Stability and Aggregation of Experimental Results

Cheng Soon Ong

Machine Learning Research Group
Data61 | CSIRO
Australian National University

10 August 2017
ICML workshop on Computational Biology

What is machine learning?

Machine learning is about prediction

Examples/features	$x_1, \ldots, x_n \sim \mathcal{X}$
Labels/annotations	$y_1, \ldots, y_n \sim \mathcal{Y}$
Predictor	$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

Estimate best predictor = training

Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.

- No mechanistic model of the phenomenon
- There is relatively large amounts of data (examples, x usually $\mathbb{R}^d$)
- The outcomes (labels, y usually binary) are well defined

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

What are good features?

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

What to measure?

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

This talk ...

Use predictive model ...

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

... to discover good biomarkers

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

- Genome Wide Association Studies
- Spearman's Correlation
- Stability and Aggregation

Genome wide association study

Case-control studies
A cohort of sick individuals (cases) and healthy individuals (controls) are genotyped and their corresponding binary phenotype are recorded.

We use the framework of hypothesis testing

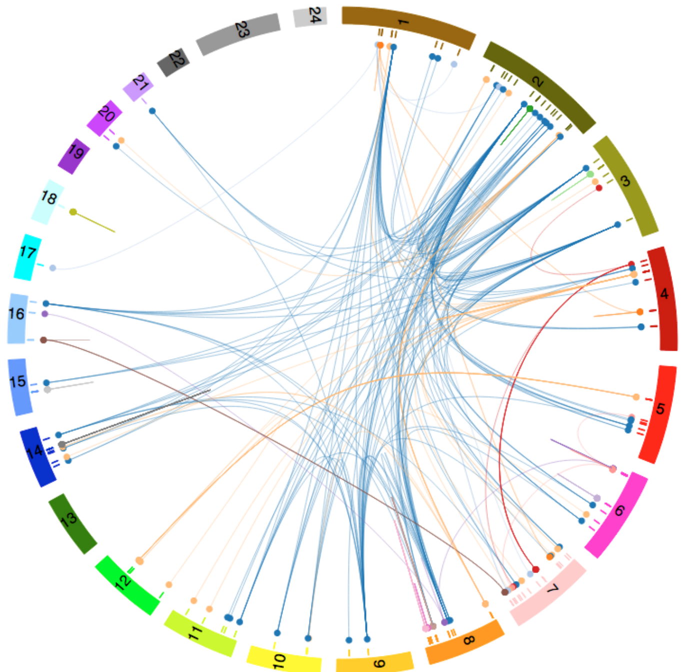
Hypothesis testing Given a case control study, test whether a particular SNP is associated with the phenotype.

Epistatic Interactions

- WTCCC data
- Need to tabulate 125 billion contingency tables
- Consider specificity and sensitivity
- Gain over univariate ROC
- CPU ($\approx$ days) and GPU ($\approx$ hours)
- Store the top 1 million pairs

Goudey,...,Ong,...,Kowalczyk, BMC Genomics, 2013

Interacting with results



D3.js

- Circular plot
- Linear plot
- Manhattan plot
- Heat map

Interaction

- Filter
- Zoom
- Drill down
- Call out

Cristovao Freitas Iglesias Junior, Stefan Sevelda

github.com/chengsoon.ong/rede

p-values

Interpreting p-values

Is 10^{-10} probability of association very significant?

Quote

... but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

Fisher, The Design of Experiments, 1947, p. 14

Stability of scoring

We consider p-values as a score of association.

- How stable is this score if we repeat the experiment?
- How do we combine scores?

Challenges

- Scores available for only the top-k examples
- Scores from different sources not calibrated

Outline

Use predictive model ...

$$\color{red}{f_{\mathbf{w}}(x)} : \mathcal{X} \to \mathcal{Y}$$

... to discover good biomarkers

$$f_{\mathbf{w}}(x) : \color{red}{\mathcal{X}} \to \mathcal{Y}$$

Stability of feature selection

How to measure overlap?

Rank aggregation

How to combine different sources of information?

Modeling using Spearman's correlation

How to represent ranks?

Multiple ways to represent ranks

- Ordered list of n objects selected from Ω
- List of values $[1, \ldots, n]$ (the ranks of the object)
- Normalised ranks $\in (0, 1)$
- Permutation mapping $R : \Omega \to (0, 1)$

Measuring Overlap

Motivation

Given a set of replicated experiments, how do we measure overlap?

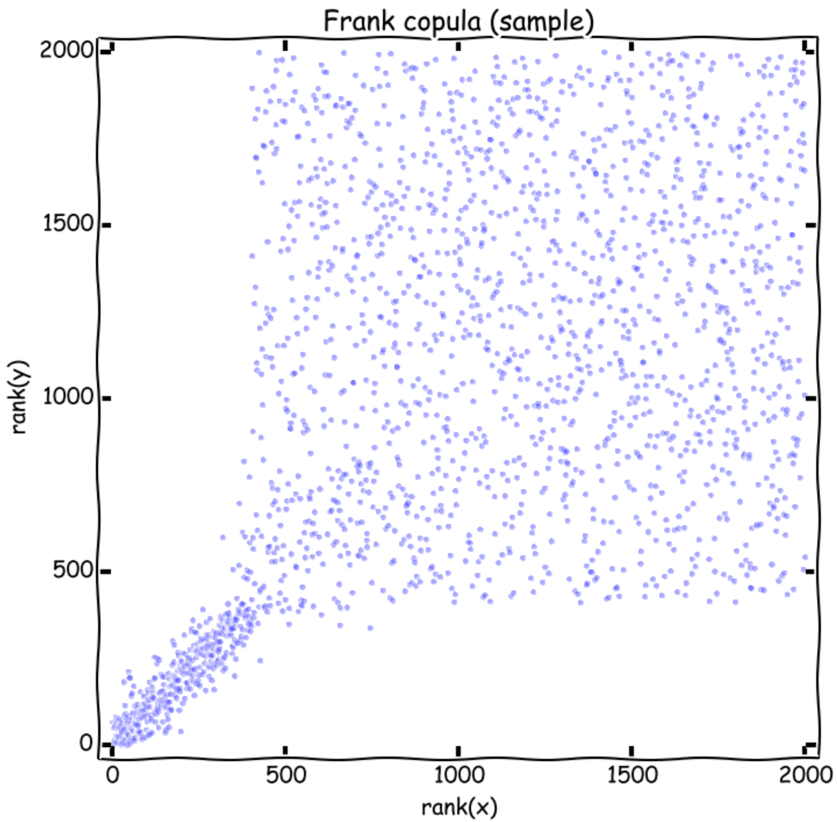
Examples

- Perform repeated splits of the data
- Experiments on different cohorts
- Multiple sources of information

Challenges

- Scores available for only the top-k examples
- Scores from different sources not calibrated

Signal and Noise



Set based overlap

Running example (6 objects)

$$A = [a, b, c, d, e, f]$$

$$B = [a, b, e, f, c, d]$$

Jaccard Index

$$\text{overlap} = \frac{|A \cap B|}{|A \cup B|}$$

Measuring stability

- Easy to compute
- Works for top-k lists
  Consider the top-3 lists from above:

$$\text{Jaccard index} = \frac{|\{a, b\}|}{|\{a, b, c, e\}|} = \frac{1}{2}$$

- Ignores the order given by scores

Spearman's ρ

- Similar to Pearson's correlation for the measure of dependence
- Spearman's ρ is a correlation measure between ranked lists

$$\rho(A,B) := \frac{\sum_i (r_A^{(i)} - \bar{r}_A)(r_B^{(i)} - \bar{r}_B)}{\sqrt{\sum_i (r_A^{(i)} - \bar{r}_A)^2 \sum_i (r_B^{(i)} - \bar{r}_B)^2}},$$

- Running example:

$$\rho([a,b,c,d,e,f],[a,b,e,f,c,d]) = 0.543$$

(Jaccard index $= 1$)

- Need the same elements in A and B

$$\rho([a,b,c],[a,b,e]) \text{ ?}$$

Spearman's ρ on top k lists

Our idea

Define Spearman's ρ for top k lists

Key observation

Any elements in list A that do not appear in list B must have a rank higher than the number of elements in B

Running example (top-3)

$$A = [a, b, c, d, e, f] \quad \text{and} \quad B = [a, b, e, f, c, d]$$

$$A_3 = [a, b, c] \quad \text{and} \quad B_3 = [a, b, e]$$

$$A_3^{\overset{B_3}{\rightarrow}} = [a, b, c, e] \quad \text{and} \quad B_3^{\overset{A_3}{\rightarrow}} = [a, b, e, c]$$

Spearman's $\rho = \rho(A_3^{\overset{B_3}{\rightarrow}}, B_3^{\overset{A_3}{\rightarrow}}) = 0.8$

Spearman's ρ on top k lists

Extend the list

We expand lists A and B to complete rankings over the same set of elements, denoting them as $A^{\overset{B}{\rightarrow}}$ and $B^{\overset{A}{\rightarrow}}$ respectively.

The missing values in the extension are given the average rank.

Running example (top-4)

$$A_4 = [a, b, c, d] \qquad \text{and} \qquad B_4 = [a, b, e, f]$$

$$A_4^{\overset{B_4}{\rightarrow}} = [1, 2, 3, 4, 5.5, 5.5] \qquad \text{and} \qquad B_4^{\overset{A_4}{\rightarrow}} = [1, 2, 5.5, 5.5, 3, 4]$$

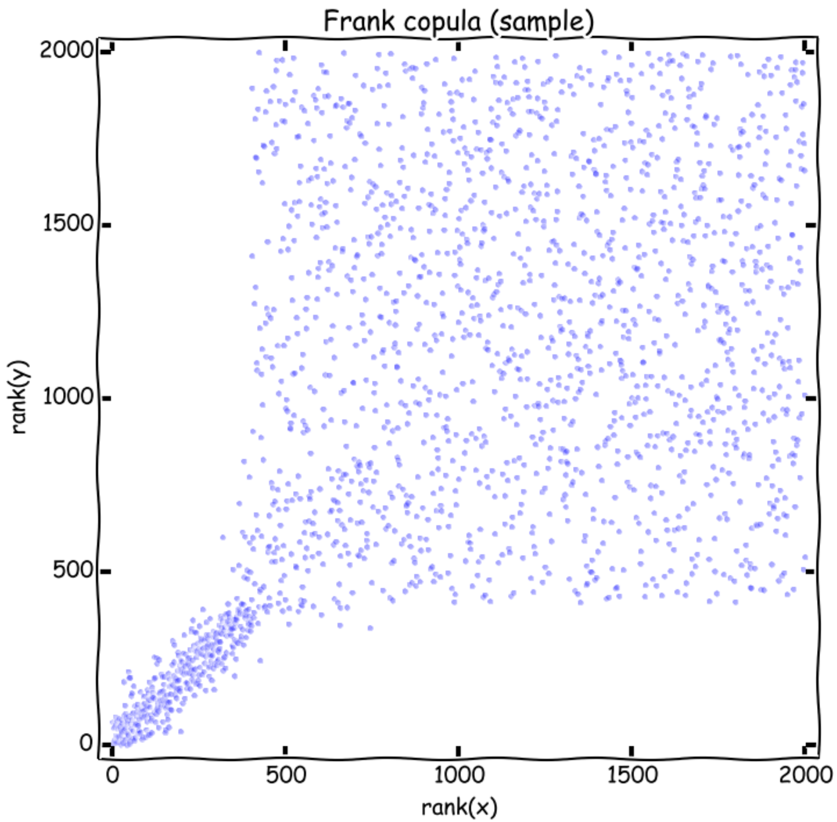
Makes no assumption about the order of the unranked objects

Other possible imputation approaches

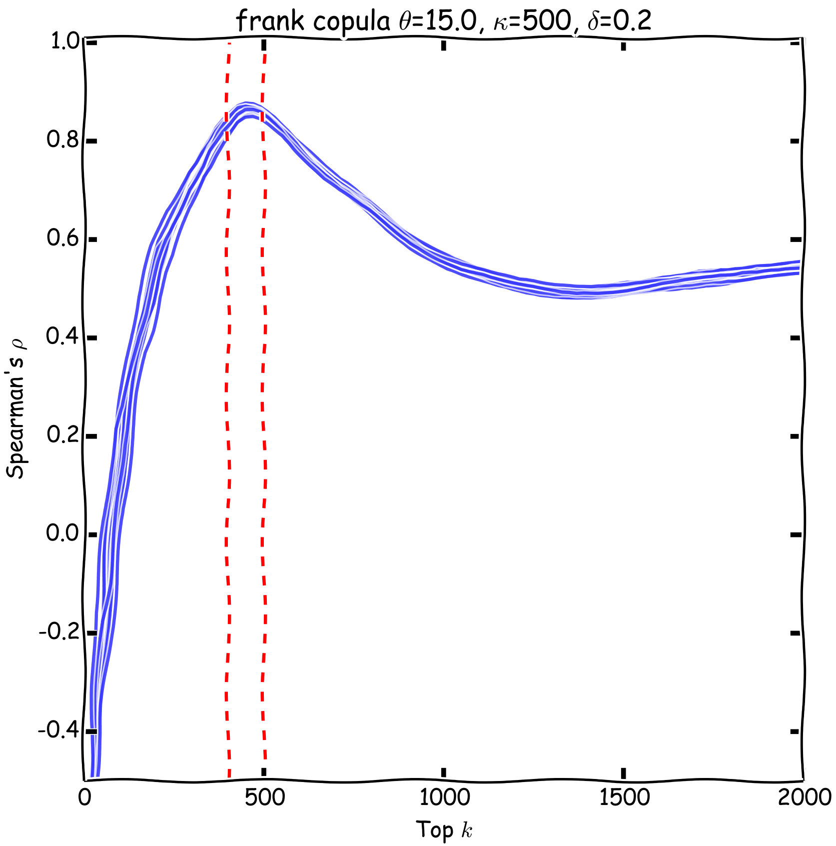
- Optimistic
- Worst case

Bedő, Rawlinson, Goudey, Ong, PLoS ONE, 2014

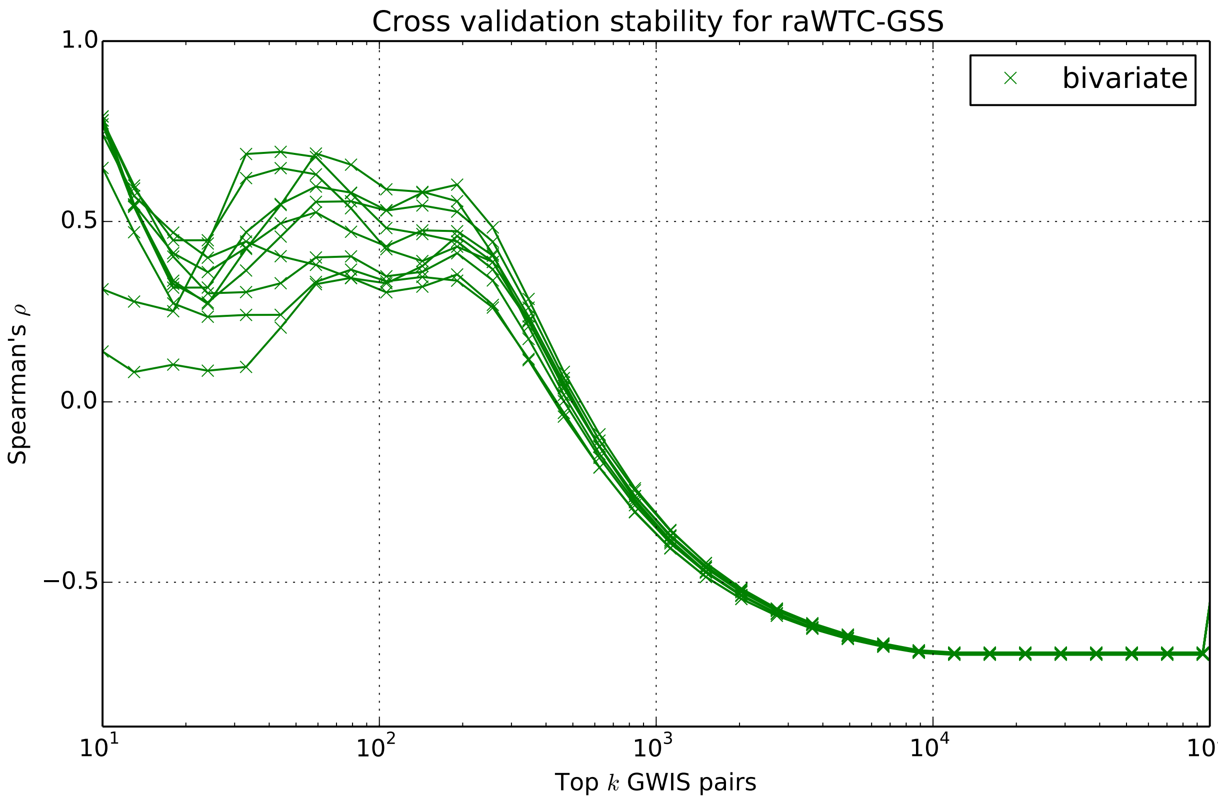
Signal and Noise



Spearman's ρ



Simulate two cohorts by splitting



Bedő, Rawlinson, Goudey, Ong, PLoS ONE, 2014

Measuring Overlap

Motivation

Given a set of replicated experiments, how do we measure overlap?

Challenges

- Scores available for only the top-k examples
- Scores from different sources not calibrated

Model

- Ranked list Instead of just using set intersection, we can use the scores from GWIS to order the results
- top k Traditional methods (Spearman's ρ) requires ranks for the whole list. We have incomplete information, but we know our ranks are the top ones.
- Multivariate Textbook Spearman's ρ is for computing correlation between two ranks. We want to compute the correlation between multiple ranked lists.

Copulas

Intuition

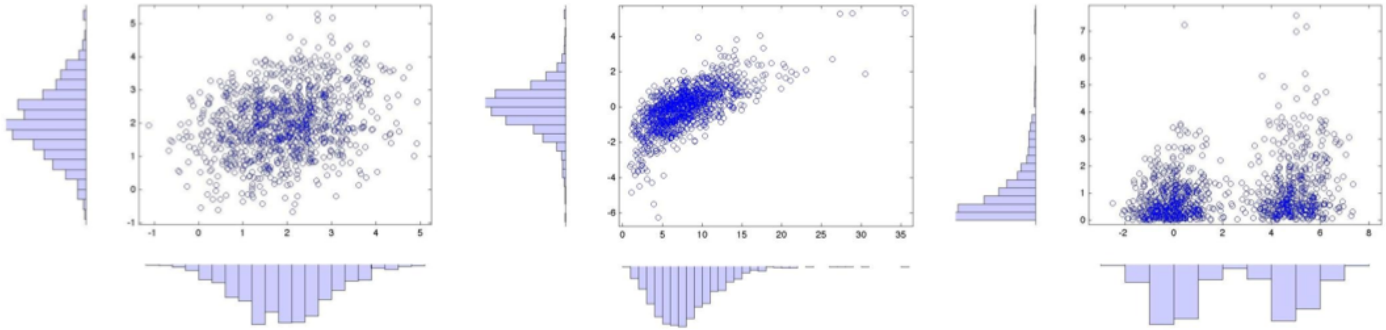
For continuous random variables, copulas model the dependence component after discounting for univariate marginal effects

Probabilistic definition

Let $U_1, \ldots, U_d$ be real random variables $\sim U([0,1])$.
A copula function $C : [0,1]^d \longrightarrow [0,1]$ is a joint distribution

$$C_\theta(u_1, \ldots, u_d) = P(U_1 \leqslant u_1, \ldots, U_d \leqslant u_d)$$

The same Gaussian copula function



Copulas and Spearman's ρ

Spearman's ρ can be expressed in terms of the copula

$$\rho(A, B) = 12 \int_{[0,1]^2} C(u, v) du dv - 3$$

Proof

$$\begin{aligned}
\rho(A, B) &= \frac{\sum_i (r_A^{(i)} - \bar{r}_A)(r_B^{(i)} - \bar{r}_B)}{\sqrt{\sum_i (r_A^{(i)} - \bar{r}_A)^2 \sum_i (r_B^{(i)} - \bar{r}_B)^2}} \\
&= \frac{\mathbb{E}[F(X)G(Y)] - \mathbb{E}[F(X)]\mathbb{E}[G(Y)]}{\mathrm{STD}[F(X)]\mathrm{STD}[G(Y)]} \\
&= \frac{\mathbb{E}[F(X)G(Y)] - \frac{1}{2}^2}{\frac{1}{12}} \\
&= 12\mathbb{E}[F(X)G(Y)] - 3 \\
&= 12 \int \int uvC(u, v) - 3 \\
&= 12 \int_{[0,1]^2} C(u, v) du dv - 3
\end{aligned}$$

Copulas and Spearman's ρ

Spearman's ρ can be expressed in terms of the copula

$$\rho(A, B) = 12 \int_{[0,1]^2} C(u, v) du dv - 3$$

Empirical copula

$$C_n(u, v) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \mathbf{1} \left(R(x) \leqslant u, S(x) \leqslant v \right)$$

Why do the math?

- Unclear how to extend formula for Spearman's correlation.
- Multivariate distributions $\Rightarrow$ multivariate copula.

Multivariate Spearman's ρ

A multivariate extension of Spearman's ρ

For a d dimensional set of random variables $\mathbf{u}$, the multivariate Spearman's ρ is given by

$$\rho(R_1, \ldots, R_d) = Q(C, \pi) = h(d) \left(2^d \int_{[0,1]^d} \pi(\mathbf{u}) \, \mathrm{d}C(\mathbf{u}) - 1 \right),$$

where

$$h(d) = \frac{d+1}{2^d - (d+1)}.$$

Empirical multivariate Spearman's corelation

$$\rho_n(R_1, \ldots, R_d) = h(d) \left[\frac{2^d}{n} \sum_x \prod_{j=1}^{d} R_j(x) - 1 \right].$$

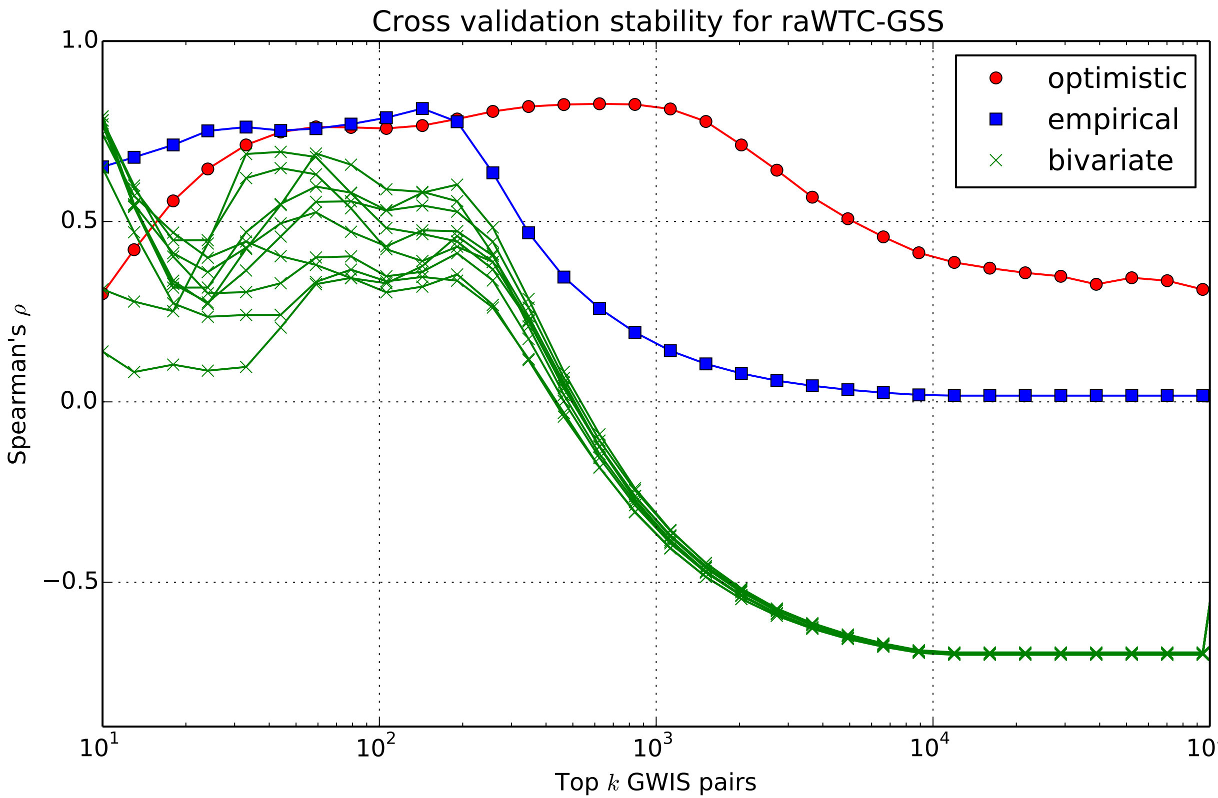
Multivariate Spearman correlation

```
n_rank /= n+1.
prod = np.prod(1.-n_rank, axis=1)
total = np.sum(prod)
s_rho = h(d) * ((2**d/float(n)) * total - 1.)
```

Empirical multivariate Spearman's corelation

$$\rho_n(R_1, \ldots, R_d) = h(d) \left[\frac{2^d}{n} \sum_x \prod_{j=1}^{d} R_j(x) - 1 \right].$$

Multiple replicates



Wait... there's more

Modeling using Spearman's correlation

Use predictive model ...

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

... to discover good biomarkers

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Stability of feature selection

How to measure overlap?

$$\rho(R_1, \ldots, R_d)$$

Rank aggregation

How to combine different sources of information?

Macintyre, Yepes, Ong, Verspoor, PeerJ, 2014

Optimal aggregator: geometric mean

How to combine different sources of information?

We maximise multivariate correlation

$$R^* = \arg\max_R \rho(R, R_1, R_2, \ldots, R_d).$$

Theorem The aggregator that maximises multivariate Spearman's correlation is the product of the normalised ranks.

Use the geometric mean

NOT pairwise correlation

Instead of decomposing the association into a combination of pairwise similarities $\rho(R, R_1), \rho(R, R_2), \ldots, \rho(R, R_d)$.

Method

1. Divide rank by number of items
2. return log average

Bedő, Ong, JMLR 17(201):1–30, 2016

Supervised learning to rank

Problem setting

We are given a ranking L of n objects which comprise our labels, and a set of d experts $\{R_j\}$.
Find a weighting of the experts.

Relaxed optimal aggregator

We solve the least squares problem

$$\min_{\omega} \sum_{x} \left(l(x) - \sum_{j=1}^{d} \omega_j r'_j(x) \right)^2 ,$$

where the outer sum is over the n examples x,
$l(x)$ is the log scaled normalised labels,
$r'_j(x)$ is the log scaled normalised completed ranks.

LETOR 4.0

We (surprisingly) perform much better than state of the art

Bedő, Ong, JMLR 17(201):1–30, 2016

Why is ML for Comp Bio hard?

Deep data

- High dimensional low sample size data
- Finite domain (we want predictions genome wide)
- Difficult to satisfy i.i.d. assumption

Prior knowledge

- Hard and soft constraints
- Weak constraints
- Causal systems with feedback and delays

Different problems/communities

- Genomics, proteomics, medical imaging, health records
- Communication challenges

Keep up the great work!

Summary

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

- Use predictive model to discover good features

Spearman's correlation - applied to GWAS

- Stability of scoring
- Rank aggregation
- Supervised learning to rank
- Imputation from top-k lists
- Multivariate correlation using copulas

The dream

Facilitating scientific knowledge discovery
through automated experimental design with machine learning

Thank You

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

- Use predictive model to discover good features

Spearman's correlation - applied to GWAS

- Stability of scoring
- Rank aggregation
- Supervised learning to rank
- Imputation from top-k lists
- Multivariate correlation using copulas

The dream

Facilitating scientific knowledge discovery
through automated experimental design with machine learning

Please make your research open

machlearn.gitlab.io/fmml/ www.ong-home.my

Scoring candidates - ABCDE

Active Learning

- Choose a particular example to label using heuristics
- Annotator assumed to provide ground truth

Bandits

- Select a choice from a set of actions
- Simple algorithms with theoretical guarantees
- Manage uncertainty with repeated sampling

Choice theory

- Aggregate set of ranks into one ordering
- Economics and social science, impossiblity theorems

Designing Experiments

- Choose a set of trials to measure
- Optimisation algorithms with theoretical analysis
- Information theory, real random variables