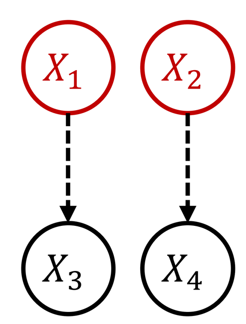
Quantifying uncertainty and stability among highly correlated predictors: a subspace perspective

Xiaozhu Zhang*, Jacob Bien†, Armeen Taeb*

Department of Statistics, University of Washington*; Department of Data Sciences and Operations, University of Southern California†

Motivation

Toy example ♠:

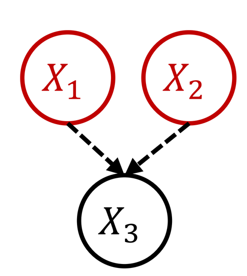


- X_1 and X_2 are independent, so are X_3 and X_4
- X_1 and X_3 are highly correlated, so are X_2 and X_4
- $y = \beta_1^\star X_1 + \beta^\star X_2 + \epsilon$

Challenges in this setting:

- How to quantify **false positive error** and false negative error?
  - If $\widehat{S} = \{3, 4\}$, did I make two false discoveries? NO!
- How to measure **stability** of selected features and sets?
  - In stability selection, no features are stable due to "vote splitting".
- How to **aggregate** stable features in to a model?
  - Should: return both sets $\{1, 2\}$ and $\{3, 4\}$.
  - Should not: include redundant features, such as $\{1, 2, 3\}$.
- How to identify **substitutes** within stable models?
  - X_1 and X_3 are substitutes; X_2 and X_4 are substitutes.

A more complex structure ♣ than pairwise correlation:



- X_1 and X_2 are independent
- $X_3 = X_1 + X_2 + \delta$;
- $y = X_1 - X_2 + \epsilon$

Subspace false positive/ negative error

Given two sets S and $\widetilde{S}$, we measure the similarity between the two **feature subspaces** $\mathrm{col}(X_S)$ and $\mathrm{col}(X_{\widetilde{S}})$!

- For the toy example ♠, $\mathrm{col}(X_{\{1,2\}}) \approx \mathrm{col}(X_{\{3,4\}})$.
- For the complex structure ♣, $\mathrm{col}(X_{\{1,2\}}) \approx \mathrm{col}(X_{\{1,3\}}) \approx \mathrm{col}(X_{\{2,3\}})$.

False positive error and false negative error

Let $S^\star$ be the true set of features and $S \subseteq \{1, 2, \dots, p\}$ be an estimated set of features. Then

$$\mathrm{TP}(\widehat{S}, S^\star) := \mathrm{trace}\left(\mathcal{P}_{\mathrm{col}(X_{\widehat{S}})}\mathcal{P}_{\mathrm{col}(X_{S^\star})}\right),$$
$$\mathrm{FPE}(\widehat{S}, S^\star) := |\widehat{S}| - \mathrm{TP}(\widehat{S}, S^\star),$$
$$\mathrm{FNE}(\widehat{S}, S^\star) := |S^\star| - \mathrm{TP}(\widehat{S}, S^\star).$$

Subspace stability

- Apply a variable selection procedure (Lasso or ℓ_0-regression) to B subsamples of size $\lfloor n/2 \rfloor$ to obtain sets $\{\widehat{S}^{(\ell)}\}_{\ell=1}^B$.
- Compute the average matrix

$$\mathcal{P}_{\mathrm{avg}} := \frac{1}{B}\sum_{\ell=1}^{B} \mathcal{P}_{\mathrm{col}(X_{\widehat{S}^{(\ell)}})}.$$

Stability of a set S

Given the average matrix $\mathcal{P}_{\mathrm{avg}}$, the stability of S is

$$\pi(S) := \sigma_{|S|}(\mathcal{P}_{\mathrm{col}(X_S)}\mathcal{P}_{\mathrm{avg}}\mathcal{P}_{\mathrm{col}(X_S)}) \in [0, 1],$$

A larger $\pi(S)$ indicates a more stable S.

We propose a Feature Subspace Stability Selection (FSSS) algorithm that returns stable model with a guaranteed $\pi(S)$.

FSSS algorithm

- Input: Matrix $\mathcal{P}_{\mathrm{avg}}$, and stability threshold $\alpha \in (1/2, 1)$
- Sequentially add variables:
  - Start from the null model $\varnothing$
  - Add X_j to the current model S if $\pi(S \cup \{j\}) \geq \alpha$
  - Stop until no new features can be added
- Output: one selection set $\widehat{S}$

- Any output $\widehat{S}$ is at least **α-stable**: $\pi(\widehat{S}) \geq \alpha$.
- Each run may give **different selection sets**:
  - For the toy example ♠, both $\{1, 2\}$ and $\{3, 4\}$ can be returned.
  - For the complex example ♣, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$ can be returned.
- **No redundant features** can be selected:
  - For the toy example ♠,
  $\pi(\{1, 2, 3\}) \approx \pi(\{1, 2, 4\}) \approx \pi(\{1, 3, 4\}) \approx \pi(\{2, 3, 4\}) \approx 0.$
  - For the complex example ♣, $\pi(\{1, 2, 3\}) \approx 0.$
- **False positive error** can be controlled:

Theorem: Under mild conditions, for any $\widehat{S}$ returned by FSSS,

$$\mathbb{E}\left[\mathrm{FPE}(\widehat{S}, S^\star)\right] \leq \frac{p(\gamma + b)^2}{2\alpha - 1} + a.$$

- The term γ is a quality term of the base procedure, with $\gamma \approx s_0/p$.
- The terms a and b are slackness terms that vanishes when features are perfectly orthogonal (a scenario reducing back to stability selection)

Subspace substitutability

- A substitutability metric based on subspace:

$$\tau(\bar{S}, \widetilde{S}) := \frac{\min\{\|\mathcal{P}_{\bar{S}}(y)\|_2, \|\mathcal{P}_{\widetilde{S}}(y)\|_2\}}{\max\{\|\mathcal{P}_{\bar{S}}(y)\|_2, \|\mathcal{P}_{\widetilde{S}}(y)\|_2\}} \cdot |\mathrm{Corr}(\mathcal{P}_{\bar{S}}(y), \mathcal{P}_{\widetilde{S}}(y))| \in [0, 1].$$

- Two more metrics $\nabla\tau(\bar{S}, \widetilde{S})$ and $\triangle\tau(\bar{S}, \widetilde{S})$ that examines the interesting level of substitutes $(\bar{S}, \widetilde{S})$.
- One algorithm that searches for substitutes among all stable models.

Experiments

- A synthetic dataset including both pairwise correlation ♠ and complex structure ♣.
- Base procedure: ℓ_0-penalized regression.