Feature Ordering by Conditional Independence

Sourav Chatterjee

(Joint work with Mona Azadkia)

Model-free variable selection

- Let $\mathbf{X} = (X_1, \ldots, X_p)$ be the vector of predictors and Y be the response.
- A subset of predictors $(X_j)_{j \in S}$ is called *sufficient* if Y and $(X_j)_{j \notin S}$ are independent given $(X_j)_{j \in S}$.
- (Also known as a *Markov blanket* in the graphical models literature. Related to the concept of *sufficient dimension reduction* in classical statistics.)
- Data: n i.i.d. copies $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ of $(Y, \mathbf{X})$.
- Goal: Using the data, find a sufficient set of predictors (preferably, a small set).
- Separate goal: After finding a sufficient set of predictors, fit a predictive model.

A new variable selection algorithm: *Feature Ordering by Conditional Independence (FOCI)*

- For any $S \subseteq \{1, \ldots, p\}$ and $1 \leq i \leq n$, let $\mathbf{X}_{i,S} = (X_{i,j})_{j \in S}$.
- Let $N(i, S) =$ the k such that $\mathbf{X}_{k,S}$ is the nearest neighbor of $\mathbf{X}_{i,S}$. (Ties broken at random.)
- Let $R_i = \#\{j : Y_j \leq Y_i\}$ and $L_i = \#\{j : Y_j \geq Y_i\}$.
- Define the *estimated predictive power* of a set S as

$$P_n(S) = \sum_{i=1}^{n} \min\{R_i, R_{N(i,S)}\}$$

  if S is nonempty, and $P_n(\emptyset) = \frac{1}{n} \sum_{i=1}^{n} L_i^2$.
- Algorithm:
  - Start with $S_0 = \emptyset$.
  - Having defined S_k, choose an index $j_{k+1} \notin S_k$ that maximizes $P_n(S_k \cup \{j_{k+1}\})$ and let $S_{k+1} = S_k \cup \{j_{k+1}\}$.
  - Stop at the first k such that $P_n(S_{k+1}) \leq P_n(S_k)$ (or $k = p$).
  - Declare $\hat{S} = S_k$ to be the chosen subset.

Reference

The FOCI algorithm was proposed in:

Mona Azadkia and Sourav Chatterjee (2019). A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327.*

Features

- No tuning parameters.
- No cross-validation.
- No modeling assumptions.
- No *a priori* restriction on p.
- Fast (each step takes time $n \log n$).
- Very simple to understand and implement (no black boxes).
- Variables can be discrete or continuous.
- Provably consistent under a sparsity assumption (later).
- *But does it actually work?*

An example

- $n = 2000$, $p = 1000$, i.i.d. $N(0, 1)$ covariates.
- Model: $Y = X_{256}X_{778} + \sin(X_{256}X_{889})$.
- Note that the model is non-linear, non-monotone in each variable, and has interactions.
- This makes variable selection very difficult by methods based on linear or additive models.
- For example, none of the following methods were successful in selecting the relevant set $\{X_{256}, X_{778}, X_{889}\}$ in repeated simulations:
  - Lasso with cross-validation.
  - Dantzig selector with cross-validation.
  - Forward stepwise (with stopping by AIC).
  - SCAD with cross-validation.
  - SPAM (sparse additive models) with cross-validation.
- On the other hand, FOCI selected *exactly this subset* almost every time (and very rarely with one or two extra variables).
- Nothing special about this particular example. Any other example of this type gives the same results.

Did any other method work for this example?

- Yes. Random forests, and forward stepwise by maximizing mutual information were able to detect the importance of $\{X_{256}, X_{778}, X_{889}\}$.
- But they were much slower.
- Random forests took 15 times as much time as FOCI. (43 seconds for FOCI, 633 seconds for random forests.) Mutual information took more than 100 times as much time as FOCI due to the difficulty of estimating MI.
- Also, these methods only give an ordering of variables by importance, instead of selecting a subset. Often it's not clear how to select a subset.

What about real data?

- We tried FOCI on a number of real data examples, all from the UCI repository, and compared with a number of other methods. Some of these are reported in the paper.
- For each method, after selecting the variables, a predictive model was fitted using random forests (because that usually gives better prediction error than linear models).
- Prediction errors were estimated by testing on a test set.
- In most examples, FOCI performed comparably with random forests, and gave better results than other methods.
- There is need for more extensive testing on real data.
- One problem we observed on rare occasions is that FOCI sometimes stops too early, missing out some important variables. It is possible that there is a better stopping rule than the one we proposed, or some better way of extracting a subset from the ordering given by FOCI.

A typical chart of prediction errors

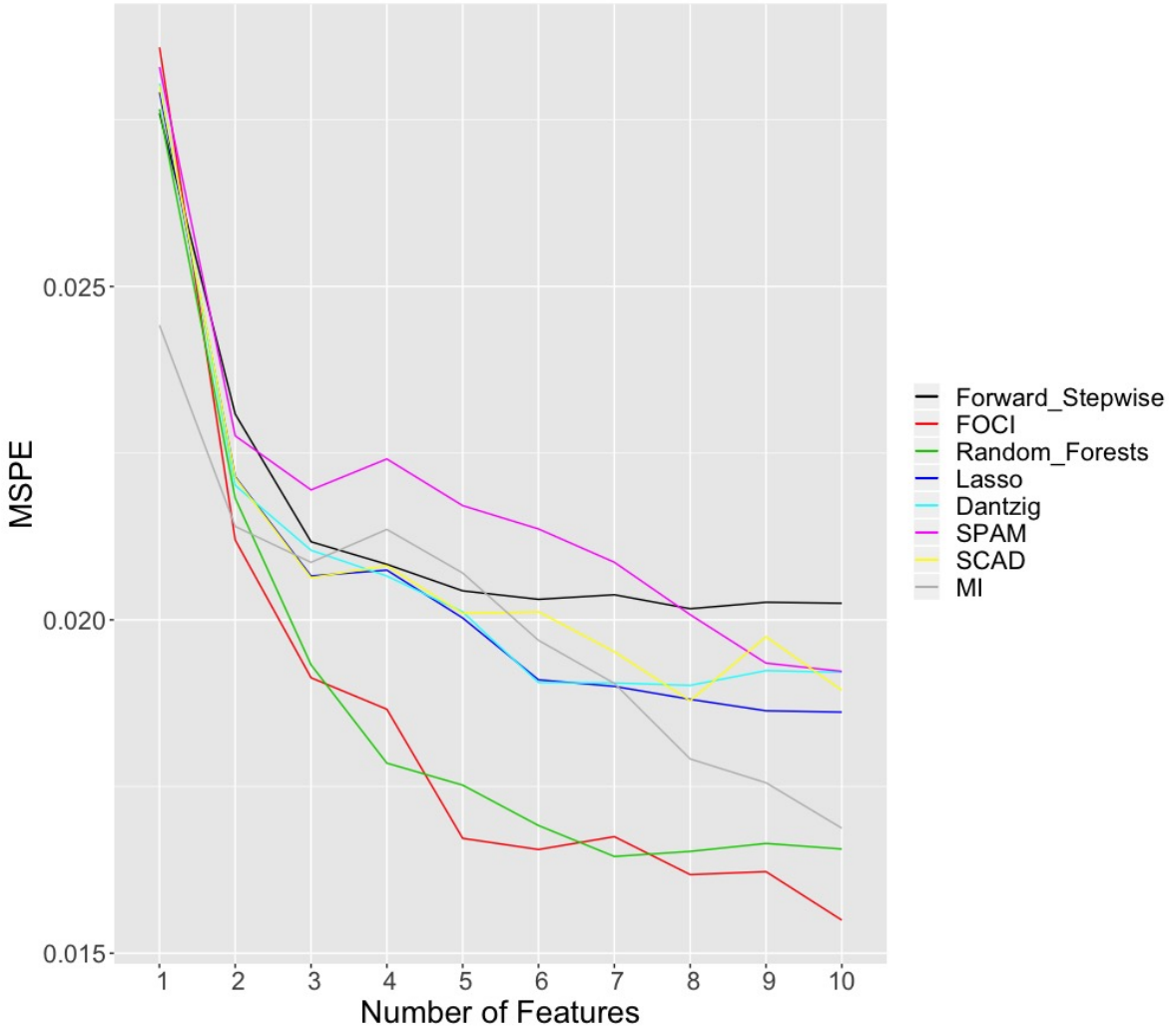

Figure: Mean squared prediction errors versus subset size for the "Polish companies bankruptcy data" from the UCI repository.

Subset size advantage

We found that in the majority of examples, FOCI selects a lesser number of variables but with better prediction error.

Table: Number of variables selected using different methods and the corresponding mean squared prediction errors in the spambase data from the UCI repository.

Method	Subset size	MSPE
FOCI	26	0.036
Forward stepwise	45	0.039
Lasso	56	0.038
Dantzig selector	51	0.038
SCAD	33	0.039

A sparsity assumption for consistency of FOCI

- For $S \subseteq \{1, \ldots, p\}$, let $\mathbf{X}_S = (X_j)_{j \in S}$.
- Let $Q(S) = \int \mathrm{Var}(\mathbb{P}(Y \geq t | \mathbf{X}_S)) d\mu(t)$, where μ is the law of Y. Note that $0 \leq Q(S) \leq 1$ always.
- Fact: If $S' \supseteq S$, then $Q(S') \geq Q(S)$, and $Q(S') = Q(S)$ if and only if Y and $\mathbf{X}_{S' \setminus S}$ are independent given $\mathbf{X}_S$.
- Thus, $Q(S') - Q(S)$ can be thought of as a measure of the amount of extra predictive power provided by $\mathbf{X}_{S' \setminus S}$ over $\mathbf{X}_S$.
- Key assumption for our consistency theorem: There is some $\delta > 0$ such that whenever S is an *insufficient* subset, there is some $j \notin S$ such that $Q(S \cup \{j\}) \geq Q(S) + \delta$.
- We implicitly assume that $\delta \not\to 0$ as $n, p \to \infty$.
- Since $Q(S) \leq 1$ for all S, the above assumption implies that *there is a sufficient subset of size* $\leq 1/\delta$. That's why this is a sparsity assumption.

Regularity assumptions

In addition to the sparsity assumption, we need two regularity assumptions:

(A1) There is a number L such that for any S of size $\leq 1/\delta$, any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^S$, and any $t \in \mathbb{R}$,

$$|\mathbb{P}(Y \leq t | \mathbf{X}_S = \mathbf{x}) - \mathbb{P}(Y \leq t | \mathbf{X}_S = \mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|.$$

(A2) There is a number B such that for any S of size $\leq 1/\delta$, the support of $\mathbf{X}_S$ has diameter $\leq B$.

Consistency of FOCI

Theorem (Azadkia and Chatterjee, 2019)

Suppose that the sparsity assumption holds for some $\delta > 0$, and that the regularity assumptions (A1) and (A2) hold with some constants L and B. Let $\hat{S}$ be the subset selected by FOCI with a sample of size n. There are positive real numbers C_1, C_2 and C_3 depending only on L, B and δ such that

$$\mathbb{P}(\hat{S} \text{ is sufficient}) \geq 1 - C_1 p^{C_2} e^{-C_3 n}.$$

Remark: Note that if L, B and δ are held fixed, the bound tends to 1 as long as n grows faster than $\log p$.

Where does the FOCI algorithm come from?

- The idea originates in the following paper:

  Sourav Chatterjee (2019). A new coefficient of correlation. *arXiv preprint arXiv:1909.10140.*

A new coefficient of correlation

- Let (X, Y) be a pair of random variables, and let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. copies of (X, Y).
- Reorder as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$, so that $X_{(1)} \leq \cdots \leq X_{(n)}$ (ties broken at random).
- Let r_i be the rank of $Y_{(i)}$, where ranks go from 1 to n, with ties broken at random.
- Let $\ell_i = \#\{j : Y_{(j)} \geq Y_{(i)}\}$.
- Define

$$\xi_n(X, Y) = 1 - \frac{n \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \sum_{i=1}^{n} \ell_i (n - \ell_i)}.$$

- When the Y_i's have no ties, this reduces to the simpler expression

$$\xi_n(X, Y) = 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}.$$

Main property of ξ_n

Theorem (Chatterjee, 2019)

Suppose that Y is not a constant. Then as $n \to \infty$, $\xi_n(X, Y)$ converges almost surely to the deterministic limit

$$\xi(X, Y) = \frac{\int \mathrm{Var}(\mathbb{E}(1_{\{Y \geq t\}}|X))d\mu(t)}{\int \mathrm{Var}(1_{\{Y \geq t\}})d\mu(t)},$$

where μ is the law of Y. This limit belongs to the interval $[0, 1]$. It is 0 if and only if X and Y are independent, and it is 1 if and only if there is a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$ almost surely.

Remarks: (1) There are no assumptions on the law of (X, Y) other than that Y is not a constant. (2) There is no other known coefficient that interpolates between independence and exact functional dependence. (No, maximal correlation does not work.)

Some examples

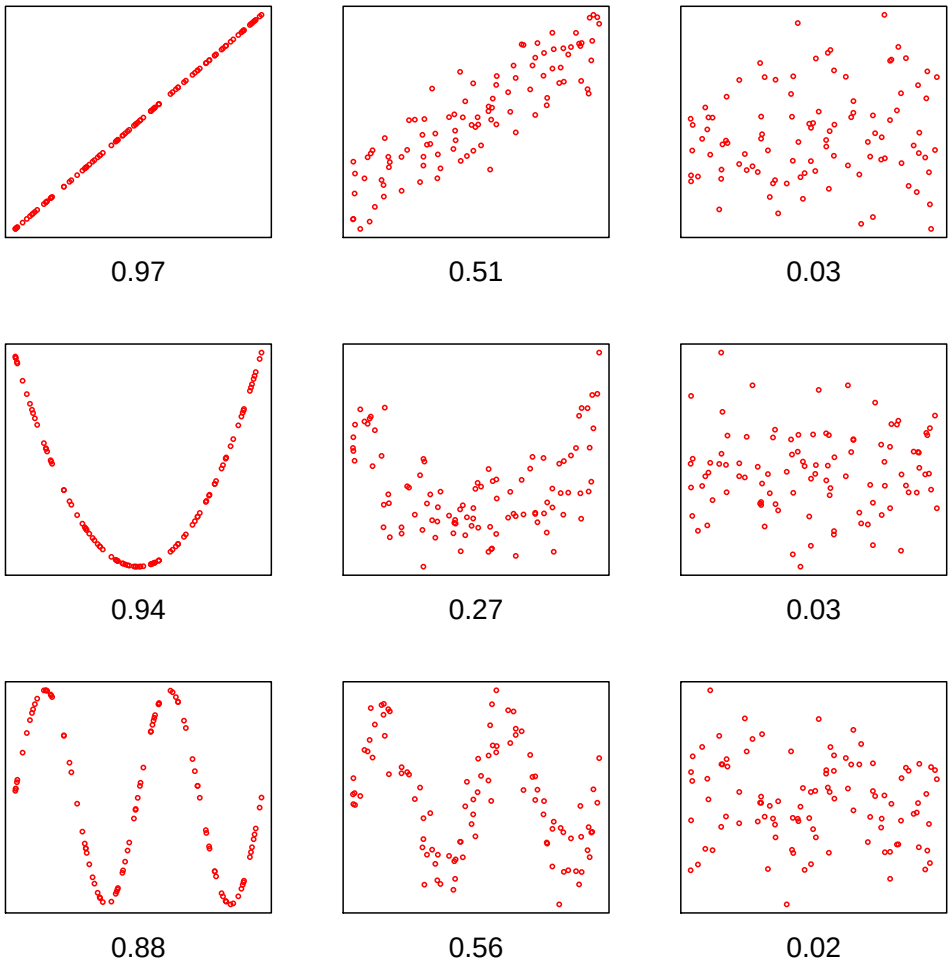

Figure: Values of ξ_n for various kinds of scatterplots, with $n = 100$. Noise increases from left to right.

From correlation coefficient to FOCI

- In Azadkia and Chatterjee (2019), this correlation coefficient was generalized to a similar coefficient $T_n(Y, \mathbf{Z}|\mathbf{X})$ that measures the *conditional dependence* between Y and $\mathbf{Z}$ given $\mathbf{X}$, based on an i.i.d. sample $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$.
- It was proved that $T_n(Y, \mathbf{Z}|\mathbf{X})$ converges to a limit $T(Y, \mathbf{Z}|\mathbf{X}) \in [0, 1]$, which is 0 if and only if Y and $\mathbf{Z}$ are independent given $\mathbf{X}$, and 1 if and only if Y is a function of $\mathbf{Z}$ given $\mathbf{X}$.
- At each step of the FOCI algorithm, we add the variable that maximizes the conditional dependence with Y given the set of variables already selected, according to the above measure.
- This is similar to ordinary forward stepwise, where we maximize the partial correlation between Y and the new variable given the variables already selected.

Definition of $T_n(Y, \mathbf{Z}|\mathbf{X})$

- Let $N(i)$ be the index j such that $\mathbf{X}_j$ is the nearest neighbor of $\mathbf{X}_i$ (ties broken at random).
- Let $M(i)$ be the index j such that $(\mathbf{X}_j, \mathbf{Z}_j)$ is the nearest neighbor of $(\mathbf{X}_i, \mathbf{Z}_i)$.
- Let $R_i = \#\{j : Y_j \leq Y_i\}$.
- Then we define

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

- The limit of $T_n(Y, \mathbf{Z}|\mathbf{X})$ is

$$T(Y, \mathbf{Z}|\mathbf{X}) = \frac{\int \mathbb{E}(\mathrm{Var}(\mathbb{P}(Y \geq t|\mathbf{Z}, \mathbf{X})|\mathbf{X}))d\mu(t)}{\int \mathbb{E}(\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{X}))d\mu(t)},$$

  where μ is the law of Y.
- $T(Y, \mathbf{Z}|\mathbf{X}) = 0$ iff Y and $\mathbf{Z}$ are independent given $\mathbf{X}$, and $T(Y, \mathbf{Z}|\mathbf{X}) = 1$ iff Y is a measurable function of $\mathbf{Z}$ given $\mathbf{X}$.

Proof sketch for $\xi_n \to \xi$

- For simplicity, assume that X and Y are continuous variables.
- Recall that r_i is the rank of $Y_{(i)}$, where $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ is a rearrangement of the data in increasing order of X_i's.
- Recall that $\xi_n = 1 - \frac{3}{n^2-1} \sum_{i=1}^{n} |r_{i+1} - r_i|$.
- Note that $r_i/n \approx F(Y_{(i)})$, where F is the c.d.f. of Y.
- Thus, $\xi_n \approx 1 - \frac{3}{n} \sum_{i=1}^{n} |F(Y_i) - F(Y_{N(i)})|$, where $N(i)$ is the index j such that X_j is immediately to the right of X_i.
- $|F(x) - F(y)| = \int (1_{\{t \le x\}} - 1_{\{t \le y\}})^2 d\mu(t)$, where μ is the law of Y.
- Since $X_i \approx X_{N(i)}$, the random variables Y_i and $Y_{N(i)}$ are approximately i.i.d. conditional on $\mathbf{X} = (X_1, \ldots, X_n)$, which gives:
- $\mathbb{E}[(1_{\{t \le Y_i\}} - 1_{\{t \le Y_{N(i)}\}})^2 | \mathbf{X}] \approx 2\mathrm{Var}(1_{\{t \le Y_i\}} | \mathbf{X}) = 2\mathrm{Var}(1_{\{t \le Y_i\}} | X_i)$.
- This gives $\mathbb{E}(1_{\{t \le Y_i\}} - 1_{\{t \le Y_{N(i)}\}})^2 \approx 2\mathbb{E}[\mathrm{Var}(1_{\{t \le Y\}} | X)]$.
- So, we get $\mathbb{E}|F(Y_i) - F(Y_{N(i)})| \approx \int 2\mathbb{E}[\mathrm{Var}(1_{\{t \le Y\}} | X)] d\mu(t)$.
- From this, it is easy to show $\mathbb{E}(\xi_n) \to \xi$. Using concentration inequalities, we then get $\xi_n \to \xi$.

Proof sketch for properties of ξ

- Recall that

$$\xi(X, Y) = \frac{\int \mathrm{Var}(\mathbb{E}(1_{\{Y \geq t\}}|X))d\mu(t)}{\int \mathrm{Var}(1_{\{Y \geq t\}})d\mu(t)},$$

  where μ is the law of Y.
- Since $\mathrm{Var}(\mathbb{E}(1_{\{Y \geq t\}}|X)) \leq \mathrm{Var}(1_{\{Y \geq t\}})$ for every t, we have $\xi \in [0, 1]$.
- It is not hard to see that $\mathrm{Var}(\mathbb{E}(1_{\{Y \geq t\}}|X)) = \mathrm{Var}(1_{\{Y \geq t\}})$ if and only if $1_{\{Y \geq t\}}$ is a measurable function of X.
- This holds for all t in the support of Y if and only if Y is a measurable function of X.
- Similarly, $\mathrm{Var}(\mathbb{E}(1_{\{Y \geq t\}}|X)) = 0$ if and only if $1_{\{Y \geq t\}}$ is independent of X.
- Again, this holds for all t in the support of Y if and only if Y and X are independent.
- This proves that $\xi = 0$ iff X and Y are independent, and $\xi = 1$ iff Y is a measurable function of X.

Package

- If you want to experiment with FOCI, there is a package available on CRAN.
- R package: FOCI
- Function: foci
- Implementation:
  - In R, let y be the vector of responses (length n).
  - Let x be the matrix of covariates (order $n \times p$).
  - The command foci(y, x) returns a vector containing the indices of the chosen variables, in the order that they were chosen by the algorithm.
- The covariates can be numerical or categorical. The program automatically converts categorical variables to numerical.