Penalized Autoregressive Conditional Betas*

Christian Francq
CREST, Institut Polytechnique de Paris and University of Lille
and
Sébastien Laurent
Aix-Marseille University (Aix-Marseille School of Economics),
CNRS & EHESS, Aix-Marseille Graduate School of Management,
IAE and Institut Universitaire de France (IUF), France
and
Julie Schnaitmann
Eberhard Karls Universität Tübingen, Germany

March 24, 2026

Abstract

We examine the estimation of a linear regression model with time-varying slope coefficients (betas), called Autoregressive Conditional Beta. This model is unidentified if some betas are constant (or zero). To address this non-identifiability issue, we employ a Lasso-type estimator. This penalized estimator simplifies the model by shrinking the estimates in favor of natural constant beta representations. We propose a multistep estimator that first captures the dynamics of the regressors before estimating the dynamics of the betas. This strategy breaks down a large-dimensional optimization problem into several lower-dimensional ones. Since we avoid making strict parametric assumptions about the innovation distributions, we use quasi-maximum likelihood estimators. The non-Markovian nature of the global model means that standard convex optimization results cannot be applied. We analyze the asymptotic distribution of the multistep Lasso estimator and its adaptive version, deriving bounds on the maximum value of the penalty term. We also propose a nonlinear coordinate-wise descent algorithm, which is demonstrated to find stationary points of the objective function. The finite-sample properties of these estimators are further explored through a Monte Carlo simulation and illustrated with an application to financial data.

Keywords: Time series, Linear model, Time-varying coefficients, Penalized likelihood

*Francq and Laurent acknowledge the research support of the French National Research Agency Grants ANR-21-CE26-0007-01 while Laurent also acknowledges the research support of the French National Research Agency Grants ANR-17-EURE-0020. Emails: christian.francq@univ-lille3.fr, sebastien.laurent@univ-amu.fr (corresponding author) and julie.schnaitmann@uni-tuebingen.de. The data that support the findings of this study are available from the corresponding author, upon reasonable request.

1 Introduction

The linear regression model is a basic tool that allows to explore and quantify relationships between different variables. These variables are often recorded over time. Under stationarity and the existence of moments conditional on past observations, the conditional beta coefficient of a linear regression is well defined, but like any other conditional moment, it is likely to be time-varying. In a general specification, the linear regression model with time-varying slope coefficients (betas) and GARCH(1,1) residual dynamics takes the form

$$y_t = \beta_{1,t}x_{1,t} + \cdots + \beta_{p,t}x_{p,t} + v_t, \qquad v_t = g_t\eta_t, \tag{1}$$

where $E(\eta_t) = 0$, $E(\eta_t^2) = 1$, and the conditional variance evolves as $g_{t+1}^2 = \omega + \alpha v_t^2 + \beta g_t^2$.

Although the regression is linear in the regressors at each point in time, allowing the slope coefficients to vary over time introduces a flexible source of nonlinearity into the overall data-generating process while maintaining a clear economic interpretation of the parameters. Such specifications are particularly well suited to settings in which the effects of explanatory variables may evolve in response to changing economic conditions. This type of model was used for instance by Engle (2016), Reh et al. (2023) and Blasques et al. (2024) to obtain time-varying weights of tracking portfolios, but also by Darolles et al. (2018) to obtain conditional covariance or precision matrices using a system of equations analogous to (1). Several methods have been used in the literature to obtain time-varying betas, including instrumental variables (Gagliardini et al. 2016), realized betas (Barndorff-Nielsen & Shephard 2004), state space models (Hamilton 1994, Durbin & Koopman 2001) and dynamic conditional beta (DCB) models (Bollerslev et al. 1988, Engle 2016).

Blasques et al. (2024), hereafter referred to as BFL, recently proposed a model for time-varying conditional betas, called the Autoregressive Conditional Beta (ACB) model, which can be viewed as an analogue of the GARCH(1,1) models for the modelling of the betas.

The ACB model is a conditional score driven model in which the time-varying conditional betas are based on the (scaled) score function of a Gaussian predictive model density and an autoregressive component. In its simplest form, the ACB model specifies the dynamics of each beta as

$$\beta_{i,t+1} = \varpi_i + \xi_i \frac{v_t x_{i,t}}{\mu_{i,t}^2 + g_{i,t}^2} + c_i \beta_{i,t}, \tag{2}$$

where $\mu_{i,t}$ and $g_{i,t}^2$ are the conditional mean and conditional variance of $x_{i,t}$, respectively. The time-varying conditional beta $\beta_{i,t}$ is driven by three main components: a constant term ϖ_i, a score-driven updating term scaled by the coefficient ξ_i and an autoregressive component governed by the persistence parameter c_i. BFL establishes the existence of a stationary solution for the ACB model, the invertibility of the filter for the time-varying betas, and the asymptotic properties of the Quasi Maximum Likelihood estimator (QMLE).

Despite its appeal, the ACB framework faces a fundamental and largely overlooked difficulty in empirical implementation. In many applications, some coefficients are in fact constant or equal to zero, even though the model allows them to vary over time. In such cases, the unrestricted ACB parametrization is no longer identified: multiple parameter configurations generate the same constant beta, while the quasi-maximum likelihood estimator (QMLE) typically fails to set the dynamic parameters exactly to zero in finite samples. As a result, the estimated beta paths may display spurious time variation, creating the appearance of economically meaningful dynamics where none exist.

This issue is not merely technical. In practice, researchers and practitioners rely on the estimated paths of conditional betas to interpret economic mechanisms, construct portfolios, or assess risk exposures. Spurious time variation may therefore lead to misleading conclusions, unstable decisions, and poor out-of-sample performance (as illustrated in the empirical application). Moreover, the problem becomes more severe as the number of po-

tential regressors increases, since the presence of irrelevant or constant coefficients may contaminate the estimation of all parameters through lack of identification.

In this paper, we introduce the Penalized Autoregressive Conditional Betas (PACB) model that uses a L^1-penalty to shrink conditional betas to a constant. The penalty is imposed on the parameters governing the conditional betas, more specifically the score update and autoregressive parameters, ξ_i and c_i. Both parameters are shrunk to zero which resolves the non-identification problem by favoring the solution $\beta_i = \varpi_i$. Additionally, the PACB model also allows for classical variable selection in the sense that it shrinks irrelevant betas to zero. This is achieved in a second penalization step. We show that it is not advisable to penalize ξ_i, c_i, and ϖ_i at once, as this creates another identification problem; rather, we must first identify the betas for which c_i is set to 0 before penalizing ϖ_i.

The PACB model enables variable selection in a linear regression framework with time-varying conditional betas and volatility. In particular, it distinguishes between time-varying coefficients $\beta_{i,t}$, constant but nonzero coefficients $\beta_i \neq 0$, and zero coefficients $\beta_i = 0$.

Since variable selection via the Least Absolute Shrinkage and Selection (Lasso) method introduced by Tibshirani (1996) is known to be a powerful technique for building parsimonious and interpretable regression models, it seems natural to employ this technique to select an PACB model. The Lasso is so extensively used for estimation and model selection that an exhaustive review of its applications is practically impossible. However, its use in time series models remains relatively limited and has primarily focused on Markovian models and Least Squares (LS) type estimators. In these settings, penalized estimators are typically defined as the solution to optimization problems of the form $Q(\boldsymbol{\vartheta}) + \lambda p(\boldsymbol{\vartheta})$, where $Q(\boldsymbol{\vartheta})$ is a quadratic loss function, and $p(\boldsymbol{\vartheta})$ is a penalty term (see Section 5 for details). Quadratic objective functions $Q(\boldsymbol{\vartheta})$ are, for instance, obtained for AutoRegressive (AR) models and AutoRegressive Conditional Heteroskedasticity (ARCH) models estimated by LS. In the framework of L^1 penalization of quadratic objective functions, Knight & Fu

(2000) derive the asymptotic behavior of the Lasso for linear regressions and autoregressions, Wang et al. (2007) consider Lasso for linear regression models with AR errors, Nardi & Rinaldo (2011) consider penalized LS estimators for AR models when the order grows with the sample size, Basu & Michailidis (2015) extend Lasso to high-dimensional vector autoregressive (VAR) models, Kock (2016) studies the Adaptive Lasso of Zou (2006) for non-stationary AR processes, and Poignard & Fermanian (2021) study grouped Lasso for multivariate ARCH models. When the objective function $Q(\boldsymbol{\vartheta})$ is quadratic, the L^1-penalized estimator can be computed without resorting to numerical optimization routines. In such cases, efficient algorithms like LARS (Efron et al. 2004) and the Shooting algorithm (Fu 1998) can be used to compute the Lasso estimator for any given penalty level λ. Lasso regressions and autoregressions can also be applied to high-dimensional settings where the number of predictors p may exceed the sample size n, as shown in Adamek et al. (2023).

In the present work, we consider a time series model that departs from the Markovian structure. Specifically, we use GARCH instead of ARCH models to better capture the persistent nature of financial return volatility. Additionally, we adopt Quasi-Maximum Likelihood (QML)-type estimators rather than LS, in order to downweight periods of high volatility and thus improve estimation accuracy. As a result, the objective function $Q(\boldsymbol{\vartheta})$ is no longer quadratic or even convex. The use of non-convex penalty functions $p(\boldsymbol{\vartheta})$ in conjunction with quadratic (and thus convex) loss functions $Q(\boldsymbol{\vartheta})$ has been explored by Loh (2017). Wang et al. (2014) study regularized robust estimators in the context of sparse, high-dimensional linear models under *i.i.d.* observations. Penalized likelihood approaches have been proposed by Fan & Li (2001) and Fan & Peng (2004) for independent data. More recently, Nielsen & Rahbek (2024) extended this analysis to accommodate time series data with temporal dependence, and to allow for parameters that lie on the boundary of the parameter space. Alami Chentoufi (2024) proposes a two-step Lasso estimation procedure for ARMA and GARCH-type models: an initial unpenalized estimation is performed, followed

by a penalized regression on the lagged values of the observed process and the innovations obtained from the first step, leading to model selection consistency.

We complement these seminal papers in several ways. First, we do not consider regressions on *i.i.d.* observations, but rather on the components of a multivariate time series. Second, we allow for time-varying beta coefficients, which is natural since conditional moments typically evolve over time. Third, as a consequence of the previous point, the objective functions we optimize are not convex, and we develop new mathematical arguments to address this issue. Fourth, we use penalized estimators not only for variable selection, but also to eliminate unidentified parameters by shrinking the estimated model toward its simplest form. Finally, to ensure that our method is easy to implement, we rely on multi-step estimation procedures. Each step estimates a subset of the parameters with QMLE. Such estimators present the advantage of being consistent in a semi-parametric framework.

Our approach requires the estimation of a relatively small number of models to identify a terminal model. This contrasts with the 'Best subset selection' method, which would require the estimation of a huge number of models before selecting the terminal model using information criteria. Indeed, in the simplest setting each conditional beta implies $8-2=6$ plausible specifications, so that for a model with p regressors, 6^p specifications are possible.[1] This leads to $46,656$ different specifications for $p=6$, to $1,679,616$ specifications for $p=8$ and to $60,466,176$ specifications for $p=10$.

Following this line of argument also illustrates why testing for the constancy of betas is conceptually difficult: the model is non-identified as soon as one conditional beta is constant but specified as in (2). In order to test both the constancy of several betas and the nullity of certain betas, a large sequence of hypothesis tests (not only t-tests but also Wald tests) would need to be used to reduce the full ACB model in line with the Autometrics method (Doornik et al. 2009). However, theoretical properties of Autometrics are very difficult to

[1] The two implausible specifications are those with ϖ_i either set to 0 or left unconstrained and $\xi_i = 0$.

derive and are mainly evaluated through Monte Carlo simulations, as shown by Hendry & Johansen (2011). Our approach does not build on testing for constancy but is rather based on variable selection via shrinkage.

Section 2 provides an informal illustration of the proposed estimator. The ACB model and the (unpenalized) QMLE are presented in Section 3. Section 4 defines a penalized Lasso-type estimator and studies its asymptotic distribution. Section 5 discusses the optimization algorithm and penalty parameter selection, while Section 6 provides Monte Carlo evidence on the finite sample performance of the penalized estimator. Section 7 contains an empirical application to financial returns. The supplementary material is composed of the assumptions and proofs, our proposed optimization algorithm, and additional tables and graphs for the simulation study and the empirical application.

2 Variable selection in application

To illustrate the usefulness of our shrinkage approach to select an appropriate model, we estimate an ACB model using the daily (excess) returns of the *Durbl* sector portfolio (durable consumption) as the dependent variable for the period from January 2000 to December 2024 (yielding 6,225 observations). The model includes 11 variables. $x_{1,t} = 1 \forall t$ so that $\beta_{1,t}$ is a time-varying intercept, $x_{2,t}, \ldots, x_{6,t}$ are the five Fama-French risk factors, FF5 (see Fama & French 1993, 2015), i.e., MKT, the market factor proxied by log returns on the S&P 500 index in excess of the risk-free rate, SMB, *small minus big*, HML, *high minus low*, RMW, *robust minus weak*, and CMA, *conservative minus aggressive*. To illustrate the problem of non-identification when $\xi_i = 0$, we augment the model by five irrelevant simulated factors $(x_{7,t}, \ldots, x_{11,t})$ generated by a DCC-GARCH(1,1) model (as described in Section 6). The last five betas are known to be constant and even equal to zero, so $\varpi_i = \xi_i = c_i = 0$ for $i = 7, \ldots, 11$. Under these population restrictions, the model

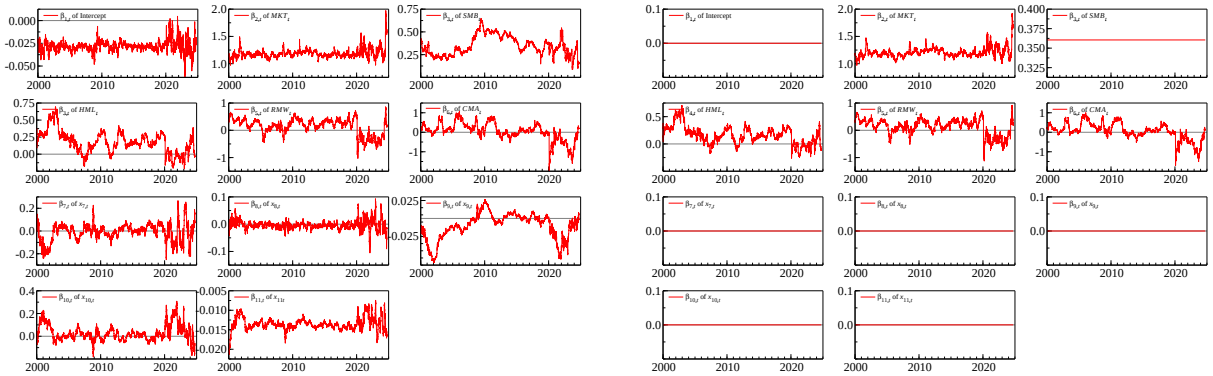

Figure 1: Time series of the conditional betas for series *Durbl* of the full ACB model estimated by QMLE on the left and the PACB on the right.

is not identified and therefore the full ACB does not consistently estimate the structural parameters of the data generating process.

In the left panel, the model is estimated by full ACB and the obtained conditional betas do not reflect that the last five betas are constant and equal to zero. While $\xi_i = 0$ and, hence, $\beta_{i,t} = \beta_i \; \forall t$ in the population for $i = 7, \ldots, 11$, the QMLEs of these parameters are different from 0, which leads to estimated time-varying conditional betas. Some of the estimated betas are rough (like $\beta_{8,t}$) while others are smoother (like $\beta_{9,t}$). Failing to detect the constancy of a specific beta leads to a non-identified model and to inconsistent estimators for all the parameters and non-constant trajectories of the conditional betas. Furthermore, it is not possible to conclude from the left panel of Figure 1 which conditional betas should be constrained to be constant and/or zero.

The right panel of Figure 1 reports the estimated conditional betas obtained using the PACB method. The results indicate that the conditional betas associated with the five simulated irrelevant factors are correctly shrunk to zero. Among the remaining coefficients, four are identified as time-varying, while one factor (SMB) is detected as constant but significantly different from zero. Overall, the proposed approach effectively identifies irrelevant regressors (i.e., $\beta_i = 0$). We extend this analysis for 12 industry portfolios in Section H in the Appendix. We find comparable patterns for all industry portfolios.

3 The Autoregressive Conditional Beta (ACB) model

This section presents the ACB model (with exogenous variables) proposed by BFL and its estimation by QML. Let $(y_t, \boldsymbol{x}_t^\top)$ be a stationary time series of $p+1$ random variables, and $\mathcal{F}_t$ be the information available at time t, given by $\{y_u, \boldsymbol{x}_u; u \leq t\}$ and possibly some vector $\boldsymbol{z}_t = (z_{1t}, \ldots, z_{qt})^\top$ of exogenous variables. Assume that, given $\mathcal{F}_{t-1}$, the conditional variance of $(y_t, \boldsymbol{x}_t^\top)$ exists and is almost surely non singular.

3.1 The ACB model in a nutshell

Assuming simple GARCH(1,1) models for the regressors and for the regression error term v_t, and using a (quasi) score-driven approach, BFL proposed the following regression model with time-varying conditional betas

$$\begin{aligned} x_{i,t} &= \mu_{0i} + \varepsilon_{i,t}, \quad \varepsilon_{i,t} = g_{i,t}\eta_{i,t}, \\ g_{i,t+1}^2 &= \omega_{0i} + \alpha_{0i}\varepsilon_{i,t}^2 + \beta_{0i}g_{i,t}^2, \\ \beta_{i,t+1} &= \varpi_{0i} + \xi_{0i}\tfrac{v_t x_{i,t}}{\mu_{0i}^2 + g_{i,t}^2} + c_{0i}\beta_{i,t} + \gamma_{01,i}z_{1,t} + \cdots + \gamma_{0q,i}z_{q,t}, \\ g_{t+1}^2 &= \omega_0 + \alpha_0 v_t^2 + \beta_0 g_t^2, \end{aligned} \tag{3}$$

where $\mu_{0i}, \varpi_{0i}, \xi_{0i}, c_{0i}, \gamma_{0ji} \in \mathbb{R}$ for $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$, $\omega_{0i} > 0$, $\alpha_{0i} > 0$ and $\beta_{0i} \geq 0$ for $i = 1, \ldots, p$, $\omega_0 > 0$, $\alpha_0 > 0$ and $\beta_0 \geq 0$, and $(\boldsymbol{\eta}_t)$, with $\boldsymbol{\eta}_t = (\eta_{1t}, \ldots, \eta_{pt}, \eta_t)^\top$, is an *i.i.d.* sequence satisfying the conditions $E(\eta_t) = E(\eta_{it}) = 0$ and $E(\eta_t^2) = E(\eta_{it}^2) = 1$.

In a financial application, BFL find evidence for time-varying conditional betas and highlight the empirical relevance of the ACB model in a portfolio and risk management empirical exercise. However, they experience difficulties in selecting constant and zero betas. The beta $\beta_{i,t}$ in (3) is constant iff under the regularity conditions of Appendix A

$$\xi_{0i} = \gamma_{01,i} = \cdots = \gamma_{0q,i} = 0. \tag{4}$$

As explained above, when this relation holds, the parameters ϖ_{01} and c_{0i} are not identifiable because $\beta_{i,t} = \beta_i \equiv \varpi_{i0}/(1 - c_{0i})$ for an infinite number of values of (ϖ_{i0}, c_{0i}). In this paper, we will define a Lasso-type estimator that solves this identifiability problem by favoring the solution $c_{0i} = 0$ when (4) holds.

3.2 Multi-step QMLE

The observations are $(y_t, \boldsymbol{x}_t, \boldsymbol{z}_t)$ for $t = 1, \ldots, n$. In a first step, the GARCH(1,1) parameters $\boldsymbol{\theta}_0^{(i)} = (\mu_{0i}, \omega_{0i}, \alpha_{0i}, \beta_{0i})^\top \in \Theta \subset \mathbb{R} \times (0, \infty)^2 \times [0, 1)$ can be estimated in parallel for $i = 1, \ldots, p$,

$$\widehat{\boldsymbol{\theta}}^{(i)} = \arg\min_{\boldsymbol{\tau} \in \Theta} \widetilde{O}_n^{(i)}(\boldsymbol{\tau}), \qquad \widetilde{O}_n^{(i)}(\boldsymbol{\tau}) = \frac{1}{n} \sum_{t=2}^{n} \widetilde{\ell}_{i,t}(\boldsymbol{\tau}),$$

by the standard QMLE, where $\boldsymbol{\tau} = (\mu, \omega, \alpha, \beta)^\top$ denotes a generic element of Θ and

$$\widetilde{\ell}_{i,t}(\boldsymbol{\tau}) = \frac{(x_{i,t} - \mu)^2}{\widetilde{g}_{i,t}^2(\boldsymbol{\tau})} + \log \widetilde{g}_{i,t}^2(\boldsymbol{\tau}), \quad \widetilde{g}_{i,t}^2(\boldsymbol{\tau}) = \omega + \alpha(x_{i,t-1} - \mu)^2 + \beta \widetilde{g}_{i,t-1}^2(\boldsymbol{\tau}),$$

with a given initial value $\widetilde{g}_{i1}^2(\boldsymbol{\tau}) = \widetilde{g}^2 \geq 0$. We then obtain an estimator $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\theta}}^{(1)\top}, \ldots, \widehat{\boldsymbol{\theta}}^{(p)\top}\right)^\top$ for the vector $\boldsymbol{\theta}_0 = \left(\boldsymbol{\theta}_0^{(1)\top}, \ldots, \boldsymbol{\theta}_0^{(p)\top}\right)^\top \in \Theta^p \subset \mathbb{R}^{d_1}$, with $d_1 = 4p$, of the GARCH(1,1) parameters for the p regressors. Under the regularity conditions in Appendix A, the QMLE is consistent and satisfies the Bahadur representation

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\boldsymbol{J}_*^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(\frac{\partial \ell_{1t}(\boldsymbol{\theta}_0^{(1)})}{\partial \boldsymbol{\tau}^\top}, \ldots, \frac{\partial \ell_{pt}(\boldsymbol{\theta}_0^{(p)})}{\partial \boldsymbol{\tau}^\top} \right)^\top + o_P(1), \tag{5}$$

where $\boldsymbol{J}_*$ is a block-diagonal matrix, and $\{\ell_{i,t}(\boldsymbol{\tau})\}$ is a stationary proxy of $\left\{\widetilde{\ell}_{i,t}(\boldsymbol{\tau})\right\}$ (see the supplementary file).

The GARCH(1,1) parameters $\boldsymbol{\vartheta}_0^{(0)} = (\omega_0, \alpha_0, \beta_0)^\top$ of the regression error term and the parameters $\boldsymbol{\vartheta}_0^{(i)} = (\varpi_{0i}, \xi_{0i}, c_{0i}, \gamma_{01,i}, \ldots, \gamma_{0q,i})^\top$ that are specific to $\beta_{i,t}$, for $i \in \{1, \ldots, p\}$, are

collected in $\boldsymbol{\vartheta}_0 = \left(\boldsymbol{\vartheta}_0^{(0)\top}, \boldsymbol{\vartheta}_0^{(1)\top}, \ldots, \boldsymbol{\vartheta}_0^{(p)\top}\right)^\top = (\vartheta_{01}, \ldots, \vartheta_{0d_2})^\top$, with $d_2 = 3 + p(3+q)$. It is not restrictive to set $c_{i0} = 0$, that is $\boldsymbol{\vartheta}_0^{(i)} = (\varpi_{0i}, 0, \ldots, 0)^\top$, when (4) holds. With that convention, the parameter $\boldsymbol{\vartheta}_0$ is uniquely defined. Denote by $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{d_2})^\top$ a generic element of the space $\Theta_\beta \subset (0,\infty)^2 \times [0,\infty) \times \mathbb{R}^{p(3+q)}$ and let the generic parameter $\boldsymbol{\varphi} = (\boldsymbol{\theta}^\top, \boldsymbol{\vartheta}^\top)^\top$ belong to the global parameter space $\Theta_G = \Theta^p \times \Theta_\beta$. In a second step, BFL estimate $\boldsymbol{\vartheta}_0$ by

$$\widehat{\boldsymbol{\vartheta}} = \left(\widehat{\boldsymbol{\vartheta}}^{(0)\top}, \widehat{\boldsymbol{\vartheta}}^{(1)\top}, \ldots, \widehat{\boldsymbol{\vartheta}}^{(p)\top}\right)^\top = \underset{\boldsymbol{\vartheta} \in \Theta_\beta}{\arg\min}\ \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}), \qquad \widetilde{O}_n(\boldsymbol{\varphi}) = \frac{1}{n}\sum_{t=2}^{n} \widetilde{\ell}_t(\boldsymbol{\varphi}),$$

where some transposes are omitted for simplicity of notation, and

$$\begin{aligned}
\widetilde{\ell}_t(\boldsymbol{\varphi}) &= \frac{\widetilde{v}_t^2(\boldsymbol{\varphi})}{\widetilde{g}_t^2(\boldsymbol{\varphi})} + \log \widetilde{g}_t^2(\boldsymbol{\varphi}), \qquad \widetilde{v}_t(\boldsymbol{\varphi}) = y_t - \sum_{i=1}^{p} \widetilde{\beta}_{i,t}(\boldsymbol{\varphi}) x_{i,t}, \\
\widetilde{g}_t^2(\boldsymbol{\varphi}) &= \omega + \alpha \widetilde{v}_{t-1}^2(\boldsymbol{\varphi}) + \beta \widetilde{g}_{t-1}^2(\boldsymbol{\varphi}), \\
\widetilde{\beta}_{i,t}(\boldsymbol{\varphi}) &= \varpi_i + \xi_i \frac{\widetilde{v}_{t-1}(\boldsymbol{\varphi}) x_{i,t-1}}{\mu_i^2 + \widetilde{g}_{i,t-1}^2(\boldsymbol{\varphi})} + c_i \widetilde{\beta}_{i,t-1}(\boldsymbol{\varphi}) + \sum_{j=1}^{q} \gamma_{j,i} z_{j,t-1},
\end{aligned}$$

with some starting values $\widetilde{\beta}_{i1}(\boldsymbol{\varphi}) = \beta^0 \in \mathbb{R}$ and $\widetilde{g}_1^2(\boldsymbol{\varphi}) = \widetilde{g}^2 \geq 0$. Note that, with some abuse of notation, we write $\widetilde{g}_{i,t}^2(\boldsymbol{\varphi}) = \widetilde{g}_{i,t}^2(\boldsymbol{\theta}) = \widetilde{g}_{i,t}^2(\boldsymbol{\theta}^{(i)})$.

By Proposition 2 in BFL we know that, under Assumptions **A1**-**A3** in Appendix A, there exists a unique strictly stationary ergodic and non anticipative solution to the system. They proved the consistency and asymptotic normality (CAN) of $\widehat{\boldsymbol{\varphi}}$ under the assumption $\xi_{0i} \neq 0$ for $i = 1, \ldots, p$, which rules out the case in (4).

The aim of the present paper is to study a penalized estimator that returns a constant estimated beta when (4) holds true. This is not the case for the multistep QMLE $\widehat{\boldsymbol{\varphi}} = \left(\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\vartheta}}^\top\right)^\top$ for which the trajectories of $\widetilde{\beta}_{i,t}(\widehat{\boldsymbol{\varphi}})$ turn out to be either erratic or converging exponentially towards a stationary solution, even when n is large and (4) holds.

4 Lasso-type estimators

In this section, we define and study Lasso-type estimators, which have the following advantages: *(i)* they are consistent in a more general framework than multi-stage QMLE; *(ii)* they lead to more parsimonious models; *(iii)* they are variable selection consistent.

4.1 Partially penalized estimator for constant beta detection

Following Tibshirani (1996), we define a Lasso estimator $\widehat{\boldsymbol{\varphi}}_n = \left(\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\vartheta}}_n^\top\right)^\top$ that encourages sparsity of specific components of $\widehat{\boldsymbol{\vartheta}}_n$ in order to allow constant $\beta_{i,t}(\widehat{\boldsymbol{\varphi}}_n)$'s. This estimator penalizes non zero estimated values of $d_3 = p(q+2)$ parameters (corresponding to $\xi_i, c_i, \gamma_{1,i}, \ldots, \gamma_{q,i}$, for $i = 1, \ldots, p$), but does not penalize the first-step estimator $\widehat{\boldsymbol{\theta}}$, the GARCH estimator $\widehat{\boldsymbol{\vartheta}}_n^{(0)}$ of the regression innovation, as well as the first components of $\widehat{\boldsymbol{\vartheta}}_n^{(i)\top}$ (corresponding to ϖ_i) for $i = 1, \ldots, p$. More precisely, let $\overline{S} = \{1, 2, 3, 4, q+7, 2(q+3)+4, \ldots, (p-1)(q+3)+4\}$ be the set of the $d_4 = 3+p$ indexes of the estimates of $\boldsymbol{\vartheta}$ that are not penalized by the estimator we are about to define. Let the complementary set $S = \{5, 6, \ldots, q+8, q+9, \ldots, (p-1)(q+3)+5, \ \ (p-1)(q+3)+6 = d_2\}$ contain the d_3 components that are penalized. Thus, consider the penalized QMLE

$$\widehat{\boldsymbol{\vartheta}}_n = \underset{\boldsymbol{\vartheta}\in\Theta_\beta}{\arg\min}\, \widetilde{Q}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}), \qquad \widetilde{Q}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) = \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) + \lambda_n p(\boldsymbol{\vartheta}), \tag{6}$$

where $\lambda_n \geq 0$ and $p(\boldsymbol{\vartheta}) = \sum_{i\in S} |\vartheta_i|$. The estimator depends on the value of the tuning parameter λ_n. Studying the asymptotic behavior of the penalized estimator is insufficient for making a practical choice of λ_n. This issue will be addressed in Section 5. The general strategy will be to start with a large value of λ_n (for example, equal to λ^s given in Proposition 1 below), which leads to the simplest model (where all the constrained coefficients are equal to zero). Then, the value of λ_n will be decreased to investigate the relevance of more complex models in a refined grid search. Note that the estimator also depends on

the scaling of the exogenous variables. Although there are no theoretical results on this topic, we recommend scaling each of the exogenous variables according to standard Lasso regression practice, matching their variances to the average variance of the regressors.

Recall that the multi-step QMLE is denoted by $\widehat{\boldsymbol{\vartheta}}$, whereas the penalized estimator is denoted by $\widehat{\boldsymbol{\vartheta}}_n = \left(\widehat{\boldsymbol{\vartheta}}_n^{(0)\intercal}, \widehat{\boldsymbol{\vartheta}}_n^{(1)\intercal}, \ldots, \widehat{\boldsymbol{\vartheta}}_n^{(p)\intercal}\right)^\intercal = \left(\widehat{\vartheta}_{n,1}, \ldots, \widehat{\vartheta}_{n,d_2}\right)^\intercal$. Note that $\widehat{\boldsymbol{\vartheta}}_n = \widehat{\boldsymbol{\vartheta}}$ when $\lambda_n = 0$ and recall that $\widehat{\boldsymbol{\vartheta}}$ is not consistent when some $\beta_{i,t}$ is constant, because, when (4) holds, the limit objective function is the same at several values. At the price of some bias, the Lasso estimator solves the lack of identifiability when $n \to \infty$, $\lambda_n \to \lambda_0 > 0$ and

$$Q_{\lambda_0}(\boldsymbol{\vartheta}) = E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) + \lambda_0 p(\boldsymbol{\vartheta}) \tag{7}$$

admits a minimum over Θ_β at some unique point $\boldsymbol{\vartheta}^\circledast$ (see Lemmas S1 and S2 in Appendix B). In (7), $\{\ell_t(\boldsymbol{\varphi})\}_{t\in\mathbb{Z}}$ denotes a stationary proxy of $\left\{\widetilde{\ell}_t(\boldsymbol{\varphi})\right\}$ (see the supplementary file).

Theorem 1 *Assume* **A1**–**A4** *and* $\mathbf{A5}(\lambda_0)$ *for some* $\lambda_0 > 0$ *in Appendix A hold. If* $\lambda_n \to \lambda_0$, *then* $\widehat{\boldsymbol{\vartheta}}_n$ *converges in probability to* $\boldsymbol{\vartheta}^\circledast = \arg\min_{\boldsymbol{\vartheta}\in\Theta_\beta} Q_{\lambda_0}(\boldsymbol{\vartheta})$.

A proof of Theorem 1 and all associated Lemmas are contained in Appendix B.

Remark 1 (FOC satisfied by $\boldsymbol{\vartheta}^\circledast$) *Let* $\partial p\left(\boldsymbol{\vartheta}^\circledast\right)$ *be the subdifferential (i.e. the set of the subgradients*[2]*) of* p *on* Θ_β *at* $\boldsymbol{\vartheta}^\circledast = (\vartheta_1^\circledast, \ldots, \vartheta_{d_2}^\circledast)$. *When the limit* $\boldsymbol{\vartheta}^\circledast$ *belongs to* $\mathring{\Theta}_\beta$, *the interior of* Θ_β, *it must satisfy the subgradient first-order condition (FOC)*

$$0 \in \partial^\circ Q_{\lambda_0}(\boldsymbol{\vartheta}^\circledast) := \left\{\frac{\partial E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}^\circledast)}{\partial \boldsymbol{\vartheta}}\right\} + \lambda_0 \partial p\left(\boldsymbol{\vartheta}^\circledast\right), \tag{8}$$

using the Minkowski sum notation. When $E\ell_1(\boldsymbol{\theta}_0, \cdot)$ *is convex,* $\partial^\circ Q_{\lambda_0}(\boldsymbol{\vartheta})$ *is the subdifferential of* $Q_{\lambda_0}(\boldsymbol{\vartheta})$ *and the FOC* (8) *is a necessary and sufficient condition that characterizes*

[2] $\boldsymbol{g}$ is a subgradient of p at $\boldsymbol{\vartheta}^\circledast$ if $p(\boldsymbol{\vartheta}) \geq p(\boldsymbol{\vartheta}^\circledast) + \boldsymbol{g}^\intercal\left(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\circledast\right)$ for all $\boldsymbol{\vartheta} \in \Theta_\beta$.

$\boldsymbol{\vartheta}^{\circledast}$. *More generally, Clarke (1975) showed that, since the functions* $E\ell_1(\boldsymbol{\theta}_0,\cdot)$ *and* $p(\cdot)$ *are locally Lipschitz,* $\partial^\circ Q_{\lambda_0}(\boldsymbol{\vartheta})$ *is the set of generalized gradients*[3] *of* $Q_{\lambda_0}(\boldsymbol{\vartheta})$*, which contains its subgradients. In general, the FOC* (8) *is thus only a necessary condition satisfied by* $\boldsymbol{\vartheta}^{\circledast}$*. When the limit* $\boldsymbol{\vartheta}^{\circledast}$ *belongs to the boundary of* Θ_β*,* (8) *still holds if* $\left\{\frac{\partial E\ell_1(\boldsymbol{\theta}_0,\boldsymbol{\vartheta}^{\circledast})}{\partial\boldsymbol{\vartheta}}\right\}$ *is replaced by the set of the generalized gradients of* $E\ell_1(\boldsymbol{\theta}_0,\cdot)$ *on* Θ_β *at* $\boldsymbol{\vartheta}^{\circledast}$*. The estimator satisfies an analogue FOC.*

Remark 2 (Sparsity of the limit) *The set* $\partial p(\boldsymbol{\vartheta})$ *consists of the vectors of the form* $\boldsymbol{u} = (u_1,\ldots,u_{d_2})^\intercal$ *where: for* $j \in \overline{S}$*,* $u_j = 0$ *for* $j \in S$*,* $u_j = sign(\vartheta_j)$ *when* $\vartheta_j \neq 0$*, where* $sign(\vartheta_j) = \vartheta_j/|\vartheta_j|$*, and* $u_j \in [-1,1]$ *when* $\vartheta_j = 0$*. It follows that if* $\boldsymbol{\vartheta}^{\circledast} \in \mathring{\Theta}_\beta$*, for all* $j \in S$*,*

$$\left|\frac{\partial E\ell_1(\boldsymbol{\theta}_0,\boldsymbol{\vartheta}^{\circledast})}{\partial\vartheta_j}\right| \leq \lambda_0 \;\; \textit{if} \;\; \vartheta_j^{\circledast} = 0, \qquad \frac{\partial E\ell_1(\boldsymbol{\theta}_0,\boldsymbol{\vartheta}^{\circledast})}{\partial\vartheta_j} = -\lambda_0 sign(\vartheta_j^{\circledast}) \;\; \textit{if} \;\; \vartheta_j^{\circledast} \neq 0.$$

Remark 3 (Sparsity of the penalized estimator) *Let the constrained QMLE be*

$$\widehat{\boldsymbol{\vartheta}}_n^c = \underset{\boldsymbol{\vartheta}\in\Theta_\beta^c}{\arg\min}\, \widetilde{O}_n(\widehat{\boldsymbol{\theta}},\boldsymbol{\vartheta}), \quad \widehat{\boldsymbol{\varphi}}_n^c = (\widehat{\boldsymbol{\theta}}^\intercal, \widehat{\boldsymbol{\vartheta}}_n^{c\intercal})^\intercal, \tag{9}$$

where Θ_β^c *denotes the set of the parameters* $\boldsymbol{\vartheta} \in \Theta_\beta$ *with* j*-th element* $\vartheta_j = 0$ *for all* $j \in S$*. It is clear that, under* **A3** *(viii), if the penalization* λ_n *is large enough, then the penalized estimator* $\widehat{\boldsymbol{\vartheta}}_n$ *defined in* (6) *is equal to the constrained estimator* $\widehat{\boldsymbol{\vartheta}}_n^c$*. More precisely, there exists* $\overline{\lambda} > 0$ *such that: (i) if* $\lambda_n > \overline{\lambda}$*,* $\widehat{\boldsymbol{\vartheta}}_n = \widehat{\boldsymbol{\vartheta}}_n^c$*, (ii) if* $\lambda_n < \overline{\lambda}$*,* $\widehat{\boldsymbol{\vartheta}}_n \neq \widehat{\boldsymbol{\vartheta}}_n^c$*.*

The FOCs of the optimization problem entail that any solution $\widehat{\boldsymbol{\vartheta}}_n \in \mathring{\Theta}_\beta$ *of* (6) *is such that, for all* $j \in S$*,*

$$\left|\frac{1}{n}\sum_{t=2}^n \frac{\partial\widetilde{\ell}_t(\widehat{\boldsymbol{\varphi}}_n)}{\partial\vartheta_j}\right| \leq \lambda_n \;\; \textit{if} \;\; \widehat{\vartheta}_{n,j} = 0, \quad \frac{1}{n}\sum_{t=2}^n \frac{\partial\widetilde{\ell}_t(\widehat{\boldsymbol{\varphi}}_n)}{\partial\vartheta_j} = -\lambda_n sign\left(\widehat{\vartheta}_{n,j}\right) \;\; \textit{if} \;\; \widehat{\vartheta}_{n,j} \neq 0.$$

[3] A generalized gradient of a locally Lipschitz function at a point $\boldsymbol{\vartheta}$ is the convex hull of the limits at $\boldsymbol{\vartheta}$ of all its differentials.

This entails that

$$\overline{\lambda} \geq \lambda^* := \max_{j \in S} \left| \frac{1}{n} \sum_{t=2}^{n} \frac{\partial}{\partial \vartheta_j} \widetilde{\ell}_t(\widehat{\boldsymbol{\varphi}}_n^c) \right|. \tag{10}$$

If the solution of (6) *is a continuous function of* λ_n*, then the previous inequality is actually an equality. An example in the Appendix shows that this continuity may fail if* $\widetilde{Q}_n(\cdot)$ *is not convex, and then we may have* $\overline{\lambda} > \lambda^*$*. We thus give another bound for* $\overline{\lambda}$*.*

Proposition 1 *Under* **A3** *(viii), let* $\Theta^{j-}_{\boldsymbol{\vartheta}_*}$ *(resp.* $\Theta^{j+}_{\boldsymbol{\vartheta}_*}$*) denote the set of the parameters* $\boldsymbol{\vartheta} \in \Theta_\beta$ *with* j*-th element* $\vartheta_j \leq 0$ *(resp.* $\vartheta_j \geq 0$*) and the other component of* $\boldsymbol{\vartheta}$ *are that of* $\boldsymbol{\vartheta}_*$*. The solution of* (6) *is* $\widehat{\boldsymbol{\vartheta}}_n^c$ *defined by* (9) *when* $\widehat{\boldsymbol{\vartheta}}_n^c \in \mathring{\Theta}_\beta$ *and*

$$\lambda_n > \lambda^s := \max_{j \in S} \max \left\{ \sup_{\boldsymbol{\vartheta} \in \Theta^{j-}_{\widehat{\boldsymbol{\vartheta}}_n^c}} \frac{1}{n} \sum_{t=2}^{n} \frac{\partial}{\partial \vartheta_j} \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}), - \inf_{\boldsymbol{\vartheta} \in \Theta^{j+}_{\widehat{\boldsymbol{\vartheta}}_n^c}} \frac{1}{n} \sum_{t=2}^{n} \frac{\partial}{\partial \vartheta_j} \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) \right\}. \tag{11}$$

If $\boldsymbol{\vartheta} \mapsto \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta})$ *is a strictly convex function and* $\widehat{\boldsymbol{\vartheta}}_n^c \in \mathring{\Theta}_\beta$*, then* $\overline{\lambda} = \lambda^* = \lambda^s$*.*

A proof of Proposition 1 can be found in Appendix B.

4.2 Adaptive penalized estimator for constant beta detection

In the spirit of the adaptive Lasso (Zou 2006), let us introduce data-driven weights

$$\widehat{\boldsymbol{\delta}} = \left(\widehat{\delta}_1, \ldots, \widehat{\delta}_{d_2}\right)^\top, \quad \widehat{\delta}_i = \frac{1}{\left|\widehat{\vartheta}_{ni}\right|} 1_{\widehat{\vartheta}_{ni} \neq 0} + \infty 1_{\widehat{\vartheta}_{ni} = 0} \text{ for } i \in S, \quad \widehat{\delta}_i = 0 \text{ for } i \in \overline{S}.$$

We then consider the adaptive penalized QMLE

$$\widehat{\boldsymbol{\vartheta}}_n^a = \arg\min_{\boldsymbol{\vartheta} \in \Theta_\beta} \widetilde{Q}_n^a(\boldsymbol{\vartheta}), \qquad \widetilde{Q}_n^a(\boldsymbol{\vartheta}) = \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) + \lambda_n^a p_{\widehat{\boldsymbol{\delta}}}(\boldsymbol{\vartheta}), \quad p_{\widehat{\boldsymbol{\delta}}}(\boldsymbol{\vartheta}) = \sum_{i=1}^{d_2} \widehat{\delta}_i |\vartheta_i|. \tag{12}$$

Let $\mathcal{A}$ be the subset of the *active* (and shrunk) components of the model (3), *i.e.* the set of indices $i \in S$ such that $\vartheta_{0i} \neq 0$. Let $\mathcal{I}$ be the subset of the *inactive* components of the

model, *i.e.* the set of indices $i \in S$ such that $\vartheta_{0i} = 0$. The three sets $\overline{S}$, $\mathcal{A}$ and $\mathcal{I}$ thus form a partition of $\{1, \ldots, d_2\}$. Let d_5 be the cardinality of $\mathcal{A}$ and $d_6 = d_4 + d_5$ be the number of active or unshrunk components. Let $\mathbb{I}_k$ be the identity matrix of size $k \times k$. We also introduce the selector matrix $\boldsymbol{A}$ which selects the active (or not shrunk) components of $\boldsymbol{\vartheta}_0$. More precisely, $\boldsymbol{A}$ is the $d_6 \times d_2$ matrix obtained by suppressing the columns of $\mathbb{I}_{d_2}$ which belong to $\mathcal{I}$. We thus have $\boldsymbol{\vartheta}_0 \in U$ where $U = \{\boldsymbol{u} \in \mathbb{R}^{d_2} : \boldsymbol{u} = \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{u}\}$. Finally, introduce the selector matrix $\boldsymbol{A}_\varphi$ which selects the active components of $\boldsymbol{\varphi}_0$:

$$\boldsymbol{A}_\varphi = \begin{pmatrix} \mathbb{I}_{d_1} & 0_{d_1 \times d_2} \\ 0_{d_6 \times d_1} & \boldsymbol{A} \end{pmatrix}.$$

Note that, under the assumptions of Theorem 1, $\widehat{\boldsymbol{\vartheta}}_n$ converges in probability to $\boldsymbol{\vartheta}^\circledast$, and thus $\widehat{\boldsymbol{\delta}} \to \boldsymbol{\delta} = (\delta_1, \ldots, \delta_{d_2})^\top$ in probability, where $\delta_i = 0$ for $i \in \overline{S}$, and under $\mathbf{A6}(\lambda_0)$ $\delta_i \in (0, +\infty)$ for $i \in \mathcal{A}$ and $\delta_i = +\infty$ for $i \in \mathcal{I}$. The sparsity of $\widehat{\boldsymbol{\vartheta}}_n$ also entails that

$$\widehat{\delta}_i = +\infty \text{ for all } i \in \mathcal{I}, \text{ with probability tending to 1.} \tag{13}$$

Let the vector $\boldsymbol{S}_t = \left(\frac{\partial \ell_{1t}(\boldsymbol{\theta}_0^{(1)})}{\partial \boldsymbol{\tau}^\top}, \ldots, \frac{\partial \ell_{pt}(\boldsymbol{\theta}_0^{(p)})}{\partial \boldsymbol{\tau}^\top}, \frac{\partial \ell_t(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\vartheta}^\top} \right)^\top$, the $(d_1 + d_2) \times (d_1 + d_2)$ matrices

$$\boldsymbol{I} = E\boldsymbol{S}_t\boldsymbol{S}_t^\top = \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\theta}} & \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\vartheta}} \\ \boldsymbol{I}_{\boldsymbol{\vartheta}\boldsymbol{\theta}} & \boldsymbol{I}_{\boldsymbol{\vartheta}} \end{pmatrix}, \quad \boldsymbol{J} = E\frac{\partial^2 \ell_t(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} = \begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\theta}} & \boldsymbol{J}_{\boldsymbol{\theta}\boldsymbol{\vartheta}} \\ \boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta}} & \boldsymbol{J}_{\boldsymbol{\vartheta}} \end{pmatrix}$$

and the $(d_1 + d_6) \times (d_1 + d_6)$ matrices

$$\boldsymbol{I}^{\boldsymbol{A}} = \boldsymbol{A}_\varphi \boldsymbol{I} \boldsymbol{A}_\varphi^\top = \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\theta}} & \boldsymbol{I}^{\boldsymbol{A}}_{\boldsymbol{\theta}\boldsymbol{\vartheta}} \\ \boldsymbol{I}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}\boldsymbol{\theta}} & \boldsymbol{I}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}} \end{pmatrix}, \quad \boldsymbol{J}^{\boldsymbol{A}} = \boldsymbol{A}_\varphi \boldsymbol{J} \boldsymbol{A}_\varphi^\top = \begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\theta}} & \boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\theta}\boldsymbol{\vartheta}} \\ \boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}\boldsymbol{\theta}} & \boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}} \end{pmatrix}.$$

The following result shows that unlike the Lasso-QMLE $\widehat{\boldsymbol{\vartheta}}_n$, which always has an asymptotic bias, the adaptive Lasso-QMLE can converge to the true parameter $\boldsymbol{\vartheta}_0$.

Theorem 2 *Let the assumptions of Theorem 1 and* $\mathbf{A6}(\lambda_0)$ *hold for some* $\lambda_0 > 0$. *If* $\lambda_n \to \lambda_0$, *if there exists* n_0 *such that* $\lambda_n^a > 0$ *for all* $n \geq n_0$, *and* $\sqrt{n}\lambda_n^a \to \lambda_0^a \geq 0$, *then* $\widehat{\boldsymbol{\vartheta}}_n^a \in U$ *(*i.e. *the components of* $\widehat{\boldsymbol{\vartheta}}_n^a$ *whose indices belong to* $\mathcal{I}$ *are zero) with probability tending to 1, and*

$$\sqrt{n}\boldsymbol{A}\left(\widehat{\boldsymbol{\vartheta}}_n^a - \boldsymbol{\vartheta}_0\right) \xrightarrow{d} \underset{\boldsymbol{u}\in\mathbb{R}^{d_6}}{\arg\min}\, V(\boldsymbol{u}),$$

where, for $\boldsymbol{u} = (u_1, \ldots, u_{d_6})^\top$, $V(\boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{W}_2 - \boldsymbol{u}^\top \boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta}}^{\boldsymbol{A}} \boldsymbol{J}_*^{-1} \boldsymbol{W}_1 + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{J}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}} \boldsymbol{u} + \lambda_0^a\, p(\boldsymbol{\delta}, \boldsymbol{\vartheta}_0, \boldsymbol{A}^\top \boldsymbol{u})$, *with* $(\boldsymbol{W}_1^\top, \boldsymbol{W}_2^\top)^\top \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}^{\boldsymbol{A}})$ *and, for* $\boldsymbol{u}_2 = (u_1, \ldots, u_{d_2})^\top$, $p(\boldsymbol{\delta}, \boldsymbol{\vartheta}_0, \boldsymbol{u}_2) = \sum_{i\in\mathcal{A}} \delta_i u_i sign(\vartheta_{0i})$.

The proof of Theorem 2 can be found in Appendix B.

Corollary 1 *Choosing in Theorem 2 a penalty term such that* $\lambda_0^a = 0$, *we obtain*

$$\sqrt{n}\boldsymbol{A}\left(\widehat{\boldsymbol{\vartheta}}_n^a - \boldsymbol{\vartheta}_0\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}}\right),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}} = \left(\boldsymbol{J}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}}\right)^{-1} \left\{\boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta}}^{\boldsymbol{A}} \boldsymbol{J}_*^{-1} \boldsymbol{I}_{\boldsymbol{\theta}} \boldsymbol{J}_*^{-1} \boldsymbol{J}_{\boldsymbol{\theta}\boldsymbol{\vartheta}}^{\boldsymbol{A}} + \boldsymbol{I}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}} - \boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta}}^{\boldsymbol{A}} \boldsymbol{J}_*^{-1} \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\vartheta}}^{\boldsymbol{A}} - \boldsymbol{I}_{\boldsymbol{\vartheta}\boldsymbol{\theta}}^{\boldsymbol{A}} \boldsymbol{J}_*^{-1} \boldsymbol{J}_{\boldsymbol{\theta}\boldsymbol{\vartheta}}^{\boldsymbol{A}}\right\} \left(\boldsymbol{J}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}}\right)^{-1}$.

The proof of Corollary 1 is contained in Appendix B.

Remark 4 (Oracle property) *The adaptive Lasso-QMLE* $\widehat{\boldsymbol{\vartheta}}_n^a$ *satisfies an oracle property in the sense of Fan & Li (2001) and Zou (2006), consisting of two parts. First, by Theorem 2, the inactive components (indices in* $\mathcal{I}$*) are correctly set to zero with probability tending to 1 — this is* selection consistency. *Second, when* $\lambda_0^a = 0$ *(Corollary 1), the active components are asymptotically normal with covariance* $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}}$, *which, by a direct extension of Theorem 3 in BFL, coincides with the asymptotic variance of the multistep QMLE computed as if the true zero components were* known a priori. *By contrast, the plain Lasso-QMLE* $\hat{\boldsymbol{\vartheta}}_n$ *does not enjoy this second property, as it retains an asymptotic bias.*

It should be noted, however, that the oracle property is a pointwise asymptotic result: it characterizes the limiting distribution at a fixed data-generating process but does not hold uniformly over the parameter space. As emphasized by Leeb & Pötscher (2008) and Hansen (2016), this means the oracle property can give a misleading picture of finite-sample behavior, particularly when true parameters lie close to zero. The Monte Carlo evidence in Section 6 therefore provides an important complement to this asymptotic result.

Remark 5 (Comparison with another penalized QMLE) *Recently, Nielsen & Rahbek (2024) studied the asymptotic behavior of a one-step penalized QMLE when some parameters may lie on the boundary of the parameter space. They allow for general penalization terms but do not allow for non-identified parameters. The authors showed the local consistency of the estimator and an oracle property in the case of a hard threshold or smoothly clipped absolute deviation (SCAD) penalty, but showed an asymptotic bias with L^1 penalization. This is not contradictory to our results because it is not the Lasso estimator, but rather the second-step adaptive estimator, which has the oracle property.*

Remark 6 (Post Lasso) *After a first-step Lasso estimation and variable selection, the so-called post-Lasso method simply consists of a second-step estimation of the components selected from the first-step Lasso (see* e.g. *Belloni & Chernozhukov 2013). Given Remark 4, the adaptive Lasso $\widehat{\boldsymbol{\vartheta}}_n^a$ has the same asymptotic distribution as the post-Lasso when $\lambda_0^a = 0$.*

As for the plain Lasso, there is an upper limit to the penalty term of the adaptive Lasso. That is, there is a $\overline{\lambda}^a > 0$ such that: i) if $\lambda_n^a > \overline{\lambda}^a$, $\widehat{\boldsymbol{\vartheta}}_n^a = \widehat{\boldsymbol{\vartheta}}_n^c$, and ii) if $\lambda_n^a < \overline{\lambda}^a$, $\widehat{\boldsymbol{\vartheta}}_n^a \neq \widehat{\boldsymbol{\vartheta}}_n^c$. An obvious adaptation of Proposition 1 provides bounds for $\overline{\lambda}^a$. In particular, similarly to (11), we have $\overline{\lambda}^a \geq \lambda_a^* := \max_{j\in S} \left|\widehat{\vartheta}_{nj}\right| \left|\frac{1}{n}\sum_{t=2}^n \frac{\partial}{\partial \vartheta_j}\widetilde{\ell}_t(\widehat{\boldsymbol{\varphi}}_n^c)\right|$, when $\widehat{\boldsymbol{\vartheta}}_n^c \in \mathring{\Theta}_\beta$.

4.3 Penalized estimator for detecting irrelevant betas

The previously defined estimators allow to detect constant betas, but not irrelevant betas, *i.e.* the indices $i \in \{1,\dots,p\}$ such that $\boldsymbol{\vartheta}_0^{(i)} = \mathbf{0}_{q+3}$. The naive solution would be to penalize all the $\boldsymbol{\vartheta}^{(i)}$'s coefficients, that is to set $S = \{4,5,\dots,d_2\}$ in (6) and (12). That would correspond to a penalization of the form $p(\boldsymbol{\vartheta}) = \sum_{i=1}^{p} \|\boldsymbol{\vartheta}^{(i)}\|_1$ where $\|\boldsymbol{\vartheta}^{(i)}\|_1 = |\varpi_i| + |\xi_i| + |c_i| + |\gamma_{1,i}| + \cdots + |\gamma_{q,i}|$. This naive approach does not work because a constant beta $\beta_{i,t+1} \equiv \bar{\beta}_i$ implies (4) and can be written as $\bar{\beta}_i = \varpi_i + \xi_i \frac{v_t x_{i,t}}{\mu_i^2 + g_{i,t}^2} + c_i \bar{\beta}_i + \gamma_{1,i} z_{1,t} + \cdots + \gamma_{q,i} z_{q,t}$, with many possibilities for $\boldsymbol{\vartheta}^{(i)}$, in particular $\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_1 := (\bar{\beta}_i, \mathbf{0}'_{q+2})'$ or $\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_2 := (\varrho\bar{\beta}_i, 0, 1-\varrho, \mathbf{0}'_q)'$, where $\varrho \in (0,1)$. Note that the solution $\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_2$ would be favored by the penalized estimator if $|\bar{\beta}_i| > 1$, because in this case $\|\boldsymbol{\vartheta}_1\|_1 > \|\boldsymbol{\vartheta}_2\|_1$. Therefore, penalizing ϖ_i along with the other parameters $\xi_i, c_i, \gamma_{1,i}, \dots, \gamma_{q,i}$ will result in an inconsistent estimator.

Having, in a first step, identified (some of) the ξ_{0i} and c_{0i} that are zero using the previous penalized estimators, one can focus on a constrained model of the form (3) with

$$\begin{aligned}
\beta_{i,t+1} &= \varpi_{0i} + \gamma_{01,i} z_{1,t} + \cdots + \gamma_{0q,i} z_{q,t}, && i = 1,\dots,p^1 \\
\beta_{i,t+1} &= \varpi_{0i} + \xi_{0i} \tfrac{v_t x_{i,t}}{\mu_{0i}^2 + g_{i,t}^2} + \gamma_{01,i} z_{1,t} + \cdots + \gamma_{0q,i} z_{q,t}, && i = p^1+1,\dots,p^1+p^2 \\
\beta_{i,t+1} &= \varpi_{0i} + \xi_{0i} \tfrac{v_t x_{i,t}}{\mu_{0i}^2 + g_{i,t}^2} + c_{0i}\beta_{i,t} + \gamma_{01,i} z_{1,t} + \cdots + \gamma_{0q,i} z_{q,t}, && i = p^1+p^2+1,\dots,p,
\end{aligned} \tag{14}$$

for $0 \le p^1 \le p^1 + p^2 \le p$, with obvious convention. This constrained model is such that the $\boldsymbol{\vartheta}_0$ coefficient has dimension $d_2 = 3 + p^1(q+1) + p^2(q+2) + p^3(q+3)$ with $p^3 = p - p^1 - p^2$. It is possible to shrink all the beta coefficients in (14), except the last ϖ_{0i} for $i = p^1 + p^2 + 1, \dots, p$. Let $S_0 = \{4,5,\dots,d_2\}$ be the set of the corresponding indices of $\boldsymbol{\vartheta}_0$. The assumptions have to be adapted. Specifically, in **A1** (iv), "$i = 1,\dots,p$" must be replaced by "$i = p^1 + p^2 + 1, \dots, p$", and since we do not need to estimate μ_{0i} and g_{i0} for $i = 1,\dots,p^1$, in **A4** (i), we can replace "$i = 1,\dots,p$" with "$i = p^1 + 1,\dots,p$" in **A4** (i). No additional assumptions are needed, since for $i = 1,\dots,p^1 + p^2$ all $\beta_{i,t}$ parameters are identifiable, even if $\beta_{i,t}$ is constant.

Theorem 3 *Under the previous modifications of* **A1** *and* **A4**, *the results of Theorem 1, Proposition 1, and Theorem 2 remain valid for the model defined by* (3) *and* (14) *when the shrinkage parameter set is* $S = S_0$.

5 Optimization algorithm

In order to extend the scope of this section, and to make it autonomous, we introduce notations slightly different to those of the other sections. Consider the optimization problem

$$\boldsymbol{\vartheta}(\lambda) = \arg\min_{\boldsymbol{\vartheta}\in\Theta} Q_\lambda(\boldsymbol{\vartheta}), \qquad Q_\lambda(\boldsymbol{\vartheta}) = Q(\boldsymbol{\vartheta}) + \lambda p(\boldsymbol{\vartheta}), \quad p(\boldsymbol{\vartheta}) = \sum_{i\in S} \delta_i |\vartheta_i|, \tag{15}$$

where $\lambda \geq 0$, $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d)^\top$, Θ is a convex compact subset of $\mathbb{R}^d$, $\delta_1, \ldots, \delta_d$ are given relative shrinkage coefficients with $\delta_i \geq 0$ ($\delta_i = 0$ when we do not want to shrink ϑ_i) and $S = \{i : \delta_i > 0\} \neq \varnothing$. We assume that $Q(\cdot)$ is two times continuously differentiable but not necessarily convex. Minimizing non convex and non differentiable objective functions like $Q_\lambda(\cdot)$ is challenging. In particular, the Newton-Raphson method may not work since the objective function is not differentiable everywhere.

We propose to generalize the "shooting algorithm" of Fu (1998), which is a coordinate-wise descent algorithm (see Friedman et al. 2007), called NLShoot (for non-linear shooting). Consider the minimization of $Q_\lambda(\boldsymbol{\vartheta})$ with respect to a single coordinate ϑ_i, the other coordinates of $\boldsymbol{\vartheta} \in \mathring{\Theta}$ being fixed. Let $Q_\lambda^{(i)}(\cdot; \boldsymbol{\vartheta}) : \mathbb{R} \to \mathbb{R}$ such that $Q_\lambda^{(i)}(\vartheta_i; \boldsymbol{\vartheta}) = Q_\lambda(\boldsymbol{\vartheta})$. Let $\Theta_{\boldsymbol{\vartheta}}^{(i)}$ be the section of Θ such that any vector of Θ with i-th component $\vartheta_i \in \Theta_{\boldsymbol{\vartheta}}^{(i)}$, the other components ϑ_j with $j \neq i$ being fixed to that of $\boldsymbol{\vartheta}$, belongs to Θ. The solution $\widetilde{\vartheta}_i = \arg\min_{\vartheta\in\Theta_{\boldsymbol{\vartheta}}^{(i)}} Q_\lambda^{(i)}(\vartheta; \boldsymbol{\vartheta})$ must satisfy the FOC

$$0 \in \left\{ \left. \frac{\partial Q(\boldsymbol{\vartheta})}{\partial \vartheta_i} \right|_{\vartheta_i = \widetilde{\vartheta}_i} \right\} + \delta_i \lambda \partial |\widetilde{\vartheta}_i|.$$

The set on the right side of the previous equation corresponds to the generalized gradients of $\partial Q_\lambda^{(i)}$ (see Clarke 1975) to which the searched subgradients belong. Let the sets

$$T_i^-(\boldsymbol{\vartheta}) = \left\{\vartheta_i < 0 : \frac{\partial Q(\boldsymbol{\vartheta})}{\partial \vartheta_i} = \delta_i \lambda\right\} \cap \Theta_{\boldsymbol{\vartheta}}^{(i)}, \qquad T_i^+(\boldsymbol{\vartheta}) = \left\{\vartheta_i > 0 : \frac{\partial Q(\boldsymbol{\vartheta})}{\partial \vartheta_i} = -\delta_i \lambda\right\} \cap \Theta_{\boldsymbol{\vartheta}}^{(i)}.$$

We thus have

$$\widetilde{\vartheta}_i = \underset{\vartheta \in T_i^-(\boldsymbol{\vartheta}) \cup T_i^+(\boldsymbol{\vartheta}) \cup \{0\}}{\arg\min} Q_\lambda^{(i)}(\vartheta; \boldsymbol{\vartheta}).$$

We propose the generalized shooting algorithm: start with an initial value $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^0$ and

$$\text{replace the } i\text{-th coordinate of } \boldsymbol{\vartheta} \text{ by } \widetilde{\vartheta}_i \text{ for } i = 1, 2, \ldots, d.$$

A discussion of the algorithm, a Proposition and Proof showing that cluster point(s) of the NLShoot algorithm are stationary points of $Q_\lambda(\cdot)$, are contained in Appendix E.1.

In the Monte Carlo simulations and the empirical application, we combine the advantages of the local quadratic approximation (LQA) algorithm proposed by Fan & Li (2001) and described in Appendix E.3 and the NLShoot algorithm to obtain parameter estimates efficiently, requiring the lowest calculation times. For each replication, we construct a grid of 25 equidistant values of λ between 0 and λ^*, as given in (10), or λ_a^*. For every value of $\lambda > 0$, the optimization of the penalized likelihood in (6) is carried out in multiple steps.

Step 1: For a given value of λ, we penalize ξ_i and c_i (for $i = 1, \ldots, p$) in (6). We use the LQA algorithm to obtain good starting values for the NLShoot algorithm and re-estimate the penalized QMLE by the NLShoot algorithm.

Based on the variable selection of the NLShoot algorithm for all elements of the grid of λs, the model is re-estimated by constrained QML by imposing the nullity of the non active parameters. The best model is chosen as the one minimizing the BIC criterion and is called Post-NLShoot in reference to the Post-Lasso method. We

perform a second round of estimations on another grid of 25 λs around the optimal λ obtained on the first grid.

Step 2: Step 2 is similar to Step 1 but the models are re-estimated using the adaptive penalized QMLE in (12), where the ϖ_i parameters of the betas for which c_i is set to 0 are also penalized (which corresponds to the first two equations of (14)). The vector of weights ($\hat{\boldsymbol{\delta}}$) is set using the estimates obtained in Step 1. Importantly, unlike Step 1, Step 2 allows to identify constant betas as well as irrelevant explanatory variables.

The two steps described above correspond to what we call the Penalized Autoregressive Conditional Beta (PACB) model estimated by Post-NLShoot.

We illustrate one penalized coefficient path for each of the two steps and the BIC of the PACB model for one simulation design ($p_{tv} = p_{cst} = p_{irr} = 2$) in Section F in the supplementary material. The figures show the penalized coefficients obtained using NLShoot for a broad range of penalty parameter values λ. In both cases, the same coefficients are detected to be active for a range of λ values. The optimal penalty parameters obtained by the grid search are the smallest λ values for which the BIC is minimized. In this simulation, the correct model is selected using two sets of grids, a coarse equidistant grid of 25 values between 0 and λ^* or λ_a^* and a finer one in the neighborhood of the smallest BIC. Given that for $p = 6$, 46,656 different specification of the model exist, this is remarkable given the relatively small grid size.

6 Monte Carlo simulations

In this section, we provide Monte Carlo simulations to investigate the finite sample properties of the penalized ACB estimator.

6.1 Data Generating Process

We simulate time-series of length $n = 4,000$ for 500 replications. The regression model is specified as in Equation (3) with $q = 0$. For the GARCH specification of the error term v_t, we set $\omega_0 = 0.05$, $\alpha_0 = 0.05$ and $\beta_0 = 0.9$ so that the unconditional variance of v_t is 1.

All regressors $\boldsymbol{x}_t$ are heteroskedastic and follow a DCC-GARCH(1,1) model (Engle 2002) with a zero mean and conditional covariance matrix $\boldsymbol{H}_t$, i.e., $\boldsymbol{x}_t = \boldsymbol{H}_t^{1/2}\boldsymbol{e}_t$, where $\boldsymbol{e}_t$ is *i.i.d.* and follows a p-dimensional standard Gaussian distribution. More specifically, the conditional covariance matrix $\boldsymbol{H}_t$ is specified as $\boldsymbol{H}_t = \mathbf{D}_t\mathbf{R}_t\mathbf{D}_t$, where $\boldsymbol{D}_t = \text{diag}\{g_{i,t}\}$, a diagonal matrix where the specification of $g_{i,t}^2$ is given in Equation (3) with parameters $\boldsymbol{\theta}_0^{(i)} = (\mu_{0i}, \omega_{0i}, \alpha_{0i}, \beta_{0i})^\top = (0, 0.05, 0.05, 0.9)^\top$ for all $i = 1, \ldots, p$. The conditional correlation matrix $\mathbf{R}_t$ is $\mathbf{R}_t = \text{diag}\left(q_{11,t}^{-1/2} \ldots q_{pp,t}^{-1/2}\right)\mathbf{Q}_t\,\text{diag}\left(q_{11,t}^{-1/2} \ldots q_{pp,t}^{-1/2}\right)$, where the $p \times p$ symmetric positive definite matrix $\mathbf{Q}_t = (q_{ij,t})$ is given by: $\mathbf{Q}_t = (1 - a - b)\bar{\mathbf{Q}} + a\boldsymbol{u}_{t-1}\boldsymbol{u}_{t-1}' + b\mathbf{Q}_{t-1}$, where $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{pt})^\top$ with $u_{it} = x_{it}/g_{it}$, for $i = 1, \ldots, p$, and a and b are set to 0.05 and 0.9, respectively. The $p \times p$ matrix $\bar{\boldsymbol{Q}}$ is the unconditional covariance matrix of $\boldsymbol{u}_t$. Its i,j-th entry is given by $\bar{\boldsymbol{Q}}_{i,j} = \rho^{|i-j|}$ for all $i, j = 1, \ldots, p$ which implies entries of 1 on the main diagonal. We set $\rho = 0.8$ so that the conditional correlations between the regressors vary over time, that two consecutive regressors are strongly correlated but the level of unconditional correlation between the variables decreases exponentially with the difference between the indices of the explanatory variables.

We consider various specifications with p_{tv} time-varying conditional betas, p_{cst} constant but non zero conditional betas and p_{irr} zero conditional betas. When the conditional betas are time-varying, they follow the specification in Equation (3). The parameters of the time-varying conditional betas are chosen as $\varpi_{0i} = 0.05$, $\xi_{0i} = 0.05$ and $c_{0i} = 0.95$. Hence, the p_{tv} time-varying conditional betas, meander around the value of 1. The p_{cst} constant (conditional) betas are set to one, i.e., $\varpi_{0i} = 1$, $\xi_{0i} = c_{0i} = 0$ while the p_{irr} zero betas are

obtained by setting $\varpi_{0i} = \xi_{0i} = c_{0i} = 0$.

We consider several specifications with $p = 6, 8$ and 10 regressors so that the total number of parameters in $\boldsymbol{\vartheta}$ are respectively $21, 27$ and 33. For each value of p, we consider a balanced case between time-varying and constant (non zero and zero) betas, i.e., $p_{tv} = (p_{cst} + p_{irr}) = p/2$ and two unbalanced cases with either $p_{tv} > p_{cst} + p_{irr}$ or $p_{tv} < p_{cst} + p_{irr}$.

6.2 Results

This section discusses the results of the Monte-Carlo simulation. The values of p_{tv}, p_{cst} and p_{irr} are reported in the first three columns of Table 1 while the percentage of correctly identified non active parameters (i.e., true zeros) and active parameters (i.e., true non zeros) are reported in columns % correct 0's and % correct !0's for Step 1 and Step 2.

For Step 1, the percentage of correctly identified non active (resp. active) parameters is computed as 100 times the average number of ξ_i and c_i parameters equal to 0 (resp. not equal to 0) in the DGP that are correctly set to 0 (resp. not equal 0) when minimizing the penalized QMLE in (6) using the NLShoot algorithm. In the two columns corresponding to Step 2, the percentage of correctly identified non active (resp. active) parameters is computed as the 100 times the average number of ϖ_i, ξ_i and c_i parameters equal to 0 (resp. not equal to 0) in the DGP that are correctly set to 0 (resp. not equal 0) when minimizing (12) using the NLShoot algorithm.

Several comments are in order. The percentage of zero coefficients correctly set to zero is larger than 96% in Step 2 demonstrating that the method produces excellent results. The percentage of non zero coefficients correctly identified as active parameters is also very high in Step 2, ranging between 90 and 94% but is much smaller in Step 1 (i.e., between about 83 and 92 %). Overall, the results are significantly better in Step 2, which shows that penalizing the intercepts (when possible) and estimating the model via the penalized adaptive QMLE in (12) improves the quality of the estimation.

Table 1: Monte-Carlo simulation results based on 500 replications for sample of 4,000 observations.

p_{tv}	p_{cst}	p_{irr}	Step 1		Step 2	
			% correct 0's	% correct !0's	% correct 0's	% correct !0's
3	1	2	89.0	90.9	97.8	92.4
4	0	2	86.6	92.3	98.7	93.6
2	2	2	91.7	90.6	98.6	94.0
4	2	2	88.8	87.8	97.0	90.4
6	0	2	82.6	92.3	99.2	93.0
2	4	2	94.7	83.4	98.3	92.1
5	3	2	85.5	85.7	96.2	90.0
7	1	2	78.9	91.2	97.4	91.3
3	5	2	92.5	83.5	98.1	91.7

Note: The number of time-varying betas (i.e., p_{tv}), constant betas (i.e., p_{cst}) and zero betas (i.e., p_{irr}) are reported in the first three columns. All explanatory variables follow a DCC-GARCH model. Columns '% correct 0's' and '% correct !0's' correspond respectively to the number of penalized parameters correctly identified as non active parameters (i.e., true zeros) and active parameters (i.e., true non zeros). In the panel 'Step 1', only the ξ and c parameters are penalized while in the panel 'Step 2', the intercepts ϖ's of the conditional betas for which the corresponding c parameter is set to 0 in Step 1 are also penalized.

In Appendix G, we provide more simulation results including a discussion on how well the separation between the different types of conditional betas performs. Across all simulation designs, the percentage of correctly detected betas is very high (above 95%). The separation between constant and zero betas in step 2 works very well as above 98% of $\beta_i \neq 0$ are correctly detected and around 90% of the case $\beta_i = 0$. To assess how well the full ACB model and our PACB model are able to capture the underlying dynamics in the conditional betas, we report the ratio of the RMSE on the conditional betas of full ACB relative to PACB for all simulation designs. All ratios are larger than one indicating that the PACB leads to smaller average RMSE on the conditional betas, even when we separate time-varying and constant betas. In the case of time-varying betas, full ACB correctly estimates all conditional betas as time-varying but because all underlying models are misspecified and due to the failure to detect the constancy and nullity of some conditional betas, the results for the time-varying betas are also distorted.

Additionally, we present in Appendix G results on the bias of the estimated parameters of our method in comparison to the (misspecified) full ACB model. In cases in which

PACB correctly selects a time-varying beta, the average bias is of similar magnitude for both methods. When a true beta is constant, the average bias of full ACB exceeds that of PACB. In addition to average biases, we additionally plot the distribution of the bias for two simulation designs. PACB outperforms full ACB in all settings. In conclusion, these simulation results show that the two-step estimation procedure performs remarkably well on a sample size that is common in financial applications.

7 Global Minimum Variance Portfolio application

This section illustrates the practical relevance of the PACB methodology in a portfolio allocation context. We consider the construction of a Global Minimum Variance Portfolio (GMVP) for a universe of k assets, where portfolio weights $\pi_{t+1} = (\pi_{t+1,1}, \ldots, \pi_{t+1,k})$ are obtained from the forecasts of the conditional precision matrix of asset returns or using one-step-ahead forecasts of the betas of a linear model, with seven competing models.

Let $r_t = (r_{1,t}, \ldots, r_{k,t})^\top$ denote the vector of returns for the k assets. In this application, the portfolio universe consists of the following 50 US stocks.[4] We use daily returns from 2005-02-28 to 2024-12-31 (4,991 observations), reserving the last 1,000 observations for out-of-sample testing.

7.1 GMVP weights via the precision matrix

Let $\Sigma^{-1}_{t+1|t}$ denote the one-step-ahead forecast of the conditional precision matrix of r_t. The one-step-ahead of the optimal GMVP weights are equal to $\pi_{t+1} = \frac{\Sigma^{-1}_{t+1|t}\iota}{\iota^\top \Sigma^{-1}_{t+1|t}\iota}$, where ι is an $k \times 1$ vector of ones. This ensures that portfolio weights sum to one while minimizing the conditional portfolio variance. The key empirical challenge with this method is the

[4]These are ADI, AMAT, ABT, AMGN, AXP, BA, BAC, BMY, CAT, C, CL, CMCSA, COST, CSCO, DE, DIS, GE, GS, HD, HON, IBM, INTC, JNJ, JPM, KMB, KO, LOW, MCD, MMM, MRK, MS, MSFT, NKE, ORCL, PEP, PFE, PG, QCOM, SBUX, T, TGT, TXN, UNH, UNP, USB, VZ, WFC, WMT, XOM, AAPL.

estimation of $\Sigma_{t+1|t}^{-1}$. We consider three competing approaches.

COV100. The first is the inverse of a rolling-window estimator based on the sample covariance matrix computed over the previous 100 trading days.

DCC and DECO. The other two methods are the inverse of the one-step ahead forecast of the covariance matrix obtained with either a DCC-GARCH(1,1) or a DECO-GARCH(1,1), as proposed by Engle (2002) and Engle & Kelly (2012), respectively. Correlation targeting is used to reduce the number of parameters to be estimated by QML.

These models are estimated on the first 3,991 observations to obtain the initial forecast of $\Sigma_{t+1|t}$. The parameters are then kept constant to obtain one-step forecasts for the next 99 observations. The models are then re-estimated every 100 days over an expanding window.

7.2 GMVP weights via linear regression models

Kempf & Memmel (2006) exploits the equivalence between the GMVP and a regression problem with constant betas, where portfolio weights can be obtained from the coefficients of a linear regression of one asset return on return differences with respect to a reference asset. Reh et al. (2023) extend this approach by assuming that all betas vary over time.

Select asset k as a reference asset and define the dependent variable and regressors as $y_t = r_{k,t}$, and $X_t = \left(1,\ r_{k,t} - r_{1,t},\ \ldots,\ r_{k,t} - r_{k-1,t}\right)^\top$. Consider the linear projection

$$y_t = X_t^\top \beta_t + \varepsilon_t, \tag{16}$$

where $\beta_t = (\beta_{0,t}, \beta_{1,t}, \ldots, \beta_{k-1,t})^\top$ denotes the vector of conditional betas. Then, the corresponding GMVP weights satisfy $\pi_{t+1,i} = \beta_{t+1|t,i}$, for $i = 1, \ldots, k-1$, and $\pi_{t+1,k} = 1 - \sum_{j=1}^{k-1} \beta_{t+1|t,j}$. We implement two static and two dynamic versions of this approach.

Lasso and Auto. For the static case, where $\beta_t = \beta \ \forall t$, model (16) is estimated on the first 3,991 observations using either Post-Lasso (with cross-validation and 5 folds) or Autometrics (Doornik et al. 2009) with a target size of 1%. The parameters are kept constant to produce the GMVP weights for the next 100 days. The models are re-estimated every 100 days over an expanding window, keeping the same set of active variables as in the first sample.

ACB and PACB. In the dynamic case, model (16) is estimated using an ACB specification (where all betas are assumed to be time-varying) or using the PACB approach (where some betas can be set to a constant or even to 0). For the PACB, the selection of the active parameters is done on the first 3,991 observations. The parameters are estimated by QML and kept constant to produce one-step-ahead forecasts of the conditional betas (i.e., $\beta_{t+1|t,i}$ for $i = 0, \ldots, k-1$) and the corresponding GMVP weights for the next 100 days. The models are then re-estimated every 100 days over an expanding window, keeping the same set of active variables as for the first sample for the PACB.

Weight dynamics and turnover. Portfolio turnover of method m over n time-periods is computed as $\text{Turnover}_t^m = \frac{1}{2}\sum_{i=1}^{k}\left|\pi_{i,t}^m - \pi_{i,t-1}^m\right|$, where $\pi_{i,t}^m$ denotes the portfolio weight of asset i at time t for method m. To smooth short-term fluctuations, we plot in Figure 2 a 20-day moving average of the turnover series for the 7 competing methods.

The static Kempf–Memmel approach (denoted Lasso and Auto) yields the lowest turnover since weights remain constant within each window of 100 observations. Dynamic covariance models such as DCC and DECO generate highly volatile weights, while rolling estimators also produce substantial fluctuations due to sampling variability. By contrast, ACB and PACB produce considerably smoother weight trajectories than the three methods based on a prediction of the precision matrix. The penalization mechanism of the PACB filters out noisy dynamics and shrinks irrelevant components, resulting in portfolios with low turnover

while preserving time variation when supported by the data.

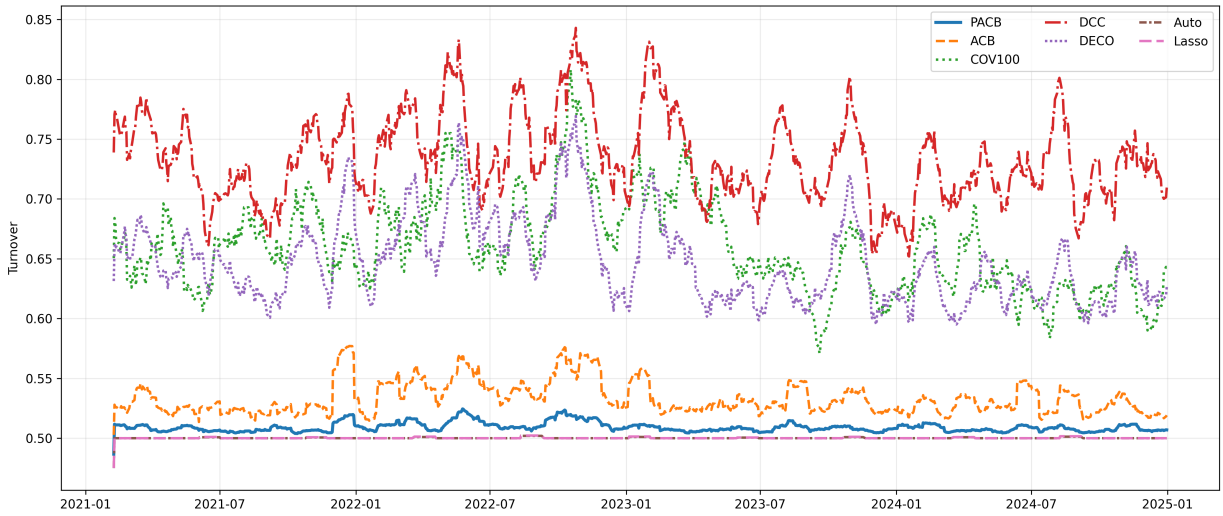

Figure 2: 20-day moving average of the turnover series.

Out-of-sample evaluation. Portfolio performance is evaluated using the empirical variance of the realized returns of the portfolio, computed over the 1,000 out-of-sample forecasts as $L_m = \frac{1}{1000}\sum_{t=1}^{1000}\left(r_{p,t+1}^m - \bar{r}_p^m\right)^2$ where $r_{p,t+1}^m = \pi_{t+1}^m r_{t+1}$ denotes the realized portfolio return of the m-th method (for $m = 1, \ldots, 7$) and $\bar{r}_p^m$ the corresponding sample mean. This criterion directly measures realized portfolio variance, which is the objective of the GMVP.

Model comparison is conducted using the Model Confidence Set (MCS) procedure of Hansen et al. (2011) and in particular the Mulcom package for Ox (Hansen et al. 2021) with 10,000 bootstrap samples and a block length of 5 observations.

Results. The results of this forecasting exercise are reported in Table 2. The number of assets receiving a non-zero portfolio weight for each strategy is reported in the second column (# Assets). This quantity provides a direct measure of portfolio sparsity and reflects the ability of a method to eliminate irrelevant assets. The third column (# TV weights) reports the number of time-varying weights. The fourth column reports the average Loss ($\times 10^3$), while the last column is the p-value of the MCS test.

Table 2: Results of the GMVP application

Method	# Assets	# TV weights	$L_m \times 10^3$	p-value
PACB	15	2	0.53868	1.0000
Lasso	32	0	0.61429	0.0000
Auto	29	0	0.62263	0.0000
ACB	50	50	0.64079	0.0086
DCC	50	50	0.65128	0.0000
COV100	50	50	0.91946	0.0000
DECO	50	50	0.93566	0.0000

Note: The number of assets receiving a non-zero portfolio weight for each strategy is reported in the second column (# Assets). The third column (# TV weights) reports the number of time-varying weights. The fourth column reports the empirical variance of the portfolios while the last column is the p-value of the MCS test.

Several important conclusions emerge from the results. First, PACB selects the smallest number of assets (15 out of 50), indicating a strong sparsity effect induced by the penalization mechanism. Furthermore, only 2, out 15 non zero weights, are time-varying. This is illustrated in the top panel of Figure 3 that displays the trajectory of $\pi_{t+1|t}$ for the non-zero weights obtained with the PACB method. The bottom panel of Figure 3 displays the trajectory of $\pi^m_{t+1|t}$ for the DCC method. It is clear that this model generates highly unstable weights for the 50 components of this portfolio and, consequently, high transaction costs.

Importantly, PACB not only generates a low-turnover portfolio, but also achieves the smallest realized variance. Indeed, PACB is the only method belonging to the MCS at conventional nominal sizes (i.e., 1, 5 or 10%). These results suggest that shrinking noisy or weakly informative exposures enhances both the stability and the risk efficiency of the GMVP. In contrast, methods generating highly unstable weights, such as DCC, COV100 and DECO, fail to translate their flexibility into improved out-of-sample risk performance as they do not belong to the MCS (as their p-values are close to 0%). Finally, the static selection approaches based on Lasso and Autometrics provide some degree of sparsity but remain dominated by PACB as their variance is statistically higher.

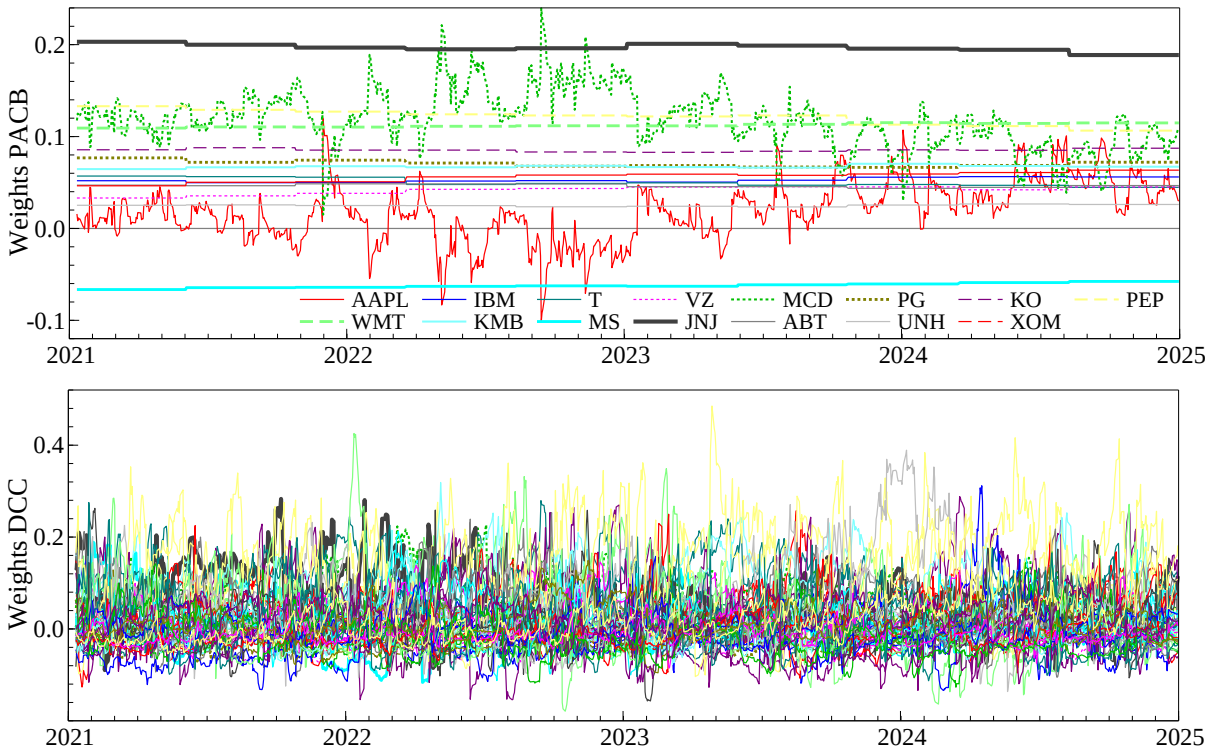

Figure 3: Time-varying GMVP weights obtained using PACB (top panel) and DCC (bottom panel)

Overall, the results highlight that combining dynamic modeling with penalization leads to parsimonious portfolios that are not only easier to implement but also exhibit superior variance performance.

8 Conclusion

In this paper, we explore the application of penalized QMLE in the context of time series regressions, focusing on the identification of time-varying, constant, and zero conditional beta coefficients in a linear regression. We introduce a Lasso-type estimator to address the non-identifiability of parameters when conditional betas are constant, which simplifies the model by shrinking the parameters driving the dynamics of the conditional betas to 0 when the betas are constant. Our method not only helps to identify constant conditional betas

but also irrelevant variables.

Our multi-step estimator strategy breaks down the large-dimensional optimization problem into several lower-dimensional ones, making the estimation process more manageable. To optimize the non-convex objective function, we introduce a nonlinear coordinate-wise descent algorithm (NLShoot) and demonstrate its effectiveness in finding stationary points of the objective function. The asymptotic properties of the estimators are analyzed, and bounds on the penalty term are derived.

Monte Carlo simulations illustrate the finite-sample properties of the proposed method for identifying time-varying and constant betas, as well as for selecting relevant regressors.

We apply the PACB model in the context of a global minimum variance portfolio on 50 series of daily returns of US stocks. Overall, the results indicate that combining dynamic modeling with penalization leads to sparse and stable portfolio allocations that improve out-of-sample risk performance.

Non-identified parameters also appear in many models other than the ACB model. Well-known examples in time series analysis include multivariate GARCH models and other models with a score-driven update. Although it is possible to use penalized QMLE for general non-identified time series models, precise results require case-by-case studies.

References

Adamek, R., Smeekes, S. & Wilms, I. (2023), 'Lasso inference for high-dimensional time series', *Journal of Econometrics* **235**(2), 1114–1143.

Alami Chentoufi, R. (2024), Penalized Convex Estimation in Dynamic Location-Scale models, MPRA Paper 123283, University Library of Munich, Germany.

Barndorff-Nielsen, O. E. & Shephard, N. (2004), 'Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics', *Econometrica* **72**(3), 885–925.

Basu, S. & Michailidis, G. (2015), 'Regularized estimation in sparse high-dimensional time series models', *The Annals of Statistics* **43**(4), 1535–1567.

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

Blasques, F., Francq, C. & Laurent, S. (2024), 'Autoregressive conditional betas', *Journal of Econometrics* **238**(2), 105630.

Bollerslev, T., Engle, R. F. & Wooldridge, J. M. (1988), 'A capital asset pricing model with time-varying covariances', *Journal of Political Economy* **96**(1), 116–131.

Clarke, F. H. (1975), 'Generalized gradients and applications', *Transactions of the American Mathematical Society* **205**, 247–262.

Darolles, S., Francq, C. & Laurent, S. (2018), 'Asymptotics of cholesky GARCH models and time-varying conditional betas', *Journal of Econometrics* **204**(2), 223–247.

Doornik, J. A. et al. (2009), 'Autometrics', *Castle, and Shephard (2009)* pp. 88–121.

Durbin, J. & Koopman, S. J. (2001), *Time Series Analysis by State Space Methods (Oxford Statistical Science Series)*, Oxford University Press.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *The Annals of Statistics* **32**(2), 407–499.

Engle, R. (2002), 'Dynamic conditional correlation - a simple class of multivariate GARCH models', *Journal of Business & Economic Statistics* **20**, 339–350.

Engle, R. F. (2016), 'Dynamic conditional beta', *Journal of Financial Econometrics* **14**(4), 643–667.

Engle, R. & Kelly, B. (2012), 'Dynamic equicorrelation', *Journal of Business & Economic Statistics* **30**(2), 212–228.

Fama, E. F. & French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* **33**(1), 3–56.

Fama, E. F. & French, K. R. (2015), 'A five-factor asset pricing model', *Journal of Financial Economics* **116**(1), 1–22.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J. & Peng, H. (2004), 'Nonconcave penalized likelihood with a diverging number of parameters', *The Annals of Statistics* **32**(3), 928–961.

Friedman, J., Hastie, T., Höfling, H. & Tibshirani, R. (2007), 'Pathwise coordinate optimization', *The Annals of Applied Statistics* **1**(2), 302–332.

Fu, W. J. (1998), 'Penalized regressions: the bridge versus the lasso', *Journal of Computational and Graphical Statistics* **7**(3), 397–416.

Gagliardini, P., Ossola, E. & Scaillet, O. (2016), 'Time-varying risk premium in large cross-sectional equity data sets', *Econometrica* **84**(3), 985–1046.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.

Hansen, B. E. (2016), 'The risk of james–stein and Lasso shrinkage', *Econometric Reviews* **35**(8-10), 1456–1470.

Hansen, P., Lunde, A. & Laurent, S. (2021), *Econometric Toolkit for Multiple Comparisons: Mulcum 3.0*, Timberlake Consultants Press.

Hansen, P., Lunde, A. & Nason, J. (2011), 'The model confidence set', *Econometrica* **79**, 453–497.

Hendry, D. F. & Johansen, S. (2011), The properties of model selection when retaining theory variables, Technical Report 11-25, University of Copenhagen Discussion Paper.

Kempf, A. & Memmel, C. (2006), 'Estimating the global minimum variance portfolio', *Schmalenbach Business Review* **58**(4), 332–348.

Knight, K. & Fu, W. (2000), 'Asymptotics for Lasso-type estimators', *The Annals of Statistics* **28**(5), 1356–1378.

Kock, A. B. (2016), 'Consistent and conservative model selection with the adaptive Lasso in stationary and nonstationary autoregressions', *Econometric Theory* **32**(1), 243–259.

Leeb, H. & Pötscher, B. M. (2008), 'Sparse estimators and the oracle property, or the return of hodges' estimator', *Journal of Econometrics* **142**(1), 201–211.

Loh, P.-L. (2017), 'Statistical consistency and asymptotic normality for high-dimensional robust M-estimators', *Annals of Statistics* **45**(2), 866–896.

Nardi, Y. & Rinaldo, A. (2011), 'Autoregressive process modeling via the lasso procedure', *Journal of Multivariate Analysis* **102**(3), 528–549.

Nielsen, H. B. & Rahbek, A. (2024), 'Penalized quasi-likelihood estimation and model selection with parameters on the boundary of the parameter space', *The Econometrics Journal* **27**(1), 107–125.

Poignard, B. & Fermanian, J.-D. (2021), 'High-dimensional penalized ARCH processes', *Econometric Reviews* **40**(1), 86–107.

Reh, L., Krüger, F. & Liesenfeld, R. (2023), 'Predicting the global minimum variance portfolio', *Journal of Business & Economic Statistics* **41**(2), 440–452.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288.

Wang, H., Li, G. & Tsai, C.-L. (2007), 'Regression coefficient and autoregressive order shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(1), 63–78.

Wang, Z., Liu, H. & Zhang, T. (2014), 'Optimal computational and statistical rates of convergence for sparse nonconvex learning problems', *The Annals of Statistics* **42**(6), 2164–2201.

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