STAT 383C: Statistical Modeling I **Fall 2016**

Lecture 3 — September 1

Lecturer: Purnamrita Sarkar *Scribe: Giorgio Paulon, Carlos Zanini*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

3.1 Multivariate Calculus and MLEs

In the previous lecture, we studied the Maximum Likelihood Estimators (MLEs) in the case of scalar random variables. Let us analyze now the case of multivariate random variables: that is, we have n data points, each of them having k different features. The generic i-th data point is then a vector of dimension k, denoted by $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ik})^T$. For the sake of simplicity, we consider in this course only the case $n > k$. In this framework, the parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ we want to estimate is a vector with dimension k as well.

The notion of first derivative has to be generalized to the multivariate case.

Definition 3.1. ***(Gradient vector).*** *Given a score function $l : \mathbb{R}^k \to \mathbb{R}$, the gradient is the vector of the partial derivatives*

$$\nabla l = \left(\frac{\partial l}{\partial \theta_1}, \dots, \frac{\partial l}{\partial \theta_k} \right)^T.$$

Analogously, we define the generalization of the second derivatives.

Definition 3.2. ***(Hessian matrix).*** *The Hessian of the function as the matrix H of the second derivatives, whose element (i, j) is given by*

$$H_{ij} = \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \quad \forall i, j \in \{1, \dots, k\}.$$

In the following, ∇l and H will always denote the gradient and the Hessian matrix of the log-likelihood function.

The Fisher information matrix is now defined as

$$I(\boldsymbol{\theta})_{ij} = -\mathbb{E}[H_{ij}] \quad \forall i, j \in \{1, \dots, k\}$$

where the expectations are taken at each entry of the corresponding matrix.

3.2 Maximization algorithms

In the previous section, we studied criteria to find the maximum of univariate functions. However, if the log-likelihood is more complex and it does not have a closed form solution, these approaches cannot be used. Suppose now the log-likelihood is convex. Several iterative algorithms can be used in order to find its optimum. In this section, we briefly discuss two algorithms: the **Gradient ascent** and **Newton-Raphson** method.

3.2.1 Gradient ascent

The gradient ascent is an iterative method used to maximize an objective function. In our case, the function of interest is the log-likelihood of the data, as a function of the parameters $\boldsymbol{\theta}$.

The method starts from an initial guess $\boldsymbol{\theta}^{(0)}$, where the superscript denotes the number of the iteration. At the generic t-th step, the updating rule is the following:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \alpha \nabla l(\boldsymbol{\theta}^{(t)}) \qquad \text{while } ||\nabla l(\boldsymbol{\theta}^{(t)})|| > \varepsilon \tag{3.1}$$

that is, the next point is chosen with a small step in the gradient direction (i.e. the direction of maximum local growth of the function). The stopping criterion simply serves to avoid useless iterations when the algorithm has already reached convergence. The step size α is a critical tuning parameter: for values of α too large the algorithm may overshoot the maximum; for values too small, it converges too slowly.

3.2.2 Newton-Raphson

The Newton-Raphson algorithm, in general, works more efficiently than the gradient ascent. This method, in fact, exploits even the information given by the second derivatives of the goal function, i.e. its curvature, to converge more rapidly. However, calculating (and inverting) the Hessian at each iteration may be computationally infeasible for large k.

Let us suppose we have a quadratic function, that is

$$l = \boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^T C \boldsymbol{\theta}.$$

This function is concave if and only if the matrix C is negative semidefinite. Mathematically speaking, $C \preceq 0$, i.e. $\forall \boldsymbol{a} \in \mathbb{R}^k$, $\boldsymbol{a}^T C \boldsymbol{a} \leq 0$.

We can maximize the function by differentiating it and by setting the result equal to 0:

$$\nabla l(\boldsymbol{\theta}) = \boldsymbol{b} + C\boldsymbol{\theta} = 0.$$

If C is negative definite (and therefore invertible), we can solve for $\boldsymbol{\theta}$ getting

$$\boldsymbol{\theta} = -C^{-1}\boldsymbol{b}.$$

Therefore, Newton-Raphson method converges in one iteration for quadratic functions.

When the log-likelihood is more complex, we can still apply this method by maximizing at each step the local quadratic approximation. According to Taylor's formula,

$$l(\boldsymbol{\theta}^{(t)} + \boldsymbol{h}) \approx l(\boldsymbol{\theta}^{(t)}) + \nabla l(\boldsymbol{\theta}^{(t)})^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T H_{\boldsymbol{\theta}^{(t)}} \boldsymbol{h}. \tag{3.2}$$

Maximizing (3.2) corresponds to taking as an ascent direction

$$\boldsymbol{h}^* = -H^{-1}_{\boldsymbol{\theta}^{(t)}} \cdot \nabla l(\boldsymbol{\theta}^{(t)}).$$

Therefore the general updating rule is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - H^{-1}_{\boldsymbol{\theta}^{(t)}} \cdot \nabla l(\boldsymbol{\theta}^{(t)})$$

where, in analogy with (3.1), a step size α can be introduced.

3.3 Convergence of multivariate random variables

In the first lecture we defined several notions of convergence for scalar random variables, in particular the convergence in probability and the convergence in distribution. We now present the extension to the multivariate case.

Definition 3.3. ***(Convergence in probability).*** *The sequence of random vectors $\boldsymbol{Y}_n$ is said to converge in probability to the random vector $\boldsymbol{Y}$, denoted by $\boldsymbol{Y}_n \xrightarrow{\mathcal{P}} \boldsymbol{Y}$ if and only if each of the components of $\boldsymbol{Y}_n$ converges in probability to the components of $\boldsymbol{Y}$, i.e.*

$$Y_{ni} \xrightarrow{\mathcal{P}} Y_i, \quad \forall i \in \{1, \ldots, k\}.$$

The same analogy between joint and componentwise convergence does not still hold for the convergence in distribution. This latter, in fact, requires an additional condition.

Definition 3.4. ***(Convergence in distribution).*** *The sequence of random vectors $\boldsymbol{Y}_n$ is said to converge in distribution to the random vector $\boldsymbol{Y}$, denoted by $\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{Y}$ if and only if*

$$\boldsymbol{t}^T \boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{t}^T \boldsymbol{Y} \quad \forall \boldsymbol{t} \in \mathbb{R}^k,$$

that is, the univariate convergence in distribution must hold not only for the components of $\boldsymbol{Y}_n$ but for each projection of the vector on any unidimensional space.

In other words, the convergence in distribution of the joint vector implies the convergence of each component, but the opposite is not true.

Example 3.1. ***(Marginal convergence does not imply joint convergence).*** *Let us consider as a counterexample the random vector $(U_n, V_n)^T$, where $U_n \sim \mathcal{N}(0,1)$ and $V_n = (-1)^n U_n$. If we consider the projection along $\boldsymbol{t} = (1,1)$, we obtain*

$$U_n + V_n \sim \begin{cases} 2U_n & \textit{if } n \textit{ is even} \\ 0 & \textit{if } n \textit{ is odd} \end{cases}$$

which clearly does not converge, even if both the components U_n and V_n converge to $\mathcal{N}(0,1)$.

3.4 Estimating the asymptotical variance

All the properties of the MLEs that we analyzed during the last lecture still hold in the multivariate case. However, calculations can become pretty hard. As an example, we consider the Delta method in the multivariate case and then we present a sampling scheme in order to obtain the same result.

3.4.1 Multivariate Delta method

The Delta method is a useful technique to calculate the asymptotic variance of some function of an estimator. In fact, if

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, M)$$

then, for $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^m$, we have

$$\sqrt{n}(\boldsymbol{g}(\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \nabla\boldsymbol{g} M \nabla\boldsymbol{g}^T).$$

However, in high-dimensional cases, the computations become tricky. Therefore we introduce the parametric bootstrap, a simple technique which aims at estimating the asymptotic variance of any estimator and of its transformations.

3.4.2 Parametric Bootstrap

The parametric bootstrap is a simulation technique used to derive the asymptotic variance of estimators.

Let us consider $X_1, \dots, X_n \sim f(x;\theta)$. Of course, as usual we do not know the real parameter θ generating the data, but we know population density f (that is why we speak here of **parametric bootstrap**).

We can then use the following technique:

1. Sample $X_1^{(1)}, \dots, X_n^{(1)} \sim f(x;\hat{\theta}) \longrightarrow$ compute $\widehat{\theta_n}^{(1)}$ MLE;

2. Sample $X_1^{(2)}, \dots, X_n^{(2)} \sim f(x;\hat{\theta}) \longrightarrow$ compute $\widehat{\theta_n}^{(2)}$ MLE;

   $\vdots$

3. Sample $X_1^{(B)}, \dots, X_n^{(B)} \sim f(x;\hat{\theta}) \longrightarrow$ compute $\widehat{\theta_n}^{(B)}$ MLE.

We obtained a sample of MLE estimators, and therefore we can estimate the variance by

$$\widehat{\mathrm{Var}(\hat{\theta})} = \frac{1}{B}\sum_{i=1}^{B}(\widehat{\theta_n}^{(i)} - \overline{\widehat{\theta_n}})^2.$$

Now, if we wanted to compute the variance of the MLE of some function of θ, say e^{θ} the strategy is straightforward. According to the invariance principle, the MLE is $e^{\hat{\theta}}$, and the bootstrap techniques yields

$$\widehat{\mathrm{Var}(e^{\hat{\theta}})} = \frac{1}{B}\sum_{i=1}^{B}(e^{\widehat{\theta_n}^{(i)}} - e^{\overline{\widehat{\theta_n}}})^2.$$

3.5 Limitations of MLEs

Let us now see an example in which the MLE does not work as a good estimator. This will lead us to define a new kind of estimators, i.e. the shrinkage estimators, which are slightly biased estimators which allow to gain a lot in terms of variance.

3.5.1 The Neyman-Scott example

Consider n clusters of gaussian samples, which differ among themselves only by the population means, i.e.

$$\begin{aligned}
&X_{11}, X_{12}, \ldots, X_{1k} \sim \mathcal{N}(\mu_1, \sigma^2) \\
&\ldots \\
&X_{n1}, X_{n2}, \ldots, X_{nk} \sim \mathcal{N}(\mu_n, \sigma^2).
\end{aligned}$$

If our goal is to estimate the variance of the whole population σ^2, it seems obvious to consider as an estimator the mean of the sample variances of the single populations, that is,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(X_{ij} - \bar{X}_i)^2}{k}.$$

In fact, this is the MLE for σ^2.

However, it can be shown that this estimator does not converge to the real parameter σ^2. This is a consequence of the fact that the number of parameters grows with n.