STAT 824 sp 2023 Lec 06 slides

Nonparametric regression: Penalized splines and trend filtering

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Table of Contents

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be indep. realizations of $(X, Y) \in [0, 1] \times \mathbb{R}$, where

$$Y = m(X) + \varepsilon, \qquad \text{for some } m : [0, 1] \to \mathbb{R},$$

where ε is independent of X with $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \sigma^2$.

Define the class of functions (which is a Sobolev space)

$$\mathcal{W}_2 = \left\{ g : [0,1] \to \mathbb{R} : g' \text{ is continuous, } \int_0^1 [g''(x)]^2 dx < \infty \right\}.$$

Note that $\mathcal{H}(2, L)$ on $[0,1]$ is contained in $\mathcal{W}_2$.

Smoothing spline estimator

The estimator

$$\hat{m}_n^{\text{sspl}} = \underset{g \in \mathcal{W}_2}{\text{argmin}} \sum_{i=1}^{n} [Y_i - g(X_i)]^2 + \lambda \int_0^1 [g''(x)]^2 dx,$$

for $\lambda > 0$ is called the *smoothing spline estimator* of m.

How do we search among all functions belonging to $\mathcal{W}_2$?

Beautiful result: $\hat{m}_n^{\text{sspl}}$ is a natural cubic spline with knots $u_i = X_i$, $i = 1, \ldots, n$.

Natural cubic splines are cubic splines constrained to be linear beyond end knots.

Can bound MSE by $C \cdot n^{-4/5}$ when $\lambda = c \cdot n^{1/5}$. See Grace Wahba's book, [4].

Sets of basis functions for natural cubic splines lack nice properties of B-splines.

Since we love B-splines, we often consider (instead of smoothing splines) this:

Penalized spline estimator

Let $\mathcal{M}_{n,3}$ be the space of cubic splines on the knots

$$u_{-3} = u_{-2} = u_{-1} = u_0 < u_1 < \cdots < u_{K_n} = u_{K_n+1} = u_{K_n+2} = u_{K_n+3}.$$

Then

$$\hat{m}_n^{\text{pspl}} = \underset{g \in \mathcal{M}_{n,3}}{\text{argmin}} \sum_{i=1}^{n} [Y_i - g(X_i)]^2 + \lambda \int_0^1 [g''(x)]^2 dx,$$

for $\lambda > 0$ is the *penalized spline estimator* of m. Nice reference is [1].

Idea is to choose K_n very large and then tune wiggliness by choosing λ.

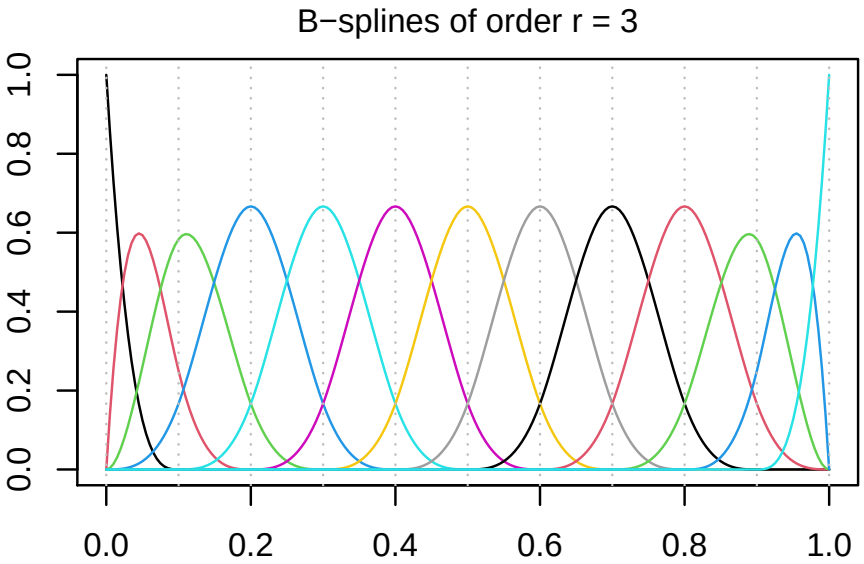
When K_n is very large, $\hat{m}_n^{\text{pspl}}$ is practically identical to $\hat{m}_n^{\text{sspl}}$.

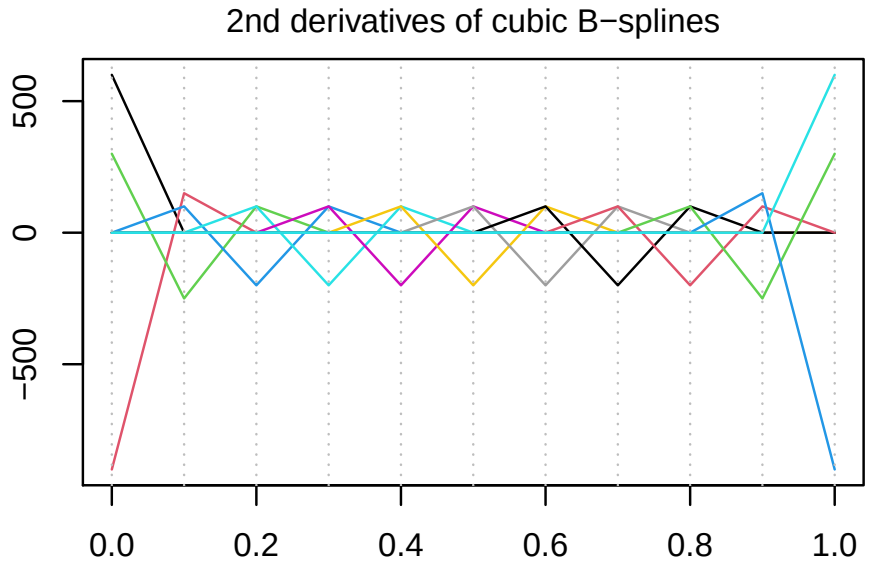
Exercise: Give a representation of $\hat{m}_n^{\text{pspl}}(x_0)$ in matrices given a basis for $\mathcal{M}_{n,3}$.

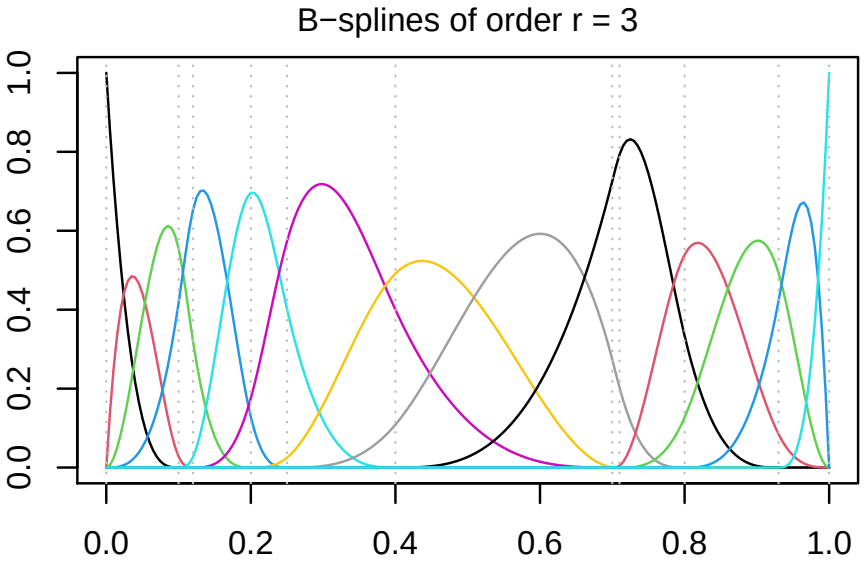
We can obtain the row vector $\mathbf{b}''(x) = (b_1''(x), \ldots, b_d''(x))$ with

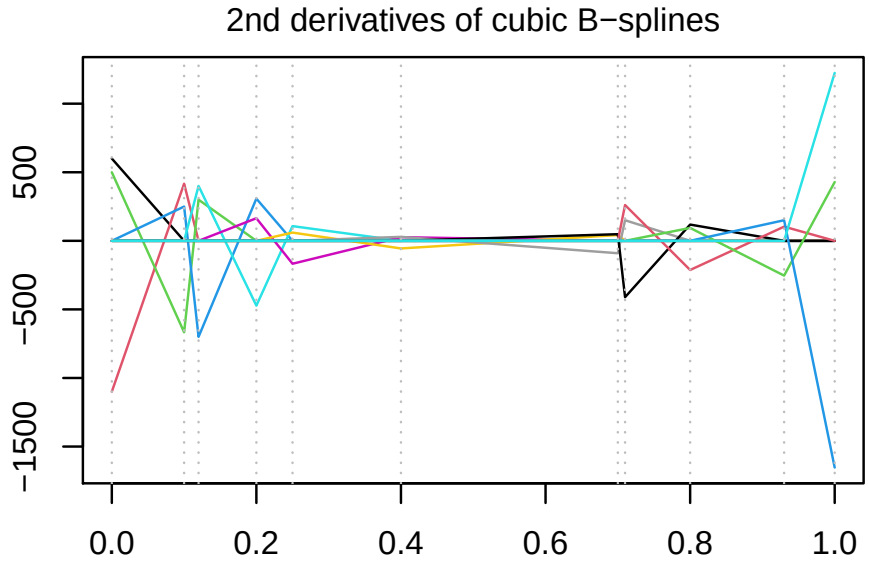
```
splineDesign(knots=knots,x=x,ord=4,derivs=rep(2,K+1))
```

where `knots` is the complete set of knots $u_{-3}, \ldots, u_{K+3}$.


B−splines of order r = 3
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


2nd derivatives of cubic B−splines
500
0
−500
0.0
0.2
0.4
0.6
0.8
1.0


B−splines of order r = 3
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


2nd derivatives of cubic B−splines
500
−500
−1500
0.0
0.2
0.4
0.6
0.8
1.0

Computation of $\boldsymbol{\Omega}$

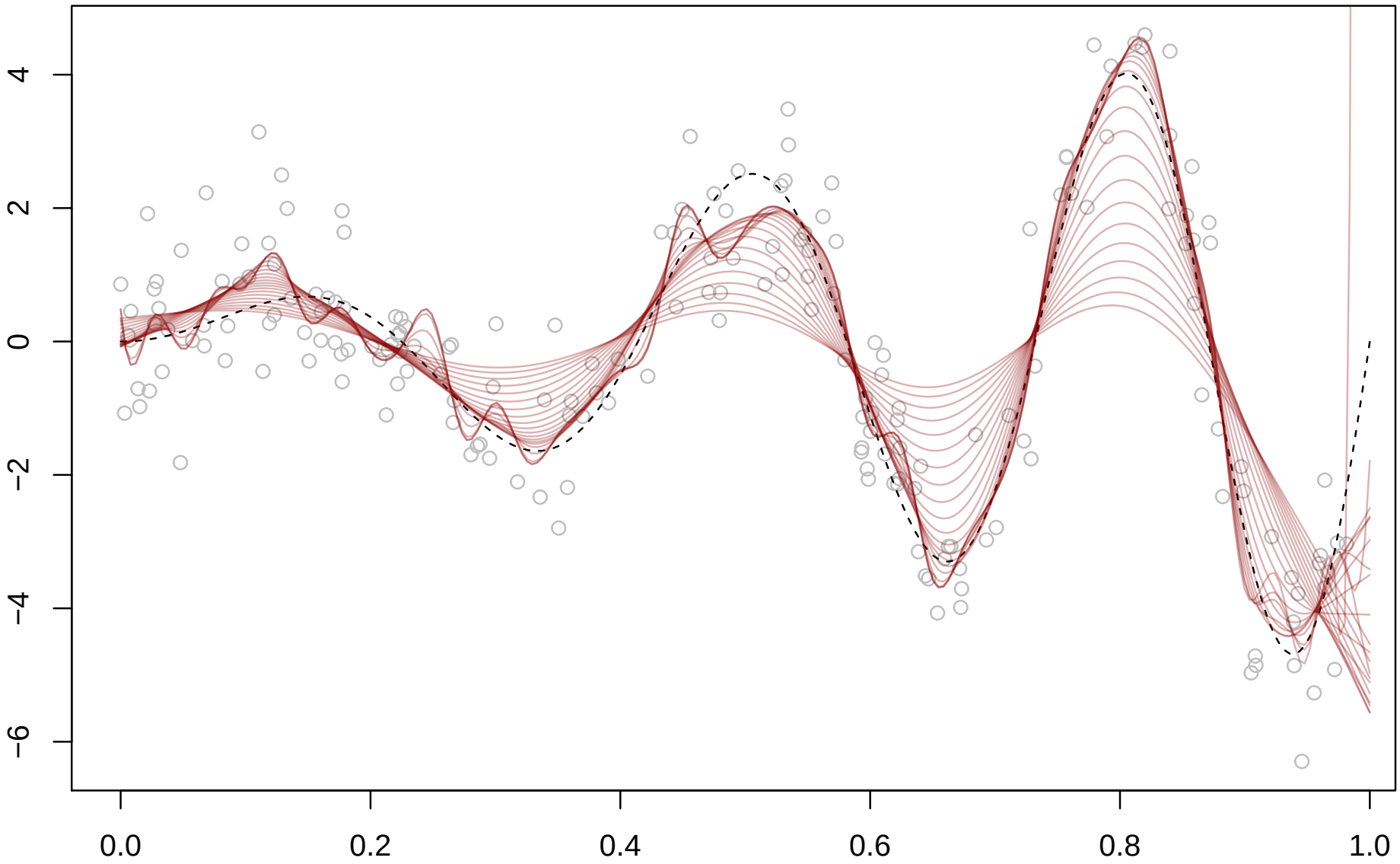
For cubic B-splines basis functions $b_1, \ldots, b_{d_n}$, $d_n = K_n + 3$, based on knots $u_{-3} = u_{-2} = u_{-1} = u_0 < u_1 < \cdots < u_{K_n} = u_{K_n+1} = u_{K_n+2} = u_{K_n+3}$, we have

$$\int_0^1 b_\ell''(x) b_j''(x) dx = \sum_{k=0}^{K-1} (u_{k+1} - u_k) \left[\frac{1}{2} (b_j''(u_k) b_\ell''(u_{k+1}) + b_j''(u_{k+1}) b_\ell''(u_k)) + \frac{1}{3} (b_\ell''(u_{k+1}) - b_\ell''(u_k))(b_j''(u_{k+1}) - b_j''(u_k)) \right]$$

for each $1 \leq j, \ell \leq d_n$.

We can derive the above using the fact that each b_ℓ'' is piecewise linear.

Exercise: Demonstrate fitting the penalized splines estimator.


4
2
0
−2
−4
−6
0.0
0.2
0.4
0.6
0.8
1.0

Note that we may write

$$(\hat{m}_n^{\text{pspl}}(X_1), \ldots, \hat{m}_n^{\text{pspl}}(X_n))^T = \mathbf{SY},$$

where $\mathbf{S} = \mathbf{B}^T(\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{B}^T$, with

$$\mathbf{B} = (b_\ell(X_i))_{1\leq i\leq n, 1\leq \ell\leq d_n} \quad \text{and} \quad \boldsymbol{\Omega} = \left(\int_0^1 b_\ell''(x)b_j''(x)dx\right)_{1\leq \ell,j\leq d}.$$

The $n \times n$ matrix $\mathbf{S}$ is called a *smoother matrix*.

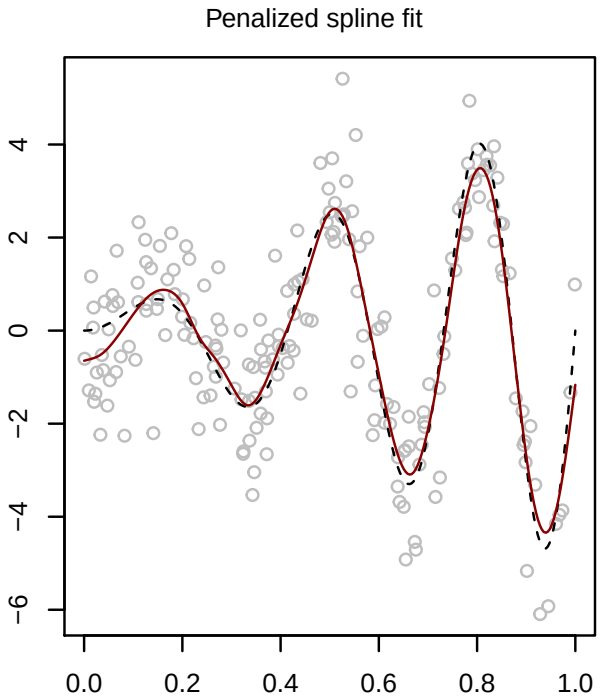
Linear estimator and smoother matrix

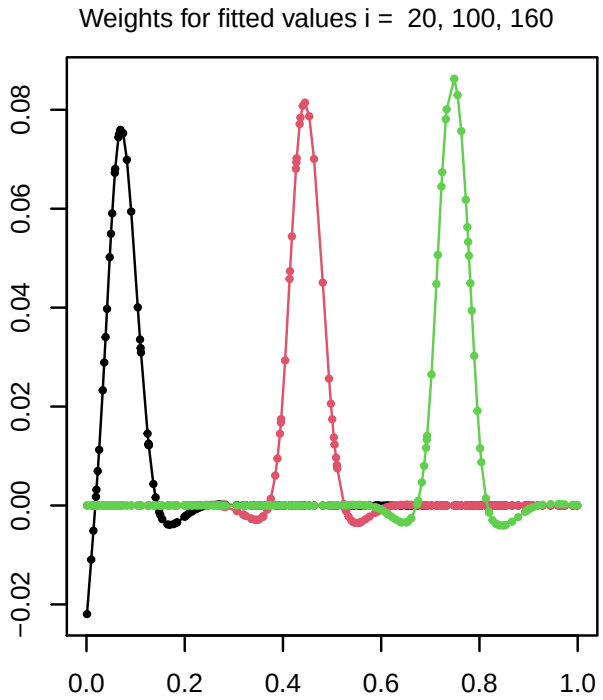
A *linear estimator* is any estimator $\hat{m}_n$ of the form

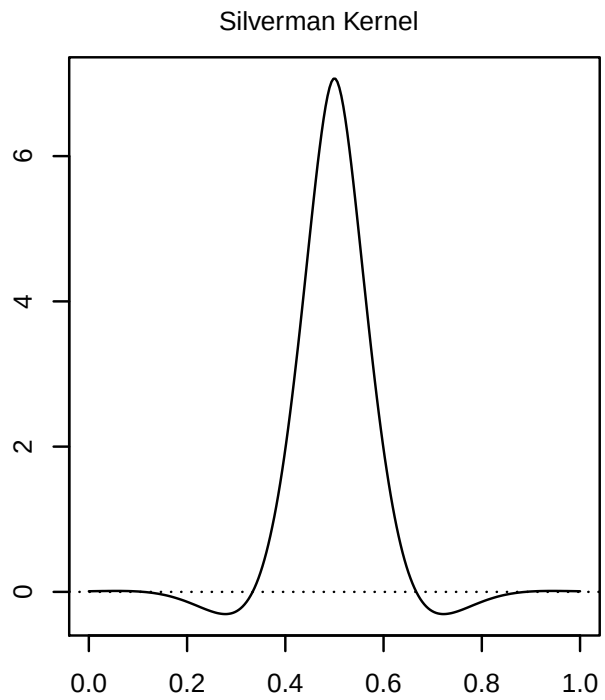
$$\hat{m}_n(x) = \sum_{i=1}^n W_{ni}(x)Y_i \quad \text{for some weights} \quad W_{n1}(x), \ldots, W_{nn}(x).$$

The *smoother matrix* associated with $\hat{m}_n$ is the matrix $\mathbf{S} = (W_{ni}(X_{i'}))_{1\leq i,i'\leq n}$.

Exercise: Plot some rows of a penalized spline smoother matrix $\mathbf{S}$. Discuss.


Penalized spline fit
4
2
0
−2
−4
−6
0.0
0.2
0.4
0.6
0.8
1.0


Weights for fitted values i = 20, 100, 160
0.08
0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0


Silverman Kernel
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0

Silverman's kernel, [2]

For $X_1, \ldots, X_n$ having the density f on $[0,1]$, the smoothing spline estimator is asymptotically equivalent to N-W under K and local bandwidth $h(x)$ given by

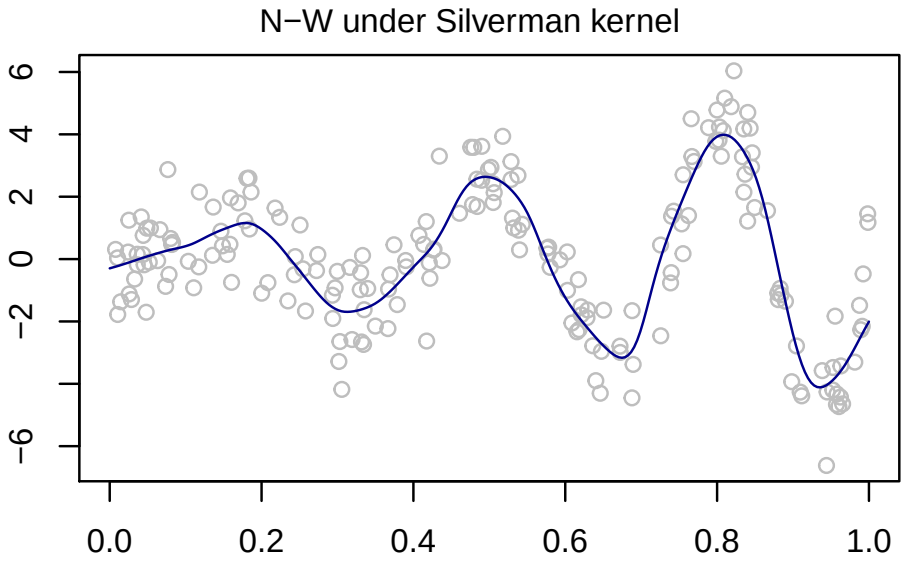
$$K(u) = \frac{1}{2} e^{-|u|/\sqrt{2}} \cdot \sin(|u|/\sqrt{2} + \pi/4) \quad \text{and} \quad h(x) = (n^{-1}\lambda / f(x))^{1/4}.$$

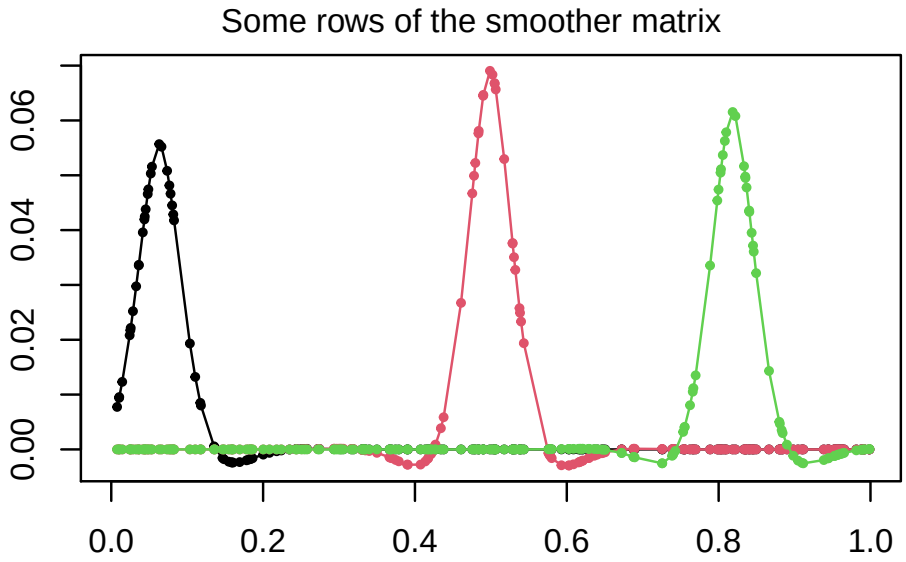
The previous slide shows a plot of the Silverman kernel function.

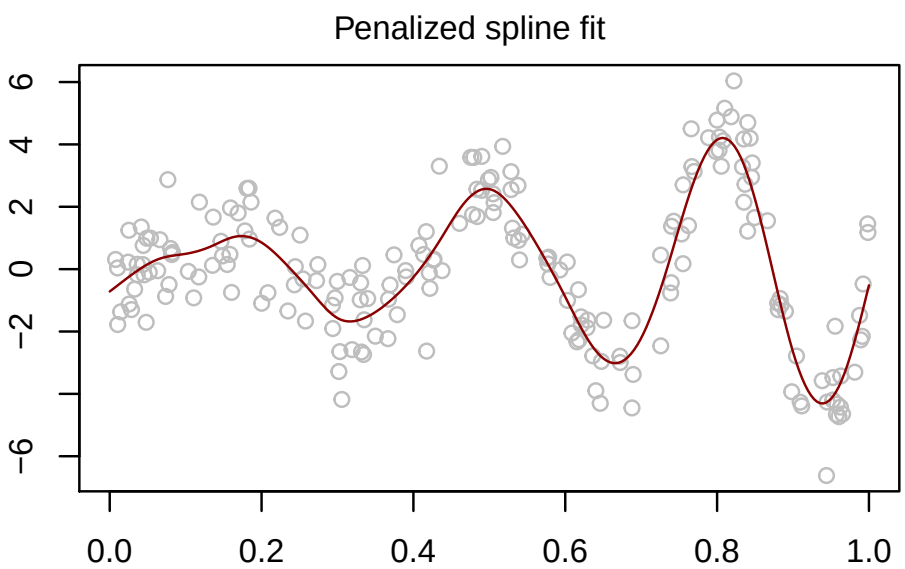
Exercise: Make some plots comparing the rows of the smoother matrices

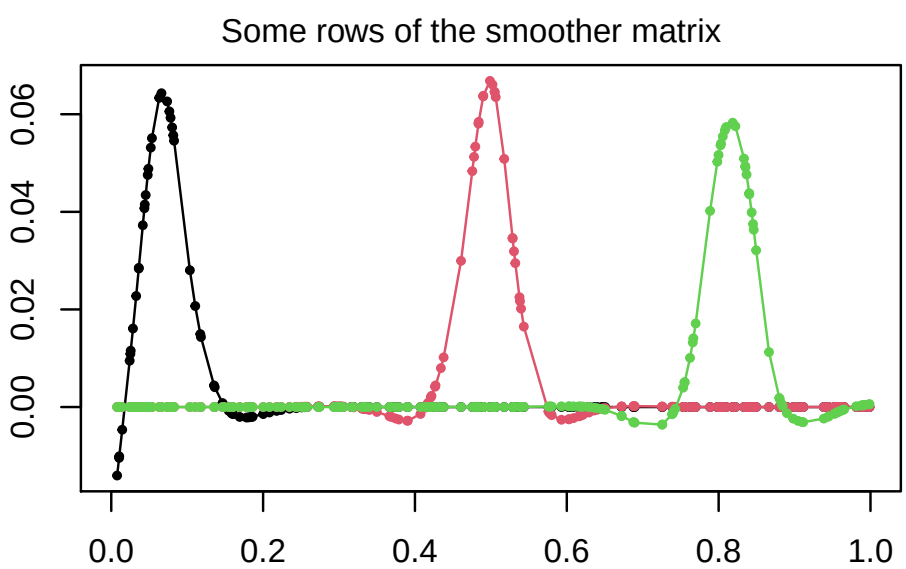
1. of a penalized spline estimator.
2. of the N-W estimator under Silverman's kernel with bandwidth $h = (\lambda/n)^{1/4}$.

For the exercise, generate $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Uniform}(0,1)$.


N−W under Silverman kernel
6
4
2
0
−2
−6
0.0
0.2
0.4
0.6
0.8
1.0


Some rows of the smoother matrix
0.06
0.04
0.02
0.00
0.0
0.2
0.4
0.6
0.8
1.0


Penalized spline fit
6
4
2
0
−2
−6
0.0
0.2
0.4
0.6
0.8
1.0


Some rows of the smoother matrix
0.06
0.04
0.02
0.00
0.0
0.2
0.4
0.6
0.8
1.0

Consider the eigendecomposition of the smoother matrix

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad \mathbf{U}\mathbf{U}^T = \mathbf{U}^\mathbf{T}\mathbf{U} = \mathbf{I}_n, \quad \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$ are the eigenvalues.

We see that the vector of fitted values can be written as

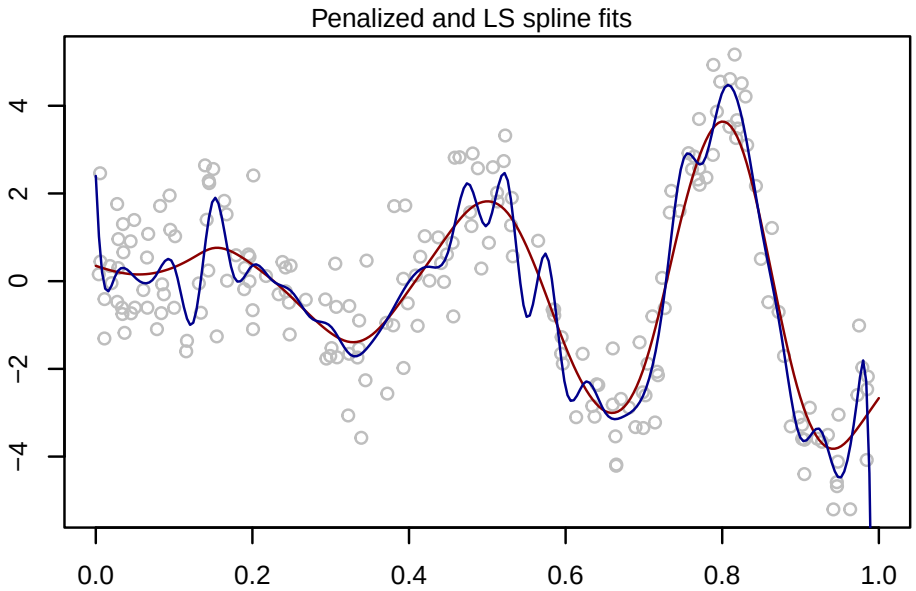
$$(\hat{m}_n^{\text{pspl}}(X_1), \ldots, \hat{m}_n^{\text{pspl}}(X_n))^T = \mathbf{S}\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y}.$$

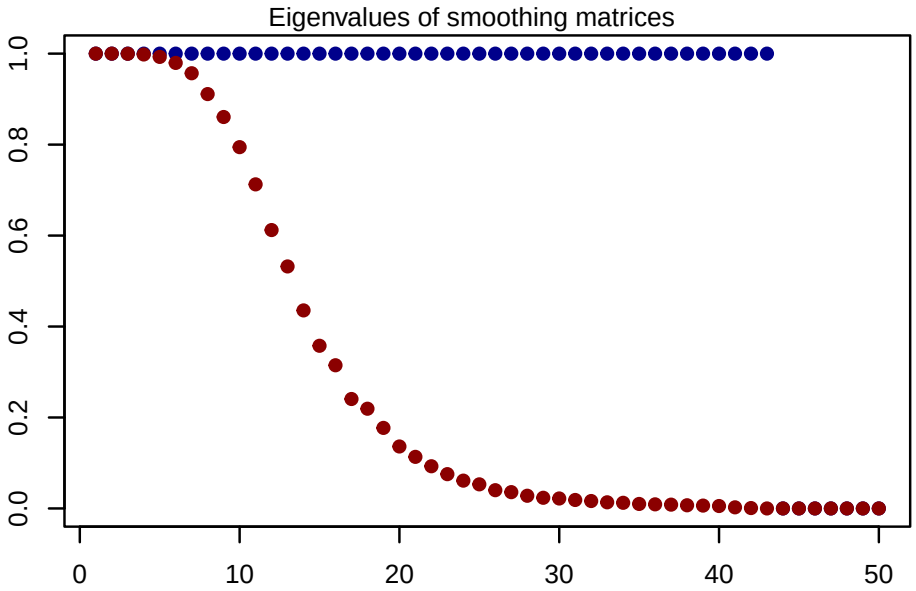
Discuss: Can we learn anything from this?

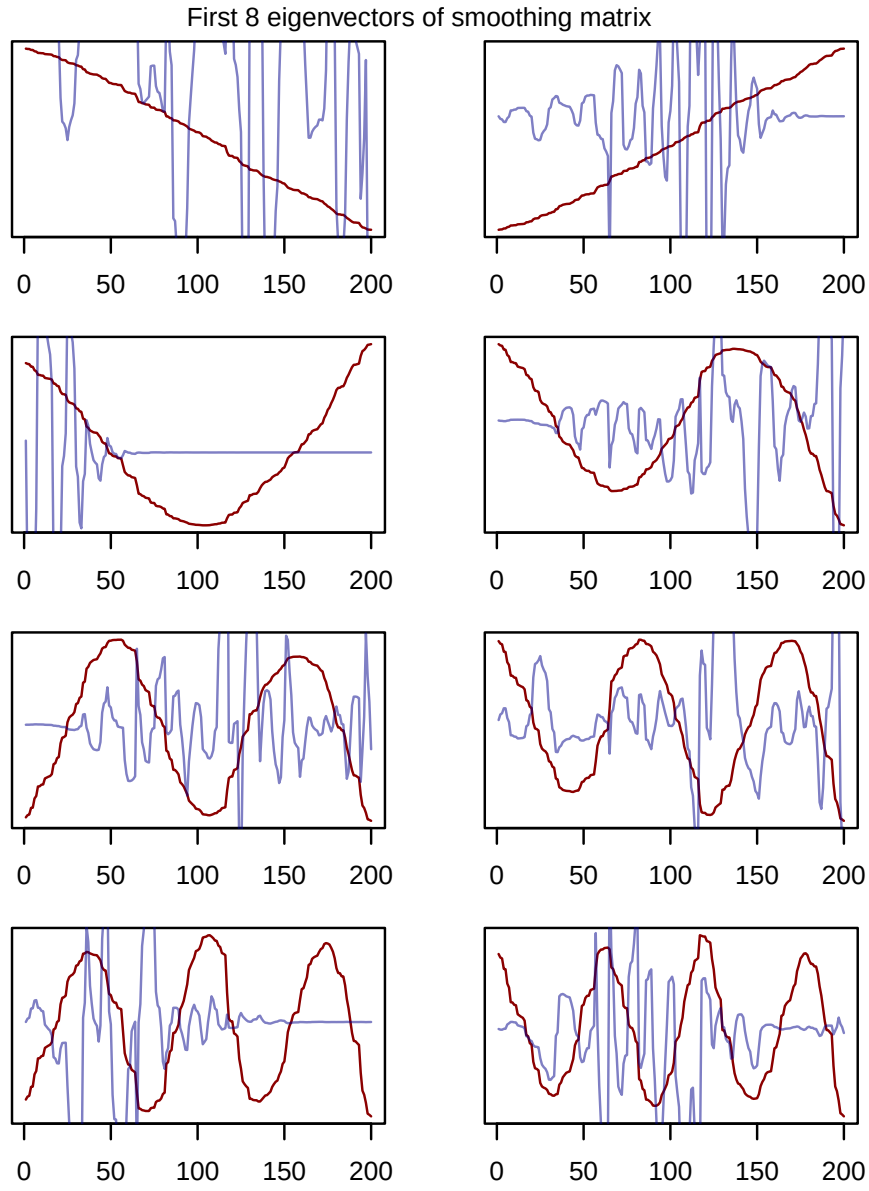
Exercise: Inspect eigenvectors/eigenvalues of the smoother matrices

1. $\mathbf{S} = \mathbf{B}^T(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$
2. $\mathbf{S} = \mathbf{B}^T(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{\Omega})^{-1}\mathbf{B}^T$

Make plots and discuss.





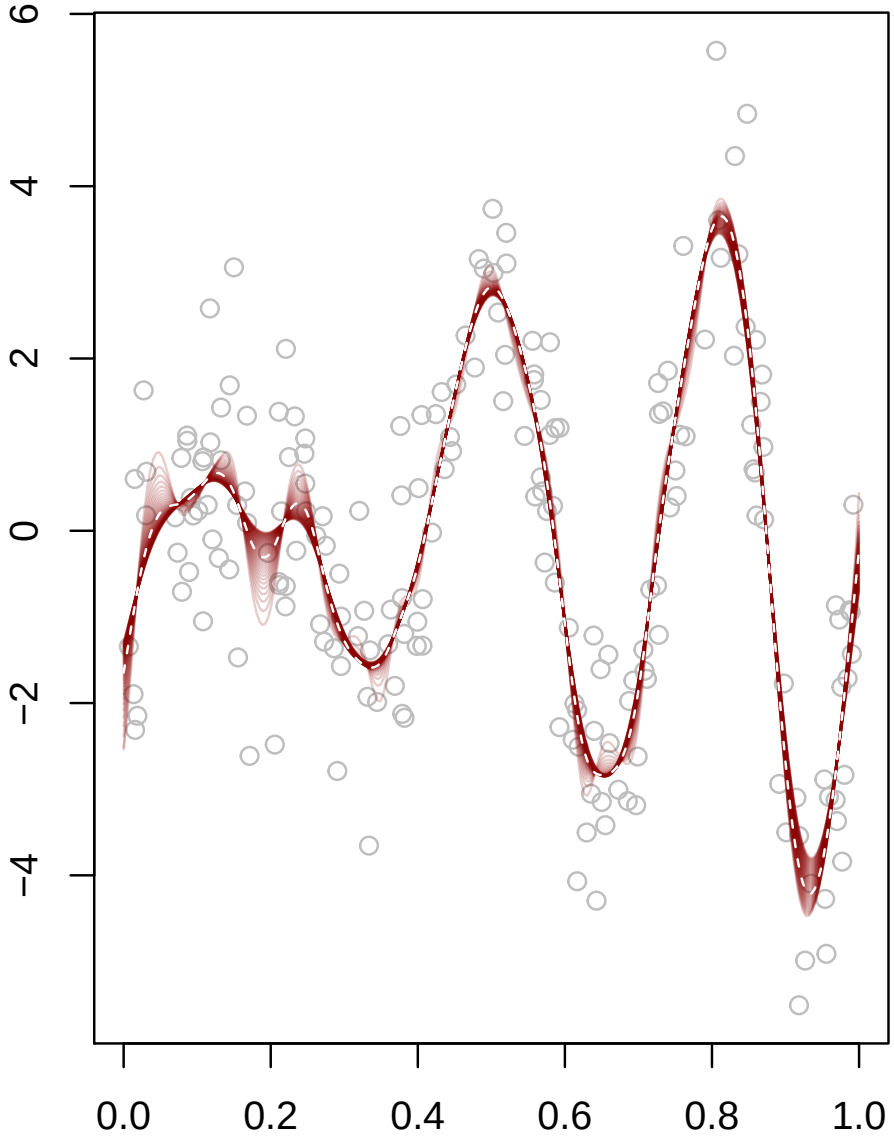


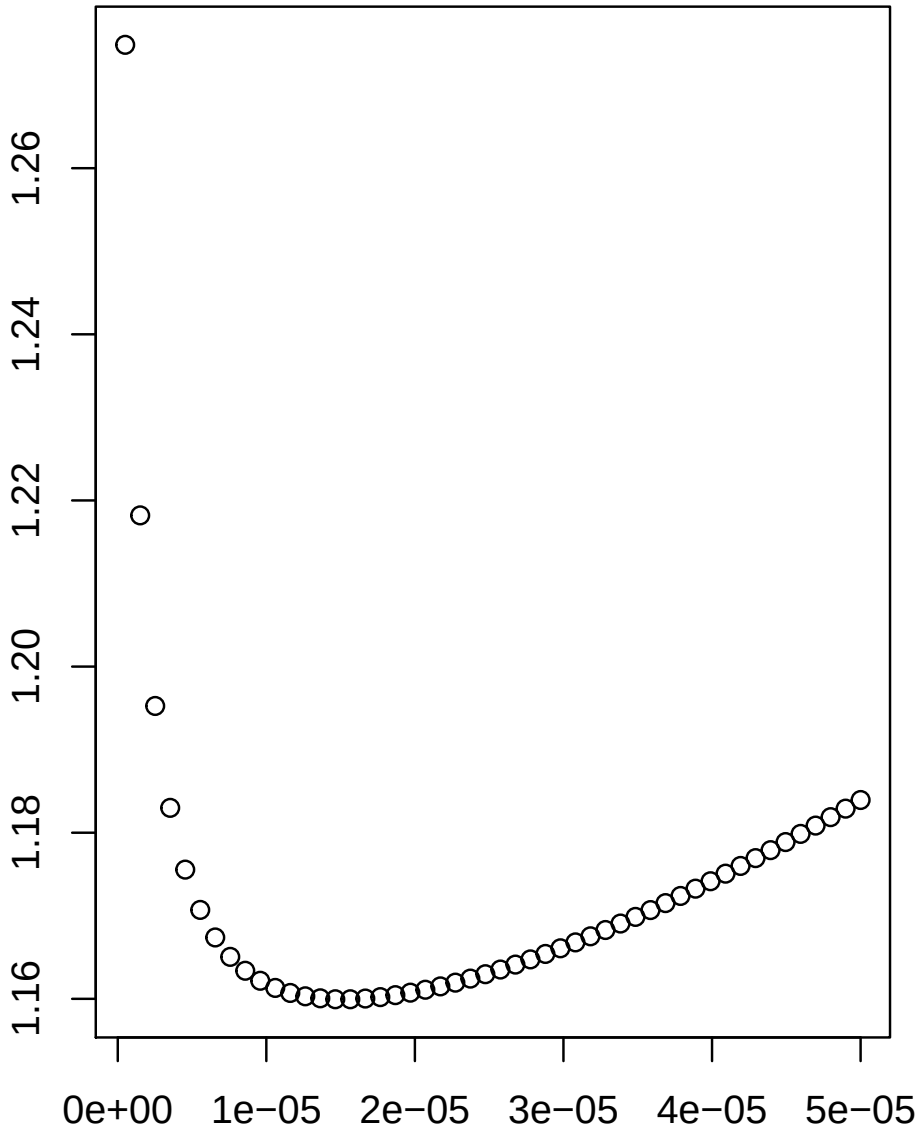
Prove: If $\underset{n \times d}{\mathbf{B}}$ full-rank, $\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ has exactly d nonzero eigenvals, all $= 1$.

Since $\hat{m}_n^{\text{pspl}}$ is a linear estimator, we have

$$\text{CV}_n(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{n,-i}^{\text{pspl}}(X_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{m}_n^{\text{pspl}}(X_i)}{1 - S_{ii}}\right]^2,$$

where S_{ii} is the element i on the diagonal of the smoother matrix $\mathbf{S}$.


6
4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0


1.26
1.24
1.22
1.20
1.18
1.16
0e+00
1e−05
2e−05
3e−05
4e−05
5e−05

Suppose $X_i = i/n$, $i = 1, \ldots, n$, for now, and let $\boldsymbol{\mu} = (m(X_1), \ldots, m(X_n))^T$.

Trend filtering estimator (good reference is Tibshirani's paper, [3])

The *trend filtering estimator* of μ is given by

$$\hat{\boldsymbol{\mu}} = \underset{\mathbf{u} \in \mathbb{R}^n}{\operatorname{argmin}} \ \|\mathbf{Y} - \mathbf{u}\|_2^2 + \lambda \|\mathbf{D}^{(k+1)}\mathbf{u}\|_1,$$

where

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}$$

and $D^{(k+1)} = D^{(1)} D^{(k)}$, $k \geq 1$, adjusting the dimension of $D^{(1)}$ as needed.

To estimate m at any $x \in [0, 1]$, we can just linearly interpolate $\boldsymbol{\mu}$.

Accommodate unequally spaced inputs with a modification to $\mathbf{D}^{(k+1)}$. See [3].

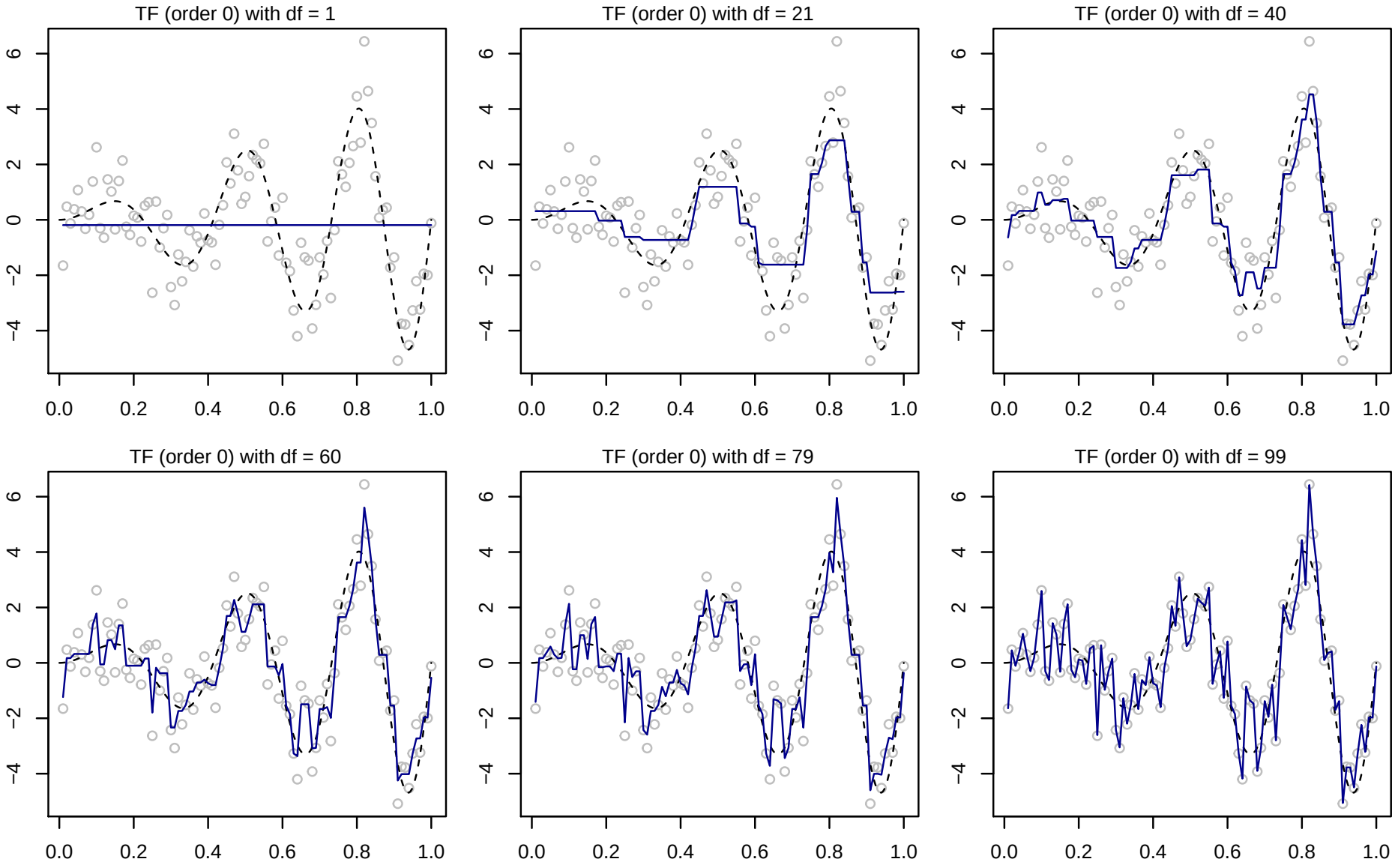
Exercise: Study the penalties $\lambda\|\mathbf{D}^{(1)}\mathbf{u}\|_1$, $\lambda\|\mathbf{D}^{(2)}\mathbf{u}\|_1$, $\lambda\|\mathbf{D}^{(3)}\mathbf{u}\|_1$, and $\lambda\|\mathbf{D}^{(4)}\mathbf{u}\|_1$ and consider their effects on $\hat{\boldsymbol{\mu}}$.

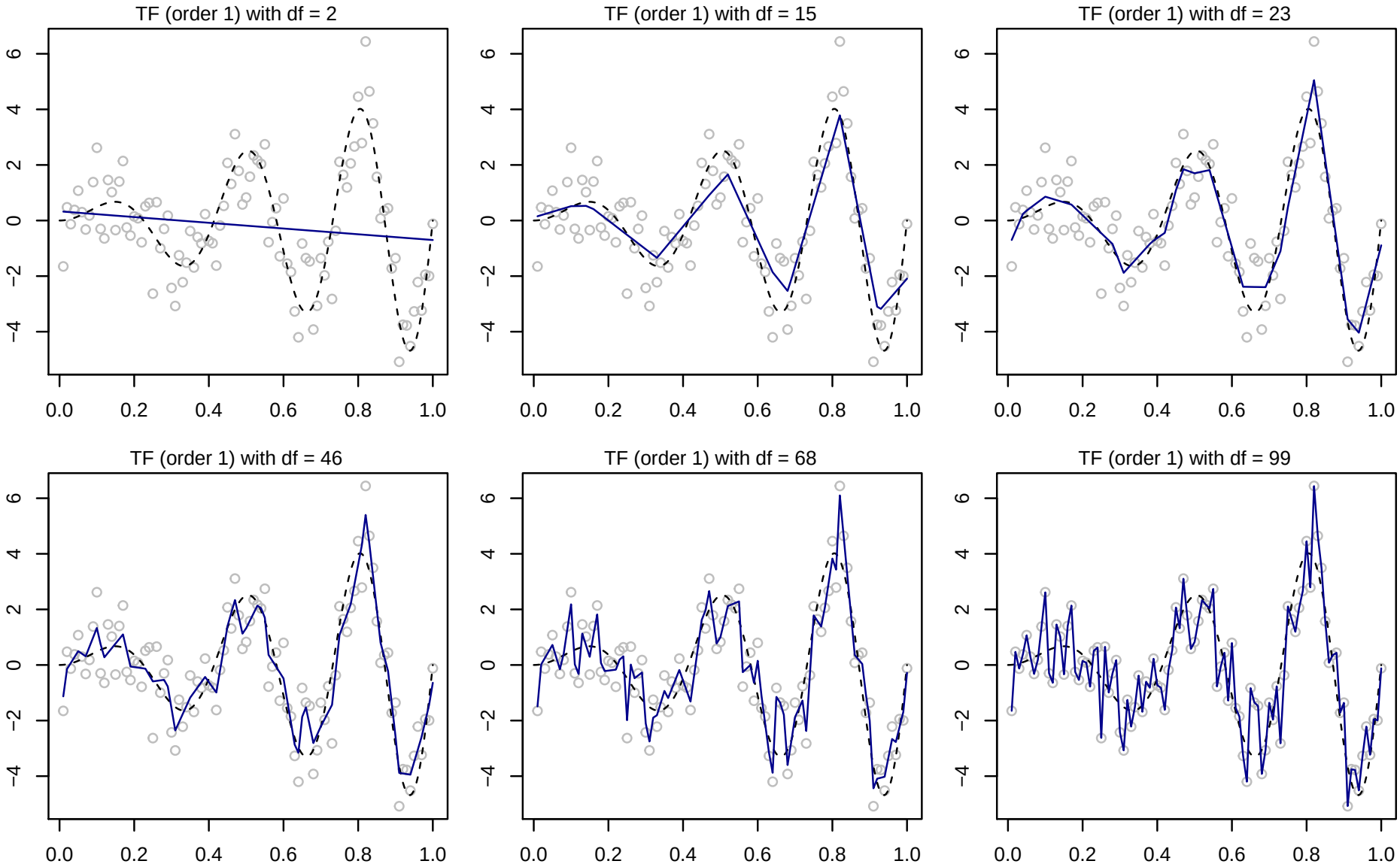
The following is known as a *generalized lasso* minimization problem:

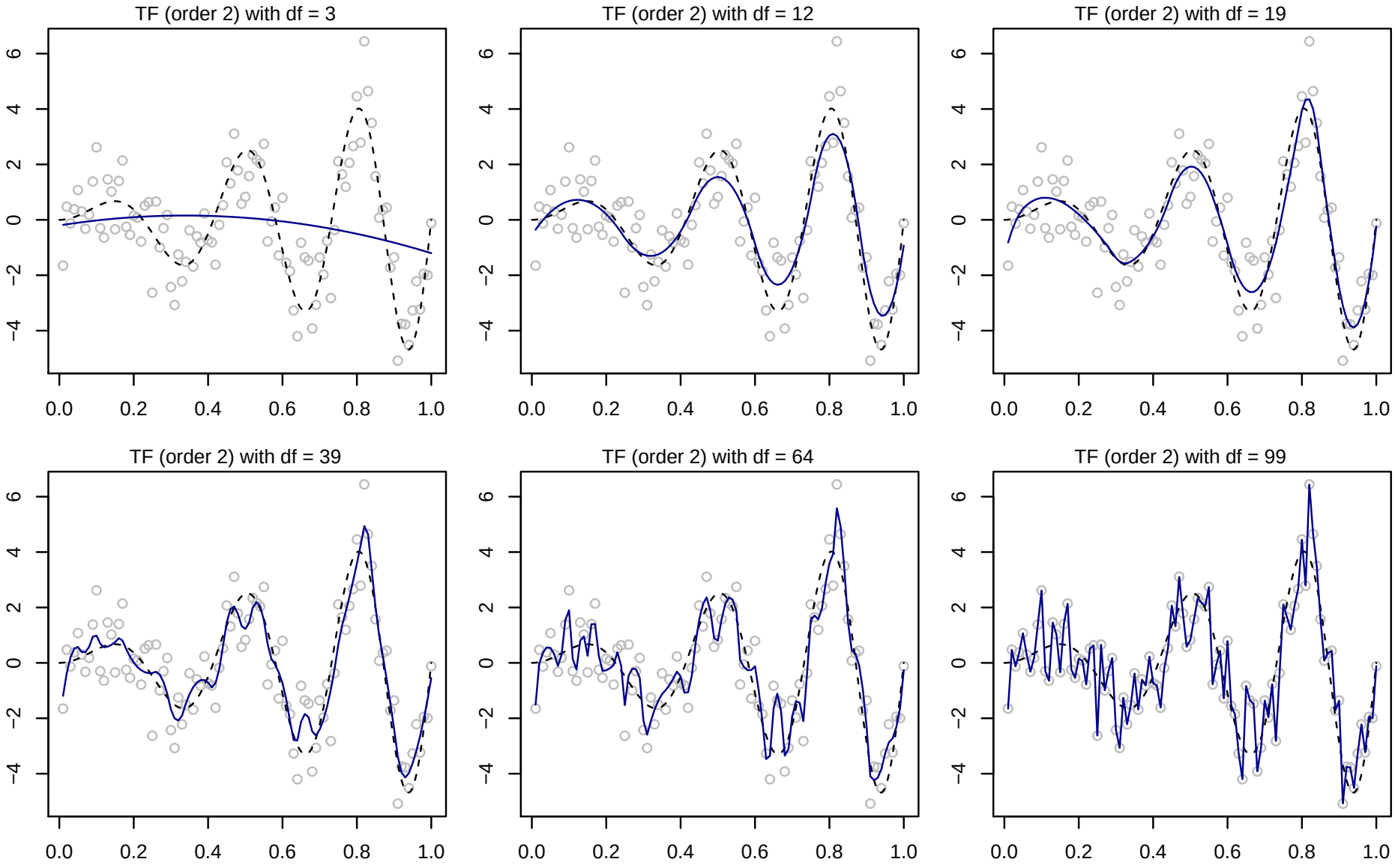
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\mathbf{D}\boldsymbol{\beta}\|_1.$$

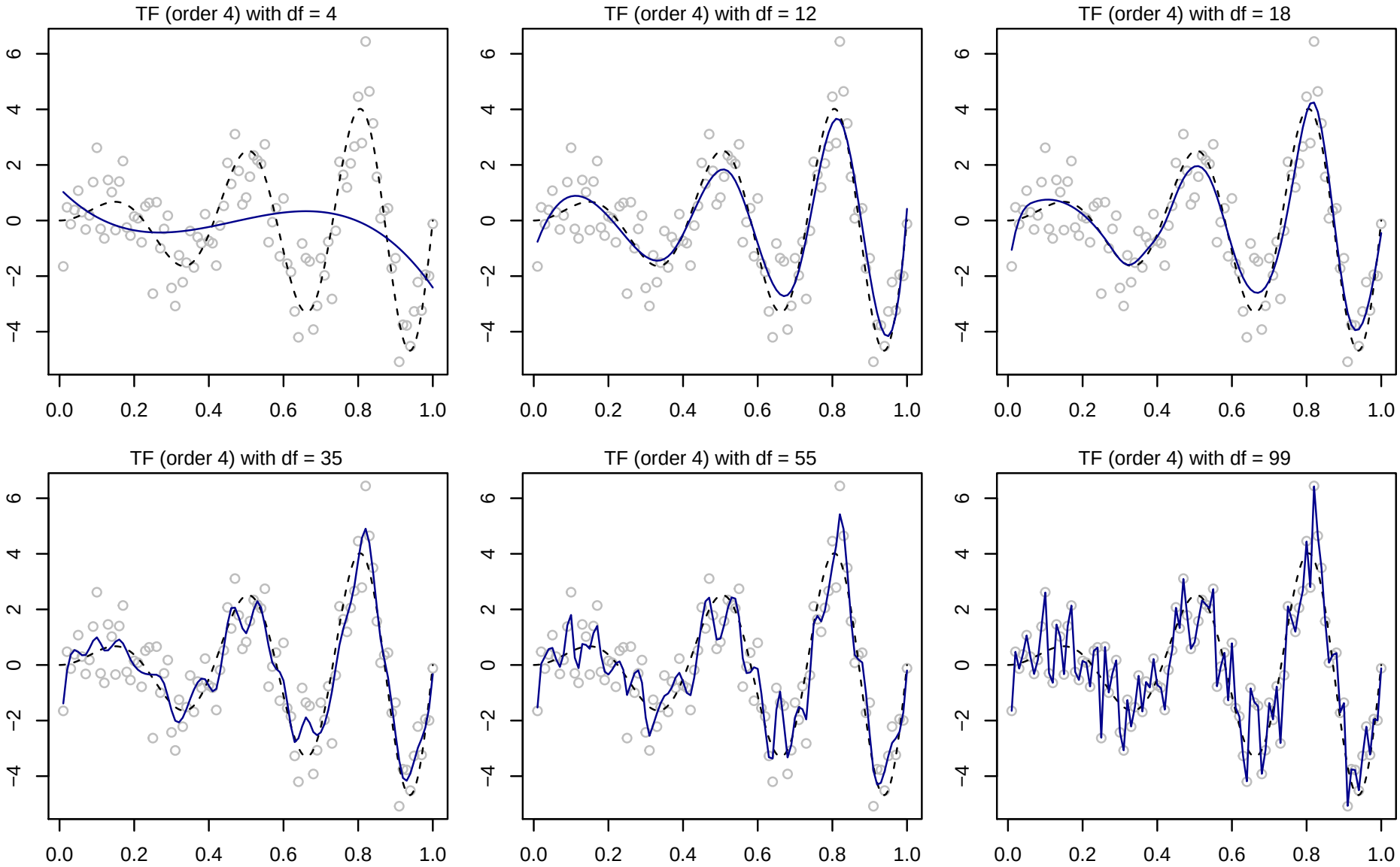
It can be solved with the `genlasso` package.

Exercise: Fit the trend filtering estimator on some data for $k = 0, 1, 2, 3$ and plot.


TF (order 0) with df = 1
TF (order 0) with df = 21
TF (order 0) with df = 40
TF (order 0) with df = 60
TF (order 0) with df = 79
TF (order 0) with df = 99
−4
−2
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0


TF (order 1) with df = 2
TF (order 1) with df = 15
TF (order 1) with df = 23
TF (order 1) with df = 46
TF (order 1) with df = 68
TF (order 1) with df = 99
6
4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0


TF (order 2) with df = 3
TF (order 2) with df = 12
TF (order 2) with df = 19
TF (order 2) with df = 39
TF (order 2) with df = 64
TF (order 2) with df = 99
−4
−2
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0


TF (order 4) with df = 4
TF (order 4) with df = 12
TF (order 4) with df = 18
TF (order 4) with df = 35
TF (order 4) with df = 55
TF (order 4) with df = 99
6
4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0

David Ruppert, Matt P Wand, and Raymond J Carroll.
Semiparametric regression.
Number 12. Cambridge university press, 2003.

Bernard W Silverman.
Spline smoothing: the equivalent variable kernel method.
The Annals of Statistics, pages 898–916, 1984.

Ryan J Tibshirani et al.
Adaptive piecewise polynomial estimation via trend filtering.
The Annals of statistics, 42(1):285–323, 2014.

Grace Wahba.
Spline models for observational data.
SIAM, 1990.