Non-parametric regression using splines, with applications

Lecture dedicated to the memory of Milcho Tsvetkov

Ognyan Kounchev and Georgi Simeonov

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

13th BSAC, Velingrad, October 3-6, 2022

ACKNOWLEDGEMENTS

Sponsored by grants with Bulgarian NSF (KP-06- N32-8, KP-06-N52-1, and KP-06N42-2), and by the Alexander von Humboldt Foundation.
Based on joint research with H. Render, Ts. Tsachev.

Applications of splines to Astronomy and Astrophysics

- Akerlof, C. et al., **Application of Cubic Splines to the Spectral Analysis of Unequally Spaced Data**, 1994.

Applications of splines to Astronomy and Astrophysics

- Akerlof, C. et al., **Application of Cubic Splines to the Spectral Analysis of Unequally Spaced Data**, 1994.
- Kundiukov, S. G. ; Nazarenko, N. B., **Application of smoothing splines in processing the amplitude-time characteristics of radio meteors**, 1988.

Applications of splines to Astronomy and Astrophysics

- Akerlof, C. et al., **Application of Cubic Splines to the Spectral Analysis of Unequally Spaced Data**, 1994.
- Kundiukov, S. G. ; Nazarenko, N. B., **Application of smoothing splines in processing the amplitude-time characteristics of radio meteors**, 1988.
- V. A. Baturin, W. Däppen, A. V. Oreshina, S. V. Ayukov and A. B. Gorshkov, **Interpolation of equation-of-state data**, A&A, Volume 626, June 2019.

Applications of splines to Astronomy and Astrophysics

- Akerlof, C. et al., **Application of Cubic Splines to the Spectral Analysis of Unequally Spaced Data**, 1994.
- Kundiukov, S. G. ; Nazarenko, N. B., **Application of smoothing splines in processing the amplitude-time characteristics of radio meteors**, 1988.
- V. A. Baturin, W. Däppen, A. V. Oreshina, S. V. Ayukov and A. B. Gorshkov, **Interpolation of equation-of-state data**, A&A, Volume 626, June 2019.
- Collin A. Politsch, Jessi Cisewski-Kehe, Rupert A. C. Croft, and Larry Wasserman, **Trend Filtering – I. A Modern Statistical Tool for Time-Domain Astronomy and Astronomical Spectroscopy**, 2020

A special non-parametric model - Cubic splines S(x) - a reminder

- $S(x)$ is a piecewise cubic polynomial in every interval (x_i, x_{i+1}), where $a = x_1$ and $b = x_n$, and the **knots** x_j satisfy

$$a = x_1 < x_2 < \cdots < x_n = b$$

A special non-parametric model - Cubic splines S(x) - a reminder

- $S(x)$ is a piecewise cubic polynomial in every interval (x_i, x_{i+1}), where $a = x_1$ and $b = x_n$, and the **knots** x_j satisfy

$$a = x_1 < x_2 < \cdots < x_n = b$$

- $S \in C^2$ on the whole interval $[a, b]$

A special non-parametric model - Cubic splines S(x) - a reminder

- $S(x)$ is a piecewise cubic polynomial in every interval (x_i, x_{i+1}), where $a = x_1$ and $b = x_n$, and the **knots** x_j satisfy

$$a = x_1 < x_2 < \cdots < x_n = b$$

- $S \in C^2$ on the whole interval $[a, b]$
- some boundary conditions at a and b are added; e.g. Natural BC – in this case the splines are called **Natural**.

A special non-parametric model - Cubic splines S(x) - a reminder

- $S(x)$ is a piecewise cubic polynomial in every interval (x_i, x_{i+1}), where $a = x_1$ and $b = x_n$, and the **knots** x_j satisfy

$$a = x_1 < x_2 < \cdots < x_n = b$$

- $S \in C^2$ on the whole interval $[a, b]$
- some boundary conditions at a and b are added; e.g. Natural BC – in this case the splines are called **Natural**.
- **THEOREM**. For every set of interpolation data $\{f_i\}_{i=1}^n$ defined at $\{x_i\}_{i=1}^n$ there exists a unique (Natural) spline $S(x)$ with breaks at $\{x_i\}$ s.t.

$$S(x_i) = f_i \qquad \text{for } i = 1, 2, ..., n.$$

It is called interpolation spline to the data $\{f_i\}$.

A special non-parametric model - Cubic splines S(x) - a reminder

- $S(x)$ is a piecewise cubic polynomial in every interval (x_i, x_{i+1}), where $a = x_1$ and $b = x_n$, and the **knots** x_j satisfy

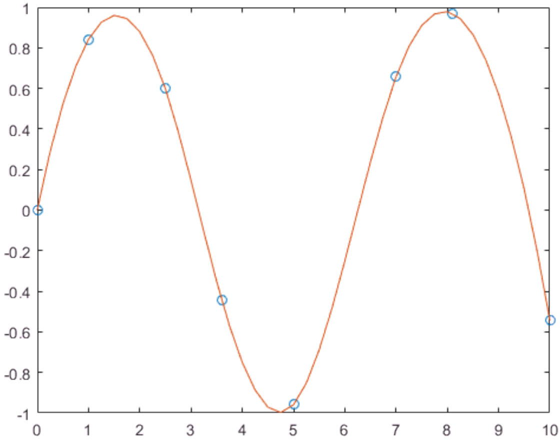
$$a = x_1 < x_2 < \cdots < x_n = b$$

- $S \in C^2$ on the whole interval $[a, b]$
- some boundary conditions at a and b are added; e.g. Natural BC – in this case the splines are called **Natural**.
- **THEOREM**. For every set of interpolation data $\{f_i\}_{i=1}^n$ defined at $\{x_i\}_{i=1}^n$ there exists a unique (Natural) spline $S(x)$ with breaks at $\{x_i\}$ s.t.

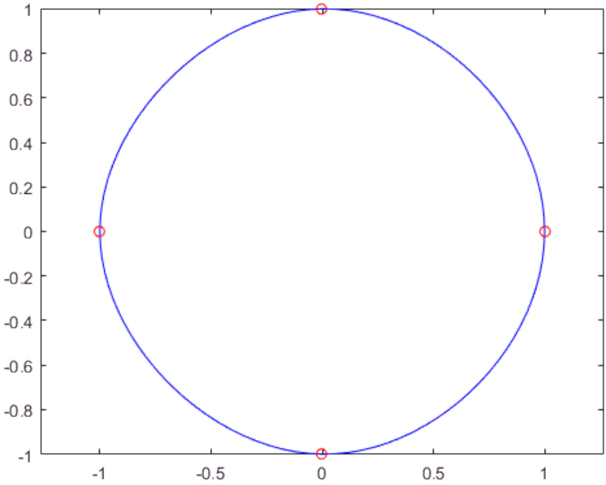
$$S(x_i) = f_i \qquad \text{for } i = 1, 2, ..., n.$$

  It is called interpolation spline to the data $\{f_i\}$.
- References: Sommerfeld (1903), de Boor (1978, 2001), Stoer-Bulirsch (1998), Green-Silverman (1994).

Why are polynomial splines good? An example - the sin function



Example - the circle



Fast algorithms for computation of interpolation cubic splines

- Fast algorithms exist for large amount of data (cf. **Wahba** 1990, **Green-Silverman** 1994)

The Smoothing cubic spline - Finding trends

- Assume data values $\mathbf{Y} = \{Y_j\}$ measured at $x_j \in [a, b]$, for $j = 1, ..., N$

The Smoothing cubic spline - Finding trends

- Assume data values $\mathbf{Y} = \{Y_j\}$ measured at $x_j \in [a, b]$, for $j = 1, ..., N$
- We consider the penalized functional

$$S(g) = \sum_{j=1}^{N} (g(x_j) - Y_j)^2 + \lambda \int_a^b |g''(t)|^2 \, dt$$

to avoid "wiggling" typical also for polynomials!!!

The Smoothing cubic spline - Finding trends

- Assume data values $\mathbf{Y} = \{Y_j\}$ measured at $x_j \in [a, b]$, for $j = 1, ..., N$
- We consider the penalized functional

$$S(g) = \sum_{j=1}^{N} (g(x_j) - Y_j)^2 + \lambda \int_a^b |g''(t)|^2 \, dt$$

to avoid "wiggling" typical also for polynomials!!!

- **THEOREM**. The solution to problem

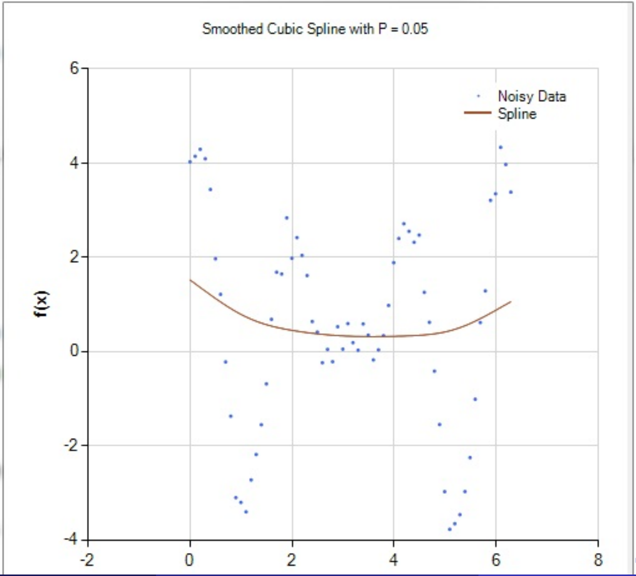
$$\min_g S(g) \qquad \text{where } g \in C^2(a, b)$$

is a cubic spline, with knots $\{x_j\}$ and interpolation data

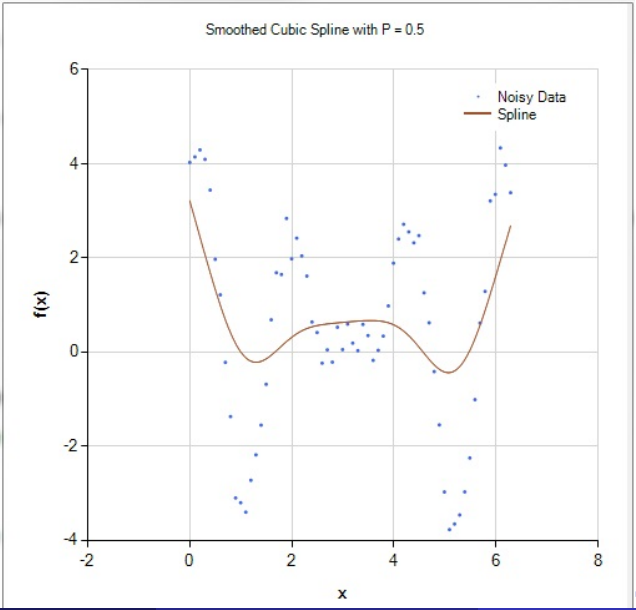
$$\mathbf{g} = (I + \lambda K)^{-1} \mathbf{Y}$$

where $K = QR^{-1}Q^T$.

Examples of smoothing splines with different lambda; here lambda = 0.95

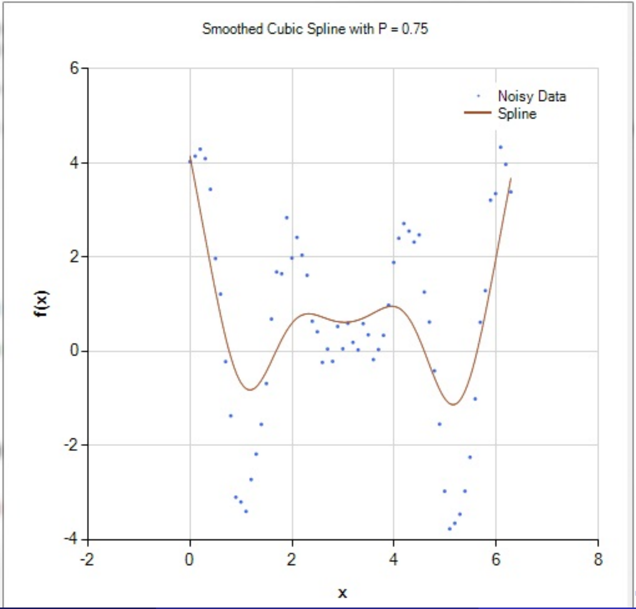


lambda is 0.5

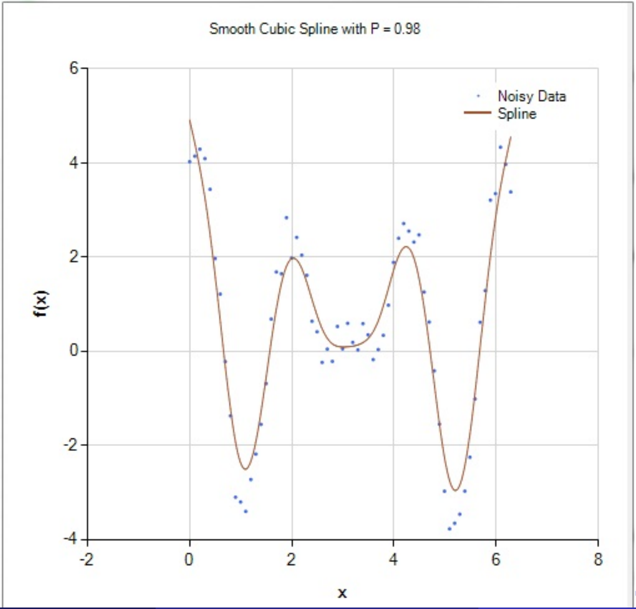


Smoothed Cubic Spline with P = 0.5
Noisy Data
Spline
f(x)
x
6
4
2
0
-2
-4
-2
0
2
4
6
8

lambda is 0.25 - more wiggling



lambda is 0.02 - very wiggling



The fast (O(n) time) Reinsch algorithm (1971)

FACT: There exists a fast algorithm of Reinsch for the computation of the smoothing splines. Reference: Stoer-Bulirsch, Numerical Analysis, Springer, 2010.

Cross Validation for finding parameter lambda

- Let $\lambda > 0$ be fixed.

Cross Validation for finding parameter lambda

- Let $\lambda > 0$ be fixed.
- Let $\widehat{g}^{(-i)}(t;\lambda)$ be a solution to the minimization problem

$$\min_{g} \sum_{j \neq i} \left(Y_j - g(t_j) \right)^2 + \lambda \int \left| g''(t) \right|^2 dt$$

Cross Validation for finding parameter lambda

- Let $\lambda > 0$ be fixed.
- Let $\widehat{g}^{(-i)}(t;\lambda)$ be a solution to the minimization problem

$$\min_{g} \sum_{j \neq i} (Y_j - g(t_j))^2 + \lambda \int |g''(t)|^2 \, dt$$

- The cross-validation (leave-one-out) score function is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{g}^{(-i)}(t_i;\lambda) \right)^2$$

Cross Validation for finding parameter lambda

- Let $\lambda > 0$ be fixed.
- Let $\widehat{g}^{(-i)}(t;\lambda)$ be a solution to the minimization problem

$$\min_g \sum_{j \neq i} (Y_j - g(t_j))^2 + \lambda \int |g''(t)|^2 \, dt$$

- The cross-validation (leave-one-out) score function is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{g}^{(-i)}(t_i;\lambda) \right)^2$$

- We **minimize** $CV(\lambda)$ to find λ.

The representation of Cross-Validation and GCV

- **THEOREM**: We have

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{g}(t_i;\lambda)}{1 - A_{ii}(\lambda)}\right)^2$$

The representation of Cross-Validation and GCV

- **THEOREM**: We have

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{g}(t_i;\lambda)}{1 - A_{ii}(\lambda)}\right)^2$$

- here the matrix

$$A(\lambda) = \left(I + \lambda Q R^{-1} Q^T\right)^{-1}$$

and its diagonal elements A_{ii} may be computed in a **FAST** way, for details see G. Wahba (1990) and Green-Silverman (1994).

The representation of Cross-Validation and GCV

- **THEOREM**: We have

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{g}(t_i; \lambda)}{1 - A_{ii}(\lambda)} \right)^2$$

- here the matrix

$$A(\lambda) = \left(I + \lambda Q R^{-1} Q^T \right)^{-1}$$

  and its diagonal elements A_{ii} may be computed in a **FAST** way, for details see G. Wahba (1990) and Green-Silverman (1994).
- Similar formula for Generalized Cross Validation - see the same references

Spline model - nodes of spline differ from data points

- This is a more complicated stuff - there may be gaps of the data

Spline model - nodes of spline differ from data points

- This is a more complicated stuff - there may be gaps of the data
- von Golitscheck - L. Schumaker

Multidimensional case

- Extremely large area of applications - Earth Observations (EO), Meteorology, Medicine, Finance (Volatility Surface), etc.

Multidimensional case

- Extremely large area of applications - Earth Observations (EO), Meteorology, Medicine, Finance (Volatility Surface), etc.
- What about Smoothing methods? error estimates, Conf. intervals, etc. ?

Multidimensional case

- Extremely large area of applications - Earth Observations (EO), Meteorology, Medicine, Finance (Volatility Surface), etc.
- What about Smoothing methods? error estimates, Conf. intervals, etc. ?
- Thin plate splines (TPS) in Wahba (1990) ;

Multidimensional case

- Extremely large area of applications - Earth Observations (EO), Meteorology, Medicine, Finance (Volatility Surface), etc.
- What about Smoothing methods? error estimates, Conf. intervals, etc. ?
- Thin plate splines (TPS) in Wahba (1990) ;
- Also, in Green-Silverman (1994):
  with Thin plate splines "**some**, but not all, of the attractive features of spline smoothing in one dimension carry over."

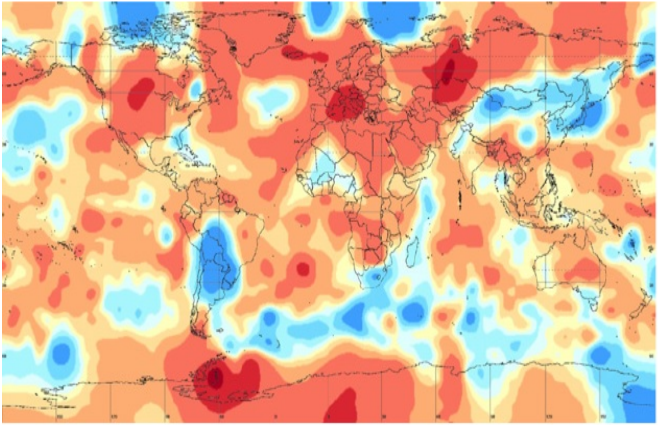
Multidimensional case

- Extremely large area of applications - Earth Observations (EO), Meteorology, Medicine, Finance (Volatility Surface), etc.
- What about Smoothing methods? error estimates, Conf. intervals, etc. ?
- Thin plate splines (TPS) in Wahba (1990) ;
- Also, in Green-Silverman (1994):
  with Thin plate splines "**some**, but not all, of the attractive features of spline smoothing in one dimension carry over."
- In Ramsay-Silverman (2005), chapter 22.2.3 Multidimensional arguments:
  "Although we have touched multivariate functions of a single argument *t*, coping with more than one dimension **in the domain** of our functions has been mainly **beyond our scope**."

Multidimensional case

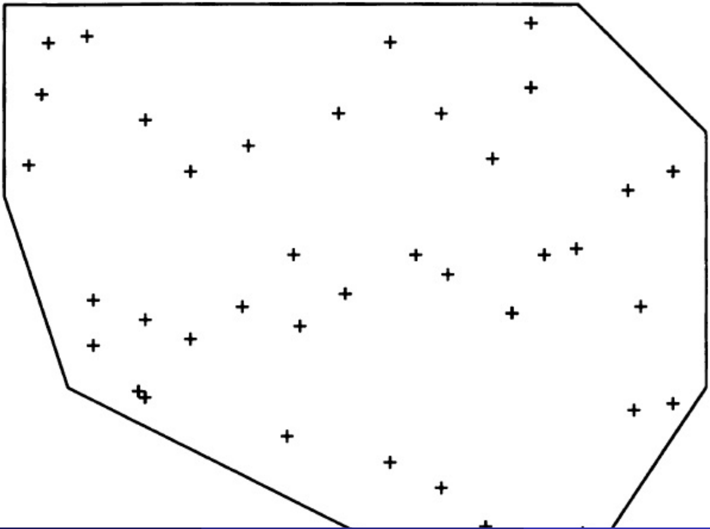
- Extremely large area of applications - Earth Observations (EO), Meteorology, Medicine, Finance (Volatility Surface), etc.
- What about Smoothing methods? error estimates, Conf. intervals, etc. ?
- Thin plate splines (TPS) in Wahba (1990) ;
- Also, in Green-Silverman (1994):
  with Thin plate splines "**some**, but not all, of the attractive features of spline smoothing in one dimension carry over."
- In Ramsay-Silverman (2005), chapter 22.2.3 Multidimensional arguments:
  "Although we have touched multivariate functions of a single argument t, coping with more than one dimension **in the domain** of our functions has been mainly **beyond our scope**."
- One may use also RBFs, Kriging, Minimum Curvature, Shepard's method, etc. And our approach - POLYSPLINES.

Smoothed data - an example



Example of Multidimensional Scattered data set

- Importance for life problems even in dimension 2 – data of Earth Observations,



The generalized L-splines - the main bricks of the Polysplines

- Instead of 1D polynomials we use piecewise exponential functions called $L-$splines. A special case: fix ξ, then the $L-$spline is defined as a piecewise solution in every interval $[x_j, x_{j+1}]$ of the equation:

$$L_\xi f(t) = 0 \qquad \text{with } L_\xi = \left(\frac{\partial^2}{\partial t^2} - \xi^2\right)^2$$

which is C^2 at the knots x_j; the basis of solutions are $e^{t\xi}, te^{t\xi}, e^{-t\xi}, te^{-t\xi}$, while for the classical case are $1, t, t^2, t^3$.

The generalized L-splines - the main bricks of the Polysplines

- Instead of 1D polynomials we use piecewise exponential functions called $L-$splines. A special case: fix ξ, then the $L-$spline is defined as a piecewise solution in every interval $[x_j, x_{j+1}]$ of the equation:

$$L_\xi f(t) = 0 \qquad \text{with } L_\xi = \left(\frac{\partial^2}{\partial t^2} - \xi^2 \right)^2$$

  which is C^2 at the knots x_j; the basis of solutions are $e^{t\xi}, te^{t\xi}, e^{-t\xi}, te^{-t\xi}$, while for the classical case are $1, t, t^2, t^3$.
- A much bigger generalization: Consider a polynomial L of degree 4 and the solutions of the related differential operator

$$L\left(\frac{\partial}{\partial t}\right) f(t) = 0$$

The generalized L-splines - the main bricks of the Polysplines

- Instead of 1D polynomials we use piecewise exponential functions called $L-$splines. A special case: fix ξ, then the $L-$spline is defined as a piecewise solution in every interval $[x_j, x_{j+1}]$ of the equation:

$$L_\xi f(t) = 0 \qquad \text{with } L_\xi = \left(\frac{\partial^2}{\partial t^2} - \xi^2\right)^2$$

which is C^2 at the knots x_j; the basis of solutions are $e^{t\xi}, te^{t\xi}, e^{-t\xi}, te^{-t\xi}$, while for the classical case are $1, t, t^2, t^3$.

- A much bigger generalization: Consider a polynomial L of degree 4 and the solutions of the related differential operator

$$L\left(\frac{\partial}{\partial t}\right) f(t) = 0$$

- In the case of real coefficients of the polynomial L with four different roots a_j the basis of all solutions is given by the exponential functions $e^{a_j t}$.

Examples of L-splines

- Interpolation and smoothing $L-$splines of the special form depending on ξ were considered exhaustively, with fast algorithms in a paper **"On a class of L-splines of order 4: fast algorithms for interpolation and smoothing"**, BIT Numerical Mathematics, 2020. They have as basis the exponential functions $e^{\xi t}, te^{\xi t}, e^{-\xi t}, te^{-\xi t}$.

Examples of L-splines

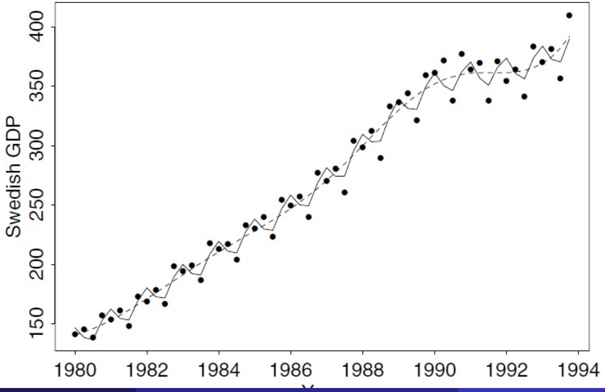
- Interpolation and smoothing $L-$splines of the special form depending on ξ were considered exhaustively, with fast algorithms in a paper **"On a class of L-splines of order 4: fast algorithms for interpolation and smoothing"**, BIT Numerical Mathematics, 2020. They have as basis the exponential functions $e^{\xi t}, te^{\xi t}, e^{-\xi t}, te^{-\xi t}$.
- These 1D $L-$splines are **important for the multidimensional theory of polysplines**.

Examples of L-splines

- Interpolation and smoothing $L-$splines of the special form depending on ξ were considered exhaustively, with fast algorithms in a paper **"On a class of L-splines of order 4: fast algorithms for interpolation and smoothing"**, BIT Numerical Mathematics, 2020. They have as basis the exponential functions $e^{\xi t}, te^{\xi t}, e^{-\xi t}, te^{-\xi t}$.
- These 1D $L-$splines are **important for the multidimensional theory of polysplines**.
- The case of more general $L-$splines of order 4 is considered in a more recent paper **"Fast algorithms for interpolation with L-splines for differential operators L of order 4 with constant coefficients"** , in ARXIV, submitted in J. Comp. and Applied Maths.

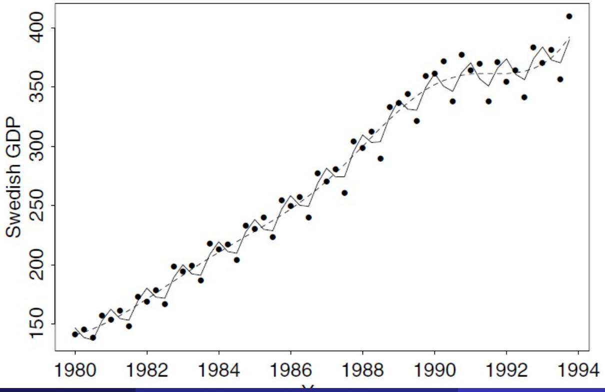
Further motivating examples to study smoothing L-splines (and exponential splines)

- GDP for Sweden with seasonal variation (in Ramsay-Silverman, 2005) – a cyclic effect superimposed on a linear development



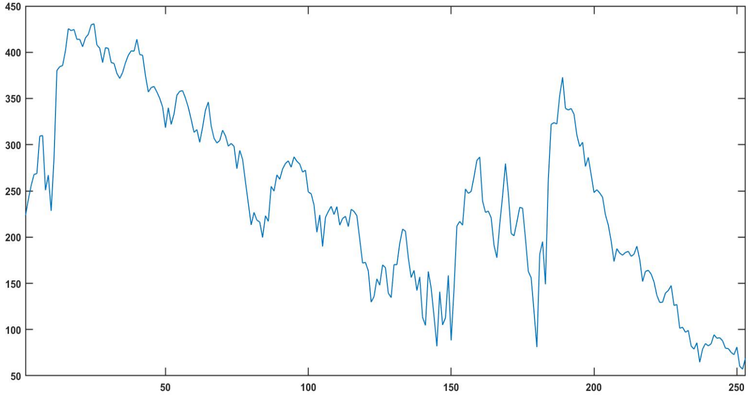
Further motivating examples to study smoothing L-splines (and exponential splines)

- GDP for Sweden with seasonal variation (in Ramsay-Silverman, 2005) – a cyclic effect superimposed on a linear development
- the dashed line is Cubic smoothing (with GCV for λ), and the solid line is a smoothing $L-$spline with $L = \left(-\gamma \frac{d}{dt} + \frac{d^2}{dt^2}\right)\left(\omega^2 + \frac{d^2}{dt^2}\right)$.



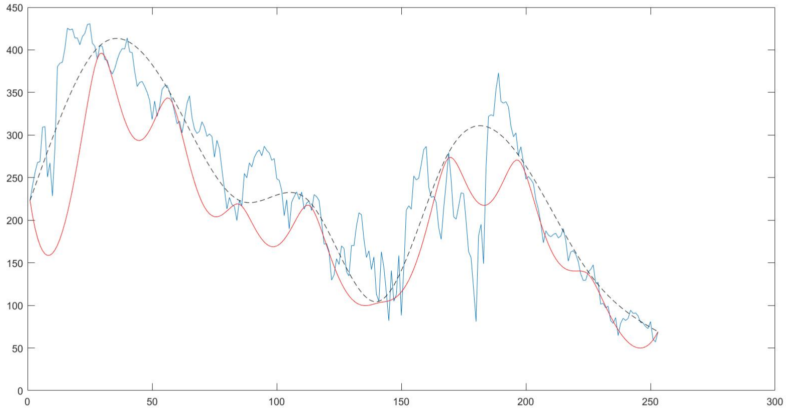
Examples of smoothing L-splines - S&P 500 data

- Daily S&P500 prices for the period 24 October, 2017 – 24 October, 2018, total 253 days.



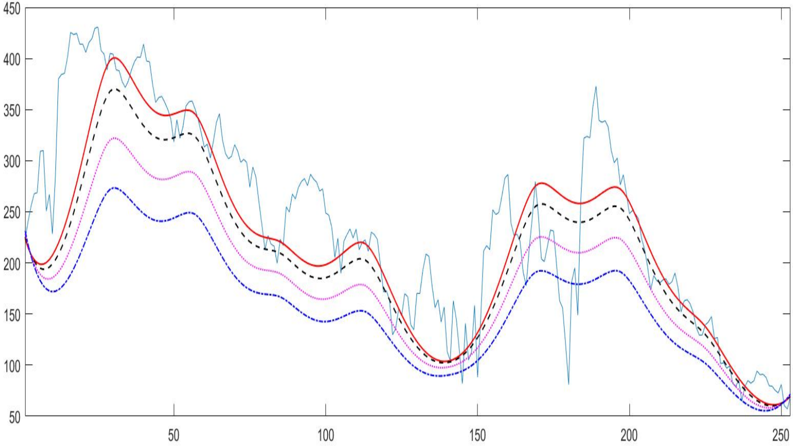
Smoothing results for the operator L_xi

- for $N = 10$ knots; $\lambda = 3$, $\xi = 0.01$ (dash) and $\xi = 0,13$:



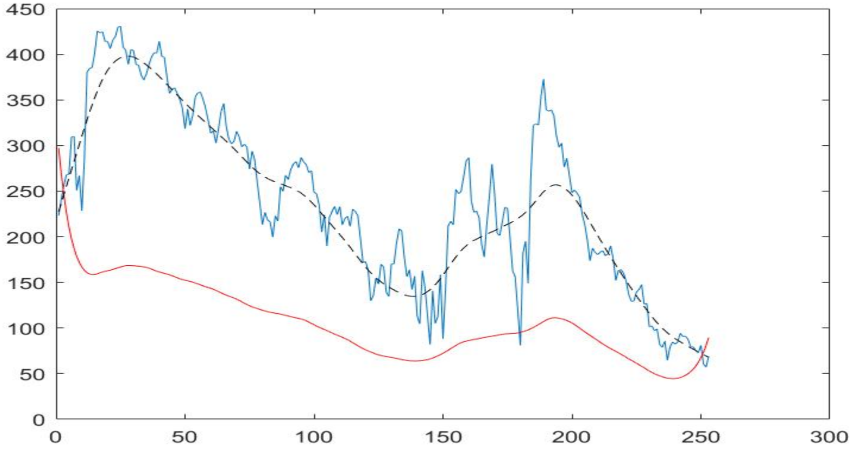
Cont'd

- for $N = 10$ knots; $\lambda = 5, 30, 80, 150$, and $\xi = 0.13$.

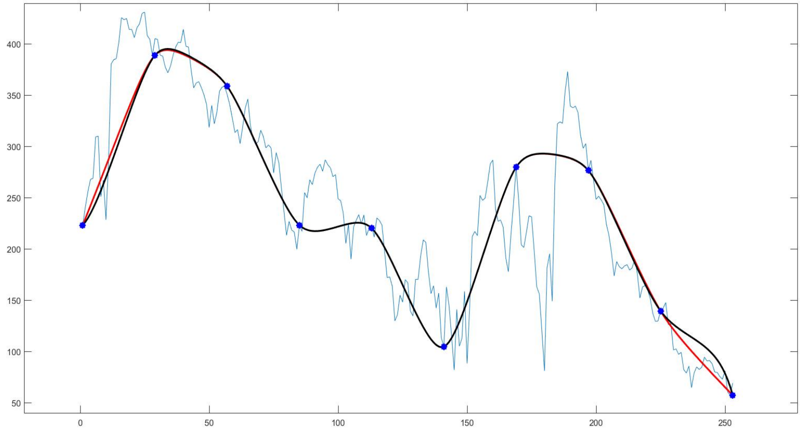


Cont'd

- for $N = 30$ knots; $\lambda = 500$, and $\xi = 0.01$ and $\xi = 0.13$:



The new $L-$splines on the S&P500 data



The new $L-$splines - some subtleties

- The splines in the Figure above are two different $L-$splines although the same differential operators.

The new $L-$splines - some subtleties

- The splines in the Figure above are two different $L-$splines although the same differential operators.
- The first polynomial L has the 4 different zeros $(-0.01; 0.01; 0.20; -0.20)$ and is "natural spline"

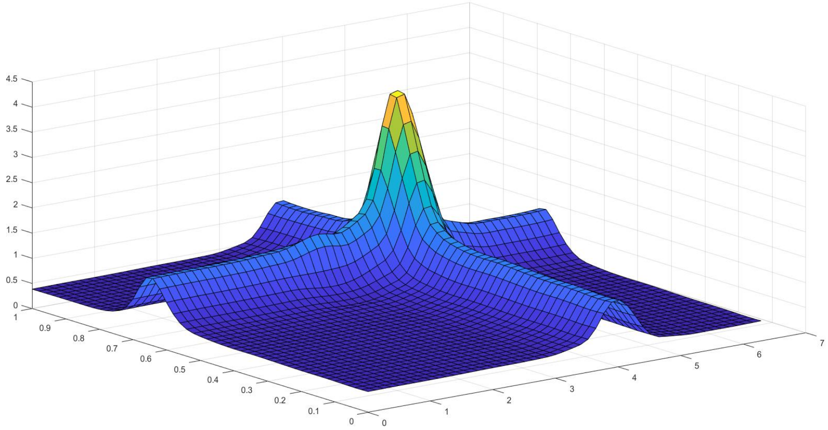
The new $L-$splines - some subtleties

- The splines in the Figure above are two different $L-$splines although the same differential operators.
- The first polynomial L has the 4 different zeros $(-0.01; 0.01; 0.20; -0.20)$ and is "natural spline"
- The second has the same set of zeros $(-0.01; 0.20; 0.01; -0.20)$ but is a different "natural spline"

The new $L-$splines - some subtleties

- The splines in the Figure above are two different $L-$splines although the same differential operators.
- The first polynomial L has the 4 different zeros $(-0.01; 0.01; 0.20; -0.20)$ and is "natural spline"
- The second has the same set of zeros $(-0.01; 0.20; 0.01; -0.20)$ but is a different "natural spline"
- Polsyplines are just one step forth

Polyspline interpolating 2D Titanium data at 70 points



References

- G. Wahba, Spline Models for Observational Data, SIAM, 1990.

References

- G. Wahba, Spline Models for Observational Data, SIAM, 1990.
- P. Green, B. Silverman, Nonparametric regression and generalized linear models, Chapman and Hall, 1994.

References

- G. Wahba, Spline Models for Observational Data, SIAM, 1990.
- P. Green, B. Silverman, Nonparametric regression and generalized linear models, Chapman and Hall, 1994.
- Ramsay, Silverman, 2005, Functional Data Analysis

References

- G. Wahba, Spline Models for Observational Data, SIAM, 1990.
- P. Green, B. Silverman, Nonparametric regression and generalized linear models, Chapman and Hall, 1994.
- Ramsay, Silverman, 2005, Functional Data Analysis
- Gu, Ch. , Smoothing Spline ANOVA Models, Springer, 2013.

References

- G. Wahba, Spline Models for Observational Data, SIAM, 1990.
- P. Green, B. Silverman, Nonparametric regression and generalized linear models, Chapman and Hall, 1994.
- Ramsay, Silverman, 2005, Functional Data Analysis
- Gu, Ch. , Smoothing Spline ANOVA Models, Springer, 2013.
- Hastie, Tibshirani, Friedman, The elements of statistical learning: Data Mining, Inference, and Prediction, 2009

The end

- THANK YOU !