Regression

Recap: linear regression
Recap: ridge regression, LASSO
bias variance trade-off
nonlinear regression by basis expansion
wavelet regression

Joachim M. Buhmann

October 10, 2019

Recap: Statistical Learning Theory and Parametric Statistics alike for linear regression

- Optimal solution of regression problem: $\arg\min_f \mathbb{E}(Y - f(X))^2$ given by conditional mean

$$f^\star(x) = \mathbb{E}\left(Y|X = x\right).$$

- Yet, $\mathbf{P}(Y|X)$ and $\mathbf{P}(X)$ unknown.
- (Parametric) maximum likelihood: Assume $Y|X \sim \mathcal{N}\left(f(X), \sigma^2\mathbf{I}\right)$.
  Solve: $\arg\max_f \sum_{i=1}^n \log \mathbf{P}(Y = y_i|X = x_i, \sigma^2)$.
- Statistical learning theory: Minimize directly empirical risk $\arg\min_f \sum_{i=1}^n (y_i - f(x_i))^2$.

$\Rightarrow$ Both approaches lead to same solution.

Recall: Linear Regression Models and Least Squares

Statistical model

Given a vector of inputs $X^{\mathsf{T}} = (X_1, \ldots, X_d)$. The output variable (also called response variable) is predicted via the model

$$Y = \beta_0 + \sum_{j=1}^{d} X_j \beta_j, \quad Y \in \mathbb{R}$$

β_0 is called bias (Machine Learning) or intercept (Statistics)

Homogeneous coordinates

Introduce a constant coordinate $X_0 = 1$. Then

$$Y = X^{\mathsf{T}} \beta, \quad X, \beta \in \mathbb{R}^{d+1}$$

Residual Sum of Squares (RSS)

Fitting data to models

For given data $\{(x_i, y_i) | i = 1, \ldots, n\} \subset \mathbb{R}^{d+1} \times \mathbb{R}$, minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - x_i^{\mathsf{T}} \beta)^2$$

Matrix notation

We have $RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\beta)$, where $\mathbf{X}$ is an $n \times (d+1)$ matrix whose rows are the input vectors $x_i \in \mathbb{R}^{d+1}$ in the training set; $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the vector of outputs in the training set.

Minimum condition

Setting the derivative $\nabla_\beta RSS(\beta) \stackrel{!}{=} 0$ leads to $\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\beta) \stackrel{!}{=} 0$

Solution for nonsingular $\mathbf{X}^{\mathsf{T}}\mathbf{X}$: $\boxed{\hat{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}}$

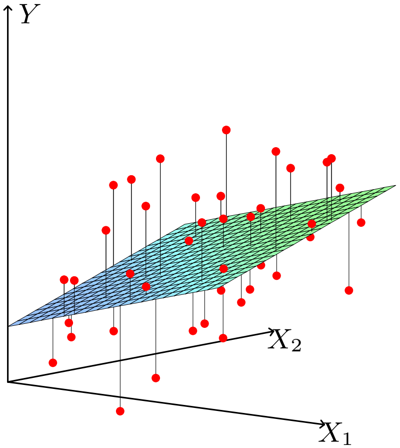


Figure 3.1: *Linear least squares fitting with* $X \in \mathbb{R}^2$. *We seek the linear function of* X *that minimizes the sum of squared residuals from* Y.

Prediction $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$

The matrix $\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}$ is sometimes called the hat matrix which is an orthogonal projection on the space spanned by the columns of $\mathbf{X}$.

Statistical assumptions: Assume that for given $(X_1, \ldots, X_d)$ we have

$$
\begin{aligned}
Y &= \mathbb{E}(Y|X_1, \ldots, X_d) + \epsilon \\
&= \beta_0 + \sum_{j=1}^{d} X_j \beta_j + \epsilon \;=\; X\beta + \epsilon
\end{aligned}
$$

with additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with conditional mean $\mathbb{E}(Y|X_1, \ldots, X_d) = X^T\beta$ depending linearly on $(1, X_1, \ldots, X_d)$.

Distribution of the estimator $\hat{\beta}$:

$$
\hat{\beta} \sim \mathcal{N}\big(\beta, (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\sigma^2\big)
$$

Optimality of Least Squares Estimate

The least squares estimate of the parameter β has the smallest variance among all linear unbiased estimates.

- Consider the problem of estimating a linear combination $\theta = a^\mathsf{T}\beta$ of the entries of β, e.g. $f(x_{n+1}) = x_{n+1}^\mathsf{T}\beta$ at a new location $x_{n+1} \in \mathbb{R}^{d+1}$.
- The estimate of θ obtained from the least squares estimate $\hat{\beta}$ is

$$\hat{\theta} = a^\mathsf{T}\hat{\beta} = a^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

- For fixed $\mathbf{X}$, the estimate $\hat{\theta}$ is a *linear* function of the response vector $\mathbf{y}$.
- $\hat{\theta} = a^\mathsf{T}\hat{\beta}$ is also *unbiased*:

$$\begin{aligned}
\mathbb{E}(a^\mathsf{T}\hat{\beta}) &= \mathbb{E}(a^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}) \\
&= a^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbb{E}(\mathbf{X}\beta + \epsilon) \\
&= a^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\beta + \underbrace{\mathbb{E}(\epsilon)}_{=0}) \;=\; a^\mathsf{T}\beta
\end{aligned}$$

Variance of $a^{\mathsf{T}}\hat{\beta}$

$$
\begin{aligned}
\mathbb{V}(a^{\mathsf{T}}\hat{\beta}) &= \mathbb{V}\left(a^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\beta+\epsilon)\right) \\
&= \mathbb{V}\left(a^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\epsilon\right) \\
&= \mathbb{E}\left(a^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\epsilon\epsilon^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}a\right) \\
&= \sigma^2 a^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}a
\end{aligned}
$$

Alternative unbiased linear estimator $\tilde{\theta} = c^{\mathsf{T}}\mathbf{y} = a^{\mathsf{T}}\hat{\beta} + a^{\mathsf{T}}\mathbf{D}\mathbf{y}$

$$
\begin{aligned}
\mathbb{E}(c^{\mathsf{T}}\mathbf{y}) &= \mathbb{E}(a^{\mathsf{T}}\hat{\beta}) + \mathbb{E}(a^{\mathsf{T}}\mathbf{D}\mathbf{y}) = a^{\mathsf{T}}\beta + \mathbb{E}\left(a^{\mathsf{T}}\mathbf{D}(\mathbf{X}\beta+\epsilon)\right) \\
&= a^{\mathsf{T}}\beta + a^{\mathsf{T}}\mathbf{D}\mathbf{X}\beta + a^{\mathsf{T}}\mathbf{D}\underbrace{\mathbb{E}(\epsilon)}_{=0} = a^{\mathsf{T}}\beta
\end{aligned}
$$

The unbiasedness condition $\mathbb{E}(c^{\mathsf{T}}\mathbf{y}) = a^{\mathsf{T}}\beta$ implies $a^{\mathsf{T}}\mathbf{D}\mathbf{X} = 0$.

Gauss Markov Theorem

Theorem (Gauss-Markov Theorem)

For any linear estimator $\tilde{\theta} = c^T\mathbf{y}$ that is unbiased for $a^T\beta$, we have

$$\mathbb{V}(a^T\hat{\beta}) \le \mathbb{V}(c^T\mathbf{y}).$$

Proof.

Let $c^{\mathsf{T}}\mathbf{y} = a^{\mathsf{T}}\hat{\beta} + a^{\mathsf{T}}\mathbf{D}\mathbf{y} = a^{\mathsf{T}}\big((\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}} + \mathbf{D}\big)\mathbf{y}$ be an unbiased estimator of $a^{\mathsf{T}}\beta$; then it follows that $a^{\mathsf{T}}\mathbf{D}\mathbf{X}\beta = 0$ which implies $a^{\mathsf{T}}\mathbf{D}\mathbf{X} = 0$ (see previous slide).

$$\begin{aligned}
\mathbb{V}(c^{\mathsf{T}}\mathbf{y}) &= \mathbb{E}[(c^{\mathsf{T}}\mathbf{y})^2] - (\mathbb{E}c^{\mathsf{T}}\mathbf{y})^2 = c^{\mathsf{T}}(\mathbb{E}\mathbf{y}\mathbf{y}^{\mathsf{T}} - \mathbb{E}\mathbf{y}\mathbb{E}\mathbf{y}^{\mathsf{T}})c = \sigma^2 c^{\mathsf{T}}c \\
&= \sigma^2 \left(a^{\mathsf{T}}((\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}} + \mathbf{D})(\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} + \mathbf{D}^{\mathsf{T}})a\right) \\
&= \sigma^2 \left(a^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}a + a^{\mathsf{T}}\mathbf{D}\mathbf{D}^{\mathsf{T}}a\right) \\
&= \mathbb{V}(a^{\mathsf{T}}\hat{\beta}) + \sigma^2 a^{\mathsf{T}}\underbrace{\mathbf{D}\mathbf{D}^{\mathsf{T}}}_{\succeq 0}a \;\ge\; \mathbb{V}(a^{\mathsf{T}}\hat{\beta})
\end{aligned}$$

Note that $\mathbf{D}\mathbf{D}^{\mathsf{T}}$ is positive semidefinite. The third "=" holds because of $a^{\mathsf{T}}\mathbf{D}\mathbf{X} = 0$. □

Is this the best we can do?

- $\hat{f}(x) = x^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$ has smallest variance of all unbiased linear models.
- Over-fitting can be a problem.
- Bias-variance trade-off:
  mean squared error $=$ bias^2 $+$ variance $+$ noise variance

Hence, $\hat{f}(x)$ is best among all unbiased linear models in the sense of minimizing the MSE.

- Goal: minimize generalization error.
- Option: Trade bias increase for variance reduction.
- Goal: Minimize bias and variance simultaneously.

Recall: Bias/Variance Dilemma - Regression I

Regression Setting: Easier to understand than classification

Data: $D = \{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d,\ y_i \in \mathbb{R}$

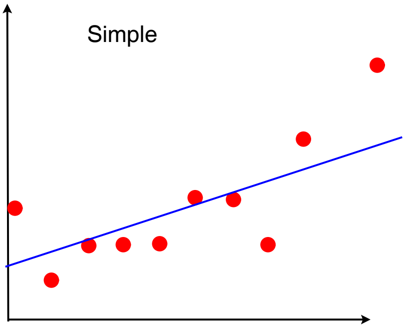
Source: (X_i, Y_i) i.i.d. $P(X, Y)$

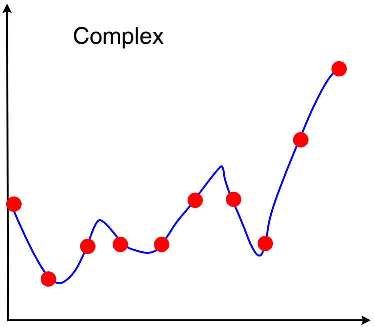
Objective: Find regression function $f \in \mathcal{C}$ such that $\mathbb{E}(Y - f(X))^2$ is minimal. $\mathcal{C}$ is the hypothesis class.

Optimum: $f^*(x) = \mathbb{E}(Y|X = x)$

Estimator: $\hat{f}(X)$ (depends on r.v. X and on data D)

Two problems: We only have a finite training set!
Complexity of hypothesis class $\mathcal{C}$ is unknown!





Trade-off

Complex $\mathcal{C}$ overfitting

Simple $\mathcal{C}$ underfitting

Objective: Find 'best balance' between the two
Split error into Bias + Variance

Bias/Variance - Regression II

Identify error components

We decompose the expected prediction error at $X = x$:

$$\begin{aligned}
&\mathbb{E}_D \, \mathbb{E}_{Y|X=x} \left(\hat{f}(x) - Y\right)^2 \\
&\quad = \mathbb{E}_D \left(\hat{f}(x) - \mathbb{E}(Y|X=x)\right)^2 + \mathbb{E}\left(Y - \mathbb{E}(Y|X=x)\right)^2 \\
&\quad = \mathbb{E}_D \left(\hat{f}(x) - \mathbb{E}_D \hat{f}(x)\right)^2 && \text{(variance)} \\
&\qquad + \left(\mathbb{E}_D \hat{f}(x) - \mathbb{E}(Y|X=x)\right)^2 && \text{(bias)}^2 \\
&\qquad + \mathbb{E}\left(Y - \mathbb{E}(Y|X=x)\right)^2 && \text{(noise)}
\end{aligned}$$

The mixed quadratic terms vanish due to the averages $\mathbb{E} \equiv \mathbb{E}_{Y|X=x}$ and $\mathbb{E}_D$.

Unbiased estimator: bias $= \mathbb{E}_D \hat{f}(x) - \mathbb{E}(Y|X=x) = 0$.

Bias/variance Tradeoff

Objective: Minimize bias and variance simultaneously - usually impossible

Tradeoff: Small data sets and large $\mathcal{C}$

variance large, bias small

Large data sets and small $\mathcal{C}$

variance small, bias large

The optimal tradeoff between bias and variance is achieved when we avoid both underfitting (large bias) and overfitting (large variance).

Outlook: Ensemble methods seem to avoid the bias/variance tradeoff since they lower variance while keeping the bias fixed. Note: The Rao-Cramer inequality defines a lower bound for variance reduction by ensemble averaging (no free lunch).

Several solutions to avoid overfitting

- Regularization: Add model complexity term to cost function:

$$\arg\min_{\theta} \sum_{i=1}^{n} l\left(f(x_i, \theta), y_i\right) + R(\theta)$$

  Adding regularization is often equivalent to choosing a prior in a Bayesian framework and using a MAP estimator:
  - $\arg\min_\theta \sum_{i=1}^n l\left(f(x_i,\theta), y_i\right) + R(\theta) = \arg\max_\theta \Pi_{i=1}^n \mathbf{P}(y_i|x_i,\theta)\mathbf{P}(\theta)$
  - $R(\theta) = -\log \mathbf{P}(\theta)$, and $l\left(f(x_i,\theta), y_i\right) = -\log \mathbf{P}(y_i|x_i,\theta)$
- Model selection based on generalization error estimate (e.g. by cross-validation).
- Ensembles of classifiers (see later).

Regularization $=$ Bayesian Maximum A Posteriori (MAP) estimates

Ridge Regression

Cost function: $RSS(\beta; \lambda) = (\mathbf{y} - \mathbf{X}\beta)^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^{\mathsf{T}}\beta$.

Bayesian view: $Y|(X, \beta) \sim \mathcal{N}\left(x^{\mathsf{T}}\beta, \sigma^2\mathbf{I}\right)$, prior on β: $\beta \sim \mathcal{N}(0, \sigma^2/\lambda\mathbb{I})$.

Solution: $\boxed{\hat{\beta}^{\text{ridge}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}}$

Tikhonov regularization $R(\beta) = \lambda\beta^{\mathsf{T}}\beta$ is also called weight decay in Neural Networks Literature.

LASSO

Cost function: $RSS(\beta; \lambda) = (\mathbf{y} - \mathbf{X}\beta)^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda\|\beta\|_1$.

Bayesian view: $Y|(X, \beta) \sim \mathcal{N}\left(x^{\mathsf{T}}\beta, \sigma^2\mathbf{I}\right)$, prior on β_i: Laplace: $p(\beta_i) = \frac{\lambda}{4\sigma^2}\exp(-|\beta|\frac{\lambda}{2\sigma^2})$.

Solution: By efficient optimization techniques (e.g. LARS). Note: $\|\beta\|_1 = \sum_{j=0}^{d}|\beta_j|$ is not differentiable.

For model selection, the complexity parameters λ or s are chosen by estimates of the generalization error, e.g. cross-validation.

Singular Value Decomposition (SVD)

Centered input matrix

Use centered $\mathbf{X}$ and perform a SVD. (Centering means that the center of mass of all points coincides with the origin.)

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$$

Here $\mathbf{U} \in \mathbb{R}^{n \times d}$ and $\mathbf{V} \in \mathbb{R}^{d \times d}$ are matrices with and orthonormal columns. $\mathbf{D}$ is a $d \times d$ diagonal matrix with entries $d_1 \geq d_2 \geq \cdots \geq d_d > 0$, the singular values of $\mathbf{X}$. (Here we assume that $rk(\mathbf{X}) = d$.)

SVD of least squares fitted vector

$$\mathbf{X}\hat{\beta}^{\mathsf{ls}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}(\mathbf{V}\mathbf{D}^2\mathbf{V}^\mathsf{T})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} = \mathbf{U}\mathbf{U}^\mathsf{T}\mathbf{y}$$

SVD of ridge regression solution

Suppression of contributions by small eigenvalues

$$
\begin{aligned}
\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} \;=\; \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda \mathbf{I})^{-1}\mathbf{D}\mathbf{U}^{\mathsf{T}}\mathbf{y} \\
&= \sum_{j=1}^{d} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j^{\mathsf{T}}\mathbf{y}
\end{aligned}
$$

Note that the shrinkage factor $\frac{d_j^2}{d_j^2+\lambda}$ is small for small singular values d_j and it approaches 1 for large singular values.

Built-in model selection.

The LASSO

Equivalent formulation: Least Absolute Shrinkage and Selection Operator

Cost function:

$$
\begin{aligned}
\hat{\beta}^{\text{LASSO}} &= \arg\min_{\beta} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{d} x_{i,j}\beta_j\Big)^2 \\
\text{subject to} &\quad \sum_{j=1}^{d} |\beta_j| \le s.
\end{aligned}
$$

Sparseness: LASSO estimates are known to be sparse with few coefficients non-vanishing.
Reason: the LSE error surface hits often the corners of the constraint surface (see fig 3.12 of Hastie et al.).

Ridge vs. LASSO Estimation

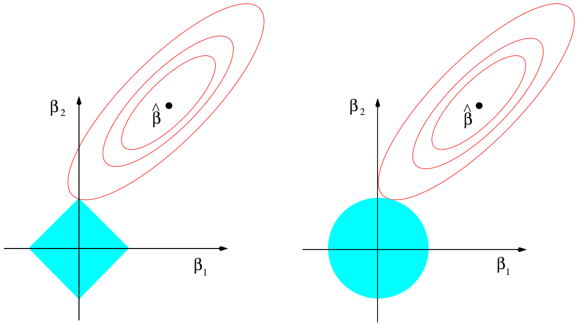


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Results of Different Regression Methods

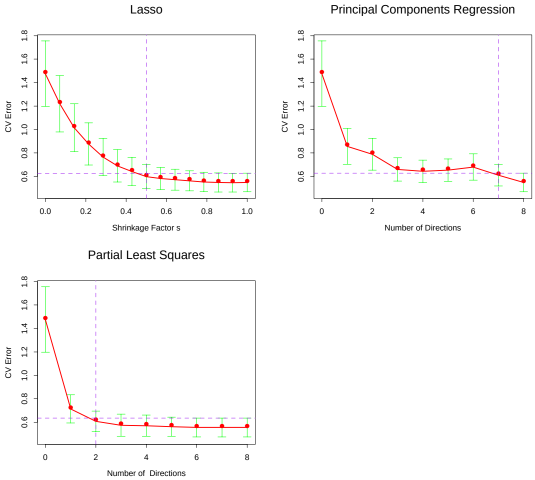


Fig. 3.6: Estimated prediction error curves and their standard errors for three selection and shrinkage methods, found by 10-fold cross-validation. (HTF'01)

Coefficient Weights and Interpretability

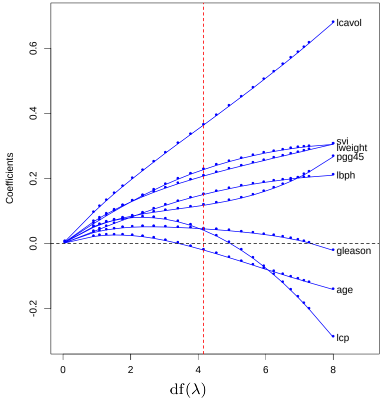


Figure 3.7: *Profiles of ridge coefficients for the prostate cancer example, as tuning parameter λ is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$, *the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 4.16$, *the value chosen by cross-validation.*

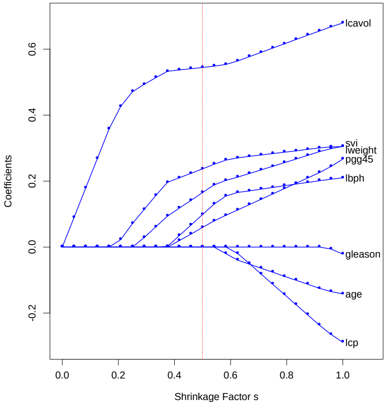


Figure 3.9: *Profiles of lasso coefficients, as tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.5$, the value chosen by cross-validation. Compare Figure 3.7 on page 7; the lasso profiles hit zero, while those for ridge do not.*

Some Remarks on Shrinkage Methods

Generalized Ridge Regression

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{d} x_{i,j}\beta_j\Big)^2 + \lambda \sum_{j=1}^{d} |\beta_j|^q \right\}.$$

This cost function models the shrinkage of the coefficients!

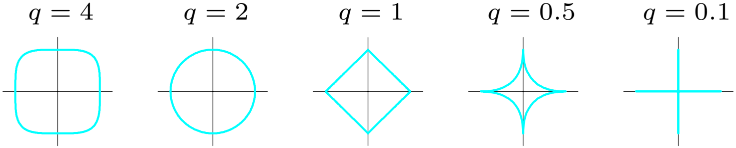


Fig. 3.13: Level set of $\sum_j |\beta_j|^q$ for given values of q. (HTF'01)

Idea behind shrinkage: When white noise is added to the data then all Fourier coefficients are increased by a constant on average. $\Rightarrow$ Shrink all coefficients by the estimated noise amount to derive a robust predictor.

Nonlinear Regression: Basis Expansion

Idea: Transform the variables X nonlinearly and fit a linear model in the resulting (feature) space.
Transformation: $h_m(X) : \mathbb{R}^d \mapsto \mathbb{R},\ 1 \leq m \leq M$

Model of response variable

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

f is linear in β but nonlinear in X!

Cubic splines are a common choice for h, e.g., for $d = 1$ with 2 knots at ξ_1, ξ_2

$$\begin{array}{lll} h_1(X) = 1, & h_3(X) = X^2, & h_5(X) = (X - \xi_1)_+^3 \\ h_2(X) = X, & h_4(X) = X^3, & h_6(X) = (X - \xi_2)_+^3 \end{array}$$

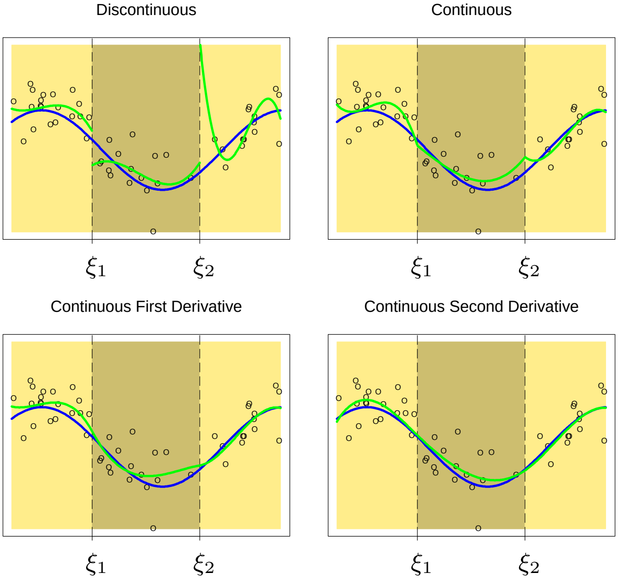


Figure 5.2: A series of piecewise-cubic polynomials. (HTF'01)

Smoothing Splines

Knot selection

Use the maximal number of knots and control the smoothness by regularization!

$$RSS(f,\lambda) = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \int \big(f''(x)\big)^2 dx$$

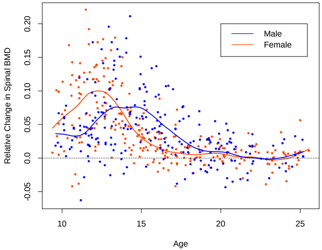


Fig. 5.6: Change in bone mineral density vs. age ($\lambda \approx 0.00022$). (HTF'01)

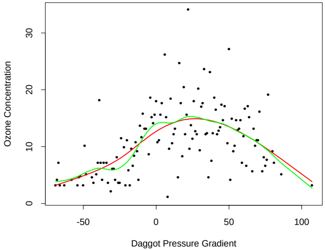


Fig. 5.7: Smoothing spline with 5, 11 degrees of freedom. (HTF'01)

Regression with Wavelets

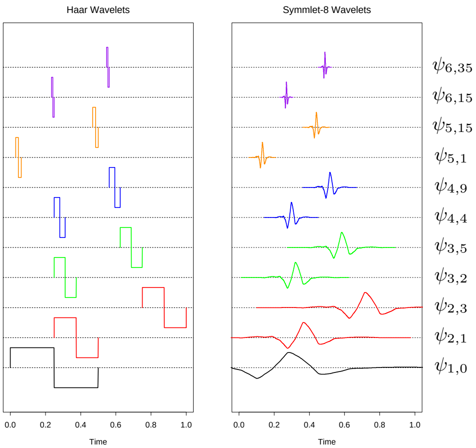


Fig. 5.13: Haar and symmlet-8 wavelets for different translations and dilations. (HTF'01)

NMR Denoising by Wavelet Shrinkage

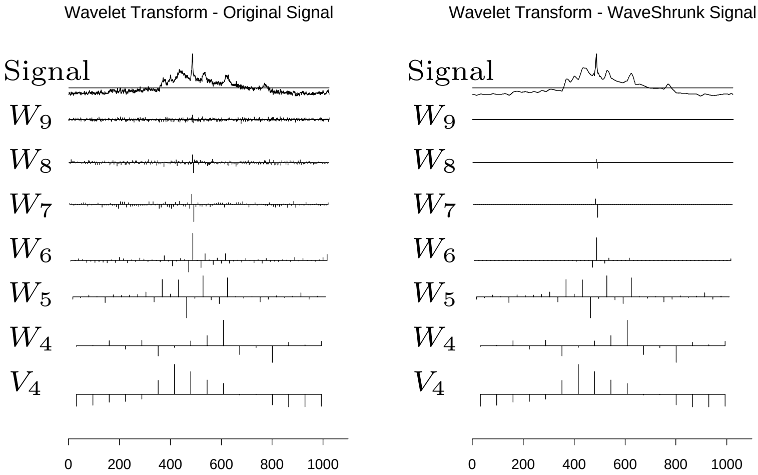


Fig. 5.14 (bottom): Wavelet transform of original signal (left) and wavelet coefficients after shrinkage (right). (HTF'01)

Denoised NMR Signal

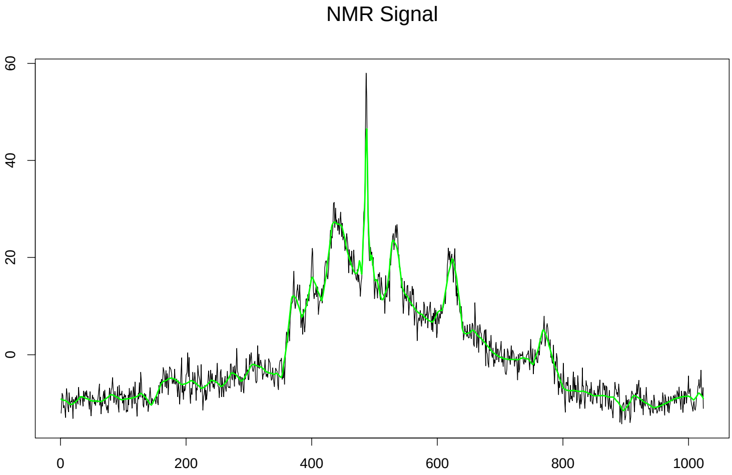


Fig. 5.14 (top): NMR signal and a wavelet-shrunk version (green). (HTF'01)

Wavelet and Spline Comparison

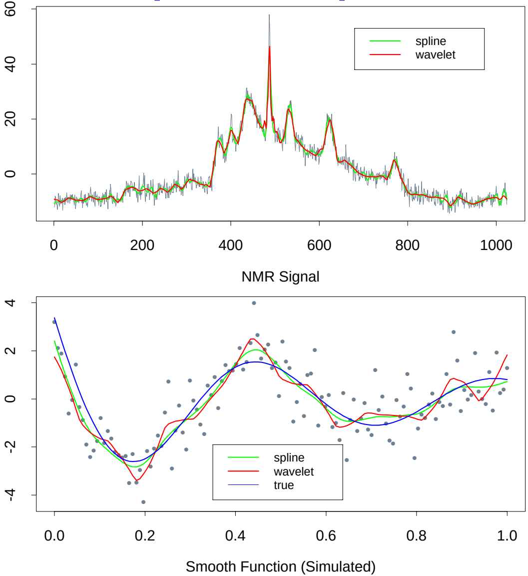


Fig. 5.16: Wavelet smoothing compared with smoothing splines on two examples. Each panel compares the SURE-shrunk wavelet fit to the crossvalidated smoothing spline fit. (HTF'01)