Perturbation for Sparse PCA: Project Overview

Daniel Ringwalt

April 7, 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). We work in this combinatorial framework, and try to improve performance of the lower and upper bounds.

We adapt Berk et. al's treatment of the Hermitian eigenproblem (PCA on the covariance matrix) to a singular value analysis, without changing the actual implementation. Sparse PCA is treated on a matrix of m observations and n variables: M, and considering different n-by-k matrices, D, where the columns are taken to be standard basis vectors (subsetting the columns of the identity matrix). Then $\max_D ||MD||_2^2$ provides us with the maximum of Hermitian eigenproblem taking k rows and columns: $\max_D ||D^T \Sigma D||$, where $\Sigma = M^T M$ is a sample covariance matrix.

Berk et. al perform a branch-and-bound [5] tree search, where the root node has D with no columns, the next level has $D_1 \in M_{n,1}(\mathbb{R})$, and has a terminal level where $D \in M_{n,k}(\mathbb{R})$. The singular value treatment immediately lends a rank-one update computational tactic to each graph edge [3], which was not pursued by Berk et. al. This tactic does not improve performance on its own, and analysis of the partial matrix MD_i and its perturbation needs to be performed in this project.

We will refer to the "trace bound", which is a basic bound on the objective from the squared column norms of M (with equality if M is rank-one). If M is unit-normed, then the trace bound is a constant $||MD||_2^2 \leq k$. For more general data, then the trace bound admits a greedy tactic for variable selection, rather than Sparse PCA, by retaining only high-variance variables. We will display results from Berk et. al showing that although unit-normed data does not admit a greedy tactic, the covariance matrix's spectrum was actually more amenable to analysis (an observation from their program's performance).

2 Lower bounds on Sparse PCA

We will leverage: $||D^T M^T u||_2 \leq ||MD||_2$. A good u will come, at a graph node that is not trivially shallow, from the first singular vector of MD_i at this node. The lower bound on the system (from projecting the system onto a subspace spanned by this vector) allows us to elide all of the stochastic, or power iteration, methods pursued by Berk et. al, but will not prove impactful on the provided code and data sets.

3 Upper bound from Frobenius norm

We will review Berk et. al's Gershgorin circle theorem bound, which appears as a useful tactic for bounding many matrices whose rows and columns are drawn each from finite possible entries (the set of full rows and columns of the covariance matrix). We will generalize this algorithm, instead of retaining one matrix row, to retain k matrix rows (after elementwise squaring) in order to bound the squared Frobenius norm $||D^T \Sigma D||_F^2$. We will present our extension of Berk et al's code, to simultaneously produce a Gershgorin circle bound and Frobenius bound.

We can construct a matrix where the Frobenius norm is contained inside the largest Gershgorin disk (a tighter bound), as well as a matrix where the Frobenius norm is beyond the Gershgorin disks on the real line (a poorly-performing bound). We will show for Sparse PCA that when we increment our k parameter, then the difference of the Frobenius norm minus the Gershgorin bound will become more negative, and using the Frobenius norm alone performs well on the provided test cases and values of k.

4 Perturbation upper bound from Davis-Kahan

We will finally treat the singular values of MD as a perturbation from MD_i. For this section, we will move to the Hermitian eigenproblem for $D^T \Sigma D$. We will justify the first singular vectors of MD_i and MD being similar, but the dimensions of MD_i possibly being very small. Instead, we will project all variables onto a one-dimensional subspace and its complement. The subspace is spanned by the first singular vector u of MD_i. We have $A = D^T M^T u u^T MD$, and $E = \Sigma - A$.

For this problem, introduce a new $\Sigma \in M_{k,k}(\mathbb{R}), \Sigma = D^T M^T MD$, and introduce binary indicator variables $d \in \{0,1\}^n, ||d||_1 = k$ representing the entries of D. We form an angular similarity between the first eigenvector of Σ to u. This is analyzed in the subspaces of Σ and of a rank-one matrix scaled up to closely match Σ: $A = D^T M^T u u^T MD$. The Davis-Kahan $\sin\theta$ theorem [4] gives us the angular similarity which we are seeking, when we analyze $||E||$ (the difference of the two matrices). E involves a projection matrix internally, and Ev (for some v) can introduce action from both the dominant eigenvalue which we started with $((\sigma_1(MD))^2)$ as well as any other eigenvalues (bounded

by $(\sigma_2(MD))^2$). We multiplied the eigenvectors so that they may no longer be orthogonal, so we are now forced to add eigenvalues together, but we may shrink the effect coming from the first eigenvalue of Σ because its eigenvector has a small sine distance from u (which spans the kernel of the projection matrix).

Davis-Kahan will provide us with an inequality of $||E||$ (any unitarily invariant norm), which we will linearize. After our analysis, this linear program's objective could be greater than that of the trace bound, or it could be a novel and tight bound for Sparse PCA. As we are already taking the min of multiple analyses, this is a useful new tool in the toolbox for this system.

We started with one tactic which was linear in the entries of D (the trace bound). We produced a projected rank-one system (lower bound), a concave Frobenius norm bound (square root of the linear expression), and a linearized (Weyl) upper bound (depending on squared Frobenius norm bound), all of which are solved by a simple greedy solver. As a tactic, we will introduce comparing the lower bound and upper bound of variables' contributions (now that the contributions are linearized), as a novel variable selection step.

5 Future Directions

Berk's collaborator, Bertsimas, has produced follow-ups using mixed-linear solvers and branch-and-cut [2]. Our Davis-Kahan analysis produces interesting linear cuts in the variables, and may be best suited to branch-and-cut rather than branch-and-bound.

References

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

[2] Bertsimas, D., Cory-Wright, R., & Pauphilet, J. (2022). Solving Large-Scale Sparse PCA to Certifiable (Near) Optimality. J. Mach. Learn. Res., 23(13), 1-35.

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

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

[5] Lawler, E. L., & Wood, D. E. (1966). Branch-and-bound methods: A survey. Operations research, 14(4), 699-719.