Nonnegative garrote in additive models using P-splines

Anneleen Verhasselt

Department of Mathematics and LStat

Katholieke Universiteit Leuven, Belgium

anneleen.verhasselt@wis.kuleuven.be

JOINT WORK WITH ANESTIS ANTONIADIS AND IRÈNE GIJBELS

Outline

1. Nonnegative garrote in additive model
2. P-spline estimator
3. Smoothing spline estimator
4. Backfitting
5. Consistency
6. COSSO-type P-spline estimator
7. Simulations and Boston housing data

Outline
- Outline

Nonnegative garrote

Initial estimators

Consistency

Cosso-type estimator

Applications

References

Nonnegative garrote

- **Multiple linear model**: response Y, d explanatory variables $X_1, \ldots, X_d$

$$Y_i = \sum_{j=1}^{d} \beta_j X_{ij} + \varepsilon_i$$

- **Nonnegative garrote** (Breiman, 1995):

$$\begin{cases} (\hat{c}_1, \ldots, \hat{c}_d) = \text{argmin}_{\vec{c}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{d} c_j \hat{\beta}_j^{\text{LS}} X_{ij} \right)^2 \\ \text{s.t. } c_j \geq 0 \ (j = 1, \ldots, d), \quad \sum_{j=1}^{d} c_j \leq s \end{cases}$$

$$\Longleftrightarrow$$

$$\begin{cases} (\hat{c}_1, \ldots, \hat{c}_d) = \text{argmin}_{\vec{c}} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{d} c_j \hat{\beta}_j^{\text{LS}} X_{ij} \right)^2 + \lambda \sum_{j=1}^{d} c_j \right\} \\ \text{s.t. } c_j \geq 0 \end{cases}$$

$$\Longrightarrow \hat{\beta}_j^{\text{NNG}} = \hat{c}_j \hat{\beta}_j^{\text{LS}}$$

Outline

Nonnegative garrote

- Nonnegative garrote
- Functional nonnegative garrote

Initial estimators

Consistency

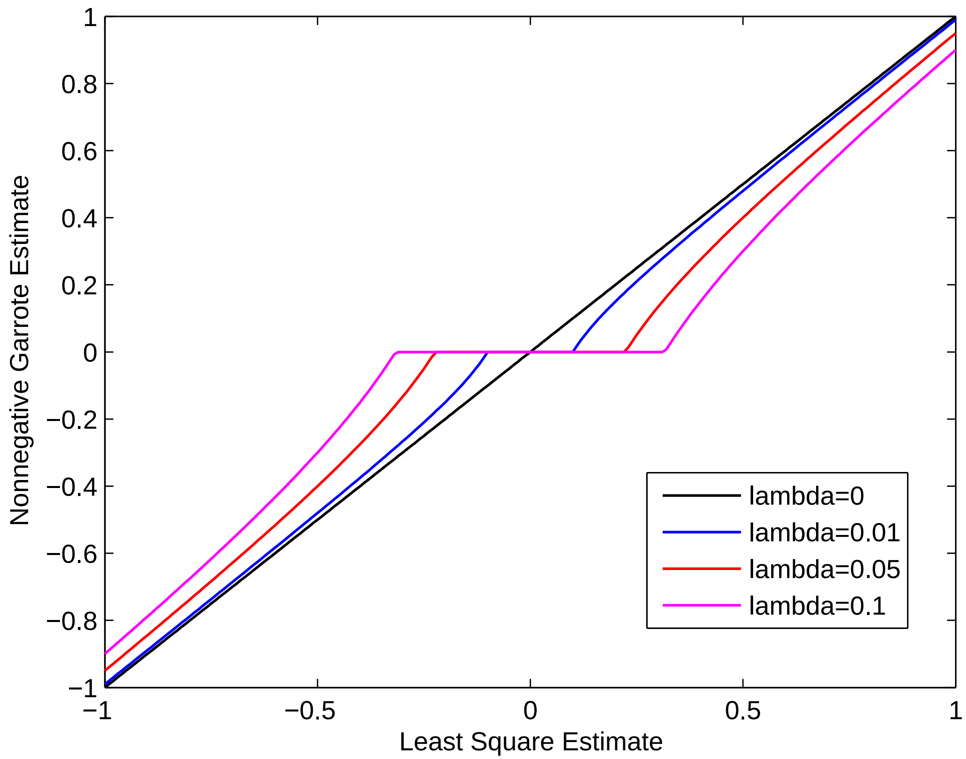
Cosso-type estimator

Applications

References

Orthogonal design: $X'X = I_{n \times n}$

$$\hat{\beta}_j^{\text{NNG}} = \left(1 - \frac{\lambda}{2(\hat{\beta}_j^{\text{LS}})^2}\right)_+ \hat{\beta}_j^{\text{LS}}$$

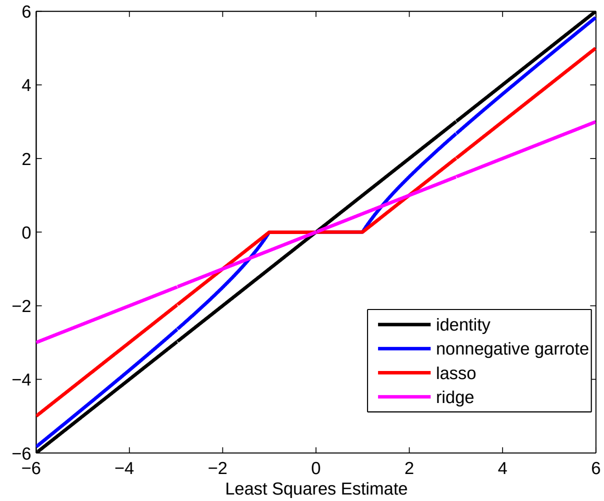


Lasso:

$$\begin{cases} (\hat\beta_1^{\text{Lasso}}, \ldots, \hat\beta_d^{\text{Lasso}}) = \text{argmin}_{\vec\beta} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^d \beta_j X_{ij} \right)^2 \\ \text{s.t. } \sum_{j=1}^d |\beta_j| \leq s \end{cases}$$

Ridge:

$$\begin{cases} (\hat\beta_1^{\text{Ridge}}, \ldots, \hat\beta_d^{\text{Ridge}}) = \text{argmin}_{\vec\beta} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^d \beta_j X_{ij} \right)^2 \\ \text{s.t. } \sum_{j=1}^d \beta_j^2 \leq s \end{cases}$$



Functional nonnegative garrote

- **Additive model**: response Y, d explanatory variables $X_1, \ldots, X_d$

$$Y_i = \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i$$

- **Functional nonnegative garrote** (Cantoni et al., 2006 and Yuan, 2007):

$$\begin{cases} (\hat{c}_1, \ldots, \hat{c}_d) = \text{argmin}_{\vec{c}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{d} c_j \hat{f}_j^{\text{init}}(X_{ij}) \right)^2 \\ \text{s.t. } c_j \geq 0 \; (j = 1, \ldots, d), \quad \sum_{j=1}^{d} c_j \leq s \end{cases}$$

$$\Longrightarrow \hat{f}_j^{\text{NNG}} = \hat{c}_j \hat{f}_j^{\text{init}}$$

P-spline estimator

- Regression of n data points (X_i, Y_i), with

$$Y_i = \mu(X_i) + \varepsilon_i$$

ε_i i.i.d. $N(0, \sigma^2)$

- Representation of $\mu(\cdot)$ in a q-th degree B-spline basis:

$$\mu(x) = \sum_{j=1}^{m} \alpha_j B_j(x; q)$$

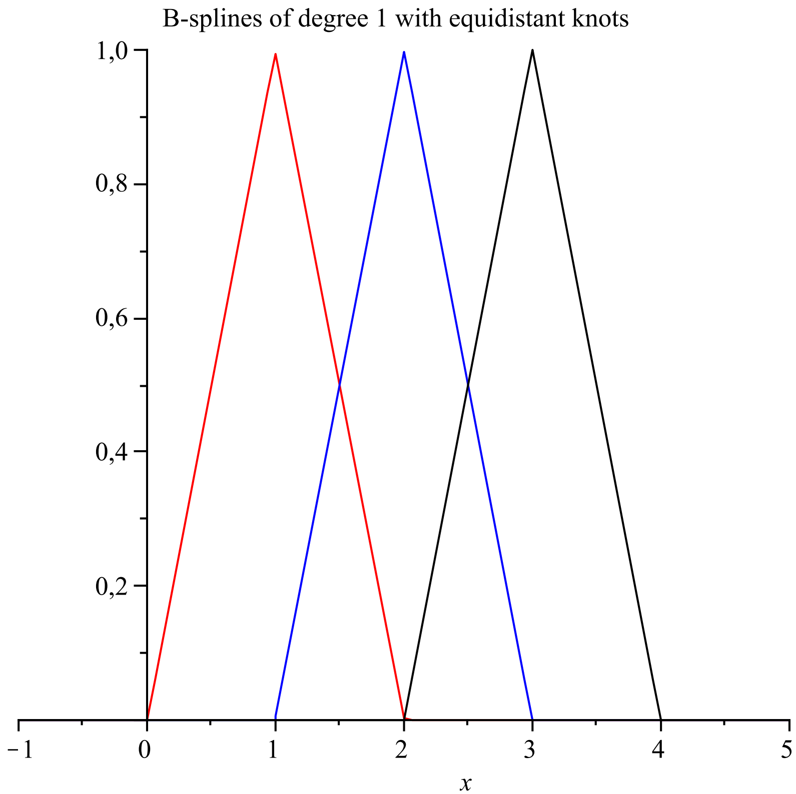
- Penalized least squares estimator $\hat{\alpha}$: minimizer of

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j B_j(X_i; q) \right)^2 + \lambda \sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2 = ||Y - B\vec{\alpha}||_2^2 + \lambda||D_k\vec{\alpha}||_2^2$$

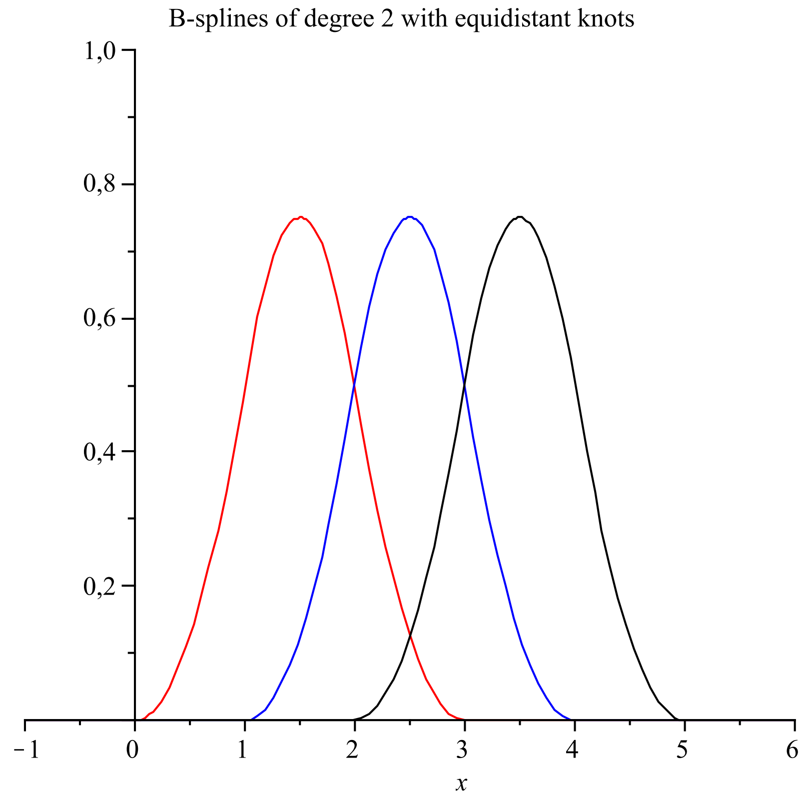
For example: $\Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$

$$\Longrightarrow \text{P-spline estimator } \hat{\mu}(x) = \sum_{j=1}^{m} \hat{\alpha}_j B_j(x; q)$$

B-splines



B-splines of degree 1 with equidistant knots
1,0
0,8
0,6
0,4
0,2
−1
0
1
2
3
4
5
x



B-splines of degree 2 with equidistant knots
1,0
0,8
0,6
0,4
0,2
−1
0
1
2
3
4
5
6
x

Smoothing spline estimator

- Regression of n data points (X_i, Y_i), with

$$Y_i = \mu(X_i) + \varepsilon_i$$

  ε_i i.i.d. $N(0, \sigma^2)$

- Smoothing spline estimator $\hat{\mu}(x)$: minimizer of

$$\sum_{i=1}^{n} \Big(Y_i - \mu(X_i)\Big)^2 + \lambda \int (\mu^{(k)}(x))^2 \, dx$$

- If $k = 2$, $\hat{\mu}(x)$ is a cubic spline on $[X_{(1)}, \ldots, X_{(n)}]$

Backfitting

$$Y_i = \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i$$

Algorithm:

1. Set $\hat{f}_j(X_{1j}, \ldots, X_{nj}) = 0$ for $j = 1, \ldots, d$
2. Repeat steps 3 to 5 until the estimates $\hat{f}_j(X_{1j}, \ldots, X_{nj})$ stop changing
3. For $j = 1, \ldots, d$ repeat steps 4 and 5
4. Calculate the partial residuals

$$e_j = Y - \sum_{h \neq j} \hat{f}_h(X_{1h}, \ldots, X_{nh})$$

5. Set $\hat{f}_j(X_{1j}, \ldots, X_{nj})$ equal to the result of smoothing e_j with respect to x_j

Consistency

- **Yuan (2007):** Assume that for some $\delta_n \to 0$

$$\frac{1}{n}\sum_{i=1}^{n}(f_j(x_{ij}) - \hat{f}_j^{\text{init}}(x_{ij}))^2 = \frac{1}{n}||f_j - \hat{f}_j^{\text{init}}||_2^2 = O_P(\delta_n^2).$$

If $\lambda \to 0$ in a fashion such that $\delta_n = o_P(\lambda)$ then

$$\sup_j \frac{1}{n}||f_j - \hat{f}_j^{\text{NNG}}||_2^2 = O_P(\lambda^2) \text{ and } P(\hat{f}_j^{\text{NNG}} = 0) \to 1 \text{ if } f_j = 0$$

- **Claeskens, Krivobokova and Opsomer (2008):**

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}((f_j(x_{ij}) - \hat{f}_j(x_{ij}))^2) = \frac{1}{n}\mathrm{E}(||f_j - \hat{f}_j||_2^2) = O(n^{-\frac{2q+2}{2q+3}})$$

- **Horowitz, Klemelä and Mammen (2006):** *each additive component can be estimated as well as it could be if the others were known*

COSSO-type P-spline estimator

- **COSSO estimator** (Lin and Zhang, 2006): function $f \in \mathcal{F}$ (with $\mathcal{F} = \{1\} \bigoplus_{j=1}^{d} \mathcal{F}_j$ a RKHS) that minimizes

$$\sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{d} f_j(X_{ij})\Big)^2 + \lambda \sum_{j=1}^{d} ||P^j f||_{\mathcal{F}}$$

- COSSO-type P-spline estimator $(\hat{\vec{\alpha}}_1, \ldots, \hat{\vec{\alpha}}_d)$: minimizer of

$$||Y - \sum_{j=1}^{d} B_j \vec{\alpha}_j||_2^2 + \sum_{j=1}^{d} \lambda_j ||D_k \vec{\alpha}_j||_2$$

$$\Longrightarrow \hat{f}(x_1, \ldots, x_d) = \sum_{j=1}^{d} \sum_{h=1}^{m} \hat{\alpha}_{hj} B_h(x_j; q)$$

Simulated data

Outline

Nonnegative garrote

Initial estimators

Consistency

Cosso-type estimator

Applications

- Simulated data
- Boston Housing data

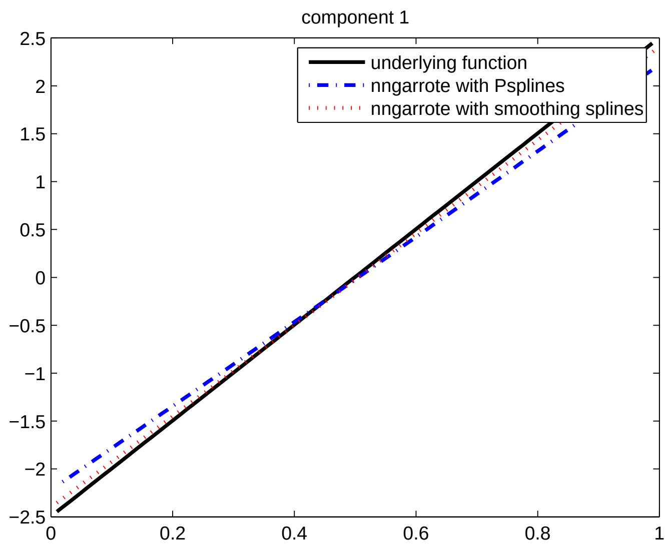
References

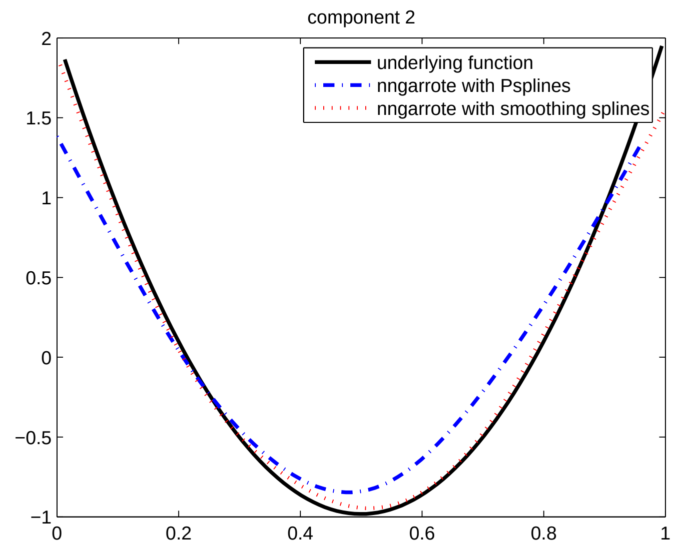
$$Y_i = f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + \varepsilon_i$$

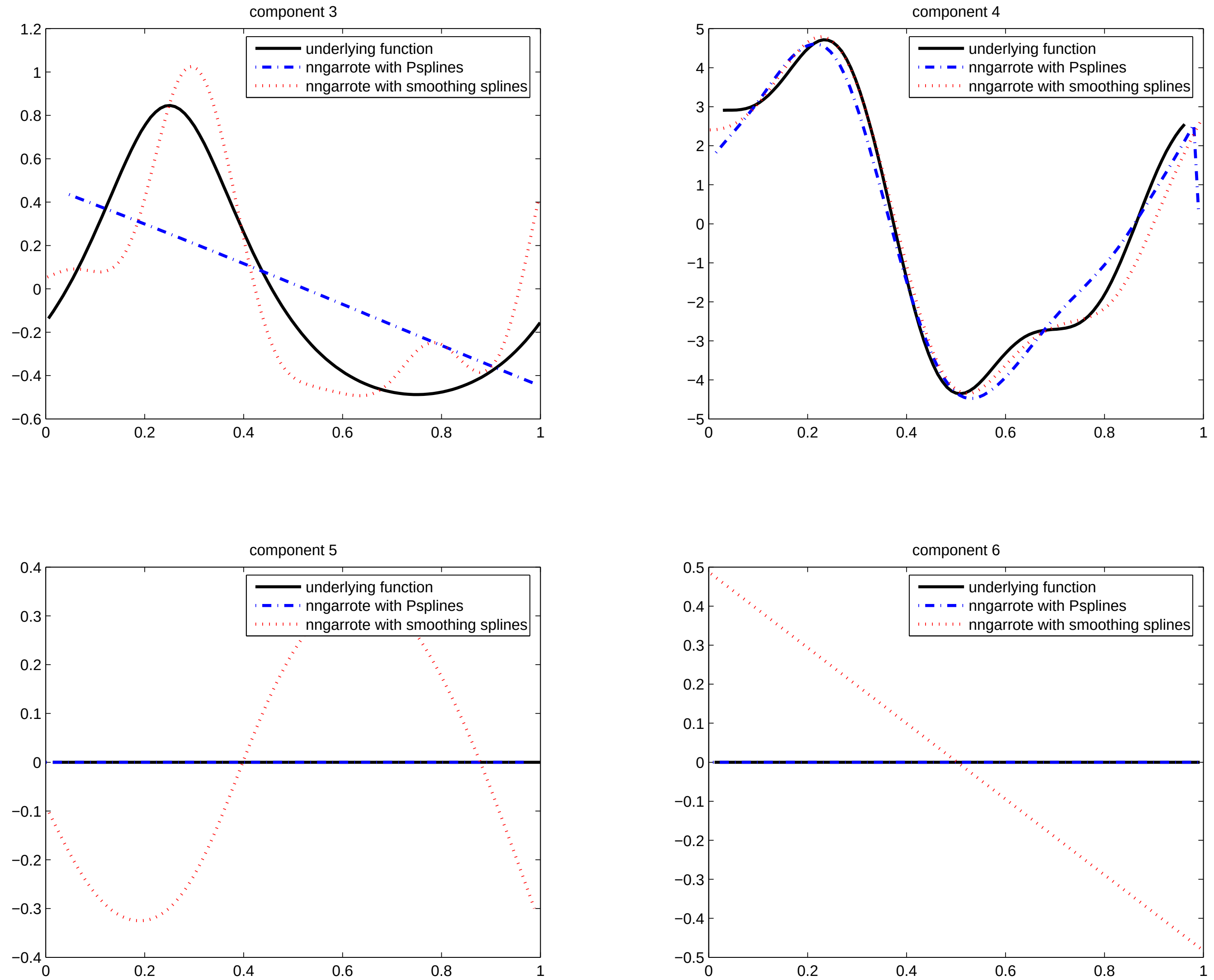
$$f_1(x) = 5x,\ f_2(x) = 3(2x-1)^2,\ f_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)}$$

$$f_4(x) = 0.6\sin(2\pi x) + 1.2\cos(2\pi x) + 1.8\sin^2(2\pi x) + 2.4\cos^3(2\pi x) + 3\sin^3(2\pi x)$$

$$AISE_j = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_j^{\text{NNG}}(X_{ij}) - f_j(X_{ij}))^2$$







Outline

Nonnegative garrote

Initial estimators

Consistency

Cosso-type estimator

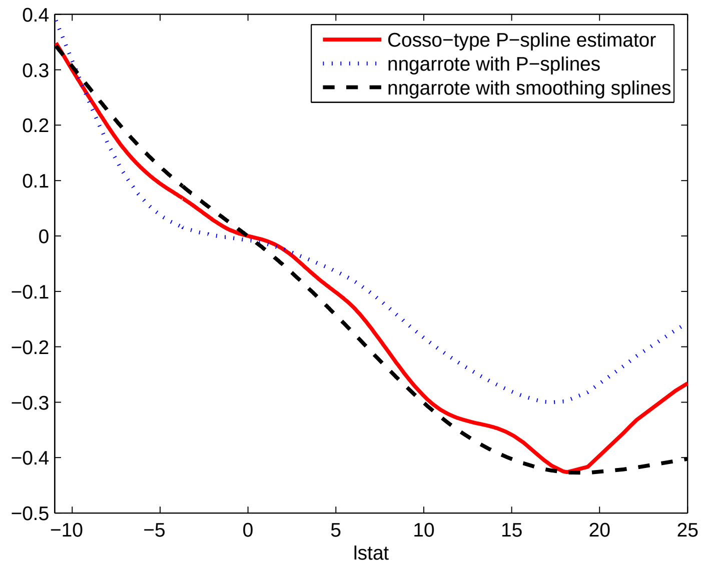
Applications

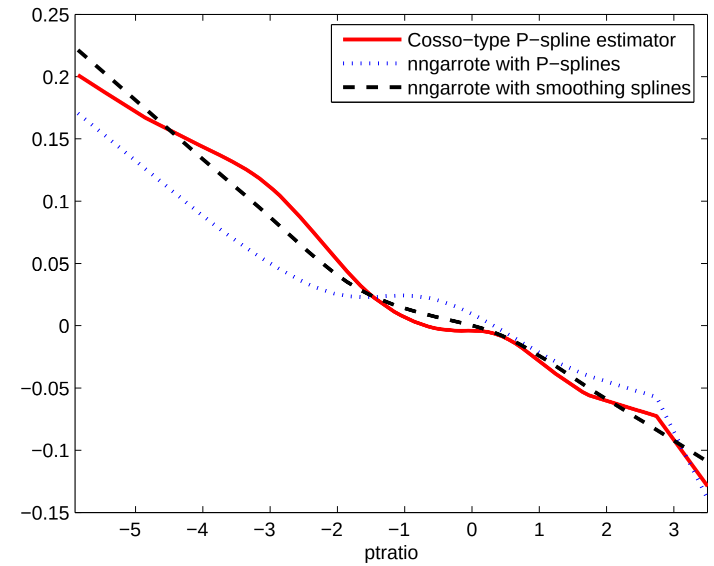
- Simulated data
- Boston Housing data

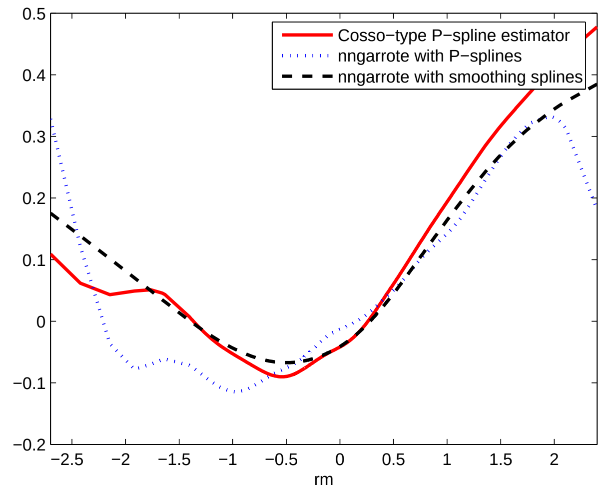
References

Boston Housing data

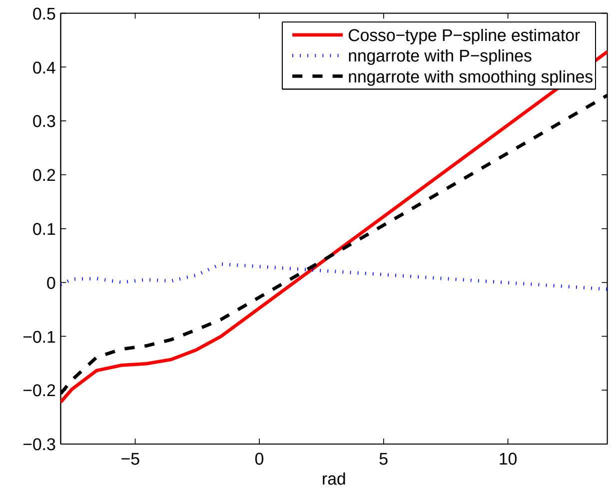
$$
\begin{aligned}
log(\text{medv}) \quad = \quad & \alpha + f_1(\text{crim}) + f_2(\text{zn}) + f_3(\text{indus}) + f_4(\text{nox}) + f_5(\text{rm}) \\
& + f_6(\text{age}) + f_7(\text{dis}) + f_8(\text{rad}) + f_9(\text{tax}) + f_{10}(\text{ptratio}) \\
& + f_{11}(\text{b}) + f_{12}(\text{lstat}) + \beta\text{chas} + \varepsilon
\end{aligned}
$$



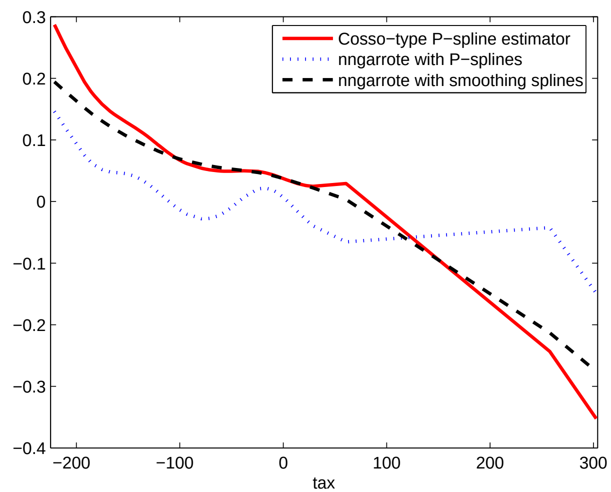




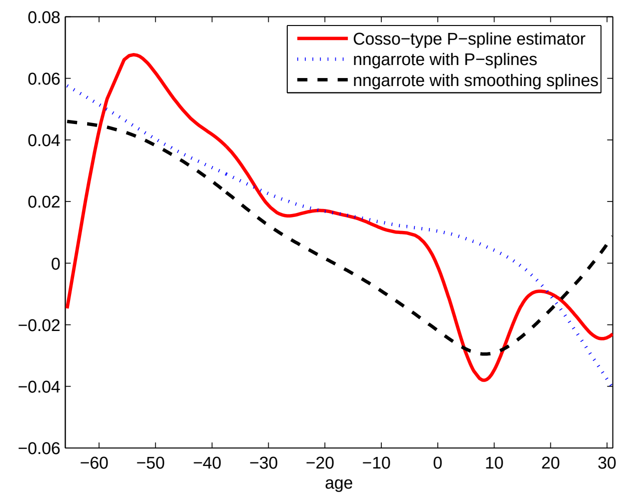
Cosso−type P−spline estimator
nngarrote with P−splines
nngarrote with smoothing splines
rm



Cosso−type P−spline estimator
nngarrote with P−splines
nngarrote with smoothing splines
rad



Cosso−type P−spline estimator
nngarrote with P−splines
nngarrote with smoothing splines
tax



Cosso−type P−spline estimator
nngarrote with P−splines
nngarrote with smoothing splines
age

Outline
Nonnegative garrote
Initial estimators
Consistency
Cosso-type estimator
Applications
References
● References

References

- Breiman, L. (1995).“Better subset regression using the nonnegative garrote", *Technometrics*, **37**: 373-384.
- Cantoni, E., Flemming, J. & Ronchetti E. (2006). “Variable Selection in Additive Models by Nonnegative Garrote", *Cahier du Département d’Econométrie 2006.02*, University of Geneva (Switzerland)
- Claeskens, G., Krivobokova, T. & Opsomer, J.D. (2008) “Asymptotic properties of penalized spline estimators".
- Eilers, P. & Marx, B. (1996): “Flexible smoothing with B-splines and penalties", *Statistical Science*, **11**: 89-102.
- Horowitz, J., Klemelä, J. & Mammen, E. (2006) “Optimal estimation in additive regression models", *Bernoulli*, **12**: 271–298.
- Lin, B. & Zhang, H. (2006). “Component selection and smoothing in multivariate nonparametric regression", *The Annals of Statistics*, **34**: 2272–2297.
- Yuan, M. (2007). “Nonnegative Garrote Component Selection in Functional ANOVA models", *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics March 21-24, 2007, San Juan, Puerto Rico*, **2**: 660-666.