Statistical Machine Learning

Lecture 7: Model Inference and Averaging

W.Q.Cui Research Group

Department of Statistics and Finance
University of Science and Technology of China

2018 Autumn

Contents

Outline

- Model Inference
  - — Maximum likelihood inference
    - EM Algorithm
  - — Bayesian inference
    - Gibbs Sampling
  - — Bootstrap
- Model Averaging and improvement
  - — Bagging
  - — Bumping

Basic Concepts

- Statistical inference
  — Using data to infer the distribution that generated the data
  - We Observe $X_1, \cdots, X_n \sim F$
  - We want to infer (or estimate or learn) F or some feature of F such as its mean.
- Statistical model
  — A set of distributions (or a set of densities) ξ
  - Parametric model
  - Non parametric model

Statistical Model

Parametric Model

- A set ξ that can be parameterized by a finite number of parameters
- E.g. Assume the data come from a normal distribution, the model is

$$\xi = \{f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{\pi}} exp(-\frac{1}{2\sigma^2}(x-\mu)^2), \mu \in R, \sigma > 0\}$$

- A parametric model takes the form

$$\xi = \{f(x;\theta) : \theta \in \Theta\}$$

Statistical Model

Non-Parametric Model

- A set ξ that cannot be parameterized by a finite number of parameters
- E.g. Assume the data comes from $\xi' = \{all_C DF's\}$

Probability density function, PDF,

$$f(x) : Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

Cumulative density function,CDF,

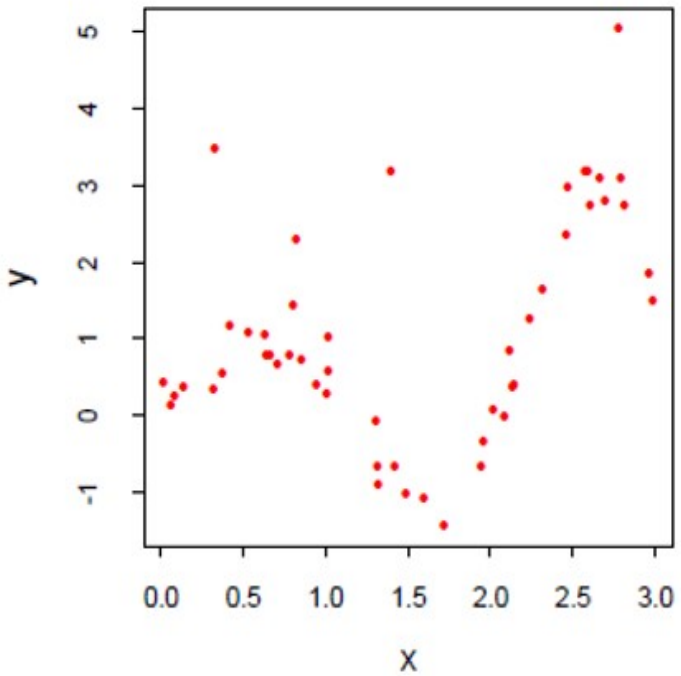
$$F(x) : Pr(X \leq x) = \int_0^x f(s)ds$$

Smoothing Example

Training Set:

$$Z = z_1, z_2, \cdots, z_N,$$

$$z_i = (x_i, y_i), N = 50$$



Smooth Splines

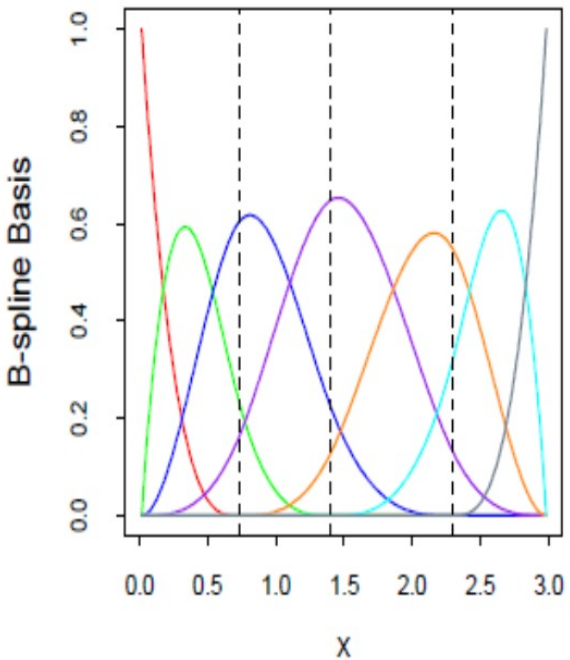


Figure: Cubic Spline

$$\mu(x) = \sum_{j=1}^{J} h_j(x)\beta_j$$

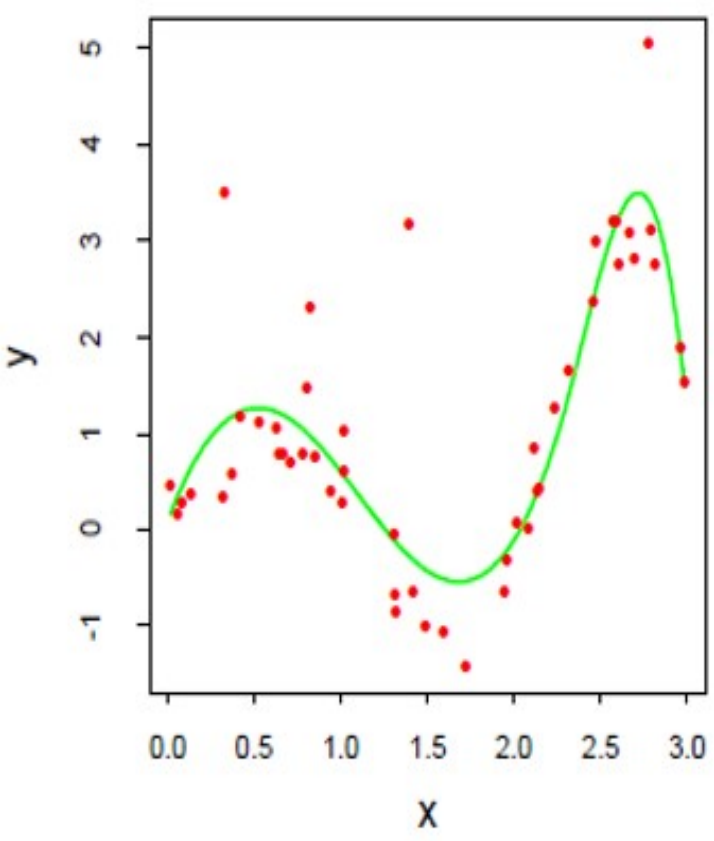
$$\hat{\beta} = (H^T H)^{-1} H^T y$$

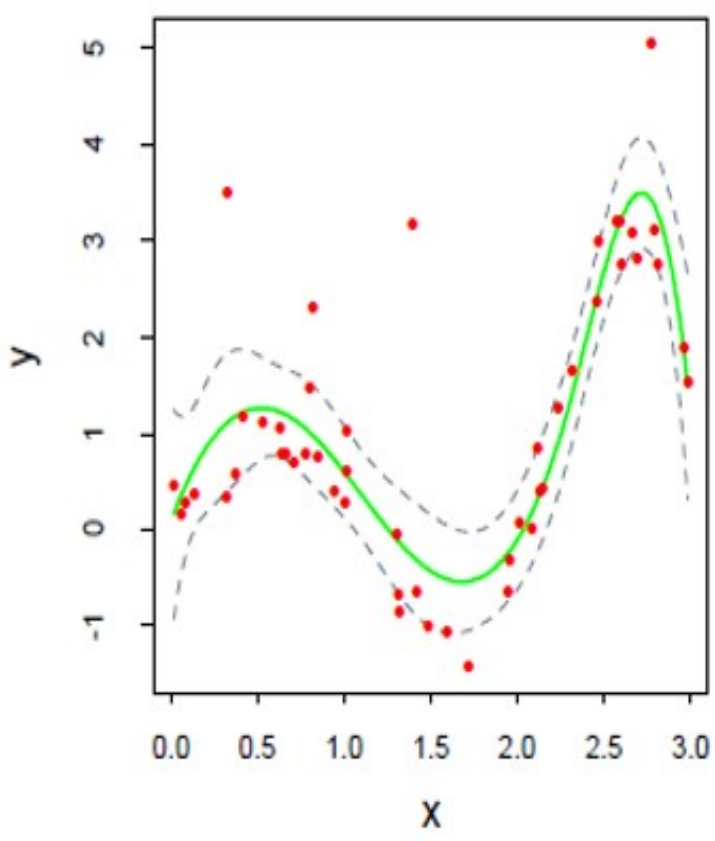
$$\hat{\sigma}^2 = \sum_{i=1}^{N} (y_i - \hat{\mu}(x_i))^2 / N$$

$$\widehat{Var}(\hat{\beta}) = (H^T H)^{-1} \hat{\sigma}^2$$

$$\widehat{se}[\hat{\mu}(x)] = [h(x)^T (H^T H)^{-1} h(x)]^{1/2} \hat{\sigma}$$

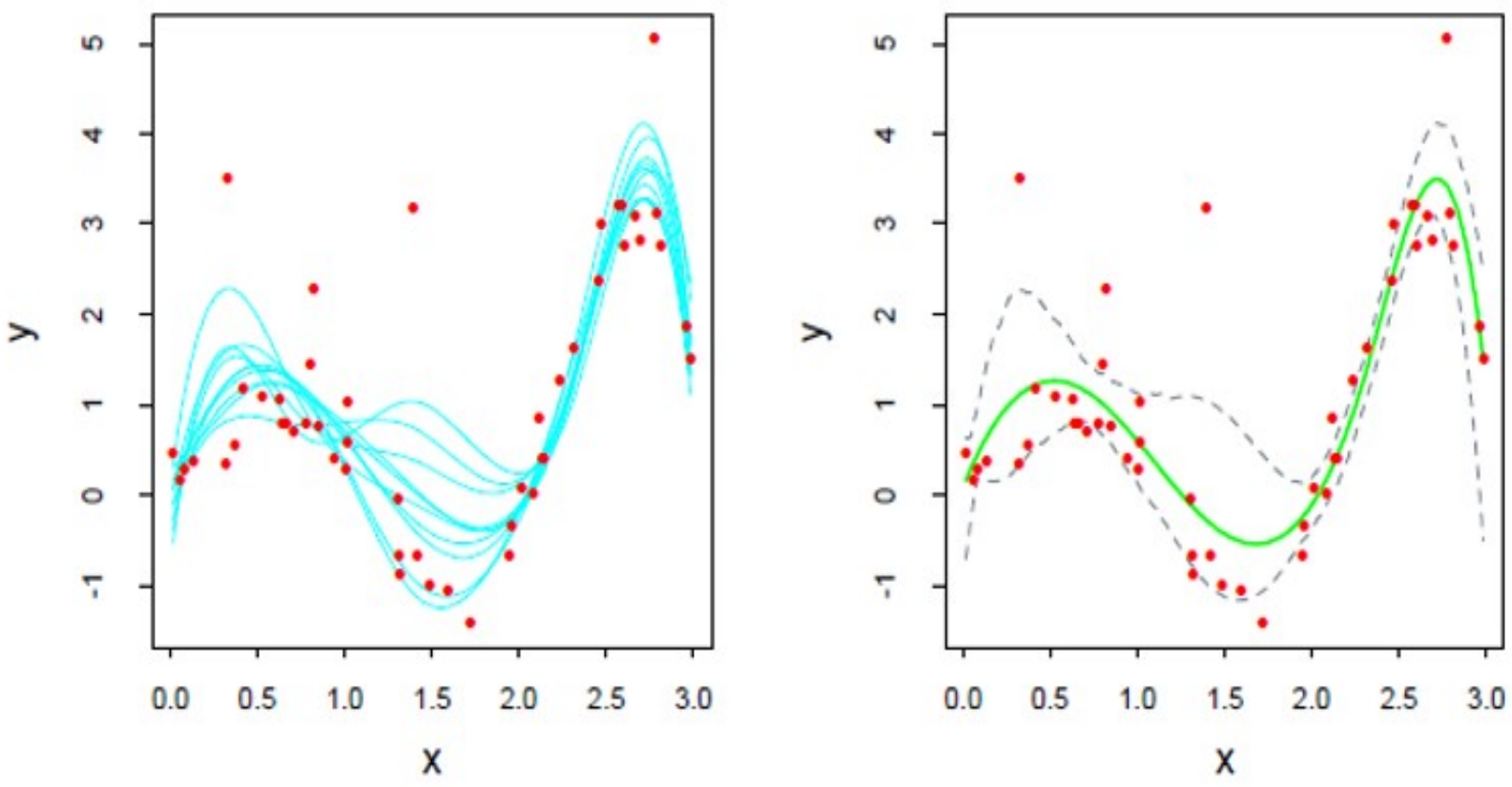
Smooth Splines Result





Nonparametric Bootstrap

Nonparametric bootstrap: replacement sampling.



Parametric Bootstrap

Parametric bootstrap: use special parametric model to generate new dataset.

$$y_i^* = \hat{\mu}(x_i) + \epsilon_i^*, \epsilon_i^* \sim N(0, \hat{\sigma}^2), i = 1, 2, \cdots, N.$$
$$\hat{\mu}^*(x) = h(x)^T (H^T H)^{-1} H^T y^*$$
$$\hat{\mu}^*(x) \sim N(\hat{\mu}(x), h(x)^T (H^T H)^{-1} h(x) \hat{\sigma}^2)$$

Parametric Inference

- Parametric Models:

$$\xi = \{f(x;\theta) : \theta \in \Theta\}$$

- The Problem of Inference
  $\rightarrow$ problem of estimating the parameter θ
- Method
  - Maximum Likelihood Inference
  - Bayesian Inference

An Example of MLE

- Suppose you have $x_1, x_2, \cdots, x_n \sim N(\mu, \sigma^2)$
- But you don't know μ of σ^2
- MLE: For which $\theta = (\mu, \sigma^2)$ is $x_1, x_2, \cdots, x_n$ most likely?

$$log Pr(x_1, x_2, \cdots, x_n | \mu, \sigma^2) = -n(log\pi + \frac{1}{2} log\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \mu_{mle} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \Rightarrow \sigma^2_{mle} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{mle})^2$$

A General MLE strategy

Suppose $\theta = (\theta_1, \theta_2, \cdots, \theta_n)^T$ is a vector of parameters.
Task: Find MLE θ for the likelihood function
$L(\theta; X) = Pr(x_1, x_2, \cdots, x_n|\theta)$

- Write Log-likelihood function: $\ell = log(L(\theta; X))$
- Work out $\frac{\partial \ell}{\partial \theta}$
- Solve the set of simultaneous equations

$$\frac{\partial \ell}{\partial \theta_1} = 0, \frac{\partial \ell}{\partial \theta_2} = 0, \cdots, \frac{\partial \ell}{\partial \theta_n} = 0$$

- Check you are at a maximum

Properties of MLE

Sampling distributions of the maximum likelihood estimator has a limiting normal distribution.

$$\hat{\theta} \to N(\theta_0, i(\theta_0)^{-1})$$

where θ_0 is true value of θ,
Fisher Information :

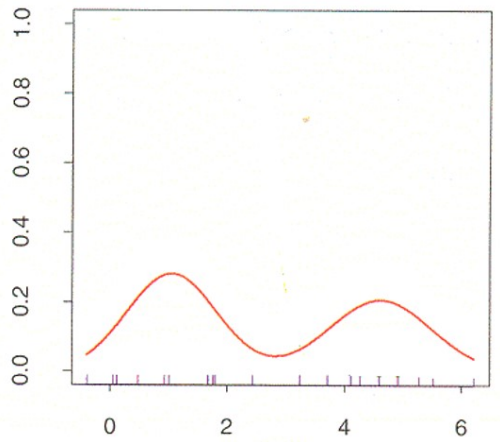
$$i(\theta) = E_\theta[I(\theta)]$$

Information Matrix :

$$I(\theta) = -\sum_{i=1}^{N} \frac{\partial^2 \ell(\theta; z_i)}{\partial\theta\partial\theta^T}$$

An Example for EM Algorithm

Model Y as a mixture of two normal distribution



$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2$$
$$Y_1 \sim N(\mu_1, \sigma_1^2)$$
$$Y_2 \sim N(\mu_2, \sigma_2^2)$$

where $\Delta \in \{0, 1\}$ with $Pr(\Delta = 1) = \pi$.

The parameters are

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

The log-likelihood based on the N training cases is

$$\ell(\theta; Z) = \sum_{i=1}^{N} log[(1 - \pi)\psi_{\theta_1}(y_i) + \pi\psi_{\theta_2}(y_i)]$$

sum of terms is inside the logarithm $\Rightarrow$ difficult to maximize it

An Example for EM Algorithm

Consider unobserved latent variables Δ_i:

$\Delta_i = 1$ while Y_i comes from model 2; otherwise from model 1.

If we know the values of Δ_i, then

$$\ell(\theta; Z) = \sum_{i=1}^{N} [(1 - \Delta_i) log\psi_{\theta_1}(y_i) + \Delta_i log\psi_{\theta_2}(y_i)]$$
$$+ \sum_{i=1}^{N} [(1 - \Delta_i) log(1 - \pi) + \Delta_i log\pi]$$

An Example for EM Algorithm

- Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma_1}^2, \hat{\mu}_2, \hat{\sigma_2}^2, \hat{\pi}$
- Expectation Step: compute

$$\hat{\gamma}_i = \frac{\hat{\pi}\psi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\psi_{\hat{\theta}_2}(y_i) + \hat{\pi}\psi_{\hat{\theta}_2}(y_i)}, i = 1, 2, \cdots, N$$

- Maximization Step: compute the values for the parameters.$\hat{\mu}_1, \hat{\sigma_1}^2, \hat{\mu}_2, \hat{\sigma_2}^2, \hat{\pi}$ which can maximize the log-likelihood given $\hat{\gamma}$
- Iterate steps 2 and 3 until convergence.

An Example for EM Algorithm

Table: Selected iterations of the EM algorithm for mixture example.

Iteration	$\hat{\pi}$
1	0.485
5	0.493
10	0.523
15	0.544
20	0.546

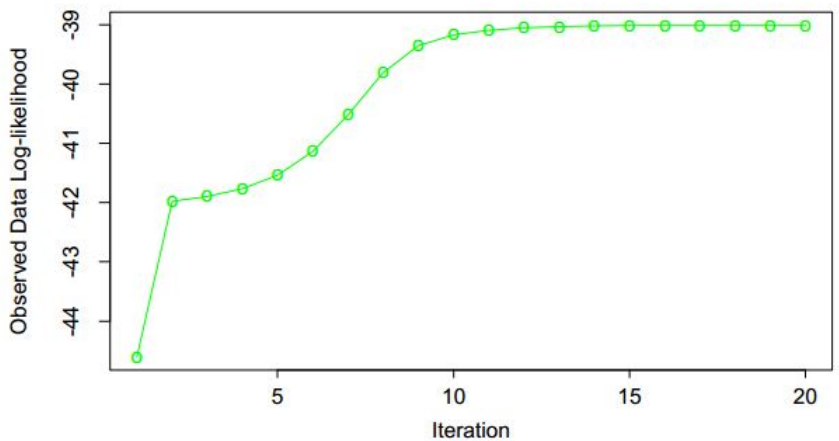


Figure: EM algorithm: observed data log-likelihood as a function of the iteration number.

Bayesian Inference

- Prior (knowledge before we see the data): $Pr(\theta)$
- Sampling model: $Pr(Z|\theta)$
- After observing data Z, we update our beliefs and form the posterior distribution

$$Pr(\theta|Z) = \frac{Pr(Z|\theta)Pr(\theta)}{\int Pr(Z|\theta)Pr(\theta)d\theta} = \frac{L_n(\theta)Pr(\theta)}{\int L_n(\theta)Pr(\theta)d\theta} \propto L_n(\theta)Pr(\theta)$$

Doesn't it cause a problem to throw away the constant?

We can always recover it, since $\int Pr(\theta|Z)d\theta = 1$

Posterior is proportional to likelihood times prior!

Prediction Using Inference

- Task: predict the values of a future observation z^{new}
- Bayesian Approach:

$$Pr(z^{new}|Z) = \int Pr(z^{new}|\theta)Pr(\theta|Z)d\theta$$

- Maximum likelihood approach $Pr(z^{new}|\hat{\theta})$

The Smoothing Example

$$\beta \sim N(0, \tau\Sigma)$$

$$K(x, x') = cov[\mu(x), \mu(x')] = \tau \cdot h(x)^T \Sigma h(x')$$

The posterior distribution for β is also Gaussian, with mean and covariance

$$E(\beta|\mathbf{Z}) = \left(H^T H + \frac{\sigma^2}{\tau}\Sigma^{-1}\right)^{-1} H^T y$$

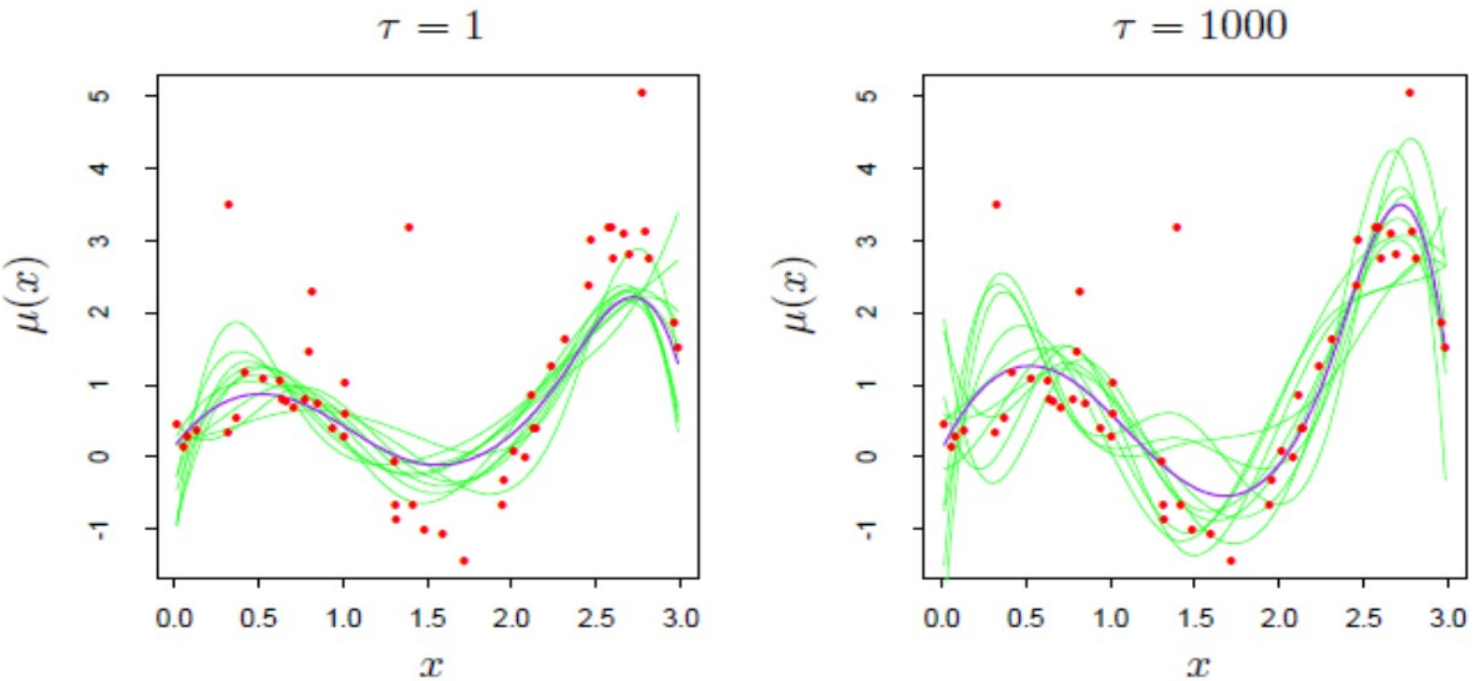
$$Cov(\beta|\mathbf{Z}) = \left(H^T H + \frac{\sigma^2}{\tau}\Sigma^{-1}\right)^{-1} \sigma^2$$

with the corresponding posterior values for $\mu(x)$,

$$E(\mu(x)|\mathbf{Z}) = h(x)^T \left(H^T H + \frac{\sigma^2}{\tau}\Sigma^{-1}\right)^{-1} H^T y$$

$$Cov[\mu(x), \mu(x')|\mathbf{Z}] = h(x)^T \left(H^T H + \frac{\sigma^2}{\tau}\Sigma^{-1}\right)^{-1} h(x')\sigma^2$$

The Smoothing Example



when $\tau \to \inf$, β is non-information prior and the posterior distribution is proportion to likelihood. The result is consistent with the maximum likelihood.

MCMC

General Problem: evaluating $E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$ can be difficult, where $\pi(\theta) = Pr(\theta|Z)$

However, if we can draw samples

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)} \sim \pi(\theta)$$

then we can estimate

$$E_\pi[h(\theta)] \approx \bar{h}_N = \frac{1}{N}\sum_{t=1}^{N} h(\theta^{(t)})$$

This is Monte Carlo (MC)integration.

MCMC

- A stochastic process is an indexed random variable $X^{(t)}$ where t maybe time and X is a random variable.
- A Markov chain is generated by sampling

$$X^{(t+1)} \sim p(x|X^{(t)}), t = 1, 2, \cdots$$

where p is the transition kernel.So, $X^{(t+1)}$ depends only on $X^{(t)}$, not on $X^{(0)}, X^{(1)}, \cdots, X^{(t-1)}$

- As $t \to \inf$, the Markov chain converges to its stationary distribution.

MCMC

Problem:

How do we construct a Markov chain whose stationary distribution is our target distribution,$\pi(\theta)$?

This is called Markov chain Monte Carlo (MCMC)

Two key objectives:

- Generate a sample from a joint probability distribution $\pi(\theta) = \pi(\theta_1, \cdots, \theta_k)$
- Estimate expectations using generated sample averages(i.e. doing MC integration)

Gibbs Sampling

- Purpose: Draw from a Joint Distribution

$$\theta = (\theta_1, \cdots, \theta_k);\ Target\ \pi(\theta)$$

- Method: Iterative Conditional Sampling

$$\forall i,\ Draw\ \theta_i \sim \pi(\theta_i|\theta_{[-i]})$$

Gibbs Sampling

- Suppose that $\theta = (\theta_1, \cdots, \theta_k)$
- Sample or update in turn:

$$\begin{aligned}
\theta_1^{(t+1)} &\sim \pi(\theta_1^{(t)} | \theta_2^{(t)}, \theta_3^{(t)}, \cdots, \theta_k^{(t)}) \\
\theta_2^{(t+1)} &\sim \pi(\theta_2^{(t)} | \theta_1^{(t)}, \theta_3^{(t)}, \cdots, \theta_k^{(t)}) \\
&\vdots \\
\theta_k^{(t+1)} &\sim \pi(\theta_1^{(t)} | \theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{k-1}^{(t)})
\end{aligned}$$

Always use the most recent values!

An Example for Conditional Sampling

- Target distribution:

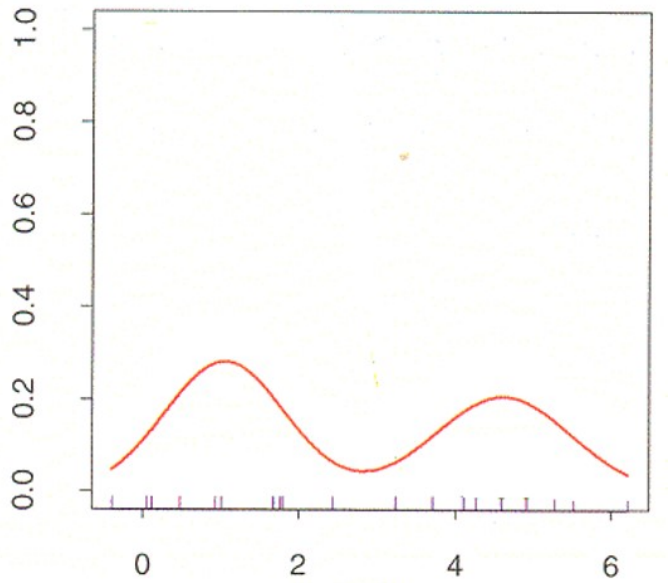
$$f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, x = 0, 1, \cdots, n; 0 \leq y \leq 1$$

- How to draw samples?

$$x \sim f(x|y) \Rightarrow Binomial(n, y)$$
$$y \sim f(y|x) \Rightarrow Beta(x+\alpha, n-x+\beta)$$

Recall: Same Example for EM

Model Y as a mixture of two normal distribution



$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2$$
$$Y_1 \sim N(\mu_1, \sigma_1^2)$$
$$Y_2 \sim N(\mu_2, \sigma_2^2)$$

where $\Delta \in \{0, 1\}$ with $Pr(\Delta = 1) = \pi$.

For simplicity, assume the parameters are $\theta = (\mu_1, \mu_2)$

Comparison between EM and Gibbs Sampling

Gibbs

- **Step 1**: Take initial guesses for the parameters $\theta^{(0)} = \{\mu_1^{(0)}, \mu_2^{(0)}\}$
- **Step 2**: Repeat for $t = 1, 2, \cdots$
  1. For $i = 1, 2, \cdots, N$ generate $\Delta_i^{(t)} \in \{0, 1\}$ with

$$Pr(\Delta_i = 1|\hat{\theta}, Z) = \frac{\hat{\pi}\psi_{\hat{\theta}_2^{t-1}}(y_i)}{(1-\hat{\pi})\psi_{\hat{\theta}_1^{(t-1)}}(y_i) + \hat{\pi}\psi_{\hat{\theta}_2^{(t-1)}}(y_i)}$$

  2. Generate $\mu_1^{(t)} \sim N(\hat{\mu}_1, \hat{\sigma}_1^2), \mu_2^{(t)} \sim N(\hat{\mu}_2, \hat{\sigma}_2^2)$
- **Step 3**: Continue step 2 until the joint distribution of $(\Delta^{(t)}, \mu_1^{(t)}, \mu_2^{(t)})$ doesn't change

Comparison between EM and Gibbs Sampling

EM

- **Step 1**: Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$
- **Step 2**: Expectation Step: compute

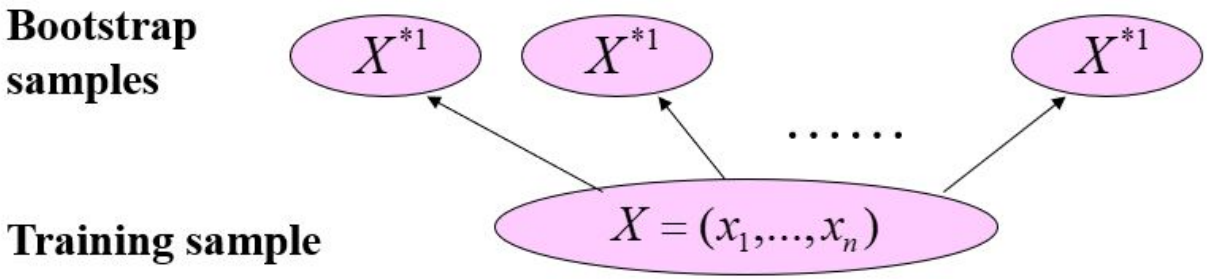
$$\begin{aligned}\hat{\gamma}_i &= E(\Delta_i|\hat{\theta}, Z) = Pr(\Delta_i = 1|\hat{\theta}, Z) \\ &= \frac{\hat{\pi}\psi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\psi_{\hat{\theta}_1}(y_i) + \hat{\pi}\psi_{\hat{\theta}_2}(y_i)}, i = 1, 2, \cdots, N\end{aligned}$$

- **Step 3**:Maximization Step: compute the values for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ which can maximize the log-likelihood given $\hat{\gamma}$
- **Step 4**: Iterate steps 2 and 3 until convergence.

Bootstrap

Basic Idea:

- Randomly draw datasets with replacement from the training data
- Each sample has the same size as the original training set



Bootstrap

- The bootstrap was introduced as a general method for assessing the statistical accuracy of an estimator.
- Data: $X_1, \cdots, X_n \sim F$
- Statistic(any function of the data): $T_n = g(X_1, \cdots, X_n)$
- We want to know $V_F(T_n)$
  Real World: $F \Rightarrow X_1, \cdots, X_n \Rightarrow T_n = g(X_1, \cdots, X_n)$
  Bootstrap World: $\hat{F} \Rightarrow X_1^*, \cdots, X_n^* \Rightarrow T_n^* = g(X_1^*, \cdots, X_n^*)$

$V_F(T_n)$ can be estimated with $V_F(T_n^*)$?

Bootstrap

Suppose we draw a sample $Y_1, \cdots, Y_B$ from a distribution F.

$$\bar{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \to \int y dF(y) = E(Y)(B \to \inf)$$

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j - \bar{Y})^2 = \frac{1}{B}\sum_{j=1}^{B} Y_j^2 - (\frac{1}{B}\sum_{j=1}^{B} Y_j)^2$$
$$\to \int y^2 dF(y) - (\int y dF(y))^2 = V(Y)$$

Bootstrap

- Real World: $F \Rightarrow X_1, \cdots, X_n \Rightarrow T_n = g(X_1, \cdots, X_n)$
- Bootstrap World: $\hat{F} \Rightarrow X_1^*, \cdots, X_n^* \Rightarrow T_n^* = g(X_1^*, \cdots, X_n^*)$
- Bootstrap Variance Estimation:
  1. Draw $X_1^*, \cdots, X_n^* \sim \hat{F}_n$
  2. Compute $T_n^* = g(X_1^*, \cdots, X_n^*)$
  3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \cdots, T_{n,B}^*$
  4. Let

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}T_{n,r}^*)^2$$

$$V_F(T_n) \approx V_{\hat{F}}(T_n^*) \approx v_{boot}$$

Bootstrap

- Non-parametric Bootstrap
  — Uses the raw data, not a specific parametric model, to generate new datasets
- Parametric Bootstrap
  — Simulate new responses by adding Gaussian noise to the predicted values
  — Example from the book
  - $\mu = \sum b_i h_i(x)$ — estimate $\hat{\mu}(x)$
  - we simulate new (x,y) by

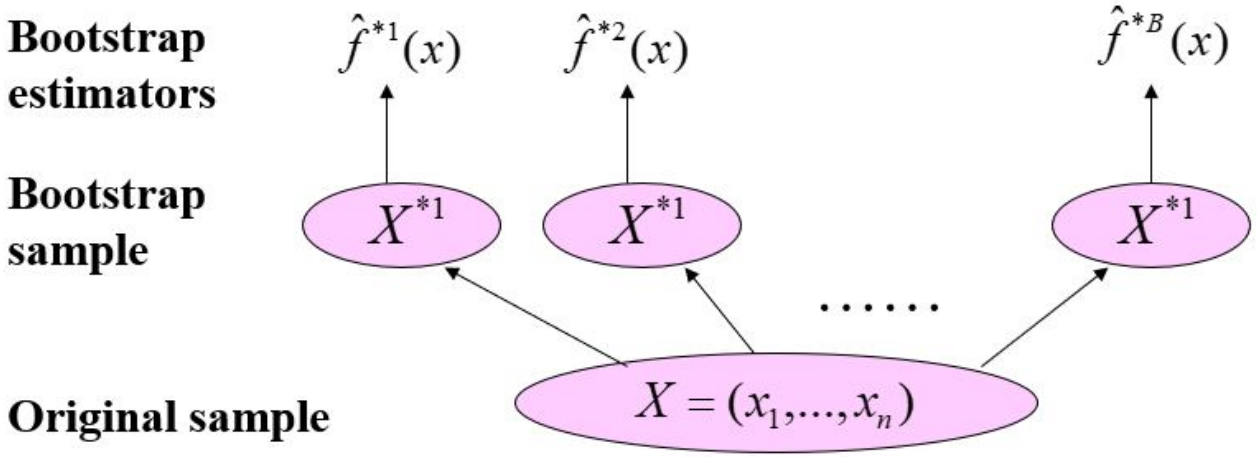
$$y_i^* = \hat{\mu}(x_i) + \epsilon_i^*, \epsilon_i^* \sim N(0, \hat{\sigma}^2)$$

Bootstrap — Summary

- Nonparametric bootstrap — No underlying distribution assumption
- Parametric bootstrap agrees with maximum likelihood
- Bootstrap distribution approximates posterior distribution of parameters with non-informative priors

Bagging

- Bootstrap
  — A way of assessing the accuracy of a parameter estimate or a prediction
- Bagging (Bootstrap Aggregating)
  — Use bootstrap samples to predict data classifiers



$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

Classification becomes majority voting.

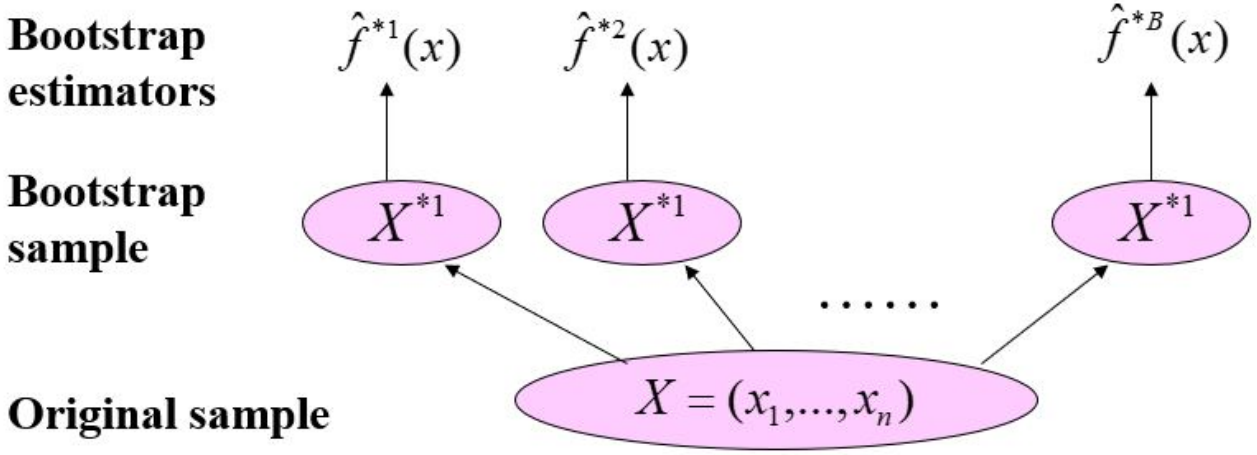
Bagging

- Pros
  - The estimator can be significantly improved if the learning algorithm is unstable.
    — Some change to training set causes large change in output hypothesis
  - Reduce the variance, bias unchanged
- Cons
  - Degrade the performance of stable procedures
  - Lose the structure after bagging

Bumping

A stochastic flavor of model selection

- Bootstrap Umbrella of Model Parameters
- Sample data set, train it, until we are satisfied or tired



$$\hat{b} = argmin_b \sum_{i=1}^{N} [y_i - \hat{f}^{*b}(x_i)]^2$$

Compare different models on the training data.

Conclusion

- Maximum Likelihood vs. Bayesian Inference
- EM vs. Gibbs Sampling
- Bootstrap
  — Bagging
  — Bumping