Imperial College Business School

London Business School

Sparse PCA with Multiple Components

Ryan Cory-Wright, Jean Pauphilet

Principal Component Analysis (PCA)

Idea

For a high-dimensional dataset X, find a few orthogonal directions (r) that explain most of the variance

Algorithm

Dataset $X \in \mathbb{R}^{n \times p}$

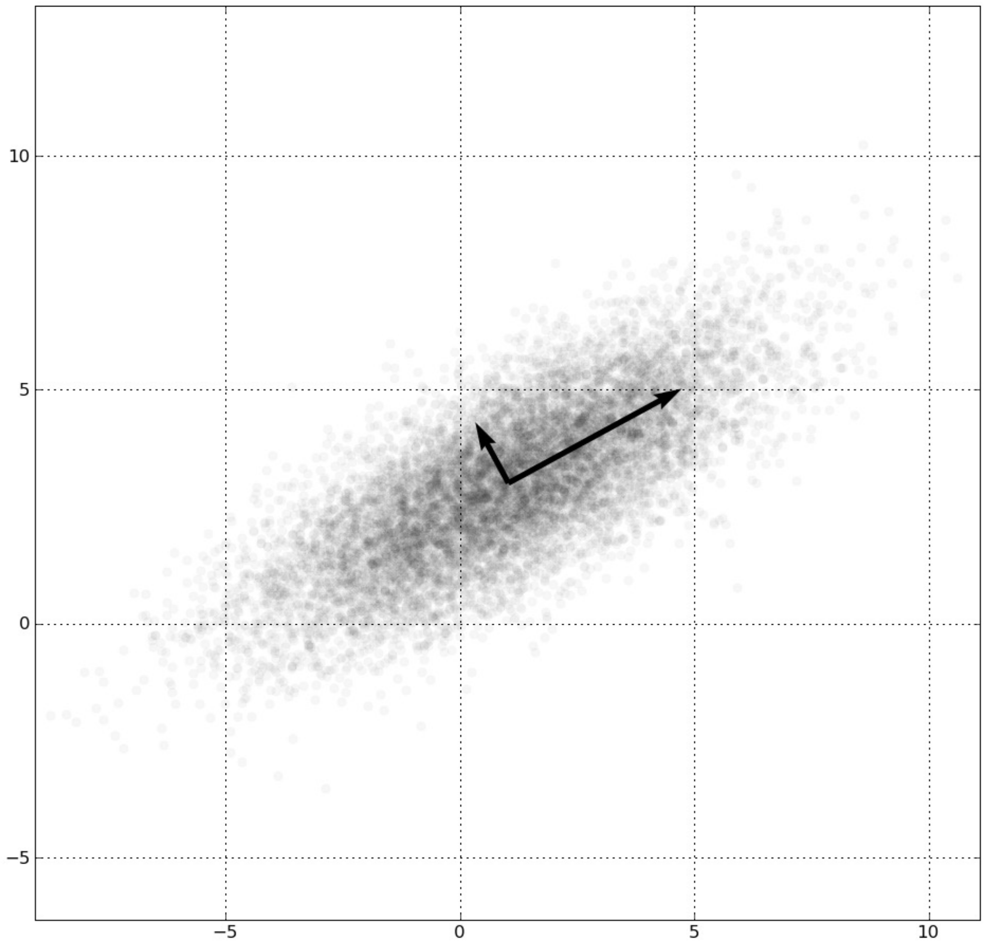
Compute its covariance/correlation matrix $\boldsymbol{\Sigma}$

Solve

$$\max_{\boldsymbol{U} \in \mathbb{R}^{p \times r}} \quad \langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{\Sigma} \rangle \text{ s.t. } \boldsymbol{U}^\top \boldsymbol{U} = \mathbb{I}$$

Compress dataset into smaller one

$$X_{\text{comp}} \coloneqq U_{[1:r]} X \in \mathbb{R}^{n \times r}$$



Optimization for PCA

$$\max_{\boldsymbol{U} \in \mathbb{R}^{p \times r}} \quad \langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{\Sigma} \rangle \text{ s.t. } \boldsymbol{U}^\top \boldsymbol{U} = \mathbb{I}$$

It looks hard…

- Maximizing a convex function
- Non-convex quadratic constraints (orthogonality)

… can be solved very efficiently

- Solution obtained via truncated SVD
- For r=1, orthogonality constraint is trivial to satisfy (scaling)
- For r > 1, greedy is optimal
  i.e., solve for r=1, deflate the covariance matrix, repeat
  Deflation naturally takes care of orthogonality!

Why sparsity in PCA?

- **Interpretability**
  New coordinates (PCs) can be dense combination of all features

- **Consistency** in high-dimensional settings
  When $p/n \to \alpha$, PCA can be inconsistent

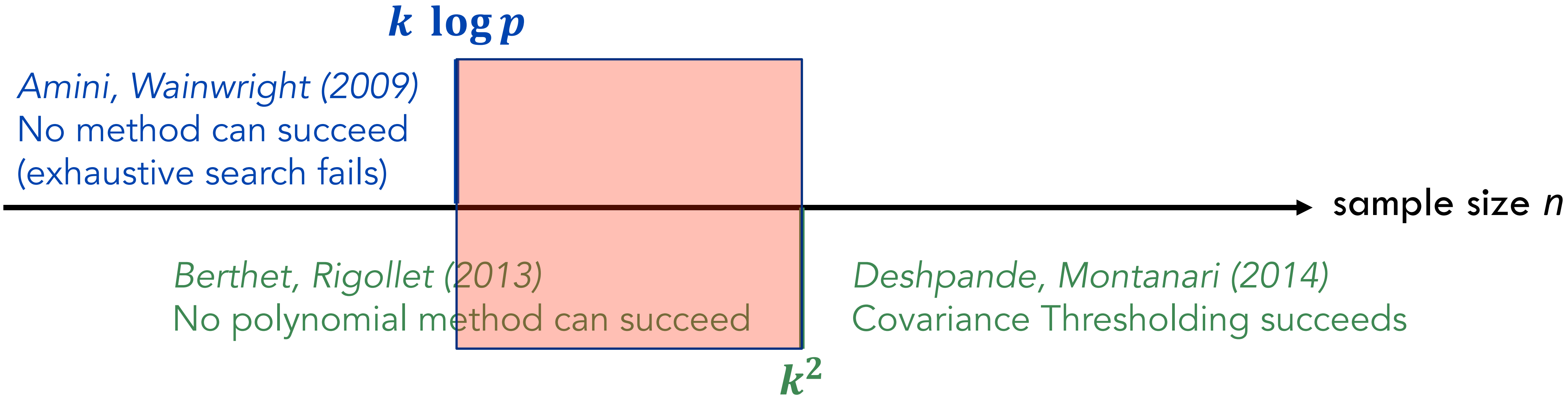
Solution Sparsity

$$\max_{\boldsymbol{u} \in \mathbb{R}^p} \quad \langle \boldsymbol{\Sigma}, \boldsymbol{u}\boldsymbol{u}^\top \rangle \text{ s.t. } \|\boldsymbol{u}\|_2^2 = 1, \|\boldsymbol{u}\|_0 \leq k$$

PCA requires in the order of $n \gtrsim p$ samples

sparse PCA requires $n \gtrsim k \log p$

Why “provably optimal” algorithms can help?



Opportunity for optimization!

From r=1 to r>1

Jungle of algorithms for sparse PCA with r=1 PC

However, for r > 1, …

- Deflation approach is no longer optimal
- Deflation no longer guarantees orthogonality (feasibility)
- Assuming perfect support detection, estimation is no longer trivial

Sparse PCA With Multiple Components

Overall problem Explain dataset using sparse and mutually orthogonal components

$$\max_{\boldsymbol{U} \in \mathbb{R}^{p \times r}} \quad \langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{\Sigma} \rangle \quad \text{s.t.} \quad \boldsymbol{U}^\top \boldsymbol{U} = \mathbb{I}, \ \|\boldsymbol{U}\|_0 \leq k.$$

OR

$$\|\boldsymbol{U}_t\|_0 \leq k_t, \ \forall t \in [r]$$

Our Contributions

1. Reformulation as as rank and sparsity constrained problem
2. Tight and scalable semidefinite relaxation w. strong valid inequalities
3. Coupled w. good heuristics, provably near-optimal solutions for *$p \approx 100$*

Reformulation

$$\max_{\boldsymbol{U}\in\mathbb{R}^{p\times r}} \quad \langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{\Sigma}\rangle \quad \text{s.t.} \quad \boldsymbol{U}^\top\boldsymbol{U} = \mathbb{I},\ \|\boldsymbol{U}\|_0 \le k.$$

$$\max_{\substack{\boldsymbol{Z}\in\{0,1\}^{p\times r}\\ \langle \boldsymbol{E},\boldsymbol{Z}\rangle\le k}} \max_{\boldsymbol{U}\in\mathbb{R}^{p\times r}} \quad \boxed{\langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{\Sigma}\rangle} \quad \text{s.t.} \quad \boxed{\boldsymbol{U}^\top\boldsymbol{U} = \mathbb{I},} \ \boxed{U_{i,t} = 0 \text{ if } Z_{i,t} = 0,}\ \forall i \in [p],\ \forall t \in [r],$$

Introduce

$$\boldsymbol{Y}^t := \boldsymbol{U}_t\boldsymbol{U}_t^\top, \quad \boldsymbol{Y} := \sum_{t\in[r]} \boldsymbol{Y}^t = \boldsymbol{U}\boldsymbol{U}^\top$$

$$\boxed{\langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{\Sigma}\rangle \to \langle \boldsymbol{Y}, \boldsymbol{\Sigma}\rangle}$$

$$\boxed{\boldsymbol{U}^\top\boldsymbol{U} = \mathbb{I} \to \operatorname{tr}(\boldsymbol{Y}^t) = 1,\ \langle \boldsymbol{Y}^t, \boldsymbol{Y}^{t'}\rangle = 0\,(t \ne t')}$$

$$\boxed{U_{i,t} = 0 \text{ if } Z_{i,t} = 0 \to Y^t_{i,j} = 0 \text{ if } Z_{i,t} = 0}$$

$$\max_{\substack{\boldsymbol{Z}\in\{0,1\}^{p\times r}:\\ \langle \boldsymbol{E},\boldsymbol{Z}\rangle\le k}} \max_{\boldsymbol{Y}\in\mathcal{S}^p, \boldsymbol{Y}^t\in\mathcal{S}^p_+} \boxed{\langle \boldsymbol{Y}, \boldsymbol{\Sigma}\rangle} \text{ s.t. } \boxed{\operatorname{tr}(\boldsymbol{Y}^t) = 1,\ \forall t \in [r], \quad \langle \boldsymbol{Y}^t, \boldsymbol{Y}^{t'}\rangle = 0, \forall t, t' \in [r], t \ne t',}$$

$$\boxed{Y^t_{i,j} = 0 \text{ if } Z_{i,t} = 0,\ \forall t \in [r], i, j \in [p]},$$

$$\boldsymbol{Y} = \sum_{t=1}^{r} \boldsymbol{Y}^t,\ \operatorname{Rank}(\boldsymbol{Y}^t) = 1,\ \forall t \in [r].$$

Reformulation

$$\max_{\substack{\boldsymbol{Z}\in\{0,1\}^{p\times r}:\\ \langle \boldsymbol{E},\boldsymbol{Z}\rangle\leq k}} \max_{\boldsymbol{Y}\in\mathcal{S}^p, \boldsymbol{Y}^t\in\mathcal{S}^p_+} \langle \boldsymbol{Y},\boldsymbol{\Sigma}\rangle \text{ s.t. } \operatorname{tr}(\boldsymbol{Y}^t)=1,\ \forall t\in[r],\quad \boxed{\boldsymbol{Y}=\sum_{t\in[r]}\boldsymbol{Y}^t\preceq\mathbb{I}}$$

$$Y^t_{i,j}=0 \text{ if } Z_{i,t}=0,\ \forall t\in[r], i,j\in[p],$$

$$\boldsymbol{Y}=\sum_{t=1}^{r}\boldsymbol{Y}^t,\ \operatorname{Rank}(\boldsymbol{Y}^t)=1,\ \forall t\in[r].$$

PROPOSITION 1. *Consider r matrices, $\boldsymbol{Y}^t\in\mathcal{S}^p_+$, such that* $\operatorname{tr}(\boldsymbol{Y}^t)=1$ *and* $\operatorname{Rank}(\boldsymbol{Y}^t)=1$. *Then,* $\sum_{t\in[r]}\boldsymbol{Y}^t\preceq\mathbb{I}$ *if and only if* $\langle\boldsymbol{Y}^t,\boldsymbol{Y}^{t'}\rangle=0\ \forall t,t'\in[r]:\ t\neq t'$.

Sparse and low-rank reformulation

$$\max_{\substack{\boldsymbol{Z}\in\{0,1\}^{p\times r}:\\ \langle \boldsymbol{E},\boldsymbol{Z}\rangle\le k}} \ \max_{\boldsymbol{Y}\in\mathcal{S}^p,\boldsymbol{Y}^t\in\mathcal{S}^p_+} \langle \boldsymbol{Y},\boldsymbol{\Sigma}\rangle \ s.t. \quad \boldsymbol{Y}\preceq \mathrm{Diag}\left(\min\left(\boldsymbol{e},\sum_t \boldsymbol{Z}_t\right)\right), \boldsymbol{Y}=\sum_{t=1}^r \boldsymbol{Y}^t,$$

$$\mathrm{tr}(\boldsymbol{Y}^t)=1,\ \forall t\in[r],\ Y^t_{i,j}=0 \ \textit{if}\ Z_{i,t}=0\ \forall t\in[r], i,j\in[p],$$

$$\mathrm{Rank}(\boldsymbol{Y}^t)=1,\ \forall t\in[r].$$

Next steps (in the paper)

- Relaxation: SDP/SOC for rank, big-M for logical constraints
- Strengthened formulation via symmetry-breaking inequalities
- Strengthened formulation via valid inequalities linking $\boldsymbol{Y}^t$ and $\boldsymbol{Z}_t$
- Relax-round-then-estimate heuristics

Valid Inequalities via Individual Sparsity Budgets

Suppose t^{th} component k_t sparse with $\sum_t k_t =: k$. Impose

$$\left(\sum_{j=1}^{p} |Y^t_{i,j}|\right)^2 \leq k_t Y^t_{i,i} Z_{i,t} \quad \forall i \in [p], \forall t \in [r]$$

$$\sum_{i \in [p]: i \neq j} {Y^t_{i,j}}^2 \leq (k_t - 1) Z_{j,t}(Z_{j,t} - Y^t_{j,j}) \quad \forall j \in [p].$$

→ Tightens relaxation substantially but requires knowledge of k_t's

Feasible Solutions via Alternating Minimization

A Lagrangean Relaxation of Sparse PCA

$$\max_{\substack{\boldsymbol{Z}\in\{0,1\}^{p\times r}:\\ \langle \boldsymbol{e},\boldsymbol{Z}_t\rangle\le k_t}} \ \max_{\boldsymbol{Y}\in\mathcal{S}^p_+,\boldsymbol{Y}^t\in\mathcal{S}^p_+} \ \sum_{t\in[r]}\langle \boldsymbol{Y}^t,\boldsymbol{\Sigma}\rangle-\lambda\sum_{t,t'\in[r]:t\neq t'}\langle \boldsymbol{Y}^t,\boldsymbol{Y}^{t'}\rangle \ \text{s.t.} \quad \mathrm{tr}(\boldsymbol{Y}^t)=1,\ \forall t\in[r],$$

$$Y^t_{i,j}=0 \text{ if } Z_{i,t}=0,\ \forall t\in[r],i,j\in[p],$$

$$\mathrm{Rank}(\boldsymbol{Y}^t)=1,\ \forall t\in[r].$$

$\lambda>0$ a penalty parameter

Fix all but one PC → Single component sparse PCA

Alternating minimization → High-quality solutions

Code available on GitHub: ryancorywright/MultipleComponentsSoftware

Application I: pitprops dataset (p=13), r=6 PCs

Lu and Zhang, MP (2011)

"we deduce that for the Pitprops data, it seems not possible to extract six highly sparse (e.g., around 60 zero loadings), nearly orthogonal and uncorrelated PCs while explaining most of variance as they may not exist."

r	k_t	Alg. 1			Alg. 2			Branch-and-Bound			
		Obj.	Viol.	T(s)	Obj.	Viol.	T(s)	Obj.	Viol.	Nodes	T(s)
6	2	**0.749**	0	1.01	0.734	0	23.31	0.740	0.001	65600	>600
	4	0.666	0	0.90	**0.814**	0.085	287.34	0.718	0.002	39200	>600
	6	0.686	0	1.54	0.836	0.153	336.99	**0.723**	0.003	39600	>600
	8	0.673	0	1.89	0.854	0.122	293.38	**0.745**	0.003	45700	>600
	10	0.645	0	1.34	0.874	0.115	206.93	**0.790**	0.004	25200	>600
ALSPCA-2			60		0.03			0.084			39.42
ALSPCA-3			63		0.00			0.222			65.97

Pitprops data

Application II: Performance on UCI Datasets

- Fix k=15,30, r=3
- Compute worst bound, best AM solution over rank-3 allocations of sparsity budget

Dataset	p	r	k	Enumerated			
			k	k_t	Obj.	Rel. gap (%)	Viol.
Pitprops	13	3	15	(7, 6, 2)	0.595	3.67%	0.002
			30	(10, 10, 10)	0.650	0.28%	0.005
Ionosphere	34	3	15	(7, 6, 2)	0.299	0%	0
			30	(11, 10, 9)	0.402	2.35%	0
Geographical	68	3	15	(6, 5, 4)	0.221	0%	0
			30	(12, 12, 6)	0.420	0%	0
Communities	101	3	15	(6, 5, 4)	0.142	0.01%	0
			30	(11, 11, 8)	0.247	0.17%	0

The average semidefinite relaxation gap is less than 1%!

Application III: US Senate Voting Patterns 2021-2022

Problem setting

Explain voting patterns in 117th senate

- n=100 data points (senators)
- p=839 features (bills/amendments/noms…)
    - Screened out p=17 unanimous votes
- Compute two 5-sparse PCs via SOC relax+greedy rounding, score senators

Results

First PC

- Initial vote on Inflation Reduction Act (IRA)
- Another vote on IRA
- Lower cost of Insulin
- Another vote on IRA
- Amendment to IRA

The July 2022 "Inflation Reduction Act"
All five votes on same day!

Second PC

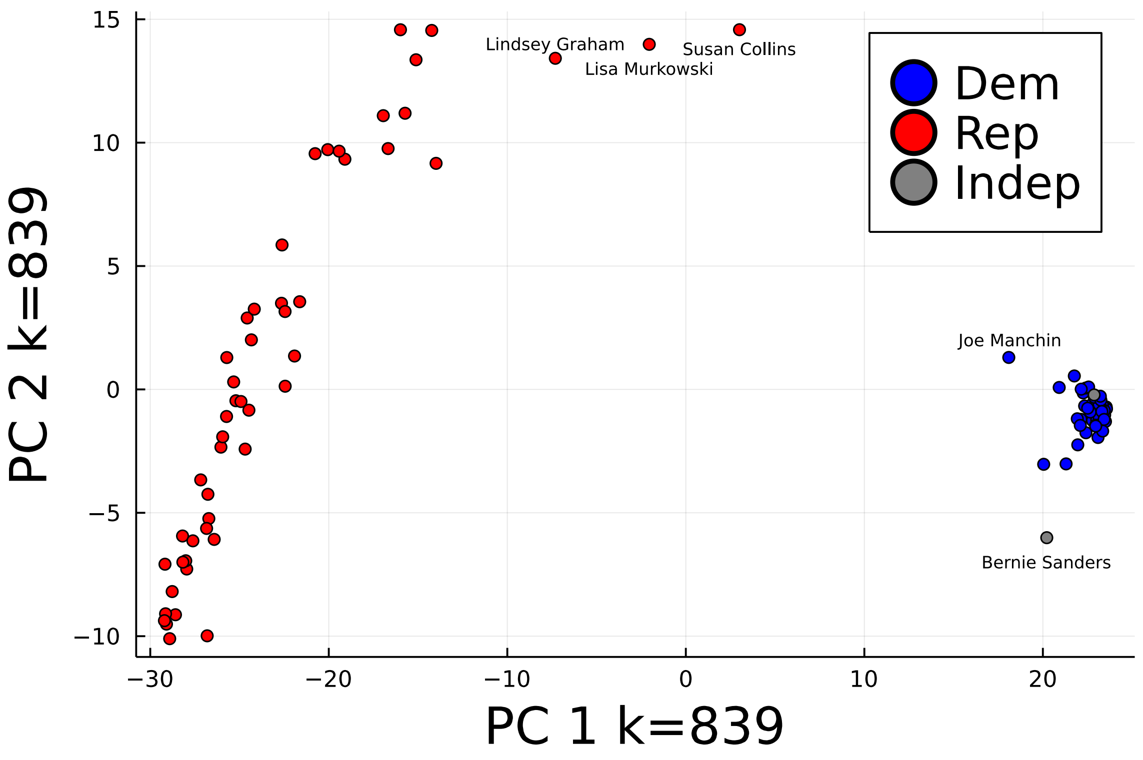
- Revenues from leasing oil/gas on federal land
- Extend Trump tax cuts
- Deficit neutral fund for catch-and-release
- Update text of 2021-22 budget
- Pass 2021-22 budget

The Feb 2021 "American Rescue Plan" and 3 proposed amendments
All five votes on same day!

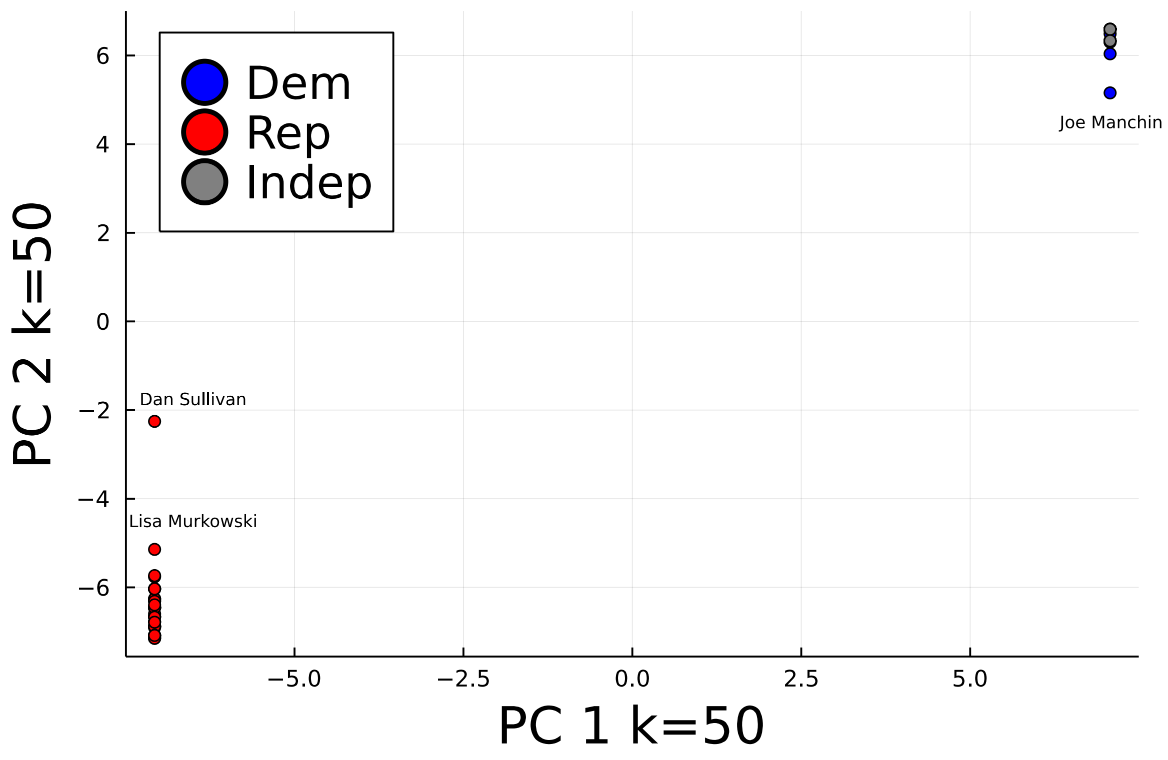
Both PCs perfect classifiers by party!
Plots not vey interesting, so increase k

Application III: US Senate Voting Patterns 2021-2022

Results no sparsity



Results k=50



Conclusion

Sparse PCA with r>1 PCs remains a largely open problem

Sparse PCA with r>1 PCs is significantly more challenging than the r=1 case

Orthogonality → rank constraints

Semidefinite Relaxation + Alternating Minimization Solves Sparse PCA to (Near) Optimality

- Bound gaps of 1%-5% in practice, depending on application
- Key is to use good valid inequalities to tighten formulations

Thank you!