1 Feb 15 notes

Suppose that the observations are a Hermitian matrix. If they are not, then start from the Gram matrix and take a matrix sqrt of this matrix to be the observation matrix.

We would like to analyze $||\Sigma||_{\text{op}}$.

$$\Sigma = D^T M^T M D$$

Where M is the Hermitian data matrix, and D is an n-by-k matrix: $\begin{pmatrix} e_{d_1} & e_{d_2} & ... & e_{d_k} \end{pmatrix}$, selecting k columns of M using multiplication on the right.

We introduced u, which can be initialized to be the left singular vector of MD for some well-performing D ($\sigma = |uMD|_2$ large). This is slightly useful, because inner products with u could be factored out to be computed using entries of Σ only, and discard/never synthesize an M.

Set $A = D^T M^T u u^T M D$ (large and rank-one), and $E = \Sigma - A$. Analyzing E for different permissible D should give us a smaller recursive problem, compared to analyzing Σ (shrinking the problem). However, we would like to produce a new upper-bound deterministically in polynomial time instead.

We found that E is similar to:

$$(I - \frac{1}{\sigma^2} M D u u^T D^T M) M^2$$

We hope that $||E||_{\text{op}}$ is small compared to $||\Sigma||_{\text{op}}$. Note that with a change-of-basis using the diagonalization of M, then the matrix in parentheses is still a rank-$n-1$ projection matrix. It could have a zero for the dominant eigenvalue of M^2, and that is the best case. We want a large sine of the angular comparison of the dominant eigenspaces of the projection matrix on the left and of M, and this sine can be bounded using the Davis-Kahan theorem.

Let $P = I - \frac{1}{\sigma^2} M D u u^T D^T M$. For Davis-Kahan, we should come up with our own upper-bound on $||P - \frac{1}{\sigma^2} M^2||_{\text{op}}$. Davis-Kahan bounds the dissimilarity in the largest eigenspaces of P and $\frac{1}{\sigma^2} M^2$, and if this dissimilarity obtains its maximum value, then the dominant eigenvalue would drop out of PM^2 and $||PM^2||_{\text{op}}$ would be driven by the second-largest eigenvalue of M^2 instead.

2 March 1 notes

We would like a large sine distance between the eigenvector spanned by the range of $\frac{1}{\sigma^2} M D u u^T D^T M$, and the dominant eigenvector of M. We expect:

$$\sin \angle (\frac{1}{\sigma^2} M D u u^T D^T M, M) = \cos \angle (I - \frac{1}{\sigma^2} M D u u^T D^T M, M)$$

We don't need to use $\sin^2 \theta + \cos^2 \theta = 1$, because we are actually swapping whether the vector or some orthogonal vector is used as sin or as cos.

To improve performance of Davis-Kahan, scale up the rank-one projection matrix so that it is $M D u u^T D^T M$. Apply Davis-Kahan:

$$|\sin\theta| \leq \frac{2||M(I - Duu^T D^T)M||_F}{\max(||MDuu^T D^T M||_{\text{op}}, \text{Gap}[M^2])}$$

We need a lower bound on Gap$[M^2]$. If our gap evaluation is intentionally simplistic (take $\lambda_2 = \text{Tr } M^2 - \lambda_1$), then the gap estimate is nondecreasing with λ_1. If we choose a putative eigenvector to evaluate λ_1 for Gap$[M^2]$, and later we increase our objective function, then our gap using the best sparse PCA solution so far will still hold as a lower bound. Therefore, we use Berk et. al's stochastic best-observed lower bound solution to compute the gap.

2.1 Follow-up notes

Referring to Theorem 2, not Theorem 1.

- This linearizes the variable selection as a linear program, which will be solved greedily (trivial optimum). For example, Berk et. al's maximizing the trace is also linear and greedy, and we can directly compare performance to that upper bound. It is becoming extremely cheap to compute these linear coefficients so will they perform reasonably well?
- Does $d = 1$ because the eigenvectors being tested are a rank-one subspace?
- Eigenvalue gap in the denominator: It is our choice whether to use Σ or $\hat{\Sigma}$ in the denominator.
- We know that $M(I - Duu^T D^T)M$ is PSD so it is the Gram matrix of some matrix, but I need a refresher on what "observation" vectors constitute this matrix. The data vectors cannot be a la $M(I - Duu^T D^T)$, because summing data matrices and then taking the covariance does not equal the sum of two covariance matrices.
- Need to add a first pass, where each variable's contribution is lower-bounded using the rank-one projection, and upper-bounded using Davis-Kahan. Remove variables immediately that can't possibly be in the optimum Sparse PCA, maybe shrinking the problem will still improve performance even with this simple objective.

3 Davis-Kahan project proposal

Suppose that we have some $\Sigma \in M_{n,n}(\mathbb{R})$ and matrix square root $M^2 = \Sigma$. For Berk et al, we want to evaluate an upper bound on k-sparse PCA, hundreds of times for different $D_0 \in M_{n,i}(\mathbb{R})$. D_0 is a sparse matrix that selects i columns; every column of D_0 is a column of the identity matrix. The upper bound is on every $\lambda(D^T \Sigma D), D \in M_{n,k}(\mathbb{R}), i < k < n$, with the constraint that every column of D_0 must appear as a column of D.

Solve for v, the leading eigenvector of $D_0^T \Sigma D_0$. We have that $D^T \Sigma D$ has the same spectrum as $MDD^T M$. Decompose this into $MDvv^T D^T M + MD(I -$

$vv^T)D^TM$. For fast computation, it is very useful that $\lambda_1(MDvv^TD^TM) = \text{Tr } MDvv^TD^TM$ is a linear combination of coefficients ($\langle Me_j, v\rangle^2$) for every column (variable) that may be selected using D. For the second term, $E = MD(I - vv^T)D^TM$, we have that $\lambda_1(E) \leq ||MD(I - vv^T)D^TM||_F$, which can be readily computed from the square root or Cholesky decomposition of this matrix, and squared row ℓ_2 norms. If we add these two norms (Weyl's inequality), then the performance is inferior to adding the contribution to the trace from each variable (Berk et al's simple upper bound from the trace).

E is similar to $(I-vv^T)MDD^TM$. We are particularly interested in the sine of the angle between v and the principal eigenvector of MDD^TM for a particular D. Apply Davis-Kahan: If we upper-bound $\lambda_1(MDD^TM), \lambda_2(MDD^TM)$ for optimal D, then the cosine of the angle of the space spanned by $I - vv^T$ and MDD^TM will shrink our upper bound of $||(I - vv^T)MDD^TM||_2$, from $\lambda_1(MDD^TM)$ towards $\lambda_2(MDD^TM)$ (which is its smallest achievable value). Fully exploring the leading eigenvalues of MDD^TM, for the particular D matrix, is NP-hard (sparse PCA). We propose a linear program based on $\langle Me_j, v\rangle^2$ coefficients (which hopefully explain much of the variance), with an additional term tacked on which is multiplied by $\lambda_1^u(MDD^TM) - \lambda_2^u(MDD^TM)$ (global upper bounds on the problem)