Model Inference and Model Averaging

Yanrong Yang

RSFAS/CBE, Australian National University

23rd August 2022

Contents of this week

Model Inference

- Review on Frenquency Statistics vs Bayesian Statistics
- Likelihood Estimation and EM algorithm
- Model Averaging in Frenquency Statistics and Bayesian Statistics

Review on Statistical Inference

Two Typical Statistical Inference

Frequentist versus Bayesian Methods

- In frequentist inference, probabilities are interpreted as long run frequencies. The goal is to create procedures with long run frequency guarantees.
- In Bayesian inference, probabilities are interpreted as subjective degrees of belief. The goal is to state and analyze your beliefs.

Distinguish Frequency from Bayes

Some differences between the frequentist and Bayesian approaches are as follows:

	Frequentist	Bayesian
Probability is:	limiting relative frequency	degree of belief
Parameter θ is a:	fixed constant	random variable
Probability statements are about:	procedures	parameters
Frequency guarantees?	yes	no

Example 1

To illustrate the difference, consider the following example. Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$. We want to provide some sort of interval estimate C for θ.

Frequentist Approach. Construct the confidence interval

$$C = \left[\overline{X}_n - \frac{1.96}{\sqrt{n}},\ \overline{X}_n + \frac{1.96}{\sqrt{n}}\right].$$

Then

$$\mathbb{P}_\theta(\theta \in C) = 0.95 \quad \text{for all } \theta \in \mathbb{R}.$$

The probability statement is about the random interval C. The interval is random because it is a function of the data. The parameter θ is a fixed, unknown quantity. The statement means that C will trap the true value with probability 0.95.

Example 2

Bayesian Approach. The Bayesian treats probability as beliefs, not frequencies. The unknown parameter θ is given a prior distributon $\pi(\theta)$ representing his subjective beliefs about θ. After seeing the data $X_1, \ldots, X_n$, he computes the posterior distribution for θ given the data using Bayes theorem:

$$\pi(\theta|X_1, \ldots, X_n) \propto \mathcal{L}(\theta)\pi(\theta) \tag{12.2}$$

where $\mathcal{L}(\theta)$ is the likelihood function. Next we finds an interval C such that

$$\int_C \pi(\theta|X_1, \ldots, X_n)d\theta = 0.95.$$

He can thn report that

$$\mathbb{P}(\theta \in C|X_1, \ldots, X_n) = 0.95.$$

Principle of Maximum Likelihood

- The function $L(\theta) = f(X \mid \theta)$ with X fixed and θ unknown is called the *likelihood function*.
- The *principle of maximum likelihood* is to estimate θ with the value $\hat{\theta}$ that maximizes $L(\theta)$.
- In practice, it is common to maximize the log-likelihood, $\ell(\theta) = \ln L(\theta)$.
- This is because X often takes the form of an independent sample so that

$$L(X) = \prod_{i=1}^{n} f(X_i \mid \theta), \qquad \ell(\theta) = \sum_{i=1}^{n} \ln f(X_i \mid \theta)$$

Example 3

- A coin has a probability θ of being a head.
- Consider tossing the coin 100 times. The probability of each single sequence with exactly x heads is $f(x \mid \theta) = p^x(1-p)^{100-x}$.
- Say we observe the sequence

$$HHTHTHHT \ldots TTH$$

  where heads appear 57 times.
- The maximum likelihood estimate is the value $\hat{\theta}$ that maximizes the function

$$L(\theta) = \theta^{57}(1-\theta)^{43},$$

  or, equivalently that maximizes

$$\ell(\theta) = 57(\ln \theta) + 43(\ln(1-\theta)) \ .$$

  Simple calculus and common sense lead to the estimate $\hat{\theta} = 0.57$.

Bayesian Inference

Bayesian Procedure

1. We choose a probability density $\pi(\theta)$ — called the prior distribution — that expresses our beliefs about a parameter θ before we see any data.
2. We choose a statistical model $p(\mathrm{x} \mid \theta)$ that reflects our beliefs about x given θ.
3. After observing data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$, we update our beliefs and calculate the posterior distribution $p(\theta \mid \mathcal{D}_n)$.

Bayes' Theorem/Formula

By Bayes' theorem, the posterior distribution can be written as

$$p(\theta \mid X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n \mid \theta)\pi(\theta)}{p(X_1, \ldots, X_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta)$$

where $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p(X_i \mid \theta)$ is the likelihood function and

$$c_n = p(X_1, \ldots, X_n) = \int p(X_1, \ldots, X_n \mid \theta)\pi(\theta)d\theta = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$$

is the normalizing constant, which is also called the evidence.

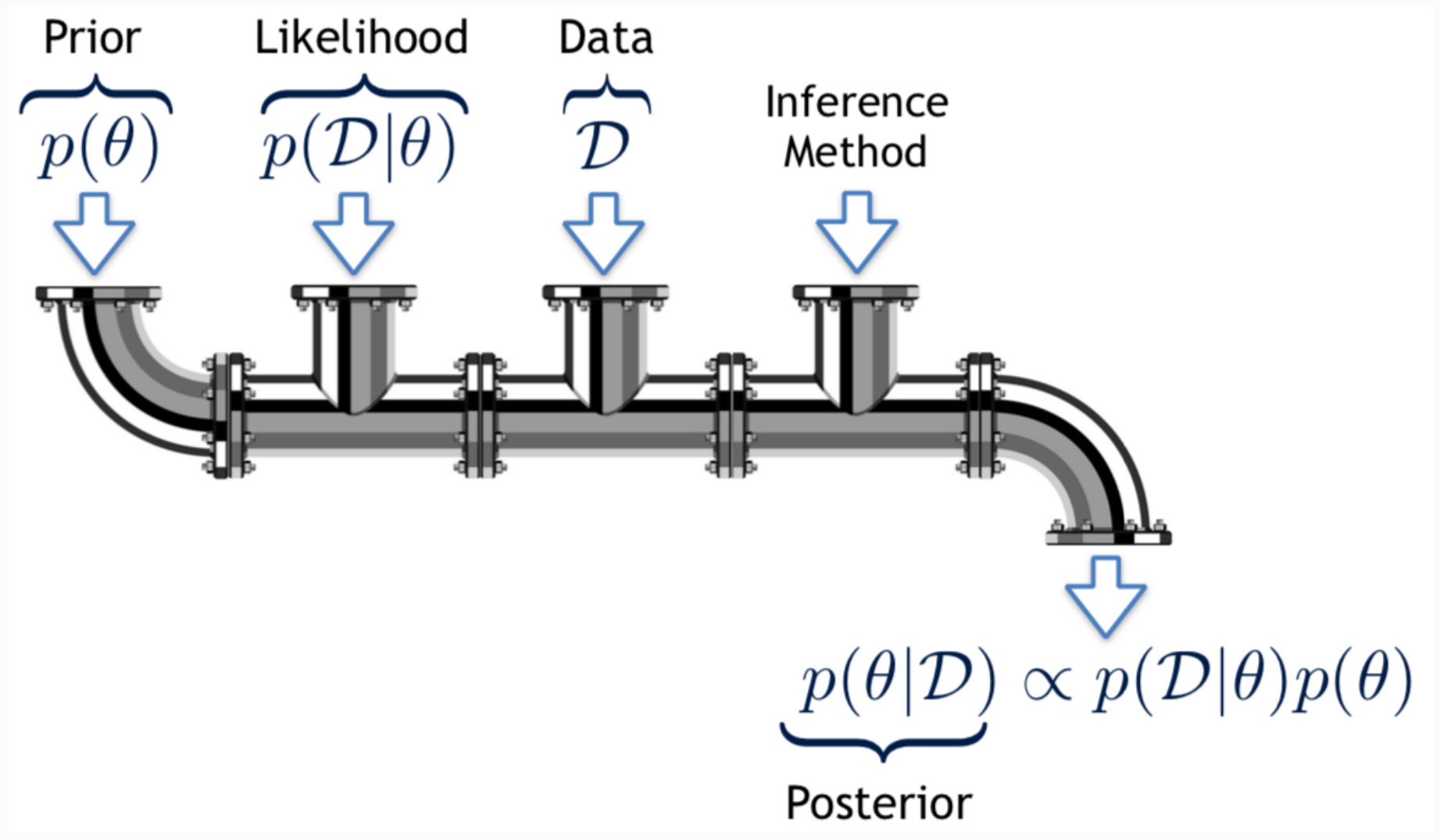
Bayesian Prediction

After the data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ have been observed, the Bayesian framework allows us to predict the distribution of a future data point X conditioned on $\mathcal{D}_n$. To do this, we first obtain the posterior $p(\theta \,|\, \mathcal{D}_n)$. Then

$$\begin{aligned} p(\mathrm{x} \,|\, \mathcal{D}_n) &= \int p(\mathrm{x}, \theta \,|\, \mathcal{D}_n) d\theta \\ &= \int p(\mathrm{x} \,|\, \theta, \mathcal{D}_n) p(\theta \,|\, \mathcal{D}_n) d\theta \\ &= \int p(\mathrm{x} \,|\, \theta) p(\theta \,|\, \mathcal{D}_n) d\theta. \end{aligned}$$

Where we use the fact that $p(\mathrm{x} \,|\, \theta, \mathcal{D}_n) = p(\mathrm{x} \,|\, \theta)$ since all the data are conditionally independent given θ. From the last line, the predictive distribution $p(\mathrm{x} \,|\, \mathcal{D}_n)$ can be viewed as a weighted average of the model $p(\mathrm{x} \,|\, \theta)$. The weights are determined by the posterior distribution of θ.

Illustration of Bayesian Inference



Example 4: Density Estimation

Example 205. Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ where $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. Suppose we take the uniform distribution $\pi(\theta) = 1$ as a prior. By Bayes' theorem, the posterior is

$$p(\theta \,|\, \mathcal{D}_n) \propto \pi(\theta)\mathcal{L}_n(\theta) = \theta^{S_n}(1-\theta)^{n-S_n} = \theta^{S_n+1-1}(1-\theta)^{n-S_n+1-1}$$

where $S_n = \sum_{i=1}^n X_i$ is the number of successes. Recall that a random variable θ on the interval $(0,1)$ has a Beta distribution with parameters α and β if its density is

$$\pi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

We see that the posterior distribution for θ is a Beta distribution with parameters $S_n + 1$ and $n - S_n + 1$. That is,

$$p(\theta \,|\, \mathcal{D}_n) = \frac{\Gamma(n+2)}{\Gamma(S_n+1)\Gamma(n-S_n+1)}\theta^{(S_n+1)-1}(1-\theta)^{(n-S_n+1)-1}.$$

We write this as

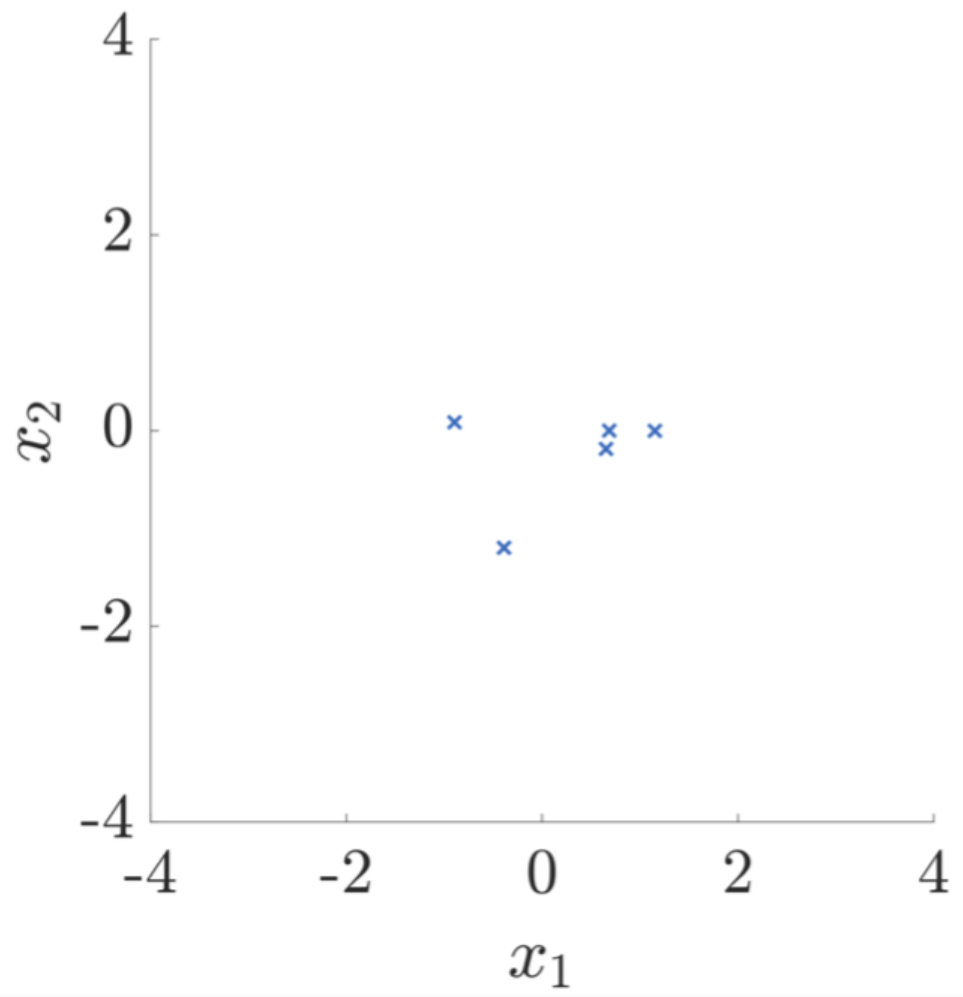
$$\theta \,|\, \mathcal{D}_n \sim \text{Beta}(S_n + 1, n - S_n + 1).$$

Example 5: Density Estimation

Presume that we decide to use an isotropic Gaussian likelihood with unknown mean θ to model the data on the right:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \mathcal{N}(x_n; \theta, I)$$

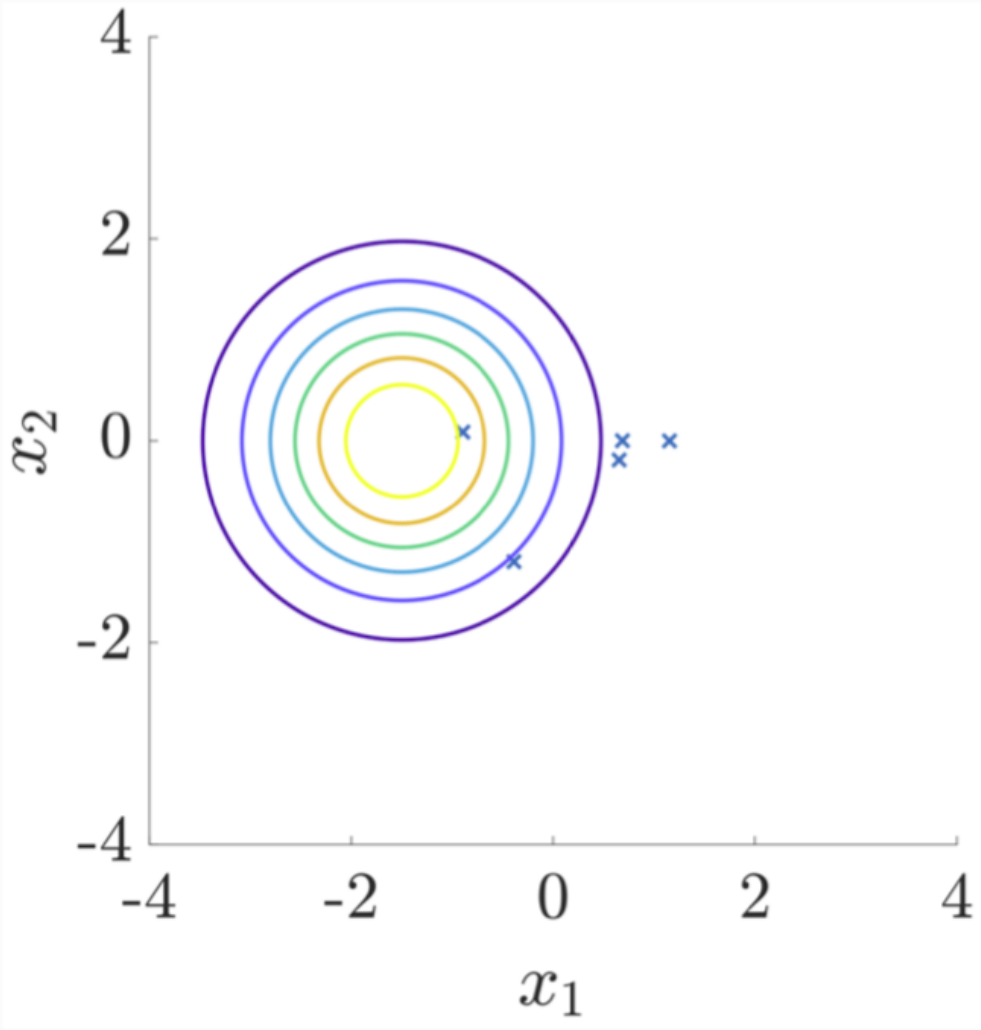
where I is a two-dimensional identity matrix



Example 5: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$$p(\mathcal{D}|\theta = [-2, 0]) = 0.00059 \times 10^{-5}$$



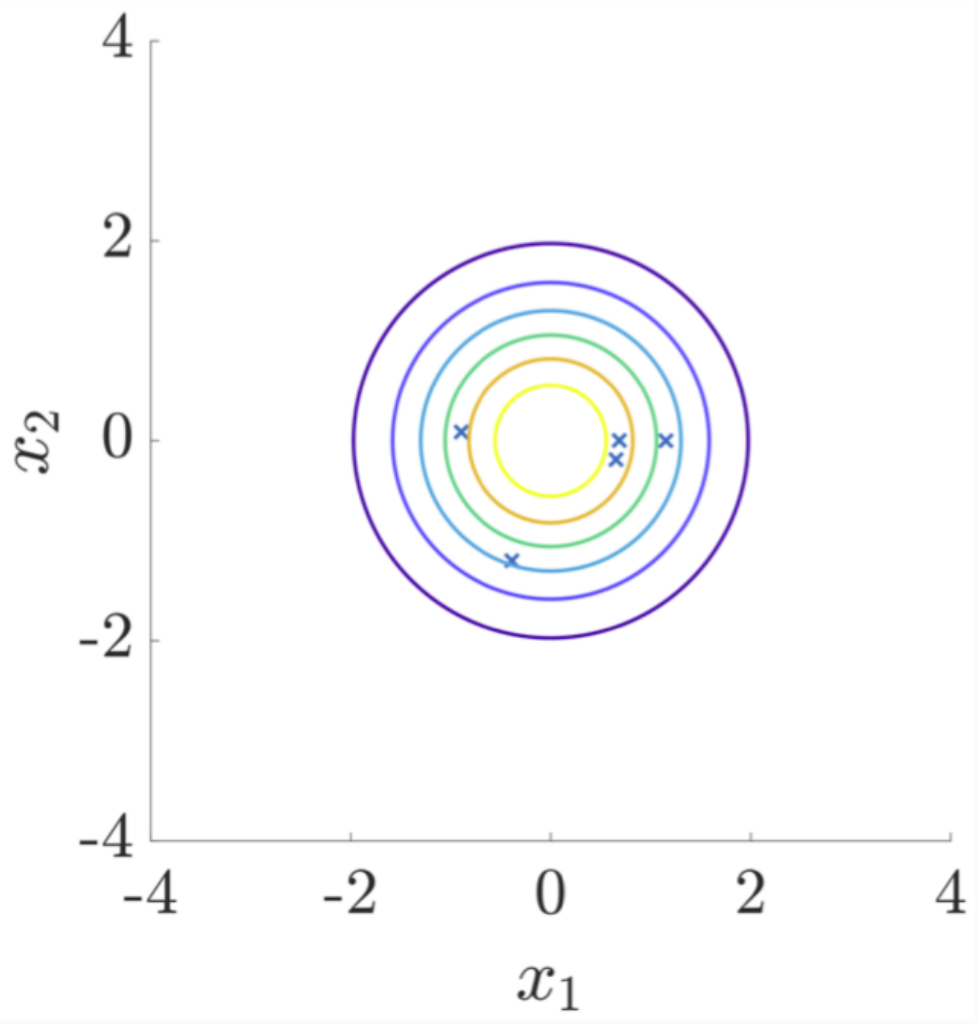
Example 5: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$$p(\mathcal{D}|\theta = [-2, 0]) = 0.00059 \times 10^{-5}$$

Hypothesis 2: $\theta = [0, 0]$

$$p(\mathcal{D}|\theta = [0, 0]) = 0.99 \times 10^{-5}$$



Example 5: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

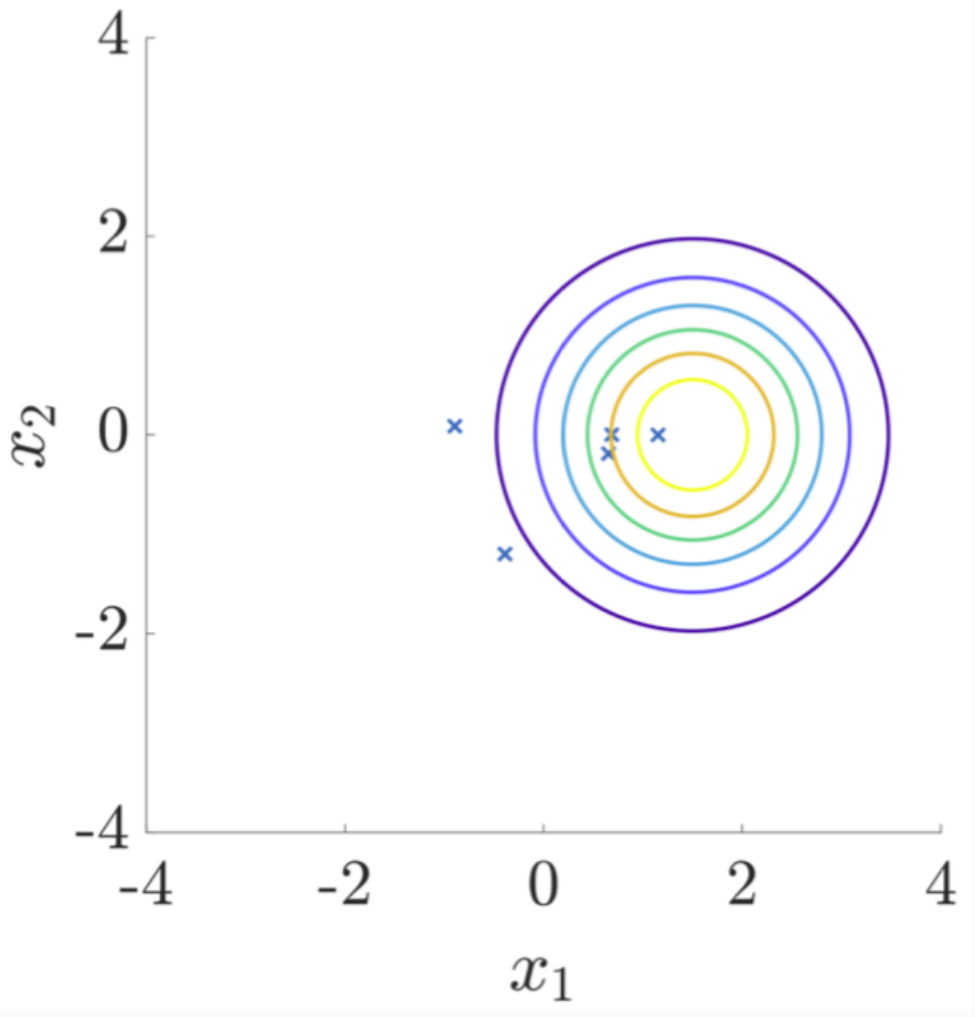
$$p(\mathcal{D}|\theta = [-2, 0]) = 0.00059 \times 10^{-5}$$

Hypothesis 2: $\theta = [0, 0]$

$$p(\mathcal{D}|\theta = [0, 0]) = 0.99 \times 10^{-5}$$

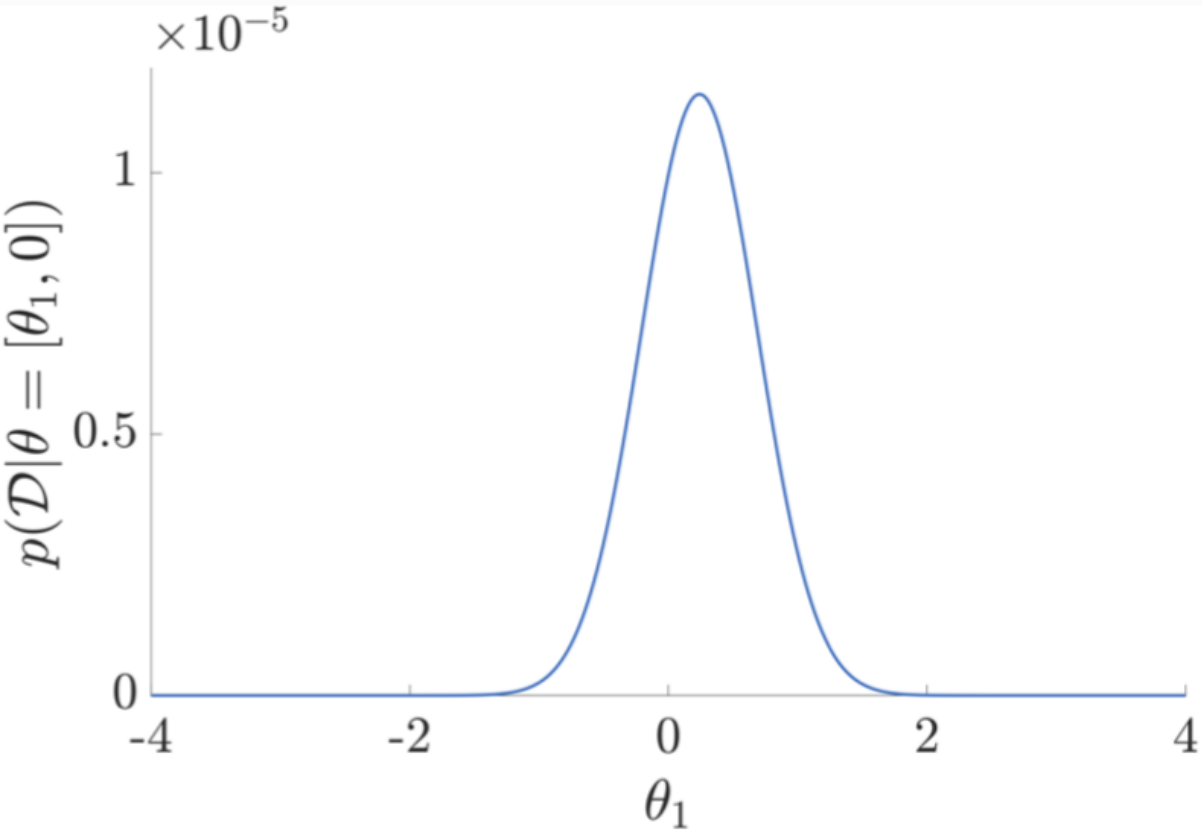
Hypothesis 3: $\theta = [2, 0]$

$$p(\mathcal{D}|\theta = [2, 0]) = 0.021 \times 10^{-5}$$



Example 5: Illustration of Density Estimation

More generally, the likelihood model is telling us how **likely** each hypothesis is to be correct given the data we observe.



Example 5: Prior Distribution

The posterior predictive distribution allows us to **average** over each of our hypotheses, weighting each by their posterior probability.

For example, in our density estimation example, lets introduce (the rather unusual but demonstrative) prior,

$$p(\theta) = \begin{cases} 0.05 & \text{if} \quad \theta = [-2, 0] \\ 0.05 & \text{if} \quad \theta = [0, 0] \\ 0.9 & \text{if} \quad \theta = [2, 0] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Example 5: Prediction

Then we have (note $d\theta$ is a counting measure below)

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\theta)p(\theta|\mathcal{D})d\theta \\
&= \frac{1}{p(\mathcal{D})}\int p(x|\theta)p(\theta,\mathcal{D})d\theta \\
&= \frac{1}{p(\mathcal{D})}\Big(\mathcal{N}(x;[-2,0],I)\times 0.05\times p(\mathcal{D}|\theta=[-2,0]) \\
&\qquad + \mathcal{N}(x;[0,0],I)\times 0.05\times p(\mathcal{D}|\theta=[0,0]) \\
&\qquad + \mathcal{N}(x;[2,0],I)\times 0.9\times p(\mathcal{D}|\theta=[2,0])\Big)
\end{aligned}
$$

Example 5: Prediction

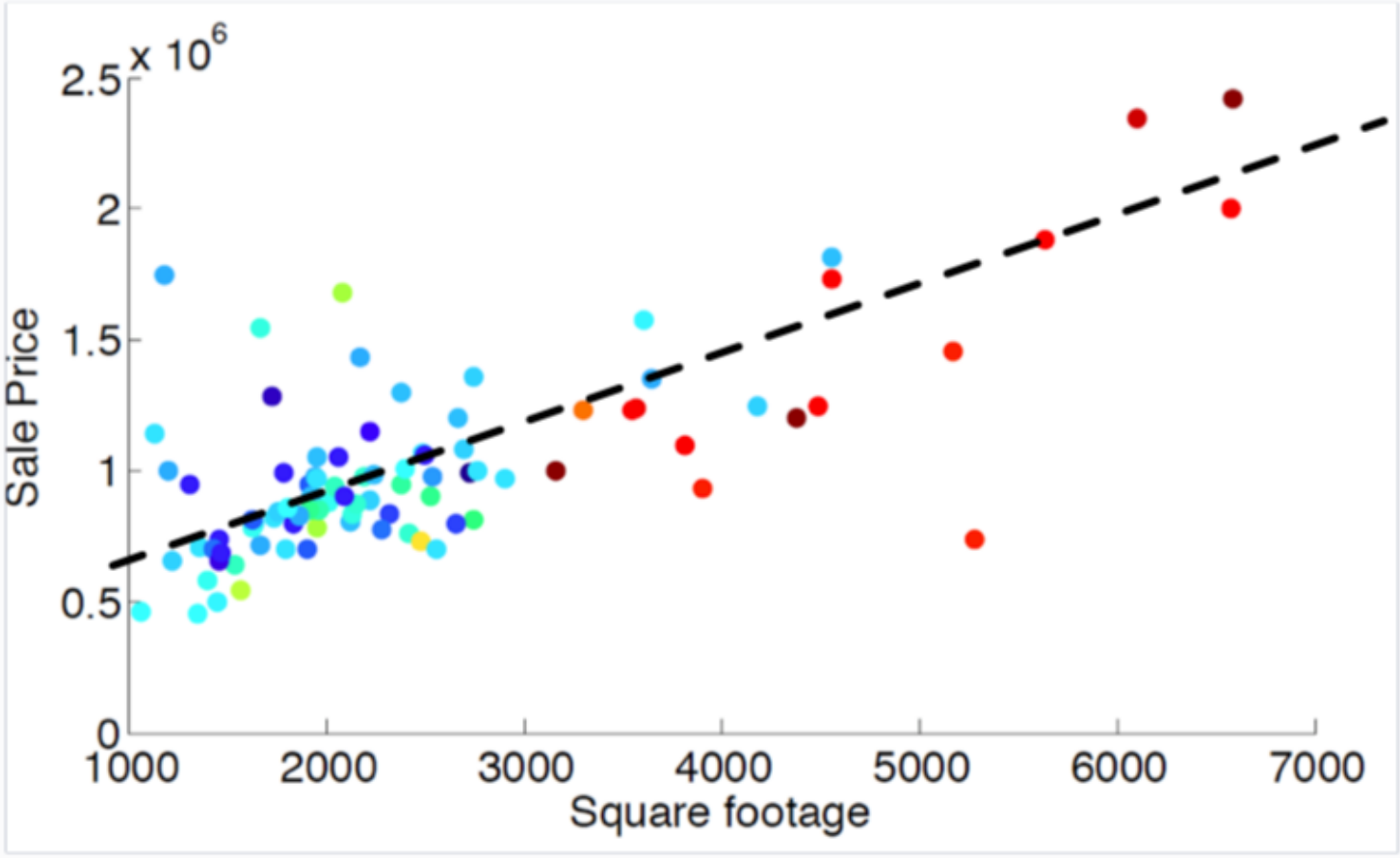
Inserting our likelihoods from earlier and trawling through the algebra now gives

$$\begin{aligned} p(x|\mathcal{D}) =& 0.0004 \times \mathcal{N}(x; [-2, 0], I) \\ &+ 0.716 \times \mathcal{N}(x; [0, 0], I) \\ &+ 0.283 \times \mathcal{N}(x; [2, 0], I) \end{aligned}$$

We thus have that the posterior predictive is a weighted sum of the three possible predictive distributions

Example 6: Bayesian Linear Model

House size is a good linear predictor for price (ignore the colors)



Example 6: Problem Setting

Here we have:

- Inputs $x \in \mathbb{R}^D$ (where $D = 1$ for this particular problem)
- Outputs $y \in \mathbb{R}$
- Data $\mathcal{D}$ comprising of N input–output pairs: $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$
- A regression model $y \approx x^T w + b$ where $w \in \mathbb{R}^D$ and $b \in \mathbb{R}$
- We can simplify this notation by redefining $x \leftarrow [1, x^T]^T$ and $w \leftarrow [b, w^T]^T$, such that we now have $y \approx x^T w$

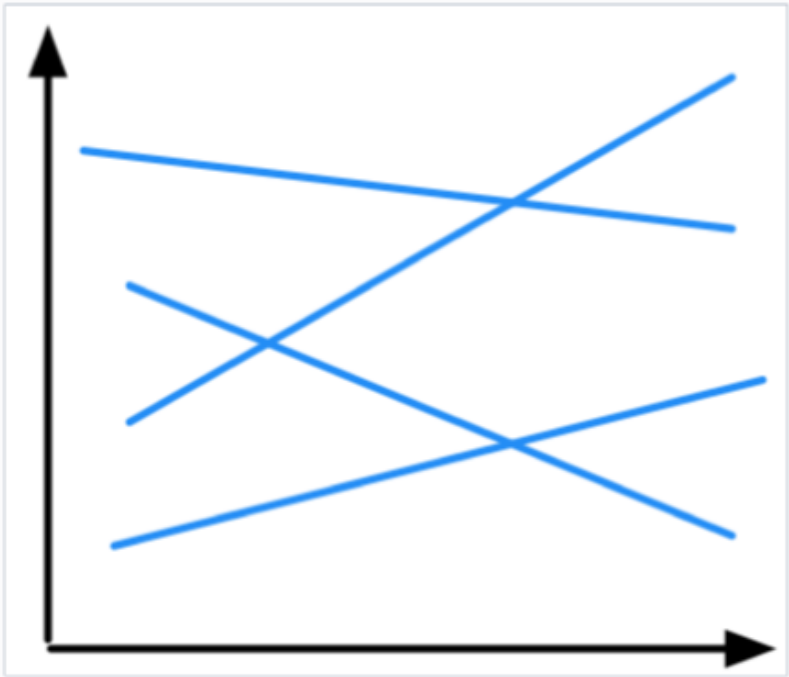
Classical least squares linear regression is a discriminative method where we aim to minimize the empirical mean squared error

$$R = \frac{1}{N} \sum_{n=1}^{N} (y_n - x_n^T w)^2.$$

Example 6: Prior Distribution

The first step to do this is to define a prior over the weights. We will use a zero-mean Gaussian with a fixed covariance matrix C:

$$p(w) = \mathcal{N}(w; 0, C) \tag{5}$$



Example 6: Likelihood

We next need to introduce a likelihood model based on these weights. We will make the standard assumption that the datapoints are independent of each other given the weights and again use a Gaussian to give

$$p(\mathbf{y}|\mathbf{x}, w) = \prod_{n=1}^{N} p(y_n|x_n, w) = \prod_{n=1}^{N} \mathcal{N}(y_n; x_n^T w, \sigma^2), \qquad (6)$$

where σ is a (fixed) standard deviation.

It is interesting to note that this likelihood is maximized by the least squares solution; this is a generalization of standard linear regression

Example 6: Posterior Distribution

We can now combine these to give the posterior using Bayes' rule:

$$p(w|\mathbf{x}, \mathbf{y}) \propto p(w)p(\mathbf{y}|\mathbf{x}, w) \tag{7}$$

$$= \mathcal{N}(w; 0, C) \prod_{n=1}^{N} \mathcal{N}(y_n; x_n^T w, \sigma^2) \tag{8}$$

We omit the necessary algebra (see C M Bishop. *Pattern recognition and machine learning*. 2006, Chapter 3), but it is reasonably straightforward to show that

$$p(w|\mathbf{x}, \mathbf{y}) = \mathcal{N}(w; m, S) \tag{9}$$

where $m = S^{-1}\mathbf{x}^T\mathbf{y}/\sigma^2$ and $S = \left(C^{-1} + \frac{\mathbf{x}^T\mathbf{x}}{\sigma^2}\right)^{-1}$.

Example 6: Prediction

Given this posterior, we can now calculate the posterior predictive as follows

$$p(\tilde{y}|\tilde{x}, \mathbf{x}, \mathbf{y}) = \int p(\tilde{y}|\tilde{x}, w) p(w|\mathbf{x}, \mathbf{y}) dw \tag{10}$$

$$= \int \mathcal{N}(\tilde{y}; \tilde{x}^T w, \sigma^2)\, \mathcal{N}(w; m, S)\, dw$$

$$= \mathcal{N}\left(\tilde{y}\,;\, \tilde{x}^T m,\, \left(\tilde{x}^T S^{-1} \tilde{x} + \frac{1}{\sigma^2}\right)^{-1}\right) \tag{11}$$

where the result is again a consequence of standard Gaussian identities and m and S are as before.

Gaussian Mixture Distribution and EM Algorithm

Review on Gaussian Distribution

If random variable X is Gaussian, it has the following PDF:

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The two parameters are μ, the mean, and σ^2, the variance (σ is called the standard deviation).

We'll use the terms "Gaussian" and "normal" interchangeably to refer to this distribution. To save us some writing, we'll write $p_X(x) = \mathcal{N}(x; \mu, \sigma^2)$ to mean the same thing (where the $\mathcal{N}$ stands for normal).

Maximum Likelihood Estimation

$$p_{X_1^n}(x_1^n) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_1-\mu)^2/2\sigma^2}$$

$$\ln p_{X_1^n}(x_1^n) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}(x_i - \mu)^2$$

$$\frac{d}{d\mu} \ln p_{X_1^n}(x_1^n) = \sum_{i=1}^{n} \frac{1}{\sigma^2}(x_i - \mu)$$

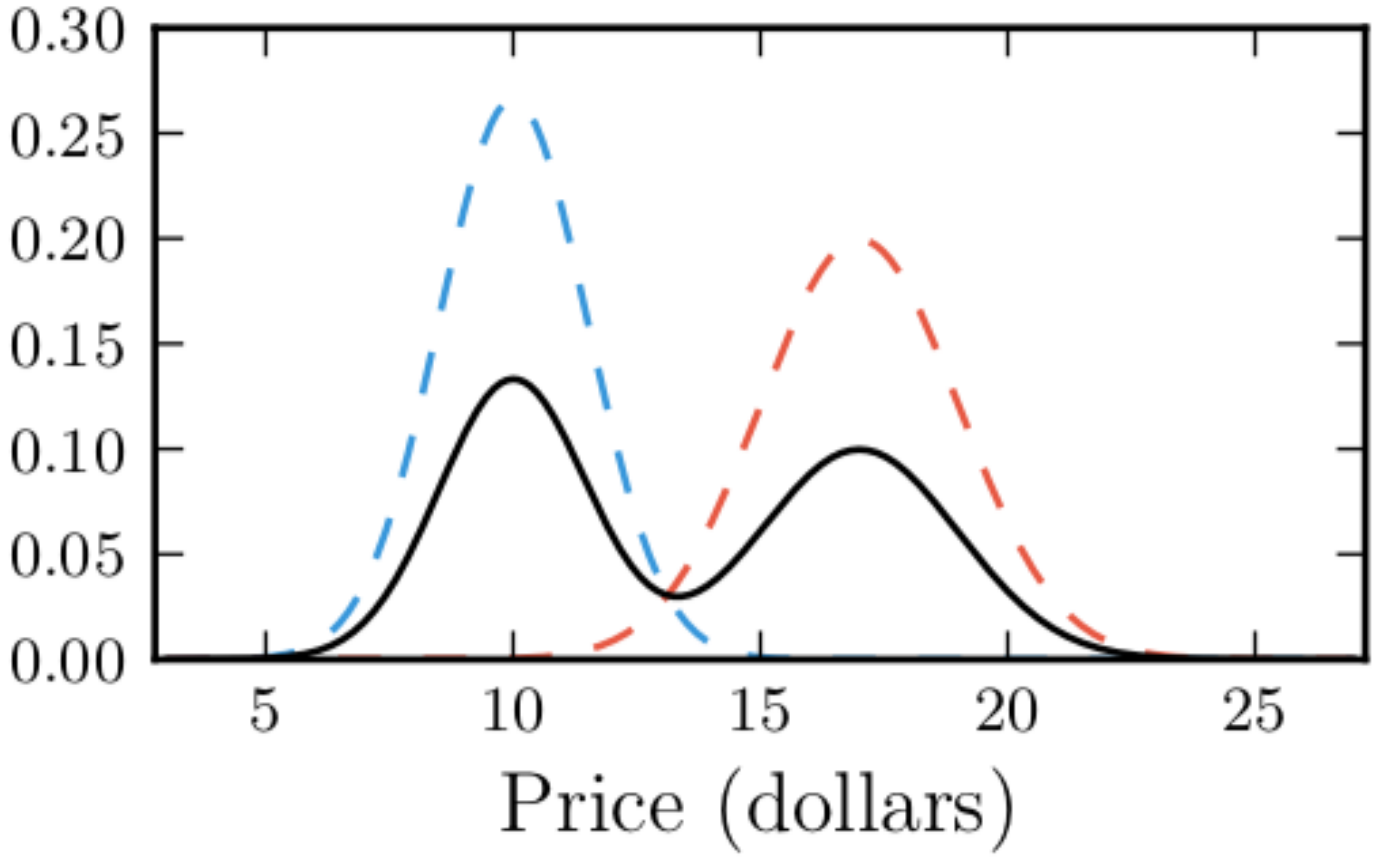
Setting this equal to 0, we see that the maximum likelihood estimate is $\widehat{\mu} = \frac{1}{N}\sum_i x_i$: it's the average of our observed samples. Notice that this estimate doesn't depend on the variance σ^2! Even though we started off by saying it was known, its value didn't matter.

Example 7: Gaussian Mixture Distribution

For example, suppose the price of a randomly chosen paperback book is normally distributed with mean \$10.00 and standard deviation \$1.00. Similarly, the price of a randomly chosen hardback is normally distributed with mean \$17 and variance \$1.50. Is the price of a randomly chosen book normally distributed?

The answer is no. Intuitively, we can see this by looking at the fundamental property of the normal distribution: it's highest near the center, and quickly drops off as you get farther away. But, the distribution of a randomly chosen book is bimodal: the center of the distribution is near \$13, but the probability of finding a book near that price is lower than the probability of finding a book for a few dollars more or a few dollars less. This is illustrated in Figure 1a.

Example 7: Illustration



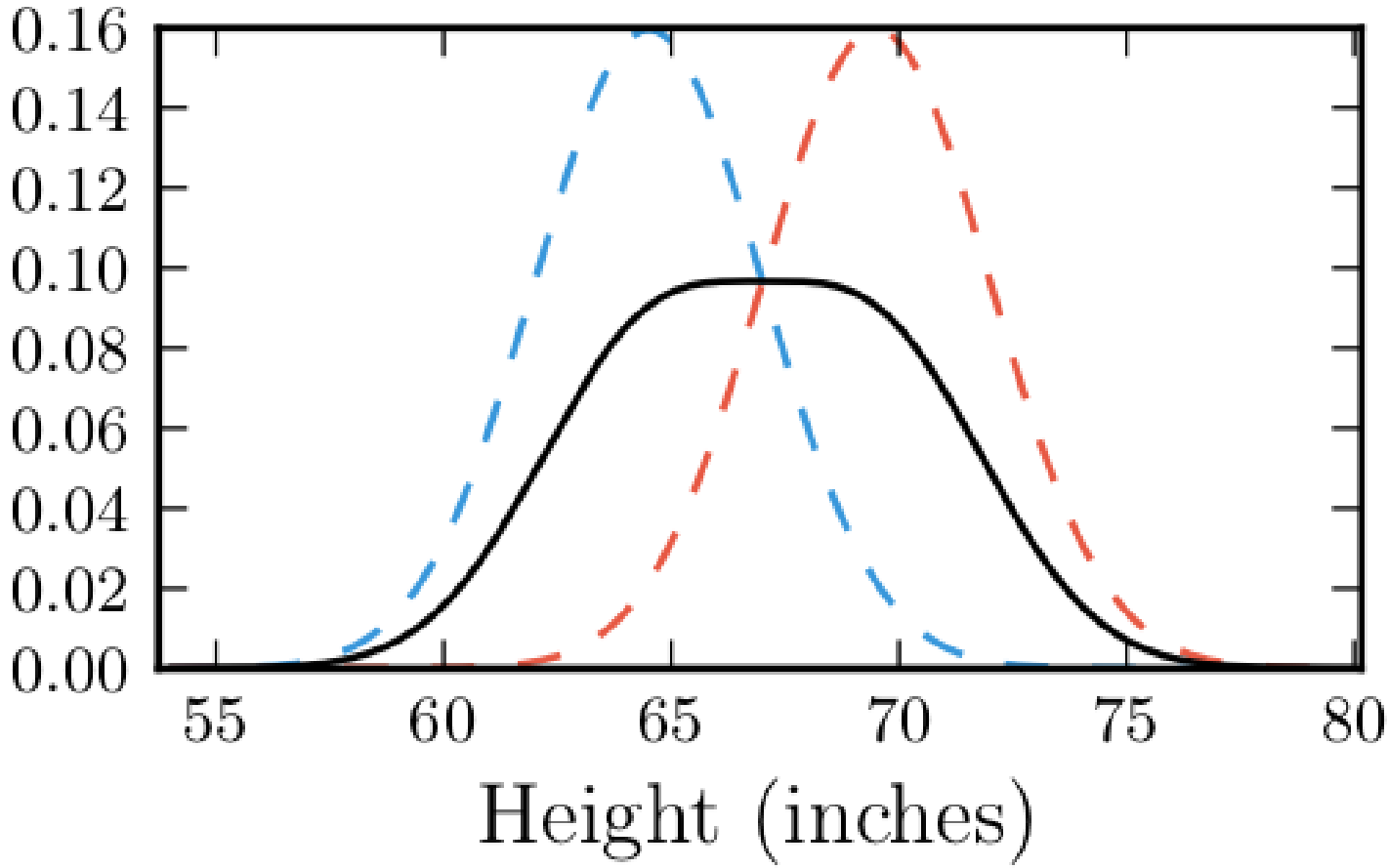
(a) Probability density for paperback books (red), hardback books (blue), and all books (black, solid)

Example 8: Gaussian Mixture Distribution

Another example: the height of a randomly chosen man is normally distributed with a mean around 5′9.5” and standard deviation around 2.5”. Similarly, the height of a randomly chosen woman is normally distributed with a mean around 5′4.5” and standard deviation around 2.5” [1] Is the height of a randomly chosen person normally distributed?

The answer is again no. This one is a little more deceptive: because there's so much overlap between the height distributions for men and for women, the overall distribution is in fact highest at the center. But it's still not normally distributed: it's too wide and flat in the center (we'll formalize this idea in just a moment). This is illustrated in Figure 1b. These

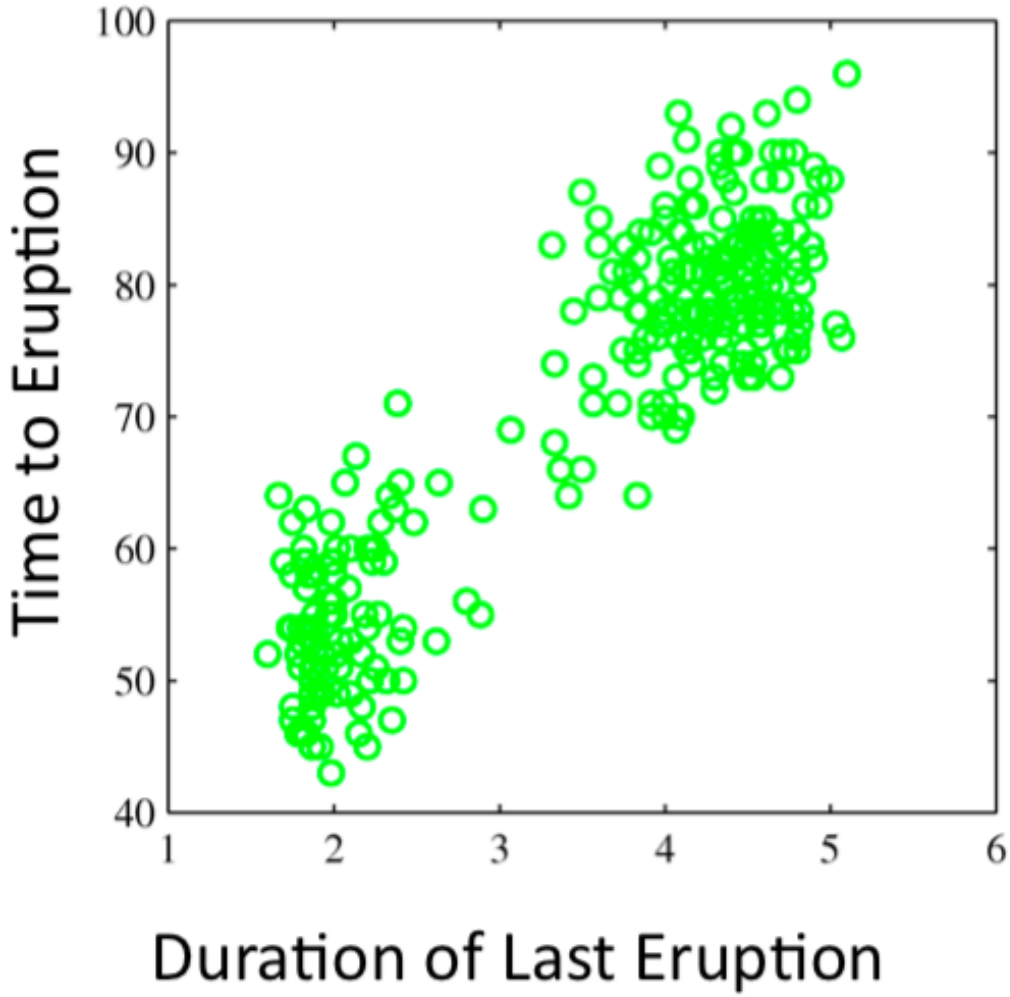
Example 8: Illustration



(b) Probability density for heights of women (red), heights of men (blue), and all heights (black, solid).

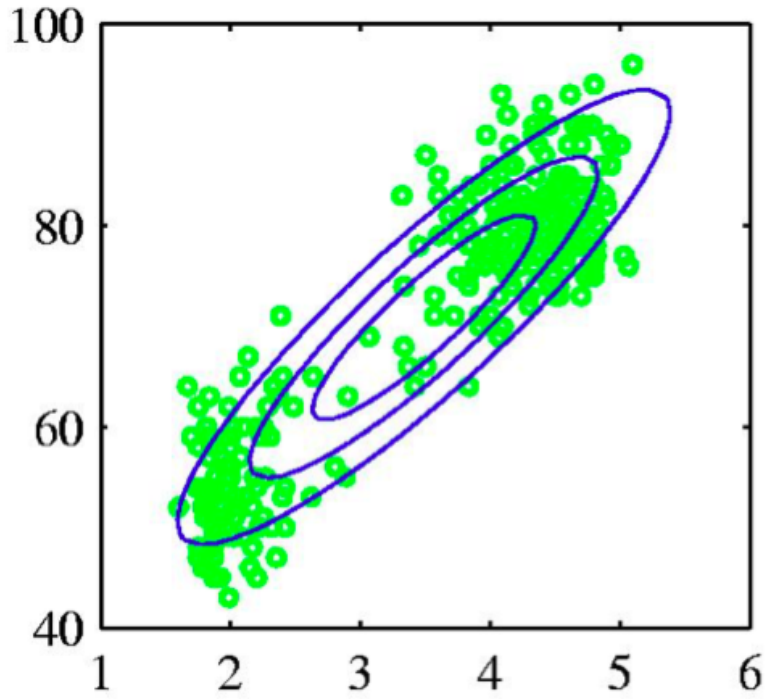
Example 9: Gaussian Mixture Distribution

Old Faithful Data Set

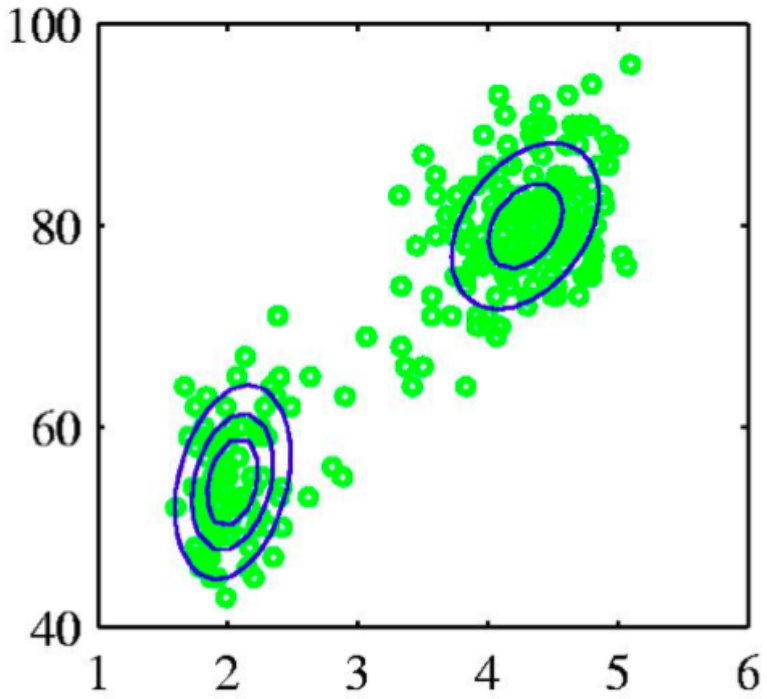


Example 9: Gaussian Mixture Distribution

Old Faithful Data Set



Single Gaussian



Mixture of two Gaussians

Gaussian Mixture Distribution

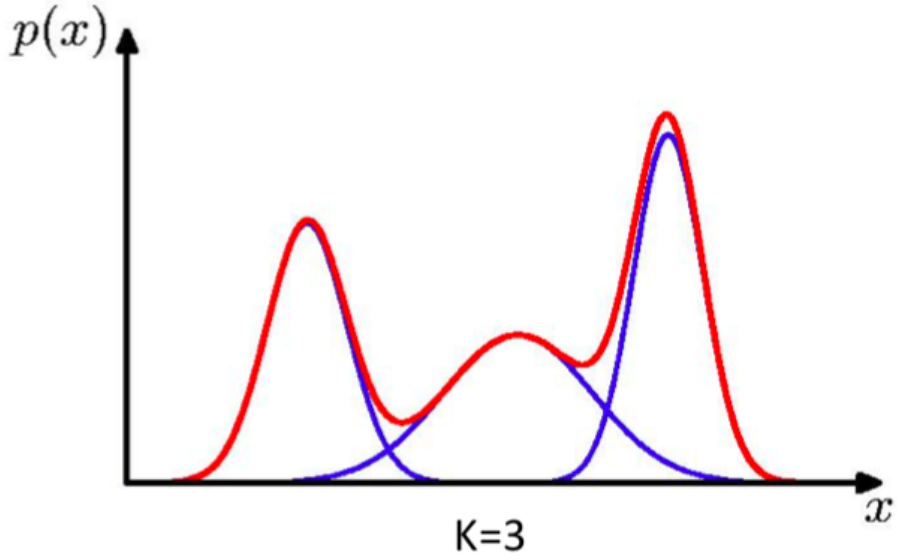
Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



Gaussian Mixture Distribution

Formally, suppose we have people numbered $i = 1, \ldots, n$. We observe random variable $Y_i \in \mathbb{R}$ for each person's height, and assume there's an unobserved label $C_i \in \{M, F\}$ for each person representing that person's gender [2]. Here, the letter c stands for "class". In general, we can have any number of possible labels or classes, but we'll limit ourselves to two for this example. We'll also assume that the two groups have the same known variance σ^2, but different unknown means μ_M and μ_F. The distribution for the class labels is Bernoulli:

$$p_{C_i}(c_i) = q^{\mathbb{1}(c_i=M)}(1-q)^{\mathbb{1}(c_i=F)}$$

We'll also assume q is known. To simplify notation later, we'll let $\pi_M = q$ and $\pi_F = 1 - q$, so we can write

$$p_{C_i}(c_i) = \prod_{c \in \{M,F\}} \pi_c^{\mathbb{1}(c_i=c)} \tag{1}$$

The conditional distributions within each class are Gaussian:

$$p_{Y_i|C_i}(y_i|c_i) = \prod_c \mathcal{N}(y_i; \mu_c, \sigma^2)^{\mathbb{1}(c_i=c)} \tag{2}$$

Challenge in MLE

Suppose we observe i.i.d. heights $Y_1 = y_1, \ldots, Y_n = y_n$, and we want to find maximum likelihood estimates for the parameters μ_M and μ_F. This is an *unsupervised learning* problem: we don't get to observe the male/female labels for our data, but we want to learn parameters based on those labels [3]

Exercise: Given the model setup in (1) and (2), compute the joint density of all the data points $p_{Y_1,\ldots,Y_N}(y_1, \ldots, y_n)$ in terms of μ_M, μ_F, σ, and q. Take the log to find the log-likelihood, and then differentiate with respect to μ_M. Why is this hard to optimize?

Challenge: Illustration

Let us start from the likelihood function.

$$
\begin{aligned}
p_{Y_i}(y_i) &= \sum_{c_i} p_{C_i}(c_i) p_{Y_i|C_i}(y_i|c_i) \\
&= \sum_{c_i} \left(\pi_c \mathcal{N}(y_i; \mu_C, \sigma^2)\right)^{\mathbb{1}(c_i=c)} \\
&= q\mathcal{N}(y_i; \mu_M, \sigma^2) + (1-q)\mathcal{N}(y_i; \mu_F, \sigma^2)
\end{aligned}
$$

Now, the joint density of all the observations is:

$$
p_{Y_1^n}(y_1^n) = \prod_{i=1}^{n} \left(q\mathcal{N}(y_i; \mu_M, \sigma^2) + (1-q)\mathcal{N}(y_i; \mu_F, \sigma^2)\right),
$$

and the log-likelihood of the parameters is then

$$
\ln p_{Y_1^n}(y_1^n) = \sum_{i=1}^{n} \ln \left(\pi_M \mathcal{N}(y_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(y_i; \mu_F, \sigma^2)\right), \tag{3}
$$

Challenge: Illustration

Before we dive into differentiating, we note that

$$\begin{aligned}\frac{d}{d\mu}\mathcal{N}(x;\mu,\sigma^2) &= \frac{d}{d\mu}\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right] \\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \frac{2(x-\mu)}{2\sigma^2} \\ &= \mathcal{N}(x;\mu,\sigma^2) \cdot \frac{(x-\mu)}{\sigma^2}\end{aligned}$$

Differentiating (3) with respect to μ_M, we obtain

$$\sum_{i=1}^{n} \frac{1}{\pi_M\mathcal{N}(y_i;\mu_M,\sigma^2) + \pi_F\mathcal{N}(y_i;\mu_F,\sigma^2)}\pi_M\mathcal{N}(y_i;\mu_M,\sigma^2)\frac{y_i-\mu_M}{\sigma^2} = 0 \tag{4}$$

At this point, we're stuck. We have a mix of ratios of exponentials and linear terms, and there's no way we can solve this in closed form to get a clean maximum likelihood expression!

Lucky Finding

Some term in (4) has another expression.

$$p_{C_i|Y_i}(c_i|y_i) = \frac{p_{Y_i|C_i}(y_i|c_i)p_{C_i}(c_i)}{p_{Y_i}(y_i)}$$
$$= \frac{\prod_{c\in\{M,F\}} \left(\pi_c \mathcal{N}(y_i;\mu_c,\sigma^2)\right)^{\mathbb{1}(c=c_i)}}{\pi_M \mathcal{N}(y_i;\mu_M,\sigma^2) + \pi_F \mathcal{N}(y_i;\mu_F,\sigma^2)} = q_{C_i}(c_i) \tag{5}$$

Let's look at the posterior probability that $C_i = M$:

$$p_{C_i|Y_i}(M|y_i) = \frac{\pi_M \mathcal{N}(y_i;\mu_M,\sigma^2)}{\pi_M \mathcal{N}(y_i;\mu_M,\sigma^2) + \pi_F \mathcal{N}(y_i;\mu_F,\sigma^2)} = q_{C_i}(M) \tag{6}$$

The Key Point: $q_{C_i}(c)$

This should look very familiar: it's one of the terms in (4)! And just like in that equation, we have to know all the parameters in order to compute this too. We can rewrite (4) in terms of q_{C_i}, and cheat a little by pretending it doesn't depend on μ_M:

$$\sum_{i=1}^{n} q_{C_i}(M) \frac{y_i - \mu_M}{\sigma^2} = 0 \tag{7}$$

$$\mu_M = \frac{\sum_{i=1}^{n} q_{C_i}(M) y_i}{\sum_{i=1}^{n} q_{C_i}(M)} \tag{8}$$

Intuitive Idea for A Possible Algorithm

- First, we fix the parameters (in this case, the means μ_M and μ_F of the Gaussians) and solve for the posterior distribution for the hidden variables (in this case, q_{C_i}, the class labels). This is done using (6).
- Then, we fix the posterior distribution for the hidden variables (again, that's q_{C_i}, the class labels), and optimize the parameters (the means μ_M and μ_F) using the expected values of the hidden variables (in this case, the probabilities from q_{C_i}). This is done using (4).

Formal Setting of Gaussian Mixture Distribution

Suppose we have observed a random variable Y. Now suppose we also have some hidden variable C that Y depends on. Let's say that the distributions of C and Y have some parameters θ that we don't know, but are interested in finding.

In our last example, we observed heights $Y = \{Y_1, \ldots, Y_n\}$ with hidden variables (gender labels) $C = \{C_1, \ldots, C_n\}$ (with i.i.d. structure over Y and C), and our parameters θ were μ_M and μ_F, the mean heights for each group.

Logic of EM Algorithm

$$\log p_Y(y;\theta) \tag{10}$$

(Marginalizing over C and introducing $q_C(c)/q_C(c)$) $$= \log\left(\sum_c q_C(c)\frac{p_{Y,C}(y,c;\theta)}{q_C(c)}\right)$$

(Rewriting as an expectation) $$= \log\left(\mathbb{E}_{q_C}\left[\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right]\right)$$

(Using Jensen's inequality) $$\geq \mathbb{E}_{q_C}\left[\log\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right] \tag{11}$$

Using definition of conditional probability $$= \mathbb{E}_{q_C}\left[\log\frac{p_Y(y;\theta)p_{C|Y}(C|y;\theta)}{q_C(C)}\right] \tag{12}$$

Now we have a lower bound on $\log p_Y(y;\theta)$ that we can optimize pretty easily. Since we've introduced q_C, we now want to maximize this quantity with respect to both θ and q_C.

We'll use (11) and (12), respectively, to do the optimizations separately. First, using (11) to find the best parameters:

$$\mathbb{E}_{q_C}\left[\log\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right] = \mathbb{E}_{q_C}\left[\log p_{Y,C}(y,C;\theta)\right] - \mathbb{E}_{q_C}[\log q_C(C)]$$

In general, q_C doesn't depend on θ, so we'll only care about the first term:

$$\boxed{\widehat{\theta} \leftarrow \operatorname*{argmax}_\theta \mathbb{E}_{q_C}\left[\log p_{Y,C}(y,C;\theta)\right]} \tag{13}$$

Logic of EM Algorithm

This is called the *M-step*: the M stands for maximization, since we're maximizing with respect to the parameters. Now, let's find the best q_C using (12).

$$\mathbb{E}_{q_C}\left[\log \frac{p_Y(y;\theta)p_{C|Y}(C|y;\theta)}{q_C(C)}\right] = \mathbb{E}_{q_C}[\log p_Y(y;\theta)] + \mathbb{E}_{q_C}\left[\log \frac{p_{C|Y}(C|y;\theta)}{q_C(C)}\right]$$

The first term doesn't depend on c, and the second term almost looks like a KL divergence:

$$\begin{aligned} &= \log p_Y(y;\theta) - \mathbb{E}_{q_C}\left[\log \frac{q_C(C)}{p_{C|Y}(C|y;\theta)}\right] \\ &= \log p_Y(y;\theta) - D(q_C(\cdot)||p_{C|Y}(\cdot|y;\theta)) \quad [6] \end{aligned} \tag{14}$$

So, when maximizing this quantity, we want to make the KL divergence as small as possible. KL divergences are always greater than or equal to 0, and they're exactly 0 when the two distributions are equal. So, the optimal q_C is $p_{C|Y}(c|y;\theta)$:

$$\boxed{\widehat{q}_C(c) \leftarrow p_{C|Y}(c|y;\theta)} \tag{15}$$

EM Algorithm

Inputs: Observation y, joint distribution $p_{Y,C}(y, c; \theta)$, conditional distribution $p_{C|Y}(c|y; \theta)$, initial values $\theta^{(0)}$

1: **function** EM$(p_{Y,C}(y, c; \theta), p_{C|Y}(c|y; \theta), \theta^{(0)})$
2: **for** iteration $t \in 1, 2, \ldots$ **do**
3: $q_C^{(t)} \leftarrow p_{C|Y}(c|y; \theta^{(t-1)})$ **(E-step)**
4: $\theta^{(t)} \leftarrow \text{argmax}_\theta \, \mathbb{E}_{q_C^{(t)}} [p_{Y,C}(y, C; \theta)]$ **(M-step)**
5: **if** $\theta^{(t)} \approx \theta^{(t-1)}$ **then**
6: return $\theta^{(t)}$

Model Averaging

A Useful Comment on Models

All models are wrong,
but some are useful
—George Box

Motivation of Model Averaging

- The purpose of a model is to help provide insights into a target problem or data and sometimes to further use these insights to make predictions
- Its purpose is **not** to try and fully encapsulate the "true" generative process or perfectly describe the data
- There are infinite different ways to generate any given dataset
  - Trying to uncover the "true" generative process is not even a well-defined problem
- In any real–world scenario, no Bayesian model can be "correct"
  - The posterior is inherently subjective
- It is still important to criticize—models can be very wrong!
  - E.g. we can use frequentist methods to falsify the likelihood

Model Averaging in Frequentist Inference

Let g be a (non-parametric) object of interest, such as a conditional mean, variance, density, or distribution function. Let $\hat{g}_m$, $m = 1, ..., M$ be a discrete set of estimators. Most commonly, this set is the same as we might consider for the problem of model selection. In linear regression, typically $\hat{g}_m$ correspond to different sets of regressors. We will sometimes call the m'th estimator the m'th "model".

Let w_m be a set of weights for the m'th estimator. Let $\mathbf{w} = (w_1, ..., w_M)$ be the vector of weights. Typically we will require

$$\begin{aligned} 0 &\leq w_m \leq 1 \\ \sum_{m=1}^{M} w_m &= 1 \end{aligned}$$

The set of weights satisfying this condition is H_M, the unit simplex in $\mathbb{R}^M$.

An averaging estimator is

$$\hat{g}(\mathbf{w}) = \sum_{m=1}^{M} w_m \hat{g}_m$$

It is commonly called a "model average estimator".

Selection estimators are the special case where we impose the restriction $w_m \in \{0, 1\}$.

Model Averaging for Linear Model

In the case of linear regression, let X_m be regressor matrix for the m'th estimator. Then the list of all regressors. Then the m'th estimator is

$$\begin{aligned}\hat{\beta}_m &= \left(X_m' X_m\right)^{-1} X_m y \\ \hat{g}_m &= X_m \hat{\beta}_m \\ &= P_m y\end{aligned}$$

where

$$P_m = X_m \left(X_m' X_m\right)^{-1} X_m$$

The averaging estimator is

$$\begin{aligned}\hat{g}(\mathbf{w}) &= \sum_{m=1}^{M} w_m \hat{g}_m \\ &= \sum_{m=1}^{M} w_m P_m y \\ &= P(\mathbf{w})\, y\end{aligned}$$

Estimation from Model Averaging

where

$$P(\mathbf{w}) = \sum_{m=1}^{M} w_m P_m$$

Let X be the matrix of all regressors. We can also write

$$\begin{aligned}
\hat{g}(\mathbf{w}) &= \sum_{m=1}^{M} w_m X_m \left(X_m' X_m\right)^{-1} X_m y \\
&= \sum_{m=1}^{M} w_m X_m \hat{\beta}_m \\
&= X \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\beta}_m \\ 0 \end{pmatrix} \\
&= X\hat{\beta}(\mathbf{w})
\end{aligned}$$

where

$$\hat{\beta}(\mathbf{w}) = \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\beta}_m \\ 0 \end{pmatrix}$$

Selection of Optimal Weights

As pointed out above, in the linear regression setting, $\hat{g}(\mathbf{w}) = P(\mathbf{w})\, y$ is a linear estimator, so falls in the class studied by Li (1987). His framework allows for estimators indexed by $\mathbf{w} \in H_M$

Under homoskedasticity, an optimal method for selection of $\mathbf{w}$ is the Mallows criterion. As we discussed before, for estimators $\hat{g}(\mathbf{w}) = P(\mathbf{w})\, y$, the Mallows criterion is

$$C(\mathbf{w}) = \hat{e}(\mathbf{w})' \hat{e}(\mathbf{w}) + 2\sigma^2 \operatorname{tr} P(\mathbf{w})$$

where

$$\hat{e}(\mathbf{w}) = y - \hat{g}(\mathbf{w})$$

is the residual.

Selection of Optimal Weights

In averaging linear regression

$$\begin{aligned}
\operatorname{tr} P(\mathbf{w}) &= \operatorname{tr} \sum_{m=1}^{M} w_m P_m \\
&= \sum_{m=1}^{M} w_m \operatorname{tr} P_m \\
&= \sum_{m=1}^{M} w_m k_m \\
&= \mathbf{w}'\mathbf{K}
\end{aligned}$$

where k_m is the number of coefficients in the m'th model, and $\mathbf{K} = (k_1, ..., k_M)'$. The penalty is twice $\mathbf{w}'\mathbf{K}$, the (weighted) average number of coefficients.

Also

$$\begin{aligned}
\hat{e}(\mathbf{w}) &= y - \hat{g}(\mathbf{w}) \\
&= \sum_{m=1}^{M} w_m (y - \hat{g}_m) \\
&= \sum_{m=1}^{M} w_m \hat{e}_m \\
&= \hat{\mathbf{e}}\mathbf{w}
\end{aligned}$$

Selection of Optimal Weights

where $\hat{e}_m$ is the $n \times 1$ residual vector from the m'th model, and $\hat{\mathbf{e}} = [\hat{e}_1, ..., \hat{e}_M]$ is the $n \times M$ matrix of residuals from all M models.

We can then write the criterion as

$$C(\mathbf{w}) = \mathbf{w}'\hat{\mathbf{e}}'\hat{\mathbf{e}}\mathbf{w} + 2\sigma^2\mathbf{w}'\mathbf{K}$$

This is quadratic in the vector $\mathbf{w}$.

The Mallows selected weight vector minimizes the criterion $C(\mathbf{w})$ over $\mathbf{w} \in H_M$, the unit simplex.

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in H_M}{\operatorname{argmin}}\, C(\mathbf{w})$$

This is a quadratic programming problem with inequality constraints, which is pre-programmed in Gauss and Matlab, so computation of $\hat{\mathbf{w}}$ is a simple command.

The Mallows selected estimator is then

$$\begin{aligned} \hat{g} &= \hat{g}(\hat{\mathbf{w}}) \\ &= \sum_{m=1}^{M} \hat{w}_m \hat{g}_m \end{aligned}$$

Linear Model under BMA

Let $\mathbf{X}$ denote the n by q matrix of all predictors under consideration. Under the full model (all predictors), the univariate multiple regression model is represented as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{u} \sim N(0, \sigma^2 I)$. In the problem of subset or variable selection among the q predictor variables, models under consideration correspond to potentially all possible subsets of the q variables, leading to a model space $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$ where $K = 2^q$ and includes the model with no predictor variables at all. Models for different subsets may be represented by a vector of binary variables, $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_q)'$ where γ_j is an indicator for inclusion of variable $\mathbf{X}_j$ under model $\mathcal{M}_k$. A convenient indexing of models in the all subset regression problem is to let $\boldsymbol{\gamma}$ denote the binary representation of k for model $\mathcal{M}_k$. Under $\mathcal{M}_k$ there are $q_{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \gamma_j$ non-zero parameters, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, with $q_{\boldsymbol{\gamma}} \times n$ design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$.

Linear Model under BMA

To incorporate model uncertainty regarding the choice of variables in the linear regression model, we build a hierarchical model (George and McCulloch 1993, 1997; Raftery et al. 1997):

$$\begin{aligned}
\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \mathcal{M}_k &\sim N(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 I_n) && (13.1)\\
\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \mathcal{M}_k &\sim p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathcal{M}_k, \sigma^2) && (13.2)\\
\sigma^2|\mathcal{M}_k &\sim p(\sigma^2|\mathcal{M}_k) && (13.3)\\
\mathcal{M}_k &\sim p(\mathcal{M}_k) && (13.4)
\end{aligned}$$

BMA Estimation

The posterior distribution over models in $\mathcal{M}$ is

$$p(\mathcal{M}_k|\mathbf{Y}) = \frac{m(\mathbf{Y}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_k m(\mathbf{Y}|\mathcal{M}_k)p(\mathcal{M}_k)}$$

where $m(\mathbf{Y}|\mathcal{M}_k)$ is the marginal distribution of the data under model $\mathcal{M}_k$,

$$m(\mathbf{Y}|\mathcal{M}_k) = \int\int p(\mathbf{Y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \mathcal{M}_k)p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \mathcal{M}_k)p(\sigma^2|\mathcal{M}_k)d\boldsymbol{\beta}_{\boldsymbol{\gamma}}d\sigma^2$$

obtained by integrating over the prior distributions for model specific parameters. The posterior distribution $p(\mathcal{M}_k|\mathbf{Y})$ provides a summary of model uncertainty after observing the data $\mathbf{Y}$. For the more general problem of model selection, the marginals are obtained similarly by integrating over all model specific parameters.

BMA Prediction

If Δ is a quantity of interest, say predicted values at a point x, then the expected value of Δ given the data $\mathbf{Y}$ is obtained by first finding the posterior expectation of Δ under each model, and then weighting each expectation by the posterior probability of the model:

$$E(\Delta|\mathbf{Y}) = \sum_k p(\mathcal{M}_k|\mathbf{Y})E(\Delta|\mathcal{M}_k, \mathbf{Y}). \tag{13.5}$$

Similarly, the posterior distribution for Δ can be represented as a mixture distribution over all models,

$$p(\Delta|\mathbf{Y}) = \sum_k p(\mathcal{M}_k|\mathbf{Y})p(\Delta|\mathbf{Y}, \mathcal{M}_k) \tag{13.6}$$

Conclusion

Understand

- Difference between Frequentist Inference and Bayesian Inference
- EM Algorithm for MLE approximation
- Model Averaging in Frequentist and Bayesian Inferences