ITERATIVE AGGREGATION METHOD FOR SOLVING PRINCIPAL COMPONENT ANALYSIS PROBLEMS

Vitaly E. Bulgakov

SemanticPro, Boston, USA

Abstract: Motivated by the previously developed multilevel aggregation method for solving structural analysis problems a novel two-level aggregation approach for efficient iterative solution of Principal Component Analysis (PCA) problems is proposed. The course aggregation model of the original covariance matrix is used in the iterative solution of the eigenvalue problem by a power iterations method. The method is tested on several data sets consisting of large number of text documents.

Keywords: principal component analysis, clustering method, power iteration method, aggregation method, eigenvalue problem

ПРИМЕНЕНИЕ ИТЕРАЦИОННОГО АГРЕГИРОВАНИЯ В МЕТОДЕ ГЛАВНЫХ КОМПОНЕНТ

В.Е. Булгаков

SemanticPro, г. Бостон, США

Аннотация. Предлагаемый метод мотивирован разработанным ранее итерационным подходом многоуровневого агрегирования для решения задач расчёта строительных конструкций. При решении задач методом главных компонент применяется метод степенных итераций на основе агрегированной модели исходной ковариантной матрицы. Метод проверен при тестировании ряда массивов данных, каждый из которых включает большое количество текстовых документов.

Ключевые слова: метод главных компонент, кластерный анализ, итерационный степенной метод, метод агрегирования, проблема собственных значений

1. INTRODUCTION

This work was envisioned as application of the *multilevel aggregation method* [1] developed by the author back in 90s to PCA problems. Multilevel aggregation method was an extension of well-known *multigrid* methods [2] from boundary value problems to general structural analysis problems which brought it to the class of *algebraic multigrid* methods. The idea of the aggregation method was to use some naturally constructed course model of the original finite element approximation of a structure which provides a fast convergence for iterative methods for solving large algebraic systems of equations. One of applications of this method was an iterative solution of large eigenvalue problems arising in structural natural vibration and buckling analyses [3]. In these problems a sought set of lowest vibration modes can be thought of as *principal components* of structure behavior. An obvious similarity with PCA was a turning point to start looking for a proper way to create an aggregation model for data matrix approximation and use it for efficient solution of PCA problems.

In this study PCA [4] is applied to and the method is tested on text analysis problems. A tested data set consists of documents each of which produces an N-dimensional vector stored as a column of a data matrix which values are term frequencies. Our raw data comes in the form of text files from data sets such as medical abstracts and news groups. The purpose of PCA

is to iteratively compute a set of highest eigenvalues and corresponding eigenvectors of the covariance matrix. Covariance matrix is never formed explicitly. The main operation is multiplication of large sparse data matrix or its transpose by a vector. The course aggregation model of the original covariance matrix is used in the iterative solution of the eigenvalue problem. Original covariance matrix and its approximation of small size assumes similarity of leading eigenvalues and eigenvectors. This fact allows fast convergence of subspace iterations at minimal additional computational cost.

For numerical experiments we use R language which is rich of linear algebra, statistical and graphical packages.

2. PCA PROBLEM FORMULATION

PCA in multivariate statistics is widely used as an effective way to perform unsupervised dimension reduction. The essence of this method lies in using Singular Value Decomposition (SVD) which provides the best low rank approximation to original data according to Eckart-Young theorem [5]. Let n data points in m dimensional space be contained in the data matrix which is assumed already centered around the origin for computational stability

$$(x_1, x_2, ..., x_n) = X \tag{1}$$

Then covariance matrix is

$$A = XX^T \tag{2}$$

Let (λ_k, φ_k) be an eigenpair of A, where eigenvectors φ_k define principal directions.

3. AGGREGATION MODEL

In order to create an aggregation model we divide the entire set of data vectors x_i into n_0 clusters using some similarity criteria where

$n_0 \ll n$. We will explain later how we do clustering. We assume that all vectors within the cluster are similar and a single representative of a cluster is an average of all vectors x_i where $i \in cluster_k$ or for $cluster_k$ we have

$$x_k^0 = \frac{1}{dim_k} * \sum_{i \in cluster_k} x_i \tag{3}$$

Transformation of matrix X to X_0 is done using matrix R which we call *aggregator*

$$X_0 = XR \tag{4}$$

where $R[i, k] =$ if $i \in cluster_k$ then $\frac{1}{dim_k}$ else 0.

X_0 is of size (m, n_0).

Approximation A_0 of covariance matrix A is

$$A_0 = X_0X_0^T = XRR^TX^T \tag{5}$$

Formally matrix A0 is of the same size as A but has a much lower rank. We do not need to use form (5) for computations. For matrix vector multiplication we rather use sparse matrix X0 which according to (3) is constructed by simple averaging of vectors inside a cluster and

$$A_0v = X_0X_0^T \tag{6}$$

Therefore A_0v requires $O(mn_0)$ operations which is much lower than $O(mn)$ operations required for Av. We also expect and this is confirmed by numerical experiments that convergence of iterative methods for solving partial eigenvalue problem for A_0 is faster than that for A.

There are quite a few clustering techniques known as computationally efficient. Besides since we need clustering as an auxiliary procedure we do not need highly accurate clustering results. In this study we use K-means clustering algorithm [6] which became very popular in data mining, unsupervised classification, etc. and which converges quickly to a local optimum.

Our experience says that the aggregated problem with a small number of clusters provides a good resemblance of the original and approximated covariance matrices in terms of first (highest) eigenvalues which is important for the iterative method described below.

In Figure 1 this resemblance is demonstrated where we show distribution of first 10 eigenvalues of both matrices where the data matrix X was obtained by processing 2014 documents of "Cardiovascular Diseases Abstracts" corpus. Matrix X_0 was obtained by K-means method with 10 clusters.
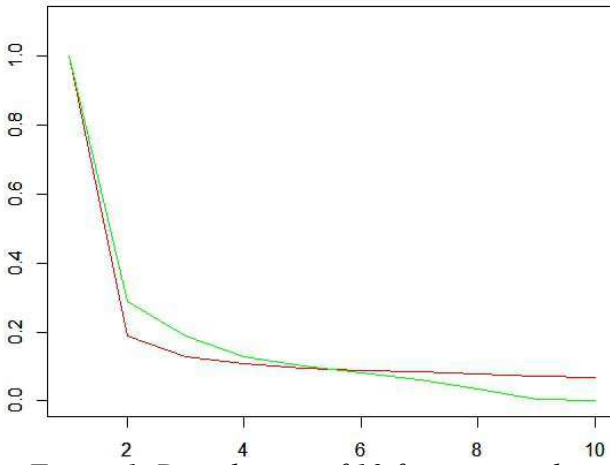

Figure 1. *Distribution of 10 first eigenvalues of the original and approximate covariance matrices for 16058 by 2014 data matrix and 10 clusters.*

4. ITERATIVE METHOD

We use power iteration method [7] for for solving auxiliary aggregated eigenvalue problem and a modified power method for solving the original eigenvalue problem. This method is also known as subspace iteration when used to simultaneously iterate a set of eigenvectors. One iteration of the power algorithm consists of the following steps:

$$for\ i = 1\ to\ l:\quad \widetilde{u}_i^{k+1} = \frac{1}{\left\| Au_i^k \right\|} * Au_i^k$$

$$u_1^{k+1},...,u_l^{k+1} = orthonorm\ (\widetilde{u}_1^{k+1},...,\widetilde{u}_l^{k+1})$$

with approximation of eigenvalues (7)

$$\lambda_i^k = \frac{\left(Au_i^k, u_i^k\right)}{\left(u_i^k, u_i^k\right)}$$

which starts with a set of l initial approximations of first eigenvectors

$$(u^0_1, u^0_2, ..., u^0_l) = U^0.$$

The key property of the power method is that if approximation u^0_i is spanned by matrix A eigenvectors subspace, then after k multiplications of matrix A by this vector the linear combination of eigenvectors will be weighted by λ_i to the power k which gives boost to terms corresponding to highest eigenvalues:

$$A^k u = \sum_k c_i \lambda_i^k \varphi_i \qquad (8)$$

In the method proposed for the first l principal directions of PCA we will need first k orthonormal eigenvectors of A_0 q_1, q_2,..., q_k where $k >= l$. These vectors can be obtained by algorithms (7). We will also need matrix Pi

$$P_i = q_i q^T_i \qquad (9)$$

Since

$$q^T_i q_j = \delta_{i,j}$$

and

$$P_i P_i = P_i,$$

it is a projector to the subspace of i -th eigenvector of A_0. We will modify method (7) using this projector in the following manner:

$$for\ i = 1\ to\ l:\ \widetilde{u}_i^{k+1} = \frac{1}{\left\| Bu_i^k \right\|} * Bu_i^k\ \ where$$

$$Bu^k_i = Au^k_i + \alpha_i * P_i A P_i u^k_i\ \ and$$

$$\alpha_i => min \left\| A\widetilde{u}_i^{k+1} - \lambda_i^k \widetilde{u}_i^{k+1} \right\| \quad (10)$$

$$u_1^{k+1},...,u_l^{k+1} = orthonorm \ (\widetilde{u}_1^{k+1},...,\widetilde{u}_l^{k+1})$$

with approximation of eigenvalues

$$\lambda_i^k = \frac{\left(Au_i^k, u_i^k\right)}{\left(u_i^k, u_i^k\right)}$$

This approach can be thought of as "help" to the power iteration method to converge on the subspace of eigenvectors of the aggregated problem. The intuition for that is similarity of first eigenvectors and eigenvalues of the original and aggregated problem if clustering is done properly. Let

$$u = \sum c_i \varphi_i$$

where φ_i are eigenvectors of the original covariance matrix A and P_k be a projector on subspace of φ_k. Then

$$Au + \alpha P_k A P_k u = \sum_{i \neq k} c_i \lambda_i^k \varphi_i + c_k \lambda_k (1+\alpha) \varphi_k$$

$$(11)$$

If α is chosen big then the second term of this expression dominates over the first term thus providing convergence for φ_k in one iteration step α can be derived from the condition stated in (10):

$$\Phi\left(\widetilde{u}_i^{k+1}\right) = \left\| A\widetilde{u}_i^{k+1} - \lambda_i^k \widetilde{u}_i^{k+1} \right\|$$

$$\alpha => min \ \Phi => \frac{d\Phi}{d\alpha} = 0 \quad (12)$$

This equation leads to the quadratic equation for α. Omitting indexes and skipping details we arrive at the following expression for α

$$\alpha = -\frac{\left(A^2u, AFu\right) - 2\lambda\left(A^2u, Fu\right) + \lambda^2\left(Au, Fu\right)}{\left(Au, AFu\right) - 2\lambda\left(Au, Fu\right) + \lambda^2\left(Fu, Fu\right)}$$

$$(13)$$

where $F = PAP$.

We note that as you can see from (12) α is chosen from the previous step to simplify computations. This can also be justified by the fact that eigenvalues converge faster than eigenvectors. Detailed algorithm discussion is out of scope of this paper. We just mention here that all operations with matrix A are reduced to the matrix vector multiplications of the sparse data matrix X or its transpose X^T.

5. NUMERICAL EXPERIMENTS

For numerical experiments we used two data sets. The fist one is "Cardiovascular Diseases Abstracts" which is a set where each abstract is an individual document. The data matrix X size is 16058 by 2014 where the first value is the total number of terms and the second one is the number of documents. We searched for 10 first eigenvalues of the covariance matrix

$$A = XX^T$$

and used 10 clusters for constructing auxiliary aggregation problem

$$A_0 = X_0 X^T{}_0.$$

So the size of this problem is more than 201 times lower than that for the original problem. The problem is solved using algorithm (10). Figure 2 shows changes of parameter _ for the first three eigenvectors. As expected the biggest contribution of projectors (9) is observed in first iterations to suppress errors caused by initial eigenvector guesses. After some number of iteration contribution of projectors is getting smaller while eigenvectors are getting more accurate.
We measure convergence of eigenvalues through

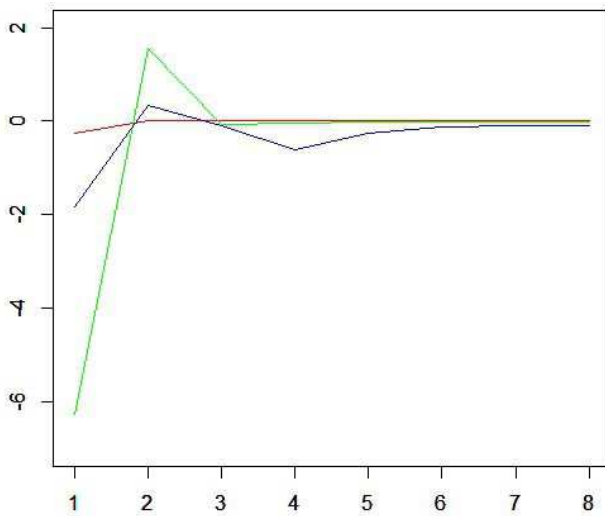$$Error_1 = \frac{\left\| \Lambda^{k+1} - \Lambda^k \right\|_F}{\left\| \Lambda^k \right\|_F}$$

Figure 2. α changes with iteration number. α1 (for eigenvector 1) - red, α2 (for eigenvector 2) – green, α3 (for eigenvector 3) – blue.

Figure 3. Eigenvalues and Eigenvectors convergence for "Cardiovascular Diseases Abstracts" data set. Error1 – red, Error2 – green.

and convergence of eigenvectors by the residual matrix through

$$Error_2 = \left\| AU^k - U^k \Lambda^k \right\|_F$$

where $\left\| \; \right\|_F$ is a matrix Frobenius norm, U^k consists of orthonormal vectors

$$u_1^k, \ldots, u_l^k$$

which are approximations of the eigenvectors and Λ^k is a diagonal matrix of approximations of eigenvalues. Errors graph is demonstrated in Figure 3.

A good convergence rate of the iterative process is demonstrated. After 40 iterations we got $Error_1 = 0.00038$ and $Error_2 = 0.0017$.

The second corpus was "talk politics" set from the news groups. Size of this problem is 13511 (terms) by 1171 (documents). We searched for 10 first eigenvalues of the covariance matrix and used 10 clusters again. The quality of the clustering aggregated model can be viewed by comparing eigenvalues of the original and aggregated covariance matrices. Figure 4 demonstrates a good resemblance of eigenvalues distribution.
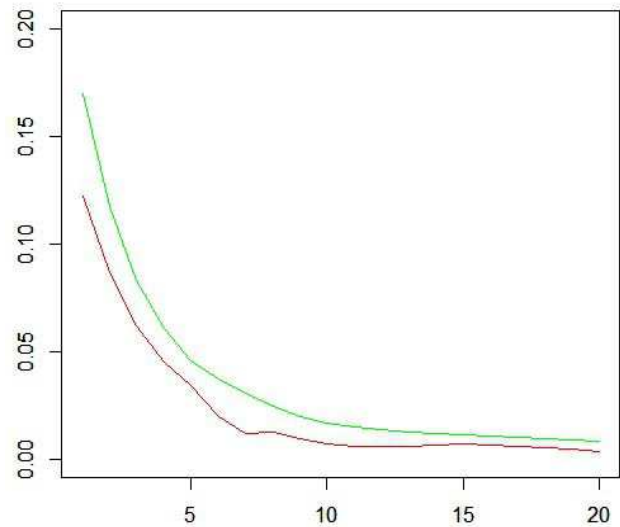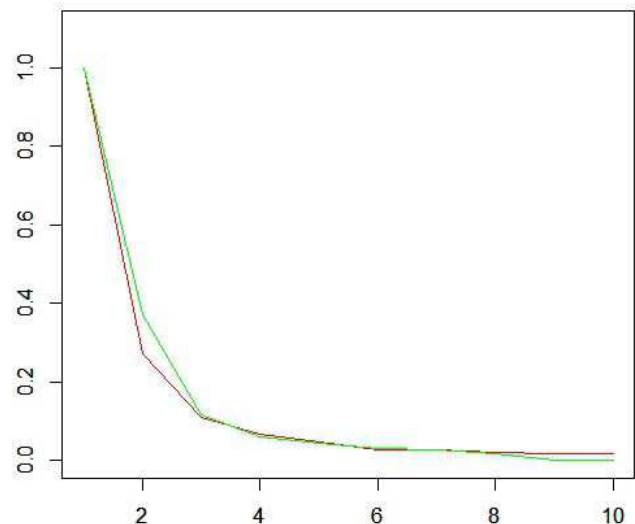

Figure 4. Distribution of 10 first eigenvalues of the original and approximate covariance matrices for 13511 by 1171 data matrix and 10 clusters.

Convergence graph is demonstrated in Figure 5. After 40 iterations we got $Error_1 = 0.00044$ and $Error_2 = 0.00049$.
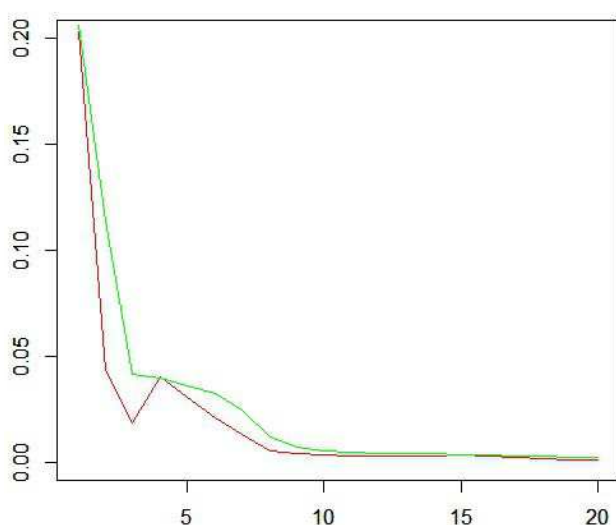
Figure 5. Eigenvalues and Eigenvectors convergence for "News Group" corpus. $Error_1$ – red, $Error_2$ – green.

REFERENCES

1. **Bulgakov V., Kuhn G.** High-performance multi-level iterative aggregation solver for large finite-element structural analysis problems. // Int. j. numer. methods eng., 1995, Vol. 38, pp. 3529-3544.
2. **Hackbush W.** Multi-Grid Methods and Applications. Springer, Berlin, 1985.
3. **Bulgakov V., Belyi M., Mathisen K.** Multilevel aggregation method for solving large-scale generalized eigenvalue problems in structural dynamics. // Int. j. numer. methods eng., 1997, Vol. 40, pp. 453-471.
4. **Jolliffe I.T.** Principal Component Analysis. Series: Springer Series in Statistics, Springer, NY, 2002, XXIX, 487 pages.
5. **Eckart C., Young G.** The approximation of one matrix by another of lower rank. // Psychometrika, Volume 1, 1936
6. **MacQueen J.B.** Some Methods for classification and Analysis of Multivariate Observations. // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281-297.
7. **Rutishauser H.** Simultaneous iteration method for symmetric matrices. // Numer. Math., 1970, Vol. 16, pp. 205-223.

СПИСОК ЛИТЕРАТУРЫ

1. **Bulgakov V., Kuhn G.** High-performance multi-level iterative aggregation solver for large finite-element structural analysis problems. // Int. j. numer. methods eng., 1995, Vol. 38, pp. 3529-3544.
2. **Hackbush W.** Multi-Grid Methods and Applications. Springer, Berlin, 1985.
3. **Bulgakov V., Belyi M., Mathisen K.** Multilevel aggregation method for solving large-scale generalized eigenvalue problems in structural dynamics. // Int. j. numer. methods eng., 1997, Vol. 40, pp. 453-471.
4. **Jolliffe I.T.** Principal Component Analysis. Series: Springer Series in Statistics, Springer, NY, 2002, XXIX, 487 pages.
5. **Eckart C., Young G.** The approximation of one matrix by another of lower rank. // Psychometrika, Volume 1, 1936
6. **MacQueen J.B.** Some Methods for classification and Analysis of Multivariate Observations. // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281-297.
7. **Rutishauser H.** Simultaneous iteration method for symmetric matrices. // Numer. Math., 1970, Vol. 16, pp. 205-223.

Vitaly Bulgakov, Professor, Dr.Sc.; SemanticPro; Boston, USA; phone +1 508-333-7178;
e-mail: contact@semanticpro.com

Булгаков Виталий Евгеньевич, профессор, доктор технических наук; SemanticPro; Boston, USA; phone +1 508-333-7178; e-mail: contact@semanticpro.com