Under TheHood

BY MADELEINE UDELL AND STEPHEN BOYD, PhD

Beyond Principal Components Analysis (PCA): Exploring Low Rank Models for Data Analysis

In many application areas, researchers seek to understand large collections of tabular data, for example, patient lab test results. The values in the table might be numerical (3.14), Boolean (yes, no), ordinal (never, sometimes, always), or categorical (A, B, O). As a practical matter, some entries in the table might also be missing.

To understand numerical data, a researcher might make a scatter plot; cluster the examples or the features; predict some of the values in the table based on others; remove (or simply identify) noisy or spurious values; or impute the values of missing entries. Many methods are available for any one of these specific tasks. By fitting a low rank model to the data, researchers can perform all of these computations simultaneously—even on large data sets containing heterogeneous values and many missing entries. Here, we describe what a low rank model is, give some examples of low rank models, and discuss how to pick a good low rank model for a particular application.

A low rank model approximates a table as the (matrix) product of two numerical matrices X and Y. Every example (e.g., patient) is represented by a row of X; every feature (e.g., lab test) is represented by a column of Y. The length of each of these rows and columns must be the same, and is called the rank of the model. A good low rank model compresses the information in the original data set using a rank that is much smaller than the number of rows or columns in the original table.

Principal Components Analysis (PCA), introduced by Karl Pearson in 1901, is a simple example of a low rank model. It finds a low rank model that minimizes the squared difference between the entries in the low rank model XY and those in the original data table.

PCA works well when the table consists only of numerical data with small, normal errors and has no missing entries. But often data does not fit these assumptions. In our lab test example, tests that have not been performed or survey questions left blank leave us with missing entries; malfunctioning sensors produce large, infrequent errors rather than small, normal errors. Moreover, PCA often returns a model that is difficult to interpret, and cannot be made to produce a model that captures our knowledge about the data, for example, it being nonnegative or sparse.

A number of methods have successfully extended PCA, each addressing one of these issues. These variations include nonnegative matrix factorization (which produces nonnegative factors), matrix completion (which handles missing data), robust PCA (which is less sensitive to noisy data), and sparse PCA (which produces factors with many zero entries).

A unified framework, which we call generalized low rank models, brings together the capabilities of these different techniques. It is able to simultaneously handle heterogeneous values, missing data, and prior beliefs about the factors. Even the well-known k-means clustering algorithm can be interpreted as a special case of a generalized low rank model. This framework makes it easier to use low rank models in everyday data analysis workflows. □

DETAILS

Madeleine Udell is a PhD candidate at Stanford University's Institute of Computational & Mathematical Engineering. She works with Stephen Boyd, PhD, professor of electrical engineering, with a focus on on convex optimization applications. The Boyd lab has developed and released a number of software packages for modeling and fitting generalized low rank models, available in different languages:

- **Julia (https://github.com/madeleineudell/LowRankModels.jl);**
- **Python (https://github.com/cehorn/GLRM); and**
- **Spark (http://git.io/glrmspark).**

The Julia and Spark packages are able to scale to datasets with billions of entries.

To learn how to use low rank models to produce scatter plots, cluster data, predict missing entries, and identify noisy or corrupted data, visit http://www.bcr.org/content/using-low-rank-models.