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 proposes a parsimoniously time varying parameter vector autoregressive model (with exogenous variables, VARX) and studies the properties of the Lasso and adaptive Lasso as estimators of this model. The parameters of the model 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. By varying the degree of parsimony our model can accommodate constant parameters, an unknown number of structural breaks, or parameters with a high degree of variation. 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 asymptotically we show that these bounds tend to zero with probability tending to one if the number of non zero increments grows slower than $\sqrt{T}$.

By simulation experiments we investigate 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 1980s and since 2008.

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 are assumed to follow a parsimonious random walk, that is, a random walk with a positive probability that an increment is exactly equal to zero. The parsimonious random walk allows the time varying parameters to be modelled non parametrically, hence 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

[☆] The authors would like to thank Anders B. Kock and Paolo Santucci de Magistris, as well as participants of the 2014 Netherlands Econometrics Study Group and the 2014 NBER-NSF time series conference for their comments and suggestions. 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)

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 time series econometrics literature, 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) who document that the monetary policy response to inflation in the US changed in the 1980s with the arrival of Paul Volker has chairman of the Federal Reserve Bank.

The contribution of this paper is to propose an estimator for VARX models with parsimoniously time-varying parameters, more precisely we would like to stress 3 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 $\alpha_T = T^{-a}$. In the case of a single variable this leads to an expected number of non-zero increments $E(s) = T^{1-a}$.

ii) We establish finite sample 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 (2014). We then turn to asymptotics and show that the errors tend to zero with probability tending to one. We also establish similar results for the adaptive Lasso of Zou (2006), and furthermore show under which conditions it possesses the oracle property, that is, the conditions under which the adaptive Lasso recovers the true model with probability tending to one. Asymptotic consistency of the Lasso requires an upper bound on the number of breaks in the parameter path in the form of the bound $a > 1/2$, implying that the number of non zero increments must grow strictly slower than $\sqrt{T}$. We also highlight the trade-off between estimation efficiency and number of breaks by showing that the speed of convergence of the estimator is in the order of $T^{1-a-1/2}$. At one extreme where the number of breaks is constant, $a = 1$, the Lasso is less efficient than OLS by a logarithmic factor, while at the other extreme where a is close to $1/2$ convergence is slow.

iii) 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 Taylor rule with inflation and output gap and find that the response to inflation has been unstable from the mid-1970s to the mid-1980s and experienced a substantial change in 2008, while the response to the output gap remained stable.

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 following 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 a parsimoniously time varying Taylor rule for US monetary policy and document substantial instability in the response of the Fed to inflation in the early 1980s and since 2008.

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 $Y_t = [y_{1t}, ..., y_{r_y t}]'$. Without loss of generality we assume that the variables have been demeaned. 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 $r \times 1$, $r = r_x + p r_y$, 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 has 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* $k = 1, ..., r_x$, $\text{Var}(X_{kt}) = \sigma_{Xk}^2 < \infty$.

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

The variances of the innovations ϵ_t and of the exogenous variables X_{kt} 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:

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

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 $p r_y \times r_y$, 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 $r_y \times T r_y$ 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*

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

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

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 $P(\zeta_{kt} = 0) > 0$, 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 $P(\zeta_{kt} = 0)$. At one extreme, when $P(\zeta_{kt} = 0) = 1$, the parameter vector

is stable, while at the other extreme, $P(\zeta_t = 0) = 0$, it follows a random walk. When we are between these two extremes we will refer to the 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:

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

We assume that $\alpha_T = T^{-a}$ to control the growth of the cardinality of the active set (the number of non-zero variables) $E(s_T) = r\alpha_T T = \mathcal{O}\left(T^{1-a}\right)$. We assume r to be fixed so that the growth rate of the active set is entirely controlled by a.[1] Consistency requirements for the Lasso estimator will impose a tighter lower bound on a, implying an upper bound on the speed with which the active set can grow. It is important to note that while assumption 3 puts no further restrictions on the path of the parsimonious random walk, then 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

[1] r could be made a function of time at the cost of a tighter upper bound on a.

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.

By assuming that the increments to the parsimonious random walk can be exactly equal to zero we assume that the parameter vector θ is sparse since all but the first r element of θ are the increments to the parsimonious random walk. This type of process 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 $\sigma_T^2 = \max\left(\sigma_\epsilon^2, \max_{1\leq k\leq rT}\sigma_k^2\right)$ 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 $\mathscr{S}_T$ as the set of indices corresponding to non-zero parameters in θ, as $\mathscr{S}_T = \left\{j \in (1, ..., rT) | \theta_j \neq 0\right\}$ and its cardinality $|\mathscr{S}_T| = s_T$. To simplify notation, when it is unambigious, we omit the subscript T. We note $\|.\|_{\ell_1}$ the ℓ_1 norm and $\|.\|$ the ℓ_2 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 $x > 0$.

3. Estimation

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

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

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. The properties of the penalty sequence λ_T and of the Lasso estimator $\widehat{\theta}$ are discussed below. 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 some finite sample properties of the estimator before considering the asymptotic behaviours of the estimation and prediction errors of the Lasso. We also derive results regarding the Lasso's variable selection properties.

This model has rT parameters and T observations so that when $r \geq 1$ 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. Before assuming the restricted eigenvalue condition, we need to ensure that we are working on a set of variables in which it is possible that the restricted eigenvalue of Ψ_T is larger than 0.

Notice that when $r > 1$ the last r columns of $Z^D W$ are $[0_r, ..., 0_r, Z_T']'$ which are by construction linearly dependent. Let W_a be the a^{th} column of W and define the set $\mathscr{A} = \{a : W_a' \iota \geq r\}$ where ι is a $Tr \times 1$ vector of ones. Define the Gram matrix $\Psi_{T\mathscr{A}} = \frac{W_{\mathscr{A}}' Z^{D'} Z^D W_{\mathscr{A}}}{T}$.

While per construction the restricted eigenvalue condition cannot be satisfied on Ψ_T it can be on $\Psi_{T,\mathscr{A}}$. In practice this means that we rule out the possibility of a change in parameter value from observation $T - r + 1$ to the end of the sample. Note that for a model with $r = 1$, $\mathscr{A} = \{1, ..., Tr\}$ so that $\Psi_{T,\mathscr{A}} = \Psi_T$.

Let $\kappa_T^2(s)$ be the smallest eigenvalue of the gramian matrix of any subset of variables with cardinality smaller or equal to that of the active set:

$$\kappa_T^2(s) = \min_{|\mathscr{S}|\leq s}\left(\min_{\delta\in\mathbb{R}^{nT}\setminus\{0\}}\left\{\frac{\delta'\Psi_{\mathscr{A}}\delta}{\left\|\delta_{\mathscr{S}}\right\|^2}\setminus\left\|\delta_{\mathscr{S}^c}\right\|_{\ell_1}\leq 3\left\|\delta_{\mathscr{S}}\right\|_{\ell_1}\right\}\right).$$

Note $\Psi = E(\Psi_T)$ the population covariance matrix and the corresponding restricted eigenvalue κ^2. We make the following assumption on the sample and population restricted eigenvalue.

Assumption 4 (Sample and population restricted eigenvalue condition)**.** *Assume that:*

i) $\kappa_T^2(s) > 0$.

ii) There exists a constant $c > 0$ *such that* $\kappa^2 \geq c$.

Assumption 4 is the restricted eigenvalue condition, it is a standard assumption in the high dimensional econometrics literature introduced by Bickel et al. (2009), and by Kock and Callot (2014) in the case of high dimensional VARs. To simplify the notation we drop the dependence on s and write the sample and population restricted eigenvalues κ_T^2 and κ^2, respectively.

Notice that the construction of $\mathscr{A}$ implies that we penalize the initial value of the parsimonious random walks, ξ_{k0} where $k = (1,...,r)$ 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* $\lambda_T = \sqrt{\frac{8\ln(1+T)^5\ln(1+r)^2\ln(1+T-r+1)^2\ln(r(T-r+1))\sigma_T^4}{T}}$ *and some constant* $A > 0$*, and under assumptions 1, 2, 3, and 4, and on a set with probability at least equal to* $1-\pi_T^{\mathscr{B}}$ *we have the following inequalities:*

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

$$\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq \frac{16s\lambda_T}{\kappa_T^2}, \tag{4}$$

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

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; holding everything else constant the probability of this inequality being satisfied decreases with λ_T.

The dependence on λ_T highlights the trade-off between selecting a larger value of λ 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 θ_{min} condition 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* $s \in \mathscr{O}(T^{1-a})$ *with* $a > \frac{1}{2}$. *Under assumptions 1, 2, 3, and 4, and as* $T \to \infty$ *we have:*

$$T^{-1} \left\| Z^D W(\theta - \hat{\theta}) \right\|^2 \to^p 0 \tag{5}$$

$$\left\| \hat{\theta} - \theta \right\|_{\ell_1} \to^p 0 \tag{6}$$

with probability tending to 1.

Theorem 2 states that the prediction and estimation errors tend to zero in probability provided the cardinality of the active set grows at a rate strictly slower than $\sqrt{T}$. This condition on the rate of growth of the active set implies that the probability of a non-zero increment $\alpha_T = T^{-a}$ tends to zero strictly faster than $\frac{1}{\sqrt{T}}$.

Theorem 2 also gives indication on the speed of convergence of the bounds. The speed with which the estimation error tends to zero is dominated by the product $s\lambda_T = \mathscr{O}\left(T^{1-a-1/2}\right)$ so that for a small number of breaks (a large value of a), the convergence speed is slightly slower than $\sqrt{T}$ while it can get extremely slow for a close to 1/2. The prediction error is dominated by $s\lambda_T^2 = \mathscr{O}\left(T^{1-a-1}\right)$ so that the speed of convergence is always greater than $\sqrt{T}$.

Were we to assume that the number of variables r grows over time at a sub-exponential rate,$r \in \mathscr{O}(e^{T^b})$ with $0 < 5b < 1$, the rate of growth of s would change to $s \in \mathscr{O}(T^{1-a-b})$. This

highlights the trade-off between the number of paths to estimate r and the number of non-zero increments. It is the total number of non-zero parameters in θ that matters so that the model can handle either a limited number of series with a lot of variation in the parameter value, or a large number of variables with only occasional breaks. 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* $\theta_{min} > \frac{16s}{qc^2}\lambda_T$ *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 improves upon the Lasso. The adaptive Lasso of Zou (2006) is based on this idea, using an initial estimator to construct adadptive 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_l by taking the inverse of the absolute value of the estimated parameters $w_l = \frac{1}{|\widehat{\theta}_{\widehat{\mathscr{S}}}|}$. The adaptive Lasso objective function is thus given by:

$$\tilde{\theta} = \text{argmin}_{\theta_{\widehat{\mathscr{S}}}} \left(\frac{1}{T} \left\| y - (Z^D W_{\widehat{\mathscr{S}}}) \theta_{\widehat{\mathscr{S}}} \right\|^2 + 2\lambda_T w_l \left\| \theta_{\widehat{\mathscr{S}}} \right\|_{\ell_1} \right). \tag{7}$$

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 (7) is of much smaller size than (2).

We study the properties of the adaptive Lasso in our setting by, as in the case of the Lasso, deriving finite sample inequalities before studying its asymptotic properties. We choose to focus on the oracle property, the ability of the adaptive Lasso to recover the exact model ($\text{sign}(\tilde{\theta}) = \text{sign}(\theta)$). Hence we work under assumptions ensuring that corollary 1 holds so that no relevant variable is discarded in the initial step. 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 estimators to compute the adaptive weights and estimators with tighter ℓ_1 bounds on the estimation error would results in tighter bounds for the adaptive Lasso.

Define $\phi_{\text{min},\mathscr{S}}$ as the smallest eigenvalue of $E\left(\frac{1}{T} Z^D W_{\mathscr{S}} W'_{\mathscr{S}} Z^{D'}\right)$, which is greater than 0 by assumption 4. We now give a finite sample probability and conditions for the adaptive Lasso to be sign consistent.

Theorem 3. *Let* $\lambda_T = \sqrt{\frac{8\ln(1+T)^5 \ln(1+r)^2 \ln(1+T-r+1)^2 \ln(r(T-r+1))\sigma_T^4}{T}}$. *Under assumptions 1, 2, 3,*

and 4, and assuming that $\theta_{\min} \geq 2\left\|\widehat{\theta}-\theta\right\|_{\ell_1}$ *and*

$$\frac{sK_T}{q\phi_{\min,\mathscr{S}}}\left(\frac{1}{2}+\frac{2}{\theta_{\min}}\right)\left\|\widehat{\theta}-\theta\right\|_{\ell_1} \leq 1 \tag{8}$$

$$\frac{\sqrt{s}}{q\phi_{\min,\mathscr{S}}}\left(\frac{\lambda_T}{2}+\frac{2\lambda_T}{\theta_{\min}}\right) \leq \theta_{\min} \tag{9}$$

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* $\text{sign}(\tilde{\theta}) = \text{sign}(\theta)$.

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. (8) 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 (8) can be interpreted as a condition on the performance of the inital estimator. (9) 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. *Under assumptions 1, 2, 3, and 4, assume that* $a>1/2$ *and define* $a_T = \ln(1+T)^{5/4}\ln(1+T-r+1)^{1/2}\ln(r(T-r+1))^{1/4}T^{-a/4}$ *and* $b_T = \ln(1+T)^{5/2}\ln(1+T-r+1)^3\ln(r(T-r+1))^{1/2}T^{1-a/2}$. *Let* $\theta_{\min} \in \Omega(\ln(T)\max(a_T,b_T))$, *then* $P(\text{sign}(\tilde{\beta}) = \text{sign}(\beta)) \to 1$.[2]

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 function of a, this restriction on the speed at which the smallest non-zero parameter may tend to zero guarantees that no relevant variable will be excluded. The speed at which $\theta_{\min}$ tends to zero is an increasing function of a, the fewer breaks in the model the faster $\theta_{\min}$ may tend to 0.

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 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 $P\left(\widehat{\mathscr{S}}_{BIC} = \mathscr{S}\right) \to 1$.

[2] $f(T) \in \Omega(g(T))$ means that there exists a constant c such that $f(T) \geq cg(T)$ for $T \geq T_0$ for a certain T_0 onwards.

3.4. Post Lasso OLS

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

iii) Belloni et al. (2013) and Kock and Callot (2014) 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 plots samples of true and estimated parameter path for different estimators to give a sense of the location and amplitude of the breaks in the estimated paths relative to the true parameter path. 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 replications of the simulations and empirical application below. 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 on commodity hardware.

4.1. Deterministic paths

		$T = 100$			$T = 1000$		
	n2s ratio $\frac{\sigma_\epsilon^2}{\sigma_X^2}$	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.

We first consider the case of a single covariate with 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 positive and negatives and true positives being uninformative in a setting with no breaks. Since the active set of the initial estimator 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 $T = 100$; 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 in proportion to 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.

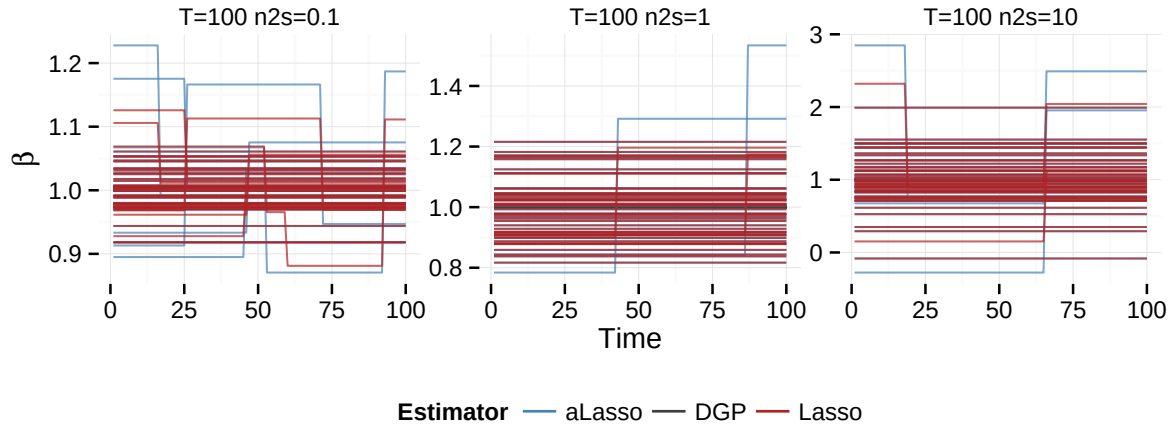

Figure 1: Sample of 50 estimated paths and the true path.

Figure 1 plots 50 estimated paths for the Lasso and adaptive Lasso (note that when no breaks were 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 1 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, in over 90% ($T = 100$) and 99% ($T = 1000$) of the iterations the estimator correctly estimates a constant path. 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.

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.

Across 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 to 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

		Break Location			Break Size			Break Number		
		10%	50%	90%	0.1	1	10	XP1	XP2	XP3
# 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.

the location of the break but is sensitive to its size, with FN < 1% when the break is of size 10 while FN ≈ 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 that 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 model the active set is larger leading to a higher number of false positive while keeping the number of true positive 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.

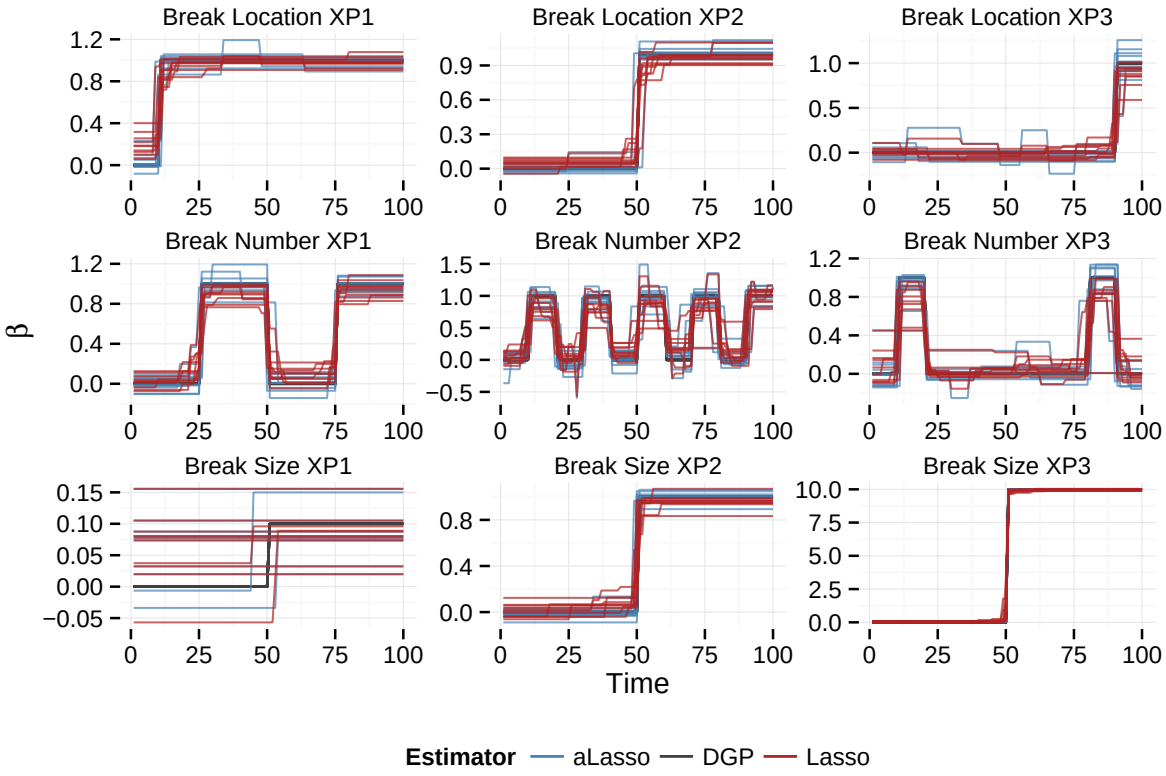

Figure 2: Structural breaks, 10 estimated paths.

Figure 2 plots 10 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 sample size is $T = 100$ and the variance of the innovations is equal to 0.1 in every experiment. The parameters follow parsimonious random walks as described by assumption 3. We vary the degree of sparsity of the model by considering $\alpha_T = 0.01, 0.1, 0.5$, for $\alpha_T = 0.01$ we expect a single break per path while when $\alpha_T = 0.5$ we expect 50 breaks. For convenience table 3 also reports the value of a corresponding to the chosen values of α_T and T, and in particular it should be noted that $\alpha_T = 0.5$ goes beyond the requirement for consistency of the Lasso given in theorem 2. 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 10% of the correct break locations. However when $\text{Var}(\eta) = 1$ or $\text{Var}(\eta) = 10$ the selected models tend to be slightly larger than the true models, and over 50% of the breaks are detected. In every experiment the adaptive Lasso

		Var(η) = 0.1			Var(η) = 1			Var(η) = 10		
	α_T	0.01	0.1	0.5	0.01	0.1	0.5	0.01	0.1	0.5
$a = -\frac{\log(\alpha_T)}{\log(T)}$		1	0.5	0.15	1	0.5	0.15	1	0.5	0.15
# breaks	DGP	0.981	9.912	49.59	0.998	9.952	49.46	0.995	9.922	49.41
	Lasso	0.993	6.354	16.74	2.218	14.85	42.27	3.156	21.29	52.18
	aLasso	0.566	3.303	9.425	0.982	7.405	26.04	1.05	8.95	30.08
FP	Lasso	0.828	4.628	6.293	1.723	9.95	15.58	2.384	13.73	17.3
	aLasso	0.44	2.15	3.116	0.582	3.667	7.082	0.408	2.79	4.393
TP	Lasso	0.165	1.726	10.44	0.495	4.896	26.69	0.772	7.55	34.88
	aLasso	0.126	1.153	6.308	0.4	3.738	18.95	0.642	6.16	25.69
FN	Lasso	0.816	8.186	39.15	0.503	5.056	22.77	0.223	2.372	14.54
	aLasso	0.856	8.759	43.28	0.598	6.214	30.51	0.352	3.762	23.73
ℓ_1 error	Lasso	0.206	0.334	0.445	0.22	0.39	0.568	0.231	0.454	0.783
	aLasso	0.241	0.331	0.453	0.237	0.373	0.576	0.237	0.415	0.792
	Post	0.211	0.358	0.509	0.228	0.445	0.717	0.242	0.531	0.898
ℓ_2 error	Lasso	0.056	0.132	0.206	0.066	0.167	0.269	0.071	0.211	0.474
	aLasso	0.076	0.131	0.21	0.075	0.156	0.271	0.073	0.185	0.471
	Post	0.058	0.133	0.205	0.065	0.164	0.268	0.068	0.185	0.369
RMSE	Lasso	0.313	0.315	0.318	0.311	0.306	0.284	0.311	0.32	0.464
	aLasso	0.305	0.308	0.31	0.305	0.301	0.283	0.312	0.324	0.473
	Post	0.311	0.306	0.297	0.307	0.287	0.239	0.306	0.278	0.314
λ	Lasso	0.024	0.019	0.012	0.022	0.01	0.005	0.02	0.01	0.008
	aLasso	0.003	0.017	0.059	0.016	0.076	0.225	0.363	1.618	3.218

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

selects models that are substantially sparser than those selected by the Lasso, and in doing so substantially decreases the number of true and false positives.

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 remarkably close to, but below, its theoretical value (≈ 0.316) for most experiments, with the exception of $\alpha_T = 0.5$ and Var(η) = 10. It 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 Var(η) increases. The estimator selects a small penalty parameter when the breaks to fit are larger; this larger penalty allows small parameters to be retained in the estimated model explaining the increase in the number of parameters retained.

Figure 3 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 3 display experiments where the variance

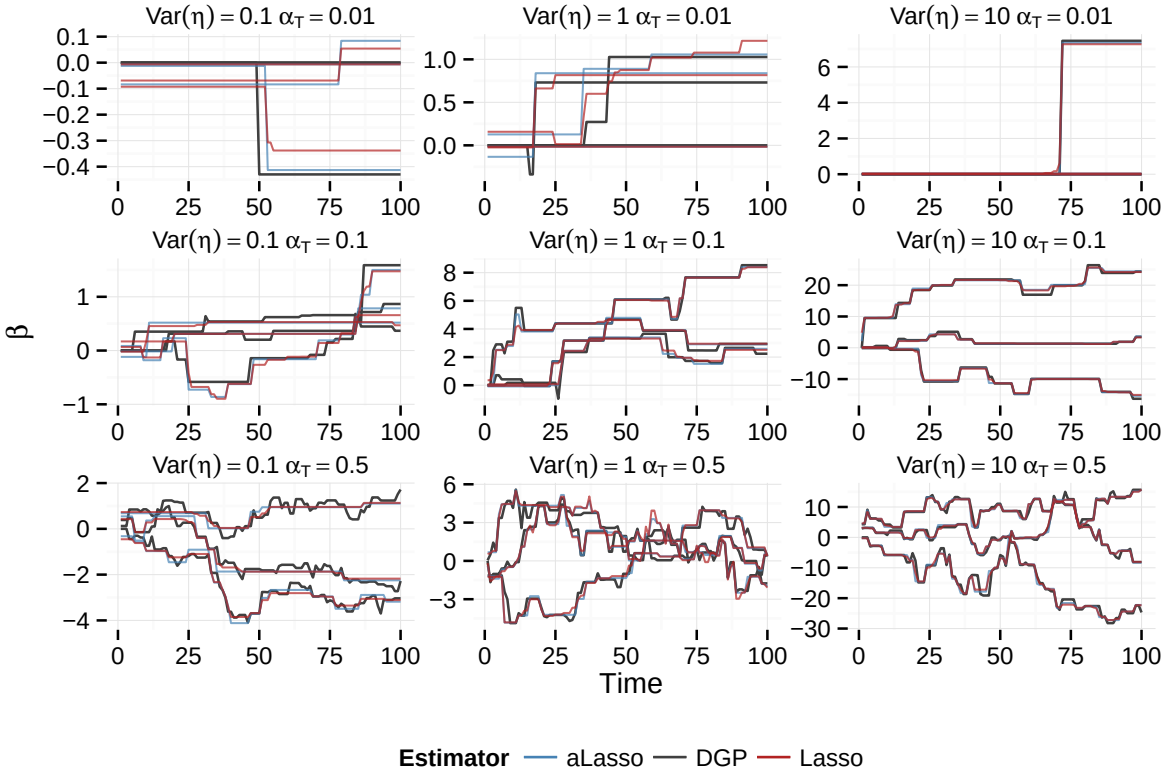

Figure 3: Parsimonious random walks, 3 estimated paths.

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 number of breaks increases (moving downward in figure 3) but more stable than the true path. When the variance of the breaks increases (moving rightwards in figure 3) 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.

5. Empirical Application

In order to illustrate the proposed methodology we revisit a well-known monetary policy problem, namely estimation of the Taylor (1993) rule. 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. Estimation of the Taylor rule is also used by Hansen, Lunde, and Nason (2011) to illustrate the model confidence set (MCS), and we choose to estimate one of the specifications included in the MCS. Specifically, we consider the general Taylor rule:

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

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). The parameters of main interest are the ones associated with the inflation and output variables: The monetary policy response to real side fluctuations is given by $\beta_{1,t} + \beta_{2,t}$; likewise response to inflation is given by $\alpha_{1,t} + \alpha_{2,t}$. The latter is of particular interest as the Taylor principle suggests that the response to inflation

should exceed 1 such that a rise in inflation results in an even larger rise in the interest rate. Compared with the specification used in Hansen et al. (2011), we let these key parameters be time-varying, i.e. we assume they follow a parsimonious random walk, thus allowing us to examine whether these responses have changed over time.

The model also contains 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. One could argue that the parameter associated with lagged interest rate, ρ, could also be time-varying. However, this would make it difficult to disentangle the time-varying nature of this smoothing parameter and the parameters associated with inflation and output, and hence we assume that it is constant. We report results for both the Lasso and the adaptive Lasso based on the methodology detailed in the previous sections. In both cases the estimator of the initial value of the parsimonious random walk is not penalized, and further, the penalty parameter, λ, is selected using the BIC.

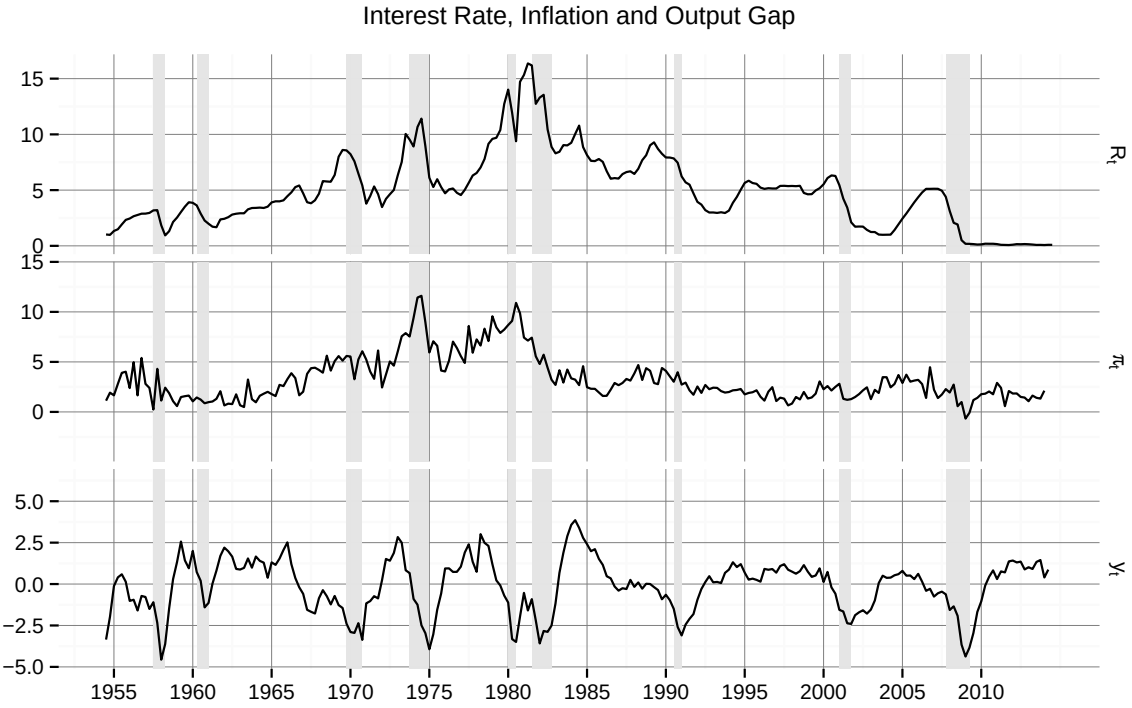

Figure 4: Plots of the data used for estimation of the Taylor rule. The variables are: Interest rate, R_t, inflation, π_t, and output gap, y_t. The vertical grey bars are the NBER recessions.

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 4.

If we assume all parameters are constant, then the model can be estimated by OLS as in Hansen et al. (2011). The resulting estimates are reported in table 4. Based on these results we would conclude that only the first lags of inflation and output are significant (at a 10%

	Estimate	Std. Error	t-value	p-value
γ	-0.4448	1.5443	-0.2881	0.7733
ρ	0.9195	0.0280	32.8038	0.0000
α_1	1.2751	0.7556	1.6876	0.0915
α_2	0.4084	0.6933	0.5891	0.5558
β_1	3.7687	1.8184	2.0725	0.0382
β_2	-1.9420	1.3231	-1.4678	0.1422

Note: The standard errors are heteroskedasticity and autocorrelation consistent.

Table 4: OLS estimation results for the Taylor rule equation under the assumption of constant parameters.

level) and that there are clear signs of interest rate smoothing (the lag of the interest rate is significant). Further, the overall response towards deviation of inflation from its target, $\alpha_1 + \alpha_2$, is clearly greater than one, thus showing that the Taylor principle is satisfied.

How does this compare to the case where we allow for time-varying parameters? Consider first figure 5 where we have plotted the time-varying estimates of $\alpha_{1,t}$, $\alpha_{2,t}$, as well as their sum. These plots show show clear signs of changes over time in the response to inflation, in several cases the estimates move outside the OLS confidence band.

The response to inflation 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. However there is clear evidence of instability from the start of the 1970s to the start of the 1980s, in line with the findings of Primiceri (2005). A first period from 1974 to the end of the 1970s is characterized by a weak monetary policy response in the face of increasing inflation. From the end of the 1970s response seems to follow a strong counter-cyclical pattern, with a strong response to inflation outside of recessions and a weak response during the two recessions of 1980 and 1983. From 2008 onwards the response to inflation drops to zero, reflecting the fact that the Fed Funds Rate has been at its lower bound since then.

Figure 6 gives the same illustration for the parameters associated with the output gap variables. The response to output gap is much more stable, and the variation is almost entirely contained within the OLS confidence band. However the response to the output gap seems to be higher from 1990 to 2008 than in the rest of the sample, this is particularly marked for the adaptive Lasso estimate, which could indicate a stronger concern for output smoothing by the Fed during that period. In general, though, there is little difference between the results for the Lasso and the adaptive Lasso with the one exception that, compared to the adaptive Lasso, the Lasso estimates are biased towards zero as we would expect. Further estimation results can be found in table 5.

The main insight gained from this analysis, compared to OLS estimation, is thus the time-varying nature of the central bank's response towards deviations of inflation from its target. In general, we have found that the large changes often coincide with the NBER recessions and, interestingly, that the Taylor principle in many cases is not satisfied. This is especially evident in the "double dip" recession of the early 1980s and the recent financial crisis. We should of course note that we do not have confidence bands for the time-varying parameters. Nonetheless, these results clearly illustrate the importance of taking structural instability into account when analysing macroeconomic relationships such as the Taylor rule, and the usefulness of our proposed methodology in this context.

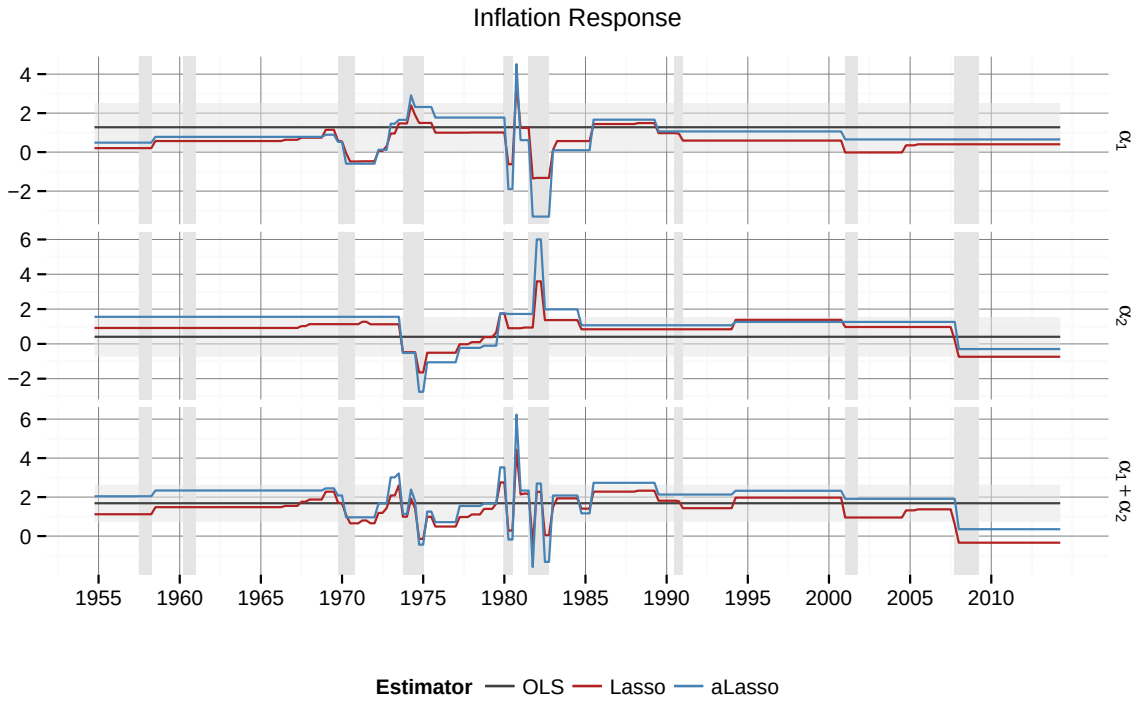

Figure 5: Parameters associated with the inflation variables. The horizontal grey bars are 90% confidence bands for the OLS estimates. The vertical grey bars are the NBER recessions.

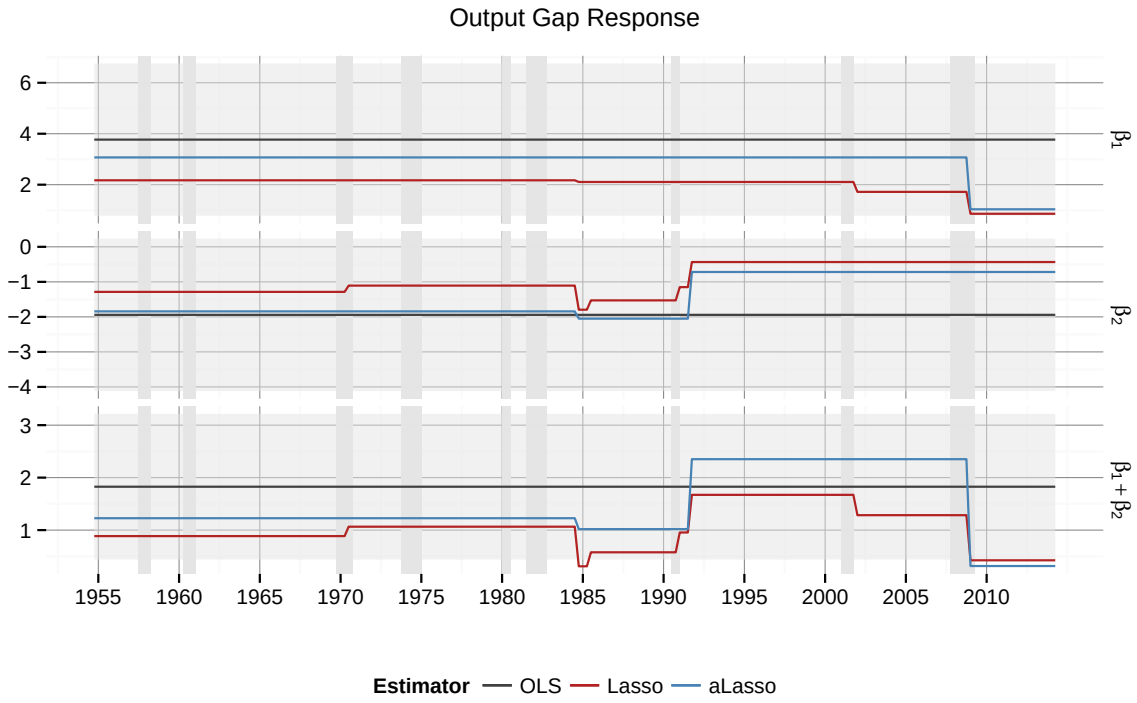

Figure 6: Parameters associated with the output gap variables. The horizontal grey bars are 90% confidence bands for the OLS estimates. The vertical grey bars are the NBER recessions.

	Avg. Est.		Min. Est.		Max. Est.		# Breaks	
	Lasso	aLasso	Lasso	aLasso	Lasso	aLasso	Lasso	aLasso
α_1	0.5732	0.7994	-1.3475	-3.3082	3.5313	4.5103	38	18
α_2	0.7263	1.0385	-1.6402	-2.7544	3.5938	6.0039	21	12
β_1	1.9789	2.8825	0.8585	1.0341	2.1720	3.0699	3	1
β_2	-0.9460	-1.4393	-1.7954	-2.0514	-0.4339	-0.7183	5	2
γ	0.5622	-0.8356						
ρ	0.8579	0.8961						

Note: As γ and ρ are not time-varying the reported average estimate is the actual estimate. Also, the selected tuning parameters, λ, are 0.0091 and 0.2049 for the Lasso and adaptive Lasso, respectively.

Table 5: Lasso and adaptive Lasso estimation results for the Taylor rule equation.

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; because of the parsimonious random walk assumption the vector of increments is sparse, and by construction it is high dimensional. We derive bounds on the precision of the Lasso in finite samples and show that the Lasso can estimate the vector of increments consistently as long as the number of non-zero parameters increases strictly slower than $\sqrt{T}$. We establish oracle results for the adaptive Lasso under the same assumptions. Because of the convexity of the Lasso's objective function, our estimator is computationally fast.

We apply our model to the estimation of a Taylor rule 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 1980s, which is consistent with previous research and historical facts, and a long lasting change in the response since 2008, driven by the fact that the Fed Funds Rate has essentially been 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 inference 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, expand the parsimoniously time varying approach to other time series models such as stochastic volatility models, score driven models, and factor models.

Appendix A. Proofs

A.1. Proofs for the Lasso

The following lemma is similar to theorem 1 in Kock and Callot (2014) and provides bounds on the prediction and estimation error without making use of the restricted eigenvalue assumption.

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

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

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

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

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

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

We can thus rewrite (A.1) as

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

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

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

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

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

which is (A.1). Note that

$$\begin{aligned} \|\hat{\theta} - \theta\|_{\ell_1} + \|\theta\|_{\ell_1} - \|\hat{\theta}\|_{\ell_1} &= \|\hat{\theta}_J - \theta_J\|_{\ell_1} + \|\theta_J\|_{\ell_1} - \|\hat{\theta}_J\|_{\ell_1} \\ &\leq 2\|\hat{\theta}_J - \theta_J\|_{\ell_1} \end{aligned}$$

using continuity of the norm, and

$$\|\hat{\theta}_J - \theta_J\|_{\ell_1} + \|\theta_J\|_{\ell_1} - \|\hat{\theta}_J\|_{\ell_1} \leq 2\|\theta_J\|_{\ell_1}$$

by sub-additivity of the norm. Using the two results above in (A.1) yields (A.2). Finally notice that (A.2) gives

$$\lambda_T \|\hat{\theta} - \theta\|_{\ell_1} \leq 4\lambda_T \|\hat{\theta}_J - \theta_J\|_{\ell_1}$$

or equivalently

$$\|\hat{\theta}_{J^c} - \theta_{J^c}\|_{\ell_1} \leq 3\|\hat{\theta}_J - \theta_J\|_{\ell_1}$$

which establishes (A.3). □

Lemma 2. *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* $A > 0$,

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

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

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

Using lemma 5 on the second term yields a first bound

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

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_{kt}, $k = 1, ..., r_X$.

Since Z_{tk} is measurable on $\sigma\left(\left\{\mathscr{F}_{Z,\epsilon,t-1}, Z_{tk}\right\}\right)$ we use lemma 4 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 $\{\epsilon_t Z_{tk} \mathbb{1}_{|\epsilon_t Z_{tk}|<L_T}\}$ is a martingale difference sequence with bounded increments. We can thus apply the Azuma-Hoeffding inequality to bound the first term.

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

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

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

□

Proof of Theorem 1. On $\mathscr{B}_T$ and under assumptions 1, 2, 3, and 4, we use equations (A.2) and (A.3) from lemma 1 and Jensen's inequality to get:

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

Rearranging yields (3). We also get

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

which is (4). Lemma 2 gives the probability of being on $\mathscr{B}_T$. □

Proof of Corollary 1. To prove this result, assume that $\hat{\theta}_j = 0$ for $j \in J$, 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. Observe that $s\lambda_T \to 0$ implies that from some step $T > T_0$ onward, $\lambda_T < 1$ so that $s\lambda_T^2 \to 0$. Also note that $P(\mathscr{B}_T) \to 1$, hence if we show that $s\lambda_T \to 0$ we can show that the bounds of theorem 1 tend to zero.

Let $a > 1/2$, then

$$s^2\lambda_T^2 \in \mathcal{O}\left(T^{2-2a}\frac{\ln(1+T)^5\ln(1+r)^2\ln(r(T-r+1)))}{T}\right)$$
$$\in \mathcal{O}\left(\ln(1+T)^5\ln(r(T-r+1))T^{1-2a}\right)$$

so that $s\lambda_T \to 0$ and $s\lambda_T^2 \to 0$.

It follows that

$$T^{-1}\left\|Z^D W(\theta-\hat{\theta})\right\|^2 \leq \frac{16s\lambda_T^2}{\kappa_T^2} \to^p 0, \tag{A.5}$$

$$\left\|\hat{\theta}-\theta\right\|_{\ell_1} \leq \frac{16s\lambda_T}{\kappa_T^2} \to^p 0, \tag{A.6}$$

which proves (5) and (6). □

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

Lemma 3. *Let $\{U_i\}$, $i = 1,...,n$, $n < \infty$ be a sequence of independent Gaussian random variables with mean zero and variances $\sigma_i^2 < \infty$; define $\sigma^2 = \max_i(\sigma_i^2)$. Then we have*

i) $\prod_i U_i$ has sub-exponential tails: $P(|\prod_i U_i| > x) \leq 2^n e^{-\frac{x}{2\sigma^2}}$.

ii) $\sum U_i$ has sub-exponential tails: $P(|\sum_i U_i| > x) \leq 2ne^{-\frac{x}{2n\sigma^2}}$.

Proof. Since U_i is Gaussian with mean zero it has sub-exponential tails (see e.g. Billingsley (1999), page 263) so that there exists constants K and C such that for every $x > 0$, $P(|X| > x) \leq Ke^{-Cx}$. Let $i, j = 1,...,n$, $i \neq j$, we first prove i)

$$\begin{aligned} P\left(|\prod_i U_i| > x\right) &\leq \prod_i P\left(|U_i| > \sqrt[n]{x}\right) \\ &\leq \prod_i 2e^{-\frac{x}{2\sigma_i^2}} \\ &\leq 2^n e^{-\frac{x}{2\sigma^2}}. \end{aligned}$$

We now prove ii)

$$\begin{aligned} P\left(|\sum U_i| > x\right) &\leq \sum P\left(|U_i| > x/n\right) \\ &\leq \sum_i 2e^{-\frac{x}{2n\sigma_i^2}} \\ &\leq 2ne^{-\frac{x}{2n\sigma^2}}. \end{aligned}$$

□

Lemma 4 ((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

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

Lemma 5 (Lemma 2 in Kock and Callot (2014))**.** *Let assumptions 1 and 2 be satisfied. Then, for some positive constant A and for any $L_T > 0$,*

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

A.2. Proofs for the adaptive Lasso.

The proofs of lemma 6 and theorem 3 are very similar to those of lemma 11 and theorem 6 in Kock and Callot (2014), hence we only sketch these proof.

Lemma 6 (Lemma 11 in Kock and Callot (2014))**.** *Let*

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

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

Proof. Proof of theorem 3 $\text{sign}(\tilde{\theta}) = \text{sign}(\theta)$ if and only if the following two conditions are met. First let $i \in \mathcal{S}^c$

$$\left|\Psi_{i,\widehat{\mathcal{S}}}\left(\Psi_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\right)^{-1}\left(\frac{(Z^D W_{\widehat{\mathcal{S}}})'\epsilon}{T} - \lambda_T \text{sign}(\theta_i)\, w_i\right) - \frac{(Z^D W_i)'\epsilon}{T}\right| \le$$
$$\left|\Psi_{i,\widehat{\mathcal{S}}}\left(\Psi_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\right)^{-1}\left(\frac{(Z^D W_{\widehat{\mathcal{S}}})'\epsilon}{T} - \lambda_T \text{sign}(\theta_i)\, w_i\right)\right| + \left|\frac{(Z^D W_i)'\epsilon}{T}\right| \le \lambda_T w_i. \tag{A.7}$$

The second condition is

$$\text{sign}\left(\theta_{\widehat{\mathcal{S}}} + \left(\Psi_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\right)^{-1}\left(\frac{(Z^D W_{\widehat{\mathcal{S}}})'\epsilon}{T} - \lambda_T \text{sign}(\theta_{\widehat{\mathcal{S}}})\, w_{\widehat{\mathcal{S}}}\right)\right) = \text{sign}(\theta_{\mathcal{S}}). \tag{A.8}$$

Theorem 6 in Kock and Callot (2014) shows that the left side of (A.7) can be bounded from above by

$$\left|\Psi_{i,\widehat{\mathcal{S}}}\left(\Psi_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\right)^{-1}\left(\frac{(Z^D W_{\widehat{\mathcal{S}}})'\epsilon}{T} - \lambda_T \text{sign}(\theta_i)\, w_i\right)\right| - \left|\frac{(Z^D W_i)'\epsilon}{T}\right| \le \frac{sK_T}{q\phi_{\min,\mathcal{S}}}\left(\frac{\lambda_T}{2} + \frac{2\lambda_T}{\theta_{\min}}\right) + \frac{\lambda_T}{2},$$

and the right side of (A.7) s bounded from below by $\left|\lambda_T w_i\right| \ge \frac{\lambda_T}{\left\|\widehat{\theta}-\theta\right\|_{\ell_1}}$. We replace these bounds in (A.7) and divide both sides by the right side bound to get (8).

For the condition (A.8) to be verified it suffices to show that

$$\left\| \left(\Psi_{\widehat{\mathscr{F}},\widehat{\mathscr{F}}} \right)^{-1} \left(\frac{(Z^D W_{\widehat{\mathscr{F}}})' \epsilon}{T} - \lambda_T \operatorname{sign}(\theta_{\widehat{\mathscr{F}}}) w_{\widehat{\mathscr{F}}} \right) \right\|_{\ell_\infty} \leq \theta_{\min} \tag{A.9}$$

which Kock and Callot (2014) shows to be satisfied if (9) is satisfied.

Lemmas 6 and 2 provide the desired bound on $P\left(\mathscr{B}_T \bigcap \mathscr{C}_T\right)$ which completes the proof. □

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 1, we then turn to the other conditions.

1. $P(\mathscr{B}_T \cap \mathscr{C}_T) \to 1$ which can be seen to hold from lemmas 6 and 2.
2. To show that $\theta_{\min} \geq 2\left\|\hat{\theta} - \theta\right\|_{\ell_1}$ is asymptotically valid, recall that from (4):

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

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

$$\frac{\left\|\hat{\theta} - \theta\right\|_{\ell_1}}{\theta_{\min}} \in O_p\left(\frac{\ln(1+T)^{5/2} \ln(1+T-r+1) \ln(r(T-r+1))^{1/2} T^{1/2-a}}{\ln(T) \ln(1+T)^{5/4} \ln(1+T-r+1)^{1/2} \ln(r(T-r+1))^{1/4} T^{-a/4}}\right) = o_p(1)$$

   so that $\theta_{\min} \geq 2\left\|\hat{\theta} - \theta\right\|_{\ell_1}$ with probability 1.
3. Recall that by assumption 4, κ_T^2 and κ^2 (and hence $\phi_{\min,\mathscr{S}}$) is bounded from below away from zero. To show that (8) holds asymptotically, we replace $\left\|\hat{\theta} - \theta\right\|_{\ell_1}$ by its upper bound from (4) and we are left to show that $s^2 K_T \lambda_T + \frac{s^2 K_T \lambda_T}{\theta_{\min}} + s\lambda_T \to 0$. Notice that $s^2 K_T \lambda_T = b_T \to 0$ which takes care of the first term. Regarding the second term: $\frac{s^2 K_T \lambda_T}{\theta_{\min}} = \frac{1}{\ln(T)} \to 0$, and for the third term: $s\lambda_T \in O_p(T^{1/2-a})$ so that $s\lambda_T \to 0$ if $a > 1/2$.
4. To show that (9) holds asymptotically we have to show that $\frac{\sqrt{s}\lambda_T}{\theta_{\min}} + \frac{\sqrt{s}\lambda_T}{\theta_{\min}^2} \to 0$. The first term:

$$\begin{aligned}\frac{\sqrt{s}\lambda_T}{\theta_{\min}} &\in O_p\left(\frac{\ln(1+T)^{5/2} \ln(1+T-r+1) \ln(r(T-r+1))^{1/2} T^{-a/2}}{\ln(T) \ln(1+T)^{5/4} \ln(1+T-r+1)^{1/2} \ln(r(T-r+1))^{1/4} T^{-a/4}}\right) \\ &\in O_p\left(\ln(1+T)^{5/4} \ln(1+T-r+1)^{1/2} \ln(r(T-r+1))^{1/4} T^{-a/4} \ln(T)^{-1}\right) = o_p(1)\end{aligned}$$

   the second term:

$$\begin{aligned}\frac{\sqrt{s}\lambda_T}{\theta_{\min}^2} &\in O_p\left(\frac{\ln(1+T)^{5/2} \ln(1+T-r+1) \ln(r(T-r+1))^{1/2} T^{-a/2}}{\ln(T)^2 \ln(1+T)^{5/2} \ln(1+T-r+1) \ln(r(T-r+1))^{1/2} T^{-a/2}}\right) \\ &\in O_p\left(\ln(T)^{-2}\right) = o_p(1).\end{aligned}$$

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.

Billingsley, P. (1999). Convergence of Probability Measures (2nd ed.). John Wiley & Sons.

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.

Hoffmann-Jørgensen, J. (1994). Probability with a View Towards Statistics. 1. CRC 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. A. Callot (2014). Oracle inequalities for high dimensional vector autoregressions. Journal of Econometrics. Conditionally accepted.

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. In press.

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

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.

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