Supplementary Material: Fast and Accurate Pseudoinverse for Real-world Sparse Matrices

Anonymous
Anonymous
Anonymous

Abstract

This document is the supplementary material for "Fast and Accurate Pseudoinverse for Real-world Sparse Matrices".

Table 1: Table of symbols.

Symbol	Definition
m	number of training instances
n	number of features
L	number of labels
$\mathbf{A} \in \mathbb{R}^{m \times n}$	input feature matrix
$\mathbf{A}^\dagger \in \mathbb{R}^{n \times m}$	pseudoinverse of the feature matrix
$\mathbf{U}_{m \times r}$	$m \times r$ left singular vectors (matrix)
$\mathbf{\Sigma}_{r \times r}$	$r \times r$ diagonal singular values matrix
$\mathbf{V}^\top_{r \times n}$	$r \times n$ right singular vectors (matrix)
α	target rank ratio in Algorithm 1 where $0 < \alpha \leq 1$
r	target rank, i.e., $r = \lceil \alpha n \rceil$ for $m \times n$ matrix when $m > n$
$\mathbf{A}_{ij} \in \mathbb{R}^{m_i \times n_j}$	(i,j)-th submatrix of reordered $\mathbf{A}$
k	hub selection ratio in Algorithm 2 where $0 < k < 1$
m_1 & n_1	number of spoke instance and feature nodes, respectively
m_2 & n_2	number of hub instance and feature nodes, respectively
B	number of rectangular blocks in $\mathbf{A}_{11}$
m_{1i} & n_{1i}	height and width of i-th block in $\mathbf{A}_{11}$, respectively
$\lvert\mathbf{A}\rvert$	number of non-zero entries in $\mathbf{A}$

1 Detailed Description of Real-world Multi-label datasets

The Bibtex dataset is from a social bookmarking system, where each instance consists of features from a bibtex item and labels are tags in the system [Katakis *et al.*, 2008]. The Eurlex dataset is from documents about European Union law, where each instance is formed by word features from a document and labels indicate categories [Mencia and Fürnkranz, 2008]. The RCV dataset is randomly sampled from an archive of newswire stories made available by Reuters, Ltd., where each instance consists of features from a document, and labels are categories [Lewis *et al.*, 2004]. The Amazon dataset is randomly sampled from a set of reviews of Amazon, where each instance is formed by word features of a review and labels are items [McAuley and Leskovec, 2013].

2 Skewed Degree Distributions in Real-world Feature Matrices

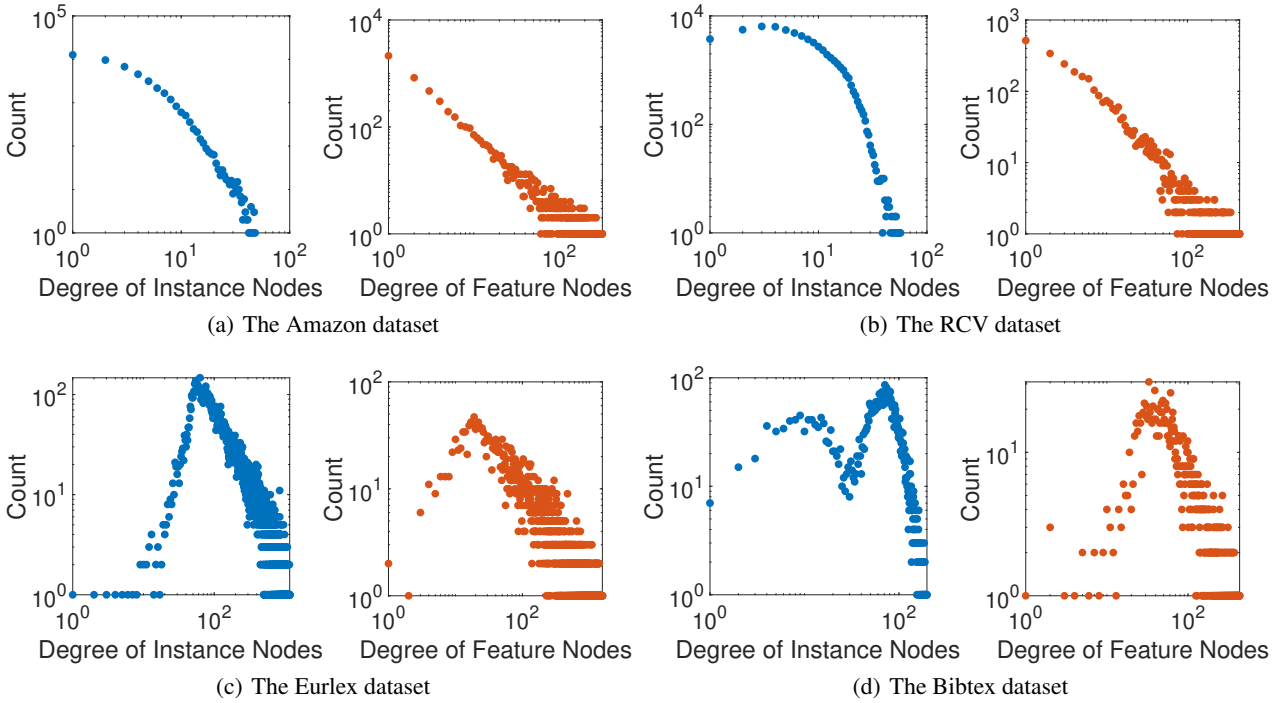
We report the degree distribution of the datasets used in the submitted paper. Figure 1 demonstrates the degree distributions of instance and feature nodes in a bipartite network derived from a real-world feature matrix. The figure shows that the degree distributions are skewed, i.e., there are few high degree nodes while a majority of nodes have low degrees. This skewness is remarkable on the Amazon and RCV datasets. Although the Eurlex and Bibtex datasets exhibit a different shape of skewness, their node degrees are not uniform and are still skewed toward a relatively low degree. Thus, many real-world feature matrices involve the skewness that a majority of data instances have few features while a few instances have many features.

3 Reordered Feature Matrices by FastPI

We report the matrix reordering results of FastPI on the datasets used in the submitted paper. Figure 2 shows that the original and reordered feature matrices. Note that after the matrix reordering, non-zero entries are concentrated at the bottom right. The empty area becomes more extensive as node degree skewness gets higher, such as in the Amazon and RCV datasets. In the submitted paper, we show that FastPI excels at fast pseudoinverse computation in such highly skewed data. On the other hand, the empty areas of the reordered matrices on the Eurlex and Bibtex datasets are not as large as those on the other datasets. Nevertheless, even such reordered structures as in Figures 2(c) and 2(d) also provide a meaningful performance improvement as presented in the submitted paper.

4 Proof for Computational Complexity of FastPI in Lemma 1

Proof. We summarize the complexity of each step of Algorithm 1 in Table 2. For this proof, we use the traditional complexity of the low-rank approximation as describe in [Gu and



(a) The Amazon dataset (b) The RCV dataset

(c) The Eurlex dataset (d) The Bibtex dataset

Figure 1: (Supplementary Section 2) Skewed degree distributions of instance and feature nodes in a bipartite network derived from a real-world feature matrix.

Table 2: Computational complexity of each step of FASTPI (Algorithm 1).

Line	Task	Computational Complexity
1	Reorder $\mathbf{A}$ using Algorithm 2	$O(T(m\log(m)+\|\mathbf{A}\|))$
2	Compute SVD of $\mathbf{A}_{11}$	$O(\sum_{i=1}^{B} m_{1i}n_{1i}s_i)$
2	Incremental Update the SVD result for $\mathbf{A}_{21}$	$O(m_1r^2+n_1r^2+m_2n_1r)$
3	Incremental Update the SVD result for $\mathbf{T}=[\mathbf{A}_{12};\mathbf{A}_{22}]$	$O(n_1r^2+mr^2+mn_2r)$
Total	$O(mr^2+n_1r^2+mn_2r+m_2n_1r+(\sum_{i=1}^{B} m_{1i}n_{1i}s_i)+T(m\log(m)+\|\mathbf{A}\|))$	

Eisenstat, 1996; Halko *et al.*, 2011]; for matrix $\mathbf{A}\in\mathbb{R}^{p\times q}$, the low-rank approximation takes $O(pqk)$ time with target rank k. For a detailed comparison, we omit the cost of the final pseudoinverse construction (line 5 in Algorithm 1) because all SVD based methods should perform the construction as a common step. The complexity of each step of the algorithm is proved as follows:

- Line 1: for each iteration, FASTPI sorts the degrees of instance and feature nodes; thus, it requires up to $O(m\log(m))$ since $m>n$. Then, it searches connected components in G' using breadth first search (BFS) algorithm in $O(|\mathbf{A}|)$ indicating the number of edges in the network. Hence, each iteration demands $O(m\log(m)+|\mathbf{A}|)$ time.
- Line 2: FASTPI computes the low-rank approximated SVD of each rectangular block in $O(m_{1i}n_{1i}s_i)$ with target rank $s_i=\lceil\alpha n_{1i}\rceil$; thus, it is $O(\sum_{i=1}^{B} m_{1i}n_{1i}s_i)$.
- Line 3: in equation (2), the low-rank approximation takes $O((m_2+s)n_1s)=O(m_2n_1s+n_1s^2)$ time. The matrix multiplication for $\mathbf{U}_{m\times s}$ takes $O(m_1s^2)$ as follows:

$$\mathbf{U}_{m\times s}=\begin{bmatrix}\mathbf{U}_{m_1\times s}\\ \mathbf{O}_{m_2\times s}\end{bmatrix}\quad \tilde{\mathbf{U}}_{(s+m_2)\times s}=\begin{bmatrix}\mathbf{U}_{m_1\times s}\tilde{\mathbf{U}}_{s\times s}\\ \tilde{\mathbf{U}}_{m_2\times s}\end{bmatrix}$$

where $\tilde{\mathbf{U}}_{(s+m_2)\times s}=\begin{bmatrix}\tilde{\mathbf{U}}_{s\times s}\\ \tilde{\mathbf{U}}_{m_2\times s}\end{bmatrix}$. Since $s\le r$, it is bounded by $O(m_1r^2+n_1r^2+m_2n_1r)$ where $s=\lceil\alpha n_1\rceil$ and $r=\lceil\alpha n\rceil$ and $n_1\le n$.
- Line 4: in equation (3), the low-rank approximation takes $O(m(n_2+s)r)=O(mn_2r+msr)$, and the matrix multiplication takes $O(n_1s^2)$ as follows:

$$\mathbf{V}_{r\times n}^{\top}=\tilde{\mathbf{V}}_{r\times(s+n_2)}^{\top}\begin{bmatrix}\mathbf{V}_{s\times n_1}^{\top} & \mathbf{O}_{s\times n_2}\\ \mathbf{O}_{n_2\times n_1} & \mathbf{I}_{n_2\times n_2}\end{bmatrix}$$
$$=[\tilde{\mathbf{V}}_{r\times s}^{\top}\mathbf{V}_{s\times n_1}^{\top}\quad \tilde{\mathbf{V}}_{r\times n_2,}^{\top}]$$

where $\tilde{\mathbf{V}}_{r\times(s+n_2)}^{\top}=[\tilde{\mathbf{V}}_{r\times s}^{\top}\ \tilde{\mathbf{V}}_{r\times n_2}^{\top}]$. Hence, it is $O(n_1r^2+mr^2+mn_2r)$ since $s\le r$.

□

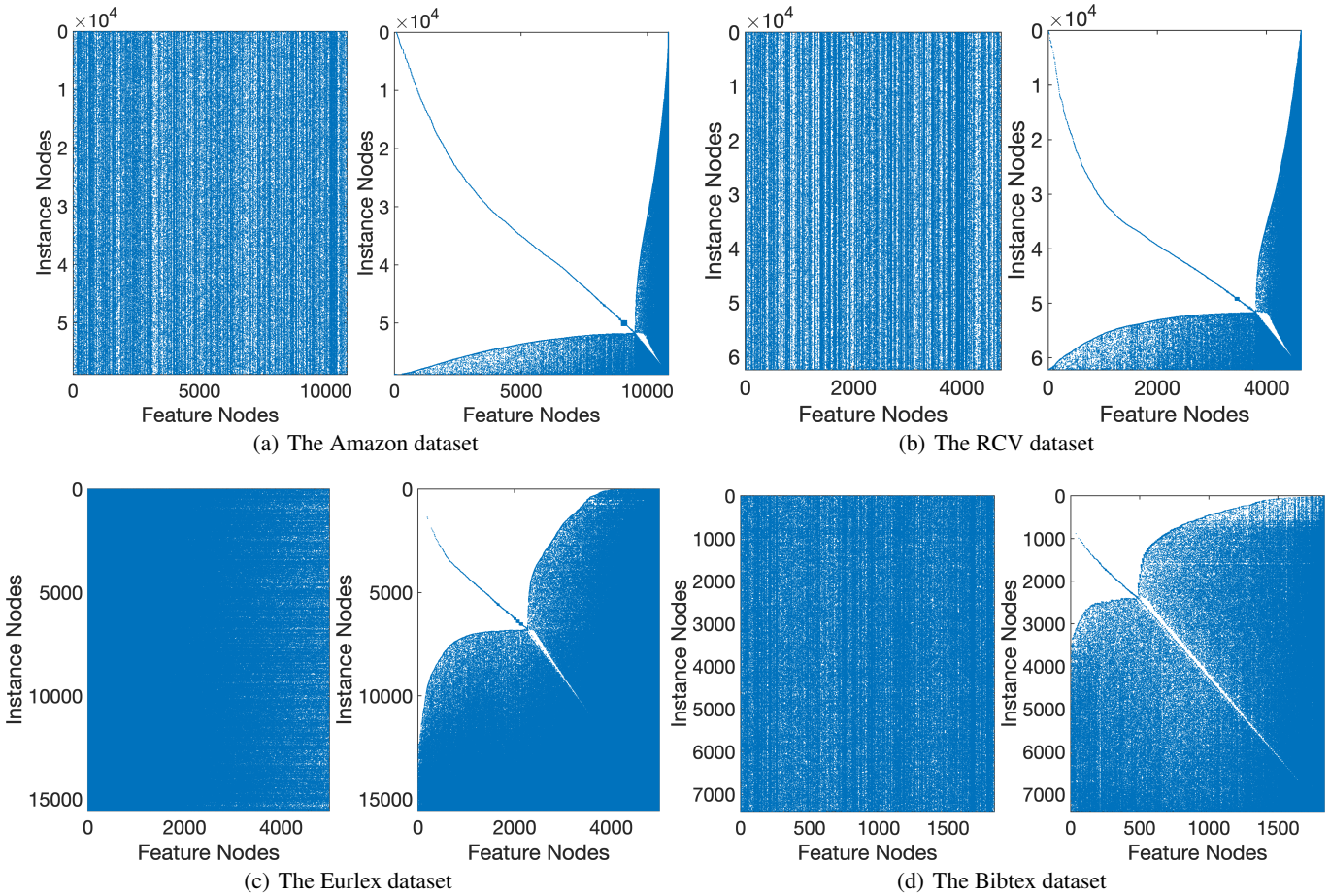


Figure 2: (Supplementary Section 3) Real-world feature matrices before and after the matrix reordering of FASTPI (Algorithm 2). For each subfigure, the left figure is the spy plot for the original matrix, and the right one is for the reordered matrix.

5 Accuracy Evaluation

In this section, we further evaluate the quality of the pseudoinverse computed by each method using NDCG@k, which is another representative metric for the multi-label linear regression task [Chen and Lin, 2012; Prabhu and Varma, 2014; Yu *et al.*, 2014]. The formal definition of NDCG@k is as follow:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\sum_{l=1}^{k} \frac{1}{\log(l+1)}} \text{ where}$$

$$\text{DCG@}k = \sum_{l \in \text{rank}_k(\hat{\mathbf{y}})} \frac{\mathbf{y}_l}{\log(l+1)}.$$

$\mathbf{y} \in \{0,1\}^L$ is a ground truth label vector, $\hat{\mathbf{y}} \in \mathbb{R}^L$ is its predictive score vector where L is the number of labels, and $\text{rank}_k(\hat{\mathbf{y}})$ returns the k largest indices of $\hat{\mathbf{y}}$ ranked in the descending order. As NDCG is higher, the predictive performance of a method is better. Figure 3 shows NDCG@3 for each dataset and method varying target rank ratio α from 0.01 to 0.1. As shown in the figure, the accuracies of all tested methods are almost the same overall datasets, similarly to P@3 in the submitted paper. This result also supports that FASTPI accurately calculates the pseudoinverse with similar accuracy as those of the competitors.

References

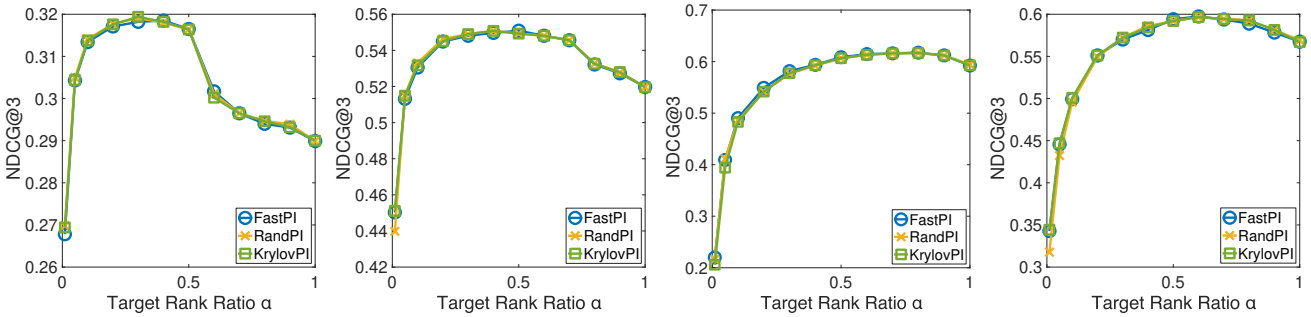
[Chen and Lin, 2012] Yao-Nan Chen and Hsuan-Tien Lin. Feature-aware label space dimension reduction for multi-label classification. In *Advances in Neural Information Processing Systems*, pages 1529–1537, 2012.

[Gu and Eisenstat, 1996] Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[Halko *et al.*, 2011] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[Katakis *et al.*, 2008] Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. In *Proceedings of the ECML/PKDD*, volume 18, 2008.

[Lewis *et al.*, 2004] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.



(a) The Amazon dataset (b) The RCV dataset (c) The Eurlex dataset (d) The Bibtex dataset

Figure 3: (Supplementary Section 5) Accuracy of the multi-label linear regression task (Application 1) in terms of NDCG@3 varying target rank ratio α from 0.01 to 1.0.

[McAuley and Leskovec, 2013] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[Mencia and Fürnkranz, 2008] Eneldo Loza Mencia and Johannes Fürnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer, 2008.

[Prabhu and Varma, 2014] Yashoteja Prabhu and Manik Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–272. ACM, 2014.

[Yu *et al.*, 2014] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International conference on machine learning*, pages 593–601, 2014.