}^{n} x_i^T \left(y_i^P - \phi' \left(x_i^T \beta_0 \right) \right) \right) u \\
|J_{n1}| &\leq \alpha_n^2 \left\| u \right\|_2 O_P(1) + \alpha_n \left\| \frac{\eta}{n} X^T \left(y^P - \phi' \left(X \beta_0 \right) \right) \right\|_2 \left\| u \right\|_2 .
\end{aligned}
$$

We can show that for a large $M > 0$ the following probability goes to zero

$$
\begin{aligned}
&\mathbb{P} \left[\left\| \frac{\eta}{n} X^T \left(y^P - \phi' \left(X \beta_0 \right) \right) \right\|_2 \geq M \left(\frac{p_n}{n} \right)^{1/2} \right] \\
&\leq \frac{n}{M^2 p_n} \mathbb{E} \left[\left\| \frac{\eta}{n} X^T \left(y^P - \phi' \left(X \beta_0 \right) \right) \right\|_2^2 \right] \\
&= \frac{n}{M^2 p_n} \mathbb{E} \left[\sum_{j=1}^{p_n} \sum_{k=1}^{d_j} \left(\frac{\eta}{n} \sum_{i=1}^{n} x_{ijk} \left(y_i^P - \phi' \left(x_i^T \beta_0 \right) \right) \right)^2 \right] \\
&= \frac{\eta^2}{M^2 p_n} \sum_{j=1}^{p_n} \sum_{k=1}^{d_j} \frac{1}{n} \mathbb{E} \left[\left(\sum_{i=1}^{n} x_{ijk} \left(y_i^P - \phi' \left(x_i^T \beta_0 \right) \right) \right)^2 \right].
\end{aligned}
$$

Then by condition (A7) we have that $\mathbb{E} \left[x_{ijk} x_{ljk} \left(y_i^P - \phi' \left(x_i^T \beta_o \right) \right) \left(y_l^P - \phi' \left(x_l^T \beta_0 \right) \right) \right] = 0$, and then

$$
\begin{aligned}
&\mathbb{P} \left[\left\| \frac{\eta}{n} X^T \left(y^P - \phi' \left(X \beta_0 \right) \right) \right\|_2 \geq M \left(\frac{p_n}{n} \right)^{1/2} \right] \\
&\leq \frac{\eta^2}{M^2 p_n} \sum_{j=1}^{p_n} \sum_{k=1}^{d_j} \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^{n} x_{ijk} \left(y_i^P - \phi' \left(x_i^T \beta_0 \right) \right)^2 \right] \\
&= \frac{1}{M^2 p_n} O_P(p_n) \to 0 \text{ for } M \to \infty.
\end{aligned}
$$

Hence,

$$
\alpha_n \left\| \frac{\eta}{n} X^T \left(y^P - \phi' \left(X \beta_0 \right) \right) \right\|_2 \left\| u \right\|_2 = \alpha_n O_p \left(\left(\frac{p_n}{n} \right)^{\frac{1}{2}} \right) \left\| u \right\|_2 = \alpha^2 \left\| u \right\|_2 O_p(1).
$$

Then we can bound $|J_{n1}|$ by

$$|J_{n1}| \leq \alpha_n^2 \|u\|_2 O_P(1).$$

The remainder of the proof follows the proof of Theorem 2.1 in Wang and Tian (2019) trivially. Hence, we have that by choosing a sufficiently large C all terms in T_{n1} and T_{n2} are dominated in size by J_{n2}, which is negative. This proves the theorem. □

D Proof of theorem 3.2

Proof. To prove this theorem we need to show that $\hat{\beta}_*$ satisfies the KKT conditions for the objective function, $Q_n\left(\beta; X, Y, Y^p\right)$, in (1) with probability approaching 1 for high n. For this, we need to show that

$$\mathbb{P}\left[\forall j \notin \{1,\ldots,k_n\}, \left\|\frac{\partial S_n\left(\hat{\beta}_*\right)}{\partial \beta_j}\right\|_2 \leq \lambda_n \left\|\tilde{\beta}_j\right\|_2^{-1}\right] \to 1,$$

which is equivalent to showing that

$$\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\frac{\partial S_n\left(\hat{\beta}_*\right)}{\partial \beta_j}\right\|_2 > \lambda_n \left\|\tilde{\beta}_j\right\|_2^{-1}\right] \to 0. \tag{9}$$

Denote $B_{nj}\left(\beta\right) = n\frac{\partial S_n(\beta)}{\partial \beta_j}$, and let $\nabla_1 B_{nj}\left(\beta\right)$ and $\nabla_1^2 B_{nj}\left(\beta\right)$ denote the first and second order partial derivatives of $B_{nj}\left(\beta\right)$ with respect to $\beta^{(1)}$. By doing a second order Taylor expansion of $B_{nj}\left(\beta\right)$ around β_0 we get

$$B_{nj}\left(\hat{\beta}_*\right) = B_{nj}\left(\beta_0\right) + \nabla_1 B_{nj}\left(\beta_0\right)^T \left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right) + \frac{1}{2}\left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right)^T \nabla_1^2 B_{nj}\left(\beta_n^*\right)\left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right), \tag{10}$$

where β_n^* is between $\hat{\beta}_*$ and β_0.

For $j \notin \{1,\ldots,k_n\}$ we have that $\left\|\tilde{\beta}_j\right\|_2 = O_P\left(\left(p_n/n\right)^{1/2}\right)$. Hence, for some large constant $L > 0$ we can rewrite (9).

$$\begin{aligned}
&\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\frac{\partial S_n\left(\hat{\beta}_*\right)}{\partial \beta_j}\right\|_2 > \lambda_n \left\|\tilde{\beta}_j\right\|_2^{-1}\right] \\
\leq &\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|B_{nj}\left(\hat{\beta}_*\right)\right\|_2 > n\lambda_n \left\|\tilde{\beta}_j\right\|_2^{-1}, \left\|\tilde{\beta}_j\right\|_2 < L\left(\frac{p_n}{n}\right)^{\frac{1}{2}}\right] \\
&+ \mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\tilde{\beta}_j\right\|_2 \geq L\left(\frac{p_n}{n}\right)^{\frac{1}{2}}\right] \\
\leq &\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|B_{nj}\left(\beta_0\right)\right\|_2 > \frac{\lambda_n n^{3/2}}{3L p_n^{1/2}}\right]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\nabla_1 B_{nj}\left(\beta_0\right)^T\left(\hat{\beta}_*^{(1)}-\beta_0^{(1)}\right)\right\|_2 > \frac{\lambda_n n^{3/2}}{3Lp_n^{1/2}}\right] \\
&+\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\frac{1}{2}\left(\hat{\beta}_*^{(1)}-\beta_0^{(1)}\right)^T \nabla_1^2 B_{nj}\left(\beta_n^*\right)\left(\hat{\beta}_*^{(1)}-\beta_0^{(1)}\right)\right\|_2 > \frac{\lambda_n n^{3/2}}{3Lp_n^{1/2}}\right] + O(1) \\
&\overset{\Delta}{=} U_1 + U_2 + U_3 + O(1).
\end{aligned}
$$

We can rewrite U_1 by bounding it by the individual probabilities and using the Markov inequality

$$
\begin{aligned}
U_1 &\leq \sum_{j=k_n+1}^{p_n} \mathbb{P}\left[\left\|B_{nj}\left(\beta_0\right)\right\|_2 > \frac{\lambda_n n^{3/2}}{3Lp_n^{1/2}}\right] \\
&\leq \sum_{j=k_n+1}^{p_n} \frac{9L^2 p_n}{\lambda_n^2 n^3}\mathbb{E}\left[\left\|B_{nj}\left(\beta_0\right)\right\|_2^2\right] \\
&\leq \left(p_n - k_n - 1\right)\frac{9L^2 p_n}{\lambda_n^2 n^3}\max_{j>k_n}\mathbb{E}\left[\left\|\sum_{i=1}^n x_{ij}\left(y_i - \phi'\left(x_i^T\beta_0\right)\right) + \eta\sum_{i=1}^n x_{ij}\left(y_i^p - \phi'\left(x_i^T\beta_0\right)\right)\right\|_2^2\right] \\
&\leq \frac{9L^2 p_n^2}{\lambda_n^2 n^3}\max_{j>k_n}\mathbb{E}\left[\sum_{k=1}^{d_j}\left(\sum_{i=1}^n\left(x_{ijk}\left(y_i - \phi'\left(x_i^T\beta_0\right)\right) + \eta x_{ijk}\left(y_i^p - \phi'\left(x_i^T\beta_0\right)\right)\right)\right)^2\right] \\
&= \frac{9L^2 p_n^2}{\lambda_n^2 n^3}\max_{j>k_n}\mathbb{E}\left[\sum_{i=1}^n\sum_{k=1}^{d_j} x_{ijk}^2\left(\varepsilon_i + \eta\varepsilon_i^p\right)^2\right] \\
&= \frac{9L^2 p_n^2}{\lambda_n^2 n^3}\max_{j>k_n}\sum_{i=1}^n\sum_{k=1}^{d_j}\mathbb{E}\left[\left(\varepsilon_i + \eta\varepsilon_i^p\right)^2\right]\mathbb{E}\left[x_{ijk}^2\right] \\
&\leq \frac{9L^2 p_n^2}{\lambda_n^2 n^2} CM \to 0.
\end{aligned}
$$

The proof that $U_2 = o(1)$ and that $U_3 \to 0$ follow the proof of theorem 2.2 of Wang and Tian (2019). □

E Proof of theorem 3.3

Proof. In order to prove sparsity we need to prove that

$$\mathbb{P}\left[\min_{j\leq k_n}\left\|\hat{\beta}_j\right\|_2 > 0\right] \to 1.$$

Using assumption (A6) and theorem 3.1 we have

$$\begin{aligned}\min_{j\leq k_n}\left\|\hat{\beta}_j\right\|_2 \geq& \theta_1 - \left\|\hat{\beta}^{(1)} - \beta_0^{(1)}\right\|_2 \\ \geq& Mn^{(c_2-1)/2} - O_p\left(\left(\frac{p_n}{n}\right)^{\frac{1}{2}}\right) \\ \geq& Mn^{(c_2-1)/2} - O_p\left(n^{(c_1-1)/2}\right) > 0.\end{aligned}$$

We will prove the limiting distribution using the Lindeberg-Feller central limit theorem. Per the definition of $\hat{\beta}$, we have that

$$\frac{\partial Q_n\left(\hat{\beta}\right)}{\partial \beta^{(1)}} = \frac{\partial S_n\left(\hat{\beta}\right)}{\partial \beta^{(1)}} - \lambda_n D_n = 0, \tag{11}$$

where we define $S_n(\beta) = \frac{1}{n}\sum_{i=1}^n \left(y_i x_i^T\beta - \phi\left(x_i^T\beta\right) + \eta\left(y_i^p x_i^T\beta - \phi\left(x_i^T\beta\right)\right)\right)$ and $D_n = \left(\left\|\tilde{\beta}_j\right\|_2^{-1}\operatorname{sgn}\left(\hat{\beta}_j\right)^T, j = 1,\ldots,k_n\right)^T$. We perform a second order Taylor expansion of $\frac{\partial S_n(\hat{\beta})}{\partial \beta^{(1)}}$ around $\beta_0^{(1)}$.

$$\begin{aligned}\frac{\partial S_n\left(\hat{\beta}\right)}{\partial \beta^{(1)}} =& \frac{\partial S_n\left(\beta_0^{(1)}\right)}{\partial \beta^{(1)}} + \nabla_1\left(\frac{\partial S_n\left(\beta_0^{(1)}\right)}{\partial \beta^{(1)}}\right)^T\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) \\ &+ \frac{1}{2}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)^T \nabla_1^2\left(\frac{\partial S_n\left(\beta_n^*\right)}{\partial \beta^{(1)}}\right)\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right).\end{aligned} \tag{12}$$

Then inserting (12) in (11) and isolating yields

$$-\nabla_1\left(\frac{\partial S_n\left(\beta_0^{(1)}\right)}{\partial \beta^{(1)}}\right)^T\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)$$

$$= \frac{\partial S_n\left(\beta_0^{(1)}\right)}{\partial \beta^{(1)}} + \frac{1}{2}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)^T \nabla_1^2 \left(\frac{\partial S_n\left(\beta_n^*\right)}{\partial \beta^{(1)}}\right)\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) - \lambda_n D_n.$$

Then using $-\nabla_1 \left(\frac{\partial S_n\left(\beta_0^{(1)}\right)}{\partial \beta^{(1)}}\right)^T \left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) = (1+\eta)\,\Sigma_{(1)}$ we get

$$\begin{aligned}
&n^{\frac{1}{2}}(1+\eta)\,\Sigma_{(1)}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) \\
&= n^{\frac{1}{2}} \frac{\partial S_n\left(\beta_0^{(1)}\right)}{\partial \beta^{(1)}} + \frac{n^{\frac{1}{2}}}{2}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)^T \nabla_1^2 \left(\frac{\partial S_n\left(\beta_n^*\right)}{\partial \beta^{(1)}}\right)\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) - \lambda_n n^{\frac{1}{2}} D_n.
\end{aligned}$$

By conditions (A3), (A4), and (A5) we have that

$$\begin{aligned}
&\left\|\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)^T \nabla_1^2 \left(\frac{\partial S_n\left(\beta_n^*\right)}{\partial \beta^{(1)}}\right)\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)\right\|_2^2 \\
&= \sum_{j=1}^{k_n} \left\| \frac{1+\eta}{n} \sum_{i=1}^{n} x_{ij}\phi^{(3)}\left(x_i^T \beta_n^*\right)\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)^T x_i^{(1)} x_i^{(1)T}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)\right\|_2^2 \\
&\leq M(1+\eta)\,\tau_{\max}^2\left(\frac{1}{n}X^{(1)T}X^{(1)}\right)\left\|\hat{\beta}^{(1)} - \beta_0^{(1)}\right\|_2^4 k_n \\
&= M(1+\eta)\,\tau_{\max}^2\left(\frac{1}{n}X^{(1)T}X^{(1)}\right) O_p\left(\frac{p_n^3}{n^2}\right) \\
&= M(1+\eta)\,\tau_{\max}^2\left(\frac{1}{n}X^{(1)T}X^{(1)}\right) O_p\left(n^{-1}\right).
\end{aligned}$$

Using that $\min_{j\in\{1,\ldots,k_n\}} \left\|\tilde{\beta}_j\right\|_2 = O_p\left(n^{(c_2-1)/2}\right)$ we have that

$$\begin{aligned}
\left\|\lambda_n n^{\frac{1}{2}} D_n\right\|_2^2 &= \sum_{j=1}^{k_n}\sum_{k=1}^{d_j}\left(\lambda_n n^{\frac{1}{2}}\left\|\tilde{\beta}_j\right\|_2^{-1}\operatorname{sgn}\left(\hat{\beta}_{jk}\right)\right)^2 \\
&= \lambda_n^2 n k_n O_p\left(n^{1-c_2}\right) \\
&= \lambda_n^2 O_p\left(n^{2-c_2+c_1}\right) \to 0,
\end{aligned}$$

since $\lambda_n n^{(2-c_2+c_1)/2} \to 0$. Then we have that

$$n^{\frac{1}{2}}\alpha_n^T \Sigma_{(1)}^{\frac{1}{2}}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) = n^{\frac{1}{2}}(1+\eta)^{-1}\alpha_n^T \Sigma_{(1)}^{-\frac{1}{2}} \frac{\partial S_n\left(\beta_0^{(1)}\right)}{\partial \beta^{(1)}} + o_p(1).$$

We verify the conditions of the Lindeberg-Feller CLT. Let

$$\begin{aligned} Z_{ni} =& n^{-\frac{1}{2}} (1+\eta)^{-1} \alpha_n^T \Sigma_{(1)}^{-\frac{1}{2}} x_i^{(1)} \left(y_i - \phi' \left(x_i^{(1)T} \beta_0^{(1)} \right) + \eta \left(y_i^p - \phi' \left(x_i^{(1)T} \beta_0^{(1)} \right) \right) \right) \\ =& n^{-\frac{1}{2}} (1+\eta)^{-1} \alpha_n^T \Sigma_{(1)}^{-\frac{1}{2}} x_i^{(1)} \left(\varepsilon_i + \eta \varepsilon_i^p \right), \end{aligned}$$

denoting $\varepsilon_i = y_i - \phi' \left(x_i^T \beta_0 \right)$ and $\varepsilon_i^p = y_i^p - \phi' \left(x_i^T \beta_0 \right)$. It holds that $\mathbb{E}\left[Z_{ni} \right] = 0$ and that

$$\begin{aligned} \mathbb{V}\left[\sum_{i=1}^n Z_{ni} \right] =& \mathbb{E}\left[\left(\sum_{i=1}^n Z_{ni} \right)^2 \right] \\ =& \mathbb{E}\left[\sum_{i=1}^n Z_{ni}^2 \right] \\ =& \mathbb{E}\left[\sum_{i=1}^n \left(n^{-\frac{1}{2}} (1+\eta)^{-1} \alpha_n^T \Sigma_{(1)}^{-\frac{1}{2}} x_i^{(1)} \left(\varepsilon_i + \eta \varepsilon_i^p \right) \right)^2 \right] \\ =& \frac{1}{n (1+\eta)^2} \mathbb{E}\left[\sum_{i=1}^n \alpha_n^T \Sigma_{(1)}^{-\frac{1}{2}} x_i^{(1)} \left(\varepsilon_i + \eta \varepsilon_i^p \right)^2 x_i^{(i)T} \Sigma_{(1)}^{-\frac{1}{2}} \alpha_n \right] = 1. \end{aligned}$$

Then we show that for any $\epsilon > 0$,

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}\left[Z_{ni}^2 I\left\{ |Z_{ni}| > \epsilon \right\} \right] &= \sum_{i=1}^n \mathbb{E}\left[Z_{ni}^2 \right] \mathbb{E}\left[I\left\{ |Z_{ni}| > \epsilon \right\} \right] \\ &\leq \left(\sum_{i=1}^n \mathbb{E}\left[Z_{ni}^2 \right]^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^n \mathbb{E}\left[I\left\{ |Z_{ni}| > \epsilon \right\} \right]^2 \right)^{\frac{1}{2}} \\ &\leq \left(\sum_{i=1}^n \mathbb{E}\left[Z_{ni}^4 \right] \right)^{\frac{1}{2}} \left(\sum_{i=1}^n \mathbb{P}\left[|Z_{ni}| > \epsilon \right] \right)^{\frac{1}{2}} = o(1). \end{aligned}$$

First by showing that

$$\begin{aligned} \sum_{i=1}^n \mathbb{P}\left[|Z_{ni}| > \epsilon \right] \leq& \frac{1}{\epsilon^2} \sum_{i=1}^n \mathbb{E}\left[Z_{ni}^2 \right] \\ =& \frac{1}{n\epsilon^2} (1+\eta)^{-2} \sum_{i=1}^n \mathbb{E}\left[\left(\varepsilon_i + \eta \varepsilon_i^p \right)^2 \right] \alpha_n^T \Sigma_{(1)}^{-\frac{1}{2}} x_i^{(1)} x_i^{(1)T} \Sigma_{(1)}^{-\frac{1}{2}} \alpha_n = O\left(k_n \right). \end{aligned}$$

And finally by showing that

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left[Z_{ni}^{4}\right] =& \frac{1}{n^{2}}\left(1+\eta\right)^{-4}\sum_{i=1}^{n}\mathbb{E}\left[\left(\varepsilon_{i}\eta+\varepsilon_{i}^{p}\right)^{4}\right]\left(x_{i}^{(1)T}\Sigma_{(1)}^{-\frac{1}{2}}\alpha_{n}\alpha_{n}^{T}\Sigma_{(1)}^{-\frac{1}{2}}x_{i}^{(1)}\right)^{2} \\ =& \frac{1}{n^{2}}\left(1+\eta\right)^{-4}\sum_{i=1}^{n}\mathbb{E}\left[\left(\varepsilon_{i}\eta+\varepsilon_{i}^{p}\right)^{4}\right]\tau_{\max}^{2}\left(\alpha_{n}\alpha_{n}^{T}\right)\tau_{\max}^{2}\left(\Sigma_{(1)}^{-1}\right)\left\|x_{i}^{(1)T}x_{i}^{(1)}\right\|_{2}^{2} \leq O\left(\frac{k_{n}^{2}}{n}\right).\end{aligned}$$

□

F Proof of theorem 3.5

The following Bernstein's inequality can be found in lemma 2.2.11 of van der Vaart and Wellner (1997).

Lemma F.0.1. Let $\gamma_1, \ldots, \gamma_n$ be independent random varables with zero mean such that for all i and some constant $M > 0$, $\mathbb{E}\left[|\gamma_i|^m\right] \leq m! M^{m-2} \mathbb{E}\left[\gamma_i^2\right]/2$ for every $m \geq 2$. Then

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} \gamma_i\right| > t\right] \leq 2\exp\left(-\frac{t^2}{2\left(\sum_{i=1}^{n} \mathbb{E}\left[\gamma_i^2\right] + Mt\right)}\right).$$

Proof. To prove this theorem we need to show that $\hat{\beta}_*$ satisfies the KKT conditions for the objective function, $Q_n\left(\beta; X, Y, Y^p\right)$, in (1) with probability approaching 1 for high n. Thus we need to show that

$$\mathbb{P}\left[\forall j \notin \{1, \ldots, k_n\}, \left\|\frac{\partial S_n\left(\hat{\beta}_*\right)}{\partial \beta_j}\right\|_2 \leq \lambda \left\|\tilde{\beta}_j\right\|_2^{-1}\right] \to 1,$$

which is equivalent to showing that

$$\mathbb{P}\left[\exists j \notin \{1, \ldots, k_n\}, \left\|\frac{\partial S_n\left(\hat{\beta}_*\right)}{\partial \beta_j}\right\|_2 > \lambda \left\|\tilde{\beta}_j\right\|_2^{-1}\right] \to 0,$$

Denote $B_{nj}\left(\beta\right) = n\frac{\partial S_n(\beta)}{\partial \beta_j}$, and let $\nabla_1 B_{nj}\left(\beta\right)$ and $\nabla_1^2 B_{nj}\left(\beta\right)$ denote the first and second order partial derivatives of $B_{nj}\left(\beta\right)$ with respect to $\beta^{(1)}$. By doing a second order Taylor expansion of $B_{nj}\left(\beta\right)$ around β_0 we get

$$B_{nj}\left(\hat{\beta}_*\right) = B_{nj}\left(\beta_0\right) + \nabla_1 B_{nj}\left(\beta_0\right)^T \left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right) + \frac{1}{2}\left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right)^T \nabla_1^2 B_{nj}\left(\beta_n^*\right)\left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right),$$

where β_n^* is between $\hat{\beta}_*$ and β_0.

For $j \notin \{1, \ldots, k_n\}$ we have from condition (B3) that $\left\|\tilde{\beta}_j\right\|_2 = O_P\left(r_n^{-1}\right)$. Hence, for some

large constant $L > 0$ we can rewrite (9).

$$
\begin{aligned}
&\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\frac{\partial S_n\left(\hat{\beta}_*\right)}{\partial \beta_j}\right\|_2 > \lambda_n \left\|\tilde{\beta}_j\right\|_2^{-1}\right] \\
\leq &\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|B_{nj}\left(\hat{\beta}_*\right)\right\|_2 > n\lambda_n \left\|\tilde{\beta}_j\right\|_2^{-1}, \left\|\tilde{\beta}_j\right\|_2 < \frac{L}{r_n}\right] \\
&+ \mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\tilde{\beta}_j\right\|_2 \geq \frac{L}{r_n}\right] \\
\leq &\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|B_{nj}\left(\beta_0\right)\right\|_2 > \frac{nr_n\lambda_n}{3L}\right] \\
&+ \mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\nabla_1 B_{nj}\left(\beta_0\right)^T\left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right)\right\|_2 > \frac{nr_n\lambda_n}{3L}\right] \\
&+ \mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\frac{1}{2}\left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right)^T \nabla_1^2 B_{nj}\left(\beta_n^*\right)\left(\hat{\beta}_*^{(1)} - \beta_0^{(1)}\right)\right\|_2 > \frac{nr_n\lambda_n}{3L}\right] + O(1) \\
\overset{\Delta}{=} &U_1 + U_2 + U_3 + O(1).
\end{aligned}
$$

Then considering U_1 and using lemma F.0.1 and conditions (B5) and (B6) we get

$$
\begin{aligned}
U_1 =&\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|B_{nj}\left(\beta_0\right)\right\|_2 > \frac{nr_n\lambda_n}{3L}\right] \\
=&\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \left\|\sum_{i=1}^n x_{ij}\left(\varepsilon_i + \eta\varepsilon_i^p\right)\right\|_2 > \frac{nr_n\lambda_n}{3L}\right] \\
=&\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \sqrt{\sum_{k=1}^{d_j}\left(\sum_{i=1}^n x_{ijk}\left(\varepsilon_i + \eta\varepsilon_i^p\right)\right)^2} > \frac{nr_n\lambda_n}{3L}\right] \\
\leq &\mathbb{P}\left[\exists j \notin \{1,\ldots,k_n\}, \sum_{k=1}^{d_j}\left|\sum_{i=1}^n x_{ijk}\left(\varepsilon_i + \eta\varepsilon_i^p\right)\right| > \frac{nr_n\lambda_n}{3L}\right] \\
\leq &p_n \max_{j\notin\{1,\ldots,k_n\}} \mathbb{P}\left[\sum_{k=1}^{d_j}\left|\sum_{i=1}^n x_{ijk}\left(\varepsilon_i + \eta\varepsilon_i^p\right)\right| > \frac{nr_n\lambda_n}{3L}\right] \\
\leq &p_n \max_{j\notin\{1,\ldots,k_n\}} \sum_{k=1}^{d_j} \mathbb{P}\left[\left|\sum_{i=1}^n x_{ijk}\left(\varepsilon_i + \eta\varepsilon_i^p\right)\right| > \frac{nr_n\lambda_n}{3Ld_b}\right] \\
\leq &2q_n \exp\left(-\frac{(nr_n\lambda_n)^2}{18L^2d_b^2nR + 2Mnr_n\lambda_n}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= O(1) \exp\left(- \log(p_n) \left(\frac{(nr_n\lambda_n)^2}{2\log(p_n)\left(9L^2 d_b^2 nR + Mnr_n\lambda_n\right)} - 1 \right) \right) \\
&= O(1) \exp\left(- \log(p_n) \left(\frac{1}{18L^2 d_b^2 R \log(p_n)/\left(nr_n^2\lambda_n^2\right) + 2M\log(p_n)/(nr_n\lambda_n)} - 1 \right) \right) \to 0.
\end{aligned}
$$

The proof for $U_2 \to 0$ and $U_3 \to 0$ is similar to the proof of theorem 2.5 in Wang and Tian (2019). □

G Proof of theorem 3.6

Proof. We need to show that

$$\mathbb{P}\left[\min_{j\in\{1,\dots,k_n\}} \left\|\hat{\beta}\right\|_2 > 0\right) \to 1.$$

From condition (B4) we get

$$\begin{aligned}
\min_{j\in\{1,\dots,k_n\}} \left\|\hat{\beta}\right\|_2 \geq & \theta_1 - \left\|\hat{\beta} - \beta_0\right\|_2 \\
\geq & Mn^{(c_4-1)/2} - O_p\left(\left(\frac{k_n}{n}\right)^{\frac{1}{2}}\right) \\
\geq & Mn^{(c_4-1)/2} - O_p\left(n^{(c_3-1)/2}\right) > 0.
\end{aligned}$$

□

Research Papers

2021-05: Stefano Grassi and Francesco Violante: Asset Pricing Using Block-Cholesky GARCH and Time-Varying Betas

2021-06: Gloria González-Rivera, Carlos Vladimir Rodríguez-Caballero and Esther Ruiz Ortega: Expecting the unexpected: economic growth under stress

2021-07: Matei Demetrescu and Robinson Kruse-Becher: Is U.S. real output growth really non-normal? Testing distributional assumptions in time-varying location-scale models

2021-08: Luisa Corrado, Stefano Grassi and Aldo Paolillo: Modelling and Estimating Large Macroeconomic Shocks During the Pandemic

2021-09: Leopoldo Catania, Alessandra Luati and Pierluigi Vallarino: Economic vulnerability is state dependent

2021-10: Søren Johansen and Anders Ryghn Swensen: Adjustment coefficients and exact rational expectations in cointegrated vector autoregressive models

2021-11: Bent Jesper Christensen, Mads Markvart Kjær and Bezirgen Veliyev: The incremental information in the yield curve about future interest rate risk

2021-12: Mikkel Bennedsen, Asger Lunde, Neil Shephard and Almut E. D. Veraart: Inference and forecasting for continuous-time integer-valued trawl processes and their use in financial economics

2021-13: Anthony D. Hall, Annastiina Silvennoinen and Timo Teräsvirta: Four Australian Banks and the Multivariate Time-Varying Smooth Transition Correlation GARCH model

2021-14: Ulrich Hounyo and Kajal Lahiri: Estimating the Variance of a Combined Forecast: Bootstrap-Based Approach

2021-15: Salman Huseynov: Long and short memory in dynamic term structure models

2022-01: Jian Kang, Johan Stax Jakobsen, Annastiina Silvennoinen, Timo Teräsvirta and Glen Wade: A parsimonious test of constancy of a positive definite correlation matrix in a multivariate time-varying GARCH model

2022-02: Javier Haulde and Morten Ørregaard Nielsen: Fractional integration and cointegration

2022-03: Yue Xu: Spillovers of Senior Mutual Fund Managers' Capital Raising Ability

2022-04: Morten Ørregaard Nielsen, Wonk-ki Seo and Dakyung Seong: Inference on the dimension of the nonstationary subspace in functional time series

2022-05: Kristoffer Pons Bertelsen: The Prior Adaptive Group Lasso and the Factor Zoo