Penalized global quantile regression using MM algorithm

Introduction

Consider a scalar random variable Y and a D-dimensional random vector X such that the following linear quantile regression model holds

$$Q(\tau|x) = x'\beta(\tau),$$

for $\tau \in (0,1)$ and $x \in \text{support}(X)$, where $\beta : (0,1) \to \mathbb{R}^D$ is a smooth function and where

$$Q(\tau|x) := \inf\{y \in \mathbb{R} : F(y|x) \geq \tau\},$$

with $F(\cdot|x)$ denoting the conditional cdf of Y given $X = x$. It is well known that, given any vector $b \in \mathbb{R}^D$ and $\tau \in (0,1)$, one has

$$\mathbb{E}\rho_\tau\big(Y - X'b\big) \geq \mathbb{E}\rho_\tau\big(Y - X'\beta(\tau)\big),$$

where $2\rho_\tau(z) := (2\tau - 1)z + |z|$ for $z \in \mathbb{R}$ and $\tau \in (0,1)$. Hence, given a grid

$$0 < \tau_1 \leq \tau_2 \leq \cdots \leq \tau_M < 1$$

of quantile levels (possibly depending on $N \in \mathbb{N}$), and writing

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1(\tau_1) & \beta_1(\tau_2) & \cdots & \beta_1(\tau_M) \\ \beta_2(\tau_1) & \beta_2(\tau_2) & \cdots & \beta_2(\tau_M) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_D(\tau_1) & \beta_D(\tau_2) & \cdots & \beta_D(\tau_M) \end{pmatrix},$$

it also holds that, for any $D \times M$ matrix B with columns $b_m \in \mathbb{R}^D$ $(m = 1, \ldots, M)$,

$$R(B) := \sum_{m=1}^{M} \mathbb{E}\rho_{\tau_m}\big(Y - X'b_m\big) \geq \sum_{m=1}^{M} \mathbb{E}\rho_{\tau_m}\big(Y - X'\beta(\tau_m)\big) = R(\boldsymbol{\beta})$$

Said another way, we have

$$R(\boldsymbol{\beta}) = \arg\min_{B} R(B)$$

where the minimum in B runs through all $D \times M$ matrices B.

Now let $\varphi_1, \varphi_2, \ldots, \varphi_L$ be linearly independent vectors in $\mathbb{R}^M$ (possibly depending on $N \in \mathbb{N}$), with $L \leq M$. Also assume that

$$\boldsymbol{\beta} \approx \boldsymbol{\alpha}\boldsymbol{\varphi} \tag{proj}$$

for some $D \times L$ matrix $\boldsymbol{\alpha}$, where $\boldsymbol{\varphi}$ is the $L \times M$ matrix with *rows* φ_ℓ. This will be the case whenever the mapping $\tau \mapsto \beta(\tau)$ is "sufficiently" regular, in the sense that, for all d, the coefficient $\beta_d(\cdot)$ can be represented in terms of a sequence of basis functions $f_1, f_2, \ldots$ as

$$\beta_d(\cdot) = \sum_{\ell=1}^{\infty} \alpha_{d\ell} f_\ell(\cdot).$$

In this setting, one could take $[\varphi_\ell]_m = f_\ell(\tau_m)$ for example. To be concrete, the vectors $\varphi_1, \ldots, \varphi_L$ can be taken from application of a Gram-Schmidt algorithm to the set $\{\tilde{\varphi}_1, \ldots, \tilde{\varphi}_L\} \subseteq \mathbb{R}^M$ given by

$$\tilde{\varphi}_\ell' = \begin{pmatrix} \tau_1^{\ell-1} & \cdots & \tau_M^{\ell-1} \end{pmatrix}, \qquad \ell = 1, \ldots, L.$$

Estimation

Given a random sample $(X_n, Y_n)_{n=1}^N$ from (X, Y), the representation in equation (proj) suggests estimating $\boldsymbol{\beta}$ by finding a $D \times L$ matrix $\widehat{\boldsymbol{\alpha}}$ to solve the optimization problem

$$\min_{\boldsymbol{a}} \sum_{m=1}^{M} \sum_{n=1}^{N} \rho_{\tau_m}(Y_n - X_n' \boldsymbol{a} \varphi_m)$$

where φ_m is the mth *column* of $\boldsymbol{\varphi}$. Letting $\boldsymbol{\eta}$ denote an $N \times M$ matrix with positive entries, we can use the eta-trick which states that

$$2|v| = \min_{\eta > 0} \frac{v^2}{\eta} + \eta$$

to write the above minimization problem alternatively as

$$\min_{\boldsymbol{a}} \min_{\boldsymbol{\eta} > 0} \widehat{R}(\boldsymbol{a}, \boldsymbol{\eta})$$

where

$$2\widehat{R}(\boldsymbol{a}, \boldsymbol{\eta}) := \sum_{m=1}^{M} \sum_{n=1}^{N} (2\tau_m - 1)(Y_n - X_n' \boldsymbol{a} \varphi_m) + \frac{1}{2} \left[\frac{(Y_n - X_n' \boldsymbol{a} \varphi_m)^2}{\eta_{nm}} + \eta_{nm} \right] \tag{1}$$

This can be extended to include group lasso-type penalties, for example by letting $\boldsymbol{\zeta}$ denote a $D \times M$ matrix and writing

$$\widehat{R}_\lambda(\boldsymbol{a}, \boldsymbol{\eta}, \boldsymbol{\zeta}) = \widehat{R}(\boldsymbol{a}, \boldsymbol{\eta}) + \lambda \sum_{d=1}^{D} \frac{1}{2} \left[\frac{\|\boldsymbol{a}_d\|^2}{\zeta_d} + \zeta_d \right] \tag{2}$$

where $\lambda > 0$ is a tuning parameter and

$$\|\boldsymbol{a}_d\| = \sqrt{\sum_{\ell=1}^{L} \boldsymbol{a}_{d\ell}^2}$$

Algorithm

0. Initialize η_{nm}^0 and ζ^0 arbitrarily.
1. Given the values of η_{nm}^{r-1} and ζ^{r-1} in the (r-1)th iteration, find the solution $\widehat{\boldsymbol{a}}^r$ that minimizes (2) w.r.t. $\boldsymbol{a}$;
2. Update $\eta_{nm}^r = |Y_n - vec(X_n \varphi_m')' vec(\boldsymbol{a})^r|$ and $\zeta = \|\boldsymbol{a}\|$;
3. Iterate steps 1-2 until the convergence criterion has been met: $\frac{abs(R_\lambda(\boldsymbol{a}^r, \eta^r, \zeta^r) - R_\lambda(\boldsymbol{a}^{r+1}, \eta^{r+1}, \zeta^{r+1}))}{R_\lambda(\boldsymbol{a}^r, \eta^r, \zeta^r)} < tol.$

Note: Use a precision parameter π to set $\eta_{nm}^r = \max\{\pi, \eta_{nm}^r\}$ and $\zeta^r = \max\{\pi, \zeta^r\}$.

Additional notes

By vectorizing the matrix $\boldsymbol{a}$, we are able to find an analyitical solution to $vec(\boldsymbol{a})$ and use the standard `optim()` function to perform step 1:

Let

$$vec(\boldsymbol{a}) = [a_{11}, a_{21}, \ldots, a_{d\ell}],$$

$$vec(X_n\varphi_m') = \begin{bmatrix} X_{n1}\varphi_{1m} \\ X_{n2}\varphi_{1m} \\ \vdots \\ X_{nd}\varphi_{1m} \\ X_{n1}\varphi_{2m} \\ X_{n1}\varphi_{2m} \\ \vdots \\ X_{n1}\varphi_{\ell m} \\ X_{n2}\varphi_{\ell m} \\ \vdots \\ X_{nd}\varphi_{\ell m} \end{bmatrix}$$

With this, we re-write $X_n'\boldsymbol{a}\varphi_m$ in equation (1) by $vec(X_n\varphi_m')'vec(\boldsymbol{a}) = vec(\boldsymbol{a})'vec(X_n\varphi_m')$. In this notation, we differentiate (2) and find the estimator for $vec(\boldsymbol{a})$ as follows:

$$\widehat{vec(\boldsymbol{a})} = \left[\sum_{n=1}^{N}\sum_{m=1}^{M}\left[-\frac{1}{\eta_{nm}}vec(X_n\varphi_m')vec(X_n\varphi_m')' + \lambda diag(vec(\frac{1}{\zeta}))\right]\right]^{-1} + \left[\sum_{n=1}^{N}\sum_{m=1}^{M}\left[(1-\tau_m)vec(X_n\varphi_m') - \frac{Y_n}{\eta_{nm}}vec(X_n\varphi_m')\right]\right]$$