Regression splines: a review

David Ruppert
School of OR & IE
Cornell University

The overheads of this lecture are available at:
`http://www.orie.cornell.edu/~davidr/ruppert.html`

Outline

1. Regression splines are linear models
2. Splines allow essentially "nonparametric" fitting
3. High dimensional parameters require
   (a) careful selection of knots before least-squares estimation, **or**
   (b) alternative estimators, e.g., penalized least-squares
4. Generalized cross validation for choosing the penalty weight
5. Estimation of derivatives
6. Bayesian interpretation
7. Additive models
8. Interactions
9. Measurement error

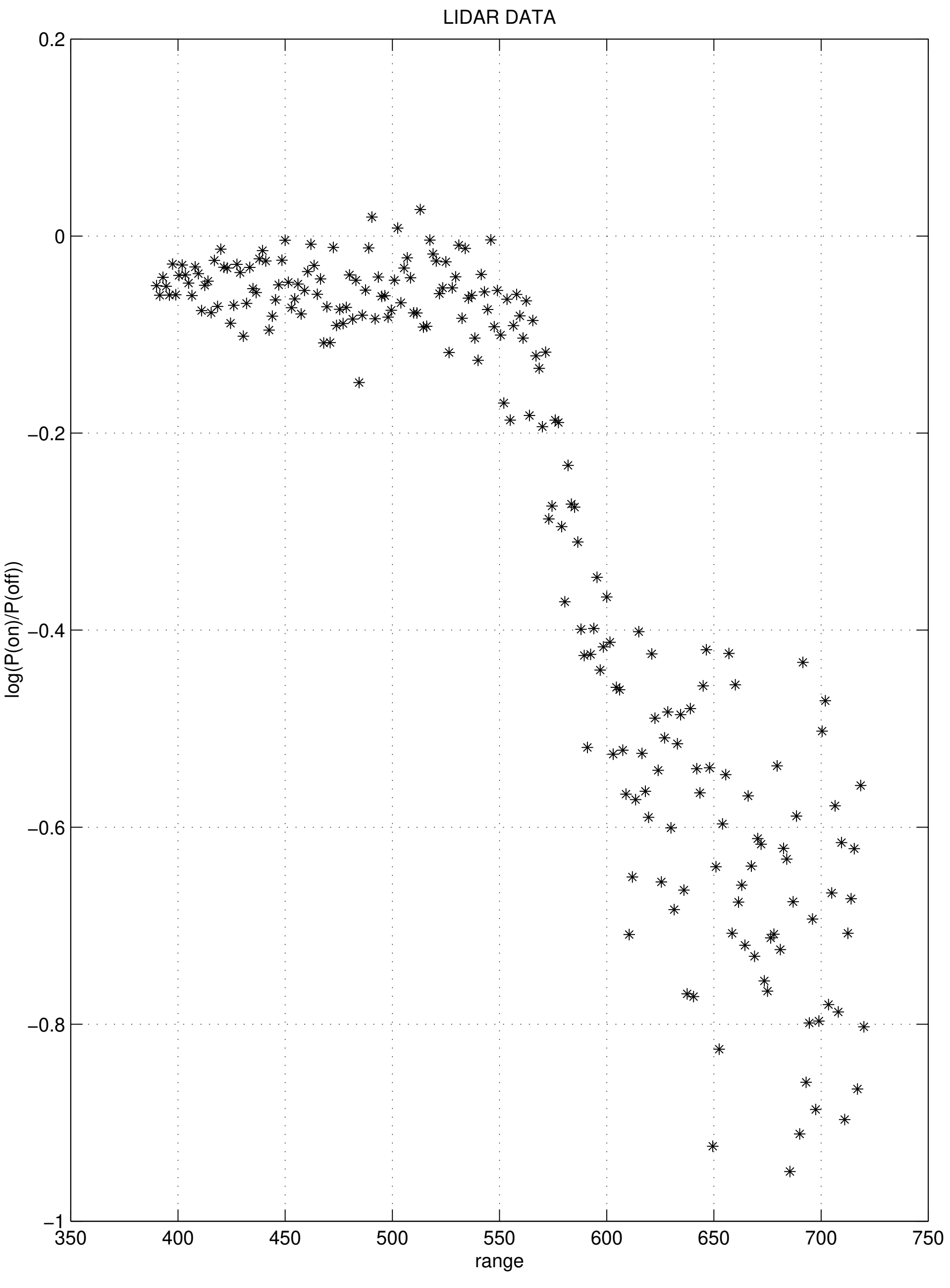


Figure 1: Example showing the need for nonparametric models

Nonparametric models

Nonparametric models are those models where the forms of the functions in the model, e.g., density functions or conditional expectations, are dictated by the data.

Two approaches:

1. Fit simple parametric models *locally*, e.g., moving averages or local polynomial estimation
2. Fit a highly complex parametric model with a complexity penalty to prevent "overfitting"

Roughly, "overfitting" means fitting the noise, not the signal, so that features are detected the would not be present in an independent replicate of the data.

Both approaches require the selection of a "smoothing parameter."

1. Local modeling needs a bandwidth to determine the degree of "localness."
2. Penalized estimation needs a penalty weight to determine the tradeoff between goodness-of-fit and complexity.

In this talk, approach 2 is used.

Linear models, polynomial regression, and splines

Linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \text{error}$$

Polynomial regression

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \text{error}$$

Linear in $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^t \Rightarrow$ a linear model.

Example: LIDAR data

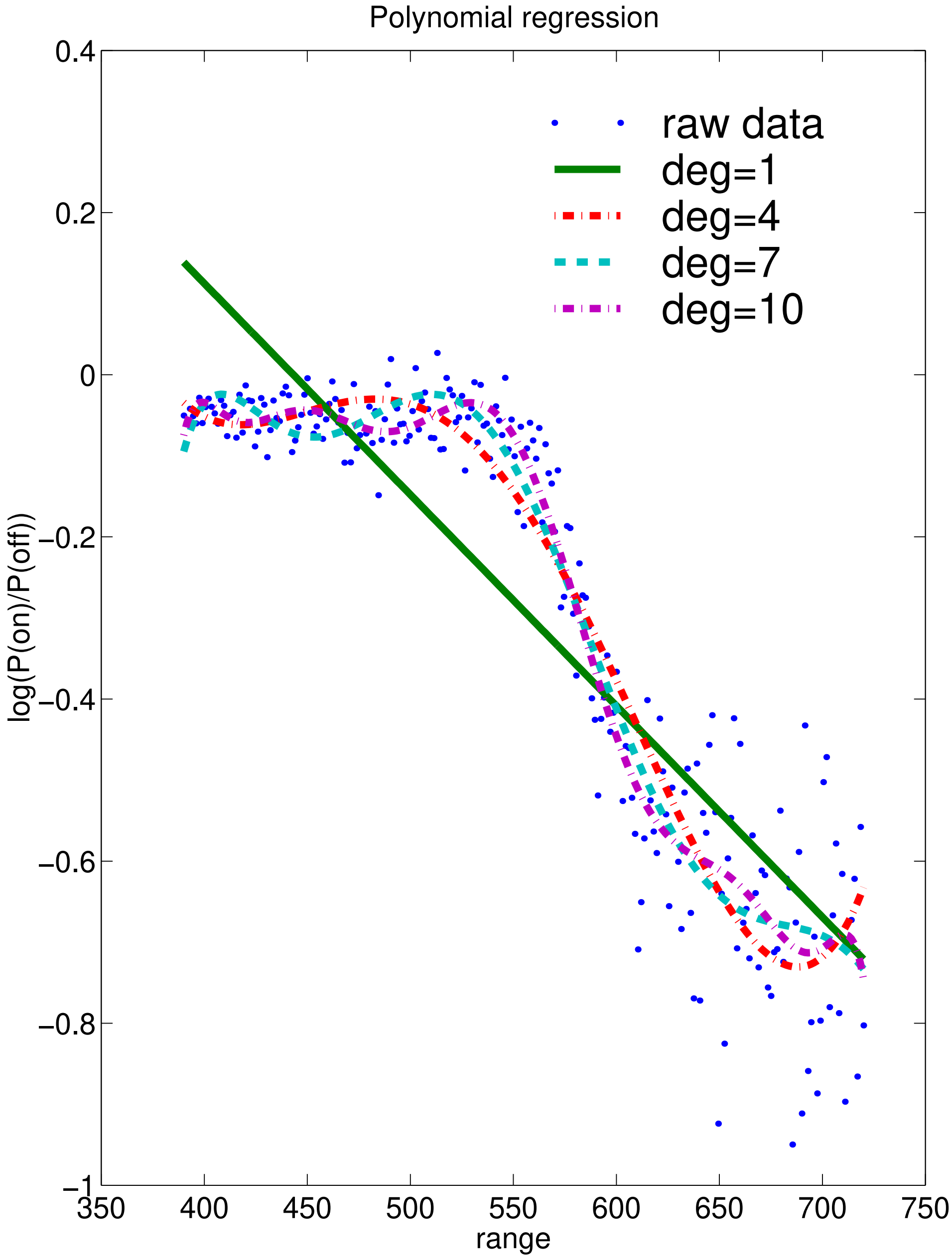


Figure 2: Lidar data: polynomial fits.

Polynomial-spline regression

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \sum_{j=1}^{K} \beta_{p+j}(X_i - t_j)_+^p + \text{error}$$

- Here $t_1 < \ldots < t_K$ are fixed "knots"
- $u_+^p = u^p$ if $u > 0$ and equals 0 otherwise.
- $(x - t)_+^p$ will be called a truncated power function.
- Linear in $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p+K})^t \Rightarrow$ again a linear model.

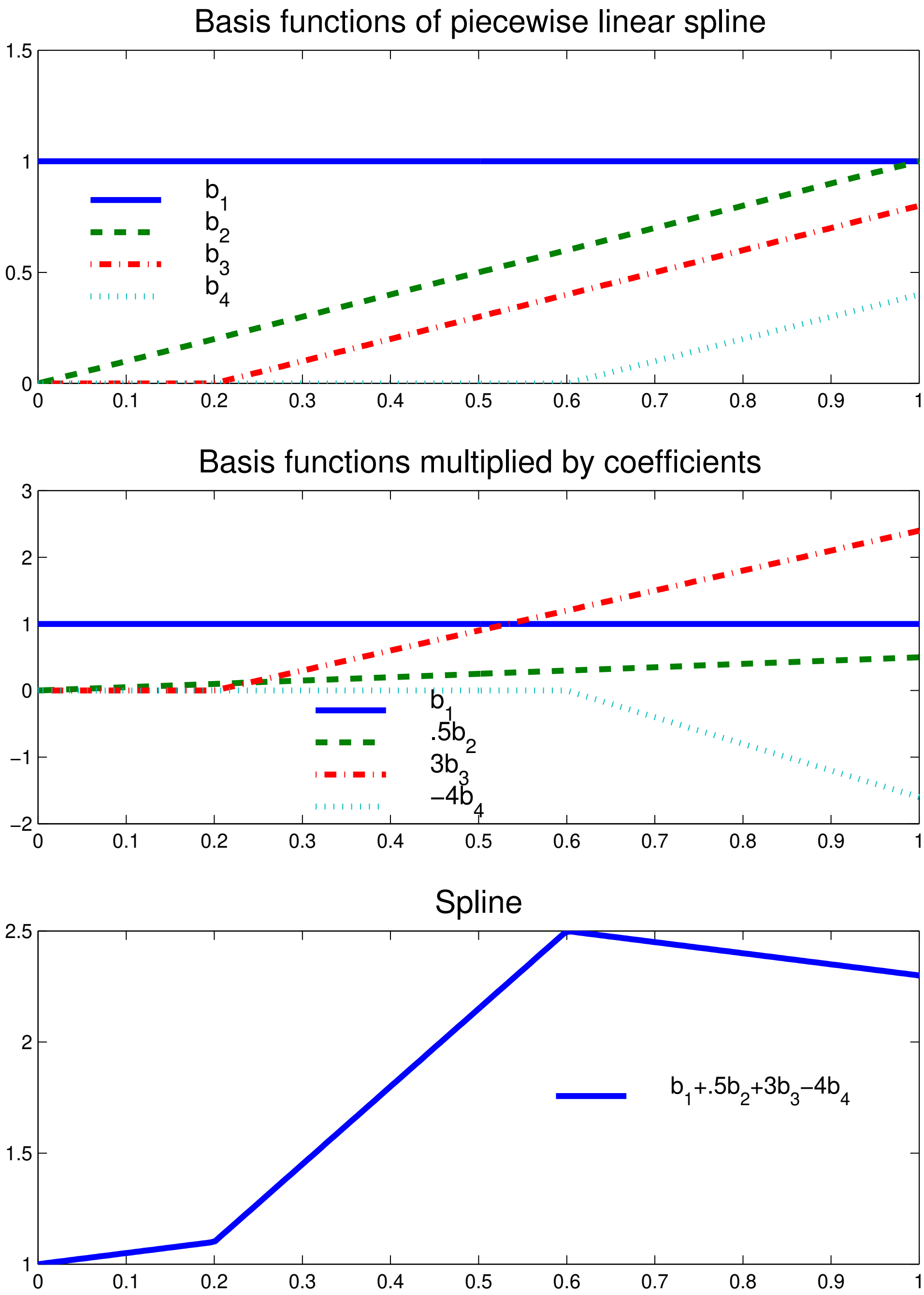


Figure 3: Piecewise linear spline and basis functions.

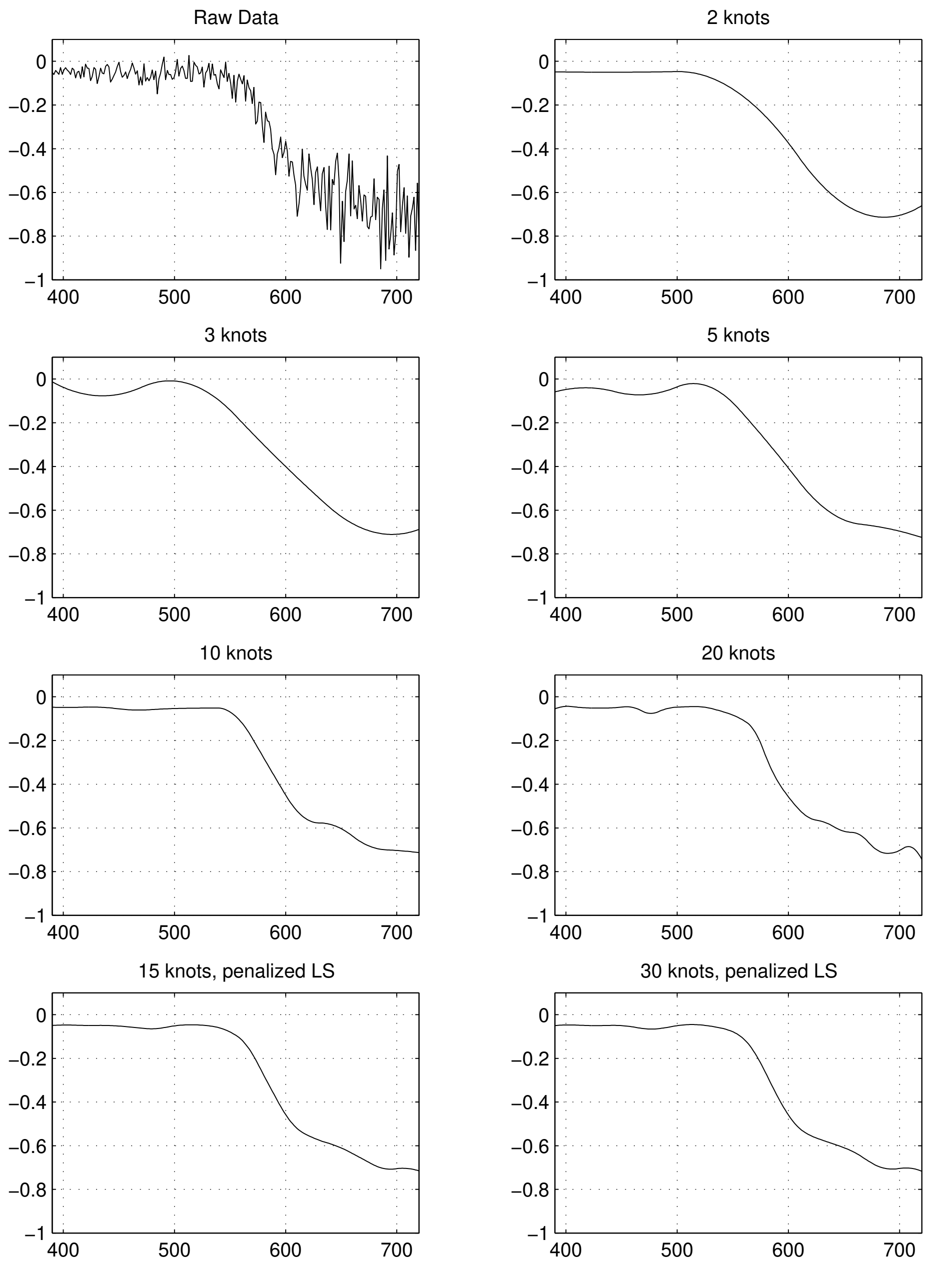


Figure 4: Lidar data: Spline fits with and without penalty

Knot placement and least-squares

References

- Friedman and Silverman (1989) — TURBO
- Friedman (1991) — MARS (**m**ultivariate **a**daptive **r**egression **s**plines)
- Stone et al. (1997)

Knots are added sequentially with the set of candidate knots depending on knots already selected.

Selection is driven by a goodness-of-fit statistic, assuming ordinary least-squares (OLS) estimation.

After a maximum number of knots has been added, knots are deleted one at a time.

The final model optimizes a statistic combining goodness-of-fit with a complexity penalty.

The final model is fit by OLS.

See Chen, Ruppert, and Shoemaker (1997) for an application of MARS.

Penalized Least Squares

$$\boldsymbol{X}_i^t = \left\{1 \; X_i \; \cdots \; X_i^p \; (X_i - t_1)_+^p \; \cdots \; (X_i - t_K)_+^p\right\}$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1^t \\ \vdots \\ \boldsymbol{X}_n^t \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p+K} \end{pmatrix}$$

$$\mathrm{PSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{Y_i - \boldsymbol{X}_i^t \boldsymbol{\beta}\right\}^2 + \alpha \sum_{j=1}^{K} \beta_{p+j}^2$$

$\widehat{\beta}(\alpha)$ minimizes $\mathrm{PSS}(\beta)$.

Here $\alpha > 0$ is a penalty weight. As $\alpha \to \infty$, the penalized spline converges to a pth degree polynomial fit. As $\alpha \to 0$, the penalized spline converges to the OLS fitted spline.

β_{p+j} is the jump in the pth derivative at the knot t_j. Thus, the penalty is on the $p+1$th derivative, interpreted as a generalized function.

Penalized Least Squares — cont.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ and $\boldsymbol{D} = \mathrm{diag}(0, \ldots, 0, 1, \ldots, 1)$ ($p+1$ zeros and K ones). Then

$$\mathrm{PSS}(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \alpha\|\boldsymbol{D}\boldsymbol{\beta}\|^2.$$

Fact: $\partial/\partial\,\boldsymbol{\beta}\|A\boldsymbol{\beta}\|^2 = 2A^2\boldsymbol{\beta}$ for any symmetric A.

Therefore, $\widehat{\boldsymbol{\beta}}(\alpha)$ solves

$$0 = \partial/\partial\boldsymbol{\beta}\ \mathrm{PSS}(\boldsymbol{\beta}) = -2\boldsymbol{X}^t(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\alpha\boldsymbol{D}\boldsymbol{\beta}$$

$\Rightarrow$

$$\widehat{\boldsymbol{\beta}}(\alpha) = (\boldsymbol{X}^t\boldsymbol{X} + \alpha\boldsymbol{D})^{-1}\boldsymbol{X}\boldsymbol{Y}.$$

(Ridge regression)

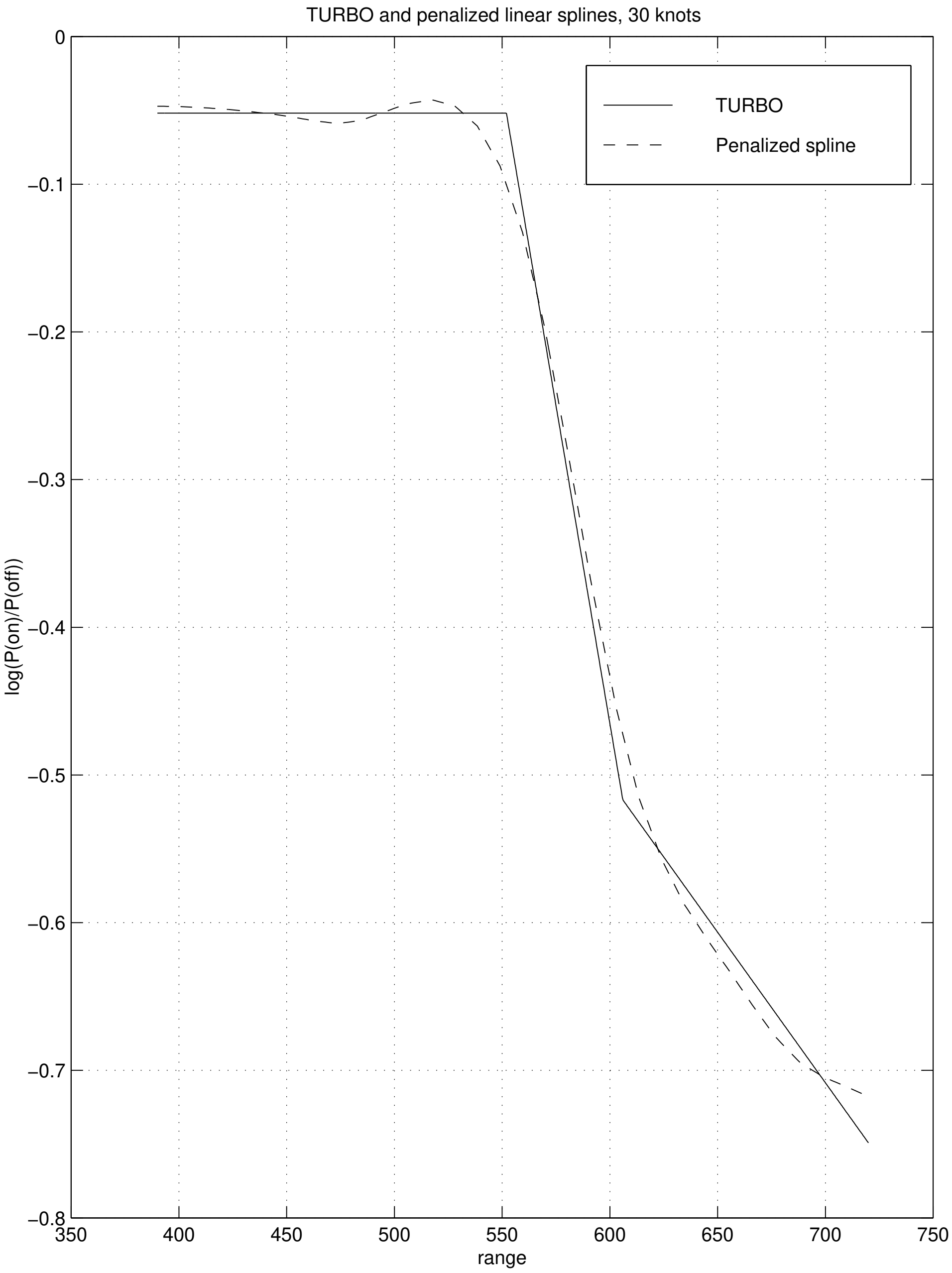


Figure 5: Lidar data: TURBO and penalized spline fits

Cross Validation (CV)

Goal: Chose α.

$m(x; \widehat{\boldsymbol{\beta}}(\alpha))$ is the fitted spline with penalty weight α.

$m_{-i}(x; \widehat{\boldsymbol{\beta}}(\alpha))$ is the fitted spline with penalty weight α and without using (X_i, Y_i).

$$\mathrm{CV}(\alpha) = \sum_{i=1}^{n} \left\{ Y_i - m_{-i}(X_i; \widehat{\boldsymbol{\beta}}(\alpha)) \right\}^2$$

Principle: Choose α that minimizes $\mathrm{CV}(\alpha)$.

Problem: CV is computationally intensive.

Solution: Use generalized cross validation (GCV), an approximation to CV. GCV would better be called ACV = approximation cross validation.

Generalized Cross Validation (GCV)

The Smoother or Hat Matrix

$\boldsymbol{H}(\alpha) := \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \alpha\boldsymbol{D})^{-1}\boldsymbol{X}^t$

$\Rightarrow$

$(m(X_i; \widehat{\boldsymbol{\beta}}(\alpha)), \ldots, m(X_n; \widehat{\boldsymbol{\beta}}(\alpha)))^t := \widehat{\boldsymbol{Y}} = \boldsymbol{H}(\alpha)\boldsymbol{Y}$

Degrees of freedom: $\mathrm{DF}(\alpha) = \mathrm{tr}(\boldsymbol{H}(\alpha))$.

GCV:

$$\mathrm{GCV}(\alpha) = \frac{\Sigma_{i=1}^{n}\left\{Y_i - m(X_i; \widehat{\boldsymbol{\beta}}(\alpha))\right\}^2}{(1 - \mathrm{DF}(\alpha)/n)^2}$$

So use $\widehat{\alpha}$ that minimizes $\mathrm{GCV}(\alpha)$.

Define $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\alpha})$.

Estimation of Derivatives

This is very simple using regression splines

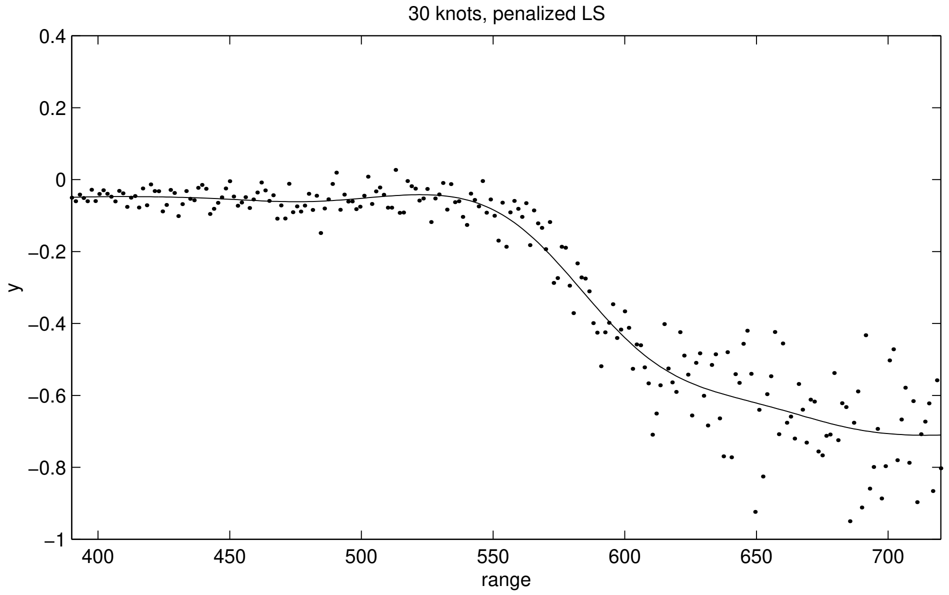
$$\frac{\partial}{\partial x}\, m(x;\boldsymbol{\beta}) = (0\ 1\ 2x\ \ \ldots p!x^{p-1}\ p!(x-t_1)_+^{p-1}\ \ \ldots\ \ p!(x-t_k)_+^{p-1})^t\boldsymbol{\beta}$$

Replace $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$

Application: Lidar

$p = 2 \ \Rightarrow$ the estimated derivative is piecewise linear.

α was chosen by GCV and may be a bit too small.



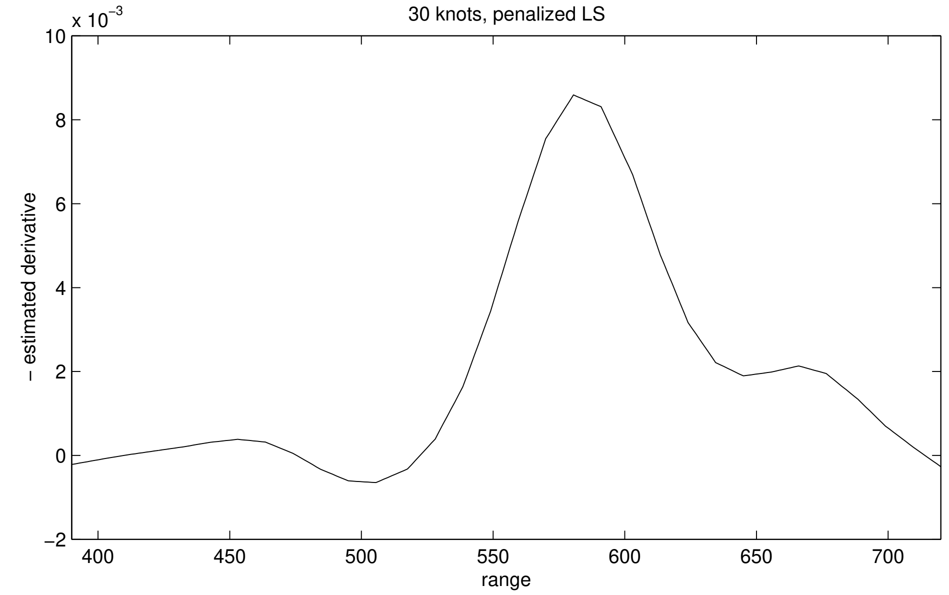


Figure 6: Lidar data: estimation of the derivative

A Bayesian Interpretation

Suppose that the errors in the model are iid $N(0, \sigma^2)$.

Put a prior on $\boldsymbol{\beta}$ such that the polynomial coefficients are $N(0, \sigma_1^2)$ and the coefficients of the truncated power functions are $N(0, \sigma_2^2)$.

Finally, let $\sigma_1 \to \infty$. This implies little prior knowledge of the polynomial coefficients.

Think of σ_2^2/σ^2 as a signal to noise ratio.

The penalized LS estimate with $\alpha = \sigma^2/\sigma_2^2$ is the Bayes estimate. This assumes that σ and σ_2 are known.

One can allow a more realistic model by putting priors on σ and σ_2. Then the Bayes estimates of σ and σ_2 automatically chose the penalty weight α. This is an area of ongoing research.

Reference: Lindley and Smith (1972)

Additive Models

Suppose that there are J predictors, $X^{(1)}, \ldots, X^{(J)}$. An observation is $(X_i^{(1)}, \ldots, X_i^{(J)}, Y_i)$.

An additive model is of the form

$$Y = \mu + m^{(1)}(X^{(1)}) + \cdots m^{(J)}(X^{(J)}) \ + \ \text{error.}$$

The usual estimation method is "backfitting" which iterates a one-at-a-time scheme for estimating the $m^{(j)}$s.

Regression spline estimates can be computed simultaneously and without iteration.

Principle: A linear model is equivalent to a vector space of functions of the predictor variable(s) and can be specified by giving a basis.

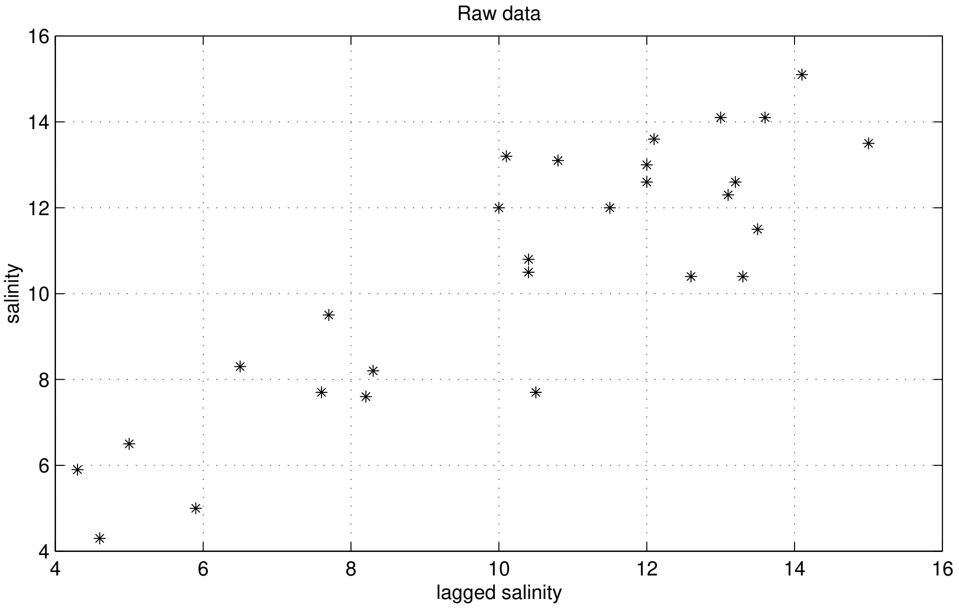
Let $\mathcal{B}^{(j)}$ be the regression spline basis for $X^{(j)}$, i.e.,

$$\mathcal{B}^{(j)} = \left\{ X^{(j)} \ \ldots \ (X^{(j)})^p \ (X^{(j)} - t_1)_+^p \ \ldots X^{(j)} - t_K)_+^p \right\}.$$

K and $t_1, \ldots, t_k$ will depend on j. Then the basis for a additive spline model is

$$\mathcal{B} = \{1\} \oplus \mathcal{B}^{(1)} \oplus \ \cdots \ \oplus \mathcal{B}^{(J)}$$

An additive spline model is a linear model with this basis and so it can be fit by OLS or by penalized LS.



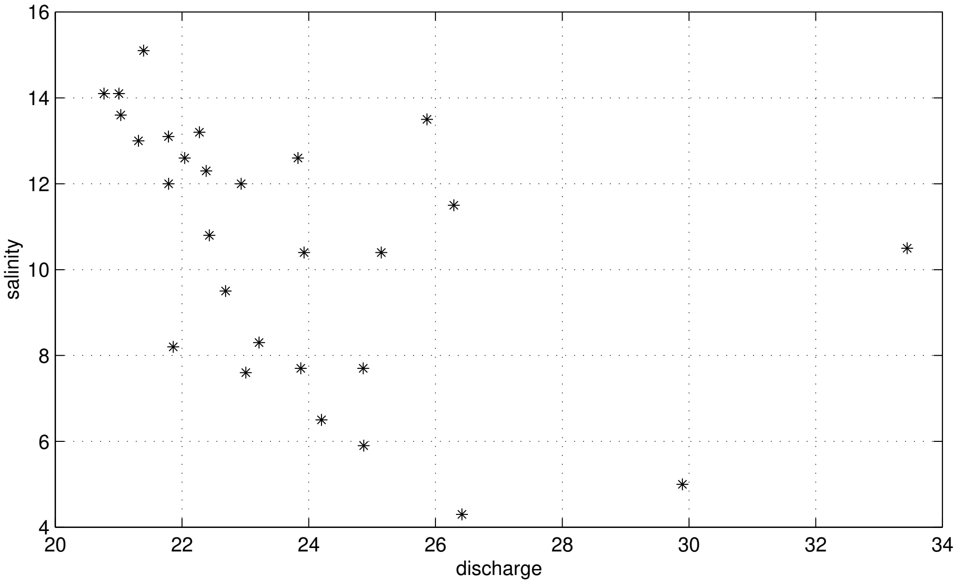


Figure 7: Salinity data: Raw data.

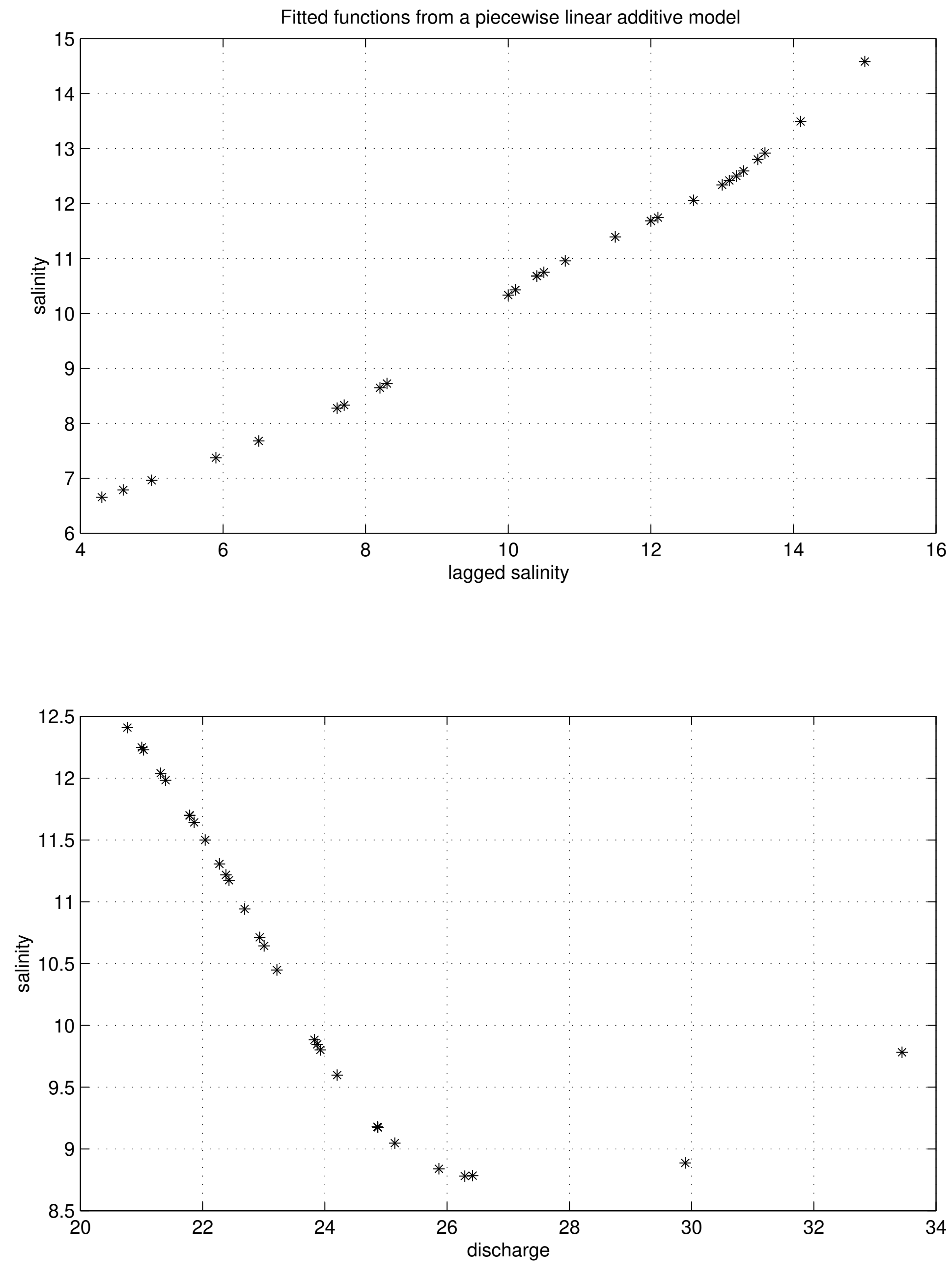


Figure 8: Salinity data: additive model.

Interactions

Suppose that we think that two predictor, say $X^{(1)}$ and $X^{(2)}$, may interact. Then we expand the additive model to

$$Y = \mu + m^{(1,2)}(X^{(1)}, X^{(2)}) + m^{(3)}(X^{(3)}) + \cdots m^{(J)}(X^{(J)}) \ + \ \text{error}. \tag{1}$$

Let $\mathcal{B}^{(1)} \otimes \mathcal{B}^{(2)} = \{b^{(1)}b^{(2)} : b^{(1)} \in \mathcal{B}^{(1)} \text{ and } b^{(2)} \in \mathcal{B}^{(2)}\}$.

$\mathcal{B}^{(1)} \otimes \mathcal{B}^{(2)}$ is called the tensor product of $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(2)}$.

Then the basis for model (1) is

$$\mathcal{B} = \mathcal{B}^{(1)} \oplus \cdots \oplus \mathcal{B}^{(J)} \oplus (\mathcal{B}^{(1)} \otimes \mathcal{B}^{(2)}).$$

Other interactions are defined analogously.

Interactions — application

Reference: Ruppert and Carroll (1996)

$Y = H_g$ concentration about an incinerator. $(X^{(1)}, X^{(2)})$ is spatial location. $X^{(3)}$ is distance from the incinerator (a function of $(X^{(1)}, X^{(2)})$).

Model 1: $\mathcal{B} = \mathcal{B}^{(1)} \oplus \mathcal{B}^{(2)} \oplus (\mathcal{B}^{(1)} \otimes \mathcal{B}^{(2)})$ (full bivariate model).

Model 2: $\mathcal{B} = \mathcal{B}^{(1)} \oplus \mathcal{B}^{(2)} \oplus \mathcal{B}^{(3)} \oplus (\mathcal{B}^{(1)} \otimes \mathcal{B}^{(2)})$.

$X^{(3)}$ is a function of $(X^{(1)}, X^{(2)})$, so the addition of $X^{(3)}$ to the model seems strange at first. The idea is that $X^{(3)}$ *alone* may explain most of the variation in H_g.

All basis functions in $\mathcal{B}^{(1)} \oplus \mathcal{B}^{(2)} \oplus (\mathcal{B}^{(1)} \otimes \mathcal{B}^{(2)})$ are penalized as are the truncated power functions in $\mathcal{B}^{(3)}$. This shrinks the fit towards the univariate model is $X^{(3)}$ alone. The result is a much smoother fit.

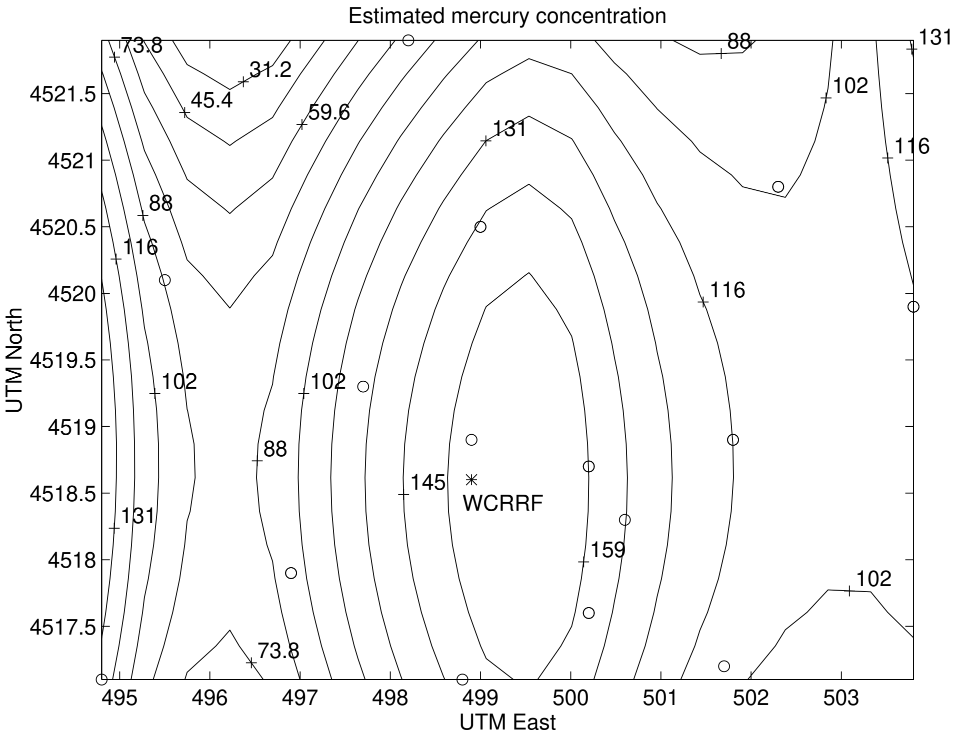


Figure 9: Mercury data: Fit from model 1.

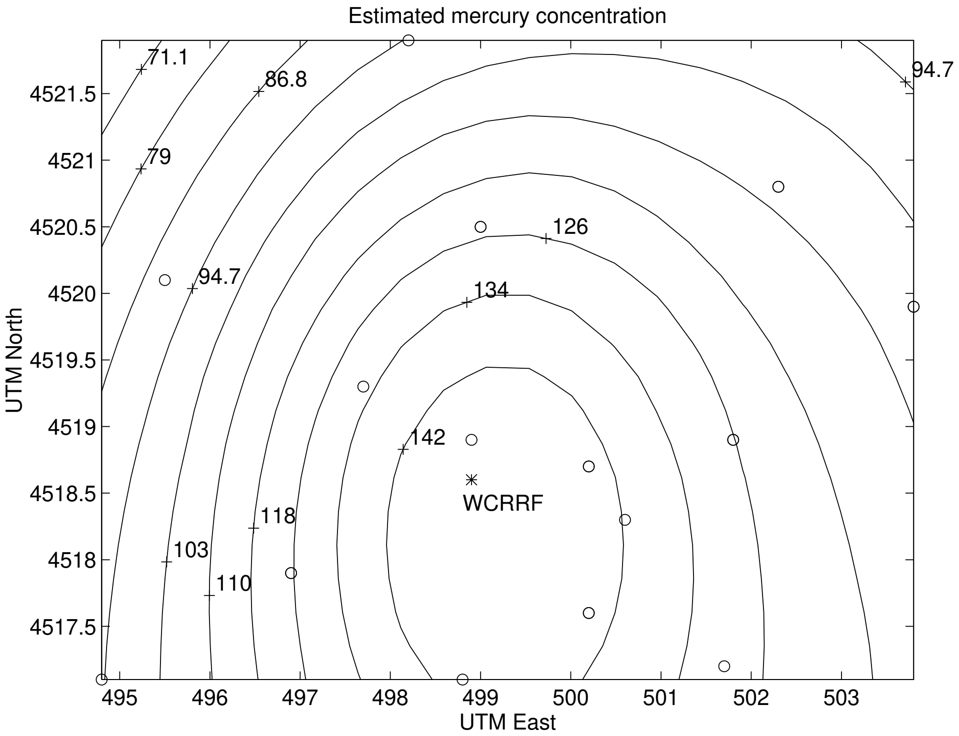


Figure 10: Mercury data: Fit from model 2.

Measurement Error

Reference: Maca, Carroll, and Ruppert (1996).

Suppose that instead of X we observe $W = X + U$ where U is $N(0, \sigma_U^2)$ measurement error.

Key idea:

$$
\begin{aligned}
E(Y_i|W_i) = \beta_0 + \beta_1 E(X_i|W_i) \ + \ \cdots + \beta_p E(X_i^p|W_i) \\
+ \sum_{j=1}^{K} \beta_{p+j} E((X_i - t_j)_+^p|W_i)
\end{aligned}
\qquad (2)
$$

If we knew the joint density of (X, W) then we could calculate $E(f(X_i)|W_i)$ for any function f. This would make (2) a model in *known* variables.

The joint density is the product of the marginal density of X and the conditional density of W given X. From the joint density, one can calculate the conditional density of X given W.

Assuming that σ_U is known, we know the density of W given X, and σ_U can be estimated from replicates.

We don't know the marginal density of X. Maca et al. (1996) adapt a method of Roeder and Wasserman (1997) to measurement error data to estimate the marginal density of X.

Thus, one can estimate the joint density of (X, W) and the conditional density of X given W.

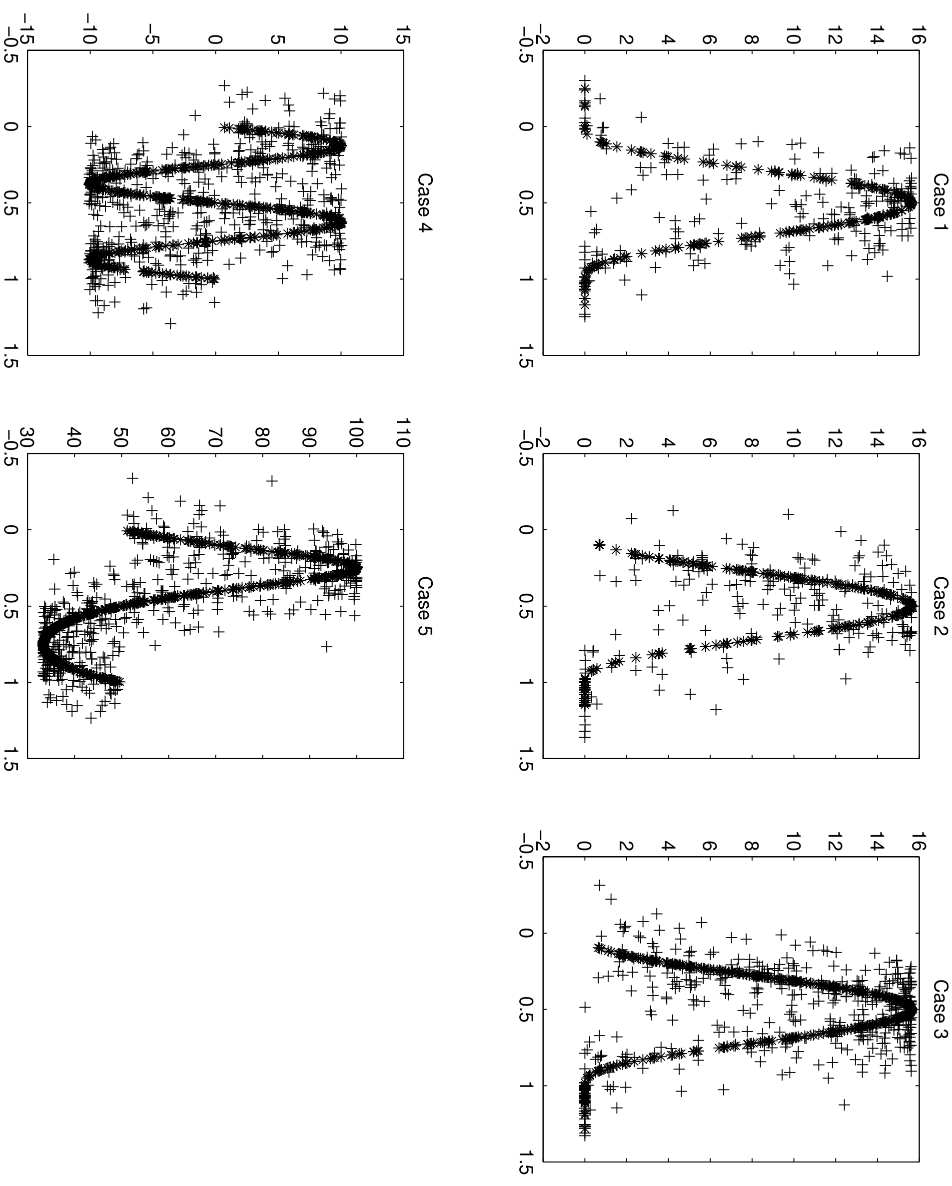


Figure 11: Cases for the measurement error simulation

	Naive Kernel	Naive Regression Spline	Naive Smoothing Spline	Simex Kernel	Simex Smoothing Spline	Simex Regression Spline	Structural Spline	Decon. Kernel
	Mean Squared Error							
Case 1	12.4	13.2	13.5	6.4	6.26	5.6	2.1	8.9
Case 2	14.4	14.6	15.1	8.1	7.3	6.6	5.5	10.9
Case 3	14.4	14.6	14.4	7.7	5.8	6.9	5.6	13.7
Case 4	32.8	31.2	31.8	18.8	15.5	14.5	7.6	20.3
Case 5	80.5	76.0	79.1	32.6	35.2	34.4	37.9	62.1

Table 1: Mean Squared Error results from the simulations.

Nonquadratic penalties

$$\mathrm{PSS}(\boldsymbol{\beta};\alpha,q) = \sum_{i=1}^{n}\Big\{Y_i - \boldsymbol{X}_i^t\boldsymbol{\beta}\Big\}^2 + \alpha\sum_{j=1}^{K}|\beta_{p+j}|^q$$

$\widehat{\beta}(\alpha,q)$ minimizes $\mathrm{PSS}(\beta;\alpha,q)$.

Here $\alpha > 0$ is a penalty weight, and $q > 0$ is another penalty parameter.

The choice of q can have very large effects in some situations, especially when m has discontinuities or discontinuous derivatives.

Current work with George Casella — data-driven selection of α and q based on a Bayesian model and a Gibbs sampler.

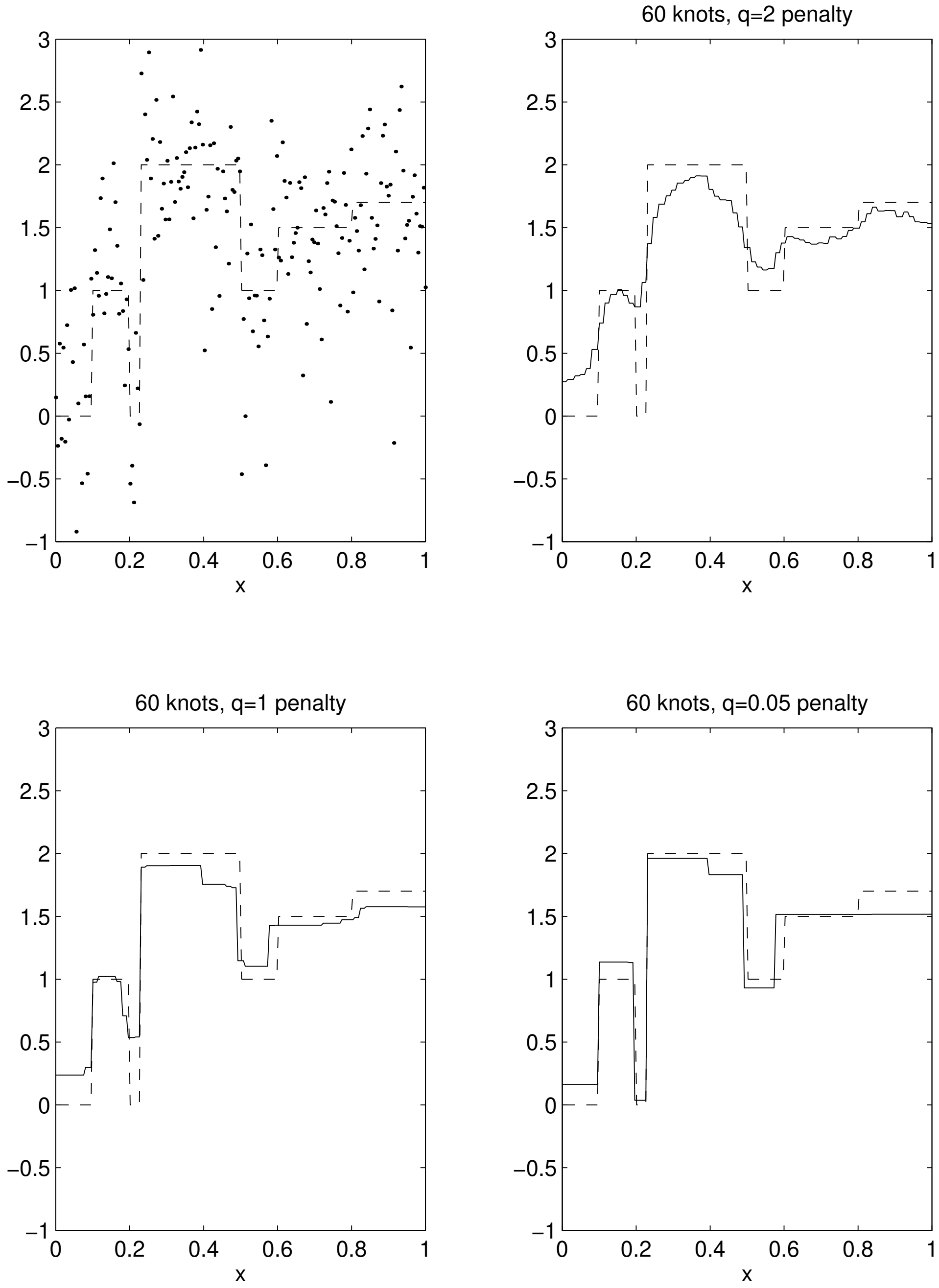


Figure 12: Jump function with different penalty norms

Summary

- Nonparametric estimators can either
  - fit simple parametric models locally, or
  - fit large parametric models with a complexity penalty
- Regression splines are linear models
- A penalty on the spline coefficients is easy to implement and the penalized estimator can be computed quickly
- The penalty weight can be chosen by GCV
- Additive models can be fit noniteratively
- Interactions can be added by using tensor products
- Regression splines are easy to fit when one only observes W which is X measured with error. However, one must estimate the conditional distribution of X given W

Summary of Notation

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \sum_{j=1}^{K} \beta_{p+j}(X_i - t_j)_+^p + \text{error}$$

Here $t_1 < \ldots < t_K$ are fixed "knots" and $u_+^p = u^p$ if $u > 0$ and equals 0 otherwise.

$$\boldsymbol{X}_i = \left\{1 \ X_i \ \cdots \ X_i^p \ (X_i - t_1)_+^p \ \cdots \ (X_i - t_K)_+^p\right\}$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1^t \\ \vdots \\ \boldsymbol{X}_n^t \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p+K} \end{pmatrix}$$

$$\text{PSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{Y_i - \boldsymbol{X}_i^t \boldsymbol{\beta}\right\}^2 + \alpha \sum_{j=1}^{K} \beta_{p+j}^2$$

$\boldsymbol{D} = \text{diag}(0, \ldots, 0, 1, \ldots, 1)$ ($p+1$ zeros and K ones)

$$\text{PSS}(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \alpha\|\boldsymbol{D}\boldsymbol{\beta}\|^2.$$

$$\widehat{\boldsymbol{\beta}}(\alpha) = (\boldsymbol{X}^t\boldsymbol{X} + \alpha\boldsymbol{D})^{-1}\boldsymbol{X}\boldsymbol{Y}.$$

$m(x; \widehat{\boldsymbol{\beta}}(\alpha))$ is the fitted spline with penalty weight α
$m_{-i}(x; \widehat{\boldsymbol{\beta}}(\alpha))$ is the fitted spline with penalty weight α and without using (X_i, Y_i)

$$\text{CV}(\alpha) = \sum_{i=1}^{n} \left\{Y_i - m_{-i}(X_i; \widehat{\boldsymbol{\beta}}(\alpha))\right\}^2$$

$\boldsymbol{H}(\alpha) := \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \alpha\boldsymbol{D})^{-1}\boldsymbol{X}^t$
$\Rightarrow$
$(m(X_i; \widehat{\boldsymbol{\beta}}(\alpha)), \ldots, m(X_n; \widehat{\boldsymbol{\beta}}(\alpha)))^t := \widehat{\boldsymbol{Y}} = \boldsymbol{H}(\alpha)\boldsymbol{Y}$
$\text{DF}(\alpha) = \text{tr}(\boldsymbol{H}(\alpha))$.

$$\text{GCV}(\alpha) = \frac{\sum_{i=1}^{n}\left\{Y_i - m(X_i; \widehat{\boldsymbol{\beta}}(\alpha))\right\}^2}{(1 - \text{DF}(\alpha)/n)^2}$$

References

Chen, V.C.P., Ruppert, D., and Shoemaker, C.A. (1997), Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming, *Operations Research*, to appear. [Application of MARS]

Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Sciences*, **11**, 89-121.

Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141. [Introduces the MARS algorithm.]

Friedman, J.H., and Stuetzle, W. (1981). Projection pursuit regression, *J. Amer. Statist. Assoc.*, **76**, 817–823.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*, Chapman and Hall: New York.

Lindley, D. V., and Smith, A. F. M. (1972), Bayes estimates for the linear model (with discussion), *J. Royal Statistical Society, Series B*, 34, 1–41.

Maca, J., Carroll, R.J., and Ruppert, D. (1996). Nonparametric kernel and regression spline estimation in the presence of measurement error, Manuscript. `http://www.orie.cornell.edu/~davidr/papers/index/index.html`

Roeder, K., and Wasserman, L. (1997). "Practical Bayesian density estimation using mixtures of normals," *J. Amer. Statist. Assoc.*, 92, 894–902.

Ruppert, D., and Carroll, D. (1997). Penalized Regression Splines. `http://www.orie.cornell.edu/~davidr/papers/index/index.html`

Stone, C.J., Hansen, M., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling, *Ann. Statist.*, 25, 1371–1470.

The overheads of this lecture are available at:
`http://www.orie.cornell.edu/~davidr/papers/index.html`