Soft Clustering with Projections: PCA, ICA, and Laplacian

David Gleich and Leonid Zhukov

***Abstract*— In this paper we present a comparison of three projection methods that use the eigenvectors of a matrix to investigate high-dimensional dataset: principal component analysis (PCA), principal component analysis followed by independent component analysis (PCA+ICA), and Laplacian projections. We demonstrate the application of these methods to a sponsored links search listings dataset and provide a comparison of the results both by examining the qualities of the projected dataset and looking at the topics represented by each soft cluster.**

***Index Terms*— PCA, ICA, Laplace Projection, Clustering**

I. Introduction

As computers have become ubiquitous in society, the amount of data available has increased enormously. One of the most useful ways of analyzing a dataset is to describe the clusters in the data, that is, grouping of the data elements which are related by some criteria. More recently, sparse high-dimensional datasets have been used to examine both textual data [1] and Internet page ranking [11].

In this paper, we consider clustering with three different eigenprojection methods: principal component analysis (PCA) [7], principal component analysis with further independent component analysis (PCA+ICA) [9], and Laplacian projections (e.g. spectral clustering) [5]. Each method uses the eigenvectors of a matrix to extract information that facilitates identifying each data point with a similar set of points. We present results from these methods on a several artificial test datasets and an advertiser-bidded keyword dataset from a pay-for-performance search engine listing market.

Spectral methods for data analysis use the eigenvectors of a data matrix. For example, PCA uses the eigenvectors of a covariance matrix to compute a set of important directions (e.g. tendencies) within the data. These directions can then be used to visualize the data [4], reduce the dimension of the dataset, or extract semantic information [6]. However, on certain datasets, PCA projection will miss clustering behavior, e.g. Figure 1(c).

David Gleich is with Harvey Mudd College
Leonid Zhukov is with Yahoo! Research Labs

More recently, Laplacian projections, which use eigenvalues of the Laplacian graph, have been used to approximate NP-complete problems in graph partitioning [5]. When applied to graphs from textual data, the results are clusters of similar terms [2]; on graphs from pixel correlation maps in images, the results are similar segments of the image [12]. Laplacian projections have the property that they separate dissimilar element, which makes them suited for these tasks.

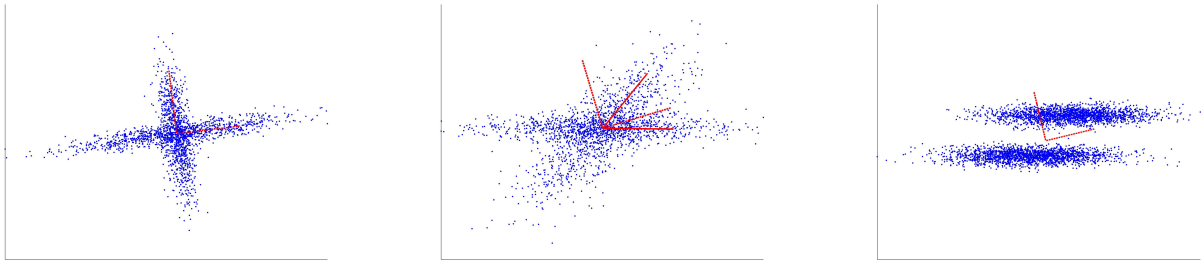
The strength of all spectral methods is that the eigenvectors depend upon the entire dataset, and thus, spectral methods are powerful global analysis techniques. However, different spectral algorithms have their own (and quite different) properties, so they must be used with care. In the remainder of the paper, we first present the details of the datasets we use to examine each of the three clustering methods. Then, we explain the three clustering algorithms as applied to our high-dimensional dataset. We finish with a discussion of our observations from each method and a summary of our conclusions.

II. Data

In this study, we use series of artificial datasets to demonstrate the differences between the three methods as well as a small, densely connected subset of Overture's United States advertiser-term data with 9,352 advertisers, 14,385 bidded search terms, and more than 250,000 bids.

The three artificial data matrices are in two dimensions and have two general "clusters" of data. Each cluster is generated by a normal distribution of 1000 points and then scaled or translated to separate the clusters. The first dataset has two clusters aligned with the horizontal and vertical axes. The second dataset has one cluster aligned with the horizontal axis and another cluster slightly shifted off the vertical axis. The third dataset has two horizontal clusters translated in the vertical direction.

Our representation for bidding data is a advertiser-term matrix, A, whose columns correspond to bidded search



(a) A good PCA dataset. (b) A good ICA dataset. (c) A good Laplacian projection dataset.

Fig. 1. Examples of PCA, PCA+ICA, and Laplacian projections on a set of artificial data designed to highlight the differences between the methods. In the first figure, PCA correctly identifies the orthogonal axes. In the second figure, PCA cannot correctly identify the axes because they are not orthogonal; however, when ICA is applied after PCA, the ICA algorithm can determine the correct axes of the data. In the final figure, applying PCA causes the clusters to collapse onto the horizontal axis. Instead, the first Laplacian projection separates the two clusters.

keywords and rows to advertisers. [1] Each non-zero entry in the matrix has the value of the advertiser's bid on that particular term. Thus, every row of the matrix shows a bidding pattern for an advertiser. Any bidded term, i.e column of the matrix, t_i can be considered as a vector (data point) in the advertiser space. This matrix is strictly non-negative and is also sparse, since the majority of the advertisers bid only on a small number of terms.

Interestingly, the data has a power law distribution in both the number of bids on each term, and the total value of bids on that term. Figure 5 confirms these observations.

III. Method

In this section, we present the three algorithms we used to cluster our high-dimensional dataset.

A. PCA

Principal component analysis uses the eigen-decomposition of the correlation matrix $M = AA^T$ to find orthogonal directions with total maximal variance of projections, $MU = \Lambda U$. PCA sorts the axis of the reduced dimensionality basis according to the total variance of the projection and retains the k largest axes, thus removing redundancy and reducing the dimensionality of the data. PCA can also be calculated using the singular value decomposition of A [8]. The columns of U form an orthogonal basis in term space, and the columns of V form an orthogonal basis in advertiser space. Since we are interested in term space, we will perform projection of a transposed A^T matrix onto V basis by $V_k^T A^T$.

$$X_k = S_k^{-1} V_k^T A^T = U_k^T$$

The covariance matrix $X_k X_k^T \approx I$ is diagonal, that is, the projection axes are uncorrelated with equal variance.[2] The matrix X_k is also an optimal reduced dimensional representation of the term vectors from A, in the least-squared distance sense.

B. PCA+ ICA

Independent component analysis finds a set of directions in the data such that when the data points are projected onto these directions, the resulting data are *statistically independent* (a much stronger condition that uncorrelated). Unlike PCA, these directions need not be orthogonal within the original space.

We start with PCA "sphered" results from the previous section $X_k X_k^T = I$. We then employ FastICA [10] algorithm to reconstruct a matrix W such that,

$$Y_k = W X_k,$$

where the rows of Y are statistically independent, and X_k is the reduced dimensional representation of the terms from PCA. Thus, after the PCA "sphering" procedure, an ICA algorithms only needs to adjust the axes.

C. Laplace projections

Laplace projections are defined by analogy with graph partitioning techniques [2], [5]. We define a diagonal

[1] Our definition of A is transpose of that typically used in latent semantic indexing literature [6], where columns correspond to documents and words (terms) to rows. Latent semantic indexing is the term for PCA applied to histograms of textual data. We choose our form since we are interested in terms behavior and not advertisers. Both representations lead to the same results.

[2] Due to the sparsity and high dimensionality of the data the row mean is already close to zero, and we chose not remove the mean from the data.

degree matrix $D_{ii} = \sum_j M_{ij}$ [3], where as before $M = AA^T$ is the correlation matrix. Then, the corresponding Laplacian matrix is given by $L = D - M$. The coordinates of data points in the reduced representation can be found as solution (eigenvectors) of the following generalized eigensystem,

$$(D - M)x = \lambda Dx.$$

This eigensystem corresponds to normalized cuts criterion in graph partitioning [12]. Due to the algebraic properties of the above system, $\lambda_1 = 0$ and x_1 is constant. We use the remaining eigenvectors, $x_2, x_3, \ldots$ as projection axes.

IV. Results

We first present the results on a series of artificial datasets designed to highlight the differences between the two methods. We then examine the three methods applied to the advertiser-term dataset.

On the artificial data, the differences between the methods are extreme. Figure 1 demonstrates how each method behaves on artificial archetypal datasets. From these datasets, we see how each method, in theory, works as a clustering tool. PCA finds the two orthogonal clusters; PCA+ICA finds the two non-orthogonal clusters; and Laplacian projections finds the two separated clusters.

When we apply these three methods to our advertiser-term dataset, we see similar phenomena. Thus, as seen from the forthcoming figures, the advertisers tends to exhibit grouping behavior in their bidding patterns, i.e form clusters corresponding to sub-markets. When we examine the data with PCA, we see these groupings along the axes, Figure 2. But at the same time PCA enforces orthogonality of the clusters, while the sub-markets formed by advertisers are not necessarily uncorrelated. ICA relaxes that restriction and allows the axes to adjust to follow the clusters, Figure 3. Then, in ICA space, points are nicely aligned along the axis and have strong projection only on one of the axis.

When we apply Laplacian projections to the data, we see that the data points group into four to five distinct clusters, Figure 4. By segmenting the dataset based on these values, we then achieve a hard clustering of the data, that is, each datum is in one and only one cluster. For PCA and PCA+ICA, if we view each axis as a cluster, then a data point may belong to many clusters, i.e. has a strong projection on many axes. In this sense, PCA and PCA+ICA is a soft clustering of the data.

[3]For unweighted graphs, this matrix would have node degrees on the diagonal.

We can further examine the properties of the soft clustering induced by PCA and PCA+ICA by attempting to label the axes in the manner of Booker et. al. [4]. The topic of the axis, or cluster, is identified using the terms with maximal projection on that axis. Table I and II shows the identity of the top three clusters for both PCA and PCA+ICA. We can use these terms to show the real-world benefit of PCA+ICA, compared with PCA. In Table III we show the terms for similar clusters from PCA and PCA+ICA. The PCA cluster appears to mix two separate clusters; whereas the PCA+ICA cluster is much cleaner. This observation reinforces T

V. Summary and Conclusions

We analyzed the properties of three spectral projection clustering algorithms on data from a pay-for-performance advertising market. The direct spectral methods are powerful techniques to both reduce the dimensionality and extract the structure from a dataset. However, the orthogonality constraint imposed with PCA is too restrictive to permit an accurate solution of the data. The PCA+ICA algorithm uses PCA to extract the important structure and then ICA relaxes the orthogonality constraint to better align with the clusters on each axis. Laplacian projections work by finding a projection of the data that separates the classes as much as possible.

We conclude that PCA+ICA is the best method to generate soft clusters from individual data points and that Laplacian projections are superior when there are unambiguous clusters within the data.

TABLE III

A COMPARISON OF THE PCA AND ICA CLUSTERS.

PCA Axis 31/50	ICA Axis 28/50
book sport	book internet sport
book internet sport	baseball betting
card credit offer	gambling sport
apr card credit low	betting online sport
best card credit	football wagering
baseball betting	

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[3] M. W. Berry and M. Browne. *Understanding search engines: mathematical modeling and text retrieval*. Society for Industrial and Applied Mathematics, 1999.

[4] A. Booker, M. Condliff, M. Greaves, F. B. Holt, A. Kao, D. J. Pierce, S. Poteet, and Y.-J. J. Wu. Visualizing text data sets. *Computing Science and Engineering*, 1(4):26–35, July/ Aug. 1999.

[5] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics*, 92, 1997.

[6] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

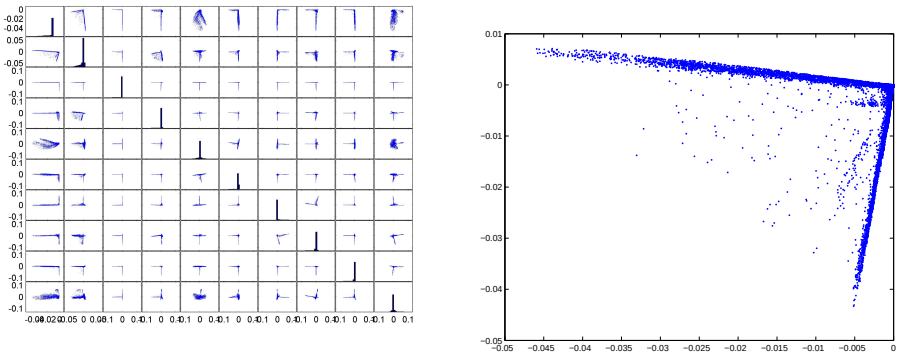
[8] G. H. Golub and C. F. V. Loan. *Matrix Computations*. John Hopkins Univ. Press, 1989.

[9] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[10] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

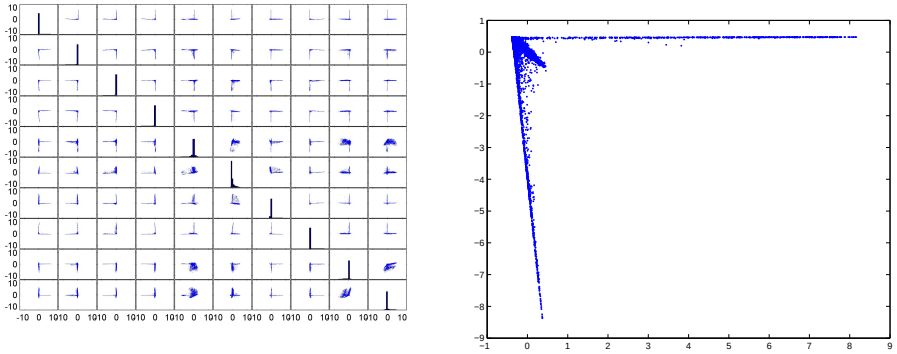
[11] J. Kleinberg and A. Tomkins. Applications of linear algebra in information retrieval and hypertext analysis. In *Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 185–193. ACM Press, 1999.

[12] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.



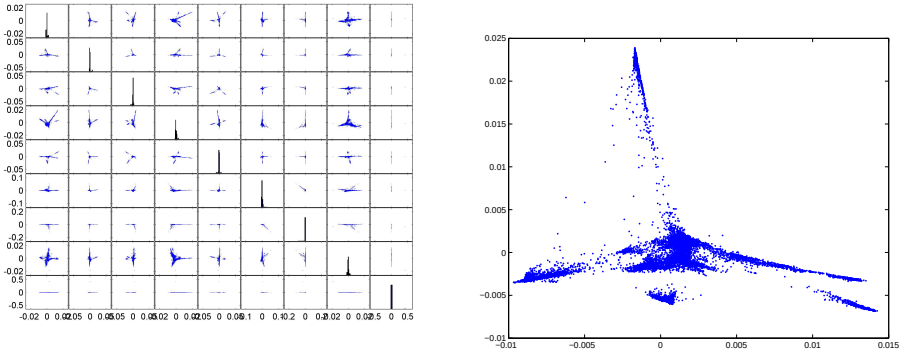
(a) Top 10 axes. (b) Top 2 axis.

Fig. 2. PCA projections. In the first figure, we plot each data point by its projection on each of the top 10 PCA axes. The second figure is an enlarged version of the dataset projected onto the top 2 PCA axes. Since some of the axes are rotated, PCA was not able to find a completely orthogonal transformation of the dataset.



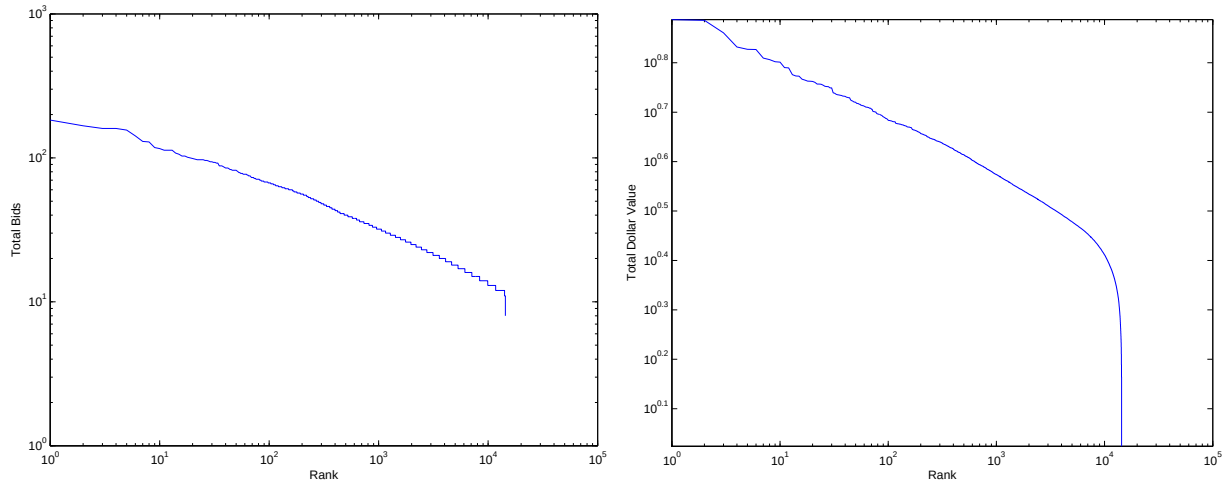
(a) Top 10 axis. (b) Top 2 axis.

Fig. 3. PCA+ICA projections. In the first figure, we plot each data point by its projection on each of the top 10 ICA axes. The second figure is an enlarged version of the dataset projected onto the top 2 ICA axes. Most of the data points have a strong projection on only one axis.



(a) Top 10 axis. (b) Top 2 axis.

Fig. 4. Laplace projections. In the first figure, we plot each data point by its projection on each of the top 10 Laplace projection axes. The second figure is an enlarged version of the dataset projected onto the top 2 Laplace projection axes. These projections emphasize clustering.



(a) Number of bids (b) Net term value

Fig. 5. Two different power law relationships in our dataset. The left figure shows the rank-ordered plot of total term bids, i.e. the number of advertisers bidding on the term. The right figure shows the rank-ordered plot (different ordering) of the total monetary value of each term, i.e. the sum of all advertiser bids on that term. In both figures, the axes are logarithmic.

TABLE I

TERMS WITH MAXIMUM PROJECTION ON THE TOP 3 PCA AXES.

PCA Axis 1/50	PCA Axis 2/50	PCA Axis 3/50
austin hotel	best hosting service web	pharmacy phentermine
beach hotel myrtle	company hosting web	cheap phentermine
albuquerque hotel	affordable cheap hosting web	online phentermine
hotel reno	domain transfer	diet phentermine bill
atlanta hotel	cheap domain hosting	buy online phentermine

TABLE II

TERMS WITH MAXIMUM PROJECTION ON THE TOP 3 ICA AXES.

ICA Axis 1/50	ICA Axis 2/50	ICA Axis 3/50
car insurance	business ecommerce internet solution	diego lodging san
car insurance quote	ecommerce page solution web	cheap diego hotel san
auto free insurance quote	development ecommerce page web	anaheim discount hotel
auto insurance	ecommerce provider solution	francisco lodging san
automobile insurance	ecommerce host	coronado hotel