Chapter 8
Transforming the Method of Least Squares to the Dataflow Paradigm

Ilir Murturi
https://orcid.org/0000-0003-0240-3834
Distributed Systems Group, TU Wien, Austria

ABSTRACT

In mathematical statistics, an interesting and common problem is finding the best linear or non-linear regression equations that express the relationship between variables or data. The method of least squares (MLS) represents one of the oldest procedures among multiple techniques to determine the best fit line to the given data through simple calculus and linear algebra. Notably, numerous approaches have been proposed to compute the least-squares. However, the proposed methods are based on the control flow paradigm. As a result, this chapter presents the MLS transformation from control flow logic to the dataflow paradigm. Moreover, this chapter shows each step of the transformation, and the final kernel code is presented.

INTRODUCTION

The Method of Least Squares (MLS) is one of the oldest modern statistic methods and a standard regression analysis approach. The well-known technique determines the best fit line to a set of data points and it is used to estimate the parameters. The MLS use can be traced to Greek mathematics, where the predecessor is considered to be Galileo. The first approach towards a modern method description was introduced by the French mathematician Adrien Merie Legendre in 1805, which has been contested by Karl F. Gauss and Pierre S. Laplace. Gauss, in his memoir published in 1809, mentioned that he had discovered MLS and used it at the beginning of 1795 in estimating the orbit of an asteroid. Harter presents in detail the history and pre-history of the MLS (Harter, 1975).

Many opinions exist related to the authorship of the method. Placket presents some conclusions related to how the technique was discovered (Placket, 1972). Over the years, there have been attempts to

DOI: 10.4018/978-1-7998-7156-9.ch008

propose techniques to determine the best fit line to a set of data points. However, the least-squares method is considered the most important among the methods used to find or estimates numerical values to fit a function to a set of given data points. The MLS exists in different variations, and the most well-known ones are Ordinary Least Squares (OLS) and Weighted Least Squares (WLS).

Data collection is essential for any modern system, and often in real-world scenarios, such data may have linear relationships. Therefore, several approaches and platforms utilize MLS to analyze the correlation between various variables or data values (e.g., see (Xu et. al., 2013)). In this context, as well-known data collection platforms are crowdsourcing systems (Murturi et.al, 2015). Data collection from the crowd is an essential part of any crowdsourcing system (Rexha et. al., 2019). Analyzing such data is a way of answering the critical questions for different processes. In crowdsourcing systems, the quality of contributions remains plagued by poor quality, and often such data is often too hard to interpret. Therefore, statistical calculations help explain crowd data into meaningful results or identify the relationships between different data. By calculating the correlation between answers in collected data useful results can be obtained. However, since several MLS implementations exist in the control flow paradigm, we present the implementation and the method's transformation with a straightforward example in this chapter. Note that we do not evaluate the MLS's performance aspects in the dataflow paradigm in this chapter. Beside that, this chapter aims to show the simplicity of transforming control flow logic applications into the dataflow paradigm.

The remainder of this chapter is organized as follows. Section II presents related work regarding the MLS. Section III describes in more detail the method by including the mathematical background. Section IV, showes the implementation of the method in the control flow paradigm. Afterward, the method transforming steps to the dataflow paradigm is described by analyzing the current implementation in CPU code given in C++. Final remarks are given in Section V.

RELATED WORK

The MLS is known to be published for the first time by Adrien Marie Legendre (Legendre, 1805). However, the method was not used as well as no mathematical proof was given. Legendre formulate the problem and starts with the linear equation of the form $E = a + bx + cy + ...,$ where the requirement determines unknown variables $\left(x, y\right)$ that E decreases to zero or a very small number for each equation. However, such equations are derived without the explicit use of calculus. The equations are generated by multiplying the linear form in the unknowns by the coefficient a, b, c. Each of the unknowns are summed over all the observations and then setting the sums equal to zero (Harter, 1975). When the results produce errors, the proposed approach rejects the equations that produce such error while determining the unknowns from the rest of the equations. On the other side, Puissant, discusses theoretical aspects of the MLS (Puissant, 1805). He also presents an application to the determination of the earth's ellipticity from measures of degrees meridian.

Several research papers use the MLS technique to determine the best fit line to a set of given data points. To compute least squares for the large sparse systems, Shi et al. provide a survey of distributed least squares in distributed networks, sketches the algorithm's skeleton first, and analyzes time-to-completion and communication cost (Shi et al., 2017). Furthermore, the study provides helpful insights into the methods that can be modified to run in a distributed manner to solve linear least-squares. MLS