TIARA Summer School on Astrostatistics: Density Estimation and Regression

Yen-Chi Chen

Department of Statistics
University of Washington

Summer 2017

Density Estimation

Density Estimation: Introduction

- Recall that a statistical model views the data as random variables $X_1, \cdots, X_n$ from an unknown distribution function $P(x)$.
- We further assume that such a distribution function has a probability density function (PDF) $p(x)$.

Density Estimation: Introduction

- Recall that a statistical model views the data as random variables $X_1, \cdots, X_n$ from an unknown distribution function $P(x)$.
- We further assume that such a distribution function has a probability density function (PDF) $p(x)$.
- In most cases, we do not know the PDF $p(x)$ but we want to reconstruct it from the data.
- The goal of density estimation is to *estimate* $p(x)$ *using* $X_1, \cdots, X_n$.
- In other words, the parameter of interest is the PDF $p(x)$.

Parametric Approach: Introduction

- The PDF $p(x)$ is a function. It may not be easily estimated.
- A simple approach is to further assume that $p(x) = p(x; \theta)$ for some parameter θ.
- Then estimating p can be done by estimating θ.
- Estimating a value is much easier than estimating the entire function.

Parametric Approach: MLE

- We can estimate the parameters by the MLE.
- Here is an example of assuming the data being normally distributed: $N(\mu, \sigma^2)$.
- In this case, we only need to estimate the two parameters μ, σ^2.

Parametric Approach: MLE

- We can estimate the parameters by the MLE.
- Here is an example of assuming the data being normally distributed: $N(\mu, \sigma^2)$.
- In this case, we only need to estimate the two parameters μ, σ^2.
- The estimators are

$$\widehat{\mu}_{MLE} = \bar{X}_n, \quad \widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

- Then the estimated density function is

$$\widehat{p}(x) = \frac{1}{\sqrt{2\pi\widehat{\sigma}^2_{MLE}}} e^{-\frac{(x-\widehat{\mu}_{MLE})^2}{2\widehat{\sigma}^2_{MLE}}}.$$

- Note that MLE is not the only approach to estimate the parameter; one can use method of moments or other approaches.

Parametric Approach: Mixture Model - 1

- Sometimes simple parametric models such as Gaussian, exponential, Gamma distribution are too restrictive to capture the complicated structure of the data.
- For instance, if the distribution has a bimodal density (two local maxima), none of these traditional model is reasonable.
- In this case, the *mixture model* may be useful.

Parametric Approach: Mixture Model - 2

- The idea of mixture model is to assume that the PDF can be written as a mixture of several 'simple' parametric densities.

Parametric Approach: Mixture Model - 2

- The idea of mixture model is to assume that the PDF can be written as a mixture of several 'simple' parametric densities.
- For instance, Gaussian mixture model assumes

$$p(x) = \omega_1 \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \cdots + \omega_K \cdot \frac{1}{\sqrt{2\pi\sigma_K^2}} e^{-\frac{(x-\mu_K)^2}{2\sigma_K^2}},$$

where $\omega_1 + \cdots + \omega_K = 1$ and $\omega_\ell > 0$.

Parametric Approach: Mixture Model - 2

- The idea of mixture model is to assume that the PDF can be written as a mixture of several 'simple' parametric densities.
- For instance, Gaussian mixture model assumes

$$p(x) = \omega_1 \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \cdots + \omega_K \cdot \frac{1}{\sqrt{2\pi\sigma_K^2}} e^{-\frac{(x-\mu_K)^2}{2\sigma_K^2}},$$

  where $\omega_1 + \cdots + \omega_K = 1$ and $\omega_\ell > 0$.
- The above model assumes that the PDF consists of K components and each component is a Gaussian.
- The quantities $\omega_1, \cdots, \omega_K$ are the mixing proportion of each component.

Parametric Approach: Mixture Model - 2

- The idea of mixture model is to assume that the PDF can be written as a mixture of several 'simple' parametric densities.
- For instance, Gaussian mixture model assumes

$$p(x) = \omega_1 \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \cdots + \omega_K \cdot \frac{1}{\sqrt{2\pi\sigma_K^2}} e^{-\frac{(x-\mu_K)^2}{2\sigma_K^2}},$$

  where $\omega_1 + \cdots + \omega_K = 1$ and $\omega_\ell > 0$.
- The above model assumes that the PDF consists of K components and each component is a Gaussian.
- The quantities $\omega_1, \cdots, \omega_K$ are the mixing proportion of each component.
- In this case, the parameters are

$$\theta = (\omega_1, \mu_1, \sigma_1^2, \cdots, \omega_K, \mu_K, \sigma_K^2).$$

Parametric Approach: Mixture Model - 3

- We can estimate a mixture model using the MLE again.
- However, performing MLE in the mixture model is often computationally difficult and in general, there is no closed form solution to the MLE.
- People often use a method called *EM algorithm* to compute the MLE.

Parametric Approach: Mixture Model - 3

- We can estimate a mixture model using the MLE again.
- However, performing MLE in the mixture model is often computationally difficult and in general, there is no closed form solution to the MLE.
- People often use a method called *EM algorithm* to compute the MLE.
- In additional to the computational challenges, the identifiability is another issue of the mixture model (different parameters lead to the same PDF).

Nonparametric Approach: Introduction

- Parametric models only require estimating a few parameters to estimate the PDF.
- However, they are either too restrictive to capture the intricate structure of the PDF or computationally infeasible.

Nonparametric Approach: Introduction

- Parametric models only require estimating a few parameters to estimate the PDF.
- However, they are either too restrictive to capture the intricate structure of the PDF or computationally infeasible.
- An alternative approach is to estimate the PDF nonparametrically.
- Namely, we directly estimate the PDF without assuming a parametric form of the PDF.

Nonparametric Approach: Histogram

- A simple nonparametric approach is the histogram.
- A histogram first bins the entire range into equal width bins and counts the number of observation within each bin.

Nonparametric Approach: Histogram

- A simple nonparametric approach is the histogram.
- A histogram first bins the entire range into equal width bins and counts the number of observation within each bin.
- To make a histogram a density estimator, we need to rescale the Y-axis a bit.
- Instead of using the count of numbers observations within each bin, we need to divide the count by the total number of observations and the width of the bin.

Nonparametric Approach: Histogram

- A simple nonparametric approach is the histogram.
- A histogram first bins the entire range into equal width bins and counts the number of observation within each bin.
- To make a histogram a density estimator, we need to rescale the Y-axis a bit.
- Instead of using the count of numbers observations within each bin, we need to divide the count by the total number of observations and the width of the bin.
- Assume our histogram has bins $B_1, \cdots, B_K$ and all bins have width L.
- For a point x within the bin B_ℓ, the density estimated by the histogram is

$$\widehat{p}_{hist}(x) = \frac{\#of\, X_1, \cdots, X_n \text{ within } B_\ell}{n \cdot L}.$$

Bias-Variance Tradeoff - 1

- Is the histogram a good estimator?

Bias-Variance Tradeoff - 1

- Is the histogram a good estimator?
- We can answer this question using the mean square error (MSE).

Bias-Variance Tradeoff - 1

- Is the histogram a good estimator?
- We can answer this question using the mean square error (MSE).
- Given x being fixed, the quantity $\widehat{p}_{hist}(x)$ is a random variable and it is the estimator of $p(x)$.
- Therefore, we compute its bias and variance to obtain the MSE.

Bias-Variance Tradeoff - 1

- Is the histogram a good estimator?
- We can answer this question using the mean square error (MSE).
- Given x being fixed, the quantity $\widehat{p}_{hist}(x)$ is a random variable and it is the estimator of $p(x)$.
- Therefore, we compute its bias and variance to obtain the MSE.
- It terms outs that when the size of bin $L \approx 0$ and sample size n is large,

$$\textbf{bias}(\widehat{p}_{hist}(x)) = O(L), \quad \text{Var}(\widehat{p}_{hist}(x)) = O\left(\frac{1}{nL}\right).$$

- Therefore, the MSE of the histogram estimator is

$$\text{MSE}(\widehat{p}_{hist}(x)) = O(L^2) + O\left(\frac{1}{nL}\right).$$

Bias-Variance Tradeoff - 2

$$\mathrm{MSE}(\widehat{p}_{hist}(x)) = \underbrace{O(L^2)}_{\text{Bias}} + \underbrace{O\left(\frac{1}{nL}\right)}_{\text{Variance}}.$$

- The MSE shows a very important pattern: it can be decomposed into the bias and the variance.

Bias-Variance Tradeoff - 2

$$\mathrm{MSE}(\widehat{p}_{hist}(x)) = \underbrace{O(L^2)}_{\text{Bias}} + \underbrace{O\left(\frac{1}{nL}\right)}_{\text{Variance}}.$$

- The MSE shows a very important pattern: it can be decomposed into the bias and the variance.
- When the bin width L is small, the bias is small but the variance is large.
- When the bin width L is large, the bias is large but the variance is small.

Bias-Variance Tradeoff - 2

$$\text{MSE}(\widehat{p}_{hist}(x)) = \underbrace{O(L^2)}_{\text{Bias}} + \underbrace{O\left(\frac{1}{nL}\right)}_{\text{Variance}}.$$

- The MSE shows a very important pattern: it can be decomposed into the bias and the variance.
- When the bin width L is small, the bias is small but the variance is large.
- When the bin width L is large, the bias is large but the variance is small.
- Such a tradeoff between bias and variance is known as the *bias-variance tradeoff*.

Bias-Variance Tradeoff - 2

$$\mathrm{MSE}(\widehat{p}_{hist}(x)) = \underbrace{O(L^2)}_{\text{Bias}} + \underbrace{O\left(\frac{1}{nL}\right)}_{\text{Variance}}.$$

- The MSE shows a very important pattern: it can be decomposed into the bias and the variance.
- When the bin width L is small, the bias is small but the variance is large.
- When the bin width L is large, the bias is large but the variance is small.
- Such a tradeoff between bias and variance is known as the *bias-variance tradeoff*.
- Moreover, it shows that we should choose L at the rate of $L \asymp n^{-1/3}$ to minimize the MSE.
- This choice leads to the optimal rate of histogram: $\mathrm{MSE}^*(\widehat{p}_{hist}(x)) = O(n^{-2/3})$.

Nonparametric Approach: Kernel Density Estimation - 1

- Here we introduce another nonparametric density estimation approach: the kernel density estimation (KDE).

Nonparametric Approach: Kernel Density Estimation - 1

- Here we introduce another nonparametric density estimation approach: the kernel density estimation (KDE).
- The KDE estimate the PDF using the following form:

$$\widehat{p}_{KDE}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $K(x)$ is a function called the kernel function and $h > 0$ is a quantity called smoothing bandwidth that controls the amount of smoothing.

Nonparametric Approach: Kernel Density Estimation - 1

- Here we introduce another nonparametric density estimation approach: the kernel density estimation (KDE).
- The KDE estimate the PDF using the following form:

$$\widehat{p}_{KDE}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

  where $K(x)$ is a function called the kernel function and $h > 0$ is a quantity called smoothing bandwidth that controls the amount of smoothing.
- Common choice of $K(x)$ includes the Gaussian $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, uniform $K(x) = \frac{1}{2} I(-1 \leq x \leq 1)$.

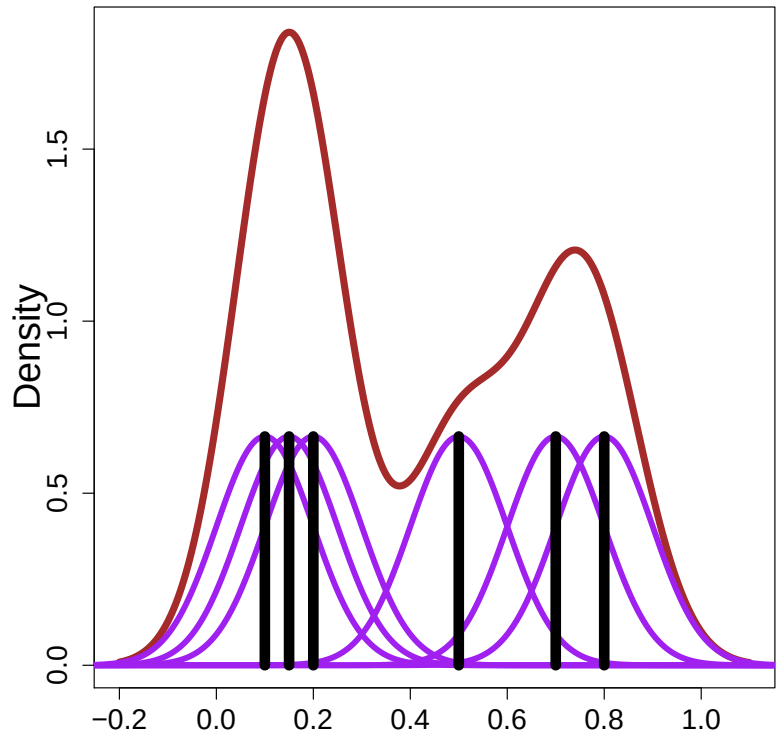
Nonparametric Approach: Kernel Density Estimation - 1

- Here we introduce another nonparametric density estimation approach: the kernel density estimation (KDE).
- The KDE estimate the PDF using the following form:

$$\widehat{p}_{KDE}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

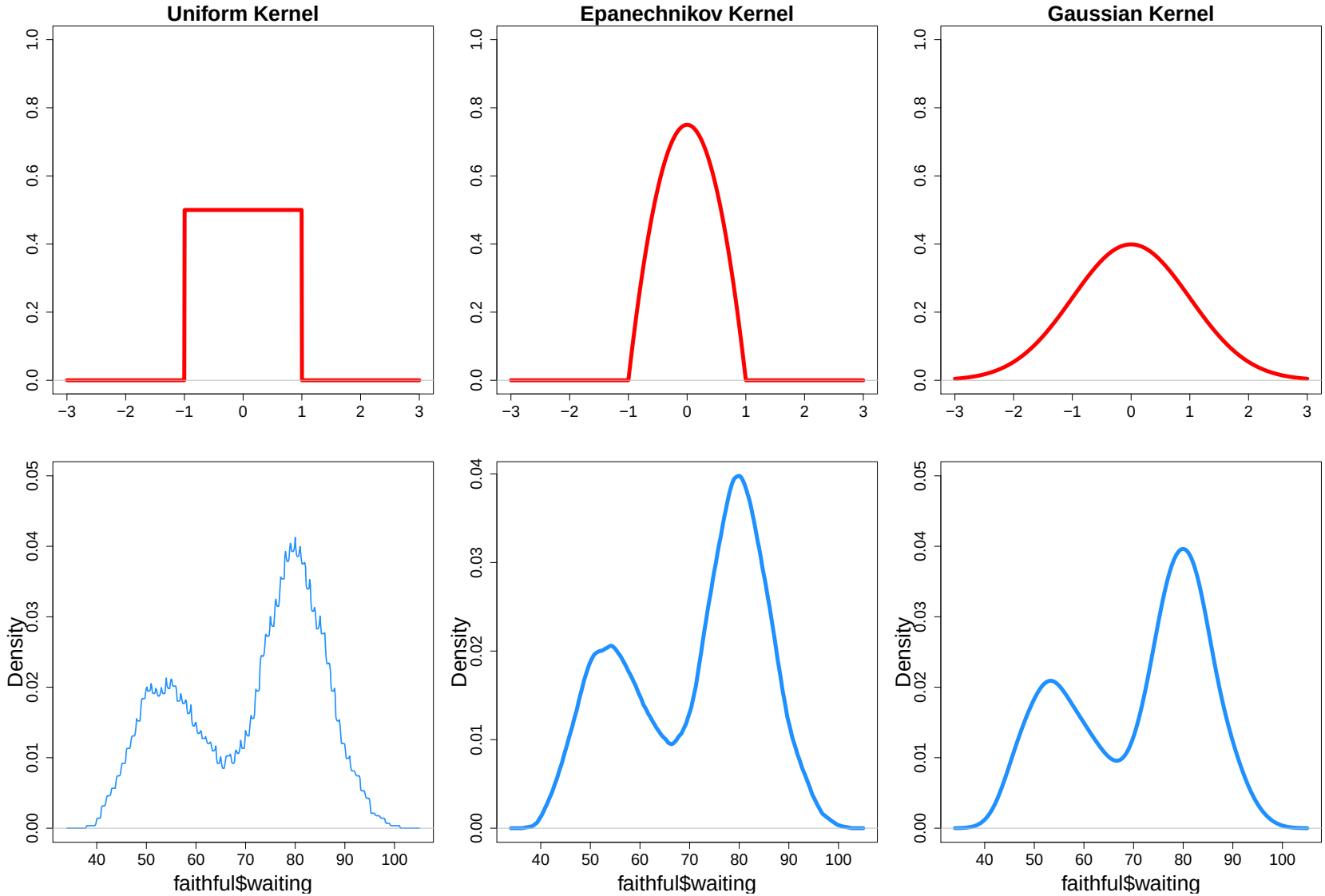
  where $K(x)$ is a function called the kernel function and $h > 0$ is a quantity called smoothing bandwidth that controls the amount of smoothing.
- Common choice of $K(x)$ includes the Gaussian $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, uniform $K(x) = \frac{1}{2} I(-1 \leq x \leq 1)$.
- The idea of KDE is: we smooth out each data point using the kernel function into small bumps and then we sum over all bumps to obtain a density estimate.

Nonparametric Approach: Kernel Density Estimation - 2



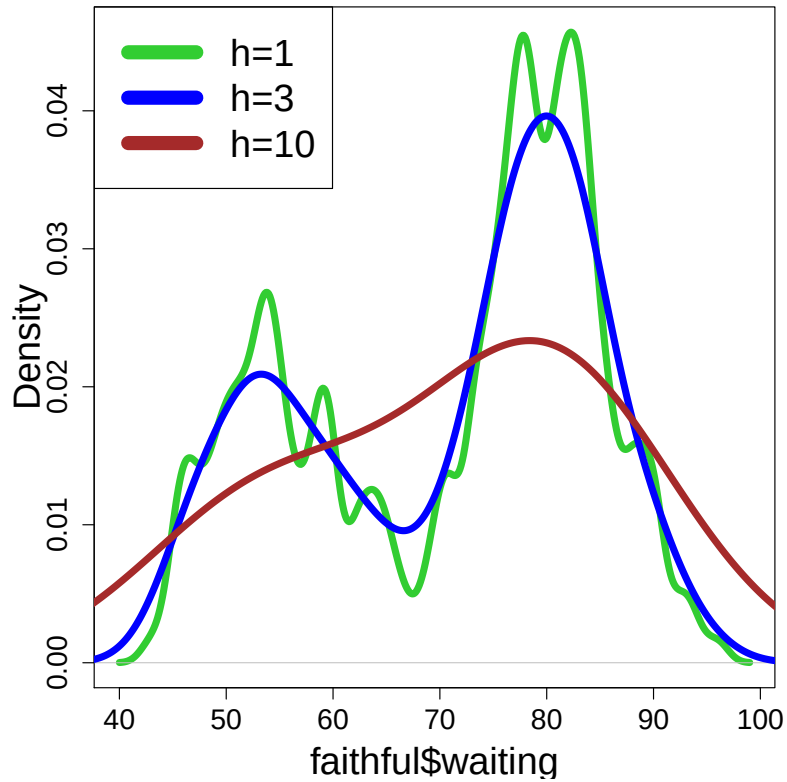
Black dots: locations of observations.
Purple bumps: the kernel function at each observation.
Brown curve: final density estimate from KDE.

Nonparametric Approach: Kernel Density Estimation - 3



The kernel function generally does not affect the density estimate too much.

Nonparametric Approach: Kernel Density Estimation - 4



The smoothing bandwidth often has a much stronger effect on the quality of estimation.

Nonparametric Approach: Kernel Density Estimation - 5

- We can also analyze the MSE of the KDE.
- When the smoothing bandwidth $h \approx 0$ and sample size n is large,

$$\textbf{bias}(\widehat{p}_{KDE}(x)) = O(h^2), \quad \text{Var}(\widehat{p}_{KDE}(x)) = O\left(\frac{1}{nh}\right).$$

Nonparametric Approach: Kernel Density Estimation - 5

- We can also analyze the MSE of the KDE.
- When the smoothing bandwidth $h \approx 0$ and sample size n is large,

$$\textbf{bias}(\widehat{p}_{KDE}(x)) = O(h^2), \quad \text{Var}(\widehat{p}_{KDE}(x)) = O\left(\frac{1}{nh}\right).$$

- Therefore, the MSE of the KDE is

$$\text{MSE}(\widehat{p}_{KDE}(x)) = O(h^4) + O\left(\frac{1}{nh}\right).$$

Nonparametric Approach: Kernel Density Estimation - 5

- We can also analyze the MSE of the KDE.
- When the smoothing bandwidth $h \approx 0$ and sample size n is large,

$$\textbf{bias}(\widehat{p}_{KDE}(x)) = O(h^2), \quad \text{Var}(\widehat{p}_{KDE}(x)) = O\left(\frac{1}{nh}\right).$$

- Therefore, the MSE of the KDE is

$$\text{MSE}(\widehat{p}_{KDE}(x)) = O(h^4) + O\left(\frac{1}{nh}\right).$$

- The optimal choice of h is $h \asymp n^{-1/5}$, leading to the optimal convergence rate

$$\text{MSE}^*(\widehat{p}_{KDE}(x)) = O(n^{-4/5}).$$

- Note that this convergence rate is faster than the rate of histogram $\text{MSE}^*(\widehat{p}_{hist}(x)) = O(n^{-2/3})$.

Nonparametric Approach: k-NN - 1

- The idea of the k-nearest neighbor (k-NN) can also be applied to density estimation.
- For a given point x (not necessarily an observation), its k-NN are the collection of observations whose distance to x is among the shortest k.

Nonparametric Approach: k-NN - 1

- The idea of the k-nearest neighbor (k-NN) can also be applied to density estimation.
- For a given point x (not necessarily an observation), its k-NN are the collection of observations whose distance to x is among the shortest k.
- Let $R_k(x)$ be the distance from x to its k-th nearest neighbor observation.
- The k-NN density estimation uses the following approximation:

$$\frac{k}{n} \approx P(X_{new} \in B(x, R_k(x))) \approx C_d R_k^d(x) \cdot p(x),$$

where d is the dimension of the data (often $d = 1, 2, 3$) and C_d is the size of d-dimensional unit ball and $p(x)$ is the PDF.

Nonparametric Approach: k-NN - 1

- The idea of the k-nearest neighbor (k-NN) can also be applied to density estimation.
- For a given point x (not necessarily an observation), its k-NN are the collection of observations whose distance to x is among the shortest k.
- Let $R_k(x)$ be the distance from x to its k-th nearest neighbor observation.
- The k-NN density estimation uses the following approximation:

$$\frac{k}{n} \approx P(X_{new} \in B(x, R_k(x))) \approx C_d R_k^d(x) \cdot p(x),$$

  where d is the dimension of the data (often $d = 1, 2, 3$) and C_d is the size of d-dimensional unit ball and $p(x)$ is the PDF.
- Thus, the k-NN density estimation is

$$\widehat{p}_{knn}(x) = \frac{k}{n} \cdot \frac{1}{C_d R_k^d(x)}.$$

Nonparametric Approach: k-NN - 2

- When $d = 1$, $C_d = 2$ so

$$\frac{k}{n} \approx 2R_k(x) \cdot p(x), \quad \widehat{p}_{knn}(x) = \frac{k}{n} \cdot \frac{1}{2R_k(x)}.$$

- When $d = 2$, $C_d = \pi$ so

$$\frac{k}{n} \approx \pi R_k^2(x) \cdot p(x), \quad \widehat{p}_{knn}(x) = \frac{k}{n} \cdot \frac{1}{\pi R_k^2(x)}.$$

- When $d = 3$, $C_d = \frac{4}{3}\pi$ so

$$\frac{k}{n} \approx \frac{4}{3}\pi R_k^3(x) \cdot p(x), \quad \widehat{p}_{knn}(x) = \frac{k}{n} \cdot \frac{3}{4\pi R_k^3(x)}.$$

- And again, there will be a bias-variance tradeoff; in the case of $d = 1$, we have:

$$MSE(\widehat{p}_{knn}(x)) = \underbrace{O\left(\left(\frac{k}{n}\right)^4\right)}_{\text{bias}} + \underbrace{O\left(\frac{1}{k}\right)}_{\text{variance}}.$$

Density Estimation: Inference - 1

- There are ways to do statistical inference for the PDF. We will comment on how to construct a CI.

Density Estimation: Inference - 1

- There are ways to do statistical inference for the PDF. We will comment on how to construct a CI.
- In the case of parametric approach, we can convert a CI of parameter into a CI of a PDF.
- In the case of nonparametric approach, generally we will use a *bootstrap* approach to construct a CI of a PDF.

Density Estimation: Inference - 1

- There are ways to do statistical inference for the PDF. We will comment on how to construct a CI.
- In the case of parametric approach, we can convert a CI of parameter into a CI of a PDF.
- In the case of nonparametric approach, generally we will use a *bootstrap* approach to construct a CI of a PDF.
- But note that there are two types of CI for a 'function'.

Density Estimation: Inference - 1

- There are ways to do statistical inference for the PDF. We will comment on how to construct a CI.
- In the case of parametric approach, we can convert a CI of parameter into a CI of a PDF.
- In the case of nonparametric approach, generally we will use a *bootstrap* approach to construct a CI of a PDF.
- But note that there are two types of CI for a 'function'.
- Pointwise CI: given a point x and confidence level $1-\alpha$, we construct an interval $C_{1-\alpha} = [\ell_{1-\alpha}, u_{1-\alpha}]$ from the data such that

$$P(\ell_{1-\alpha} \leq p(x) \leq u_{1-\alpha}) \approx 1-\alpha.$$

Density Estimation: Inference - 1

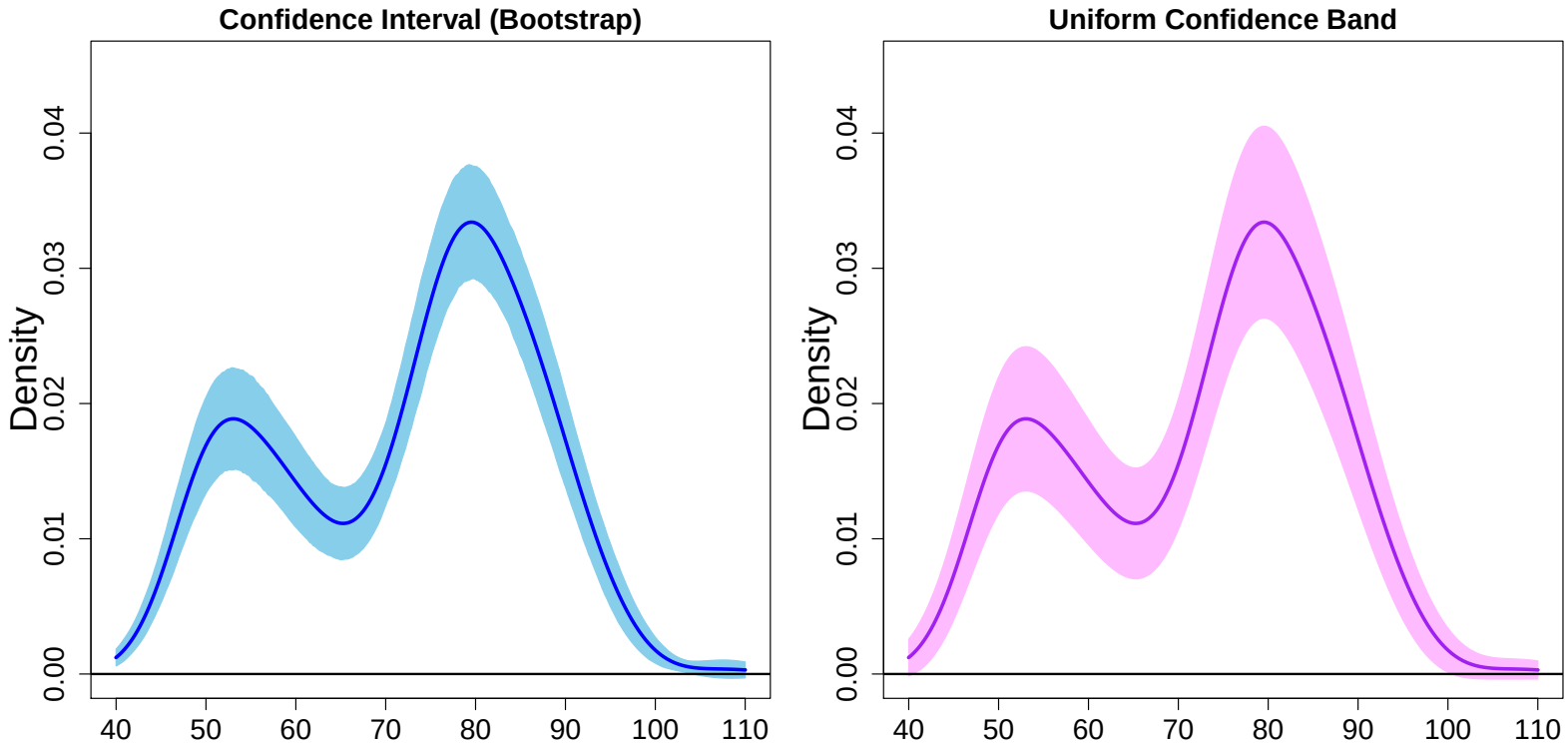
- There are ways to do statistical inference for the PDF. We will comment on how to construct a CI.
- In the case of parametric approach, we can convert a CI of parameter into a CI of a PDF.
- In the case of nonparametric approach, generally we will use a *bootstrap* approach to construct a CI of a PDF.
- But note that there are two types of CI for a 'function'.
- Pointwise CI: given a point x and confidence level $1-\alpha$, we construct an interval $C_{1-\alpha} = [\ell_{1-\alpha}, u_{1-\alpha}]$ from the data such that

$$P(\ell_{1-\alpha} \leq p(x) \leq u_{1-\alpha}) \approx 1-\alpha.$$

- Simultaneous CB (confidence band): given α, we construct a band $C_{1-\alpha}(x) = [L_{1-\alpha}(x), U_{1-\alpha}(x)]$ from the data such that

$$P(L_{1-\alpha}(x) \leq p(x) \leq U_{1-\alpha}(x) \text{ for all } x) \approx 1-\alpha.$$

Density Estimation: Inference - 2



Pointwise CI (left) and simultaneous CB (right)[1].

[1] A tutorial on this topic is in: **https://arxiv.org/abs/1704.03924**

Regression

Regression: Introduction

- Regression is an approach to study the relationship between a *response* variable Y and a *covariate* X.
- The covariate is also called a *feature*, a *predictor*, or an *independent variable*.
- Note that the covariate X can be multivariate.

Regression: Introduction

- Regression is an approach to study the relationship between a *response* variable Y and a *covariate* X.
- The covariate is also called a *feature*, a *predictor*, or an *independent variable*.
- Note that the covariate X can be multivariate.
- A traditional way to summarize the relationship via the *regression function*:

$$r(x) = \mathbb{E}(Y|X = x) = \int y \cdot f(y|x)dy.$$

- The goal of regression is to estimate $r(x)$ using the random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$.

Linear Regression - 1

- Linear regression is a parametric approach that models the function $r(x)$ as a linear function:

$$r(x) = \beta_0 + \beta_1 x.$$

- In many case, we will make further assumption on the noise and rewrite the linear model as

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{signal}} + \underbrace{\epsilon_i}_{\text{noise}},$$

where $\mathbb{E}(\epsilon_i|X_i) = 0$ and $\text{Var}(\epsilon_i|X_i) = \sigma^2$.

Linear Regression - 2

- In the linear regression model, there are two parameters: intercept β_0 and slope β_1.

Linear Regression - 2

- In the linear regression model, there are two parameters: intercept β_0 and slope β_1.
- To estimate them, a classical approach is the least squares (LS):

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 ,$$

  where the notation $\text{argmin}_{\beta_0,\beta_1}$ means finding the value of β_0, β_1 that minimizes the followings.
- You can solve the above LS criterion and find a closed form solution to the estimate:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}, \quad \widehat{\beta}_0 = \bar{Y}_n - \widehat{\beta}_1 \bar{X}_n.$$

Linear Regression - 3

- Using the LS estimator (LSE), we will predict the value of Y_i as

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i.$$

Linear Regression - 3

- Using the LS estimator (LSE), we will predict the value of Y_i as
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i.$$
- The difference between predicted and observed value is called *residual*
$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i.$$
- The *residual sums of squares* $RSS = \sum_{i=1}^{n} e_i^2$ measures how our estimate fits the data.

Linear Regression - 3

- Using the LS estimator (LSE), we will predict the value of Y_i as
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i.$$
- The difference between predicted and observed value is called *residual*
$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i.$$
- The *residual sums of squares* $RSS = \sum_{i=1}^{n} e_i^2$ measures how our estimate fits the data.
- You can interpret the LS approach as finding the best linear model to minimize RSS.
- Note that the noise level σ^2 can be estimated by $\widehat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$.

Linear Regression - 4

- The LSE has nice theoretical properties:

$$\textbf{bias}(\widehat{\beta}_0|X_1,\cdots,X_n)=0, \quad \textbf{bias}(\widehat{\beta}_1|X_1,\cdots,X_n)=0$$

$$\text{Var}(\widehat{\beta}_0|X_1,\cdots,X_n)=\frac{\sigma^2}{ns_X^2}\frac{1}{n}\sum_{i=1}^{n}X_i^2$$

$$\text{Var}(\widehat{\beta}_1|X_1,\cdots,X_n)=\frac{\sigma^2}{ns_X^2},$$

where $s_X^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$.

Linear Regression - 4

- The LSE has nice theoretical properties:

$$\textbf{bias}(\widehat{\beta}_0|X_1,\cdots,X_n)=0, \quad \textbf{bias}(\widehat{\beta}_1|X_1,\cdots,X_n)=0$$

$$\text{Var}(\widehat{\beta}_0|X_1,\cdots,X_n)=\frac{\sigma^2}{ns_X^2}\frac{1}{n}\sum_{i=1}^{n}X_i^2$$

$$\text{Var}(\widehat{\beta}_1|X_1,\cdots,X_n)=\frac{\sigma^2}{ns_X^2},$$

  where $s_X^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$.
- Moreover, central limit theorem implies that the LSE converges to a normal distribution under appropriate conditions.
- Thus, we can construct CI for β_0 and β_1 using the standard errors of $\widehat{\beta}_0, \widehat{\beta}_1$.

Linear Regression - 5

$$\text{Var}(\widehat{\beta}_0|X_1,\cdots,X_n) = \frac{\sigma^2}{ns_X^2}\frac{1}{n}\sum_{i=1}^{n}X_i^2$$

$$\longrightarrow SE(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n}X_i^2}{n}}$$

$$\text{Var}(\widehat{\beta}_1|X_1,\cdots,X_n) = \frac{\sigma^2}{ns_X^2}$$

$$\longrightarrow SE(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}.$$

Thus, a $1-\alpha$ CI will be

$$\widehat{\beta}_0 \pm z_{\alpha/2}SE(\widehat{\beta}_0), \quad \widehat{\beta}_1 \pm z_{\alpha/2}SE(\widehat{\beta}_1)$$

for β_0 and β_1 respectively.

Linear Regression: Multiple Covariates - 1

- In the case of the multiple covariates $x = (x_1, \cdots, x_p)$, the linear regression can be easily extended:

$$r(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

- Let $Y = (Y_1, \cdots, Y_n)$ be the vector of responses and

$$\mathbb{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ 1 & X_{2,1} & \cdots & X_{2,p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix}$$

be the $n \times (p+1)$ data matrix (each row is an observation).

- The multiple linear regression can be written as the follows:

$$Y = \mathbb{X}\beta + \epsilon,$$

where $\beta = (\beta_0, \cdots, \beta_p)$ is the parameter vector and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$ is the noise.

Linear Regression: Multiple Covariates - 2

- The LS method is to find

$$\widehat{\beta} = \text{argmax}_{\beta} \| Y - \mathbb{X}\beta \|^2$$

- And it has a closed form solution:

$$\widehat{\beta} = \left(\mathbb{X}^T \mathbb{X} \right)^{-1} \mathbb{X}^T Y.$$

Linear Regression: Multiple Covariates - 2

- The LS method is to find

$$\widehat{\beta} = \text{argmax}_{\beta} \| Y - \mathbb{X}\beta \|^2$$

- And it has a closed form solution:

$$\widehat{\beta} = \left(\mathbb{X}^T \mathbb{X} \right)^{-1} \mathbb{X}^T Y.$$

- The LSE has a nice property that

$$\widehat{\beta} \approx N \left(\beta, \sigma^2 \left(\mathbb{X}^T \mathbb{X} \right)^{-1} \right).$$

- Actually, you can show that the LSE is an unbiased estimator and the variance is $\sigma^2 \left(\mathbb{X}^T \mathbb{X} \right)^{-1}$. The above expression further suggests that we can use it to construct a CI for β.

Linear Regression: Remarks

- The linear regression is an important topic in statistics. It can be a course for an entire semester!
- You can search online to learn more about it.
- Here are a few key words related to it: ANOVA, R^2, outliers, leverage points.
- Note that the idea of LS approach can be applied to 'non-linear' model as well. For instance, we can model

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 exp(-\beta_4 x)$$

and apply LS approach to find the parameters.

Logistic Regression - 1

- In some special case, the response Y may take only two possible values, say 0 and 1.
- For instance, our response Y may be the type of galaxy and $Y = 1$ if it is a spiral galaxy and $Y = 0$ if it is an elliptical galaxy.

Logistic Regression - 1

- In some special case, the response Y may take only two possible values, say 0 and 1.
- For instance, our response Y may be the type of galaxy and $Y = 1$ if it is a spiral galaxy and $Y = 0$ if it is an elliptical galaxy.
- In this special case,

$$\mathbb{E}(Y|X = x) = P(Y = 1|X = x) = r(x)$$

is a probability.

Logistic Regression - 1

- In some special case, the response Y may take only two possible values, say 0 and 1.
- For instance, our response Y may be the type of galaxy and $Y = 1$ if it is a spiral galaxy and $Y = 0$ if it is an elliptical galaxy.
- In this special case,

$$\mathbb{E}(Y|X = x) = P(Y = 1|X = x) = r(x)$$

  is a probability.
- If we naively model it as a linear function, then we may obtain a negative probability or a probability greater than 1, both are not reasonable.
- The logistic regression uses a smart way to model such a probability.

Logistic Regression - 2

- The quantity $r(x) = P(Y = 1|X = x) \in [0, 1]$ because it is a probability.

Logistic Regression - 2

- The quantity $r(x) = P(Y = 1|X = x) \in [0, 1]$ because it is a probability.
- We first consider the odds:

$$o(x) = \frac{r(x)}{1 - r(x)} = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} \in [0, \infty).$$

Logistic Regression - 2

- The quantity $r(x) = P(Y = 1|X = x) \in [0, 1]$ because it is a probability.
- We first consider the odds:

$$o(x) = \frac{r(x)}{1 - r(x)} = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} \in [0, \infty).$$

- However, the odds is not symmetric with respect to x, so we take logarithm of it:

$$\ell(x) = \log o(x) = \log \left(\frac{r(x)}{1 - r(x)} \right) \in (-\infty, \infty).$$

  This quantity is more symmetric – it can take values anywhere in the real line.

Logistic Regression - 2

- The quantity $r(x) = P(Y = 1|X = x) \in [0, 1]$ because it is a probability.
- We first consider the odds:

$$o(x) = \frac{r(x)}{1 - r(x)} = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} \in [0, \infty).$$

- However, the odds is not symmetric with respect to x, so we take logarithm of it:

$$\ell(x) = \log o(x) = \log \left(\frac{r(x)}{1 - r(x)} \right) \in (-\infty, \infty).$$

  This quantity is more symmetric – it can take values anywhere in the real line.
- The logistic regression models the log odds as a linear function of x:

$$\ell(x) = \beta_0 + \beta_1 x.$$

Logistic Regression - 3

- The model

$$\ell(x) = \log o(x) = \log\left(\frac{r(x)}{1 - r(x)}\right) = \beta_0 + \beta_1 x$$

leads to the following form of $r(x)$:

$$r(x) = P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

- Using this probability model, we can then apply the MLE to find β_0 and β_1.

Logistic Regression - 3

- The model

$$\ell(x) = \log o(x) = \log\left(\frac{r(x)}{1-r(x)}\right) = \beta_0 + \beta_1 x$$

leads to the following form of $r(x)$:

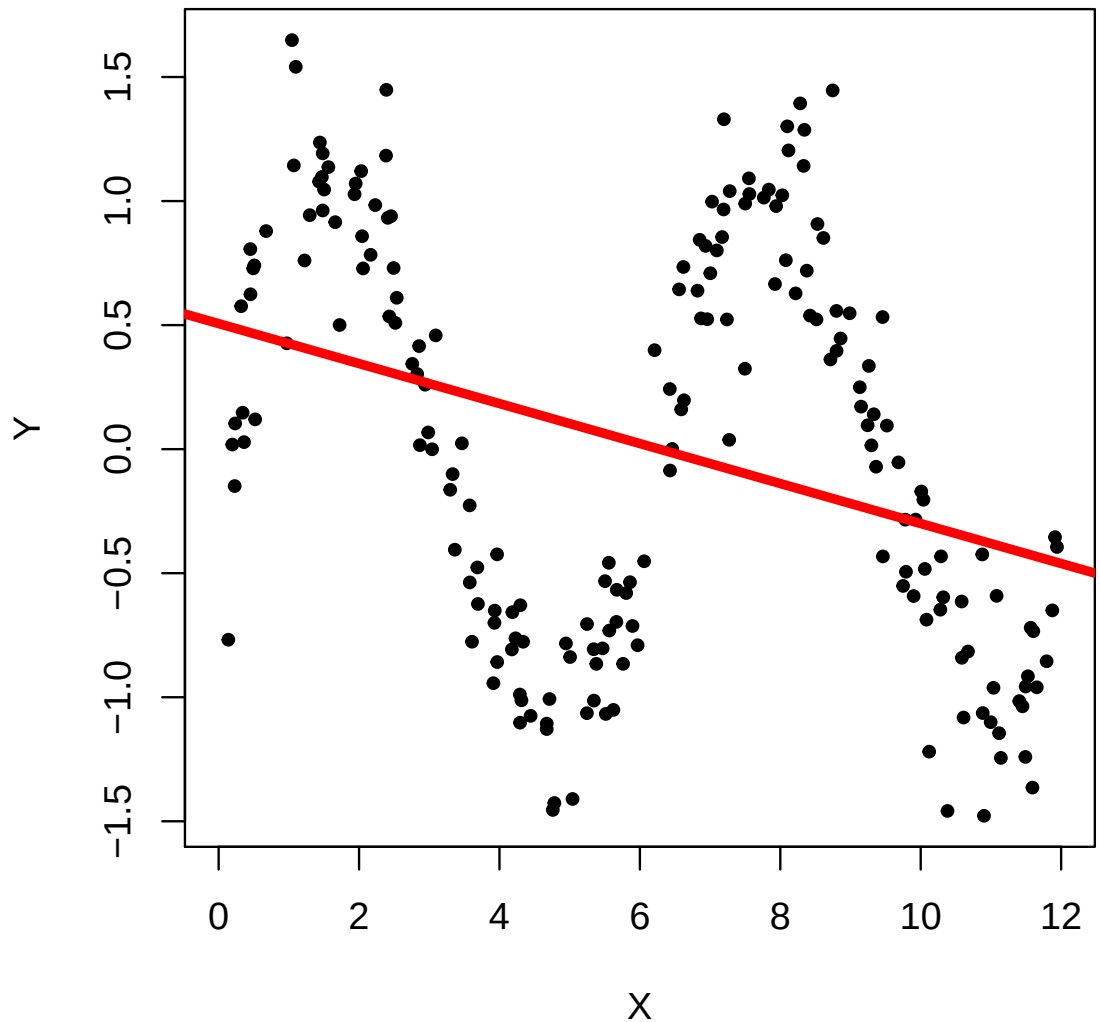
$$r(x) = P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

- Using this probability model, we can then apply the MLE to find β_0 and β_1.
- Note that the MLE does not have a closed form solution but one can find it using numerical methods such a gradient descent approach.

Nonparametric Regression

- A problem of parametric regression is: the actual regression function may not have the desired form.
- When the parametric form is mis-specified, the result can be very bad.
- Nonparametric regression attempts to directly estimate the regression function without assuming a parametric form of it.
- We will talk about three popular methods: regressogram (binning), kernel regression, and spline approach.

Nonparametric Regression: Example



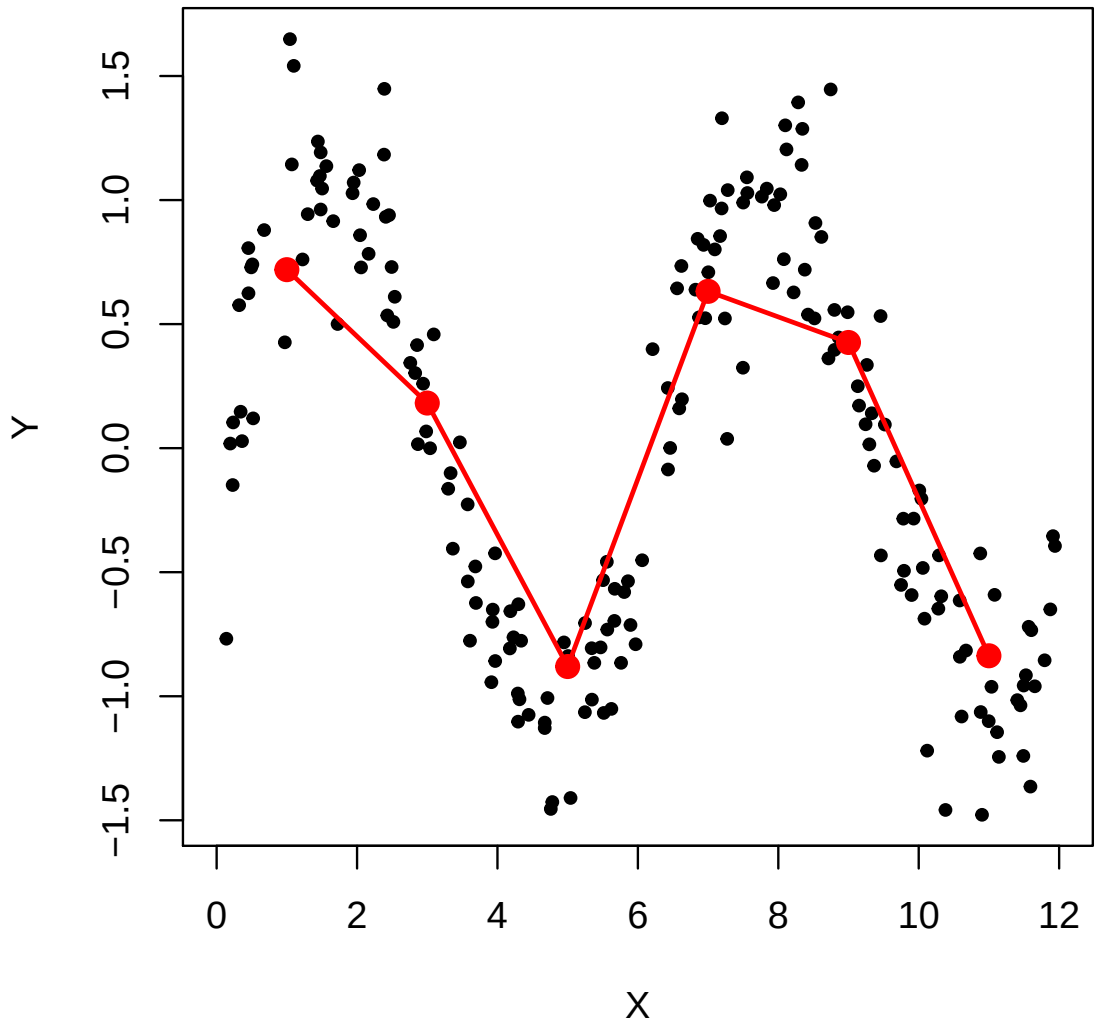
Regressogram (Binning)

- The regressogram (binning) might be one of the most popular regression approach but very few people know its name.
- The regressogram = regression + histogram.

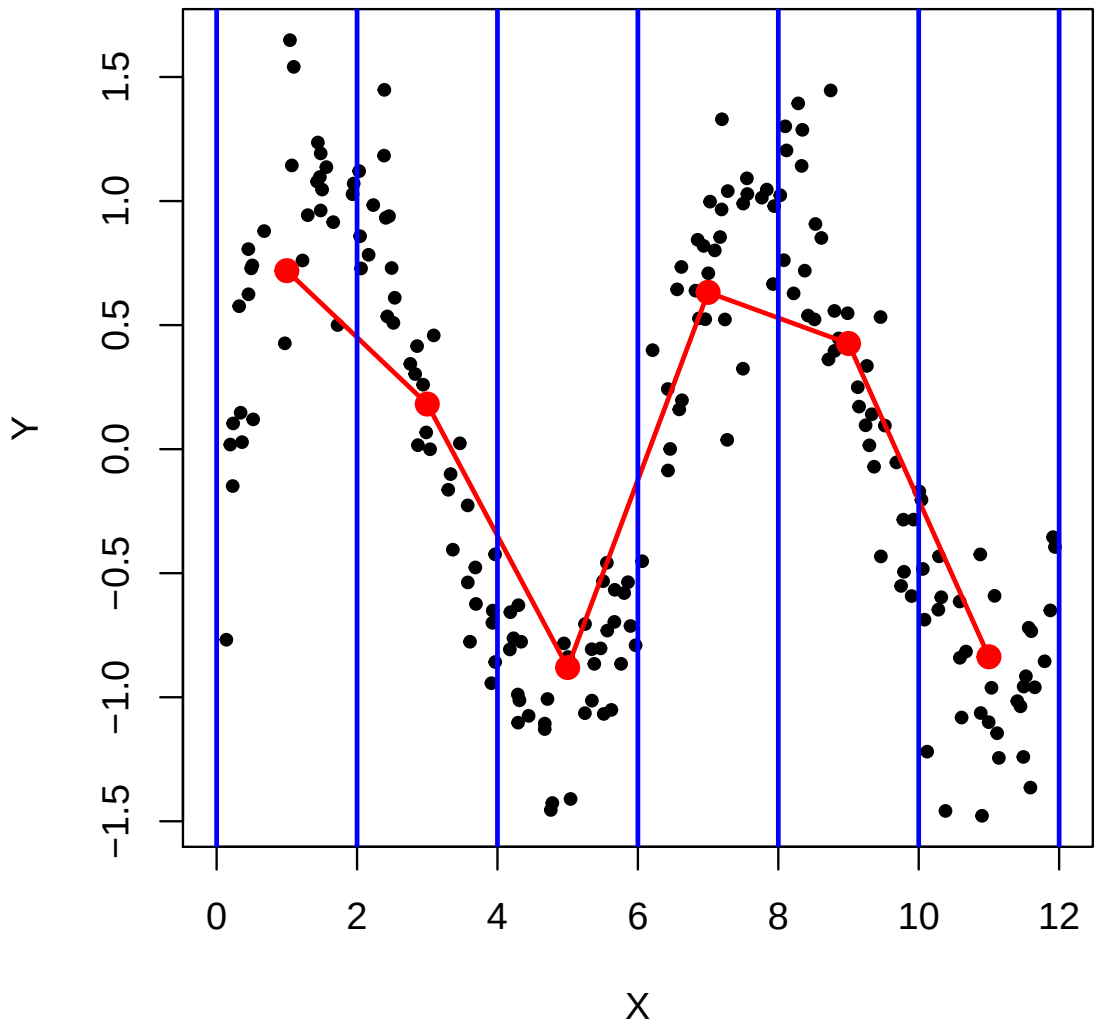
Regressogram (Binning)

- The regressogram (binning) might be one of the most popular regression approach but very few people know its name.
- The regressogram = regression + histogram.
- The idea is: we bin the range of covariates into several intervals.
- We then use the average of the responses for observations within the same interval as the estimated value.

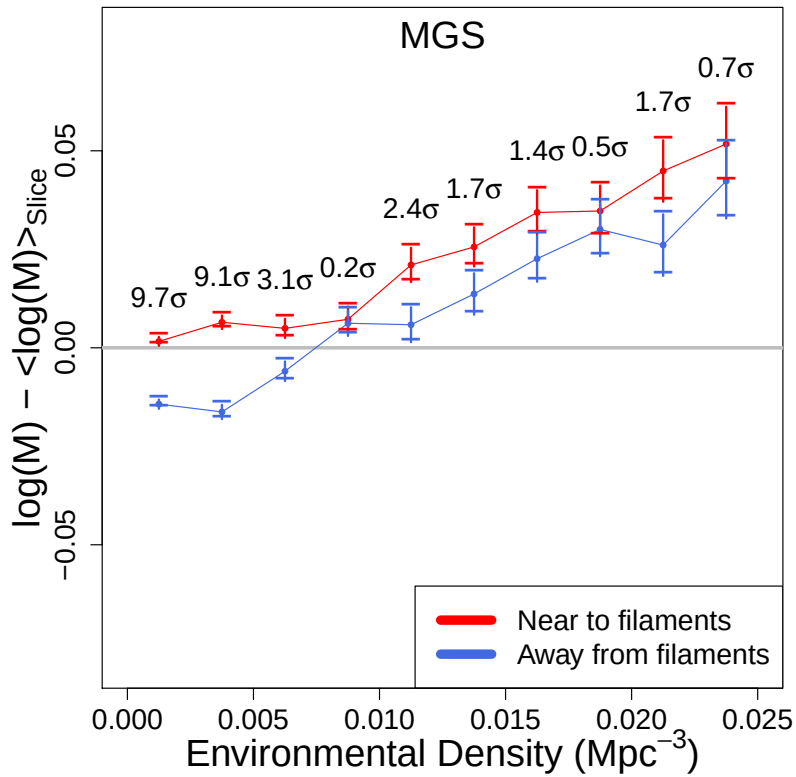
Regressogram: Example - 1



Regressogram: Example - 1



Regressogram: Example - 2



Kernel Regression - 1

- The kernel regression is another nonparametric regression estimator.
- The kernel regression uses an estimator of the form

$$\begin{aligned}\widehat{r}_{ker}(x) &= \sum_{i=1}^{n} W_i(x) Y_i \\ &= \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}\end{aligned}$$

  where

$$W_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}.$$

- The function $K(x)$ is again the kernel function we talk about in the KDE.

Kernel Regression - 2

- The quantity

$$W_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}.$$

  satisfies $\sum_{i=1}^{n} W_i(x) = 1$ and $W_i(x) \geq 0$.
- Namely, it behaves like a *weight* of each Y_i.

Kernel Regression - 2

- The quantity

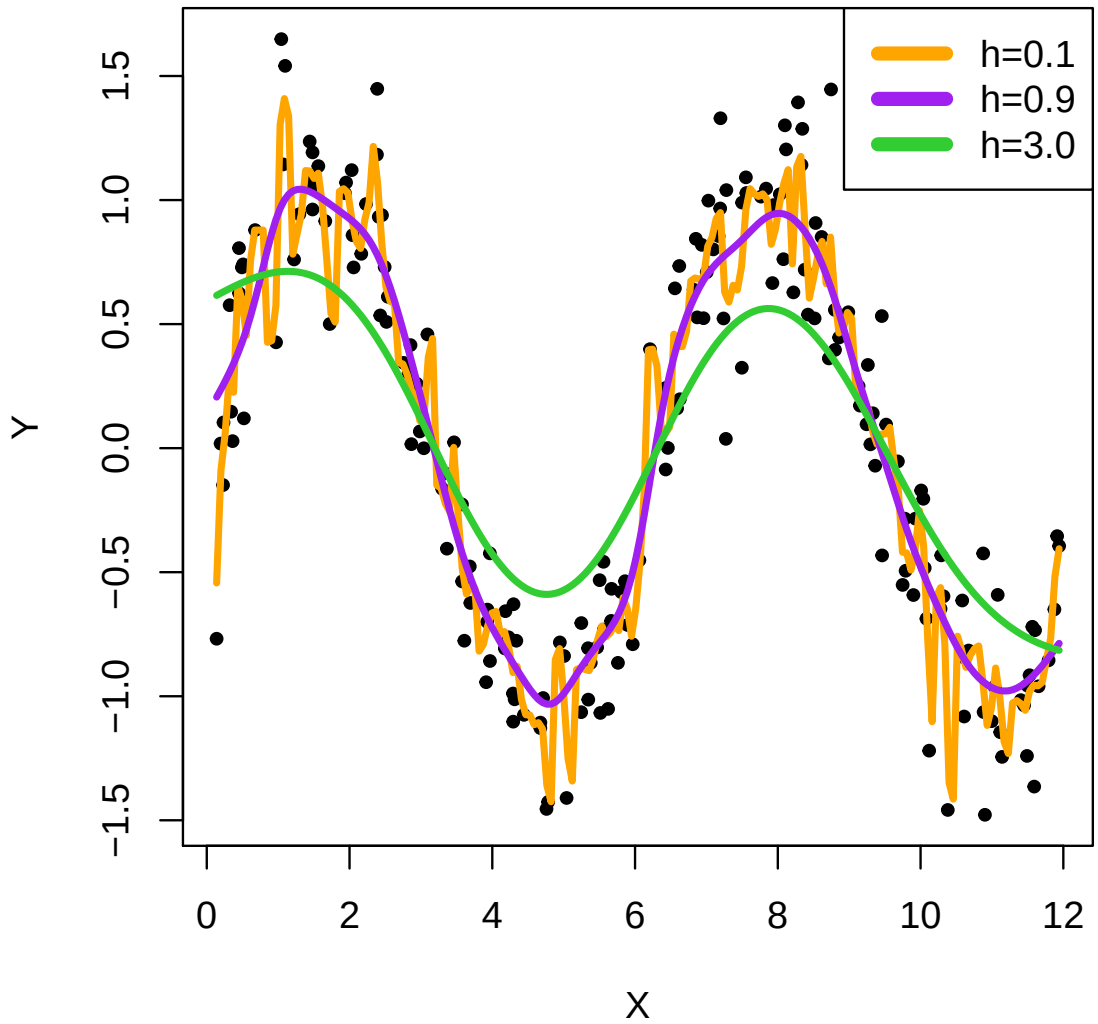
$$W_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}.$$

  satisfies $\sum_{i=1}^{n} W_i(x) = 1$ and $W_i(x) \geq 0$.
- Namely, it behaves like a *weight* of each Y_i.
- The estimator $\widehat{r}_{ker}(x) = \sum_{i=1}^{n} W_i(x) Y_i$ can be interpreted as follows.
- To estimate the regression function at $X = x$, we use a *weighted average* of all responses such that observations close to x will be given a higher weight ($W_i(x)$ will be large if X_i is close to x).

Kernel Regression - 2

- The quantity
$$W_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}.$$
satisfies $\sum_{i=1}^{n} W_i(x) = 1$ and $W_i(x) \geq 0$.
- Namely, it behaves like a *weight* of each Y_i.
- The estimator $\widehat{r}_{ker}(x) = \sum_{i=1}^{n} W_i(x) Y_i$ can be interpreted as follows.
- To estimate the regression function at $X = x$, we use a *weighted average* of all responses such that observations close to x will be given a higher weight ($W_i(x)$ will be large if X_i is close to x).
- The kernel function determines how we are going to give weights to the nearby points.
- The smoothing bandwidth h controls the range of influence from each observation (the degree of smoothing).

Kernel Regression: Example



Cross-Validation Approach - 1

- How can we choose the smoothing bandwidth?
- There are many ways to do that but a simple principle is: we want to choose it to optimize the *prediction accuracy*.

Cross-Validation Approach - 1

- How can we choose the smoothing bandwidth?
- There are many ways to do that but a simple principle is: we want to choose it to optimize the *prediction accuracy*.
- For an estimator $\widehat{m}$, a prediction accuracy is

$$R = \mathbb{E}(|Y_{new} - \widehat{m}(X_{new})|^2),$$

  where (X_{new}, Y_{new}) is a new observation.
- In the case of kernel regression, the prediction accuracy depends on h so

$$R(h) = \mathbb{E}(|Y_{new} - \widehat{m}_{ker}(X_{new})|^2).$$

Cross-Validation Approach - 1

- How can we choose the smoothing bandwidth?
- There are many ways to do that but a simple principle is: we want to choose it to optimize the *prediction accuracy*.
- For an estimator $\widehat{m}$, a prediction accuracy is

$$R = \mathbb{E}(|Y_{new} - \widehat{m}(X_{new})|^2),$$

  where (X_{new}, Y_{new}) is a new observation.
- In the case of kernel regression, the prediction accuracy depends on h so

$$R(h) = \mathbb{E}(|Y_{new} - \widehat{m}_{ker}(X_{new})|^2).$$

- We want to pick the smoothing bandwidth

$$h^* = \text{argmin}_h R(h).$$

Cross-Validation Approach - 2

- The quantity $R(h) = \mathbb{E}(|Y_{new} - \widehat{m}_{ker}(X_{new})|^2)$ is unknown to us – we need to estimate it.
- However, $R(h)$ involves *two* expectations: one for the estimator $\widehat{m}_{ker}$ and the other for the *new observation*.

Cross-Validation Approach - 2

- The quantity $R(h) = \mathbb{E}(|Y_{new} - \widehat{m}_{ker}(X_{new})|^2)$ is unknown to us – we need to estimate it.
- However, $R(h)$ involves *two* expectations: one for the estimator $\widehat{m}_{ker}$ and the other for the *new observation*.
- We know that we can use sample average to estimate the expectation.

Cross-Validation Approach - 2

- The quantity $R(h) = \mathbb{E}(|Y_{new} - \widehat{m}_{ker}(X_{new})|^2)$ is unknown to us – we need to estimate it.
- However, $R(h)$ involves *two* expectations: one for the estimator $\widehat{m}_{ker}$ and the other for the *new observation*.
- We know that we can use sample average to estimate the expectation.
- Thus, a simple approach to consistently estimate $R(h)$ is to split the data into two parts: we use one part to construct $\widehat{m}_{ker}$ and the other part of the data as the *new observations*.
- This idea is called *data splitting*.

Cross-Validation Approach - 2

- The quantity $R(h) = \mathbb{E}(|Y_{new} - \widehat{m}_{ker}(X_{new})|^2)$ is unknown to us – we need to estimate it.
- However, $R(h)$ involves *two* expectations: one for the estimator $\widehat{m}_{ker}$ and the other for the *new observation*.
- We know that we can use sample average to estimate the expectation.
- Thus, a simple approach to consistently estimate $R(h)$ is to split the data into two parts: we use one part to construct $\widehat{m}_{ker}$ and the other part of the data as the *new observations*.
- This idea is called *data splitting*.
- The *cross-validation* is a modified approach of data splitting that repeat the splitting procedure multiple times and then use the average as the final estimate of $R(h)$.

Cross-Validation Approach - 3

- In practice, we will split the data into several subsets and treat part of them as *training set* (the part of data used to computed the estimator $\widehat{m}_{ker}$) and the other part as *validation set* (the set treated as future observations).

Cross-Validation Approach - 3

- In practice, we will split the data into several subsets and treat part of them as *training set* (the part of data used to computed the estimator $\widehat{m}_{ker}$) and the other part as *validation set* (the set treated as future observations).
- We often choose one subset as the validation set and the others as the training set.

Cross-Validation Approach - 3

- In practice, we will split the data into several subsets and treat part of them as *training set* (the part of data used to computed the estimator $\widehat{m}_{ker}$) and the other part as *validation set* (the set treated as future observations).
- We often choose one subset as the validation set and the others as the training set.
- After evaluating the prediction risk, then we use another subset as the validation set and others as the training set.

Cross-Validation Approach - 3

- In practice, we will split the data into several subsets and treat part of them as *training set* (the part of data used to computed the estimator $\widehat{m}_{ker}$) and the other part as *validation set* (the set treated as future observations).
- We often choose one subset as the validation set and the others as the training set.
- After evaluating the prediction risk, then we use another subset as the validation set and others as the training set.
- We repeat this process until all subsets have been used as the validation set.

Cross-Validation Approach - 3

- In practice, we will split the data into several subsets and treat part of them as *training set* (the part of data used to computed the estimator $\widehat{m}_{ker}$) and the other part as *validation set* (the set treated as future observations).
- We often choose one subset as the validation set and the others as the training set.
- After evaluating the prediction risk, then we use another subset as the validation set and others as the training set.
- We repeat this process until all subsets have been used as the validation set.
- We then use the average of all these prediction risks as an estimate of the prediction.

Cross-Validation Approach - 3

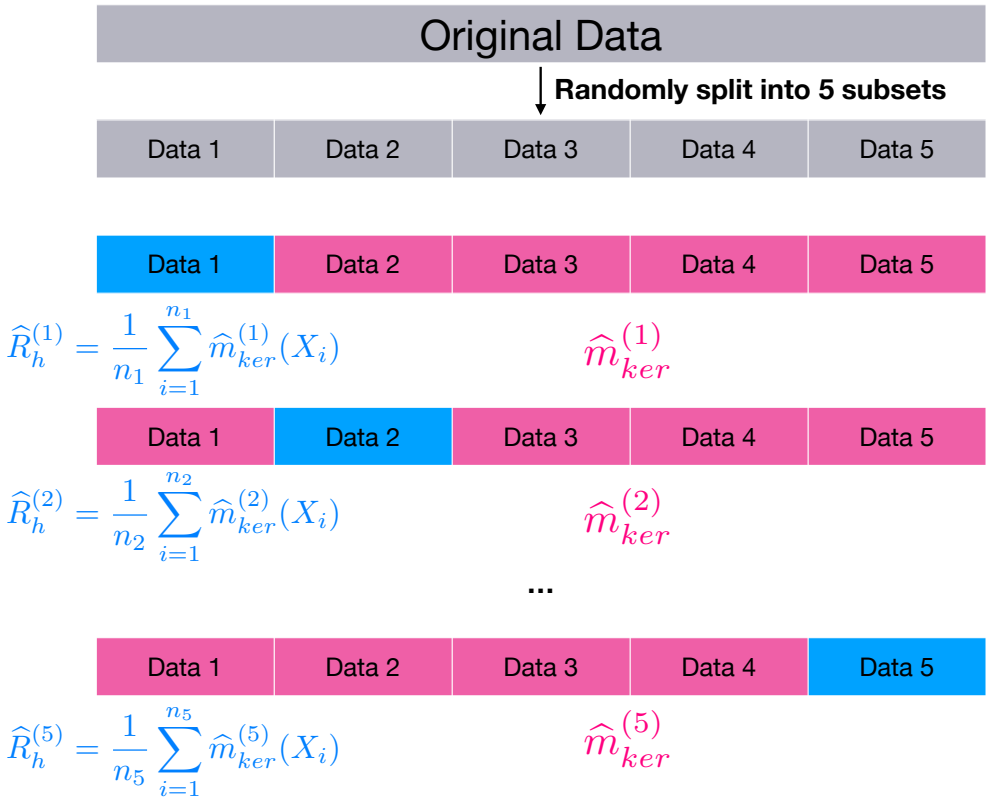
- In practice, we will split the data into several subsets and treat part of them as *training set* (the part of data used to computed the estimator $\widehat{m}_{ker}$) and the other part as *validation set* (the set treated as future observations).
- We often choose one subset as the validation set and the others as the training set.
- After evaluating the prediction risk, then we use another subset as the validation set and others as the training set.
- We repeat this process until all subsets have been used as the validation set.
- We then use the average of all these prediction risks as an estimate of the prediction.
- We often repeat the above procedure several times and take the total average as the the final risk estimate.

Cross-Validation Approach - 3

- In practice, we will split the data into several subsets and treat part of them as *training set* (the part of data used to computed the estimator $\widehat{m}_{ker}$) and the other part as *validation set* (the set treated as future observations).
- We often choose one subset as the validation set and the others as the training set.
- After evaluating the prediction risk, then we use another subset as the validation set and others as the training set.
- We repeat this process until all subsets have been used as the validation set.
- We then use the average of all these prediction risks as an estimate of the prediction.
- We often repeat the above procedure several times and take the total average as the the final risk estimate.
- Note: if we split the data into k subset, we call this approach the k-fold cross validation.

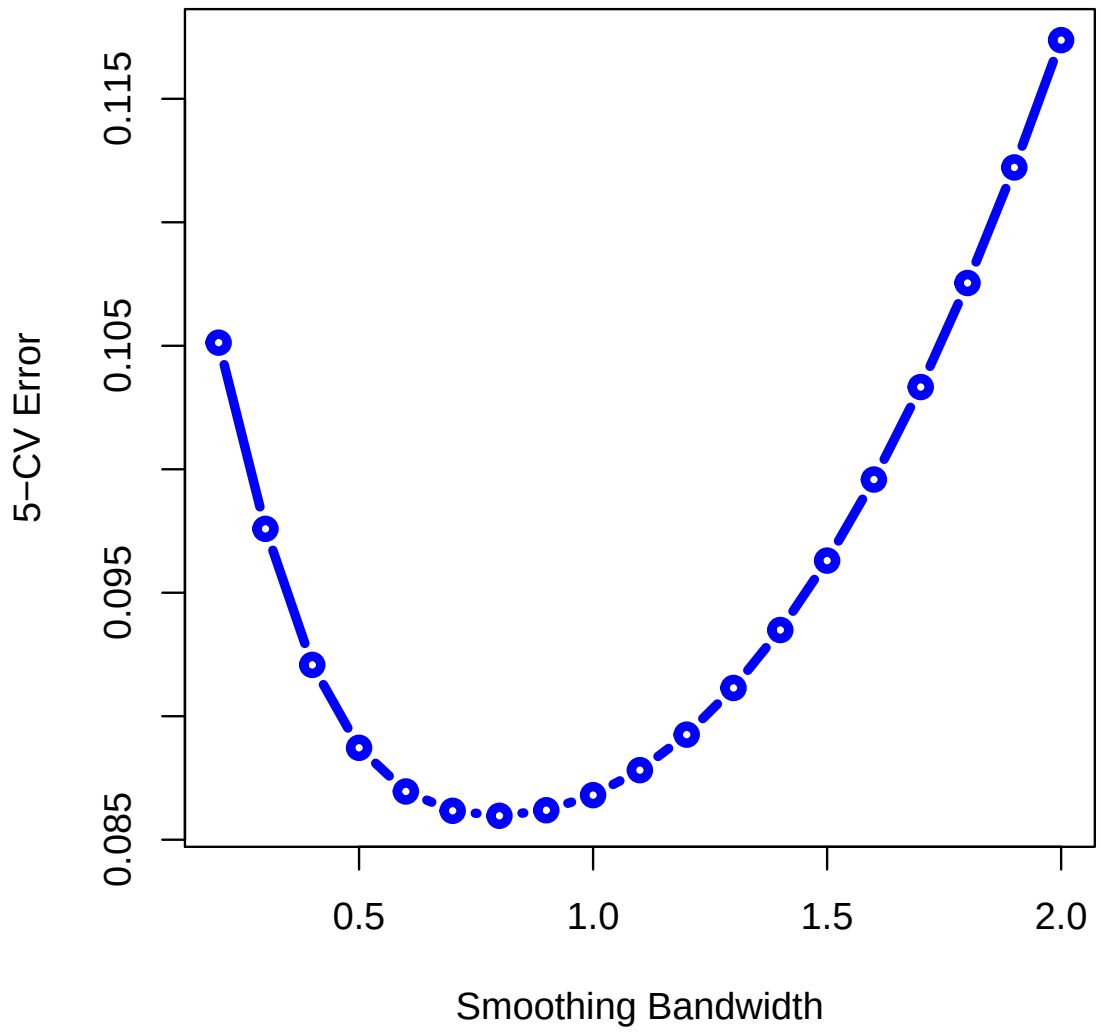
Cross-Validation Approach: 5-fold CV

- Here is an illustration for 5-fold CV:



- Validation Set. Training Set.
- We use the average $\widehat{R}(h) = \frac{1}{5}\sum_{\ell=1}^{5} \widehat{R}^{(\ell)}(h)$ as a risk estimate.
- In practice, we repeat this procedure for several times and take the total average of them.

5-fold Cross-Validation: Example



Overfitting and underfitting

- Why we cannot use the same data twice for both training set and validation set?

Overfitting and underfitting

- Why we cannot use the same data twice for both training set and validation set?
- From a theoretical point of view, this leads to a *biased* estimator of the prediction risk.
- Sometimes people call this overfitting – a more complex model you are using, you may seemly fit the data better but actually the prediction error gets worse.

Overfitting and underfitting

- Why we cannot use the same data twice for both training set and validation set?
- From a theoretical point of view, this leads to a *biased* estimator of the prediction risk.
- Sometimes people call this overfitting – a more complex model you are using, you may seemly fit the data better but actually the prediction error gets worse.
- As an extreme example: consider $h \approx 0$, then the kernel regression passes every data point. If we use the training set as the validation set, this leads to a prediction risk $= 0$!

Overfitting and underfitting

- Why we cannot use the same data twice for both training set and validation set?
- From a theoretical point of view, this leads to a *biased* estimator of the prediction risk.
- Sometimes people call this overfitting – a more complex model you are using, you may seemly fit the data better but actually the prediction error gets worse.
- As an extreme example: consider $h \approx 0$, then the kernel regression passes every data point. If we use the training set as the validation set, this leads to a prediction risk $= 0$!
- Note that: an opposite case is called underfitting – you fit a too easy model so it cannot capture the complicated structure of the data. When we apply the linear regression to the example of a wave-form data, we suffer from underfitting.

Spline Approach - 1

- Spline approach is a *penalized regression* method.
- The goal is to find a function f such that it fits the data well and f is smooth.

[2] https://en.wikipedia.org/wiki/Smoothing_spline

Spline Approach - 1

- Spline approach is a *penalized regression* method.
- The goal is to find a function f such that it fits the data well and f is smooth.
- To quantify smoothness, the spline approach places a penalty on the curvature – the second derivative of f.

[2] https://en.wikipedia.org/wiki/Smoothing_spline

Spline Approach - 1

- Spline approach is a *penalized regression* method.
- The goal is to find a function f such that it fits the data well and f is smooth.
- To quantify smoothness, the spline approach places a penalty on the curvature – the second derivative of f.
- In more details, the spline approach attempts to find $\widehat{f}_{sp}$ such that

$$\widehat{f}_{sp} = argmin_f \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda \int_{X_{min}}^{X_{max}} |f''(s)|^2 ds,$$

where $\lambda > 0$ is a parameter determines how smooth we want.

[2] https://en.wikipedia.org/wiki/Smoothing_spline

Spline Approach - 1

- Spline approach is a *penalized regression* method.
- The goal is to find a function f such that it fits the data well and f is smooth.
- To quantify smoothness, the spline approach places a penalty on the curvature – the second derivative of f.
- In more details, the spline approach attempts to find $\widehat{f}_{sp}$ such that

$$\widehat{f}_{sp} = argmin_f \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda \int_{X_{min}}^{X_{max}} |f''(s)|^2 ds,$$

  where $\lambda > 0$ is a parameter determines how smooth we want.
- There are some smart ways[2] to find such a minimal function $\widehat{f}_{sp}$.

[2] https://en.wikipedia.org/wiki/Smoothing_spline

Spline Approach - 2

$$\widehat{f}_{sp} = argmin_f \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2}_{\text{fitting to the data}} + \underbrace{\lambda \int_{X_{min}}^{X_{max}} |f''(s)|^2 ds}_{\text{smoothness penalty}},$$

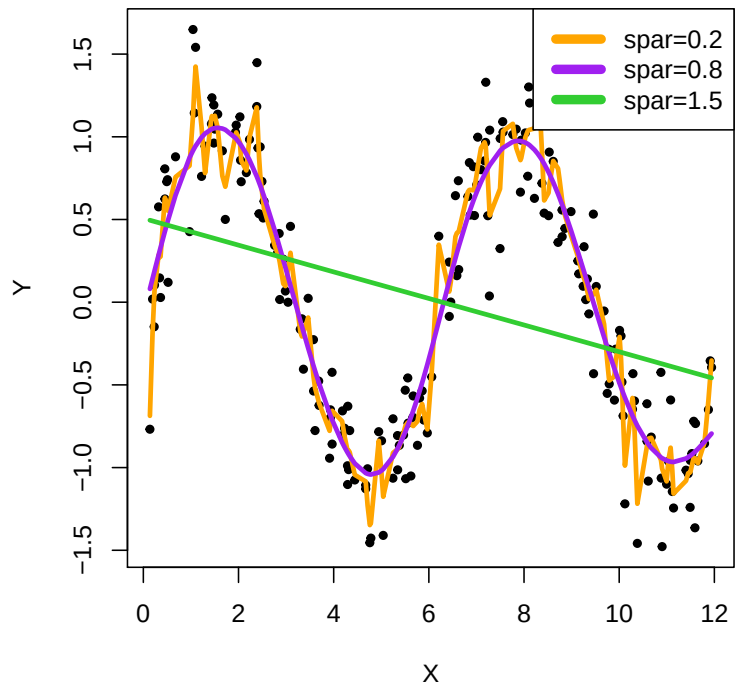
- A large λ leads to a smooth function $\widehat{f}_{sp}$.
- A small λ yields a more wiggly function.

Spline Approach - 2

$$\widehat{f}_{sp} = argmin_f \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2}_{\text{fitting to the data}} + \underbrace{\lambda \int_{X_{min}}^{X_{max}} |f''(s)|^2 ds}_{\text{smoothness penalty}},$$

- A large λ leads to a smooth function $\widehat{f}_{sp}$.
- A small λ yields a more wiggly function.
- The choice of λ determines how we want to weight the fitting quality and smoothness.
- We often use cross-validation to choose λ.

Spline Approach: Example



spar: a quantity in R related to λ.

Useful References

- **All of statistics: a concise course in statistical inference.** Larry Wasserman. Springer Science & Business Media, 2013.
- **All of nonparametric statistics.** Larry Wasserman. Springer, 2006.
- **Multivariate density estimation: theory, practice, and visualization.** David Scott. John Wiley & Sons, 2015.
- **Applied Linear Regression.** Sanford Weisberg. Wiley Series in Probability and Statistics, 2005.