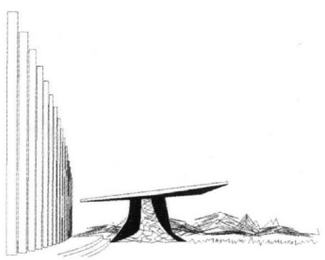
IBM **Research**



Sobolev Independence Criterion

Youssef Mroueh

Tom Sercu

Mattia Rigotti

Inkit Padhi

Cicero Nogueira dos Santos

Motivation: Reproducibility Crisis and


6C THE TUSCALOOSA NEWS FRIDAY, DECEMBER 8, 1995

Spaghetti sauce and pizza fight cancer

By PAUL RECER
The Associated Press

WASHINGTON — Men who eat at least 10 servings a week of tomato-based foods are up to 45 percent less likely to develop prostate cancer, Harvard University researchers report.

A six-year study of the dietary habits of 47,000 men found that pizza, spaghetti sauce and other foods rich in tomato substantially lowered the risk of prostate cancer. A report on the study will be published Wednesday in the Journal of the National Cancer Institute.

Dr. Edward Giovannucci, a researcher at the Harvard School of Public Health, said that tomato-based products and strawberries were the only foods of 46 fruits and vegetables checked that seemed to have a protective effect against prostate cancer. And the benefits of tomatoes came from several forms of the food: sauce, juice, raw and even when cooked into pizza.

"We found that more was better," said Giovannucci. He said men who had 10 or more servings a week had a 45 percent reduction in the rate of prostate cancer, while those who ate only four to seven servings of the tomato-based food had a 20 percent reduction in the cancer.

Spaghetti sauce was the most common tomato-based food eaten by the men in the study group.

Giovannucci cautioned that the findings should not be interpreted to mean that men should load up on tomato products.

"These findings support the idea that people should eat a variety of fruits and vegetables," he said. Nutrients in other foods may be protective against other types of cancers, said Giovannucci.

The researcher said that tomatoes are rich in an anti-oxidant called lycopene. In fact, tomatoes and tomato products accounted for almost 90 percent of the lycopene in the diet of the men studied.

Giovannucci said the study found that other nutrients, such as beta carotene and vitamin A, had no effect on the rate of prostate cancer. However, he said, these nutrients may be protective against other types of cancers as some research has suggested.

The protective effects of tomatoes had been suggested in some earlier studies of prostate cancer rates. Giovannucci said that the research had found that prostate cancer was less common in southern Mediterranean countries, such as Italy and Greece, where tomato-based foods are a major part of the diet.

In the new study, Giovannucci said that cooked tomato products seemed to be more protective than either juice or raw tomatoes. It could be, he said, that when tomatoes are heated during cooking, the cells burst and release more lycopene.

The study is based on dietary survey taken of 47,000 males in the health professions between the ages of 40 and 75. The first survey was taken in 1986 and the men were followed and periodically re-examined.

At the end of the study, in 1992, there were 812 cases of prostate cancer among those in the study group. The researchers then compared the dietary history of those who developed prostate cancer with those who did not.

U.S. medical illiteracy widespread, hazardous

By LINDSEY TANNER
The Associated Press

CHICAGO — Warning: Inability to read prescription bottle labels could be hazardous to your health.

That's the alert researchers have issued after finding that a disturbing number of patients are medically illiterate — unable to read basic written medical instructions.

Of 2,659 patients studied at two large public hospitals, nearly 30 percent had inadequate functional health literacy. The percentage was much higher among elderly patients.

The researchers cited a man who suffered serious side effects after taking asthma medicine a pharmacist mistakenly gave

Illiterate patients harbor a tremendous amount of shame and may try to hide the problem from even their closest relatives, he said. Doctors may be able to ease the problem by simply asking patients compassionately about their reading skills, Williams said.

One doctor involved in the study tells of a male diabetic whose blood sugar levels were always out of control. When the doctor finally asked, "Can you read?" the man broke down and cried" and said he couldn't follow instructions he'd received in a diabetes education class, Williams said.

"Physicians and nurses need to be sensitized to the problem and look for it," he said.

Study subjects were queried in

Educational screenings

The Attention and Cognitive Disorders (ACD) Research Foundation will conduct an educational screening clinic, testing for dyslexia, attention deficit hyperactivity disorder, and other types of learning disabilities, from 9 a.m. to 7 p.m. Dec. 16 at Holy Spirit Catholic School.

If a child is found to need an individualized approach to education, then you can enroll him/her in the 13-week Shedd After School Tutorial program beginning in January 1996. They will conduct another screening at 4 p.m. Dec. 19 at Holy Spirit.

Dr. Paul Thompson, director of the Shedd Dyslexia Foundation in Mayfield, Ky., will direct this testing. More information on either screening is available by calling the ACD Research Foundation office at 333-7733.

Thoughtful gifts seniors can use all year long

DEAR READERS: Well, it seems as though we just finished polishing off the Thanksgiving leftovers, and it's time to start

ABIGAIL VAN BUREN
DEAR ABBY

incomes would appreciate a basket of practical goodies. How about small tins of tuna, chicken and ham? Also, crackers, instant cof-

cards, condolence cards and "congratulations on the new baby" cards.

Should you be tempted to recycle a lovely but useless gift still in its original box, make sure the card to you is not still in the box.

Never give a pet to anyone unless you are absolutely certain that new

Another good idea for those living alone on a fixed income: a gift certificate for some kind of service such as window washing, carpet cleaning, taxi rides, barber shop, beauty parlor, or dinner or lunch at their favorite place. And (don't laugh) a gift certificate entitling

BRIDGE

"A heretic . . . is a fellow who disagrees with you regarding something neither of you know anything about." — William Cowper Brann

"I had to guess which ace you held," explained West. "Had I switched to clubs, we would have beaten him. How was I to know?"

"I gave you my lowest spade at trick one," offered East. "I was hoping you might read it as a suit-preference signal."

Why were both views off base?

Smith took his spade 10 and dislodged West's heart ace on the second round. West could see his only chance was to find a minor-suit ace with East; how was he to tell which suit to lead?

That East had played major-suit deuces at every turn had nothing to do with suit preference. His play in both suits was aimed at giving count. One should avoid interpreting any card as a suit-preference signal unless the card can have no other meaning in the minds of both players.

Guessing wrong, East led a diamond at trick four. At least he wouldn't be risking giving South a free finesse. He didn't, but he did give South his game. South took his red-suit winners and gratefully conceded the rest.

Why should West have "guessed" correctly? Had South held the A-Q of clubs, the extra club trick would not have surrendered game (one spade, four hearts and two clubs). However, if South held the diamond ace, the diamond switch would (and did) concede the game (one spade, four hearts and four diamonds).

Often, thorough analysis and logical thought convert a guess into a sure thing.

ANSWER: Three no-trump. With no apparent major-suit fit, play in the nine-trick no-trump game.

Send bridge questions to The Aces, P.O. Box 12363, Dallas, Texas, 75225, with SASE for reply.

NORTH
♠7 3
♥K Q 10 9 7
♦K Q J 6
♣8 3

WEST
♠A Q 9 6 5
♥A 4
♦10 5 4
♣K 6 2

EAST
♠8 4 2
♥6 5 2
♦3 2 2
♣A 7 5 4

SOUTH
♠K J 10
♥J 8 3
♦A 10 7
♣Q J 10 9

Vulnerable: Both
Dealer: West

The bidding:

West	North	East	South
1♠	2♥	Pass	2♠*
Pass	3♦	Pass	3 NT
Pass	Pass	Pass	

*forcing, asking definition

Opening lead: Spade six

BID WITH THE ACES

South holds:
♠A Q 9 6 5
♥A 4
♦10 5 4
♣K 6 2

North	South
1♥	1♠
1NT	?

- Amgen could only replicate 6 of 53 studies they considered landmarks in basic cancer science
- HealthCare could only replicate about 25% of 67 seminal studies
- Systematic attempts to replicate widely cited priming experiments have failed

Early report (Kaplan, '08)

50% of Phase III FDA studies ended in failure

Causality, not only simple linear correlations !

Need to control for False Discoveries

[Pictures/ slides: from Emmanuel Candes Presenta

Feature Selection
with False Discovery Rate Control

- Given a high dimensional input $x \in \mathbb{R}^d$ and response y
- **Goal:** Find a subset of features $S \subset \{1 \dots d\}$ such that

$$x_{S^c} \perp y \,|\, x_S$$

- S is a markov blanket
- Find S in $\text{TPR} := \mathbb{E}\left[\frac{\#\{i : i \in \hat{S} \cap \mathcal{S}\}}{\#\{i : i \in \mathcal{S}\}}\right]$ $\text{FDR} := \mathbb{E}\left[\frac{\#\{i : i \in \hat{S} \backslash \mathcal{S}\}}{\#\{i : i \in \hat{S}\}}\right]$ control

$\hat{S}$ Candidate Set

$\mathcal{S}$ Ground truth Set

Feature Selection
with Linear Models and Sparsity

$$\min_{\beta} ||Y - \beta X||_F^2 + \lambda ||\beta||_{\ell_0}$$

$$\min_{\beta} ||Y - \beta X||_F^2 + \lambda_1 ||\beta||_{\ell_1} + \lambda_2 ||\beta||_{\ell_2}^2$$

Elastic Net

Use $|\beta_j|$ in feature selection ;

Problem: this models only linear relationships between x and y

[Hastie et al 2001]

Feature Selection
an Information theoretic point view

- Let D be an Integral Probability Metric associated with a function space $\mathscr{F}$, i.e for two distributions p, q:

$$D(p, q) = \sup_{f \in \mathscr{F}} \mathbb{E}_{x \sim p} f(x) - \mathbb{E}_{x \sim q} f(x).$$

- With $p = p_{xy}$ and $q = p_x p_y$ this becomes a generalized definition of Mutual Information.

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

Feature selection as finding a sparse gate that maximizes "mutual information "

$$\sup_{w, \|w\|_{\ell_0} \leq s} D(p_{w \odot x, y}, p_{w \odot x} p_y),$$

$$\sup_{w} \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(w \odot x, y) - \mathbb{E}_{p_x p_y} f(w \odot x, y) - \lambda ||w||_0$$

This amounts to controlling the sparsity of gradients $\frac{\partial f(x)}{\partial x_j}$ on the support of a dominant measure μ

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

$$\text{(SIC):} \quad \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \lambda P_S(f),$$

where $P_S(f)$ is a penalty that controls the sparsity of the gradient of the witness function f on the support of the measures.

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

$$\text{(SIC):} \quad \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \lambda P_S(f),$$

where $P_S(f)$ is a penalty that controls the sparsity of the gradient of the witness function f on the support of the measures.

$$\Omega_{\ell_0}(f) = \#\{j | \mathbb{E}_{(x,y)\sim\mu} \left| \frac{\partial f(x,y)}{\partial x_j} \right|^2 = 0\}$$

ℓ_0

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

$$\text{(SIC):} \quad \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \lambda P_S(f),$$

where $P_S(f)$ is a penalty that controls the sparsity of the gradient of the witness function f on the support of the measures.

$$\Omega_{\ell_0}(f) = \#\{j | \mathbb{E}_{(x,y)\sim\mu} \left| \frac{\partial f(x,y)}{\partial x_j} \right|^2 = 0\}$$

ℓ_0

[Rosasco et al 2013]

$$\Omega_S(f) = \sum_{j=1}^{d_x} \sqrt{\mathbb{E}_{(x,y)\sim\mu} \left| \frac{\partial f(x,y)}{\partial x_j} \right|^2}.$$

ℓ_1

Sobolev Independence Criterion

Non Linear Sparsity Inducing Penalty

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

Mutual information **Gradient Sparsity** **L2 penalty**

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

Mutual information **Gradient Sparsity** **L2 penalty**

For $\mathscr{F} = \{f(x) = \langle \beta, x \rangle\}$ we obtain elastic Net Regularization

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

Mutual information **Gradient Sparsity** **L2 penalty**

For $\mathscr{F} = \{f(x) = \langle \beta, x \rangle\}$ we obtain elastic Net Regularization

$$\Omega_S(f) = \sum_{j=1}^{d_x} \sqrt{\mathbb{E}_{(x,y)\sim\mu} \left| \frac{\partial f(x,y)}{\partial x_j} \right|^2}.$$

We use $\mu = p_x p_y$

Sobolev Independence Criterion
Non Linear Sparsity Inducing Penalty

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

Mutual information **Gradient Sparsity** **L2 penalty**

For $\mathscr{F} = \{f(x) = \langle \beta, x \rangle\}$ we obtain elastic Net Regularization

$$\Omega_S(f) = \sum_{j=1}^{d_x} \sqrt{\mathbb{E}_{(x,y)\sim\mu} \left| \frac{\partial f(x,y)}{\partial x_j} \right|^2}.$$

We use $\mu = p_x p_y$

Use non linear functions and $\tilde{\beta}_j = \sqrt{\mathbb{E}_{(x,y)\sim\mu} \left| \frac{\partial f(x,y)}{\partial x_j} \right|^2}$ in feature selection!

Sobolev Independence Criterion

η-trick and a tractable optimization problem

Sobolev Independence Criterion

η-trick and a tractable optimization problem

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

Mutual information **Gradient Sparsity** **L2 penalty**

Sobolev Independence Criterion

η-trick and a tractable optimization problem

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

Mutual information **Gradient Sparsity** **L2 penalty**

- Problems with the cost function: The Gradient penalty has an expectation preceded by square root non linearity. This will introduce biases in stochastic optimization !

- Square root is non smooth at zero (derivative discontinuous at zero)

- Alleviate those issues using **perturbation** and the η **-trick** *[Bach 2011]*

Sobolev Independence Criterion

η-trick and a tractable optimization problem

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

- η **-trick**

$$\left(\Omega_{S,\varepsilon}(f)\right)^2 = \inf\{\sum_{j=1}^{d_x} \frac{\mathbb{E}_{p_x p_y} \left|\frac{\partial f(x,y)}{\partial x_j}\right|^2 + \varepsilon}{\eta_j} : \eta, \eta_j > 0, \sum_{j=1}^{d_x} \eta_j = 1\},$$

For $\varepsilon \to 0$,Optimize on f and η! All expectations are linear!

Sobolev Independence Criterion

η-trick and a tractable optimization problem

$$\mathrm{SIC}_{(L_1)^2}(p_{xy}, p_x p_y) = \sup_{f \in \mathscr{F}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y) - \frac{\lambda}{2} \left(\Omega_S(f)\right)^2 - \frac{\rho}{2} \mathbb{E}_\mu f^2(x,y).$$

- η **-trick**

$$\left(\Omega_{S,\varepsilon}(f)\right)^2 = \inf\{\sum_{j=1}^{d_x} \frac{\mathbb{E}_{p_x p_y} \left|\frac{\partial f(x,y)}{\partial x_j}\right|^2 + \varepsilon}{\eta_j} : \eta, \eta_j > 0, \sum_{j=1}^{d_x} \eta_j = 1\},$$

For $\varepsilon \to 0$,Optimize on f and η! All expectations are linear!

η_j defines influence scores of the coordinates that we can use for feature selection!

At optimum

$$\eta_j = \frac{\sqrt{\mathbb{E}_{p_x p_y} \left|\frac{\partial f(x,y)}{\partial x_j}\right|^2 + \varepsilon}}{\sum_{k=1}^{d_x} \sqrt{\mathbb{E}_{p_x p_y} \left|\frac{\partial f(x,y)}{\partial x_k}\right|^2 + \varepsilon}}$$

Sobolev Independence Criterion

η-trick and a tractable optimization problem

$$\mathrm{SIC}_{(L_1)^2,\varepsilon}(p_{xy}, p_x p_y) = -\inf\{L_\varepsilon(f,\eta) : f \in \mathscr{F}, \eta_j, \eta_j > 0, \sum_{j=1}^{d_x} \eta_j = 1\}$$

where $L_\varepsilon(f,\eta) = -\Delta(f, p_{xy}, p_x p_y) + \frac{\lambda}{2}\sum_{j=1}^{d_x} \frac{\mathbb{E}_{p_x p_y}\left|\frac{\partial f(x,y)}{\partial x_j}\right|^2 + \varepsilon}{\eta_j} + \frac{\rho}{2}\mathbb{E}_{p_x p_y} f^2(x,y)$,
and $\Delta(f, p_{xy}, p_x p_y) = \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x p_y} f(x,y)$.

Sobolev Independence Criterion
Empirical Estimate

Given samples $\{(x_i, y_i), i = 1, \ldots, N\}$ from the joint probability distribution p_{xy} and iid samples $\{(x_i, \tilde{y}_i), i = 1, \ldots, N\}$ from $p_x p_y$,

$$\widehat{\mathrm{SIC}}_{(L_1)^2,\varepsilon}(p_{xy}, p_x p_y) = -\inf\{\hat{L}_\varepsilon(f, \eta) : f \in \mathscr{F}, \eta_j, \eta_j > 0, \sum_{j=1}^{d_x} \eta_j = 1\}$$

where $\hat{L}_\varepsilon(f, \eta) = -\hat{\Delta}(f, p_{xy}, p_x p_y) + \frac{\lambda}{2} \sum_{j=1}^{d_x} \frac{\frac{1}{N} \sum_{i=1}^{N} \left| \frac{\partial f(x_i, \tilde{y}_i)}{\partial x_j} \right|^2 + \varepsilon}{\eta_j} + \frac{\rho}{2} \sum_{i=1}^{N} f^2(x_i, \tilde{y}_i)$,
and main the objective $\hat{\Delta}(f, p_{xy}, p_x p_y) = \frac{1}{N} \sum_{i=1}^{N} f(x_i, y_i) - \frac{1}{N} \sum_{i=1}^{N} f(x_i, \tilde{y}_i)$.

Convex Sobolev Independence Criterion

Estimate in RKHS, Joint Convexity and Interpretability of SIC

$$\mathscr{F} = \{f | f(x,y) = \langle u, \Phi_\omega(x,y) \rangle, \|u\|_2 \leq \gamma\},$$

where $\Phi_\omega : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$ is a fixed finite dimensional feature map.

Theorem 1 (Interpretability of Convex SIC). *SIC $_\varepsilon$ is jointly convex in (u, η). Let (u^*, η^*) be the unique solution (limit as $\varepsilon \to 0$). Define $f^*(x,y) = \langle u^*, \Phi_\omega(x,y) \rangle$, and $\|f^*\|_{\mathscr{F}} = \|u^*\|$. We have that*

$$\begin{aligned} SIC_{(L^1)^2}(p_{xy}, p_x p_y) &= \frac{1}{2}\left(\mathbb{E}_{p_{xy}} f^*(x,y) - \mathbb{E}_{p_x p_y} f^*(x,y)\right) \\ &= \frac{\lambda}{2}\left(\sum_{j=1}^{d_x} \sqrt{\mathbb{E}_{p_x p_y} |\frac{\partial f^*(x,y)}{\partial x_j}|^2}\right)^2 + \frac{\rho}{2}\mathbb{E}_{p_x p_y} f^{*,2}(x,y) + \frac{\tau}{2}||f^*||^2_{\mathscr{F}}. \end{aligned}$$

Moreover, $\sqrt{\mathbb{E}_{p_x p_y} |\frac{\partial f^*(x,y)}{\partial x_j}|^2} = \eta_j^* \Omega_{S,L_1}(f^*)$ and $\sum_{j=1}^{d_x} \eta_j = 1$. The terms η_j^* can be seen as quantifying how much dependency as measured by SIC can be explained by a coordinate j. Ranking of η_j^* can be used to rank influence of coordinates.

Convex Sobolev Independence Criterion

Estimate in RKHS, Joint Convexity and Interpretability of SIC

- SIC is then a form of mutual information that decomposes on the contribution of coordinates (influence scores)!

- **SIC is an interpretable measure of Mutual Information !**

- For convex SIC thanks to smoothness and joint convexity , Alternating optimization on u and eta , or Block Coordinate Descent methods are convergent!

Neural Sobolev Independence Criterion

Gradient Regularization of Deep ReLU Networks with no Biases

$\mathscr{F}_{ReLu} = \{f | f(x,y) = \langle u, \Phi_\omega(x,y) \rangle, \text{ where } \Phi_\omega(x,y) = \sigma(W_L \dots \sigma(W_2 \sigma(W_1[x,y]))), u \in \mathbb{R}^m, \Phi_\omega : \mathbb{R}^{d_x+d_y} \to \mathbb{R}^m\}$, where $\sigma(t) = \max(t,0), W_j$ are linear weights.

- $f_\theta \in \mathscr{F}_{ReLu}$ is 1 homogeneous in its inputs
- By Euler Theorem we have:

$$f_\theta(x,y) = \langle \nabla_x f_\theta(x,y), x \rangle + \langle \nabla_y f_\theta(x,y), y \rangle$$

- Controlling the gradient the sparsity for function that are deep Relu networks (with no biases) controls input sparsity as well !

Neural Sobolev Independence Criterion

Stochastic Block Coordinate Descent Algorithm

Algorithm 1 *(non convex)* Neural SIC(X, Y) (Stochastic BCD)

Inputs: X, Y dataset $X \in \mathbb{R}^{N \times d_x}, Y \in \mathbb{R}^{N \times d_y}$, such that $(x_i = X_{i,.}, y_i = Y_{i,.}) \sim p_{xy}$
Hyperparameters: ε,λ, τ, ρ, $\alpha_\theta, \alpha_\eta$ (learning rates)
Initialize $\eta_j = \frac{1}{d_x}, \forall j$, $\text{Softmax}(z) = e^z / \sum_{j=1}^{d_x} e^{z_j}$
for $i = 1 \dots \text{Maxiter}$ **do**
  Fetch a minibatch of size N $(x_i, y_i) \sim p_{xy}$
  Fetch a minibatch of size N $(x_i, \tilde{y}_i) \sim p_x p_y$ $\{\tilde{y}_i$ obtained by permuting rows of $Y\}$
  Stochastic Gradient step on θ:
  $\theta \leftarrow \theta - \alpha_\theta \frac{\partial \hat{L}(f_\theta, \eta)}{\partial \theta}$ {We use ADAM}
  Mirror Descent η :
  $\text{logit} \leftarrow \log(\eta) - \alpha_\eta \frac{\partial \hat{L}(f_\theta, \eta)}{\partial \eta}$
  $\eta \leftarrow \text{Softmax}(\text{logit})$ {stable implementation of softmax}
end for
Output: f_θ, η

FDR control with SIC- HRT:

SIC as a Statistics in FDR Control with Holdout Randomization Tests

How to decide which features to keep based on the ranking of η_j? How can we make this decision in a reliable way , while assessing the statistical significance and controlling the False discovery rate?

Claim 2: Hold Out Randomization Testing With Generative models and SIC

$$H_0 : x_j \perp\!\!\!\perp y \mid x_{-j}$$

$$\Updownarrow$$

$$p_{xy} = p_{x_j|x_{-j}}\ p_{y|x_{-j}}\ p_{x_{-j}}\ ?$$

The null Hypothesis: the response is conditional independent of feature j

Compare the joint to the Null hypothesis

Claim 2: Hold Out Randomization Testing With Generative models and SIC

$$H_0 : x_j \perp\!\!\!\perp y \mid x_{-j}$$

$$\Updownarrow$$

$$p_{xy} = \boxed{p_{x_j|x_{-j}}}\, p_{y|x_{-j}}\, p_{x_{-j}} \quad ?$$

The null Hypothesis: the response is conditional independent of feature j

Compare the joint to the Null hypothesis

Problem we don't have access to the conditional distribution of features!

FDR control with SIC- HRT:

SIC as a Statistics in FDR Control with Holdout Randomization Tests

For each feature j test for conditional independence:

- Run SIC and obtain the critic f
- Compute on a holdout set the expectation of the critic **on the joint** (H_1)
- Now we need to test the critic against the Null Hypothesis
- $H_0 : x_j \perp y | x_{-j}$, conditional independence
- We need to simulate observations from the null : **Use pertained conditional generators** $G(x_j|x_{-j})$
- For each feature j compute p-values p_j on the mean of the critic on the simulated null versus H_1
- Use Benjamini-Hochberg for accepting/rejecting the dependent tests

Hold Out Randomization Testing

Null Hypothesis Simulation with Generative models

$$H_0 : x_j \perp\!\!\!\perp y \mid x_{-j}$$

$$\Updownarrow$$

$$p_{xy} = p_{x_j|x_{-j}}\ p_{y|x_{-j}}\ p_{x_{-j}}\ ?$$

The null Hypothesis: the response is conditional independent of feature j

Compare the joint to the Null hypothesis

Hold Out Randomization Testing

Null Hypothesis Simulation with Generative models

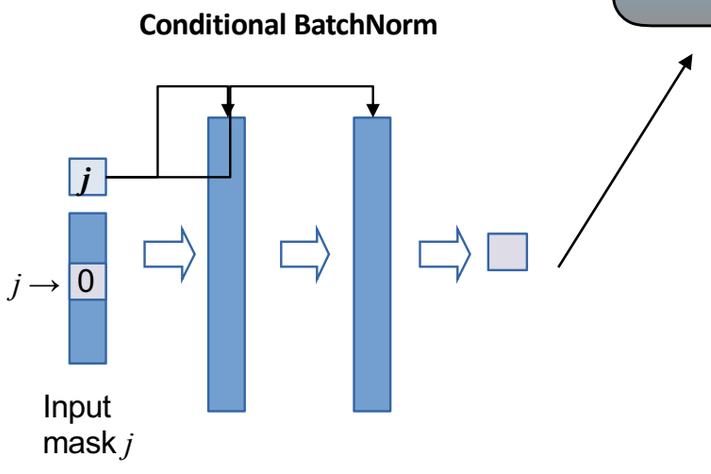
$$H_0 : x_j \perp\!\!\!\perp y \mid x_{-j}$$

$$\Updownarrow$$

$$p_{xy} = p_{x_j|x_{-j}}\ p_{y|x_{-j}}\ p_{x_{-j}}\ ?$$

The null Hypothesis: the response is conditional independent of feature j

Compare the joint to the Null hypothesis



[de Vries et al. 2017]

Hold Out Randomization Testing

Null Hypothesis Simulation with Generative models

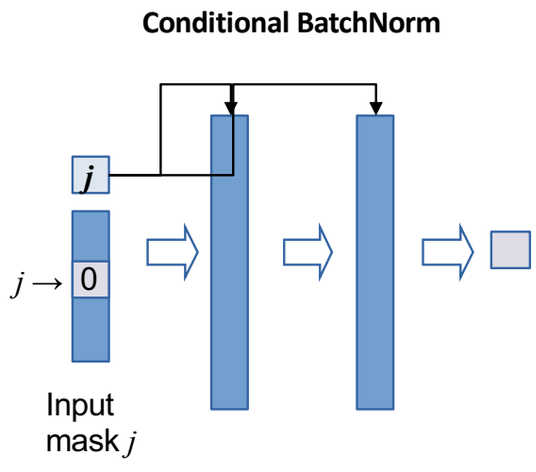
$$H_0 : x_j \perp\!\!\!\perp y \mid x_{-j}$$

$$\Updownarrow$$

$$p_{xy} = p_{x_j|x_{-j}}\ p_{y|x_{-j}}\ p_{x_{-j}}\ ?$$

The null Hypothesis: the response is conditional independent of feature j

Compare the joint to the Null hypothesis



[de Vries et al. 2017]

Train a conditional Generative model to sample from the null Hypothe

Hold Out Randomization Testing

Null Hypothesis Simulation with Generative models

$$\text{For each feature } j \text{ , for } r = 1 \dots M \text{ repetitions do :}$$

Joint Holdout set

(x_i, y_i)

Simulate H_0 with Generative model

$(x_{-j,i}, G(x_{-j,i}, j), y_i)$

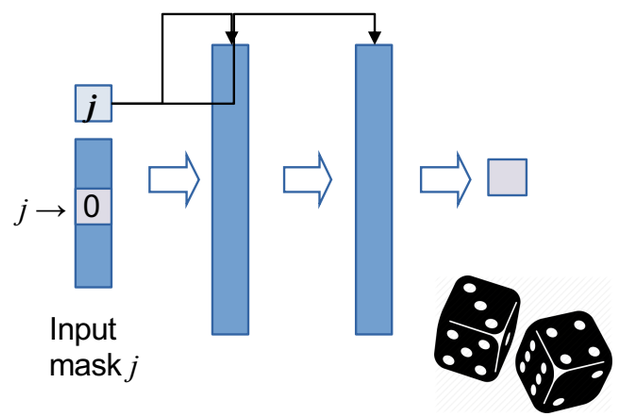
f(x,y)

Pre-trained SIC witness function

S_j=mean of f on holdout

S_j^r=mean of f on H_0

Conditional BatchNorm



[de Vries et al. 2017]

[HRT introduced in Tansey et al 2018]

Our innovation is using it with Generative models and SIC

Hold Out Randomization Testing

Null Hypothesis Simulation with Generative models

Joint Holdout set

(x_i, y_i)

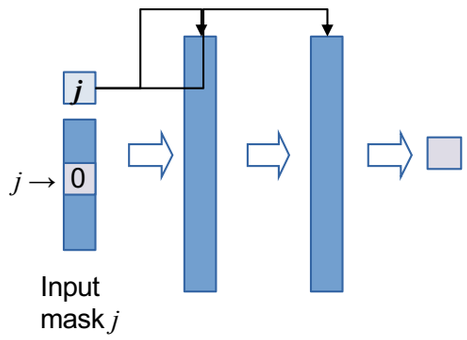
Simulate H_0
with Generative model

$(x_{-j,i}, G(x_{-j,i}, j), y_i)$

f(x,y)

Pre-trained SIC witness function

Conditional BatchNorm



[de Vries et al. 2017]

Hold Out Randomization Testing

Null Hypothesis Simulation with Generative models

Joint Holdout set

(x_i, y_i)

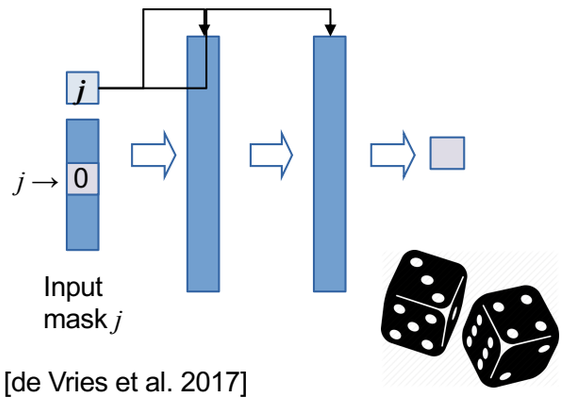
Simulate H_0
with Generative model

$(x_{-j,i}, G(x_{-j,i}, j), y_i)$

f(x,y)

Pre-trained SIC witness function

Conditional BatchNorm



[de Vries et al. 2017]

Hold Out Randomization Testing

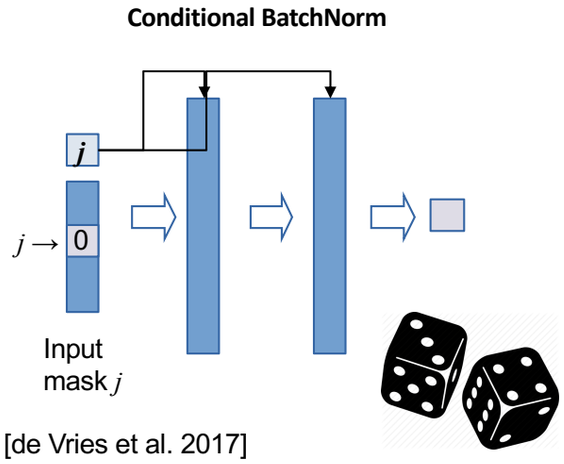
Null Hypothesis Simulation with Generative models

Joint Holdout set

(x_i, y_i) → f(x,y) *Pre-trained SIC witness function* → S_j=mean of f on holdout

Simulate H_0 with Generative model

$(x_{-j,i}, G(x_{-j,i}, j), y_i)$ → f(x,y) *Pre-trained SIC witness function* → S_j^r=mean of f on H_0



[de Vries et al. 2017]

Hold Out Randomization Testing

Null Hypothesis Simulation with Generative models

Joint Holdout set

(x_i, y_i)

Simulate H_0
with Generative model

$(x_{-j,i}, G(x_{-j,i}, j), y_i)$

f(x,y)
Pre-trained SIC witness function

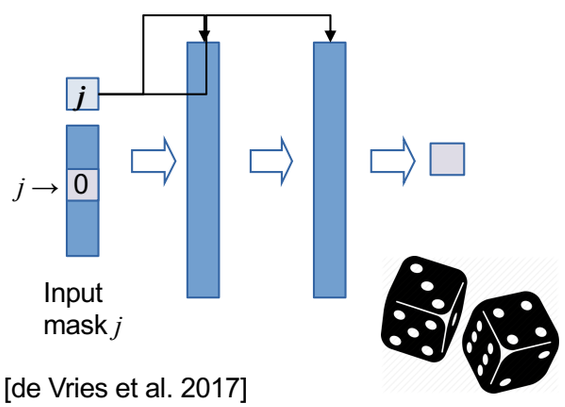
S_j=mean of f on holdout

S_j^r=mean of f on H_0

$p_j = \frac{1}{M+1}(1 + \sum_{r=1}^{M} \mathbb{1}_{S_j^r > S_j})$

p-values

Conditional BatchNorm



[de Vries et al. 2017]

Hold Out Randomization Testing

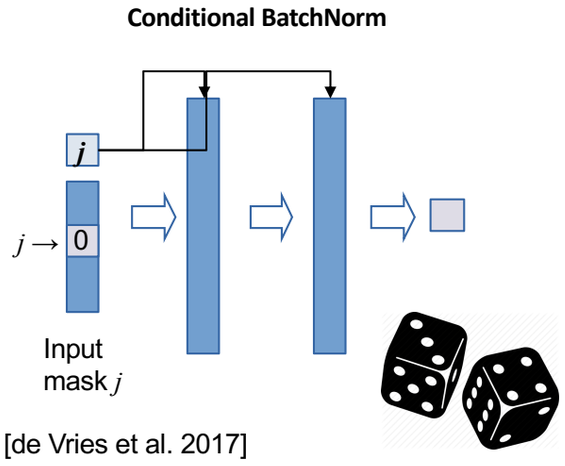
Null Hypothesis Simulation with Generative models

Joint Holdout set

(x_i, y_i)

Simulate H_0
with Generative model

$(x_{-j,i}, G(x_{-j,i}, j), y_i)$



[de Vries et al. 2017]

f(x,y)
Pre-trained SIC witness function

S_j=mean
of f on holdout

S_j^r=mean
of f on H_0

$p_j = \frac{1}{M+1}(1 + \sum_{r=1}^{M} \mathbb{1}_{S_j^r > S_j})$

p-values

Benjamini Hochberg Procedure

Accept **Reject**

FDR control with SIC- HRT:

[HRT introduced in Tansey et al 2018

SIC as a Statistics in FDR Control with Holdout Randomization Tests

Algorithm 1 HRT With SIC (X, Y)

Inputs: $D_{train} = (X_{tr}, Y_{tr})$, a Heldout set $D_{\text{Holdout}} = (X, Y)$, features Cutoff K
SIC: $(f_{\theta^*}, \eta_*) = \text{SIC}(D_{train})$
Score of witness on Hold out : $S^* = \text{MEAN}(f_{\theta^*}(X, Y))$
Conditional Generators Pre-trained conditional Generator : $G(x_{-j}, j)$ predicts $X_j|X_{-j}$
Shortlist : $I = \text{INDEXTOPK}(\eta)$
$\{p-$ values for $j \in I$; randomizations tests$\}$
for $j \in I$ **do**
  for $r = 1 \dots R$ **do**
    Construct $\tilde{X}$, $\tilde{X}_{.,k} = X_{.,k} \forall k \neq j$ and $\tilde{X}_{.,j} = G(X_{-j}, j)$ {Simulate Null Hyp.}
    $S_{j,r} = \text{MEAN}(f_{\theta^*}(\tilde{X}, Y))$ {Score of witness function on the Null}
  end for
  $p_j = \frac{1}{R+1}\left(1 + \sum_{r=1}^{R} 1_{S_j^r \geq S^*}\right)$
end for
discoveries =**BH**(p,targetFDR) {Benjamini-Hochberg Procedure}
Output: discoveries

FDR control with SIC- Knockoffs

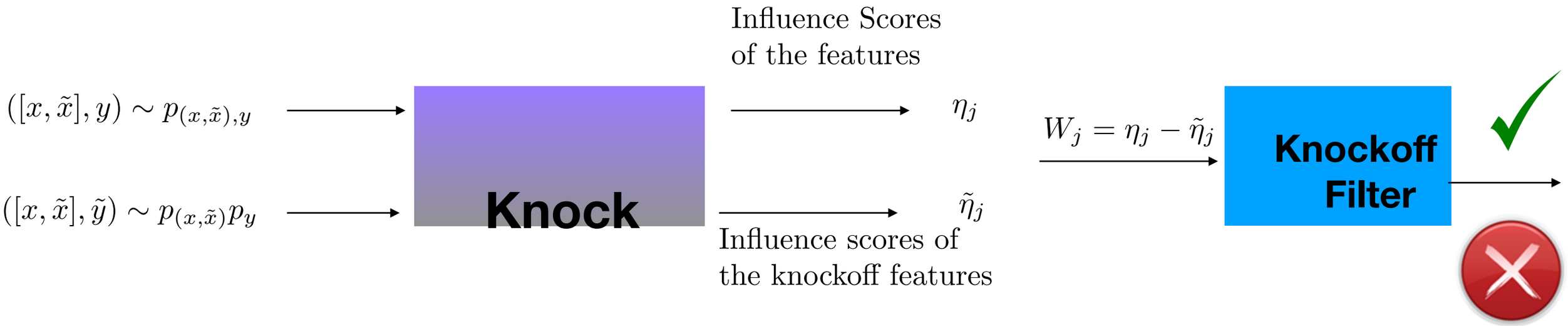
SIC as a Statistics in FDR Control with Knockoffs

- $(x, \tilde{x})_{\text{Swap}(j)} = (x, \tilde{x})$ (in distribution)
- $\tilde{x} \perp y | x$
- $\tilde{x}$ is a control variable : metaphor from biology a genome is a control for another, here we create fake genomes for control

[Knockoffs introduced in Candes et al 2018]

FDR control with SIC- Knockoffs

SIC as a Statistics in FDR Control with Knockoffs



FDR control with SIC- Knockoffs

SIC as a Statistics in FDR Control with Knockoffs

Algorithm 1 Model-X Knockoffs FDR control with SIC

Inputs: $D_{train} = (X_{tr}, Y_{tr})$, Model-X knockoff features $\tilde{X} \sim \text{ModelX}(X_{tr})$, target FDR q

Train SIC: $(f_{\theta^*}, \eta) = \text{SIC}([X_{tr}, \tilde{X}], Y)$, where $[X_{tr}, \tilde{X}]$ is the concatenation of X_{tr} and knockoffs $\tilde{X}$

for $j = 1, \ldots, d_X$ **do**

Compute importance score of j feature: $W_j = \eta_j - \eta_{j+d_x}$, where η_{j+d_X} is the η of feature knockoff $\tilde{X}_j$

end for

Compute threshold $\tau > 0$ by setting

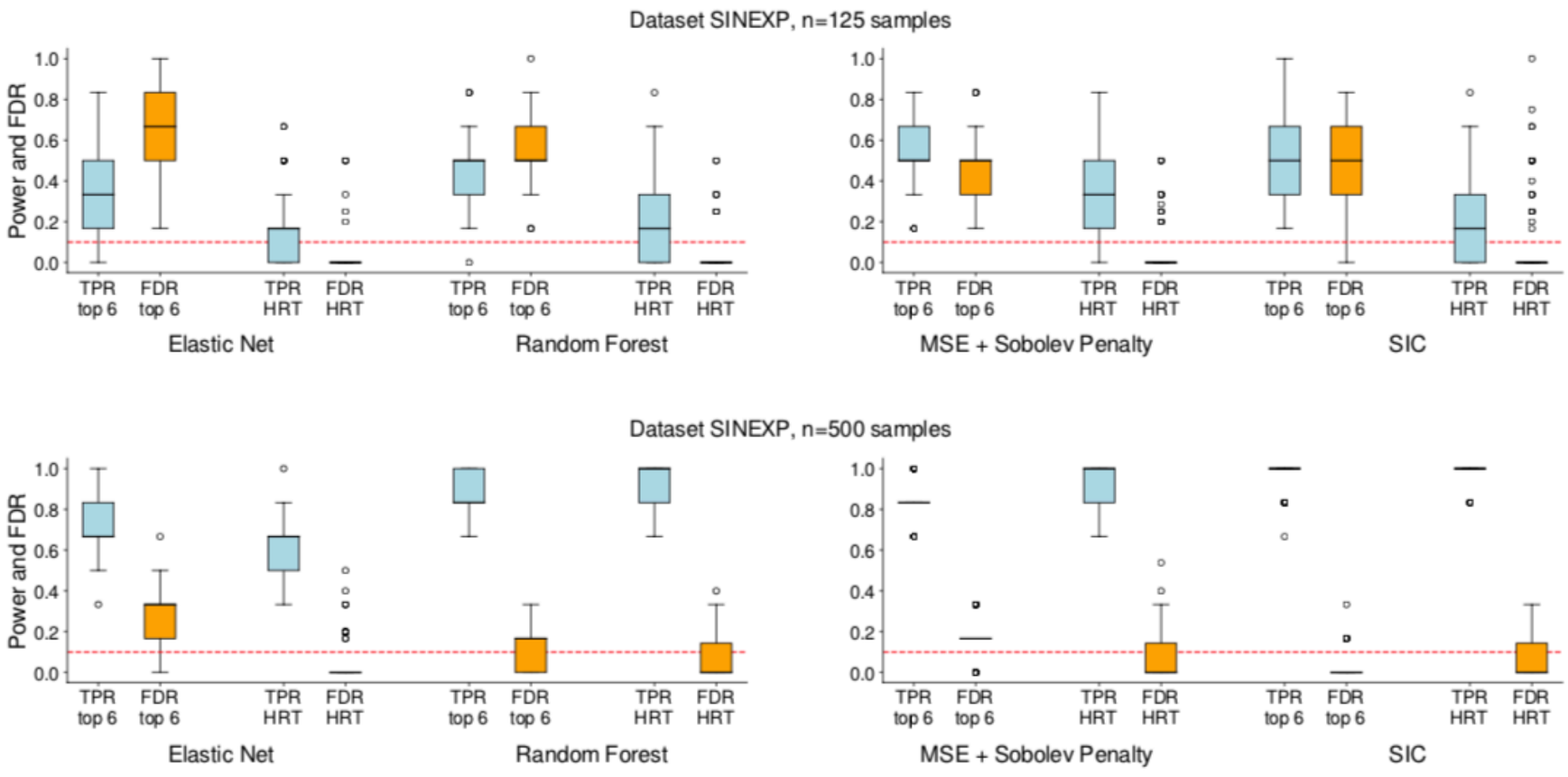
$\tau = \min \left\{ t > 0 : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q \right\}$

Output: discoveries $\{j : W_j > \tau\}$

Synthetic Experiments

$$y = \sin(x_1(x_1 + x_2)) \cos(x_3 + x_4 x_5) \sin(e^{x_5} + e^{x_6} - x_2).$$

$x = (x_1, \ldots, x_{50})$, x_j uniform correlated with correlation $\rho = 0.5$



CCLE Experiments: Drug Response

	NN	RF
All 7251 features	1.160 ± 3.990	0.783 ± 0.167
Elastic-Net1 [36] top-7	0.864 ± 0.432	0.931 ± 0.215
Elastic-Net2 [8] top-10	$\mathbf{0.663 \pm 0.161}$	0.830 ± 0.190
SIC top-7	0.728 ± 0.166	0.856 ± 0.189
SIC top-10	0.706 ± 0.158	$\mathbf{0.817 \pm 0.173}$
SIC top-15	0.734 ± 0.168	0.859 ± 0.202

Table 1: CCLE results on downstream regression task. Heldout MSE for drug PLX4720 prediction based on selected features. Columns: neural network (NN) and random forest (RF) regressors.

HIV-1Drug Resistance SIC-Knockoffs

- Detecting mutations associated with resistance to a drug type.
- For our experiments we use all the three classes of drugs: Protease Inhibitors (PIs), Nucleoside Reverse Transcriptase Inhibitors (NRTIs), and Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs).
- Gaussian Knockoffs
- Geometric mean of η for a SIC ensemble : Boosted SIC

HIV-1Drug Resistance SIC-Knockoffs

Drug Class	Drug Type	Knockoff with GLM			Boosted SIC Knockoff		
		TD	FD	FDP	TD	FD	FDP
PIs	APV	19	3	**0.13**	17	5	0.22
	ATV	22	8	0.26	19	1	**0.05**
	IDV	19	12	0.38	15	3	**0.16**
	LPV	16	1	**0.05**	14	2	0.12
	NFV	24	7	0.22	19	5	**0.21**
	RTV	19	8	0.29	12	2	**0.20**
	SQV	17	4	**0.19**	14	8	0.36
NRTIs	X3TC	0	0	0	7	0	**0**
	ABC	10	1	0.09	11	1	**0.08**
	AZT	16	4	**0.2**	12	5	0.29
	D4T	6	1	0.14	8	0	**0**
	DDI	0	0	0	8	0	**0**
NNRTIs	DLV	10	13	0.56	8	10	**0.55**
	EFV	11	11	0.5	11	10	**0.47**
	NVP	7	10	**0.58**	7	11	0.611

Table 2: Comparison of applying (knockoff filter + GLM) and (Knockoff filter+Boosted SIC). For each <drug-class, drug-type> we compared the True Discoveries (TD), False Discoveries(FD) and False Discovery Proportion (FDP). Knockoff with Boosted SIC keeps FDP under control without compromising power, and succeeds in making true discoveries that GLM with knockoffs doesn't find.

Perspectives

- Learning Knockoffs and SIC jointly
- Generative models for Knockoff generation
- Auto-regressive models for learning Knockoffs
- SIC for sequences and Markov or graphical Models
- Mutual information between time series or stochastic processes in general ?