# StreamFitter: A Real Time Linear Regression Analysis System for Continuous Data Streams

Chandima Hewa Nadungodage, Yuni Xia, Fang Li, Jaehwan John Lee, Jiaqi Ge
Indiana University-Purdue University Indianapolis, USA

**Abstract.** In this demo, we present the StreamFitter system for real-time regression analysis on continuous data streams. In order to perform regression on data streams, it is necessary to continuously update the regression model parameters while receiving new data. In this demo, we will present two approaches for on-line, multi-dimensional linear regression analysis of stream data, namely Incremental Mathematical Stream Regression (IMSR) and Approximate Stream Regression (ASR). These methods dynamically recompute the regression function parameters, considering not only the data records of the current window, but also the synopsis of the previous data. Therefore, the refined parameters more accurately model the entire data stream. The demo will show that the proposed methods are not only efficient in time and space, but also generate better fitted regression functions compared to the traditional sliding window methods and well-adapted to data changes.

**Keywords:** Linear regression, Data streams, Sliding windows.

## 1 Introduction

With the emergence of network and sensor technology in recent times, there is a growing need of data stream management and mining techniques for collecting, querying, and analyzing data streams in real time [1]. Regression analysis is a widely used technique for the modeling and analysis of the relationship between dependent variables and independent variables [4]. In the recent years there is a focus on applying regression analysis techniques to model and predict the behavior of the stream data. In this context, it is required to continuously update the regression model as new data streams in, on the other hand, it is impossible to scan the entire data set multiple times due to the huge volume of the data. Therefore, it is necessary to incrementally reconstruct the regression model using one-scan algorithms. One widely used approach is to consider the current window of data to construct the regression model. Although this approach is efficient in terms of time and space, it has poor performance when accuracy of the model is considered [3].

In this demo, we will present StreamFitter system which will facilitate real-time regression analysis on continuous data streams. It demonstrates two new methods for on-line, multi-dimensional linear regression analysis of stream data namely IMSR (Incremental Mathematical Stream Regression) and ASR (Approximate Stream Regression). Both methods use a window based approach to prevent multiple scans of the entire data set, nevertheless they are able to maintain the accuracy of the regression model.

## 2 Linear Regression Analysis on Data Streams

We implemented the StreamFitter System using C++ language. In this demo, we will show and compare three regression approaches: IMSR regression, ASR regression and the original Sliding Window (SW) regression approach. We will use real stream data such as the financial data from Tokyo Stock Exchange (TSE) [5] and sea surface temperature data from Tropical Atmosphere Ocean project (TAO) [6]. StreamFitter has a convenient user interface which allows users to specify the data source, window size, half life (which is used in ASR regression), number of independent variables in the data set (IV Count), and the regression methods. If the user needs to see a comparison between different regression methods, multiple methods can be selected. User can also choose to input offline data stream recorded in a file or receive a data stream giving a network IP and a port number.

### 2.1 Incremental Mathematical Stream Regression (IMSR)

Figures 1 show IMSR on sea surface temperature (SST) measurements gathered hourly from January 2000 to December 2000, by a moored buoy located in the tropical pacific [6]. There were around 8700 records and we used a window size of 1000 records. Data points are plotted in yellow and IMSR regression line is plotted in blue. Figure 2 shows how IMSR regression line has dynamically adjusted over the time, as the new data streamed in.

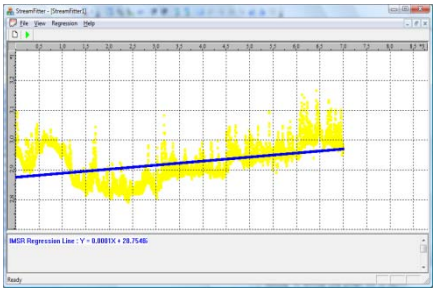

**Fig. 1** IMSR regression over SST data

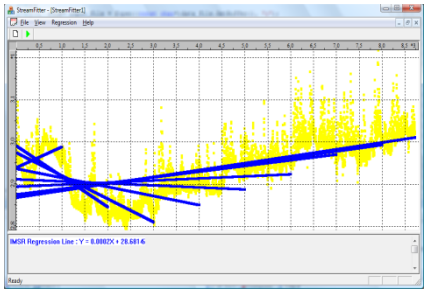

**Fig. 2.** Dynamic adjustment of IMSR regression line

Assume the window size is $n$ data records and the number of independent variables (IVs) is $p$. Thus the values of IVs is a $n*p$ matrix, denoted by $X$. Values of dependent variable is a $n*1$ vector, denoted by $y$. Values of regression coefficients is $p*1$ vector, denoted by $\beta$. This $\beta$ can be calculated using Ordinary Least Squares (OLS) method as $\beta = (X'X)^{-1}(X'y)$. Hence the regression parameters for the $k^{th}$ window will be calculated as $\beta_k = (X_k'X_k)^{-1}(X_k'y_k)$. However, this calculation is based only on the records in current window. In order to improve the accuracy of the model previous data records should also be considered. Therefore, we should maintain an effective synopsis of pervious data tuples which can be used later.

The idea of IMSR is as follows: If values of the IVs for the $1^{st}$ window is $X1$, values of the IVs for the $2^{nd}$ window is $X2$, and the total data set available so far is $X$,

it can be proved that *(X'*X) = (X1'*X1 + X2'*X2)*. Similarly, if values of the dependent variable for the 1st window is *y1*, values of the dependent variable for the 2nd window is *y2*, and the total data set available so far is *y*, it can be proved that *(X'*y) = (X1'*y1 + X2'*y2). X'*y* is always a vector of size *p**1. Therefore to calculate refined $\beta$ values for a particular window, we only have to maintain the sum of *X'X* products for previous windows and sum of *X'y* products for previous windows. Let us refer the sum of *X'X* products as M and sum of *X'y* products as *V*, then the refined parameter vector for the $k^{th}$ window can be computed as

### 2.2 Approximate Stream Regression

The second regression method we will present is the approximate stream regression (ASR), which refines the values of the parameter vector $\beta$ for a particular window considering the previous data records. The Brown's Simple Exponential Smoothing or Exponential Weighted Moving Average (EWMA) [2] has been widely used method for time series prediction for a long time. We use this EWMA method to refine the $\beta$ vector for the current window, considering the $\beta$ vector calculated from the previous window. For example, suppose we know the values of the $\beta$ vector for the $(k-1)^{th}$ window, using that we can calculate the value of the $\beta$ vector for the $k^{th}$ window as

[1] Expanding this equation we get,

Refined value is a weighted combination of the previous values and the current value. By varying the smoothing factor $\alpha$, we can adjust the importance of the historical data. The smaller $\alpha$, the more weight is assigned to the current data and the less historical data will affect the regression function parameters. For $\alpha = 0.5$ current value and past history are equally weighted.

In this demo, we will show that ASR performs well for non-stationary data when the smoothing factor is adjusted to give more weight to the recent samples. Figures 3-5 show ASR regression on daily REIT index of TSE [5] from Jan 2003 - Feb 2010. There were around 1700 records in this data set, we used a window size of 100 records. Data points are plotted in yellow, and ASR regression line is plotted in green. It is visible how the ASR regression line has dynamically adjusted over the time, as the new data streams in. As visible in figure 5, TSE REIT index has drastically dropped after mid 2007. So the nature of the data stream has significantly changed. ASR is capable of adjusting to this kind of fluctuations in the data stream as it gradually discards the historical records. Here we used half-life , so $\alpha$ = 0.25, therefore the historical records were rapidly discarded favoring the recent samples. Figure 6 shows a comparison of IMSR and ASR with traditional sliding window (SW) method. It is visible that both IMSR (blue) and ASR (green) has produced better fitted regression lines compared to SW (red).

---

[1] $\alpha$ – Smoothing factor, a constant value where $0 \leq \alpha \leq 1$. $\beta_k$ - Parameter vector for the $k^{th}$ window calculated considering only the data records of $k^{th}$ window. $\beta_{k-1}$ - Parameter vector for the $(k-1)^{th}$ window calculated considering all data records seen up to $(k-1)^{th}$ window. $\beta_k$ - Refined parameter vector for the $k^{th}$ window calculated considering all data records seen up to $k^{th}$ window.

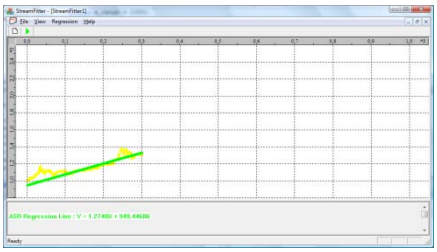

**Fig. 3** ASR at time t

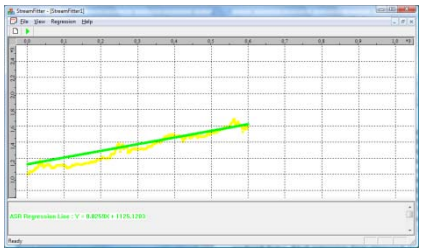

**Fig. 4** ASR at time t+3n

**Fig. 5** ASR at time t+kn

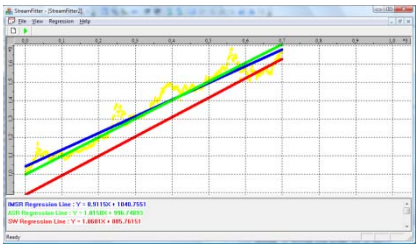

**Fig. 6** Comparison of IMSR, ASR, and SW

## 3 Conclusion

In the demo we will show StreamFitter, a real time linear regression analysis tool for stream data. We will show how the data streams are processed by StreamFitter, how the regression functions are generated in realtime and how the functions are dynamically adjusted as new data streams in. We will also compare and visualize various techniques in terms of function fitness and adaptiveness of concept drift.

## References


[1] B. Babcock, S. Babu, M. Datar, R. Motwani, and J.Widom, "Models and Issues in Data Stream Systems," PODS, March 2002.

[2] Brown, R. G. and Meyer, R. F. 1961, "The fundamental theorem of exponential smoothing," Operations Research, Vol. 9, No. 5, pp. 673-685.

[3] E. Keogh, S. Chu, D. Hart, and M. Pazzani. “Segmenting Time Series: A Survey And Novel Approach,” Data Mining in Time Series Databases, World Scientific Publishing Company, 2004.

[4] R. A. Berk. Regression Analysis: A Constructive Critique. Sage Publications, 2004.

[5] Tokyo Stock Exchange, http://www.tse.or.jp/english/market/topix/data/index.html

[6] Tropical Atmosphere Ocean project, http://www.pmel.noaa.gov/tao/index.shtml