Sparse PCA for Text Corpus Summarization and Exploration

Brian Gawalt, Youwei Zhang, and Laurent El Ghaoui
Department of EECS
University of California, Berkeley
Berkeley, CA 94720
{gawalt, zyw, elghaoui}@eecs.berkeley.edu

Abstract

Low-rank matrix approximation can be used not just for greater computational efficiency or robustness, but also increasing data interpretability. We propose using sparse principal component analysis (PCA) for summarizing large corpora of text documents. When made substantially sparse, i.e. with cardinalities of no more than ten features, the principal components of a Salton matrix can define a set of key tokens which describe the collection of documents. Additionally, the user can quickly find the documents most exemplary of the corpus by selecting those whose vector representations are best approximated by projection into the range of these sparse principal components. The ability to safely eliminate many features from consideration when performing sparse PCA, we hypothesize, should greatly decrease computational costs for fast and iterative exploration. We propose comparing the performance of sparse PCA on these tasks to the more traditional latent semantic indexing or latent Dirichlet allocation approaches.

1 Corpus summarization

The volume of content is ever increasing. When an investigator faces a large body of documents, two useful objectives appear:

1. It would be nice to select a small number of documents which typify the corpus, documents which are the least "outlying" from the corpus overall.
2. The corpus's many documents are themselves written in a vocabulary of many words and multi-word phrases. It would be nice to select a small number these word (or phrase) vocabulary tokens which explain much of the diversity in the corpus.

Low-rank matrix approximation can help achieve these two objectives. Specifically, we propose that sparse principal component analysis (PCA) provides a complete framework for accomplishing both tasks in a way that keeps the investigator fully cognizant of the meaning of the results. Many familiar metrics, e.g., fraction of total variance explained, may suffer given that we are bound by these goals to produce exceedingly simple models. We hypothesize that this sparse formulation presents a good candidate for these objectives and conditions.

The Salton matrix A corpus can be represented as a matrix $X = \{x_{ij}\}$, with each row $i = 1, \ldots, m$ corresponding to a particular document, each column $j = 1, \ldots, n$ corresponding to a particular text token, and each element encoding patterns of how each token appears in each document, resulting in a large, sparse matrix. This encoding of value x_{ij} is open to a variety of methods. One could use a count of the number of times token j appeared in document i, or a binary

matrix where x_{ij} is a 0/1 indicator of which tokens appeared in which documents. The TF-iDF approach [1] is a popular representation for many text processing applications, and the effects of it and of several other candidate representations are tested in [2]. It's important to remember that choice of this encoding can greatly impact our results. This paper refers to this matrix (under any encoding) as the "Salton matrix", and it is with low-rank approximations of this matrix that we hope to achieve our corpus summarization and exploration goals from above.

Latent semantic indexing and its descendents Latent semantic indexing (LSI) [3] is an application of linear dimensionality reduction via principal component analysis to the Salton matrix. By projecting the document vectors onto the first k principal components, we build a "reduced matrix." The columns of this reduced matrix now correspond, not to particular text tokens, but to specific linear combinations of tokens.

This technique is a commendable addition to many text processing algorithms, but given our goals, there is an acute drawback. The components produced are too dense to function as a keyword list or ranking. For example, the model parameters are each sensitive to all the tokens, and interpreting them in light of this is frustrating. Though LSI has inspired more improved models such as probabilistic LSI [4], the latent dirichlet allocation (LDA) [5], etc., there remains an implicit constraint being violated. The number of tokens used by any model attempting to achieve this paper's goals need to be few enough in number – perhaps dozens, at the most – for an investigator to conceive of them all together at once.

Sparse PCA Principal component analysis can be reformulated to find a low-rank approximation of a Salton matrix (by finding directions of greatest variance) alongside a penalty for non-zero model parameter values. Finding the solution to this regularized objective can be posed as a semi-definite program [6]. This sparsity regularization will necessarily bias the solution away from the best possible approximation, but to the benefit of a human-interpretable model. We can increase the regularization parameter until the model found has sufficiently few tokens in our components. When comparing how much variance is explained by selecting a few tokens in this fashion versus, e.g., how much variance is explained if one uses only the top few components derived by normal PCA, the benefits of sparse PCA become clearer.

Finding this approximation can be greatly sped up by using safe feature elimination. It can be shown that any token feature with sample variance less than the regularization parameter must be zero in the final solution, and so can be dropped from the Salton matrix when calculating the sparse principal components. This holds for many cases of l_1 regularization, as seen in [7]. The number of features so dropped is a conservative lower bound on the total number of zero-weight features in the final solution. While dropping low-variance features is a common first step in ordinary PCA, this necessarily means the final solution reached must be an approximation; that is not the case here. Dropping low variance features in sparse PCA has no downside; an exact final solution can be calculated much faster than in the dense case.

2 Data Exploration

The keywords included in the sparse principal components are interesting and informative, but the model also presents more exploratory opportunities:

Visualizing the corpus When two or three principal components are used in the approximation, the document data points can visualized in a plot. Because of the few tokens involved in producing the principal components, the user has an immediate and complete grasp of the word use distinguishing documents from different areas of the plot. The same visualization for dense components introduces more ambiguity: is a high value along a component due to the use of a few strongly-weighted tokens, or the accumulation of many weakly-weighted ones? This is an easier question to answer when it's known the number of tokens utilized is small.

Well-approximated documents The span of the k sparse principal components of the Salton matrix form a subspace of the token-space defining the original document data points. A residual can be found representing the difference between the original data point and its projection into the sparse-PC span. It is then easy to identify and highlight documents with small-magnitude residuals.

PCA has been used to implement anomaly detection by identifying points which are far from this subspace [8]; our hypothesis merely takes the corrolary that points close to the subspace should be broadly representative.

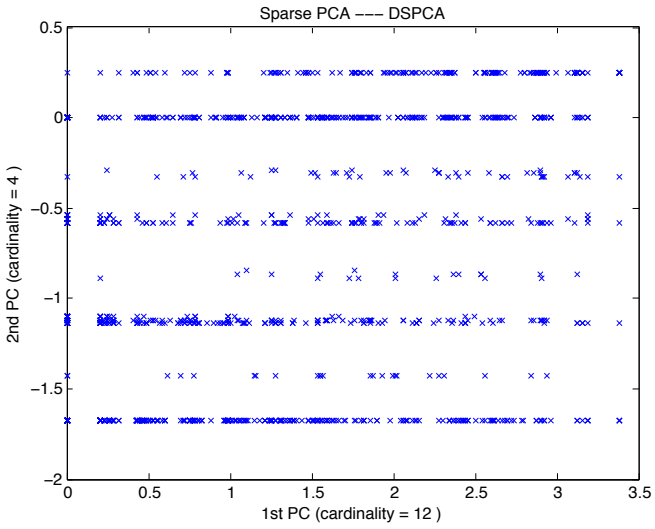
3 Demonstration

As a demonstration, we ran this procedure over the collection of the 1,288 news articles published by the *New York Times*'s International section mentioning the word "China." We tokenized by unigrams, removed no stop words, and performed no stemming. The Salton matrix encodes the binary 1/0 appearance/non-appearance of tokens. Table 1 contains the tokens and respective weights used in the two dominant principal component. The first component implies much variance occurs in terms of China's international standing, especially vis a vis the United States and the United Nations; the second component of keywords have a domestic component, countered by another international agent (Russia). Table 2 shows the headlines for articles closest to their sparse-PCA projections, along with the magnitude of this residual. It is interesting that all these articles are of short length.

Table 1: Sparse principal components for articles mentioning "China"

1st comp. token	Weight	2nd comp. token	Weight
states	0.3929	chinese	-0.5788
united	0.3913	beijing	-0.5578
american	0.3195	chinas	-0.5394
obama	0.3155	russia	0.2507
president	0.2858		
washington	0.2792		
countries	0.2633		
nations	0.2618		
administration	0.2457		
international	0.2271		
would	0.2058		
nuclear	0.1976		

Figure 1: Corpus as projected onto its two sparse principal components



4 Validation

Much prior work in [9], [10], etc., suggests using human survey respondents to evaluate how well an approach such as ours works. Respondents could be instructed to read many dozen example

Table 2: Best approximated articles

Res. Mag.	Article Headline
66.593085	"World Briefing — Europe: Vatican: New Language for Web Site"
75.671066	"Aid Sought for Students"
78.128291	"World Briefing — Asia: Taiwan: Agreement With China Opens a Rare Diplomatic Door"
79.107418	"World Briefing — Asia: China: Use of Controversial Software to Filter Web Is Optional, Official Says"
79.108141	"World Briefing — Asia: China: Political Site Is Shut Down"
80.052708	"World Briefing — Europe: Global Arms Spending Up, Study Shows"
80.132263	"World Briefing — Asia: China: Court Upholds Sentences Stemming From Riots"
80.532884	"World Briefing — Asia: China: Ceiling Collapse Kills at Least 11 Workers"
80.723861	"World Briefing — Asia: China: Border to Korea Reopens"
81.543736	"World Briefing — Asia: India: Dalai Lama on Hacking"

documents (drawn at random), then solicited for scores of a) how reasonable they find the word lists generated by looking at the sparse principal component features and b) how representative they feel the well-approximated documents are, given their reading. We then repeat for word lists and exemplary documents generated by another method, whether thresholded LSI or LDA.[1]

It will be interesting to compare the results of this survey to other, more quantitative metrics. We naturally expect a rather poor approximation (using only a few tokens in our projection is going to mean only a low percentage of the overall data variance is explained), but it would be intriguing to see what relationships obtain between the survey respondent scores and the effectiveness of the sparse approximation error.

References

[1] G. Salton and M. McGill (editors). *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[2] Preslav Nakov, Antonia Popova, and Plamen Mateev. Weight functions impact on LSA performance. In *EuroConference RANLP'2001 (Recent Advances in NLP*, pages 187–193, 2001.

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

[4] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation of sparse PCA using semidefinite programming. *SIAM Review*, 49(3), 2007.

[7] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. Technical Report UC/EECS-2010-126, EECS Dept., University of California at Berkeley, September 2010.

[8] Anukool Lakhina, Mark Crovella, and Christophe Diot. Diagnosing network-wide traffic anomalies. In *SIGCOMM '04: Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 219–230, New York, NY, USA, 2004. ACM.

[9] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632, New York, NY, USA, 2010. ACM.

[10] J. Chang, J. Boyd-Graber, C. Wang S. Gerrish, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 2009.

[1] It is of course important to get validation as well for all these approximation methods run on different Salton representations, as mentioned in the above section.