Foundations of the MVUE algorithm

Pushpendre Rastogi
v1: 13 Mar 2021
last update: March 14, 2021
Inspiration: The concepts used in these notes were inspired from Prof. Charles Elkan's notes for CSE291 (2005) available at http://cseweb.ucsd.edu/~elkan/291winter2005/lect0[2-6].pdf

1 How do we measure how good is a parameter estimate?

We assume that data is generated from p_θ. Given this assumption we want to estimate either θ itself or some function of θ, like $g(\theta)$. An estimator E for $g(\theta)$ in general is a function/algorithm that maps a dataset $\mathcal{D}$ to a value that is "close" to $g(\theta)$. The actual value of $\mathrm{E}(\mathcal{D}) \triangleq \hat{g}$ is obviously a random variable and there are different criteria for analysing the quality of E. For example, one way of quantifying the quality of E is through its **Mean-Squared-Error (MSE)** under the assumed distribution, i.e. $\mathbb{E}[(\hat{g}(\mathcal{D})-g(t))^2]$. Another criterion is called unbiasedness that $\mathbb{E}[(\hat{g}(\mathcal{D})] = g(t)$ and a third one is consistency that as the dataset size increases the estimator's bias goes to zero – As $|\mathcal{D}| \to \infty$ then $\mathbb{E}[(\hat{g}(\mathcal{D})] \to g(t)$. This is a property of *eventual* unbiasedness, a.k.a. an asymptotic property.

Another important direction for quantifying the quality of an estimator is by quantifying the **variance of an estimator** for a given sample size, or equivalently by giving **high probability bounds on the deviation** of the estimator from the true quantity. These are typically finite sample bounds.

Another type of description is of the asymptotic distribution followed by the deviation between the estimator and the true quantity. These types of descriptions are given by the central limit theorems.

2 Sufficient Partitions and their Sufficient Statistics

A statistic is called sufficient if it preserved all information from x that is relevant for estimating which distribution P_θ generated x.

In order to understand the concept of sufficient statistics we need to think about sufficient statistics. That's the key and then basically a sufficient statistic just becomes the indicator of the partition that a sample falls into.

Let's say we have a sample space of our dataset. For example, if we toss a coin n times then $\{0,1\}^n$ is the sample space. This can be partitioned into a few sets, e.g. the dataset of n coin tosses can be partitioned into $n+1$ sets according to whether the number of 1's was 0, or 1, or 2, ..., or n. Now this partition is "sufficient" if the likelihood of different parameters is the same within a set of the partition for all sets in the partition.

More formally, for a given partition, i.e. a set of sets that covers the sample space, $\{A\}$ if for every A, $P_\theta(x|x \in A) = P_{\theta'}(x|x \in A)$ for all θ, θ' then $\{A\}$ is called a sufficient partition.

The punch line is that if the level-sets (the contours) of a function induce a sufficient partition then that function of the data is a sufficient statistic.

Alternative characterization An alternative way of describing a sufficient statistic is that t is a sufficient statistic wrt to the family of distributions $\{p_\theta\}$ if the conditional pdf $P(\mathcal{D}|t(\mathcal{D}) =$ some value$)$ is invariant with θ.

3 The Rao Blackwell Theorem

The RBT says the following

Rao-Blackwell Theorem

Let $\{P_\theta\}$ be a family of distributions on a sample space X. Suppose $\tilde{g}(\mathcal{D})$ be an unbiased estimator of $g(\theta)$. Let $t(\mathcal{D})$ be a sufficient statistic. Let $\mathcal{D}$ be give and let $t_0 = t(\mathcal{D})$.
Then $\hat{g}(\mathcal{D}) = \mathbb{E}_{\mathcal{D}'}[\tilde{g}(\mathcal{D}') \mid t(\mathcal{D}') = t_0]$ has three properties:
1. $\hat{g}$ is a function of only the dataset $\mathcal{D}$.
2. $\hat{g}$ is unbiased.
3. $\hat{g}$ has lower variance than $\tilde{g}$.

4 The MVUE Algorithm

The algorithm to obtain a MVUE has four steps

1. Find a sufficient statistic t, given Θ and the sample space $\mathcal{X}$ and the dataset $\mathcal{D} \in \mathcal{X}$.
2. Show that the family of distributions of t is complete.
3. Find a crude unbiased estimator $\tilde{g}(\mathcal{D})$
4. Evaluate $\hat{g}(t(x)) = \mathbb{E}_\theta[\tilde{g}(\mathcal{D}') \mid t(\mathcal{D}') = t(\mathcal{D})]$

Factorization theorem: A trick for identifying sufficient statistics

A statistic t is sufficient for a family of distributions P_θ iff. we can write $P_\theta(\mathcal{D}) = f(\theta, t(\mathcal{D}))h(\mathcal{D})$.
Often a dataset $\mathcal{D}$ is collected from iid observations therefore $P_\theta(\mathcal{D}) = \prod P_\theta(x)$, therefore if $P_\theta(x) = f(\theta, t(x))h(x)$ then the above requirement is also satisfied.

What is a **complete** family of distributions?
A family of distributions Θ is complete if $\mathbb{E}_\theta[f(y)] = 0 \; \forall \theta \in \Theta \implies f = 0 \; a.e.$.
In other words Θ is big enough that if $f \neq 0$ $a.e.$ then there exists some $\theta \in \Theta$ that $\mathbb{E}_\theta[f(y)] \neq 0$.