A Primer in Econometric Theory

Lecture 13: Regularization

John Stachurski
Lectures by Akshay Shanker

March 26, 2017

Nonparametric Density Estimation

Nonparametric density estimation an application of regularization to the problem of recovering distributions from data

- combine the data with a prior belief that probability mass most likely falls in places other than just the sample points observed so far

We review parametric density estimation and then proceed to nonparametric methods

Parametric Estimation

Suppose data consist of IID observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from unknown distribution P on $\mathbb{R}^d$

- assume P is absolutely continuous
- aim is to estimate the density of P, denoted below by f

In a parametric setting, for example:

- assume f belongs to the class of normal densities on $\mathbb{R}$, so that $f = f(\cdot\,; \mu, \sigma) =$ the normal density for distribution $\mathrm{N}(\mu, \sigma^2)$
- MLEs of the parameters are $\hat{\mu}_N := \bar{x}_N$ and $\hat{\sigma}_N := s_N$
- Plug MLEs into f gives density estimate $f(\cdot\,; \bar{x}_N, s_N)$

Since $\bar{x}_N$ and s_N are consistent, the random density $f(\cdot\,; \bar{x}_N, s_N)$ will be close to $f(\cdot\,; \mu, \sigma)$ with high probability for large N

Now we extend notion of consistency from vectors to densities

For any $\mathscr{B}$-measurable function $f\colon \mathbb{R}^d \to \mathbb{R}$ and $p \geq 1$, set

$$\|f\|_p := \left\{ \int |f|^p \right\}^{1/p} := \left\{ \int |f(\mathbf{s})|^p \, \mathrm{d}\mathbf{s} \right\}^{1/p} \tag{1}$$

Integration is over all of $\mathbb{R}^d$

If above expression is finite, then we write $f \in L_p$

For densities f and g on $\mathbb{R}^d$, define the L_p distance

$$d_p(f,g) := \|f - g\|_p \tag{2}$$

The norm $\|\cdot\|_p$ satisfies most of the properties of the norms we've met so far

E.g., the triangle inequality

$$\|f - g\|_p \leq \|f - h\|_p + \|h - g\|_p$$

for all $p \geq 1$ and $f, g, h \in L_p$

Specializing $p = 1$ gives the L_1 distance, which sums over absolute deviation

- L_1 distance is arguably a better choice for studying deviation between densities
- L_1 distance between densities is always well-defined

Specializing to $p = 2$ gives the popular L_2 distance — a variation of the L_2 distance we used in §5.2.2

Fact. (14.1.1) **Scheffés lemma** If $\{f_n\}$ and f are densities on $\mathbb{R}^d$, then

$$f_n(\mathbf{s}) \to f(\mathbf{s}) \text{ for all } \mathbf{s} \text{ in } \mathbb{R}^d \implies \|f_n - f\|_1 \to 0$$

Fact. (14.1.2) For any densities f, g and h on $\mathbb{R}^d$ we have

1. $\|f - g\|_1 \leq \sqrt{2D(f,g)}$, where $D(f,g)$ is the KL deviation defined in (8.34), and
2. $\|f - g\|_1 = 2 \sup_{B \in \mathscr{B}(\mathbb{R}^d)} \left| \int_B f - \int_B g \right|$.

The bound in 1. is called **Pinsker's inequality**, while 2. is called **Scheffé's identity**

We say a sequence $\{\hat{f}_N\}$ of random densities on $\mathbb{R}^d$ is L_p-**consistent** for a density f on $\mathbb{R}^d$ if

$$\|\hat{f}_N - f\|_p \xrightarrow{p} 0 \quad \text{as} \quad N \to \infty$$

Example. Let $\hat{f}_N = f(\cdot\,; \bar{x}_N, s_N)$ be the Nth element of the sequence of normal densities described above

- $x_1, \ldots, x_N$ are independent draws from a normal density $f = f(\cdot\,; \mu, \sigma)$
- $\bar{x}_N$ and s_N the sample mean and standard deviation respectively

This sequence of densities is L_1-consistent for f (exercise 14.4.3)

Failure of Consistency

Parametric class may not contain the density generating the data or any good approximation — parametric approach is not consistent

If we estimate f with parametric class $\{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$, then the L_p deviation between our estimate and f is bounded below by

$$\delta(f) := \inf_{\boldsymbol{\theta}\in\Theta} \|f - f_{\boldsymbol{\theta}}\|_p \tag{3}$$

Example. Assume in the example above that the true density f not Gaussian

Either:

- the sequence $\hat{f}_N$ is not L_1-consistent for any density
- or, the sequence is L_1-consistent for some density, but that density is not f

The reason is that $\delta(f)$ in (3) is always positive when the parametric class is Gaussian and f is not

- the set of normal densities is closed under the taking of limits in L_1

Kernel Density Estimation

Sometimes we can make good choices for parametric classes:

- using descriptive statistics
- appealing to some theory with sharp quantitative implications

When the above is difficult, best to use a nonparametric approach that is consistent under weaker assumptions

Suppose we have IID data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ generated from unknown density f on $\mathbb{R}^d$

To estimate f using the data, employ a **kernel density estimator** (KDE), which takes the form

$$\hat{f}_N(\mathbf{s}) := \frac{1}{Nh^d} \sum_{n=1}^{N} K\left(\frac{\mathbf{s} - \mathbf{x}_n}{h}\right) \tag{4}$$

Here K is the **kernel function** of the estimator, and h is the **bandwidth**

The kernel function K required to be a density on $\mathbb{R}^d$

The bandwidth h is any positive number

The function $\hat{f}_N$ is always a density (Exercise 14.4.4)

Consider a simple instance created from just three data points x_1, x_2, x_3 on $\mathbb{R}$

For K we take the standard normal density

Since $N = 3$, the function $\hat{f}_N$ is the sum of three individual functions

The nth function:

$$g_n(s) = \frac{1}{Nh} K\left(\frac{s - x_n}{h}\right) = \frac{1}{Nh} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(s - x_n)^2}{2h^2}\right\}$$

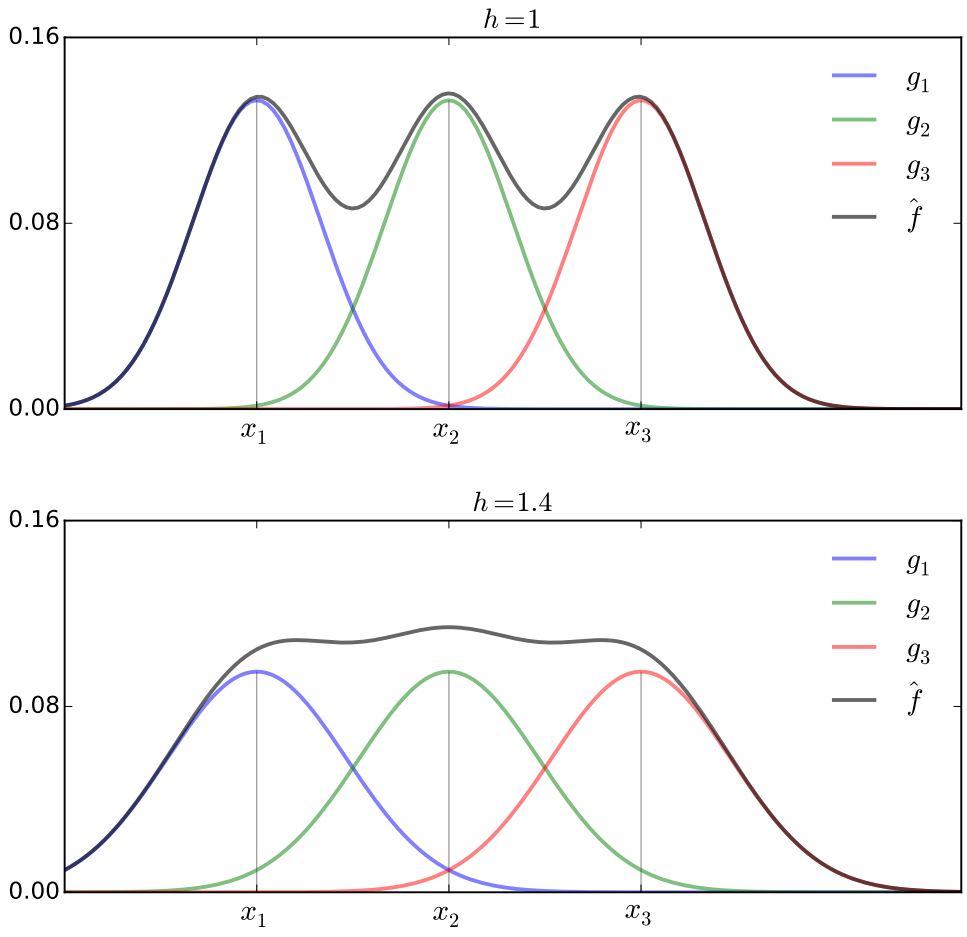

Figure: Nonparametric KDE, different bandwidths

Bandwidth adds smoothing to empirical distribution

However, trade-off associated with smoothing

- as the bandwidth goes to zero, kernel density becomes similar to the empirical distribution — overfitting
- at the same time, excessive smoothing adds too much bias, hiding features of the true distribution

Optimal bandwidth in terms of minimizing L_p deviation depends on the unknown density f.

Two approaches:

- make assumptions on f and choose the bandwidth accordingly
- cross-validation (we review below)

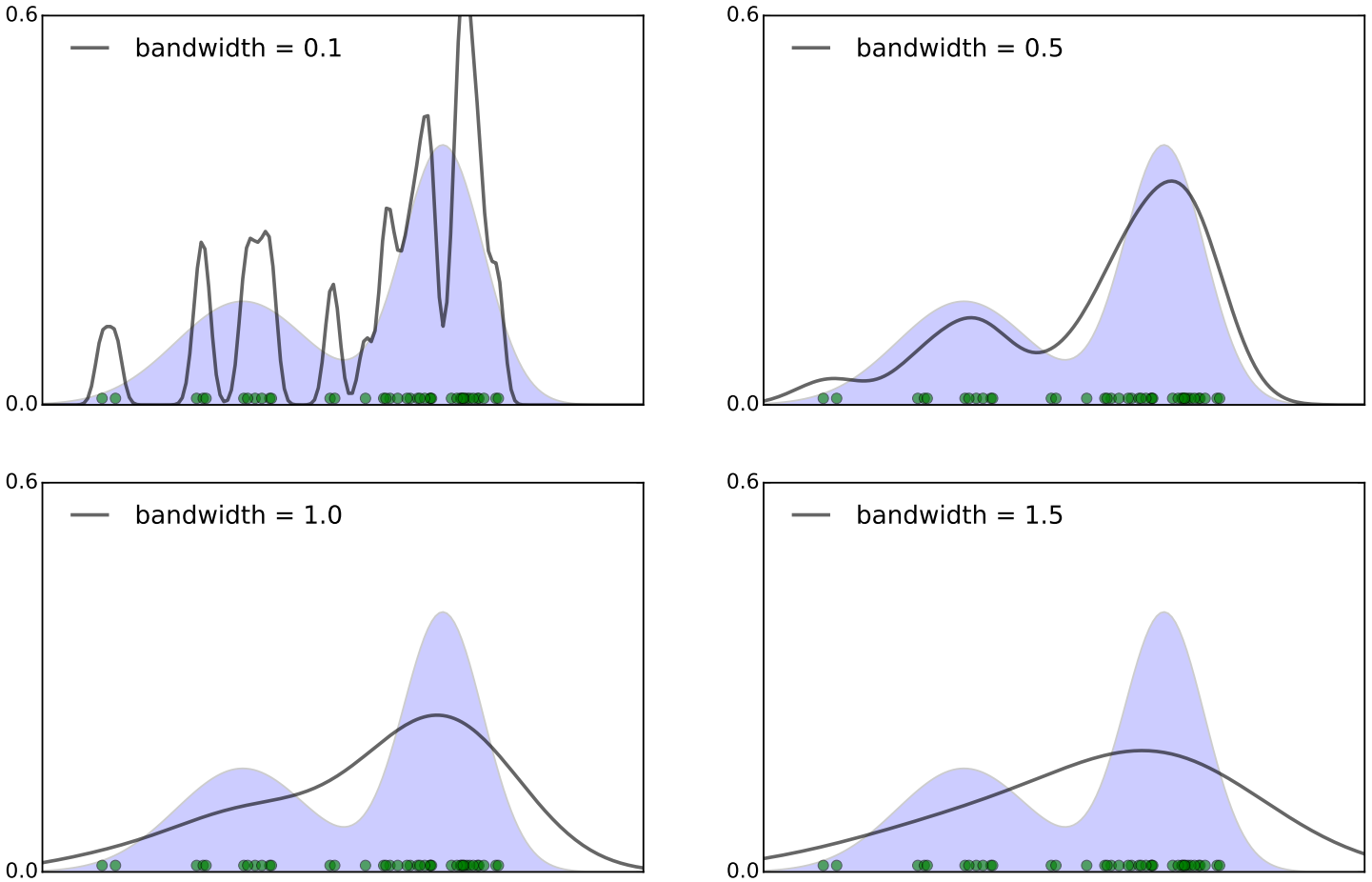

Figure: Effect of changing the bandwidth

Theory of Kernal Density Estimation

Convolution of an arbitrary distribution Q and a density K on $\mathbb{R}^d$ is the density on $\mathbb{R}^d$ defined by

$$(K \star Q)(\mathbf{s}') = \int K(\mathbf{s}' - \mathbf{s}) Q(\mathrm{d}\mathbf{s}) \qquad (\mathbf{s}' \in \mathbb{R}^d) \tag{5}$$

Fact. (14.1.3) For any density K and arbitrary distribution Q on $\mathbb{R}^d$, the density $K \star Q$ equals $\mathcal{L}(\mathbf{x} + \mathbf{y})$ when $\mathbf{x}$ and $\mathbf{y}$ are independent with $\mathcal{L}(\mathbf{x}) = K$ and $\mathcal{L}(\mathbf{y}) = Q$

Example. If $K = \mathrm{N}(0, \sigma^2)$ for some $\sigma > 0$ and Q is a distribution on $\mathbb{R}$ that puts mass q_n on points $s_1, \ldots, s_N$, then by (5) and the rule for integrating over discrete distributions

$$(K \star Q)(s') = \sum_{n=1}^{N} K(s' - s_n) q_n \tag{6}$$

This distribution is a mixture of normals

We are interested in convolutions induced by densities of the form

$$K_h(\mathbf{s}) := \frac{1}{h^d} K\left(\frac{\mathbf{s}}{h}\right) \tag{7}$$

where K is any density and $h > 0$ is a parameter

The density K_h in (7) is the density of $h\mathbf{x}$ when $\mathbf{x}$ is a random vector on $\mathbb{R}^d$ with density K

Example. Figure on next slide shows the convolution of a tent-shaped distribution Q and K_h when K is the standard normal density and h takes different values

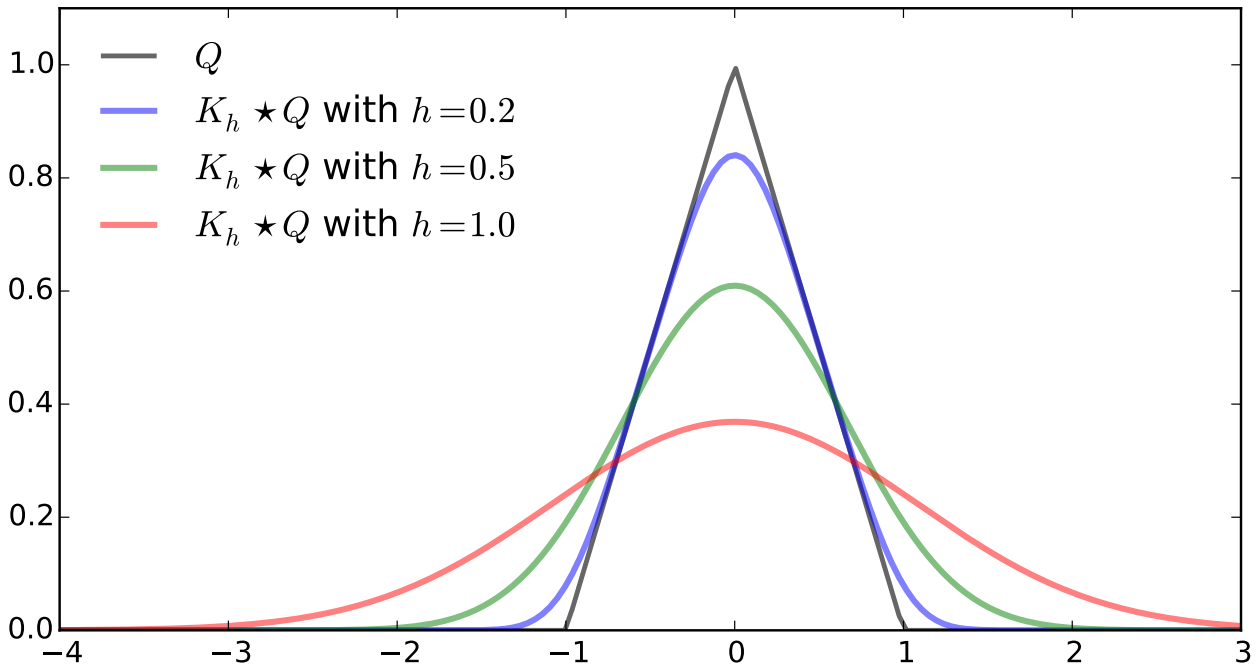

Figure: Smoothing induced by convolutions

Fact. (14.1.4) Let f and K be any densities on $\mathbb{R}^d$, and let K_h be defined from K via (7). If $f \in L_p$, then $K_h \star f \in L_p$, and

$$\lim_{h \downarrow 0} \|K_h \star f - f\|_p = 0$$

Convolution and KDEs

Let h be a positive number and let K_h be as defined in (7)

Rewrite the KDE in (4) as $\hat{f}_N(\mathbf{s}) = \frac{1}{N}\sum_{n=1}^{N} K_h(\mathbf{s} - \mathbf{x}_n)$

$\hat{P}_N$ denotes the empirical distribution of the sample

Recalling expression for integrals with respect to $\hat{P}_N$, we can write

$$\hat{f}_N(\mathbf{s}') = \int K_h(\mathbf{s}' - \mathbf{s})\hat{P}_N(\mathrm{d}\mathbf{s})$$

or, more simply,

$$\hat{f}_N = K_h \star \hat{P}_N \tag{8}$$

Adds smoothing to estimation of density using the sample analogue principle

For *any* density f and kernel K, the nonparametric kernel density estimator $\hat{f}_N$ is L_1-consistent for f

Here we'll state and prove the corresponding L_2 result:

Theorem. (14.1.1) Let f and K be densities on $\mathbb{R}^d$ and elements of L_2. If

1. $\{\mathbf{x}_n\}_{n\geq 1}$ is an IID sequence of draws from f, and
2. the bandwidth sequence $\{h_N\}$ satisfies $h_N \to 0$ and $Nh_N^d \to \infty$ as $N \to \infty$,

then the sequence of density estimates $\{\hat{f}_N\}$ defined in (4) is L_2-consistent for f.

For the remainder of this section, $\|\cdot\|$ is the L_2 norm and $h := h_N$

Using the expression for $\hat{f}_N$ in (8) and the triangle inequality in (5), we have

$$\|\hat{f}_N - f\| \leq \|K_h \star \hat{P}_N - K_h \star f\| + \|K_h \star f - f\| \tag{9}$$

The first term is called the **estimation error**

- we only observe the empirical distribution $\hat{P}_N$ rather than the true distribution f

The second term called the **approximation error** or **bias**

- caused by the smoothing we deliberately added to control the estimation error

Condition 2. of theorem 14.1.1

- $h_N \to 0$ shrinks the amount of smoothing as sample size increases — reducing approximation error as sample size increases
- $Nh_N^d \to \infty$ ensures smoothing not reduced too quickly, controls estimation error

Proof. [Proof of theorem 14.1.1]

By fact 14.1.4, the approximation error converges to zero under the conditions of theorem 14.1.1

We now show the estimation error converges in probability to zero

Proof.(cont.) Fix $\delta > 0$; by Chebyshev's inequality, we have

$$\mathbb{P}\left\{\|K_h \star \hat{P}_N - K_h \star f\| \geq \delta\right\}$$

$$= \mathbb{P}\left\{\|K_h \star \hat{P}_N - K_h \star f\|^2 \geq \delta^2\right\} \leq \frac{\xi_N}{\delta^2}$$

where

$$\xi_N := \mathbb{E}\left\{\|K_h \star \hat{P}_N - K_h \star f\|^2\right\}$$

To complete the proof, we need show ξ_N converges to zero. Let

$$\bar{K}_N(\mathbf{s}) := (K_h \star \hat{P}_N)(\mathbf{s}) - (K_h \star f)(\mathbf{s})$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left\{K_h(\mathbf{s}-\mathbf{x}_n) - \mathbb{E}\left[K_h(\mathbf{s}-\mathbf{x}_n)\right]\right\}$$

Proof.(cont.) We can then write

$$\xi_N = \mathbb{E}\left\{\int [\bar{K}_N(\mathbf{s})]^2 \, \mathrm{d}\mathbf{s}\right\} = \int \mathbb{E}\left\{[\bar{K}_N(\mathbf{s})]^2\right\} \mathrm{d}\mathbf{s} \tag{10}$$

(The interchange of order of expectation and integration is valid for nonnegative integrands)

Since $\bar{K}_N(\mathbf{s})$ is the sample mean of N IID zero-mean random variables, we have

$$\mathbb{E}\left\{[\bar{K}_N(\mathbf{s})]^2\right\} = \mathrm{var}\left[\bar{K}_N(\mathbf{s})\right] = \frac{1}{N}\mathrm{var}[K_h(\mathbf{s} - \mathbf{x}_n)]$$

Proof.(cont.) Moreover

$$\begin{aligned}\operatorname{var}[K_h(\mathbf{s}-\mathbf{x}_n)] &= \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\} - \left\{\mathbb{E}\left[K_h(\mathbf{s}-\mathbf{x}_n)\right]\right\}^2 \\ &\leq \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\}\end{aligned}$$

In summary,

$$\xi_N \leq \frac{1}{N}\int \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\} d\mathbf{s}$$

Switching the order of integration,

$$\int \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\} d\mathbf{s} = \int \left\{\int [K_h(\mathbf{s}-\mathbf{s}')]^2 \, d\mathbf{s}\right\} f(\mathbf{s}') \, d\mathbf{s}'$$

Proof.(cont.) From the definition of K_h and a change of variable argument,

$$\int [K_h(\mathbf{s}-\mathbf{s}')]^2 \, \mathrm{d}\mathbf{s} = \frac{1}{h^{2d}} \int \left[K\left(\frac{\mathbf{s}-\mathbf{s}'}{h} \right) \right]^2 \mathrm{d}\mathbf{s} = \frac{1}{h^d} \int [K(\mathbf{u})]^2 \, \mathrm{d}\mathbf{u}$$

Putting this together with (28) gives the bound

$$\xi_N \leq \int \frac{1}{Nh^d} \|K\|^2 f(\mathbf{s}') \, \mathrm{d}\mathbf{s}' = \frac{1}{Nh^d} \|K\|^2$$

The term $\|K\|^2 := \|K\|_2^2$ is finite by assumption

Recalling $Nh^d = Nh_N^d \to \infty$, we see $\xi_N \to 0$ as required □

Parametric techniques excel when we have knowledge about parametric classes and specific functional forms — from underlying theory

- this is more difficult in economics and other social sciences — makes non-parametric techniques more attractive

Nonparametric methods do not solve all our problems

- theoretical results presented above are purely asymptotic
- strong finite sample results require strict assumptions on the target density
- no uniform rate of convergence for all target densities

Nonparametric methods have relatively little structure in the form of prior knowledge and hence require abundant data

Ridge Regresion

Turn attention to finite sample properties

Ridge regression — a popular method of estimation in both econometrics and machine learning

Ridge regression connects to ideas at the heart of finite sample theory

- complexity
- prior knowledge
- bias–variance trade-off

Start with the OLS setting with assumptions 12.1.2–12.1.4

Recall the usual OLS estimator $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\intercal \mathbf{b})^2$$

OLS estimator minimizes the empirical risk under quadratic loss when the hypothesis space is the set of linear functions

Under our assumptions, the OLS estimator:

- unbiased for $\boldsymbol{\beta}$
- has the lowest variance among all linear unbiased estimators of $\boldsymbol{\beta}$

More natural way to evaluate estimators is to consider their mean squared error

Define the MSE of an estimator $\hat{\mathbf{b}}$ of $\boldsymbol{\beta}$ as

$$\text{mse}(\hat{\mathbf{b}}, \boldsymbol{\beta}) := \mathbb{E}\left\{\|\hat{\mathbf{b}} - \boldsymbol{\beta}\|^2\right\}$$

Alternatively,

$$\text{mse}(\hat{\mathbf{b}}, \boldsymbol{\beta}) = \mathbb{E}\left\{\| \hat{\mathbf{b}} - \mathbb{E}[\hat{\mathbf{b}}] \|^2\right\} + \| \mathbb{E}[\hat{\mathbf{b}}] - \boldsymbol{\beta} \|^2 \qquad (11)$$

Minimization of MSE involves a trade-off between

1. variance term
2. bias term

There exists a biased linear estimator with lower mean squared error than $\hat{\boldsymbol{\beta}}$

The estimator is the solution to the modified least squares problem

$$\min_{\mathbf{b}\in\mathbb{R}_K}\left\{\sum_{n=1}^{N}(y_n-\mathbf{x}_n^{\intercal}\mathbf{b})^2+\lambda\|\mathbf{b}\|^2\right\} \tag{12}$$

where $\lambda \geq 0$ is called the **regularization parameter**

Minimizing the empirical risk plus a term that penalizes large values of $\|\mathbf{b}\|$

The solution to (12) is

$$\hat{\boldsymbol{\beta}}_\lambda := (\mathbf{X}^{\intercal}\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^{\intercal}\mathbf{y} \tag{13}$$

The estimator $\hat{\boldsymbol{\beta}}_\lambda$ is called the **ridge regression estimator**. Note:

1. $\hat{\boldsymbol{\beta}}_\lambda$ is the OLS estimator when $\lambda = 0$ and
2. $\hat{\boldsymbol{\beta}}_\lambda$ is biased whenever $\lambda > 0$ (ex. 14.4.9)

Theorem. (14.2.1) Under the OLS assumptions 12.1.2–12.1.4, there exists a $\lambda > 0$ such that

$$\operatorname{mse}\left(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}\right) < \operatorname{mse}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right)$$

Proof is in Hoerl and Kennard (1970)

Traditional view of ridge regression:

- OLS assumptions valid, however, instances where $\mathbf{X}^\mathsf{T}\mathbf{X}$ is almost singular due to strong correlation between regressors
- in this case, inverting $\mathbf{X}^\mathsf{T}\mathbf{X}$ is numerically unstable
- stabilize the inversion by adding some positive value of λ

Here's another view of ridge regression:

- standard OLS assumptions are implausible
- obtaining the regression function f^* (an infinite dimensional object) with a finite amount of data is ill-posed
- regularization term in ridge regression manages complexity of the candidate functions used to approximate the regression function

Tikhonov Regularization

Least squares estimator the solution to an overdetermined system of equations

Solving ill-posed linear systems in high dimensions: *any attempt to back out or infer a complex object by solving a system about which we have limited information requires a degree of regularization*

To illustrate, suppose

1. $\mathbf{Ab} = \mathbf{c}$ is an overdetermined system, where $\mathbf{A}$ is $N \times K$ with $N > K$
2. due to measurement error, we only observe an approximation $\mathbf{c}_0$ of $\mathbf{c}$
3. $\mathbf{b}^*$ is the (unobservable) least squares solution $\text{argmin}_{\mathbf{b}} \|\mathbf{Ab} - \mathbf{c}\|^2$

Natural approach to approximating $\mathbf{b}^*$ is to solve $\mathbf{Ab} = \mathbf{c}_0$ by least squares

Another approach to solve the above problem:

$$m(\lambda) := \|\mathbf{Ab} - \mathbf{c}_0\|^2 + \lambda\|\mathbf{b}\|^2$$

for some small but positive λ

This approach called **Tikhonov regularization**

- minimizing least squares plus a penalty term

Simulation where $\mathbf{A}$ chosen stochastically but with a tendency towards multicollinearity

- first set $\mathbf{b}^* := (10, 10, \ldots, 10)^\mathsf{T}$
- then set $\mathbf{c} := \mathbf{A}\mathbf{b}^*$

By construction, $\mathbf{b}^*$ is a solution to the system $\mathbf{A}\mathbf{b}^* = \mathbf{c}$, and also the least squares solution

The figure on next slide shows 10 solutions each for the ordinary and regularized solutions, corresponding to 10 draws of $\mathbf{c}_0$

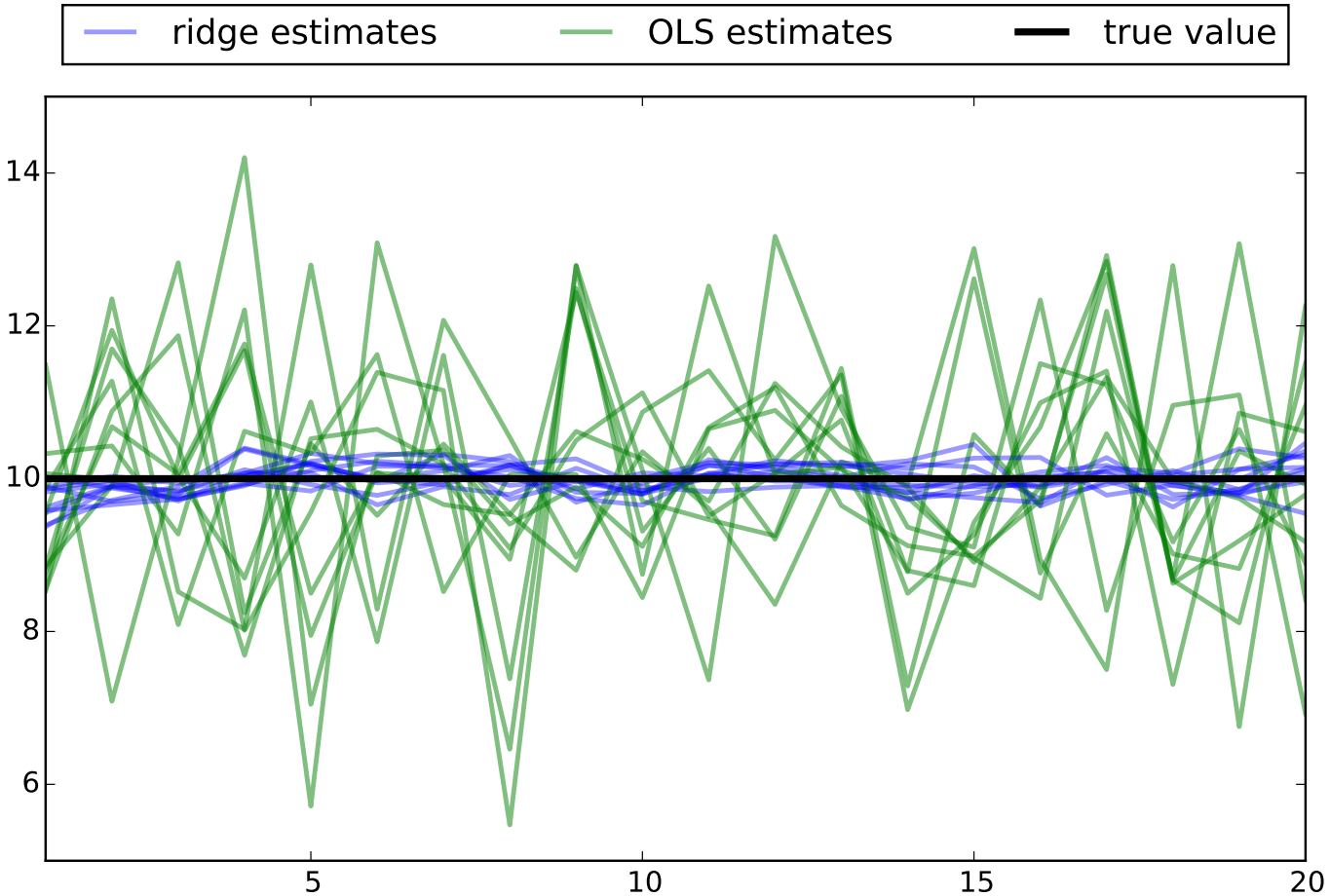

Figure: Effect of Tikhonov regularization, $\lambda = 1$

Subset Selection and Ridge Regression

Problem of **subset selection**

- which variables to include in a regression?
- choose the right set of basis functions

Subset selection a version of the empirical risk minimization problem

Suppose we have output y and inputs $\mathbf{x} \in \mathbb{R}^K$ — $\mathbf{x}$ contains K candidate regressors

To include all regressors, minimize empirical risk over $\mathcal{H}_\ell$, the hypothesis space of linear functions from $\mathbb{R}^K$ to $\mathbb{R}$

To exclude some subset of regressors, set $I \subset \{1, \ldots, K\}$ to be the set of indices of the regressors we want to exclude and regress y on the remainder

Equivalent to minimizing the empirical risk over the hypothesis space

$$\mathcal{H}_{-I} := \left\{ \text{ all functions } f(\mathbf{x}) = \mathbf{b}^\mathsf{T}\mathbf{x} \text{ with } b_k = 0 \text{ for all } k \in I \right\}$$

The subset selection problem has been tackled by many researchers

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Mallow's C_p statistic

K regressors means 2^K subsets— to avoid computational burden, use ridge regression:

- regularization term leads us to choose an estimate with smaller norm
- the coefficients of less helpful regressors are driven towards zero
- reduces model selection problem to tuning a single parameter

Reconsider the problem of mapping a single co-variate x into $\phi(x) = (1, x, x^2, \ldots, x^d)$ and regressing y on $\phi(x)$

The hypothesis spaces were the sets $\mathscr{P}_d$ of degree d polynomials for different values of d

For each d we minimized the empirical risk over $\mathscr{P}_d$, which translates into solving

$$\min_{\mathbf{b}} \sum_{n=1}^{N} \left[y_n - \mathbf{b}^\mathsf{T} \boldsymbol{\phi}(x_n) \right]^2 \quad \text{where} \quad \boldsymbol{\phi}(x) = (x^0, x^1, \ldots, x^d)$$

Choosing the right d isomorphic to subset selection

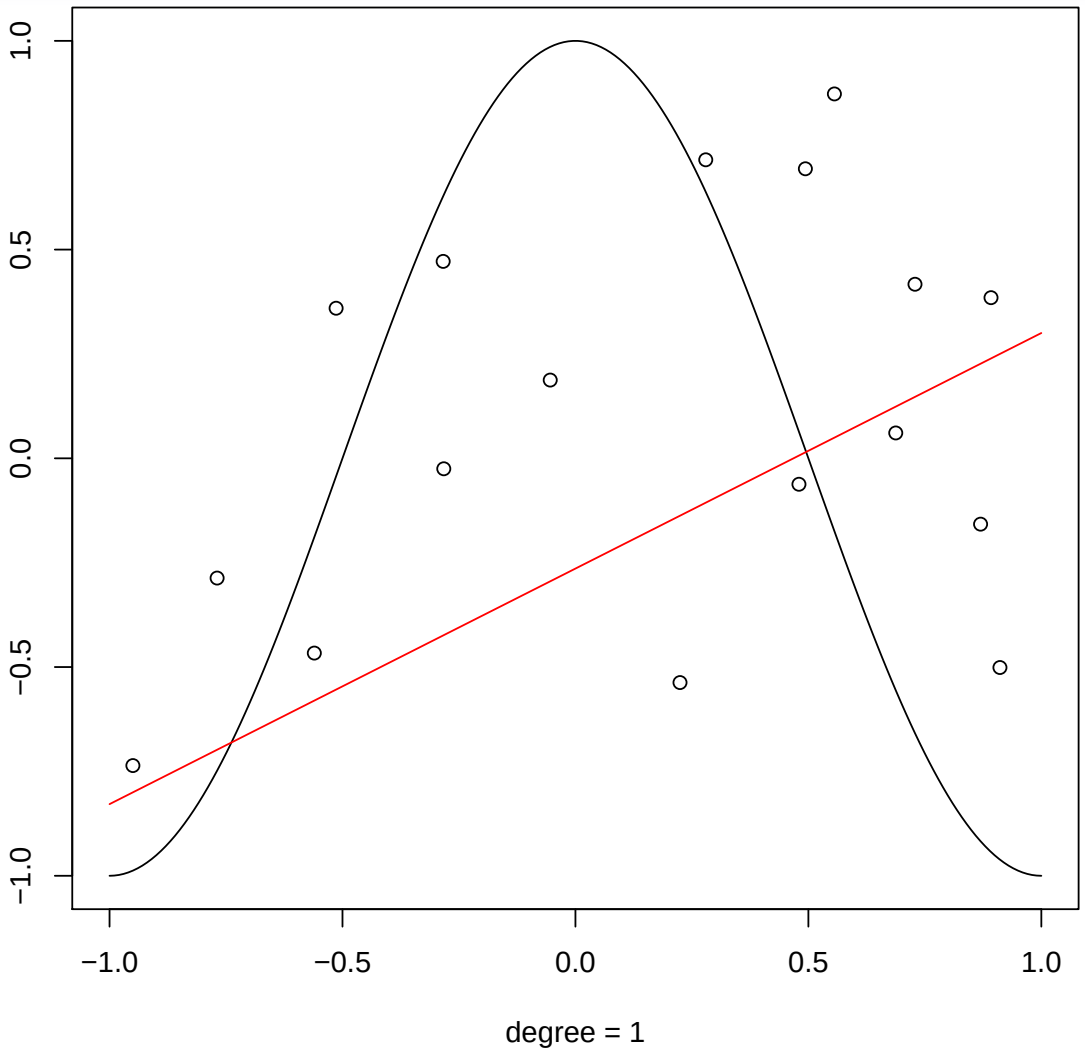

Figure: Fitted polynomial, $d = 1$

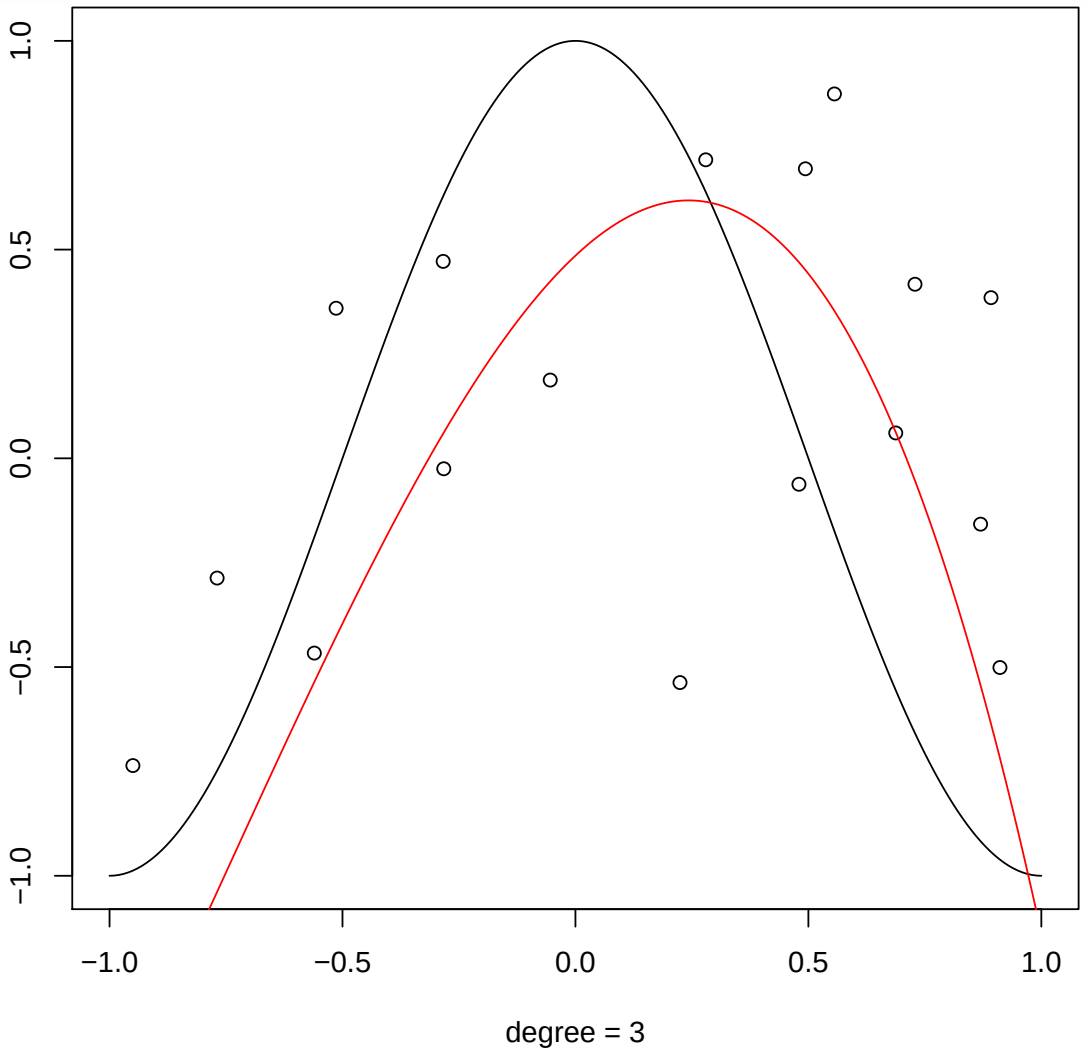

Figure: Fitted polynomial, $d = 3$

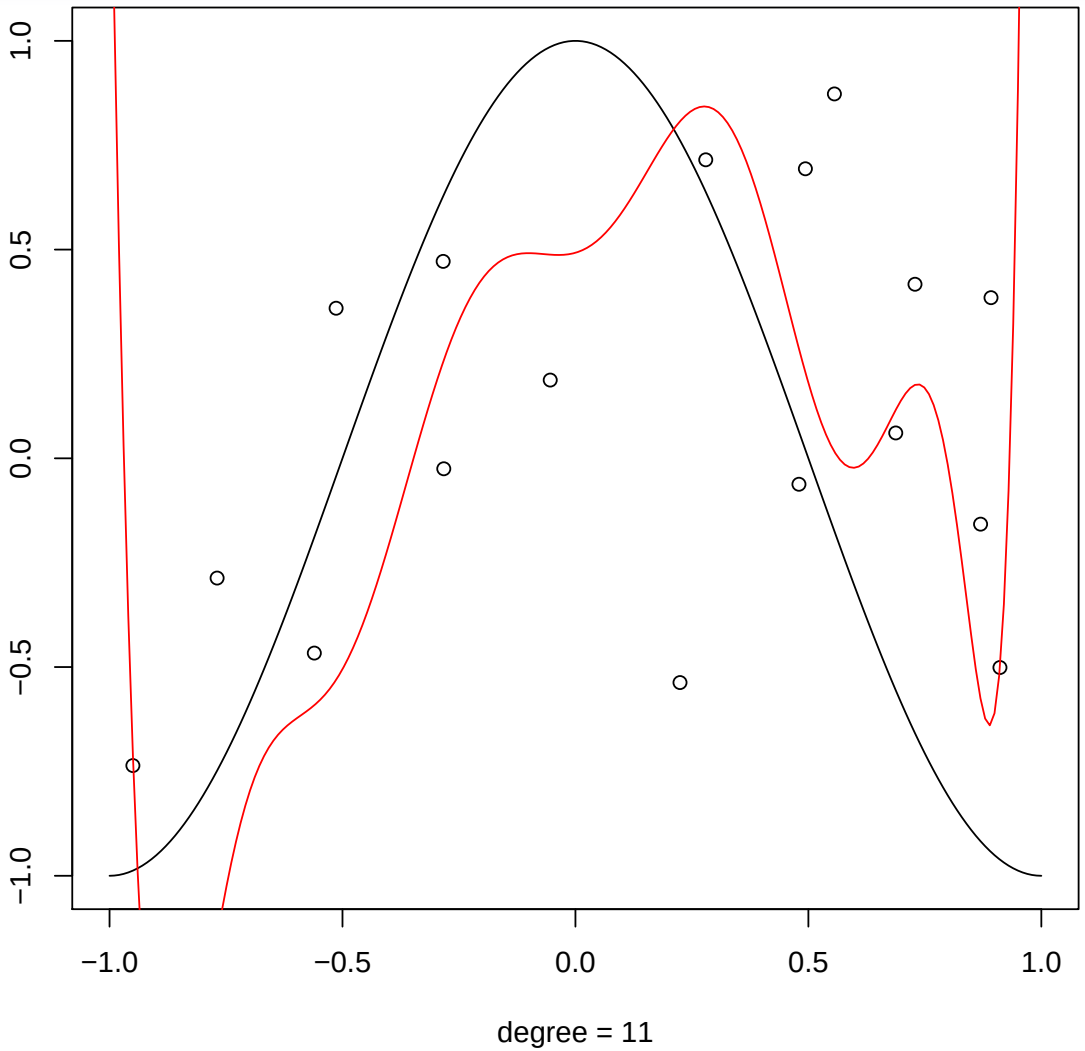

Figure: Fitted polynomial, $d = 11$

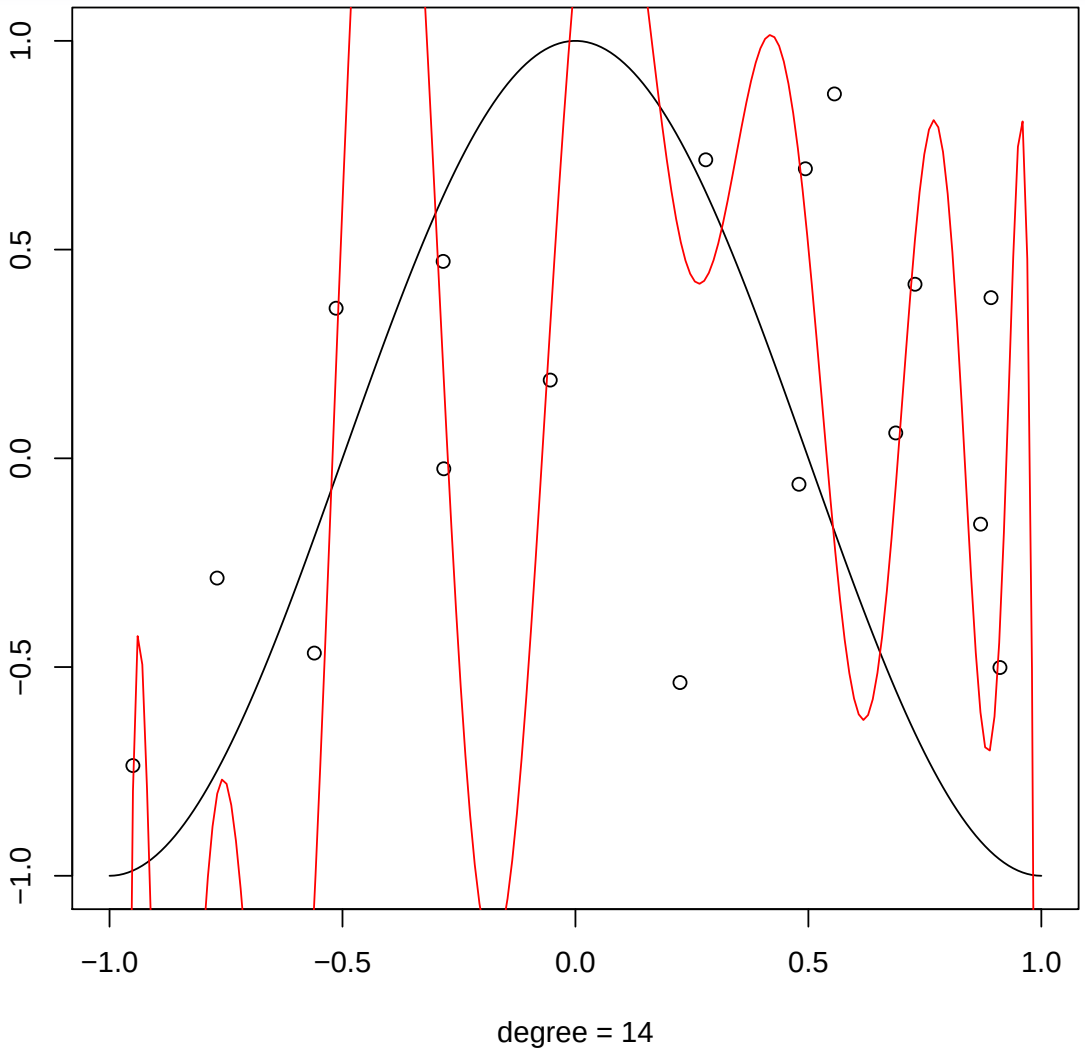

Figure: Fitted polynomial, $d = 14$

To use ridge regression for the problem

First, take $\mathscr{P}_{14}$ as our hypothesis space

- large enough to provide a good fit to the data, but we have overfitting

Solve the regularization problem

$$\min_{\mathbf{b}} \sum_{n=1}^{N} \left\{ \left[y_n - \mathbf{b}^{\mathsf{T}} \boldsymbol{\phi}(x_n) \right]^2 + \lambda \|\mathbf{b}\|^2 \right\}$$

for different values of λ

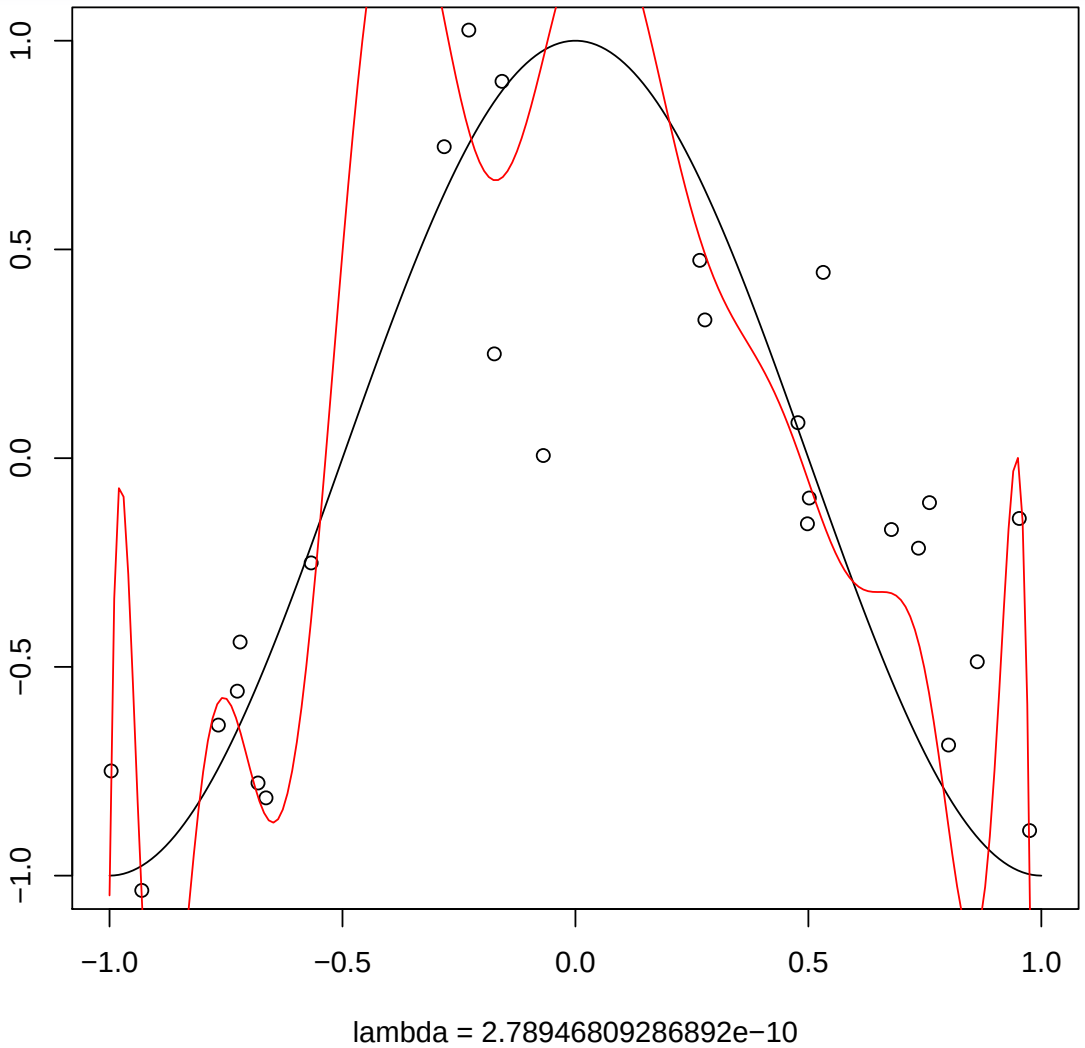

Figure: Fitted polynomial, $\lambda \approx 0$

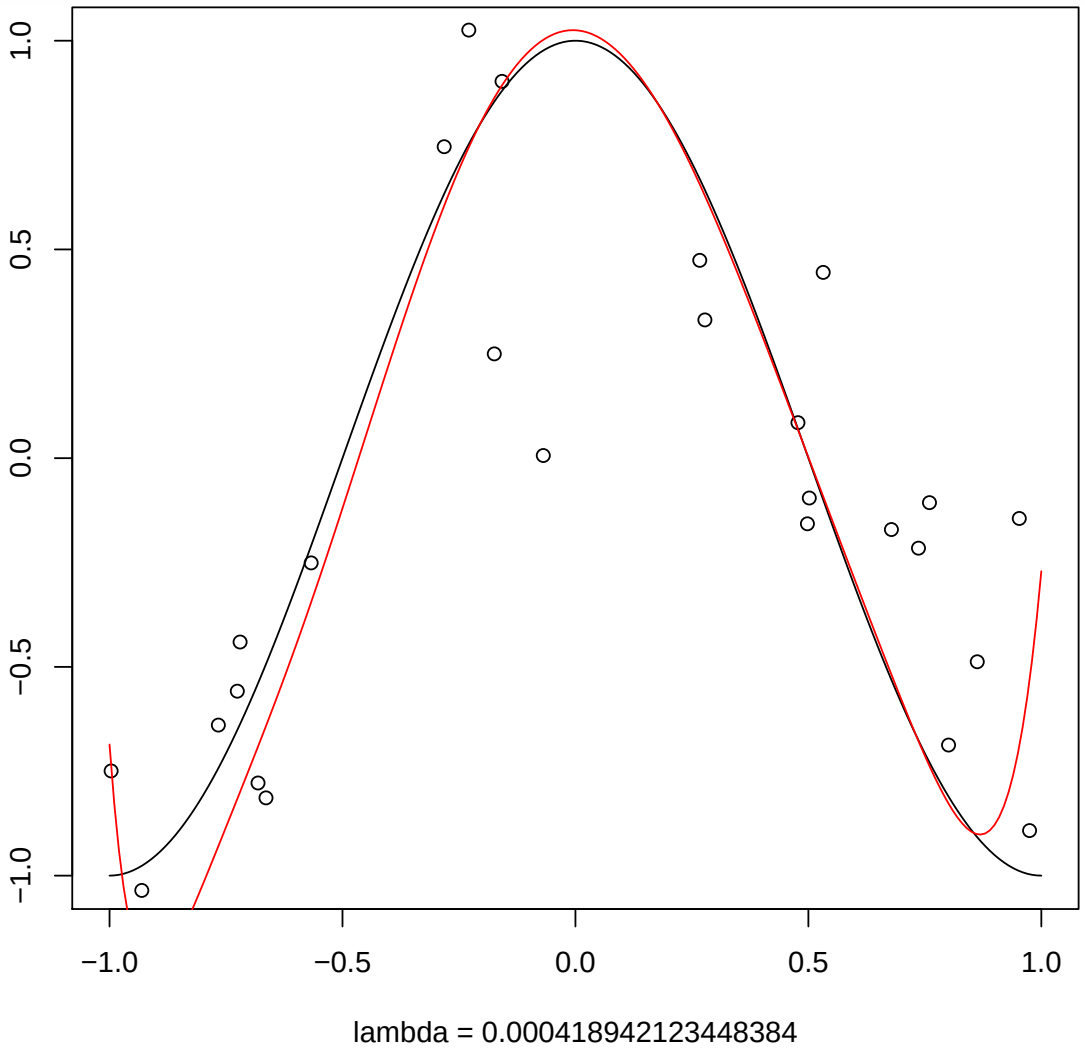

Figure: Fitted polynomial, $\lambda \approx 0.0004$

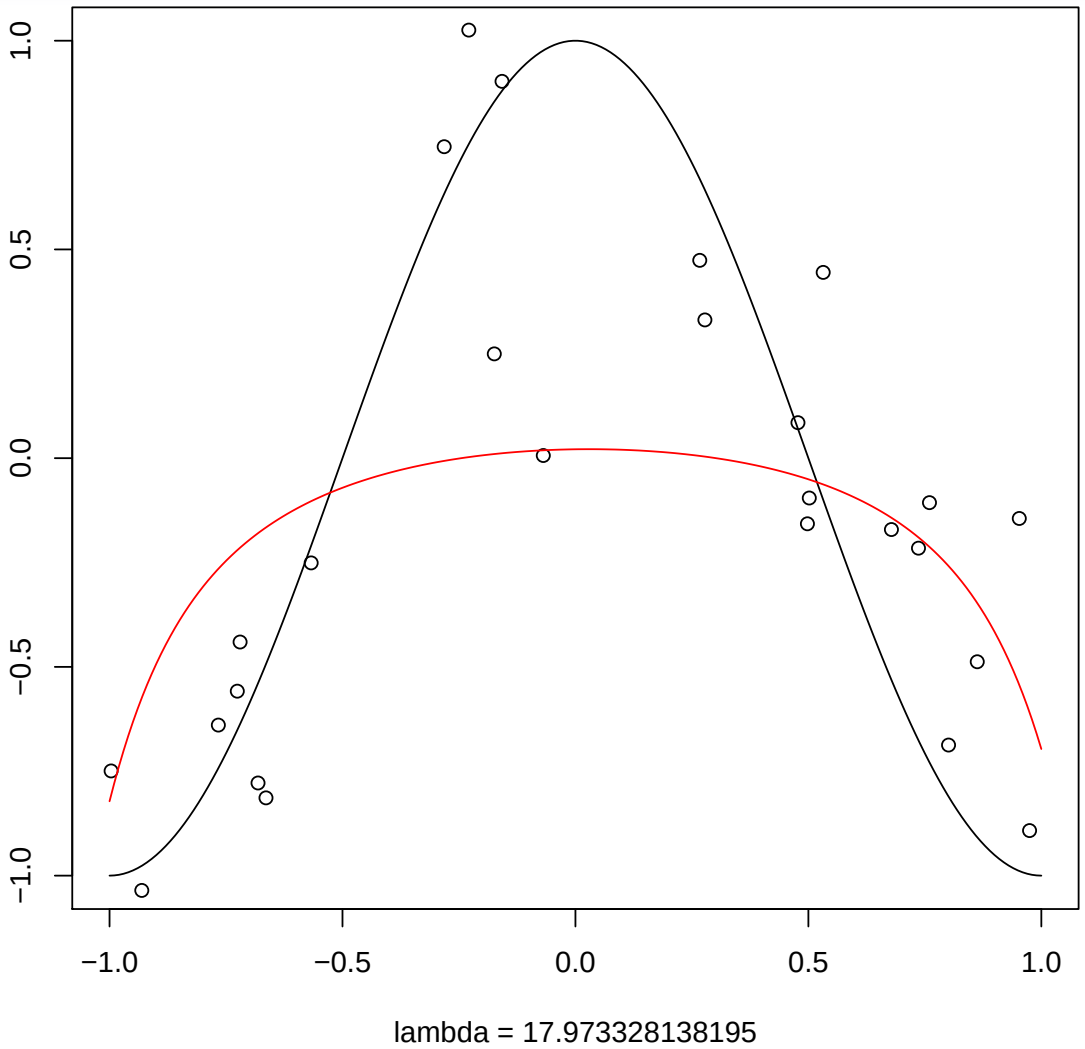

Figure: Fitted polynomial, $\lambda \approx 18$

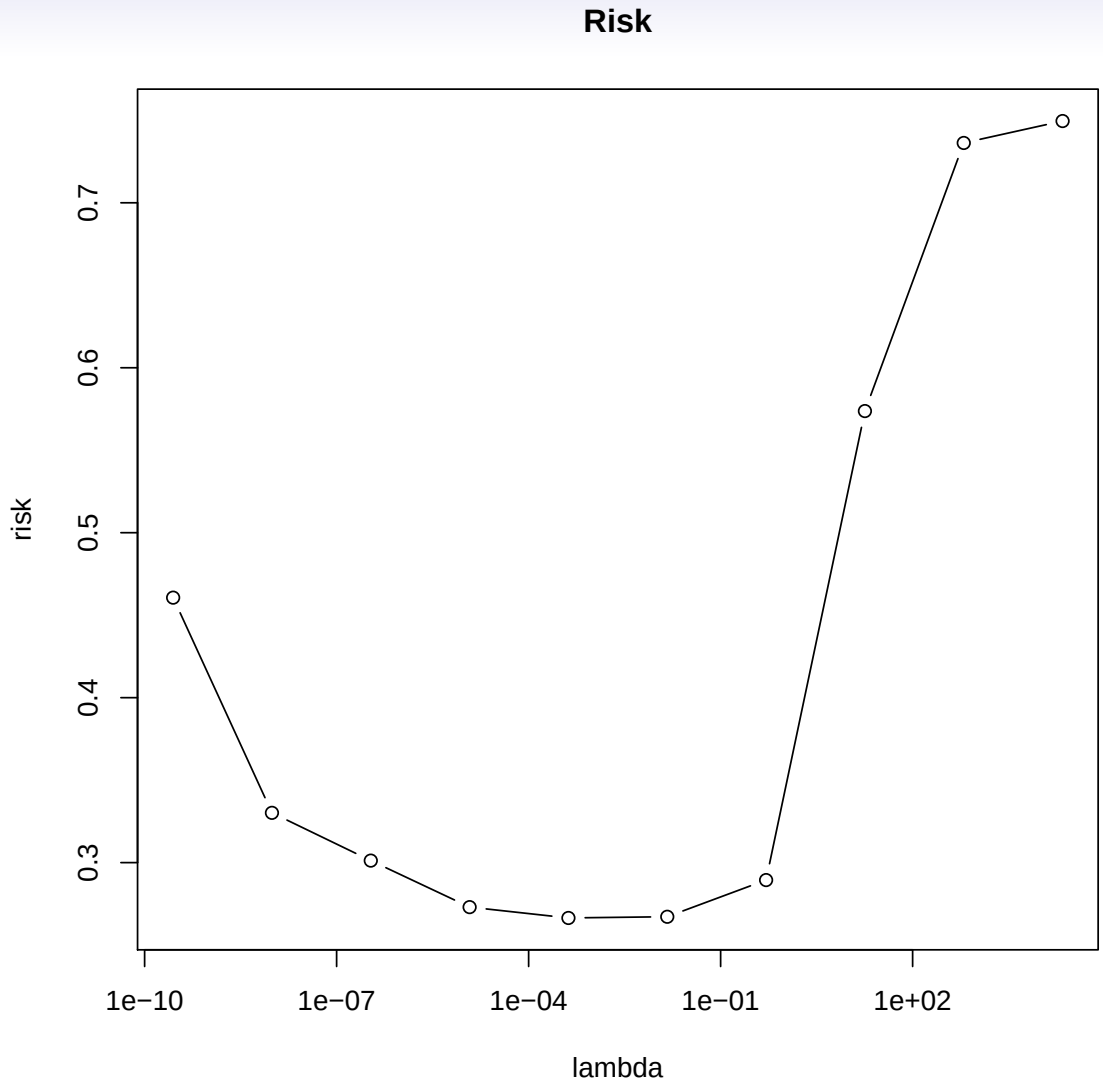

Figure: Risk of $\hat{f}_\lambda$ plotted against λ

Bayesian Methods and Regularization

Bayesian analysis provides method for including prior information into statistical estimation

Prior information can be

- guidance from economic theory on which regressors to include
- which functional forms to use
- which values of our regularization parameter to choose

We now compare Bayesian linear regression with ridge regression

Suppose

- regression data take the linear form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
- for simplicity, $\mathbf{X}$ is nonrandom (Taking $\mathbf{X}$ to be random leads to the same conclusions but with a longer derivation)
- $\mathbf{u}$ is random and unobservable

Bayesian perspective — take $\boldsymbol{\beta}$ to be random and unobservable as well

Take the priors to be $\mathcal{L}(\mathbf{u}) = \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathcal{L}(\boldsymbol{\beta}) = \mathrm{N}(\mathbf{0}, \tau^2\mathbf{I})$

Given our model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, our prior on $\mathbf{u}$ implies that the density of $\mathbf{y}$ given $\boldsymbol{\beta}$ is $\mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$

Write our distributions as

$$p(\mathbf{y} \mid \boldsymbol{\beta}) = \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \text{and} \quad p(\boldsymbol{\beta}) = \mathrm{N}(\mathbf{0}, \tau^2\mathbf{I}) \tag{14}$$

Applying Bayes' law to the pair $(\mathbf{y}, \boldsymbol{\beta})$, we obtain

$$p(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y})} \tag{15}$$

The left-hand side is the posterior density of $\boldsymbol{\beta}$ given the data $\mathbf{y}$

The maximizer of the posterior is called the **maximum a posteriori** (MAP) probability estimate

Taking logs of (15) and dropping the term that does not contain $\boldsymbol{\beta}$, it can be expressed as

$$\hat{\boldsymbol{\beta}}_M := \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \{\ln p(\mathbf{y} \mid \boldsymbol{\beta}) + \ln p(\boldsymbol{\beta})\} \tag{16}$$

Insert our functional forms into (14), drop constant terms and multiply by -1 to obtain

$$\hat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} (y_n - \mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta})^2 + \frac{\sigma^2}{\tau^2} \|\boldsymbol{\beta}\|^2 \right\} \tag{17}$$

This is the penalized least squares problem (12)

The regularization parameter λ is equal to $(\sigma/\tau)^2$

The solution:

$$\hat{\beta}_M := (\mathbf{X}^\mathsf{T}\mathbf{X} + (\sigma/\tau)^2\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

Bayesian analysis provides the same effect as Tikhonov regularization, but now regularization arises out of combining prior knowledge with the data

- $(\sigma/\tau)^2$ part of our prior knowledge, hence no model selection problem

In practice, we can question that we have prior knowledge to pin down $\lambda := (\sigma/\tau)^2$

- back at the model selection problem

Now we forgo the assumption that strong prior knowledge available and consider a more automated approach to choosing λ

Cross-Validation

Recall, given loss function L and a system producing input–output pairs $(\mathbf{x}, y) \in \mathbb{R}^{K+1}$ with joint distribution P, the prediction risk of a function $f\colon \mathbb{R}^K \to \mathbb{R}$ is

$$R(f) := \mathbb{E}\left[L(y, f(\mathbf{x}))\right] = \int \int L(t, f(\mathbf{s}))P(\mathrm{d}t, \mathrm{d}\mathbf{s})$$

Suppose we observe N IID input–output pairs $\mathbf{z}_{\mathcal{D}} := \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_N, y_N)\}$

Given a selection of models, we would like to find the one that takes this data set and returns a predictor $\hat{f}$ such that $\hat{f}$ has lower prediction risk than the predictors returned by the other models

Take the data set as given, see how well we can predict new values as evaluated by expected loss

Define the prediction risk of $\hat{f}$ to be the expected loss taking $\mathbf{z}_{\mathcal{D}}$ (and hence $\hat{f}$) as given:

$$R(\hat{f} \mid \mathcal{D}) := \mathbb{E}[L(y, \hat{f}(\mathbf{x})) \mid \mathcal{D}] = \int \int L(t, \hat{f}(\mathbf{s}))P(\mathrm{d}t, \mathrm{d}\mathbf{s})$$

If we have a collection of models M indexed by m, and $\hat{f}_m$ is the predictor produced by fitting model m with data $\mathcal{D}$, then we would like to find the model m^* such that

$$R(\hat{f}_{m^*} \mid \mathcal{D}) \leq R(\hat{f}_m \mid \mathcal{D}) \quad \text{for all } m \in M$$

Problem: we do not know P

Could we approximate $R(\hat{f} \mid \mathcal{D})$ by

$$\frac{1}{N} \sum_{n=1}^{N} L(y_n, \hat{f}(\mathbf{x}_n))$$

where the pairs $(\mathbf{x}_n, y_n)$ are from the data set $\mathbf{z}_{\mathcal{D}}$?

- this is just the empirical risk, and the empirical risk is a highly biased estimator of the risk

In essence, the problem is that we are using the data $\mathbf{z}_{\mathcal{D}}$ twice, for conflicting objectives

1. we are using it to fit the model, producing $\hat{f}$
2. using it to evaluate the predictive ability of $\hat{f}$ on new observations

So what we really need is fresh data!

If we had J new observations $(y_j^v, \mathbf{x}_j^v)$, then estimate the risk by

$$\frac{1}{J}\sum_{j=1}^{J} L(y_j^v, \hat{f}(\mathbf{x}_j^v))$$

Not a genuine solution because we don't have any new data in general

Take $\mathbf{z}_{\mathcal{D}}$ and split it into two disjoint subsets, called the **training set** and the **validation set**

- training set is used to fit $\hat{f}$
- validation set is used to estimate the risk of $\hat{f}$

Repeat for all models and choose the one with lowest estimated risk

Cross-validation attempts to use the whole data set for both fitting the model and estimating the risk

Suppose we partition the data set into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$

- use $\mathcal{D}_1$ as the training set and $\mathcal{D}_2$ as the validation set
- next, use $\mathcal{D}_2$ as the training set, and $\mathcal{D}_1$ as the validation set

Estimate of the risk is the average of the estimates of the risk produced in these two steps

Divide data into more than two sets — extreme is to partition the data into N subsets; called **leave-one-out cross-validation**

Letting $\mathcal{D}_{-n} := \mathbf{z}_{\mathcal{D}} \setminus \{(\mathbf{x}_n, y_n)\}$, the data set with just the nth data point $(\mathbf{x}_n, y_n)$ omitted, the leave-one-out cross validation algorithm:

1: **for** $n = 1, \ldots, N$ **do**
2: fit $\hat{f}_{-n}$ using data $\mathcal{D}_{-n}$
3: set $r_n := L(y_n, \hat{f}_{-n}(\mathbf{x}_n))$
4: **end for**
5: return the risk estimate $r := \frac{1}{N} \sum_{n=1}^{N} r_n$

In terms of model selection

- run each model through the cross-validation procedure
- select the one that produces the lowest value of r

Consider the ridge regression procedure applied to the problem of subset selection we looked at above

- set of models is indexed by λ, the regularization parameter in the ridge regression

For each λ, the fitted function $\hat{f}_\lambda$ is

$$\hat{f}_\lambda(x) = \hat{\boldsymbol{\beta}}_\lambda^\mathsf{T} \boldsymbol{\phi}(x)$$

$$\text{where} \quad \hat{\boldsymbol{\beta}}_\lambda := \operatorname*{argmin}_{\mathbf{b}} \sum_{n=1}^{N} \left\{ (y_n - \mathbf{b}^\mathsf{T} \boldsymbol{\phi}(x_n))^2 + \lambda \|\mathbf{b}\|^2 \right\}$$

Recall

- $\boldsymbol{\phi}(x) = (x^0, x^1, \ldots, x^d)$ with d fixed at 14
- we are fitting a polynomial of degree 14 to the data by minimizing regularized least squares error
- amount of regularization is increasing in λ
- intermediate values of λ produced the best fit in terms of minimizing risk

We used the fact that we knew the underlying model to evaluate the risk, and hence the values of λ that produce low risk

- however, risk is unobservable, and we need to choose λ on the basis of the data alone (assuming no priors)

Experiment: for each λ in the grid `exp(seq(-22, 10, length=10))`, perform leave-one-out cross-validation

The fit at each step within the loop is via ridge regression, omitting the nth data point, and the resulting polynomial is used to predict y_n from x_n

For each λ in the grid, we use the following to estimate risk:

1: **for** $n = 1, \ldots, N$ **do**
2: set $\hat{\boldsymbol{\beta}}_{\lambda,-n} := \operatorname{argmin}_{\mathbf{b}} \sum_{i \neq n} \{(y_i - \mathbf{b}^\mathsf{T} \boldsymbol{\phi}(x_i))^2 + \lambda \|\mathbf{b}\|^2\}$
3: set $r_{\lambda,n} := (y_n - \hat{\boldsymbol{\beta}}_{\lambda,-n}^\mathsf{T} \boldsymbol{\phi}(x_n))^2$
4: **end for**
5: return $r_\lambda := \frac{1}{N} \sum_{n=1}^{N} r_{\lambda,n}$

The value of λ producing the smallest estimated risk r_λ is around 0.015

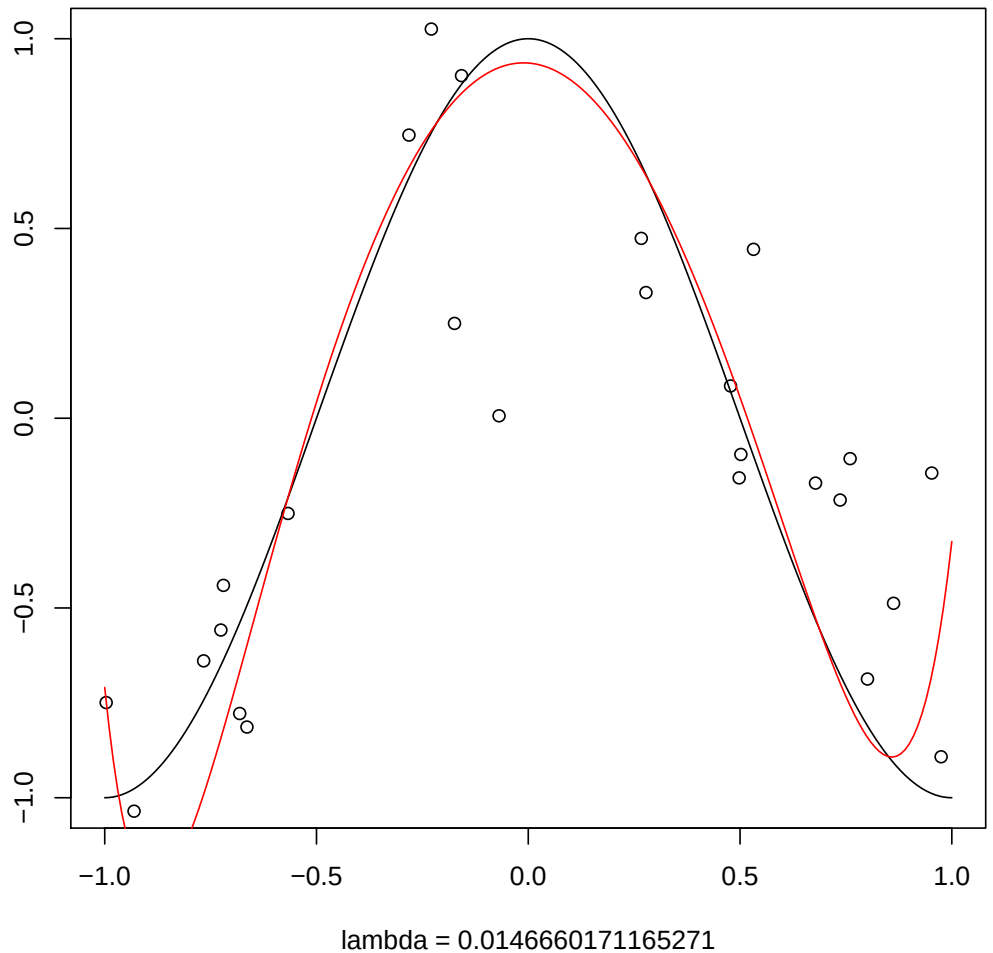

Figure: Fitted polynomial, $\lambda \approx 0.015$