Penalized QMLEs of time series regressions

Christian Francq, Sébastien Laurent and Julie Schnaitmann

ECODEP CONFERENCE

September 30-October 1, 2024, IHP

Framework

We consider a linear regression model applied to the components of a time series, with the aim to

- identify the constant conditional beta coefficients;
- identify the zero conditional betas.

To address the non-identifiability of the parameters when a conditional beta is constant, and to reduce the numerical complexity of the estimation procedure

- we employ a lasso-type* multi-step† QMLE.

*this penalized estimator simplifies the model by shrinking the estimates to their simplest form when the beta is constant.

†which first captures the dynamics of the regressors before estimating the dynamics of the betas.

Outline

Regressions with time-varying betas

Let $(y_t, \boldsymbol{x}_t^\top)$ be a time series of $1+p$ random variables.

Under stationarity and existence of second-order moments, it makes sense to predict y_t at a long horizon by a time-constant mean or regression:

$$Ey_1 \quad \text{or} \quad Ey_1 + \boldsymbol{\beta}^\top(\boldsymbol{x}_t - E\boldsymbol{x}_1).$$

Short-term predictions can incorporate the information $\mathcal{F}_t$ available at time t (given by $\{y_u, \boldsymbol{x}_u; u \leq t\}$ and possibly some vector $\boldsymbol{z}_t$ of exogenous variables): given $\mathcal{F}_{t-1}$, y_t is better predicted by

$$E_{t-1}y_t := E(y_t \mid \mathcal{F}_{t-1}) \quad \text{or} \quad E_{t-1}y_t + \boldsymbol{\beta}_t^\top(\boldsymbol{x}_t - E_{t-1}\boldsymbol{x}_t)$$

with

$$\boldsymbol{\beta}_t = \mathrm{Cov}_{t-1}(y_t, \boldsymbol{x}_t)\mathrm{Var}_{t-1}^{-1}\boldsymbol{x}_t.$$

Regressions with time-varying betas

Let $(y_t, \boldsymbol{x}_t^\top)$ be a time series of $1+p$ random variables.

Under stationarity and existence of second-order moments, it makes sense to predict y_t at a long horizon by a time-constant mean or regression:

$$Ey_1 \quad \text{or} \quad Ey_1 + \boldsymbol{\beta}^\top(\boldsymbol{x}_t - E\boldsymbol{x}_1).$$

Short-term predictions can incorporate the information $\mathcal{F}_t$ available at time t (given by $\{y_u, \boldsymbol{x}_u; u \leq t\}$ and possibly some vector $\boldsymbol{z}_t$ of exogenous variables): given $\mathcal{F}_{t-1}$, y_t is better predicted by

$$E_{t-1}y_t := E(y_t \mid \mathcal{F}_{t-1}) \quad \text{or} \quad E_{t-1}y_t + \boldsymbol{\beta}_t^\top(\boldsymbol{x}_t - E_{t-1}\boldsymbol{x}_t)$$

with

$$\boldsymbol{\beta}_t = \mathrm{Cov}_{t-1}(y_t, \boldsymbol{x}_t)\mathrm{Var}_{t-1}^{-1}\boldsymbol{x}_t.$$

Which time series model for a conditional beta ?

Blasques, Francq and Laurent (2024) (BFL) propose the autoregressive conditional beta (ACB) model

$$\begin{cases} x_{it} & = \mu_{0i} + \varepsilon_{it}, \quad \varepsilon_{it} = g_{it}\eta_{it}, \\ g^2_{i,t+1} & = \omega_{0i} + \alpha_{0i}\varepsilon^2_{it} + \beta_{0i}g^2_{it}, \\ y_t & = \beta_{1t}x_{1t} + \cdots + \beta_{pt}x_{pt} + v_t, \quad v_t = g_t\eta_t, \\ \beta_{i,t+1} & = \varpi_{0i} + \xi_{0i}\frac{v_t x_{it}}{\mu^2_{0i}+g^2_{it}} + c_{0i}\beta_{it} + \gamma_{01i}z_{1t} + \cdots + \gamma_{0qi}z_{qt}, \\ g^2_{t+1} & = \omega_0 + \alpha_0 v^2_t + \beta_0 g^2_t, \end{cases}$$

with obvious notations ($\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})^\top$ are the regressors and $\boldsymbol{z}_t = (z_{1t}, \ldots, z_{qt})^\top$ is a vector of exogenous variables).

▸ More about ACB

Unidentifiability of the constant betas

The beta β_{it} is constant if (and only if, under some regularity conditions)

$$\xi_{0i} = \gamma_{01i} = \cdots = \gamma_{0qi} = 0.$$

Note that, when this relation holds the parameter c_{0i} is not well defined because the model remains the same for all values of (ϖ_{i0}, c_{0i}) such that $\varpi_{i0}/(1 - c_{0i})$ is fixed:

$$\beta_i = \varpi_{0i} + 0 \times \frac{v_t x_{it}}{\mu_{0i}^2 + g_{it}^2} + c_{0i}\beta_i$$

First-step estimator of the regressor GARCH models

The observations are $(y_t, \boldsymbol{x}_t, \boldsymbol{z}_t)$ for $t = 1, \ldots, n$. The unknown parameter is

$$\boldsymbol{\varphi}_0 = (\boldsymbol{\theta}_0^\top, \boldsymbol{\vartheta}_0^\top)^\top, \qquad \underbrace{\boldsymbol{\theta}_0 = \left(\boldsymbol{\theta}_0^{(1)\top}, \ldots, \boldsymbol{\theta}_0^{(p)\top}\right)^\top}_{\text{GARCH parameters of the regressors}} .$$

The regressor GARCH(1,1) models can be estimated in parallel by the standard QMLE $\widehat{\boldsymbol{\theta}}^{(i)}$ involving $(x_{i1}, \ldots, x_{in})$.

Let

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\theta}}^{(1)\top}, \ldots, \widehat{\boldsymbol{\theta}}^{(p)\top}\right)^\top .$$

The remaining parameters

Let

$$\boldsymbol{\vartheta}_0^{(0)} = (\omega_0, \alpha_0, \beta_0)^\top$$

be the GARCH(1,1) parameters of the regression error term. Let

$$\boldsymbol{\vartheta}_0^{(i)} = (\varpi_{0i}, \xi_{0i}, c_{0i}, \gamma_{01i}, \ldots, \gamma_{0qi})^\top$$

the parameters that are specific to β_{it}, for $i \in \{1, \ldots, p\}$. The vector of the remaining parameters is thus

$$\boldsymbol{\vartheta}_0 = \left(\boldsymbol{\vartheta}_0^{(0)\top}, \boldsymbol{\vartheta}_0^{(1)\top}, \ldots, \boldsymbol{\vartheta}_0^{(p)\top}\right)^\top \in \mathbb{R}^{d_2}.$$

Denote by $\boldsymbol{\vartheta}$ be a generic element of the parameter space $\Theta_{\boldsymbol{\vartheta}} \subset (0, \infty) \times [0, \infty)^2 \times \mathbb{R}^{p(3+q)}$ and let the generic parameter

$$\boldsymbol{\varphi} = (\boldsymbol{\theta}^\top, \boldsymbol{\vartheta}^\top)^\top.$$

Second-step estimator of the beta dynamics

We estimate $\boldsymbol{\vartheta}_0$ by

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

where, with standard notation,

$$\widetilde{\ell}_t(\boldsymbol{\varphi}) = \frac{\widetilde{v}_t^2(\boldsymbol{\varphi})}{\widetilde{g}_t^2(\boldsymbol{\varphi})} + \log \widetilde{g}_t^2(\boldsymbol{\varphi}), \qquad \widetilde{v}_t(\boldsymbol{\varphi}) = y_t - \sum_{i=1}^{p} \widetilde{\beta}_{it}(\boldsymbol{\varphi}) x_{it},$$

$$\widetilde{g}_t^2(\boldsymbol{\varphi}) = \omega + \alpha \widetilde{v}_{t-1}^2(\boldsymbol{\varphi}) + \beta \widetilde{g}_{t-1}^2(\boldsymbol{\varphi}),$$

$$\widetilde{\beta}_{it}(\boldsymbol{\varphi}) = \varpi_i + \xi_i \frac{\widetilde{v}_{t-1}(\boldsymbol{\varphi}) x_{i,t-1}}{\mu_i^2 + \widetilde{g}_{i,t-1}^2(\boldsymbol{\theta})} + c_i \widetilde{\beta}_{i,t-1}(\boldsymbol{\varphi}) + \sum_{j=1}^{q} \gamma_{ji} z_{j,t-1}.$$

Motivation for penalized estimators

BFL showed the consistency and asymptotic normality (CAN) of

$$\widehat{\boldsymbol{\varphi}} = (\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\vartheta}}^\top)^\top$$

under the assumption $\xi_{0i} \neq 0$ for $i = 1, \ldots, p$ (and regularity conditions).

We define and study Lasso-type estimators which present the advantages:

- of being CAN under a more general framework than the multi-step QMLE;
- to lead to sparsity of the parameters, and thus to more parsimonious models;
- of being variable selection consistent (impossibility to use BIC for, say $p \geq 5$ which corresponds to more than 32000 models).

The penalized components

We want to penalize non-zero estimated values of the parameters $\xi_i, c_i, \gamma_{1i}, \ldots, \gamma_{qi}$, for $i = 1, \ldots, p$, to solve the non-identifiability problem by favoring the solution

$$\widehat{\boldsymbol{\vartheta}}_n^{(i)} = (\widehat{\varpi}_i, 0, 0, 0, \ldots, 0)^\top$$

if β_{it} is constant.

Let $S = \{5, \ldots, q+6, q+8, \ldots, d_2\}$ be the set of the components that we want to shrink.

Partially penalized estimator for constant beta detection

Thus consider the penalized QMLE

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

where $\lambda_n \geq 0$ and $p(\boldsymbol{\vartheta}) = \sum_{i\in S} |\vartheta_i|$.

The Lasso multi-step QMLE $\widehat{\boldsymbol{\varphi}}_n = \left(\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\vartheta}}_n^\top\right)^\top$ encourages sparsity of specific components of $\widehat{\boldsymbol{\vartheta}}_n$ in order to allow constant $\beta_{it}(\widehat{\boldsymbol{\varphi}}_n)$'s.

Consistency to a biased model

At the price of some bias, the Lasso estimator solves the lack of identifiability when $\lambda_n \to \lambda_0 > 0$ and

A(λ_0)**:**

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

admits a minimum over Θ_ϑ at some unique point $\boldsymbol{\vartheta}^\star$.

Assume standard regularity conditions and $\mathbf{A}(\lambda_0)$ for some $\lambda_0 > 0$. If $\lambda_n \to \lambda_0$, then $\widehat{\boldsymbol{\vartheta}}_n \to \boldsymbol{\vartheta}^\star$ in probability as $n \to \infty$.

Comments on $\mathbf{A}(\lambda_0)$

- Note that $\mathbf{A}(\lambda_0)$ 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.
- It can be shown that $\mathbf{A}(\lambda_0)$ generally holds true, at least when $\lambda_0 > 0$ is sufficiently small.
- Note however that one can not take $\lambda_0 = 0$, that is $\mathbf{A}(0)$ does not hold true when a beta is constant.
- Note also that, in general $\boldsymbol{\vartheta}^\star \neq \boldsymbol{\vartheta}_0$. Thus the penalization introduces an asymptotic bias, but it can be shown that the bias is small when $\lambda_0 > 0$ is small.

▸ Illustration

FOC satisfied by $\boldsymbol{\vartheta}^\star$

Let $\boldsymbol{\partial} p(\boldsymbol{\vartheta}^\star)$ be the set of the subgradients of p on Θ_ϑ at $\boldsymbol{\vartheta}^\star = (\vartheta_1^\star, \ldots, \vartheta_{d_2}^\star)$. When the limit $\boldsymbol{\vartheta}^\star$ belongs to $\mathring{\Theta}_\vartheta$, the interior of Θ_ϑ, it must satisfy the subgradient first-order condition

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

When $E\ell_1(\boldsymbol{\theta}_0, \cdot)$ is convex, $\boldsymbol{\partial}^\circ Q_{\lambda_0}(\boldsymbol{\vartheta})$ is the subdifferential of $Q_{\lambda_0}(\boldsymbol{\vartheta})$ (the FOC is necessary and sufficient).

More generally, Clarke (1975) showed that, since the functions $E\ell_1(\boldsymbol{\theta}_0, \cdot)$ and $p(\cdot)$ are locally Lipschitz, $\boldsymbol{\partial}^\circ Q_{\lambda_0}(\boldsymbol{\vartheta})$ is the set of the generalized gradients of $Q_{\lambda_0}(\boldsymbol{\vartheta})$, which contains its subgradients (the FOC is necessary).

Sparsity of $\boldsymbol{\vartheta}^\star$

The set $\boldsymbol{\partial} p(\boldsymbol{\vartheta})$ consists of the vectors $\boldsymbol{u} = (u_1, \ldots, u_{d_2})^\top$ where:

- for $i \in \overline{S}$, $u_i = 0$;
- for $i \in S$,
  - $u_i = \text{sign}(\vartheta_i)$ when $\vartheta_i \neq 0$,
  - $u_i \in [-1, 1]$ when $\vartheta_i = 0$.

It follows that if $\boldsymbol{\vartheta}^\star \in \mathring{\Theta}_\vartheta$, for all $i \in S$,

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

▸ Estimator's FOC

Upper bound for the penalty term

If $\lambda_n > \overline{\lambda}$ for some sufficiently large $\overline{\lambda}$, then the penalized estimator $\widehat{\boldsymbol{\vartheta}}_n$ is equal to the constrained QMLE $\widehat{\boldsymbol{\varphi}}_n^c = (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\vartheta}}_n^c)$ such that

$$\widehat{\boldsymbol{\vartheta}}_n^c = \underset{\boldsymbol{\vartheta} \in \Theta_\vartheta^c}{\arg\min}\, \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}),$$

where Θ_ϑ^c denotes the set of the parameters $\boldsymbol{\vartheta} \in \Theta_\vartheta$ with i-th element $\vartheta_i = 0$ for all $i \in S$.

The FOC entails that

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

Another bound for $\overline{\lambda}$

Let $\Theta^{j-}_{\boldsymbol{\vartheta}_*}$ (resp. $\Theta^{j+}_{\boldsymbol{\vartheta}_*}$) denotes the set of the parameters $\boldsymbol{\vartheta} \in \Theta_\vartheta$ with j-th element $\vartheta_j \leq 0$ (resp. $\vartheta_j \geq 0$) and the other component of $\boldsymbol{\vartheta}$ are that of $\boldsymbol{\vartheta}_*$. Assume $\widehat{\boldsymbol{\vartheta}}^c_n \in \mathring{\Theta}_\vartheta$.
One can take

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

If $\boldsymbol{\vartheta} \mapsto \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta})$ is a strictly convex, then $\overline{\lambda} = \lambda^i = \lambda^s$. ▸ case $\lambda^i < \lambda^s$

Adaptive estimator for constant beta detection

Let the data-driven weights

$$\widehat{\delta}_i = \frac{1}{\left|\widehat{\vartheta}_{ni}\right|} 1_{\widehat{\vartheta}_{ni}\neq 0} + \infty 1_{\widehat{\vartheta}_{ni}=0} \text{ for } i \in S,$$

and the adaptive penalized QMLE

$$\widehat{\boldsymbol{\vartheta}}_n^a = \underset{\boldsymbol{\vartheta}\in\Theta_\vartheta}{\arg\min}\, \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) + \lambda_n^a \sum_{i\in S} \widehat{\delta}_i |\vartheta_i|.$$

Let $\mathcal{A} = \mathcal{A}(\boldsymbol{\vartheta}_0)$ be the subset of the active (and shrunk) components of the model of parameter $\boldsymbol{\vartheta}_0$, *i.e.* the set of indices $i \in S$ such that $\vartheta_{0i} \neq 0$. Let $\mathcal{I} = S \cap \overline{\mathcal{A}(\boldsymbol{\vartheta}_0)}$ be the subset of the inactive components. Let $\boldsymbol{A}$ be the $d_6 \times d_2$ selector matrix which selects the active (or not shrunk) components of $\boldsymbol{\vartheta}_0$.

Asymptotics of the adaptive penalized multi-step QMLE

Let the previous assumptions and $\mathcal{A}(\boldsymbol{\vartheta}_0) = \mathcal{A}(\boldsymbol{\vartheta}^\star)$ for some $\lambda_0 > 0$. If $\lambda_n \to \lambda_0$, if there exists n_0 such that $\lambda_n^a > 0$ for all $n \geq n_0$, and $\sqrt{n}\lambda_n^a \to \lambda_0^a \geq 0$, then the components of $\widehat{\boldsymbol{\vartheta}}_n^a$ whose indices belong to $\mathcal{I}$ are zero with probability tending to 1, and

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

where, for $\boldsymbol{u} = (u_1, \ldots, u_{d_6})^\top$,

$$V(\boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{W}_2 - \boldsymbol{u}^\top \boldsymbol{J}_{\boldsymbol{\vartheta}\boldsymbol{\theta}}^{\boldsymbol{A}} \boldsymbol{J}_*^{-1} \boldsymbol{W}_1 + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{J}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}} \boldsymbol{u} + \lambda_0^a\, p(\boldsymbol{\delta}, \boldsymbol{\vartheta}_0, \boldsymbol{A}^\top \boldsymbol{u}),$$

with $\left(\boldsymbol{W}_1^\top, \boldsymbol{W}_2^\top\right)^\top \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}^{\boldsymbol{A}})$ and, for $\boldsymbol{u}_2 = (u_1, \ldots, u_{d_2})^\top$,

$$p(\boldsymbol{\delta}, \boldsymbol{\vartheta}_0, \boldsymbol{u}_2) = \sum_{i\in\mathcal{A}} \delta_i u_i \mathsf{sign}\left(\vartheta_{0i}\right).$$

Particular case with oracle property

Choosing a penalty term such that $\lambda_0^a = 0$, we obtain

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

where $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^{\boldsymbol{A}}$ is the asymptotic distribution of the multistep QMLE that we would obtained if we would know which are the constant betas and all the non active components of the model.

After a first-step Lasso variable selection, the post-Lasso method which simply consists of a second-step estimation of the components selected from the first-step Lasso has the same asymptotic distribution.

This kind of property is interpreted as an oracle property by Fan and Li (2001) and Zou (2006), but the interpretation is strongly criticized by Leeb and Pötscher (2008) and Hansen (2016).

Penalizing the constant does not work

The previously defined estimators do not allow to detect the irrelevant betas, *i.e.* the indices i such that $\boldsymbol{\vartheta}_0^{(i)} = \mathbf{0}_{q+3}$.
The naive solution would be to penalize all the $\boldsymbol{\vartheta}^{(i)}$'s coefficients, does not work because a constant beta $\beta_{i,t+1} \equiv \beta$ can be written as

$$\beta_{i,t+1} = \varpi_i + \xi_i \frac{v_t x_{it}}{\mu_i^2 + g_{it}^2} + c_i \beta_{it} + \gamma_{1i} z_{1t} + \cdots + \gamma_{qi} z_{qt},$$

with many possibilities for $\boldsymbol{\vartheta}^{(i)}$, in particular

$$\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_1 := (\beta, \mathbf{0}'_{q+2})' \qquad \text{or} \qquad \boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_2 := (0, 0, 1, \mathbf{0}'_q)'.$$

Note that the solution $\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}_2$ would be favored by the penalized estimator if $|\beta| > 1$, because in this case $\|\boldsymbol{\vartheta}_1\|_1 > \|\boldsymbol{\vartheta}_2\|_1$.
Therefore, penalizing ϖ_i along with the other parameters $\xi_i, c_i, \gamma_{1i}, \ldots, \gamma_{qi}$ is likely to result in an inconsistent estimator.

First detecting the zero c_i's and then the irrelevant betas

Having, in a first step, identified (some of) the ξ_{0i} and c_{0i} that are zero, we obtain the second-step model

$$\begin{cases} \beta_{i,t+1} &= \varpi_{0i} + \gamma_{01i} z_{1t} + \cdots + \gamma_{0qi} z_{qt}, \quad i = 1, \ldots, p^1 \\ \beta_{i,t+1} &= \varpi_{0i} + \xi_{0i} \frac{v_t x_{it}}{\mu_{0i}^2 + g_{it}^2} + \gamma_{01i} z_{1t} + \cdots + \gamma_{0qi} z_{qt}, \\ & \quad i = p^1 + 1, \ldots, p^1 + p^2 \\ \beta_{i,t+1} &= \varpi_{0i} + \xi_{0i} \frac{v_t x_{it}}{\mu_{0i}^2 + g_{it}^2} + c_{0i} \beta_{it} + \gamma_{01i} z_{1t} + \cdots + \gamma_{0qi} z_{qt}, \\ & \quad i = p^1 + p^2 + 1, \ldots, p, \end{cases}$$

for $0 \leq p^1 \leq p^1 + p^2 \leq p$, with obvious convention. It is possible to shrink all the beta coefficients for $i = 1, \ldots, p^1 + p^2$ (but not the ϖ_{0i}'s for $i = p^1 + p^2 + 1, \ldots, p$.)

General optimization problem

Consider the optimization problem

$$\boldsymbol{\vartheta}(\lambda) = \arg\min_{\boldsymbol{\vartheta}\in\Theta} Q_\lambda(\boldsymbol{\vartheta}), \qquad Q_\lambda(\boldsymbol{\vartheta}) = Q(\boldsymbol{\vartheta}) + \lambda p(\boldsymbol{\vartheta}),$$

with $p(\boldsymbol{\vartheta}) = \sum_{i\in S} \delta_i |\vartheta_i|$, where $\lambda \geq 0$, $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d)^\top$, Θ is a convex compact subset of $\mathbb{R}^d$, $\delta_1, \ldots, \delta_d$ are given relative shrinkage coefficients with $\delta_i \geq 0$ and $S = \{i : \delta_i > 0\} \neq \emptyset$.

Assume that $Q(\cdot)$ is two times continuously differentiable but not necessarily convex. The Newton-Raphson method may not work since $Q_\lambda(\cdot)$ is not differentiable everywhere.

Nonlinear shooting (NLShoot) algorithm

▸ Illustration ▸ LQA

For the Lasso-LSE of linear regressions Fu (1998) proposes the "shooting algorithm", which is a coordinate-wise descent algorithm.

We first define set of negative $T_i^-(\boldsymbol{\vartheta})$ and positive $T_i^+(\boldsymbol{\vartheta})$ points, real functions $Q_\lambda^{(i)}(\cdot;\boldsymbol{\vartheta})$, and

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

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

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

We show that the cluster point(s) of this NLShoot algorithm are stationary points of $Q_\lambda(\cdot)$.

Monte-Carlo design

The explanatory $p_{tv} + p_{cst}$ variables follow a DCC-GARCH model. For each replication, the penalized model is evaluated on a grid of 15 equidistant values of λ between 0 and $\overline{\lambda}$, using the nonlinear shooting (NLShoot) algorithm initialized with the LQA algorithm.

The next table reports the number of time-varying betas, of constant betas, the percentage of correctly selected models, of correctly identified time-varying betas, of constant conditional betas, of correctly identified non active parameters and active parameters.

In Step 1, only the ξ and c parameters are penalized. In Step 2 , the intercepts ϖ's of the conditional betas identified as constant or for which the c parameter is 0 in Step 1 are also penalized.

100 replications for sample of $4,000$ observations

More than 1 billion of possible models when $p = 10$

p_{tv}	p_{cst}	% correct p_{tv}	% correct p_{cst}	% correct 0's	% correct !0's
			Step 1		
3	3	93.3	66.0	82.3	94.3
6	2	94.0	45.5	71.7	95.8
2	6	88.4	78.3	88.0	90.6
5	5	92.4	66.4	81.7	93.9
7	3	94.8	52.3	75.0	96.1
3	7	88.0	76.3	86.6	90.0
			Step 2		
3	3	92.6	100.0	97.4	95.9
6	2	93.0	99.5	96.2	94.0
2	6	77.5	99.3	97.5	91.4
5	5	88.8	99.0	95.6	94.0
7	3	92.0	99.3	95.6	95.3
3	7	87.0	97.9	95.6	94.9

The data set

We regress 30 daily Dow Jones stock returns from January 5th, 2010 to June 29th, 2023 (3,394 values) on nine sector ETFs (potentially more than 1 billion of models for each stock).

For example, the BP series, which belongs to the Oil & Gas sector, has a penalized ACB model that includes only 3 ETFs:

- XLB (Material) and XLF (Financial services) with cst betas;
- XLE (Energy) with $(\varpi, \xi, c) = (0.005, 0.009, 0.995)$;
- and an error term with vol $(\omega, \alpha, \beta) = (0.079, 0.089, 0.846)$.

In general, the ETF to which the stock is linked always appears as relevant regressor, and the error term of the regression is always conditionally heteroscedastic.

Conclusion

- The Lasso can be used to solve the non-identifiability of our beta time series model, at the price of some bias.
- The adaptive version is able to suppress the bias.
- The Lasso Multi-Step QMLE is tractable even though the optimization is nonlinear, nonconvex, and involves a large number of parameters.
- The penalized estimator allows to select the constant betas in a first step, and the relevant regressors in a second term.

Thank you!

The autoregressive conditional beta (ACB) model

We focus on simple GARCH models for simplicity, but there are no major conceptual or practical issues in considering more general time series models for the regressors x_{it} and for the error term v_t of the regression.

The term $\frac{v_t x_{it}}{\mu_{0i}^2 + g_{it}^2}$ can be interpreted as an update variable for the dynamics of the betas.
This term has been obtained by using the score-driven (SD) approach ▸ SD approach with the aim to obtain a simple beta model, analogous to an ARMA or a GARCH for the first conditional moments.

Interpretation of the update variable $\frac{v_t x_{it}}{\mu_{0i}^2 + g_{it}^2}$

The presence of the product $v_t x_{it}$ allows dynamic monitoring of the orthogonality condition between the error term of the regression v_t and the regressors x_{it}.
Indeed, if $v_t x_{it} \simeq 0$ and there is no exogenous variable ($q = 0$), then we have the constant solution $\beta_{it} \equiv \beta_i$. Now, if β_{it} is negative and very small, then $v_t x_{it} = y_t x_{it} - \sum_{j \neq i} \beta_{jt} x_{it} x_{jt} - \beta_{it} x_{it}^2$ tends to be positive. Therefore if β_{it} is small and $\xi_i > 0$, then $\beta_{i,t+1}$ tends to be larger than β_{it}.

Also note that since the volatility g_{it}^2 is in the denominator of the update variable, the time variation of the beta is smaller in periods of high volatility in the regressors.

Score Driven (GAS and Beta-t-GARCH) models

Assume that y_t follows a conditional density $p(y_t|f_t, \mathcal{F}_{t-1}, \theta)$, where f_t is a time-varying parameter of interest.

Harvey and Chakravarty (2008) and Creal, Koopman and Lucas (2012) proposed the score-driven (SD) model for f_t:

$$f_{t+1} = \varpi + \xi S(f_t) \frac{\partial \log p(y_t|f_t, \mathcal{F}_{t-1}, \theta)}{\partial f_t} + c f_t,$$

where ϖ, ξ and c are unknown parameters and $S(f_t)$ is the inverse of the conditional information matrix.

$\rightarrow$ The scaled score is the updating mechanism in this approach.

Examples of SD models

The SD model for m_t in the location model

$$y_t = m_t + \eta_t, \quad \eta_t \text{ iid } \mathcal{N}(0, \sigma^2),$$

is the ARMA model

$$m_{t+1} = \varpi + \xi(y_t - m_t) + cm_t.$$

The SD model for g_t^2 in the scale model

$$y_t = g_t \eta_t, \quad \eta_t \text{ iid } \mathcal{N}(0, 1),$$

is the GARCH model

$$g_{t+1}^2 = \varpi + \xi(y_t^2 - g_t^2) + cg_t^2.$$

$\rightarrow$ The SD approach provides benchmark models for Gaussian conditional distributions with location-scale parameters.

Applying the SD approach for the beta parameters

Let us define a SD model for β_{it} in the regression model

$$y_t = \beta_{1t}x_{1t} + \cdots + \beta_{pt}x_{pt} + v_t,$$

where $v_t = g_t\eta_t$, $g_t^2 = E_{t-1}v_t^2$, η_t iid $\mathcal{N}(0,1)$. Let $E_{t-1}x_{it} = \mu_{it}$ and $E_{t-1}x_{it}^2 = g_{it}^2 + \mu_{it}^2$.

We have $l_t := \log p(y_t|\mathcal{F}_{t-1},\theta) = \frac{-1}{2}\left\{\frac{v_t^2}{g_t^2} + \log g_t^2\right\}$,

$$\frac{\partial l_t}{\partial \beta_{it}} = \frac{v_t x_{it}}{g_t^2}, \quad S(\beta_{it}) = -\left(E_{t-1}\frac{\partial^2 l_t}{\partial^2 \beta_{it}}\right)^{-1} = \frac{g_t^2}{\mu_{it}^2 + g_{it}^2}.$$

Therefore the updating mechanism $S(\beta_{it})\frac{\partial l_t}{\partial \beta_{it}} = \frac{v_t x_{it}}{\mu_{it}^2+g_{it}^2}$.

ACB with exogenous variables and special cases

For simplicity, assume $\mu_{it} = \mu_i$ and a GARCH(1,1) volatility g_{it}.
Let the ACB with q additional exogenous variables $z_{1t}, \ldots, z_{qt}$

$$\beta_{it+1} = \varpi_i + \xi_i \frac{v_t x_{it}}{\mu_i^2 + g_{it}^2} + c_i \beta_{it} + \gamma_{1i} z_{1t} + \cdots + \gamma_{qi} z_{qt}.$$

1. No GARCH effects in x_{it}: $\alpha_i = \beta_i = 0 \rightarrow g_{it+1}^2 = \omega_i$.

$$\beta_{it+1} = \varpi_i + \xi_i v_t x_{it} + c_i \beta_{it} + \gamma_{1i} z_{1t} + \cdots + \gamma_{qi} z_{qt}.$$

2. $x_{it} = 1$ (intercept): β_{it} time varying unless $\xi_i = 0$ and $q = 0$

$$\beta_{it+1} = \varpi_i + \xi_i v_t + c_i \beta_{it} + \gamma_{1i} z_{1t} + \cdots + \gamma_{qi} z_{qt}.$$

3. Assume that $p = 1$, $q = 0$ and that $x_{1t} = 1$
$\rightarrow y_t = \varpi + (\xi - c) v_{t-1} + c y_{t-1} + v_t$ =ARMA(1,1).

Illustration of $\mathbf{A}(\lambda_0)$

Let

$$Q_\lambda(\boldsymbol{\vartheta}) = Q(\boldsymbol{\vartheta}) + \lambda \|\boldsymbol{\vartheta}\|_1$$

with $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^\top$,

$$Q(\boldsymbol{\vartheta}) = \mathcal{P}(\vartheta_1 - \vartheta_2), \quad \mathcal{P}(x) = \frac{x^4}{4} - \frac{2x^3}{3} - x^2 + 1.$$

Note that $Q(\boldsymbol{\vartheta})$ is minimal for all $\boldsymbol{\vartheta}$ such that $\vartheta_2 = \vartheta_1 - 2$.

- If $\lambda > 0$ is very small but not zero, $\boldsymbol{\vartheta}^\star = \arg\min Q_\lambda(\boldsymbol{\vartheta})$ is unique and is close to $(1, -1)$;
- if λ is very large $\boldsymbol{\vartheta}^\star$ is also unique and equal to $(0, 0)$;
- if $\lambda = 2.03$ the function reaches its minimum at 2 points
- $\lambda = 0$ the minimum is not well defined (*i.e.* not unique).

$\mathbf{A}(\lambda_0)$ always satisfied, except for $\lambda_0 = 0$ and $\lambda_0 = 2.03$

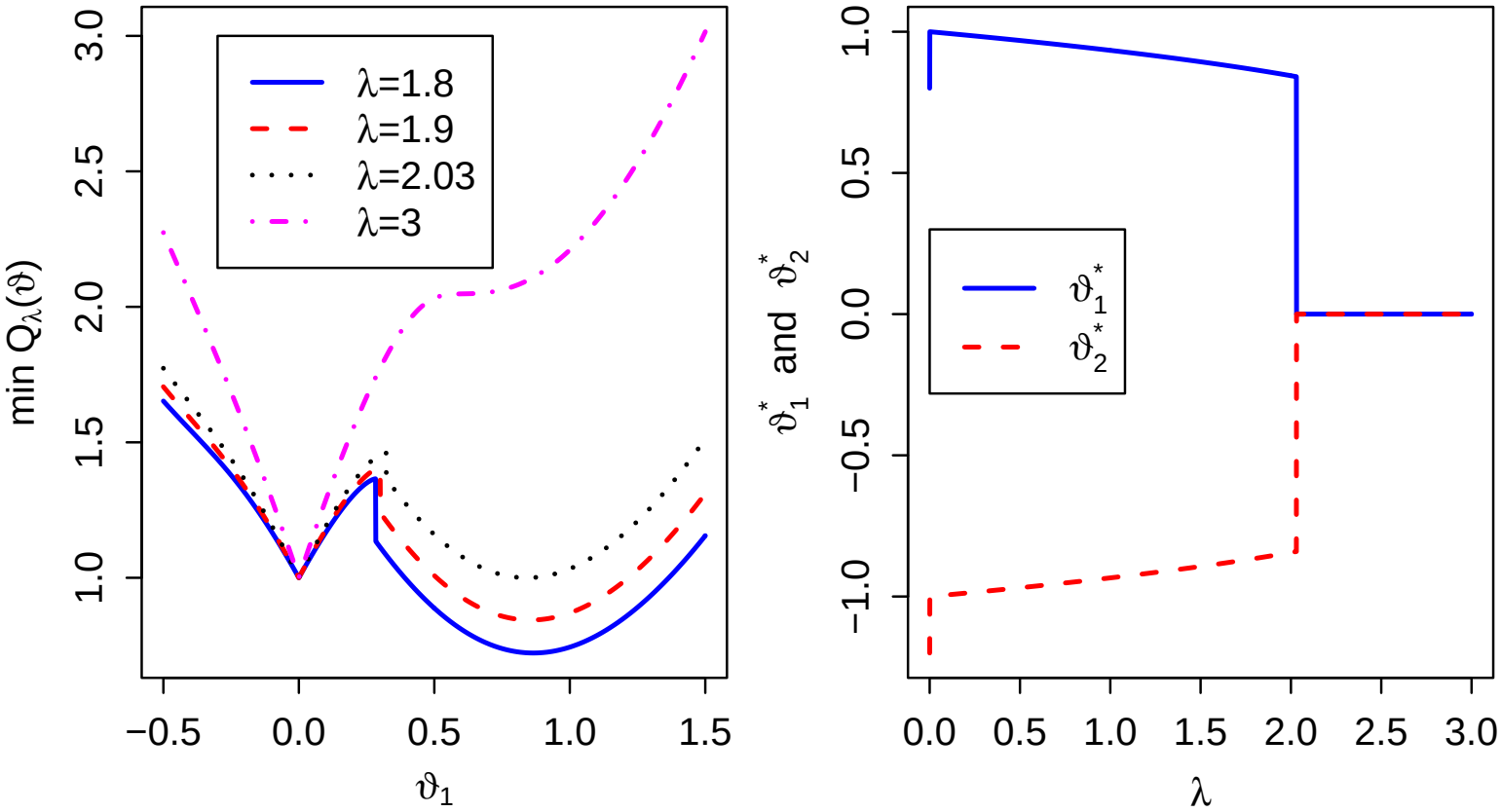

Figure: Left graph: $\min Q_\lambda(\boldsymbol{\vartheta})$ as function of ϑ_1 for several values of λ. Right graph: $(\vartheta_1^\star, \vartheta_2^\star) = \arg\min Q_\lambda(\boldsymbol{\vartheta})$ as function of λ.

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

Let

$$Q_\lambda(\boldsymbol{\vartheta}) = Q(\boldsymbol{\vartheta}) + \lambda|\vartheta_1|, \quad \boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^\top,$$

with $Q(\boldsymbol{\vartheta}) = \sum_{i=1}^2 \vartheta_i^4 - \frac{2}{3}\vartheta_i^3 - 2\vartheta_i^2 + 2\vartheta_i + 1$. Note that $\boldsymbol{\vartheta}^\star := \arg\min Q_\lambda(\boldsymbol{\vartheta}) = (\vartheta_1^\star, -1)^\top$, and that for $\boldsymbol{\vartheta}^c := (0, -1)^\top$ we have

$$\lambda^i = \partial Q(\boldsymbol{\vartheta}^c)/\partial\vartheta_1 = 2.$$

However, $\arg\min Q_2(\boldsymbol{\vartheta}) = (-0.781, -1)^\top \neq \boldsymbol{\vartheta}^c$. We have $\boldsymbol{\vartheta}^\star = \boldsymbol{\vartheta}^c$ for $\lambda \geq 2.746$, which is in agreement with

$$\lambda^s = \sup_{\vartheta_1 \leq 0} 4\vartheta_1^3 - 2\vartheta_1^2 - 4\vartheta_1 + 2 = 3.032.$$

Example where $\lambda^i < \lambda^s$

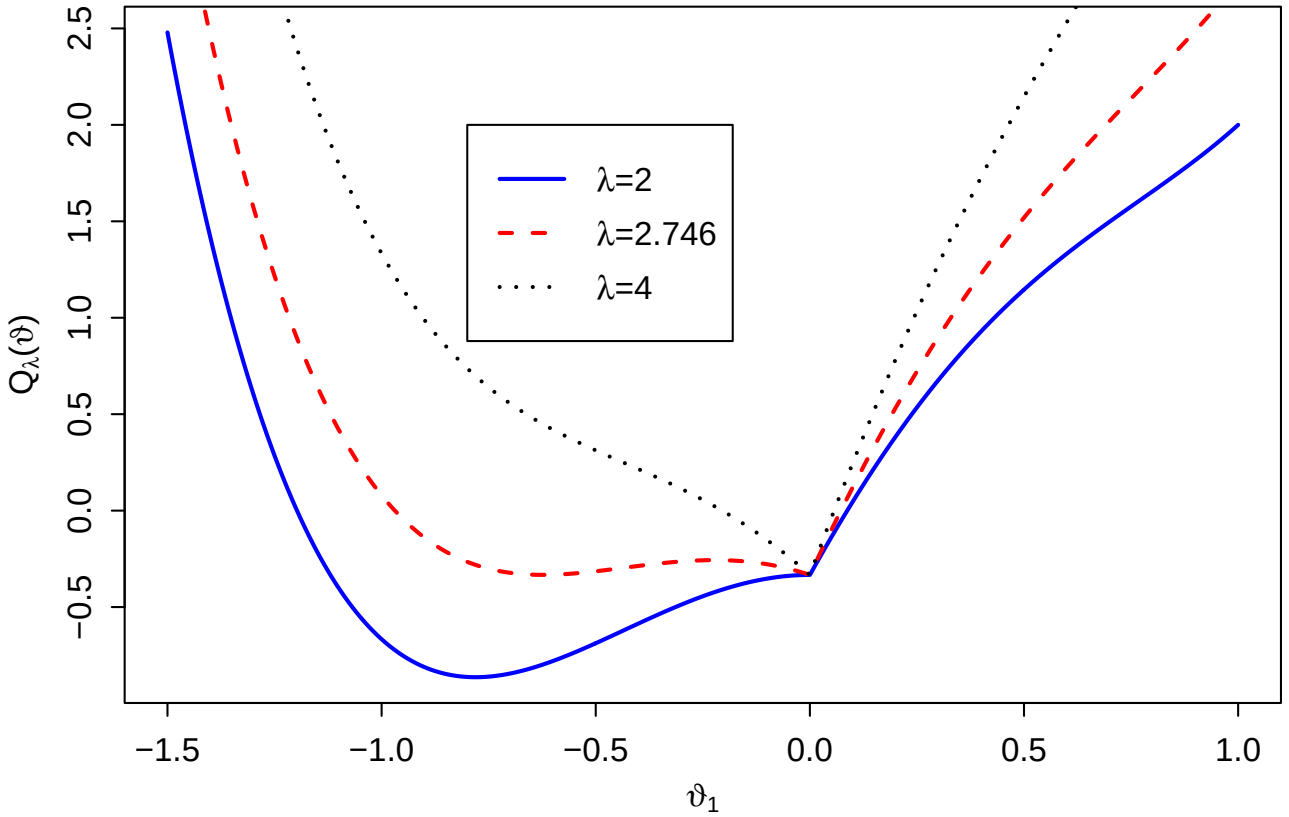

Figure: Function $\vartheta_1 \mapsto Q_\lambda(\vartheta_1, -1) = Q(\vartheta_1, -1) + \lambda|\vartheta_1|$. For $\boldsymbol{\vartheta}^c = (0, -1)$, we have $\partial Q(\boldsymbol{\vartheta}^c)/\partial\vartheta_1 = 2$, but $\arg\min_{\boldsymbol{\vartheta}} Q_\lambda(\boldsymbol{\vartheta}) = \boldsymbol{\vartheta}^c$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

Example of computation of λ^i and λ^s

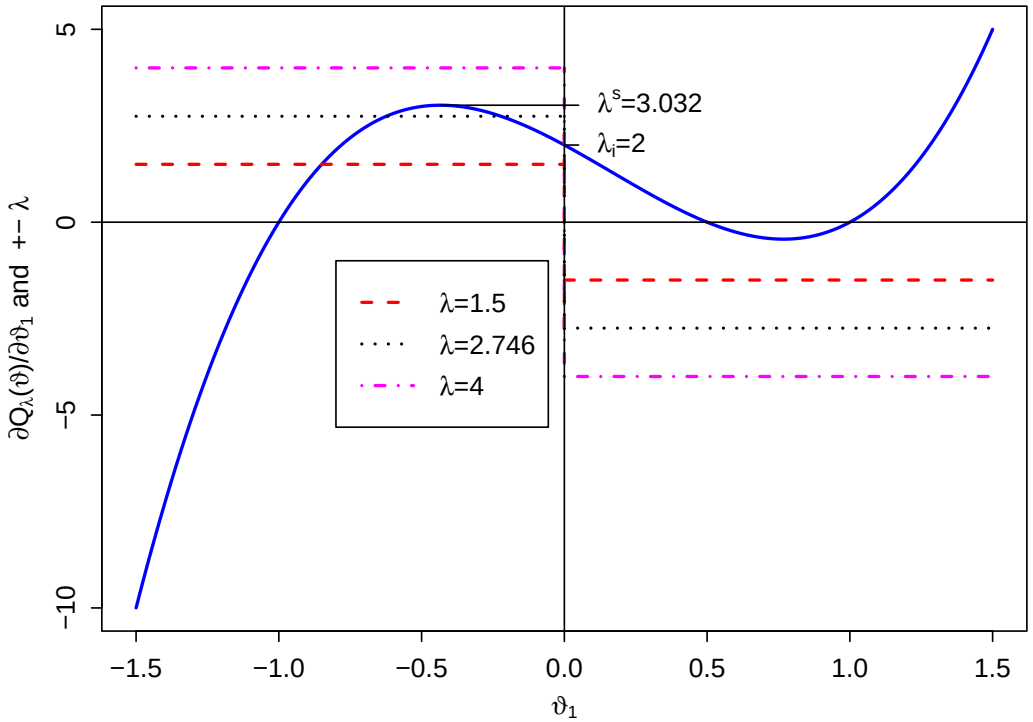

Figure: Graphical representation of λ^i and λ^s.

NLShoot for a simple optimization

Let us use the NLshoot algorithm to compute

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

NLShoot reduces to a single application of the algorithm. The figure below shows 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.

The objective function

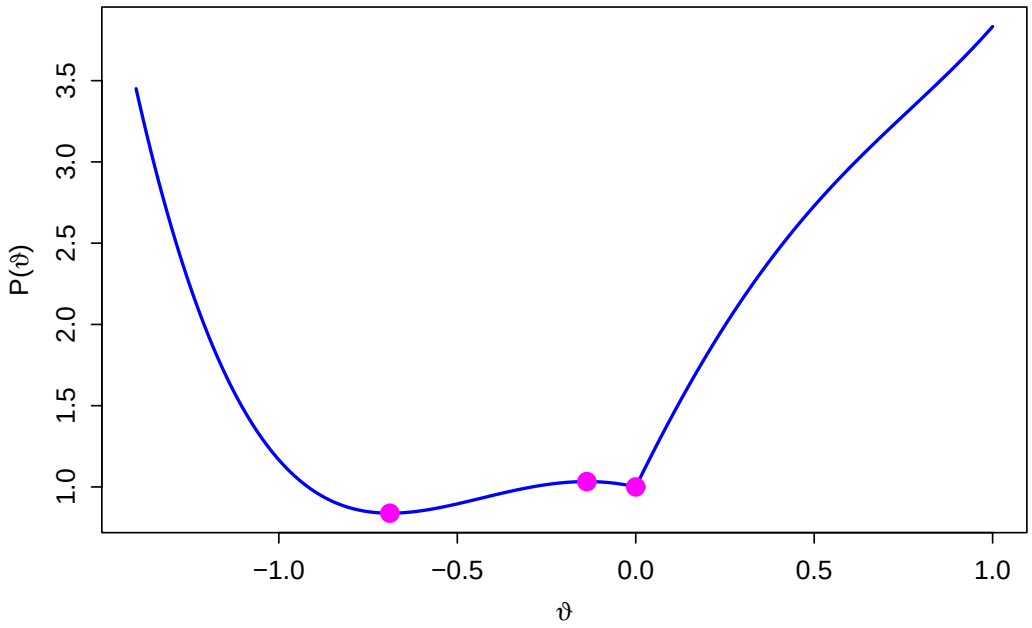

Figure: Function $P(\vartheta) + 2.5|\vartheta|$ with a minimum at $\vartheta = -0.689$, and two other "critical points" at $\vartheta = -0.137$ and $\vartheta = 0$.

NLShoot in a simple (degenerate) case

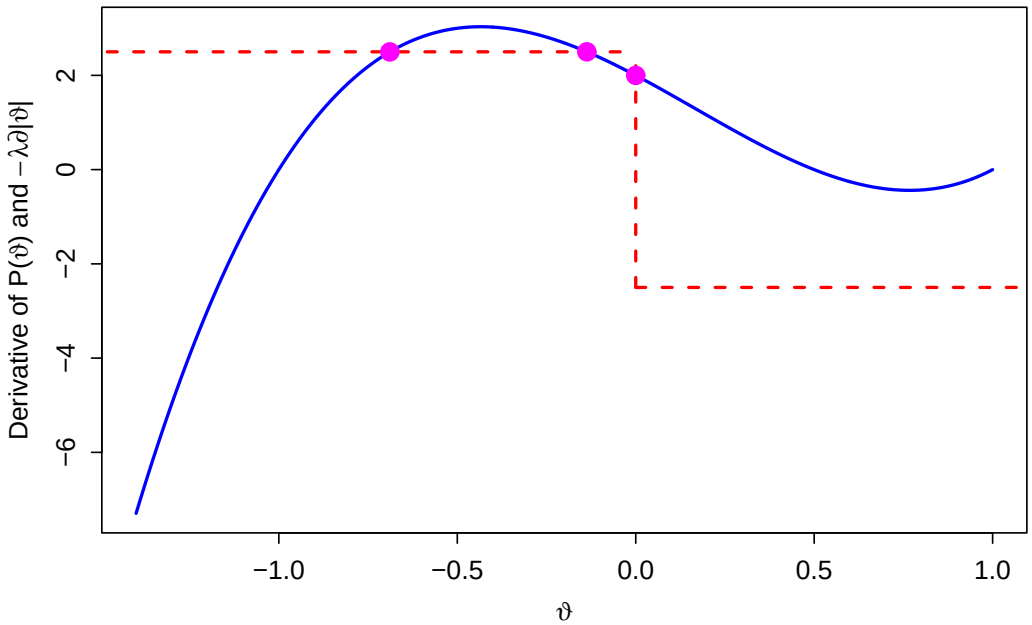

Figure: NLShoot finds the 3 critical points.

FOC satisfied by $\widehat{\boldsymbol{\vartheta}}_n$ and sparsity of the estimator

We have

$$0 \in \left\{ \frac{\partial \widetilde{O}_n(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\vartheta}}_n)}{\partial \boldsymbol{\vartheta}} \right\} + \lambda_n \boldsymbol{\partial} p\left(\widehat{\boldsymbol{\vartheta}}_n\right).$$

Therefore, for all $i \in S$,

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

The LQA algorithm

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

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

until convergence, where $\boldsymbol{\vartheta}^{(0)} \in \Theta$ is an initial value.

Fan and 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 $\varepsilon > 0$, but the choice of ε matters.