Draft 12/04/2009. Printing date 12/7/2009.

Singular Value Decomposition and Missing Values

Paul Purdom, Computer Science, Indiana University
Daniel Maki, Mathematics, Indiana University

Abstract

The *standard singular value decomposition problem* is to represent (or approximate) a given matrix as the sum of a small number of rank-one matrices. Given a matrix Y and a rank r, find matrices U, V, and D such that

$$y_{ij} = \sum_{1\le k\le r} u_{ik} d_{kk} v_{kj},$$

where the columns of U and the rows of V are orthonormal. This is a well studied problem that can be solved with high quality iterative algorithms. The related problem where some of Y is missing has received much less study. This problem can be generalized to approximating Y with

$$\sum_{1\le k\le r} w_{ij} u_{ik} d_{kk} v_{kj},$$

where W is a matrix of non-negative weights and where U and V must obey the constrains above. If element y_{ij} is missing then w_{ij} is zero. We discuss the alternating least squares method [4] for solving this problem. This algorithm has two difficulties: (1) for some problems and initializations, it approaches a non-optimum solution and (2) when it does converge, the rate is only linear. These limitations not-withstanding, it usually performs well enough to be extremely useful. We give a new iteration that often avoids the non-optimum solutions. We briefly consider the situation when there are known constraints on the value of unknown elements. When Y is an m by n array with missing values and the known elements of Y comes from a rank r matrix, it requires at least $(m+n-r)r$ elements to reconstruct all of Y. Also, each row and column of Y must have at least r elements. In favorable cases, this is all that is needed. Without a bound on r or some similar constraint, there is no way to reliably reconstruct Y. The paper has two parts: (1) a discussion of the theory and (2) sample calculations. Three of the sample problems are small problems concerning determining depth information from multiple photographs of a scene. Missing values arise in a natural way: some parts of the scene are not visible from some camera positions. The fourth sample problem has a matrix of voting results with almost three thousand rows and over a thousand columns. Knowing that the data has a low rank representation is a key part of the problem. The first problem is like this. The second and third have a low rank approximation. For the fourth problem there is only empirical evidence that the data has a low rank approximation.

1 Introduction

Problems in data analysis often involve a large matrix of results. In many cases the matrix either has low rank or at least can be well approximated by a low rank matrix. When this happens, factoring the data into the product of two *skinny matrices* often reveals important information about the problem. (We call a matrix skinny if it has just a few rows or if it has just a few columns.) Factoring the data into three factors where the middle factor is diagonal and the other two obey orthonormal conditions permits ordering the various rank-one contributions to the data in order of their importance.

For simple data sets, rank-one may be enough. Thus, if y_{ij} is the heat produced by burning a_i units of a fuel that produces b_j units of heat per unit, we expect to have

$$y_{ij} = a_i b_j, \qquad 1$$

perhaps with some small measurement errors. Other processes have a low rank representation, but with a rank that is higher than one. For example, if you photograph a scene using distant cameras at several angles, the position of the various features in the various photographs can be represented by the product of a matrix with three columns that represents the angle and magnification of the various cameras times a matrix with three rows that represent the three dimensional position of the various features. In this example, missing