False Discovery Control in Low-rank Estimation

Armeen Taeb (ETH Zürich)

joint with: Parikshit Shah (Facebook), Venkat Chandrasekaran (Caltech)

Inference in contemporary data analysis

Approaches based on false discovery (rate) control have had substantial impact on how variable or feature selection is done in practice

False discovery control techniques useful primarily in model selection problems of a 'discrete' nature

Agenda: assess and control false discoveries (false positives) in settings where the decision space is 'more complicated'

- e.g. low-rank estimation, ranking, causal inference

Row/column spaces signify discoveries

Imaging spectroscopy: signature materials of scene

Radar: direction of moving targets

Phase retrieval: phase of an underlying signal

Recommender systems: latent spaces of user preferences and item attributes

Imaging spectroscopy

Data: reflectance properties of a scene across multiple wavelengths

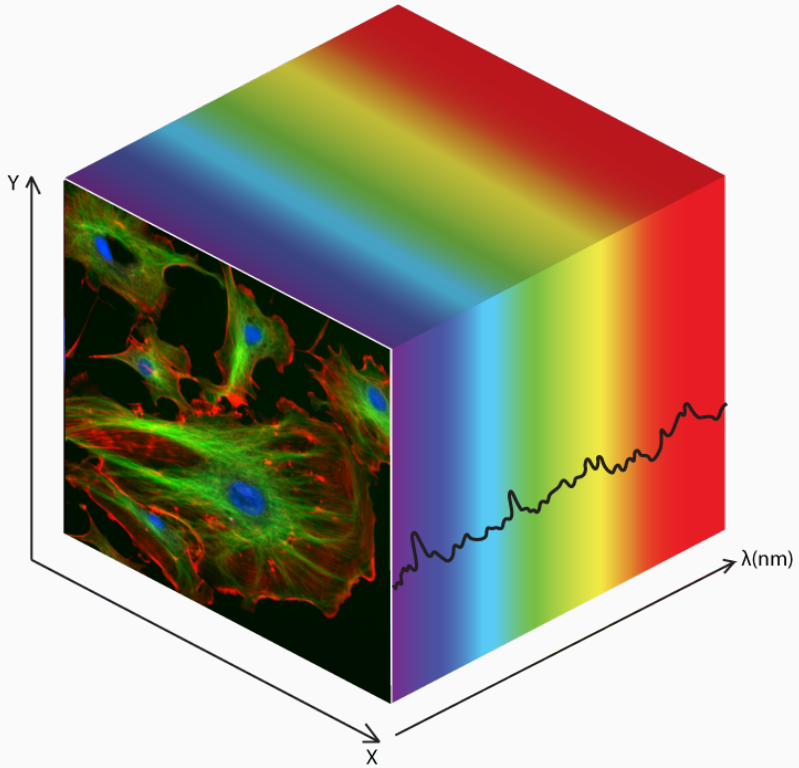
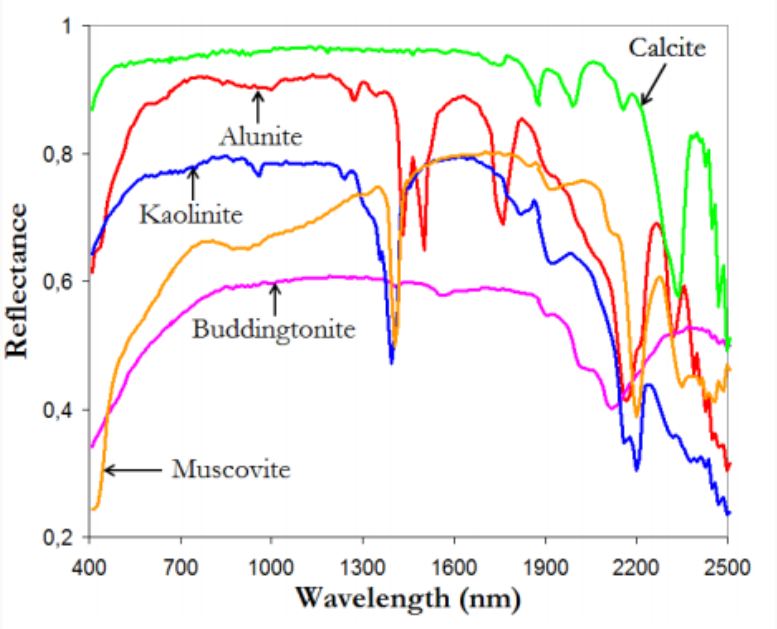


image collection comprises of a mix of material signatures

Imaging spectroscopy

Goal: identify materials present in the scene (unmixing)

Challenge: both the material signatures and mixing coefficients are unknown



structure: # materials $\ll$ # of wavelengths

Subspace discovery

Reflectance matrix: $Y \in \mathbb{R}^{p \times n}$ (p channels, n pixels)

Low-rank decomposition: $Y \approx WH$

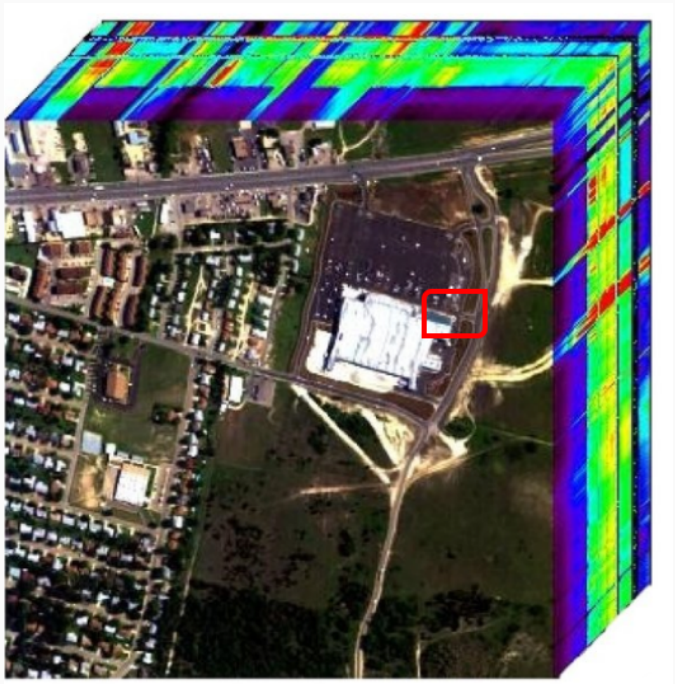
- $W \in \mathbb{R}^{p \times k}$: material matrix ; $H \in \mathbb{R}^{k \times n}$: mixing matrix
- column-space(W) : linear span of materials

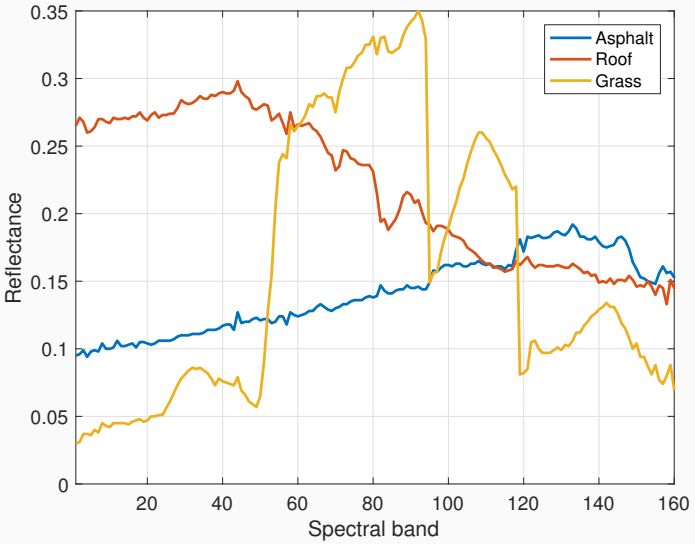
Subspace discovery = col-space(W)

- subsequent task: project data into subspace and check correlation with test material (e.g. methane) at each location

Implications of mistakes in column-space

Urban dataset

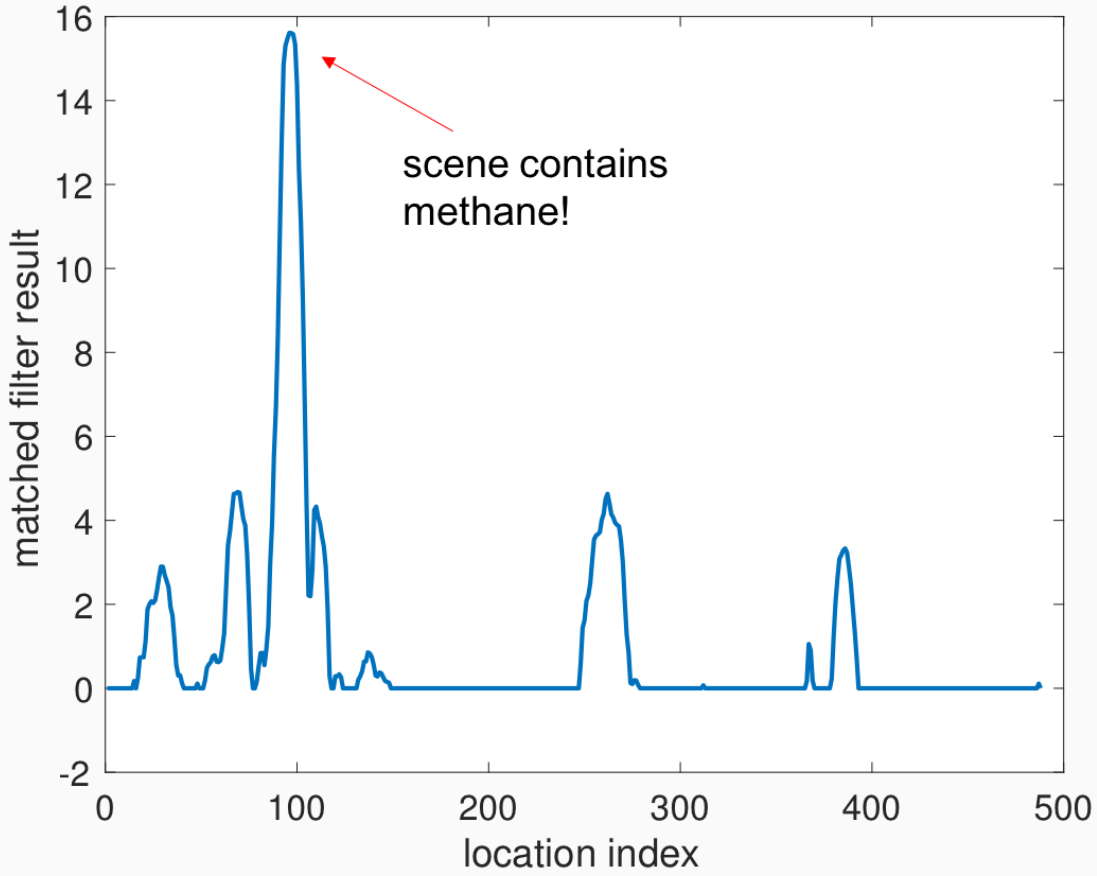




reflectance data $Y \in \mathbb{R}^{p \times n}$ with $p = 162$, $n = 94000$

Implications of mistakes in column-space

We estimate a column-space with false discovery proportion * $= 0.45$



methane is incorrectly labeled as present

Objective

Assess & control for false discoveries in row/column spaces

Prior work:

- significance testing of the singular values of an observed matrix (Choi et al (2017); Song & Shin (2018); . . .)
- shortcoming: do not control for deviation of row/column space
- shortcoming: rely on full observations of underlying matrix

Reminder: false and true discoveries in variable selection

Discovery: $\widehat{\text{subset}}$

Size of true discoveries: $\left|\widehat{\text{subset}} \cap \text{subset}^\star\right|$

Size of false discoveries: $\left|\widehat{\text{subset}}\right|$ – size of true discoveries

Neyman Pearson: $\max \left|\widehat{\text{subset}}\right|$ subject-to size of false discoveries $\leq \delta$

How to assess true/false discoveries: first attempt...

Discovery: row or column space $\hat{\mathcal{C}}$

- analogous to $\widehat{\text{subset}}$ in variable selection

Size of true discoveries: $\dim\left(\hat{\mathcal{C}} \cap \mathcal{C}^\star\right)$

- analogous to $\left|\widehat{\text{subset}} \cap \text{subset}^\star\right|$ in variable selection

Size of false discoveries: $\dim(\hat{\mathcal{C}})$ – size of true discoveries

Shortcoming: size of true discoveries $= 0$ (generically)

Geometric reformulation in variable selection

Estimate and population subsets $\hat{S} \subseteq \{1, 2, \ldots, p\}$; $S^\star \subseteq \{1, 2, \ldots, p\}$

Estimated (discovery) and population subspaces

$$T(\hat{S}) = \text{span}(\{e_i\}) \text{ for all } i \in \hat{S} \quad ; \quad T(S^\star) = \text{span}(\{e_i\}) \text{ for all } i \in S^\star$$

Size of true discoveries: $\dim\left(T(\hat{S}) \cap T(S^\star)\right) = \text{trace}\left(\mathcal{P}_{T(\hat{S})}\mathcal{P}_{T(S^\star)}\right)$

Size of false discoveries: $\dim\left(T(\hat{S}) \cap T(S^\star)^\perp\right) = \text{trace}\left(\mathcal{P}_{T(\hat{S})}\mathcal{P}_{T(S^\star)^\perp}\right)$

False and true discoveries: low-rank estimation

Discovery: row or column space $\hat{\mathcal{C}}$

Size of true discoveries: $\mathrm{trace}\,(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^\star})$

Size of false discoveries: $\mathrm{trace}\,(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^{\star\perp}})$

These notions can be derived based on 'axioms':

1. size of true discoveries $+$ size of false discoveries $=$ dim$(\hat{\mathcal{C}})$
2. invariant to simultaneous isometric linear transformation
   (i.e. simultaneous rotation of $\hat{\mathcal{C}}, \mathcal{C}^\star$)

Formal definitions

For estimate $\hat{\mathcal{C}}$ and population $\mathcal{C}^\star$:

$$\begin{aligned} \mathrm{FD} &= \mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^{\star\perp}}\right)\right] \\ \mathrm{PW} &= \mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^{\star}}\right)\right] \\ \mathrm{FDR} &= \mathbb{E}\left[\frac{\mathrm{trace}\left(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^{\star\perp}}\right)}{\dim(\hat{\mathcal{C}})}\right] \end{aligned}$$

where expectation is w.r.t randomness of the data

Properties:

1. $0 \leq \mathrm{FD} \leq \dim(\mathcal{C}^{\star\perp})$ & $0 \leq \mathrm{PW} \leq \dim(\mathcal{C}^\star)$
2. $\mathrm{FD} + \mathrm{PW} = \mathbb{E}\left[\dim(\hat{\mathcal{C}})\right]$ & $0 \leq \mathrm{FDR} \leq 1$

Is there a unifying model selection perspective?

Yes! the subspaces are tangent spaces to algebraic varieties

- variable selection: tangent spaces to sparse variety
- low-rank estimation: tangent spaces to quotients of low-rank variety

Tangent space perspective also enables a natural manner to assess both row and column spaces simultaneously

Algorithm to control FD

Inspired by stability selection [Meinshausen & Bühlmann '10]

- produce many bags of your data
- for each bag, use any variable selection procedure to select significant variables
- choose variables that appear often across subsamples

Theory: control on the size of false discoveries

- further work by Shah & Samworth '12

Our algorithm: subspace stability selection

Ingredients for subspace stability selection

1. Bagging: compute tangent spaces for each bag

2. Aggregate: fuse information from all tangent spaces

3. Output: produce a tangent space well-aligned to aggregate

   (e.g. select direction that appear often)

Aggregation step

Bagging: collection of tangent spaces $\{\hat{T}^{(i)}\}_{i=1}^{B}$

Aggregate: compute *average projection operator*

$$\mathcal{P}_{\text{avg}} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{P}_{\hat{T}^{(i)}}$$

Intuition: most of energy in $\mathcal{P}_{\text{avg}}$ is in $T^\star$

Properties:

- $\mathcal{P}_{\text{avg}}$ is self-adjoint
- eigenvalues of $\mathcal{P}_{\text{avg}}$ lie in $[0, 1]$

Output step

Given fixed $\alpha \in (0, 1)$

Output: solve the optimization problem

$$\max \dim(T) \quad \text{s.t.} \quad \sigma_{\min}\left(\mathcal{P}_T \mathcal{P}_{\text{avg}} \mathcal{P}_T\right) \geq \alpha \tag{1}$$

Remarks:

- T well-aligned with $\mathcal{P}_{\text{avg}}$
- efficient approach to solve (1) , e.g. *top* singular vectors of $\mathcal{P}_{\text{avg}}$
- computational cost of algorithm Bp^3

Subspace stability selection in variable selection

Subspace stability selection $=$ stability selection

Proof:

- $\mathcal{P}_{T(S)}$: diagonal with $\{0, 1\}$
- $\mathcal{P}_{\text{avg}}$: diagonal; elements encode frequency of variables
- **Key**: these two matrices commute

$$T(S) \text{ such that } \quad \sigma_{\min}(\mathcal{P}_{T(S)}\mathcal{P}_{\text{avg}}\mathcal{P}_{T(S)}) \geq \alpha$$

$$\Leftrightarrow$$

$$\text{for all } i \in S \quad : \quad (\mathcal{P}_{\text{avg}})_{i,i} \geq \alpha$$

Theoretical support

Some intuitions

When is subspace stability selection effective?

One scenario:

- tangent space estimates contain many directions around $T^\star$ (true signal)
- remaining components spread out over all other directions

Too stringent as the collection of subspaces is smooth – 'many direction' around $T^\star$; a less stringent scenario:

- tangent space estimates contains their energy around $T^\star$ or $T^{\star\perp}$
- formalized mathematically via commutators

Commutator

Def: For self-adjoint operators A, B, commutator: $[A, B] = AB - BA$

Two commutator terms in our analysis:

$$\kappa_{\text{bag}} = \mathbb{E}\left[\sqrt{\frac{1}{B}\sum_{j=1}^{B} \left\|[\mathcal{P}_{\hat{T}(j)}, \mathcal{P}_{T^\star \perp}]\right\|_F^2}\right]$$

$$\kappa_{\text{indiv}} = \mathbb{E}\left\|\left[\mathcal{P}_{\hat{T}(n/2)}, \mathcal{P}_{\text{span}(M)}\right]\right\|_F \;;\; M \text{ rank-1} \in T^{\star\perp}$$

Remarks:

- $\left\|[\mathcal{P}_{\hat{T}(j)}, \mathcal{P}_{T^\star \perp}]\right\|_F^2 = \sum_i \sin(2\theta_i)^2$; θ_i: principal angles
- $\kappa_{\text{bag}} = \kappa_{\text{indiv}} = 0$ for variable selection

Assumptions

Conditions on the estimator and the data generation process

- $\hat{T}(n/2)$: tangent space from $[n/2]$ observations

Assumption 1: "better than random guessing"

$$\text{normalized false discovery} \leq \text{normalized power}$$

Assumption 2: exchangeability in rank-1 directions of $T^{\star\perp}$

$$\text{distribution of } \|\mathcal{P}_{\hat{T}(n/2)}(M)\|_F \text{ is the same } \forall\ M \in T^{\star\perp}$$
$$\text{with } \text{rank}(M) = 1 \ \ \& \ \ \|M\|_F = 1$$

Assumptions

Assumptions: "better than random guessing", exchangeability in rank-1 directions of $T^{\star\perp}$

Natural model ensembles and estimators satisfy both assumptions

- e.g. matrix denoising with Gaussian noise
- e.g. linear measurements with Gaussian design & noise

Remarks:

- reduce to [Meinshausen & Bühlmann '12] for variable selection
- a less interpretable bound without these assumptions in the paper

Theoretical results

Theorem

Given n i.i.d data points, and input $\alpha \in (0, 1)$

- population $T^\star$ of $p_1 \times p_2$ low-rank matrix
- assumptions 1 & 2 satisfied

Let $q \triangleq \mathbb{E}[\dim(\hat{T}(n/2))]$. Then for any T s.t. $\sigma_{\min}\left(\mathcal{P}_T \mathcal{P}_{\text{avg}} \mathcal{P}_T\right) \geq \alpha$

$$\text{FD} \leq \frac{q^2}{p_1 p_2} + (1-\alpha)\mathbb{E}[\dim(T)] + f(\kappa_{\text{bag}}, \kappa_{\text{indiv}}),$$

where $f(\kappa_{\text{bag}}, \kappa_{\text{indiv}}) = p_1 p_2 \kappa_{\text{indiv}}^2 + 2q\kappa_{\text{indiv}} + \frac{4\sqrt{1-\alpha}}{\alpha}\sqrt{q\kappa_{\text{bag}}}$

Remark: a tighter (but less interpretable) bound in paper

Theoretical results

Theorem bound:

$$\mathrm{FD} \leq \frac{q^2}{p_1 p_2} + 2(1-\alpha)\mathbb{E}[\dim(T)] + f(\kappa_{\mathsf{bag}}, \kappa_{\mathsf{indiv}})$$

Remarks:

1. large p_1, p_2 reduce false discovery
2. assumption free bound on $\mathbb{E}[\dim(T)] \leq \frac{q}{\alpha}$
3. α chosen close to 1 reduces false discovery

Theoretical results

Theorem bound:

$$\mathrm{FD} \leq \frac{q^2}{p_1 p_2} + 2(1-\alpha)\mathbb{E}[\dim(T)] + f(\kappa_{\text{bag}}, \kappa_{\text{indiv}}),$$

where $f(\kappa_{\text{bag}}, \kappa_{\text{indiv}}) = p_1 p_2 \kappa_{\text{indiv}}^2 + 2q\kappa_{\text{indiv}} + \frac{4\sqrt{1-\alpha}}{\alpha}\sqrt{q\kappa_{\text{bag}}}$

Remarks:

1. f: increasing function of arguments with $f(0,0) = 0$.
2. influence of # bags via κ_{bag} with $\kappa_{\text{bag}} \overset{\text{assump free}}{\leq} \frac{q}{2}$; κ_{indiv} empirically bounded
3. $f(\kappa_{\text{bag}}, \kappa_{\text{indiv}}) = 0$ for variable selection

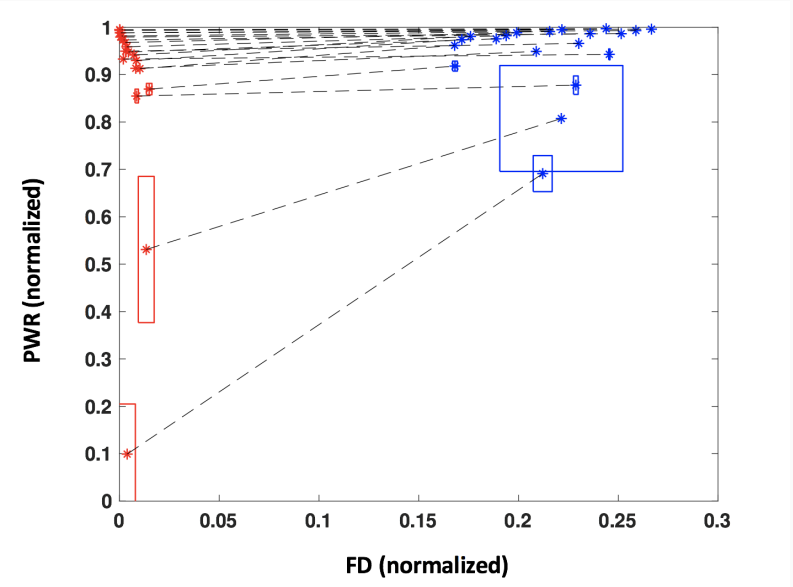
   refined analysis replaces $\frac{q^2}{p_1 p_2} + 2(1-\alpha)\mathbb{E}[\dim(T)] \to \frac{q^2}{p_1 p_2(2\alpha-1)}$

Experiments

Synthetic simulations

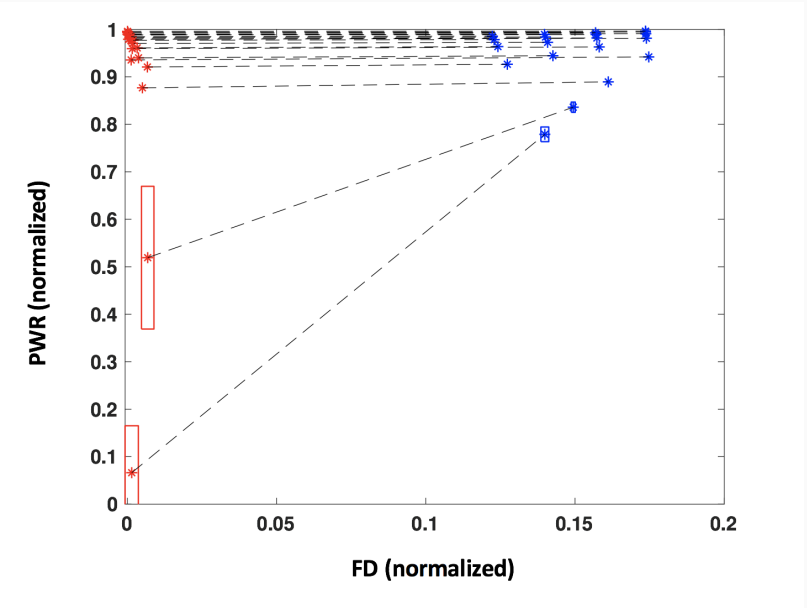
Matrix Completion

- dimension $= 50$, rank $= \{1, 2, 3, 4\}$
- SNR $= [1, 5]$, 10% obs.



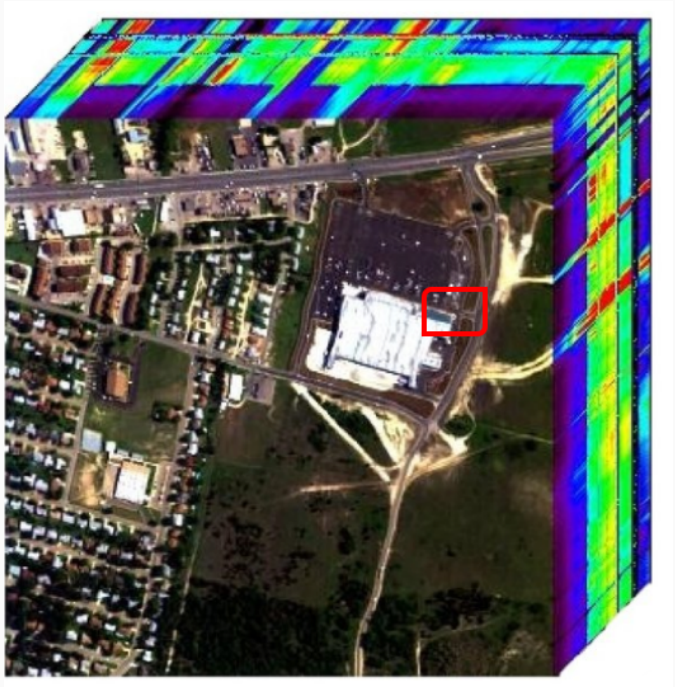
Linear measurements

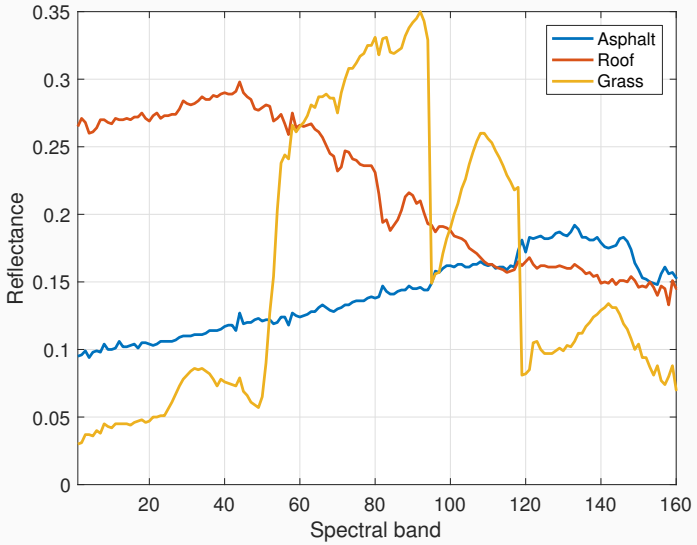
- dimension $= 50$, rank $= \{1, 2, 3, 4\}$
- SNR $= [1, 5]$, 10% meas.



Imaging spectroscopy

Urban Dataset





experiment: randomly subsample 10% data

Imaging spectroscopy

Factorization $Y \approx WH$

Known: col-space($W^\star$)

Estimator: alternating least-squares

$$(\hat{W}\ ,\ \hat{H}) = \underset{W \in \mathbb{R}^{p\times k}, H \in \mathbb{R}^{q\times k}}{\arg\min} \|(Y - WH^T)_{\text{obs}}\|_F^2 + \lambda\ (\|W\|_F^2 + \|H\|_F^2)$$

Result: for CV λ

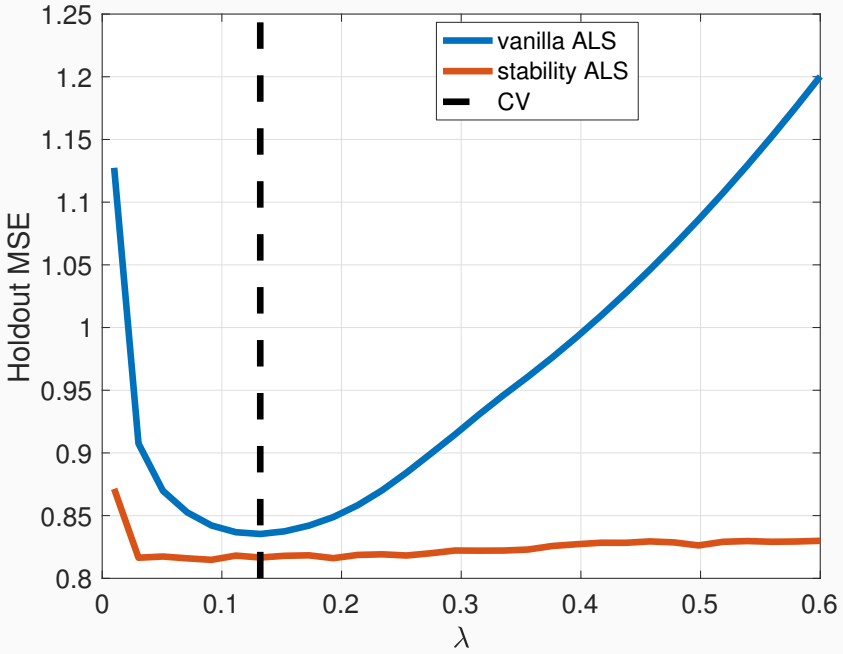
- ALS: $\text{rk} = 20;\ \frac{\text{FD}}{\dim(\mathcal{C}^{\star\perp})} \approx 0.1\ \&\ \frac{\text{PW}}{\dim(\mathcal{C}^\star)} \approx 0.98$
- Stability + ALS: $\text{rk} = 3;\ \frac{\text{FD}}{\dim(\mathcal{C}^{\star\perp})} \approx 0.0005\ \&\ \frac{\text{PW}}{\dim(\mathcal{C}^\star)} \approx 0.96$

$$\frac{\text{FD}}{\dim(\mathcal{C}^{\star\perp})} = 0.003 \text{ when rank (ALS)} = 3$$

Recommendation system: amazon book

Dataset: 1245 users, 1054 items, 6.1% observed

Estimator: ALS with fixed embedding dimension $k = 80$



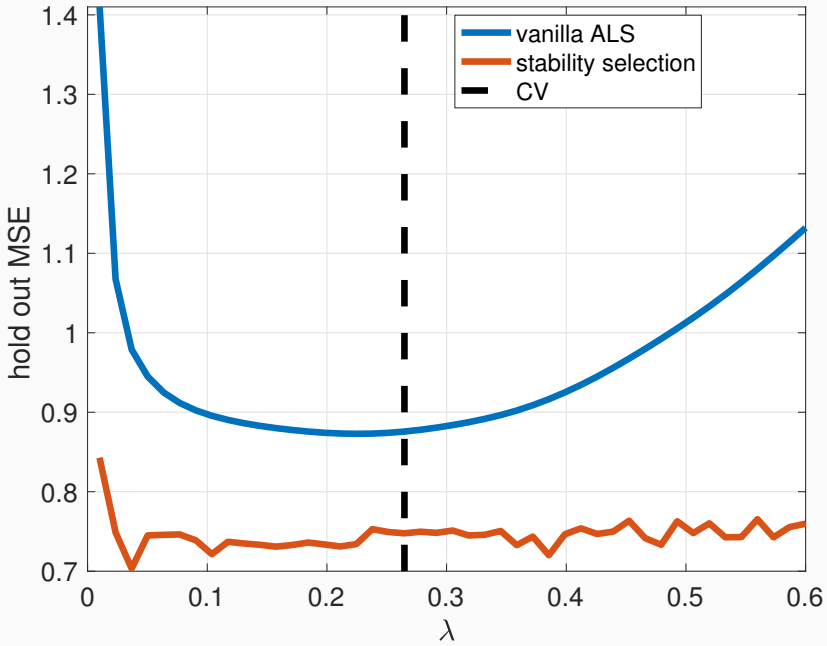
Result: for CV λ

- ALS: $\text{rk} = 80$, $\text{sing}(\hat{L})_{1:3} = 4300, 125, 63$
- Stability + ALS: $\text{rk} = 2$; performance boost of 2.4%

Recommendation system: amazon video games

Dataset: 482 users, 520 items, 3.5% observed

Estimator: ALS with fixed embedding dimension $k = 80$



Result: for CV λ

- ALS: $\text{rk} = 39$, $\text{sing}(\hat{L})_{1:5} = 913, 49, 43, 28, 27$
- Stability + ALS: $\text{rk} = 4$; performance boost of 17%

Deconfounding in Generalized linear graphical models

Convex optimization procedure to deconfound latent effects in a graphical model for GLM's (coming soon)

$$\begin{aligned} &\underset{K \in \mathbb{R}^{p \times p}, L \in \mathbb{R}^{p \times n}}{\arg\min} && \ell(X; KX + L) + \lambda(\|K\|_1 + \gamma \|L\|_\star) \\ &\text{subject-to} && K \text{ symmetric} \;\; ; \;\; K \text{ zeros on diagonal} \end{aligned}$$

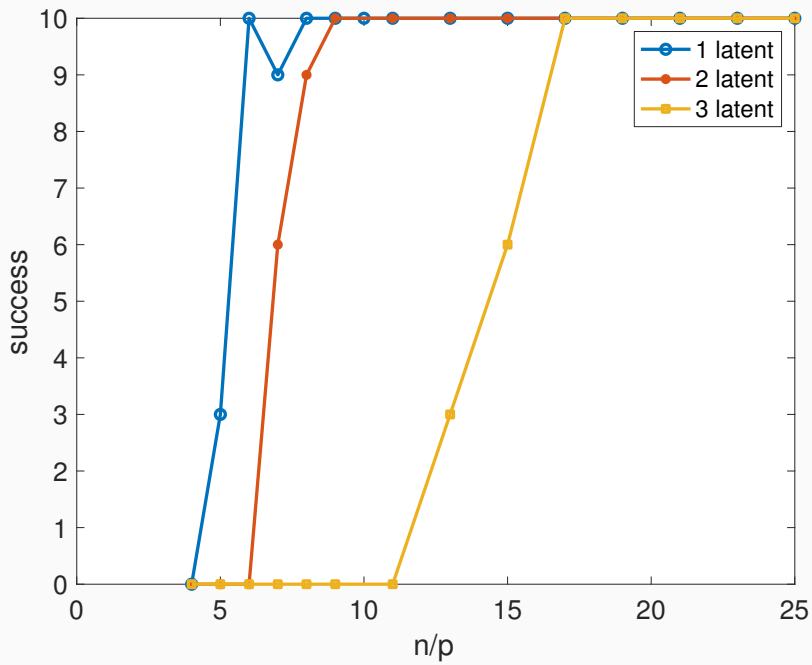
- $X \in \mathbb{R}^{p \times n}$: input data matrix
- ℓ: loss function of the generalized linear model
- K: graph structure, L: latent effect

Selecting (λ, γ) via cross-validation fails ($K = 0$, L full rank)

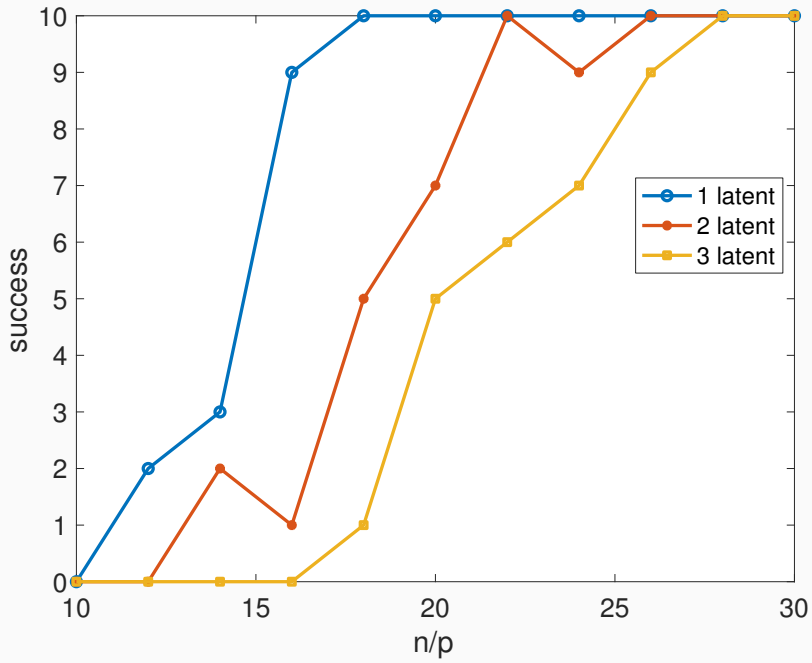
- Criteria: stable and low complexity model ; selection procedure based on subspace stability selection

Synthetic demonstrations

varying # latent variables



Poisson – Bernoulli
cycle graph, $p = 60$



Bernoulli – Gaussian
erdős – rényi$(60, 0.05)$

Summary

Agenda: testing for 'complicated' decision spaces

- low-rank estimation in this talk
- proposed subspace stability selection to control false discoveries

Future: false discovery rate control

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\mathrm{trace}\left(\mathcal{P}_{\hat{T}}\mathcal{P}_{T^\star{}^\perp}\right)}{\dim(\hat{T})}\right]$$

http://www.its.caltech.edu/ ataeb/index.html