NOTES FROM LECTURE 4

Example 0.1 (Exponential with Gamma prior)**.** Let $X_1, \dots, X_n$ be independent and identically distributed (i.i.d.), each with pdf

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0\,, \\ 0 & x < 0\,. \end{cases}$$

Suppose also that λ has a Gamma(α, β) density, so that

$$f(\lambda;\alpha,\beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \qquad \lambda \geq 0,.$$

Find the posterior distribution of λ given $x_1, \dots, x_n$.

Solution. We write $\boldsymbol{x} = (x_1, \dots, x_n)$. Bayes Theorem tells us that

$$f(\lambda;\boldsymbol{x}) = \frac{f(\boldsymbol{x};\lambda) f(\lambda;\alpha,\beta)}{\int_0^\infty f(\boldsymbol{x};\lambda) f(\lambda;\alpha,\beta)\, d\lambda}\,. \tag{1}$$

The integral is cumbersome to calculate, but can be avoided.

First, note that by independence,

$$f(\boldsymbol{x};\lambda) = \prod_{i=1}^n f(x_i;\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}\,.$$

For convenience, write $s = \sum_{i=1}^n x_i$.

Let us just take the numerator in (1). Any quantity that does note depend on λ (but could depend on $\boldsymbol{x}, \alpha, \beta$) we don't keep track of, and we push them into constants, which we write as $c_1, c_2, \dots$.

$$\begin{aligned} f(\boldsymbol{x};\lambda) f(\lambda) &= \lambda^n e^{-\lambda s} \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \\ &= c_1 \lambda^{n+\alpha-1} e^{-(\beta+s)\lambda} \end{aligned}$$

Note that the integral in the denominator of (1) does not depend on λ, as it is integrated out. So

$$f(\lambda;\boldsymbol{x}) = c_2 \lambda^{n+\alpha-1} e^{-(\beta+s)\lambda}\,.$$

Now, if the above is to be a probability density, it must integrate to 1 over λ. Thus,

$$1 = c_2 \int \lambda^{n+\alpha-1} e^{-(\beta+s)\lambda}\, d\lambda\,.$$

Notice that the form of the integrand is almost a (different) Gamma density, up to normalizing constants. Indeed,

$$1 = c_2 \frac{\Gamma(n+\alpha)}{(\beta+s)^{n+\alpha-1}} \int \frac{(\beta+s)^{n+\alpha-1}}{\Gamma(n+\alpha)} \lambda^{n+\alpha-1} e^{-(\beta+s)\lambda} d\lambda\,.$$

The integrand is now a probability density, whence it integrates to 1, and we obtain

$$1 = c_2 \frac{\Gamma(n+\alpha)}{(\beta+s)^{n+\alpha-1}}\,,$$

identifying the constant c_2.

We conclude that

$$\frac{(\beta+s)^{n+\alpha-1}}{\Gamma(n+\alpha)} \lambda^{n+\alpha-1} e^{-(\beta+s)\lambda}.$$

□

1. Gaussian Distribution

The d-dimensional (non-degenerate) Gaussian distribution is a random vector $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(d)})$ in $\mathbb{R}^d$ specified by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and by a covariance matrix Σ which is a positive definite matrix. The density for $\boldsymbol{X}$ is

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T \Sigma^{-1} (\boldsymbol{x}-\mu)\right).$$

We have $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\operatorname{cov}(X^{(i)}, X^{(j)}) = \Sigma_{i,j}$.

The maximum likelihood estimators of $\boldsymbol{\mu}$ and Σ are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T.$$

(Here, unless stated otherwise, all vectors $\boldsymbol{x}$ are treated as column vectors in matrix operations.)

2. Mixtures of Gaussians

We now consider a distribution that arised by taking a weighted average of Gaussians. Let π_j be a probability distribution on $\{1, \ldots, K\}$. Let $(\boldsymbol{\mu}_j, \Sigma_j)$ be the mean and covariance matrix for a Gaussian, where $j = 1, \ldots, K$.

Now consider the density

$$f(\boldsymbol{x}; (\boldsymbol{\mu}_j, \Sigma_j)_{j=1}^K) = \sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x}; \boldsymbol{\mu}_j, \Sigma_j)\,. \tag{2}$$

The density f_j is a Normal$(\boldsymbol{\mu}_j, \Sigma_j)$ density.

We can consider a random vector $\boldsymbol{X}$ having the density above as being generated in two stages: First select Z from $\{1,\ldots,K\}$ using the probabilities $\pi_1,\ldots,\pi_K$, and next given $Z=j$, generate a random variable from the $N(\boldsymbol{\mu}_j,\Sigma_j)$ distribution.

As an exercise, convince yourself that this has the mixture density in (2)

Note that an observer cannot see Z, but only is allowed to observe $\boldsymbol{X}$.

The goal is to compute the MLE (maximum likelihood estimator) for all these parameters, given an i.i.d. sample from (2). Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be such a sample. Associated to each $\boldsymbol{X}_i$ is a **latent** variable Z_i which we do not observe.

Define, for given parameter values:

$$\gamma^{(i)}(j) = \mathbb{P}(Z_i = j \mid \boldsymbol{X}_i; \boldsymbol{\mu}_j, \Sigma_j) = \frac{\pi_j f_j(\boldsymbol{x}; \boldsymbol{\mu}_j, \Sigma_j)}{\sum_{\ell=1}^K \pi_\ell f_\ell(\boldsymbol{x}; \boldsymbol{\mu}_\ell, \Sigma_\ell)}$$

$$N_j = \sum_{i=1}^n \gamma^{(i)}(j) = \mathbb{E}[C_j \mid (\boldsymbol{\mu}_j, \Sigma_j)_{j=1}^K],$$

where $C_j = \sum_{i=1}^n \mathbf{1}\{Z_i = j\}$ is the number of data points with $Z_i = j$, i.e. which came from f_j.

We derived (see slides) the following system of equations that a critical point should satisfy

$$\boldsymbol{\mu}_j = \frac{1}{N_j}\sum_{i=1}^n \gamma_j^{(i)} x_i$$

$$\Sigma_j = \frac{1}{N_j}\sum_{i=1}^n \gamma^{(i)}(j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T$$

$$\pi_j = \frac{N_j}{n}.$$

This system of equation is not a closed-form solution, since the γ's depend on the parameters that appear on the left-hand sides. Nonetheless, we can use an iterative algorithm to approximate solutions:

- Initialize with some values $\boldsymbol{\mu}_j^0, \Sigma_j^0, \pi_j^0$ for $j=1,\ldots,K$.
- Compute the $\gamma_j^{(i)}$'s using the initial parameters.
- Use the equations above to produce new parameter estimates, $\boldsymbol{\mu}_j^{(1)}, \Sigma_j^{(1)}, \pi_j^{(1)}$.
- Use the newly obtained parameter estimates to recompute the γ's.
- Use the newly recomputed γ's to obtain new parameter estimates.
- Continue until the estimates stop changing.

This algorithm is guaranteed to increase the likelihood function at each step. However, it could become stuck at local maximum.