Final Project Proposal

Eric Tay, Adithya Vellal, Grace Tian

March 2, 2022

Introduction and Background

The problem of estimating optimal data-generating parameters from an observed dataset is ubiquitous in statistics. Given a dataset of independent and identically distributed (i.i.d) observations $x_1, \ldots, x_n$, we are tasked with determining the probability distribution parameters θ that best represent the joint distribution $p(x_1, \ldots, x_n|\theta)$. Oftentimes this is formulated as a simple maximization problem:

$$\max_{\Theta} p(x_1, \ldots, x_n|\theta)$$

Depending on the choice of distribution used to model the data and the specific parameters being optimized, this optimization varies in difficulty. In most cases, there is no closed form solution for θ. We must thus resort to numerical, iterative methods in order to obtain an estimate for θ.

This project will investigate various procedures for maximum likelihood estimation (MLE). Although there are many procedures available for MLE, we will focus on the following methods: gradient descent, Newton-Raphson, Broyden-Fletcher-Goldfarb-Shanno (BFGS), and Fisher's scoring.

We plan to implement these methods in Python with various probability distributions and generated data for the purpose of demonstrating situations in which one method may be preferred over another.

Methods

The procedures that we focus on are iterative. We start with a guess $\hat{\theta}_1$ and update successive estimates until we converge at a satisfactory $\hat{\theta}_r$:

$$\hat{\theta}_{r+1} = \hat{\theta}_r + \eta_r \mathbf{d_r}(\hat{\theta})$$

$\mathbf{d_r}(\hat{\theta})$ is the function we choose to signal the direction in which to step, while η_r is the step size.

Various methods use different combinations of $\mathbf{d}$ and η_r to solve this problem:

- Gradient Descent: $\mathbf{d}_r(\hat{\theta}) = \Delta l(\hat{\theta}_r; y)$
- Newton-Raphson: $\eta_r = 1$ and $\mathbf{d}_r(\theta) = -\mathbf{H}_r^{-1}(\hat{\theta})\mathbf{s}_r(\hat{\theta})$ where $\mathbf{s}_r(\theta)$ is the gradient of the log-likelihood with respect to θ and $\mathbf{H}_r^{-1}$ is the inverse Hessian of the log-likelihood.
- BFGS: $B_{k+1} = B_k + \frac{y_k y_k^\top}{y_k^\top s_k} - \frac{B_k s_k s_k^\top B_k^\top}{s_k^\top B_k s_k}$, where $y_k = \nabla \ell(x_k + s_k) - \nabla \ell(x_k)$ and $s_k = x_{k+1} - x_k . s_k = x_{k+1} - x_k$.
- Fisher's scoring: $\eta_r = 1$ and $\mathbf{d}_r(\hat{\theta}) = -\mathcal{I}^{-1}(\hat{\theta})\mathbf{s}_r(\hat{\theta})$ where $\mathcal{I}(\theta) = \mathbb{E}[\mathbf{H}_r(\hat{\theta})]$ is the Fisher information matrix.

Outline

We shall first introduce the different methods in consideration, motivating their conception and discussing theory when appropriate. We will also mention potential applications of the methods and their uses in recent research.

We will then evaluate each algorithm experimentally, highlighting their strengths and weaknesses in various scenarios. The experimental procedure we will use is as follows:

1. Choose a data generating distribution
2. Set true parameters θ and generate data accordingly
3. Obtain MLE estimates $\hat{\theta}$ using each of the methods listed above
4. Evaluate both the accuracy and computational efficiency of each method

References

[1] Lam, Adrian. "BFGS in a Nutshell: An Introduction to Quasi-Newton Methodsv." Medium, Towards Data Science, 22 Feb. 2022, https://towardsdatascience.com/bfgs-in-a-nutshell-an-introduction-to-quasi-newton-methods-21b0e13ee504.

[2] Schworer, Andrew and Hovey, Peter, "Newton-Raphson Versus Fisher Scoring Algorithms in Calculating Maximum Likelihood Estimates" (2004). Undergraduate Mathematics Day, Electronic Proceedings. Paper 1. http://ecommons.udayton.edu/mth$_e pumd/1$

[3] Young, G. A. (2019) Mathematical Statistics: An Introduction to Likelihood Based Inference Richard J. Rossi John Wiley Sons, 2018, ebook ISBN: 978-1-118-77104-4, LCCN 2018010628 (ebook). International Statistical Review, 87: 178– 179. https://doi-org.proxy.lib.duke.edu/10.1111/insr.12315.