Recap

- We are discussing Bayesian estimation of densities.
- We start with a prior distribution on the parameter and use the data to transform it into a posterior distribution.
- The posterior distribution is essentially the Bayesian estimate

Recap

- We have

$$f(\theta \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \theta) f(\theta)}{\int f(\mathcal{D} \mid \theta) f(\theta)\, d\theta}$$

  $f(\theta)$ – prior desity
  $f(\theta \mid \mathcal{D})$ – posterior density
  $f(\mathcal{D} \mid \theta) = \prod f(x_i \mid \theta)$ – data likelihood
- Conjugate Prior ensures that prior and posterior have same parametric form

Recap

- There are different options regarding using the posterior density as an estimate
- MAP estimate: $\hat{\theta}_{\text{MAP}} = \max_\theta f(\theta \mid \mathcal{D})$
- Mean of posterior can also be used as the estimate
- Or we can get the density model as

$$f(x \mid \mathcal{D}) = \int f(x \mid \theta) f(\theta \mid \mathcal{D}) \, d\theta$$

Recap

- We discussed Bayesian estimate for normal distribution with known variance. The density model is

$$f(x \mid \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- The conjugate prior is normal and the prior density is

$$f(\mu|\mu_0, \sigma_0) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

- As we saw, posterior is also normal:

$$f(\mu \mid \mathcal{D}) = \mathcal{N}(\mu_n, \sigma_n^2)$$

We calculated μ_n and σ_n.

Recap

$$\begin{aligned} \sigma_n^2 &= \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2} \\ \mu_n &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\, \bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\, \mu_0 \end{aligned}$$

($\bar{\mu}_n$ is the ML estimate)

- μ_n is a convex combination of $\bar{\mu}_n$ and μ_0. Both prior and data have a role to play.
- For large n, $\mu_n \approx \bar{\mu}_n$ and σ_n becomes very small.
- 'Large n' means $n\sigma_0^2 >> \sigma^2$

Recap

- We can take μ_n (which is the mean and mode of the posterior) as the estimate.
- Thus, we can use $\mathcal{N}(\mu_n, \sigma^2)$ as the estimated density.
- Or we can compute

$$f(x \mid \mathcal{D}) = \int_{-\infty}^{\infty} f(x \mid \mu) f(\mu \mid \mathcal{D}) \, d\mu$$

- We saw that this gives us $\mathcal{N}(\mu_n, \sigma^2 + \sigma_n^2)$ as the estimated density.

Recap

- We also considered estimating the parameter of the Bernoulli density.
- The conjugate prior here is the beta density.
- If we use Beta(a, b) as the prior then the posterior density is Beta$(\sum x_i + a,\ n + b - \sum x_i)$.

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{p}_B = \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}$$

- We can say we have started with $a + b$ 'fictitous' trials of which a were successes.
- This is how our 'prior beliefs' affect final estimate.
- As n becomes large, Bayes estimate is same as ML.

Density of Discrete Random variables

- Consider the example of estimating the mass function of a discrete random variable, Z, which can take one of M values, say, $a_1, \cdots, a_M$.
- Let $p_i = P[Z = a_i]$.
- We want to estimate $p_i,\ i = 1, \cdots, M$, given a sample of n *iid* realizations of Z.
- We have earlier seen how to get these estimates using maximum likelihood method.
- We will now derive the Bayesian estimates.

- As earlier, we represent any realization of Z by an M-dimensional Boolean vector,

$$X = [X^1, \cdots, X^M]^T,\ X^i \in \{0, 1\},\ \sum_i X^i = 1$$

  and we have $P[X^i = 1] = p_i$.
- Now the mass function for X can be written as

$$f(x \mid p) = \prod_{i=1}^{M} p_i^{x^i},$$

- As usual, the data is $\mathcal{D} = \{x_1, \cdots, x_n\}$.
- Note that
  $x_i = [x_i^1, \cdots, x_i^M]^T,\ x_i^j \in \{0, 1\},\ \sum_j x_i^j = 1,\ \forall i$
- Given such data, we want to estimate
  $p_i,\ i = 1, \cdots, M$.
- The first question is: what is the conjugate prior?
- For this, we should examine the form of the data likelihood.

The data likelihood is given by

$$
\begin{aligned}
f(\mathcal{D} \mid p) &= \prod_{i=1}^{n} f(x_i) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{M} p_j^{x_i^j} \\
&= \prod_{j=1}^{M} p_j^{n_j}, \quad \text{where} \quad n_j = \sum_i x_i^j
\end{aligned}
$$

- The likelihood is

$$f(\mathcal{D} \mid p) = \prod_{j=1}^{M} p_j^{n_j}, \quad \text{where} \quad n_j = \sum_i x_i^j$$

- Hence the prior density over p should have a form

$$f(p) \propto \prod_{j=1}^{M} p_j^{a_j}$$

(where $p = [p_1, \cdots, p_M]^T$ with $p_i \geq 0$ and $\sum_i p_i = 1$)

- Such a density is the Dirichlet density given by (note $p = [p_1, \cdots, p_M]^T$)

$$f(p) = \frac{\Gamma(a_1 + a_2 + \cdots + a_M)}{\Gamma(a_1) \cdots \Gamma(a_M)} \prod_{j=1}^{M} p_j^{a_j - 1},$$

  where $a_j \geq 1$ are the parameters of the density.
- The density is zero except for p that satisfy $p_i \geq 0$, $\sum_i p_i = 1$.

Dirichlet Density

- The Dirichlet density is

$$f(p) = \frac{\Gamma(a_1 + a_2 + \cdots + a_M)}{\Gamma(a_1) \cdots \Gamma(a_M)} \prod_{j=1}^{M} p_j^{a_j - 1},$$

- The Dirichlet density is the conjugate prior here.
- When $M = 2$ this density becomes the Beta density.

Dirichlet Density

- Suppose $p_1, \cdots, p_M$ have joint density that is Dirichlet with parameters a_j. Then

$$\begin{aligned} \text{the mode is at} \quad p_j &= \frac{a_j - 1}{a_0 - M} \\ E[p_j] &= \frac{a_j}{a_0} \\ \text{Var}[p_j] &= \frac{a_j(a_0 - a_j)}{a_0^2(a_0 + 1)} \\ \text{Cov}(p_i, p_j) &= -\frac{a_i a_j}{a_0^2(a_0 + 1)} \end{aligned}$$

where $a_0 = a_1 + \cdots + a_M$.

- Now, taking the prior as Dirichlet, the posterior density can be obtained as

$$\begin{aligned} f(p \mid \mathcal{D}) &\propto f(\mathcal{D} \mid p)\, f(p) \\ &\propto \prod_{j=1}^{M} p_j^{n_j} \prod_{j=1}^{M} p_j^{a_j - 1} \\ &\propto \prod_{j=1}^{M} p_j^{n_j + a_j - 1} \end{aligned}$$

- Thus posterior is Dirichlet with parameters $n_j + a_j$.

- The posterior is Dirichlet with parameters $n_j + a_j$.
- If we want the MAP estimate,

$$\hat{p}_j = \frac{n_j + a_j - 1}{n + a_0 - M}$$

- Recall that the MLE for p_j is $\frac{n_j}{n}$.
- If $a_j = 1, \forall j$, then MAP estimate is same as MLE. (Flat prior).

- As earlier, we can calculate $f(x|\mathcal{D})$.
- Recall that x takes only M values such as $[1\ 0\ \cdots]^T, [0\ 1\ 0\ \cdots]^T$ etc.
- Call them $e_1, e_2, \cdots, e_M$.
- $f(x|\mathcal{D})$ gives us $f(z|\mathcal{D})$: $P[Z = b_j] = P[X = e_j]$.

- Now we get

$$\begin{aligned} P[X = e_j] &= \int P[X = e_j|p]\ f(p|\mathcal{D})dp \\ &= \int p_j\ f(p|\mathcal{D})dp \\ &= Ep_j \\ &= \frac{n_j + a_j}{n + a_0} \end{aligned}$$

- Hence we can take the mean of posterior as our final Bayesian estimate:

$$\hat{p}_j = \frac{n_j + a_j}{n + a_0}$$

- Recall that the ML estimate is $\hat{p}_j = \frac{n_j}{n}$.
- Our choice of prior decides on the values of a_j.
- The nature of the Bayesian estimate is same as in the case on Bernoulli.

Estimating variance of a Gaussian

- Consider estimating variance of a normal distribution with mean known.
- We take $\nu = \frac{1}{\sigma^2}$ as the parameter. (ν is called precision). Then the density model is

$$f(x \mid \nu) = \frac{\sqrt{\nu}}{\sqrt{2\pi}} \exp\left(-\frac{\nu}{2}(x-\mu)^2\right)$$

  where we assume μ is known.
- Note that we have $\nu > 0$.

- Now the likelihood is given by

$$\begin{aligned} f(\mathcal{D} \mid \nu) &= \prod_{i=1}^{n} f(x_i \mid \nu) \\ &= (2\pi)^{-\frac{n}{2}} \nu^{\frac{n}{2}} \exp\left(-\frac{\nu}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right) \end{aligned}$$

- Hence conjugate prior should be proportional to product of a power of ν and exponential of a linear function of ν.
- Such a prior would be the gamma density:

$$f(\nu) = \frac{1}{\Gamma(a)} b^a \nu^{a-1} e^{-b\nu}, \ \nu \geq 0$$

where a, b are parameters of the gamma density.

- The mean of gammadensity is $\frac{a}{b}$ and its mode is $\frac{a-1}{b}$.
- We take the prior as gamme density with parameters a_0, b_0.

Now we can get the posterior density as

$$
\begin{aligned}
f(\nu \mid \mathcal{D}) &\propto f(\mathcal{D} \mid \nu)\, f(\nu) \\
&\propto \nu^{\frac{n}{2}} \exp\left(-\frac{\nu}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \nu^{a_0-1}\exp(-b_0\nu) \\
&\propto \nu^{a_0+\frac{n}{2}-1} \exp\left(-b_0\nu - \frac{\nu}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)
\end{aligned}
$$

- As expected, the posterior density is gamma.

- Thus the posterior density for ν is gamma with parameters a_n and b_n where

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= b_0 + \frac{n}{2}\,\hat{\sigma}^2_{\mathrm{ML}}, \quad \text{where } \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

- Recall that $\hat{\sigma}^2_{\mathrm{ML}}$ is the ML estimate for variance.

- If we take mean of the posterior as our final estimate then

$$\hat{\nu} = \frac{a_0 + \frac{n}{2}}{b_0 + \frac{n}{2}\,\hat{\sigma}^2_{\mathrm{ML}}}$$

- The a_0 and b_0 are determined by our choice of prior.
- As $n \to \infty$, we have $\hat{\nu} \to \left(\hat{\sigma}^2_{\mathrm{ML}}\right)^{-1}$
- Also, note that the variance of the posterior, a_n/b_n^2, goes to zero as $n \to \infty$.

- We can once again calculate $f(x|\mathcal{D})$. We have

$$f(x|\mathcal{D}) = \int f(x|\nu)\, f(\nu|\mathcal{D})\, d\nu$$

- Here, x is gaussian with precision ν (with known mean, μ) and the posterior of ν is Gamma with parameters a_n, b_n.
- Thus, $f(x|\mathcal{D})$ would be a density that depends on μ and the parameters of the gamma density and is given by

$$\begin{aligned} f(x|\mu, a, b) &= \int_0^\infty \left(\frac{\nu}{2\pi}\right)^{0.5} \exp\left(-0.5\nu(x-\mu)^2\right) \\ &\qquad \frac{1}{\Gamma(a)} b^a \nu^{a-1} \exp(-b\nu)\, d\nu \end{aligned}$$

- One can show this to be (with $\tau = 2a$ and $\lambda = a/b$)

$$f(x|\mu, \lambda, \tau) = \frac{\Gamma(0.5(1+\tau)}{\Gamma(0.5\tau)} \sqrt{\frac{\lambda}{\pi\tau}} \left(1 + \frac{\lambda(x-\mu)^2}{\tau}\right)^{-0.5(\tau+1)}$$

- This is called Student's t-distribution. It has heavier tails than Gaussian.
- As $\tau \to \infty$, it becomes Gaussian with mean μ and precision λ.

- We looked at Bayesian estimation either for the mean or for the variance of a Gaussian.
- Suppose both are unknown. We write the density model as

$$f(x \mid \mu, \nu) = \frac{\sqrt{\nu}}{\sqrt{2\pi}} \exp\left(-\frac{\nu}{2}(x-\mu)^2\right)$$

- Now the prior needed is a joint density on μ, ν.

- Then the conjugate prior would be a Gaussian-Gamma density.

$$\begin{aligned} f(\mu, \nu) &= f(\nu)\, f(\mu \mid \nu) \\ &= \nu^{a_0 - 1} \exp(-b_0 \nu) \exp\left(-\frac{c_0 \nu}{2} (\mu - \mu_0)^2\right) \end{aligned}$$

- That is, the marginal for ν is a gamma density and the conditional density of μ conditioned on ν is Gaussian.
- The algebra is a little more complicated; but final estimates are similar.

- So far we have been considering only one dimensional case.
- Consider data from multidimensinal Gaussian with Σ known.
- Then the conjugate prior would be $f(\mu) = \mathcal{N}(\mu_0, \Sigma_0)$.
- By same techniques as earlier, the posterior gaussian density can be obtained.

- Now consider the case where μ is known and Σ is unknown. We take $\Lambda = \Sigma^{-1}$ as the parameter.
- Now the conjugate prior would be Wishart distribution:

$$\mathcal{W}(\Lambda|W,\nu) = B|\Lambda|^{(\nu-d-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(W^{-1}\Lambda)\right)$$

  where B is the normalizing constant.
- The density is zero except for Λ that are symmetric and positive definite.

- If x_i are iid $\mathcal{N}(0, \Sigma)$, then $\sum_i^n x_i x_i^T$ has Wishart distribution with $W = \Sigma$ and $\nu = n$
- ν is called the degrees of freedom and W is called the scale matrix.
- For a general case of Gaussian density with unknown mean and covariance matrix, the conjugate prior is a Gauss-Wishart distribution!

- We can similarly obtain Bayesian estimates for many standard densities.
- As we saw, the conjugate prior would depend on form of $f(x \mid \theta)$.
- The procedure is a little more involved than ML method.
- As we saw through examples, prior allows us to incorporate any knowledge we have of the parameter and the Bayesian method also allows us to take care of small sample cases.

Exponential family of densities

- Exponential family:
  any density with a (vector) parameter η

$$\begin{aligned} f(x \mid \eta) &= h(x)\, g(\eta)\, \exp(\eta^T u(x)) \\ &= \exp\left[\eta^T u(x) + \ln(h(x)) + \ln(g(\eta))\right] \end{aligned}$$

  where $u(x)$ is, in general, a vector function.
- Many standrad densities such as Bernoulli, binomial, poisson, gamma, beta, Gaussian etc can be put in this form

Examples of exponential family

- Consider the Bernoulli distribution

$$
\begin{aligned}
f(x \mid p) &= p^x (1-p)^{1-x} \\
&= \exp\left[\ln\left(p^x (1-p)^{1-x}\right)\right] \\
&= \exp[x\, ln(p) + (1-x)\, \ln(1-p)] \\
&= (1-p)\, \exp\left[x\, \ln\, \frac{p}{1-p}\right] \\
&= \frac{1}{1+\frac{p}{1-p}}\, \exp\left[x\, \ln\, \frac{p}{1-p}\right] \\
&= \frac{1}{1+\exp(\eta)}\, \exp[\eta\, x], \quad \eta = \ln\left(\frac{p}{1-p}\right)
\end{aligned}
$$

- Thus the Bernoulli mass function can be written as

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

where $\eta = \ln \frac{p}{1-p}$, and

$$h(x) = 1, \quad g(\eta) = \frac{1}{1 + \exp(\eta)} \quad \text{and } u(x) = x$$

- Thus Bernoulli belongs to the exponential family.
- Sometimes, this η is called the 'natural parameter' for Bernoulli.

Similarly, for Gaussian density, we can show

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \\ &= \frac{1}{\sqrt{\pi}} \sqrt{-\eta_2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \exp(\eta_1 u_1(x) + \eta_2 u_2(x)) \end{aligned}$$

where

$$\eta_1 = \frac{\mu}{\sigma^2} \quad \eta_2 = -\frac{1}{2\sigma^2} \quad u_1(x) = x \quad u_2(x) = x^2$$

- This is once again in the form $h(x)\, g(\eta)\, \exp(\eta^T u(x))$.
- Similarly we can show that many standard densities belong to the exponential family of densities.

- Consider ML estimation of the parameter vector η for a density from the exponential family

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

- The data likelihood would be

$$f(\mathcal{D} \mid \eta) = \left(\prod_{i=1}^{n} h(x_i) \right) (g(\eta))^n \exp \left[\eta^T \sum_{i=1}^{n} u(x_i) \right]$$

- The loglikelihood is

$$l(\eta \mid \mathcal{D}) = K + n \ln(g(\eta)) + \eta^T \sum_{i=1}^{n} u(x_i)$$

  Where K is a term that does not depend on η.
- To find η to maximize log likelihood we need to equate gradient (with respect to η) to zero:

$$n \nabla_\eta \ln(g(\hat{\eta})) + \sum_{i=1}^{n} u(x_i) = 0$$

- Hence, the ML estimate, $\hat{\eta}_{\mathrm{ML}}$, satisfies

$$-\nabla_\eta \ln(g(\hat{\eta}_{\mathrm{ML}})) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)$$

or

$$-\frac{1}{g(\hat{\eta}_{\mathrm{ML}})}\nabla_\eta g(\hat{\eta}_{\mathrm{ML}}) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)$$

- Thus we have simple and uniform procedure for estimating all densities in the exponential family.

- For exponential family, the density is given by

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

- Since $\int f(x \mid \eta) = 1$, we have

$$g(\eta) \int h(x)\, \exp(\eta^T u(x))\, dx = 1$$

- By differentiating the above w.r.t η, we get

$$\nabla g(\eta) \int h(x)\, \exp(\eta^T u(x))\, dx +$$

$$\int \left(g(\eta) h(x)\, \exp(\eta^T u(x))\right) u(x)\, dx \quad = \quad 0$$

$$\Rightarrow \quad \nabla g(\eta) \frac{1}{g(\eta)} + E[u(x)] = 0$$

- Thus, for an exponential family density, the η satisfies

$$\nabla g(\eta) \frac{1}{g(\eta)} = -E[u(x)]$$

- We saw that the ML estimate satisfies

$$\frac{1}{g(\hat{\eta}_{\mathrm{ML}})} \nabla_\eta g(\hat{\eta}_{\mathrm{ML}}) = -\frac{1}{n} \sum_{i=1}^{n} u(x_i)$$

- Shows that ML estimate is consistent for exponential family densities.

- For all exponential family densities, we can also easily find a conjugate prior.
- Recall the density is

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

- The conjugate prior for η can be written as

$$f(\eta|a, b) = h_1(a, b) g(\eta)^b \exp(b\eta^T a)$$

(the vector a and scalar b are hyperparameters)

- We can easily see this to be conjugate prior. We have

$$f(\mathcal{D}|\eta) = \left(\prod_{i=1}^{n} h(x_i)\right) (g(\eta))^n exp\left(\eta^T (\sum_{i=1}^{n} u(x_i))\right)$$

- The prior is

$$f(\eta|a,b) = h_1(a,b) g(\eta)^b \exp(b\eta^T a)$$

Hence

$$f(\mathcal{D}|\eta) f(\eta) \propto (g(\eta))^{n+b} \exp\left(\eta^T (\sum_{i=1}^{n} u(x_i) + ba)\right)$$

- Thus posterior is of the same form as prior.

- Consider Bernoulli:
  $g(\eta) = (1 + e^{\eta})^{-1}$; $u(x) = x$; $\eta = \ln(p/(1-p))$
- Since $e^{\eta} = p/(1-p)$, we get $\exp(c\eta) = (p/(1-p))^{c}$
- Note $g(\eta) = \left(1 + \frac{p}{1-p}\right)^{-1} = (1-p)$.
- Thus

$$\begin{aligned} f(\eta|a,b) &\propto (g(\eta))^{b} \exp(b\eta a) \\ &= (1-p)^{b} \left(\frac{p}{1-p}\right)^{ab} \\ &= (1-p)^{b(1-a)} p^{ab} \end{aligned}$$

- This is the beta distribution

- The ML estimate for all exponential family satisfies

$$-\frac{1}{g(\hat{\eta}_{\mathrm{ML}})}\nabla_{\eta} g(\hat{\eta}_{\mathrm{ML}}) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)$$

- There is an interesting aspect of these equations.
- To obtain the ML estimate we do not explicitly need all the data.
- We need only $\sum u(x_i)$.
- We say that such $\sum u(x_i)$ are **sufficient statistic** for η.

Sufficient Statistic

- A statistic is any function of the data, $\mathcal{D} = \{x_1, \cdots, x_n\}$
- A statistic S is said to be **sufficient** for parameter θ if $f(\mathcal{D} \mid S, \theta)$ is not a function of θ.
- That is, the conditional density of data, given S is not dependent on θ.
- As we saw, in the ML estimation we do not need all data explicitly; the sufficient statistic would do.

- In the Bayesian framework, if S is sufficient, then $f(\mathcal{D} \mid S, \theta) = f(\mathcal{D} \mid S)$.
- The posterior density now is

$$\begin{aligned} f(\theta \mid S, \mathcal{D}) &= \frac{f(\mathcal{D} \mid S, \theta)\, f(\theta \mid S)}{f(\mathcal{D} \mid S)} \\ &= f(\theta \mid S) \end{aligned}$$

- Thus if we are given S we do not need data $\mathcal{D}$ (when S is a sufficient statistic).

Example of sufficient statistic

- Consider Poisson distribution with parameter λ.

$$f(x \mid \lambda) = \frac{1}{x!}\, \lambda^x\, e^{-\lambda}$$

- Let $S = \sum_i x_i$ be a statistic. Let $s = S(\mathcal{D})$.
- Let us show S is sufficient for λ.
- For this we need to look at $f(\mathcal{D} \mid S, \lambda)$.
- To be precise in our notation, we will denote this conditional mass function as $f_{\mathbf{X}|S}(\mathbf{x} \mid s)$.

$$
\begin{aligned}
f_{\mathbf{X}|S}(\mathbf{x} \mid s) &= \frac{P[X_1 = x_1, \cdots, X_n = x_n, S = s]}{P[S = s]} \\
&= \frac{P[X_1 = x_1, \cdots, X_n = x_n, \sum_{i=1}^{n} X_i = s]}{P[S = s]} \\
&= \frac{P[X_1 = x_1, \cdots, X_{n-1} = x_{n-1}, X_n = s - \sum_{i=1}^{n-1} x_i]}{P[S = s]} \\
&= \frac{P[X_1 = x_1, \cdots, X_{n-1} = x_{n-1}, X_n = s - \sum_{i=1}^{n-1} x_i]}{\sum_{x_i} P[X_1 = x_1, \cdots, X_{n-1} = x_{n-1}, X_n = s - \sum_{i=1}^{n-1} x_i}
\end{aligned}
$$

We now have

$$
\begin{aligned}
& P\left[X_1 = x_1, \cdots, X_{n-1} = x_{n-1}, X_n = s - \sum_{i=1}^{n-1} x_i\right] \\
&= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_{n-1}}}{x_{n-1}!} \frac{e^{-\lambda}\lambda^{s-\sum_{i=1}^{n-1} x_i}}{(s-\sum_{i=1}^{n-1} x_i)!} \\
&= e^{-n\lambda} \frac{1}{x_1! \cdots x_{n-1}!} \lambda^{\sum_{i=1}^{n-1} x_i} \frac{1}{(s-\sum_{i=1}^{n-1} x_i)!} \lambda^{(s-\sum_{i=1}^{n-1} x_i)} \\
&= e^{-n\lambda}\, \lambda^s\, h(\mathbf{x})
\end{aligned}
$$

where $h(\mathbf{x})$ is a term that depends on x_i but not on λ.

Putting this in the earlier equation

$$\begin{aligned} f_{\mathbf{X}|S}(\mathbf{x} \mid s) &= \frac{e^{-n\lambda}\, \lambda^{s}\, h(\mathbf{x})}{\sum_{x_i} e^{-n\lambda}\, \lambda^{s}\, h(\mathbf{x})} \\ &= \frac{h(\mathbf{x})}{\sum_{x_i} h(\mathbf{x})} \end{aligned}$$

which is not dependent on λ.

- This shows that $\sum x_i$ is a sufficient statistic for λ in a Poisson distribution.

Factorization Theorem

- The following theorem characterizes a sufficient statistic.
- **Theorem**: A statistic S is sufficient for θ if and only if the likelihood function can be factorized as

$$f(\mathcal{D} \mid \theta) = g(s, \theta)\, h(\mathcal{D}), \qquad \text{where} \quad s = S(\mathcal{D})$$

- In our example, we saw how similar factorization happens when S is a sufficient statistic.

- Consider any density in the exponential family

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

- Now

$$f(\mathcal{D} \mid \eta) = \left[(g(\eta))^n \exp(\eta^T \sum_i u(x_i)) \right] \prod_i h(x_i)$$

- Thus, if we take the (vector) statistic as $S = \sum_i u(x_i)$, we have the needed factorization.

Proof of Factorization Theorem

- We now sketch the proof of factorization theorem when X is a discrete random variable.
- Let the data be $\mathcal{D} = \{x_1, \cdots, x_n\}$.
- Let $s = S(x_1, \cdots, x_n)$.
- First, assume that S is sufficient for θ. We show that the likelihood function factorizes as needed.

We note that, since $s = S(x_1, \cdots, x_n)$,

$$P[X_1 = x_1, \cdots, X_n = x_n] = P[X_1 = x_1, \cdots, X_n = x_n, S = s]$$

$$
\begin{aligned}
f(\mathcal{D} \mid \theta) &= P[X_1 = x_1, \cdots, X_n = x_n \mid \theta] \\
&= P[X_1 = x_1, \cdots, X_n = x_n, S = s \mid \theta] \\
&= P[S = s|\theta]\, P[X_1 = x_1, \cdots, X_n = x_n|S = s, \theta] \\
&= P[S = s \mid \theta]\, P[X_1 = x_1, \cdots, X_n = x_n \mid S = s] \\
&\qquad\qquad \text{(since } S \text{ is sufficient)} \\
&= g(s,\, \theta)\, h(\mathcal{D})
\end{aligned}
$$

This completes the proof of first part of the theorem.

- Now we assume that the likelihood function can be factorized and show that S is sufficient.
- As earlier, since $s = S(x_1, \cdots, x_n)$, we have

$$P[X_1 = x_1, \cdots, X_n = x_n] = P[X_1 = x_1, \cdots, X_n = x_n, S = s]$$

- To show S is sufficient, we have to show that $f(\mathcal{D} \mid s, \theta)$ is not dependent on θ.

$$
\begin{aligned}
f(\mathcal{D} \mid s, \theta) &= \frac{P[X_1 = x_1, \cdots, X_n = x_n, S = s \mid \theta]}{P[S = s \mid \theta]} \\
&= \frac{P[X_1 = x_1, \cdots, X_n = x_n, S = s \mid \theta]}{\sum_{x_i} P[X_1 = x_1, \cdots, X_n = x_n, S = s \mid \theta]}
\end{aligned}
$$

where the summation is only over those $x_1, \cdots, x_n$ such that $S(x_1, \cdots x_n) = s$

- Since the likelihood function factorizes, we have

$$
P[X_1 = x_1, \cdots, X_n = x_n, S = s \mid \theta] = g(s, \theta)\, h(\mathcal{D})
$$

- Hence we get

$$\begin{aligned} f(\mathcal{D} \mid s, \theta) &= \frac{g(s, \theta)\, h(\mathcal{D})}{\sum_{x_i} g(s, \theta)\, h(\mathcal{D})} \\ &= \frac{h(\mathcal{D})}{\sum_{x_i} h(\mathcal{D})} \end{aligned}$$

  which is not dependent on θ, thus showing that S is sufficient.
- This completes the proof of the theorem.

- Sufficient statistics give us good compression of the data (for parameter estimation).
- The factorization we talked about is not unique. For example, given any $g_1(s)$,

$$\begin{aligned} f(\mathcal{D} \mid \theta) &= g(s, \theta)\, h(\mathcal{D}) \\ &= [g_1(s)\, g(s, \theta)]\, \frac{h(\mathcal{D})}{g_1(s)} \end{aligned}$$

- Often, to avoid this, one takes $\tilde{g}(s, \theta) = \frac{g(s,\theta)}{\int\, g(s,\theta)\, d\theta}$.
- Sufficient statistics are also useful in finding UMVUE.

Recursive estimates

- So far, when we derived ML or Bayesian estimates, we essentially assumed that all data is with us.
- Thus these are 'batch' versions of the estimation methods.
- We could also have these estimates in an 'incremental' or 'recursive' fashion.
- Here we assume that we get data samples one-by-one and we do not store all data.
- Using estimate after $n-1$ samples and the n^{th} sample, we derive the estimate with n samples.

- For example consider the sample mean which is the ML estimate for mean of a, e.g., normal density. We can rewrite this as

$$\begin{aligned}
\hat{\mu}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n-1} x_i\right) + \frac{1}{n}x_n \\
&= \frac{n-1}{n}\hat{\mu}_{n-1} + \frac{1}{n}x_n
\end{aligned}$$

- Most ML estimates can be written in such a recursive manner.
- We can actually get a kind of general recursive form for ML estimates.
- We can rewrite the sample mean estimate as

$$\begin{aligned}\hat{\mu}_n &= \frac{n-1}{n}\,\hat{\mu}_{n-1} + \frac{1}{n}\,x_n \\ &= \hat{\mu}_{n-1} + \frac{1}{n}\,(x_n - \hat{\mu}_{n-1})\end{aligned}$$

  which is like an 'error-correcting' or 'optimization' algorithm.
- We can view the above as a gradient descent optimization algorithm.

- We can get the expectation through the optimization problem: $\min_\theta \ \frac{1}{2}E(x-\theta)^2$
- A gradient descent algorithm for this is:

$$\theta_n = \theta_{n-1} - \eta\nabla\frac{1}{2}\left(E(x-\theta_{n-1})^2\right) = \theta_{n-1} + \eta E[x-\theta_{n-1}]$$

- But we do not know the expectation. We only have a 'noisy' version of it, namely $(x_n - \theta_{n-1})$.
- This is what we used in our recursive algorithm for sample mean:

$$\hat{\mu}_n = \hat{\mu}_{n-1} + \frac{1}{n}(x_n - \hat{\mu}_{n-1})$$

- We can generalize the idea as follows.

- We want to optimize $g(\theta)$.
- But, we do not know $g(\theta)$. What we can observe is the random variable $Z(\theta)$ such that $E[Z(\theta)] = \nabla_\theta g(\theta)$
- Question: can be use $Z(\theta)$ in a gradient descent algorithm?
- That is, can we solve for $E[Z(\theta)] = \nabla_\theta g(\theta) = 0$ based only on observations of $Z(\theta)$.
- A classical algorithm for this is called Robbins-Munro algorithm.
- It is a special case of so called stochastic approximation algorithms.

- The algorithm is

$$\theta_n = \theta_{n-1} - a_n Z(\theta_{n-1})$$

Where the step-size parameter $a_n > 0$ should satisfy

$$\lim_{n\to\infty} a_n = 0; \quad \sum_n a_n = \infty; \quad \sum_n a_n^2 < \infty$$

- A possible step-size is $a_n = \frac{C}{n}$ where C is a constant.
- This is the recursive equation we had for the sample mean estimate.

- The ML estimate maximizes log likelihood and hence satisfies

$$\frac{\partial}{\partial \theta}\left[\frac{1}{n}\sum_{i=1}^{n} \ln(f(x_i \mid \theta)\right] = 0$$

- We can take the derivative inside the summation and in the limit of large n can replace 'sample mean' by expectation.
- Thus, we can say, ML estimate satisfies

$$h(\theta) = E_x\left[\frac{\partial}{\partial \theta} \ln(f(x \mid \theta)\right] = 0$$

- This fits in with Robbins Munro algorithm with

$$Z(\theta) = \frac{\partial}{\partial \theta} \ln(f(x \mid \theta)$$

- We want to find zero of $h(\theta) = E[Z(\theta)]$ based on observations of $Z(\theta)$.
- Thus we get an incremental algorithm for ML estimation using the general framework of Robbins-Munro algorithm.

Recursive Bayesian Estimation

- The Bayesian estimation is inherently recursive.
- Let $\mathcal{D}^n = \{x_1, \cdots, x_n\}$ denote data of n samples.
- Now, we can write the likelihood as

$$f(\mathcal{D}^n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = f(x_n \mid \theta)\, f(\mathcal{D}^{n-1} \mid \theta)$$

- This allows us to write the posterior density in a recursive form.

$$
\begin{aligned}
f(\theta \mid \mathcal{D}^n) &= \frac{f(\mathcal{D}^n \mid \theta)\, f(\theta)}{\int\ f(\mathcal{D}^n \mid \theta')\, f(\theta')\, d\theta'} \\
&= \frac{f(x_n \mid \theta)\, f(\mathcal{D}^{n-1} \mid \theta)\, f(\theta)}{\int\ f(x_n \mid \theta')\, f(\mathcal{D}^{n-1} \mid \theta')\, f(\theta')\, d\theta'}
\end{aligned}
$$

We also have

$$
f(\theta \mid \mathcal{D}^{n-1}) = \frac{f(\mathcal{D}^{n-1} \mid \theta)\, f(\theta)}{\int\ f(\mathcal{D}^{n-1} \mid \theta'')\, f(\theta'')\, d\theta''}
$$

We have

$$f(\theta \mid \mathcal{D}^n) = \frac{f(x_n \mid \theta)\, f(\mathcal{D}^{n-1} \mid \theta)\, f(\theta)}{\int f(x_n \mid \theta')\, f(\mathcal{D}^{n-1} \mid \theta')\, f(\theta')\, d\theta'}$$

$$f(\theta \mid \mathcal{D}^{n-1}) = \frac{f(\mathcal{D}^{n-1} \mid \theta)\, f(\theta)}{\int f(\mathcal{D}^{n-1} \mid \theta'')\, f(\theta'')\, d\theta''}$$

This gives us

$$\begin{aligned} f(\theta \mid \mathcal{D}^n) &= \frac{f(x_n \mid \theta)\, \frac{f(\mathcal{D}^{n-1} \mid \theta)\, f(\theta)}{\int f(\mathcal{D}^{n-1} \mid \theta'')\, f(\theta'')\, d\theta''}}{\int f(x_n \mid \theta')\, \frac{f(\mathcal{D}^{n-1} \mid \theta')\, f(\theta')}{\int f(\mathcal{D}^{n-1} \mid \theta'')\, f(\theta'')\, d\theta''}\, d\theta'} \\ &= \frac{f(x_n \mid \theta)\, f(\theta \mid \mathcal{D}^{n-1})}{\int f(x_n \mid \theta')\, f(\theta' \mid \mathcal{D}^{n-1})\, d\theta'} \end{aligned}$$

- The bayesian estimate in recursive form is

$$f(\theta \mid \mathcal{D}^n) = \frac{f(x_n \mid \theta)\, f(\theta \mid \mathcal{D}^{n-1})}{\int f(x_n \mid \theta')\, f(\theta' \mid \mathcal{D}^{n-1})\, d\theta'}$$

- After seeing $n-1$ samples, we have $f(\theta \mid \mathcal{D}^{n-1})$ which becomes the 'current prior' while calculating the posterior when we see the n^{th} sample.
- This is the recursive form for any general Bayesian estimate.
- Often termed Bayesian learning of densities.

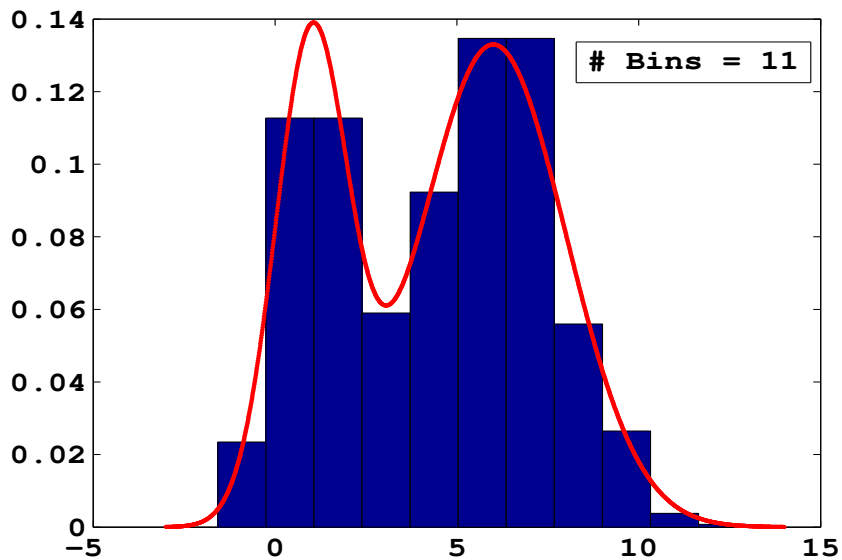
Non-parametric Estimation

- So far we have considered parametric estimation techniques.
- We assumed: $\mathcal{D} = \{x_1, \cdots, x_n\},\ x_i \sim f(x|\theta)$
- Then we can use ML or Bayesian methods for density estimation.
- We now consider the case where we do not want to assume any parametric form for the density.

- Consider a one dimensional case.
- We are given samples, x_i, $i = 1, \cdots, n$.
- We need to find the density function $f(x)$ and we do not know form of f.
- One simple idea is to learn a piece wise constant approximation to f.

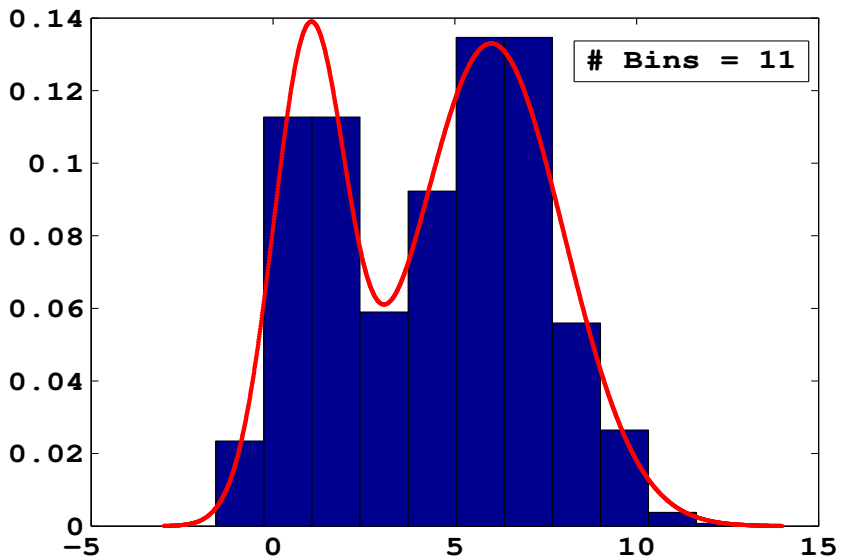
- For this, we cut the x-axis into small intervals and build a function that is constant in each of these intervals.
- If $f(x) = K$ over an interval $[a,\ b]$, then $P[a \leq X \leq b] = K(b - a)$.
- The probability above is well approximated by the fraction of data points that fall in that interval.
- Thus we can approximate f by the histogram of the data.

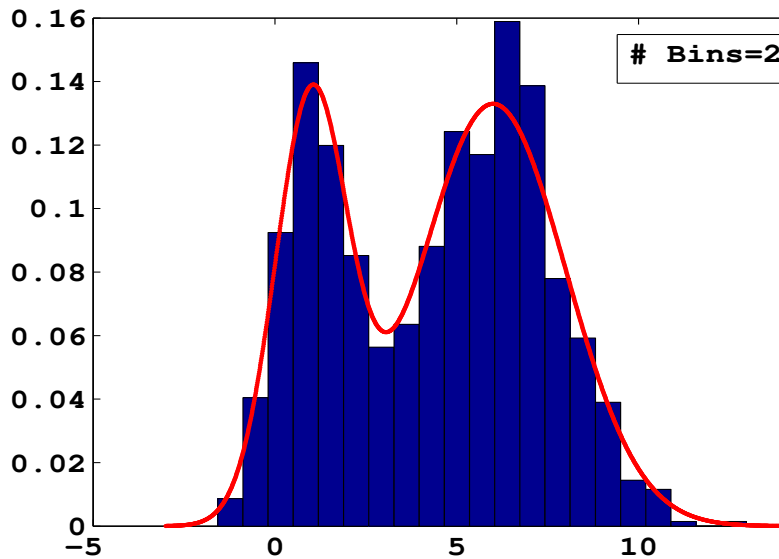
- Here is a simple example



- The quality of approximation depends on the size of intervals.

- We can get better approximation by having finer intervals





- But now, the memory needed to store $\hat{f}$ increases.

- If we extend this idea (in a simple-minded fashion) to d dimensions, the number of bins grows rapidly.
- Also, many bins may be empty.
- Curse of Dimensionality!
- However, this basic idea can be made to work.
- Essentially, we erect bins only where needed.

- Let $B(x)$ be a region (e.g., ball of some small radius) around x. Let
$$\rho = \int_{B(x)} f(x')\, dx'$$
- If f is nearly constant over $B(x)$, then $\rho \approx f(x)\, V$, where V is 'volume' of $B(x)$. Thus,
$$f(x) \approx \frac{\rho}{V}$$

- Suppose out of the n *iid* sample, k samples fall in $B(x)$.
- Then k is binomial with parameter n and ρ.
- Since, for large n, binomial distribution sharply peaks around its mean,

$$k \approx n\,\rho \qquad \text{or} \qquad \rho \approx \frac{k}{n}$$

- Combining these two, we get

$$f(x) \approx \frac{\rho}{V} \approx \frac{k}{nV}$$

- This is the basic idea of finding an approximation of f.
- At any x, we take a small volume V around x and count the number of data samples that fall in this region. This gives approximate value of $f(x)$ as above.

- Choice of V affects the quality of approximation.
- For the approximation $\rho \approx f(x)\, V$ to be good, we need V to be small.
- But if V is very small, unless n is very large, k may be zero most of the time.
- So, choice of size of V is a compromise between these two requirements.

- Let V_n denote the volume when we have n examples and let $f_n(x)$ and k_n denote the corresponding values. ($f_n(x) = \frac{k_n/n}{V_n}$)
- Then, for $f_n \to f$, as $n \to \infty$ we must have

$$V_n \to 0, \qquad k_n \to \infty, \qquad \frac{k_n}{n} \to 0$$

- We need $V_n \to 0$ to get correct estimates.
- If $f(x) \neq 0$, then we need $k_n \to \infty$.
- Finally, $\frac{k_n}{n} \to 0$ is needed to get proper estimate. (We need $nV_n \to \infty$)

- In practice we have only finite data. We choose size of V based on n.
- Actually we have a choice of two approaches.
- We can fix a V and then calculate k. Known as Parzen Window or Kernel density estimate.
- Or, we can fix k and calculate V. Known as k-nearest neightbour method.