Perturbation bound on Sparse PCA

Daniel Ringwalt

March 14, 2023

1 Background

Berk et. al [1] created a modular, best-first search design for combinatorial Sparse PCA. Nodes in the search tree have a lower bound (generally stochastic method for a k-sparse eigenvalue approximation), as well as an upper bound (which is more challenging to bound).

Online singular value decomposition [2] is seen as a time-saving optimization, but constructing a k-subset SVD via column updates may not affect our big-O time complexity for the dimensions found here. The outer product is taken of the new column of data, and results in a symmetric rank-one matrix being added to the $k-1$-subset covariance matrix (logically including a zero row and column in the original subset matrix). We propose that once the Sparse PCA algorithm commits to including a variable (in the current tree being searched), then it should eagerly update the online SVD. This solved sub-problem should produce a variety of useful findings.

In a graph search, we will already have i variables committed to a new covariance computation, and there are still $\binom{n-i}{k-i}$ possible results. Upper bounding the k-sparse SVD solution is still challenging. Current bounds (using trace and Gershgorin circle on all possible subsets) avoid any decomposition or partial solution of the problem. We suggest that the diagonalization of the subproblem (i x i covariance matrix) is the best starting point for a more intensive search for an upper bound. There ought to be a less unwieldy, and better performing, bound on this partially decomposed problem, than the approach of successive updates shown here.

2 Sparse PCA problem instance

Berk et. al use $\Sigma \succeq 0$ as a dense, expected high-rank, covariance matrix input. An alternate mode uses the decomposition $\Sigma = M^T M$ in the case of few observations in M, for low-rank inputs, but we will focus on dense Σ. In either case, the estimator for the covariance has alreaday been computed. We may work with Σ with unbiased variance estimate on the diagonals, on $M = \sqrt{\Sigma} = V\text{Diag}(\sigma)V^T$ (where $\Sigma = V\text{Diag}(\sigma^2)V^T$), but not on original observations.

The sparse PCA solution will be a matrix $D : n$ x k, where columns are taken from $e_1, \ldots, e_n$ (selecting k variables). Then we want D which maximizes $||MD||_2^2$. Sparse PCA branch-and-bound will immediately start removing columns from M, so we cannot analyze the eigenvalues/singular values of M as a Hermitian eigenproblem.

Importantly, the majority of iterations of Sparse PCA branch-and-bound provide $D_i : n$ x $i, i < k$, where every column in D_i must appear in D.

3 Rank-one updates

Several treatments of sparse PCA/sparse recovery have appended rank-one or low-rank data to a subset of the data matrix, and desired an online SVD. However, fast perturbations of the subsetted covariance matrix are still a novel algorithmic detail.

Consider a subset of $k-1$ variables, and an existing positive definite diagonalized matrix $\Sigma^{k-1} = V^{k-1}\Lambda^{k-1}\left(V^{k-1}\right)^T$. We should treat the matrices as having an additional kth row and column (zeros in the covariance matrix and diagonal matrix, and the basis vector e_k in V).

Next, consider appending the variable to the subset, which will go at location k. For the matrix $\frac{1}{n-1}DD^T$, we have a rank-one update $\frac{1}{n-1}DD^T + \frac{1}{n-1}d_k d_k^T$. This matrix has the same spectrum as the updated covariance matrix Σ^k.

Using the singular vectors of D, we can produce another matrix which has the same spectrum: $\Lambda^{k-1} + \frac{1}{n-1}\left(V^{k-1}\right)^T \sigma'_k((\left(V^{k-1}\right)^T \sigma'_k)^T$, where σ'_k is a newly constructed column vector and the first $k-1$ elements come from the kth column of Σ, taken for the $k-1$ subset of covariates seen so far. $(\sigma'_k)_k$ is constructed to preserve the trace:

$$\mathrm{Tr}\left(\frac{1}{n-1}DD^T + \frac{1}{n-1}d_k d_k^T\right) = \mathrm{Tr}\left(\Lambda^{k-1} + \frac{1}{n-1}\left(V^{k-1}\right)^T \sigma'_k((\left(V^{k-1}\right)^T \sigma'_k)^T\right)$$

Consider the covariance matrix Σ of the subsets of variables Σ^{k-1}, Σ^k. The change in trace is due to inserting a new variance on the diagonal: Σ_{kk}.

$$\mathrm{Tr}\left(\frac{1}{n-1}\sigma'_k\left(\sigma'_k\right)^T\right) = \frac{1}{n-1}\sum_{i=1}^{k}\left(\sigma'_k\right)^2 = \Sigma_{kk}$$

$$\left(\sigma'_k\right)_k = \sqrt{\Sigma_{kk} - \sum_{i=1}^{k-1}\left(\sigma'_k\right)_i^2}$$

3.1 Solving the rank-one update

New eigenvalues are found analytically, by expanding the characteristic polynomial. The characteristic polynomial after the rank-one update is divided through, to produce a sum of k rational functions which have no zero, and one pole at exactly one original eigenvalue. After the sum, we have k zeroes, and every zero is separated from the next zero by one of the k poles. The updated eigenvalues (at the zeroes) are at least slightly perturbed from the original locations. The equations are solved by a root-finding approximation, but quickly converging up to machine epsilon, using the LAPACK `slaed9` helper function.

For the eigenvalue numbered i, the updated eigenvector is:

$$x_i v_i = V^{k-1}(\Lambda^{k-1} - I\left(\Lambda^k\right)_{ii})^{-1}\sigma'_k$$

This is divided through by a scale factor x_i determinied by taking the norm.

This formula performs stably for infinitesimal updates. The inverted diagonal matrix is ill-conditioned due to the large diagonal entry at i, so the result would be infinitesimally close to extracting one original eigenvector using e_i.

This technique is under-utilized so far, for accelerating this combinatorial linear algebra problem. Generally, combinatorial sparse PCA has not moved from research into practicum (productionizing and optimizing a particular algorithm). Time and space complexity is dominated by tracking and multiplying through a k x k eigenvector matrix, and big-O performance of each step in a combinatorial procedure has not changed.

4 Lower bound on Sparse PCA from D_i

At each iteration, for $MD_i = U_i \text{Diag}(\sigma_i)V_i^T$ (SVD), we have σ_i^2 and V_i provided to us from the Hermitian eigenproblem (rank-one updates). U_i^1 (dominant left eigenvector) ought to start approaching the linear combination of k variables which will follow from the optimal Sparse PCA solution (U_k^1). The Sparse PCA branch-and-bound lower and upper bounds will refer to a dominant left singular vector estimate u, which here will be implemented using U_k^1.

Our lower bound on the Hermitian eigenproblem M^TM follows from the factor M, projecting columns of M onto a smaller subspace. Use uu^T as our projection matrix. Then, $||(uu^TMD)^T(uu^TMD)||_2 = ||(uu^TMD)||_2^2 = ||(MD)^Tu||_2^2$, which is a lower bound linear in the columns of M being selected. Similar to the upper bound from the trace, this is a greedy optimization problem.

5 Upper bound on Sparse PCA from D_i

Upper bound will require a detailed analysis of the perturbation from our rank-one projection to the full-rank matrix Σ.

$$D^T\Sigma D = A + E \text{ where } A = D^TMuu^TMD, E = D^TM(I - uu^T)MD$$

In the optimal case, u would be the dominant singular vector of MD_k, which is a matrix which spans a range containing the range of MD_i. For any u, we have:

$$||A||_2 \leq \left(\sigma_1(MD)\right)^2 , ||E||_2 \geq \left(\sigma_2(MD)\right)^2$$

Additionally, Weyl's inequalities apply, and they may produce a reasonable (not tight) upper bound:

$$||D^T \Sigma D||_2 \leq ||A||_2 + ||E||_2$$

References

[1] Berk, Lauren, & Dimitris Bertsimas. "Certifiably optimal sparse principal component analysis." Mathematical Programming Computation 11.3 (2019): 381-420.

[2] Bunch, James R., & Christopher P. Nielsen. "Updating the singular value decomposition." Numerische Mathematik 31.2 (1978): 111-129.

[3] Zhu, L., Peng, X. & Liu, H. Rank-one perturbation bounds for singular values of arbitrary matrices. J Inequal Appl 2019, 138 (2019).