SUPPLEMENTARY MATERIAL to 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

This document provides some supplementary material to the paper "Penalized Autoregressive Conditional Betas". More specifically, we provide the maintained assumptions, proofs of the theorems and additional lemmas, as well as illustrations of some of the assumptions, a description of our proposed NLShoot algorithm for the PACB model, additional Monte Carlo results, and a detailed treatment of the multi-factor model introduced as methodological validation.

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 and julie.schnaitmann@uni-tuebingen.de

A Technical assumptions

In this section, we provide and discuss the assumptions imposed for the Penalized Autoregressive Conditional Beta (PACB) model.

A1: (i) $E\log\left(\alpha_{0i}\eta_{i,t}^2+\beta_{0i}\right)<0$; (ii) $E\log(\alpha_0\eta_1^2+\beta_0)<0$; (iii) $(\eta_t,\boldsymbol{x}_t^\top,\boldsymbol{z}_t^\top)$ is stationary and ergodic and $E\|\boldsymbol{z}_t\|^s<\infty$ for some $s>0$; (iv) $|c_{0i}|<1$ for $i=1,\ldots,p$.

Assumptions (i) and (ii) are necessary and sufficient for strict stationarity of the GARCH equations, (iii) is required because the dynamics of the covariates are not given by the model ($\boldsymbol{z}_t$ is considered exogenous), and the constraints (iv) on the persistence parameters of the betas are required to avoid explosiveness. It has been shown that **A1** is sufficient for the existence of a stationary and ergodic solution to the entire system (3).

The sigma field generated by the past observations $\{y_u,\boldsymbol{x}_u^\top,\boldsymbol{z}_u^\top,u\le t\}$ is denoted by $\mathcal{F}_t^*$. For $\boldsymbol{\varphi}\in\Theta_G$, we now study the existence of stationary and $\mathcal{F}_t^*$-measurable solutions

$$\boldsymbol{Y}_t(\boldsymbol{\varphi}):=\left(g_{1,t+1}^2(\boldsymbol{\varphi}),\ldots,g_{p,t+1}^2(\boldsymbol{\varphi}),g_{t+1}^2(\boldsymbol{\varphi}),\beta_{1,t+1}(\boldsymbol{\varphi}),\ldots,\beta_{p,t+1}(\boldsymbol{\varphi}),v_t(\boldsymbol{\varphi})\right)$$

such that, for all t,

$$\begin{aligned}
g_{i,t+1}^2(\boldsymbol{\varphi}) &= g_{i,t+1}^2(\boldsymbol{\tau})=\omega_i+\alpha_i(x_{i,t}-\mu_i)^2+\beta_i g_{i,t}^2(\boldsymbol{\varphi}),\\
v_t(\boldsymbol{\varphi}) &= y_t-\beta_{1,t}(\boldsymbol{\varphi})x_{1,t}-\cdots-\beta_{p,t}(\boldsymbol{\varphi})x_{p,t},\\
\beta_{i,t+1}(\boldsymbol{\varphi}) &= \varpi_i+\xi_i\tfrac{v_t(\boldsymbol{\varphi})x_{i,t}}{\mu_i^2+g_{i,t}^2(\boldsymbol{\varphi})}+c_i\beta_{i,t}(\boldsymbol{\varphi})+\gamma_{1,i}z_{1,t}+\cdots+\gamma_{q,i}z_{q,t},\\
g_{t+1}^2(\boldsymbol{\varphi}) &= \omega+\alpha v_t^2(\boldsymbol{\varphi})+\beta g_t^2(\boldsymbol{\varphi}).
\end{aligned}\tag{S1}$$

A2: (i) for all $\boldsymbol{\varphi}\in\Theta_G$, $\omega>0$, $0\le\beta<1$ and $\omega_i>0$, $0\le\beta_i<1$ for $i=1,\ldots,p$; (ii) there exists an operator norm $\|\cdot\|$ and an integer $k\ge1$ such that

$$E\log\sup_{\boldsymbol{\varphi}\in\Theta_G}\left\|\prod_{i=1}^{k}\Lambda_{t-i}(\boldsymbol{\varphi})\right\|<0,$$

where $\Lambda_t(\boldsymbol{\varphi})$ is the $p \times p$ matrix whose entries are $-\frac{\xi_i x_{i,t} x_{j,t}}{\mu_i^2 + g_{i,t}^2(\boldsymbol{\varphi})} + c_i 1_{\{i=j\}}$.

The invertibility condition (i) of **A2** ensures the existence of stationary processes $\left\{g_{i,t}^2(\boldsymbol{\varphi})\right\}_{t\in\mathbb{Z}}$ and $\left\{\ell_{i,t}^2(\boldsymbol{\varphi})\right\}_{t\in\mathbb{Z}}$, where $\ell_{i,t}(\boldsymbol{\varphi}) = \ell_{i,t}(\boldsymbol{\tau}) = \frac{(x_{i,t}-\mu)^2}{g_{i,t}^2(\boldsymbol{\tau})} + \log g_{i,t}^2(\boldsymbol{\tau})$, satisfying (S1), and shows that, uniformly in $\boldsymbol{\varphi}$, $g_{i,t}^2(\boldsymbol{\varphi}) - \widetilde{g}_{i,t}^2(\boldsymbol{\varphi}) \to 0$ and $\ell_{i,t}(\boldsymbol{\varphi}) - \widetilde{\ell}_{i,t}(\boldsymbol{\varphi}) \to 0$ at an exponential rate as $t \to \infty$. The second invertibility condition (ii) is less explicit, but BFL have discussed how to empirically check invertibility, that is the existence of $\varrho \in (0,1)$ such that, as $t \to \infty$, a.s.

$$\varrho^{-t} \sup_{\boldsymbol{\varphi}\in\Theta_G} \left\|\widetilde{\boldsymbol{\beta}}_t(\boldsymbol{\varphi}) - \boldsymbol{\beta}_t(\boldsymbol{\varphi})\right\| \to 0, \tag{S2}$$

and

$$\varrho^{-t} \sup_{\boldsymbol{\varphi}\in\Theta_G} \left\{\left|\widetilde{g}_t^2(\boldsymbol{\varphi}) - g_t^2(\boldsymbol{\varphi})\right| + \sum_{i=1}^{p} \left|\widetilde{g}_{i,t}^2(\boldsymbol{\varphi}) - g_{i,t}^2(\boldsymbol{\varphi})\right| + \left|\widetilde{v}_t(\boldsymbol{\varphi}) - v_t(\boldsymbol{\varphi})\right|\right\} \to 0. \tag{S3}$$

It follows that, for all $\boldsymbol{\varphi} \in \Theta_G$, the process $\{\ell_t(\boldsymbol{\varphi})\}_{t\in\mathbb{Z}}$ defined by

$$\ell_t(\boldsymbol{\varphi}) = \frac{v_t^2(\boldsymbol{\varphi})}{g_t^2(\boldsymbol{\varphi})} + \log g_t^2(\boldsymbol{\varphi})$$

is stationary, ergodic, $\mathcal{F}_t^*$-measurable, and that, uniformly in $\boldsymbol{\varphi}$, $\ell_t(\boldsymbol{\varphi}) - \widetilde{\ell}_t^2(\boldsymbol{\varphi}) \to 0$ at an exponential rate as $t \to \infty$. Denote by $\mathring{A}$ the interior of a set A. Assume the following identifiability condition.

A3: (i) for $i \in \{1,\ldots,p\}$, the support of the distribution of η_{it} conditional on $\{\boldsymbol{\eta}_u^\top, \boldsymbol{z}_u^\top, u < t\}$, contains at least 3 points, and Θ is a compact subset of $\mathbb{R} \times (0,\infty) \times [0,\infty) \times [0,1)$ such that $\boldsymbol{\theta}_0^{(i)} \in \mathring{\Theta}$; (ii) the support of the distribution of η_1 contains at least 3 points, Θ_β is a convex compact set and $\boldsymbol{\vartheta}_0 \in \mathring{\Theta}_\beta$; (iii) η_t is independent of the sigma field generated by $\{\boldsymbol{\eta}_u^\top, \boldsymbol{z}_u^\top, u < t\}$ and $\boldsymbol{x}_t$; (iv) for $i \in \{1,\ldots,p\}$, the conditional distribution of η_{it} given $\{\boldsymbol{\eta}_u^\top, \boldsymbol{z}_u^\top, u < t\}$ and $\{\eta_{j,t}, j \neq i\}$ is not degenerate; (v) η_t does not belong to the sigma field generated by $\mathcal{F}_{t-1}$, $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$; (vi) for $i \in \{1,\ldots,p\}$, η_{it} does not belong to the sigma field generated by $\mathcal{F}_{t-1}$ and $\boldsymbol{z}_t$; (vii) for $i \in \{1,\ldots,q\}$, z_{it} does not

belong to the sigma field generated by $\mathcal{F}_{t-1}$ and $\{z_{j,t}, j \neq i\}$; (viii) if $\boldsymbol{\vartheta} \in \mathring{\Theta}_\beta$ then the parameter obtained by replacing ϑ_i with 0 for all $i \in S$ also belongs to $\mathring{\Theta}_\beta$.

It is well-known that the parameter of a GARCH model is not identifiable when the square of the noise is degenerated (*i.e.* equal to 1). This is why we require conditional distributions that take at least 3 values. It is clear that **A3** (iii)-(vii) are satisfied when $\boldsymbol{\eta}_t$ and $(\boldsymbol{z}_t)$ are independent and when $\boldsymbol{\eta}_t$ and $\boldsymbol{z}_t$ have non singular variances. Note also that **A3** (iv) avoids the multicollinearity of the regressors. This assumption even guarantees that if $\boldsymbol{\lambda}_t^\top \boldsymbol{x}_t = 0$ a.s. for some $\boldsymbol{\lambda}_t \in \mathcal{F}_{t-1}$ then $\boldsymbol{\lambda}_t = \mathbf{0}$. Finally **A3** (viii) implies that $\widehat{\vartheta}_{ni} = 0$ for all $i \in S$ if λ_n is large enough in (6).

Note that, under **A2**, $g_{i,t}^2(\cdot)$ $g_t^2(\cdot)$, $v_t(\cdot)$, $\beta_{i,t}(\cdot)$ and $\ell_t(\cdot)$ admit derivatives of any order uniformly in $\boldsymbol{\varphi} \in \Theta_G$ (by the arguments given at the beginning of Section 3.5 of BFL). We need the following moment conditions.

A4: (i) we have $E(\varepsilon_{it}^4) < \infty$ for all $i \in \{1, \ldots, p\}$; (ii) for all $\boldsymbol{\varphi}_1 \in \Theta_G$ there exist a neighborhood $V(\boldsymbol{\varphi}_1)$ of $\boldsymbol{\varphi}_1$ such that $E(\sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_1)} |\ell_t(\boldsymbol{\varphi})|) < \infty$; (iii) there exists a neighborhood $V(\boldsymbol{\varphi}_0)$ of $\boldsymbol{\varphi}_0$ and (conjugate) numbers $p_i \geq 1$ and $q_i \geq 1$ satisfying $1/p_i + 1/q_i = 1$ for $i = 1, \ldots, 4$ such that

$$
\begin{array}{ll}
\left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \frac{v_t^2(\boldsymbol{\varphi})}{g_t^2(\boldsymbol{\varphi})} \right\|_{p_1} < \infty, & \left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \left\| \frac{1}{g_t^2(\boldsymbol{\varphi})} \frac{\partial g_t^2(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right\| \right\|_{2q_1} < \infty, \\
\left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \frac{v_t^2(\boldsymbol{\varphi})}{g_t^2(\boldsymbol{\varphi})} \right\|_{p_2} < \infty, & \left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \left\| \frac{1}{g_t^2(\boldsymbol{\varphi})} \frac{\partial^2 g_t^2(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} \right\| \right\|_{q_2} < \infty, \\
\left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \frac{|v_t(\boldsymbol{\varphi})|}{g_t(\boldsymbol{\varphi})} \right\|_{p_3} < \infty, & \left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \left\| \frac{1}{g_t(\boldsymbol{\varphi})} \frac{\partial^2 v_t^2(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} \right\| \right\|_{q_3} < \infty \\
\left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \frac{|v_t(\boldsymbol{\varphi})|}{g_t(\boldsymbol{\varphi})} \right\|_{p_4} < \infty, & \left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \left\| \frac{1}{g_t^2(\boldsymbol{\varphi})} \frac{\partial g_t^2(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \frac{1}{g_t(\boldsymbol{\varphi})} \frac{\partial v_t(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}^\top} \right\| \right\|_{q_4} < \infty, \\
\left\| \sup_{\boldsymbol{\varphi} \in V(\boldsymbol{\varphi}_0)} \frac{1}{g_t(\boldsymbol{\varphi})} \left\| \frac{\partial v_t(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}^\top} \right\| \right\|_2 < \infty. &
\end{array}
$$

Note that **A4** (i) is required for the asymptotic normality of the QMLE $\widehat{\boldsymbol{\theta}}^{(i)}$ because the

GARCH equations contain the intercept μ_i (see Example 7.2 in Francq & Zakoïan 2019). It is easy to see that $E\ell_t(\boldsymbol{\varphi}_0) = 1 + E\log g_t^2(\boldsymbol{\varphi}_0) < \infty$ under **A1** and that $E\ell_t^-(\boldsymbol{\varphi}) < \infty$ under **A1**–**A2**. However, it is a priori possible to have $E\ell_t^+(\boldsymbol{\varphi}_1) = +\infty$ for some $\boldsymbol{\varphi}_1 \neq \boldsymbol{\varphi}_0$. It is easy to show that we then have $\widetilde{O}_n(\boldsymbol{\varphi}_1) \to \infty$, and thus $\widehat{\boldsymbol{\varphi}}_n$ cannot be equal to $\boldsymbol{\varphi}_1$, for n large enough. Such points can therefore be excluded from the parameter space. This is what **A4**(ii) does. Under this assumption, the application $\boldsymbol{\varphi} \mapsto E\ell_1(\boldsymbol{\varphi})$ is well defined and continuous. Assumption **A4** (iii) is also made in BFL. It implies consistency of empirical estimates of Fisher information matrices.

We introduce additional notations. From (4) it is seen that the parameters ϖ_i and c_i are not uniquely defined when $\boldsymbol{\vartheta}^{(i)}$ is of the form

$$\boldsymbol{\vartheta}^{(i)} = (\varpi_i, 0, c_i, \mathbf{0}_q)^\top. \tag{S4}$$

The simplest equivalent form of such a $\boldsymbol{\vartheta}^{(i)}$ is $\breve{\boldsymbol{\vartheta}}^{(i)} = (\varpi_i/(1-c_i), \mathbf{0}_{q+2})^\top$. When $\boldsymbol{\vartheta}^{(i)}$ is not of the form (S4), we set $\breve{\boldsymbol{\vartheta}}^{(i)} = \boldsymbol{\vartheta}^{(i)}$. Let also $\breve{\boldsymbol{\vartheta}} = \left(\boldsymbol{\vartheta}^{(0)\top}, \breve{\boldsymbol{\vartheta}}^{(1)\top}, \ldots, \breve{\boldsymbol{\vartheta}}^{(p)\top}\right)^\top$. Note that $p\left(\breve{\boldsymbol{\vartheta}}^{(i)}\right) < p\left(\boldsymbol{\vartheta}^{(i)}\right)$, and thus $Q_n(\breve{\boldsymbol{\vartheta}}) < Q_n(\boldsymbol{\vartheta})$ when (S4) holds with $c_i \neq 0$. Note also that we have assumed $\boldsymbol{\vartheta}_0 = \breve{\boldsymbol{\vartheta}}_0$, and that it is not possible to impose this constraint over Θ_β, which is required to be compact and convex.

We now need an assumption implying the asymptotic uniqueness of the penalized estimator.

A5(λ_0)**:** There exists (a unique) $\boldsymbol{\vartheta}^\circledast \in \Theta_\beta$, such that $Q_{\lambda_0}(\boldsymbol{\vartheta}) > Q_{\lambda_0}(\boldsymbol{\vartheta}^\circledast)$ for all $\boldsymbol{\vartheta} \in \Theta_\beta$, $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}^\circledast$.

Note that **A5**(λ_0) would be satisfied for all $\lambda_0 \geq 0$ if the function $\boldsymbol{\vartheta} \mapsto E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta})$ where strictly convex. Here, the objective function is not convex, but Lemmas S1 and S2 below, as well as an illustrative example given in the supplementary file, show that **A5**(λ_0) generally holds true, at least when $\lambda_0 > 0$ is sufficiently small. This is because $\boldsymbol{\vartheta} \mapsto E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta})$ is minimum for all $\boldsymbol{\vartheta}$ such that $\breve{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}_0$, and when $\breve{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}_0$ we have $p(\boldsymbol{\vartheta}) > p(\boldsymbol{\vartheta}_0)$.

Note however that one can not take $\lambda_0 = 0$, that is $\mathbf{A5}(0)$ does not hold true when (4) holds, because in this case there exist several $\boldsymbol{\vartheta}$ such that $\breve{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}_0$ and thus the solution of the penalized optimization is not unique. *i.e.* β_{it} is constant for some $i \in \{1, \ldots, p\}$. Note also that, in general $\boldsymbol{\vartheta}^{\circledast} \neq \boldsymbol{\vartheta}_0$. Thus the penalization introduces a bias, but Lemma S2 below shows that the bias is small when $\lambda_0 > 0$ is small.

Let $\mathcal{A}$ (resp. $\mathcal{A}^{\circledast}$) be the set of indices $i \in S$ such that $\vartheta_{0i} \neq 0$ (resp. $\vartheta^{\circledast}_{0i} \neq 0$). When $\lambda_0 > 0$ is small, Lemma S2 below shows that $\boldsymbol{\vartheta}^{\circledast}$ is close to $\boldsymbol{\vartheta}_0$, and thus the following assumption holds when λ_0 is sufficiently small.

A6(λ_0)**:** $\mathcal{A}^{\circledast} = \mathcal{A}$.

B Complementary proofs and results

We first recall that the penalized estimator (6) can be interpreted as a constrained estimator. Indeed, note that

$$\min_{\boldsymbol{\vartheta}} E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) + \lambda p(\boldsymbol{\vartheta}) \tag{S5}$$

subject to $\lambda \geq 0$, is the Lagrangian form of the constrained optimization problem

$$\min_{\boldsymbol{\vartheta} \in \Theta_\beta} E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) \quad \text{subject to} \quad p(\boldsymbol{\vartheta}) \leq p^\star. \tag{S6}$$

If $\boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_1(\lambda)$ is solution of (S5) then it is also solution of (S6) with $p^\star = p(\boldsymbol{\vartheta}_1)$. Conversely, (S6) is associated with the Lagrangian $\mathcal{L}(\boldsymbol{\vartheta}, \nu) = E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) + \nu \{p(\boldsymbol{\vartheta}) - p^\star\}$ where $\nu \geq 0$ (penalizing $\boldsymbol{\vartheta}$ when the inequality in (S6) is not satisfied). If $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(p^\star)$ is solution of (S6) then $E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}(p^\star)) \geq \min_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta}, \nu)$ for all $\nu \geq 0$ (by the max-min inequality). If there is strong duality then the Karush-Kuhn-Tucker (KKT) conditions are satisfied and $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(p^\star)$ is solution of (S5) with $\lambda = \arg\max_\nu \min_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta}, \nu)$. The strong duality holds true when $\boldsymbol{\vartheta} \mapsto E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta})$ is convex, but when this is not the case, the constrained optimization

(S6) and Lagrangian optimization (S5) are not necessarily equivalent.

Lemma S1 *For $p^\star = p(\boldsymbol{\vartheta}_0)$, under Assumptions* **A1**-**A4**, *the constrained optimization problem* (S6) *admits a unique solution, which is $\boldsymbol{\vartheta}_0$.*

Remark S1 (Uniqueness of the solution) *When* (4) *holds and $p^\star > p(\boldsymbol{\vartheta}_0)$ the constrained optimization problem* (S6) *does not have a unique solution. This is because there exists $\boldsymbol{\vartheta}$ such that $\breve{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}_0$ (and thus $E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) = E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}_0)$) and $\boldsymbol{\vartheta}^{(i)} = (\varpi_i, 0, c_i, \mathbf{0}_q^\top)^\top$ with $c_i > 0$ such that $p(\boldsymbol{\vartheta}_0) < p\left(\breve{\boldsymbol{\vartheta}}\right) < p^\star$. If $p^\star < p(\boldsymbol{\vartheta}_0)$, the optimization problem* (S6) *can have a unique solution, but there is no guarantee because $E\ell_1(\boldsymbol{\theta}_0, \cdot)$ is not a convex function in general.*

Proof of Lemma S1. From BFL (see iii) of the proof of their Theorem 1) we already know that, in the case $\xi_{0i} \neq 0$ for $i = 1, \ldots, p$, $E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) > E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}_0)$ when $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}_0$. Now consider the case where $\boldsymbol{\vartheta}_0$ is such that $\xi_{0i} = 0$ for some $i \in \{1, \ldots, p\}$ and assume $E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) = E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}_0)$. We then have $v_t(\boldsymbol{\varphi}) = v_t$ and $g_t^2(\boldsymbol{\varphi}) = g_t^2$ a.s. By **A3** (iv) it follows that $\beta_{it}(\boldsymbol{\varphi}) = \beta_{it}$ a.s. In view of (3) and (S1) we thus have

$$v_t x_{i,t} \frac{\xi_i}{\mu_i^2 + g_{i,t}^2} \in \sigma\{\eta_{u-1}, \boldsymbol{x}_u, \boldsymbol{z}_u : u \leq t\}$$

where $\sigma\{S\}$ denotes the sigma-field generated by a set S of random variables. By **A3** (v) and (vi), it follows that $\xi_i = 0$. Now $\beta_{i,t}(\boldsymbol{\varphi}) = \beta_{i,t}$ a.s. with $\xi_i = \xi_{0i} = 0$ entails

$$\varpi_{0i} - \varpi_i + (c_i - c_{i0})\beta_{i,t-1} + \sum_{j=1}^{q} (\gamma_{j,i} - \gamma_{0j,i}) z_{j,t-1} = 0 \quad \text{a.s.}$$

By **A3** (vii), we thus have $\gamma_{j,i} = \gamma_{0j,i}$ for $j = 1, \ldots, q$. If $\gamma_{0j,i} \neq 0$ for some $j \in \{1, \ldots, q\}$ then, again by **A3** (vii), $\beta_{i,t-1}$ is not degenerate and we must have $\varpi_{0i} = \varpi_i$ and $c_i = c_{i0}$, and then $\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_0^{(i)}$. Now, if $\gamma_{0ji} = 0$ for all $j \in \{1, \ldots, q\}$ then (4) holds, $\beta_{i,t-1}$ and $\beta_{i,t}(\boldsymbol{\varphi})$ are

constant, and $\varpi_i/(1-c_i) = \varpi_{0i}/(1-c_{0i})$. For i such that $\xi_{0i} \neq 0$, by the arguments given in BFL, $\beta_{i,t}(\boldsymbol{\varphi}) = \beta_{i,t}$ a.s. implies $\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_0^{(i)}$. It is also clear that $v_t(\boldsymbol{\varphi}) = v_t$ and $g_t^2(\boldsymbol{\varphi}) = g_t^2$ a.s. entails $\boldsymbol{\vartheta}_0 = \boldsymbol{\vartheta}_0^{(0)}$.

We, thus, have shown that $E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) \geq E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}_0)$ and $E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) = E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}_0)$ if and only if $\breve{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}_0$. Since $p(\boldsymbol{\vartheta}) > p(\boldsymbol{\vartheta}_0)$ for all $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}_0$ such that $\breve{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}_0$, the conclusion follows. □

Lemma S2 *Under Assumptions* **A1**-**A4**, *for all neighborhood* V_0 *of* $\boldsymbol{\vartheta}_0$, *there exists* $\lambda > 0$ *such that for all* $\lambda_0 \in (0, \lambda)$ *we have*

$$\left\{ \arg\min_{\boldsymbol{\vartheta} \in \Theta_\beta} Q_{\lambda_0}(\boldsymbol{\vartheta}) \right\} \subset V_0.$$

Proof of Lemma S2. The compact set Θ_β is covered by the union of V_0 and a finite number of open balls $V_1, \ldots V_{k_1+k_2}$ such that the center $\boldsymbol{\vartheta}_k$ of V_k satisfies $\breve{\boldsymbol{\vartheta}}_k \neq \boldsymbol{\vartheta}_0$ for $k = 1, \ldots, k_1$, $\breve{\boldsymbol{\vartheta}}_k = \boldsymbol{\vartheta}_0$ for $k = k_1 + 1, \ldots, k_1 + k_2$, and

$$m_k := \inf_{\boldsymbol{\vartheta} \in V_k} E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}) > m_0 := E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}_0) = \inf_{\boldsymbol{\vartheta} \in V_j} E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta}),$$

for all $k = 1, \ldots, k_1$ and $j = k_1 + 1, \ldots, k_1 + k_2$. Let $\overline{p} := \sup_{\boldsymbol{\vartheta} \in \Theta_\beta} p(\boldsymbol{\vartheta}) \in (0, \infty)$. By choosing $0 < \lambda_0 < \inf_{k=1,\ldots,k_1}(m_k - m_0)/\overline{p}$, we have

$$Q_{\lambda_0}(\boldsymbol{\vartheta}_0) < m_0 + \lambda_0 \overline{p} < m_k \leq \inf_{\boldsymbol{\vartheta} \in V_k} Q_{\lambda_0}(\boldsymbol{\vartheta})$$

for all $k = 1, \ldots, k_1$. Now recall that $p(\boldsymbol{\vartheta}_k) > p(\boldsymbol{\vartheta}_0)$ if $\breve{\boldsymbol{\vartheta}}_k = \boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta}_k \neq \boldsymbol{\vartheta}_0$. By continuity of $p(\cdot)$ we also have $\inf_{\boldsymbol{\vartheta} \in V_k} p(\boldsymbol{\vartheta}) > p(\boldsymbol{\vartheta}_0)$ when V_k is sufficiently small, which can be assumed without loss of generality by compactness of Θ_β. For $\lambda_0 > 0$, we then have

$$Q_{\lambda_0}(\boldsymbol{\vartheta}_0) = m_0 + \lambda_0 p(\boldsymbol{\vartheta}_0) < \inf_{\boldsymbol{\vartheta} \in V_k} Q_{\lambda_0}(\boldsymbol{\vartheta})$$

for all $k = k_1 + 1, \ldots, k_1 + k_2$. □

Lemma S3 *Under Assumptions* **A1**-**A4**, *if* $\lambda_n \to \lambda_0$ *we have*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\vartheta}\in\Theta_\beta} \left|\widetilde{Q}_n(\boldsymbol{\vartheta}) - Q_{\lambda_0}(\boldsymbol{\vartheta})\right| = 0 \ \textit{a.s.}$$

Proof of Lemma S3. As in i) of the proof of Theorem 1 in BFL, using (S2)-(S3), we have

$$\lim_{n\to\infty} \sup_{\boldsymbol{\varphi}\in\Theta_G} \left|\widetilde{O}_n(\boldsymbol{\varphi}) - O_n(\boldsymbol{\varphi})\right| = 0 \text{ a.s.}$$

where

$$O_n(\boldsymbol{\varphi}) = \frac{1}{n}\sum_{t=1}^{n} \ell_t(\boldsymbol{\varphi}), \qquad \ell_t(\boldsymbol{\varphi}) = \frac{v_t^2(\boldsymbol{\varphi})}{g_t^2(\boldsymbol{\varphi})} + \log g_t^2(\boldsymbol{\varphi}).$$

Let $V_k(\boldsymbol{\varphi})$ be the intersection of the ball of center $\boldsymbol{\varphi}$ and radius $1/k$ with the parameter space Θ_G. For all $\boldsymbol{\varphi}^* \in \Theta_G$, we have

$$\sup_{\boldsymbol{\varphi}\in V_k(\boldsymbol{\varphi}^*)} |O_n(\boldsymbol{\varphi}) - E\ell_t(\boldsymbol{\varphi})| \le \frac{1}{n}\sum_{t=1}^{n} \sup_{\boldsymbol{\varphi}\in V_k(\boldsymbol{\varphi}^*)} |\ell_t(\boldsymbol{\varphi}) - \ell_t(\boldsymbol{\varphi}^*)| + \left|\frac{1}{n}\sum_{t=1}^{n} \ell_t(\boldsymbol{\varphi}^*) - E\ell_t(\boldsymbol{\varphi}^*)\right|$$
$$+ \sup_{\boldsymbol{\varphi}\in V_k(\boldsymbol{\varphi}^*)} |E\ell_t(\boldsymbol{\varphi}) - E\ell_t(\boldsymbol{\varphi}^*)|.$$

The almost sure limit as $n \to \infty$ of the first term of the right-hand side of the inequality can be made arbitrarily small when k is large, using the ergodic theorem, the continuity of $\ell_t(\cdot)$ and Beppo Levi's theorem. The second term converges almost surely to zero by the ergodic theorem. The third term is arbitrarily small when k is large, by the continuity of $E\ell_t(\cdot)$. For all neighborhood V of $(\boldsymbol{\theta}_0, \boldsymbol{\vartheta})$ (transposes are omitted for simplicity of notation), the consistency of $\widehat{\boldsymbol{\theta}}$ entails

$$\left|O_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) - O_n(\boldsymbol{\theta}_0, \boldsymbol{\varphi})\right| \le \frac{1}{n}\sum_{t=1}^{n} \sup_{\boldsymbol{\varphi}\in V} |\ell_t(\boldsymbol{\varphi}) - \ell_t(\boldsymbol{\theta}_0, \boldsymbol{\vartheta})|$$

which is arbitrarily small when $n \to \infty$ and V is small enough. The conclusion follows by compactness of Θ_G. □

The next result comes from the argmax theorem (see Theorem 3.2.2 of Van Der Vaart & Wellner 1996). It is also close to Lemma 2.2 in Davis et al. (1992) and its Remark 1. For the sake of completeness we give the proof.

Lemma S4 *Let $(\boldsymbol{U}, d)$ be a metric space, and let stochastic processes $\{V_n(\boldsymbol{u}), \boldsymbol{u} \in \boldsymbol{U}\}$ and $\{V(\boldsymbol{u}), \boldsymbol{u} \in \boldsymbol{U}\}$ continuous and with values in $\mathbb{R}$. Assume that $(\widehat{\boldsymbol{u}}_n)$ is a sequence of random elements such that $V_n(\widehat{\boldsymbol{u}}_n) \leq V_n(\boldsymbol{u})$ for all $\boldsymbol{u} \in \boldsymbol{U}$ and suppose there exists a random element $\widehat{\boldsymbol{u}}$ such that almost surely $V(\widehat{\boldsymbol{u}}) < V(\boldsymbol{u})$ for all $\boldsymbol{u} \in \boldsymbol{U}$ such that $\boldsymbol{u} \neq \widehat{\boldsymbol{u}}$. If (a) $\sup_{\boldsymbol{u} \in \boldsymbol{K}} |V_n(\boldsymbol{u}) - V(\boldsymbol{u})| = o_P(1)$ for all compact set $\boldsymbol{K}$, and (b) the sequence $(\widehat{\boldsymbol{u}}_n)$ is tight, then $d(\widehat{\boldsymbol{u}}_n, \widehat{\boldsymbol{u}}) = o_P(1)$ as $n \to \infty$.*

Proof of Lemma S4. Let some arbitrary $\epsilon > 0$. By (b) there exists a compact set $\boldsymbol{K}$ such $P(\widehat{\boldsymbol{u}}_n \in \boldsymbol{K}) > 1 - \epsilon$ for all n. It is not restrictive to choose $\boldsymbol{K}$ such $P(\widehat{\boldsymbol{u}} \in \boldsymbol{K}) > 1 - \epsilon$. For every $\iota > 0$, we have $\psi(\iota) := \inf_{\boldsymbol{u} \in \boldsymbol{K}, d(\boldsymbol{u}, \widehat{\boldsymbol{u}}) \geq \iota} V(\boldsymbol{u}) - V(\widehat{\boldsymbol{u}}) > 0$. Note that

$$
\begin{aligned}
\{d(\widehat{\boldsymbol{u}}_n, \widehat{\boldsymbol{u}}) \geq \iota\} &\subset \{\widehat{\boldsymbol{u}}_n \notin \boldsymbol{K}\} \cup \{V(\widehat{\boldsymbol{u}}_n) - V(\widehat{\boldsymbol{u}}) \geq \psi(\iota)\} \\
&\subset \{\widehat{\boldsymbol{u}}_n \notin \boldsymbol{K}\} \cup \{V(\widehat{\boldsymbol{u}}_n) - V(\widehat{\boldsymbol{u}}) + V_n(\widehat{\boldsymbol{u}}) - V_n(\widehat{\boldsymbol{u}}_n) \geq \psi(\iota)\} \\
&\subset \{\widehat{\boldsymbol{u}}_n \notin \boldsymbol{K}\} \cup \{\widehat{\boldsymbol{u}} \notin \boldsymbol{K}\} \cup \left\{\sup_{\boldsymbol{u} \in \boldsymbol{K}} |V_n(\boldsymbol{u}) - V(\boldsymbol{u})| \geq \psi(\iota)/2\right\}.
\end{aligned}
$$

The conclusion follows. □

Proof of Theorem 1. The result follows from Lemmas S3 and S4. To check (b) in Lemma S4, we note that, since Θ_β is compact, the sequence $(\widehat{\boldsymbol{\vartheta}}_n)$ is tight. □

Proof of Proposition 1. Let $j \in S$. By the arguments given in Remark 3, any solution

of (6) satisfies

$$\frac{1}{n}\sum_{t=2}^{n}\frac{\partial\widetilde{\ell}_t(\widehat{\boldsymbol{\varphi}}_n)}{\partial\vartheta_j}=\lambda_n \text{ if } \widehat{\vartheta}_{nj}<0, \quad \frac{1}{n}\sum_{t=2}^{n}\frac{\partial\widetilde{\ell}_t(\widehat{\boldsymbol{\varphi}}_n)}{\partial\vartheta_j}=-\lambda_n \text{ if } \widehat{\vartheta}_{nj}>0.$$

Therefore $\widehat{\vartheta}_{nj}=0$ if (11) holds with $\widehat{\boldsymbol{\vartheta}}_n^c=\widehat{\boldsymbol{\vartheta}}_n$. Now, if $\boldsymbol{\vartheta}\mapsto\widetilde{O}_n(\widehat{\boldsymbol{\theta}},\boldsymbol{\vartheta})$ is a strictly convex function, then the function $\vartheta_i\mapsto\frac{1}{n}\sum_{t=2}^{n}\frac{\partial\widetilde{\ell}_t(\widehat{\boldsymbol{\theta}},\boldsymbol{\vartheta})}{\partial\vartheta_j}$ is strictly increasing, which implies that (10) is then an equality (see Figure S3 for an illuminating graphical illustration). □

Proof of Theorem 2. By (13) and the assumption $\lambda_n^a>0$, with probability tending to 1 we have $\lambda_n^a\widehat{\delta}_i=+\infty$ for all $i\in\mathcal{I}$. We thus have $|\partial\widetilde{O}_n(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\vartheta}}_n^a)/\partial\vartheta_i|<\lambda_n^a\widehat{\delta}_i$ for all $i\in\mathcal{I}$, and the FOC entails $\widehat{\boldsymbol{\vartheta}}_n^a\in U$, or equivalently $\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}}_n^a-\boldsymbol{\vartheta}_0\right)\in U$, with probability tending to 1.

Let $\boldsymbol{u}_2=\sqrt{n}(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0)$. Note that

$$\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}}_n^a-\boldsymbol{\vartheta}_0\right)=\underset{\boldsymbol{u}_2\in\mathbb{R}^{d_2}}{\arg\min}\,W_n(\boldsymbol{u}_2),$$

with $W_n(\boldsymbol{u}_2)=\sum_{t=2}^{n}\widetilde{\ell}_t(\widehat{\boldsymbol{\theta}},\frac{\boldsymbol{u}_2}{\sqrt{n}}+\boldsymbol{\vartheta}_0)+n\lambda_n^a\left\{p_{\widehat{\boldsymbol{\delta}}}(\frac{\boldsymbol{u}_2}{\sqrt{n}}+\boldsymbol{\vartheta}_0)-p_{\widehat{\boldsymbol{\delta}}}(\boldsymbol{\vartheta}_0)\right\}-\sum_{t=2}^{n}\widetilde{\ell}_t(\widehat{\boldsymbol{\theta}},\boldsymbol{\vartheta}_0)$. Using $\sqrt{n}\lambda_n^a\to\lambda_0^a$, $\lambda_n^a>0$ and (13) we obtain[1] the convergence in probability

$$n\lambda_n^a\left\{p_{\widehat{\boldsymbol{\delta}}}(\frac{\boldsymbol{u}_2}{\sqrt{n}}+\boldsymbol{\vartheta}_0)-p_{\widehat{\boldsymbol{\delta}}}(\boldsymbol{\vartheta}_0)\right\}\to\lambda_0^a\,p(\boldsymbol{\delta},\boldsymbol{\vartheta}_0,\boldsymbol{u}_2)\quad\text{if}\quad\boldsymbol{u}_2\in U$$

and, with probability tending to 1,

$$n\lambda_n^a\left\{p_{\widehat{\boldsymbol{\delta}}}(\frac{\boldsymbol{u}_2}{\sqrt{n}}+\boldsymbol{\vartheta}_0)-p_{\widehat{\boldsymbol{\delta}}}(\boldsymbol{\vartheta}_0)\right\}=+\infty\quad\text{if}\quad\boldsymbol{u}_2\notin U.$$

We thus have, with probability tending to 1,

$$\sqrt{n}\boldsymbol{A}\left(\widehat{\boldsymbol{\vartheta}}_n^a-\boldsymbol{\vartheta}_0\right)=\underset{\boldsymbol{u}\in\mathbb{R}^{d_6}}{\arg\min}\,U_n(\boldsymbol{u}),\quad U_n(\boldsymbol{u})=W_n(\boldsymbol{A}^\top\boldsymbol{u}).$$

[1] using $\sqrt{n}\left(\left|a+\frac{u}{\sqrt{n}}\right|-|a|\right)=u\times\text{sign}(a)1_{a\neq0}+|u|1_{a=0}$ for n large enough.

A second-order Taylor expansion yields

$$\sum_{t=2}^{n} \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}}, \frac{\boldsymbol{A}^\top \boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\vartheta}_0) = \sum_{t=2}^{n} \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}_0) + \boldsymbol{u}^\top \boldsymbol{A} \frac{1}{\sqrt{n}} \sum_{t=2}^{n} \frac{\partial \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}_0)}{\partial \boldsymbol{\vartheta}} + \frac{1}{2} \boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{J}_{\boldsymbol{\vartheta},n} \boldsymbol{A}^\top \boldsymbol{u},$$

where the row i of the $d_2 \times d_2$ matrix $\boldsymbol{J}_{\boldsymbol{\vartheta},n}$ is of the form $\partial^2 \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \frac{\boldsymbol{u}_i^*}{\sqrt{n}} + \boldsymbol{\vartheta}_0)/\partial \boldsymbol{\vartheta}^\top \partial \boldsymbol{\vartheta}_i$ for some $\boldsymbol{u}_i^*$ between $\boldsymbol{A}^\top \boldsymbol{u}$ and $\boldsymbol{0}_{d_2}$. Another Taylor expansion gives

$$\frac{1}{\sqrt{n}} \sum_{t=2}^{n} \frac{\partial \ell_t(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}_0)}{\partial \boldsymbol{\vartheta}} = \frac{1}{\sqrt{n}} \sum_{t=2}^{n} \frac{\partial \ell_t(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\vartheta}} + \boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta},n} \sqrt{n} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right),$$

where the row i of the $d_2 \times d_1$ matrix $\boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta},n}$ is of the form $n^{-1} \sum_{t=2}^{n} \partial^2 \ell_t(\boldsymbol{\theta}_i^*, \boldsymbol{\vartheta}_0)/\partial \boldsymbol{\vartheta} \partial \boldsymbol{\theta}^\top$ for some $\boldsymbol{\theta}_i^*$ between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Let

$$V_n(\boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{A} \left\{ \frac{1}{\sqrt{n}} \sum_{t=2}^{n} \frac{\partial \ell_t(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\vartheta}} + \boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta}} \sqrt{n} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \right\} + \frac{1}{2} \boldsymbol{u}^\top \boldsymbol{J}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}} \boldsymbol{u} + \lambda_0^a \, p(\boldsymbol{\delta}, \boldsymbol{\vartheta}_0, \boldsymbol{A}^\top \boldsymbol{u}).$$

It is easy to show that for all compact sets $\boldsymbol{K} \subset \mathbb{R}^{d_6}$ we have $\sup_{\boldsymbol{u} \in \boldsymbol{K}} |U_n(\boldsymbol{u}) - V_n(\boldsymbol{u})| = o_P(1)$. By (5), the central limit theorem for martingale differences and standard arguments, we have the weak convergence[2]

$$V_n(\cdot) \xrightarrow{d} V(\cdot) \quad \text{on} \quad C(\mathbb{R}^{d_6}).$$

By the argument given in the proof of Theorem 2 in BFL, it can be shown that $\boldsymbol{J}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}}$ is not singular. It follows that, almost surely, $V(\cdot)$ is strictly convex, and thus admits a unique minimum. By Skorokhod's representation theorem (see the beginning of the proof of Lemma 2.2 in Davis et al. (1992), it is not restrictive to assume that (a) of Lemma S4 holds. The convergence in distribution is then obtained from Lemmas S4 and S5. □

Lemma S5 *Under Assumptions of Theorem 2,* $\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}}_n^a - \boldsymbol{\vartheta}_0\right) = O_P(1)$.

[2]Since the processes are convex, it is sufficient to notice the convergence of the finite dimensional distributions.

Proof of Lemma S5. With the notations and arguments of the beginning of the proof of Theorem 2, it suffices to show that $\arg\min_{\boldsymbol{u}\in\mathbb{R}^{d_6}} V_n(\boldsymbol{u}) = O_P(1)$. We will show that for any $\epsilon > 0$ there exists a large constant C, such that when $\|\boldsymbol{u}\| \geq C$, $P(V_n(\boldsymbol{u}) > 0) \geq 1-\epsilon$. Noting that $V_n(\mathbf{0}) = 0$, this will show that, with probability at least $1-\epsilon$, $\widehat{u}_n := \arg\min_{\boldsymbol{u}\in\mathbb{R}^{d_6}} V_n(\boldsymbol{u})$ is such that $\widehat{u}_n \leq C$, which is the announced result.

Since $V_n(\boldsymbol{u}) = V(\boldsymbol{u}) + o_P(1)$, it suffices to note that for any small constant $\iota > 0$ and large constant C_2, we have $V(\boldsymbol{u}) > \iota$ when $\|\boldsymbol{W}_2\| \leq C_2$ and $\|\boldsymbol{u}\| \geq C$ with C large enough. Indeed, since we have $V(\boldsymbol{u}) \geq \lambda_1\|\boldsymbol{u}\|^2 - \|\boldsymbol{u}\|\|\boldsymbol{W}_2\| - \overline{\delta}\|\boldsymbol{u}\|$ where $\lambda_1 > 0$ is the smallest eigenvalue of $\boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}}$ and $\overline{\delta} = \lambda_0^a \max_{i\in\mathcal{A}} \delta_i$, one can take $C = (C_2 + \overline{\delta})/2\lambda_1 + \sqrt{\iota + (C_2+\overline{\delta})^2/4\lambda_1^2}$. □

Proof of Corollary 1. When $\lambda_0^a = 0$, $\widehat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}\in\mathbb{R}^{d_6}} V(\boldsymbol{u})$ is such that

$$\boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}}\widehat{\boldsymbol{u}} = \boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}\boldsymbol{\theta}}\boldsymbol{J}_*^{-1}\boldsymbol{W}_1 - \boldsymbol{W}_2 \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}}\boldsymbol{\Sigma}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}}\boldsymbol{J}^{\boldsymbol{A}}_{\boldsymbol{\vartheta}}\right).$$

□

C Illustration of A5(λ_0)

We illustrate the assumption when $Q_\lambda(\boldsymbol{\vartheta}) = Q(\boldsymbol{\vartheta}) + \lambda\|\boldsymbol{\vartheta}\|_1$ with $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^\intercal$, $Q(\boldsymbol{\vartheta}) = \mathcal{P}(\vartheta_1 - \vartheta_2)$ and $\mathcal{P}(x) = x^4/4 - x^3/3 - x^2 + 1$. Note that $Q(\boldsymbol{\vartheta})$ is minimal for all $\boldsymbol{\vartheta}$ such that $\vartheta_2 = \vartheta_1 - 2$. Therefore, similar to $\boldsymbol{\vartheta} \mapsto E\ell_1(\boldsymbol{\theta}_0, \boldsymbol{\vartheta})$ in (7) when (4) holds true, the minimum of the function $Q(\boldsymbol{\vartheta})$ is not well identified, but a small penalization λ solves the problem. Indeed, if $\lambda > 0$ is very small but not zero, $\boldsymbol{\vartheta}^{\circledast} = \arg\min Q_\lambda(\boldsymbol{\vartheta})$ is unique and is close to $(1, -1)$, which is the element of $\arg\min Q(\boldsymbol{\vartheta})$ with the smallest norm. On the other hand, when λ is very large $\boldsymbol{\vartheta}^{\circledast} = \arg\min Q_\lambda(\boldsymbol{\vartheta})$ is also unique and equal to $(0, 0)$. The left graph in Figure S1 in I shows the function $\vartheta_1 \mapsto \min_{\vartheta_2} Q_\lambda(\vartheta_1, \vartheta_2)$ for $\lambda \in \{1.8, 1.9, 2.03, 3\}$. For $\lambda < 2.03$, the minimum of the function is obtained for ϑ_1 close to 1 (but less than 1). For

$\lambda > 2.03$, the minimum is obtained for ϑ_1 exactly equal to zero. For $\lambda = 2.03$ the function (black dotted line in the figure) reaches its minimum at 2 points. The right plot shows $\boldsymbol{\vartheta}^{\circledast} = (\vartheta_1^\star, \vartheta_2^\star) = \arg\min Q_\lambda(\boldsymbol{\vartheta})$ as function of λ. For $\lambda = 0$ the minimum is not well defined (*i.e.* not unique). For a small $\lambda > 0$, the minimum of $Q_\lambda(\boldsymbol{\vartheta})$ is obtained for a unique $\boldsymbol{\vartheta}^{\circledast}$ closed to $(1, -1)$ and this point where the function reaches its minimum shrinks slightly toward (0,0) as λ increases. When $\lambda > 2.03$, the point suddenly shrinks to (0,0). Finally, in this example, the assumption $\mathbf{A5}(\lambda_0)$ is always satisfied, except for $\lambda_0 = 0$ (where the minimum is reached at an infinite number of points) and for $\lambda_0 = 2.03$ (where the minimum is reached at two points).

D Example showing that we can have $\lambda^i \neq \lambda^s$

We now give an example showing that the inequality (10) can be strict when the objective function $\widetilde{Q}_n(\cdot)$ is not convex. Let $Q_\lambda(\boldsymbol{\vartheta}) = Q(\boldsymbol{\vartheta}) + \lambda|\vartheta_1|$ with $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^\top$, and $Q(\boldsymbol{\vartheta}) = \sum_{i=1}^2 \vartheta_i^4 - \frac{2}{3}\vartheta_i^3 - 2\vartheta_i^2 + 2\vartheta_i + 1$. Note that $Q(\boldsymbol{\vartheta})$ is a non convex function and that $\boldsymbol{\vartheta}^{\circledast} := \arg\min Q_\lambda(\boldsymbol{\vartheta})$ is of the form $\boldsymbol{\vartheta}^{\circledast} = (\vartheta_1^\star, -1)^\top$. If $\boldsymbol{\vartheta}$ is constrained to be of the form $\boldsymbol{\vartheta} = (0, \vartheta_2)^\top$, the minimum of $Q(\boldsymbol{\vartheta})$ is obtained at $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^c := (0, -1)^\top$ and we have $\lambda^i = \partial Q(\boldsymbol{\vartheta}^c)/\partial\vartheta_1 = 2$. However, for $\lambda = 2$ we have $\boldsymbol{\vartheta}^{\circledast} = (-0.781, -1)^\top \neq \boldsymbol{\vartheta}^c$. This is because $\boldsymbol{\vartheta}^c$ is only a local minimum (see Figure S2). We have $\boldsymbol{\vartheta}^{\otimes} = \boldsymbol{\vartheta}^c$ for $\lambda \geq 2.746$ (see Figure S2). It can be shown that $\lambda^s = \sup_{\vartheta_1 \leq 0} 4\vartheta_1^3 - 2\vartheta_1^2 - 4\vartheta_1 + 2 = 3.032$ which, in accordance with Proposition 1 is larger than 2.746.

Figure S3 represents $\vartheta_1 \mapsto \partial Q(\vartheta_1, -1)/\partial\vartheta_1$ in blue. Using the necessary FOC, the first coordinate $\vartheta_1^\star$ of $\boldsymbol{\vartheta}^{\otimes}$ is the abscissa of one of the points at the intersection of the blue curve and the set $(-\infty, 0) \times \{\lambda\} \cup \{0\} \times [-\lambda, \lambda] \cup (0, +\infty) \times \{-\lambda\}$, displayed in color for 3 different values of λ. When $\lambda < \lambda^i = 2$, we cannot have $\vartheta_1^\star = 0$, and thus $\boldsymbol{\vartheta}^{\otimes} \neq \boldsymbol{\vartheta}^c$. When $\lambda > \lambda^s = 3.032$, we have $\vartheta_1^\star = 0$, and thus $\boldsymbol{\vartheta}^{\otimes} = \boldsymbol{\vartheta}^c$.

E Optimization algorithms

This section provides a full discussion on the NLShoot algorithm including a proof of its validity in Section E.1. Section E.2 presents an example of the NLShoot algorithm for a non convex objective function. The local quadratic approximation (LQA) algorithm introduced by Fan & Li (2001) and adapted to our optimization problem is included in Section E.3. LQA is used to obtain more suitable starting values for the NLShoot algorithm.

E.1 The NLShoot algorithm

This section describes our proposed NLShoot algorithm. In the following, the algorithm is reintroduced for better readability. Suppose Θ is large enough so that $\boldsymbol{\vartheta}(\lambda)$ defined by (15) belongs to the interior of Θ. We also assume that all the shrunk components can be shrunk without restriction: if $\boldsymbol{\vartheta} \in \mathring{\Theta}$ then the parameter obtained by replacing ϑ_i with 0 for all $i \in S$ also belongs to $\mathring{\Theta}$. First 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

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

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 thus propose the following generalized shooting algorithm, which we call NLShoot: start with an initial value $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^0$ and

$$\text{replace the } i\text{-th coordinate of } \boldsymbol{\vartheta} \text{ by } \widetilde{\vartheta}_i \text{ for } i = 1, 2, \ldots, d. \tag{S7}$$

A stationary point of $Q_\lambda(\cdot)$ is a point $\boldsymbol{\vartheta}$ at which all the directional derivatives which exist[3] are positive or zero. All the solutions of (15) are stationary points of $Q_\lambda(\cdot)$.

Proposition 1 *The cluster point(s) of the NLShoot algorithm are stationary points of* $Q_\lambda(\cdot)$.

Proof of Proposition 1. The proof is inspired by the appendix of Wu & Lange (2008). The function $Q(\cdot)$ is differentiable, and thus it is Gateaux-differentiable on $\mathring{\Theta}$. At a point $\boldsymbol{\vartheta}$ belonging to the boundary of Θ, the function $Q_\lambda(\cdot)$ admits directional derivatives along directions crossing the interior of Θ. It follows that $Q_\lambda(\cdot)$ also admits directional derivatives. In fact, the directional derivative $Q_\lambda(\cdot)$ at $\boldsymbol{\vartheta}$ along the direction $\boldsymbol{d} = (d_1, \ldots, d_d)^\top$ is

$$Q'_\lambda(\boldsymbol{\vartheta}, \boldsymbol{d}) := \lim_{\tau \downarrow 0} \frac{Q_\lambda(\boldsymbol{\vartheta} + \tau \boldsymbol{d}) - Q_\lambda(\boldsymbol{\vartheta})}{\tau} = \left\langle \frac{\partial Q(\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}, \boldsymbol{d} \right\rangle + \lambda \sum_{i \in S} \delta_i \left\{ d_i 1_{\vartheta_i > 0} - d_i 1_{\vartheta_i < 0} + |d_i| 1_{\vartheta_i = 0} \right\}.$$

If we denote by $\boldsymbol{e}_i$ the i-th column of the identity matrix of size d, then we have

$$Q'_\lambda(\boldsymbol{\vartheta}, \boldsymbol{d}) = \sum_{d_i > 0} d_i Q'_\lambda(\boldsymbol{\vartheta}, \boldsymbol{e}_i) - \sum_{d_i < 0} d_i Q'_\lambda(\boldsymbol{\vartheta}, -\boldsymbol{e}_i) \geq 0$$

if $\boldsymbol{\vartheta}$ is a stationary point coordinate-wise, *i.e.* if $Q'_\lambda(\boldsymbol{\vartheta}, \boldsymbol{e}_i) \geq 0$ and $Q'_\lambda(\boldsymbol{\vartheta}, -\boldsymbol{e}_i) \geq 0$ for all $i \in \{1, \ldots, d\}$.

[3] If $\boldsymbol{\vartheta}$ stays on the boundary of Θ some directional derivatives may not exist.

Denote by $\boldsymbol{\vartheta}^k = \left(\vartheta_1^k, \ldots, \vartheta_d^k\right)^\top$ the value the value of the parameters in the generalized shooting algorithm at step $k \geq 0$. Since the sequence $\{Q_\lambda(\boldsymbol{\vartheta}^k)\}_k$ is decreasing and bounded below, it converges. By compactness of Θ, there exists a subsequence of $(\boldsymbol{\vartheta}^k)_k$ converging to some $\boldsymbol{\vartheta}^\infty = \left(\vartheta_1^\infty, \ldots, \vartheta_d^\infty\right)^\top \in \Theta$. It suffices to show that $\boldsymbol{\vartheta}^\infty$ is a stationary point coordinate-wise. Assume that $Q'_\lambda(\boldsymbol{\vartheta}^\infty, \boldsymbol{e}_i) < 0$. The existence of this directional derivative implies that $\vartheta_i^\infty + \tau \in \mathring{\Theta}^{(i)}$ for small enough $\tau > 0$. For infinitely many k we also have $Q'_\lambda(\boldsymbol{\vartheta}^k, \boldsymbol{e}_i) \leq c < 0$. Note that

$$Q'_\lambda(\boldsymbol{\vartheta}, \boldsymbol{e}_i) = \frac{\partial Q_0^{(i)}(\vartheta_i; \boldsymbol{\vartheta})}{\partial \vartheta} + \lambda \delta_i \left\{1_{\vartheta_i \geq 0} - 1_{\vartheta_i < 0}\right\}.$$

For ϑ_i of the same sign as ϑ_i^k if $\vartheta_i^k \neq 0$ and for $\vartheta_i > 0$ if $\vartheta_i^k = 0$, with ϑ^* between ϑ_i and ϑ_i^k, a Taylor expansion yields

$$\begin{aligned} Q_\lambda^{(i)}(\vartheta_i; \boldsymbol{\vartheta}^k) &= Q_\lambda^{(i)}(\vartheta_i^k; \boldsymbol{\vartheta}^k) + Q'_\lambda(\boldsymbol{\vartheta}^k, \boldsymbol{e}_i)(\vartheta_i - \vartheta_i^k) + \frac{1}{2}\frac{\partial^2 Q^{(i)}(\vartheta^*; \boldsymbol{\vartheta}^k)}{\partial \vartheta^2}(\vartheta_i - \vartheta_i^k)^2 \\ &\leq Q_\lambda^{(i)}(\vartheta_i^k; \boldsymbol{\vartheta}^k) + c(\vartheta_i - \vartheta_i^k) + \frac{1}{2} b (\vartheta_i - \vartheta_i^k)^2, \end{aligned}$$

with $b = \sup_{\boldsymbol{\vartheta} \in \Theta} \left\| \frac{\partial^2 Q(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top} \right\| < \infty$. First assume that $b \neq 0$. It follows that

$$Q_\lambda(\boldsymbol{\vartheta}^{k+1}) = \min_{\vartheta_i \in \Theta^{(i)}} Q_\lambda^{(i)}(\vartheta_i; \boldsymbol{\vartheta}^k) \leq \min_{\vartheta_i \in \Theta^{(i)}} Q_\lambda(\boldsymbol{\vartheta}^k) + c(\vartheta_i - \vartheta_i^k) + \frac{1}{2} b (\vartheta_i - \vartheta_i^k)^2 = Q_\lambda(\boldsymbol{\vartheta}^k) - \frac{c^2}{2b},$$

if c/b satisfies the condition $\vartheta_i^k - c/b \in \Theta^{(i)}$, and also $\vartheta_i^k - c/b < 0$ if $\vartheta_i^k < 0$. If c/b does not satisfy one of the conditions, one can replace c by $\tilde{c} \in (c, 0)$ such that $\vartheta_i^\infty - \tilde{c}/b \in \mathring{\Theta}^{(i)}$ and also $\vartheta_i^\infty - \tilde{c}/b < 0$ if $\vartheta_i^k < 0$. We then have $Q_\lambda(\boldsymbol{\vartheta}^{k+1}) \leq Q_\lambda(\boldsymbol{\vartheta}^k) - \frac{\tilde{c}^2}{2b}$ for k large enough. Since $\frac{\tilde{c}^2}{2b} > 0$ and the inequality holds for infinitely many k, this contradicts the fact that $Q_\lambda(\boldsymbol{\vartheta}^\infty) > -\infty$. Similar (even simpler) arguments hold when $b = 0$. So we have shown by contradiction that $Q'_\lambda(\boldsymbol{\vartheta}^\infty, \boldsymbol{e}_i) \geq 0$. Similarly, it can be shown that $Q'_\lambda(\boldsymbol{\vartheta}^\infty, -\boldsymbol{e}_i) \geq 0$. □

E.2 NLShoot algorithm usage examples

Let us use the NLShoot algorithm to compute

$$\arg\min_{\vartheta} P(\vartheta) + \lambda|\vartheta|, \quad P(x) = x^4 - \frac{2}{3}x^3 - 2x^2 + 2x.$$

Since the optimization is univariate, NLShoot reduces to a single iteration of (5). Figures S4 and S5 show that, for $\lambda = 2.5$, the set $\{\vartheta < 0 : P'(\vartheta) = \lambda\}$ contains 2 points, that 0 is also a point at which a generalized gradient is zero, and that the set $\{\vartheta > 0 : P'(\vartheta) = -\lambda\}$ is empty. To find the minimum, it is then sufficient to compare the value of the penalized function at these 3 points.

Now consider a more complex optimization where the objective function has multiple interdependent variables. Let $Q_\lambda(\boldsymbol{\vartheta}) = Q(\boldsymbol{\vartheta}) + \lambda\|\boldsymbol{\vartheta}\|_1$ with $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_9)^\top$,

$$Q(\boldsymbol{\vartheta}) = \sum_{i=1}^{9} P\left(\vartheta_i - 1 - \frac{(-1)^i}{i}\right) + \prod_{i=1}^{9}\left(\vartheta_i - \frac{(-1)^i}{i}\right)^2$$

and $P(x) = x^4 - \frac{2}{3}x^3 - 2x^2 + 2x$. Figure S6 displays the minimum of this function obtained with the generalized shooting algorithm applied to a fine grid of values of λ.

E.3 Local Quadratic Approximation (LQA) algorithm

For penalty terms that are non convex and non differentiable, Fan & Li (2001) proposed the local quadratic approximation (LQA) algorithm. In our framework the LQA algorithm consists of repeatedly solving

$$\boldsymbol{\vartheta}^{(k+1)} = \arg\min_{\boldsymbol{\vartheta}\in\Theta}\left\{Q(\boldsymbol{\vartheta}) + \lambda\sum_{i\in S}\delta_i\frac{\operatorname{sign}(\vartheta_i^{(k)})}{2\vartheta_i^{(k)}}\vartheta_i^2\right\} \tag{S8}$$

until convergence, where $\boldsymbol{\vartheta}^{(0)} \in \Theta$ is an initial value (*e.g.* the unpenalized minimum $\boldsymbol{\vartheta}^{(0)} = \boldsymbol{\vartheta}(0)$). Of course, in (S8) we need $\vartheta_i^{(k)} \neq 0$. Fan & Li (2001) suggested to set $|\vartheta_i^{(\cdot)}| \equiv 0$ and to remove this component from the optimization problem if $|\vartheta_i^{(k)}| < \varepsilon$ for some small $\varepsilon > 0$. As noted by Zou & Li (2008) the drawback of the algorithm is the choice of ε which potentially affects the degree of sparsity of the solution as well as the speed of convergence. Moreover, it does not deal with non convex objective functions $Q(\boldsymbol{\vartheta})$. Hence, we only use the algorithm to obtain suitable starting values for the NLShoot algorithm.

F Penalty parameter selection

In the estimation of the PACB model, penalty parameters must be selected in both Step 1 and Step 2. As described in Section 5, each step follows a multi-stage procedure combining a penalized estimation, implemented via the NLShoot algorithm, and a subsequent constrained estimation by QMLE (Post-NLShoot), conditional on the set of active parameters identified in the penalization stage.

The penalty parameter λ is selected by minimizing the Bayesian Information Criterion (BIC) computed from the Post-NLShoot estimates across the set of candidate models. Because different values of λ may lead to the same active set, the mapping from λ to the selected model is not necessarily one-to-one. As a result, multiple values of λ can yield identical BIC values and correspond to the same optimal specification.

In this section, we illustrate the penalized coefficient paths and the corresponding terminal models over a grid of penalty values λ for the simulation design with $p_{tv} = 2$, $p_{cst} = 2$, and $p_{irr} = 2$.

Across all simulation settings, the number of distinct active parameter sets is smaller than the number of λ values in the grid. Consequently, the number of models re-estimated by constrained QMLE is strictly lower than the number of grid points.

We discuss Step 1 and Step 2 separately. In both steps, we start from an equidistant grid of 25 values of λ over the interval $[0, \lambda_{max}]$. In Step 1 (resp. Step 2), we set $\lambda_{\max} = \lambda^*$ (resp. $\lambda_{\max} = \lambda_a^*$), where λ^* and λ_a^* are defined in Equation (10).

In Step 1, we perform the penalized estimation of the objective function in Equation (6) using the NLShoot algorithm. In this step, only ξ_i and c_i are penalized, while ϖ_i remains unpenalized. As a result, the procedure discriminates between time-varying and constant betas, but does not allow for the identification of irrelevant regressors.

Figure S7 reports the BIC values associated with the terminal models obtained from Post-NLShoot (i.e., constrained QMLE) over a grid of λ values between 0 and 0.250. A total of seven distinct terminal models (i.e., sets of active parameters) are identified. Values of λ in the range $[0.02, 0.04]$ yield the same active set and attain the lowest BIC.

In practice, both in the simulation study and in the empirical applications, we refine the selection by constructing a second grid centered around the preliminary optimal value and repeating the procedure.

The regularization paths for the six conditional betas are displayed in Figure S8. The values of the penalty parameter that minimize the BIC for the PACB model are highlighted by green circles. For the coefficients associated with constant betas (i.e., $\beta_{3,t}, \ldots, \beta_{6,t}$), the score and autoregressive parameters, ξ_i and c_i, are shrunk to zero over most of the penalty grid. In contrast, these parameters remain nonzero over a substantial range of λ values for the truly time-varying betas (i.e., $\beta_{1,t}$ and $\beta_{2,t}$).

Importantly, for no value of λ does an autoregressive parameter c_i remain active when the corresponding score parameter ξ_i is inactive, ensuring that only identified model specifications are considered. The intercept ϖ_i is not penalized at this stage and behaves accordingly: it remains small when ξ_i and c_i are active, but increases toward one when they are shrunk to zero. For $\beta_{3,t}$ and $\beta_{4,t}$, the intercept is close to one, whereas it is close to zero for $\beta_{5,t}$ and $\beta_{6,t}$.

In Step 2, we further distinguish between time-varying, constant, and zero betas by additionally penalizing the intercept ϖ_i whenever c_i is inactive in Step 1; see Section 4.3. Figure S9 reports the BIC as a function of λ. In this step, only four distinct terminal models are identified, and the minimum BIC is attained for λ values in the range $[0.03, 0.06]$. Once again, a second lambda grid could be defined around these optimal values, but we have not done so in this illustration.

The corresponding penalized coefficient paths are shown in Figure S10. In this step, all three parameters are penalized for $\beta_{3,t}$–$\beta_{6,t}$, whereas no additional penalization is applied to $\beta_{1,t}$ and $\beta_{2,t}$ when c_i remains active. Among the constant betas, the procedure successfully distinguishes between nonzero constants ($\beta_{3,t}$ and $\beta_{4,t}$) and zero betas ($\beta_{5,t}$ and $\beta_{6,t}$), with the latter being set to zero as soon as $\lambda > 0$. For the selected penalty level, the correct terminal model is recovered, comprising two time-varying betas, two constant betas equal to one, and two zero betas.

In summary, the grid-search procedure for selecting the penalty parameter performs well in our setting. The combination of an initial coarse grid to identify a relevant region, followed by a refined grid, delivers reliable results while avoiding the need for excessively fine global grids and thereby reducing computational cost.

G Monte Carlo simulation results

This section provides additional details on the Monte Carlo experiments. We first report, in Table S2, the percentage of correctly classified conditional betas across all simulation designs described in Section 6. The first three columns indicate the number of time-varying betas (p_{tv}), constant and nonzero betas (p_{cst}), and zero betas (p_{irr}).

Several patterns emerge. First, the detection rate of constant betas (both nonzero and zero) is very high across all designs. Misclassification of constant betas as fully time-varying

is rare. When errors occur for nonzero constants, they typically correspond to cases where the score parameter ξ_i remains active while the autoregressive parameter c_i is inactive, so that the identification issue is effectively mitigated. Second, in Step 2, the procedure reliably distinguishes between nonzero and zero constant betas, as reflected in the high detection rate of irrelevant regressors. Third, the method occasionally over-shrinks, with correct classification rates for time-varying betas between 80% and 90%. A finer grid for the penalty parameter would further improve these results.

We next evaluate how well the PACB and full ACB models capture the dynamics of the conditional betas. Table S3 reports the ratio of RMSEs for the conditional betas under full ACB relative to PACB across all simulation designs. Column *All* reports averages across all betas, while columns *TV* and *CST* focus on time-varying and constant betas, respectively. All reported ratios exceed one, indicating that PACB achieves lower RMSEs in all cases. Importantly, this remains true even for time-varying betas. Although full ACB treats all betas as time-varying, the failure to detect constant and zero coefficients leads to misspecification, which in turn deteriorates the estimation of genuinely time-varying betas. As expected, the performance gap is particularly pronounced for constant betas.

We further report detailed results on bias and RMSE for the estimated parameters underlying the conditional betas in Tables S4–S8. The results are presented separately for time-varying and constant betas. The first set of tables focuses on correctly selected models, while Table S8 considers cases in which PACB fails to detect a constant beta. The main findings are as follows. When the true beta is time-varying (Tables S4 and S6), PACB and full ACB exhibit similar bias and RMSE for the dynamic parameters, conditional on correct selection. When the true beta is constant (Tables S5 and S7), PACB substantially outperforms full ACB, which is misspecified and therefore inconsistent in this setting. Table S8 documents the performance loss when PACB incorrectly classifies a constant beta as time-varying.

To complement these summary statistics, we analyze the distribution of estimation errors. Figures S11 and S12 present histograms and kernel density estimates of the bias for the parameters governing the beta dynamics. Results are shown for a representative design with $p_{tv} = 2$, $p_{cst} = 2$, and $p_{irr} = 2$, aggregated over replications and coefficients. For time-varying betas, the bias distributions under PACB (conditional on correct selection) are centered around zero with moderate dispersion, while full ACB exhibits more pronounced skewness due to misspecification. For constant betas, full ACB produces highly dispersed and unstable estimates, reflecting the underlying identification problem, whereas PACB yields tightly concentrated distributions around zero. When PACB fails to detect constancy, the misspecification typically manifests through a small but active score parameter.

Finally, we illustrate the implications for the estimated beta paths. Figure S15 displays representative trajectories for one beta of each type in the benchmark design. When PACB correctly selects the model, the estimated betas closely track the true underlying processes. In contrast, full ACB produces erratic dynamics for constant and zero betas and exhibits spillover effects on the time-varying betas due to model misspecification. Figure S16 shows a case where PACB incorrectly classifies a time-varying beta as constant. Even in this case, and more generally whenever at least one beta is constant or zero, PACB remains preferable to full ACB in terms of overall performance, as also reflected in the RMSE comparisons.

H Methodological validation using a multi-factor model

In this section, we complement the illustration presented in the introduction by considering daily (excess) returns of 12 US industry portfolios (denoted y_t^j for $j = 1, \ldots, 12$). The data set is spanning a period from January 2000 to December 2024, resulting in 6,225 observations.

Recall that we investigate a multi-factor model, with an intercept, five observed Fama-French risk factors (i.e., MKT, SMB, HML, RMW and CMA) as well as five irrelevant simulated factors $x_{i,t}$ (for $i = 7, \ldots, 11$) generated by the same DCC-GARCH(1,1) model used in Section 6, i.e.,

$$\begin{aligned} y_t^j &= \beta_{int,t}^j + \beta_{MKT,t}^j x_{MKT,t} + \beta_{SMB,t}^j x_{SMB,t} + \beta_{HML,t}^j x_{HML,t} + \beta_{RMW,t}^j x_{RMW,t} \\ &\quad + \beta_{CMA,t}^j x_{CMA,t} + \beta_{7,t}^j x_{7,t} + \cdots + \beta_{11,t}^j x_{11,t} + v_t. \end{aligned}$$

This model has up to eleven time-varying betas following each an ACB specification but for which we expect at least $\beta_{i,t}^j = 0 \forall t$ for $i = 7, \ldots, 11$.

While we do not know what to expect for the conditional betas associated with FF5 and the intercept, the last five betas should be constant and equal to zero. This entails that $\xi_{0i} = 0$ for the last five betas, see Equation (4). In population, the model is not identified and full ACB is an inconsistent estimator of the underlying parameters.

Table S1 presents the parameter estimates obtained with the PACB model using the Post-NLShoot method for all 12 industry portfolios. The main findings can be summarized as follows:

- For ten out of twelve industry portfolios, none of the conditional betas associated with the simulated factors remain active, whereas for two portfolios, one coefficient remains active. The percentage of coefficients set to zero out of all the coefficients known to be zero is 98.9%, suggesting that our proposed Post-NLShoot method works well on real data. The two exceptions occur for industry portfolios $NoDur$ and $Utils$ for which one coefficient remains active. In both cases, the selection in which the respective coefficient is shrunken to zero is not on the penalization path for the grids of the penalization parameter. Therefore, the Post-NLShoot penalized ACB model with "correct" selection was not evaluated. When we enforce the coefficient to be zero,

Table S1: Industry portfolios

		Buseq	Chems	Durbl	Enrgy	Hlth	Manuf	Money	NoDur	Other	Shops	Telcm	Utils
int	ϖ	0.0099											
	ξ					0.0611			0.0260			0.0371	
	c												
MKT	ϖ	0.0053	0.0069	0.0169	0.0060	0.0021	0.0029	0.0083	0.0073	0.0050	0.0046	0.0027	0.0083
	ξ	0.0175	0.0172	0.0124	0.0165	0.0097	0.0106	0.0231	0.0178	0.0121	0.0128	0.0138	0.0294
	c	0.9950	0.9924	0.9858	0.9945	0.9977	0.9973	0.9920	0.9908	0.9952	0.9950	0.9972	0.9888
SMB	ϖ	-0.0001		0.3604	-0.0005		0.0019	-0.0323	-0.0002	0.0004	0.0771	-0.1281	-0.1111
	ξ	0.0031			0.0083	0.0301	0.0136		0.0064	0.0037			
	c	0.9997			0.9929		0.9915		0.9973	0.9971			
HML	ϖ	-0.0013	-0.0561	0.0010	0.0018	-0.0011	0.0004	0.0028	-0.0005	0.0013	-0.0005	0.0003	0.0002
	ξ	0.0129		0.0097	0.0160	0.0118	0.0100	0.0181	0.0119	0.0109	0.0118	0.0142	0.0114
	c	0.9962		0.9950	0.9963	0.9969	0.9964	0.9961	0.9969	0.9899	0.9969	0.9982	0.9992
RMW	ϖ		0.3070	0.0018	-0.0007	-0.0005	0.0023	-0.0002	0.0019	-0.0000	0.0016	-0.0001	0.0005
	ξ	0.0172		0.0148	0.0206	0.0139	0.0133	0.0097	0.0140	0.0033	0.0118	0.0125	0.0115
	c	0.9989		0.9932	0.9985	0.9969	0.9887	0.9991	0.9936	0.9987	0.9953	0.9965	0.9968
CMA	ϖ	-0.0010	0.0047	0.0002	0.0011	0.0020	0.0018	-0.0013	0.0033	0.0003	-0.0001	0.0022	0.0039
	ξ	0.0267	0.0202	0.0105	0.0192	0.0230	0.0164	0.0235	0.0174	0.0138	0.0098	0.0172	0.0176
	c	0.9982	0.9897	0.9969	0.9977	0.9969	0.9899	0.9944	0.9934	0.9944	0.9976	0.9932	0.9915
x_1	ϖ												
	ξ												
	c												
x_2	ϖ												
	ξ												
	c												
x_3	ϖ								0.0235				
	ξ												
	c												
x_4	ϖ												
	ξ												
	c												
x_5	ϖ												
	ξ												-0.0223
	c												
	ω	0.0019	0.0020	0.0040	0.0042	0.0042	0.0011	0.0009	0.0013	0.0007	0.0010	0.0013	0.0053
	α	0.0613	0.0508	0.0558	0.0575	0.0746	0.0465	0.0522	0.0429	0.0444	0.0403	0.0330	0.0723
	β	0.9269	0.9425	0.9414	0.9397	0.9131	0.9473	0.9426	0.9511	0.9505	0.9563	0.9642	0.9201

Note: This table reports the parameter estimates of our penalized ACB approach given by Post NLShoot QMLE for the 12 industry portfolios. The factors model consists of an intercept, the five Fama-French (2015) factors, MKT, SMB, HML, RMW and CMA as described in Section 2 and the five simulated factors x_1 to x_5 estimated on daily data spanning from Jan 2000 to Dec 2024. If the entry in a cell is zero, the parameter is shrunken to zero.

the BIC in both cases decreases (*NoDur* PACB: 1.1811 vs. when $\varpi_3 = 0$ enforced: 1.1803 and *Utils* PACB: 2.2164 vs. when $\xi_5 = 0$ enforced: 2.1255).

- There is clear empirical evidence that our model performs better than a GARCH model with constant slopes or a full ACB model in terms of the BIC for all industry portfolios. We also estimated the FF5-model (without the five simulated factors) using a GARCH model with constant slopes and dynamics in all (six) betas and compared them with the results of the PACB model. We find PACB to be preferred in terms of the BIC in all cases (the results are not presented here in the interests of saving space). Hence, there is clear evidence that some of the conditional betas are

in fact time-varying but allowing all of the betas to be time-varying results in worse specifications.

- For all test assets, if a beta is selected to be time-varying, the associated estimates of the autoregressive parameters, c_i, are close to one but smaller than one (stationary filters) and the estimated lagged scaled scores, ξ_i, are close to zero but positive.

- There is evidence that for several industry portfolios the conditional beta associated with the size factor, SMB, is constant. When estimating a GARCH with constant slopes for the pure FF5 factor model, the constant betas associated with SMB is often close to zero and in 1/3 of the cases statistically insignificant.

- For eight out of twelve portfolio, $\beta^j_{int,t} = 0 \ \forall t$ suggesting that their alphas are zero not only unconditionally but also conditionally. The conditional intercept of $Buseq$ is found to be constant but not zero, with a p-value of the ϖ coefficient of 3.2%. For $Hlth$, $NoDur$ and $Telcm$, the intercept (alpha) is zero on average (i.e., $\varpi = 0$) but displays short-term dynamics as $\xi \neq 0$ and $c = 0$. The p-values for ξ_{Hlth}, ξ_{NoDur} and ξ_{Telcm} are respectively $< 0.01\%$, 7.9% and 1.1%.

- For the portfolio $Hlth$, the lagged score coefficient associated to SMB, i.e. ξ_{SMB}, is active and highly significant, whereas the other two coefficients, ϖ_{SMB} and c_{SMB} are not active and therefore set to 0. This result is discussed in Section H in the supplementary material.

Additionally, we provide detailed results on the industry portfolios $Durbl$ (Durable Consumption) and $Hlth$ (Health). The estimated parameters of both methods and their standard errors for $Durbl$ are reported in Table S9 in Section J. Note that the grids of λs using in the NLShoot algorithm contain 50 values and not 25 as in the simulations and that the best specification is chosen by minimizing the BIC criterion.

Figure S17 is the same as Figure 1, except that it is for the $Hlth$ portfolio instead of the $Durbl$ portfolio. As expected the betas of the simulated factors are all set to zero. The betas associated with the MKT, HML, RMW and CMA factors are clearly significant and time-varying. It can also be seen that the dynamics of $\beta^{Hlth}_{SMB,t}$ fluctuate around 0 and are very erratic. This is because our penalized QMLE provides an estimated non-zero value of ξ_{SMB}, whereas the estimates of ϖ_{SMB} and c_{SMB} are exactly zero. However, the significance of the estimated non-zero value of the ξ_{SMB} parameter is questionable. A comparison of the BIC criteria with and without the factor SMB favors keeping SMB. Moreover, Figure S18 shows that $\beta^{Hlth}_{SMB,t}$ has some very large spikes on days when both $x_{SMB,t}$ and the excess return of the $Hlth$ portfolio display outliers, suggesting that SMB carries relevant information for $Hlth$.

In summary, the empirical results show that our Post-NLShoot penalized ACB model is able to consistently select relevant regressors and shrink the conditional betas associated with irrelevant regressors to zero. The selected models are superior to a GARCH model with constant slopes and to a full ACB model estimated for this linear factor model. The full ACB model likely leads to non-identified models and, therefore, to inconsistent parameter estimates. Moreover, there is empirical evidence that among the five observed Fama-French factors some of the conditional betas should be included but estimated as constant. These features of PACB model are of even more importance if the relevance of the regressors is difficult to assess for a researcher. In the linear asset pricing context, this would be the case when less established "anomaly" factors are introduced in the model.

I Figures

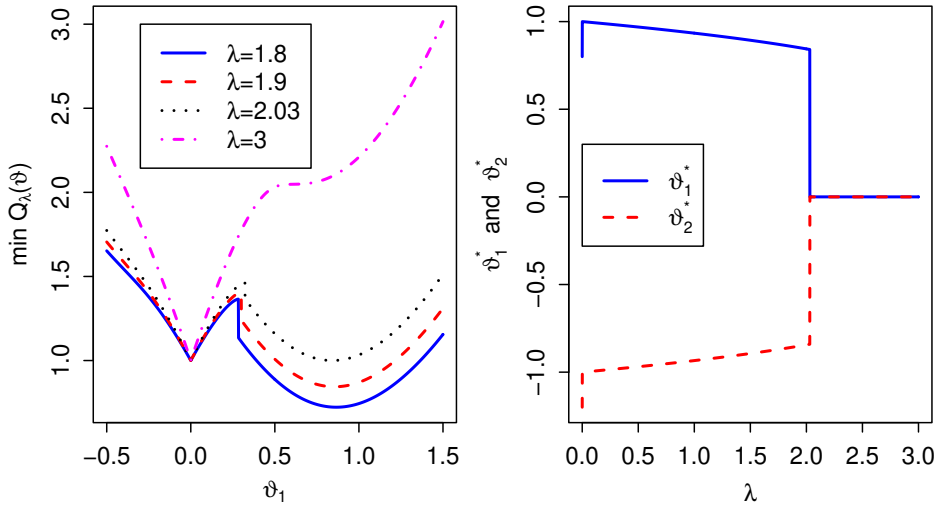

Figure S1: Left graph: $\min Q_\lambda(\boldsymbol{\vartheta})$ as function of ϑ_1 for several values of λ. Right graph: $(\vartheta_1^\star, \vartheta_2^\star) = \arg\min Q_\lambda(\boldsymbol{\vartheta})$ as function of λ.

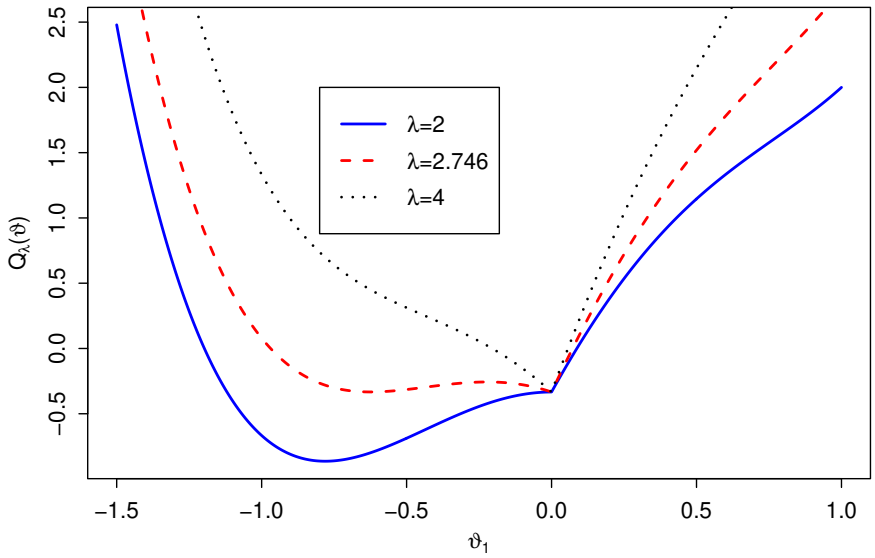

Figure S2: Function $\vartheta_1 \mapsto Q_\lambda(\vartheta_1, -1) = Q(\vartheta_1, -1) + \lambda|\vartheta_1|$. For $\boldsymbol{\vartheta}^c = (0, -1)$, we have $\lambda^i = \partial Q(\boldsymbol{\vartheta}^c)/\partial\vartheta_1 = 2$, but $\arg\min_{\boldsymbol{\vartheta}} Q_\lambda(\boldsymbol{\vartheta}) = \boldsymbol{\vartheta}^c$ only for $\lambda \geq 2.746$ and not for $\lambda = 2$.

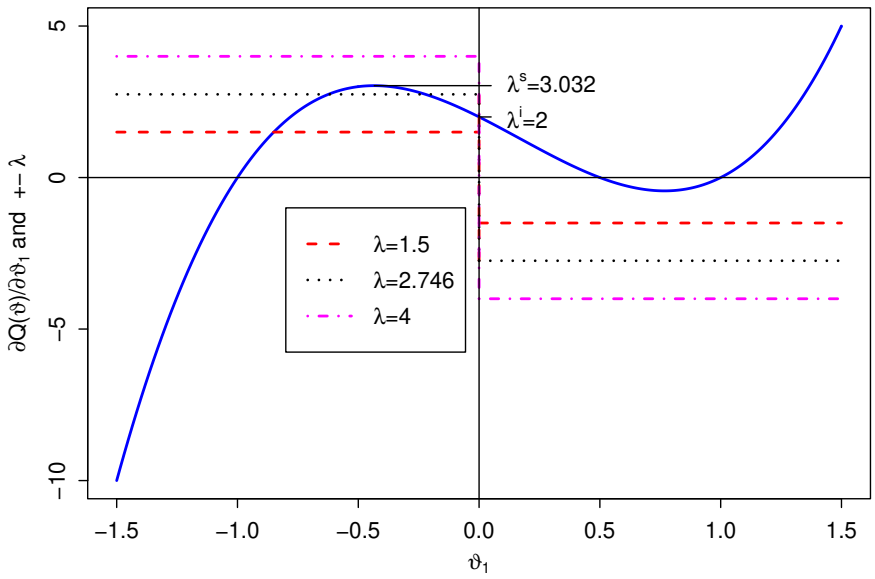

Figure S3: Graphical representation of λ^* and λ^s.

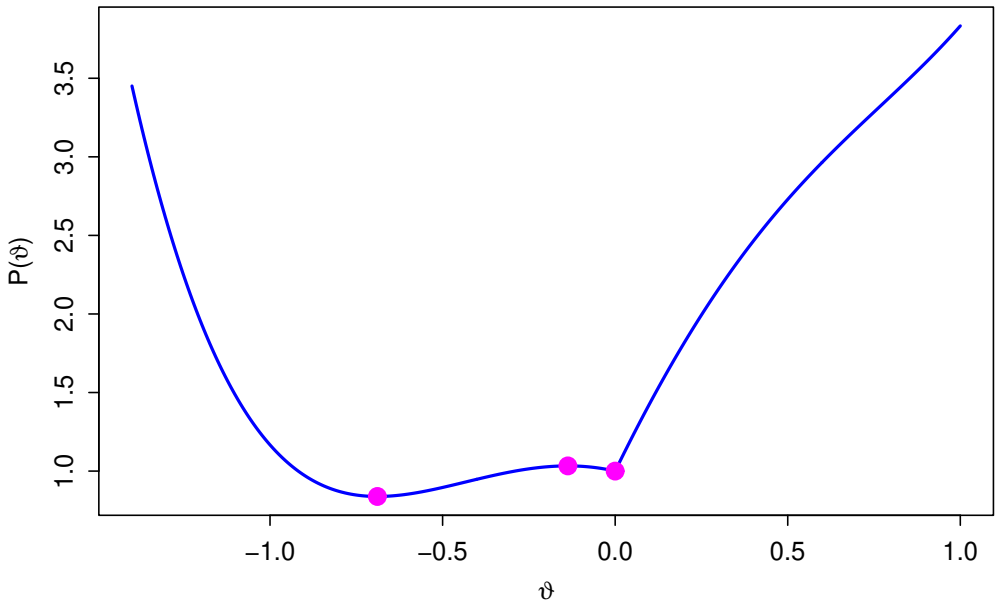

Figure S4: Function $P(\vartheta) + 2.5|\vartheta|$ with a minimum at $\vartheta = -0.689$, and two other critical points at $\vartheta = -0.137$ and $\vartheta = 0$.

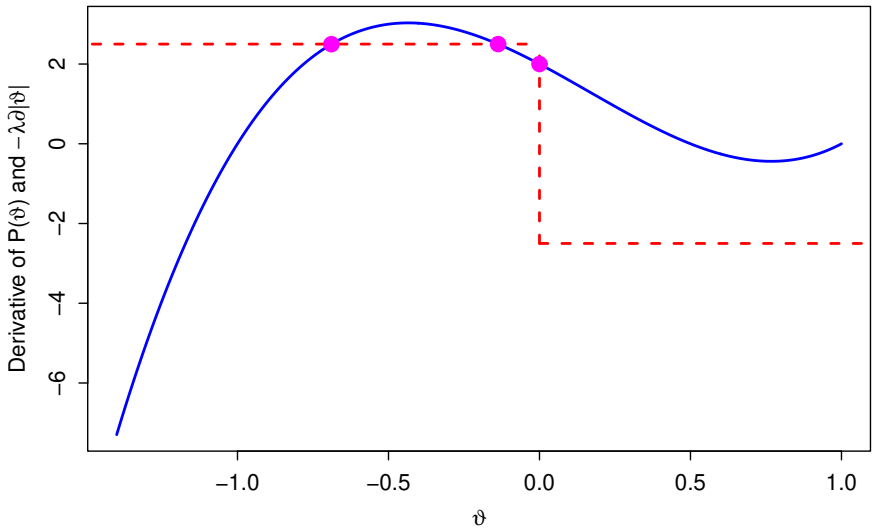

Figure S5: NLShoot finds the 3 critical points.

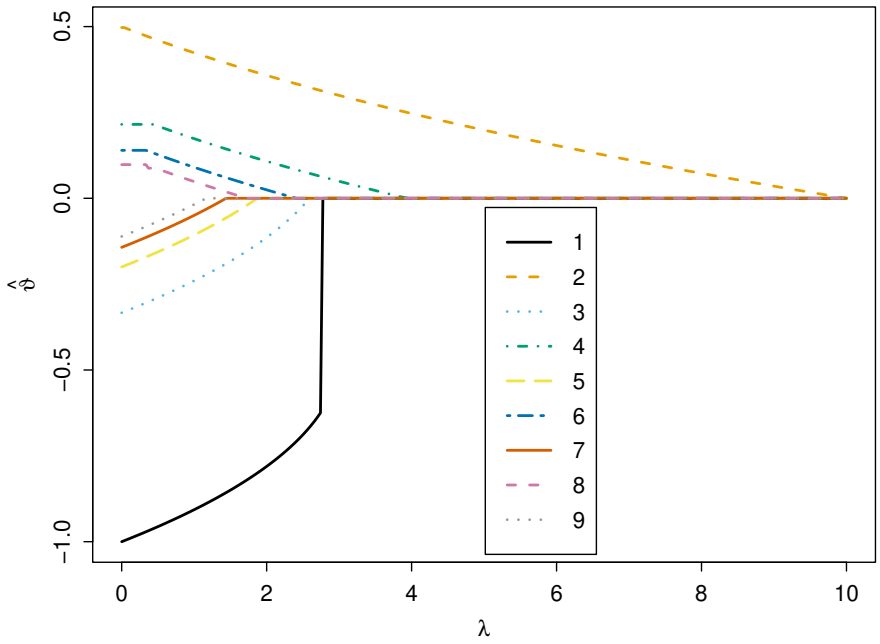

Figure S6: Function $\lambda \mapsto \arg\min_{\boldsymbol{\vartheta}} Q_\lambda(\boldsymbol{\vartheta})$ obtained by NLShoot.

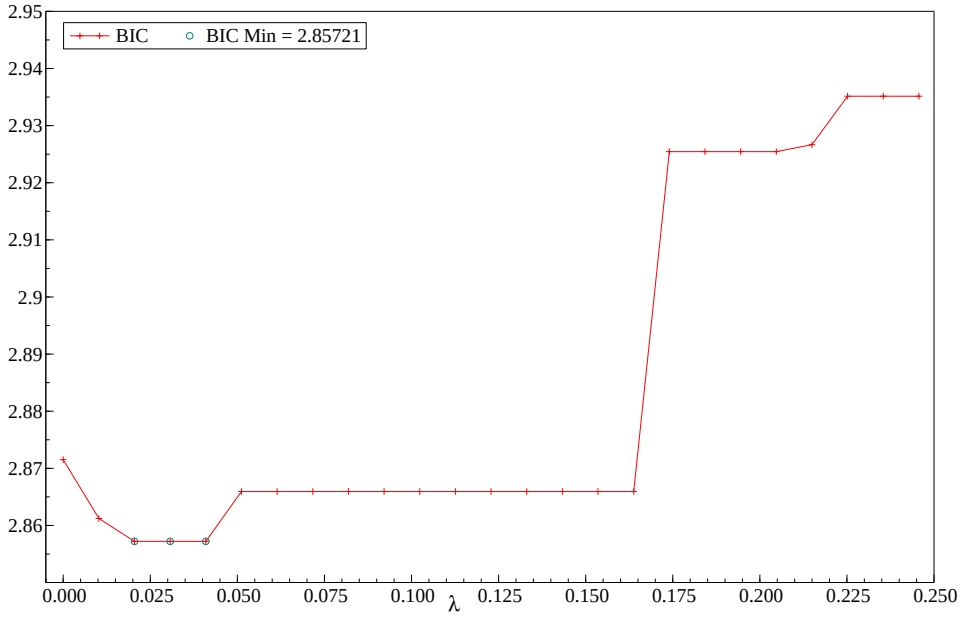

Figure S7: This graph shows the information criterion BIC of the PACB model (Post-NLShoot) for Step 1 for an equidistant grid of 25 penalty parameters between 0 and λ^*.

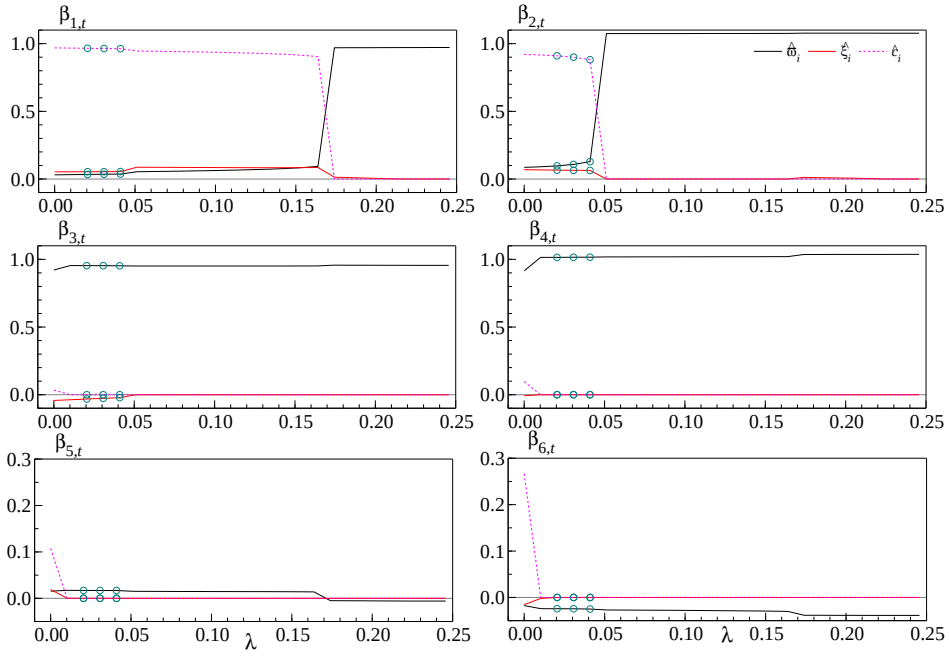

Figure S8: This graph shows the penalized coefficient path for Step 1 (only ξ_i and c_i are penalized) for an equidistant grid of 25 penalty parameters between 0 and λ^*.

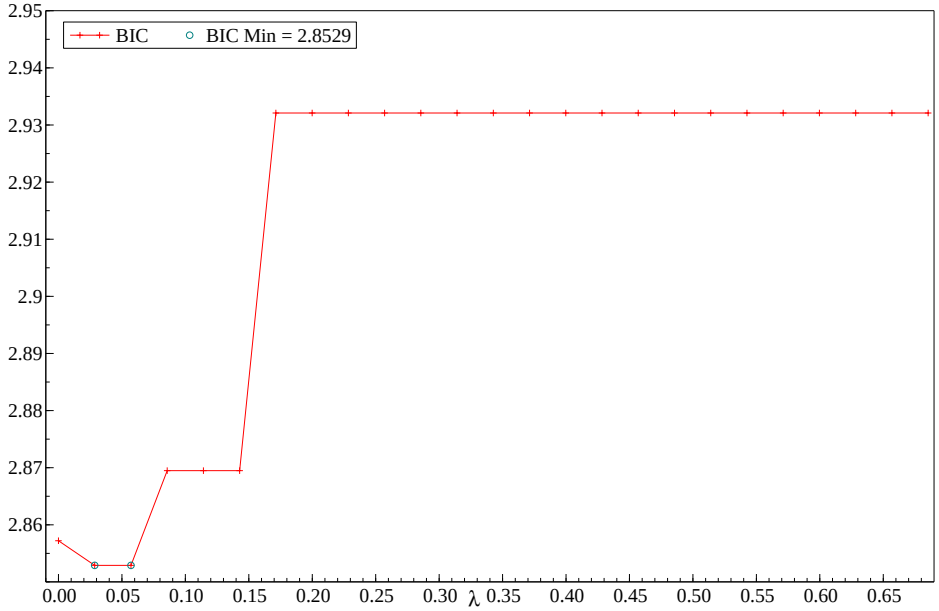

Figure S9: This graph shows the information criterion BIC of the PACB model (Post-NLShoot) for Step 2 for an equidistant grid of 25 penalty parameters between 0 and λ_a^*.

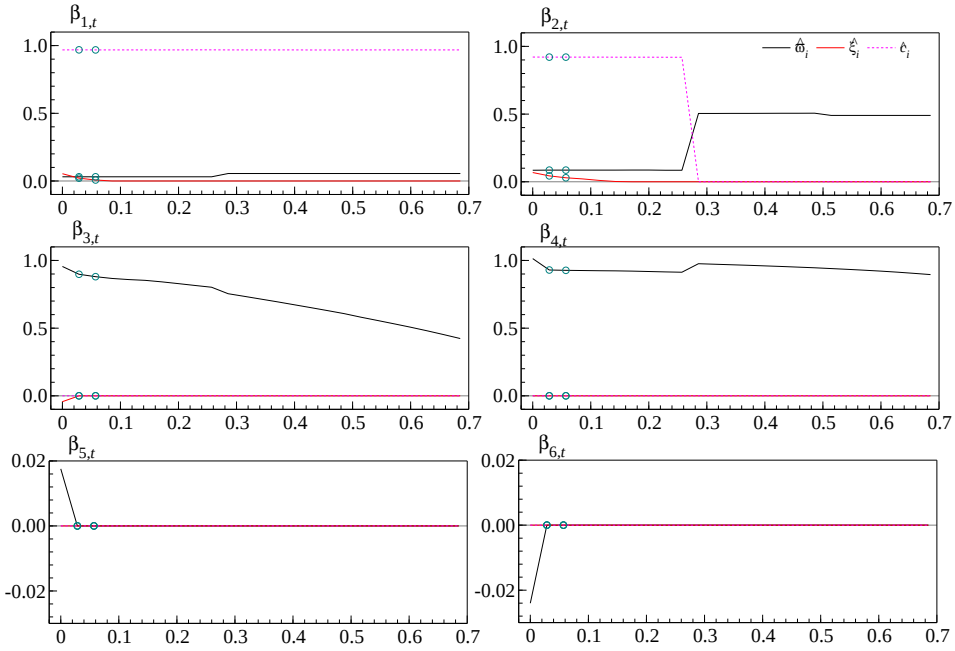

Figure S10: This graph shows the penalized coefficient path for Step 2 (if $c_i = 0$ in Step 1 for $\beta_{i,t}$, all three coefficients are penalized) for an equidistant grid of 25 penalty parameters between 0 and λ_a^*.

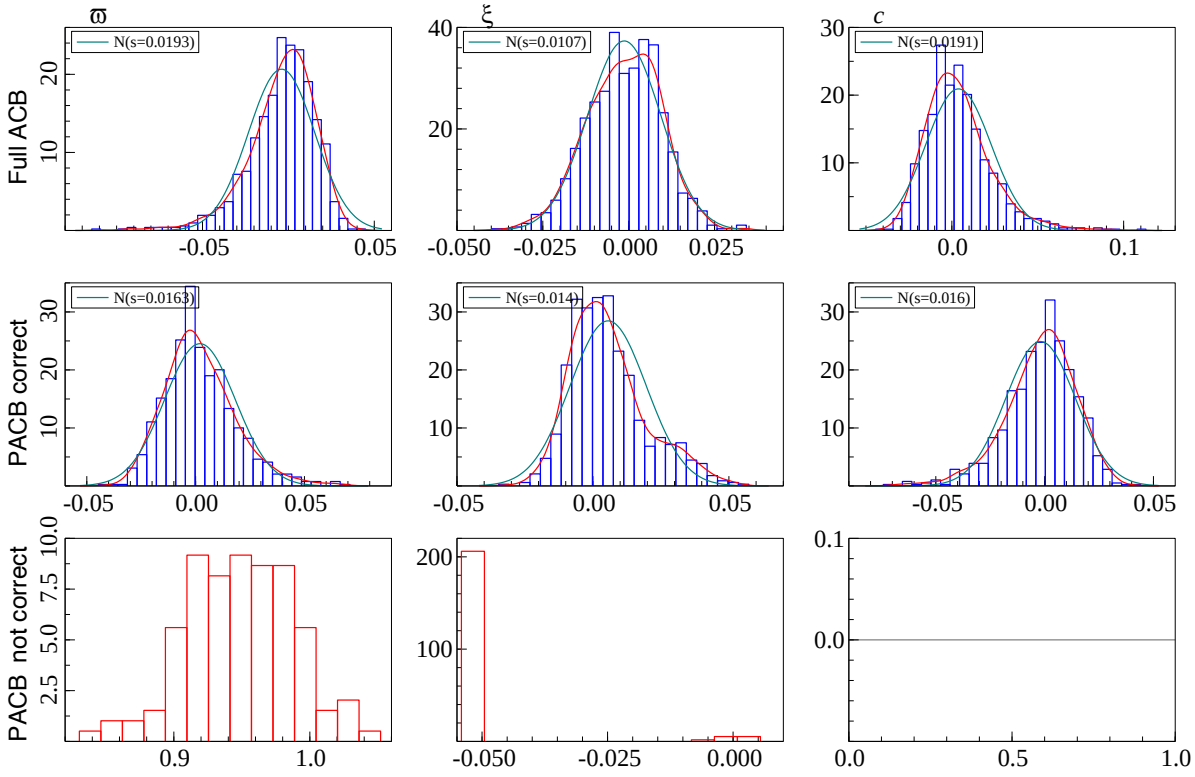

Figure S11: This graph shows the bias of the parameter estimates, ϖ_i, ξ_i and c_i, aggregated over all (true) time-varying conditional betas for the case of p_{tv} = 2 time-varying betas and 500 replications. The true parameter values are given by ϖ_i = 0.05, ξ_i = 0.05 and c_i = 0.95. The first row plots the bias for the model estimated by the full ACB model. The second row includes the bias for the conditional betas (correctly) selected as time-varying using PACB. The last row presents the bias of the conditional betas (incorrectly) selected as constant by PACB.

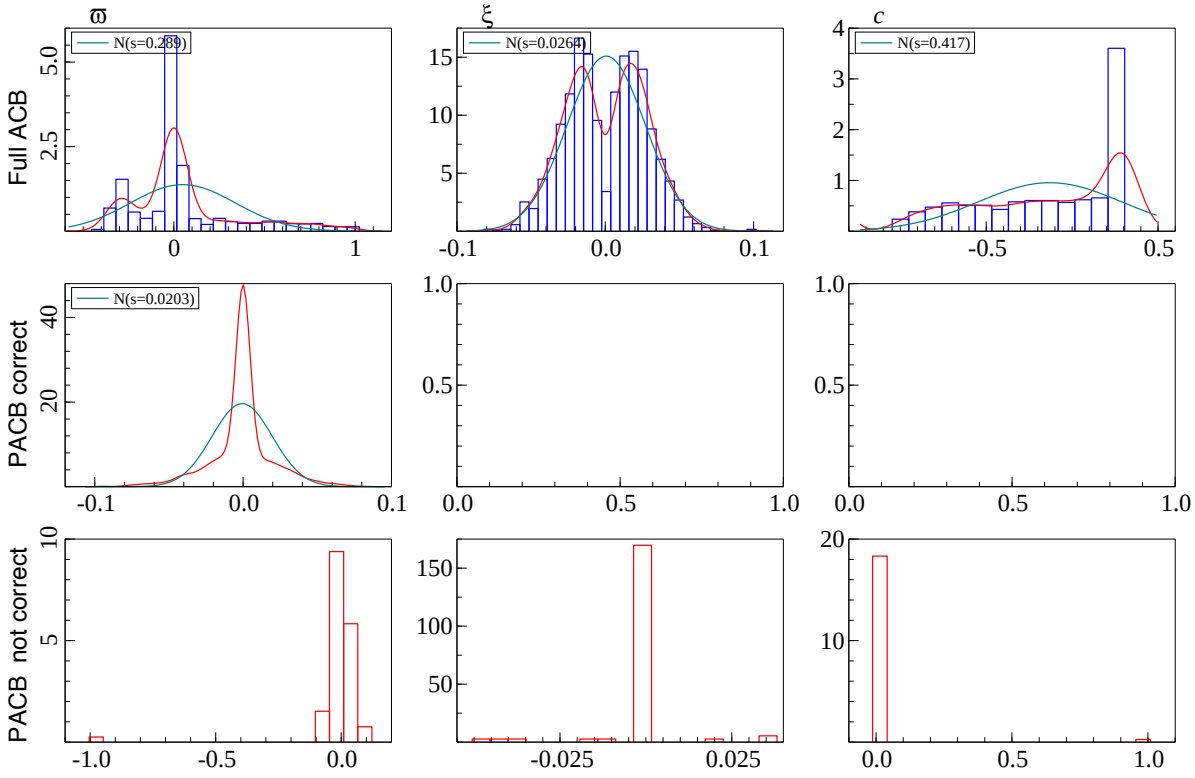

Figure S12: This graph shows the bias of the parameter estimates, ϖ_i, ξ_i and c_i, aggregated over all (true) constant conditional betas for the case of $p_{cst} = 4$ constant betas and 500 replications. The true parameter values are given by $\varpi_i = 1$, $\xi_i = 0$ and $c_i = 0$ for two of the constant parameters and $\varpi_i = 0$, $\xi_i = 0$ and $c_i = 0$ for the other two. The first row plots the bias for the model estimated by the full ACB model. The second row includes the bias for the conditional betas (correctly) selected as constant using PACB. The last row presents the bias of the conditional betas (incorrectly) selected as time-varying by PACB.

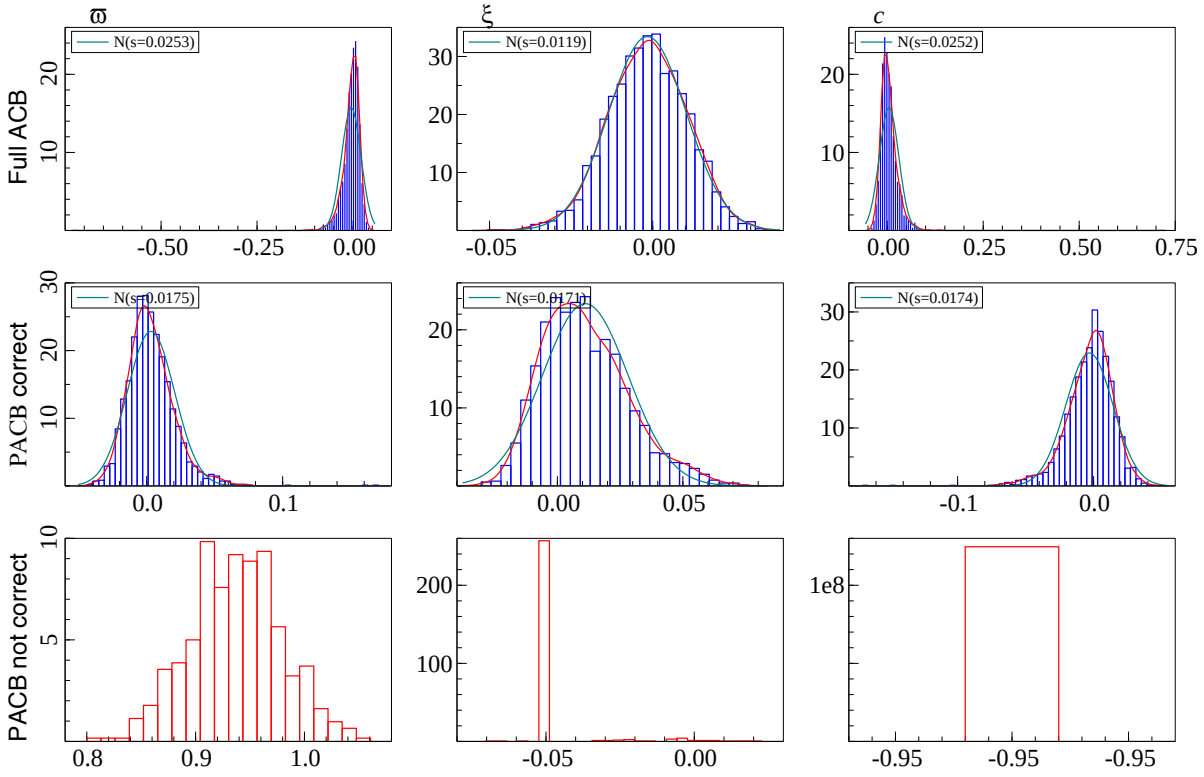

Figure S13: This graph shows the bias of the parameter estimates, ϖ_i, ξ_i and c_i, aggregated over all (true) time-varying conditional betas for the case of p_{tv} = 5 time-varying betas and 500 replications. The true parameter values are given by ϖ_i = 0.05, ξ_i = 0.05 and c_i = 0.95. The first row plots the bias for the model estimated by the full ACB model. The second row includes the bias for the conditional betas (correctly) selected as time-varying using PACB. The last row presents the bias of the conditional betas (incorrectly) selected as constant by PACB.

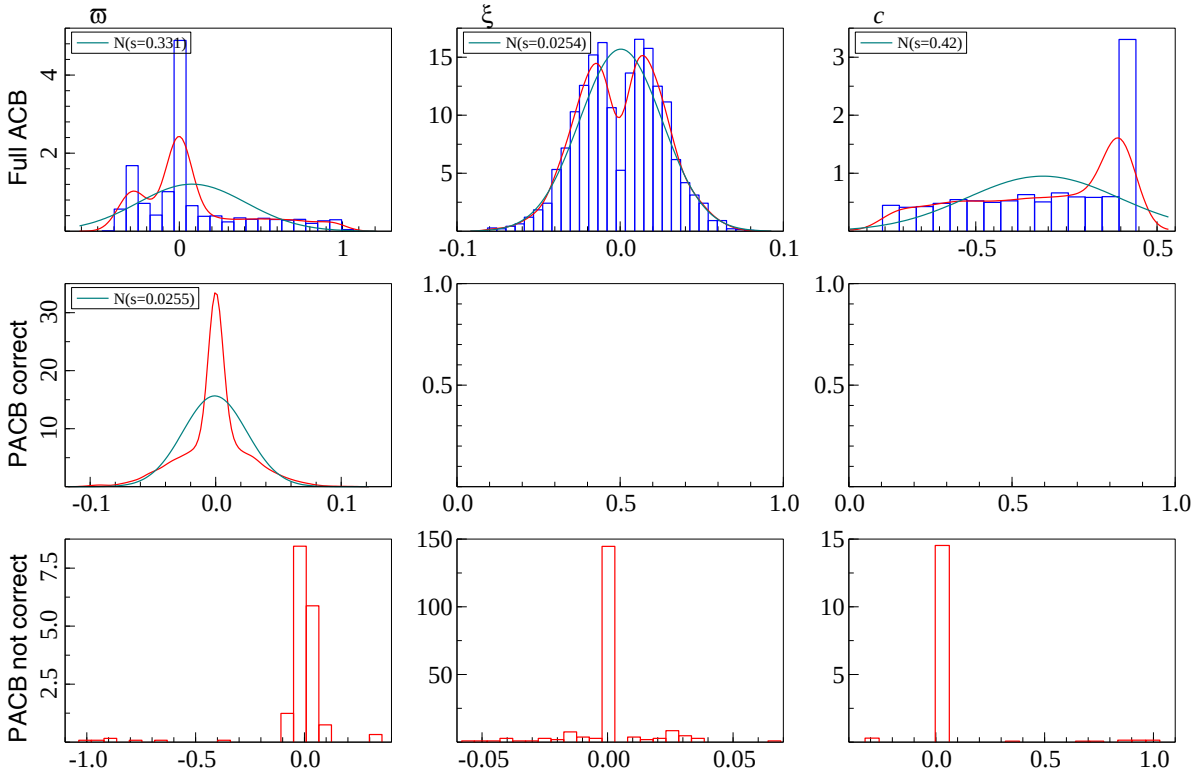

Figure S14: This graph shows the bias of the parameter estimates, ϖ_i, ξ_i and c_i, aggregated over all (true) constant conditional betas for the case of p_{cst} = 5 constant betas and 500 replications. The true parameter values are given by ϖ_i = 1, ξ_i = 0 and c_i = 0 for three of the constant parameters and ϖ_i = 0, ξ_i = 0 and c_i = 0 for the other two. The first row plots the bias for the model estimated by the full ACB model. The second row includes the bias for the conditional betas (correctly) selected as constant using PACB. The last row presents the bias of the conditional betas (incorrectly) selected as time-varying by PACB.

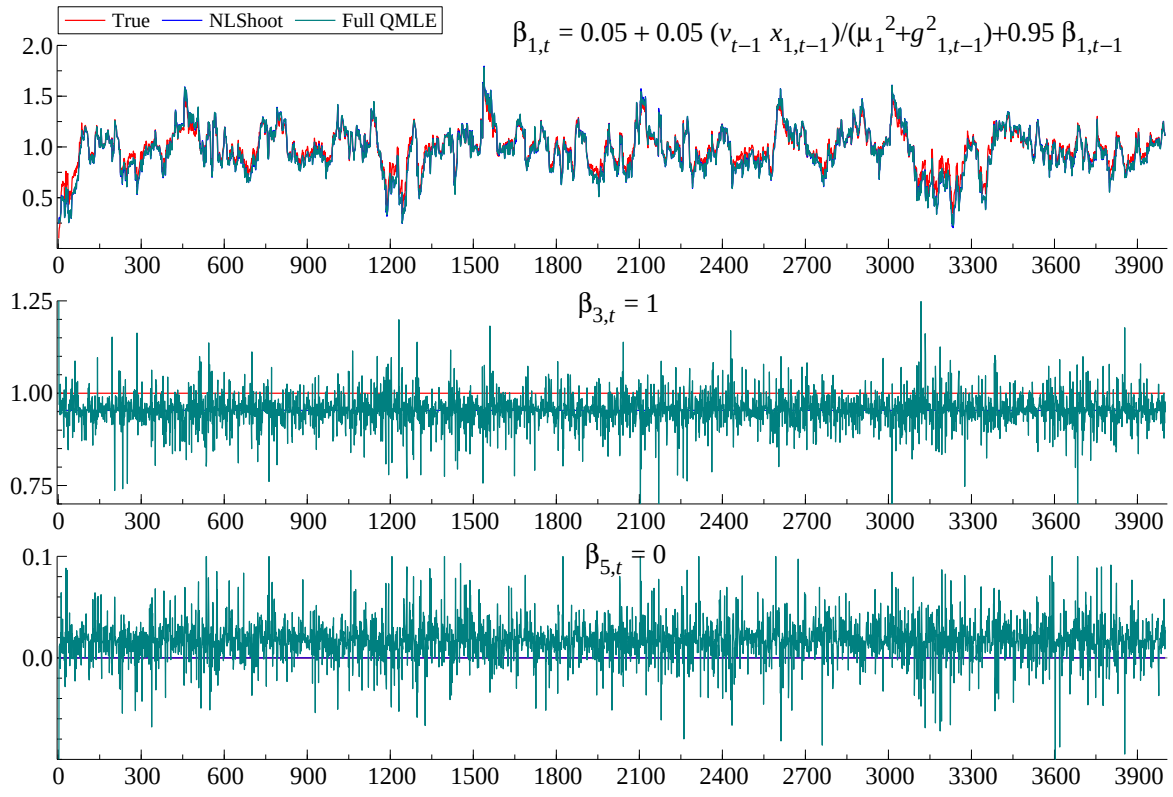

Figure S15: This graph shows three conditional betas obtained from full ACB (turquois) and PACB (blue), along with the true conditional betas for a simulation setting of p_{tv} = 2, p_{cst} = 2 and p_{irr} = 2. It shows one of each type of betas. In this case, PACB has selected the correct model (happens in 93.8% of the simulations, see Table S2). While the conditional betas obtained by PACB are closely aligned with the true conditional betas for all three types, full ACB results in conditional betas that are very rough for the two constant betas.

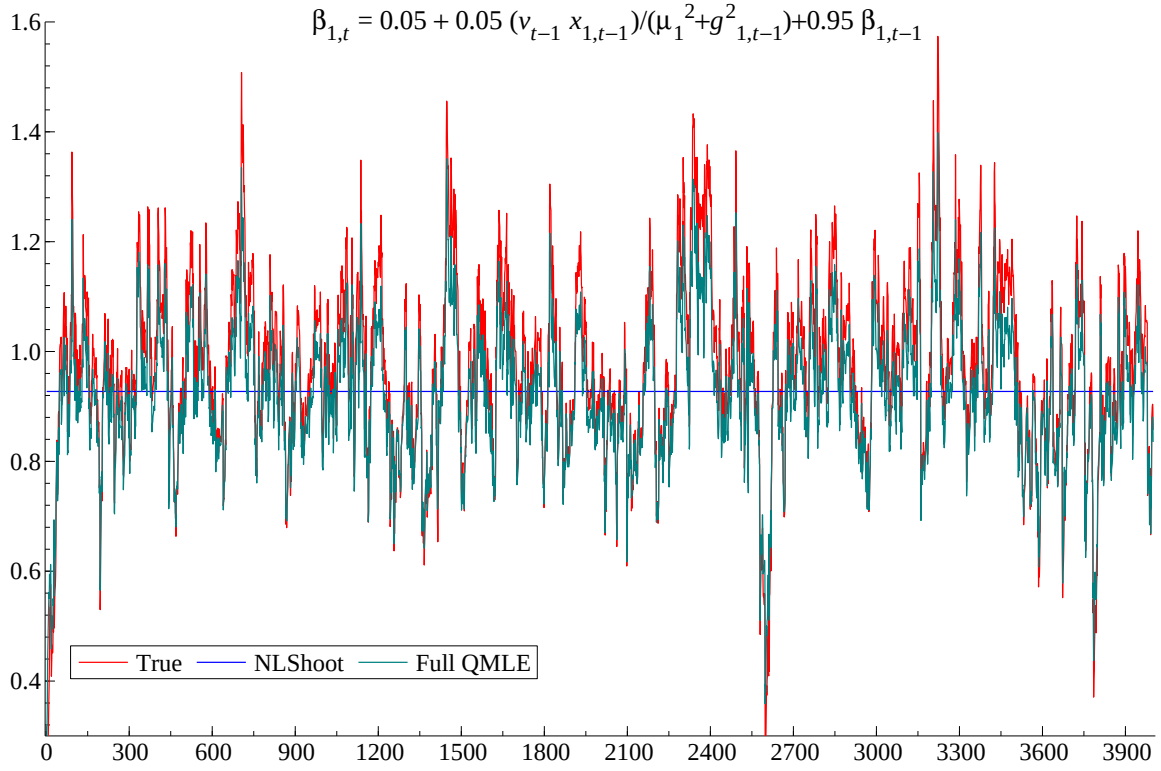

Figure S16: This graph shows a simulation run and one conditional beta for which NLShoot fails to select the true model. The true conditional beta (red) is time-varying but PACB (blue) selects it to be constant for a simulation setting of $p_{tv} = 2$, $p_{cst} = 2$ and $p_{irr} = 2$. In this case, PACB has failed to select the correct model (happens in 12.4% of the simulations, see Table S2).

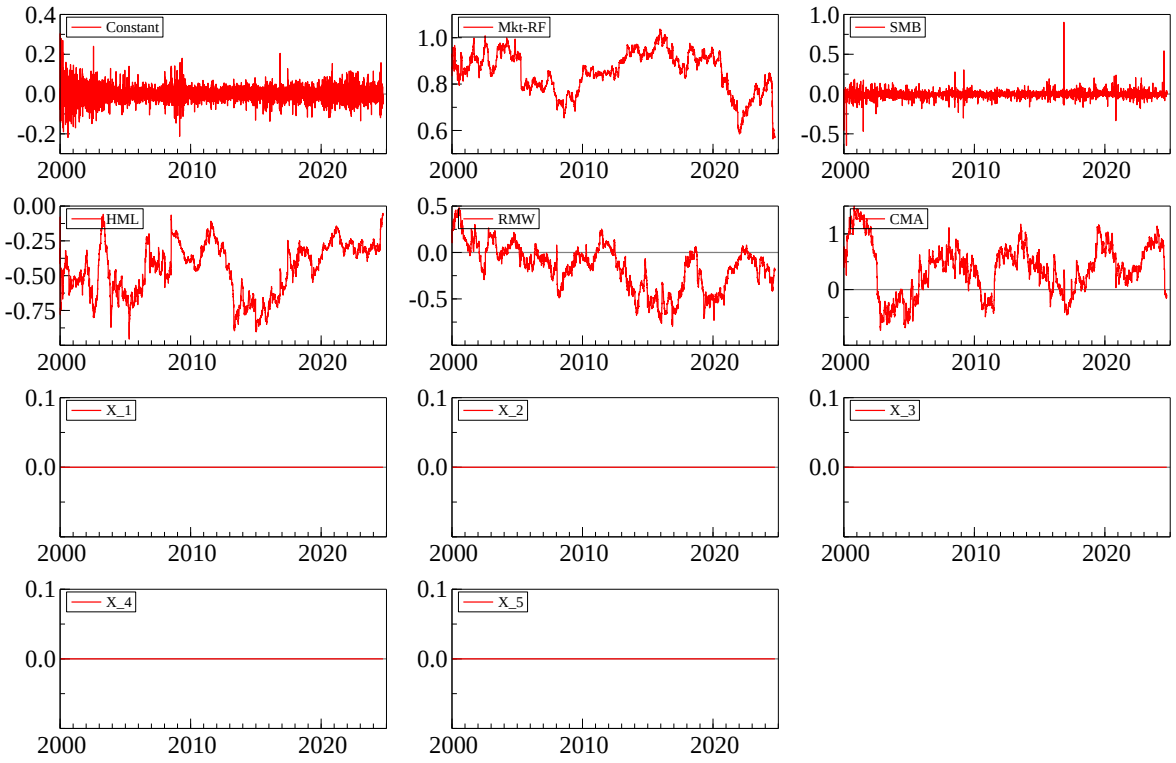

Figure S17: Time series graph of the conditional betas of series *Hlth* for ACB model estimated using the PACB method.

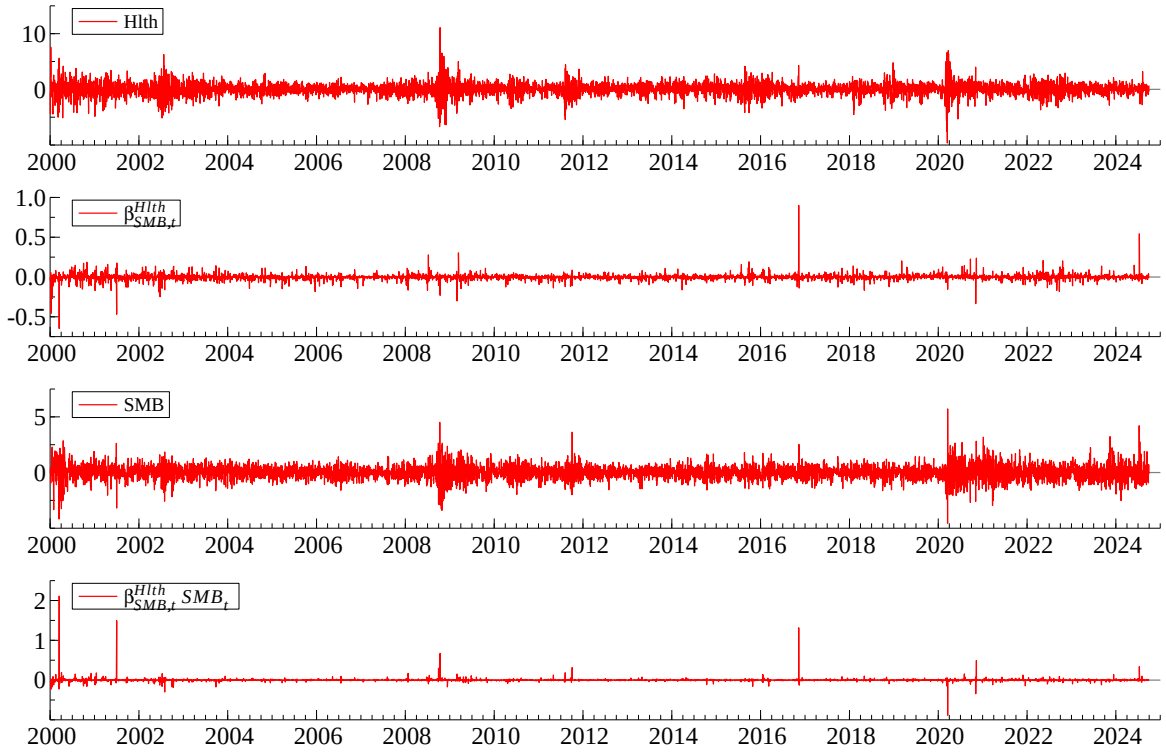

Figure S18: Time series graph of the daily returns of series $Hlth$, $\beta^{Hlth}_{SMB,t}$, SMB and $\beta^{Hlth}_{SMB,t}SMB_t$.

J Tables

Table S2: Percentage of correctly specified conditional betas

p_{tv}	p_{cst}	p_{irr}	$\%\, correct\ \beta_{i,t}$	$\%\, correct\ \beta_i \neq 0$	$\%\, correct\ \beta_i = 0$
3	1	2	86.9	98.6	92.0
4	0	2	90.1		96.5
2	2	2	87.6	99.2	93.8
4	2	2	82.5	96.9	89.0
6	0	2	89.3		97.6
2	4	2	79.2	99.0	91.3
5	3	2	81.0	97.1	83.5
7	1	2	86.0	97.2	91.8
3	5	2	79.2	98.8	88.0

Note: The number of time-varying betas (i.e., p_{tv}), the number of constant and nonzero betas (i.e., p_{cst}) and the number of constant and zero betas (i.e., p_{irr}) are reported in the first three columns. All explanatory variables follow a DCC-GARCH model. Columns '% correct $\beta_{i,t}$', '% correct $\beta_i \neq 0$' and '% correct $\beta_i = 0$' correspond respectively to the percentage of correctly selected time-varying, constant and irrelevant conditional betas.

Table S3: RMSE on the conditional betas: RMSE full ACB/PACB

p_{tv}	p_{cst}	p_{irr}	All	TV	CST
3	1	2	12.313	6.9549	122.90
4	0	2	11.541	8.3502	346.29
2	2	2	17.562	7.0913	88.780
4	2	2	9.579	5.8836	71.971
6	0	2	9.661	7.8816	792.00
2	4	2	14.219	5.9879	60.651
5	3	2	8.620	5.1788	52.124
7	1	2	8.4377	6.5634	112.19
3	5	2	12.368	4.9874	51.667

Note: This table reports the ratio of the RMSE for the conditional betas of full ACB / PACB. The results are separated into all betas (All) and purely the time-varying (TV) and constant (CST) ones. Remarkably, the ratios are above one for simulated cases indicating that PACB has a smaller RMSE on the obtained conditional betas.

Table S4: Bias of time-varying conditional betas - correctly selected by PACB

p_{tv}	p_{cst}	p_{irr}	ϖ	ξ	c	%correct betas
PACB						
3	1	2	0.002	0.007	-0.002	86.867
4	0	2	0.002	0.006	-0.002	90.100
2	2	2	0.002	0.006	-0.002	87.600
4	2	2	0.002	0.010	-0.002	82.450
6	0	2	0.003	0.006	-0.003	89.300
2	4	2	0.001	0.008	-0.001	79.200
5	3	2	0.003	0.011	-0.003	81.000
7	1	2	0.003	0.008	-0.003	86.029
3	5	2	0.002	0.012	-0.002	79.200
full ACB						
3	1	2	0.003	0.001	-0.003	100.000
4	0	2	0.003	0.001	-0.003	100.000
2	2	2	0.004	0.001	-0.004	100.000
4	2	2	0.004	0.001	-0.004	100.000
6	0	2	0.004	0.001	-0.004	100.000
2	4	2	0.003	-0.000	-0.003	100.000
5	3	2	0.005	0.002	-0.005	100.000
7	1	2	0.004	0.001	-0.004	100.000
3	5	2	0.004	0.002	-0.004	100.000

Note: This table reports the average Bias of the estimated parameters (ϖ_i, ξ_i and c_i) of the conditional betas for truly time-varying betas based on PACB (top panel) and full ACB (bottom panel). The table only contains cases, in which the conditional beta is selected as time-varying by the respective method. The number of time-varying betas (i.e., p_{tv}), the number of constant and nonzero betas (i.e., p_{cst}) and the number of constant and zero betas (i.e., p_{irr}) are reported in the first three columns. The last column states the percentage of correctly detected betas.

Table S5: Bias for correctly identified constant conditional betas by PACB

p_{tv}	p_{cst}	p_{irr}	ϖ	ξ	c	%correct betas
PACB						
3	1	2	-0.001	0.000	0.000	94.200
4	0	2	0.000	0.000	0.000	96.500
2	2	2	-0.001	0.000	0.000	96.500
4	2	2	-0.001	0.000	0.000	92.950
6	0	2	0.000	0.000	0.000	97.600
2	4	2	-0.000	0.000	0.000	96.400
5	3	2	-0.001	0.000	0.000	91.680
7	1	2	-0.000	0.000	0.000	93.600
3	5	2	-0.000	0.000	0.000	95.743
full ACB						
3	1	2	-0.039	-0.001	0.146	0.000
4	0	2	-0.000	-0.002	0.140	0.000
2	2	2	-0.050	-0.001	0.129	0.000
4	2	2	-0.065	-0.000	0.136	0.000
6	0	2	-0.001	-0.001	0.125	0.000
2	4	2	-0.086	-0.000	0.135	0.000
5	3	2	-0.078	-0.000	0.129	0.000
7	1	2	-0.031	-0.000	0.109	0.000
3	5	2	-0.094	-0.000	0.132	0.000

Note: This table reports the average Bias of the estimated parameters (ϖ_i, ξ_i and c_i) of the conditional betas for truly constant conditional betas based on PACB (top panel) and full ACB (bottom panel). The table only contains cases, in which the conditional beta is selected as constant by PACB in the top panel. The number of time-varying betas (i.e., p_{tv}), the number of constant and nonzero betas (i.e., p_{cst}) and the number of constant and zero betas (i.e., p_{irr}) are reported in the first three columns. The last column states the percentage of correctly detected betas.

Table S6: RMSE of time-varying conditional betas - correctly selected by NLShoot

p_{tv}	p_{cst}	p_{irr}	ϖ	ξ	c	%correct betas
PACB						
3	1	2	0.017	0.015	0.016	86.867
4	0	2	0.017	0.014	0.017	90.100
2	2	2	0.016	0.015	0.016	87.600
4	2	2	0.017	0.020	0.017	82.450
6	0	2	0.019	0.016	0.019	89.300
2	4	2	0.015	0.018	0.015	79.200
5	3	2	0.018	0.020	0.018	81.000
7	1	2	0.018	0.018	0.018	86.029
3	5	2	0.015	0.020	0.015	79.200
full ACB						
3	1	2	0.020	0.011	0.020	100.000
4	0	2	0.020	0.011	0.020	100.000
2	2	2	0.020	0.011	0.019	100.000
4	2	2	0.024	0.012	0.024	100.000
6	0	2	0.022	0.012	0.022	100.000
2	4	2	0.020	0.011	0.020	100.000
5	3	2	0.026	0.012	0.026	100.000
7	1	2	0.024	0.012	0.024	100.000
3	5	2	0.019	0.011	0.020	100.000

Note: This table reports the average RMSE of the estimated parameters (ϖ_i, ξ_i and c_i) of the conditional betas for truly time-varying betas based on PACB (top panel) and full ACB (bottom panel). The table only contains cases, in which the conditional beta is selected as time-varying by the respective method. The number of time-varying betas (i.e., p_{tv}), the number of constant and nonzero betas (i.e., p_{cst}) and the number of constant and zero betas (i.e., p_{irr}) are reported in the first three columns. The last column states the percentage of correctly detected betas.

Table S7: RMSE of constant conditional betas - correctly selected by NLShoot

p_{tv}	p_{cst}	p_{irr}	ϖ	ξ	c	%correct betas
PACB						
3	1	2	0.016	0.000	0.000	94.200
4	0	2	0.000	0.000	0.000	96.500
2	2	2	0.020	0.000	0.000	96.500
4	2	2	0.022	0.000	0.000	92.950
6	0	2	0.000	0.000	0.000	97.600
2	4	2	0.026	0.000	0.000	96.400
5	3	2	0.025	0.000	0.000	91.680
7	1	2	0.016	0.000	0.000	93.600
3	5	2	0.028	0.000	0.000	95.743
full ACB						
3	1	2	0.249	0.025	0.448	0.000
4	0	2	0.028	0.024	0.444	0.000
2	2	2	0.294	0.026	0.437	0.000
4	2	2	0.309	0.027	0.443	0.000
6	0	2	0.027	0.023	0.436	0.000
2	4	2	0.360	0.027	0.445	0.000
5	3	2	0.340	0.025	0.440	0.000
7	1	2	0.244	0.023	0.426	0.000
3	5	2	0.370	0.026	0.441	0.000

Note: This table reports the average RMSE of the estimated parameters (ϖ_i, ξ_i and c_i) of the conditional betas for truly constant conditional betas based on PACB (top panel) and full ACB (bottom panel). The table only contains cases, in which the conditional beta is selected as constant by PACB in the top panel. The number of time-varying betas (i.e., p_{tv}), the number of constant and nonzero betas (i.e., p_{cst}) and the number of constant and zero betas (i.e., p_{irr}) are reported in the first three columns. The last column states the percentage of correctly detected betas.

Table S8: RMSE for constant conditional betas incorrectly selected by NLShoot

p_{tv}	p_{cst}	p_{irr}	ϖ	ξ	c	% incorrect betas
PACB						
3	1	2	0.034	0.013	0.000	5.800
4	0	2	0.049	0.022	0.152	3.500
2	2	2	0.126	0.012	0.120	3.500
4	2	2	0.169	0.016	0.166	7.050
6	0	2	0.039	0.018	0.000	2.400
2	4	2	0.103	0.018	0.163	3.600
5	3	2	0.161	0.014	0.179	8.320
7	1	2	0.116	0.012	0.180	6.400
3	5	2	0.114	0.016	0.109	4.257
full ACB						
3	1	2	0.249	0.025	0.448	100.000
4	0	2	0.028	0.024	0.444	100.000
2	2	2	0.294	0.026	0.437	100.000
4	2	2	0.309	0.027	0.443	100.000
6	0	2	0.027	0.023	0.436	100.000
2	4	2	0.360	0.027	0.445	100.000
5	3	2	0.340	0.025	0.440	100.000
7	1	2	0.244	0.023	0.426	100.000
3	5	2	0.370	0.026	0.441	100.000

Note: This table reports the average Bias of the estimated parameters (ϖ_i, ξ_i and c_i) of the conditional betas for truly constant conditional betas based on PACB (top panel) and full ACB (bottom panel). The table only contains cases, in which the conditional beta is (incorrectly) selected as time-varying by PACB in the top panel. The number of time-varying betas (i.e., p_{tv}), the number of constant and nonzero betas (i.e., p_{cst}) and the number of constant and zero betas (i.e., p_{irr}) are reported in the first three columns. The last column states the percentage of incorrectly detected betas.

Table S9: full ACB and PACB estimation results for the industry portfolio *Durbl*

		full ACB		PACB	
		$\hat{\boldsymbol{\vartheta}}$	*s.e.*	$\hat{\boldsymbol{\vartheta}}$	*s.e.*
int	ϖ	-0.0010	0.0008		
	ξ	0.0013	0.0037		
	c	0.9642	0.0292		
MKT	ϖ	0.0292	0.0362	0.0169	0.0090
	ξ	0.0147	0.0056	0.0124	0.0042
	c	0.9750	0.0316	0.9858	0.0076
SMB	ϖ	0.0003	0.0002	0.3604	0.0224
	ξ	0.0038	0.0024		
	c	0.9989	0.0021		
HML	ϖ	0.0007	0.0005	0.0010	0.0006
	ξ	0.0069	0.0031	0.0097	0.0037
	c	0.9973	0.0017	0.9950	0.0028
RMW	ϖ	0.0017	0.0022	0.0018	0.0011
	ξ	0.0121	0.0074	0.0148	0.0048
	c	0.9950	0.0056	0.9932	0.0043
CMA	ϖ	0.0002	0.0006	0.0002	0.0005
	ξ	0.0119	0.0037	0.0105	0.0029
	c	0.9967	0.0021	0.9969	0.0019
x_1	ϖ	0.0001	0.0001		
	ξ	-0.0096	0.0049		
	c	0.9911	0.0067		
x_2	ϖ	-0.0002	0.0004		
	ξ	0.0039	0.0069		
	c	0.9618	0.0248		
x_3	ϖ	-0.0000	0.0001		
	ξ	-0.0010	0.0033		
	c	0.9978	0.0046		
x_4	ϖ	0.0000	0.0002		
	ξ	0.0082	0.0050		
	c	0.9932	0.0092		
x_5	ϖ	-0.0002	0.0002		
	ξ	0.0002	0.0041		
	c	0.9882	0.0062		
	ω	0.0039	0.0020	0.0040	0.0021
	α	0.0556	0.0163	0.0558	0.0170
	β	0.9421	0.0167	0.9414	0.0176

Note: This table reports the QMLE results of the ACB model with dynamics on all betas (column 'full ACB') and PACB for the industry portfolio *Durbl*. The model consists eleven variables, i.e., an intercept, the five Fama-French (2015) factors, MKT, SMB, HML, RMW and CMA as described in Section 2 and the five simulated factors x_1 to x_5 estimated on daily data spanning from Jan 2000 to Dec 2024. Columns '$\hat{\boldsymbol{\vartheta}}$' contains the parameter estimates while columns '*s.e.*' contains the standard errors. An empty cell in columns '$\hat{\boldsymbol{\vartheta}}$' and '*s.e.*' means that the corresponding parameter has been set to zero by the NLShoot algorithm.

References

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

Davis, R. A., Knight, K. & Liu, J. (1992), 'M-estimation for autoregressions with infinite variance', *Stochastic Processes and Their Applications* **40**(1), 145–180.

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.

Francq, C. & Zakoïan, J.-M. (2019), *GARCH Models - Structure, Estimation and Finance Applications*, John Wiley.

Van Der Vaart, A. W. & Wellner, J. A. (1996), *Weak convergence*, Springer.

Wu, T. T. & Lange, K. (2008), 'Coordinate descent algorithms for Lasso penalized regression', *Annals of Applied Statistics* **2**(1), 224–244.

Zou, H. & Li, R. (2008), 'One-step sparse estimates in nonconcave penalized likelihood models', *The Annals of Statistics* **36**(4), 1509.