1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameters that maximize the likelihood of the observations. Concretely, given observations $y_1, y_2, \ldots, y_n$ distributed according to $p_\theta(y_1, y_2, \ldots, y_n)$ (here p_θ can be a probability mass function for discrete observations or a density for continuous observations), the likelihood function is defined as $L(\theta) = p_\theta(y_1, y_2, \ldots, y_n)$ and the MLE is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta L(\theta).$$

We often make the assumption that the observations are *independent and identically distributed* or iid, in which case $p_\theta(y_1, y_2, \ldots, y_n) = p_\theta(y_1) \cdot p_\theta(y_2) \cdot \cdots \cdot p_\theta(y_n)$.

(a) Your friendly TA recommends maximizing the log-likelihood $\ell(\theta) = \log L(\theta)$ instead of $L(\theta)$. Why does this yield the same solution $\hat{\theta}_{\text{MLE}}$? Why is it easier to solve the optimization problem for $\ell(\theta)$ in the iid case? Given the observations $y_1, y_2, \ldots, y_n$, write down both $L(\theta)$ and $\ell(\theta)$ for the Gaussian $f_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$ with $\theta = (\mu, \sigma)$.

(b) The Poisson distribution is $f_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!}$. Let $Y_1, Y_2, \ldots, Y_n$ be a set of independent and identically distributed random variables with Poisson distribution with parameter λ. Find the joint distribution of $Y_1, Y_2, \ldots, Y_n$. Find the maximum likelihood estimator of λ as a function of observations $y_1, y_2, \ldots, y_n$.

2 Independence and Multivariate Gaussians

As described in lecture, a covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$ for a random variable $X \in \mathbb{R}^N$ with the following values, where $\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$ is the covariance between the i-th and j-th elements of the random vector X:

$$\Sigma = \begin{bmatrix} \text{cov}(X_1, X_1) & \dots & \text{cov}(X_1, X_n) \\ \dots & & \dots \\ \text{cov}(X_n, X_1) & \dots & \text{cov}(X_n, X_n) \end{bmatrix}. \tag{1}$$

Recall that the density of an N dimensional Multivariate Gaussian Distribution $\mathcal{N}(\mu, \Sigma)$ is defined as follows when Σ is positive definite:

$$f(x) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}. \tag{2}$$

Here, $|\Sigma|$ denotes the determinant of the matrix Σ.

(a) Consider the random variables X and Y in $\mathbb{R}$ with the following conditions.

(i) X and Y can take values $\{-1, 0, 1\}$.

(ii) When X is 0, Y takes values 1 and -1 with equal probability ($\frac{1}{2}$). When Y is 0, X takes values 1 and -1 with equal probability ($\frac{1}{2}$).

(iii) Either X is 0 with probability ($\frac{1}{2}$), or Y is 0 with probability ($\frac{1}{2}$).

Are X and Y uncorrelated? Are X and Y independent? Prove your assertions. *Hint:* Write down the joint probability of (X, Y) for each possible pair of values they can take.

(b) For $X = [X_1, \cdots, X_n]^\top \sim \mathcal{N}(\mu, \Sigma)$, **verify that if X_i, X_j are independent (for all $i \neq j$), then Σ must be diagonal, i.e., X_i, X_j are uncorrelated.**

(c) Let $N = 2$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$. Suppose $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$. **Show that X_1, X_2 are independent if $\beta = 0$.** Recall that two continuous random variables W, Y with joint density $f_{W,Y}$ and marginal densities f_W, f_Y are independent if $f_{W,Y}(w, y) = f_W(w) f_Y(y)$.

(d) Consider a data point x drawn from an N-dimensional zero mean Multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, as shown above. Assume that Σ^{-1} exists. **Prove that there exists a matrix $A \in \mathbb{R}^{N\times N}$ such that $x^\top \Sigma^{-1} x = \|Ax\|_2^2$ for all vectors x. What is the matrix A?**

3 Least Squares (using vector calculus)

(a) In ordinary least-squares linear regression, we typically have $n > d$ so that there is no $\mathbf{w}$ such that $\mathbf{X}\mathbf{w} = \mathbf{y}$ (these are typically overdetermined systems — too many equations given the number of unknowns). Hence, we need to find an approximate solution to this problem. The residual vector will be $\mathbf{r} = \mathbf{X}\mathbf{w} - \mathbf{y}$ and we want to make it as small as possible. The most common case is to measure the residual error with the standard Euclidean ℓ^2-norm. So the problem becomes:

$$\min_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$$

Where $\mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{w} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^n$. Derive using vector calculus an expression for an optimal estimate for $\mathbf{w}$ for this problem assuming $\mathbf{X}$ is full rank.

(b) How do we know that $\mathbf{X}^\top \mathbf{X}$ is invertible?

(c) What should we do if $\mathbf{X}$ is not full rank?