A Geometric Perspective on False Discovery Control

Armeen Taeb (Caltech)

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

What is the Goal in Model Selection?

A common objective – **prediction**

Other related but distinct objectives

- *Discovery* – useful in scientific contexts
- Fairness – useful to mitigate bias across subpopulations
- Privacy – relevant in contexts in which anonymity is important

Need model selection approaches that address these latter objectives

A Case Study – Variable Selection

From a collection of p features, identify those that influence a response

- e.g., response may be a disease and features may be genes
- p is often very large

Goal: Estimate subset $\hat{S} \subseteq \{1, 2, \ldots, p\}$ from data so that $\hat{S} \approx S^\star$

- $S^\star \subseteq \{1, 2, \ldots, p\}$ is the true collection of relevant features
- Control for *false discoveries* $|\hat{S} \cap {S^\star}^c|$
- Neyman-Pearson: maximize $|\hat{S}|$ s.t. $\mathbb{E}|\hat{S} \cap {S^\star}^c| \leq \delta$

Controlling False Discoveries in Other Settings

Previous approach relevant in 'discrete' model selection problems

- e.g., graph structure estimation

How about controlling false discoveries / false positives more generally?

- What might discoveries (true or false) even mean?

Test case for today: low-rank estimation

- Ubiquitous in many applications
- As we'll see, generalizes variable selection nicely

Hyperspectral Imaging

Data: Reflectance properties of a scene across multiple wavelengths

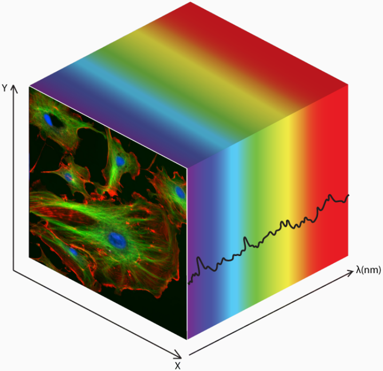
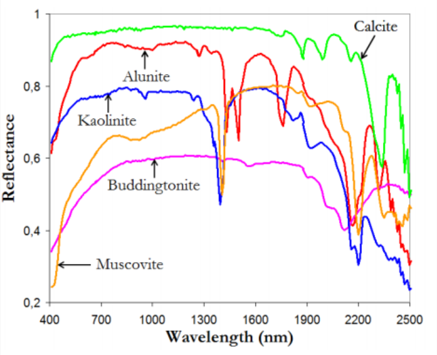


image collection comprises of a mix of **material signatures**

Hyperspectral Imaging

Goal: Identify materials present in the scene (unmixing)

Challenge: Both the material signatures and mixing coefficients are unknown



Structure: $\#$ materials $\ll$ $\#$ of wavelengths

Unmixing in Hyperspectral Imaging

Form reflectance matrix:

$$Y \in \mathbb{R}^{p \times n}$$

p : # of wavelengths ; n : # of pixels

Find 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

Unmixing in Hyperspectral Imaging

Form reflectance matrix:

$$Y \in \mathbb{R}^{p \times n}$$

p : # of wavelengths ; n : # of pixels

Find low-rank decomposition:

$$Y \approx WH$$

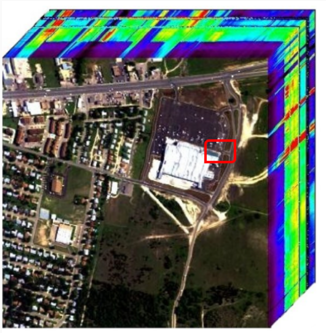
$W \in \mathbb{R}^{p \times k}$: material matrix ; $H \in \mathbb{R}^{k \times n}$: mixing matrix

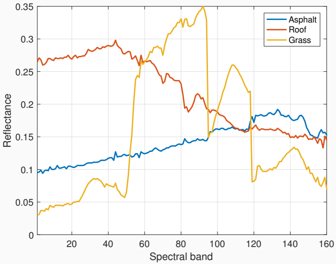
column-space(W) : linear span of materials

column-space is the key structural attribute of interest

Implications of Mistakes in Column-Space

Urban Hyperspectral Imaging Dataset

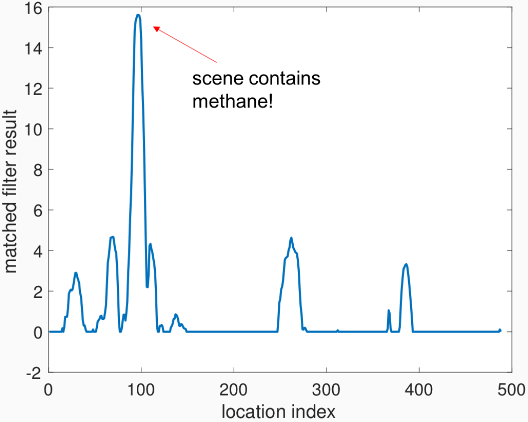




Reflectance data $Y \in \mathbb{R}^{p \times n}$ with $p = 162$, $n = 94000$

Implications of Mistakes in Column-Space

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



Methane is incorrectly labeled as present

Row/Column Spaces Signify Discoveries

Radar

- Row/column spaces: direction of moving targets

Phase Retrieval

- Row/column spaces: phase of underlying signal

Recommender systems

- Row/column spaces: latent spaces of user preferences and item attributes

Objective

Evaluate & 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

False Discovery Framework

Variable Selection

Discovery:

estimated subset

Number of True Discoveries:

$$|\text{estimated subset} \cap \text{population subset}|$$

Number of False Discoveries:

$$|\text{estimated subset}| - \text{number of true discoveries}$$

Question: *Generalize to row/column spaces of low-rank matrices?*

A First Attempt...

Discovery:

estimated row/column space

Number of True Discoveries:

$$\dim\left(\text{estimated row/column space} \cap \text{population row/column space}\right)$$

Number of False Discoveries:

$$\dim(\text{estimated row/column space}) - \text{number of true discoveries}$$

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

Geometric Picture: Variable Selection

Estimated/population subset

$$\hat{S} \subseteq \{1, 2, \ldots, p\} \; ; \; S^\star \subseteq \{1, 2, \ldots, p\}$$

Estimated/population subspaces

$$\begin{aligned} T(\hat{S}) &= \{x \in \mathbb{R}^p \mid x_i = 0 \text{ for all } i \in \hat{S}^c\} \\ T(S^\star) &= \{x \in \mathbb{R}^p \mid x_i = 0 \text{ for all } i \in {S^\star}^c\} \end{aligned}$$

Geometric Picture: Variable Selection

Estimated/population subset

$$\hat{S} \subseteq \{1, 2, \ldots, p\} \;;\; S^\star \subseteq \{1, 2, \ldots, p\}$$

Estimated/population subspaces

$$\text{Discovery: } T(\hat{S}) \quad ; \quad \text{Population: } T(S^\star)$$

Number of True Discoveries

$$\dim\left(T(\hat{S}) \cap T(S^\star)\right)$$

Number of False Discoveries

$$\dim\left(T(\hat{S}) \cap T(S^\star)^\perp\right)$$

Geometric Picture: Variable Selection

Estimated/population subset

$$\hat{S} \subseteq \{1, 2, \ldots, p\} \ ; \ S^\star \subseteq \{1, 2, \ldots, p\}$$

Estimated/population subspaces

$$\text{Discovery: } \ T(\hat{S}) \quad ; \quad \text{Population: } \ T(S^\star)$$

Number of True Discoveries

$$\dim\Big(T(\hat{S}) \cap T(S^\star)\Big) = \text{trace}\Big(\mathcal{P}_{T(\hat{S})}\mathcal{P}_{T(S^\star)}\Big)$$

Number of False Discoveries

$$\dim\Big(T(\hat{S}) \cap T(S^\star)^\perp\Big) = \text{trace}\Big(\mathcal{P}_{T(\hat{S})}\mathcal{P}_{T(S^\star)^\perp}\Big)$$

How to Think of the Map $S \leftrightarrow T(S)$

Variety of sparse vectors $\mathcal{V}(k)$ where $k = |S|$:

$$\mathcal{V}(k) = \{z \mid |\text{support}(z)| \leq k\}$$

$T(S) =$ **Tangent space** w.r.t. $\mathcal{V}(k)$ at x with support$(x) = S$

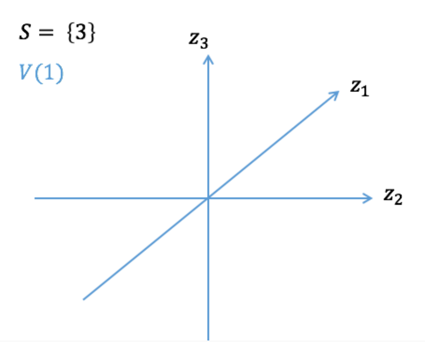
How to Think of the Map $S \leftrightarrow T(S)$

Variety of sparse vectors $\mathcal{V}(k)$ where $k = |S|$:

$$\mathcal{V}(k) = \{z \mid |\text{support}(z)| \leq k\}$$

$T(S) =$ **Tangent space** w.r.t. $\mathcal{V}(k)$ at x with support$(x) = S$

Example:



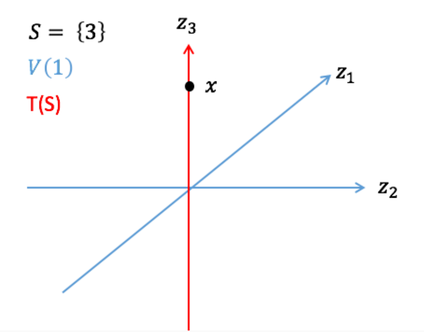
How to Think of the Map $S \leftrightarrow T(S)$

Variety of sparse vectors $\mathcal{V}(k)$ where $k = |S|$:

$$\mathcal{V}(k) = \{z \mid |\text{support}(z)| \leq k\}$$

$T(S) =$ **Tangent space** w.r.t. $\mathcal{V}(k)$ at x with support$(x) = S$

Example:



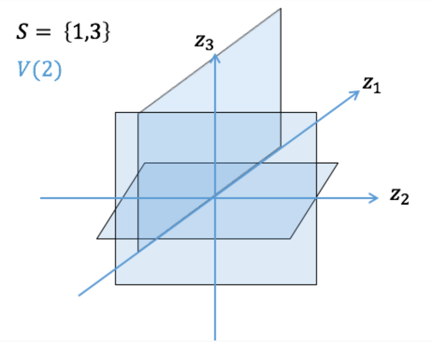
How to Think of the Map $S \leftrightarrow T(S)$

Variety of sparse vectors $\mathcal{V}(k)$ where $k = |S|$:

$$\mathcal{V}(k) = \{z \mid |\text{support}(z)| \leq k\}$$

$T(S) =$ **Tangent space** w.r.t. $\mathcal{V}(k)$ at x with $\text{support}(x) = S$

Example:



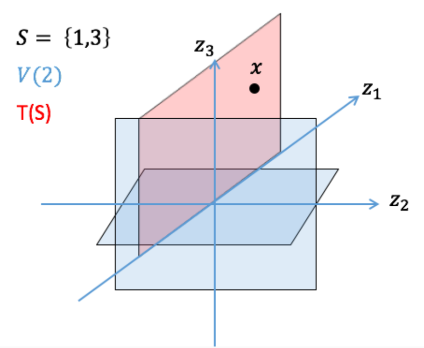
How to Think of the Map $S \leftrightarrow T(S)$

Variety of sparse vectors $\mathcal{V}(k)$ where $k = |S|$:

$$\mathcal{V}(k) = \{z \mid |\text{support}(z)| \leq k\}$$

$T(S) =$ **Tangent space** w.r.t. $\mathcal{V}(k)$ at x with support$(x) = S$

Example:



A Geometric Recipe

Takeaway: recipe to assess true/false discoveries

1. Identify structured *variety*
2. Determine *tangent space* w.r.t. variety
3. Compute inner-product between associated *projection matrices*

Generalization to Low-Rank Matrices

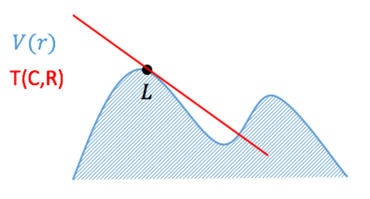
Variety: *determinantal* variety (space of low-rank matrices)

$$\mathcal{V}(r) \triangleq \{Z \mid \mathrm{rk}(Z) \leq r\}$$

Tangent Space: at a rank-r L w.r.t. $\mathcal{V}(r)$:

$$T(\mathcal{C},\mathcal{R}) : \{\mathcal{P}_{\mathcal{C}} Y_1 + Y_2 \mathcal{P}_{\mathcal{R}}\} \quad ; \quad \text{where } \mathcal{C} = \text{col-space} \; ; \; \mathcal{R} = \text{row-space}$$

One-to-one mapping between $(\mathcal{C},\mathcal{R})$ and $T(\mathcal{C},\mathcal{R})$



False/True Discoveries: Low Rank Estimation

Discovery: $T(\hat{\mathcal{C}}, \hat{\mathcal{R}})$

"Number" of True Discoveries: $\mathrm{trace}\left(\mathcal{P}_{T(\hat{\mathcal{C}},\hat{\mathcal{R}})}\mathcal{P}_{T(\mathcal{C}^\star,\mathcal{R}^\star)}\right)$

"Number" of False Discoveries: $\mathrm{trace}\left(\mathcal{P}_{T(\hat{\mathcal{C}},\hat{\mathcal{R}})}\mathcal{P}_{T(\mathcal{C}^\star,\mathcal{R}^\star)^\perp}\right)$

False/True Discoveries: Low Rank Estimation

Discovery: $T(\hat{\mathcal{C}}, \hat{\mathcal{R}})$

"Number" of True Discoveries: $\mathrm{trace}\left(\mathcal{P}_{T(\hat{\mathcal{C}},\hat{\mathcal{R}})}\mathcal{P}_{T(\mathcal{C}^\star,\mathcal{R}^\star)}\right)$

"Number" of False Discoveries: $\mathrm{trace}\left(\mathcal{P}_{T(\hat{\mathcal{C}},\hat{\mathcal{R}})}\mathcal{P}_{T(\mathcal{C}^\star,\mathcal{R}^\star)^\perp}\right)$

Interpretations of False Discovery

- Total energy of discovery in bad directions
- $\sum \cos(\angle T^{\star\perp}, \hat{T})^2$
- Specializes to variable selection if matrices are diagonal

False/True Discoveries: Column-space

What if we care only about column-space?

- e.g. hyperspectral imaging

Tangent spaces with respect to *quotients* of determinantal variety

- Quotient $\mathcal{V}(r)/[L]$ where $[L]$ is equivalence class

Using same machinery:

- “Number" of True Discoveries: $\mathrm{trace}\left(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^\star}\right)$
- “Number" of False Discoveries: $\mathrm{trace}\left(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^{\star\perp}}\right)$

Formal Definitions

Definition: Let $\hat{T}$ be tangent space estimate of a population tangent space $T^\star$. Then,

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

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

Formal Definitions

Definition: Let $\hat{T}$ be tangent space estimate of a population tangent space $T^\star$. Then,

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

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

Properties:

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

Formal Definitions

Definition: Let $\hat{T}$ be tangent space estimate of a population tangent space $T^\star$. Then,

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

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

Properties:

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

Algorithm

Stability Selection [Meinshausen & Bühlmann '12]

Algorithm: General-purpose approach based on bagging

Inputs: your favorite variable selection method; a threshold $\alpha \in (0, 1)$

1. *Bagging:* use variable selection method to estimate significant subset for each bag
2. *Aggregate:* compute frequency with which each variable is selected
3. *Output:* choose **all** variables that are selected with frequency at least α (maximize number of discoveries)

Theory: False discovery control as a function of α

- Further analysis by Shah & Samworth (2013)

Stability Selection [Meinshausen & Bühlmann '12]

Algorithm: General-purpose approach based on bagging

Inputs: your favorite variable selection method; a threshold $\alpha \in (0, 1)$

1. *Bagging:* use variable selection method to estimate significant subset for each bag
2. *Aggregate:* compute frequency with which each variable is selected
3. *Output:* choose **all** variables that are selected with frequency at least α (maximize number of discoveries)

Theory: False discovery control as a function of α

- Further analysis by Shah & Samworth (2013)

Develop a geometric analog:
Subspace Stability Selection

Ingredients for Subspace Stability Selection

Subspace Stability Selection Algorithm:

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

Aggregation Step

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

Aggregate: compute *average projection operator*

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

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

Properties:

1. $\mathcal{P}_{\mathrm{avg}}$ is self-adjoint
2. Eigenvalues of $\mathcal{P}_{\mathrm{avg}}$ lie in $[0, 1]$

Output Step

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

Output: choose largest tangent space T to determinantal variety s.t.:

$$\sigma_{\min}\left(\mathcal{P}_T \mathcal{P}_{\mathsf{avg}} \mathcal{P}_T\right) \geq \alpha$$

Remarks:

1. T well-aligned with $\mathcal{P}_{\mathsf{avg}}$
2. Efficient approach to find T that satisfies criterion

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

$$
\begin{array}{ccc}
T(S) \text{ such that} & & \sigma_{\min}(\mathcal{P}_{T(S)}\mathcal{P}_{\text{avg}}\mathcal{P}_{T(S)}) \geq \alpha \\
& \Leftrightarrow & \\
\text{for all } i \in S & : & (\mathcal{P}_{\text{avg}})_{i,i} \geq \alpha
\end{array}
$$

Theoretical Support

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"

$$\underbrace{\frac{\mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{T^{\star\perp}}\mathcal{P}_{\hat{T}(n/2)}\right)\right]}{\dim(T^{\star\perp})}}_{\text{normalized false discovery}} \leq \underbrace{\frac{\mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{T^{\star}}\mathcal{P}_{\hat{T}(n/2)}\right)\right]}{\dim(T^{\star})}}_{\text{normalized power}}$$

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

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"

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

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

Assumptions

Assumption 1: “better than random guessing"

Assumption 2: 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 sparse variety
- A less interpretable bound without these assumptions in the paper

Commutator

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

Important quantities in our analysis:

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

$$\kappa_{\mathsf{indiv}} = \min_i \mathbb{E}\left\|\left[\mathcal{P}_{\hat{T}(n/2)}, \mathcal{P}_{\mathrm{span}(M_i)}\right]\right\|_F \;;\; \{M_i\}_{i=1}^{\dim(T^{\star\perp})} \text{ rank-1 basis for } T^{\star\perp}$$

Remarks:

1. $\|[\mathcal{P}_{\hat{T}^{(j)}}, \mathcal{P}_{T^{\star\perp}}]\|_F^2 = \sum_i \sin(2\theta_i)^2$; θ_i: principal angles
2. $\kappa_{\mathsf{bag}} = \kappa_{\mathsf{indiv}} = 0$ for sparse variety

Theoretical Results

Theorem

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

- population tang. space $T^\star$ of $p_1 \times p_2$ low-rank matrix
- final output based on aggregating estimates from bags of size $\lfloor n/2 \rfloor$
- Assumptions 1 & 2

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

$$\mathrm{FD} \leq \frac{q^2}{p_1 p_2} + \frac{2(1-\alpha)}{\alpha} q + f(\kappa_{\mathrm{bag}}, \kappa_{\mathrm{indiv}}),$$

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

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

Theoretical Results

Theorem bound:

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

Remarks:

1. large p_1, p_2 reduce false discovery
2. α chosen close to 1 reduces false discovery

Theoretical Results

Theorem bound:

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

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:

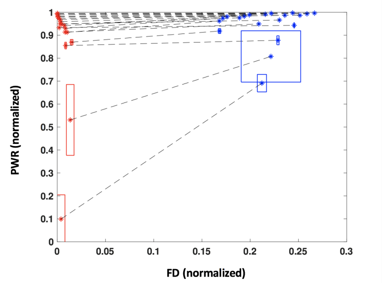
1. f: increasing function of arguments with $f(0,0) = 0$.
2. influence of $\#$ bags via κ_{bag}; $\kappa_{\text{bag}} \leq \frac{q}{2}$ for bag independent bound
3. $f(\kappa_{\text{bag}}, \kappa_{\text{indiv}}) = 0$ for variable selection
   refined analysis replaces $\frac{q^2}{p_1 p_2} + \frac{2(1-\alpha)}{\alpha} q \to \frac{q^2}{p_1 p_2 (2\alpha - 1)}$

Experiments

Synthetic Simulations

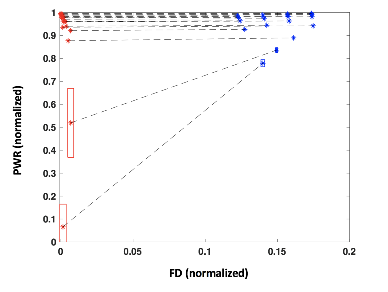
Matrix Completion

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



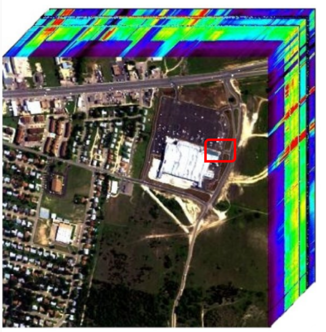
Linear measurements

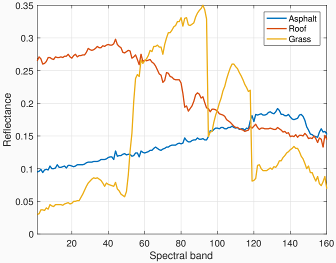
- dimension $= 50$, rank $= \{1,2,3,4\}$
- SNR $= [1,5], 10\%$ meas.



Imaging Spectroscopy

Urban Hyperspectral Imaging Dataset





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

Experiment: randomly subsample 10% data

Imaging Spectroscopy

Factorization of $Y \approx WH$ from incomplete obs.

Known: column-space of $W^\star$

Estimator: alternating least-squares

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

Imaging Spectroscopy

Factorization of $Y \approx WH$ from incomplete obs.

Known: column-space of $W^\star$

Estimator: alternating least-squares

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

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: $\text{rk} = 3$; $\frac{\text{FD}}{\dim(\mathcal{C}^{\star\perp})} \approx 0.0005$ & $\frac{\text{PW}}{\dim(\mathcal{C}^\star)} \approx 0.96$

Imaging Spectroscopy

Factorization of $Y \approx WH$ from incomplete obs.

Known: column-space of $W^\star$

Estimator: alternating least-squares

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

Result: for CV λ

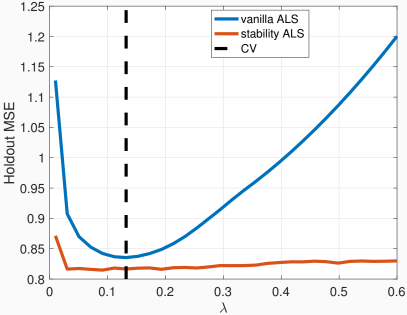
- ALS: $\text{rk} = 20;\ \frac{\text{FD}}{\dim(\mathcal{C}^{\star\perp})} \approx 0.1\ \&\ \frac{\text{PW}}{\dim(\mathcal{C}^\star)} \approx 0.98$
- Stability + ALS: $\text{rk} = 3;\ \frac{\text{FD}}{\dim(\mathcal{C}^{\star\perp})} \approx 0.0005\ \&\ \frac{\text{PW}}{\dim(\mathcal{C}^\star)} \approx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

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

Recommendation System: Amazon Book

Dataset: 1245 users, 1054 items, 6.1% observed

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



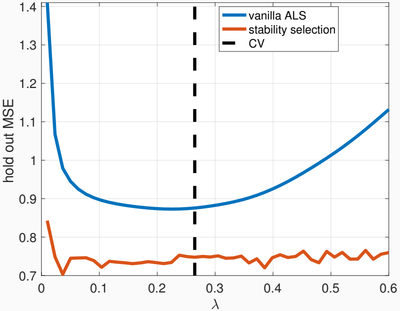
Result: for CV λ

- ALS: rk $= 80$, $\text{sing}(\hat{L})_{1:3} = 4300, 125, 63$
- Stability + ALS: 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: rk $= 39$, $\text{sing}(\hat{L})_{1:5} = 913, 49, 43, 28, 27$
- Stability + ALS: rk $= 4$; performance boost of 17%

Summary

Agenda: testing for continuous decision spaces

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

Future: perspective immediately useful in related problems

- latent variable graphical modeling; tensors; manifold learning

Future: False Discovery Rate control

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

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

Selecting a Tangent Space

Goal: Select T from the determinantal variety s.t.

$$\sigma_{\min}\left(\mathcal{P}_T \mathcal{P}_{\text{avg}} \mathcal{P}_T\right) \geq \alpha$$

Compute average projection row/column spaces

$$\mathcal{P}^C_{\text{avg}} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{P}_{\hat{\mathcal{C}}^{(i)}} \qquad \mathcal{P}^R_{\text{avg}} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{P}_{\hat{\mathcal{R}}^{(i)}}$$

Find closest row/col space

$$\hat{\mathcal{C}}(r) \quad = \quad \underset{\mathcal{C} \text{ subspace of dimension } r}{\arg\max} \sigma_{\min}\left(\mathcal{P}_{\mathcal{C}} \mathcal{P}^C_{\text{avg}} \mathcal{P}_{\mathcal{C}}\right)$$

Output largest r so that $T \triangleq T(\hat{\mathcal{C}}(r), \hat{\mathcal{R}}(r))$ satisfies criterion.