On Measuring Conditional Dependence

Mona Azadkia

London School of Economics and Political Sciences

Summary of Contribution

A new coefficient of conditional dependence:

1. It has a simple expression
2. It is fully non-parametric
3. It has no tuning parameters
4. It does not rely on estimating conditional densities or conditional characteristic functions or mutual information
5. There is absolutely no assumption on the laws of the random variables
6. It can be estimated from data very quickly, $O(n \log n)$

A New Measure of Conditional Dependence:

CODEC

(conditional dependence coefficient)

Simple Case: No Conditioning

- Y is a random variable
- $Z = (Z_1, \ldots, Z_q)$ is a random vector $(q \geq 1)$
- μ is the probability law of Y

CODEC: unconditional

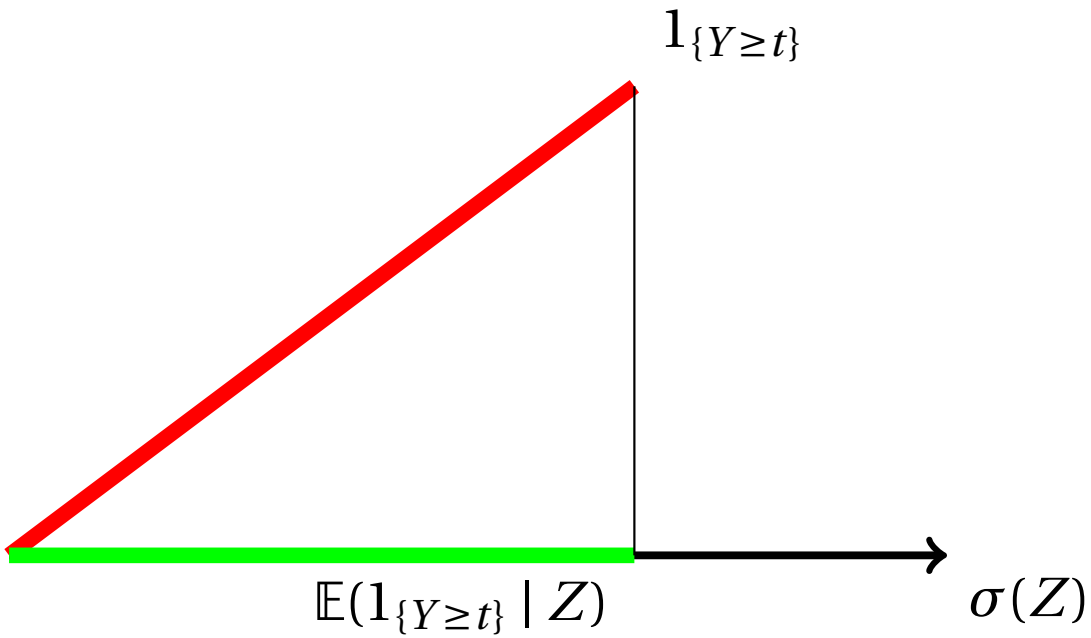
$T(Y, Z)$ gives the measure of dependence of Y on Z:

$$T(Y,Z) := \frac{\int_{\mathbb{R}} \mathrm{var}(\mathbb{E}(1_{\{Y \geq t\}} \mid Z)) d\mu(t)}{\int_{\mathbb{R}} \mathrm{var}(1_{\{Y \geq t\}}) d\mu(t)}$$

Conditional Expectation as Projection

μ is the probability law of Y.

$$T(Y,Z) = \frac{\int_{\mathbb{R}} \mathrm{var}(\mathbb{E}(1_{\{Y \geq t\}} \mid Z)) d\mu(t)}{\int_{\mathbb{R}} \mathrm{var}(1_{\{Y \geq t\}}) d\mu(t)}$$



Simple Case: A Closer Look

μ is the probability law of Y.

$$T(Y,Z) = \frac{\int_{\mathbb{R}} \mathrm{var}(\mathbb{E}(1_{\{Y \geq t\}} \mid Z)) d\mu(t)}{\int_{\mathbb{R}} \mathrm{var}(1_{\{Y \geq t\}}) d\mu(t)}$$

Conditioning **does not increase** the variance

$$\mathrm{var}(\mathbb{E}(1_{\{Y \geq t\}} \mid Z)) \quad \leq \quad \mathrm{var}(1_{\{Y \geq t\}})$$

- $T(Y,Z) \in [0,1]$
- $T(Y,Z) = 0$ **if and only if** $Y \perp Z$
- $T(Y,Z) = 1$ **if and only if** Y is a function of Z
- For $W = (W_1, \cdots, W_{q'})$ another random vector

$$T(Y,Z) \leq T(Y,(Z,W))$$

- $T(Y,Z)$ is **invariant** under one-to-one transformations of Y and Z
- $T(Y,Z)$ is not symmetric, (consider $Y = Z^2$)

CODEC: Conditional Dependence Coefficient

- Y is a random variable
- $Z = (Z_1, \ldots, Z_q)$ is a random vector ($q \geq 1$)
- $X = (X_1, \ldots, X_p)$ is a random vector ($p \geq 0$)
- μ be the probability law of Y

CODEC: general case

When Y is not a function of X,

$$T(Y, Z \mid X) := \frac{\int_{\mathbb{R}} \mathbb{E}(\mathrm{var}(\mathbb{E}(1_{\{Y \geq t\}} \mid Z, X) \mid X)) d\mu(t)}{\int_{\mathbb{R}} \mathbb{E}(\mathrm{var}(1_{\{Y \geq t\}} \mid X)) d\mu(t)}$$

Properties

Y is not a function of X, and μ be the probability law of Y

$$T(Y,Z \mid X) := \frac{\int_{\mathbb{R}} \mathbb{E}(\mathrm{var}(\mathbb{E}(1_{\{Y \geq t\}} \mid Z,X) \mid X))d\mu(t)}{\int_{\mathbb{R}} \mathbb{E}(\mathrm{var}(1_{\{Y \geq t\}} \mid X))d\mu(t)}$$

- $T(Y,Z \mid X) \in [0,1]$
- $T(Y,Z \mid X) = 0$ **if and only if** $Y \perp Z \mid X$
- $T(Y,Z \mid X) = 1$ **if and only if** Y is a function of Z given X
- $T(Y,Z \mid X)$ is **invariant** under one-to-one transformations
- $T(Y,Z \mid X)$ is a **non-random** quantity that depends on the joint law of (Y,Z,X)
- If $p = 0$, $T(Y,Z \mid X) = T(Y,Z)$, unconditional dependence

The Estimator

The Estimator

- Sample of n i.i.d. copies $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ of the triple (Y, X, Z)
- $X_{N(i)}$ is the closest neighbor of X_i w.r.t. Euclidean distance
- $(Z_{M(i)}, X_{M(i)})$ is the closest neighbor of (Z_i, X_i) w.r.t. Euclidean distance
- $R_i = \sum_{j=1}^{n} 1_{\{Y_j \leq Y_i\}}$ is the **rank** of Y_i, the number of j such that $Y_j \leq Y_i$
- Let $T_n(Y, Z \mid X)$ be the estimate of $T(Y, Z \mid X)$

$$T_n(Y, Z \mid X) := \frac{\sum_{i=1}^{n} (\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\})}{\sum_{i=1}^{n} (R_i - \min\{R_i, R_{N(i)}\})}$$

T_n: Easy and Fast!

$$T_n(Y, Z \mid X) := \frac{\sum_{i=1}^{n} (\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\})}{\sum_{i=1}^{n} (R_i - \min\{R_i, R_{N(i)}\})}.$$

- Nearest neighbors indices $N(i)$ and $M(i) \Rightarrow O(n \log n)$ (dimension is fixed)
- Ranks $R_i \Rightarrow O(n \log n)$
- No knowledge of law of (Y, X, Z) is needed
- No need to estimate the densities

Consistency of The Estimator

Theorem

Suppose that Y is not a function of X. Then as $n \to \infty$, $T_n(Y, Z \mid X) \to T(Y, Z \mid X)$ almost surely.

There are no hidden assumptions.

Rate of convergence

Theorem

Suppose that $p \geq 1$ and that the assumptions (A1) and (A2) hold with some β and C. Then, as $n \to \infty$,

$$T_n - T = O_P\left(\frac{(\log n)^{p+q+\beta+1}}{n^{1/(p+q)}}\right).$$

There are nonnegative real numbers β and C, C_1, C_2 such that

(A1) for any $t \in \mathbb{R}$, $x, x' \in \mathbb{R}^p$ and $z, z' \in \mathbb{R}^q$,

$$\begin{aligned}&|P(Y \geq t|X = x, Z = z) - P(Y \geq t|X = x', Z = z')| \\ &\leq C(1 + \|x\|^\beta + \|x'\|^\beta + \|z\|^\beta + \|z'\|^\beta)(\|x - x'\| + \|z - z'\|),\end{aligned}$$

and

$$|P(Y \geq t|X = x) - P(Y \geq t|X = x')| \leq C(1 + \|x\|^\beta + \|x'\|^\beta)\|x - x'\|.$$

(A2) for any $t > 0$, $\mathbb{P}(\|X\| \geq t)$ and $\mathbb{P}(\|Z\| \geq t) \leq C_1 e^{-C_2 t}$.

An Application: Variable Selection

Feature Ordering By Conditional Independence (FOCI)

A fully model-free forward step-wise algorithm.

- $Y \in \mathbb{R}$ is the response variable
- $X = (X_1, \cdots, X_p)$ is the vector of features

Algorithm

- $\hat{S} = \emptyset$ (set of selected variables)
- $j_1 = \text{argmax}_{i \in \{1,\cdots,p\}} T_n(Y, X_i) \Rightarrow \hat{S} = \{j_1\}$
- $j_2 = \text{argmax}_{i \neq j_1} T_n(Y, X_i \mid X_{j_1}) \Rightarrow \hat{S} = \{j_1, j_2\}$
- $j_3 = \text{argmax}_{i \neq j_1, j_2} T_n(Y, X_i \mid X_{j_1}, X_{j_2}) \Rightarrow \hat{S} = \{j_1, j_2, j_3\}$
- $\cdots$
- $(k+1)$th step is the first time s.t. $T_n(Y, X_i \mid X_{j_1}, \cdots, X_{j_k}) \leq 0 \Rightarrow$ stop!
- $\hat{S} = \{j_1, \cdots, j_k\}$.

Efficacy of FOCI

Efficacy of FOCI

- $S \subseteq \{1, \cdots, p\}$ is sufficient if Y and X_{S^c} are conditionally independent given X_S. Sufficient sets are also known as Markov Blanket
- For each set S define

$$Q(S) = \int_{\mathbb{R}} \text{var}(\mathbb{E}(1_{\{Y \geq t\}} \mid X_S)) d\mu(t)$$

- Q is monotone, If $S \subseteq S'$, then $Q(S) \leq Q(S')$
- Let δ be the smallest number such that for any **insufficient** subset S, there exist some $j \notin S$ such that $Q(S \cup \{j\}) \geq Q(S) + \delta$
- Think of δ as the smallest prediction power that be achieved by increasing the size of an insufficient set

Efficacy of FOCI

(B1) There are nonnegative real numbers β and C such that for any S of size $\leq 1/\delta + 2$ and any $t \in \mathbb{R}$ for any $t \in \mathbb{R}$, and any $x, x' \in \mathbb{R}^S$

$$|P(Y \geq t | X_S = x) - P(Y \geq t | X_S = x')|$$
$$\leq C(1 + \|x\|^\beta + \|x'\|^\beta)\|x - x'\|$$

(B2) There are positive numbers C_1 and C_2 such that for any set S of size $\leq 1/\delta + 2$ and any $t > 0$, $\mathbb{P}(\|X_S\| \geq t) \leq C_1 e^{-C_2 t}$.

Theorem

Suppose that $\delta > 0$, and that the assumptions (B1) and (B2) hold. Let $\hat{S}$ be the subset selected by FOCI with a sample of size n. There are positive real numbers L_1, L_2 and L_3 depending only on C, β, C_1, C_2 and δ such that $\mathbb{P}(\hat{S}$ is sufficient$) \geq 1 - L_1 p^{L_2} e^{-L_3 n}$.

If δ is not too close to zero, and $n \gg \log(p)$, then with high probability, FOCI chooses a sufficient set of predictors.

Example: Variable Selection, Simulated Data

Sample of size $n = 1000$ of $X = (X_1, \cdots, X_{1500})$ where X_i's are i.i.d. $N(0, 1)$ and

$$Y = X_1 X_2 + \sin(X_1 X_3)$$

Method	Selected variables
FOCI	1, 2, 3.
Forward stepwise	247 variables were selected, but 1, 2, and 3 were not in the list.
Lasso	28, 43, 68, 95, 96, 189, 241, 262, 275, 292, 351, 362, 387, 403, 490, 514, 526, 537, 560, 578, 583, 623, 635, 675, 787, 814, 834, 914, 965, 968.
Dantzig selector	No variables were selected.
SCAD	28, 43, 68, 241, 262, 292, 351, 387, 403, 537, 583, 623, 675, 814, 834, 968.

CLT under the null

Theorem (Shi, Drton, Han; 2021)

Assume $Y \in \mathbb{R}$ is continuous and independent of $\mathbb{X} \in \mathbb{R}^p$, which is absolutely continuous. Then as $n \to \infty$

$$\sqrt{n}T_n \xrightarrow{d} \mathcal{N}(0, \frac{2}{5} + \frac{2}{5}q_p + \frac{4}{5}o_p),$$

with q_p and o_p constants that only depend on p.

CLT in general

Theorem (Lin, Han; 2022)

As long as $Y \in \mathbb{R}$ is not a measurable function of $\mathbb{X}$ and both are continuous

$$(T_n - \mathbb{E}[T_n]) / \sqrt{\mathrm{Var}[T_n]} \xrightarrow{d} \mathcal{N}(0,1).$$

Boosting the power

- Lin, Han (2022) proposed using M nearest neighbor to improved the power
- Work under progress: we are trying to work out the use of general kernels to boost the power!

Conclusion

- A new measure of conditional dependence, CODEC
- Model-free
- Non-parametric
- Consistent estimator with $O(n \log n)$ computational time
- Variable selection algorithm with a stopping criteria, FOCI
- R-package FOCI is CRAN
- Work continues to tackle more problems!

Thank you!

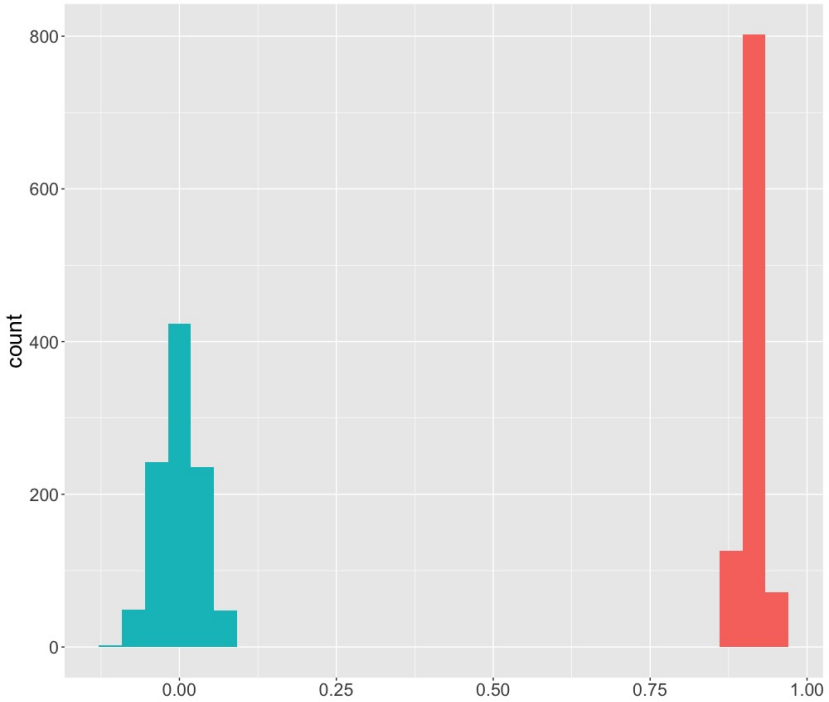
Extra Slides

Example: CODEC

Let X_1 and X_2 be i.i.d. uniform in $[0,1]$ and $Y = (X_1 + X_2) \pmod 1$

$$Y \perp X_2, \qquad Y \text{ is a function of } X_2 \mid X_1.$$

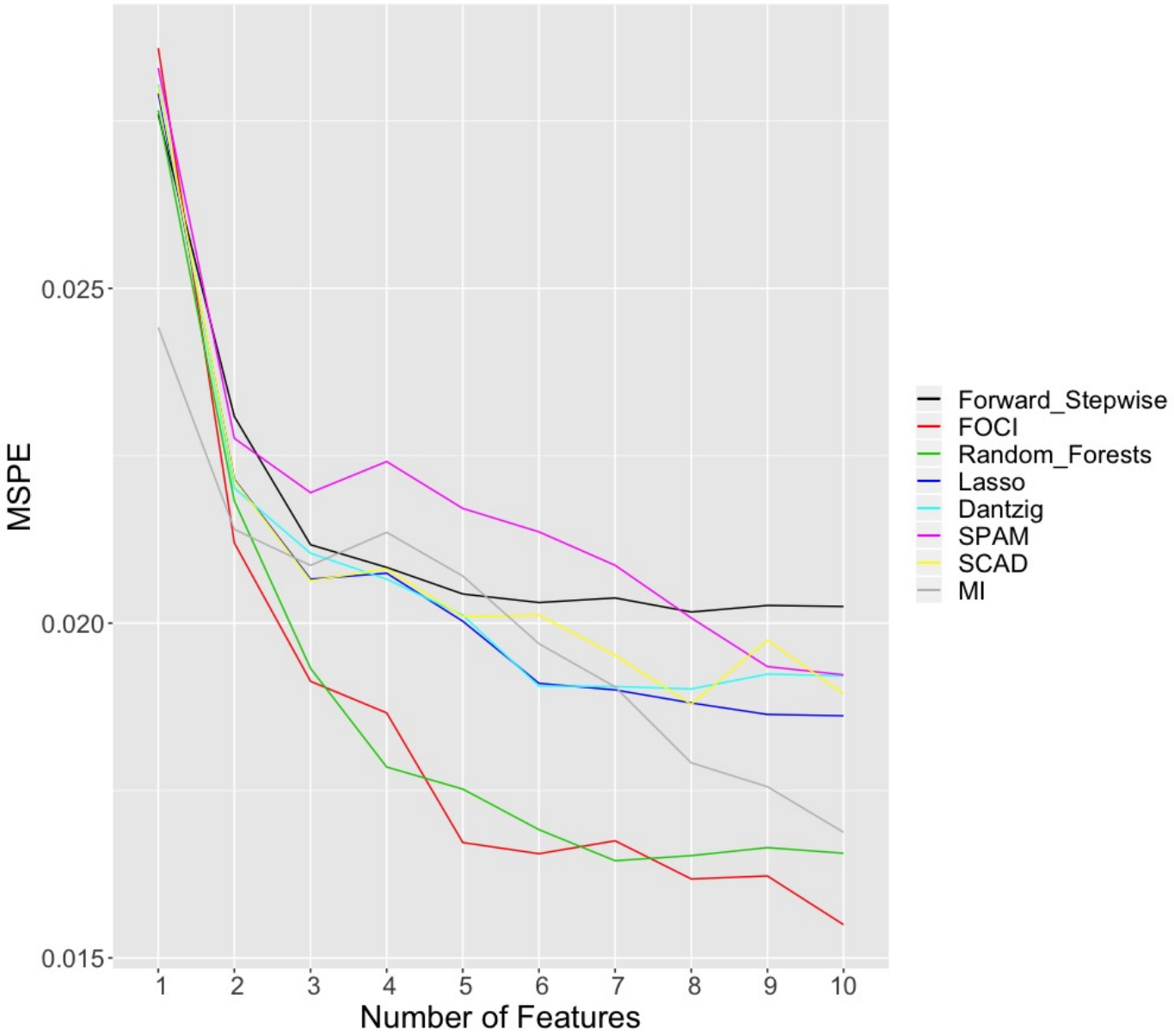
Histograms of $T_n(Y, X_2)$ and $T_n(Y, X_2 \mid X_1)$ for 1000 independent samples of size $n = 1000$.



Example: Polish Companies Bankruptcy

- Response variable Y is binary (a company is bankrupt or not)
- Sample size $n = 19967$ of $p = 64$ features
- Data has been splitted in half at random to training and test sets
- For $k \in \{1, \cdots, 10\}$ we have selected subsets of size k of the features using different variable selection techniques
- For each selected set we predicted values of Y on the test set by Random Forests

Example: Polish Companies Bankruptcy



Example: Polish Companies Bankruptcy

Sample size $n = 19967$ of $p = 64$ features and response variable Y is binary.

Method	Subset size	MSPE
FOCI	10	0.015
Forward stepwise	24	0.016
Lasso	48	0.017
Dantzig selector	27	0.017
SCAD	3	0.021

- Considered only methods with stopping rule
- Prediction method is Random Forests