Perturbation bound on Sparse PCA

Daniel Ringwalt

March 27, 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 $||MD||_2^2 = ||D^T \Sigma D||_2$. 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 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_1^2 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 $M^T M$ follows from the factor M, projecting columns of M onto a smaller subspace. Use uu^T as our projection matrix. Then, $||(uu^T MD)^T (uu^T MD)||_2 = ||(uu^T MD)||_2^2 = ||(MD)^T u||_2^2$, 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.

After greedily selecting k variables in the subspace, then we can take those k rows and columns in the original Hermitian eigenproblem. This gives us a concrete solution D_k and a new lower bound on the objective value (not the rank-one problem, but a new k-by-k Hermitian eigenproblem): $(\sigma_1^\ell(MD_k))^2$.

4 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.

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

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:

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

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

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

Set $\cos\theta = \langle v_1, u\rangle$, where we expect an eigenspace with multiplicity 1: $(D^T\Sigma D)v_1 = (\sigma_1(MD))^2 v_1$. Apply Davis-Kahan to the dominant eigenvector v_1 of $D^T\Sigma D$, and to the range of $D^T Muu^T MD$:

$$|\sin\theta| \le \frac{||D^T\Sigma D - D^T Muu^T MD||_F}{||D^T\Sigma D||_2} = \frac{||E||_F}{(\sigma_1(MD))^2}$$

We are testing an angular distance between principal eigenvectors. A sine of zero has the consequence that u is the principal eigenvector of $D^T\Sigma D$, aligning with $||E||_F$ minimized. Then we would have $||E||_2 = (\sigma_2(MD))^2$ (the first singular value was eliminated without otherwise affecting the SVD). There is action on the first singular value of $(I-uu^T)MD$ given by $\sin\theta$, and action on the remaining singular values (upper-bounded by $\sigma_2(MD)$) given by $\cos\theta$. These projections of $(I-uu^T)MD$ onto further smaller subspaces have orthogonal ranges, so we have:

$$\sigma_1((I-uu^T)MD) \le \sqrt{(\sigma_1(MD))^2\sin^2\theta + (\sigma_2(MD))^2\cos^2\theta}$$

Now, our bound on $||E||_2$ depends on $||E||_F^2$. If $||E||^2$ can be derived (or bounded) as a linear function in the indicator variables for the entries of D, then we will produce a new linear program to maximize $(\sigma_1(MD))^2 \le ||A|| + ||E||$. The coefficients should produce a tighter bound than the linear program which optimizes Sparse PCA using the trace of the submatrix. As part of linearizing the problem, $\sigma_2(MD)$ will be upper-bounded using an upper bound on the trace and the lower bound on $\sigma_1(MD)$: $(\sigma_2(MD))^2 \le \mathrm{Tr}\ D^T\Sigma D - (\sigma_1(MD))^2$.

4.1 Nonlinear optimization from matrix contraction

We can treat these entries as a 2-by-1 matrix, a contraction of $(I-uu^T)MD$ which upper-bounds its action (the operator norm). Without affecting the operator norm, we can accept inputs which are in the same two-dimensional contracted space as the output which we create. Call the contracted matrix C.

$$C = \begin{pmatrix} \sigma_1(MD)\sin^2\theta & \sigma_1(MD)\sin\theta\cos\theta \\ \sigma_2(MD)\sin\theta\cos\theta & \sigma_2(MD)\cos^2\theta \end{pmatrix}$$

In this vector space, our rank-one matrix $uu^T MD$ is modeled as:

$$B = \begin{pmatrix} ||D^T Mu||_2 & 0 \\ 0 & 0 \end{pmatrix}$$

Notably, $\sigma_1((I-uu^T)MD) \le ||C||_F$ is a useful inequality, preserved from the square root expression above.

$$C^T C = \begin{pmatrix} \sigma_1^2(x^2 + x(1-x)) & \sigma_1\sigma_2\sqrt{x(1-x)} \\ \sigma_1\sigma_2(x+1-x)\sqrt{x(1-x)} & \sigma_2^2((1-x)^2 + x(1-x)) \end{pmatrix}$$

where $x = \sin^2\theta$.

Simplify:

$$C^T C = \begin{pmatrix} \sigma_1^2 x & \sigma_1\sigma_2\sqrt{x(1-x)} \\ \sigma_1\sigma_2\sqrt{x(1-x)} & -\sigma_2^2 x \end{pmatrix}$$

Apply Gershgorin circle theorem to $B^2 + C^T C$. The bound on the spectral radius can be implemented as a conic program, where σ_1, σ_2 are our analytic upper bounds on the problem by other methods.

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.