The Union of Intersections (UoI) Method for Interpretable Data Driven Discovery and Prediction

Kristofer E. Bouchard\textsuperscript{1,2,3*}, Alejandro Bujan\textsuperscript{3}, Farbod Roosta\textsuperscript{4}, Prabhat\textsuperscript{5}, Antoine Snijdes\textsuperscript{1}, J-H Mao\textsuperscript{1}, Edward F. Chang\textsuperscript{6,7}, Michael Mahoney\textsuperscript{8}, Sharmodeep Bhattacharyya\textsuperscript{9}
Understanding the fundamentals of Nature
New technologies generate vast amounts of data
Some desired properties of statistical-machine learning

- **predictive**
  allow prediction of the response variable
- **scalable**
  return an answer on large data sets
Some desired properties of statistical-machine learning

- **predictive**
  - allow prediction of the response variable

- **scalable**
  - return an answer on large data sets

- **selective**
  - only features that influence the response variable are selected

- **accurate**
  - estimated parameters are as close to the “real” value as possible

- **stable**
  - return the same values on multiple runs

### Industry | Science
---|---
| X | X |
| X | X |
|  | X |
|  | X |

**Interpretable**
Identify features for intervention/synthesis
Linear regression: the canonical statistical estimation problem

\[
Y = X\beta + \varepsilon
\]  

where, \( Y = (Y_1, \ldots, Y_n) \), \( X \) is the \( n \times p \) design matrix and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \) are iid random noise terms with \( \varepsilon \sim N(0, \sigma^2 I_n) \). Let \( S \) be the set of non-zero coefficients of \( \beta \), \( N \) be the set of zero coefficients of \( \beta \) and \( |S| = s \).

\[
L(\beta) = ||Y - X\beta||^2
\]  

(2)
Linear regression: priors are a double-edged sword

\[ Y = X\beta + \varepsilon \]  

(1)

where, \( Y = (Y_1, \ldots, Y_n) \), \( X \) is the \( n \times p \) design matrix and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \) are iid random noise terms with \( \varepsilon \sim N(0, \sigma^2 I_n) \). Let \( S \) be the set of non-zero coefficients of \( \beta \), \( N \) be the set of zero coefficients of \( \beta \) and \(|S| = s\).

\[ L(\beta, \lambda) = ||Y - X\beta||^2 + \lambda||\beta||_1 \]  

(2)

Wainwright, ARSA 2014
Linear regression: priors are a double-edged sword

$$Y = X\beta + \varepsilon$$  \hspace{1cm} (1)

where, $Y = (Y_1, \ldots, Y_n)$, $X$ is the $n \times p$ design matrix and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ are iid random noise terms with $\varepsilon \sim N(0, \sigma^2 I_n)$. Let $S$ be the set of non-zero coefficients of $\beta$, $N$ be the set of zero coefficients of $\beta$ and $|S| = s$.

$$L(\beta, \lambda) = ||Y - X\beta||^2 + \lambda||\beta||_1$$  \hspace{1cm} (2)

Wainwright, ARSA 2014
Regularization: feature compression
Ensemble methods: feature expansion
Union of Intersections (UoI) Method

\[
\hat{S}_{UoI} = \bigcup_{k=1}^{B_2} \Lambda \left( \bigcap_{l=1}^{B_1} \hat{S}_l^{\lambda q} \right)_{k/q}
\]

Bach, BoLasso, ICML, ‘08
Breiman, Bagging, ML, ’96
Union of Intersections (UoI) Method

Data Matrix (P x N) Samples
- Predictors (i.e., features, X)
- Bootstrap Samples
- Regression (OLS)
- Average
- Regularized Regression (e.g., Lasso)
- Regularized Parameters
- Response Variable
- Bagged Parameters
- Bagged Selected Parameters

Step 1: Feature Compression (Intersection (n) of bootstraps)
- Step 2: Feature Expansion (Union (U) of bootstraps and supports)
- Predict
- Max
- U(N) Features
- Union of Intersections (S)
UoI Lasso: Distributed algorithm for high-dimensional regression

![Diagram of data processing steps]

\[
S_{UoI} = \bigcup_{k=1}^{B_2, \Lambda} \bigcup_{q=1}^{B_1} \bigcap_{l=1}^{\bar{\lambda}_q} S_l^k
\]
UoLasso: Distributed algorithm for high-dimensional regression

Slope = 1.2

2046 cores
Python/mpi
UoILasso: Simulation Results
Uo\textsuperscript{Lasso} outperforms other methods
UolLasso out performs other methods
Theorem 1.1. Consider the model in (1). Assume that the distribution of \( \{ 1 \left( k \in S_{Uol}^\lambda \right), k \in N \} \) is exchangeable and that the UoI\(_{Lasso} \) procedure is not worse than random guessing for \( \lambda \). Also consider that the SRC assumption is satisfied. For large \( n \), let \( \lambda_{min} = \frac{2\sigma \sqrt{2C_{max}/c_{min}} \log(p)/n}{} \). Then,

(i) On a set \( \Omega_1 \) such that \( P(\Omega_1) \geq (1 - (3/p)^{B_1})^{B_2} \), for \( \lambda \geq \lambda_{min} \),

\[
N \cap S_{Uol}^\lambda = \emptyset \quad (3)
\]

(ii) Also, on a set \( \Omega_2 \) such that \( P(\Omega_2) \geq 1 - (1 - \exp(-3B_1/p))^{B_2} \), for \( \lambda \geq \lambda_{min} \), we have,

\[
(S \setminus S_{small}^\lambda) \subseteq S_{Uol}^\lambda \quad (4)
\]

where, \( S_{small}^\lambda = \{ \beta_k : |\beta_k|^2 \leq 5Cs\lambda^2 \} \).

(iii) On a set \( \Omega_A = \Omega_1 \cap \Omega_2 \) such that \( P(\Omega_A) \geq 1 - \max[(1 - \exp(-3B_1/p))^{B_2}, 1 - (1 - (3/p)^{B_1})^{B_2}] \), both (3) and (4) are satisfied.

(iv) On the same set \( \Omega_A \), we have,

\[
\|X(\hat{\beta}_{Uol-Lasso} - \beta)\|^2 \leq nc_1\lambda^2 \quad (5)
\]
Neural Recordings Directly from Human Cortex
Speech is generated by spatio-temporal patterns of cortical activity.
UoILasso extracts sparse, meaningful graphs
UoILasso extracts sparse, meaningful graphs
UoILasso extracts sparse, meaningful graphs

Change in Pred. Acc. ($\Delta R^2$)

Time (ms)

Syllable Production

UoILasso

SCAD
Identification of genetic basis of complex phenotypes

Need accurate prediction with small number of factors

Very hard:
Correlated factors
P >> N

Collaborative Cross Mice
(90% of genetic variability)
UoILasso identifies a small number genes contributing to behavioral phenotype
UoI\text{Lasso} identifies a small number of highly predictive parameters

<table>
<thead>
<tr>
<th>Weight</th>
<th>$R^2$</th>
<th>Selection Ratio</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>0.72</td>
<td>$3250 \times 10^{-4}$</td>
<td>13397</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.72</td>
<td>$360 \times 10^{-4}$</td>
<td>1298</td>
</tr>
<tr>
<td>UoI\text{Lasso}</td>
<td>0.73</td>
<td>$8.2 \times 10^{-4}$</td>
<td>-126</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speed</th>
<th>$R^2$</th>
<th>Selection Ratio</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>0.65</td>
<td>$350 \times 10^{-4}$</td>
<td>1327</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.65</td>
<td>$290 \times 10^{-4}$</td>
<td>1101</td>
</tr>
<tr>
<td>UoI\text{Lasso}</td>
<td>0.66</td>
<td>$6.9 \times 10^{-4}$</td>
<td>-129</td>
</tr>
</tbody>
</table>
### Drug Discovery

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prediction Accuracy</th>
<th>Selection Ratio (PSR)</th>
<th>Parsimony (BIC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dorothea</td>
<td>93%</td>
<td>53 x 10^{-5}</td>
<td>456</td>
</tr>
<tr>
<td>L₁-Logistic</td>
<td>93%</td>
<td>3 x 10^{-5}</td>
<td>174</td>
</tr>
<tr>
<td>UoI-L₁Logistic</td>
<td>93%</td>
<td>3 x 10^{-5}</td>
<td>174</td>
</tr>
</tbody>
</table>

### Cancer Prediction

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prediction Accuracy</th>
<th>Selection Ratio (PSR)</th>
<th>Parsimony (BIC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcene</td>
<td>66%</td>
<td>59 x 10^{-4}</td>
<td>437</td>
</tr>
<tr>
<td>L₁-Logistic</td>
<td>66%</td>
<td>5 x 10^{-4}</td>
<td>280</td>
</tr>
<tr>
<td>UoI-L₁Logistic</td>
<td>66%</td>
<td>5 x 10^{-4}</td>
<td>280</td>
</tr>
</tbody>
</table>

### Parkinsons Prediction

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prediction Accuracy</th>
<th>Selection Ratio (PSR)</th>
<th>Parsimony (BIC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parkinsons</td>
<td>65%</td>
<td>0.69</td>
<td>533</td>
</tr>
<tr>
<td>L₁Logistic</td>
<td>65%</td>
<td>0.69</td>
<td>533</td>
</tr>
<tr>
<td>UoI-L₁Logistic</td>
<td>68%</td>
<td>0.31</td>
<td>478</td>
</tr>
</tbody>
</table>
Summary/Conclusions

1) **UoI**: balance feature compression/expansion to maximize prediction accuracy
2) **UoI**: predictive, scalable, selective, accurate, stable
3) **UoILasso**: Sparse solutions with no prior
4) **UoILasso**: regression/ **UoIL1Logistic**: classification
5) Improved sparse prediction in neuroscience, genetics, disease prediction, drug discovery, …

**UoI**: scalable and accurate approximation of $L_0$-norm (an NP-hard problem)
Future Directions

**Supervised Learning**

- Generalized Linear Models
  - Regression
  - Classification

- Multiple Input-Multiple Outputs
  - CCA

- Non-linear Functions
  - Random Forests
  - Random Intersection Trees
  - MARS

- Time Series Analysis
  - Vector Autoregressive
  - Sparse Identification of Non-linear Dynamics

**Unsupervised Learning**

- Non-negative matrix Factorization

- CUR matrix decompositions

- Convolutional Sparse Coding

- ADMM/RLA/sub-sampling schemes
  - non-orthogonal designs