1 Darren's approach

Suppose $\mathbb{X} = UV^\top$. Then a plug-in PCR-based estimate of β is $\tilde{\beta}_d = V_d \Lambda_d^{-1} U^\top Y$. We can make a plug-in estimate (using the notation from Ding and McDonald) $\hat{\beta}_d = U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top \mathbb{X}^\top Y$. Also, write $U = [U_d\ U_{p-d}]$ to be the decomposition of U into the first d columns and last $p-d$ columns and write $\tilde{U}_d$ and $\tilde{U}_{p-d}$ to indicate padding with zeros to give the same dimension as U. Now, $U_d(F) \approx V_d$, $\Lambda_d(F)^{-1} \approx \Lambda^{-2}$, and $U_d(F)^\top V \approx [I_d\ 0]$. So, Ill proceed as usual to distill this expression down to these parts[1]

- I'm thinking it will be better to do this under the model in the paper. This means $U_{p-d} = 0 \Rightarrow R_d = 0$, but no harm in keeping it around for now.
- Rewriting from the beginning:

$$
\begin{aligned}
\left\|\hat{\beta}_d - \tilde{\beta}_d\right\| &= \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top \mathbb{X}^\top Y - V_d\Lambda_d^{-1}U^\top Y\right\| && (1)\\
&= \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top V\Lambda[U_d\ U_{p-d}]^\top Y - V_d\Lambda_d^{-1}U^\top Y\right\| && (2)\\
&= \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top V\Lambda(\tilde{U}_d^\top Y + \tilde{U}_{p-d}^\top Y - V_d\Lambda_d^{-1}U^\top Y\right\| && (3)\\
&\leq \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top V\Lambda\tilde{U}_d^\top Y\right\| && (4)\\
&\quad + \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top V\Lambda\tilde{U}_{p-d}^\top Y\right\| && (5)\\
&\leq \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top V_d\Lambda_d - V_d\Lambda_d^{-1}\right\| \left\|U_d^\top Y\right\| + R_d && (6)\\
&= \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top V_d\Lambda_d - V_d\Lambda_d^{-1}\right\| M_d + R_d \text{ (dropping } R_d \text{ now)} && (7)\\
&\leq \left\|U_d(F)\Lambda_d(F)^{-1}U_d(F)^\top V_d\Lambda_d - U_d(F)\Lambda_d(F)^{-1}\Lambda_d\right\| M_d && (8)\\
&\quad + \left\|U_d(F)\Lambda_d(F)^{-1}\Lambda_d - V_d\Lambda_d^{-1}\right\| M_d && (9)\\
&\leq \left\|U_d(F)\Lambda_d(F)^{-1}\right\| \left\|U_d(F)^\top V_d - I\right\| \|\Lambda_d\| M_d && (10)\\
&\quad + \left\|U_d(F)\Lambda_d(F)^{-1/2}\Lambda_d(F)^{-1/2}\Lambda_d - V_d\Lambda_d^{-1}\right\| M_d && (11)\\
&\leq \left\|U_d(F)\Lambda_d(F)^{-1}\right\| \left\|U_d(F)^\top V_d - I\right\| \|\Lambda_d\| M_d && (12)\\
&\quad + \left\|U_d(F)\Lambda_d(F)^{-1/2}\right\| \left\|\Lambda_d(F)^{-1/2}\Lambda_d - I\right\| M_d + \left\|U_d(F)\Lambda_d(F)^{-1/2} - V_d\Lambda_d^{-1}\right\| M_d && (13)
\end{aligned}
$$

- Is there a relationship between $\|\Lambda_d\|$ and $\|\Lambda_d(F)\|$? This would be nice.
- M_d seems like it will be a pain: $\Theta(n)$.

[1] Needing repeated constant 2 for $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$.

2 Another option

- My thinking (up to now) had been to mimic Paul, Bair, et. al:
  1. Show that $\|\sin(\mathcal{E}, \mathcal{F})\|$ is small where $\mathcal{E}$ is the span of V_d and $\mathcal{F}$ is the span of $U_d(F)$.
  2. Show that $\|\Lambda(F)_d - \Lambda_d\|$ is small.
  3. See whether this gives anything about $\hat{\beta}_d$.
- Start with the model:

$$\mathbb{X} = U_G \Lambda_G V_G^\top + \sigma_0 E \tag{14}$$
$$Y = U_K \Theta + \sigma_1 Z \tag{15}$$

  where $\mathbb{X} \in \mathbb{R}^{n \times p}$, $G \ll p$, $G < n$, $K \leq G$, $U_g \sim N(0, I_n)$, $E_{ij} \sim N(0,1)$, and $Z \sim N(0, I_n)$.
- An implication of this model is:

$$\Sigma_{xx} = \mathbb{E}\left[x_i x_i^\top\right] = V_G \Lambda_G^2 V_G^\top + \sigma_0^2 I_p. \tag{16}$$

- There are a few things we could do at this point, but I think we should make the following assumption: $\|V_G\|_{2,0} = \#\{\|V_{j,G}\|_2 \neq 0\} \leq p_* < n < p$. This is a "row sparsity" assumption as in Vu and Lei (2013). It also the corresponds to the set $\mathcal{D} = \{j : \|\Lambda_G V_{j,G}\|_2 \neq 0\}$ in Paul et al. (2008) via the inequality $\|\Lambda_G V_{j,G}\|_2 \leq \|V_{j,G}\|_2 \|\Lambda_G\|_2$. Essentially this means that only p_* variables actually provide information about $col(V_G)$.
- "Row sparsity" also matches how we generated data for the simulations. Actually, we used more than this: we set $V_{G,j} = 0$ for many j.
- Finally, we have that the selected variables in the marginal regression screening step are a subset of $\mathcal{D}$ and hence, we correctly recover some of the necessary variables.
- Now, the beginning of the idea:
  1. Set $\sigma_0 = 0$ and $\sigma_1 = 0$, the no-noise model.
  2. Suppose marginal regression recovers q of the p_* relevant variables.
  3. Can we characterize $\left\|U_K(F)^\top V_K - I\right\|$?
  4. For the first step, this would amount to examining a function of $V_K V_K^\top - U_K(F) U_K(F)^\top$. I was thinking with Lemma 4.2 or Corollary 4.1 in Lei and Vu's sparse PCA paper. Although, this again is just a different way of measuring the approximation accuracy of $U_K(F)$. So maybe this is already done?
- Can we restate Darren's version in terms of this model?

3 Thoughts

My thoughts on the target journal here is JCGS. To that end, I think we need some or all of the following:

1. Minor theoretical contributions along the lines above. Get as far as we can before it gets painful, likely under strong assumptions.

2. Do the Nystrom version as well. (Already done in simulations, it's a bit worse, though not terrible)
3. Implement GLMs.

References

Paul, D., Bair, E., Hastie, T., and Tibshirani, R. (2008), "'Preconditioning' for feature selection and regression in high-dimensional problems," *The Annals of Statistics*, **36**(4), 1595–1618.

Vu, V. Q., and Lei, J. (2013), "Minimax sparse principal subspace estimation in high dimensions," *Annals of Statistics*, **41**, 2905–2947.