Convex Formulations for Fair Principal Component Analysis

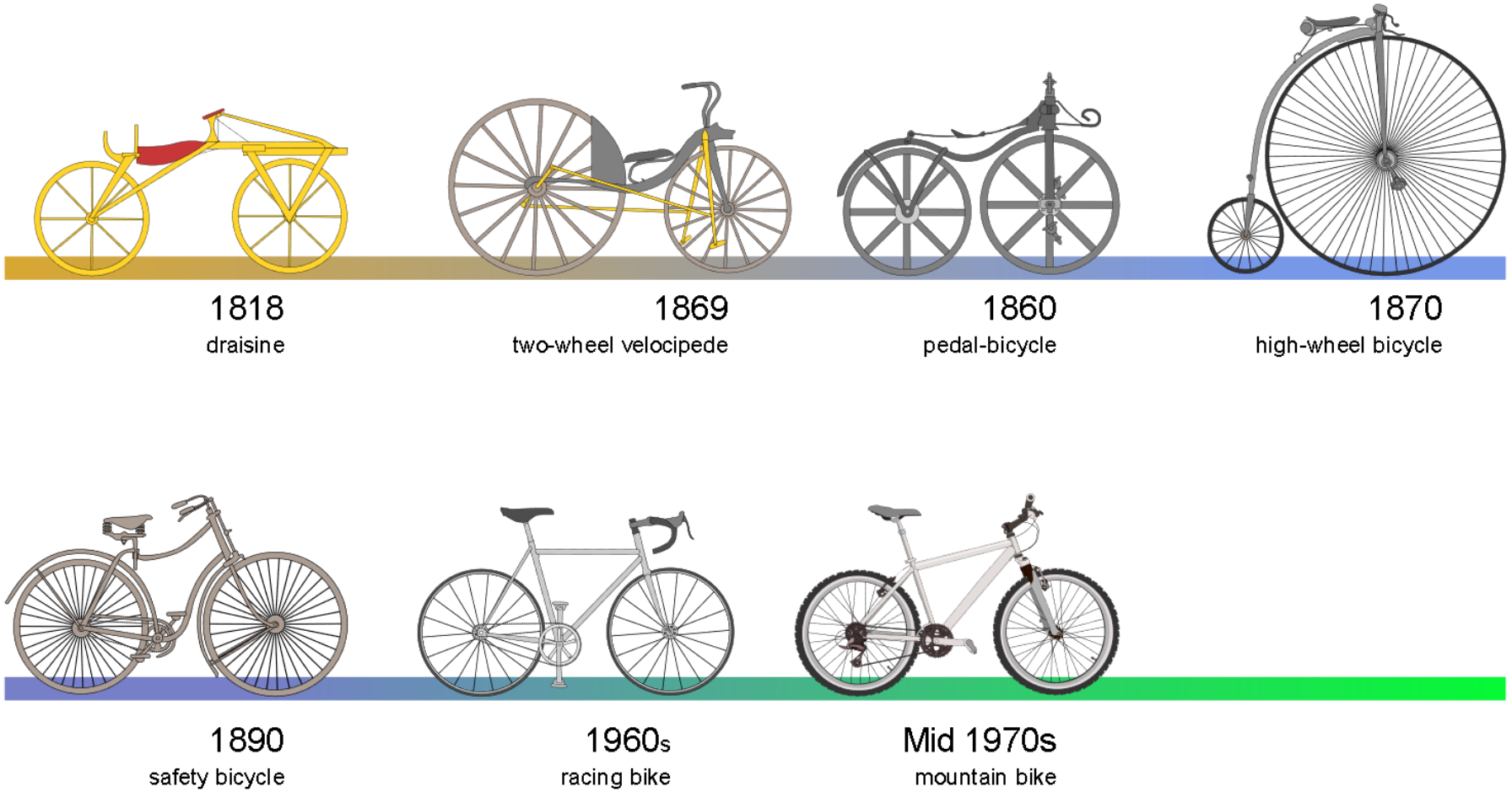
Matt Olfat & Anil Aswani

11/13/2018

https://arxiv.org/pdf/1802.03765.pdf

BERKELEY
IEOR
INDUSTRIAL ENGINEERING
& OPERATIONS RESEARCH

Black Box vs. Socio-technical Systems

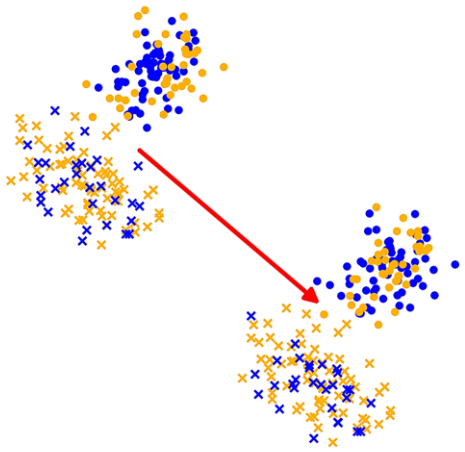


- Technological change driven by social needs
  - Steering
  - Pedals
  - Gears
- Machine learning also at a crossroads
  - Robustness
  - Interpretability
  - Fairness

Potential for Codifying Bias

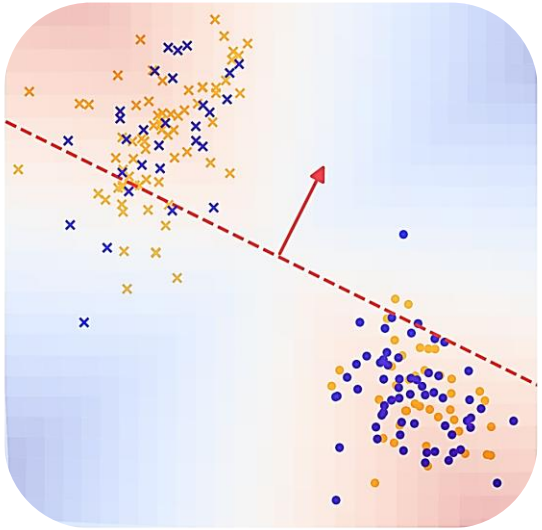
- Societal biases make their way into data
  - Women undertreated for pain (NYT, 2013)
  - Racial underrepresentation in clinical drug trials (ProPublica, 2018)
- "Fairness through ignorance" insufficient
  - Misleading correlations in data misinterpreted as signal
  - Systematized bias further corrupts data…
- Anecdotal examples
  - Biased algorithms to predict recidivism (Propublica, 2016)
  - Gender bias in LinkedIn job ads (The Seattle Times, 2016)
  - Facebook censorship of hate speech (Propublica, 2017)
  - Removing fake news with certifiable nonpartisanship

Prior Approaches



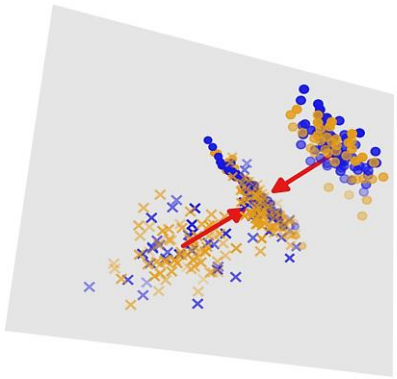
Relabeling

- Pedreschi et al (2008)
- Kamiran & Calders (2009)
- Luong et al (2011)
- Hardt et al (2016)



Regularization

- Calders & Verwer (2010)
- Kamishima et al (2011)
- Zliobaite et al (2015)
- Zafar et al (2017)



Feature Modification

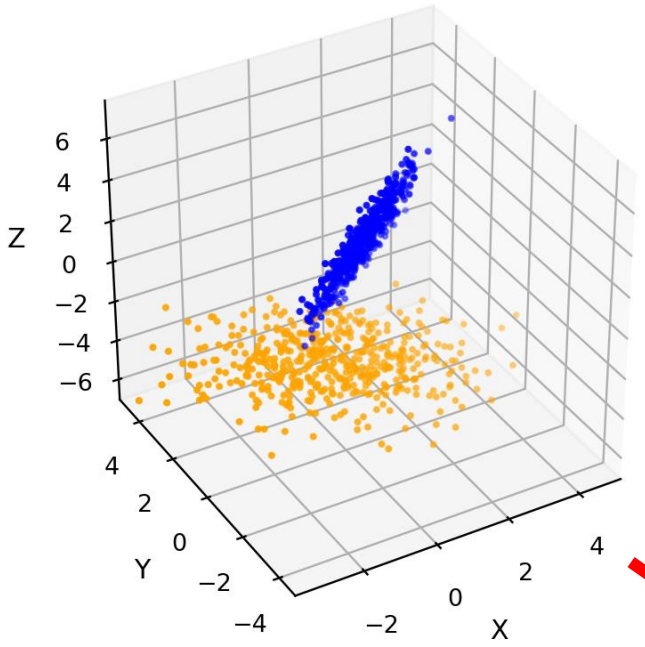
- Calders et al (2011)
- Dwork et al (2011,2012)
- Zemel et al (2014)
- Calmon et al (2017)

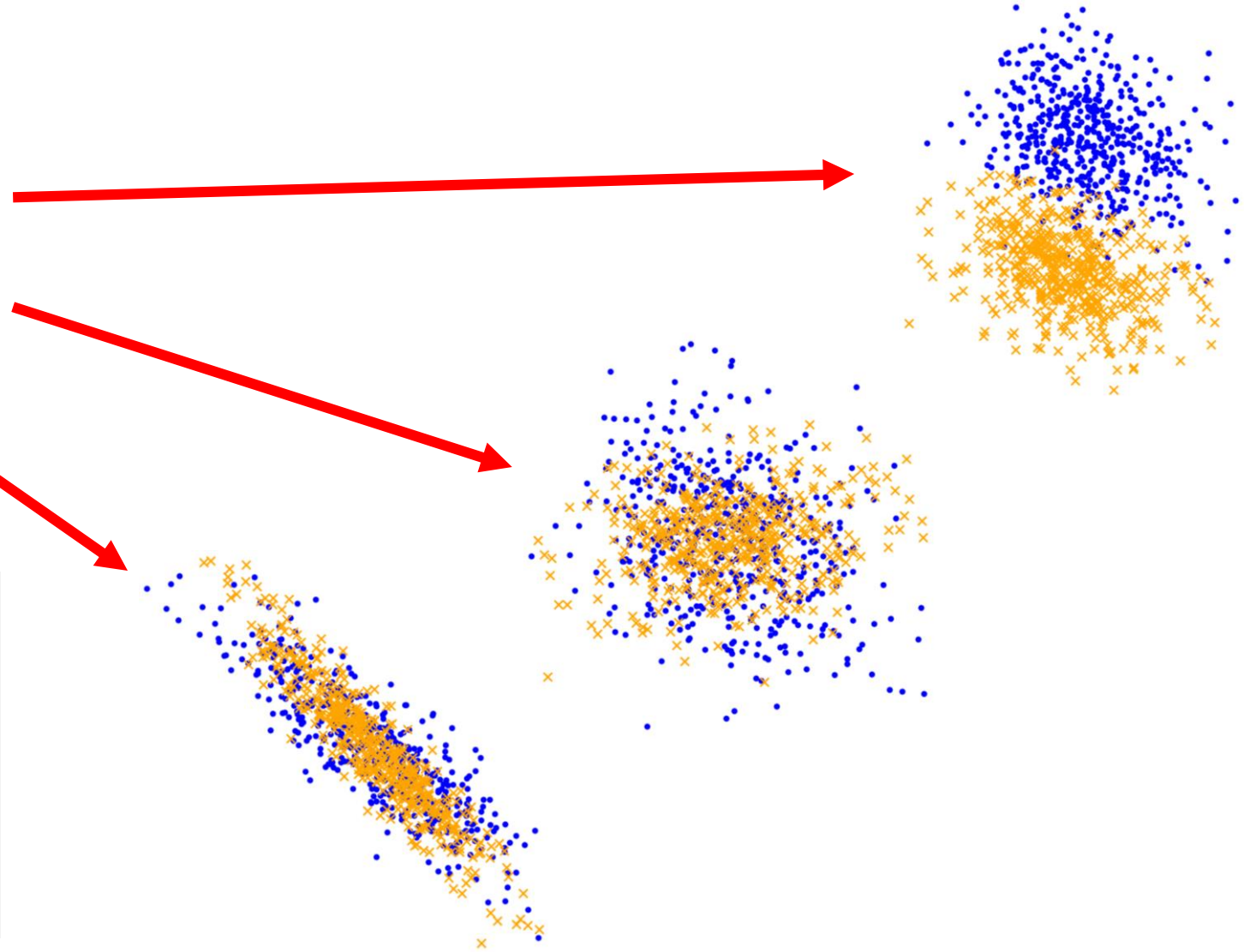
- Chierichetti et al (2017) approximate NP-hard preprocessing step, but specific to clustering

Relevance of Fairness to PCA

- Difficulty: no ordered output to map to good/bad outcomes
  - i.e. lending, censorship, recidivism, etc.
  - So what does fairness mean
- Dimensionality reduction as preprocessing step
  - Commonly used to avoid curse of dimensionality
  - Flexible to choice of algorithm, unsupervised algorithms not easy to modify
  - Recent applications in natural language processing
- Also interesting for gaining insights
  - PCA common technique for data visualization

Fairness?





Can we find a projection to a lower-dimensional space that mixes the protected class?

Fairness Notion

- Say that a mapping Π is $\Delta(h)$-fair for some classifier $h: \mathbb{R}^d \times \mathbb{R} \to \{-1, +1\}$ if the following holds $\forall t \in \mathbb{R}$:

$$\left| \mathbb{P}\left[h(\Pi(x_i), t) = +1 \middle| z = +1\right] - \mathbb{P}\left[h(\Pi(x_i), t) = +1 \middle| z = -1\right] \right| \leq \Delta(h)$$

True-positive rate

False-positive rate

- May approximate above with empirical estimate:

$$\left| \frac{1}{\#P} \sum_{i \in P} \mathbf{1}\big(h(\Pi(x), t) = +1\big) - \frac{1}{\#N} \sum_{i \in N} \mathbf{1}\big(h(\Pi(x), t) = +1\big) \right|$$

  - Say Π is *empirically* $\Delta(h)$-fair

- For class of classifiers $\mathcal{F}$, say Π is $\Delta(\mathcal{F})$-fair if $\Delta(h)$-fair $\forall h \in \mathcal{F}$

Non-asymptotic convergence:

Let $\mathcal{F}$ be the set of linear classifiers, and let Π be an emp. $\Delta_{\text{emp}}(\mathcal{F})$-fair mapping. Then Π is $\Delta(\mathcal{F})$-fair with probability $\geq 1 - \exp(-\frac{n\delta^2}{2})$, where:

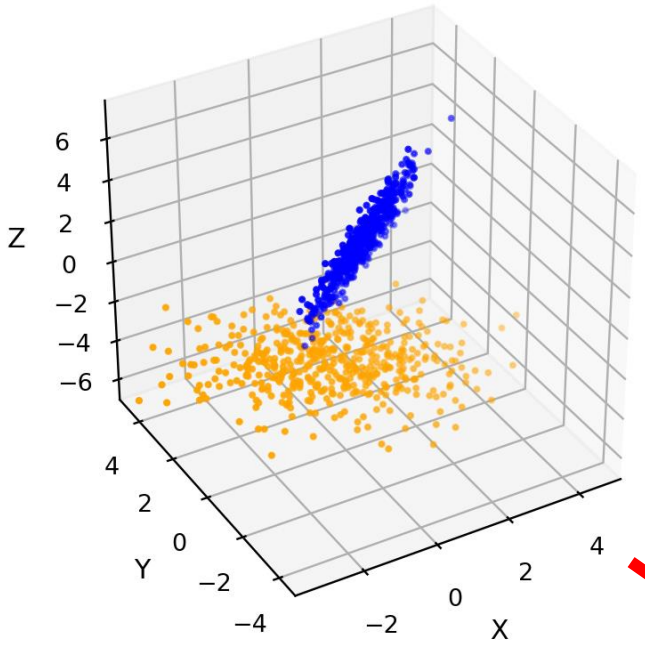
$$\Delta(\mathcal{F}) = \Delta_{\text{emp}}(\mathcal{F}) + 8\sqrt{\frac{(d+1)}{n}} + \delta$$

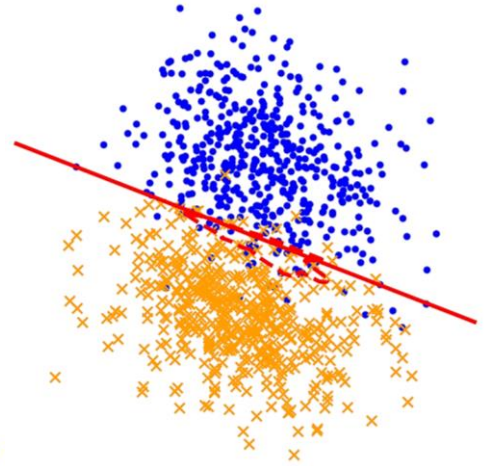
– Robust to adversarial user

– But infinite-dimensional

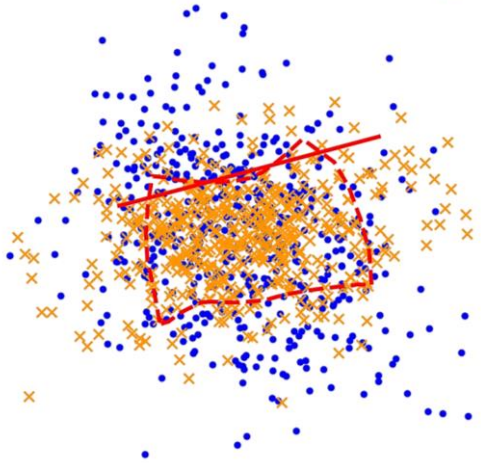
– Nonconvex even for fixed t

Fairness!

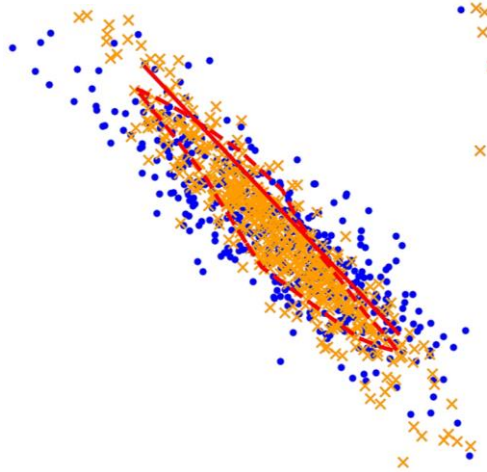




$\widehat{\Delta}(\text{Linear SVM}) = 0.904$
$\widehat{\Delta}(\text{Gaussian SVM}) = 0.906$



$\widehat{\Delta}(\text{Linear SVM}) = 0.172$
$\widehat{\Delta}(\text{Gaussian SVM}) = 0.346$



$\widehat{\Delta}(\text{Linear SVM}) = 0.076$
$\widehat{\Delta}(\text{Gaussian SVM}) = 0.316$

Depending on sophistication of $\mathcal{F}$, projection can ensure varying levels of fairness

SDP Formulation

- Want to solve the following:
  - Usually solved via power iterations and deflation

$$\max_{v_i \in \mathbb{R}^p} \sum_{i=1}^{d} v_i^\mathsf{T} X^\mathsf{T} X v_i = \max_{S \in \mathbb{R}^{p \times d}} S^\mathsf{T} X^\mathsf{T} X S = \max_{S \in \mathbb{R}^{p \times d}} \langle X^\mathsf{T} X, S S^\mathsf{T} \rangle$$

- Can tractably write as SDP
  - Extending d'Aspremont, El Ghaoui, Jordan & Lanckriet (2005)
  - Will return optimal solution if eigenvalues of $X^T X$ separable

$$\begin{array}{lll} \max & \langle X^\mathsf{T} X, D \rangle & \\ \text{s.t.} & \text{rank}(D) & = \quad d \\ & D & \succeq \quad 0 \end{array} \implies \begin{array}{lll} \max & \langle X^\mathsf{T} X, D \rangle & \\ \text{s.t.} & \text{trace}(D) & = \quad d \\ & I - D & \succeq \quad 0 \\ & D & \succeq \quad 0 \end{array}$$

Benefits:

- Non-iterative, allows for constraining eigenvectors
- Avoids deflation, can use non-greedy approach
- Easily modeled using existing solvers

Drawbacks:

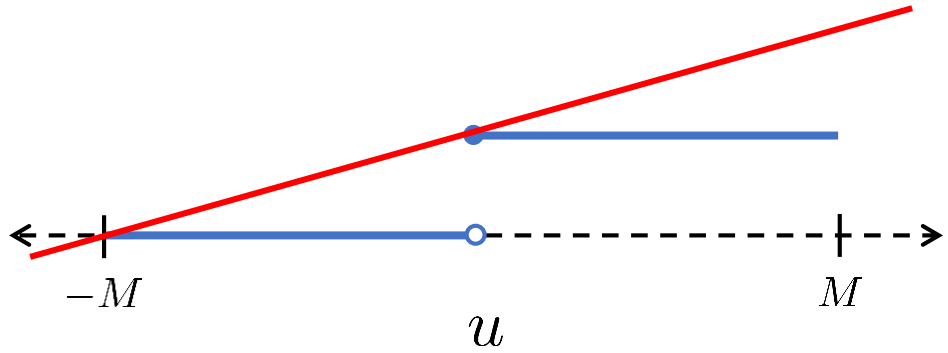
- Still have to get eigenvectors of D
- Less stable for eigenvectors with smaller eigenvalues

Convex relaxation

- Recall, need to bound the following

$$\max_{t\in\mathbb{R}} \left| \frac{1}{\#P} \sum_{i\in P} \mathbf{1}\big(h(\Pi(x),t) = +1\big) - \frac{1}{\#N} \sum_{i\in N} \mathbf{1}\big(h(\Pi(x),t) = +1\big) \right|$$

  - Difficult to control
  - Proceed with assumption that h is a margin classifier
- Upper-bounding sign functions simplifies problem
  - Assume $-M \le u \le M$
  - $\mathbf{1}\{\text{sign}(u) = +1\} \le 1 + \frac{u}{M}$
  - $\mathbf{1}\{\text{sign}(u) = -1\} \le -\frac{u}{M}$



Initial Convex Relaxation

- Convex upper bounding procedure gives approximation
  - For some fixed linear classifier $h_\beta(x,t) = \text{sign}(\beta^T x + t)$

$$-\delta \le \beta^\mathsf{T}\left(\frac{1}{\#P}\sum_{i\in P}\Pi(x_i) - \frac{1}{\#N}\sum_{i\in N}\Pi(x_i)\right) \le \delta$$

$$\implies -\delta \le \beta^\mathsf{T} S^\mathsf{T}\left(\frac{1}{\#P}\sum_{i\in P} x_i - \frac{1}{\#N}\sum_{i\in N} x_i\right) \le \delta$$

- To bound for all h_β, take the maximum
  - Let $f = \frac{1}{\#P}\boldsymbol{e}^T X_+ - \frac{1}{\#N}\boldsymbol{e}^T X_-$

$$\max_\beta \frac{\left|\beta^\mathsf{T} S^\mathsf{T} f\right|}{\|\beta\|_2} = \|S^\mathsf{T} f\|_2 \implies f^\mathsf{T} S S^\mathsf{T} f \le \delta^2$$

Benefits:

- Convenient interpretation as mean-matching constraint over all basis vectors of range of Π
- Easily included in SDP framework
- Also guaranteed to have rank d optimal solution

$$\begin{array}{lrcl} \max & \langle X^\mathsf{T} X, D\rangle & & \\ \text{s.t.} & \text{trace}(D) & = & d \\ & \langle f f^T, D\rangle & \le & \delta^2 \\ & I - D & \succeq & 0 \\ & D & \succeq & 0 \end{array}$$

Covariance Bound

- Can likely improve fairness by covariance matching
  - Pinsker's inequality → exact fairness for Gaussian data
  - Want $\mathrm{Var}[\beta^T \Pi(x_i)|z_i = +1] \approx \mathrm{Var}[\beta^T \Pi(x_i)|z_i = -1]$

$$\begin{aligned}
&\mathrm{Var}\left[\beta^\mathsf{T}\Pi(x)\middle|z=+1\right] \\
&\quad = \beta^\mathsf{T}\Big(\mathbb{E}\left[\Pi(x)\Pi(x)^\mathsf{T}\middle|z=+1\right]\Big)\beta \\
&\quad = \beta^\mathsf{T} S^\mathsf{T}\Sigma_+ S\beta
\end{aligned}$$

- Can generalize by bounding the following:
  - Choose $\varepsilon > \|\Sigma_+ - \Sigma_-\|_2$

$$\max_{\|\beta\|_2=1} \beta^\mathsf{T} SS^\mathsf{T}(\Sigma_+ - \Sigma_-)SS^\mathsf{T}\beta$$

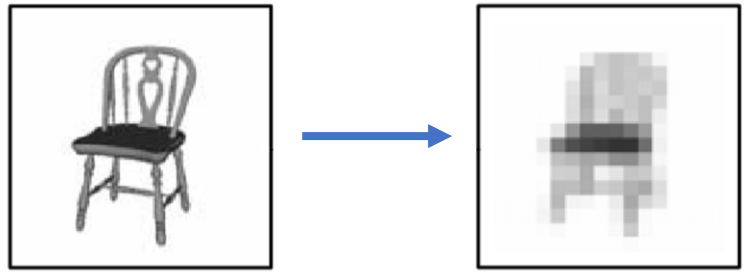
$$= \max\left\{\|SS^\mathsf{T}(\Sigma_+ - \Sigma_- + \varepsilon I)SS^\mathsf{T}\|_2, \|SS^\mathsf{T}(\Sigma_- - \Sigma_+ + \varepsilon I)SS^\mathsf{T}\|_2\right\} - \varepsilon$$

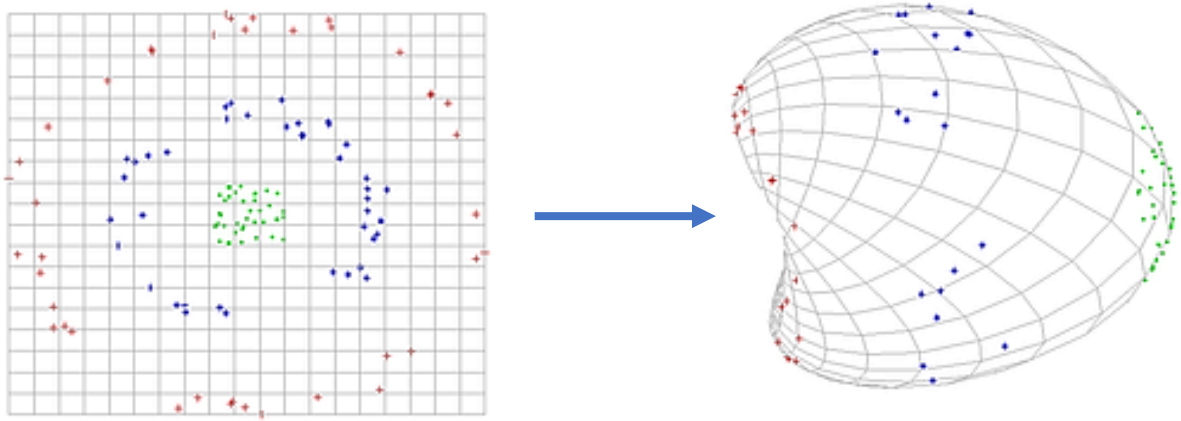
- Using Schur complement, can also represent as SDP constraint
- Let $M_i M_i^T = i(\Sigma_+ - \Sigma_-) + \varepsilon I$

$$\begin{aligned}
\max \quad & \langle X^T X, D\rangle - \mu t & & \\
\text{s.t.} \quad & \mathrm{Tr}(D) & = & d \\
& \langle ff^T, D\rangle & \leq & \delta^2 \\
& \begin{bmatrix} tI & DM_{+1} \\ M_{+1}^\mathsf{T} D & I \end{bmatrix} & \succeq & 0 \\
& \begin{bmatrix} tI & DM_{-1} \\ M_{-1}^\mathsf{T} D & I \end{bmatrix} & \succeq & 0 \\
& I - D & \succeq & 0 \\
& D & \succeq & 0
\end{aligned}$$

Extension to Kernel PCA

- Given transformation $\phi: \mathbb{R}^p \to \mathbb{R}^m$ and associated kernel $K: \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that $K(x, x') = \phi(x)^T \phi(x)$, want principal components in transformed space
  - Recovered as $v = \sum_{i=1}^{n} a_i \phi(x_i)$ such that $n\lambda a = K(X, X)a$
- Fit easily into our framework:
  - $X^T X \to K(X, X)$
  - $f \to f_K = \frac{1}{\#P} K(X, X_+)\boldsymbol{e} - \frac{1}{\#N} K(X, X_-)\boldsymbol{e}$
  - $\Sigma_+ - \Sigma_- \to \frac{1}{\#P} K(X, X_+)K(X_+, X) - \frac{1}{\#N} K(X, X_-)K(X_-, X)$





Dimensionality Reduction on Real Datasets

Table 4. Δ-fairness levels for both linear and Gaussian kernel SVM.

Data Set	Unconstrained			Linear Constraint			Both Constraints		
	%var	Lin.	Gaus.	%var	Lin.	Gaus.	%var	Lin.	Gaus.
Adult Income	11.39	0.53	0.54	9.20	0.18	0.35	3.59	**0.07**	**0.16**
Biodeg	30.88	0.54	0.60	23.69	0.27	0.53	3.73	**0.18**	**0.46**
E. Coli	61.30	0.92	0.91	50.62	0.63	0.91	23.61	**0.26**	**0.41**
Energy	84.24	0.13	0.24	51.74	**0.10**	0.29	25.63	0.12	**0.21**
German Credit	11.19	0.19	0.34	10.85	0.18	**0.29**	6.25	**0.16**	0.46
Image	59.21	1.00	0.99	17.77	0.34	0.70	0.01	**0.21**	**0.49**
Letter	43.50	0.14	0.27	33.53	0.10	0.28	3.28	**0.08**	**0.22**
Magic	58.09	0.39	0.41	49.22	0.13	0.31	4.77	**0.09**	**0.18**
Parkinson's	67.35	0.12	0.36	64.31	0.20	0.43	0.01	**0.06**	**0.14**
Pima	48.66	0.35	0.39	44.33	0.20	0.44	28.75	**0.16**	**0.36**
Recidivism	56.22	0.24	0.25	43.48	**0.07**	**0.15**	16.6	0.11	0.25
Skillcraft	41.20	0.12	0.20	29.89	**0.08**	**0.19**	9.64	**0.08**	**0.19**
Statlog	87.82	0.98	0.99	67.72	0.16	0.45	0.14	**0.11**	**0.23**
Steel	45.21	0.21	0.37	27.40	**0.14**	0.44	0.01	**0.14**	**0.28**
Taiwan Credit	45.45	0.09	0.14	28.70	**0.06**	0.16	3.03	**0.06**	**0.15**
Wine Quality	50.22	0.97	0.97	37.31	0.20	0.50	6.70	**0.06**	**0.16**

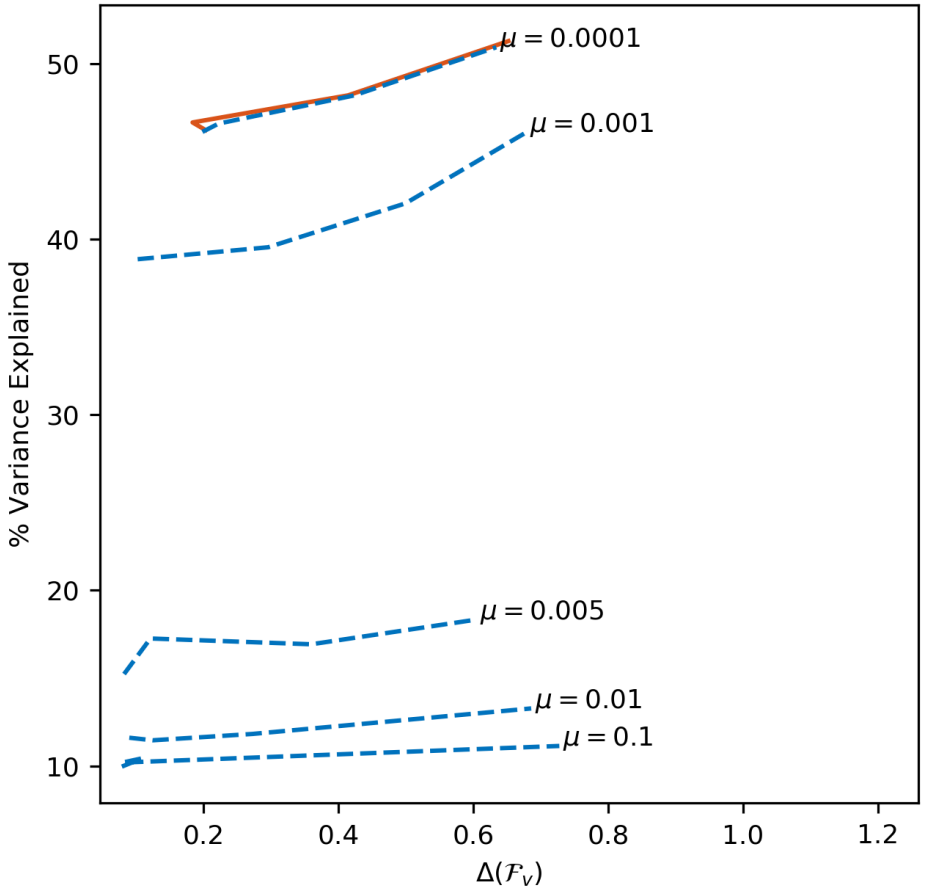
Classification Task on Real Datasets

Table 5: Comparison of accuracy and fairness on classification task using linear SVM. Results shown for linear SVM after dimensionality-reduction via PCA, FPCA with just the mean constraint and FPCA with both constraints, and are compared to the FSVM method of Olfat and Aswani (2017) (run with $\delta = 0, \mu = 0.1$ on non-dimensionality-reduced data) and the non-parametric method of Calmon et al.. Best fairness results are bolded.

	FSVM (NO PCA)		UNCONSTRAINED		FPCA - MEAN		FPCA - BOTH		CALMON ET AL.	
DATA SET	AUC	Δ	AUC	Δ	AUC	Δ	AUC	Δ	AUC	Δ
ADULT INCOME	0.86	0.13	0.66	0.17	0.69	**0.07**	0.57	0.08	0.51	0.23
BIODEG	0.85	0.12	0.82	0.20	0.81	0.13	0.79	**0.11**	0.60	0.14
ECOLI	0.74	**0.17**	0.84	0.50	0.69	0.23	0.72	0.29	0.63	0.30
ENERGY	0.55	0.09	0.51	0.09	0.56	0.08	0.55	**0.07**	0.54	0.13
GERMAN CREDIT	0.76	0.11	0.62	0.11	0.57	**0.10**	0.58	0.14	0.63	0.11
IMAGE SEG	0.99	0.19	0.99	0.16	0.99	0.19	0.98	**0.15**	0.79	0.20
LETTER REC	0.72	**0.07**	0.58	0.60	0.50	0.09	0.49	0.10	0.65	0.19
MAGIC	0.83	0.13	0.74	0.14	0.82	0.13	0.72	**0.12**	0.65	0.13
PIMA DIABETES	0.80	0.14	0.75	0.21	0.73	**0.11**	0.76	0.15	0.54	0.15
RECIDIVISM	0.54	0.08	0.69	0.24	0.54	**0.06**	0.52	0.07	0.55	0.08
SKILLCRAFT	0.82	0.06	0.85	0.10	0.82	**0.05**	0.80	**0.05**	0.62	0.07
STATLOG	0.99	0.31	1.00	0.33	0.99	0.33	0.85	0.18	0.67	**0.16**
STEEL	0.73	0.15	0.53	0.37	0.62	0.19	0.61	**0.12**	0.55	0.15
TAIWANESE CREDIT	0.73	**0.07**	0.60	0.11	0.60	0.09	0.64	**0.07**	0.75	**0.07**
WINE QUALITY	0.78	0.10	0.69	0.75	0.69	0.19	0.67	**0.05**	0.66	0.09

Sensitivity to Hyperparameters δ, μ

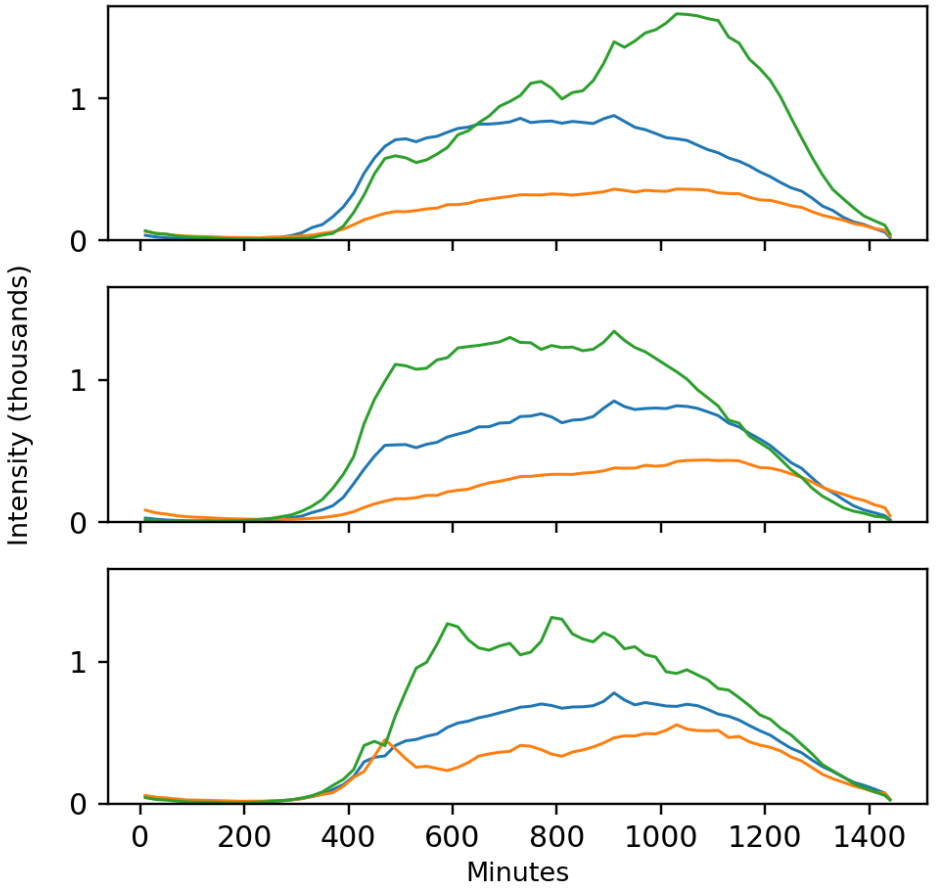
- Seek tradeoff between fairness, explanatory power
  - Data shown for Wine Quality dataset
  - 5-fold cross-validation
  - Each line shows $\delta \in \{0, 0.1, 0.3, 0.5\}$
  - Red line only mean constraint, blue lines both
- Most gain at small values of δ, μ
  - But specific to the dataset
  - Closer to Gaussian data → less cost to fairness



Feature Engineering Application in Health Care

- Insurance rates often set based on activity
  - But only have access to broad outlines of activity
  - Legal restrictions on effective discrimination
  - Need specific features to target
- Use NHANES data on physical activity
  - Minute-by-minute intensity levels for 6000 women for a week
  - Projected onto 5 principal components, then clustered
  - Fairness implemented with regards to age (>=40 vs. <40)
  - 36.05% of all respondents 40 or older
- Less discrimination in activity between 8:00 a.m. – 5:00 p.m.

	UNC.	CONVEX	BOTH
Cluster 1	43.18%	33.32%	35.61%
Cluster 2	36.11%	38.64%	32.94%
Cluster 3	8.71%	33.54%	37.28%



Thank you

BERKELEY IEOR
INDUSTRIAL ENGINEERING & OPERATIONS RESEARCH

https://arxiv.org/pdf/1802.03765.pdf

Clustering Task on Real Datasets

Table 4: Average squared distance from cluster center, as well as standard deviation of the proportion of each cluster that is of a certain protected class, for PCA, FPCA and the method of Calmon et al.. Best fairness results for each dataset are bolded.

	Unconstrained		FPCA - Mean		FPCA - Both		Calmon et al.	
Data Set	Score	Std. Dev	Score	Std. Dev	Score	Std. Dev	Score	Std. Dev
Adult Income	0.19	12.43	0.23	7.57	0.29	**2.28**	0.05	11.32
Biodeg	0.27	6.87	0.27	6.16	0.27	**5.34**	0.16	5.49
Ecoli	0.08	19.66	0.05	12.2	0.09	**10.69**	0.18	11.78
Energy	0.08	3.99	0.13	3.75	0.13	**3.57**	0.10	5.02
German Credit	0.25	6.4	0.25	4.82	0.28	**3.88**	0.03	4.16
Image Seg	0.10	8.46	0.09	**4.82**	0.11	5.95	0.12	10.85
Letter Rec	0.27	16.33	0.25	3.38	0.23	**3.28**	0.37	8.65
Magic	0.20	9.26	0.31	**5.15**	0.35	5.42	0.18	8.77
Pima Diabetes	0.24	9.09	0.27	6.36	0.26	5.96	0.28	**5.72**
Recidivism	0.26	7.6	0.17	**3.7**	0.19	3.8	0.05	4.69
SkillCraft	0.21	4.57	0.21	**2.27**	0.24	2.88	0.38	3.21
Statlog	0.09	21.99	0.23	16.06	0.31	**10.18**	0.13	11.12
Steel	0.16	18.49	0.19	9.85	0.24	**4.22**	0.22	17.97
Taiwanese Credit	0.17	3.85	0.24	2.99	0.29	**2.67**	0.03	3.64
Wine Quality	0.22	22.41	0.29	11.77	0.35	**2.11**	0.34	11.70