Knot deletion for robust penalized spline regression

Steffen Liebscher and Thomas Kirschstein

Martin-Luther-University Halle-Wittenberg, Germany
Institute of Economics and Business Studies
Chair of Statistics

Rio de Janeiro, July 28, 2015

Introduction

- penalized spline regression is a popular method to fit a smooth spline function to a set of observations, see Eilers and Marx (1996) and Ruppert et al. (2003) for details
- applications e.g. in environmental modeling, classification, in the context of functional and longitudinal data analysis, etc.
- robust variants are available, see Lee and Oh (2007) and Tharmaratnam et al. (2010)
- crucial parameters: penalty parameter & the number of knots and their position

Fundamentals

- regression model

$$Y = m(x, \boldsymbol{\beta}) + \epsilon \tag{1}$$

where $m(x, \boldsymbol{\beta})$ is a smooth regression function with

$$m(x, \boldsymbol{\beta}) = \sum_{i=1}^{n} \beta_i f_i(x) \tag{2}$$

with basis $f_1(x), \ldots, f_n(x)$ and coefficients $\beta_1, \ldots, \beta_n$

- popular choices include a B-spline basis or a truncated polynomial basis

Fundamentals cont'd

- given a sample $(Y_1, x_1), \ldots, (Y_J, x_J)$, coefficients are estimated by solving

$$\widehat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{j=1}^{J} (Y_j - m(x_j, \boldsymbol{\beta}))^2 + \lambda \cdot pen(\beta_{p+1}, \ldots, \beta_{p+K}) \right\} \quad (3)$$

  with penalty parameter $\lambda > 0$ and some penalty measure $pen(\cdot)$ which depends on $\beta_{p+1}, \ldots, \beta_{p+K}$

- common choices for the penalty measure are the sum of squares of the betas or the sum of squares of differences between adjacent betas
- to robustify (3) replace the RSS component by a robust alternative

Basis functions

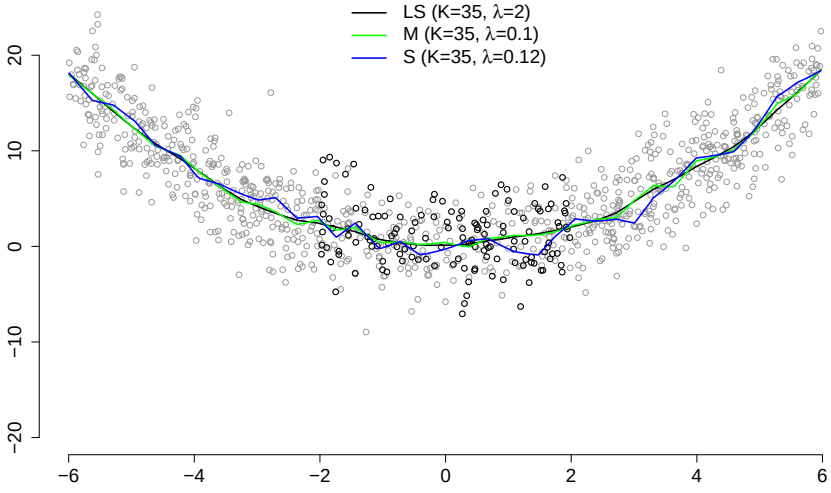
- e.g. a truncated polynomial basis of order p

$$m(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k} \max(x - \kappa_k, 0)^p \qquad (4)$$

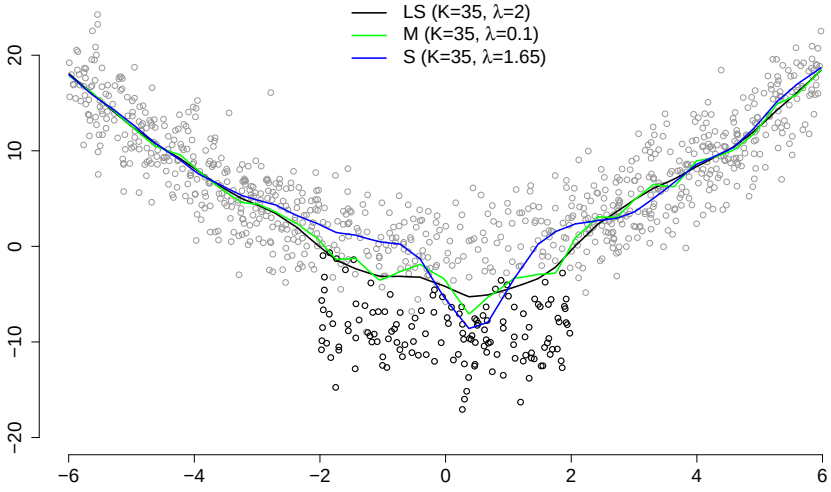
  where p is the order of the basis system (usually predefined $p = 3$) and $\kappa_1, \ldots \kappa_K$ is a set of knots (usually with $K \in \{5, \ldots, 35\}$)

- common choices for knot positioning: either use sample quantiles or equally-spaced

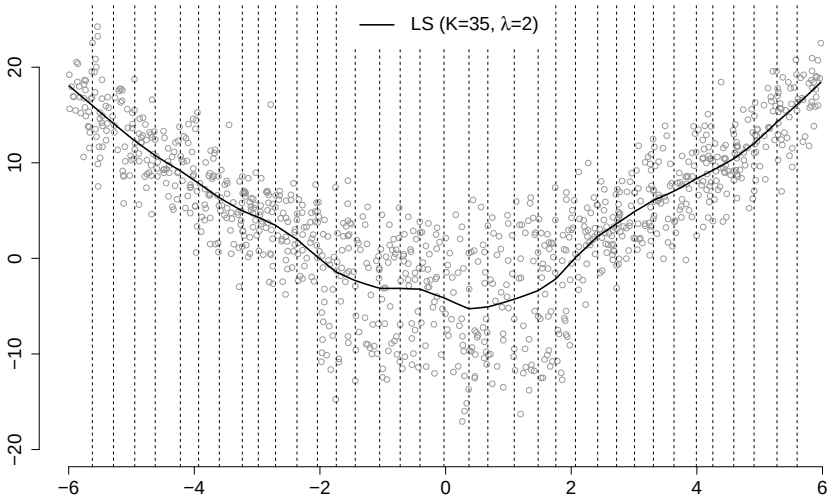
Motivation: artificial data (1,000 obs.) and results of LS-, S-, and M-est.



Motivation: results with 45% (144) random outliers in [-2,2]

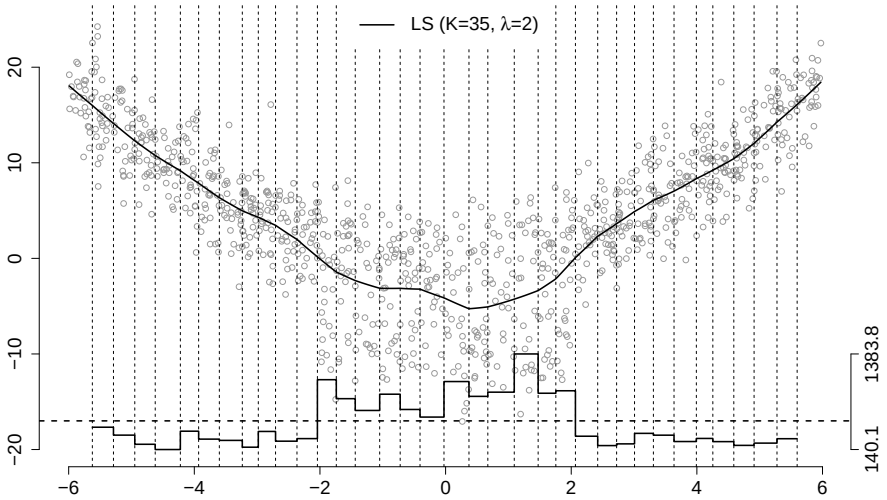


Knot deletion: knots based on sample quantiles (status quo)

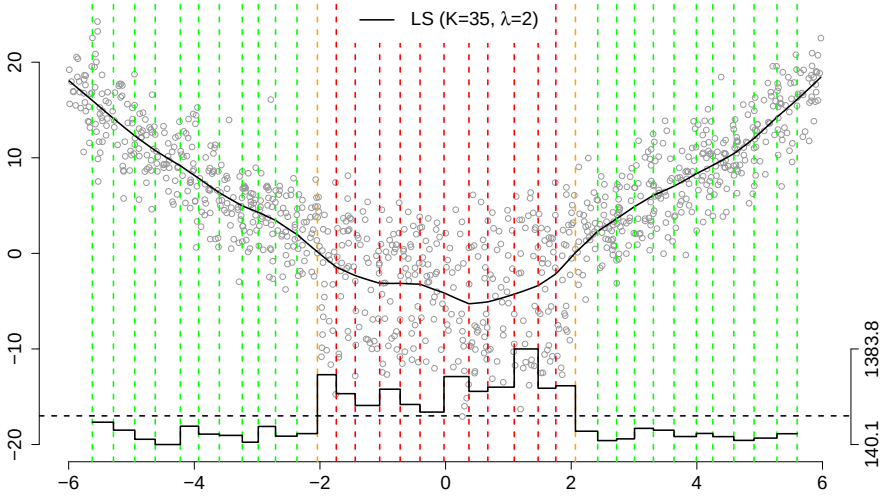


Knot deletion: calculating the RSS within each quantile region

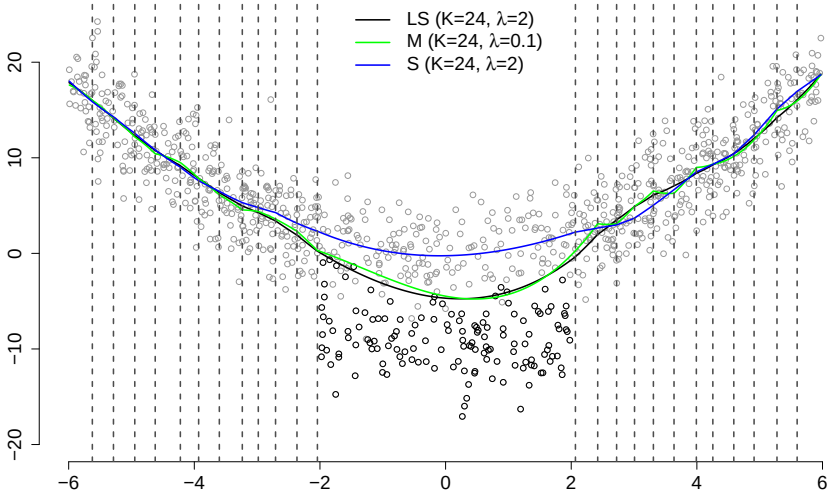
cut-off: $cheby = median(RSS_1, \ldots, RSS_{K+1}) + \sqrt{\frac{K^2+2K}{(K+1)^2 \cdot 0.95 - K - 1}} \cdot MAD(RSS_1, \ldots, RSS_{K+1})$



Knot deletion: potential contaminated regions



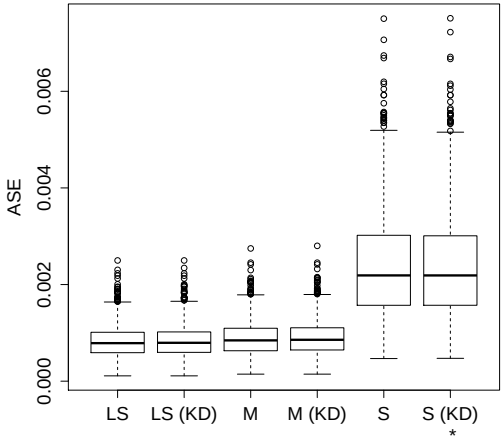
Knot deletion: final results after knot deletion



Simulation setting

true function:	$m = \sin(\pi \cdot x)$
number of observations:	$n = 1{,}000$ in $[-1, 1]$
error distribution:	$\mathcal{N}(0, 0.3^2)$
outlier interval:	$[-0.3, 0.3]$
outlier distribution:	$\mathcal{N}(\{-1, -2\}, 0.1^2)$
contamination rate:	$\{0\%, 45\%, 55\%\}$
basis system:	truncated polynomial basis of order $p = 3$
number of knots:	$K = 35$
replications per setting:	$1{,}000$
estimators:	LS, M, S
lambda:	robust (generalized) cross validation
performance measure:	average squared error (ASE)

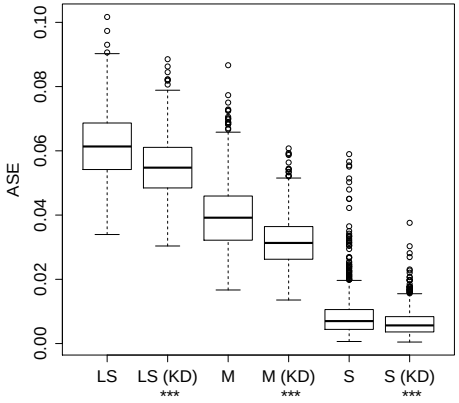
Results: no contamination



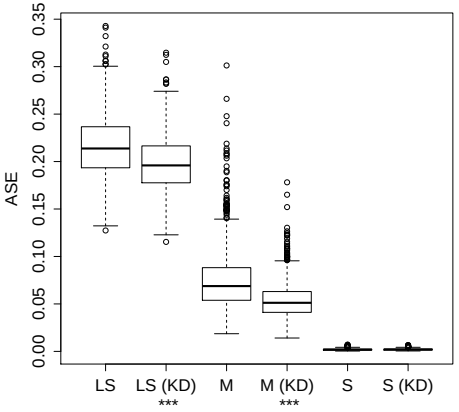
* indicates a significant result ($p < 0.05$) of Wilcoxon's signed rank test

Results: 45% contamination

mean level of outliers: -1



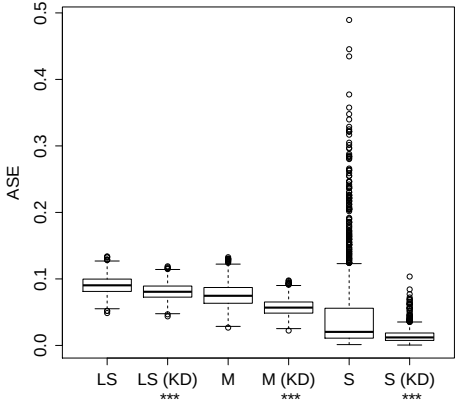
mean level of outliers: -2



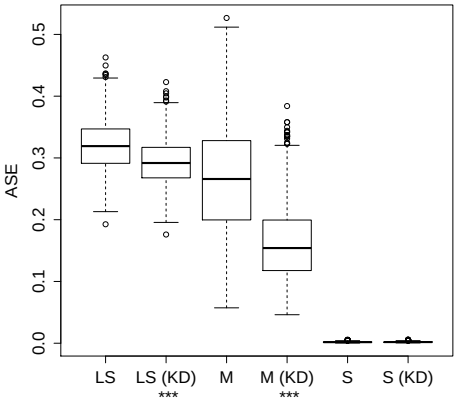
*** indicates a highly significant result ($p < 0.001$) of Wilcoxon's signed rank test

Results: 55% contamination

mean level of outliers: -1



mean level of outliers: -2



*** indicates a highly significant result ($p < 0.001$) of Wilcoxon's signed rank test

Conclusion/Further Developments

- spline fit can be significantly improved (in contaminated situations) by applying a simple knot deletion scheme
- no adverse effects in situations without contamination
- almost no additional computational costs
- similar results (not shown in here) are achieved when using knot repositioning instead of knot deletion
- further research: combination of knot deletion and knot adding (see e.g. Yao and Lee 2008 for the latter)

Questions

Literature

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):pp. 89–102.

Lee, T. and Oh, H.-S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics*, 22(1):159–171.

Liebscher, S. and Kirschstein, T. (2015). Efficiency of the pmst and rdela location and scatter estimators. *AStA Advances in Statistical Analysis*, 99(1):63–82.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Tharmaratnam, K., Claeskens, G., Croux, C., and Salibian-Barrera, M. (2010). S-estimation for penalized regression splines. *Journal of Computational and Graphical Statistics*, 19(3):609–625.

Yao, F. and Lee, T. C. (2008). On knot placement for penalized spline regression. *Journal of the Korean Statistical Society*, 37(3):259 – 267.