Contents

1 Table for matrix sketching results

	Covariance estimation ($\mathbb{E}\|\hat{\Sigma}-\Sigma\|_2^2$ or canonical angle)	Eigenvalues	Eigenvectors	PCR ($\hat{Y}$)	PCR ($\hat{\beta}$)
Nystrom					
CS		CS_2	CS_3		CS_5
Martinsson					

Table 1: Note that the CS results apply to AIMER.

2 Preliminary notation, definitions, and statistical model

2.1 Notation

- $\mathbb{X} \in \mathbb{R}^{n\times p}$
- $\mathbf{x}_j = [X_{i1}, \ldots, X_{ip}]^\top \in \mathbb{R}^p$ is the j^{th} column of $\mathbb{X}$ and an i.i.d. sample from $x_j \sim N(0, \Sigma(j,j))$.
- $\mathcal{P} = \{1, \ldots, p\}$
- $\mathcal{A} = \{$ of active covariates $\}$
- $\mathcal{S} = \{$ nonzero marginal covariance $\}$ (using $\mathcal{S}$ as it is the 'selected' model)
- $\mathcal{D} = \mathcal{A} \setminus \mathcal{S}$ (to be the difference between active and selected covariates)
- $\mathcal{T} = \{$ nonzero θ $\}$ (using $\mathcal{T}$ due to... whatever)
- For any subsets A, B of $\mathcal{P}$ and matrix $\mathbb{A}$, the submatrix with rows A and columns B is $\mathbb{A}_{A,B}$

2.2 Definitions

- The underlying machinery of these supervised PCA papers is a suite of estimators of the form $\hat{\Sigma}_{A,B}$, where $A, B \subseteq \mathcal{P}$. In the SPCA paper, they choose $\hat{\Sigma}_{\mathcal{S},\mathcal{S}}$. Using F is tantamount to using $\hat{\Sigma}_{\mathcal{P},\mathcal{S}}$. This protects somewhat against $\mathcal{S} \subset \mathcal{A}$. If we had a good estimator of $\mathcal{T}$ we would/could use $\hat{\Sigma}_{\mathcal{T},\mathcal{S}}$ instead. Perhaps this estimator should be investigated as well...
- $F = \mathbb{X}^\top \mathbb{X}_1 = V(F)\Lambda(F)U(F)^\top$ (note, I think this reversed order makes much more sense at we are looking at approximating $\mathbb{X}^\top\mathbb{X} = VD^2V^\top$...) I haven't included any normalization by a function of n, which is surely necessary to get convergence. In particular, the sample covariance would be $n^{-1}\mathbb{X}^\top\mathbb{X}$, so defining $F \leftarrow n^{-1}F$ would seemingly make sense.

2.3 Model

- Let the covariance matrix for X be

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{1}$$

where $\Sigma_1 \equiv \Sigma_{\mathcal{A},\mathcal{A}} = \Theta\Lambda\Theta^\top + \sigma^2 I = \sum_{m=1}^M \lambda_m \theta_m \theta_m^\top + \sigma^2 I$. This should be equivalent to the model in the next bullet if $\Sigma_2 = \sigma^2 I$. We can probably generalize this model to let Σ_2 have its own eigenvector structure (with eigenvalues strictly smaller than Λ.

- $X_{ij} = \sum_{m=1}^M \lambda_m^{1/2} \eta_{im}\theta_{jm} + \sigma z_{ij}$ where $\|\theta_m\|_2^2 = 1$ and $\theta_m \equiv \theta_{\cdot m}$, $\langle \theta_m, \theta_{m'} \rangle = 0$ if $m \neq m'$, and $\theta_{jm} = 0$ if $j \notin \mathcal{A}$. All z_{ij} and η_{im} are standard normals and mutually independent.
- The regression model:

$$Y_i = \beta_0 + \sum_{m=1}^{\tilde{M}} \beta_m \eta_{im} + W_i. \tag{2}$$

  Here, I write $\tilde{M}$ to indicate the this may be different than M.

2.4 Assumptions

- $(\sum_{m=1}^M \lambda_m \theta_{jm}\theta_{km})^2 \leq \gamma_n$ for $k \in \mathcal{D}$.
- We can estimate σ^2 well so we consider it known (really, just to simplify things so we can just subtract off the diagonal component before hand)
- $\lambda_{\max} \leq C_\Lambda$ independent of n
- Probably eventually will need to codify the rates for size of some of the above sets (e.g. $|\mathcal{A}| \asymp a_n$)

3 Showing CS_2

3.1 Overview

Suppose $A \in \mathbb{R}^{m\times n}$ and $\tilde{A} = A + E$. Also, $U^\top AV = \begin{bmatrix} D \\ 0 \end{bmatrix}$ and $\tilde{U}^\top \tilde{A}\tilde{V} = \begin{bmatrix} \tilde{D} \\ 0 \end{bmatrix}$. There is a main theorem for bounding the distance between D and $\tilde{D}$:

Theorem 3.1 (Mirsky)**.**

$$\|\tilde{D} - D\| \leq \|E\|$$

where the norm can be any unitarily equivalent norm (e.g. $\|\cdot\|_2$ *or* $\|\cdot\|_F$*).*

Ultimately, we will probably use the following $\forall k$:

$$|D_k - \tilde{D}_k| \leq \|E\|_F$$

3.2 The result

To start, write $M_{ij} := \sum_{m=1}^{M} \lambda_m^{1/2}\eta_{im}\theta_{jm}$. Then $M_{ij} = 0$ if $j \notin \mathcal{A}$.

$$\begin{aligned}
F &= \left[\sum_{i=1}^{n} X_{ij}X_{ik}\right]_{j\in\mathcal{P},k\in\mathcal{S}} & (3)\\
&= \left[\sum_{i=1}^{n}\left(\sum_{m=1}^{M} \lambda_m^{1/2}\eta_{im}\theta_{jm} + \sigma z_{ij}\right)\left(\sum_{m=1}^{M} \lambda_m^{1/2}\eta_{im}\theta_{jm} + \sigma z_{ij}\right)\right]_{j\in\mathcal{P},k\in\mathcal{S}} & (4)\\
&= \begin{bmatrix} \sum_{i=1}^{n}(M_{ij} + \sigma z_{ij})(M_{ik} + \sigma z_{ik}) \\ \sum_{i=1}^{n}(\sigma z_{ij})(M_{ik} + \sigma z_{ik}) \end{bmatrix} & (5)\\
&= \begin{bmatrix} \sum_{i=1}^{n}(M_{ij}M_{ik} + \sigma z_{ij}M_{ik} + \sigma z_{ik}M_{ij} + \sigma^2 z_{ij}z_{ik}) \\ \sum_{i=1}^{n}(\sigma z_{ij}M_{ik} + \sigma^2 z_{ij}z_{ik}) \end{bmatrix}, & (6)
\end{aligned}$$

where the top block has $j \in \mathcal{A}$ and the bottom block has $j \in \mathcal{A}^c$ (this convention will persist for the rest of this proof).

Using the result from Theorem 3.1, write $A = \tilde{F}$ and $E = \Sigma_{\mathcal{P},\mathcal{D}} - \tilde{F}$, where $\tilde{F} = [F|0] \in \mathbb{R}^{p\times|\mathcal{A}|}$. The nonzero singular values of F and $\tilde{F}$ are identical. Hence, the approximation error in the estimation of the singular values of Σ_1 will be encoded in the difference $\Sigma_{\mathcal{P},\mathcal{D}} - \tilde{F}$.

Writing $E = \tilde{\Sigma}_{\mathcal{P},\mathcal{S}} - \tilde{\Sigma}_{\mathcal{P},\mathcal{S}} + \Sigma_{\mathcal{P},\mathcal{D}} - \tilde{F}$, where

$$\tilde{\Sigma}_{\mathcal{P},\mathcal{S}} = \begin{bmatrix} \Sigma_{\mathcal{A},\mathcal{S}} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{p\times|\mathcal{A}|},$$

then

$$\|E\|_F \leq \|\tilde{\Sigma}_{\mathcal{P},\mathcal{S}} - \tilde{F}\|_F + \|\Sigma_{\mathcal{P},\mathcal{D}} - \tilde{\Sigma}_{\mathcal{P},\mathcal{S}}\|_F.$$

We should have $\mathbb{E}\tilde{F} = n\tilde{\Sigma}_{\mathcal{P},\mathcal{S}}$ and hence should be able to control the first term with concentration or convergence results. The second term will have an irreducible error given by

$$\|\Sigma_{\mathcal{P},\mathcal{D}} - \tilde{\Sigma}_{\mathcal{P},\mathcal{S}}\|_F^2 = \sum_{j\in\mathcal{A},k\in\mathcal{D}} \Sigma_{j,k}^2 = \sum_{j\in\mathcal{A},k\in\mathcal{D}} \left(\sum_{m=1}^{M} \lambda_m\theta_{jm}\theta_{km}\right)^2 \leq |\mathcal{A}||\mathcal{D}|\gamma_n$$

under the assumptions in Section 2.4.

3.3 Old material assuming one latent factor (needs updating)

Start: delete this later. I'm including it to facilitate later translation.

The goal here is to extend the results of the SPCA paper by including the possibility in Σ_1 that we have missed some important features (hence their Σ_1 corresponds to our Σ_{11}. The model for X is then (here I'm writing/thinking about a single latent factor model, I'm presuming that complexifying that to multi-factor will be a matter of notation):

$$X_{ij} = v_i \theta_j + \sigma z_{ij}$$

where v_i, z_{ij} are all mutually independent standard normals and $\theta_j \neq 0$ iff $j \in \mathcal{S}$.
End: delete

Note the expectation of F:

$$\mathbb{E}F = \begin{bmatrix} n\theta\theta^\top + n\sigma^2 I \\ 0 \end{bmatrix}.$$

Hence,

$$E = \begin{bmatrix} F - \mathbb{E}F \end{bmatrix}. - \begin{bmatrix} n\theta\theta^\top + n\sigma^2 I \\ 0 \end{bmatrix}.$$

So, up to the $n\sigma^2 I$ factor, we have to bound the norm difference between a random matrix and it's expectation.

$$F - \mathbb{E}F - \begin{bmatrix} n\theta\theta^\top + n\sigma^2 I \\ 0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n v_i^2 \theta_j \theta_k - n\theta_j\theta_k \\ 0 \end{bmatrix} + \begin{bmatrix} \sigma \sum_{i=1}^n v_i (z_{ij}\theta_k + z_{ik}\theta_j) \\ \sigma \sum_{i=1}^n v_i z_{ij} \theta_k + \sigma^2 \sum_{i=1}^n z_{ij} z_{ik} \end{bmatrix} + \quad (7)$$

$$+ \begin{bmatrix} \sigma^2 \sum_{i=1}^n z_{ij} z_{ik} - n\sigma^2 I \\ 0 \end{bmatrix} \quad (8)$$

$$= (i) + (ii) + (iii). \quad (9)$$

Now,

$$(i) = (\sum_{i=1}^n v_i^2 - n) \begin{bmatrix} \theta\theta^\top \\ 0 \end{bmatrix}$$

and

$$(iii) = \sigma^2 \begin{bmatrix} \sum_{i=1}^n Z_{i\mathcal{S}} Z_{i\mathcal{S}}^\top - nI \\ 0 \end{bmatrix},$$

where $Z_{i\mathcal{S}} = [z_{ij}]_{j \in \mathcal{S}}$. Similarly, $Z_{i\mathcal{S}^c} = [z_{ij}]_{j \in \mathcal{S}^c}$. Hence,

$$(ii) = \sigma \sum_{i=1}^n v_i \begin{bmatrix} Z_{i\mathcal{S}} \theta^\top \\ Z_{i\mathcal{S}^c} \theta^\top \end{bmatrix} + \sigma \sum_{i=1}^n v_i \begin{bmatrix} \theta Z_{i\mathcal{S}}^\top \\ 0 \end{bmatrix} + \sigma^2 \sum_{i=1}^n \begin{bmatrix} 0 \\ Z_{i\mathcal{S}} Z_{i\mathcal{S}^c}^\top \end{bmatrix}.$$

Now, concentration results can be used to show E is small in Frobenius norm. Presumably, we can find some results about spectral norm as well (which would probably be more useful as it would allow us to say $|d_j - \tilde{d}_j| \leq \|E\|_2$ for all j).

4 Showing CS_3

Using the result that

$$\|\hat{v} - v\|_2^2 \leq 2 \sin(\angle(\hat{v}, v))$$

we can do the following. Supposing that $\Sigma = [\tilde{\Sigma}_{\mathcal{S}}|\tilde{\Sigma}_{\mathcal{S}^c}]$, $F = V(F)D(F)U(F)^\top$, $\tilde{\Sigma} = VDU^\top$, and $\Sigma = \Theta\Lambda\Theta^\top$, then for $k \in \mathcal{S}$,

$$\|v_q(F) - \theta_q\|_2 \leq \|v_q(F) - v_q\|_2 + \|v_q - \theta_q\|_2 \leq \sqrt{2}\left(\sin(\angle(v_q(F), v_q)) + \sin(\angle(v_q, \theta_q))\right).$$

So, by Yu et al. (2015)[1], Theorem 3

$$\sin(\angle(v_q(F), v_q)) \leq 2\frac{(2d_{\max} + \|F - \tilde{\Sigma}\|_{op})\min\{\|F - \tilde{\Sigma}\|_{op}, \|F - \tilde{\Sigma}\|_F\}}{\tilde{\delta}_q},$$

where $\tilde{\delta}_q = \min\{d_{q-1} - d_q, d_q - d_{q+1}\}$, which will be controlled by assumption on Σ (for instance, $d_{\max} \leq \lambda_{\max}$).

Now, looking at $\|F - \tilde{\Sigma}\|_F^2$ component wise for $j \in p$ and $k \in \mathcal{S}$

$$(F(j,k) - \tilde{\Sigma}(j,k))^2 = (\mathbf{x}_j^\top \mathbf{x}_k - \mathbb{E}x_j x_k)^2.$$

This will be controllable via asymptotics or concentration.

There will be nonzero approximation bias if $\mathcal{D} \neq \emptyset$. Using the same result as above

$$\sin(\angle(v_q, \theta_q)) \leq 2\frac{(2\lambda_{\max} + \|\tilde{\Sigma} - \Sigma\|_{op})\min\{\|\tilde{\Sigma} - \Sigma\|_{op}, \|\tilde{\Sigma} - \Sigma\|_F\}}{\delta_q},$$

where $\delta_q = \min\{\lambda_{q-1} - \lambda_q, \lambda_q - \lambda_{q+1}\}$. This quantity will again be controlled by assumption on Σ.

Now, looking at $\|\tilde{\Sigma} - \Sigma\|_F^2$ component wise for $j, k \in \mathcal{P}$

$$(\tilde{\Sigma}(j,k)) - \Sigma(j,k))^2 = \begin{cases} 0 & \text{if } k \in \mathcal{S} \\ (\sum_{m=1}^M \lambda_m \theta_{jm}\theta_{km})^2 & \text{if } j \in \mathcal{A}, k \in \mathcal{D} \\ (\sum_{m=1}^M \lambda_m \theta_{jm}\theta_{km} + \sigma^2)^2 & \text{if } j = k \in \mathcal{D} \\ (\sigma^2)^2 & \text{if } j = k \notin \mathcal{A} \end{cases}$$

Now, we might make some assumptions about the size of this "residual" components, due to a norm constraint on these components implying a norm constraint on the β's.

Some such assumptions are listed in Section 2.4. Then

$$\|\tilde{\Sigma} - \Sigma\|_F^2 \leq |\mathcal{A}||\mathcal{D}|\gamma_n,$$

which implies that

$$\sin(\angle(v_q, \theta_q)) \leq 2\frac{(2\lambda_{\max} + |\mathcal{A}||\mathcal{D}|\gamma_n)|\mathcal{A}||\mathcal{D}|\gamma_n}{\delta_q},$$

5 Showing CS_5

1. Show that $v_m(F)$ is close to θ_m (the PC loadings) and $\lambda_m(F)$ is close to λ_m
   (a) This is the topic of the document "convergenceSingularVectorsValues.pdf". We need show that $v_m(F)$ converges to θ_m. So, perhaps, $v_m(F) = \theta_m + \delta_m$, where $\|\delta_m\|$ is small (note: we need to formalize the connection between bounded sin(canonical angles) of singular vectors and writing them in the fashion. Perhaps the asymptotic expansion is more amenable?)

[1] http://www.statslab.cam.ac.uk/~yy366/index_files/Biometrika-2015-Yu-biomet_asv008.pdf

2. The regression part of the procedure regresses Y onto the PC scores, which are the coordinates in the PC, given by $\hat{u}_m = \mathbb{X} v_m(F) \lambda_m^{-1/2}(F)$. We need to show that these coordinates aren't too far from the coordinates created by inner product with $\theta_{m'}$:

$$\left\langle \sum_{m=1}^{M} \eta_{im} \theta_m, \theta_{m'} \right\rangle = \eta_{i,m'} \lambda_{m'} \tag{10}$$

(a) This can be done via inserting the model for X in for $\mathbb{X}$ in the definition of $\hat{u}_k$.

$$\mathbb{X} v_k(F) = \begin{bmatrix} \sum_{j=1}^{p} \left(\sum_{m=1}^{M} \lambda_m^{1/2} \eta_{1m} \theta_{jm} + \sigma z_{1j} \right) v_{jk}(F) \\ \vdots \\ \sum_{j=1}^{p} \left(\sum_{m=1}^{M} \lambda_m^{1/2} \eta_{nm} \theta_{jm} + \sigma z_{nj} \right) v_{jk}(F) \end{bmatrix} = \sum_{m=1}^{M} \lambda_m^{1/2} \theta_m^\top v_k(F) \begin{bmatrix} \eta_{1m} \\ \vdots \\ \eta_{nm} \end{bmatrix} + \sigma \begin{bmatrix} z_1^\top v_k(F) \\ \vdots \\ z_n^\top v_k(F) \end{bmatrix}. \tag{11}$$

Using the approximation: $v_k(F) = \theta_k + \delta_k$,

$$\eta_{im} \theta_m^\top v_k(F) = \eta_{im} \theta_m^\top (\theta_k + \delta_k) = \eta_{im} (\theta_m^\top \theta_k + \theta_m^\top \delta_k) = \begin{cases} \eta_{ik}(1 + \theta_k^\top \delta_k) & \text{if } k = m \\ \eta_{im}(\theta_m^\top \delta_k) & \text{if } k \neq m \end{cases} \tag{12}$$

i. Fix $k \neq m$:

$$\eta_{im} \lambda_m^{1/2} \theta_m^\top v_k(F) \lambda_k^{-1/2}(F) = \left(\frac{\lambda_m}{\lambda_k(F)} \right) \eta_{im} (\theta_m^\top \delta_k) \tag{13}$$

So, we need the ratio of eigenvalues to be bounded and then perhaps

$$|\theta_m^\top \delta_k| \leq \|\delta_k\|_2 = o(\text{some rate}). \tag{14}$$

ii. Fix $k = m$:

$$\eta_{ik} \lambda_k^{1/2} \theta_k^\top v_k(F) \lambda_k^{-1/2}(F) = \left(\frac{\lambda_k}{\lambda_k(F)} \right) \eta_{ik} (1 + \theta_k^\top \delta_k) \tag{15}$$

Now, we need the ratio of eigenvalues to go to one (implied by the perturbation bound?) and using the above bound in equation (14):

$$\left(\frac{\lambda_k}{\lambda_k(F)} \right) \eta_{ik} (1 + \theta_k^\top \delta_k) \to \eta_{ik} \tag{16}$$

(b) Combining (i) and (ii)

$$\sum_{m=1}^{M} \lambda_m^{1/2} \theta_m^\top v_k(F) \begin{bmatrix} \eta_{1m} \\ \vdots \\ \eta_{nm} \end{bmatrix} = \begin{bmatrix} \eta_{1k} \\ \vdots \\ \eta_{nk} \end{bmatrix} + o(\text{some other rate}) \tag{17}$$

(c) Lastly, we need to show that the measurement error term is bounded:

$$\sigma \begin{bmatrix} z_1^\top v_k(F) \\ \vdots \\ z_n^\top v_k(F) \end{bmatrix}.$$

This needs to be addressed with care as z and v are dependent.

3. We need to write down the form of the estimator: $\hat{U}_{\tilde{M}}^\top Y$. Plug in the regression model for Y (equation (2)):

$$\hat{\beta}_m = \hat{u}_m^\top Y = \beta_0 \hat{u}_m^\top \mathbf{1} + \sum_{m=1}^{\tilde{M}} \beta_m \hat{u}_m^\top \eta_m + \hat{u}_m^\top W = \text{(a)} + \text{(b)} + \text{(c)} \tag{18}$$

we need to write the regression model for Y in terms of these estimated coordinates:

(a) Maybe we can get rid of this via a max norm bound?

$$|\hat{u}_m^\top \mathbf{1}| \leq \|\hat{u}_m\|_1 \|\mathbf{1}\|_\infty = \|\hat{u}_m\|_1 \tag{19}$$

There should be something like a $n^{-1/2}$ running around. So, this would require that $\|\hat{u}_m\|_1 = o(n^{1/2})$, which isn't that likely.

(b) Apply the above results that show that $\hat{u}_m \approx \eta_m$ and hence

$$\beta_m \hat{u}_m^\top \eta_m \approx \beta_m \|\eta_m\|_2^2$$

So, if we have a n^{-1} floating around, then $n^{-1}\|\eta_m\|_2^2 \to 1$ and

$$\beta_m \|\eta_m\|_2^2 \to \beta_m.$$

(c) $\hat{u}_m$ and W are independent, so this can be shown to be small using a concentration bound (mean zero)