Discussion 4 **Alex Francis**
CS 189 Email: afrancis@berkeley.edu
02/14/2017 OH: Wednesday 10:00am - 12:00pm (283E Soda)

1 Lecture on Maximum Likelihood Estimates

Probability distributions can be terrific models for real-world data. However, just using the techniques of probability theory, we have no way of "fitting a curve to data" in a probabilistic sense. Given a dataset, we can hypothesize that the process underlying the generation of the data leads to some distributional form (for example, the Gaussian or Exponential distributions from the previous discussion worksheets). We assume in CS189 that the distribution is given; in practice, this is typically due to artful data analysis, and an intimate understanding of the problem that one is solving. However, for an assumed distributional form, there are always *parameters* that determine the specific version of the distributional family in question (think, altering the mean of the normal distribution μ changes the location of the peak, and altering the variance σ^2 changes the spread of the symmetric distribution). There are at least three methods for estimating parameters of a probability distribution that are utilized today.

1. Maximum Likelihood Estimation - maximize the "likelihood" of the data using calculus or numerical optimization methods.
2. Method of Moments - find the theoretical "moments" (mean, variance, etc.). Since these moments are expressed as a function of the parameters, we can reverse engineer and solve for the parameters as a function of the *sample moments*.
3. Least Squares.

We only discuss the first one in this course, since it is the most well-used in statistics and machine learning for certain technical reasons.[1]

Define the likelihood function as,

$$\mathcal{L}(\theta) = p(x_1, \ldots, x_n \mid \theta)$$

Where θ is a vector of parameters. Assuming that $x_i, i \in \{1, \ldots, n\}$ i.i.d.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

It is natural to desire an estimate of θ that maximizes this expression, as then we can reason with intuition that the distribution with θ as the parameter vector is more likely than the distribution with any other set of parameters. Since using calculus is convenient, and taking derivatives of products is cumbersome, we define the *log-likelihood* function,

$$\begin{aligned}\ell(\theta) &= \log\left(\prod_{i=1}^{n} p(x_i \mid \theta)\right) \\ &= \sum_{i=1}^{n} \log(p(x_i \mid \theta))\end{aligned}$$

[1] To elaborate on this statement, the maximum likelihood estimate has two properties called *consistency* (meaning the estimate tends to be correct in the limit) and *sufficiency*. See *Mathematical Statistics and Data Analysis*, by Rice, for more.

The maximizing θ for ℓ is the maximizing θ for $\mathcal{L}$, since the logarithm function is monotonically increasing. Using mechanical techniques from calculus on this simple derivation allows us to solve almost any MLE problem that has closed-form solutions (note: in some cases, like Homework 2, Problem 8, the optimization problem is constrained, meaning we need more mathematical machinery like Lagrange Multipliers). In the general case, we can approximate maximizing parameter vector of the log-likelihood function by running a numerical optimization algorithm, like gradient ascent.

2 Multivariate Gaussian MLE

We draw three points from a Multivariate Gaussian (1, 0), (0, 1), and (2, 2). Find the MLE for μ and Σ.

Unfortunately, you will not have seen the derivation for the MLE of the multivariate Gaussian before this section (see Homework 3). As a result, I do not plan to teach this in section, unless there is extra time. Note that the MLEs into which you would plug values are,

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

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

3 Multivariate Gaussian

1. True or False:

   (a) If X_1 and X_2 are both normally distributed and independent, then (X_1, X_2) must have multivariate normal distribution.

   Solution. True.

   Proof. Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then,

$$\begin{aligned}
p(x_1, x_2) &= p(x_1)p(x_2) \\
&= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right) \cdot \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right) \\
&= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right)\right) \\
&= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}\right)\right) \\
&= \mathcal{N}(\mu, \Sigma)
\end{aligned}$$

   In the final formula above, we have,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

   Since the inverse of a diagonal matrix means taking the reciprocal of the diagonal elements, and it can be confirmed that the determinant of Σ is $\sigma_1^2\sigma_2^2$. □

(b) If (X_1, X_2) has multivariate normal distribution, then X_1 and X_2 are independent.

Solution. False. The off-diagonal elements in the covariance matrix $\Sigma_{ij}, i \neq j$ are filled with the quantities $\text{Cov}(X_i, X_j)$. These may be non-zero quantities, indicating that there is dependence between these two random variables by definition, without breaking anything about the formulation of the multivariate Gaussian.

2. Affine Transformation. $X = \begin{bmatrix} X_1, X_2, \ldots, X_n \end{bmatrix}$ is a n-dimensional random vector which has multivariate normal distribution. If $X \sim \mathcal{N}(\mu, \Sigma)$ and $Y = BX + c$ is an affine transformation of X, where c is a constant $m \times 1$ vector and B is a constant $m \times n$ matrix, what is the expectation and variance of Y?

Solution. Expectation and variance on random vectors always confuses me, so here are some important intermediary results that use only definitions and expectation/variance characteristics of random variables and scalars.

LEMMA 1. For any random vector $X \in \mathbb{R}^m$ and constant vector $c \in \mathbb{R}^m$, $\mathbb{E}(X + c) = \mathbb{E}(X) + c$.

Proof.

$$\mathbb{E}(X + c) = \begin{bmatrix} \mathbb{E}(X_1 + c_1) \\ \vdots \\ \mathbb{E}(X_n + c_n) \end{bmatrix} = \mathbb{E}(X + c) = \begin{bmatrix} \mathbb{E}(X_1) + c_1 \\ \vdots \\ \mathbb{E}(X_n) + c_n \end{bmatrix} = \mathbb{E}(X) + c$$

□

LEMMA 2. For any constant matrix $B \in \mathbb{R}^{m,n}$ and any random vector $X \in \mathbb{R}^m$, $\mathbb{E}(BX) = B\mathbb{E}(X)$.

Proof.

$$\begin{aligned}
\mathbb{E}(BX) &= \begin{bmatrix} \mathbb{E}(b_1^\top x) \\ \vdots \\ \mathbb{E}(b_n^\top x) \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}\left(\sum_{i=1}^n b_{1i} x_i\right) \\ \vdots \\ \mathbb{E}\left(\sum_{i=1}^n b_{mi} x_i\right) \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}\left(\sum_{i=1}^n b_{1i} x_i\right) \\ \vdots \\ \mathbb{E}\left(\sum_{i=1}^n b_{mi} x_i\right) \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n b_{1i} \mathbb{E}(x_i) \\ \vdots \\ \sum_{i=1}^n b_{1i} \mathbb{E}(x_i) \end{bmatrix} \\
&= \begin{bmatrix} b_1^\top \mathbb{E}(x) \\ \vdots \\ b_n^\top \mathbb{E}(x) \end{bmatrix} \\
&= B\mathbb{E}(X)
\end{aligned}$$

□

This is enough to compute the expectation of Y. We have,

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(BX + c) \\ &= B\mathbb{E}(X) + c \\ &= B\mu + c\end{aligned}$$

LEMMA 3. For any random vector $X \in \mathbb{R}^m$ and constant vector $c \in \mathbb{R}^m$, $\text{Cov}(BX + c) = B^\top \text{Cov}(X)B$.

Proof. Note first that the definition of any random-valued vector X is,

$$\text{Cov}(X) = \mathbb{E}\left((X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top\right)$$

Which is analogous to the standard definition of covariance on random variables. Applying this definition,

$$\begin{aligned}\text{Cov}(BX + c) &= \mathbb{E}\left((BX + c - \mathbb{E}(BX + c))(BX + c - \mathbb{E}(BX + c))^\top\right) \\ &= \mathbb{E}\left((BX + c - B\mu - c)(BX + c - B\mu - c)^\top\right) \\ &= \mathbb{E}\left((B(X - \mu))(B(X - \mu))^\top\right) \\ &= \mathbb{E}\left(B(X - \mu)(X - \mu)^\top B^\top\right) \\ &= B\mathbb{E}\left((X - \mu)(X - \mu)^\top\right) B^\top \\ &= B\Sigma B^\top\end{aligned}$$

The proof that you can factor $B^\top$ out of the expectation expression above resembles that of LEMMA 2 (try it!). □

4 MLE of the Laplace Distribution

Let X have a Laplace distribution with density[2]

$$p(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

Suppose that n samples $x_1, \ldots, x_n$ are drawn independently according to $p(x; \mu, b)$.

1. Find the maximum likelihood estimate of μ.

   Solution. First, we determine the likelihood function $\mathcal{L}$,

   $$\mathcal{L}(\mu, b) = \prod_{i=1}^{n} (x_i; \mu, b) = \prod_{i=1}^{n} \frac{1}{2b} \exp\left(-\frac{|x_i - \mu|}{b}\right)$$

[2]The Laplace Distribution was originally proposed in the late 1700s as an "error curve," that is as a probabilistic explanation for the phenomenon of measurement error, or the "noisy process." Galileo and others had been thinking about this problem in the context of astronomy for centuries. Another distribution of similar form, the Gaussian, overtook the Laplace as the scientific choice for error modeling when it was introduced by Gauss in the early 19th century.

Taking logs gives the log-likelihood, which is easier to optimize,

$$\begin{aligned}\ell(\mu, b) &= \sum_{i=1}^{n} \log\left(\frac{1}{2b}\exp\left(-\frac{|x_i-\mu|}{b}\right)\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{2b}\right) + \log\left(\exp\left(\frac{-|x_i-\mu|}{b}\right)\right) \\ &= -n\log(2b) - \frac{1}{b}\sum_{i=1}^{n}|x_i-\mu| \end{aligned} \tag{1}$$

Recall that we can define the absolute value function piecewise,

$$|x_i-\mu| = \begin{cases} x_i-\mu & x_i \geq \mu \\ \mu - x_i & x_i < \mu \end{cases}$$

In order to optimize, take the derivative with respect to μ, and set equal to zero. Note that in the below, we take the derivative of the absolute value function with respect to μ, which yields the piecewise result.

$$\frac{\partial}{\partial\mu}|x_i-\mu| = \begin{cases} -1 & x_i \geq \mu \\ 1 & x_i < \mu \end{cases} = \operatorname{sgn}(x_i < \mu)$$

$$\begin{aligned} \frac{\partial\ell(\mu,b)}{\partial\mu} &= 0 \\ \frac{\partial}{\partial\mu}\left(-n\log(2b) - \frac{1}{b}\sum_{i=1}^{n}|x_i-\mu|\right) &= 0 \\ -\frac{1}{b}\left(\sum_{i=1}^{n}\operatorname{sgn}(\mu-x_i)\right) &= 0 \end{aligned}$$

This is true when μ is the *median* of the data.

2. Find the maximum likelihood estimate of b.

Solution. Take the derivative of (1) with respect to b, then set this expression equal to zero, in order to obtain the optimum.

$$\begin{aligned} \frac{\partial\ell(\mu,b)}{\partial b} &= 0 \\ \frac{\partial}{\partial b}\left(-n\log(2b) - \frac{1}{b}\sum_{i=1}^{n}|x_i-\mu|\right) &= 0 \\ -\frac{n}{b} + \frac{1}{b^2}\sum_{i=1}^{n}|x_i-\mu| &= 0 \\ \frac{n}{b} &= \frac{1}{b^2}\sum_{i=1}^{n}|x_i-\mu| \\ b_{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^{n}|x_i-\mu| \end{aligned}$$

This is the average of *absolute deviations* from the mean.

3. Assume that μ is given. Show that b_{MLE} is an unbiased estimator (to show that the estimator is unbiased, show that $\mathbb{E}[b_{\text{MLE}} - b] = 0$.

   Solution.

$$\begin{aligned}\mathbb{E}[b_{\text{MLE}} - b] &= \mathbb{E}[b_{\text{MLE}}] - \mathbb{E}[b] \\ &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|\right) - b \\ &= \mathbb{E}(|X - \mu|) - b\end{aligned}$$

   In order to compute the former term, we define a new random variable $Z = X - \mu$ must be Laplacian. Since the Laplacian is symmetric, we can double the density for the positive numbers. Therefore, the density of $|Z|$ is,

$$p(z; \mu, b) = \frac{1}{b}\exp\left(-z/b\right)$$

   Then, we can compute

$$\begin{aligned}\mathbb{E}(|Z|) &= \mathbb{E}\left(\frac{1}{b}\exp\left(-z/b\right)\right) \\ &= \frac{1}{b}\int_0^\infty z\exp\left(-z/b\right)dz \\ &= b\end{aligned}$$

   Which completes the proof.

5 Transforming a Standard Normal Multivariate Gaussian

We are given a 2 dimensional multivariate Gaussian random variable Z, with mean 0 and covariance I. We want to transform this into something cooler. Find the covariance of a multivariate Gaussian such that the axes x_1 and x_2 of the isocontours of the density are elliptically shaped with major/minor axis lengths in a 4:3 ratio, and the axes are rotated 45 degrees counterclockwise.

Solution. First, we state a result from lecture that has significant bearing on this problem. It is known that,

$$|Ax|^2 = x^\top A^2 x = 1$$

is an ellipsoid with orthonormal eigenvectors $v_1, \ldots, v_n$ and radii $1/\lambda_1, \ldots, 1/\lambda_n$. This gives something into which to sink our teeth. We transform the standard basis into the basis in question by rotation each eigenvalue by 45 degrees. It is possible, of course, to compute this rotation using a rotation matrix, but this seems silly, since we can just note that the direction of $\begin{bmatrix}1 & 0\end{bmatrix}$ becomes $\begin{bmatrix}1 & 1\end{bmatrix}$, and that the direction of $\begin{bmatrix}0 & 1\end{bmatrix}$ becomes $\begin{bmatrix}1 & -1\end{bmatrix}$. These vectors are orthogonal, but not orthonormal, so we normalize the vectors to obtain the matrix of eigenvectors,

$$U = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1 \\ 1 & -1\end{bmatrix}$$

The eigenvalues must be 4 and 3, respectively, so we can solve,

$$\begin{aligned}\frac{1}{\lambda_1} &= 4 \\ \lambda_1 &= \frac{1}{4} \\ \frac{1}{\lambda_2} &= 3 \\ \lambda_2 &= \frac{1}{3}\end{aligned}$$

Since the quadratic form contains a squared term, which in this case is Σ^{-1}, the eigendecomposition $U\Lambda U^\top$ is really for $\Sigma^{-1/2}$!. Now that we have this matrix, we can compute Σ by squaring $\Sigma^{-1/2}$ and then taking the inverse, which gives us the correct result.

$$\begin{aligned}\Sigma^{-1} &= (\Sigma^{-1/2})^2 \\ &= U\Lambda U^\top U\Lambda U^\top \\ &= U\Lambda^2 U^\top \\ &= \begin{bmatrix} 25/144 & 7/144 \\ 7/124 & 25/144 \end{bmatrix} \\ \Sigma &= (\Sigma^{-1})^{-1} \\ &= \left(\begin{bmatrix} 25/144 & 7/144 \\ 7/124 & 25/144 \end{bmatrix}\right)^{-1} \\ &= \frac{1}{2}\begin{bmatrix} 25 & -7 \\ -7 & 25 \end{bmatrix}\end{aligned}$$