Perturbation bound on Sparse PCA

Daniel Ringwalt

April 12, 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 $||\Sigma||_F^2$

This bound comes from the Hermitian eigenproblem. For each row of Σ elementwise squared, put the entries into a linear program objective where our vector $d \in \mathbb{R}^n, 0 \leq d_i \leq 1, ||d||_1 = k$. If we introduce linear cuts (produced in a loop by proceeding to the next section), then we can tighten the objective value. Otherwise, this is solved greedily. Berk et. al pre-sorted the rows of Σ, improving big-O performance.

Next, select k partial row sums from the vector of row sums, again with the possibility of introducing linear cuts. We upper-bounded $||D^T \Sigma D||_F^2$. This objective value, after square root, has more sensitivity/fewer tied objective values in the tree, and is tighter, compared to the absolute value row sums ∞-norm (Berk et. al, Gershgorin circle theorem). This approach (when linear programming is avoided) has the same asymptotic runtime as Berk et. al's Gershgorin routine.

5 Upper bound on Sparse PCA from D_i (Golub)

Try completing the data matrix MD by adding $k-i$ columns to D_i (appending $k-i$ columns to an SVD problem). Multiply MD on the left by an orthonormal matrix (from Gram-Schmidt), where the first i rows are the left singular vectors of MD_i: $u_1, ..., u_i$. When we add a column to MD_i, we know that the first entry in the column after multiplication is $\langle u_1, M_i \rangle$. The ℓ_2 norm of the added column is already known, and the perturbation to the first singular vector is maximized if the remaining mass in the column is placed in the second component, where u_2 has the second-largest singular value.

In terms of the Hermitian eigenproblem, we are proposing adding to the leading 2x2 top-left block of the matrix $U^T D_i^T M^2 D_i U$.

The update is maximized when the columns added in the perturbation are identical to one another, so apply the rank-one update for one such column M_j:

$$||U^T D_i^T M^2 D_i U + U^T M_j M_j^T U|| \leq ||U^T D_i^T M^2 D_i U + vv^T||$$

$$v = \begin{bmatrix} \langle u_1, M_j \rangle \\ \sqrt{||M_j||^2 - \langle u_1, M_j \rangle^2} \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

Apply Golub:

$$\begin{aligned}
w(\lambda) &= 1 + \frac{v_1^2}{(\sigma_1(MD_i))^2 - \lambda} + \frac{v_2^2}{(\sigma_2(MD_i))^2 - \lambda} \\
&= 0 \\
0 &= \left((\sigma_1(MD_i))^2 - \lambda\right)\left((\sigma_2(MD_i))^2 - \lambda\right) + v_1^2\left((\sigma_2(MD_i))^2 - \lambda\right) + v_2^2\left((\sigma_1(MD_i))^2 - \lambda\right) \\
0 &= (\sigma_1(MD_i))^2 (\sigma_2(MD_i))^2 - \left((\sigma_1(MD_i))^2 + (\sigma_2(MD_i))^2\right)\lambda + \lambda^2 \\
&\quad + v_1^2 (\sigma_2(MD_i)) + v_2^2 (\sigma_1(MD_i)) - \left(v_1^2 + v_2^2\right)\lambda \\
\lambda &= \frac{1}{2}\left(b \pm \sqrt{b^2 - 4(\sigma_1(MD_i))^2 (\sigma_2(MD_i))^2 - v_1^2 (\sigma_2(MD_i))^2 - v_2^2 (\sigma_1(MD_i))^2}\right) \\
b &= (\sigma_1(MD_i))^2 + (\sigma_2(MD_i))^2
\end{aligned}$$

Columns in D will be selected greedily based on $\langle u_1, M_j \rangle$, as this basis vector contributes more to the eigenvalue perturbation than the other eigenvectors do. The $k-i$ columns of M should iteratively update the first and second eigenvalues of the system, applied in any order.

6 Upper bound on Sparse PCA from D_i (Davis-Kahan)

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

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

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:

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

Returning to the singular value problem, analyze $||D^T M v||_2^2$ for some vector v. After change-of-basis onto the left singular vectors of MD, the first component of the vector that we will analyze is equal to $\langle v, u \rangle$. The squared norm of the matrix-vector multiplication squares this entry and multiplies by σ_1^2, squares a next entry and multiplies by σ_2^2, and so on until σ_k^2. We can compute various norms of $D^T M$ using the diagonal matrix of the singular values (from SVD). However, when we apply a projection $D^T M (I - u u^T)$, then the singular values can change dramatically, by mapping orthogonal left singular vectors of MD to vectors which are no longer orthogonal.

After SVD, as multiplication on the left of MD is analyzed using a diagonal matrix after SVD, then multiplication on the left of $(I - u u^T) MD$ will be bounded using a single-column (or single-row) matrix. Construct a vector in $\mathbb{R}^2$:

$$\mathbf{v} \to \begin{bmatrix} v_\top \\ v_\perp \end{bmatrix} = \begin{bmatrix} \langle \mathbf{u}, \mathbf{v} \rangle \\ \sqrt{||\mathbf{v}||_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2} \end{bmatrix}$$

$$||D^T M v||_2 \leq \left|\left| B \begin{bmatrix} v_\top \\ v_\perp \end{bmatrix} \right|\right|_2 = \left|\left| \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{bmatrix} v_\top \\ v_\perp \end{bmatrix} \right|\right|_2$$

$$||(I - u u^T) D^T M v||_2 \leq \left| C \begin{bmatrix} |v_\top| \\ |v_\perp| \end{bmatrix} \right| \leq \left| \begin{pmatrix} \sigma_1 & \sigma_2 \end{pmatrix} \begin{bmatrix} |v_\top| \\ |v_\perp| \end{bmatrix} \right|$$

However, the entries of C will actually be smaller than the singular values which we started with, because $I - u u^T$ is a projection matrix with a null space spanned by u. We find the first entry of C by multiplying by $||(I - u u^T) v_1||$ (right singular vector of MD), and we find the second entry of C by multiplying by an upper-bound on $||(I - u u^T) v_i||, 1 < i \leq k$ (subspace complementary to

the first singular vector). The bounds on these subspace relations are found using the Davis-Kahan $\sin\theta$ theorem [3]:

$$C = \begin{pmatrix} \sigma_1 \frac{||E||}{||\Sigma||} & \sigma_2 \frac{||E||}{||A|| - \sigma_2^2} \end{pmatrix}$$

Relating the singular value problem to the Hermitian eigenproblem, $\sigma_1^2 = ||\Sigma||$. We will try to make the denominator in the second entry a constant, by taking a lower-bound on $||A||$ from any particular D' matrix and $||MD'u||_2^2$. NOTE: This may not actually be a lower bound, since a better $||\Sigma||$ could still exist with a smaller component along our choice of u...

$$C = \begin{pmatrix} \frac{1}{\sigma_1} & \frac{\sigma_2}{||MD'u||^2 - \sigma_2^2} \end{pmatrix} ||E||$$

Now, our bound on $||E||_2$ comes from $||C||^2$, which depends on $||E||^2$ (to be implemented by bounding $||E||_F^2$). To linearize the problem, the 1-by-2 matrix must be held constant and its singular value upper-bounded. 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 $(\sigma_1(MD))^2 \leq ||A|| + ||E||$. 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 $\sigma_1(MD)$: $(\sigma_2(MD))^2 \leq \text{Tr } D^T \Sigma D - (\sigma_1(MD))^2$.

Analyze $||MDu||_2^2 + ||C^T C||_2 \geq ||\Sigma||$ (Weyl's inequalities). For both expressions, we produced an upper bound which is linear in the entries of D (indicator variables).

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] Davis, C., & Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. SIAM Journal on Numerical Analysis, 7(1), 1-46.