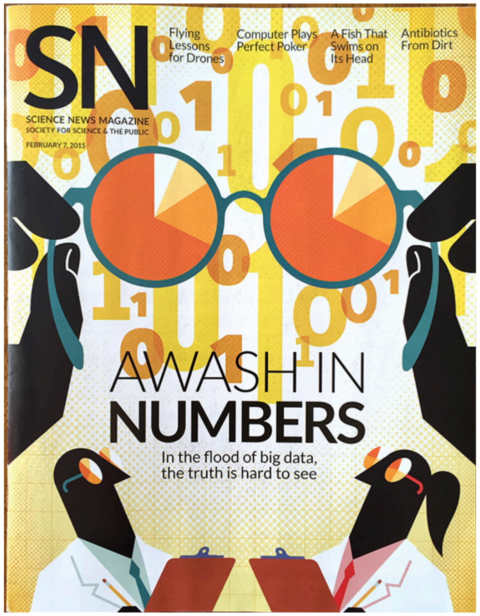
Fused Lasso Additive Model

Ashley Petersen
UMN Biostatistics

Joint Work with Noah Simon & Daniela Witten

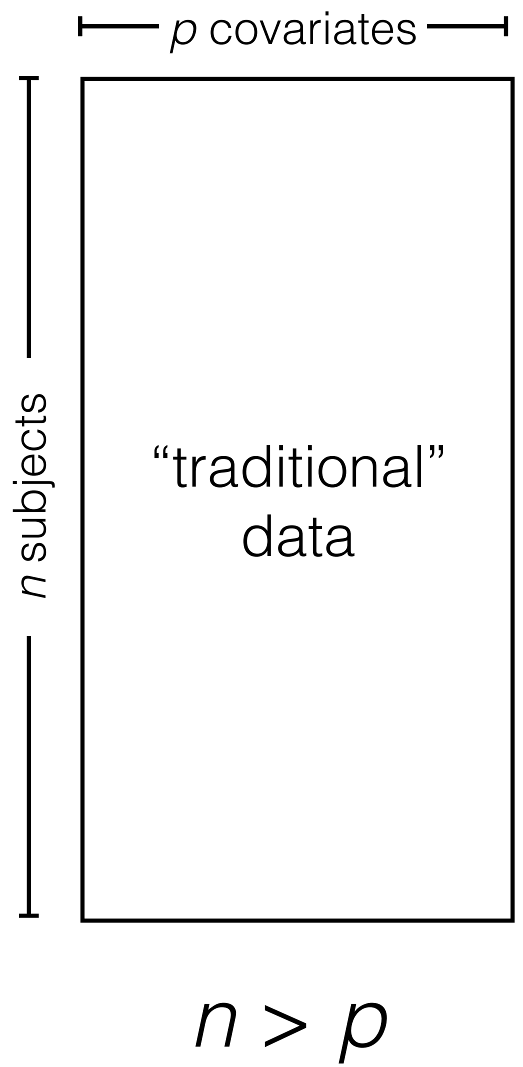


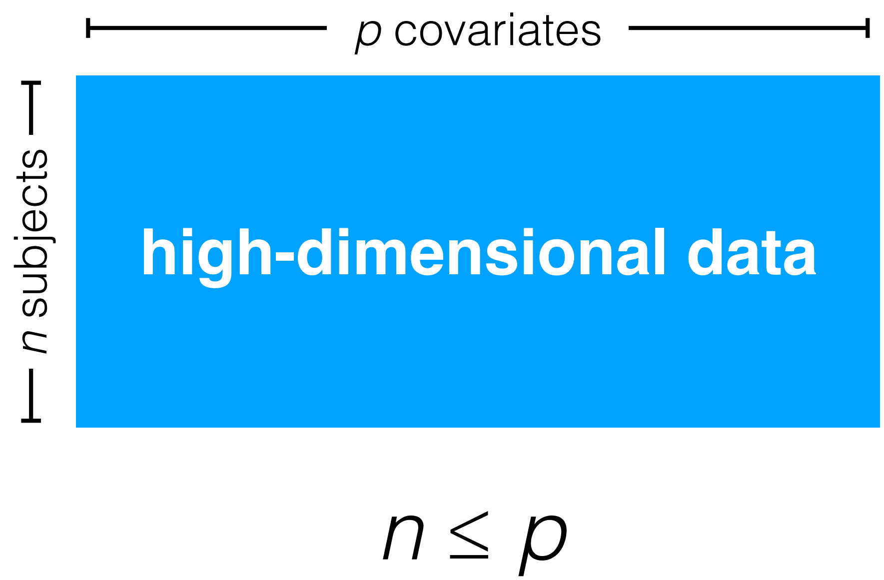
“data tsunami”

“drowning in data”

“flood of data”

What is the structure of the data?





Flexible and interpretable regression modeling

Goal: Fit the model

$$y = \sum_{j=1}^{p} f_j(x_j) + \epsilon$$

in a way that is simultaneously flexible, interpretable, and suitable for high-dimensional data.

Modeling decisions

- Which predictors should be included in the model?
- What functional forms should be used for the non-linear functions?

Modeling decisions

- Which predictors should be included in the model?
- What functional forms should be used for the non-linear functions?

Make these decisions in a data-adaptive way!

FLAM: fused lasso additive model

Fused lasso additive model

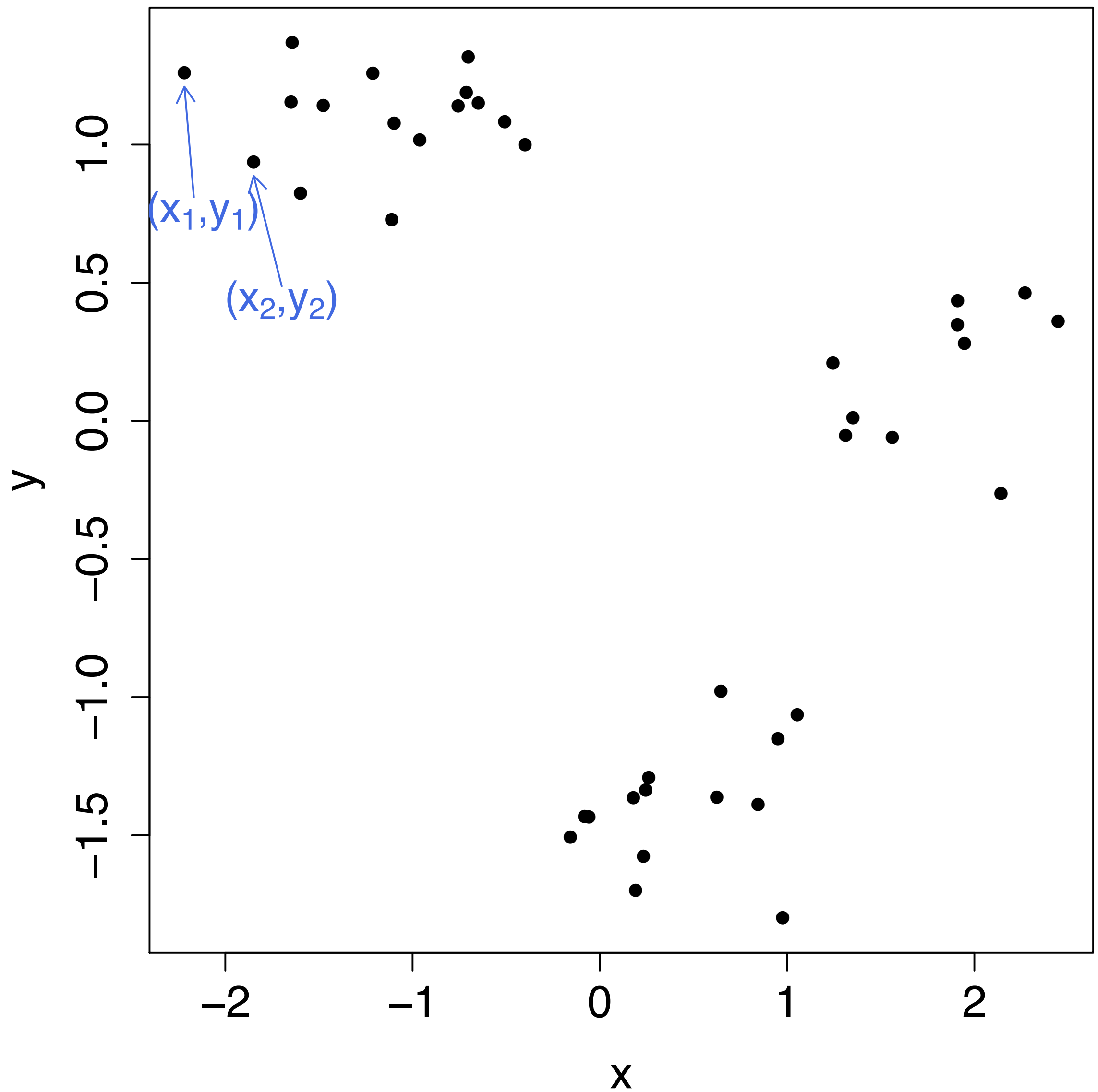
Goal: Fit the model

$$y = \sum_{j=1}^{p} f_j(x_j) + \epsilon$$

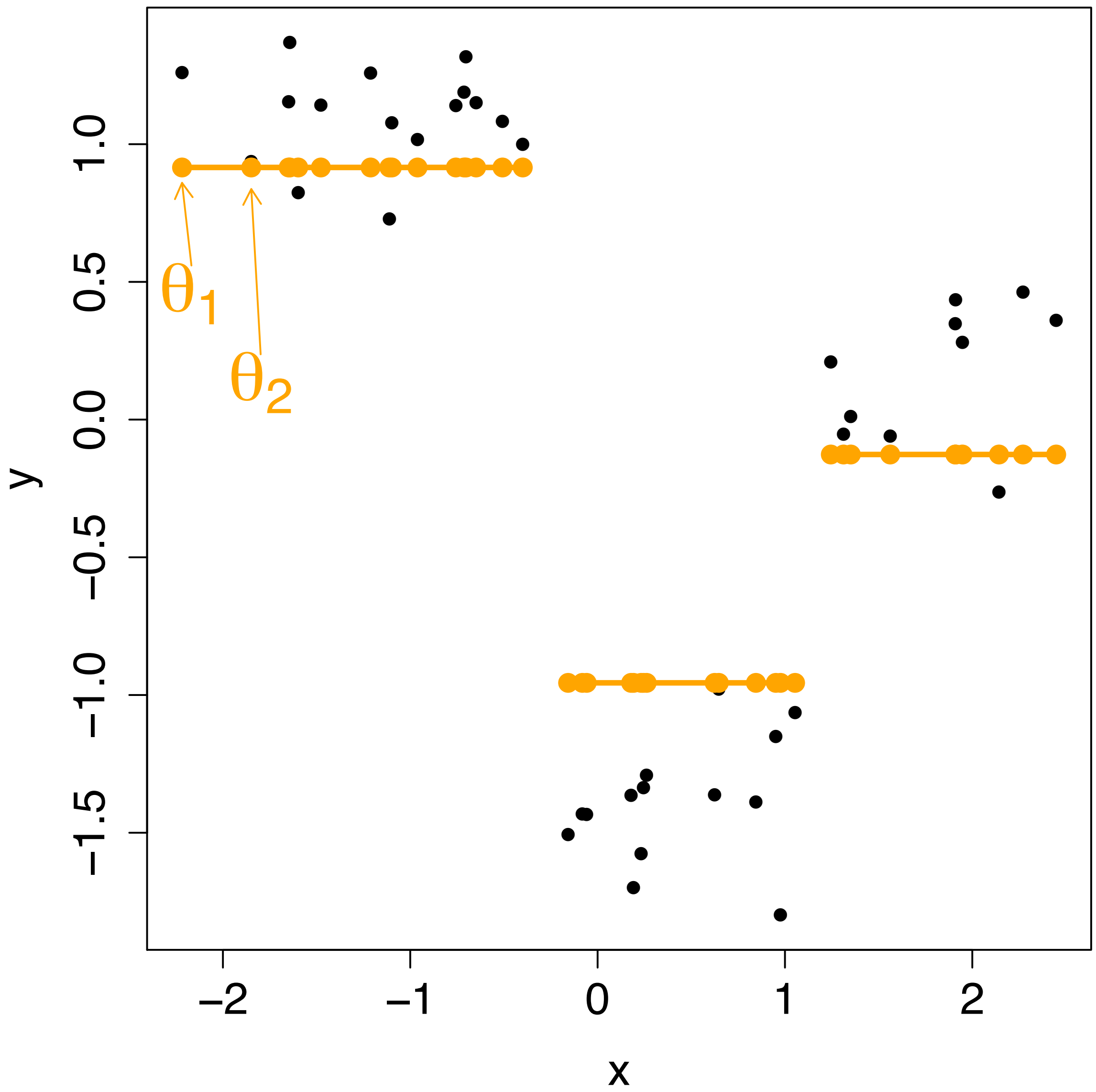
in a way that is simultaneously flexible and interpretable.

Estimate $f_1, \ldots, f_p$ to each be piecewise constant with a small number of adaptively-chosen knots

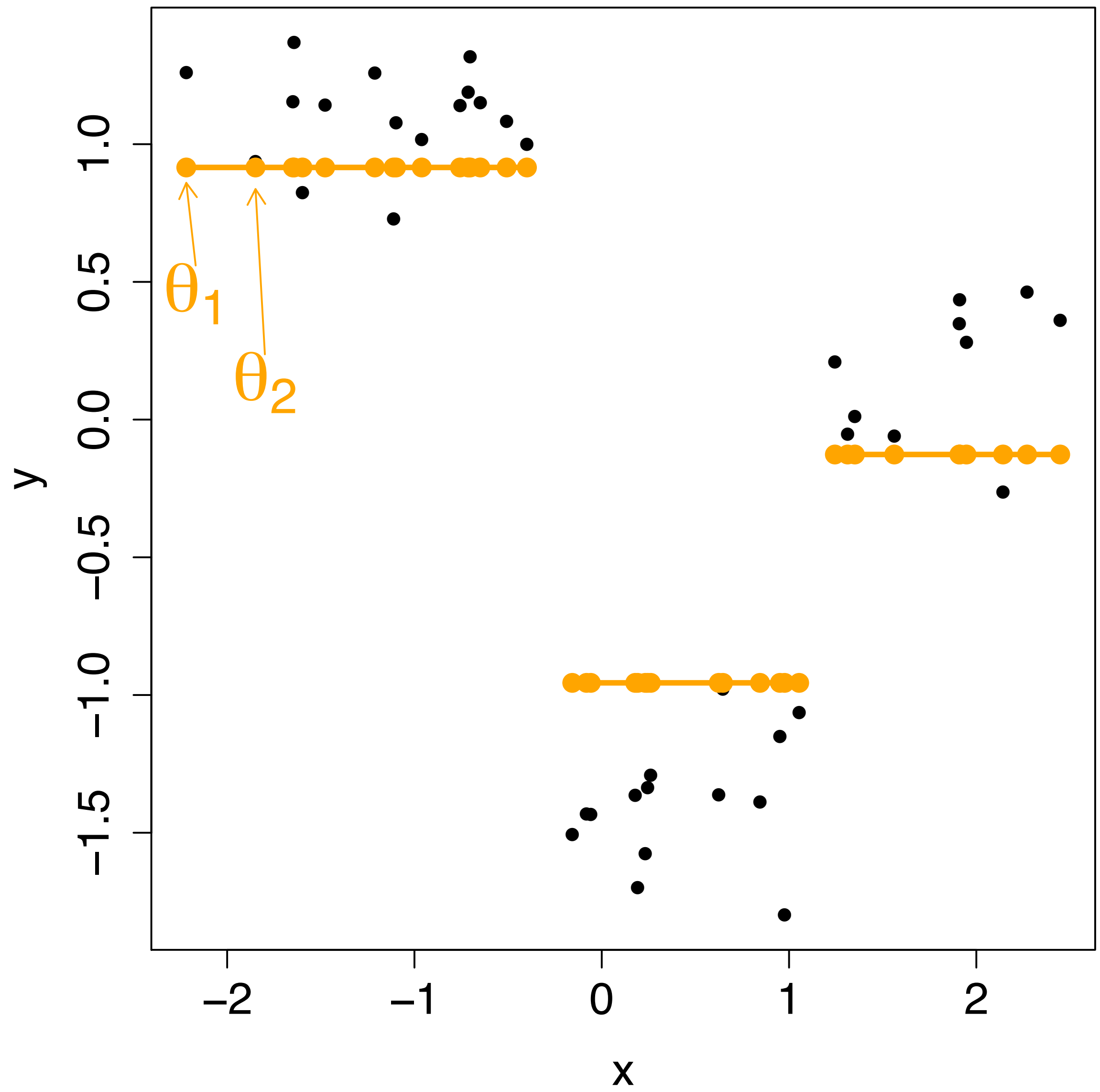
What if we only had one covariate?



What if we only had one covariate?

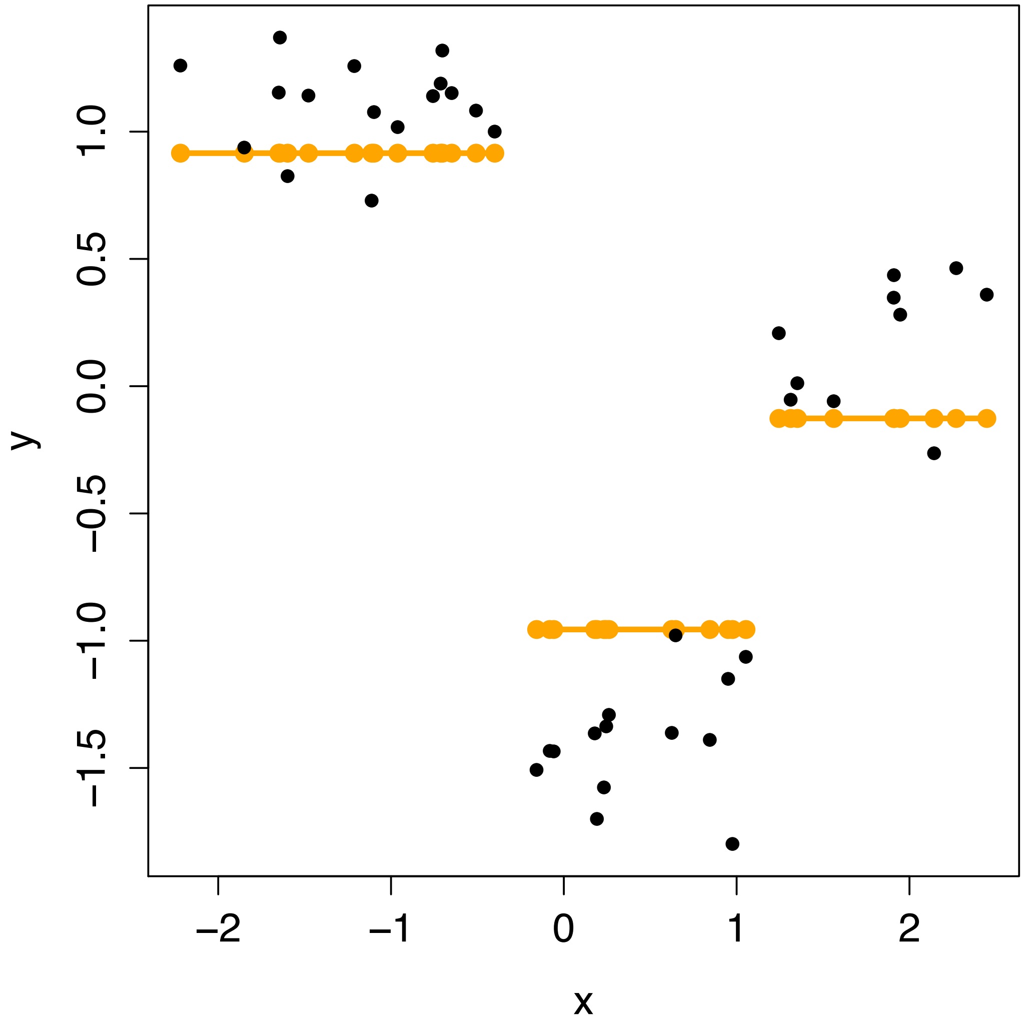


What if we only had one covariate?



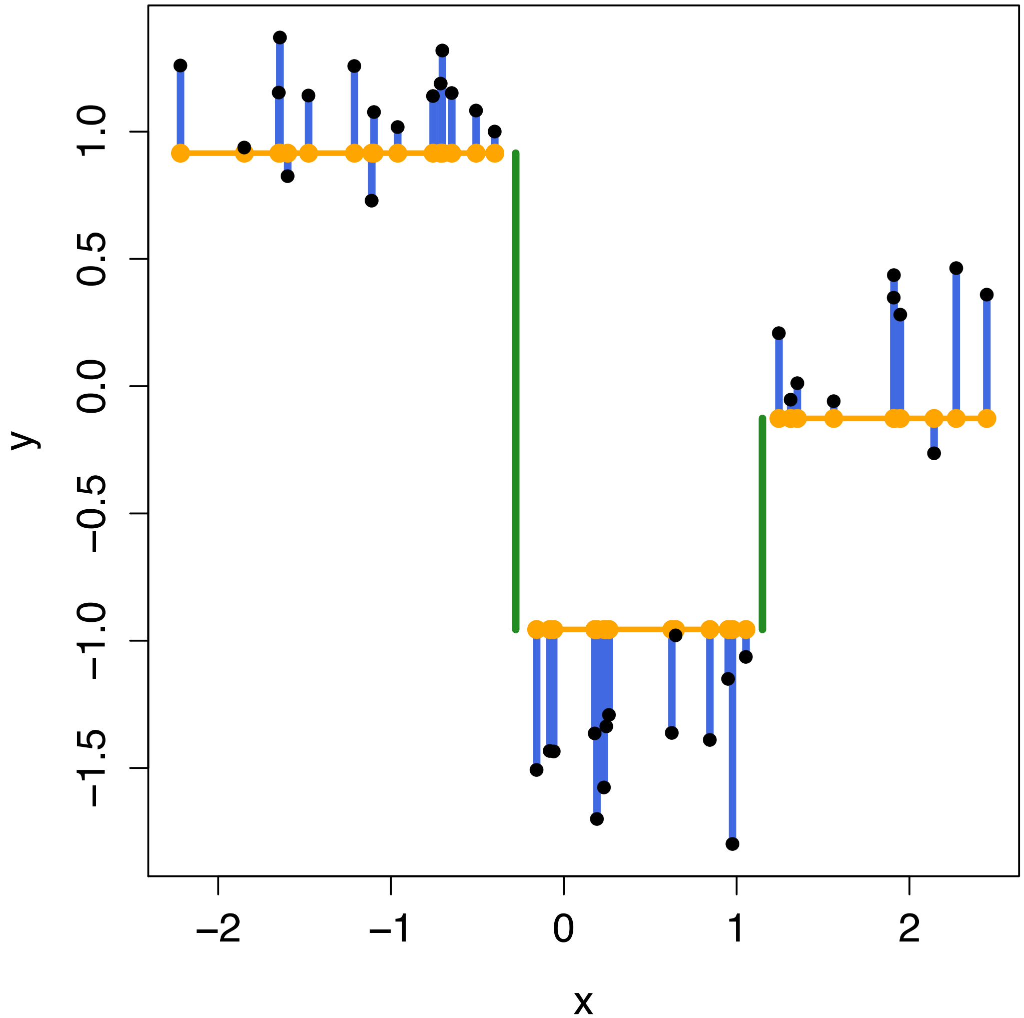
Knots occur when $\theta_i \neq \theta_{i+1}$

Estimating θ



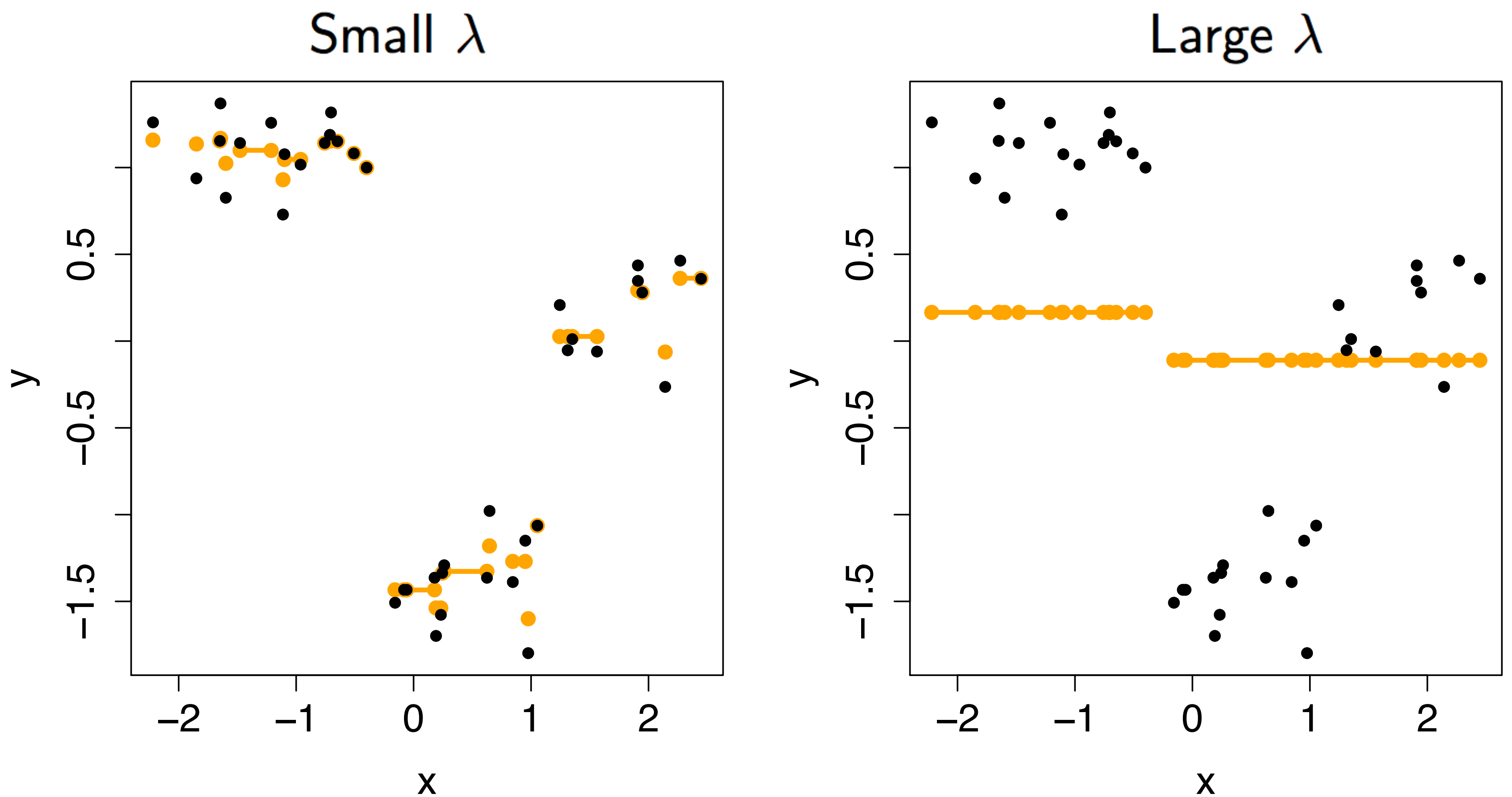
$$\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda\sum_{i=1}^{n-1}|\theta_j-\theta_{j+1}|$$

Estimating θ



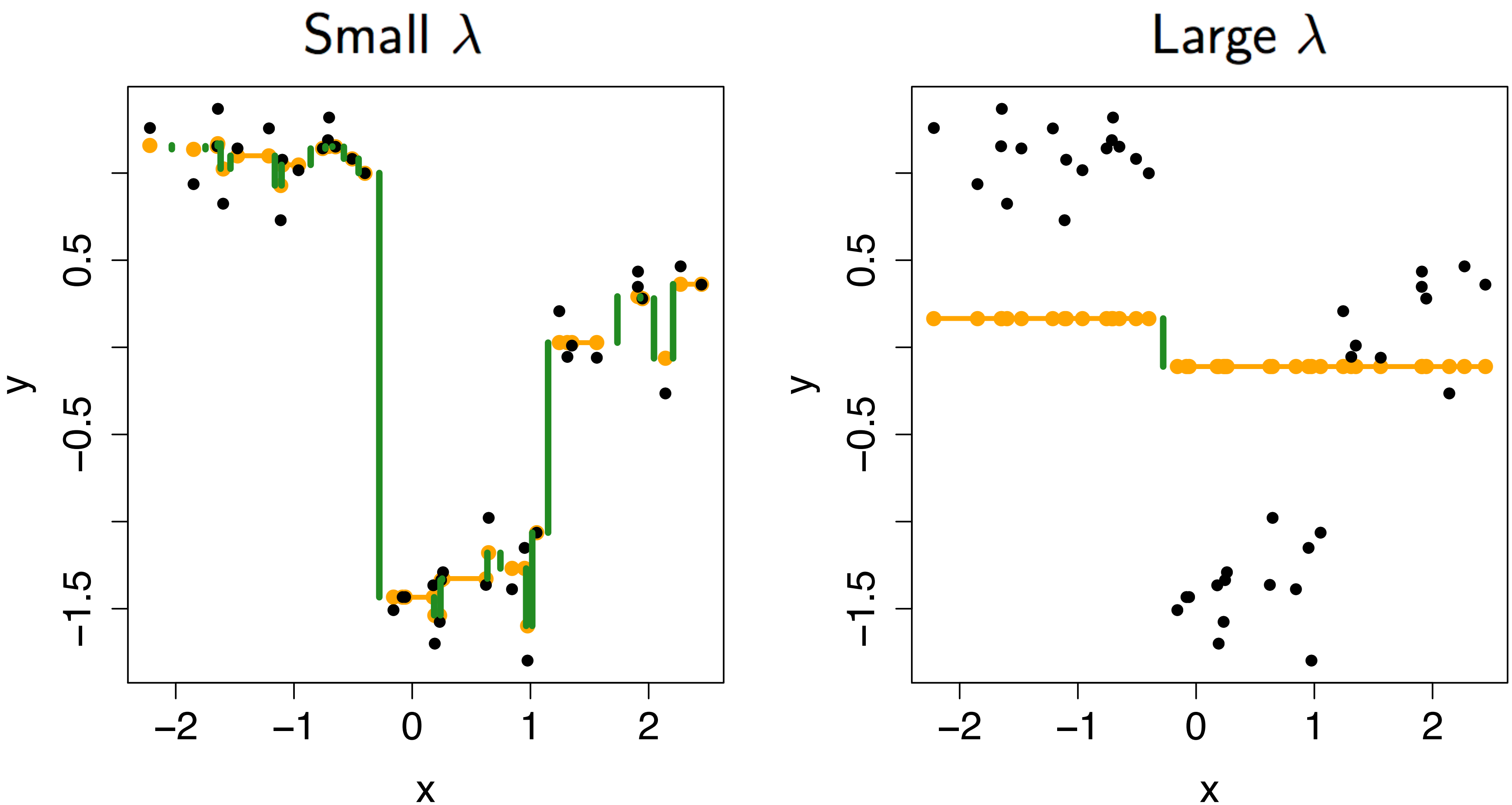
$$\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}_i)^2 + \lambda\sum_{i=1}^{n-1}|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j+1}|$$

Controlling the number of knots



$$\underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-1} |\theta_j - \theta_{j+1}|$$

Controlling the number of knots



$$\underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-1} |\boldsymbol{\theta_j - \theta_{j+1}}|$$

Optimization problem with one covariate

Solve

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \| y - \boldsymbol{\theta} \|_2^2 + \lambda \| D\boldsymbol{\theta} \|_1$$

where

$$D\boldsymbol{\theta} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix} = \begin{pmatrix} \theta_1 - \theta_2 \\ \theta_2 - \theta_3 \\ \vdots \\ \theta_{n-1} - \theta_n \end{pmatrix}$$

the **non-zero elements**
correspond to knots

Extending to multiple covariates

Single (ordered) covariate:

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|y-\boldsymbol{\theta}\|_2^2 + \lambda\|D\boldsymbol{\theta}\|_1$$

Multiple covariates:

$$\underset{\theta_0\in\mathbb{R},\boldsymbol{\theta}_j\in\mathbb{R}^n,1\le j\le p}{\text{minimize}} \quad \frac{1}{2}\left\|y-\sum_{j=1}^{p}\boldsymbol{\theta}_j-\theta_0\mathbf{1}\right\|_2^2 + \lambda\sum_{j=1}^{p}\|DP_j\boldsymbol{\theta}_j\|_1$$

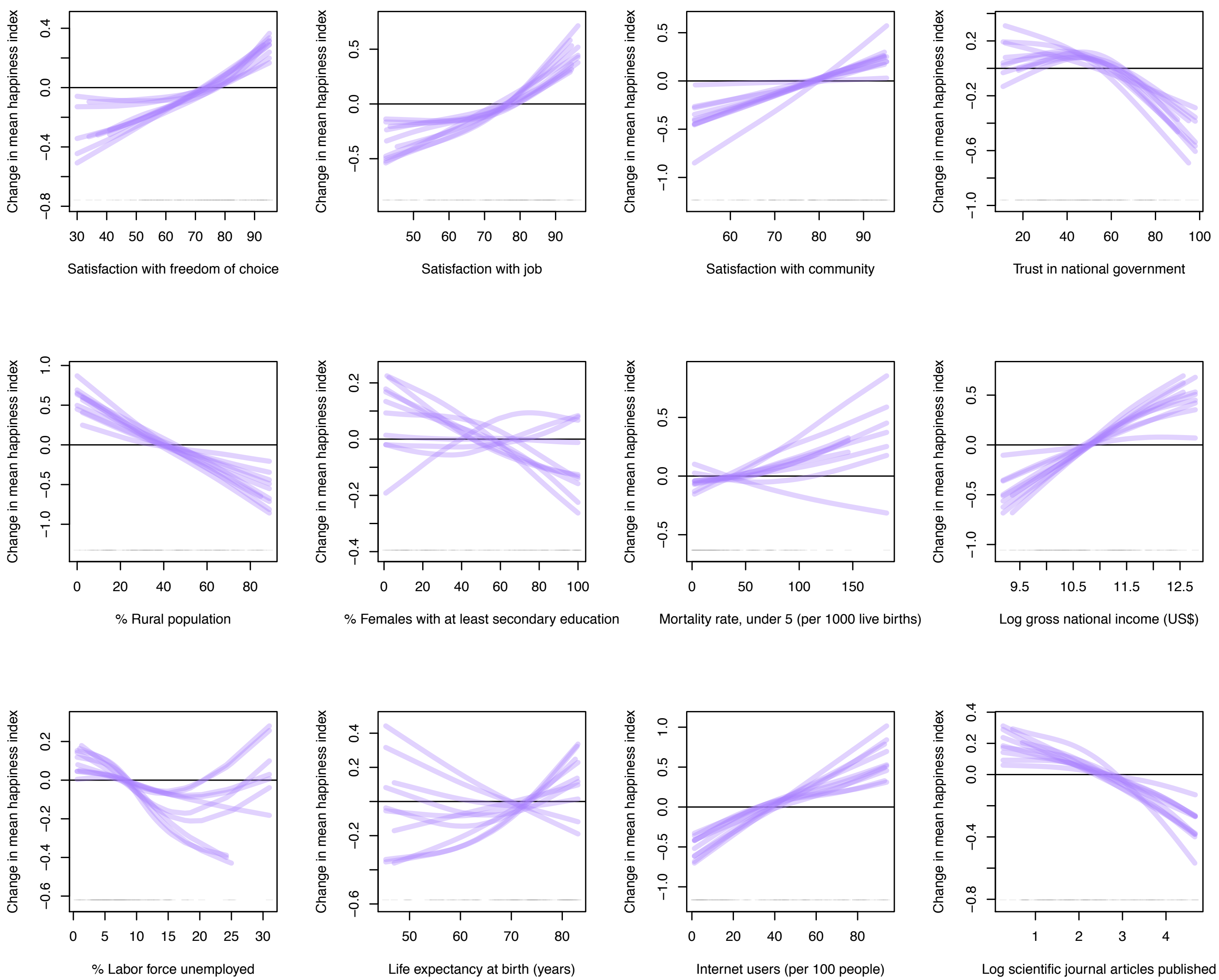
where P_j is the permutation matrix that orders x_j from least to greatest

Do wealth and publishing papers make you happy?*

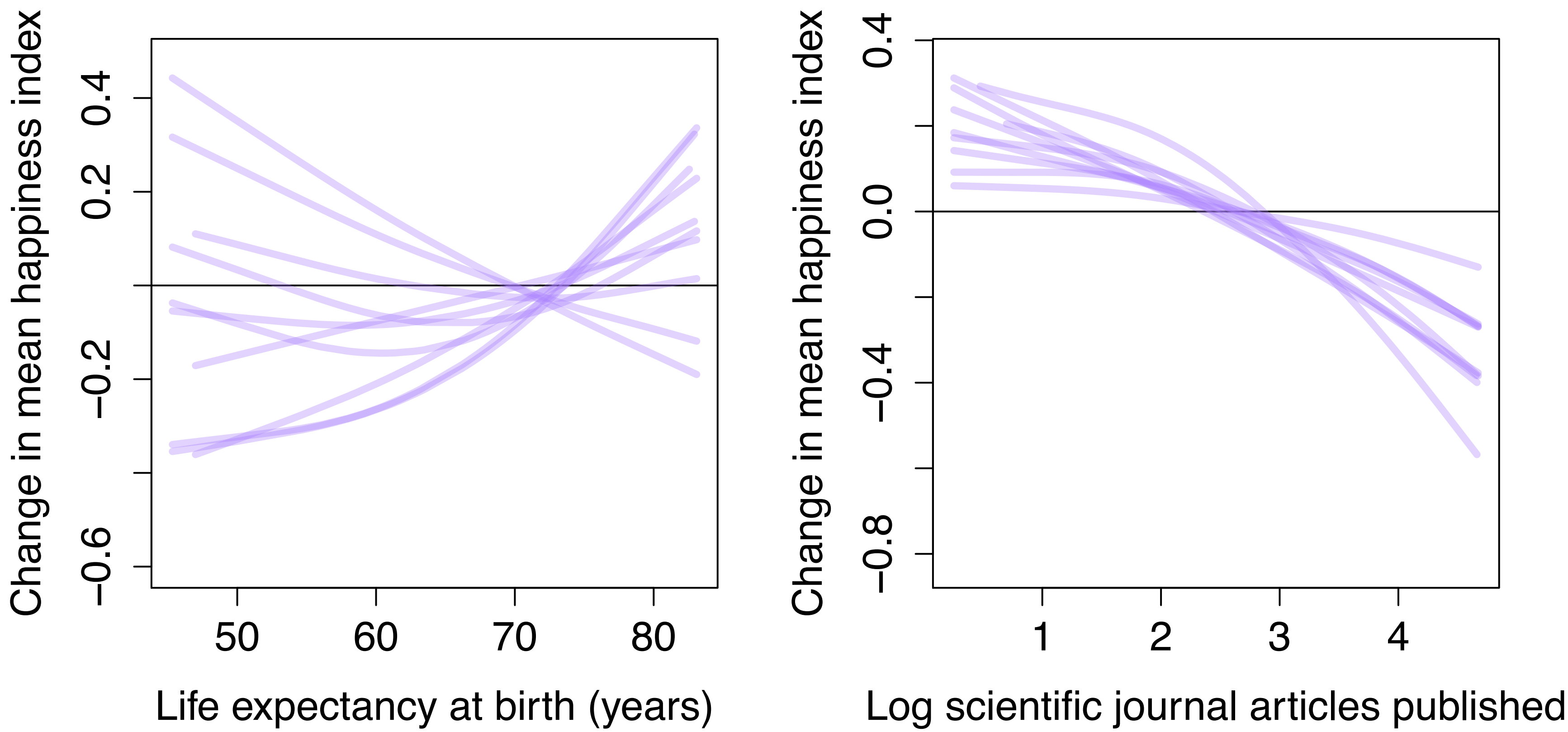
- Country-level data on 109 countries
- Outcome: happiness index from Cantril Scale
- Twelve predictors:
  - Log gross national income
  - Log scientific journal articles published
  - Percent satisfied with freedom of choice
  - Percent satisfied with job
  - Percent satisfied with community
  - Percent trusting in national government
  - Percent rural population
  - Percent females with secondary education
  - Mortality rate, under five
  - Life expectancy at birth
  - Percent Internet users
  - Percent labor force unemployed

**Probably, but we can't quite answer that question with our data*

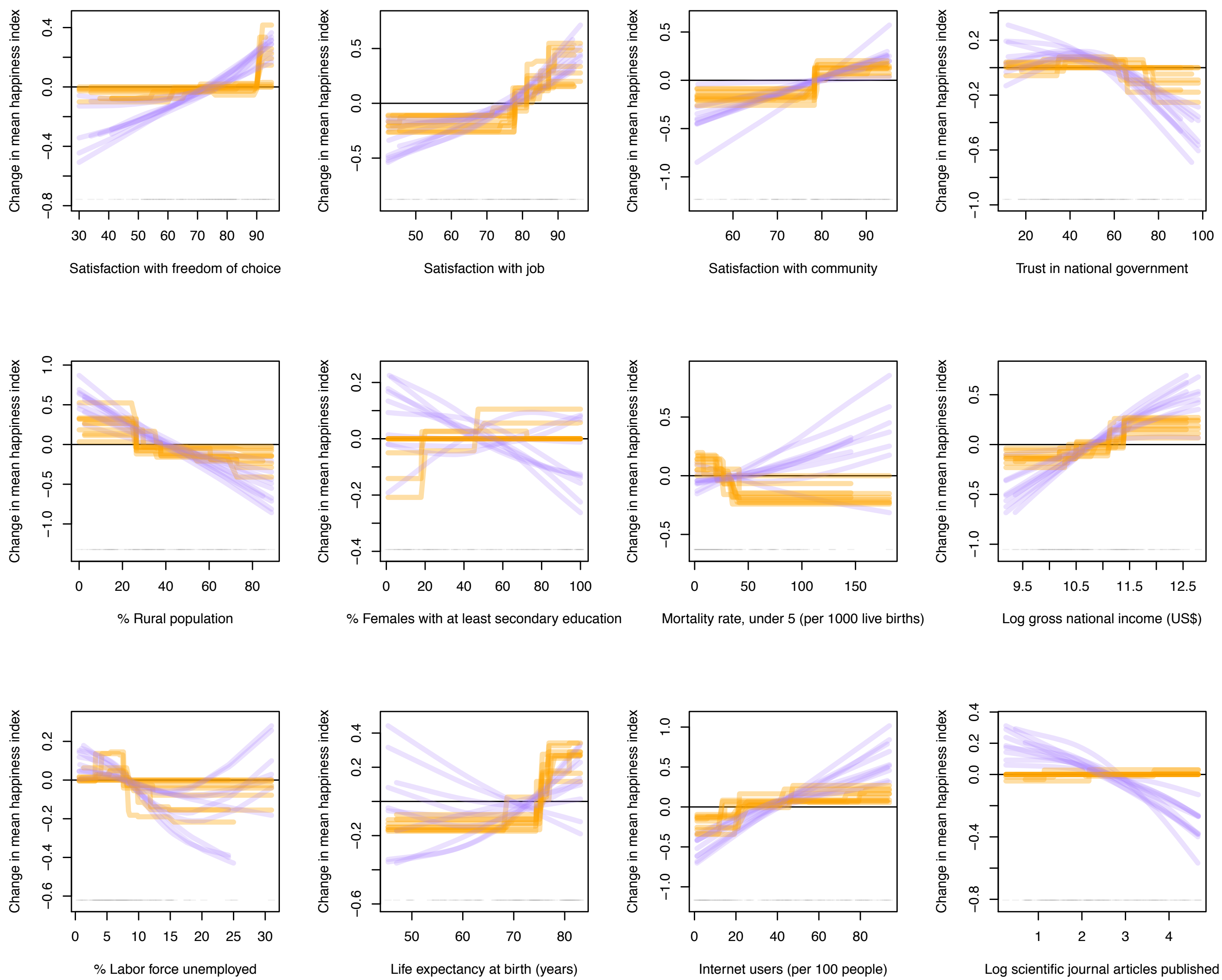
Additive model using smoothing splines



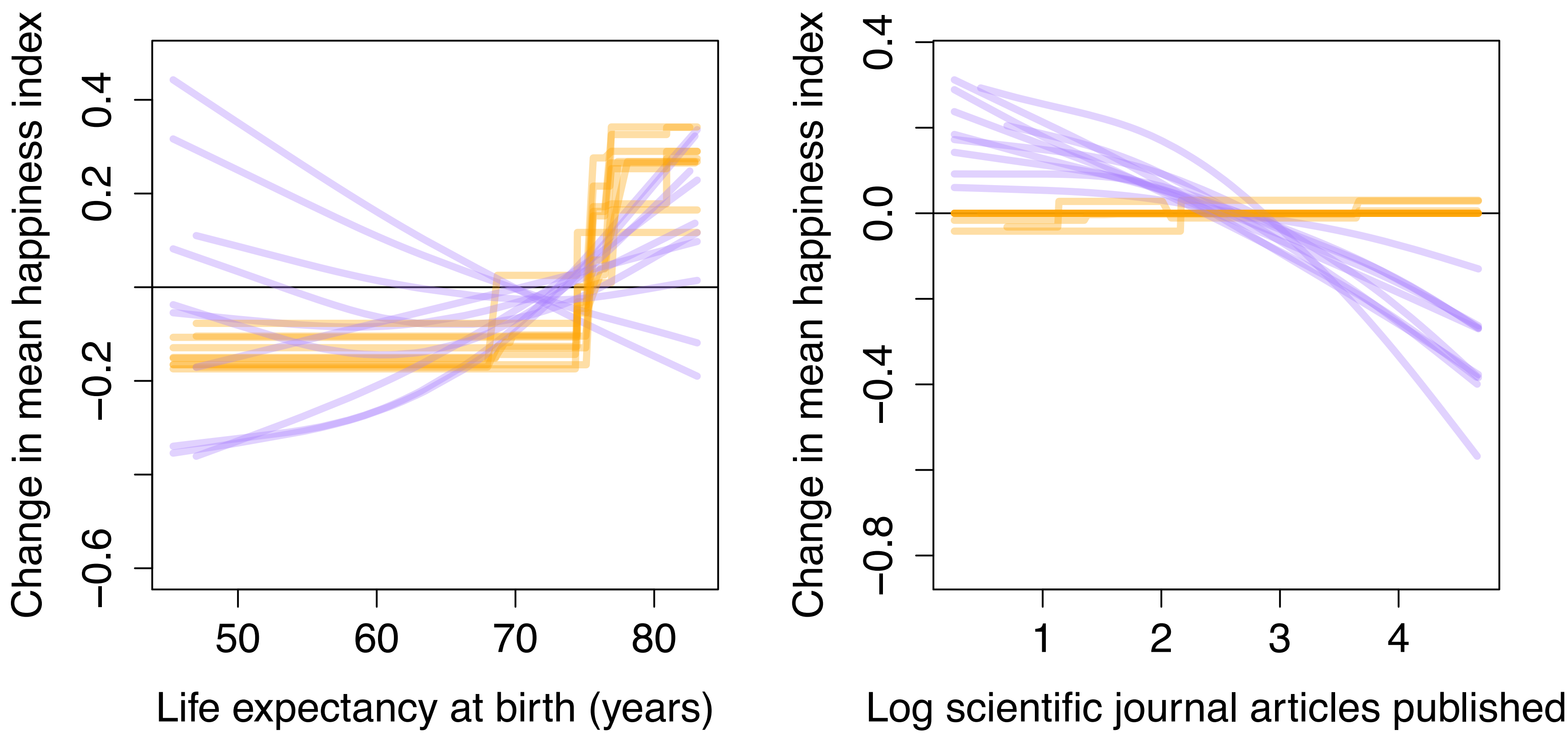
Additive model using smoothing splines



Using FLAM to predict happiness



Using FLAM to predict happiness



Inducing sparsity

- The World Bank and the United Nations don't just measure twelve covariates about countries
- There are **countless possible covariates** — many of which don't matter for predicting happiness
- Want to **induce sparsity**, i.e., estimate many of the $\theta_1, \ldots, \theta_p$ to be the zero vector

Inducing sparsity

- The World Bank and the United Nations don't just measure twelve covariates about countries
- There are **countless possible covariates** — many of which don't matter for predicting happiness
- Want to **induce sparsity**, i.e., estimate many of the $\theta_1, \ldots, \theta_p$ to be the zero vector

Add a second penalty to **induce sparsity**

$$\underset{\theta_0 \in \mathbb{R}, \boldsymbol{\theta}_j \in \mathbb{R}^n, 1 \leq j \leq p}{\text{minimize}} \quad \frac{1}{2} \left\| y - \sum_{j=1}^{p} \boldsymbol{\theta}_j - \theta_0 \mathbf{1} \right\|_2^2 + \alpha\lambda \sum_{j=1}^{p} \| D P_j \boldsymbol{\theta}_j \|_1 + (\mathbf{1} - \boldsymbol{\alpha}) \boldsymbol{\lambda} \sum_{j=1}^{p} \| \boldsymbol{\theta}_j \|_2$$

Solving FLAM (with α = 1)

Initialize $\hat{\theta}_j = \mathbf{0}$ for all j and $\hat{\theta}_0 = 0$. Cyclically iterate until convergence and for each $j = 1, \ldots, p$ perform the following:

1. Compute the residual $r_j = y - \sum_{j' \neq j} \hat{\theta}_{j'} - \hat{\theta}_0$.
2. Solve the optimization problem
$$\underset{\theta_j}{\text{minimize}} \quad \frac{1}{2} \| r_j - \theta_j \|_2^2 + \lambda \| D P_j \theta_j \|_1$$
using an algorithm for the fused lasso.
3. Compute the intercept, $\hat{\theta}_0 \leftarrow \hat{\theta}_0 + \text{mean}(\hat{\theta}_j)$, and center, $\hat{\theta}_j \leftarrow \hat{\theta}_j - \text{mean}(\hat{\theta}_j)$.

Solving FLAM

Initialize $\hat{\theta}_j = \mathbf{0}$ for all j and $\hat{\theta}_0 = 0$. Cyclically iterate until convergence and for each $j = 1, \ldots, p$ perform the following:

1. Compute the residual $r_j = y - \sum_{j' \neq j} \hat{\theta}_{j'} - \hat{\theta}_0$.
2. Solve the optimization problem
$$\underset{\boldsymbol{\theta}_j}{\textbf{minimize}} \quad \frac{1}{2} \|r_j - \boldsymbol{\theta}_j\|_2^2 + \alpha\boldsymbol{\lambda} \|DP_j\boldsymbol{\theta}_j\|_1 + (1-\alpha)\boldsymbol{\lambda}\|\boldsymbol{\theta}_j\|_2$$
using ??.
3. Compute the intercept, $\hat{\theta}_0 \leftarrow \hat{\theta}_0 + \text{mean}(\hat{\theta}_j)$, and center, $\hat{\theta}_j \leftarrow \hat{\theta}_j - \text{mean}(\hat{\theta}_j)$.

A useful result!

$$\boxed{\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|y-\boldsymbol{\theta}\|_2^2 + \alpha\lambda\,\|D\boldsymbol{\theta}\|_1} + (1-\alpha)\lambda\,\|\boldsymbol{\theta}\|_2$$

Solution $\hat{\boldsymbol{\theta}}$ obtained using algorithm for fused lasso

A useful result!

$$\boxed{\boxed{\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|y-\boldsymbol{\theta}\|_2^2 + \alpha\lambda\,\|D\boldsymbol{\theta}\|_1} + (1-\alpha)\lambda\,\|\boldsymbol{\theta}\|_2}$$

Solution $\hat{\boldsymbol{\theta}}$ obtained using algorithm for fused lasso

Solution is

$$\left(1-\frac{(1-\alpha)\lambda}{\|\hat{\boldsymbol{\theta}}\|_2}\right)_+ \hat{\boldsymbol{\theta}}$$

Solving FLAM

Initialize $\hat{\theta}_j = \mathbf{0}$ for all j and $\hat{\theta}_0 = 0$. Cyclically iterate until convergence and for each $j = 1, \ldots, p$ perform the following:

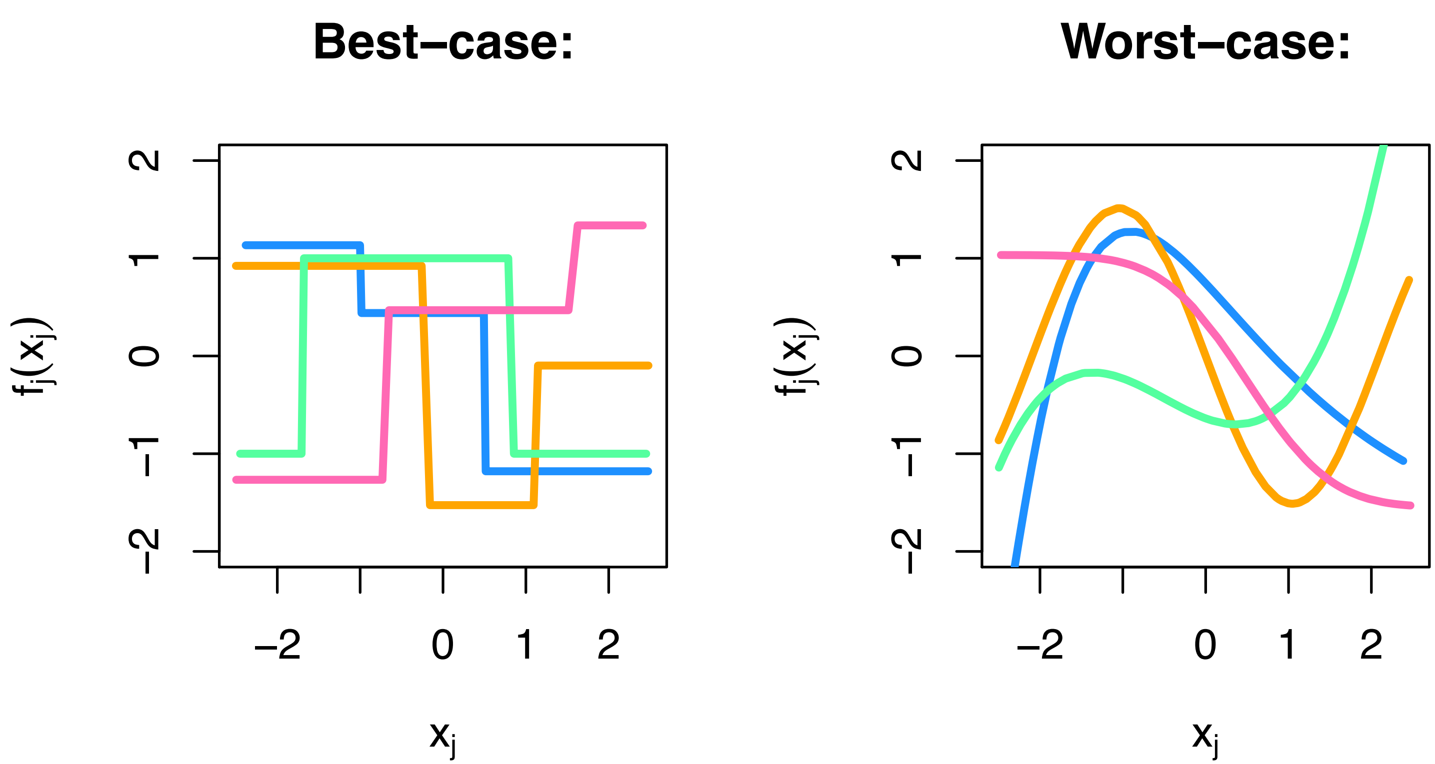
1. Compute the residual $r_j = y - \sum_{j' \neq j} \hat{\theta}_{j'} - \hat{\theta}_0$.
2. Solve the optimization problem
$$\underset{\theta_j}{\text{minimize}} \quad \frac{1}{2} \|r_j - \theta_j\|_2^2 + \alpha\lambda \|DP_j\theta_j\|_1$$
using an algorithm for the fused lasso.
3. Compute the intercept, $\hat{\theta}_0 \leftarrow \hat{\theta}_0 + \text{mean}(\hat{\theta}_j)$, and center, $\hat{\theta}_j \leftarrow \hat{\theta}_j - \text{mean}(\hat{\theta}_j)$.
4. Soft-scale the estimate: $\hat{\theta}_j \leftarrow \left(1 - \frac{(1-\alpha)\lambda}{\|\hat{\theta}_j\|_2}\right)_+ \hat{\theta}_j$.

Does FLAM work?

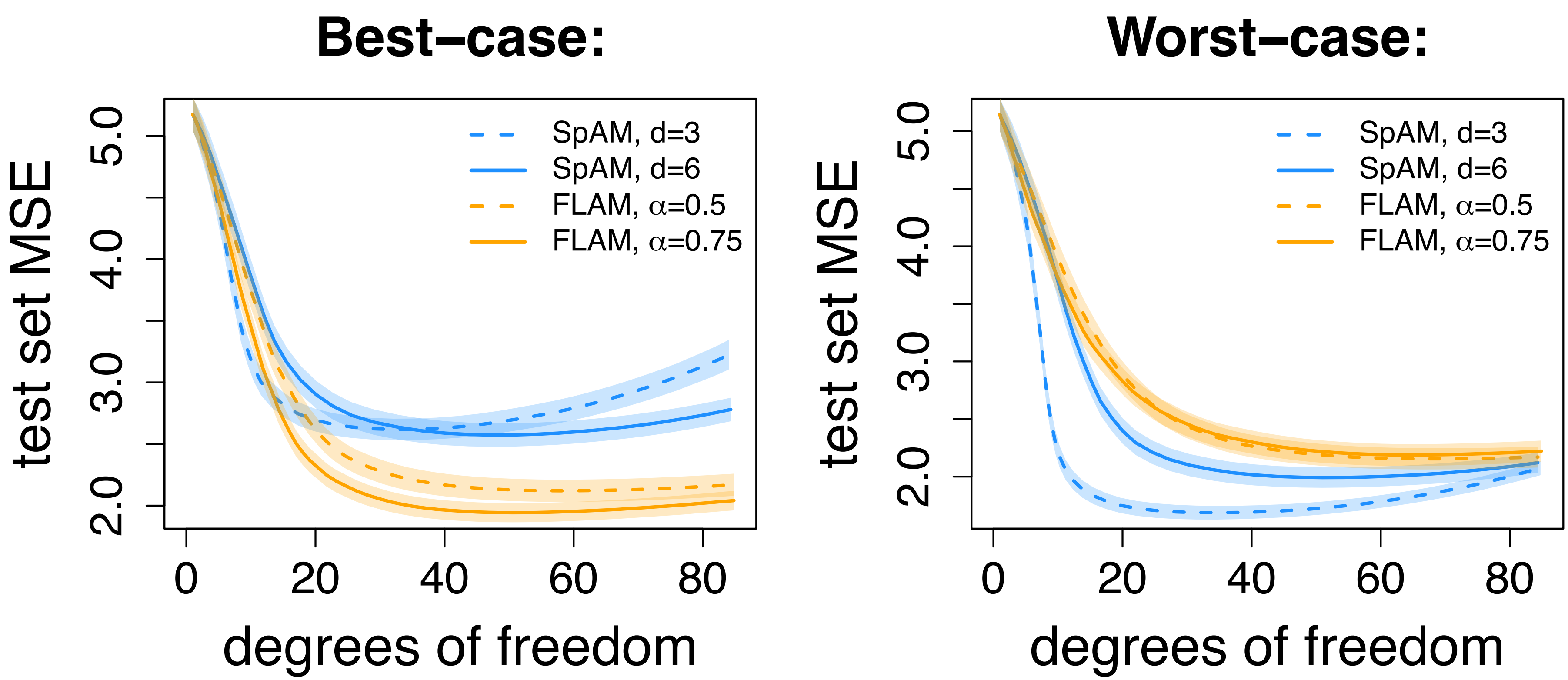
- Generate 100 observations for the training and test sets:

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \text{ with } \epsilon_i \sim N(0,1)$$

- Four non-zero f_j and ninety-six $f_j = 0$
- Compare FLAM to sparse additive model (SpAM)



Simulation results



SPLAT: sparse partially linear additive trend filtering

Sparse partially linear additive trend filtering

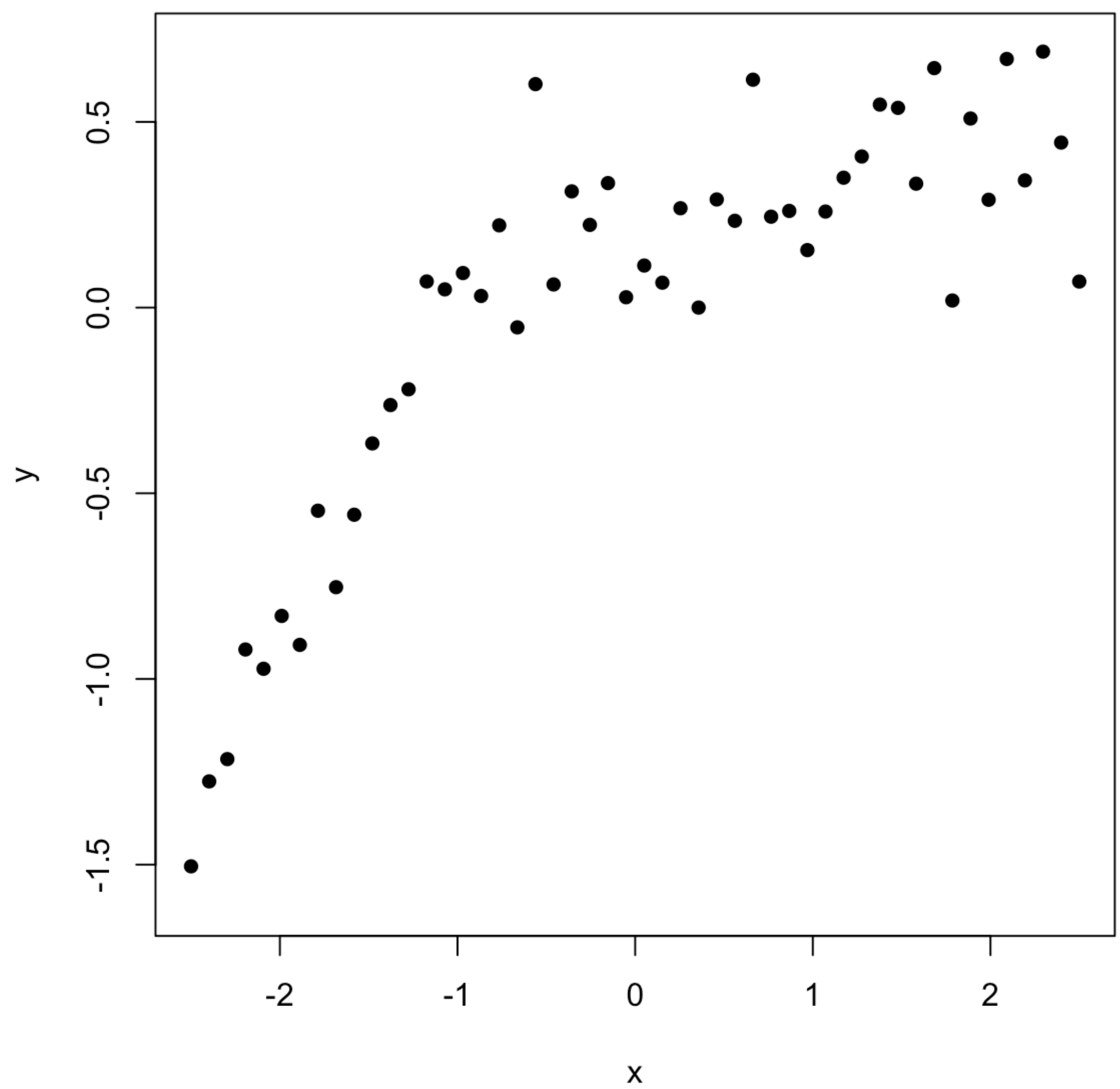
Goal: Fit the model

$$y = \sum_{j=1}^{p} f_j(x_j) + \epsilon$$

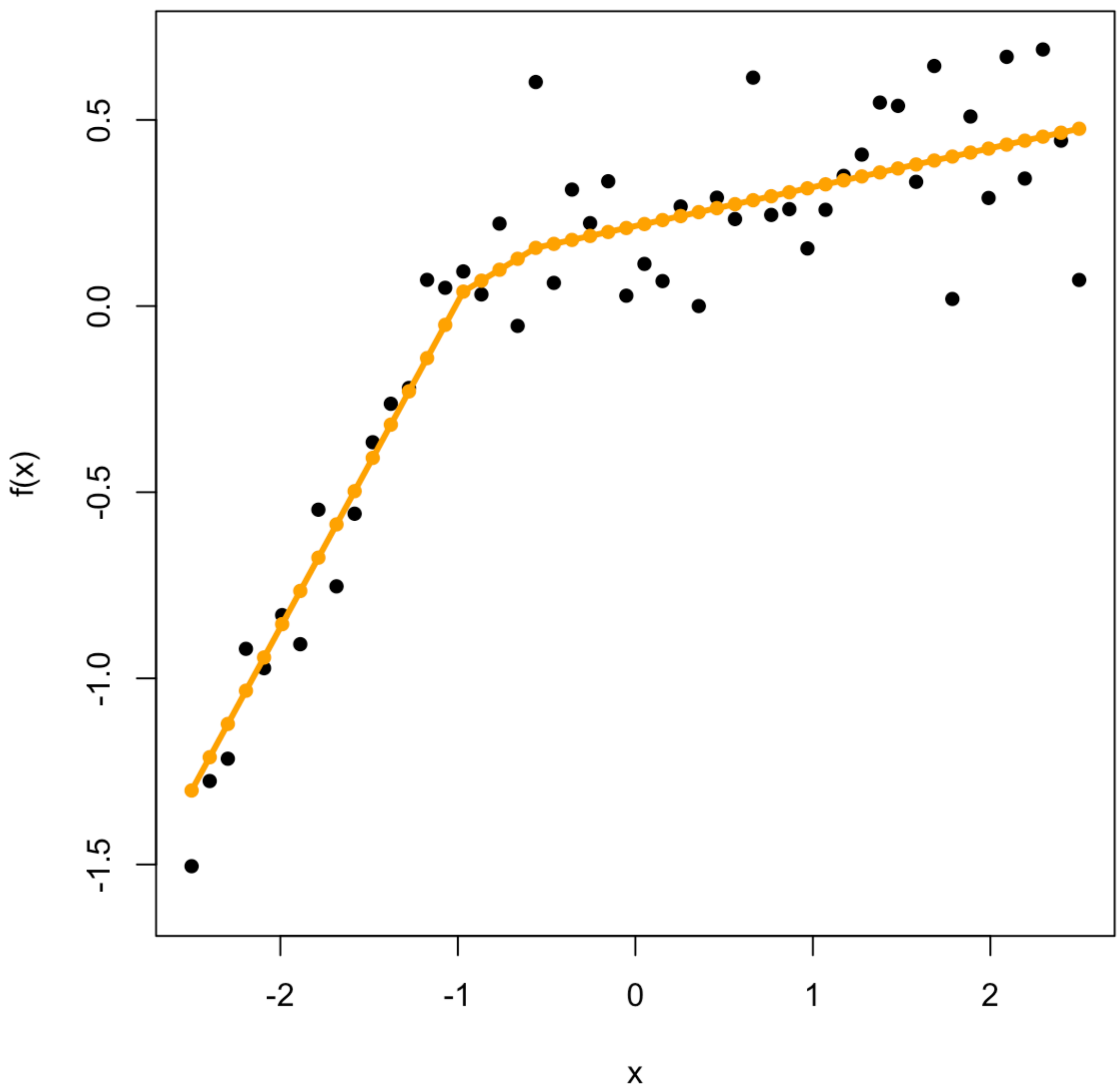
in a way that is simultaneously flexible and interpretable.

Estimate $f_1, \ldots, f_p$ to each be either linear or piecewise polynomial with a small number of adaptively-chosen knots

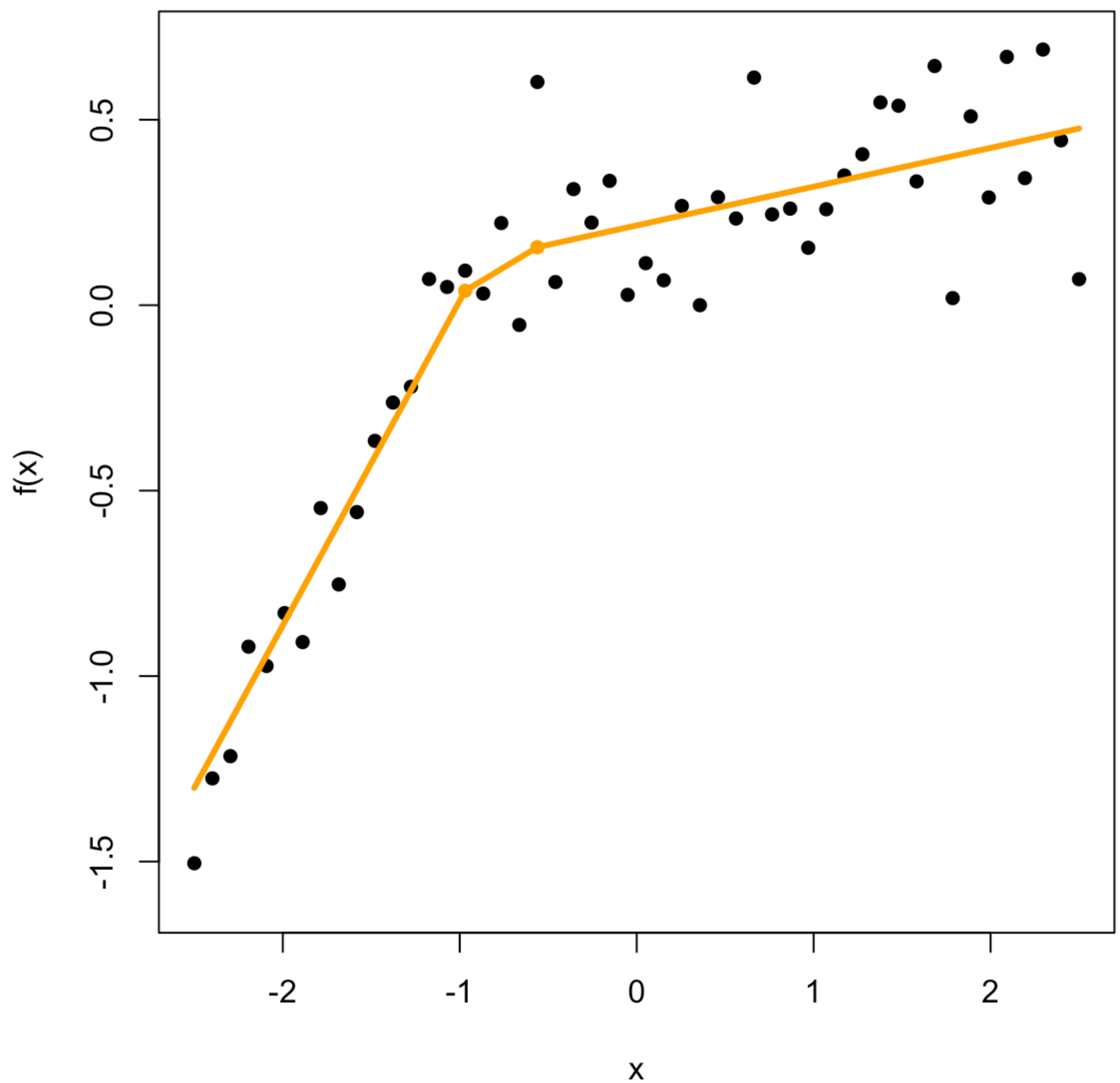
Working with a single covariate



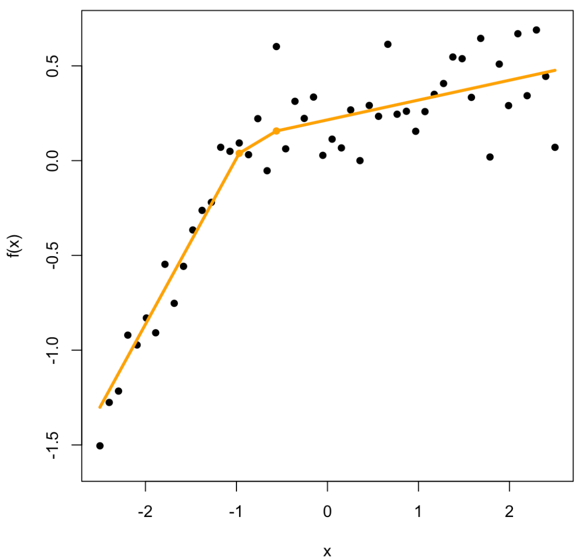
Working with a single covariate



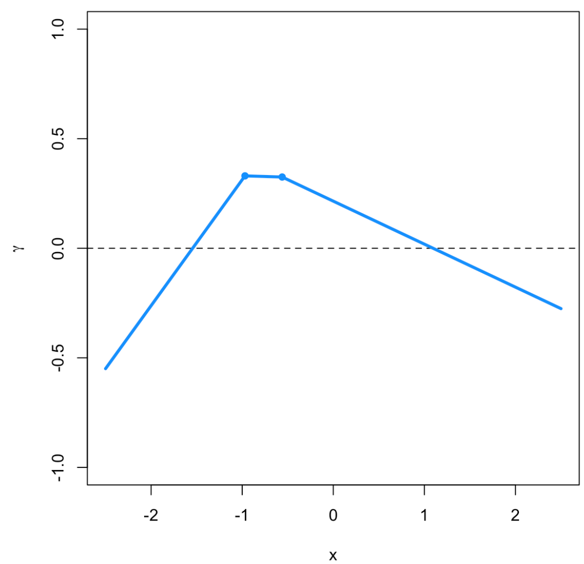
Working with a single covariate



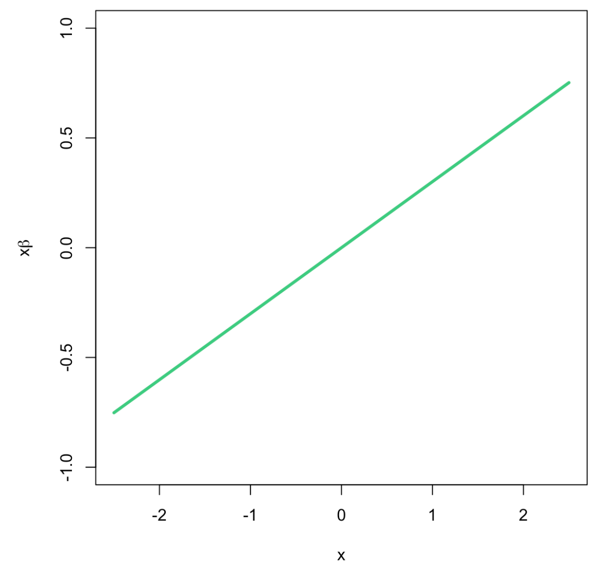
Decomposition of fit



=



+

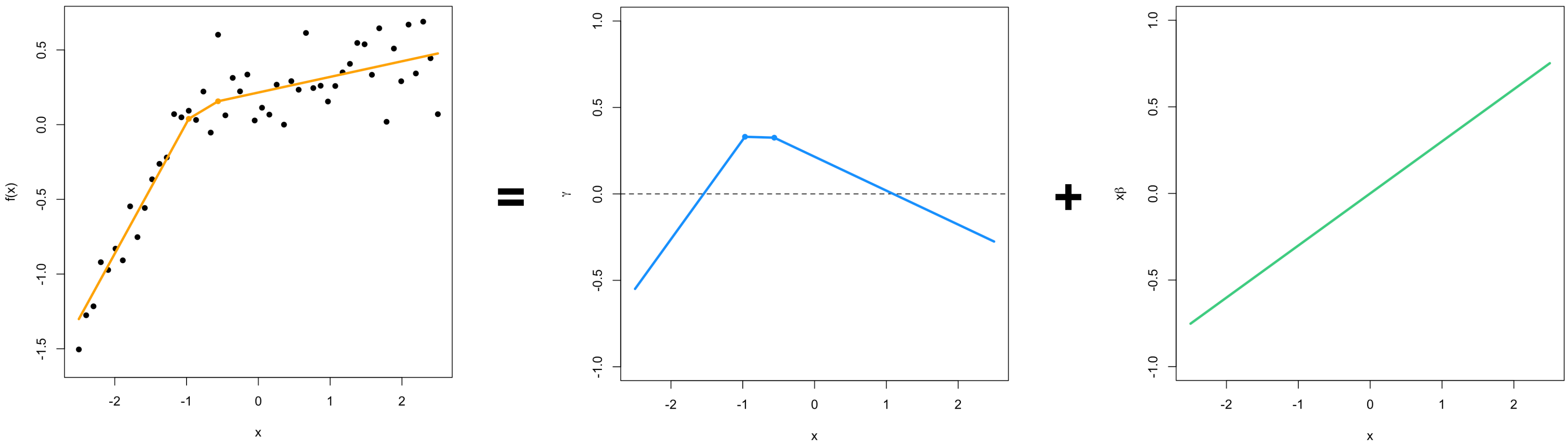


overall fit

non-linear fit

linear fit

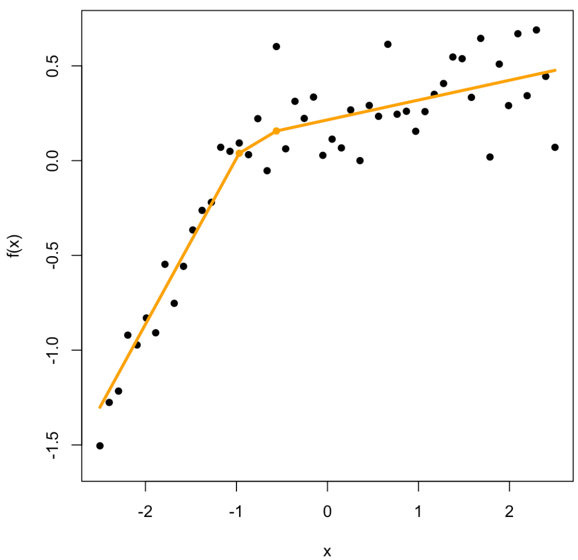
Optimization problem for single covariate



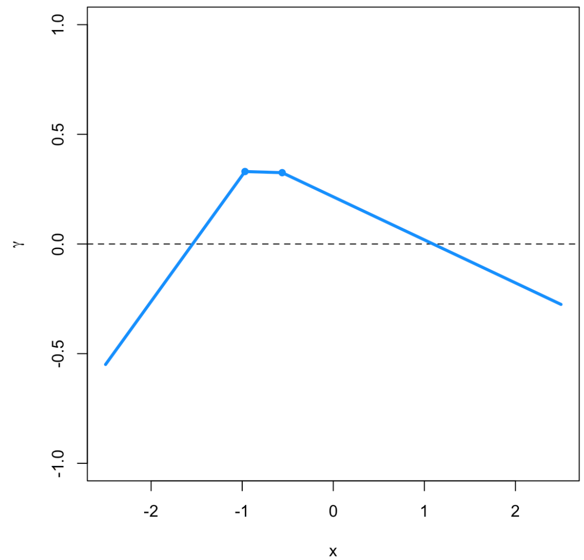
overall fit: θ **non-linear fit: γ** **linear fit: xβ**

$$\underset{\boldsymbol{\theta},\boldsymbol{\gamma}\in\mathbb{R}^n,\beta\in\mathbb{R}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|_2^2 + \alpha\lambda \left\|\boldsymbol{D}^{(k+1)}\boldsymbol{\gamma}\right\|_1 + (1-\alpha)\lambda\|\boldsymbol{\gamma}\|_2 \quad \text{subject to} \quad \boldsymbol{\theta} = \boldsymbol{x}\beta+\boldsymbol{\gamma}$$

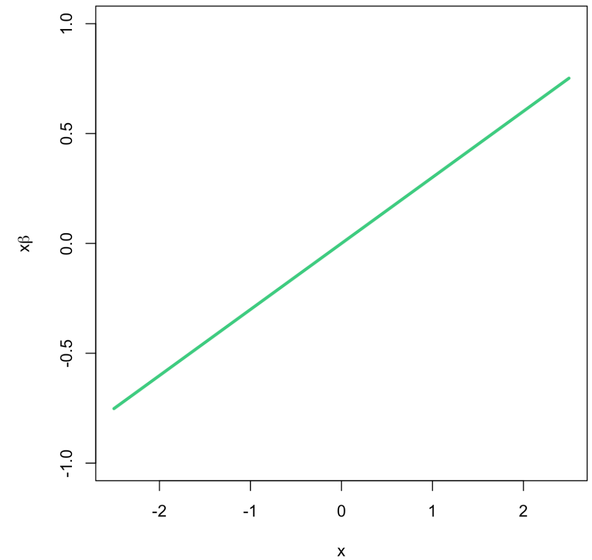
Optimization problem for single covariate



=



+

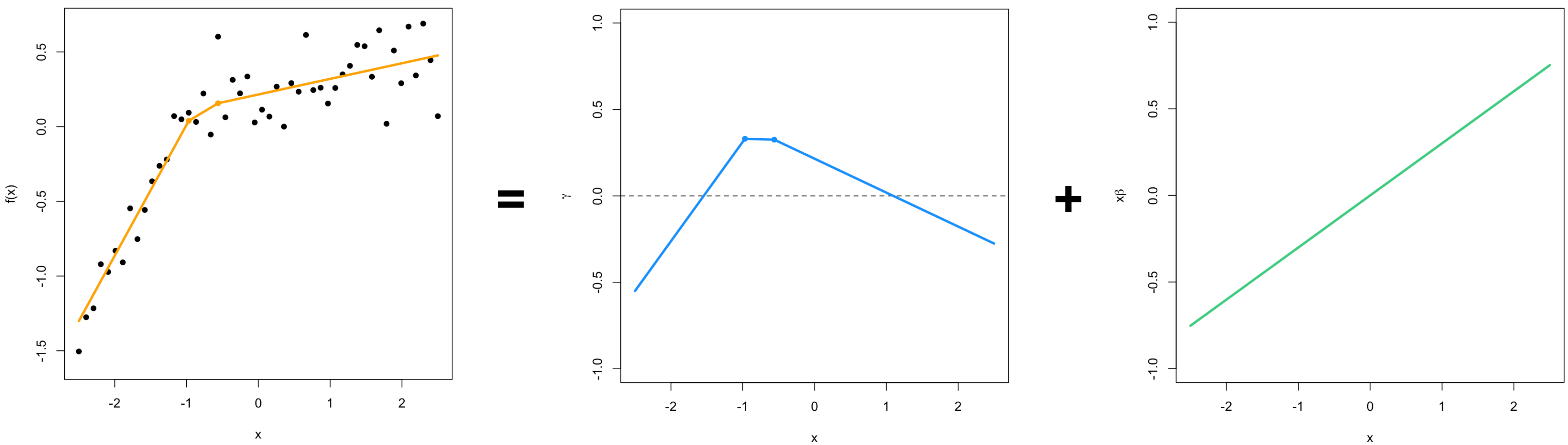


overall fit: θ **non-linear fit: γ** **linear fit: xβ**

$$\underset{\boldsymbol{\theta},\boldsymbol{\gamma}\in\mathbb{R}^n,\beta\in\mathbb{R}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|_2^2 + \alpha\lambda \boxed{\left\|\boldsymbol{D}^{(k+1)}\boldsymbol{\gamma}\right\|_1} + (1-\alpha)\lambda\|\boldsymbol{\gamma}\|_2 \quad \text{subject to} \quad \boldsymbol{\theta} = \boldsymbol{x}\beta + \boldsymbol{\gamma}$$

limits number of knots

Optimization problem for single covariate



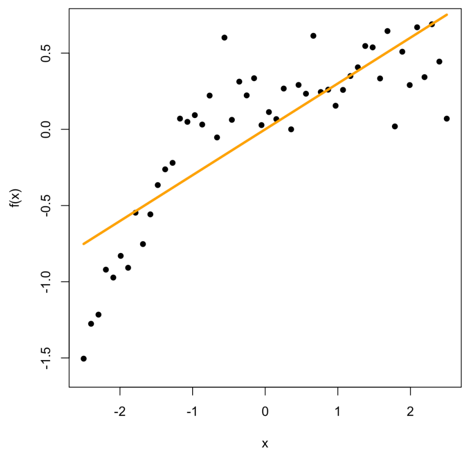
overall fit: θ **non-linear fit: γ** **linear fit: xβ**

$$\underset{\boldsymbol{\theta},\boldsymbol{\gamma}\in\mathbb{R}^n,\beta\in\mathbb{R}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|_2^2+\alpha\lambda\left\|\boldsymbol{D}^{(k+1)}\boldsymbol{\gamma}\right\|_1+(1-\alpha)\lambda\boxed{\|\boldsymbol{\gamma}\|_2} \quad \text{subject to} \quad \boldsymbol{\theta}=\boldsymbol{x}\beta+\boldsymbol{\gamma}$$

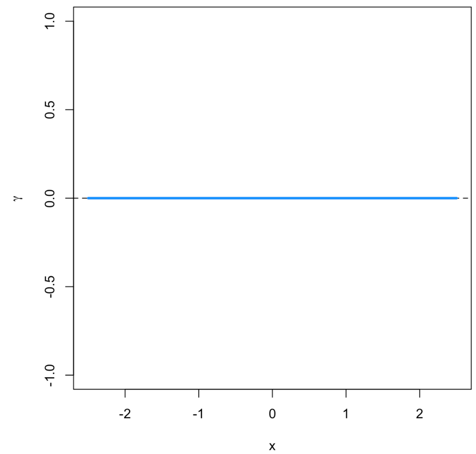
encourages linear fit

Impact of λ

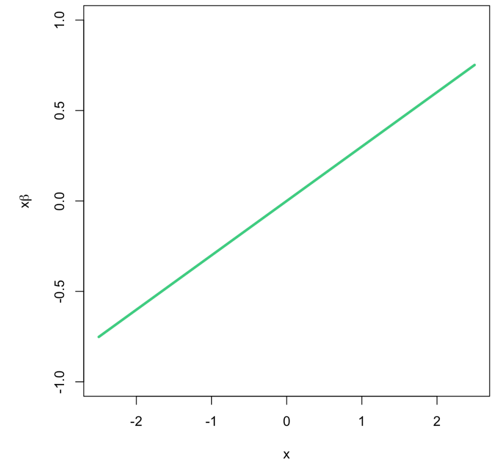
Large λ



=

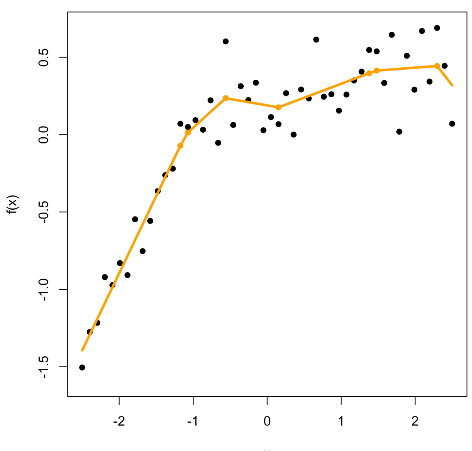


+

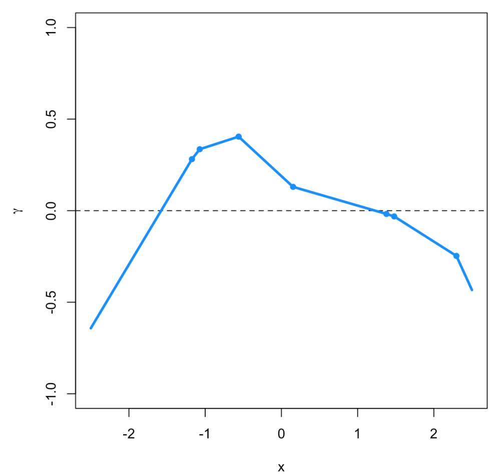


overall fit: θ **non-linear fit: γ** **linear fit: xβ**

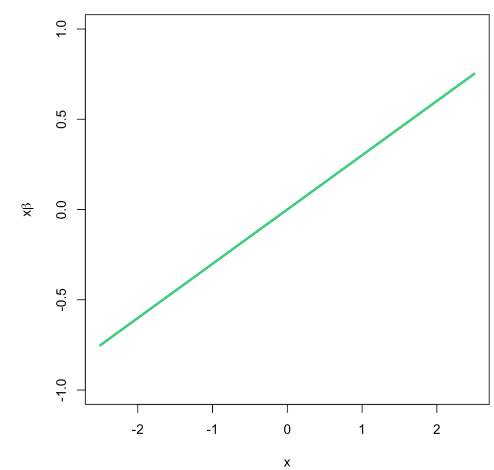
Small λ



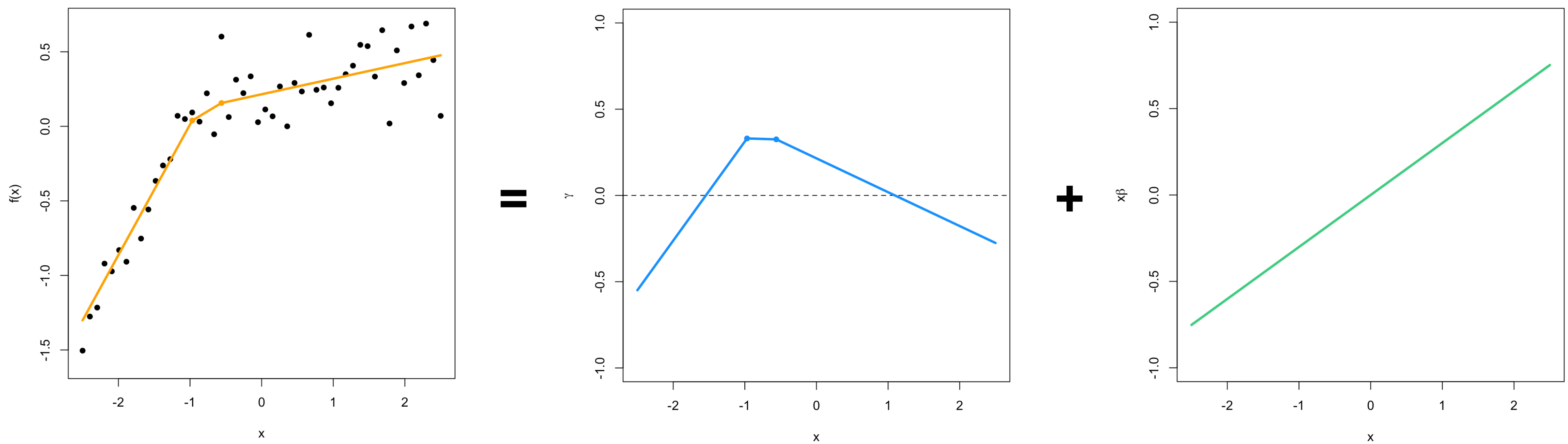
=



+



SPLAT penalties



overall fit: θ **non-linear fit: γ** **linear fit: xβ**

$$\underset{\boldsymbol{\theta}_j,\boldsymbol{\gamma}_j\in\mathbb{R}^n,1\le j\le p;\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\left\|\boldsymbol{y}-\sum_{j=1}^{p}\boldsymbol{\theta}_j\right\|_2^2 + \alpha\lambda\boxed{\sum_{j=1}^{p}\left\|\boldsymbol{D}^{(\boldsymbol{P}_j\boldsymbol{x}_j,k+1)}\boldsymbol{P}_j\boldsymbol{\gamma}_j\right\|_1} + (1-\alpha)\lambda\boxed{\sum_{j=1}^{p}\left\|\boldsymbol{\gamma}_j\right\|_2} + \tilde{\lambda}\boxed{\sum_{j=1}^{p}\left\|\boldsymbol{\theta}_j\right\|_2}$$

$$\text{subject to} \quad \boldsymbol{\theta}_j = \boldsymbol{x}_j\beta_j + \boldsymbol{\gamma}_j \quad \forall j,$$

controls complexity of non-linear fits

allows a linear or non-linear fit

performs variable selection

Solving SPLAT

Optimization problem for p = 1:

$$\underset{\boldsymbol{\theta},\boldsymbol{\gamma}\in\mathbb{R}^n,\beta\in\mathbb{R}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|_2^2 + \alpha\lambda\left\|\boldsymbol{D}^{(\boldsymbol{P}\boldsymbol{x},k+1)}\boldsymbol{P}\boldsymbol{\gamma}\right\|_1 + (1-\alpha)\lambda\|\boldsymbol{\gamma}\|_2 + \tilde{\lambda}\|\boldsymbol{\theta}\|_2 \quad \text{subject to} \quad \boldsymbol{\theta} = \boldsymbol{x}\beta + \boldsymbol{\gamma}.$$

We prove that the solution is:

$$\left(1 - \frac{\tilde{\lambda}}{\left\|\boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{x}^\top\boldsymbol{y} + \left(1-\frac{(1-\alpha)\lambda}{\|\tilde{\boldsymbol{\gamma}}\|_2}\right)_+\tilde{\boldsymbol{\gamma}}\right\|_2}\right)_+ \left(\boldsymbol{x}(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{x}^\top\boldsymbol{y} + \left(1-\frac{(1-\alpha)\lambda}{\|\tilde{\boldsymbol{\gamma}}\|_2}\right)_+\tilde{\boldsymbol{\gamma}}\right)$$

where $\tilde{\boldsymbol{\gamma}}$ is the solution to a trend filtering problem

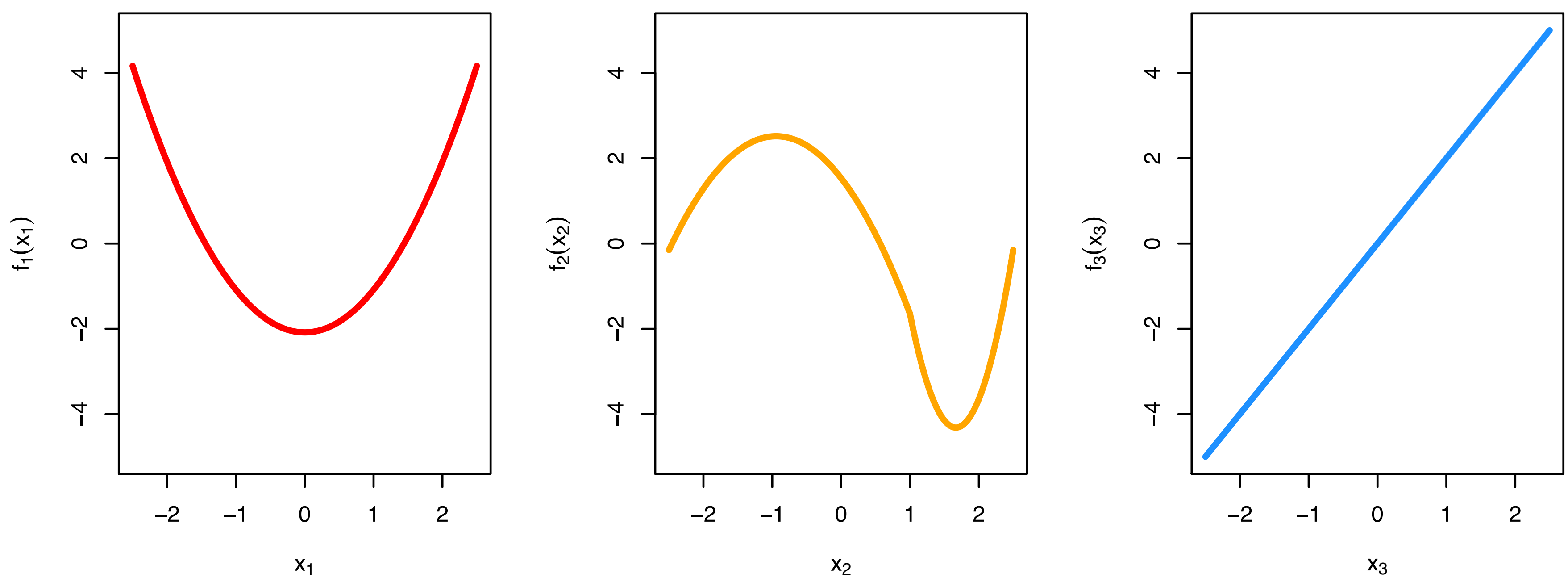
The solution is just a known function of $\tilde{\boldsymbol{\gamma}}$

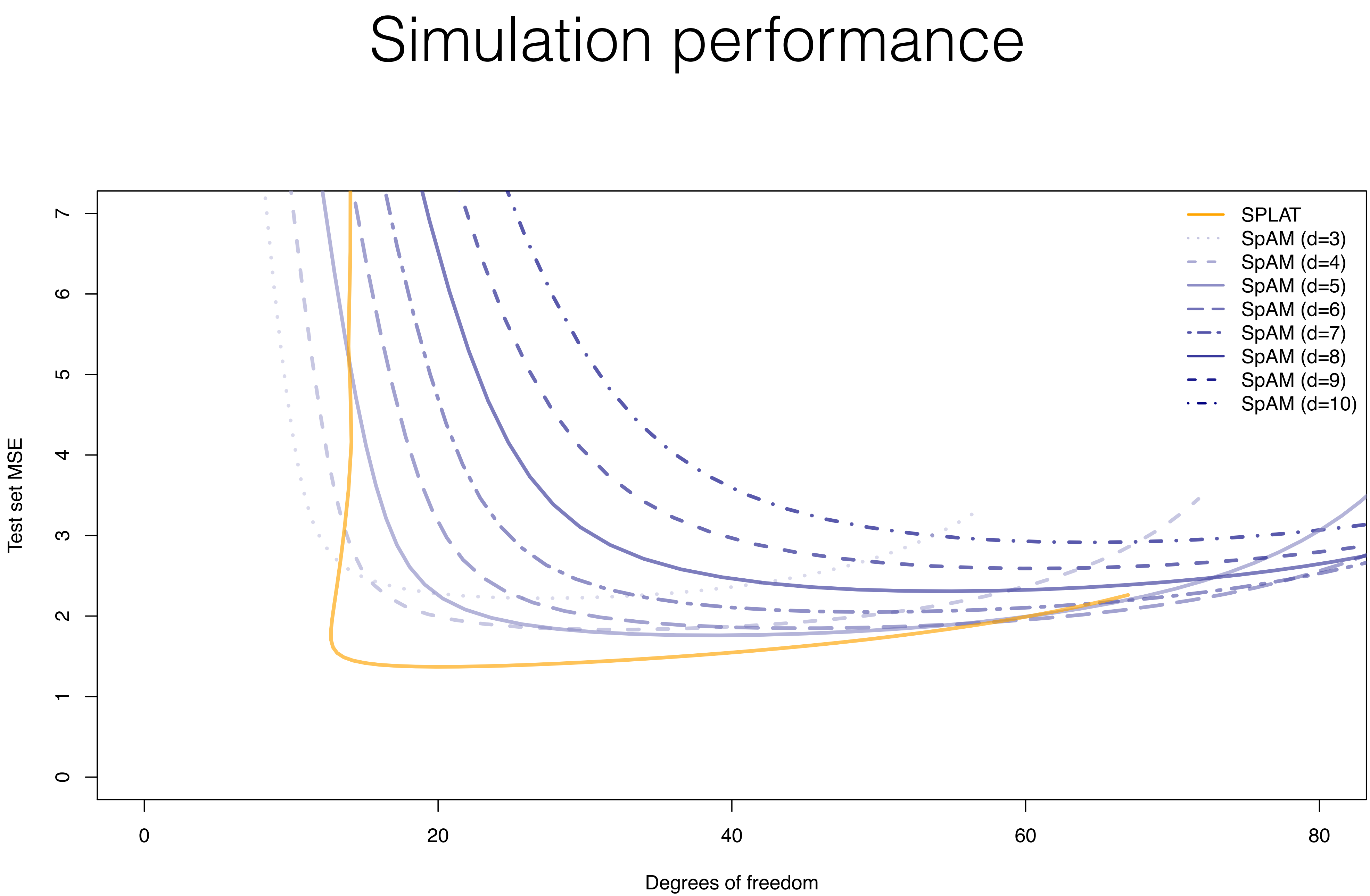
Testing out SPLAT's performance

- Generate 100 observations for the training, test, and validation sets:

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \text{ with } \epsilon_i \sim N(0,1)$$

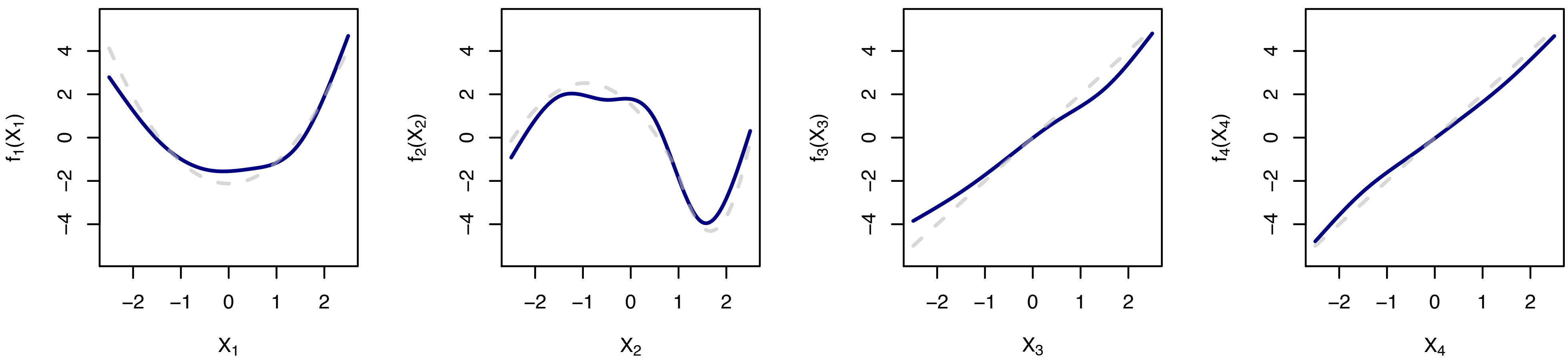
- Two non-linear f_j, two linear f_j, and sixteen $f_j = 0$
- Compare SPLAT to SpAM



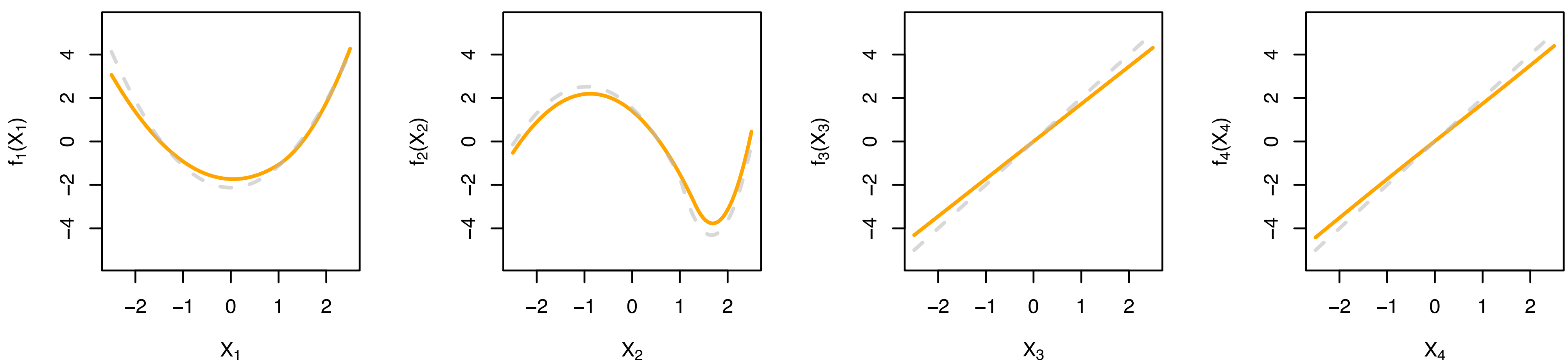

Simulation performance
Test set MSE
Degrees of freedom
0
1
2
3
4
5
6
7
0
20
40
60
80
SPLAT
SpAM (d=3)
SpAM (d=4)
SpAM (d=5)
SpAM (d=6)
SpAM (d=7)
SpAM (d=8)
SpAM (d=9)
SpAM (d=10)

Individual covariate fits

SpAM



SPLAT



overview of adaptive additive modeling

FLAM

Covariate fit = piecewise constant with adaptively chosen knots

- **Flexible:** adaptive selection of covariates and knots
- **Interpretable:** simple piecewise constant fits
- Applicable when $p > n$

FLAM

Covariate fit = piecewise constant with adaptively chosen knots

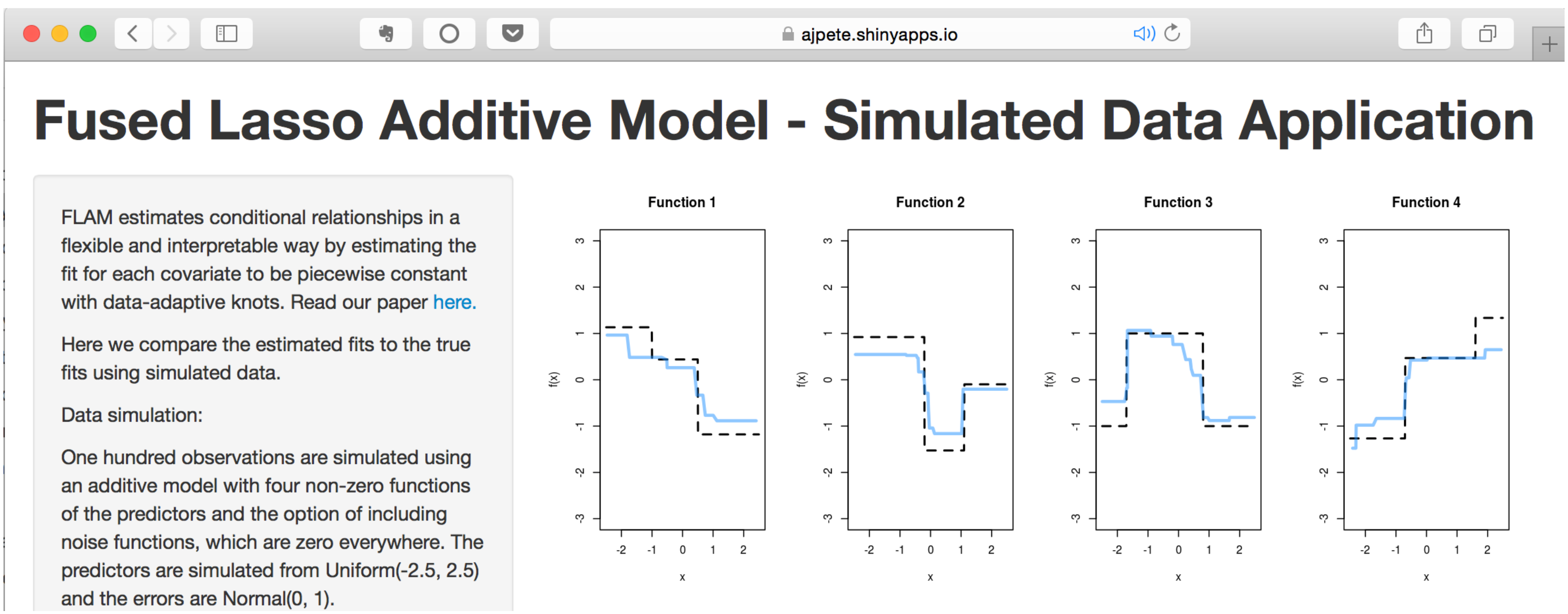
- **Flexible:** adaptive selection of covariates and knots
- **Interpretable:** simple piecewise constant fits
- Applicable when $p > n$

SPLAT

- **higher-order piecewise fits**
- **adaptive selection of exactly linear fits**

Find out more

- FLAM is published in ***Journal of Computational and Graphical Statistics***
- R package flam available on CRAN
- Shiny apps for FLAM at **ajpete.com**
- Resources for SPLAT coming soon



Questions?