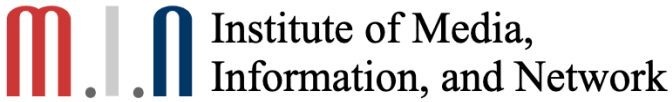


交叉创新课程

Model Inference and Averaging

Wenrui Dai
戴文睿
http://min.sjtu.edu.cn

计算机科学工程系
上海交通大学

2021 年 04 月 12 日

Outline

- The Bootstrap and Maximum Likelihood Methods
- Bayesian Methods
- The EM Algorithm

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

The Bootstrap and Maximum Likelihood Methods

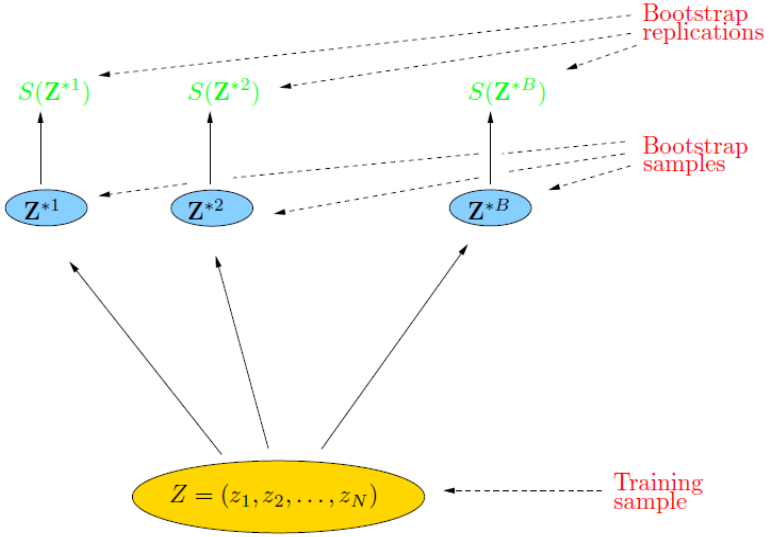
Bootstrap Methods

- ***Basic idea:*** The basic idea is to randomly draw datasets with ***replacement*** from the training data, each sample ***the same size*** *as* the original training set.

- *Training set:* $\boldsymbol{Z} = (z_1, z_2, \ldots, z_N)$ and $z_i = (x_i, y_i)$.

- $\boldsymbol{B}$ *times* ($\boldsymbol{B} = 100$ say): producing $\boldsymbol{B}$ bootstrap datasets

- Refit the model to each of the bootstrap datasets, and examine the behavior of the fits over the B replications.



Schematic of the bootstrap process

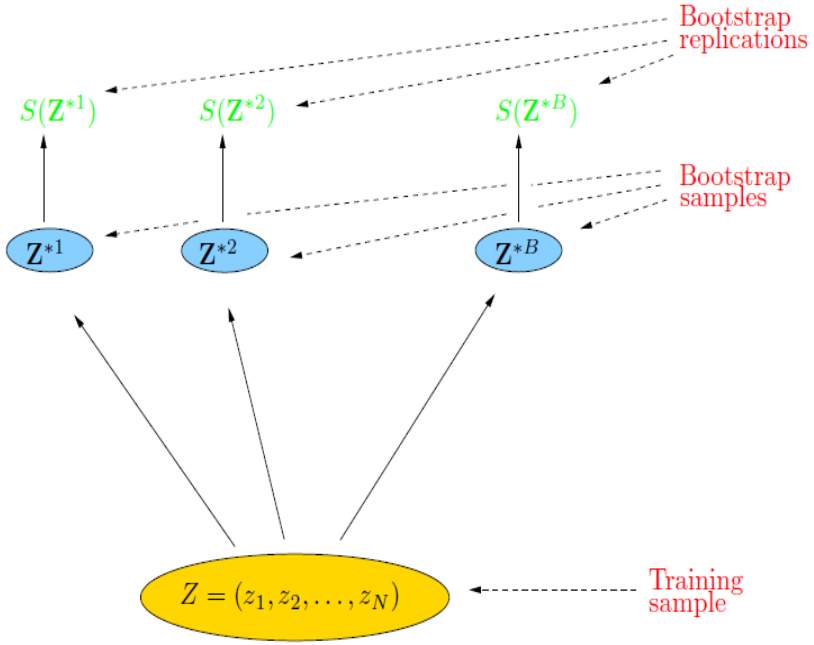
Bootstrap Methods

- $S(\boldsymbol{Z})$ is any quantity computed from the data $\boldsymbol{Z}$, for example, the prediction at some input point.

$$\bar{S}^* = \frac{1}{B}\sum_{b=1}^{B} S(\boldsymbol{Z}^{*b})$$

$$\widehat{\mathrm{Var}}[S(\boldsymbol{Z})] = \frac{1}{B-1}\sum_{b=1}^{B}\left(S(\boldsymbol{Z}^{*b}) - \bar{S}^*\right)^2$$

- Monte Carlo estimate for the distribution for the data $\boldsymbol{Z}$.



Schematic of the bootstrap process

Bootstrap Methods

- Estimating the prediction error with the bootstrap error

$$\widehat{\mathrm{Err}}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

where $\hat{f}^{*b}(x_i)$ is the predicted value at x_i, from the model fitted to the bth bootstrap dataset.

- Leave-one-out bootstrap estimate of prediction error by mimicking cross-validation

$$\widehat{\mathrm{Err}}_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

A Smoothing Example: Bootstrap by Basis Expansions

Fit a cubic spline to the data, with three knots placed at the quartiles of the X values.

- A linear expansion of B-spline $\mu(x) = \sum_{j=1}^{7} \beta_j h_j(x)$.

$h_j(x), j = 1,2,\ldots,7$ are the seven functions shown in the right panel.

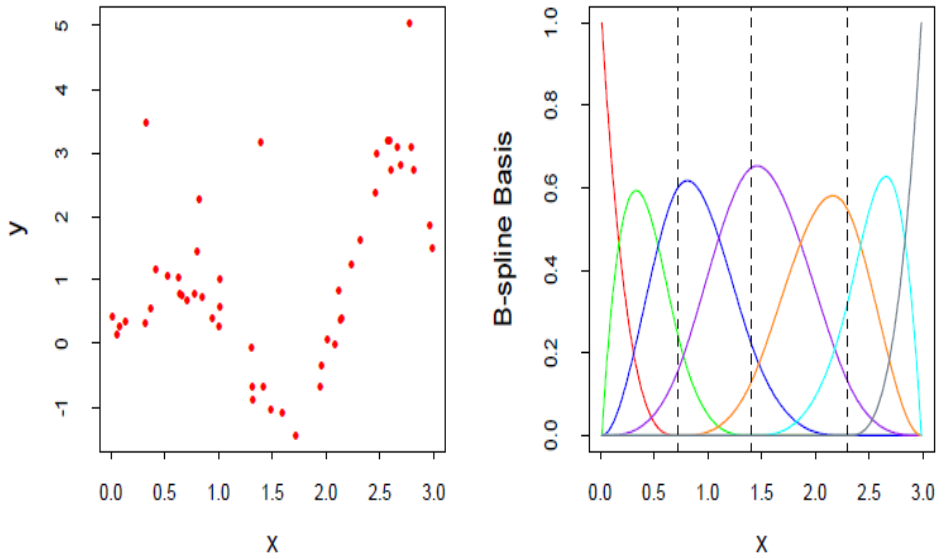
Let $\boldsymbol{H}$ be the N $\times$7 matrix with $ijth$ element $h_j(x_i)$.

- The least square error solution: $\hat{\beta} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T y$

$$\hat{\mu}(x) = \sum_{j=1}^{7} \hat{\beta}_j h_j(x)$$

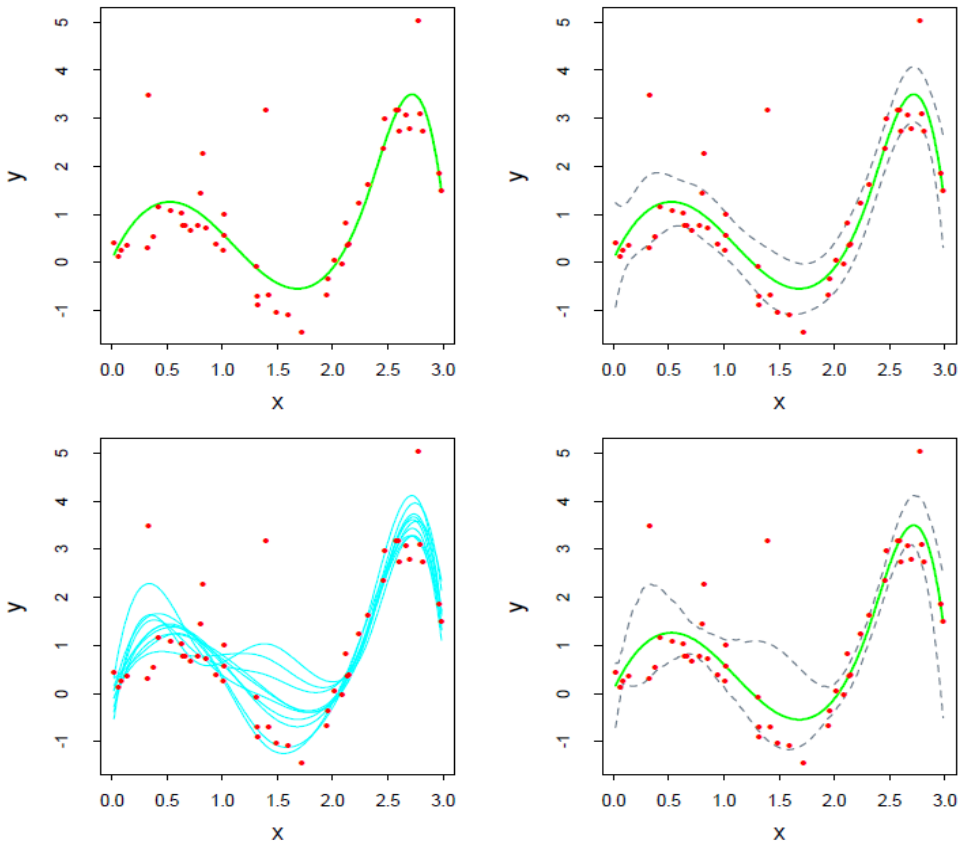
- The Covariance of β: $\mathrm{Cov}(\hat{\beta}) = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\hat{\sigma}^2$

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



(Left panel): Data for smoothing example. (Right panel:) Set of seven B-spline basis functions. The broken vertical lines indicate the placement of the three knots.

A Smoothing Example: Bootstrap by Basis Expansions



- The top left is the B-spline smooth of data.
- The top right is the B-spline smooth plus and minus 1.96× standard error bands.
- We draw B datasets each of size $N = 50$ with replacement from our training data, the sampling unit being the pair $z_i = (x_i, y_i)$. To each bootstrap dataset Z^* we fit a cubic spline $\hat{\mu}^*(x)$; the fits from ten such samples are shown in the bottom left panel.
- Using $B = 200$ bootstrap samples, we can form a 95% pointwise confidence band from the percentiles at each x: we find the $2.5\% \times 200 =$ fifth largest and smallest values at each x. These are plotted in the bottom right panel.

Bootstrap Methods

- Suppose we further assume that the model errors are Gaussian,

$$Y = \mu(X) + \epsilon;\ \epsilon \sim N(0, \sigma^2), \qquad \mu(x) = \sum_{j=1}^{7} \beta_j h_j(x)$$

Nonparametric bootstrap

- The bootstrap method described above, in which we sample with replacement from the training data, is called the ***nonparametric bootstrap*** means that the method is "model-free," since it uses the raw data, not a specific parametric model, to generate new datasets.

Parametric bootstrap

- Consider a variation of the bootstrap, called the ***parametric bootstrap***, in which we simulate new responses by adding Gaussian noise to the predicted values:

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

Maximum Likelihood Inference

- Suppose we are trying to measure the true value of some quantity (x_T).
  - We make repeated measurements of this quantity $\{x_1, x_2, \dots, x_n\}$.
  - The standard way to estimate x_T from our measurements is to calculate the mean value:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and set $x_T = \mu_x$.

DOES THIS PROCEDURE MAKE SENSE?

The maximum likelihood method (MLM) answers this question and provides a general method for estimating parameters of interest from data.

The Maximum Likelihood Method (MLM)

Statement of the Maximum Likelihood Method

- Assume we have made N measurements of x $\{x_1, x_2, \dots, x_n\}$.
- Assume we know the probability distribution function that describes x: $f(x, \boldsymbol{\alpha})$
- Assume we want to determine the parameter $\boldsymbol{\alpha}$.

MLM: pick a to maximize the probability of getting the measurements (the x_i's) we did!

The MLM Implementation

- The probability of measuring x_1 is $f(x_1, \alpha)dx$
- The probability of measuring x_2 is $f(x_2, \alpha)dx$
- The probability of measuring x_n is $f(x_n, \alpha)dx$

- If the measurements are independent, the probability of getting the measurements we did is:

$$L = f(x_1, \alpha)dx \cdot f(x_2, \alpha)dx \cdots f(x_n, \alpha)dx$$

$$= f(x_1, \alpha) \cdot f(x_2, \alpha) \cdots f(x_n, \alpha)[dx^n]$$

L is called the likelihood Function:

$$\boldsymbol{L(\alpha) = \prod_{i=1}^{N} f(x_i, \alpha)}$$

Log Maximum Likelihood Method

- **Maximizes $L(\alpha)$**

$$\left.\frac{\partial L(\alpha)}{\partial \alpha}\right|_{\alpha=\alpha^*} = 0$$

  – Often easier to maximize the log-likelihood $\ln L(\alpha)$

  – $L(\alpha)$ and $\ln L(\alpha)$ are both maximum at the same location.

- $\ln L(\alpha)$ **converts the product into a summation.**

$$\ln L(\alpha) = \sum_{i=1}^{N} \ln f(x_i, \alpha)$$

Log Maximum Likelihood Method

- The new maximization condition is:

$$\left.\frac{\partial lnL(\alpha)}{\partial \alpha}\right|_{\alpha=\alpha^*} = \sum_{i=1}^{N} \left.\frac{\partial}{\partial \alpha} \ln f(x_i, \alpha)\right|_{\alpha=\alpha^*} = 0$$

- α could be an array of parameters (*e.g.*, slope and intercept) or just a single variable.

- Resultant equations: simple linear equations or coupled non-linear equations.

Log Maximum Likelihood Method

An Example: Gaussian

- Let $f(x, \alpha)$ be given by a Gaussian distribution function.
- Let $\alpha = \mu$ be the mean of the Gaussian. Use observed data + MLM to find the mean.
- To find the best estimate of α from our set of n measurements $\{x_1, x_2, \dots, x_n\}$.

- Gaussian PDF

$$f(x_i, \alpha) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x_i-\alpha)^2}{2\sigma^2}}$$

- The likelihood function for this problem is:

$$L = \prod_{i=1}^{n} f(x_i, \alpha) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x_i-\alpha)^2}{2\sigma^2}}$$

Log Maximum Likelihood Method

An Example: Gaussian

- The Log likelihood function for this problem is:

$$\ln L = \ln \prod_{i=1}^{n} f(x_i, \alpha) = \ln\left(\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^n e^{-\sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{2\sigma^2}} \right)$$

$$= n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\left(-\sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{2\sigma^2}\right)$$

- We want to find the a that maximizes the log likelihood function:

$$\frac{\partial \ln L}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left[n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{2\sigma^2} \right] = 0$$

$$\frac{\partial}{\partial \alpha}\sum_{i=1}^{n}(x_i-\alpha)^2 = 0; \qquad \sum_{i=1}^{n} 2(x_i-\alpha)(-1) = 0 \qquad \alpha = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- If σ are different for each data point then α is just the weighted average:

$$\alpha = \frac{\sum_i^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}$$

weighted average

Log Maximum Likelihood Method

An Example: Poisson

- Let $f(x, \alpha)$ be given by a Poisson distribution.
- Let α be the mean of the Poisson.
- We want the best estimate of α from our set of n measurements $\{x_1, x_2, \dots, x_n\}$.
- Poisson PDF: $$f(x, \alpha) = \frac{e^{-\alpha}\alpha^x}{x!}$$

- The likelihood function for this problem is:

$$L = \prod_{i=1}^{n} f(x_i, \alpha) = \prod_{i=1}^{n} \frac{e^{-\alpha}\alpha^{x_i}}{x_i!}$$

$$= \frac{e^{-\alpha}\alpha^{x_1}}{x_1!}\frac{e^{-\alpha}\alpha^{x_2}}{x_2!}\cdots\frac{e^{-\alpha}\alpha^{x_n}}{x_n!} = \frac{e^{-n\alpha}\alpha^{\sum_{i=1}^{n} x_i}}{x_1!x_2!\cdots x_n!}$$

Log Maximum Likelihood Method

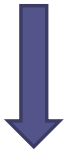
An Example: Poisson

- Find the a that maximizes the log likelihood function:

$$\frac{\partial \ln L}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(-n\alpha + \ln \alpha \cdot \sum_{i=1}^{n} x_i - \ln(x_1! \, x_2! \cdots x_n!)\right)$$

$$= -n + \frac{1}{\alpha}\sum_{i=1}^{n} x_i$$

$$= 0$$



Average $\alpha = \frac{1}{n}\sum_{i=1}^{n} x_i$

General Properties of MLM

- For large data samples (large n) the likelihood function, L, approaches a Gaussian distribution.

- Maximum likelihood estimates are usually ***consistent.***
  – For large n the estimates converge to the true value of the parameters we wish to determine.

- Maximum likelihood estimates are usually ***unbiased***.
  – For all sample sizes the parameter of interest is calculated correctly.

- Maximum likelihood estimate is ***efficient***: the estimate has the smallest variance.
- Maximum likelihood estimate is ***sufficient***: it uses all the information in the observations (the x_i's).
- The solution from MLM is unique.

- Bad news: we must know the correct probability distribution for the problem at hand!

Maximum Likelihood Inference

- In general, the parametric bootstrap agrees with maximum likelihood.
- The probability density or probability mass function for observations Z is $z_i \sim g_\theta(z)$.
- We maximize the likelihood function: $L(\theta; \boldsymbol{Z}) = \prod_{i=1}^{N} g_\theta(z_i)$
- Log-likelihood function: $\ell(\theta; \boldsymbol{Z}) = \log L(\theta; \boldsymbol{Z}) = \sum_{i=1}^{N} \log g_\theta(z_i)$

$$= \sum_{i=1}^{N} \ell(\theta; z_i)$$

where $\ell(\theta; z_i) = \log g_\theta(z_i)$.

Maximum Likelihood Inference

Score function

- Assess the precision of θ using the **score function**

where $\dot{\ell}(\theta; Z) = \frac{\partial \ell(\theta; z_i)}{\partial \theta}$

$$\dot{\ell}(\theta; \boldsymbol{Z}) = \sum_{i=1}^{N} \dot{\ell}(\theta; z_i)$$

- Assume that L takes its maximum in the interior parameter space. Then $\dot{\ell}(\hat{\theta}; \boldsymbol{Z}) = 0$.

Fisher Information

- Negative sum of second derivatives is the information matrix $\boldsymbol{I}(\theta) = -\sum_{i=1}^{N} \frac{\partial^2 \ell(\theta; z_i)}{\partial \theta \partial \theta^T}$

  is called the observed information, should be greater than 0.

- Fisher information (expected information) is $\boldsymbol{i}(\theta) = E_\theta[\boldsymbol{I}(\theta)]$

Maximum Likelihood Inference

Sampling Theory

- Assume that θ_0 is the true value of θ. The sampling distribution of the max-likelihood estimator approaches the following normal distribution, as $N \to \infty$

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

when we sample independently from $g_{\theta_0}(z)$.

- This suggests that the sampling distribution of $\hat{\theta}$ may be approximated by

$$N(\hat{\theta}, \mathbf{i}(\hat{\theta})^{-1}) \text{ or } N(\hat{\theta}, \mathbf{I}(\hat{\theta})^{-1})$$

where $\hat{\theta}$ represents the maximum likelihood estimate from the observed data.

Error Bound

- The corresponding error estimates are obtained from

$$\sqrt{\mathbf{i}(\hat{\theta})^{-1}_{jj}} \text{ and } \sqrt{\mathbf{I}(\hat{\theta})^{-1}_{jj}}$$

- The confidence points have the form

$$\hat{\theta}_j - z^{(1-\alpha)} \cdot \sqrt{\mathbf{i}(\hat{\theta})^{-1}_{jj}}$$

$$\text{and } \hat{\theta}_j - z^{(1-\alpha)} \cdot \sqrt{\mathbf{I}(\hat{\theta})^{-1}_{jj}}$$

$z^{(1-\alpha)}$ is the 1-α percentile of the normal distribution

The connection between MLE and Bootstrap

The smoothing Example $\theta = (\beta, \sigma^2)$

- The log-likelihood: $$l(\theta) = -\frac{N}{2}\log\sigma^2 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \beta^T h(x_i)\right)^2$$
- Estimating equations: $$\frac{\partial l(\theta)}{\partial \beta} = 0; \quad \frac{\partial l(\theta)}{\partial \sigma^2} = 0.$$

Giving

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

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

- The information matrix $$\mathbf{I}(\beta) = (\boldsymbol{H}^T\boldsymbol{H})/\sigma^2 \qquad \mathrm{Cov}(\hat{\beta}) = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\hat{\sigma}^2$$

which are the same as the usual estimates given before.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Bayesian Method

Bayesian Method

- Given a sampling model $\Pr(\boldsymbol{Z}|\theta)$ and a prior $\Pr(\theta)$ for the parameters, estimate the posterior probability

$$\Pr(\theta|\boldsymbol{Z}) = \frac{\Pr(\boldsymbol{Z}|\theta) \cdot \Pr(\theta)}{\int \Pr(\boldsymbol{Z}|\theta) \cdot \Pr(\theta)\, d\theta}$$

- By drawing samples or estimating its mean or parameters
- Differences to counting (frequentist approach)
  - Prior: allow for uncertainties present before seeing the data
  - Posterior: allow for uncertainties present after seeing the data

Bayesian Method

- The posterior distribution affords also a predictive distribution of seeing future values Z^{new}

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

- In contrast, the max-likelihood approach would predict future data on the basis of the data density $\Pr(z^{new}|\hat{\theta})$ evaluated at the maximum likelihood estimate not accounting for the uncertainty in the parameters

Bayesian Method

The smoothing Example

- Consider a linear expansion $Y = \mu(X) + \varepsilon; \varepsilon \sim N(0, \sigma^2)$

$$\mu(x) = \sum_{j=1}^{7} \beta_j h_j(x)$$

- The least square error solution $\beta \sim N(0, \tau\Sigma)$

$$p(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{p}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

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

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

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

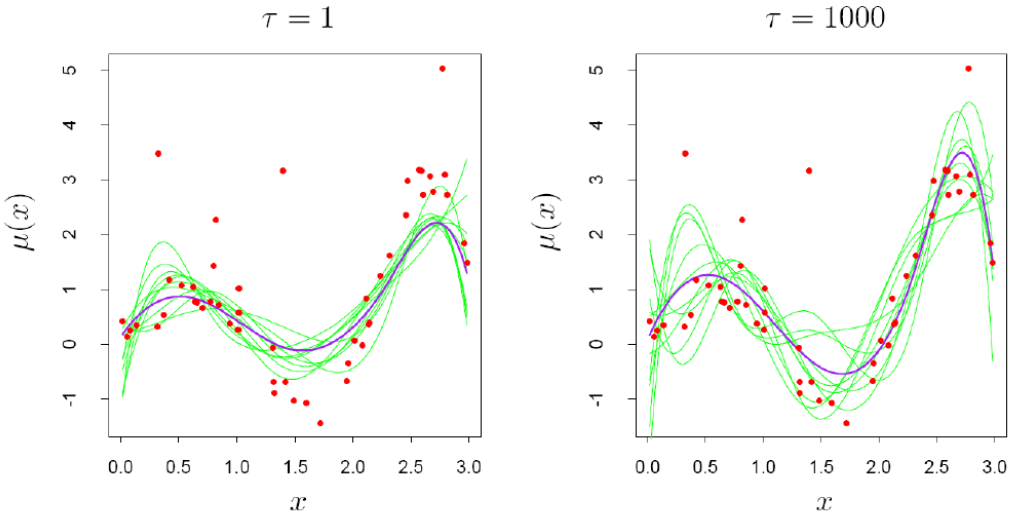
Bayesian Method

The smoothing Example

- The corresponding posterior values for $\mu(x) = \sum_{j=1}^{N} \beta_j h_j(x)$

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

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



Smoothing example: Ten draws from the posterior distribution for the function $\mu(x)$, for two different values of the prior variance τ . The purple curves are the posterior means.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

The EM Algorithm

The EM Algorithm

- In statistics, an expectation–maximization (EM) algorithm is an iterative method to find (local) ***maximum likelihood*** or ***maximum a posteriori (MAP)*** estimates of parameters in statistical models, where the model depends on unobserved ***latent variables***.

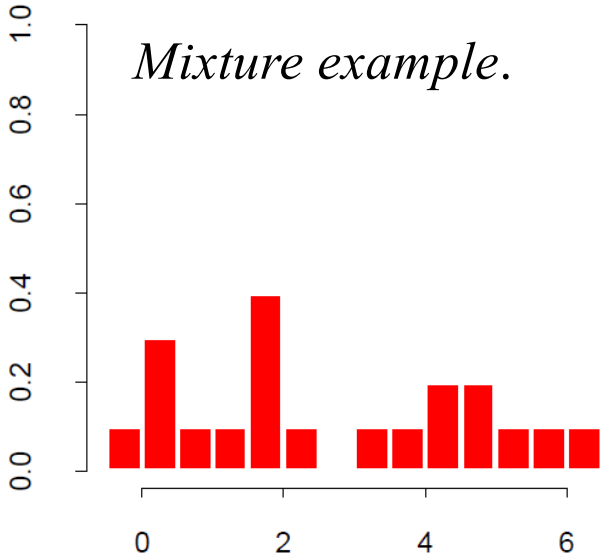
- The EM iteration alternates between performing ***an expectation (E) step***, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and ***a maximization (M) step***, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

The EM Algorithm

Two-Component Mixture Model

TABLE 8.1. *Twenty fictitious data points used in the two-component mixture example in Figure 8.5.*

-0.39	0.12	0.94	1.67	1.76	2.44	3.72	4.28	4.92	5.53
0.06	0.48	1.01	1.68	1.80	3.25	4.12	4.60	5.28	6.22



Mixture example.

- The left panel shows a histogram of the 20 fictitious data points in Table 8.1.
- We would like to model the density of the data points, and due to the apparent bi-modality, a Gaussian distribution would not be appropriate.

The EM Algorithm

Two-Component Mixture Model

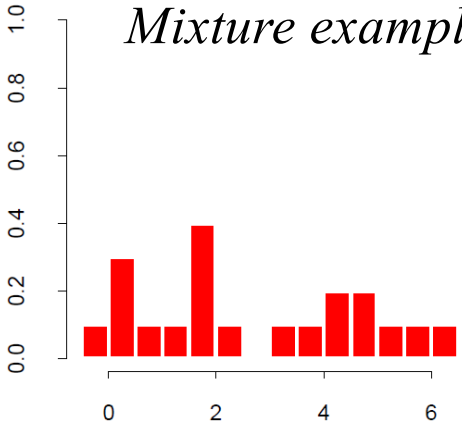
- There seems to be two separate underlying regimes, so instead we model Y as a mixture of two normal distributions:

$$Y_1 \sim N(\mu_1, \sigma_1^2) \quad Y_2 \sim N(\mu_2, \sigma_2^2)$$

$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2$$

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

Mixture example.



- Let $\phi_\theta(x)$ denote the normal density with parameters $\theta = (\mu, \sigma^2)$. Then the density of Y is

$$g_Y(y) = (1 - \pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y)$$

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

The EM Algorithm

Two-Component Mixture Model

- The log-likelihood based on the N training cases is

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

→ **Bad**

- Suppose we observe **Latent Binary Δ_i**

$$\ell_0(\theta; \boldsymbol{Z}, \Delta) = \sum_{i=1}^{N}\left[(1-\Delta_i)\log\phi_{\theta_1}(y_i) + \Delta_i \log\phi_{\theta_2}(y_i)\right]$$

$$+ \sum_{i=1}^{N}\left[(1-\Delta_i)\log(1-\pi) + \Delta_i \log\pi\right]$$

→ **Good**

Notation: $L(\theta, \boldsymbol{Z}, \Delta) = \prod_{i=1}^{n}\left[(1-\pi)\phi_{\theta_1}(y_i)\right]^{1-\Delta_i}\left[\pi\phi_{\theta_2}(y_i)\right]^{\Delta_i}$

The maximum likelihood estimates of μ_1 and σ_1 would be the sample mean and variance for those data with $\Delta_i = 0$, and similarly those for μ_2 and σ_2 would be the sample mean and variance of the data with $\Delta_i = 1$. The estimate of π would be the proportion of $\Delta_i = 1$.

The EM Algorithm

Two-Component Mixture Model

- Since the values of the Δ_i's are actually unknown, we proceed in an iterative fashion, substituting for each Δ_i in the log-likelihood its expected value

$$\gamma_i(\theta) = E(\Delta_i|\theta, \boldsymbol{Z}) = \Pr(\Delta_i = 1|\theta, \boldsymbol{Z})$$

also called the ***responsibility*** of model 2 for observation i.

Algorithm 8.1 *EM Algorithm for Two-component Gaussian Mixture.*

1. Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ (see text).

2. *Expectation Step*: compute the responsibilities **do a soft assignment of each observation to each model**

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \ldots, N. \qquad (8.42)$$

3. *Maximization Step*: compute the weighted means and variances:

update the estimates of parameters

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

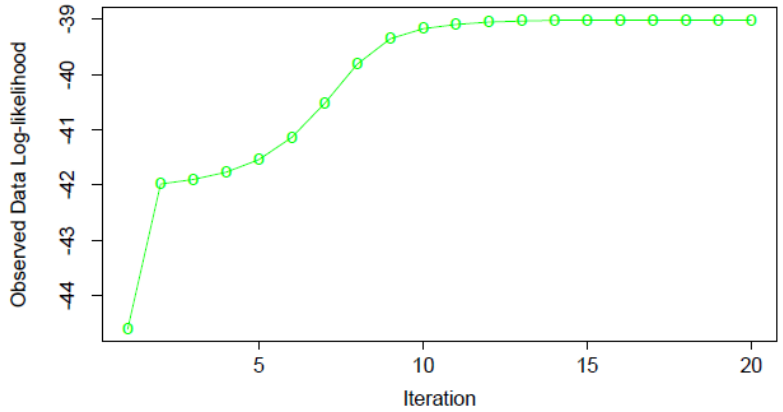
$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.

The EM Algorithm

Two-Component Mixture Model



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

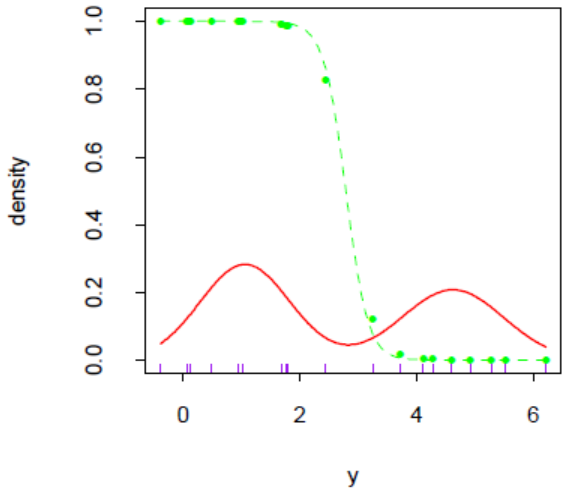
Selected iterations of the EM algorithm for mixture example.

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

$$\hat{\mu}_1 = 4.62, \quad \hat{\sigma}_1^2 = 0.87$$
$$\hat{\mu}_2 = 1.06, \quad \hat{\sigma}_2^2 = 0.77$$

Solid red curve: the estimated Gaussian mixture density from this procedure

dotted green curve: The responsibilities.



The EM Algorithm in General

Z: input data, with log-likelihood $\ell(\theta, \boldsymbol{Z})$

Z^m: latent data (in our example Δ_i)

$T = (Z, Z^m)$: complete data with log-likelihood $\ell_0(\theta; \boldsymbol{T})$

$$\Pr(\boldsymbol{Z}^m|\boldsymbol{Z}, \theta') = \frac{\Pr(\boldsymbol{Z}^m|\boldsymbol{Z}, \theta')}{\Pr(\boldsymbol{Z}|\theta')} \longrightarrow \Pr(\boldsymbol{Z}|\theta') = \frac{\Pr(\boldsymbol{T}|\theta')}{\Pr(\boldsymbol{Z}^m|\boldsymbol{Z}, \theta')}$$

- In terms of log-likelihoods, we have

$$\ell(\theta'; \boldsymbol{Z}) = \ell_0(\theta'; \boldsymbol{T}) - \ell_1(\theta'; \boldsymbol{Z}^m|\boldsymbol{Z})$$

where ℓ_1 is based on the conditional density $\Pr(\boldsymbol{Z}^m|\boldsymbol{Z}, \theta')$.

The EM Algorithm in General

- Taking conditional expectations with respect to **the distribution of $\boldsymbol{T|Z}$** governed by parameter θ gives

$$\ell(\theta';\mathbf{Z}) = E(\ell_0(\theta';\mathbf{T})|\mathbf{Z},\theta) - E(\ell_1(\theta';\mathbf{Z}^m|\mathbf{Z})|\mathbf{Z},\theta)$$

$$= Q(\theta';\theta) - R(\theta';\theta)$$

- Note that $R(\theta^*,\theta)$ is the expectation of a log-likelihood of a density (indexed by θ^*), with respect to the same density indexed by θ, and hence (by Jensen's inequality) is maximized as a function of θ^*, when $\theta^* = \theta$.

$$\begin{aligned} R(\theta';\theta) &= E\left(\ell_1(\theta';\mathbf{Z}^m \,|\mathbf{Z})\,|\mathbf{Z},,\theta\right) \\ &= \int \log\left(p(\mathbf{Z}^m \,|\mathbf{Z},\theta')\right) p(\mathbf{Z}^m \,|\mathbf{Z},\theta)\, d\mathbf{Z}^m \\ &\le \int \log\left(p(\mathbf{Z}^m \,|\mathbf{Z},\theta)\right) p(\mathbf{Z}^m \,|\mathbf{Z},\theta)\, d\mathbf{Z}^m \\ &= R(\theta;\theta) \end{aligned}$$

The EM Algorithm in General

$\ell(\theta'; \boldsymbol{Z}) = Q(\theta'; \theta) - R(\theta'; \theta)$

- In the M step, the EM algorithm maximizes $Q(\theta', \theta)$ over θ', rather than the actual objective function $\ell(\theta'; \boldsymbol{Z})$.

Why does it succeed in maximizing $\ell(\theta'; \boldsymbol{Z})$?

- If θ' maximizes $Q(\theta', \theta)$, we see that

$$\begin{aligned}\ell(\theta'; \boldsymbol{Z}) - \ell(\theta; \boldsymbol{Z}) &= Q(\theta'; \theta) - R(\theta'; \theta) \\ &= Q(\theta'; \theta) - Q(\theta; \theta) - (R(\theta'; \theta) - R(\theta; \theta)) \\ &\geq 0\end{aligned}$$

Hence the EM iteration never decreases the log-likelihood.

The EM Algorithm in General

Algorithm 8.2 *The EM Algorithm.*

1. Start with initial guesses for the parameters $\hat{\theta}^{(0)}$.
2. *Expectation Step*: at the jth step, compute

$$Q(\theta', \hat{\theta}^{(j)}) = \mathrm{E}(\ell_0(\theta'; \mathbf{T})|\mathbf{Z}, \hat{\theta}^{(j)}) \tag{8.43}$$

   as a function of the dummy argument θ'.
3. *Maximization Step*: determine the new estimate $\hat{\theta}^{(j+1)}$ as the maximizer of $Q(\theta', \hat{\theta}^{(j)})$ over θ'.
4. Iterate steps 2 and 3 until convergence.

A full maximization in the M step is not necessary: we need only to find a value $\hat{\theta}^{(j+1)}$ so that $Q(\theta', \hat{\theta}^{(j)})$ increases as a function of the first argument, that is, $Q(\hat{\theta}^{(j+1)}, \hat{\theta}^{(j)}) > Q(\hat{\theta}^{(j)}, \hat{\theta}^{(j)})$. Such procedures are called GEM (generalized EM) algorithms.

EM as a Maximization–Maximization Procedure

- Here is a different view of the EM procedure, as a joint maximization algorithm.

$$F(\theta', \tilde{P}) = E_{\tilde{P}}[\ell_0(\theta'; \boldsymbol{T})] - E_{\tilde{P}}[log\tilde{P}(\boldsymbol{Z}^m)].$$

where $\tilde{P}(\boldsymbol{Z}^m)$ is any distribution over the latent data $\boldsymbol{Z}^m$.

- **The EM algorithm can be viewed as a joint maximization method for F over θ' and $\tilde{P}(\boldsymbol{Z}^m)$, by fixing one argument and maximizing over the other.**

- The maximizer over $\tilde{P}(\boldsymbol{Z}^m)$ for fixed θ' can be shown to be

$$\tilde{P}(\boldsymbol{Z}^m) = \Pr(\boldsymbol{Z}^m | \boldsymbol{Z}, \theta')$$

This is the distribution computed by the E step.

EM as a Maximization–Maximization Procedure

- In the M step, we maximize $F(\theta', \tilde{P})$ over θ' with $\tilde{P}$ fixed: this is the same as maximizing the first term $E_{\tilde{P}}[\ell_0(\theta'; \boldsymbol{T})|\boldsymbol{Z}, \theta]$ since the second term does not involve θ'.

- Finally, since $F(\theta', \tilde{P})$ and the observed data log-likelihood agree when $\tilde{P}(\boldsymbol{Z}^m) = \Pr(\boldsymbol{Z}^m|\boldsymbol{Z}, \theta')$, maximization of the former accomplishes maximization of the latter.

- This view of the EM algorithm leads to ***alternative maximization procedures***. For example, one does not need to maximize with respect to all of the latent data parameters at once, but could instead maximize over one of them at a time, alternating with the M step.

Institute of Media, Information, and Network

生命科学技术学院
School of Life Sciences and Biotechnology

Q & A

Many Thanks