Making Good Prediction

A Theoretical Framework

Adeline Lo
with H. Chernoff, T. Zheng, S. Lo

IMC, November 10, 2017

Motivation

Prediction is important

- Predictions of violent events: civil wars, revolutions, local conflicts

Motivation

Prediction is important

- Predictions of oncoming recessions

Motivation

Prediction is important

- GWAS-level predictions of disease-status

Motivation

Prediction used in causal inference techniques

- Selecting amongst many and possibly weak instrumental variables in 2SLS (Belloni et al. 2014)
- Creating synthetic controls in DID designs (Xu 2015)
- Collecting interactions in conjoint analysis using LASSOplus (Ratkovic & Tingley 2015)
- Collecting fewer interactions of factors in factorial experiments using LASSO (Egami & Imai)
- Predicting case/control labels in text data

Motivation

Prediction used in causal inference techniques

- Selecting amongst many and possibly weak instrumental variables in 2SLS (Belloni et al. 2014)
- Creating synthetic controls in DID designs (Xu 2015)
- Collecting interactions in conjoint analysis using LASSOplus (Ratkovic & Tingley 2015)
- Collecting fewer interactions of factors in factorial experiments using LASSO (Egami & Imai)
- Predicting case/control labels in text data

Motivation

Prediction used in causal inference techniques

- Selecting amongst many and possibly weak instrumental variables in 2SLS (Belloni et al. 2014)
- Creating synthetic controls in DID designs (Xu 2015)
- Collecting interactions in conjoint analysis using LASSOplus (Ratkovic & Tingley 2015)
- Collecting fewer interactions of factors in factorial experiments using LASSO (Egami & Imai)
- Predicting case/control labels in text data

Motivation

Prediction used in causal inference techniques

- Selecting amongst many and possibly weak instrumental variables in 2SLS (Belloni et al. 2014)
- Creating synthetic controls in DID designs (Xu 2015)
- Collecting interactions in conjoint analysis using LASSOplus (Ratkovic & Tingley 2015)
- Collecting fewer interactions of factors in factorial experiments using LASSO (Egami & Imai)
- Predicting case/control labels in text data

Motivation

Prediction used in causal inference techniques

- Selecting amongst many and possibly weak instrumental variables in 2SLS (Belloni et al. 2014)
- Creating synthetic controls in DID designs (Xu 2015)
- Collecting interactions in conjoint analysis using LASSOplus (Ratkovic & Tingley 2015)
- Collecting fewer interactions of factors in factorial experiments using LASSO (Egami & Imai)
- Predicting case/control labels in text data

Motivation

Prediction used in causal inference techniques

- Selecting amongst many and possibly weak instrumental variables in 2SLS (Belloni et al. 2014, Lo & Levy 2015)
- Creating synthetic controls in DID designs (Xu 2015)
- Collecting interactions in conjoint analysis using LASSOplus (Ratkovic & Tingley 2015)
- Collecting fewer interactions of factors in factorial experiments using LASSO (Egami & Imai)
- Predicting case/control labels in text data

Not nearly enough attention on creating a framework from which to theoretically consider predictivity

Motivation

Prediction-based framework to theoretically consider predictivity
Why does this matter?

Motivation

Prediction-based framework to theoretically consider predictivity

Why does this matter?

- Creation of measures of predictivity specifically through the lens of maximizing theoretical correct prediction rates (minimizing theoretical error rates) might be important for better prediction

Motivation

Prediction-based framework to theoretically consider predictivity

Why does this matter?

- Creation of measures of predictivity specifically through the lens of maximizing theoretical correct prediction rates (minimizing theoretical error rates) might be important for better prediction
- Can we just use models/methods that seem to predict very well, regardless of how they were motivated?

Motivation

Prediction-based framework to theoretically consider predictivity

Why does this matter?

- Creation of measures of predictivity specifically through the lens of maximizing theoretical correct prediction rates (minimizing theoretical error rates) might be important for better prediction
- Can we just use models/methods that seem to predict very well, regardless of how they were motivated?
  - One motivation for this project was I-score (measure we suggest here as a good measure for predictivity) performance in complex data

Motivation

Prediction-based framework to theoretically consider predictivity

Why does this matter?

- Creation of measures of predictivity specifically through the lens of maximizing theoretical correct prediction rates (minimizing theoretical error rates) might be important for better prediction
- Can we just use models/methods that seem to predict very well, regardless of how they were motivated?
  - One motivation for this project was I-score (measure we suggest here as a good measure for predictivity) performance in complex data
  - We believe I-score strong performance is because the score itself is related to theoretical correct prediction rate

Motivation

Prediction-based framework to theoretically consider predictivity
Why does this matter?

- Creation of measures of predictivity specifically through the lens of maximizing theoretical correct prediction rates (minimizing theoretical error rates) might be important for better prediction
- Can we just use models/methods that seem to predict very well, regardless of how they were motivated?
  - One motivation for this project was I-score (measure we suggest here as a good measure for predictivity) performance in complex data
  - We believe I-score strong performance is because the score itself is related to theoretical correct prediction rate
  - There may be logic to and benefits from creation of prediction measures from a prediction framework

Current Approaches to Prediction

Variable Selection (VS)

via

- Significant variables (theory)

Current Approaches to Prediction

Variable Selection (VS)

via

- Significant variables (theory)
- Out of sample testing/Cross-validation (error rates)

Current Approaches to Prediction

Variable Selection (VS) via

- Significant variables (theory)
  - Highly significant variables are not necessarily highly predictive and vice versa (*Lo et al. 2015*)
- Out of sample testing/Cross-validation (error rates)

Current Approaches to Prediction

Variable Selection (VS) via

- Significant variables (theory)
  - Highly significant variables are not necessarily highly predictive and vice versa (*Lo et al. 2015*)
- Out of sample testing/Cross-validation (error rates)
  - No theory based measure (such as significance measures) for underlying predictivity

Current Approaches to Prediction

Variable Selection (VS) via

- Significant variables (theory)
  - Highly significant variables are not necessarily highly predictive and vice versa (*Lo et al. 2015*)
- Out of sample testing/Cross-validation (error rates)
  - No theory based measure (such as significance measures) for underlying predictivity
  - Why is it that certain approaches perform better than others in some scenarios can be hard to ascertain; what is the benchmark against which to compare?

Current Approaches to Prediction

Variable Selection (VS) via

- Significant variables (theory)
  - Highly significant variables are not necessarily highly predictive and vice versa (*Lo et al. 2015*)
- Out of sample testing/Cross-validation (error rates)
  - No theory based measure (such as significance measures) for underlying predictivity
  - Why is it that certain approaches perform better than others in some scenarios can be hard to ascertain; what is the benchmark against which to compare?
- Both approaches additionally suffer from curse of dimensionality constraints vis a vis joint/interactive variables as variable size grows.

Contribution

We provide a theoretical framework behind prediction

Contribution

We provide a theoretical framework behind prediction

- **Set up the framework**: define our objective function (what are we trying to get at? levels of predictivity)

Contribution

We provide a theoretical framework behind prediction

- **Set up the framework**: define our objective function (what are we trying to get at? levels of predictivity)
- **Maximize our objective function** and **find solution** (what solves our objective function? the variables that provide the maximal level of predictivity)

Contribution

We provide a theoretical framework behind prediction

- **Set up the framework**: define our objective function (what are we trying to get at? levels of predictivity)
- **Maximize our objective function** and **find solution** (what solves our objective function? the variables that provide the maximal level of predictivity)
- **Identify sample-appropriate measures for measuring predictivity that match the theoretical solution**

Contribution

We provide a theoretical framework behind prediction

- **Set up the framework**: define our objective function (what are we trying to get at? levels of predictivity)
- **Maximize our objective function** and **find solution** (what solves our objective function? the variables that provide the maximal level of predictivity)
- **Identify sample-appropriate measures for measuring predictivity that match the theoretical solution**
  - Solution doesn't actually have usable sample analog form. Our second major contribution stems from considering an alternative solution with a sample analog that is useable.

What is a Good Measure for Predictivity?

What, theoretically, does "maximizing predictivity" mean? What should an influence measure that measures predictivity be able to do?

What is a Good Measure for Predictivity?

What, theoretically, does "maximizing predictivity" mean? What should an influence measure that measures predictivity be able to do?

- Reflect predictive power of a given variable set

What is a Good Measure for Predictivity?

What, theoretically, does "maximizing predictivity" mean? What should an influence measure that measures predictivity be able to do?

- Reflect predictive power of a given variable set
- Handle groups of variables

What is a Good Measure for Predictivity?

What, theoretically, does "maximizing predictivity" mean? What should an influence measure that measures predictivity be able to do?

- Reflect predictive power of a given variable set
- Handle groups of variables
- Be able to differentiate between truly influential variables and noisy variables

Maximizing Predictivity: example

Suppose we are in the simplest of worlds:

Maximizing Predictivity: example

Suppose we are in the simplest of worlds:

- 1 explanatory variable X, Democracy, and 1 outcome variable Y, Civil War. Both X and Y are binary.

Maximizing Predictivity: example

Suppose we are in the simplest of worlds:

- 1 explanatory variable X, Democracy, and 1 outcome variable Y, Civil War. Both X and Y are binary.
- X is defined on a space, Π_X, with density $p_{\boldsymbol{X}}(x)$

Maximizing Predictivity: example

Suppose we are in the simplest of worlds:

- 1 explanatory variable X, Democracy, and 1 outcome variable Y, Civil War. Both X and Y are binary.
- X is defined on a space, Π_X, with density $p_{\boldsymbol{X}}(x)$
- "civil war" observations are n_d and "no civil war" observations are n_c, each with two probabilities: $p_{\boldsymbol{X}_d}$ and $p_{\boldsymbol{X}_c}$

Maximizing Predictivity: example

Suppose we are in the simplest of worlds:

- 1 explanatory variable X, Democracy, and 1 outcome variable Y, Civil War. Both X and Y are binary.
- X is defined on a space, Π_X, with density $p_{\boldsymbol{X}}(x)$
- "civil war" observations are n_d and "no civil war" observations are n_c, each with two probabilities: $p_{\boldsymbol{X}_d}$ and $p_{\boldsymbol{X}_c}$
- the expected correct prediction rate c using variable X Democracy: $constant \cdot [|p_{X_d}(x=1) - p_{X_c}(x=1)| + |p_{X_d}(x=0) - p_{X_c}(x=0)|]$

Maximizing Predictivity: example

Suppose we are in the simplest of worlds:

- 1 explanatory variable X, Democracy, and 1 outcome variable Y, Civil War. Both X and Y are binary.
- X is defined on a space, Π_X, with density $p_{\boldsymbol{X}}(x)$
- "civil war" observations are n_d and "no civil war" observations are n_c, each with two probabilities: $p_{\boldsymbol{X}_d}$ and $p_{\boldsymbol{X}_c}$
- the expected correct prediction rate c using variable X Democracy: $constant \cdot [|p_{X_d}(x=1) - p_{X_c}(x=1)| + |p_{X_d}(x=0) - p_{X_c}(x=0)|]$
  - We can also call this the **predictivity** of variable X for variable Y

Maximizing Predictivity: example

Suppose we are in the simplest of worlds:

- 1 explanatory variable X, Democracy, and 1 outcome variable Y, Civil War. Both X and Y are binary.
- X is defined on a space, Π_X, with density $p_{\boldsymbol{X}}(x)$
- "civil war" observations are n_d and "no civil war" observations are n_c, each with two probabilities: $p_{\boldsymbol{X}_d}$ and $p_{\boldsymbol{X}_c}$
- the expected correct prediction rate c using variable X Democracy: $constant \cdot |p_{X_d}(x) - p_{X_c}(x)|$
  - We can also call this the **predictivity** of variable X for variable Y

Imagine now we have many X variables to consider. Then what we are looking for is the set of X variables that maximize the correct prediction rate, c.

Maximizing Predictivity: general

Basic Bayesian set up

Correct prediction rate:

Maximizing Predictivity: general

Basic Bayesian set up

Correct prediction rate:

- $\boldsymbol{X}$ is discrete random vector defined on space Π_x, density of $p_{\boldsymbol{X}}(x)$

Maximizing Predictivity: general

Basic Bayesian set up

Correct prediction rate:

- $\boldsymbol{X}$ is discrete random vector defined on space Π_x, density of $p_{\boldsymbol{X}}(x)$
- n_d cases, n_c controls independently selected from two discrete probabilities: $p_{\boldsymbol{X}_d}(x)$ and $p_{\boldsymbol{X}_c}(x)$ ($\equiv p(x|w=d)$ and $p(x|w=c)$)

Maximizing Predictivity: general

Basic Bayesian set up

Correct prediction rate:

- $\boldsymbol{X}$ is discrete random vector defined on space Π_x, density of $p_{\boldsymbol{X}}(x)$
- n_d cases, n_c controls independently selected from two discrete probabilities: $p_{\boldsymbol{X_d}}(x)$ and $p_{\boldsymbol{X}_c}(x)$ ($\equiv p(x|w = d)$ and $p(x|w = c)$)
- $\{\boldsymbol{X}_d, \boldsymbol{X}_c\}$ always arrive as pair when $\boldsymbol{X}$ variables fixed (fixing $\Pi_x = \{x = (x_1, x_2, ..., x_m)\}$);

Maximizing Predictivity: general

Basic Bayesian set up

Correct prediction rate:

- $\boldsymbol{X}$ is discrete random vector defined on space Π_x, density of $p_{\boldsymbol{X}}(x)$
- n_d cases, n_c controls independently selected from two discrete probabilities: $p_{\boldsymbol{X_d}}(x)$ and $p_{\boldsymbol{X}_c}(x)$ ($\equiv p(x|w = d)$ and $p(x|w = c)$)
- $\{\boldsymbol{X}_d, \boldsymbol{X}_c\}$ always arrive as pair when $\boldsymbol{X}$ variables fixed (fixing $\Pi_x = \{x = (x_1, x_2, ..., x_m)\}$);
- $\boldsymbol{X}_d$ and $\boldsymbol{X}_c$ defined on common partition space, Π_x

Maximizing Predictivity: general

Basic Bayesian set up

Correct prediction rate:

- $\boldsymbol{X}$ is discrete random vector defined on space Π_x, density of $p_{\boldsymbol{X}}(x)$
- n_d cases, n_c controls independently selected from two discrete probabilities: $p_{\boldsymbol{X_d}}(x)$ and $p_{\boldsymbol{X}_c}(x)$ ($\equiv p(x|w=d)$ and $p(x|w=c)$)
- $\{\boldsymbol{X}_d, \boldsymbol{X}_c\}$ always arrive as pair when $\boldsymbol{X}$ variables fixed (fixing $\Pi_x = \{x = (x_1, x_2, ..., x_m)\}$);
- $\boldsymbol{X}_d$ and $\boldsymbol{X}_c$ defined on common partition space, Π_x
- If new observation has 50% chance to be case/control, expected correct prediction rate/error of adopting this rule is:

Maximizing Predictivity: general

Basic Bayesian set up

Correct prediction rate:

- $\boldsymbol{X}$ is discrete random vector defined on space Π_x, density of $p_{\boldsymbol{X}}(x)$
- n_d cases, n_c controls independently selected from two discrete probabilities: $p_{\boldsymbol{X_d}}(x)$ and $p_{\boldsymbol{X}_c}(x)$ ($\equiv p(x|w = d)$ and $p(x|w = c)$)
- $\{\boldsymbol{X}_d, \boldsymbol{X}_c\}$ always arrive as pair when $\boldsymbol{X}$ variables fixed (fixing $\Pi_x = \{x = (x_1, x_2, ..., x_m)\}$);
- $\boldsymbol{X}_d$ and $\boldsymbol{X}_c$ defined on common partition space, Π_x
- If new observation has 50% chance to be case/control, expected correct prediction rate/error of adopting this rule is:

$$c = c[p_{\boldsymbol{X}_d},\ p_{\boldsymbol{X}_c}] = 1 - e[p_{\boldsymbol{X}_d},\ p_{\boldsymbol{X}_c}] = \frac{1}{2} \sum_{x \in \Pi_x} \max\{p_{\boldsymbol{X}_d}(x),\ p_{\boldsymbol{X}_c}(x)\}$$

$$c[p_{\boldsymbol{X}_d},\ p_{\boldsymbol{X}_c}] = \frac{1}{2} + \frac{1}{4} \sum_{x \in \Pi_x} |p_{\boldsymbol{X}_d}(x) - p_{\boldsymbol{X}_c}(x)| \qquad (1)$$

Problems with Sample Analog

$$c[p_{\boldsymbol{X}_d}, p_{\boldsymbol{X}_c}] = \frac{1}{2} + \frac{1}{4} \sum_{x \in \Pi_x} |p_{\boldsymbol{X}_d}(x) - p_{\boldsymbol{X}_c}(x)|$$

Sample analog of equation (1) is always increasing in variables and favors ever-increasing the variable set with both truly influential as well as noisy and un-influential variables.

Suggested Alternative Solution

Lemma 1

Let a_1, a_2, $a_3 ... a_k$ be k nonnegative numbers. Then $\sum_{i=1}^{k} a_i \geq \sqrt{\sum_{i=1}^{k} a_i^2}$. If we replace a_i by $|p(i|d) - p(i|c)|$ $\forall i$, $1 \leq i \leq k$, it is clear that by maximizing $\sum_{i=1}^{k} (p_i(d) - p_i(c))^2$ over possible pairs will have the parallel effect of encouraging selection of probability pairs that satisfy the maximization in Equation 1, yielding a better predictor. We can show that the I-score can be seen asymptotically as precisely the maximization of the term up to a constant
$A(\pi_x) = \sum_{i=1}^{k} (p_i(d) - p_i(c))^2$.

Since $\sum_{i=1}^{k} |p_i(d) - p_i(c)| \geq \sqrt{\sum_{i=1}^{k} (p_i(d) - p_i(c))^2}$, a strategy that seeks for a variable set with larger value of $A(\pi_x)$ will automatically have the effect of seeking for the variable set with a better prediction rate.

Intuition behind Suggested Alternative Solution

Intuitive explanation:

Intuition behind Suggested Alternative Solution

Intuitive explanation:

- Absolute difference strictly positive, linear, increasing in positive space of integers

Intuition behind Suggested Alternative Solution

Intuitive explanation:

- Absolute difference strictly positive, linear, increasing in positive space of integers
- Squared term allows for existence of maximum

Suggested Alternative Measure: I-score

Suggested Alternative Measure: I-score

- n observations of Y and large number S of **X**s, $X_1, X_2, ..., X_S$.

Suggested Alternative Measure: I-score

- n observations of Y and large number S of **X**s, $X_1, X_2, ..., X_S$.
- Randomly select small group, m, of the **X**s. Call this m X_j, $j = 1, ..., m$ that take values 0, 1, *and* 2 (here, discrete example)

Suggested Alternative Measure: I-score

- n observations of Y and large number S of $\mathbf{X}$s, $X_1, X_2, ..., X_S$.
- Randomly select small group, m, of the $\mathbf{X}$s. Call this m X_j, $j = 1, ..., m$ that take values 0, 1, *and* 2 (here, discrete example)
- $m_1 = 3^m$ possible values for each set of X's.

Suggested Alternative Measure: I-score

- n observations of Y and large number S of **X**s, $X_1, X_2, ..., X_S$.
- Randomly select small group, m, of the **X**s. Call this m X_j, $j = 1, ..., m$ that take values 0, 1, *and* 2 (here, discrete example)
- $m_1 = 3^m$ possible values for each set of X's.
- Partition n observations into m_1 cells according to values of m **X** variables and refer to this partition as Π.

Suggested Alternative Measure: I-score

- n observations of Y and large number S of **X**s, $X_1, X_2, ..., X_S$.
- Randomly select small group, m, of the **X**s. Call this m X_j, $j = 1, ..., m$ that take values 0, 1, *and* 2 (here, discrete example)
- $m_1 = 3^m$ possible values for each set of X's.
- Partition n observations into m_1 cells according to values of m **X** variables and refer to this partition as Π.
- I-score designed to place greater weight on cells with more observations:

Suggested Alternative Measure: I-score

- n observations of Y and large number S of **X**s, $X_1, X_2, ..., X_S$.
- Randomly select small group, m, of the **X**s. Call this m X_j, $j = 1, ..., m$ that take values 0, 1, *and* 2 (here, discrete example)
- $m_1 = 3^m$ possible values for each set of X's.
- Partition n observations into m_1 cells according to values of m **X** variables and refer to this partition as Π.
- I-score designed to place greater weight on cells with more observations:

$$I_\Pi = \sum_{k=1}^{m_1} \frac{n_k}{n} \cdot \frac{(\bar{Y}_k - \bar{Y})^2}{\frac{s^2}{n_k}} = \frac{\sum_{k=1}^{m_1} n_k^2 (\bar{Y}_k - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \quad (2)$$

where $s^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

I-score

- Asymptotics
- Data simulations

I-score Asymptotics

I-score asymptotics

I-score Asymptotics

- As $n \to \infty$ the I-score decomposes to a term that looks like:

I-score asymptotics

I-score Asymptotics

- As $n \to \infty$ the I-score decomposes to a term that looks like:

$$constant \cdot \sum_{j \in \Pi} [p(j|d) - p(j|c)]^2 \tag{3}$$

I-score asymptotics

I-score Asymptotics

- As $n \to \infty$ the I-score decomposes to a term that looks like:

$$constant \cdot \sum_{j \in \Pi} [p(j|d) - p(j|c)]^2 \tag{3}$$

- Recall c:

$$c[p_{\boldsymbol{X}_d}, p_{\boldsymbol{X}_c}] = \frac{1}{2} + \frac{1}{4} \sum_{x \in \Pi_x} |p_{\boldsymbol{X}_d}(x) - p_{\boldsymbol{X}_c}(x)|$$

I-score asymptotics

I-score Asymptotics

- As $n \to \infty$ the I-score decomposes to a term that looks like:

$$constant \cdot \sum_{j \in \Pi} [p(j|d) - p(j|c)]^2 \tag{3}$$

- Recall c:

$$c[p_{\mathbf{X}_d}, p_{\mathbf{X}_c}] = \frac{1}{2} + \frac{1}{4} \sum_{x \in \Pi_x} |p_{\mathbf{X}_d}(x) - p_{\mathbf{X}_c}(x)|$$

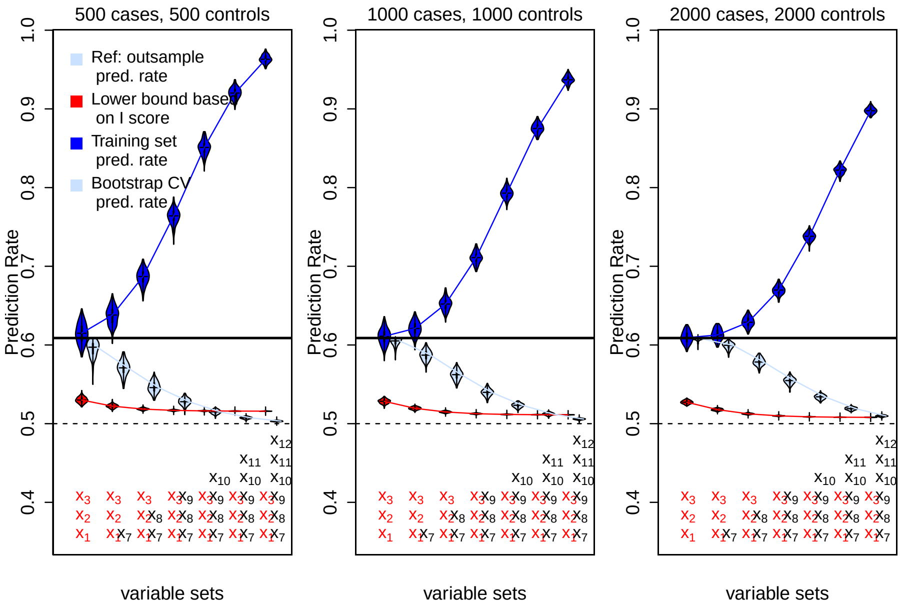
$$= \frac{1}{2} + \frac{1}{4} \sum_{i \in \Pi} |p(j|d) - p(j|c)|$$

I-score asymptotics

I-score with data

How does I-score fare with data (sample constrained world)?

I-score with simulated data



I-score with real data

	Systematic name	Gene name	Marginal p-value
1	Contig45347_RC	KIAA1683	0.008
2	NM_005145	GNG7	0.54
3	Z34893	ICAP-1A	0.15
4	NM_006121	KRT1	0.9
5	NM_004701	CCNB2	0.003
	Joint *I*-score: 2.89	**Joint p-value:** 0.005	Family-wise threshold: $6.98\text{x}10^{-5}$

Table: **Real data example** vant Veer

Desirable properties of the I-score

Desirable properties of the I-score

- **No model specification**: Requires no specification of a model for joint effect of $\{X_1, X_2, ..., X_m\}$ on Y. I captures discrepancy between the conditional means of Y on $\{X_1, X_2, ..., X_m\}$ and mean of Y.

Desirable properties of the I-score

- **No model specification**: Requires no specification of a model for joint effect of $\{X_1, X_2, ..., X_m\}$ on Y. I captures discrepancy between the conditional means of Y on $\{X_1, X_2, ..., X_m\}$ and mean of Y.

- **Differentiation between noisy and influential variables**: I doesn't monotonically increase with the addition of any variables (as would the sample analog form of Eqn 1). Rather, given a variable set of size m with $m-1$ truly influential variables, the I is higher under the influential $m-1$ variables than under all m variables. Dropping to $m-2$ variables leads to decrease in I. I has natural tendency to "peak" at variable set(s) that lead to the maximum predictive rate in the face of noisy variables, under the current sample size.

Desirable properties of the I-score

- **No model specification**: Requires no specification of a model for joint effect of $\{X_1, X_2, ..., X_m\}$ on Y. I captures discrepancy between the conditional means of Y on $\{X_1, X_2, ..., X_m\}$ and mean of Y.

- **Differentiation between noisy and influential variables**: I doesn't monotonically increase with the addition of any variables (as would the sample analog form of Eqn 1). Rather, given a variable set of size m with $m-1$ truly influential variables, the I is higher under the influential $m-1$ variables than under all m variables. Dropping to $m-2$ variables leads to decrease in I. I has natural tendency to "peak" at variable set(s) that lead to the maximum predictive rate in the face of noisy variables, under the current sample size.

- **Approximation towards theoretical maximization of prediction rate**: I approximates maximizing Eqn 1 by identifying the cluster of variables that maximize the term $\sum_{i=1}^{k}(P_i(D) - P_i(C))^2$, which is directly related to maximization of the correct prediction rate less the error rate (Lemma 1).

Desirable properties of the I-score

- **No model specification**: Requires no specification of a model for joint effect of $\{X_1, X_2, ..., X_m\}$ on Y. I captures discrepancy between the conditional means of Y on $\{X_1, X_2, ..., X_m\}$ and mean of Y.
- **Differentiation between noisy and influential variables**: I doesn't monotonically increase with the addition of any variables (as would the sample analog form of Eqn 1). Rather, given a variable set of size m with $m-1$ truly influential variables, the I is higher under the influential $m-1$ variables than under all m variables. Dropping to $m-2$ variables leads to decrease in I. I has natural tendency to "peak" at variable set(s) that lead to the maximum predictive rate in the face of noisy variables, under the current sample size.
- **Approximation towards theoretical maximization of prediction rate**: I approximates maximizing Eqn 1 by identifying the cluster of variables that maximize the term $\sum_{i=1}^{k}(P_i(D) - P_i(C))^2$, which is directly related to maximization of the correct prediction rate less the error rate (Lemma 1).
- **Interactions**: I shown elsewhere to be able to handle interactions

Discussion

- We have to rethink how we approach prediction; VS cannot be accomplished through significance criteria

Discussion

- We have to rethink how we approach prediction; VS cannot be accomplished through significance criteria
- We require new measures and new criteria that are prediction-based

Discussion

- We have to rethink how we approach prediction; VS cannot be accomplished through significance criteria
- We require new measures and new criteria that are prediction-based
- Preliminarily offer the I-score, which has the following nice properties:

Discussion

- We have to rethink how we approach prediction; VS cannot be accomplished through significance criteria
- We require new measures and new criteria that are prediction-based
- Preliminarily offer the I-score, which has the following nice properties:
  - Theoretical relationship to maximizing prediction rates

Discussion

- We have to rethink how we approach prediction; VS cannot be accomplished through significance criteria
- We require new measures and new criteria that are prediction-based
- Preliminarily offer the I-score, which has the following nice properties:
  - Theoretical relationship to maximizing prediction rates
  - Seems to predict well in simulated and real applications

Discussion

- We have to rethink how we approach prediction; VS cannot be accomplished through significance criteria
- We require new measures and new criteria that are prediction-based
- Preliminarily offer the I-score, which has the following nice properties:
  - Theoretical relationship to maximizing prediction rates
  - Seems to predict well in simulated and real applications
  - Can distinguish between noisy and influential variables

Discussion

- We have to rethink how we approach prediction; VS cannot be accomplished through significance criteria
- We require new measures and new criteria that are prediction-based
- Preliminarily offer the I-score, which has the following nice properties:
  - Theoretical relationship to maximizing prediction rates
  - Seems to predict well in simulated and real applications
  - Can distinguish between noisy and influential variables
  - Can handle interactions

Acknowledgements

Special thanks to Justin Esarey and the hosts of IMC for the invitation. This work benefited from the comments of Michael Newton, participants of HNG Workshop, and the participants of the Quantitative Social Science Colloquium at Princeton.

I-score Asymptotics

As $n \to \infty$, it can be shown that the I-score decomposes to two terms that converge to 0 in probability and a third term, call it B_n that approximates Equation 1 (correct prediction rate) via Lemma 1:

$$\frac{B_n}{n^2} = \lambda^2(1-\lambda)^2 \sum_{j \in \Pi} [p(j|d) - p(j|c)]^2$$

(Where $lim_n \frac{n_d}{n} = \lambda$, a fixed constant between 0 and 1)
Ignoring the constant term above, the I-score is exactly trying to search for the X partitions which maximize the summation term $\sum_{j \in \Pi} [p(j|d) - p(j|c)]^2$. Recall c:

$$c[p_{\mathbf{X}_d}, p_{\mathbf{X}_c}] = \frac{1}{2} + \frac{1}{4} \sum_{x \in \Pi_x} |p_{\mathbf{X}_d}(x) - p_{\mathbf{X}_c}(x)| \tag{4}$$

back