ESTIMATION

Lecture Notes

Ahmet Ademoglu, *PhD*
Bogazici University
Institute of Biomedical Engineering

Bayes Theorem

$$p(\theta|\mathbf{x}) = \frac{\overbrace{p(\mathbf{x}|\theta)}^{\textit{likelihood}}\ \overbrace{p(\theta)}^{\textit{prior}}}{\underbrace{p(\mathbf{x})}_{\textit{evidence}}} \propto p(\mathbf{x}|\theta)p(\theta)$$

If we are to estimate a set of parameters θ, given an observation $\mathbf{x}$, the best we can do is to maximize its *log-likelihood function*, if we have no idea about its prior distribution;

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)$$

$$\frac{\partial}{\partial\theta}\mathcal{L} = \frac{\partial}{\partial\theta}\log p(\mathbf{x}|\theta)$$

Assuming that **X** is a random variable having a Bernoulli distribution, what is the ML estimation of μ if we have a set of observations $\{x_1, \ldots, x_N\}$?

$$p(x|\mu) = \begin{cases} \mu^x (1-\mu)^{1-x} & \textit{if } x = \{0,1\} \\ 0 & \textit{otherwise} \end{cases}$$

$$\mathcal{L} = \prod_{i=1}^{N} \mu^{x_i} (1-\mu)^{1-x_i}$$

$$\frac{\partial}{\partial \mu} \log \mathcal{L} = \frac{\partial}{\partial \mu} \sum_{i=1}^{N} [x_i \log \mu + (1-x_i) \log(1-\mu)] = 0$$

$$= \sum_{i=1}^{N} \left[\frac{x_i}{\mu} - \frac{1-x_i}{1-\mu} \right] = 0 = \left(\frac{1}{\mu} + \frac{1}{1-\mu} \right) \sum_{i=1}^{N} x_i - \frac{1}{1-\mu} \sum_{i=1}^{N} 1$$

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

The number of marriages bu individuals are modeled by Poisson distributed random variables X and it is assumed that they are a linear function of age *i.e.* $\lambda = A\lambda_0$. Using the observations below, determine the parameter λ_0.

(# of times married) X	(Age) A
0	12
0	50
2	30
2	36
7	97

$$p_X(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

$$\log \mathcal{L} = \log p(X|\lambda) = \log \prod_{i=1}^{N} p(x_i|\lambda) = \sum_{i=1}^{N}(-\lambda + x_i \log \lambda - \log x_i!)$$

$$= \sum_{i=1}^{N}(-A_i\lambda_0 + x_i \log(A_i\lambda_0) - \log x_i!)$$

$$\frac{\partial}{\partial \lambda_0}\mathcal{L} = \frac{\partial}{\partial \lambda_0} = \sum_{i=1}^{N}\left(-A_i + \frac{1}{\lambda_0}x_i\right) = 0$$

$$\lambda_0 = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} A_i} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i}{\frac{1}{N}\sum_{i=1}^{N} A_i} = \frac{16}{225}$$

Maximum Likelihood Estimation: Multivariable Case

Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate distribution, we can estimate the parameters of the distribution by maximizing its likelihood;

$$\frac{\partial}{\partial \mathbf{a}} \ln p(\mathbf{X}|\mathbf{a}) = 0$$

Determine the maximum likelihood estimation of the mean of a multivariate Gaussian distribution based on observations $\{\mathbf{x}_n\}$.

$$\ln p(\mathbf{X}|\mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln|\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mu)^T \Sigma^{-1} (\mathbf{x}_n - \mu)$$

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{X}|\mu, \Sigma) = \Sigma^{-1} \sum_{n=1}^{N} (\mathbf{x}_n - \mu) = 0$$

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \tag{1}$$

Determine the maximum likelihood estimation of the covariance of a multivariate Gaussian distribution based on observations $\{\mathbf{x}_n\}$.

$\frac{\partial}{\partial \Sigma} \ln p(\mathbf{X}|\mu, \Sigma) = \frac{N}{2} \frac{\partial}{\partial \Sigma^{-1}} \ln |\Sigma^{-1}| - \frac{1}{2} \frac{\partial}{\partial \Sigma^{-1}} \sum_{n=1}^{N} Tr[\Sigma^{-1}(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T] = 0$

$\frac{N}{2} \Sigma^T - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T = 0$

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T$$

Find a sequential method for the maximum likelihood of μ_{ML}.

$\mu_{ML}^{(N)} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N} \mathbf{x}_N + \frac{1}{N} \sum_{n=1}^{N-1} \mathbf{x}_n$

$$\mu_{ML}^{(N)} = \frac{1}{N} \mathbf{x}_N + \frac{N-1}{N} \sum_{n=1}^{N-1} \mu_{ML}^{(N-1)} = \underbrace{\mu_{ML}^{(N-1)}}_{\text{old estimate}} + \frac{1}{N} \underbrace{(\mathbf{x}_N - \mu_{ML}^{(N-1)})}_{\text{correction}}$$

Bias and Consistency of Estimation

Bias

$Bias = \theta - E[\hat{\theta}_N] \longrightarrow$ If an estimator is unbiased then $\theta = E[\hat{\theta}_N]$.

Consistency

$$\lim_{N\to\infty} = Var[\hat{\theta}_N] = \lim_{N\to\infty} E[|\hat{\theta}_N - E[\hat{\theta}_N]|^2] = 0$$

Determine the bias of ML estimate of μ and Σ of multivariate Gaussian distribution.

$E\{\mu_{ML}\} = \frac{1}{N}\sum_{n=1}^{N} E\{\mathbf{x}_n\} = \frac{1}{N}\sum_{n=1}^{N}\mu = \mu \longrightarrow \mu_{ML}$ is an unbiased estimator.

$$E\{\Sigma_{ML}\} = \frac{1}{N}\sum_{n=1}^{N} E\{(\mathbf{x}_n - \mu_{ML})(\mathbf{x}_n - \mu_{ML})^T\}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left\{E\left\{\mathbf{x}_n\mathbf{x}_n^T\right\} - E\left\{\mathbf{x}_n\mu_{ML}^T\right\} - E\left\{\mu_{ML}\mathbf{x}_n^T\right\} + E\left\{\mu_{ML}\mu_{ML}^T\right\}\right\}$$

$$E\left\{\mathbf{x}_n\mathbf{x}_n^T\right\} = \Sigma + \mu\mu^T$$

$$E\left\{\mathbf{x}_n\mu_{ML}^T\right\} = E\left\{\mathbf{x}_n\frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m^T\right\} = \frac{1}{N}E\{\mathbf{x}_n\mathbf{x}_n^T + \underbrace{\mathbf{x}_n\mathbf{x}_1^T + \cdots + \mathbf{x}_n\mathbf{x}_N^T}_{N-1\ times}\}$$

$$= \frac{1}{N}(\Sigma + \mu\mu^T + (N-1)\mu\mu^T) = \frac{1}{N}\Sigma + \mu\mu^T$$

$$E\left\{\mu_{ML}\mu_{ML}^T\right\} = E\left\{\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \cdot \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m^T\right\} =$$

$$\frac{1}{N^2}E\{\underbrace{\mathbf{x}_1\mathbf{x}_1{}^T + \cdots + \mathbf{x}_N\mathbf{x}_N{}^T}_{N\ times} + \mathbf{x}_1(\mathbf{x}_2 + \ldots \mathbf{x}_N) + \cdots + \underbrace{\mathbf{x}_N(\mathbf{x}_1 + \cdots + \mathbf{x}_{N-1})}_{N-1\ times}\}$$

$$= \frac{1}{N^2}\left\{(N(\Sigma + \mu\mu^T) + N\cdot(N-1))\mu\mu^T\right\} = \frac{1}{N}\Sigma + \mu\mu^T$$

$$E\{\Sigma_{ML}\} = \frac{1}{N}\sum_{n=1}^{N}\left\{\Sigma + \mu\mu^T - 2(\frac{1}{N}\Sigma + \mu\mu^T) + \frac{1}{N}\Sigma + \mu\mu^T\right\} = \frac{N-1}{N}\Sigma$$

Bayesian Estimation

Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate distribution, we can estimate the parameters θ of the distribution by maximizing the posterior distribution $p(\theta|\mathbf{X})$ if the prior distribution of θ is known.

$$p(\theta|\mathbf{X}) = \frac{\overbrace{p(\mathbf{X}|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathbf{X})}_{\text{evidence}}} \propto p(\mathbf{X}|\theta)p(\theta)$$

We wish estimate the mean, of $\mathbf{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ with $\mathbf{x}$ having a distribution $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ in which Σ is known. Determine the corresponding posterior distribution $p(\mu|\mu_N, \Sigma_N)$ if prior is $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$.

Equating the exponential terms of both sides;

$$-\frac{1}{2}\mu^T\Sigma_N^{-1}\mu + \mu^T\Sigma_N^{-1}\mu_N = -\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^T\Sigma^{-1}(\mathbf{x}_n-\mu)$$

$\mu^T(.)\mu$ terms : $\Sigma_0^{-1} + N\Sigma^{-1}$ and $\mu^T(.)$ terms : $\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{n=1}^{N}\mathbf{x}_n$

$$\mu_N = \Sigma_N(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{n=1}^{N}\mathbf{x}_n), \qquad \Sigma_N = (\Sigma_0^{-1} + N\Sigma^{-1})^{-1}$$

$$\mu_N = (\Sigma_0^{-1} + N\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + N\Sigma^{-1}\mu_{ML})$$

Conjugate Priors

Conjugate priors are such distributions that they lead to posterior distributions having the same functional form as the prior.

We wish estimate the precision $\lambda = \frac{1}{\sigma^2}$, of a posterior distribution $p(\lambda|\mathbf{x},\mu)$ using $\mathbf{x} = \{x_1, \dots, x_N\}$ with $\mathbf{x}$ having $p(\mathbf{x}) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\lambda)$. Determine the corresponding posterior distribution $p(\lambda|a_N, b_N)$ if prior is $p(\lambda) = Gam(\lambda|a_0, b_0)$.

posterior distribution of λ : $p(\lambda|\mathbf{x},\mu) \propto \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\lambda) Gam(\lambda|a_0, b_0)$

$$\prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\lambda) Gam(\lambda|a_0, b_0) = \left(\frac{\lambda}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{1}{2}\lambda \sum_{n=1}^{N}(x_n-\mu)^2} \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0-1} e^{-b_0\lambda}$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} \frac{1}{\Gamma(a)} b_0^{a_0} \lambda^{a_0+\frac{N}{2}-1} e^{-(b_0+\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2)\lambda} \propto Gam(\lambda|a_N, b_N)$$

where $a_N = a_0 + \frac{N}{2}$ and $b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma_{ML}^2$.

Determine the posterior distribution $p(\mu, \lambda|\mathbf{x})$ using $\mathbf{x} = \{x_1, \dots, x_N\}$ with $\mathbf{x}$ having $p(\mathbf{x}) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda)$ if both μ and λ are unknown and their prior is a Gaussian-Gamma distribution, $p(\lambda, \mu) \propto \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1}) Gam(\lambda|a, b)$.

$$p(\mu, \lambda|\mathbf{x}) \propto \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{1}{2}\lambda(x_n-\mu)^2} \left(\frac{\beta\lambda}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\beta\lambda(\mu-\mu_0)^2} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}$$

$$\propto \left(\frac{\lambda}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{1}{2}\lambda \sum_{n=1}^{N}(x_n-\mu)^2} \left(\frac{\beta\lambda}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\beta\lambda(\mu-\mu_0)^2} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}$$

$$\propto \lambda^{\frac{N}{2}} e^{(-\frac{1}{2}\sum_{n=1}^{N} x_n^2 + \mu \sum_{n=1}^{N} x_n - \frac{N}{2}\mu^2)\lambda} \lambda^{\frac{1}{2}} e^{(-\frac{\beta\lambda}{2}(\mu^2 - 2\mu\mu_0 + \mu_0^2))} b^a \lambda^{a-1} e^{-b\lambda}$$

$$\lambda \text{ terms}: \lambda^{a+\frac{N}{2}-1} e^{-\left(b+\frac{1}{2}\sum_{n=1}^{N} x_n^2 + \frac{\beta}{2}\mu_0^2\right)\lambda} \longrightarrow a_N = a + \frac{N}{2},\ b_N = b + \frac{1}{2}\sum_{n=1}^{N} x_n^2 + \frac{\beta}{2}\mu_0^2 + ?$$

$$\mu \text{ terms}: \lambda^{\frac{1}{2}} e^{-\lambda\left(\frac{N+\beta}{2}\mu - \beta\mu_0 - \sum_{n=1}^{N} x_n\right)\mu}$$

$$= \frac{(\lambda(N+\beta))^{\frac{1}{2}}}{(N+\beta)^{\frac{1}{2}}} e^{-\left(\frac{\lambda(N+\beta)}{2}\left(\mu^2 - \frac{2\mu}{N+\beta}\left(\beta\mu_0 + \sum_{n=1}^{N} x_n\right) + \left(\frac{\beta\mu_0 + \sum_{n=1}^{N} x_n}{N+\beta}\right)^2\right)\right)} e^{\frac{\lambda(\beta\mu_0 + \sum_{n=1}^{N} x_n)^2}{2(N+\beta)}}$$

$$\mu_N = \frac{\beta\mu_0 + \sum_{n=1}^{N} x_n}{N+\beta} \longrightarrow \frac{(\lambda(N+\beta))^{\frac{1}{2}}}{} e^{-\left(\frac{\lambda(N+\beta)}{2}(\mu^2 - 2\mu\mu_N + \mu_N^2)\right)} e^{\frac{\lambda\mu_N^2(N+\beta)}{2}?} / (N+\beta)^{\frac{1}{2}}$$

$$b_N = b + \frac{1}{2}\sum_{n=1}^{N} x_n^2 + \frac{\beta}{2}\mu_0^2 - \mu_N^2 \frac{(N+\beta)}{2}$$

$$p(\mu, \lambda|\mathbf{x}) \propto \mathcal{N}(\mu|\mu_N, (\lambda[N+\beta])^{-1}) Gam(\lambda|a_N, b_N)$$

$$\propto (\lambda(N+\beta))^{\frac{1}{2}} e^{-\left(\frac{\lambda(N+\beta)}{2}(\mu-\mu_N)^2\right)} \lambda^{a_N-1} e^{-b_N\lambda}$$

Considering the multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\mu, \Lambda^{-1})$ for a $D-$dimensional variable $\mathbf{x}$, show that, for known mean and unknown precision matrix Λ, the conjugate prior is the Wishart distribution given by $\mathcal{W}(\Lambda|\mathbf{W}, \nu) = B|\Lambda|^{(\nu-D-1)/2} e^{-\frac{1}{2}Tr(\mathbf{W}^{-1}\Lambda)}$.

$$p(\Lambda|\mathbf{W}, \nu) = \mathcal{W}(\Lambda|\mathbf{W}, \nu).$$

$$\propto |\Lambda|^{\frac{N}{2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Lambda(x_n-\mu)} |\Lambda|^{(\nu-D-1)/2} e^{-\frac{1}{2}Tr(\mathbf{W}^{-1}\Lambda)}$$

$$\propto |\Lambda|^{(N+\nu-D-1)/2} e^{-\frac{1}{2}Tr[\sum_{n=1}^{N}(x_n-\mu)(x-\mu)^T+\mathbf{W}^{-1})\Lambda]}$$

$$\propto \mathcal{W}(\Lambda|\bar{\mathbf{W}}, \bar{\nu}) \text{ with } \qquad \bar{\mathbf{W}} = \sum_{n=1}^{N}(x_n-\mu_0)(x_n-\mu_0)^T + \mathbf{W}^{-1}, \qquad \bar{\nu} = N + \nu$$

Show that, for known precision matrix Λ and unknown mean, the conjugate prior is the Gaussian distribution given by $\mathcal{N}(\mu|\mu_\mu, \Lambda_{\mu_\mu}^{-1})$.

$$p(\mu|\bar{\mu}, \bar{\Lambda}) = \mathcal{N}(\mathbf{x}|\mu, \Lambda^{-1})\mathcal{N}(\mu|\mu_\mu, \Lambda_\mu^{-1}) \propto$$

$$|\Lambda|^{\frac{N}{2}} e^{-\frac{1}{2}(\sum_{n=1}^{N}(x_n-\mu)^T\Lambda(x_n-\mu))} |\Lambda_\mu|^{\frac{1}{2}} e^{-\frac{1}{2}(\mu-\mu_\mu)^T\Lambda_\mu(\mu-\mu_\mu)}$$

$$\mu^T(\bar{\Lambda})\mu : \mu^T(N\Lambda+\Lambda_\mu)\mu, \qquad \mu^T\bar{\Lambda}\bar{\mu} = \mu^T(\Lambda\sum_{n=1}^{N}x_n + \Lambda_\mu\mu_\mu)$$

$$\bar{\Lambda} = N\Lambda + \Lambda_\mu \text{ and } \bar{\mu} = (N\Lambda+\Lambda_\mu)^{-1}(\Lambda\sum_{n=1}^{N}x_n + \Lambda_\mu\mu_\mu)$$