Implement the CONCORD algorithm

This project was completed by Syed Rahman. The project was his own idea which he brought to us and completed as a summer intern in 2015.

The CONCORD algorithm implemented by Syed Rahman

The CONCORD algorithm is a method to estimate the true population of a co-variance matrix. The co-variance matrix is a summary of the relationship between every pair of fields in the data. Co-variance values close to zero indicate that the fields don't have a relationship. Values close to 1 indicate a positive relationship and values close to –1 indicate an inverse relationship.

In classic statistics there are many more observations than fields. In this case, the co-variance matrix of the sample is a good estimate for the true co-variance matrix.

Unfortunately, in big data, there any many cases where the number of fields exceeds the number of observations or may be close to the number of observations. It is the case that the sample co-variance matrix is a very poor estimate for the true co-variance matrix.

Read Syed's blog to find out more about his progress and experience and view his commits on github.

It's clear that Syed's addition to our Machine Learning Library is an important improvement, providing a way to getting more reliable results in this area.

Syed presented about his project on Community Day at the 2015 HPCC Systems® Engineering Summit at the end of September this year. His presentation demonstrates how this algorithm works and why it is a better method of getting the true population of a co-variance matrix. Watch his presentation: Syed Rahman and Kshitij Khare - Presenting about The CONCORD Algorithm (starts around the 30.00 mark). The presentation slides are also available.

For further details please refer to the following JIRA issue for this project.

In 2016, Syed was a returning student intern who completed a machine learning project which is related to this algorithm.

- Find out more about his second project to implement the Convex Sparse Cholesky Selection (CSCS) machine learning algorithm.
- View Syed's technical poster presentation on the CSCS algorithm displayed on Community Day at the HPCC Systems Engineering Summit in 2016 where he was a 3rd place winner.
- Watch a recording of his presentation on Understanding High-dimensional Networks for Continuous Variables Using ECL, on Community Day at the HPCC Systems Engineering Summit in 2016 or view the presentation slides.