Vector Autoregressions with Parsimoniously Time-Varying Parameters and an Application to Monetary Policy[☆]

Laurent Callot[a,b,c], Johannes Tang Kristensen[c,d]

[a] *Department of Econometrics and OR, VU University Amsterdam.*
[b] *the Tinbergen Institute.*
[c] *CREATES, Aarhus University.*
[d] *Department of Business and Economics, University of Southern Denmark.*

Abstract

This paper studies vector autoregressive models with parsimoniously time-varying parameters. The parameters are assumed to follow parsimonious random walks, where parsimony stems from the assumption that increments to the parameters have a non-zero probability of being exactly equal to zero. We estimate the sparse and high-dimensional vector of changes to the parameters with the Lasso and the adaptive Lasso.

The parsimonious random walk allows the parameters to be modelled non parametrically, so that our model can accommodate constant parameters, an unknown number of structural breaks, or parameters varying randomly. We characterize the finite sample properties of the Lasso by deriving upper bounds on the estimation and prediction errors that are valid with high probability, and provide asymptotic conditions under which these bounds tend to zero with probability tending to one. We also provide conditions under which the adaptive Lasso is able to achieve perfect model selection.

We investigate by simulations the properties of the Lasso and the adaptive Lasso in settings where the parameters are stable, experience structural breaks, or follow a parsimonious random walk. We use our model to investigate the monetary policy response to inflation and business cycle fluctuations in the US by estimating a parsimoniously time varying parameter Taylor rule. We document substantial changes in the policy response of the Fed in the 1970s and 1980s, and since 2007, but also document the stability of this response in the rest of the sample.

JEL codes: C01, C13, C32, E52.

Keywords: Parsimony, time varying parameters, VAR, structural break, Lasso.

1. Introduction

This paper proposes a parsimoniously time-varying vector autoregressive model (with exogenous variables, VARX). The parameters of this model are assumed to follow a parsimonious random walk, that is, a random walk with a positive probability that an increment is

[☆] The authors would like to thank Anders B. Kock, Paolo Santucci de Magistris, and seminar participants at Maastricht University for their comments and suggestions. Earlier versions of this paper were presented at the 2014 Netherlands Econometrics Study Group and the 2014 NBER-NSF time series conference. Furthermore, support from CREATES, Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation is gratefully acknowledged.

Email addresses: l.callot@vu.nl (Laurent Callot), johannes@sam.sdu.dk (Johannes Tang Kristensen)

exactly equal to zero. The parsimonious random walk allows the time varying parameters to be modelled non parametrically so that the parameters can follow a wide range of classical time varying processes. We use the Lasso of Tibshirani (1996) to estimate the vector of increments to the parameters which is sparse under the parsimonious random walk assumption, and is high dimensional in the sense of being at least as large as the sample size. For a general review of the Lasso in high-dimensional settings see Bühlmann and van de Geer (2011) and Belloni and Chernozhukov (2011). We begin this introduction by contextualizing our model within the literature on time varying parameter models, and then detail our contributions before turning to the specifics of our model and estimation method.

There exists a substantial literature on time varying parameter models in every domain of time series econometrics. Using a Bayesian approach, Koop and Korobilis (2013) estimate large time varying-parameter VARs using forgetting factors to render the estimation of their model computationally feasible, while Bitto and Frühwirth-Schnatter (2014) uses shrinkage for the same purpose. Likelihood driven models such as state space models (Durbin and Koopman, 2012), and more recently generalised autoregressive score models (Creal, Koopman, and Lucas, 2013), are routinely used to allow the parameters to vary over time guided by the data.

In the models discussed above the parameters do vary at every point in time; another strand of literature investigates models with a finite number of changes in the parameters, or a finite number of possible values the parameters may take over time. One example of such models is regime switching models (see Hamilton (2008) for a review). These are typically used in the empirical literature to model systems experiencing a succession of recessive and expansive regimes, or any other finite number of regimes, with the probability of switching between regimes being data dependent. Another example is the issue of structural breaks, i.e. cases where the parameters experience a small and finite number of changes over time, for instance in response to a policy change. The structural breaks literature is extensive, covering a breadth of models and methods. From the perspective of this paper the most relevant part is the treatment of linear regression models in e.g. Bai (1997) and Bai and Perron (1998), and VAR models in e.g. Bai (2000) and Qu and Perron (2007). For a general review see Perron (2006). The problem of structural breaks has also been addressed using shrinkage methods: In an autoregressive setting Chan, Yau, and Zhang (2014) uses the group Lasso to estimate clusters of parameters with identical values over time, and Qian and Su (2014) considers the problem of estimating time series models with endogenous regressors and an unknown number of breaks using the group fused Lasso.

Evidence of the importance of allowing for the parameters of a model to vary over time are widespread in the literature. Of particular interest for our empirical application are Primiceri (2005); Boivin and Giannoni (2006); Sims and Zha (2006) who document, using a wide range of models and estimators, that the monetary policy response to inflation in the US changed in the early 1980s. This evidence is controversial though, Sims and Zha (2006) reviews research finding no evidence of instability in US monetary policy. We use our framework to investigate this issue by estimating parsimoniously time varying Taylor rules (Taylor, 1993).

The main contribution of this paper is to propose an estimator for VARX models with parsimoniously time-varying parameters. We detail below a few novel aspects of this paper.

i) In order to model the potential time variations of the parameters of the VARX in a flexible way we propose the parsimonious random walk process. This process has two advantages. First, by allowing the increments to be exactly equal to zero with some positive probability it allows us to consider models with structural breaks or even constant parameters. Second, by allowing the parameters to behave as a random walk it allows us to model

the path of the parameter vector in a non parametric way. In this paper we assume the probability α_T for an increment to be different from zero to depend on the sample length T, specifically $\alpha_T = k_\alpha T^{-a}$, where k_α and a are positive constants. In the case of a single variable this leads to an expected number of non-zero increments $E(s) = k_\alpha T^{1-a}$.

ii) In a model similar to ours in structure, Harchaoui and Lévy-Leduc (2010) shows that consistent estimation with the Lasso is not possible under standard assumptions. We introduce the concept of asymptotic distinctness of the changes in the parameters, specifically we assume that the distance between breaks increases at a rate $\mathcal{O}(T^d)$, $0 < d \leq 1$. Under this assumption, we are able to prove consistency of the Lasso provided conditions on d are satisfied.

iii) We establish finite sample upper bounds on the ℓ_1 norm of estimation error and the ℓ_2 norm of the prediction error of the Lasso, and show that they hold with high probability, building on results from Kock and Callot (2015). We then turn to asymptotics and show that the upper bounds tend to zero with probability tending to one under conditions such that $a \geq 3/4$ and $d > 3/4$.

iv) We establish conditions under which the adaptive Lasso (Zou, 2006) possesses the oracle property, that is, the conditions under which the adaptive Lasso recovers the true model. Specifically we establish a finite sample probability of perfect selection under conditions on the size of the smallest non-zero parameter. We then derive conditions on a and d under which the adaptive Lasso recovers the true model with probability tending to one.

v) To illustrate the relevance of our model we provide an application investigating the monetary policy response to inflation in the US from 1954 to 2014. More specifically we estimate a set of parsimoniously time-varying Taylor rules in which the Fed's effective fund rate depends on past inflation and output gap as well as its own lag. We find that the response to inflation has been unstable in the 1970s and the 1980s but that, in the two decades preceding and following this period, the policy response was stable. We also find weaker evidence for a persistent change starting in 2007, coinciding with the period at which the Fed fund rate reached its lower bound.

In the next section we formally introduce the model and our assumptions. Section 3 contains the finite sample and asymptotic theorems describing the behaviour of the Lasso. The subsequent section is dedicated to investigating the properties of our estimator in Monte Carlo experiments. Finally, we illustrate the practical relevance of the proposed model by estimating several specifications of a parsimoniously time-varying parameter Taylor rule for US monetary policy and document substantial instability in the response of the Fed to inflation in the 1970s and 1980s, and since 2007.

2. Model

We consider a VARX(p) model with parsimoniously time-varying parameters including r_x exogenous variables X_t, and p lags of the r_y dependent variables $Y_t = [y_{1t}, ..., y_{r_y t}]'$. Since this model will be estimated equation by equation, we restrict our focus to equation i, $i = 1, ..., r_y$

$$\begin{aligned} y_{it} &= \beta_{it}' X_t + \sum_{l=1}^{p} \gamma_{ilt}' Y_{t-l} + \epsilon_{it} \\ &= \xi_{it}' Z_t + \epsilon_{it} \end{aligned} \tag{1}$$

where $Z_t = [X_t', Y_{t-1}', \ldots, Y_{t-p}']'$ is of dimension $r \times 1$, $r = r_x + pr_y$, and $\xi_{it} = [\beta_{it}', \gamma_{i1t}', \ldots, \gamma_{ipt}']'$. In order to lighten up the notation we drop the equation subscript i henceforth, y_t should be understood as being any element of Y_t.

In order to establish finite sample bounds on the performance of the Lasso we make use of concentration inequalities on averages of products of the elements of the model. These inequalities are valid if the tails of the entries are sub-exponential, to ensure this we need to make a series of independence and Gaussianity assumptions.

Assumption 1 (Covariates and innovations)**.** *Assume that:*

i) $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ *is a sequence of* $i.i.d$ *innovation terms,* $\sigma_\epsilon^2 < \infty$.

ii) $X_t \sim \mathcal{N}(0, \Omega_X^2)$. *For all* $k = 1, ..., r_x$, $\mathrm{Var}(X_{kt}) = \sigma_{Xk}^2 < \infty$.

iii) $E(\epsilon' X) = 0$.

The variances of the innovations ϵ_t and of the exogenous variables X_{tk} could be assumed to be heteroskedastic; for our purpose we only require that these variables are sequences of independent Gaussian random variables with finite variances. We also require y_t to be a Gaussian random variable with finite variance. The linearity of the model and assumption 1 ensures Gaussianity, but we need an extra assumption on the dynamics of the model to ensure that the variances remain finite.

Define the parameter matrices of the full VARX(p): $\Gamma_{lt} = [\gamma_{1lt}, ..., \gamma_{r_y lt}]'$ and $B_t = [\beta_{1t}, ..., \beta_{r_x t}]'$, which are of dimensions $r_y \times r_y$ and $r_y \times r_x$ respectively. We write the VARX(p) in companion form:

$$
\begin{aligned}
Y_t &= B_t X_t + \sum_{l=1}^{p} \Gamma_{lt} Y_{t-l} + \epsilon_t \\
\mathbf{Y}_t &= A_t \mathbf{Y}_{t-1}' + \Sigma_t
\end{aligned}
$$

where $\mathbf{Y}_t = \left[Y_t, Y_{t-1}, ..., Y_{t-p+1} \right]'$ and $\Sigma_t = \left[\epsilon_t + B_t X_t, 0, ..., 0 \right]'$ are matrices of dimensions $pr_y \times r_y$, and A_t is the companion matrix:

$$
A_t = \begin{bmatrix}
\Gamma_{1t} & \cdots & \cdots & \Gamma_{pt} \\
I_{r_y} & \cdots & \cdots & 0 \\
 & \ddots & \vdots & \vdots \\
0 & \cdots & I_{r_y} & 0
\end{bmatrix}.
$$

Now further define the $r_y \times Tr_y$ selection matrix $J = [I_{r_y}, 0, ..., 0]$, and let $\Phi_{jt} = J \left(\prod_{k=0}^{j-1} A_{t-k} \right) J'$. A standard results for VAR models with time varying coefficients, see for example (Lütkepohl, 2007, section 17.2.1), gives the covariance matrix of Y_t:

$$
E(Y_t Y_t') = \sum_{j=0}^{\infty} \Phi_{jt} E(\Sigma_{t-j}) \Phi_{jt}'.
$$

We can now state our assumption on the dynamics of the VAR ensuring that the variance of Y_t is finite.

Assumption 2. *(VAR dynamics) Let*

$$\mathrm{Var}\left(Y_t\right) = \left[\sigma^2_{y_1 t}, ..., \sigma^2_{y_{r_y} t}\right] = diag\left(\sum_{j=0}^{\infty} \Phi_{jt} E(\Sigma_{t-j}) \Phi'_{jt}\right).$$

For some positive constant $M < \infty$ and for all $t = 1, ..., T$ and $k = 1, ..., r_y$, we have $\sigma^2_{y_k t} \leq M$.

We now turn our attention to the process driving the parameters. The structuring assumptions of this paper is that the change in the value of the parameter vector, ξ_t, for the r variables of the model at time t, $1 \leq t \leq T$ is defined as the element-by-element product (noted $\odot$) of two random variables $\eta_t \in \mathbb{R}^r$ and $\zeta_{tk} = 0$ or 1, $k = 1, ..., r$. If $P(\zeta_{tk} = 0) > 0$, then the vector of increments to the parameters $(\eta_1 \odot \zeta_1, \eta_2 \odot \zeta_2, ..., \eta_T \odot \zeta_T)$ is sparse, and the sparsity of this vector is controlled by $P(\zeta_{tk} = 0)$, whereas if $P(\zeta_{tk} = 0) = 0$ then the parameters follow random walks. When $P(\zeta_{tk} = 0) > 0$ we refer to this process as a parsimonious random walk. For a low probability of non-zero increments the parsimonious random walk can generate parameter paths that are akin to those considered in the structural break literature, while for a higher probability of non-zero increments the paths can be akin to regime switches or other paths with a high degree of variation. The process is formally defined in assumption 3 below:

Assumption 3 (Parsimonious random walk)**.** *Assume that the parameters follow a parsimonious random walk with ξ_0 given.*

$$\xi_t = \xi_{t-1} + \zeta_t \odot \eta_t.$$

η_t and ζ_t are vectors of length r with the following properties:

$$\begin{aligned}
\alpha_T &= k_\alpha T^{-a}, \ 0 \leq a \leq 1, \ k_\alpha > 0 \\
\zeta_{jt} &= \begin{cases} 1, & w.p. \ \alpha_T \\ 0, & w.p. \ 1 - \alpha_T \end{cases} \quad j \in 1, ..., r \\
\eta_t &= \mathcal{N}(0, \Omega_\eta) \\
E(\eta'_t \eta_u) &= 0 \, if \, t \neq u \\
E(\eta'_t \zeta_u) &= 0 \ \forall \, t, u \in 1, ..., T
\end{aligned}$$

We assume that $\alpha_T = k_\alpha T^{-a}$, which controls the sparsity of the vector of increments thereby controlling the number (and rate of growth) of non-zero parameters which we seek to estimate. The constant k_α scales the probability α_T and must be such that $0 \leq \alpha_T \leq 1$. If k_α satisfies this restriction for some T_0, it will satisfy it for any $T \geq T_0$ since $a \geq 0$. Consistency requirements for the Lasso estimator will impose a tighter lower bound on a. It is important to note that while assumption 3 puts no further restrictions on the path of the parsimonious random walk, we do rule out paths that violate assumption 2, i.e. paths that cause the variance of Y_t to be unbounded.

Continuing to the task of setting up the estimation problem we start by noting that by multiplying the diagonalized matrix of covariates Z^D by a selection matrix W,

$$Z^D = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_T \end{bmatrix}, W = \begin{bmatrix} I_r & 0 & \cdots & 0 \\ I_r & I_r & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ I_r & I_r & \cdots & I_r \end{bmatrix}, Z^D W = \begin{bmatrix} Z_1 & 0 & \cdot & 0 \\ Z_2 & Z_2 & \cdot & 0 \\ \vdots & \vdots & \ddots & \vdots \\ Z_T & Z_T & \cdots & Z_T \end{bmatrix},$$

we are able to write our parsimoniously time-varying VARX model (1) as a simple regression model

$$y = Z^D W\theta + \epsilon$$

where the parameter vector $\theta' = [\xi_0' + \zeta_1' \odot \eta_1', \zeta_2' \odot \eta_2', ..., \zeta_T' \odot \eta_T']$ has length rT, and $y = (y_1, ..., y_T)'$, $\epsilon = (\epsilon_1, ..., \epsilon_T)'$. The matrix $Z^D W$ contains T observations for rT covariates constructed from the original r covariates. The first r elements of θ are the sum of the initial value of the parsimonious random walk ξ_0 and the first increment $\zeta_1 \odot \eta_1$. The subsequent elements of θ are the increments of the parsimonious random walk $\zeta_t \odot \eta_t,\ t > 1$ so that by cumulating the entries of θ we can recover the full path of the parameters.

The sparsity of the vector of increments from assumption 3 implies sparsity of the vector of parameters θ. Let s_T be the number of non-zero parameters, the expected number of non-zero parameters is $E(s_T) = rk_\alpha \alpha_T T \in \mathcal{O}(T^{1-a})$. The growth rate of the expected number of non-zero parameters is entirely controlled by a, r being fixed. The specification of the parsimonious random walk allows for great flexibility. For $a < 1$ we allow the number of parameters, and thus the number of increments, to grow with the sample size. For $a = 1$ the expected number of parameters is constant (equal to rk_α) and thus covers the cases of stable parameters, or of a fixed number of structural breaks. When estimating the model we estimate a particular realization of the stochastic processes followed by the parameters where the degree of sparsity of this realization is unknown. Results similar to ours would hold if we were to assume the parameters are fixed quantities with an unknown (finite) number of breaks.

Since the parameter vector is sparse the estimation problem requires a sparse estimator; we choose to use the Lasso to estimate θ, and we discuss the properties of the Lasso estimator in this setting in the next section.

2.1. Notation

Before proceeding further we introduce some notation. Let $\left[\sigma_1^2, ..., \sigma_{rT}^2\right] = \text{diag}(\text{Var}(Z^D W))$ and $\sigma_T^2 = \max\left(\sigma_\epsilon^2, \max_{1 \le k \le rT} \sigma_k^2\right)$ where σ_ϵ^2 is the variance of ϵ and σ_k^2 is the variance of the k^{th} column in $Z^D W$. Define the active set $\mathcal{S}_T$ as the set of indices corresponding to non-zero parameters in θ, $\mathcal{S}_T = \left\{j \in (1, ..., rT) | \theta_j \neq 0\right\}$, and its cardinality $|\mathcal{S}_T| = s_T$. To simplify notation, when it is unambiguous, we omit the subscript T. We note $\|.\|_{\ell_1}$ the ℓ_1 norm, $\|.\|$ the ℓ_2 norm, and $\|.\|_\infty$ the maximum norm. The sign function is defined as $\text{sign}(x) = -1$ if $x < 0$, $\text{sign}(x) = 0$ if $x = 0$, and $\text{sign}(x) = 1$ if $x > 0$. For a matrix A let $\underline{\phi}(A)$ denote the smallest eigenvalue of A and $\overline{\phi}(A)$ the largest. Let $f(T) \in \Omega(g(T))$ mean that there exists a constant $c > 0$ such that $f(T) \ge cg(T)$ for $T \ge T_0$ for a certain T_0 onwards, and $f(T) \in \mathcal{O}(g(T))$ mean that there exists a constant $c > 0$ such that $f(T) \le cg(T)$ for $T \ge T_0$ for a certain T_0 onwards. Similarly, let $\Omega_p(\cdot)$ and $\mathcal{O}_p(\cdot)$ define their probabilistic counterparts.

3. Estimation

The Lasso estimator $\widehat{\theta}$ minimizes the following convex objective function:

$$\widehat{\theta} = \text{argmin}_\theta \left(\frac{1}{T} \left\| y - Z^D W\theta \right\|^2 + 2\lambda_T \left\| \theta \right\|_{\ell_1} \right). \tag{2}$$

Because the objective function (2) is convex, finding the solution to (2) for a given value of λ_T is an easy problem from a computational standpoint making the estimation of this

model fast. Our model is high dimensional by construction, in the sense that the number of parameters to estimate is at least as large as the sample size; the number of non-zero parameters is of a smaller order than the sample size, however. To investigate the properties of the Lasso in this model we start by deriving finite sample bounds for the estimation and prediction errors before considering their asymptotic behaviour. We also derive results regarding the Lasso's variable selection properties, as well as conditions under which the adaptive Lasso achieves perfect variable selection. Before doing so we introduce an extra assumption.

This model has rT parameters and T observations so that when $r > 1$ its Gram matrix $\Psi_T = \frac{(W'Z^{D'})(Z^D W)}{T}$ is singular. In this setting the ordinary least squares estimator is infeasible, but Bickel, Ritov, and Tsybakov (2009) shows that the Lasso can have attractive properties as long as a weaker condition on the Gram matrix, the restricted eigenvalue condition, is satisfied. Consider the sample restricted eigenvalue:

$$\kappa_T^2(\Psi_T) = \min_{\delta} \left\{ \frac{\delta' \Psi_T \delta}{\left\| \delta_{\mathscr{S}} \right\|^2} : \delta \in \mathbb{R}^{rT} \setminus \{0\}, \left\| \delta_{\mathscr{S}^c} \right\|_{\ell_1} \leq 3 \left\| \delta_{\mathscr{S}} \right\|_{\ell_1}, |\mathscr{S}| \leq s \right\}$$

The restricted eigenvalue condition, $\kappa_T^2(\Psi_T) > 0$, implies that square submatrices of size $\leq 2s$ of the Gram matrix have positive eigenvalues. We must ensure that we are working with a set of variables on which the restricted eigenvalue condition can be satisfied. This cannot be the case when using the entire set of constructed variables $Z^D W$ since when $r > 1$ the last r columns of $Z^D W$, $[0_r, ..., 0_r, Z_T']'$, are by construction linearly dependent. So are all the sets of r columns of the form $[0_r, ..., 0_r, Z_{T_0}', ..., Z_T']'$ where $T_0 > T - r + 1$, so we must always rule out the possibility of a change in parameter values after time $T - r + 1$.

In turns out that for the asymptotic analysis this is not sufficient. Harchaoui and Lévy-Leduc (2010) consider the problem of estimating change points in a piecewise constant signal observed with noise, a model similar in structure to ours, and show that in such a setting the restricted eigenvalue is of the order T^{-1}. This would also be the case in our setting were we not to make further assumptions. It is not possible to consistently select break locations unless we assume they are asymptotically distinct by which we mean that, while the number of breaks can increase with the sample size, the distance between breaks must also increase.

In order to allow breaks to occur only at a restricted set of points in time, define the set of time indices $\mathscr{T} \subset \{1, \ldots, T\}$ with cardinality $\left|\mathscr{T}\right| = m_T$. The selection matrix W can be written as $W = \tilde{W}_T \otimes I_r$ where $\tilde{W}_T$ is a $T \times T$ matrix with ones on and below the diagonal. We then define $W_{\mathscr{T}} = \tilde{W}_{T\mathscr{T}} \otimes I_r$ where $\tilde{W}_{T\mathscr{T}}$ is a matrix containing the columns of $\tilde{W}_T$ corresponding to indexes in $\mathscr{T}$. In other words by using $W_{\mathscr{T}}$ instead of W we only look for breaks at points in time corresponding to the indexes in $\mathscr{T}$. Having defined $\mathscr{T}$ we can correspondingly define a new Gram matrix $\Psi_{T\mathscr{T}} = \frac{W_{\mathscr{T}}' Z^{D'} Z^D W_{\mathscr{T}}}{T}$, and the corresponding restricted eigenvalue $\kappa_T^2(\Psi_{T\mathscr{T}})$. In the following we provide a set of sufficient conditions under which we can control the rate of decay of $\kappa_T^2(\Psi_{T\mathscr{T}})$.

Assumption 4. *Assume that:*

i) The breaks are asymptotically distinct: Let the points in time where we check for breaks be given by $\mathscr{T} = \{t_1, t_2, \ldots, t_{m_T}\}$. Then assume that the distance between two points is $t_{i+1} - t_i = \tau_i \geq D_T$ where $D_T = cT^d$ for some positive constants c and d where $d \in (0, 1]$.

ii) Let $\check{Z}_i$ be the observations between two grid points t_i and t_{i+1}, i.e. $\check{Z}_i' = [Z_{t_i+1}, Z_{t_i+2}, \ldots, Z_{t_{i+1}}]$. Then $\underline{\phi}(T^{-d}\check{Z}_i'\check{Z}_i) > 0$ with probability approaching one.

Assumption 4 *i*) imposes that the distance between breaks is at least $D_T = cT^d$; we can always choose c so that $D_T = 1$ thereby not imposing any restrictions on the location of the breaks when estimating the model.

Lemma 1. *Under assumptions 1, 3, and 4 we have $\kappa_T^2(\Psi_{T\mathcal{T}}) \in \Omega_p(T^{d-1})$ for $d \in (0,1]$.*

Lemma 1 shows that if we assume the breaks are asymptotically distinct, the rate of decay of $\kappa_T^2(\Psi_{T\mathcal{T}})$ is a function of the distance between two breaks. We will use this property in the asymptotic analysis of our estimators to establish under which conditions on d we can achieve consistency, and perfect variable selection for the adaptive Lasso.

As assumption 4 merely provides sufficient conditions for $\kappa_T^2(\Psi_{T\mathcal{T}}) \in \Omega_p(T^{d-1})$, we will not work under this assumptions directly and instead make the following assumption.

Assumption 5 (Restricted eigenvalue condition)**.** *Assume that the index set $\mathcal{T}$ is constructed such that:*

i) $\kappa_T^2(\Psi_{T\mathcal{T}}) > 0$.

ii) $\kappa_T^2(\Psi_{T\mathcal{T}}) \in \Omega_p(T^{d-1})$ *for some* $d \in (0,1]$.

If $r = 1$, assumption 5 *i*) is (almost surely) satisfied by construction of $Z^D W$. With $r > 1$ it is a data dependent issue whether assumption 5 *i*) is satisfied as long as we rule out the possibility of changes in the parameter value after time $T - r + 1$. This restriction is explicitly taken into account in the results below. To establish the finite sample results we do not impose a minimal distance between the breaks, and assumption 5 *ii)* suffices for the asymptotic analysis. For this reason and to simplify notations we omit the subscript $\mathcal{T}$ on the Gram matrix and also write $\kappa_T^2 := \kappa_T^2(\Psi_{T\mathcal{T}})$.

By penalizing every entry of θ, we penalize the initial value of the parsimonious random walks, ξ_{k0} $(k = 1, \ldots, r)$ together with the initial increments $\eta_{k1}\zeta_{k1}$. In doing so we make it possible for the initial value of the parsimonious random walk to be set to zero by the Lasso and therefore, if all further increments are also set to zero, to exclude altogether an irrelevant variable. Alternatively it is possible not to penalize $\xi_0 + \eta_1 \odot \zeta_1$ in which case, if all further increments are set to zero by the Lasso, the value of the parsimonious random walk at any point in time is equal to the OLS estimator of $y = Z\Xi + \epsilon$. This also implies that the estimate of the initial value is not biased towards zero. Choosing either alternative has a negligible influence on the results below since it only involves the penalization (or lack thereof) of a single parameter.

3.1. The Lasso

We can now state our first theorem on the estimation and prediction errors of the Lasso.

Theorem 1. *For $\lambda_T = \sqrt{\frac{8\ln(1+T)^5\ln(1+r)^2\ln(r(T-r+1))\sigma_T^4}{T}}$ and some constant $A > 0$, under assumptions 1, 2, 3, and 5, and on the set $\mathcal{B}_T$ with probability at least equal to $1 - \pi_T^{\mathcal{B}}$ we have the following inequalities:*

$$\frac{1}{T}\left\| Z^D W(\hat{\theta} - \theta) \right\|^2 \leq \frac{16 s \lambda_T^2}{\kappa_T^2},$$
$$\left\| \hat{\theta} - \theta \right\|_{\ell_1} \leq \frac{16 s \lambda_T}{\kappa_T^2}, \tag{3}$$

with $\pi_T^{\mathscr{B}} = 2(1+T)^{-1/A} + 2(r(T-r+1))^{1-\ln(1+T)}$.

The bounds given in theorem 1 hold on a set that has probability at least $1-\pi_T^{\mathscr{B}}$ for a given value of λ_T. These bounds are valid for any value of the penalty parameter as long as $\left\| T^{-1}\varepsilon' Z^D W \right\|_\infty \leq \lambda_T/2$ is satisfied, that is, if we are on $\mathscr{B}_T$; holding everything else constant the probability of $\mathscr{B}_T$ decreases with λ_T.

The dependence on λ_T highlights the trade-off between selecting a larger value of λ_T to increase the probability of $\left\| T^{-1}\varepsilon' Z^D W \right\|_\infty \leq \lambda_T/2$ to be satisfied, and selecting a lower value to reduce the upper bounds of the estimation and prediction errors. The bounds depend linearly on the size of the active set so that more break points imply larger upper bounds. They also depend indirectly on the variance of $\eta_t \odot \zeta_t$ through σ_T^2 which enters the expression of λ_T.

If we assume that the smallest non-zero increment is larger than the estimation error, we can show that no relevant variables are rejected, or equivalently, no break point goes undetected. Let $\theta_{\min} = \min_{j\in\mathscr{S}} \left\{ |\theta_j| \right\}$ be the smallest non-zero parameter.

Corollary 1. *If* $\theta_{\min} > \left\| \hat{\theta} - \theta \right\|_{\ell_1}$ *then* $\widehat{\mathscr{S}} \bigcap \mathscr{S} = \mathscr{S}$.

The Lasso cannot surely distinguish between parameters that are smaller than the estimation error and parameters that are truly zero. There is a risk of misclassification for small non-zero parameters. Similar results are used in the literature to claim that the Lasso possess the oracle property. This result is stated as a corollary as it requires an extra condition to be met relative to theorem 1. We stress that even when the Lasso does not possess the oracle property, the properties of the Lasso in terms of overall estimation error of the path of the parameters are still valid. If the condition on $\theta_{\min}$ is violated, the Lasso cannot surely detect the precise location of every change point in the parsimonious random walk but can still approximate it well.

We now turn to an asymptotic setting to show consistency of our estimator and, importantly, to get a sense of the number of changes in the parsimonious random walks that our estimator can handle in the form of a bound on the rate of growth of s. Theorem 2 below provides an asymptotic counterpart to theorem 1.

Theorem 2. *Let* a *and* d *be scalars with* $a, d \leq 1$, $1-a+d \leq 1$, *and* $\frac{3}{2}-a-d<0$. *Then under assumptions 1, 2, 3, and 5, and as* $T \to \infty$ *we have:*

$$\frac{1}{T}\left\| Z^D W(\hat{\theta}-\theta) \right\|^2 \to^p 0 \tag{4}$$

$$\left\| \hat{\theta}-\theta \right\|_{\ell_1} \to^p 0 \tag{5}$$

Theorem 2 states that the prediction and estimation errors tend to zero in probability provided a set of conditions involving jointly the speed of growth of the cardinality of the active set ($\mathscr{O}(T^{1-a})$) and the speed at which the minimal distance between breaks grows $\mathscr{O}(T^d)$. The condition $1-a+d \leq 1$ simply ensures that the number of breaks multiplied by the distance between them is not of a larger order than the sample size. The condition $\frac{3}{2}-a-d<0$ ensures that $\frac{s\lambda}{\kappa_T^2} \in \mathscr{O}_p(T^{3/2-a-d})$, the upper bound from (3), tends to 0. $\frac{s\lambda}{\kappa_T^2} \to^p 0$ in turn implies $\frac{s\lambda^2}{\kappa_T^2} \to^p 0$ so that the prediction error tends to 0. The admissible region for a and d is plotted in figure 1.

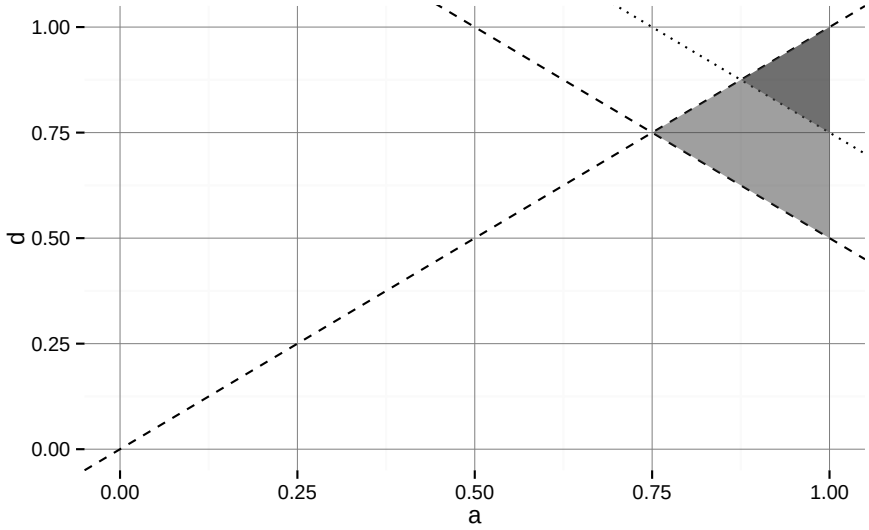

Figure 1: Admissible region for a and d. The whole gray area is the admissible region for the Lasso, with the constraints on a and d represented by the dashed lines. In darker gray is the admissible region for the adaptive Lasso, with the constraint represented by the dotted line.

The speed with which the estimation error tends to zero is $\frac{s\lambda}{\kappa_T^2} \in \mathcal{O}_p(T^{3/2-a-d})$, the line $3/2-a-d=0$ is the top-left to bottom-right dashed line in figure 1. Figure 1 shows clearly that the constraints impose $a \geq \frac{3}{4}$ (in which case we must have $d > \frac{3}{4}$) implying that the number of breaks may grow at most as $T^{1/4}$ in order to have convergence. Moving orthogonally away from the constraint $3/2-a-d=0$ the highest speed of convergence, $T^{-1/2}$, is reached for $a=d=1$ in which case the number of breaks is constant and the distance between them increases at rate T.

For completeness we state an asymptotic counterpart to corollary 1.

Corollary 2. *With probability tending to one, no relevant variables is excluded if there exists a* $T_0 \geq 1$ *such that* $\theta_{\min} > \frac{16s}{\kappa_T^2}\lambda_T$ *for all* $T \geq T_0$.

Corollary 2 is similar to corollary 1 in that it gives a lower bound for the smallest non-zero parameter above which no relevant variables are excluded. This bound tends to zero at the same speed as the estimation error.

3.2. The adaptive Lasso

If we were to penalize more heavily the parameters that are truly equal to zero than those that are different from it, instead of penalizing all parameters by λ_T, we could construct an estimator that is more accurate than the Lasso. The adaptive Lasso of Zou (2006) is based on this idea, using an initial estimator to construct adaptive penalties for each of the parameters. In this setting we use the Lasso both as a screening device and as the initial estimator. The variables that were excluded by the Lasso are not retained in the second stage. We note $(Z^D W_{\widehat{\mathscr{S}}})$ the set of variables retained by the Lasso and $\widehat{\theta}_{\widehat{\mathscr{S}}}$ the corresponding set of estimated parameters, and we construct the adaptive weights w by taking the inverse of the absolute value of the estimated parameters $w = \frac{1}{|\widehat{\theta}_{\widehat{\mathscr{S}}}|}$. The adaptive Lasso objective function is thus

given by:

$$\widetilde{\theta} = \text{argmin}_{\theta_{\widehat{\mathscr{S}}}} \left(\frac{1}{T} \left\| y - (Z^D W_{\widehat{\mathscr{S}}}) \theta_{\widehat{\mathscr{S}}} \right\|^2 + 2\lambda_T \sum_{i \in \hat{\mathscr{S}}} w_i \left| \theta_{\hat{\mathscr{S}},i} \right| \right). \tag{6}$$

The adaptive Lasso objective function is convex and hence fast to minimize, furthermore since the initial estimator discards a large amount of irrelevant variables the adaptive Lasso problem (6) is of much smaller size than (2).

We study the properties of the adaptive Lasso by, as in the case of the Lasso, establishing finite sample results before turning to asymptotic analysis. We focus on the oracle property of the adaptive Lasso, the ability for the estimator to recover the exact model, $\text{sign}(\widetilde{\theta}) = \text{sign}(\theta)$. We make use of the ℓ_1 bound on the estimation error of the Lasso to derive the properties of the adaptive Lasso; we could use other consistent estimators to compute the adaptive weights, the efficiency of the adaptive Lasso depends on that of the initial estimator via the ℓ_1 estimation error $\left\| \widehat{\theta} - \theta \right\|_{\ell_1}$. We now give a finite sample probability and conditions for the adaptive Lasso to be sign consistent.

Theorem 3. *Let* $\lambda_T = \sqrt{\frac{8\ln(1+T)^5 \ln(1+r)^2 \ln(r(T-r+1))\sigma_T^4}{T}}$. *Under assumptions 1, 2, 3, and 5, and assuming that* $\theta_{\min} \geq 2\left\| \widehat{\theta} - \theta \right\|_{\ell_1}$ *and*

$$\left[\frac{sK_T}{\kappa_T^2} \left(\frac{1}{2} + \frac{2}{\theta_{\min}} \right) + \frac{1}{2} \right] \left\| \widehat{\theta} - \theta \right\|_{\ell_1} \leq 1 \tag{7}$$

$$\frac{\sqrt{s}}{\kappa_T^2} \left(\frac{\lambda_T}{2} + \frac{2\lambda_T}{\theta_{\min}} \right) \leq \theta_{\min} \tag{8}$$

with $K_T = \ln(1 + r(T-r+1))^2 \ln(T)\sigma_T^2$. *For some constant* $A > 0$, *on a set with probability at least* $1 - \pi_T^{\mathscr{B}} - \pi_T^{\mathscr{C}}$, *with* $\pi_T^{\mathscr{B}}$ *is as in theorem 1 and* $\pi_T^{\mathscr{C}} = 2T^{-1/A}$, *we have* $\text{sign}(\widetilde{\theta}) = \text{sign}(\theta)$.

The condition $\theta_{\min} \geq 2\left\| \widehat{\theta} - \theta \right\|_{\ell_1}$ ensures that the initial estimator has not discarded any relevant variables, this condition is stronger than necessary, and indeed the 2 could be replaced by some $q > 1$ at the price of more involved notations. (7) illustrates the dependence of the adaptive Lasso on the performance of the initial estimator in the form of $\left\| \widehat{\theta} - \theta \right\|_{\ell_1}$, and indeed (7) can be interpreted as a condition on the performance of the initial estimator. (8) is a condition on $\theta_{\min}$ to ensure that no break is so small as to go unnoticed by the adaptive Lasso.

We now turn to an asymptotic counterpart to theorem 3, where we show that the probability that the adaptive Lasso recovers the correct model tends to one.

Theorem 4. *Let a and d be scalars with* $a, d \leq 1$, $1 - a + d \leq 1$ *and* $7/2 - 2a - 2d < 0$. *Define*

i) $a_T = \ln(T)\ln(1+T)^{5/2}\ln(r(T-r+1))^{1/2}\ln(1+r(T-r+1))^2 T^{7/2-2a-2d}$,

ii) $b_T = \ln(1+T)^{5/4}\ln(r(T-r+1))^{1/4} T^{1/2-a/4-d/2}$,

and let $\theta_{\min} \in \Omega(\ln(T)\max(a_T, b_T))$. *Then, under assumptions 1, 2, 3, and 5, we have:*

$$P(\text{sign}(\tilde{\theta}) = \text{sign}(\theta)) \to 1.$$

Theorem 4 states the conditions under which the adaptive Lasso possesses the oracle property. The rate at which $\theta_{\min}$ is allowed to tend to 0 is bounded from below by functions of a and d so that achieving consistency requires stronger constraints on the parameters than for consistency of the Lasso in theorem 2. The admissible region for the adaptive Lasso has a quarter the area of that of the Lasso as shown in figure 1. The adaptive Lasso can achieve perfect selection with probability tending to one only when the upper bound on the ℓ_1 parameter estimation error of the Lasso, our initial estimator, tends to zero at least as fast as $T^{-1/4}$. This also implies that we require $\theta_{\min}$ to tend to zero slower for the adaptive Lasso than for the Lasso so as to be certain to rule out classification errors.

3.3. Penalty parameter selection

The theorems above give analytical expressions and rates of growth for the penalty parameter λ_T, but do not provide a practical way of selecting it. We suggest selecting the value of λ_T that minimizes the Bayesian Information Criterion (BIC), given by:

$$BIC(\lambda) = T \times \log\left(\frac{\hat{\epsilon}'_\lambda \hat{\epsilon}_\lambda}{T}\right) + \left|\widehat{\mathscr{S}_\lambda}\right| \log(T).$$

BIC is a convenient way to select the penalty parameter since it is easily computable making it fast to find the minimizer of the BIC among the sequence of values of λ_T selected by the estimation algorithm. Let $\widehat{\mathscr{S}_{BIC}}$ denote the set of variables retained by the (adaptive) Lasso using the value of λ selected by the BIC, then theorem 2 in Kock (2014) shows that, in an autoregressive setting, choosing λ_T by BIC leads to consistent variable selection in the sense that $P\left(\widehat{\mathscr{S}_{BIC}} = \mathscr{S}\right) \to 1$. Wang, Li, and Leng (2009) shows similar results in a high dimensional *i.i.d.* setting.

3.4. Post Lasso OLS

By construction the Lasso will select an active set $\widehat{\mathscr{S}}$ for which the smallest eigenvalue of $\frac{W'_{\widehat{\mathscr{S}}} Z^{D'} Z^D W_{\widehat{\mathscr{S}}}}{T}$ is strictly positive, implying that the cardinality of the set of selected variables $s = \left|\widehat{\mathscr{S}}\right|$ is smaller than the number of observations. Hence $Z^D W_{\widehat{\mathscr{S}}_T}$ has rank s and the model $y = Z^D W_{\widehat{\mathscr{S}}} \dot{\theta} + \dot{\epsilon}$ can be estimated by ordinary least squared. This post Lasso OLS has several desirable properties

i) The Lasso biases the estimated non-zero parameters towards zero, the post Lasso provides unbiased and $\sqrt{T}$-consistent estimates of the variables selected by the Lasso. See Belloni, Chernozhukov, et al. (2013) for a formal analysis of the post Lasso OLS.

ii) Standard errors can be computed for the non-zero parameters, however they do not account for the uncertainty in the Lasso step.

iii) Belloni et al. (2013) and Kock and Callot (2015) documents by simulation that the post Lasso OLS improves marginally on the Lasso in terms of estimation and prediction errors.

4. Monte Carlo

In this section we explore the empirical properties of our model using simulated data. We compute 8 statistics for each estimator and experiment, and average them across iterations. A first group of 4 statistics focus on variable selection, a second group of 4 focuses on estimation:

i) The number of breaks (non-zero parameters) estimated, noted # breaks.

ii) The number of variables incorrectly selected (false positive), noted FP.

iii) The number of variables correctly selected (true positive), noted TP.

iv) The number of breaks missed (false negative), noted FN.

v) The estimation error of the path of the parameter $\left\|\widehat{\theta}-\theta\right\|_{\ell_1}$, noted ℓ_1 error.

vi) The prediction error $\left\|Z^D W(\widehat{\theta}-\theta)\right\|$, noted ℓ_2 error.

vii) The root mean square error $\|\hat{\epsilon}\|$ which, in a well specified model, converges towards the variance of the innovations, noted RMSE.

viii) The size of the penalty parameter λ, noted λ.

We report tables with the 8 statistics enumerated above for a variety of experiments. We also plot the true parameter path and a sample of estimated parameter paths for different estimators to give a sense of the location and amplitude of the breaks in the estimated paths relative to the true parameter paths. In these experiments we choose not to penalize the estimator of the initial value.

The estimators we consider are the Lasso, the adaptive Lasso with the Lasso as initial estimator, and the post Lasso OLS. The penalty parameter λ_T for both the Lasso and the adaptive Lasso is selected by minimizing the BIC. The data generating process for the simulations is $y = \beta X + \epsilon$ where X is generated by drawing from a standard normal distribution, ϵ is Gaussian with mean 0, variance 0.1 (except when specified otherwise), and is independent from X.

All the computations are carried out using `R` and the `parsimonious` package which permits easy replication[1] of the simulations and empirical application. The estimation of these models is fast, each iteration takes in the order of 10^{-3} seconds in most cases and around 0.5 second for the hardest model, using commodity hardware.

4.1. Deterministic paths

We first consider the case of a constant parameter equal to 1. For this experiment we consider 2 sample sizes, $T = 100$ and $T = 1000$, and 3 variances for the residuals, $\sigma_\epsilon^2 = 0.1, 1, 10$. This experiment allows us to investigate the behaviour of our estimators in a setting with a constant parameter, and investigate the effect of modifying the noise to signal (n2s) ratio on the estimators.

Table 1 reports the value of 5 out of the 8 statistics, the number of false positives, false negatives, and true positives being uninformative in a setting with no breaks. Since the active set of the initial estimator (Lasso) is often empty, no breaks are detected, the adaptive Lasso frequently cannot be estimated so we do not report results for this estimator. This table reveals that the Lasso incorrectly selects on average 0.1 breaks per models when $T = 100$ (0.01 when $T = 1000$), implying that at least in the order of 90% of the models (99% for $T = 1000$) correctly estimate a constant parameter. The number of breaks selected is not very sensitive to the noise to signal ratio in contrast to the error measures. The RMSE is close to, but on average smaller than, the standard error of the innovations (the true values are $\approx 0.316, 1, \approx 3.16$) for

[1] Replication files can be found at `https://github.com/lcallot/ptv-var`.

	n2s ratio $\frac{\sigma_\epsilon^2}{\sigma_X^2}$	$T=100$			$T=1000$		
		0.1	1	10	0.1	1	10
# breaks	DGP	0	0	0	0	0	0
	Lasso	0.138	0.093	0.095	0.006	0.012	0.013
	aLasso	0.121	0.086	0.088	0.006	0.011	0.012
ℓ_1 error	Lasso	0.153	0.272	0.483	0.083	0.147	0.264
	aLasso	-	-	-	-	-	-
	Post	0.157	0.28	0.498	0.083	0.148	0.266
ℓ_2 error	Lasso	0.028	0.088	0.279	0.008	0.026	0.083
	aLasso	-	-	-	-	-	-
	Post	0.031	0.097	0.308	0.008	0.026	0.084
RMSE	Lasso	0.311	0.985	3.108	0.316	0.998	3.158
	aLasso	-	-	-	-	-	-
	Post	0.311	0.984	3.106	0.316	0.998	3.158
λ	Lasso	0.023	0.073	0.228	0.008	0.026	0.083
	aLasso	-	-	-	-	-	-

Table 1: Constant parameter, varying sample size: 10000 iterations. The adaptive Lasso results are not reported since the Lasso often excludes every variables preventing us from estimating the adaptive Lasso.

$T=100$; the RMSE is closer to its theoretical value when $T=1000$. This under-evaluation of the RMSE, overfitting, can be attributed to the spurious inclusions of breaks in the estimated parameter path. The noise to signal ratio has a large influence on the ℓ_1 and ℓ_2 errors, they both increase with the noise-to-signal ratio but fall when going from $T=100$ to $T=1000$. Interestingly while the RMSE of the post Lasso OLS is identical or slightly smaller than that of the Lasso, it appears that the ℓ_1 and ℓ_2 errors of the post Lasso OLS are marginally larger than those of the Lasso.

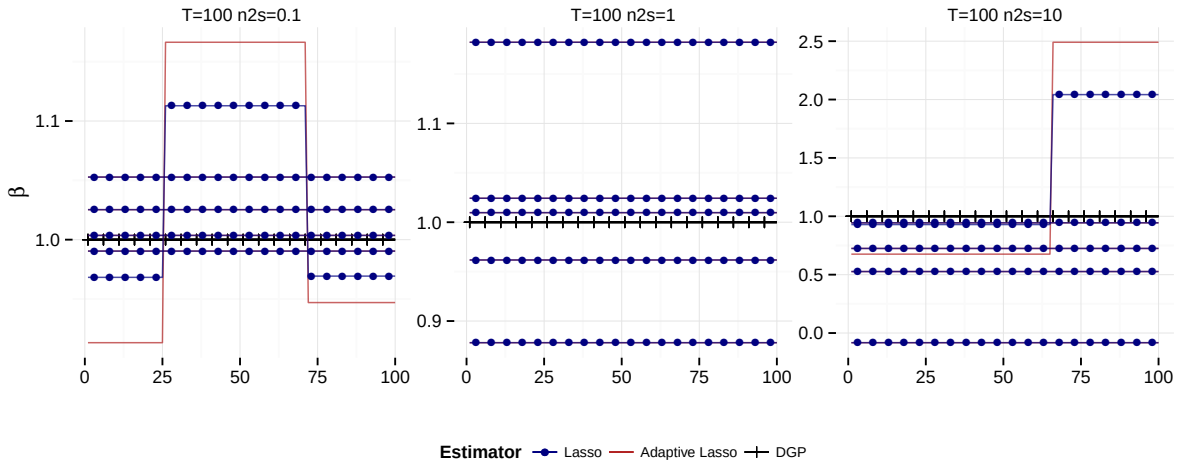

Figure 2: DGP parameter path (black crossed line) and a sample of 5 paths estimated with the Lasso (blue dotted line) with corresponding adaptive Lasso (red solid line) when feasible.

Figure 2 plots 5 estimated paths for the Lasso and adaptive Lasso (note that when no breaks are selected in the first step the adaptive Lasso is not estimated) highlighting that the vast majority of the estimated paths are constant. This figure also shows that despite the downward bias introduced by the Lasso, the estimated paths cluster around the true value, with few instances of large estimation errors on both sides of the true value. Figure 2 also reveals that when the Lasso incorrectly selects a break in the parameter path, it often selects more than one, this is consistent with the selection of a low penalty parameter λ in these iterations. This implies that the average number of breaks is an upper bound on the number of estimated paths with non-constant parameters. The adaptive Lasso tends to reduces the number of irrelevant breaks selected by the Lasso but only marginally since the breaks incorrectly retained are large.

		Break Location			Break Size			Break Number		
		10%	50%	90%	0.1	1	10	3	9	4
# breaks	DGP	1	1	1	1	1	1	3	9	4
	Lasso	3.66	3.315	3.386	0.322	3.325	3.901	8.397	19.45	10.45
	aLasso	1.501	1.305	1.423	-	1.314	1.005	3.913	11.71	5.354
FP	Lasso	2.895	2.529	2.666	0.304	2.538	2.91	6.062	12.74	7.519
	aLasso	0.882	0.673	0.808	-	0.682	0.131	1.987	5.968	2.858
TP	Lasso	0.765	0.786	0.72	0.017	0.787	0.991	2.335	6.71	2.931
	aLasso	0.619	0.632	0.615	-	0.632	0.874	1.927	5.74	2.497
FN	Lasso	0.235	0.214	0.28	0.983	0.213	0.009	0.665	2.29	1.069
	aLasso	0.381	0.368	0.385	-	0.368	0.126	1.073	3.26	1.503
ℓ_1 error	Lasso	0.249	0.256	0.248	0.22	0.256	0.285	0.333	0.439	0.353
	aLasso	0.214	0.212	0.213	-	0.212	0.249	0.279	0.397	0.31
	Post	0.262	0.253	0.254	0.234	0.252	0.27	0.343	0.501	0.383
ℓ_2 error	Lasso	0.088	0.079	0.087	0.056	0.079	0.089	0.123	0.187	0.145
	aLasso	0.065	0.058	0.064	-	0.058	0.066	0.094	0.162	0.117
	Post	0.08	0.073	0.078	0.062	0.073	0.074	0.113	0.18	0.136
RMSE	Lasso	0.31	0.309	0.31	0.313	0.309	0.313	0.306	0.3	0.307
	aLasso	0.301	0.307	0.302	-	0.307	0.316	0.299	0.278	0.293
	Post	0.303	0.304	0.303	0.312	0.303	0.304	0.29	0.264	0.286
λ	Lasso	0.017	0.023	0.017	0.027	0.023	0.028	0.013	0.007	0.009
	aLasso	0.007	0.008	0.002	-	0.008	3.844	0.05	0.082	0.026

Table 2: Structural breaks experiments, $T = 100$, 10000 iterations. We do not report the adaptive Lasso estimator for the experiment with a break of size 0.1 since the initial estimator often discard all variables.

We now turn to the case of deterministic breaks (structural breaks) in the parameters and consider 3 types of experiments. In the first experiment we consider a single break in the parameter path occurring at either 10%, 50%, or 90% of the sample. In the second series of experiments a single break, located in the middle of the sample, varies in size, the size of the break being either 0.1, 1, or 10. In the third series of experiments we vary the number of structural breaks in the path. The parameter value switches between 0 and 1, this can be seen as a minimalistic regime switching process. In these series of experiments we hold the sample size constant ($T = 100$ throughout) as well as the variance of the innovations $\sigma_\epsilon^2 = 0.1$

while the covariates are still drawn from a standard normal distribution. Notice that the first 4 blocks of rows of table 2 now show detailed variable selection statistics.

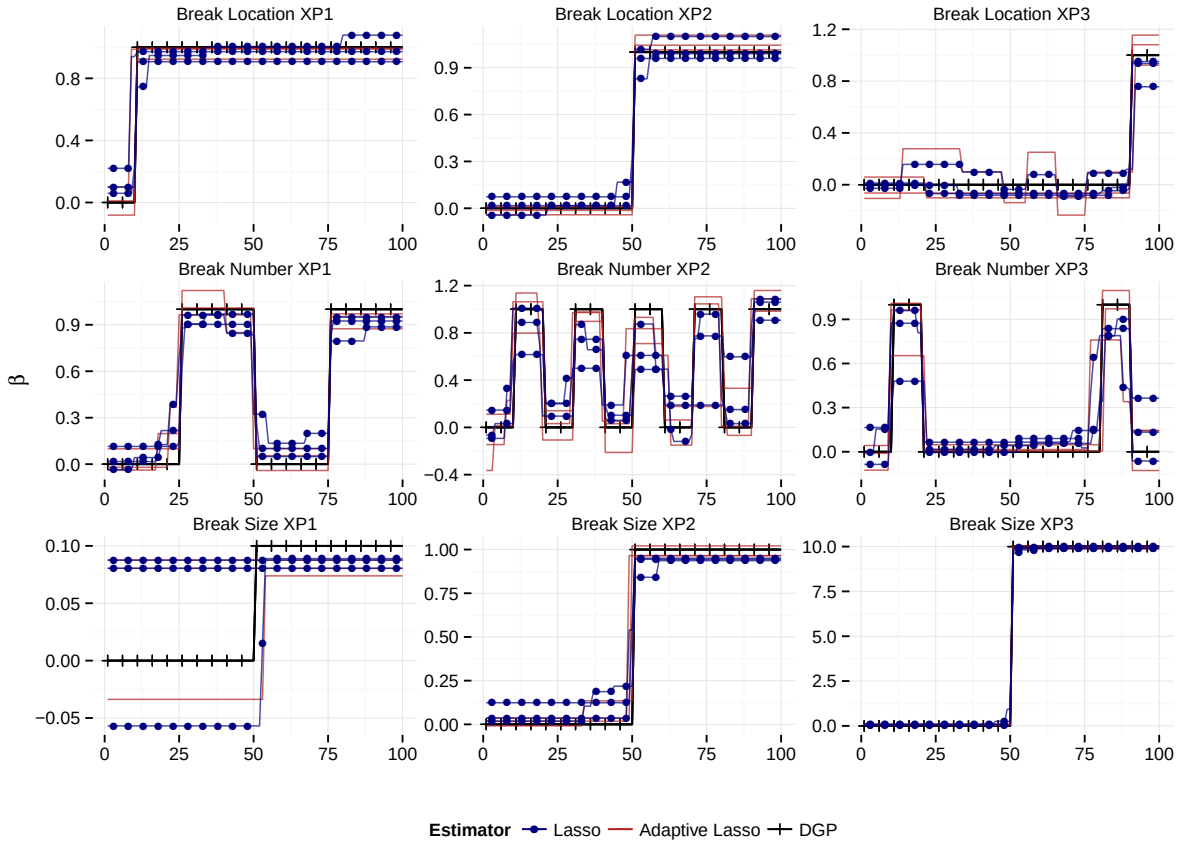

Figure 3: Structural breaks. DGP parameter path (black crossed line) and a sample of 3 paths estimated with the Lasso (blue dotted line) with corresponding adaptive Lasso (red solid line) when feasible.

In all experiments the Lasso selects, on average, models that are larger than the true model, except in the case when the break size is 0.1. The adaptive Lasso further reduces the model. As the results of the experiments on break locations and sizes illustrate, the (adaptive) Lasso is not very sensitive to the location of the break point but is sensitive to its amplitude. When the break is of size 10, the Lasso and adaptive Lasso detect a break in the correct location in 99% and 87% of the iterations. These rates fall below 2% when the break is of size 0.1. The rate of rejection of relevant variables (false negative, FN) is similarly not very sensitive to the location of the break but is sensitive to its size, with FN $< 1\%$ when the break is of size 10 while FN $\approx 98\%$ when it is of size 0.1.

The break size and location experiments also reveal that the Lasso is an efficient screening device, out of 98 irrelevant variables the number of true negatives TN $= 98 -$ FP is greater than 95 for the Lasso. In these experiments the estimated models contain on average fewer than 4 variables (fewer than 2 for the adaptive Lasso); this set contains the true location of the break in over 70% of the iterations in most settings.

The ℓ_1 and ℓ_2 errors are comparable across experiments, neither the location nor the amplitude of the break seem to have a systematic impact on these measures. Both the adaptive Lasso and the post Lasso OLS reduce the prediction and estimation errors in most experiments but these improvements are marginal. The RMSE is stable across experiments and estimators, being always close to its theoretical minimum of $\sqrt{0.1} \approx 0.316$.

The experiments varying the number of breaks, right columns of table 2, show that when we increase the number of breaks in the DGP the selected model is larger leading to a higher number of false positives while keeping the number of true positives close, but inferior, to the true number of breaks. In these settings the Lasso is not as efficient at discarding irrelevant variables as it was in the previous, sparser, experiment; the adaptive Lasso is here a useful second step since it further reduces the size of the active set and improves upon the Lasso on all the error measures. However, this comes at the price of a slight decline in the true positive rate.

Figure 3 plots 3 estimated paths for each experiment. The top 3 panels illustrate the break location experiments, the stability of the estimated paths away from the region of the break is striking. The figure also reveals that some paths follow a gradual adjustment with several breaks instead of a single one. The lower 3 panels show the break size experiments in which it appears that when the break is small it is often ignored (bottom left panel), whereas a very large break will be often detected and adjusted to in a single step even though evidence of gradual adjustment for some paths persists.

4.2. Stochastic paths

We now turn to simulations with stochastic paths, the results are reported in table 3. The parameters follow parsimonious random walks as described by assumption 3. We vary the degree of sparsity of the model by considering $\alpha_T = 0.01, 0.031, 0.1$ (corresponding to $a = 1, 0.75, 0.5$), where α_T is the probability of a break at each point. We also consider 3 variances for the non zero increments: $\text{Var}(\eta) = 0.1, 1, 10$, for a total of 9 experiments.

In the experiments with $\text{Var}(\eta) = 0.1$ the Lasso tends to select models that are sparser than the DGP, and the Lasso only detects around 15% of the correct break locations. However when $\text{Var}(\eta) = 1$ or $\text{Var}(\eta) = 10$ the selected models tend to be larger than the true models, and over 50% of the breaks are detected. When the models selected by the Lasso are larger than the true model, the models selected by the adaptive Lasso in the second step have dimensions close to those of the DGP. When the models selected by the Lasso are sparser than the DGP, the adaptive Lasso tends to select models that are even more sparse.

The ℓ_1 and ℓ_2 errors do increase with the variance of η and with the number of breaks, and the adaptive Lasso and post Lasso OLS are not consistently better or worse than the Lasso on these measures. The RMSE is close to, but below, its theoretical value (≈ 0.316) for most experiments, showing that overfitting is not excessive despite the flexible nature of the model. The RMSE is in most instances lower for the adaptive Lasso and the post Lasso OLS.

Interestingly the chosen penalty parameter λ decreases while α_T increases for the Lasso, but increases with α_T for the adaptive Lasso. This can be explained by the fact that the number of potential parameters is constant for the Lasso, while it is increasing with α_T for the adaptive Lasso since the Lasso selects increasingly larger models. For the Lasso the selected penalty parameter also decreases when $\text{Var}(\eta)$ increases.

Figure 4 provides complementary information on the dynamics of the selected models, it displays a sample of 3 true and estimated parameter paths from each of the Monte Carlo experiments in table 3. The left side panels of figure 4 display experiments where the variance of the innovations to the parameters is low. The Lasso tends to discard a large amount of small breaks only adjusting to large, and persistent, changes in the parameter value. The estimated paths are increasingly time varying when the number of breaks increases (moving downward in figure 4) but more stable than the true path. When the variance of the breaks increases (moving rightwards in figure 4) the paths are increasingly close to the true path, displaying a high degree of time variation when this is the case for the true path.

		Var(η) = 0.1			Var(η) = 1			Var(η) = 10		
	α_T	0.01	0.031	0.1	0.01	0.031	0.1	0.01	0.031	0.1
$a = -\frac{\log(\alpha_T)}{\log(T)}$		1	0.75	0.5	1	0.75	0.5	1	0.75	0.5
# breaks	DGP	0.992	3.063	9.975	1.008	3.063	9.859	0.983	3.102	9.823
	Lasso	0.992	2.614	6.33	2.251	6.002	14.77	3.091	8.641	21.05
	aLasso	0.561	1.388	3.307	0.992	2.681	7.347	1.038	2.984	8.744
FP	Lasso	0.824	2.1	4.597	1.752	4.479	9.877	2.323	6.243	13.61
	aLasso	0.436	1.021	2.15	0.584	1.482	3.605	0.393	1.006	2.689
TP	Lasso	0.168	0.515	1.733	0.499	1.523	4.888	0.768	2.397	7.442
	aLasso	0.126	0.367	1.157	0.408	1.2	3.742	0.645	1.978	6.055
FN	Lasso	0.824	2.548	8.242	0.509	1.541	4.971	0.215	0.704	2.381
	aLasso	0.867	2.696	8.818	0.6	1.864	6.117	0.338	1.124	3.767
ℓ_1 error	Lasso	0.205	0.263	0.335	0.22	0.292	0.387	0.23	0.32	0.453
	aLasso	0.24	0.267	0.332	0.236	0.276	0.371	0.235	0.287	0.415
	Post	0.209	0.27	0.352	0.227	0.311	0.448	0.24	0.349	0.518
ℓ_2 error	Lasso	0.056	0.089	0.133	0.066	0.108	0.166	0.071	0.124	0.21
	aLasso	0.076	0.092	0.131	0.075	0.098	0.155	0.072	0.106	0.184
	Post	0.058	0.089	0.133	0.065	0.105	0.162	0.067	0.114	0.184
RMSE	Lasso	0.312	0.313	0.316	0.311	0.309	0.306	0.31	0.31	0.32
	aLasso	0.304	0.305	0.308	0.305	0.305	0.301	0.31	0.315	0.325
	Post	0.31	0.308	0.306	0.307	0.3	0.287	0.305	0.296	0.279
λ	Lasso	0.024	0.023	0.019	0.022	0.017	0.01	0.02	0.015	0.01
	aLasso	0.004	0.004	0.016	0.013	0.031	0.096	0.228	2.917	8.53

Table 3: Parsimonious random walks, $T = 100$, 10000 iterations.

4.3. Autoregressions

We finally consider an AR(1) model with a break in the autoregressive parameter. The data is generated using $y_t = \gamma_t y_{t-1} + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, 0.1)$, where $\gamma_t = \gamma^0$ if $1 \leq t \leq T/2$ and $\gamma_t = \gamma^1$ if $T/2 < t \leq T$ and $T = 100$. We evaluate the performances of our estimator using every combinations of $\gamma^0 = 0, 0.5, 0.9$ and $\gamma^1 = 0.5, 0.9$. Two out of the six experiments have a stable autoregressive parameter either equal to 0.5 or to 0.9, which can be seen as the benchmarks for the autoregressive experiments.

When estimating the AR(1) models, in a non negligible number of iterations the Lasso selects a saturated model ($|\hat{\mathscr{S}}| = T$). To remedy this problem we impose that $|\hat{\mathscr{S}}| \leq T/2 = 50$. This relatively large upper bound has the effect of eliminating the instances where the Lasso selects a saturated model without affecting the other estimates.

The results of the experiments with AR(1) models are reported in table 4. In the experiments where no breaks is present the Lasso mistakenly selects breaks in some iterations. More false positives occur when the persistence is high, 0.51 breaks on average when $\gamma_t = 0.9 \forall t$, than when $\gamma_t = 0.5 \forall t$ (0.08 breaks on average). It is interesting to contrast these false positive rates with those found in the leftmost column of table 1 where the model has the same dimensions as in the autoregressive case and the innovations follow the same distribution. In the setting of table 1 where the covariates are exogenous, false positives occur more frequently

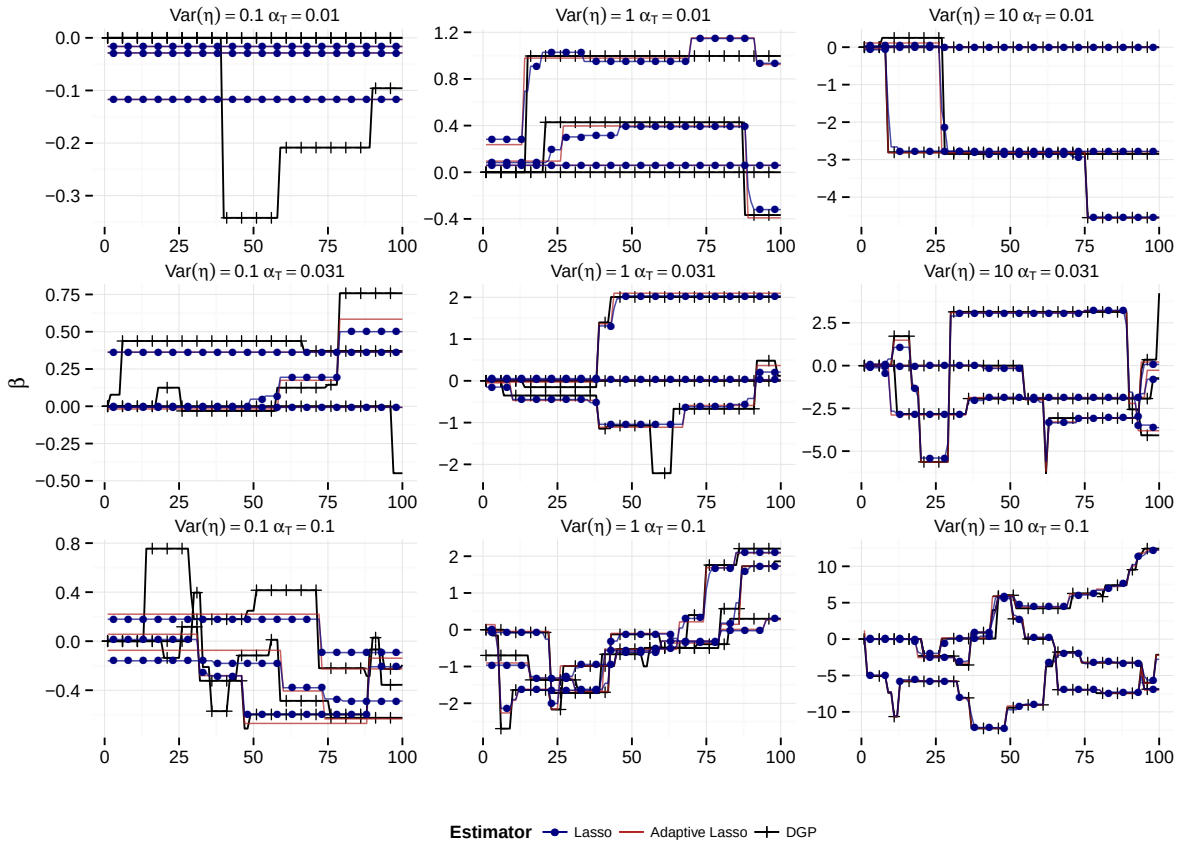

Figure 4: Parsimonious random walks. DGP parameter path (black crossed line) and a sample of 3 paths estimated with the Lasso (blue dotted line) with corresponding adaptive Lasso (red solid line) when feasible.

than in the AR case when $\gamma_t = 0.5 \forall\, t$.

The Lasso has a low rate of true positives (at most 0.29) and tends to underestimate the number of breaks in models where breaks do occur, the exception being $\gamma^0 = 0,\ \gamma^1 = 0.9$ in which case the Lasso selects 2.6 breaks. In that case the adaptive Lasso helps reduce the dimension of the model to 1.3 on average, whereas for the other models with a break the adaptive Lasso increases the underevaluation of the number of changes in the autoregressive parameter.

5. Empirical application

Whether US monetary policy in the second half of the 20$^{\text{th}}$ century has followed a stable model or not is a controversial issue, with empirical evidence supporting both views as discussed in Sims and Zha (2006). We use the parsimoniously time varying model introduced in this paper to investigate this issue. To do so we use the Taylor rule (Taylor, 1993) as describing the monetary policy response to economic conditions. According to the Taylor rule, the policy rate of the central bank can be decomposed into two parts: a response to changes in the inflation rate; and a response to deviations of output from its trend.

5.1. Models

Hansen, Lunde, and Nason (2011) illustrate the model confidence set (MCS) by estimating a large number of specifications of (backward looking) Taylor rules. We choose to estimate

		$\gamma^0=0$ $\gamma^1=0.5$	$\gamma^0=0$ $\gamma^1=0.9$	$\gamma^0=0.5$ $\gamma^1=0.5$	$\gamma^0=0.5$ $\gamma^1=0.9$	$\gamma^0=0.9$ $\gamma^1=0.5$	$\gamma^0=0.9$ $\gamma^1=0.9$
# breaks	DGP	1	1	0	1	1	0
	Lasso	0.752	2.647	0.086	0.729	0.993	0.408
	aLasso	0.558	1.326	0.08	0.537	0.721	0.336
FP	Lasso	0.688	2.361	-	0.684	0.868	-
	aLasso	0.507	1.129	-	0.499	0.606	-
TP	Lasso	0.064	0.286	-	0.045	0.125	-
	aLasso	0.051	0.197	-	0.037	0.115	-
FN	Lasso	0.936	0.714	-	0.955	0.875	-
	aLasso	0.949	0.803	-	0.963	0.885	-
ℓ_1 error	Lasso	0.433	0.432	0.251	0.398	0.391	0.213
	aLasso	0.367	0.353	0.468	0.336	0.33	0.398
	Post	0.445	0.407	0.258	0.411	0	0
ℓ_2 error	Lasso	0.074	0.088	0.028	0.083	0.084	0.037
	aLasso	0.062	0.069	0.1	0.067	0.078	0.118
	Post	0.079	0.078	0.031	0.087	0	0
RMSE	Lasso	0.314	0.311	0.312	0.315	0.314	0.31
	aLasso	0.305	0.306	0.297	0.305	0.303	0.287
	Post	0.312	0.305	0.311	0.313	0.311	0.309
λ	Lasso	0.01	0.012	0.009	0.015	0.017	0.016
	aLasso	0	0.005	0	0	0	0.014

Table 4: Autoregressive process, T=100, 10000 iterations.

parsimoniously time varying parameter versions of two specifications of the Taylor rule included in the MCS, resulting in the following 3 models:

$$R_t = (1-\rho)\left[\gamma + \alpha_t \pi_{t-1} + \beta_t y_{t-1}\right] + \rho R_{t-1} + v_t, \tag{9}$$

$$R_t = (1-\rho_t)\left[\gamma + \alpha_t \pi_{t-1} + \beta_t y_{t-1}\right] + \rho_t R_{t-1} + v_t, \tag{10}$$

$$R_t = (1-\rho)\left[\gamma + \alpha_{1,t} \pi_{t-1} + \alpha_{2,t} \pi_{t-2} + \beta_{1,t} y_{t-1} + \beta_{2,t} y_{t-2}\right] + \rho R_{t-1} + v_t, \tag{11}$$

where R_t denotes the short-term nominal interest rate, π_t is inflation, and y_t is deviations of output from its trend (i.e. the output gap). Notice that the difference between (9) and (10) is whether ρ is allowed to vary over time or not.

The parameters of main interest are the ones associated with the response of interest rates to the inflation and output variables. The response to inflation is given by α_t or $\alpha_{1,t} + \alpha_{2,t}$. The Taylor principle suggests that the response to inflation should exceed 1 such that a rise in inflation results in comparatively larger rise in the interest rate. The monetary policy response to real side fluctuations is given by β_t or $\beta_{1,t} + \beta_{2,t}$; this response should be positive so when output is below trend, the interest rate decreases. In all our specifications we let these key parameters be time-varying to examine whether these responses have changed over time.

All specifications contain the lagged interest rate which, as discussed by Hansen et al. (2011), can be interpreted as interest rate smoothing by the central bank or alternatively as

a proxy for unobserved determinants of the interest rate. When estimating (9), (10), or (11), the parameter associated with the lagged interest rate, ρ or ρ_t, also enters the parameters associated with inflation and the output gap making it difficult to disentangle time variations stemming from changes in the response to economic conditions from time variations stemming from changes in the persistence of R_t. We choose to assume that ρ does not vary over time in (9) and (11) in order to focus on changes in the response to inflation and the output gap. We allow ρ_t to vary in (10) for comparison and to examine whether changes in the persistence of monetary policy occurred during the sample we investigate.

5.2. Estimation and data

We estimate models (9) to (11) using the Lasso, and OLS assuming constant parameters. We do not report the adaptive Lasso estimates as they are similar to the Lasso. We also report results for models (9) to (11) estimated with the Lasso under the constraint that there are at most 16 ($\approx \sqrt{T}$) changes in the parameter values. This constraint is added to reduce the number of false positives that might be induced by heteroskedasticity. We refer to this estimator as constrained Lasso henceforth.

The estimator of the initial value of the parsimonious random walk is not penalized so that when no breaks is selected in the model, the Lasso is identical to OLS. Confidence intervals for the OLS estimates are based on heteroskedasticity and autocorrelation consistent standard errors. The penalty parameter, λ, is selected using the BIC. We always include an non-penalized intercept in the estimated model, we experimented with a parsimoniously time-varying intercept but in all specifications no breaks were found in the intercept and hence we do not report results for this case.

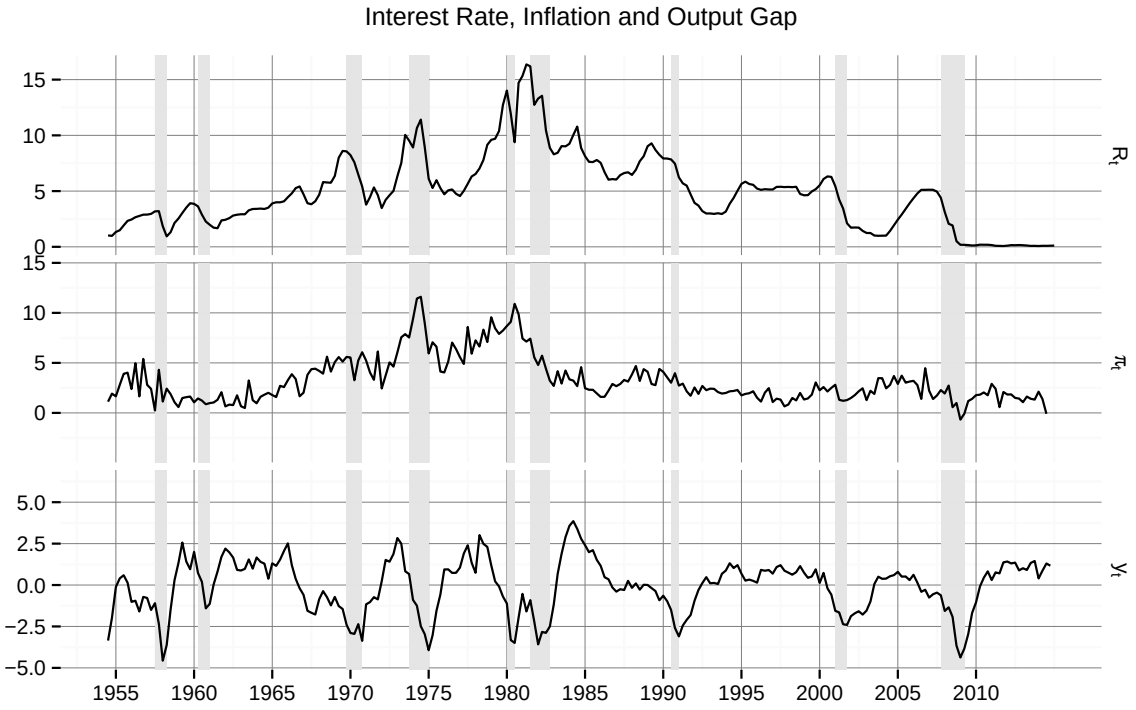

Figure 5: Plots of the data used for estimation of the Taylor rule. The variables are: Interest rate, R_t, inflation, π_t, and output gap, y_t. The vertical grey bars are the NBER recessions.

We use the same variables as Hansen et al. (2011), but for a longer timespan covering 1954:Q4–2014:Q2. For the dependent variable we use the *Effective Fed Funds Rate* aggregated

to quarterly frequency and measured at an annual rate, $R_{\text{effr},t}$, and then define: $R_t = 100 \times \log(1 + R_{\text{effr},t}/100)$. The inflation measure is based on the seasonally adjusted *Implicit GDP Deflator*, P_t, with inflation defined as: $\pi_t = 400 \times \log(P_t/P_{t-1})$. Finally, the output gap measure is based on *Real GDP in Billions of Chained 2009 Dollars*, Q_t, where $y_t = \log Q_t - \text{trend } Q_t$ and trend Q_t is obtained by applying a one-sided Hodrick-Prescott filter to $\log Q_t$. All data have been obtained from the FRED database at the Federal Reserve Bank of St. Louis, and plots of the variables are given in figure 5.

5.3. *Results*

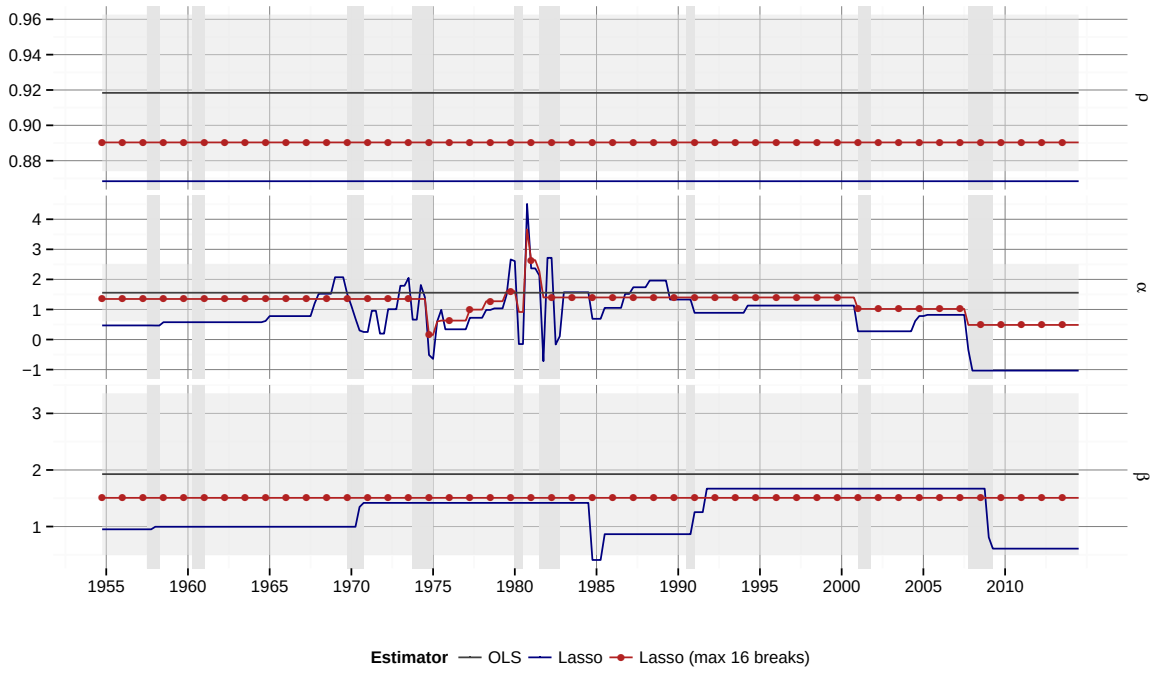

Figure 6: Parameters of the Taylor rule (9). The horizontal grey bars are 90% confidence bands for the OLS estimates. The vertical grey bars are the NBER recessions.

We first consider the Lasso and OLS estimates of model (9) plotted in figure 6. The upper panel shows $\hat{\rho}$, which we assumed to be constant over time, and indicates that R_t is very persistent. The middle panel shows the estimated response to inflation. The unconstrained Lasso estimates of α_t (blue solid line) exhibits a high degree of time variation in the 1970s and 1980s as does, over a more restricted period, the Lasso constrained to a maximum of 16 breaks (red dotted line). These estimates indicate that the response to inflation became weaker in the second half of the 1970s, weaker than the response warranted by the Taylor principle. A strengthening of the response to inflation takes place from the late 1970s, reaching its peak in 1981 before mostly stabilizing from the mid 1980s onward. Both Lasso estimators find a weakening of the response to inflation starting in 2007, which is consistent with the fact that interest rates have been close to the lower bound from that time to the end of our sample.

The response to the output gap (lower panel of figure 6) is more stable, and even found to be constant by the constrained Lasso. The unconstrained Lasso finds a drop in the response to the output gap starting during the 2007–2009 recession, corresponding with the period where interest rates fell to 0.

Figure 7 report the estimates of model (10), where ρ_t is allowed to vary over time. To isolate changes occurring in the parameter associated with the lagged interest rate to those

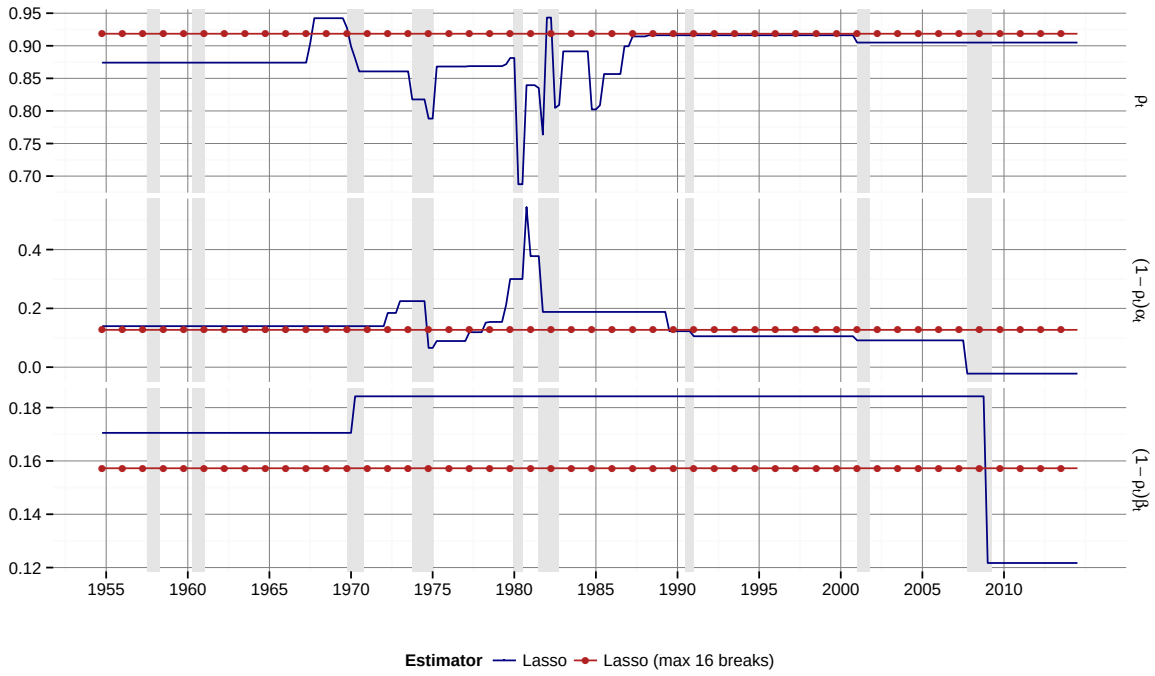

Figure 7: Parameters of the Taylor rule (10) with time varying ρ. Note that we report $\hat{\rho}_t$ as well as $\widehat{(1-\rho_t)\alpha}_t$ and $\widehat{(1-\rho_t)\beta}_t$ to separate breaks in the persistence parameter from breaks in the responses to inflation and the output gap. The vertical grey bars are the NBER recessions.

occurring in the parameters associated with inflation and the output gap, figure 7 reports the untransformed estimates of our model, that is, $\hat{\rho}_t$, $\widehat{(1-\rho_t)\alpha}_t$, and $\widehat{(1-\rho_t)\beta}_t$. The constrained Lasso selects no breaks so that it is identical to the OLS estimates. The unconstrained Lasso finds a drop in the persistence of the interest rate parameter from 1970 to 1985 indicating that monetary policy has been less persistent during that period. The patterns found for the other variables are similar to those found in figure 6. The response to inflation becomes weaker during the mid 1970s, becomes strong in the early 1980s, and stabilizes afterwards, and the response to the output gap is mostly stable and drops at the end of the sample.

Figure 8 and 9 report the estimated responses to inflation and the output gap for model (11). The top and middle panels show the parameters associated with the first and second lags while the bottom panel shows the sum of both parameters, thus giving the estimate of the total response to inflation (figure 8) or to the output gap (figure 9).

The response to inflation follows a pattern similar to those of figures 6 and 7, it is found to be stable from the beginning of the sample to the start of the 1970s, and again from the mid 1980s until 2008, with a response $\alpha_{1,t} + \alpha_{2,t}$ close to 2. There is clear evidence of instability from the start of the 1970s to the start of the 1980s, both with the Lasso and the constrained Lasso. A first period during the 1970s is characterized by a weak monetary policy response in the face of increasing inflation, in particular in the second half of that decade where it is below 1. The response then increases sharply before stabilizing close to 2.

Figure 9 gives the same illustration for the parameters associated with the output gap variables. The response to output gap is much more stable, and indeed entirely stable in the case of the constrained Lasso. The response to the output gap seems to be higher from 1990 to 2008 than it is in the rest of the sample

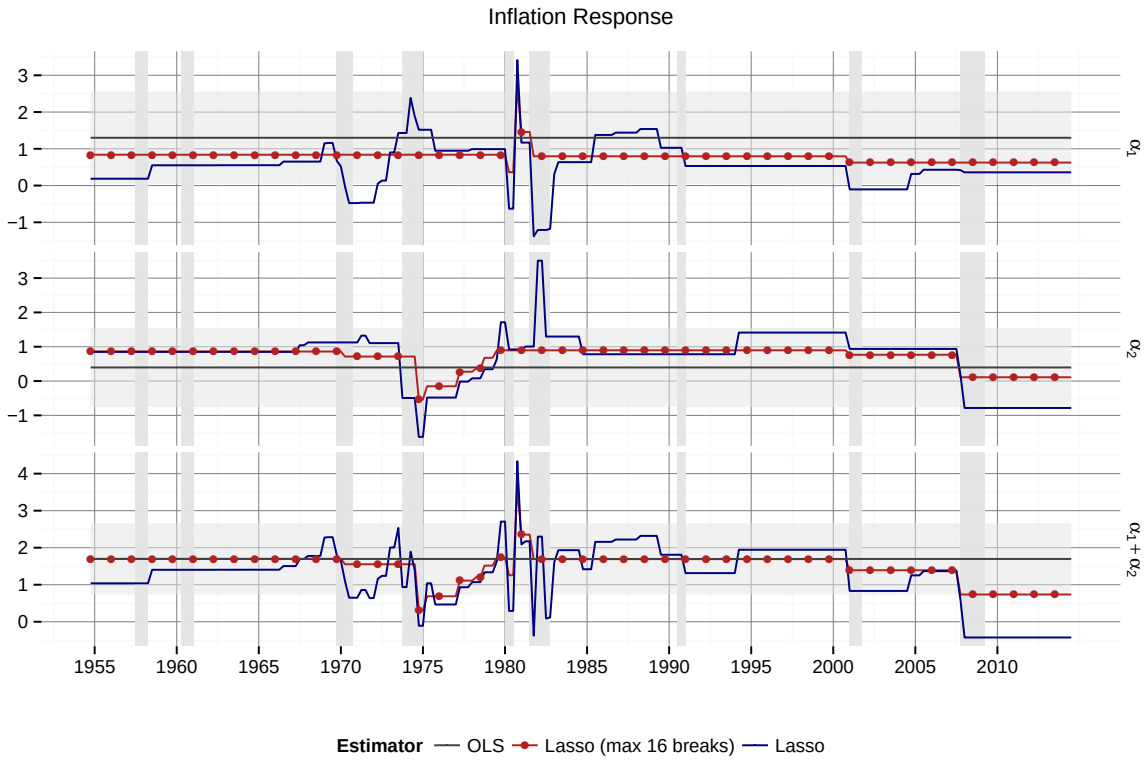

Figure 8: Parameters associated with the inflation variables in the Taylor rule (11). The horizontal grey bars are 90% confidence bands for the OLS estimates. The vertical grey bars are the NBER recessions.

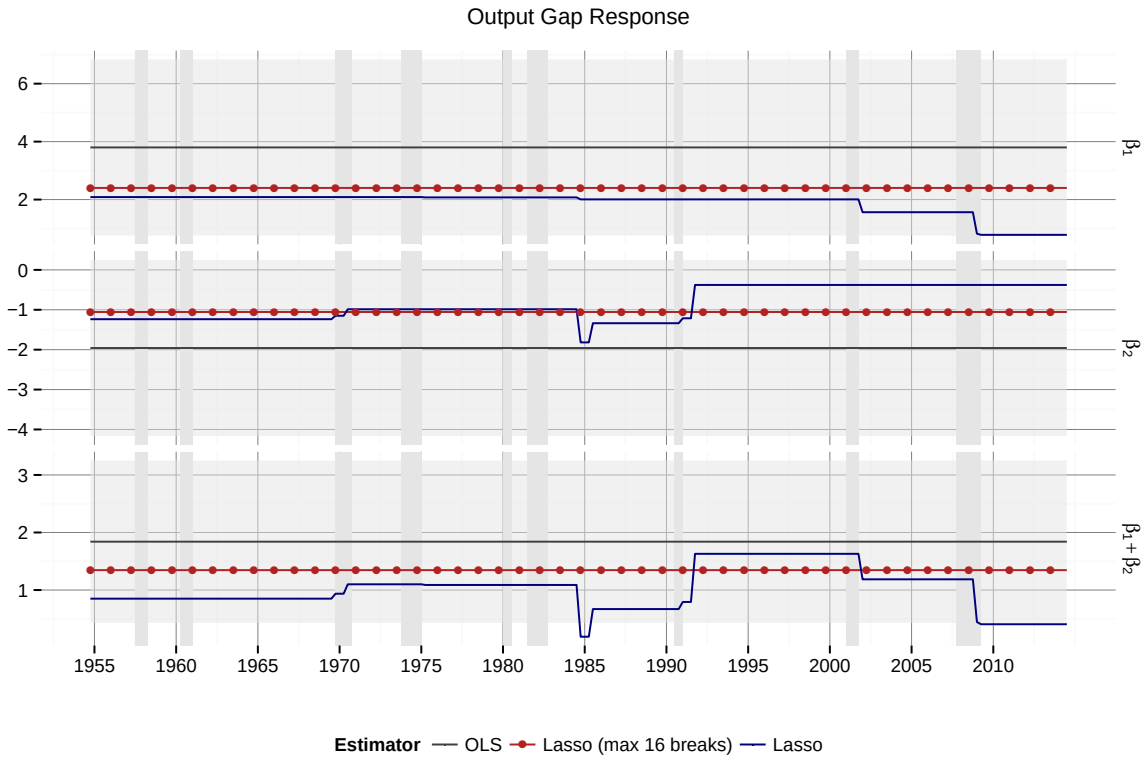

Figure 9: Parameters associated with the output gap variables in the Taylor rule (11). The horizontal grey bars are 90% confidence bands for the OLS estimates. The vertical grey bars are the NBER recessions.

5.4. Interpretation

A few patterns emerge from our results, most importantly we find robust evidence of instability in the response of interest rates to changes in inflation with a period of weak response, sufficiently weak in fact that the Taylor principle is not satisfied, starting around 1973 and lasting for 5 to 8 years. Monetary policy was gradually tightened before reaching extremes after the recession of 1980. After that decade of instability, as before it, monetary policy appears to have been mostly stable according to the Taylor rule.

Instability in US monetary policy during the 1970s and 1980s is well documented. Primiceri (2005) finds that the systematic component of the monetary policy response to inflation has grown more aggressive from the 1960 onward with considerable instability around this trend, particularly in the 1970s and 1980s. Our findings are also consistent with Boivin and Giannoni (2006) who find that monetary policy was more stable in the post 1980 period than in the two preceding decades. Using a regime switching model Sims and Zha (2006) documents frequent regime switches in the 1970s and 1980s, and in particular around 1980 where we also find the policy response to have been extremely aggressive for a short period.

While empirical evidence in favour of substantial changes occurring in monetary policy in the second half of the 20^{th} century is numerous, evidence to the contrary also abounds as reviewed by Sims and Zha (2006). Our empirical evidences could be viewed as conciliating both views as, even though we find substantial instability in the 1970s and 1980s, the parameters of the models are very similar, and stable, in the periods preceding and following the period of instability.

Another pattern emerging from our analysis is the disconnection between the interest rate and both inflation and the output gap from 2007 captured by the drop in the value of the parameters relating to economic conditions. This reflects the inability of the Taylor rule to describe monetary policy when interest rates are at the zero lower bound.

6. Conclusion

This paper proposes the parsimoniously time-varying parameter VARX model, and investigates the properties of the Lasso as an estimator for this model. We propose a process for the parameters, the parsimonious random walk, where the probability of an increment to the random walk being equal to 0 is greater than 0. This process can accommodate time-varying paths that are constant, exhibit structural breaks, or a large number of changes.

We estimate the vector of increments to the parameters which is sparse by the parsimonious random walk assumption, and high dimensional by construction. We derive bounds on the precision of the Lasso in finite samples, and conditions for asymptotic consistency. We also provide finite sample and asymptotic results on the probability that the adaptive Lasso recovers the true model. Because of the convexity of the Lasso's objective function, our estimator is computationally fast.

We apply our model to the estimation of Taylor rules to investigate the US monetary policy response to inflation from 1954 to 2014. We find evidence of substantial instability in the policy response in the 1970s and 1980s, which is consistent with previous research and historical facts. We also observe a long lasting change in the monetary policy response since 2007, driven by the fact that the Fed Funds Rate has been close to zero since that time. The simulations and empirical results in this paper can easily be replicated using the `parsimonious` package for `R`.

To further develop the parsimoniously time-varying parameter model we see a few directions for future research. First, develop a inferential framework for this model taking advantage

of the construction of the variables ($Z^D W$) to get an accurate estimator for the covariance matrix. Second, render the estimator more robust to heteroskedasticity by constructing adaptive weights that are a function of the local variance of the innovations.

Appendix A. Proofs

A.1. Proofs for the restricted eigenvalue condition

Lemma 2. *For two matrices A and B we have $\underline{\phi}(A'A)\underline{\phi}(B'B) \leq \underline{\phi}(B'A'AB)$*

Proof.

$$\underline{\phi}(A'A) = \min_{u_1} \frac{u_1' A' A u_1}{u_1' u_1} \tag{A.1}$$

so for any vector u_2 we have

$$\begin{aligned} \underline{\phi}(A'A) &\leq \frac{u_2' A' A u_2}{u_2' u_2} \\ \underline{\phi}(A'A) u_2' u_2 &\leq u_2' A' A u_2 \end{aligned} \tag{A.2}$$

specifically choose $u_2 = B u_3$ for any vector u_3

$$\underline{\phi}(A'A) u_3' B' B u_3 \leq u_3' B' A' A B u_3 \tag{A.3}$$

Now, applying the same argument, i.e. (A.1)–(A.2), to $B'B$ we get

$$\underline{\phi}(B'B) u_3' u_3 \leq u_3' B' B u_3$$

Combining this with (A.3) we get

$$\begin{aligned} \underline{\phi}(A'A)\underline{\phi}(B'B) u_3' u_3 &\leq u_3' B' A' A B u_3 \\ \underline{\phi}(A'A)\underline{\phi}(B'B) &\leq \frac{u_3' B' A' A B u_3}{u_3' u_3} \end{aligned}$$

and since this must hold for any vector u_3 it must also hold for the eigenvector associated with the smallest eigenvalue of $B'A'AB$ so we get

$$\underline{\phi}(A'A)\underline{\phi}(B'B) \leq \underline{\phi}(B'A'AB)$$

□

Proof of Lemma 1. First note that we can rewrite the design matrix $Z^D W_{\mathcal{T}}$ as $\tilde{Z}^D(\tilde{W}_{m_T} \otimes I_r)$

where $\tilde{W}_{m_T}$ is a $m_T \times m_T$ matrix with ones on and below the diagonal and

$$\tilde{Z}_D = \begin{bmatrix} Z_1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \\ Z_{t_1} & 0 & 0 & \cdots \\ 0 & Z_{t_1+1} & 0 & \cdots \\ \vdots & \vdots & \vdots & \\ 0 & Z_{t_2} & 0 & \cdots \\ \vdots & \vdots & Z_{t_2+1} & \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

which is $T \times r m_T$. Under assumption 4(i) the largest m_T is attained when $\tau_i = D_T \forall i$ in which case $m_T = T/D_T = c^{-1}T^{1-d}$. Hence in general we have that $m_T \leq c^{-1}T^{1-d}$ implying that the number of grid points grows slower than the sample size. Therefore for sufficiently large T we have that $r m_T \leq T$, so if we can bound the smallest eigenvalue of $(\tilde{W}_{m_T} \otimes I_r)' \tilde{Z}^{D\prime} \tilde{Z}^D (\tilde{W}_{m_T} \otimes I_r)/T$ we also have a bound on κ_T^2.

Notice that $\tilde{Z}^{D\prime}\tilde{Z}^D/T^d$ is a block diagonal matrix which smallest eigenvalue will be bounded away from zero with probability approaching one by assumption 4(ii). Next, for $\tilde{W}_{m_T}$ consider

$$(\tilde{W}'_{m_T}\tilde{W}_{m_T})^{-1} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & -1 & 2 \end{bmatrix}$$

We can bound the eigenvalues of this matrix using the Gershgorin circle theorem (Horn and Johnson, 1985, Thm. 6.1.1) and get $\overline{\phi}((\tilde{W}'_{m_T}\tilde{W}_{m_T})^{-1}) \leq 4$, this implies that $\underline{\phi}(\tilde{W}'_{m_T}\tilde{W}_{m_T}) \geq 1/4$. As $\tilde{W}'_{m_T}\tilde{W}_{m_T}$ and $(\tilde{W}_{m_T} \otimes I_r)'(\tilde{W}_{m_T} \otimes I_r) = \tilde{W}'_{m_T}\tilde{W}_{m_T} \otimes I_r$ have the same eigenvalues we get $\underline{\phi}((\tilde{W}_{m_T} \otimes I_r)'(\tilde{W}_{m_T} \otimes I_r)/T^{1-d}) \geq 1/(4T^{1-d})$.

By lemma 2

$$\begin{aligned} \underline{\phi}(W'_{\mathcal{T}} Z^{D\prime} Z^D W_{\mathcal{T}}/T) &= \underline{\phi}((\tilde{W}_{m_T} \otimes I_r)' \tilde{Z}^{D\prime} \tilde{Z}^D (\tilde{W}_{m_T} \otimes I_r)/T) \\ &\geq \underline{\phi}(\tilde{Z}^{D\prime}\tilde{Z}^D/T^d)\underline{\phi}((\tilde{W}_{m_T} \otimes I_r)'(\tilde{W}_{m_T} \otimes I_r)/T^{1-d}) \end{aligned}$$

Hence putting it together we have that $\underline{\phi}(W'_{\mathcal{T}} Z^{D\prime} Z^D W_{\mathcal{T}}/T) \in \Omega_p(T^{d-1})$ implying that $\kappa_T^2 \in \Omega_p(T^{d-1})$. □

A.2. Proofs for the Lasso

Before proving the main results we state some useful lemmas.

Lemma 3 ((6.8.14) in Hoffmann-Jørgensen (1994))**.** *Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be measurable such that $|f(U,V)|$ is integrable and $f(U,v)$ is integrable for P_V almost all $v \in \mathbb{R}$ (here P_V denotes the distribution of V), and let $\phi(v) = E(f(U,v))$. If, for a sigma field $\mathcal{G}$, V is measurable with respect to $\mathcal{G}$ and U is independent of $\mathcal{G}$, then we have*

$$E(f(U,V)|\mathcal{G}) = \phi(V) \text{ } P\text{-almost surely}$$

Lemma 4 (Lemma 2 in Kock and Callot (2015))**.** *Let assumptions 1 and 2 be satisfied. Then, for some positive constant A and for any $L_T > 0$,*

$$P\left(\max_{1\leq t\leq T}\max_{1\leq k\leq r}|Z_{kt}\epsilon_t| \geq L_T\right) \leq 2\exp\left(\frac{-L_T}{A\ln(1+T)\ln(1+r)\sigma_T^2}\right).$$

The following lemma provides bounds on the prediction and estimation error without making use of the restricted eigenvalue assumption.

Lemma 5. *Assuming that* $\left\|T^{-1}\epsilon' Z^D W\right\|_\infty \leq \lambda_T/2$, *then*

$$T^{-1}\left\|Z^D W\theta - Z^D W\hat{\theta}\right\|^2 + \lambda_T\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq 2\lambda_T\left(\left\|\hat{\theta}-\theta\right\|_{\ell_1} + \|\theta\|_{\ell_1} - \|\hat{\theta}\|_{\ell_1}\right) \tag{A.4}$$

$$T^{-1}\left\|Z^D W\theta - Z^D W\hat{\theta}\right\|^2 + \lambda_T\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq 4\lambda_T\left(\left\|\hat{\theta}_{\mathscr{S}}-\theta_{\mathscr{S}}\right\|_{\ell_1} \wedge \|\theta_{\mathscr{S}}\|_{\ell_1}\right) \tag{A.5}$$

$$\left\|\hat{\theta}_{\mathscr{S}^c}-\theta_{\mathscr{S}^c}\right\|_{\ell_1} \leq 3\left\|\hat{\theta}_{\mathscr{S}}-\theta_{\mathscr{S}}\right\|_{\ell_1} \tag{A.6}$$

Proof. Since $\hat{\theta}$ is the minimizer of the objective function (2) we have:

$$T^{-1}\left\|y - Z^D W\hat{\theta}\right\|^2 + 2\lambda_T\|\hat{\theta}\|_{\ell_1} \leq T^{-1}\left\|y - Z^D W\theta\right\|^2 + 2\lambda_T\|\theta\|_{\ell_1} \tag{A.7}$$

We can thus rewrite (A.7) as

$$T^{-1}\left\|Z^D W(\hat{\theta}-\theta)\right\|^2 - \frac{2}{T}\epsilon' Z^D W(\hat{\theta}-\theta) + 2\lambda_T\|\hat{\theta}\|_{\ell_1} \leq 2\lambda_T\|\theta\|_{\ell_1}$$

Using $y = Z^D W\theta + \epsilon$ we can write $\frac{2}{T}\epsilon' Z^D W(\hat{\theta}-\theta) \leq 2\left\|T^{-1}\epsilon' Z^D W\right\|_\infty\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq \lambda_T\left\|\hat{\theta}-\theta\right\|_{\ell_1}$. We now have

$$T^{-1}\left\|Z^D W(\hat{\theta}-\theta)\right\|^2 \leq \lambda_T\left\|\hat{\theta}-\theta\right\|_{\ell_1} + 2\lambda_T\left(\|\theta\|_{\ell_1} - \|\hat{\theta}\|_{\ell_1}\right)$$

so adding $\lambda_T\left\|\hat{\theta}-\theta\right\|_{\ell_1}$ yields

$$T^{-1}\left\|Z^D W(\hat{\theta}-\theta)\right\|^2 + \lambda_T\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq 2\lambda_T\left(\left\|\hat{\theta}-\theta\right\|_{\ell_1} + \|\theta\|_{\ell_1} - \|\hat{\theta}\|_{\ell_1}\right)$$

which is (A.4). Note that

$$\begin{aligned}\|\hat{\theta}-\theta\|_{\ell_1} + \|\theta\|_{\ell_1} - \|\hat{\theta}\|_{\ell_1} &= \|\hat{\theta}_{\mathscr{S}}-\theta_{\mathscr{S}}\|_{\ell_1} + \|\theta_{\mathscr{S}}\|_{\ell_1} - \|\hat{\theta}_{\mathscr{S}}\|_{\ell_1} \\ &\leq 2\|\hat{\theta}_{\mathscr{S}}-\theta_{\mathscr{S}}\|_{\ell_1}\end{aligned}$$

using continuity of the norm, and

$$\|\hat{\theta}_{\mathscr{S}}-\theta_{\mathscr{S}}\|_{\ell_1} + \|\theta_{\mathscr{S}}\|_{\ell_1} - \|\hat{\theta}_{\mathscr{S}}\|_{\ell_1} \leq 2\|\theta_{\mathscr{S}}\|_{\ell_1}$$

by sub-additivity of the norm. Using the two results above in (A.4) yields (A.5). Finally notice that (A.5) gives

$$\lambda_T\|\hat{\theta}-\theta\|_{\ell_1} \leq 4\lambda_T\|\hat{\theta}_{\mathscr{S}}-\theta_{\mathscr{S}}\|_{\ell_1}$$

or equivalently

$$\left\|\hat{\theta}_{\mathscr{S}^c} - \theta_{\mathscr{S}^c}\right\|_{\ell_1} \leq 3\left\|\hat{\theta}_{\mathscr{S}} - \theta_{\mathscr{S}}\right\|_{\ell_1}$$

which establishes (A.6). □

Lemma 6. *Let assumptions 1 and 2 be satisfied and define:*

$$\mathscr{B}_T = \left\{ \max_{1\leq k\leq r} \max_{1\leq s\leq T-r+1} \left| \frac{1}{T}\sum_{t=s}^{T} \epsilon_t Z_{tk} \right| < \frac{\lambda_T}{2} \right\}.$$

Then, for $\lambda_T = \sqrt{\frac{8\ln(1+T)^5 \ln(1+r)^2 \ln(r(T-r+1))\sigma_T^4}{T}}$ *and some constant* $A > 0$,

$$P\left(\mathscr{B}_T\right) = P\left(\left\{\left\| T^{-1}\epsilon' Z^D W \right\|_\infty < \lambda_T/2\right\}\right) \geq 1 - 2(1+T)^{-1/A} + 2(r(T-r+1))^{1-\ln(1+T)}.$$

Proof. For any $L_T > 0$, and using sub-additivity of the probability measure,

$$\begin{aligned}
&P\left(\max_{1\leq k\leq r} \max_{1\leq s\leq T-r+1} \left| \frac{1}{T}\sum_{t=s}^{T} \epsilon_t Z_{tk} \right| \geq \frac{\lambda_T}{2} \right) \\
&= P\left(\bigcup_{k=1}^{r} \bigcup_{s=1}^{T-r+1} \left\{ \left| \frac{1}{T}\sum_{t=s}^{T} \epsilon_t Z_{tk} \right| \geq \frac{\lambda_T}{2} \right\} \right) \\
&\leq P\left(\bigcup_{k=1}^{r} \bigcup_{s=1}^{T-r+1} \left\{ \left| \frac{1}{T}\sum_{t=s}^{T} \epsilon_t Z_{tk} \right| \geq \frac{\lambda_T}{2} \right\} \cap \left\{ \bigcap_{t=1}^{T} \bigcap_{k=1}^{r} \left\{ |\epsilon_t Z_{tk}| < L_T \right\} \right\} \right) + P\left(\left\{ \bigcap_{t=1}^{T} \bigcap_{k=1}^{r} \left\{ |\epsilon_t Z_{tk}| < L_T \right\} \right\}^c \right) \\
&\leq \sum_{k=1}^{r} \sum_{s=1}^{T-r+1} P\left(\left| \frac{1}{T}\sum_{t=s}^{T} \epsilon_t Z_{tk} \right| \geq \frac{\lambda_T}{2}, \bigcap_{t=1}^{T} \{ |\epsilon_t Z_{tk}| < L_T \} \right) + P\left(\max_{1\leq t\leq T} \max_{1\leq k\leq r} |\epsilon_t Z_{tk}| \geq L_T \right)
\end{aligned}$$

Using lemma 4 on the second term yields a first bound

$$P\left(\max_{1\leq t\leq T} \max_{1\leq k\leq r} |\epsilon_t Z_{tk}| \geq L_T \right) \leq 2\exp\left(\frac{-L_T}{A\ln(1+T)\ln(1+r)\sigma_T^2} \right).$$

Note that in the first term we are considering the probability of a sum of random variables on a set on which the summands are bounded by L_T. Now consider the sequence $\left\{\epsilon_t Z_{tk} \mathbb{1}_{|\epsilon_t Z_{tk}| < L_T}\right\}$ and the filtration $\mathscr{F}_{Z,\epsilon,t} = \sigma\left(\left\{\epsilon_i Z_i, i = 1, ..., t\right\}\right)$ and the conditional expectation

$$\begin{aligned}
E\left(\epsilon_t Z_{tk} \mathbb{1}_{|\epsilon_t Z_{tk}| < L_T} | \mathscr{F}_{Z,\epsilon,t-1}\right) &= E\left(E\left(\epsilon_t Z_{tk} \mathbb{1}_{|\epsilon_t Z_{tk}| < L_T} | \sigma\left(\left\{\mathscr{F}_{Z,\epsilon,t-1}, Z_{tk}\right\}\right)\right) | \mathscr{F}_{Z,\epsilon,t-1} \right) \\
&= E\left(Z_{tk} E\left(\epsilon_t \mathbb{1}_{|\epsilon_t Z_{tk}| < L_T} | \sigma\left(\left\{\mathscr{F}_{Z,\epsilon,t-1}, Z_{tk}\right\}\right)\right) | \mathscr{F}_{Z,\epsilon,t-1} \right).
\end{aligned}$$

If Z_{tk} belongs to the set of lagged variables $y_{i,t-l}$ $i = 1, ..., r_y$, $l = 1, ..., p$, $\sigma\left(\left\{\mathscr{F}_{Z,\epsilon,t-1}, Z_{tk}\right\}\right) =$

$\mathscr{F}_{Z,\epsilon,t-1}$ making the equations above redundant. This is not the case when Z_{tk} belongs to the set of contemporaneous exogenous variables X_{tk}, $k = 1, ..., r_X$.

Since Z_{tk} is measurable on $\sigma\left(\left\{\mathscr{F}_{Z,\epsilon,t-1}, Z_{tk}\right\}\right)$ we use lemma 3 with $f(\epsilon_t, Z_{tk}) = \epsilon_t \mathbb{1}_{|\epsilon_t Z_{tk}|<L_T}$ such that for all $v \in \mathbb{R}$ we get

$$E(\epsilon_t \mathbb{1}_{|\epsilon_t v|<L_T} | \sigma(\{\mathscr{F}_{\epsilon,Z,t-1,k}, Z_{tk}\})) = E(\epsilon_t \mathbb{1}_{|\epsilon_t|<\frac{L_T}{|v|}} | \sigma(\{\mathscr{F}_{\epsilon,Z,t-1,k}, Z_{tk}\})) = 0.$$

This argument holds for $v \neq 0$, for the case where $v = 0$ the results follows from noting that $E(\epsilon_t \mathbb{1}_{|\epsilon_t v|<L_T} | \sigma(\{\mathscr{F}_{\epsilon,Z,t-1,k}, Z_{tk}\})) = E(\epsilon_t | \sigma(\{\mathscr{F}_{\epsilon,Z,t-1,k}, Z_{tk}\})) = 0$. The sequence $\left\{\epsilon_t Z_{tk} \mathbb{1}_{|\epsilon_t Z_{tk}|<L_T}\right\}$ is a martingale difference sequence with bounded increments. We can thus apply the Azuma-Hoeffding inequality to bound the first term.

$$\begin{aligned}
&\sum_{k=1}^{r} \sum_{s=1}^{T-r+1} P\left(\left|\frac{1}{T}\sum_{t=s}^{T} \epsilon_t Z_{tk}\right| \geq \frac{\lambda_T}{2}, \bigcap_{t=1}^{T} \{|\epsilon_t Z_{tk}| < L_T\}\right) \\
&\leq r(T-r+1) 2 \exp\left(\frac{-\frac{\lambda_T^2}{4} T^2}{2TL_T^2}\right) \\
&\leq r(T-r+1) 2 \exp\left(-\frac{T\lambda_T^2}{8L_T^2}\right)
\end{aligned}$$

Let $L_T = \ln(1+T)^2 \ln(1+r)\sigma_T^2$, and gather the two bounds found above,

$$P\left(\left\|\frac{1}{T}\epsilon' Z^D W\right\|_\infty \geq \frac{\lambda_T}{2}\right) \leq 2(r(T-r+1))^{1-\ln(1+T)} + 2(1+T)^{-1/A}.$$

□

Proof of Theorem 1. On $\mathscr{B}_T$ and under assumptions 1, 2, 3, and 5, we use equations (A.5) and Jensen's inequality to get:

$$\frac{1}{T}\left\|Z^D W(\hat{\theta}-\theta)\right\|^2 \leq 4\lambda_T \left\|\hat{\theta}_{\mathscr{S}} - \theta_{\mathscr{S}}\right\|_{\ell_1} \leq 4\lambda_T \sqrt{s} \left\|\hat{\theta}_{\mathscr{S}} - \theta_{\mathscr{S}}\right\| \leq 4\lambda_T \sqrt{s} \frac{\left\|Z^D W(\hat{\theta}-\theta)\right\|}{\kappa_T \sqrt{T}}.$$

Note that the restricted eigenvalue condition applies due to (A.6). Rearranging yields (1). We also get

$$\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq 4\left\|\hat{\theta}_{\mathscr{S}} - \theta_{\mathscr{S}}\right\|_{\ell_1} \leq 4\sqrt{s}\left\|\hat{\theta}_{\mathscr{S}} - \theta_{\mathscr{S}}\right\| \leq 4\sqrt{s} \frac{\left\|Z^D W(\hat{\theta}-\theta)\right\|}{\kappa_T \sqrt{T}} \leq \frac{16}{\kappa_T^2} s\lambda_T.$$

which is (3). Lemma 6 gives the probability of being on $\mathscr{B}_T$. □

Proof of Corollary 1. To prove this result, assume that $\hat{\theta}_j = 0$ for $j \in \mathscr{S}$, then $|\theta_{\min}| \leq \left\|\hat{\theta}-\theta\right\|_{\ell_1}$. Hence if $|\theta_{\min}| > \left\|\hat{\theta}-\theta\right\|_{\ell_1}$ no relevant variables are excluded. □

Proof of Theorem 2. From lemma 6, $P(\mathscr{B}_T) \to 1$. Note that $\kappa_T^2 \in \Omega_p(T^{d-1}) \iff \kappa_T^{-2} \in \mathscr{O}_p(T^{1-d})$.

$$\frac{16s\lambda_T}{\kappa_T^2} \in \mathcal{O}_p\left(T^{1-a-1/2-d+1}\right) \in \mathcal{O}_p\left(T^{3/2-a-d}\right)$$
$$\frac{16s\lambda_T^2}{\kappa_T^2} \in \mathcal{O}_p\left(T^{1-a-1-d+1}\right) \in \mathcal{O}_p\left(T^{1-a-d}\right)$$

so that $\frac{16s\lambda_T}{\kappa_T^2} \to^p 0$ and $\frac{16s\lambda_T^2}{\kappa_T^2} \to^p 0$.

It follows that

$$T^{-1}\left\|Z^D W(\hat{\theta}-\theta)\right\|^2 \leq \frac{16s\lambda_T^2}{\kappa_T^2} \to^p 0,$$
$$\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq \frac{16s\lambda_T}{\kappa_T^2} \to^p 0,$$

which proves (4) and (5). □

Proof of Corollary 2. The proof of corollary 2 follows from the fact proven in corollary 1 that on $\mathscr{B}_T$ and if $|\theta_{\min}| > \frac{16s\lambda_T}{\kappa_T^2}$, no relevant variables are excluded. Noticing that $P(\mathscr{B}_T) \to 1$ and $\frac{16s\lambda_T}{\kappa_T^2} \to^p 0$ completes the proof. □

A.3. Proofs for the adaptive Lasso.

The proofs of lemma 7 and theorem 3 are very similar to those of lemma 11 and theorem 6 in Kock and Callot (2015), hence we simply state lemma 7 and sketch the proof of theorem 3. The proof of theorem 4 differs from previous results in the literature due to the fact that the restricted eigenvalue tends to zero in our setting, hence we prove that result in greater details.

Lemma 7 (Lemma 11 in Kock and Callot (2015))**.** *Let*

$$\mathscr{C}_T = \left\{ \max_{1\leq i,j\leq r(T-r+1)} \left| \frac{1}{T}(Z^D W_i)'(Z^D W_j) \right| < K_T \right\}$$

for $K_T = \ln(1+r(T-r+1))^2 \ln(T)\sigma_T^2$. *Then* $P\left(\mathscr{C}_T\right) \geq 1 - 2T^{-1/A}$ *for some constant* $A > 0$.

Proof of theorem 3. The proof of theorem 3 is very similar to the proof of theorem 6 in Kock and Callot (2015), hence we only give the main steps of this proof and refer the reader to Kock and Callot (2015) for details. Let $\Psi_{i,\widehat{\mathscr{S}}} = \frac{(Z^D W_i)'(Z^D W_{\widehat{\mathscr{S}}})}{T}$ and $\Psi_{\widehat{\mathscr{S}},\widehat{\mathscr{S}}} = \frac{(Z^D W_{\widehat{\mathscr{S}}})'(Z^D W_{\widehat{\mathscr{S}}})}{T}$. van de Geer, Bühlmann, Zhou, et al. (2011) (and Kock and Callot (2015) in the VAR case) show that $\text{sign}(\tilde{\theta}) = \text{sign}(\theta)$ if and only if the following two conditions are met for every $i \in \mathscr{S}^c$:

$$\left| \Psi_{i,\widehat{\mathscr{S}}}\left(\Psi_{\widehat{\mathscr{S}},\widehat{\mathscr{S}}}\right)^{-1}\left(\frac{(Z^D W_{\widehat{\mathscr{S}}})'\epsilon}{T} - \lambda_T \,\text{sign}(\theta_i)\, w_i\right) - \frac{(Z^D W_i)'\epsilon}{T} \right| \leq \lambda_T w_i, \tag{A.8}$$

and

$$\text{sign}\left(\theta_{\widehat{\mathscr{S}}} + \left(\Psi_{\widehat{\mathscr{S}},\widehat{\mathscr{S}}}\right)^{-1}\left(\frac{(Z^D W_{\widehat{\mathscr{S}}})'\epsilon}{T} - \lambda_T \,\text{sign}(\theta_{\widehat{\mathscr{S}}})\, w_{\widehat{\mathscr{S}}}\right)\right) = \text{sign}(\theta_{\mathscr{S}}). \tag{A.9}$$

Kock and Callot (2015) shows that on the set $\mathscr{B}_T \cap \mathscr{C}_T$ the left side of (A.8) can be bounded from above by:

$$\left| \Psi_{i,\widehat{\mathscr{F}}} \left(\Psi_{\widehat{\mathscr{F}},\widehat{\mathscr{F}}} \right)^{-1} \left(\frac{(Z^D W_{\widehat{\mathscr{F}}})' \epsilon}{T} - \lambda_T \operatorname{sign}(\theta_i) w_i \right) - \frac{(Z^D W_i)' \epsilon}{T} \right| \leq \frac{s K_T}{\kappa_T^2} \left(\frac{\lambda_T}{2} + \frac{2\lambda_T}{\theta_{\min}} \right) + \frac{\lambda_T}{2}.$$

The right side of (A.8) is bounded from below by $\left|\lambda_T w_i\right| \geq \frac{\lambda_T}{\left\|\hat{\theta}-\theta\right\|_{\ell_1}}$. We replace these bounds in (A.8) and multiply both sides by $\frac{\left\|\hat{\theta}-\theta\right\|_{\ell_1}}{\lambda_T}$ to get (7). If (7) holds, so does (A.8).

For the condition (A.9) to be verified it suffices to show that

$$\left\| \left(\Psi_{\widehat{\mathscr{F}},\widehat{\mathscr{F}}} \right)^{-1} \left(\frac{(Z^D W_{\widehat{\mathscr{F}}})' \epsilon}{T} - \lambda_T \operatorname{sign}(\theta_{\widehat{\mathscr{F}}}) w_{\widehat{\mathscr{F}}} \right) \right\|_{\infty} \leq \theta_{\min}$$

which Kock and Callot (2015) shows to be satisfied if (8) is satisfied.

Lemmas 7 and 6 provide the desired bound on $P\left(\mathscr{B}_T \bigcap \mathscr{C}_T\right)$ which completes the proof. □

Proof of theorem 4. To prove theorem 4 we have to prove that the conditions in theorem 3 are valid asymptotically. We work on the set $\mathscr{B}_T \cap \mathscr{C}_T$ which we begin by showing holds with probability tending to 1, we then turn to the other conditions.

1. $P(\mathscr{B}_T \cap \mathscr{C}_T) \to 1$ can be seen to hold from lemmas 7 and 6 since $P(\mathscr{B}_T^c) \to 0$ and $P(\mathscr{C}_T^c) \to 0$.

2. To show that $\theta_{\min} \geq 2\left\|\hat{\theta}-\theta\right\|_{\ell_1}$ is asymptotically valid, recall that from (3):

$$\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq \frac{16 s \lambda_T}{\kappa_T^2} \in \mathscr{O}_p\left(\ln(1+T)^{5/2} \ln(r(T-r+1))^{1/2} T^{3/2-a-d}\right),$$

   and since $\theta_{\min} \in \Omega\left(\ln(T) a_T\right)$ we have:

$$\begin{aligned} \frac{\left\|\hat{\theta}-\theta\right\|_{\ell_1}}{\theta_{\min}} &\in \mathscr{O}_p\left(\frac{\ln(1+T)^{5/2} \ln(r(T-r+1))^{1/2} T^{3/2-a-d}}{\ln(T)^2 \ln(1+T)^{5/2} \ln(r(T-r+1))^{1/2} \ln(1+r(T-r+1))^2 T^{7/2-2a-2d}} \right) \\ &\in \mathscr{O}_p\left(\frac{T^{a+d-2}}{\ln(T)^2 \ln(1+r(T-r+1))^2} \right) \in o_p(1) \end{aligned}$$

   so that $\theta_{\min} \geq 2\left\|\hat{\theta}-\theta\right\|_{\ell_1}$ with probability approaching 1.

3. Recall that by assumption 5, $\kappa_T^{-2} \in \mathscr{O}_p\left(T^{1-d}\right)$. To show that (7) holds asymptotically, we replace $\left\|\hat{\theta}-\theta\right\|_{\ell_1}$ by its upper bound from (3) and we are left to show that $\frac{s^2 K_T \lambda_T}{\kappa_T^4} + \frac{s^2 K_T \lambda_T}{\kappa_T^4 \theta_{\min}} + \frac{s\lambda_T}{\kappa^2} \to^p 0$.

   i) Consider the first term, $\frac{s^2 K_T \lambda_T}{\kappa_T^4} \in \mathscr{O}_p(a_T) \in o_p(1)$.

   ii) Consider now the second term $\frac{s^2 K_T \lambda_T}{\kappa_T^4 \theta_{\min}}$, which is equal to the first term divided by $\theta_{\min}$. Since $\theta_{\min} \in \Omega\left(\ln(T) a_T\right)$, the second term tends to zero.

iii) The third term is proportional to the estimation error of the Lasso, which by theorem 2 tends to zero, so that $\frac{s\lambda_T}{\kappa^2} \to^p 0$.

4. To show that (8) holds asymptotically we have to show that $\frac{\sqrt{s}\lambda_T}{\kappa_T^2}\left(\frac{1}{2\theta_{\min}} + \frac{2}{\theta_{\min}^2}\right) \to^p 0$. Notice that $\frac{\sqrt{s}\lambda_T}{\kappa_T^2} \in \mathcal{O}_p(b_T^2)$ and recall that $\theta_{\min} \in \Omega\left(\ln(T)b_T\right)$ so that $\frac{\sqrt{s}\lambda_T}{\kappa_T^2\theta_{\min}^2} \to^p 0$ implying that $\frac{\sqrt{s}\lambda_T}{\kappa_T^2\theta_{\min}} \to^p 0$.

This completes the proof. □

References

Bai, J. (1997). Estimation of a change point in multiple regression models. Review of Economics and Statistics 79(4), 551–563.

Bai, J. (2000). Vector autoregressive models with structural changes in regression coefficients and in variance-covariance matrices. Annals of Economics and Finance 1(2), 303–339.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. Econometrica 66(1), 47–78.

Belloni, A. and V. Chernozhukov (2011). High dimensional sparse econometric models: An introduction. In P. Alquier, E. Gautier, and G. Stoltz (Eds.), Inverse Problems and High-Dimensional Estimation, Lecture Notes in Statistics, pp. 121–156. Springer Berlin Heidelberg.

Belloni, A., V. Chernozhukov, et al. (2013). Least squares after model selection in high-dimensional sparse models. Bernoulli 19(2), 521–547.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 37(4), 1705–1732.

Bitto, A. and S. Frühwirth-Schnatter (2014). Time-varying parameter models–achieving shrinkage and variable selection. Unpublished manuscript.

Boivin, J. and M. P. Giannoni (2006). Has monetary policy become more effective? The Review of Economics and Statistics 88(3), 445–462.

Bühlmann, P. and S. van de Geer (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.

Chan, N. H., C. Y. Yau, and R.-M. Zhang (2014). Group lasso for structural break time series. Journal of the American Statistical Association 109(506), 590–599.

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. Journal of Applied Econometrics 28(5), 777–795.

Durbin, J. and S. J. Koopman (2012). Time Series Analysis by State Space Methods (2nd ed.). Oxford University Press.

Hamilton, J. D. (2008). Regime-switching models. In S. Durlauf and L. Blume (Eds.), The New Palgrave Dictionary of Economics, Volume 2. Palgrave Macmillan.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. Econometrica 79(2), 453–497.

Harchaoui, Z. and C. Lévy-Leduc (2010). Multiple change-point estimation with a total variation penalty. Journal of the American Statistical Association 105(492), 1480–1493.

Hoffmann-Jørgensen, J. (1994). Probability with a View Towards Statistics. 1. CRC Press.

Horn, R. A. and C. R. Johnson (1985). Matrix Analysis. Cambridge University Press.

Kock, A. B. (2014). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. Working paper.

Kock, A. B. and L. Callot (2015). Oracle inequalities for high dimensional vector autoregressions. Journal of Econometrics. In press.

Koop, G. and D. Korobilis (2013). Large time-varying parameter VARs. Journal of Econometrics 177(2), 185–198.

Lütkepohl, H. (2007). New Introduction to Multiple Time Series Analysis. Springer.

Perron, P. (2006). Dealing with structural breaks. In T. C. Mills and K. Patterson (Eds.), Palgrave Handbook of Econometrics, Volume 1, pp. 278–352. Palgrave Macmillan.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. The Review of Economic Studies 72(3), 821–852.

Qian, J. and L. Su (2014). Structural change estimation in time series regressions with endogenous variables. Economics Letters 125(3), 415–421.

Qu, Z. and P. Perron (2007). Estimating and testing structural changes in multivariate regressions. Econometrica 75(2), 459–502.

Sims, C. A. and T. Zha (2006). Were there regime switches in US monetary policy? The American Economic Review 96(1), 54–81.

Taylor, J. B. (1993). Discretion versus policy rules in practice. In Carnegie-Rochester Conference Series on Public Policy, Volume 39, pp. 195–214. Elsevier.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267–288.

van de Geer, S., P. Bühlmann, S. Zhou, et al. (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). Electronic Journal of Statistics 5, 688–749.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476), 1418–1429.