Example (from Bishop, PRML)

Estimating the mean of a univariate Gaussian

(Assume the variance is known)

In 2D, this is like estimating a probability distribution for the center of a dart board from where the darts go.

We assume a Guassian model for the distribution of darts given the mean.

Bayes rule provided a way to consider the model given the data.

Then we optimized that formula to derive one estimate for the mean (the most likely one, i.e., the max of the p.d.f.)

No one was surprised that the mean of the data was the answer.

Review of the math from previous lecture

Example (from Bishop, PRML)

Estimating the mean of a univariate Gaussian

(Assume the variance is known)

$$p(u|\{x_i\}) \propto p(\{x_i\}|u) \quad \text{(assuming uniform prior)}$$

$$p(\{x_i\}|u) = \prod_i p(x_i|u)$$

$$\propto \prod_i e^{-\frac{(x_i-u)^2}{2\sigma^2}}$$

We can maximize the likelihood by minimizing the negative log

$$-\log\left(p(u|\{x_i\})\right) = -\log\left(\prod_i e^{-\frac{(x_i-u)^2}{2\sigma^2}}\right) \propto \sum_i (x_i-u)^2$$

$$u_{ML} = \arg\min_u \left(\sum_i (x_i-u)^2\right)$$

Estimating the mean of a univariate Gaussian

(Additional comments---well worth understanding)

- Recall the trick of maximizing the p.d.f. by minimizing the negative log
- The Gaussian form for the likelihood lead to a least-squares problem
- Least-squares solutions are tightly connected to assuming Gaussian distribution for the random effects (noise)
- If the random part is not Gaussian, then squared error may not make sense
- Squared error and Gaussian assumptions are mathematically very convenient but they are very sensitive to this assumption
- The least-squares solution leads to the average as being the “best” way to characterize a group of independent numbers, but there are other answers.
  - Minimum absolute value for error
  - Median

Example in class: (0,8,10)

Example (from Bishop, PRML)

Estimating the mean of a univariate Gaussian

Assume that the variance is known.

Given data points x_i, what is the "best" estimate for the mean?

The maximum likelihood estimate is $\mu_{ML} = \frac{1}{N}\sum_i x_i$

But what if the number of points is small?

Lets consider the case where we want to incorporate prior information.

IE, let's do Bayes.

$$p(\mu \mid \{x_i\}) \propto p(\mu) p(\{x_i\} \mid \mu)$$
$$= p(\mu) \prod_i p(\{x_i\} \mid \mu)$$
$$\propto p(\mu) \prod_i \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

What should we use for $p(\mu)$?

$$p(\mu \mid \{x_i\}) \propto p(\mu) \prod_i \exp\left(-(x_i - \mu)^2\right)$$

By inspection, if $p(\mu) \propto \exp\left(-(\mu_0 - \mu)^2\right)$ then the form of the posterior is the same as the prior.

IE, given known variance, a conjugate prior for the mean of the Gaussian is a Gaussian.

Conjugacy is convenient for several reasons, but one motivating observation is Bayesian updating whereby yesterday's posterior is used for today's prior.

Quick aside one (Bayesian update)

Consider two successive groups of observations that are conditionally independent given the model

$$p(\theta, \mathbf{x}_2, \mathbf{x}_1) = p(\mathbf{x}_2 \mid \theta) p(\mathbf{x}_1 \mid \theta) p(\theta)$$
$$= p(\mathbf{x}_2 \mid \theta) p(\theta \mid \mathbf{x}_1) p(\mathbf{x}_1)$$

SO

$$p(\theta, \mathbf{x}_2 \mid \mathbf{x}_1) = p(\mathbf{x}_2 \mid \theta) \underbrace{p(\theta \mid \mathbf{x}_1)}_{\substack{\text{updated prior,} \\ \text{after seeing } \mathbf{x}_1}}$$

Quick aside two (Conjugacy)

Informal definition: Given a likelihood function $l(\theta, x) = p(x \mid \theta)$ (we reverse θ and x when we call it a likelihood function) a (prior) distribution is natural distribution where the posterior, $p(\theta \mid x) \propto p(x \mid \theta) p(\theta)$, has the same form as $p(\theta)$.

Back to our problem.

$$p(\mu \mid \{x_i\}) \propto \underbrace{\exp\left(-\frac{(\mu_0-\mu)^2}{\sigma_0^2}\right)}_{\text{conjugate prior for the likelihood}} \underbrace{\prod_i \exp\left(-\frac{(x_i-\mu)^2}{\sigma^2}\right)}_{\text{likelihood}}$$

To find the MAP (maximum a posteriori) estimate, we maximize.

Again, maximizing is the same as minimizing the negative log.

$$-\log\left(p(\mu \mid \{x_i\})\right) = \frac{(\mu_0-\mu)^2}{\sigma_0^2} + \sum_i \frac{(x_i-\mu)^2}{\sigma^2}$$

differentiating and setting derivatives to zero gives

$$-2\left(\frac{\mu_0-\mu}{\sigma_0^2}\right) - 2\sum_i \frac{x_i-\mu}{\sigma^2} = 0$$

and dividing by -2 and collecting terms with u on the RHS gives

$$\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_i x_i = \frac{\mu}{\sigma_0^2} + \frac{N\mu}{\sigma^2}$$

and setting up for the next step

$$\frac{\mu_0}{\sigma_0^2} + \frac{N}{\sigma^2}\mu_{ML} = u\left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)$$

$$\frac{\mu_0}{\sigma_0^2} + \frac{N}{\sigma^2}\mu_{ML} = u\left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)$$

and further algebra reveals that

$$\mu_{MAP} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{N}{\sigma^2}\mu_{ML}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$$

$$= \frac{\frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}} + \frac{\frac{N}{\sigma^2}\mu_{ML}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$$

$$= \left(\frac{\sigma^2}{\sigma^2 + N\sigma_0^2}\right)\mu_0 + \left(\frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}\right)\mu_{ML}$$

$$\mu_{MAP} = \frac{\sigma^2}{\sigma^2 + N\sigma_0^2}\mu_0 + \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}\mu_{ML}$$

Study this in terms of small and large $\frac{\sigma_0^2}{\sigma^2}$

Alternative treatment (not done in class in 2013) that also shows explicitly that the posterior has the same form as the conjugate prior.

$$p(\mu \mid \{x_i\}) \propto \underbrace{\exp\left(-\frac{(\mu_0-\mu)^2}{2\sigma_0^2}\right)}_{\text{conjugate prior for the likelihood}} \underbrace{\prod_i \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}_{\text{likelihood}}$$

We combine the product of the two normals by "completing the square"

$$= \exp\left(-\frac{1}{2}\left(\frac{\mu_0^2}{2\sigma_0^2} - \frac{2\mu_0\mu}{\sigma_0^2} + \frac{\mu^2}{\sigma_0^2} + \frac{\sum x_i^2}{\sigma^2} - \frac{2\mu\sum x_i}{\sigma^2} + \frac{N\mu^2}{\sigma^2}\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left(\mu^2\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right) - 2\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{N\mu_{ML}}{\sigma^2}\right)+\ldots\right)\right) \quad \text{(ignoring constant terms)}$$

$$= \exp\left(-\frac{1}{2\left/\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\right.}\left(\mu^2 - 2\mu\left(\frac{\frac{\mu_0}{\sigma_0^2}+\frac{N\mu_{ML}}{\sigma^2}}{\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)}\right)+\ldots\right)\right) \quad \text{(ignoring constant terms)}$$

$$= \exp\left(-\frac{1}{2\sigma_N^2}\left(\mu^2 - 2\mu\left(\frac{\sigma^2\mu_0 + N\sigma_0^2\mu_{ML}}{\sigma^2+\sigma_0^2 N}\right)+\ldots\right)\right) \quad \left(\text{where } \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma_N^2}(\mu-\mu_{MAP})^2\right)$$

Example (from Bishop, PRML)

Unknown variance or mean and variance

Similar stories can be told if the mean is known and the variance is not, or both are unknown. We will only set up the problem to have a look at the conjugate priors.

Simplify things by using the inverse of the covariance matrix which is called the precision matrix.

In the univariate case this is simply: $\lambda = \frac{1}{\sigma^2}$

Example (from Bishop, PRML)

Known mean, unknown variance

$$p(\{x_i\} \mid \lambda) = \prod_{i=1}^N \mathbb{N}\left(x_i \mid \mu, 1/\lambda\right)$$

$$= \prod_{i=1}^N \left\{\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right)\right\} \quad (u \text{ is constant})$$

$$\propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_i (x_i-\mu)^2\right\}$$

constant

Inspection reveals that multiplying this by a gamma distribution

$$\text{Gam}(\lambda \mid a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

yields a posterior of the same form. The normalization constant, $\Gamma(a)$ is the "gamma" function, which extends the concept of factorial to real numbers. $\Gamma(n) = (n-1)!$, for postive integers n. Also $\Gamma(x+1) = x\Gamma(x)$ for postive reals.

"Inspection"

$$p(\{x_i\} \mid \lambda) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_i (x_i-\mu)^2\right\} = \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}K\right\}$$

$$\text{Gam}(\lambda \mid a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \propto \lambda^{a-1} \exp(-b\lambda)$$

$$p(\{x_i\} \mid \lambda)\text{Gam}(\lambda \mid a,b) \propto \lambda^{N/2} \lambda^{a-1} \exp\left\{-\frac{\lambda}{2}K\right\} \exp\{-b\lambda\}$$

$$= \lambda^{((N/2)+a-1)} \exp\left\{-\lambda\left((K/2)+b\right)\right\}$$

Gamma distribution illustrated (*)

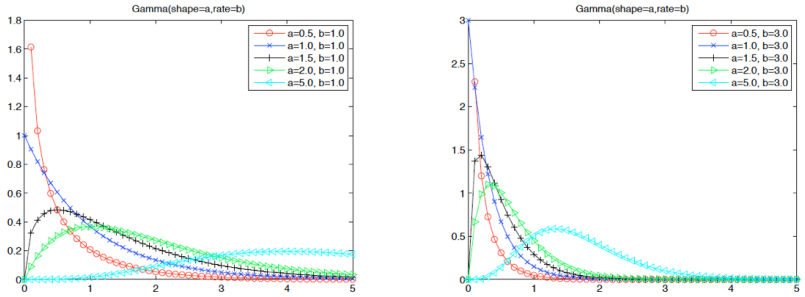


Figure 1: Some $Ga(a,b)$ distributions. If $a<1$, the peak is at 0. As we increase b, we squeeze everything leftwards and upwards. Figures generated by `gammaDistPlot2`.

* From an on-line note by Kevin Murphy (*www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall07/reading/NG.pdf*)

Example (from Bishop, PRML)

Unknown mean and variance

Indicates optional material

$$p(\mu,\lambda\,|\,\{x_i\}) \propto p(\mu,\lambda)\prod_i p(\{x_i\}\,|\,\mu,\lambda)$$

We can show that the form

$$\begin{aligned} p(u,\lambda) &= p(u\,|\,\lambda)p(\lambda) \\ &= N\left(u|u_o,(\beta\lambda)^{-1}\right)Gam(\lambda\,|\,a,b) \end{aligned}$$

where a,b,β are constants leads to the same form in the posterior. This is the normal-gamma (Gaussian-gamma) distribution.

(Derivation in notes for completeness)

$$\begin{aligned} p(\{x_i\}\,|\,u,\lambda) &\propto \lambda^{N/2}\exp\left\{-\frac{\lambda}{2}\sum_i x_i^2 + \lambda\mu\sum_i x_i - \frac{N\lambda}{2}\mu^2\right\} \\ &= \lambda^{N/2}\exp\left\{-\frac{\lambda}{2}D + \lambda\mu C - \frac{N\lambda}{2}\mu^2\right\} \\ &= \lambda^{N/2}\exp\left\{-\frac{N\lambda}{2}\mu^2 + \lambda\mu C - \frac{\lambda}{2}D\right\} \\ &= \lambda^{N/2}\exp\left((-\frac{\lambda N}{2})\left(\mu^2 - \frac{2\mu C}{N} + \frac{D}{N}\right)\right) \end{aligned}$$

$$p(\{x_i\}\,|\,u,\lambda) \propto \lambda^{N/2}\exp\left((-\frac{\lambda N}{2})\left(\mu^2 - \frac{2\mu C}{N} + \frac{D}{N}\right)\right)$$

$$\left(\mu^2 - \frac{2\mu C}{N} + \frac{D}{N}\right) = \left(\mu - \left(\frac{C}{N}\right)\right)^2 - \left(\frac{C}{N}\right)^2 + \left(\frac{D}{N}\right)$$

so,

$$p(\{x_i\}\,|\,u,\lambda) \propto \lambda^{N/2}\exp\left(-\lambda\left(\frac{C^2}{N} + D\right)\right) \underbrace{\exp\left((-\frac{\lambda N}{2})\left(\mu - \left(\frac{C}{N}\right)\right)^2\right)}$$

Multiplying this by $\mathbb{N}\left(\mu|\mu_0,(\lambda\beta)^{-1}\right)$ gives $\mathbb{N}\left(\mu|\mu_1,(\lambda\beta_1)^{-1}\right)\exp(-\lambda k)$, with $\exp(-\lambda k)$ coming from the constant when completing the sqaure.

So, multiplying by the Gauss-gamma conjugate prior, $\mathbb{N}\left(\mu\,|\,\mu_0,(\lambda\beta)^{-1}\right)Gam(\lambda\,|\,a,b)$ will give a posterior of the same form as the prior. As in the note below the last factor, completing the square leads to an extra factor that is absorbed into the new *Gam*().

Beta (and Dirichlet) distributions

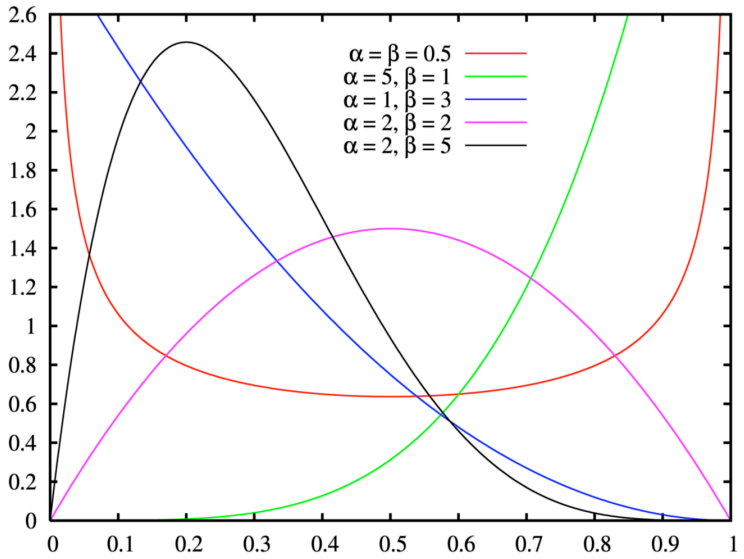
Beta (binary case)

Conjugate prior for the Bernoulli and binomial distributions

Dirichlet (multi-outcome case)

Conjugate priors for the multi outcome Bernoulli and multinomial distributions

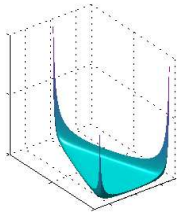
$$Beta(u \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$$



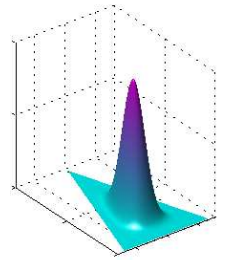
And for completeness ...

$$Dirr(u \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} u_k^{\alpha_k - 1}$$

where $\alpha_0 = \prod_{k=1}^{K} \alpha_k$



3 sided coin, likely loaded, but no idea which way



3 sided coin, close to fair

$$Beta(u \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$$

$$Bern(x \mid \mu) = \mu^{x}(1-\mu)^{(1-x)}$$

(You should be able to tell the rest of the story ...)

More on priors

If we leave off the prior, then we are completely ignorant.

Note that the prior might be the uniform distribution over all numbers

This is not a PDF!

Such priors are called improper.

A more interesting example is p(k)=1/k.

Everything can work out fine if the posterior is a PDF.