Selective inference with kernels

Makoto Yamada

http://www.makotoyamada-ml.com/

High-dimensional Statistical Modeling Team

Feature selection

Selecting important features from data.

By Guillaume Paumier, from https://commons.wikimedia.org/wiki/File:DNA_microarray.svg

Select $m < d$ features from $\boldsymbol{x} \in \mathbb{R}^d$

Nonlinear feature selection

Biological data cannot be modeled by linear models.

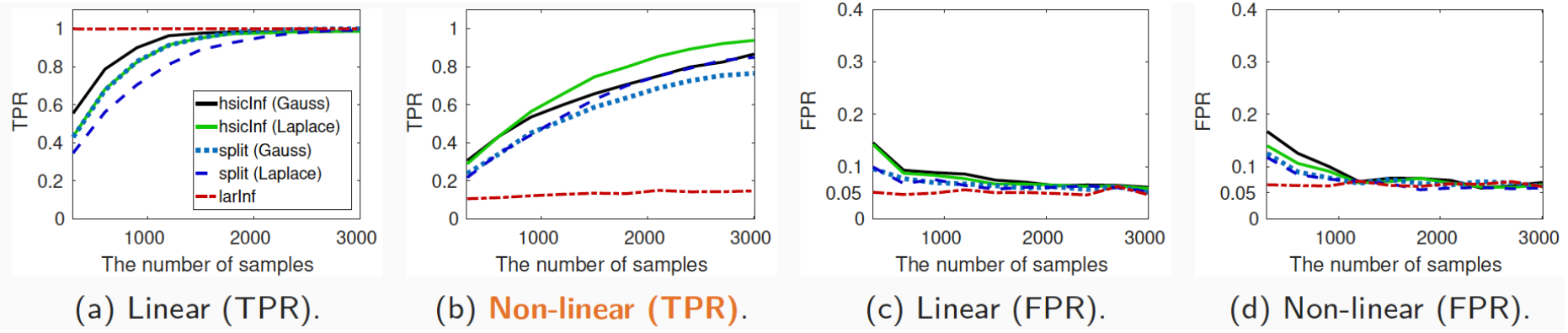
- **Interpretability**
  - Want to select small number of features.
  - The confidence of the selected features (p-values).
- **Challenge**
  - Nonlinear model is complex

    Scientist does not want to use it
  - No selective inference for nonlinear feature selection Strong assumption is needed

Selective inference with kernels

We established a kernel based selective inference framework!

- Selective independence test (AISTATS 2018, AISTATS 2020)
- Selective two-sample test (ICLR 2019)
- Selective goodness-of-fit test (NeruIPS 2019)



(a) Linear (TPR). (b) Non-linear (TPR). (c) Linear (FPR). (d) Non-linear (FPR).

Post selection inference for Lasso

- Least absolute shrinkage and selection operator (Lasso)

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1, \boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Lasso is a least-square regression method with L1 regularization. We can obtain sparse model.
- In scientific research, we want to know the confidence of the selected features (i.e., non-zero coefficient of Lasso)
- If we do feature selection and inference from the same dataset, we need to get rid of the selection bias.

Polyhedral Lemma

Lee et al. Annals of Stats 2016

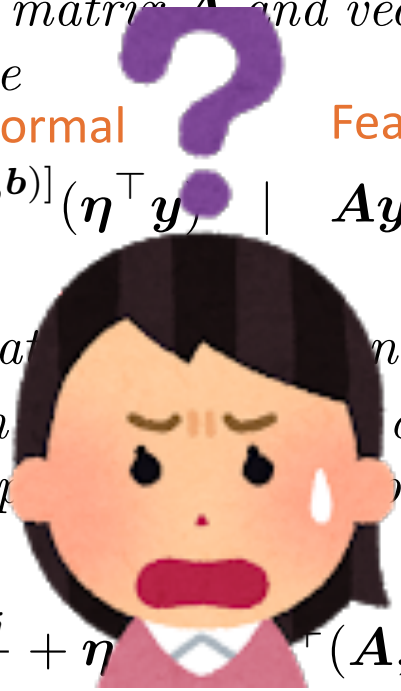
Theorem 1 *Suppose that* $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *and the feature selection event can be expressed as* $\boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b}$ *for some matrix* $\boldsymbol{A}$ *and vector* $\boldsymbol{b}$, *then for any given feature represented by* $\boldsymbol{\eta} \in \mathbb{R}^n$ *we have*

CDF the uni-variate truncated normal

Feature selection event

$$F^{[V^-(\boldsymbol{A},\boldsymbol{b}),V^+(\boldsymbol{A},\boldsymbol{b})]}_{\boldsymbol{\eta}^\top\boldsymbol{\mu},\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta}}(\boldsymbol{\eta}^\top\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b} \sim \mathrm{Unif}(0,1),$$

where $F^{[a,b]}_{\mu,\sigma^2}(x)$ *is the cumulat* *n function (CDF) of a truncated normal distribution with mean* σ^2 *truncated at [a,b]. Given that* $\boldsymbol{\alpha} = \boldsymbol{A}\frac{\boldsymbol{\Sigma}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta}}$, *the lower and up* *oints are given by*

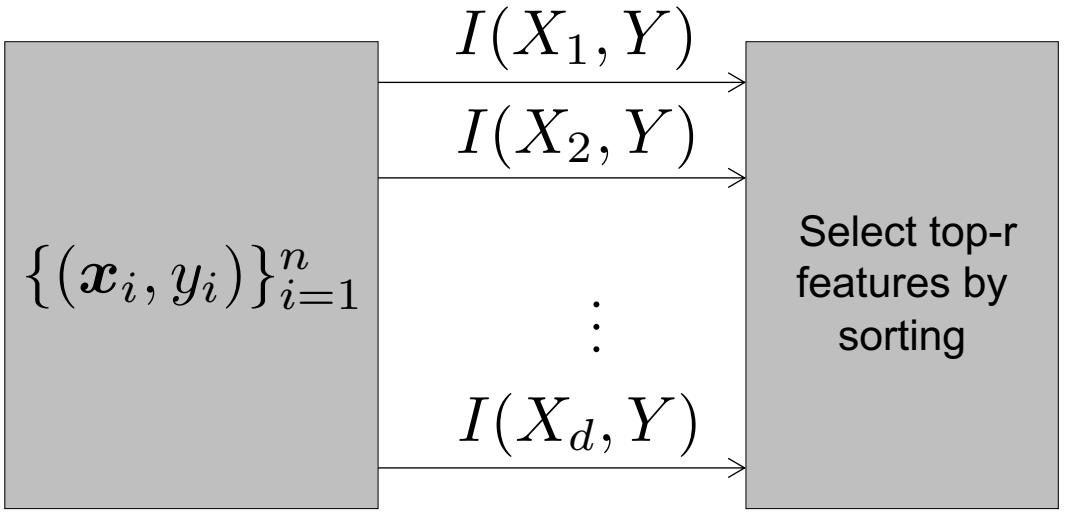


$$V^-(\boldsymbol{A},\boldsymbol{b}) = \max_{j:\boldsymbol{\alpha}_j<0} \frac{b_j - (\boldsymbol{A}\boldsymbol{y})_j}{\boldsymbol{\alpha}_j} + \boldsymbol{\eta}^\top\boldsymbol{y}, \quad V^+(\boldsymbol{A},\boldsymbol{b}) = \min_{j:\boldsymbol{\alpha}_j>0} \frac{\boldsymbol{b}_j - (\boldsymbol{A}\boldsymbol{y})_j}{\boldsymbol{\alpha}_j} + \boldsymbol{\eta}^\top\boldsymbol{y}.$$

Lower and upper truncation points

Sure Independence Screening (SIS)
(Fan, Lv, JRSSB 2008)

- Compute association score between each feature and its output and rank them.
  - Linear correlation, Mutual information and kernel based independence measures are used.
  - Easy to implement and scalable



$$\max_{\boldsymbol{\beta}\in\{0,1\}^d} \sum_{k=1}^{d} \beta_k I(X_k, Y),$$

$$\text{s.t.} \sum_{k=1}^{d} \beta_k = r$$

Selective inference with screening

- We employ an estimate of the independence measure:

$$\widehat{I}(X_i, Y)$$

with

$$\boldsymbol{y} = (\widehat{I}(X_1, Y), \widehat{I}(X_2, Y), \ldots, \widehat{I}(X_d, Y))^\top \in \mathbb{R}^d$$

following a multi-variate normal distribution:

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Illustration of Polyhedral Lemma

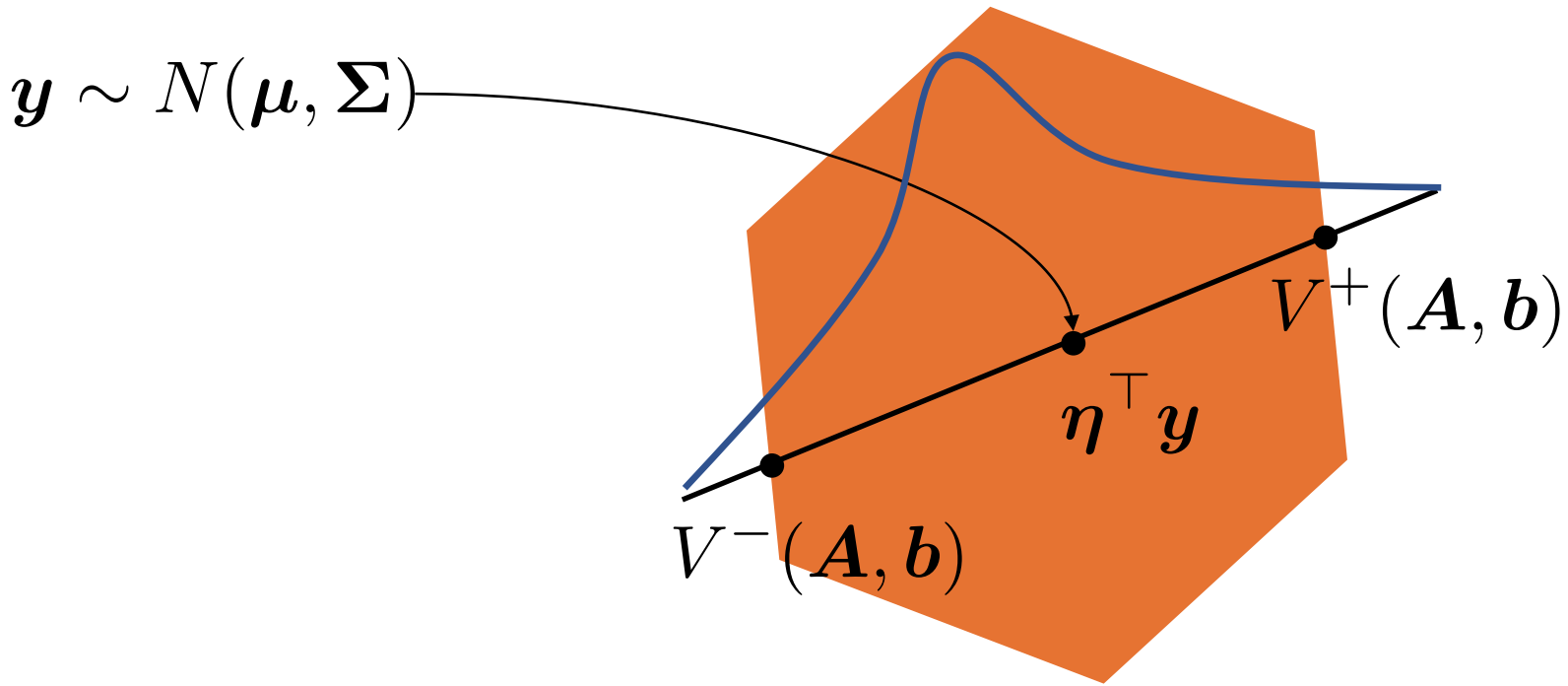
- Assumptions
  - Output follows multivariate normal $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - The feature selection event can be written as $\boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b}$
- Example (Feature ranking, SIS), we assume $\boldsymbol{y} \in \mathbb{R}^d$
  - Selected set $i \in \mathcal{S}$, non-selected set $j \in \bar{\mathcal{S}}$
  - Selection event is $y_i \geq y_j \rightarrow y_j - y_i \leq 0$ $y_i = I(\boldsymbol{X}_i, Y)$

$$[0 \quad \dots \quad 0 \quad \underbrace{-1}_{i} \quad 0 \quad \dots \quad 0 \quad \underbrace{1}_{j} \quad 0 \quad \dots \quad 0]\boldsymbol{y} \leq 0$$

- Lasso case, the selection event can be derived from the KKT condition. In this case, $\boldsymbol{y} \in \mathbb{R}^n$

Illustration of Polyhedral Lemma

- The feature selection event is polytope $\boldsymbol{Ay} \leq \boldsymbol{b}$
- Key idea: $\boldsymbol{\eta}^\top \boldsymbol{y}$ follows normal distribution



Polyhedral Lemma

Lee et al. Annals of Stats 2016

Theorem 1 *Suppose that* $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *and the feature selection event can be expressed as* $\boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b}$ *for some matrix* $\boldsymbol{A}$ *and vector* $\boldsymbol{b}$, *then for any given feature represented by* $\boldsymbol{\eta} \in \mathbb{R}^n$ *we have*

CDF the uni-variate truncated normal

Feature selection event

$$F^{[V^-(\boldsymbol{A},\boldsymbol{b}),V^+(\boldsymbol{A},\boldsymbol{b})]}_{\boldsymbol{\eta}^\top\boldsymbol{\mu},\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta}}(\boldsymbol{\eta}^\top\boldsymbol{y}) \quad | \quad \boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b} \sim \mathrm{Unif}(0,1),$$

where $F^{[a,b]}_{\mu,\sigma^2}(x)$ *is the cumulative distribution function (CDF) of a truncated normal distribution with mean* μ *and variance* σ^2 *truncated at [a,b]. Given that* $\boldsymbol{\alpha} = \boldsymbol{A}\frac{\boldsymbol{\Sigma}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta}}$, *the lower and upper truncation points are given by*

$$V^-(\boldsymbol{A},\boldsymbol{b}) = \max_{j:\boldsymbol{\alpha}_j<0} \frac{b_j - (\boldsymbol{A}\boldsymbol{y})_j}{\boldsymbol{\alpha}_j} + \boldsymbol{\eta}^\top\boldsymbol{y}, \quad V^+(\boldsymbol{A},\boldsymbol{b}) = \min_{j:\boldsymbol{\alpha}_j>0} \frac{\boldsymbol{b}_j - (\boldsymbol{A}\boldsymbol{y})_j}{\boldsymbol{\alpha}_j} + \boldsymbol{\eta}^\top\boldsymbol{y}.$$

Lower and upper truncation points

Selective Inference with Kernels

- Lasso is a linear model; we cannot select nonlinearly related features.
- Using kernel based statistics measure
  - Hilbert-Schmidt Independence Criterion (Independence test)
  - Maximum Mean Discrepancy (two-sample test)
  - Kernel Stein Discrepancy (goodness of fit test)

$$H_{0,m} : \mathrm{HSIC}(X_m, Y) = 0 \mid \mathcal{S} \text{ was selected},$$
$$H_{1,m} : \mathrm{HSIC}(X_m, Y) \neq 0 \mid \mathcal{S} \text{ was selected}.$$

Can we directly combine kernel based estimators and the polyhedral lemma? NO!

Hilbert-Schmidt Independence Criterion Inference (hsicInf) Yamada et al, AISTATS 2018

- HSIC based PSI algorithm HSIC SIS + Selective Inference
  - Non-parametric approach (can handle nonlinearly related data)
  - Can deal with Multi-variate output, multi-class, multi-label.
  - Extremely simple!

$$H_{0,m} : \mathrm{HSIC}(X_m, Y) = 0 \mid \mathcal{S} \text{ was selected},$$
$$H_{1,m} : \mathrm{HSIC}(X_m, Y) \neq 0 \mid \mathcal{S} \text{ was selected}.$$

- We set $\boldsymbol{\mu} = \mathbf{0}$ and estimate the covariance $\boldsymbol{\Sigma}$ from data.

Hilbert-Schmidt Independence Criterion (HSIC) Gretton et al. ALT 2005

- Definition of HSIC (population)

$$\begin{aligned}\mathrm{HSIC}(X,Y) = &\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',\boldsymbol{y},\boldsymbol{y}'}[K(\boldsymbol{x},\boldsymbol{x}')L(\boldsymbol{y},\boldsymbol{y}')] \\ &+ \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[K(\boldsymbol{x},\boldsymbol{x}')]\mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'}[L(\boldsymbol{y},\boldsymbol{y}')] \\ &- 2\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{x}'}[K(\boldsymbol{x},\boldsymbol{x}')]\mathbb{E}_{\boldsymbol{y}'}[L(\boldsymbol{y},\boldsymbol{y}')]\right]\end{aligned}$$

- Estimator of HSIC
  - U-statistics (unbiased, Song et al. JMLR 2012)

$$\widehat{\mathrm{HSIC}}_U(X,Y) = \frac{1}{n(n-3)}\left[\mathrm{tr}(\bar{\boldsymbol{K}}\bar{\boldsymbol{L}}) + \frac{\mathbf{1}_n^\top \bar{\boldsymbol{K}}\mathbf{1}_n\mathbf{1}_n^\top \bar{\boldsymbol{L}}\mathbf{1}_n}{(n-1)(n-2)} - \frac{2}{n-2}\mathbf{1}_n^\top \bar{\boldsymbol{K}}\bar{\boldsymbol{L}}\mathbf{1}_n\right]$$

$$\bar{\boldsymbol{K}} = \boldsymbol{K} - \mathrm{diag}(\boldsymbol{K})$$

U-statistics is not normal under the null

Block HSIC estimator
(Zhang et al. Statistics and computing 2018)

- Disjointly divided block data $\{\{(\boldsymbol{x}_i^{(b)}, \boldsymbol{y}_i^{(b)})\}_{i=1}^{B}\}_{b=1}^{n/B}$
- Block HSIC estimation

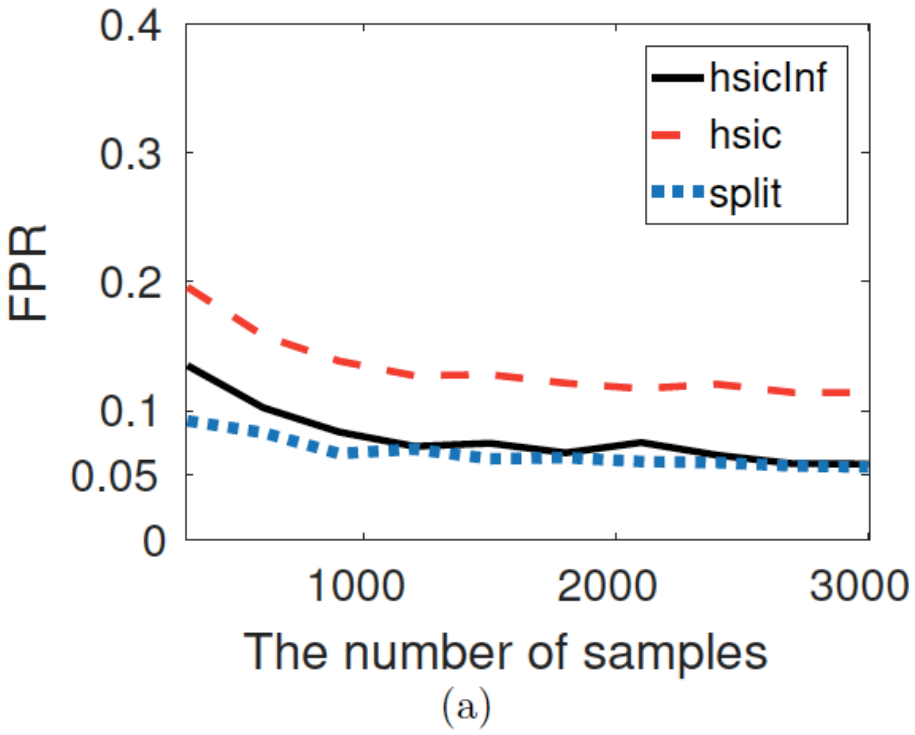
$$\widehat{\mathrm{HSIC}}(X,Y) = \frac{B}{n}\sum_{b=1}^{n/B} \widehat{\mathrm{HSIC}}_U^{(b)}$$

$\widehat{\mathrm{HSIC}}_U^{(b)}$: U-statistics of the b-th block

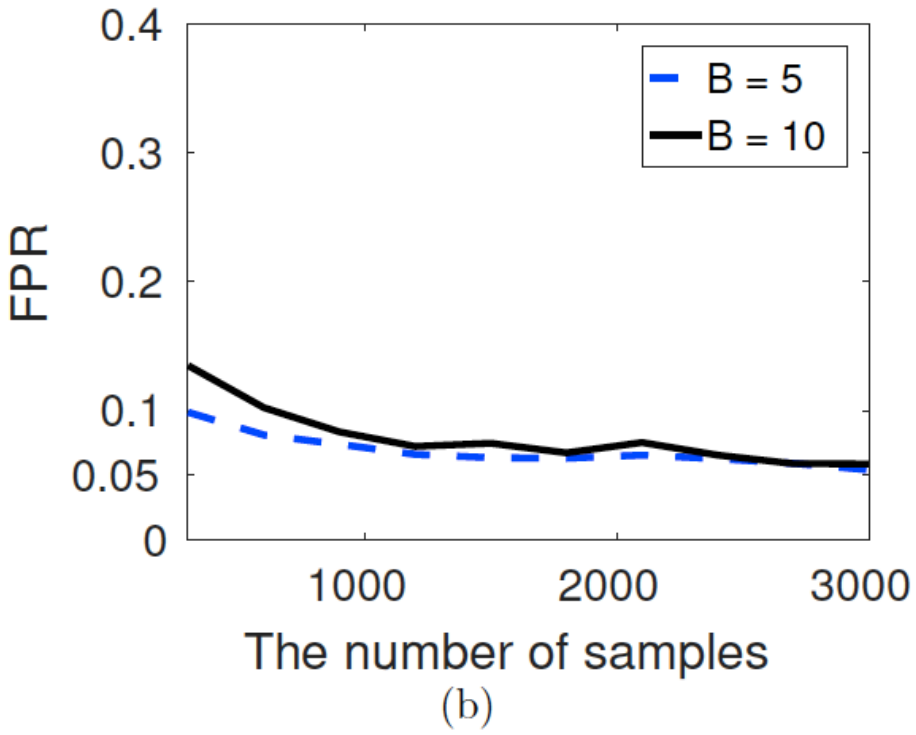
- Block HSIC is asymptotically normal by CLT!
- We can also use incomplete U-stats (ICLR 2019)

False Discovery Control (hsicInf)

- Input and output are independent
- Significance level (5%)



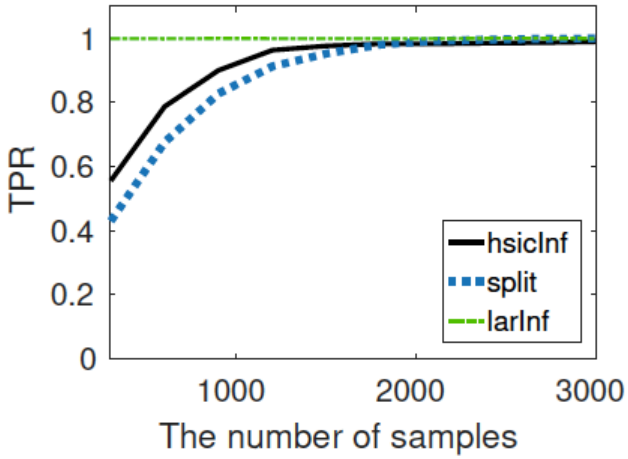
(a)



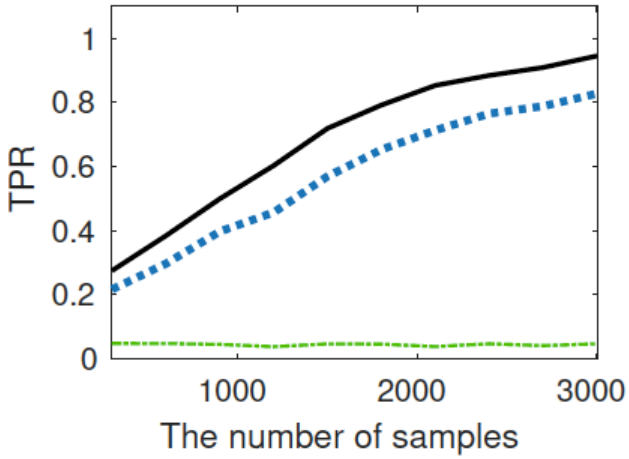
(b)

True Positive Rate comparison

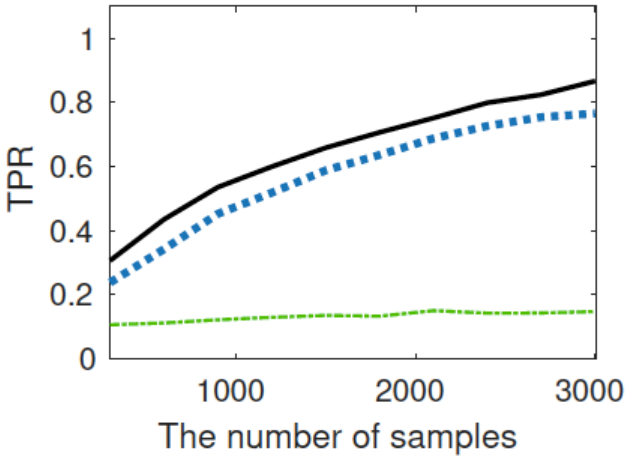
Linear data | Additive Nonlinear | Non-additive Nonlinear



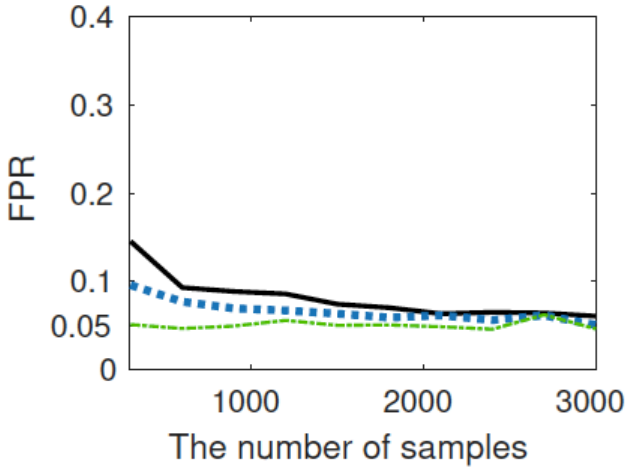
(a) Linear.



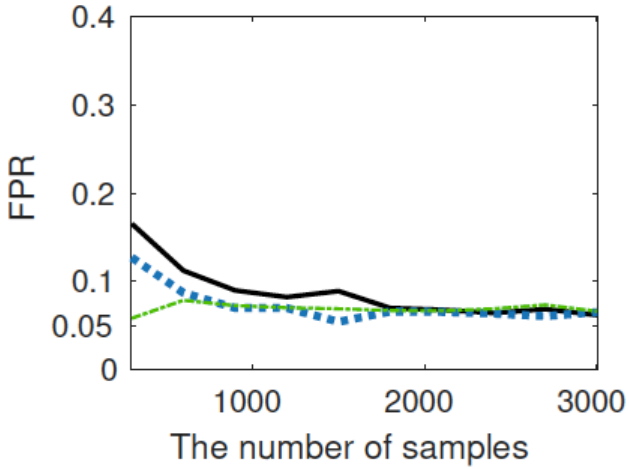
(b) Additive Non-linear.



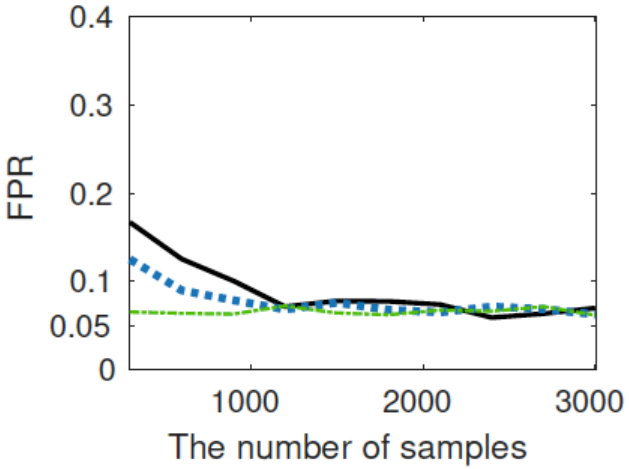
(c) Non-additive Non-linear.



(d) Linear.



(e) Additive Non-linear.



(f) Non-additive Non-linear.

Real-data Analysis 1

Yamada et al. AISTATS 2018

- Turkey student evaluation dataset
  - Output's are 1,2,…,5 (non-Gaussian)

https://archive.ics.uci.edu/ml/datasets/Turkiye+Student+Evaluation

Feature description	p-value
Q28: The Instructor treated all students in a right and objective manner.	< 0.001
Q17: The Instructor arrived on time for classes.	0.033
Q13: The Instructor's knowledge was relevant and up to date.	0.018
Q22: The Instructor was open and respectful of the views of students about the course.	0.042
Q21: The Instructor demonstrated a positive approach to students.	0.033
Q18: The Instructor has a smooth and easy to follow delivery/speech.	0.186
Q23: The Instructor encouraged participation in the course.	0.037
Q26: The Instructor's evaluation system effectively measured the course objectives.	0.176
Q2: The course aims and objectives were clearly stated at the beginning of the period.	0.452
Q20: The Instructor explained the course and was eager to be helpful to students.	0.463

Real-data Analysis 2

Lim et al. AISTATS 2020

- Divorce Predictors data set Data Set
  https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set

	p-values	
	MultiSel-HSIC	PolySel-HSIC
My argument with my wife is not calm.	<0.01	<0.01
Fights often occur suddenly.	<0.01	0.41
I can insult my spouse during our discussions.	<0.01	0.09
When fighting with my spouse, I usually use expressions such as you always or you never.	<0.01	0.17
We're compatible with my wife about what love should be.	<0.01	0.43
My wife and most of our goals are common.	<0.01	0.25
I feel aggressive when I argue with my wife.	<0.01	0.22
We're starting a fight before I know what's going on.	<0.01	<0.01
I can use negative statements about my wife's personality during our discussions.	<0.01	0.05
I hate my wife's way of bringing it up.	<0.01	<0.01
I enjoy our holidays with my wife.	0.12	0.27
When we fight, I remind her of my wife's inadequate issues.	0.13	0.04
When I argue with my wife, it will eventually work for me to contact him.	0.16	0.14
I know my wife's hopes and wishes.	0.77	0.56
I can use offensive expressions during our discussions.	0.94	0.89

Maximum Mean Discrepancy Inference (mmdInf) Yamada et al. ICLR 2019

- Two sample test with feature selection based on MMD
  - Testing whether two distributions are the same
- Feature selection: One distribution from class 1 and the other from class 2.

$$H_{0,m} : \mathrm{MMD}(X_m, X'_m) = 0 \mid \mathcal{S} \text{ was selected},$$
$$H_{1,m} : \mathrm{MMD}(X_m, X'_m) \neq 0 \mid \mathcal{S} \text{ was selected}.$$

- Dataset selection:

$$H_{0,m} : \mathrm{MMD}(X^*, X_m) = 0 \mid \mathcal{S} \text{ was selected},$$
$$H_{1,m} : \mathrm{MMD}(X^*, X_m) \neq 0 \mid \mathcal{S} \text{ was selected}.$$

Incomplete MMD estimator

- Maximum Mean Discrepancy (Population)

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \mathbf{E}_{\boldsymbol{x},\boldsymbol{x}'}[k(\boldsymbol{x}, \boldsymbol{x}')] - 2\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}}[k(\boldsymbol{x}, \boldsymbol{y})] + \mathbf{E}_{\boldsymbol{y},\boldsymbol{y}'}[k(\boldsymbol{y}, \boldsymbol{y}')]$$

- U-statistics: $\mathrm{MMD}_u^2[\mathcal{F}, \boldsymbol{X}, \boldsymbol{Y}] = \dfrac{1}{m(m-1)} \sum_{i \neq j} h(\boldsymbol{u}_i, \boldsymbol{u}_j),$

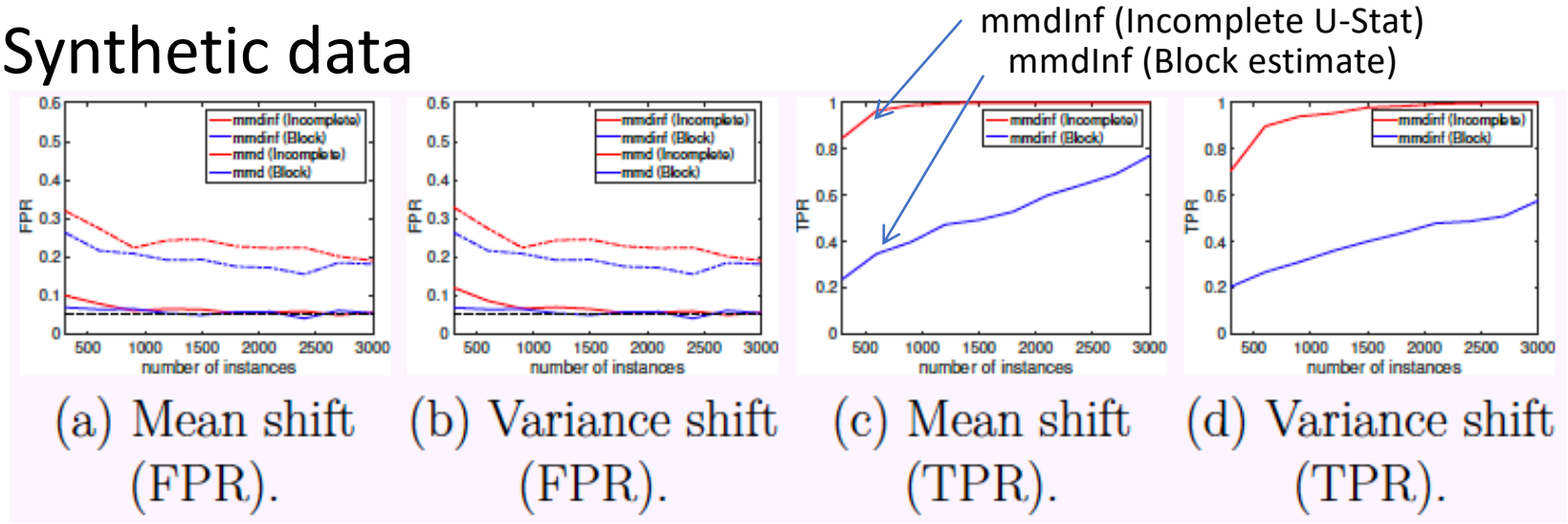
U-stat kernel $h(\boldsymbol{u}, \boldsymbol{u}') = k(\boldsymbol{x}, \boldsymbol{x}') + k(\boldsymbol{y}, \boldsymbol{y}') - k(\boldsymbol{x}, \boldsymbol{y}') - k(\boldsymbol{x}', \boldsymbol{y})$

- Incomplete U-statistics:

$$\mathrm{MMD}_{inc}^2[\mathcal{F}, \boldsymbol{X}, \boldsymbol{Y}] = \frac{1}{\ell} \sum_{(i,j) \in \mathcal{D}} h(\boldsymbol{u}_i, \boldsymbol{u}_j)$$

Post Selection Inference with MMD

- Synthetic data



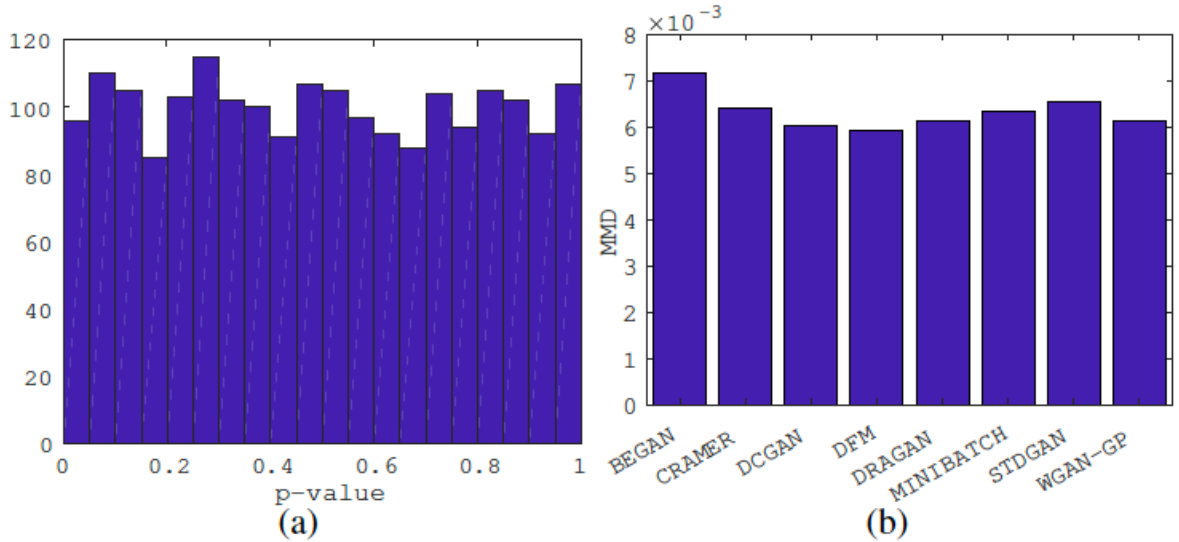
- Feature selection (binary classification)

Table 1: Post selection inference experimental results for real-world datasets. The average TPR and FPR over 200 trials are reported.

Datasets	d	n	Linear-Time		Block						Incomplete					
					$B = 5$		$B = 20$		$B = 50$		$r = 0.5$		$r = 5$		$r = 10$	
			TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Diabetis	8	768	0.05	0.05	0.06	0.05	0.12	0.06	0.21	0.11	0.07	0.07	0.41	0.07	0.55	0.07
Wine (Red)	11	4898	0.15	0.06	0.14	0.02	0.29	0.05	0.33	0.07	0.21	0.06	0.64	0.06	0.74	0.06
Wine (White)	11	1599	0.09	0.06	0.09	0.05	0.13	0.06	0.18	0.06	0.09	0.06	0.43	0.06	0.53	0.06
Australia	7	690	0.08	0.06	0.08	0.05	0.15	0.07	0.39	0.23	0.08	0.05	0.52	0.06	0.69	0.07

Selecting the best GAN family

- Which GAN is good? $\{\boldsymbol{x}_i^{(\ell)}\}_{i=1}^{n_\ell}$ $\{\boldsymbol{x}_i^*\}_{i=1}^{n}$
  - We compare the samples generated from GAN and the original images.
  - BEGAN, DCGAN, STDGAN, Cramer GAN, DFM, DRAGAN, Minibatch Discrimination, WGAN-GP



Kernel Stein Discrepancy Inference

Lim et al. NeurIPS 2019

Jenning

- Goodness of fit test based on Kernel Stein Discrepancy

  Testing whether the samples are fit to model

$$H_{0,m} : \mathrm{KSD}(P_i, R) > \mathrm{KSD}(P_m, R) \mid P_m \text{ was selected},$$
$$H_{1,m} : \mathrm{KSD}(P_i, R) < \mathrm{KSD}(P_m, R) \mid P_m \text{ was selected}.$$

- Code: https://github.com/jenninglim/model-comparison-test

Conclusion

- Selective inference (Feature selection + inference)
- We have proposed several selective testing framework
  - Selective independence test (AISTATS 2018, 2020)
  - Selective two-sample test (ICLR 2019)
  - Selective goodness of fit test (NeurIPS 2019)
- Many testing framework can be written by selective inference.

Acknowledgement

These works cannot be done without the wonderful collaborators and interns! Thanks!!

Reference

- Yamada, M., Umezu, Y., Fukumizu, K., & Takeuchi, I. Post Selection Inference with Kernels, AISTATS 2018
- Yamada, M., Wu, D., Tsai Y-H-H., Hirofumi Ota, Salakhutdinov, R., Takeuchi, I & Fukumizu, K. Post Selection Inference with Incomplete Maximum Mean Discrepancy Estimator, ICLR 2019
- Lim, J., Yamada, M., Schoelkopf, B., Jitkrittum, W. Kernel Stein Tests for Multiple Model Comparison. NeurIPS 2019.
- Lim, J., Yamada, M., Jitkrittum, W., Terada, Y., Matsui, S., Shimodaira, H. More Powerful Selective Kernel Tests for Feature Selection. AISTATS 2020
- https://github.com/jenninglim/multiscale-features
- https://github.com/jenninglim/model-comparison-test