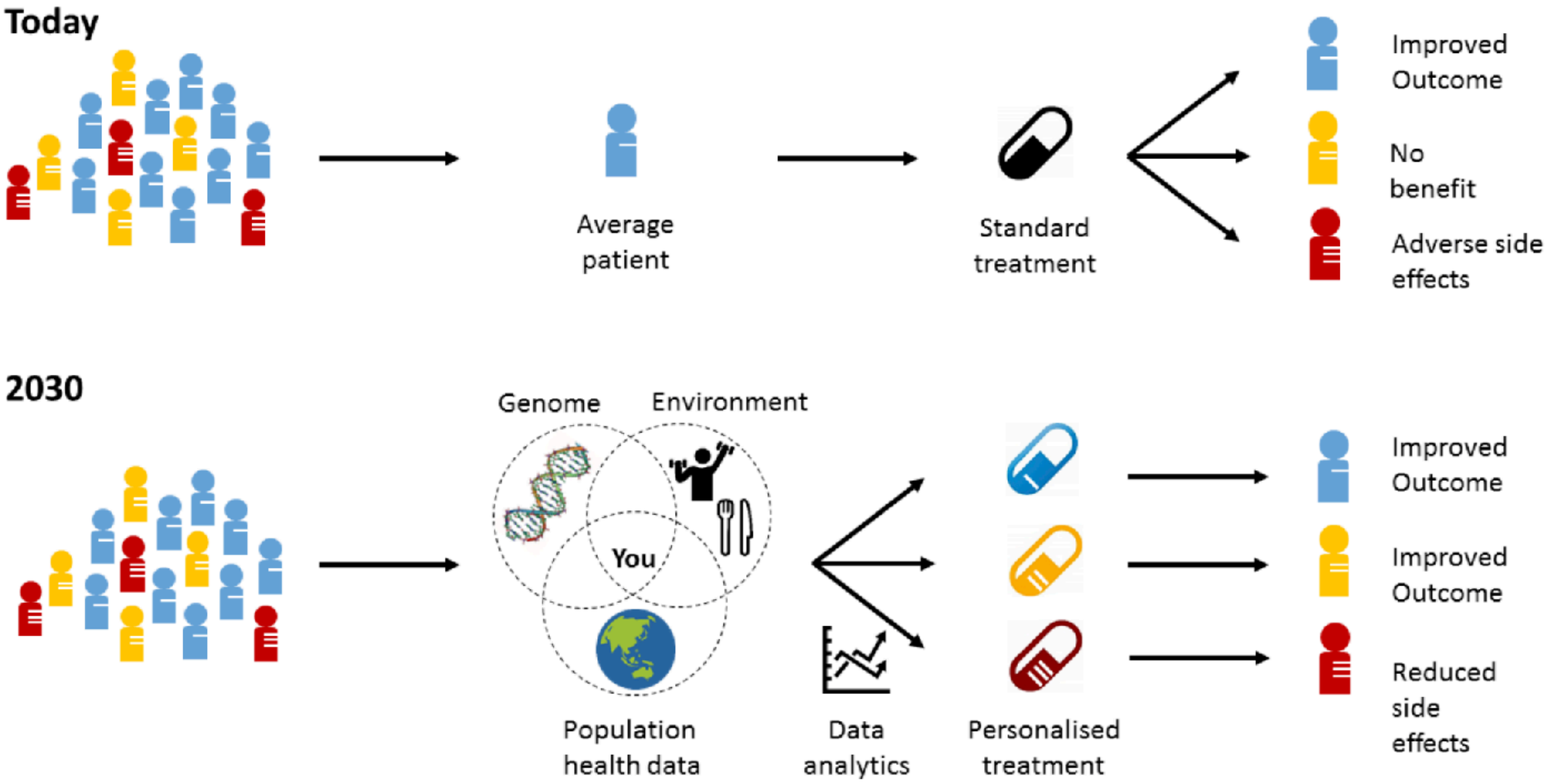
Kevin Wang

From linear regression to precision medicine

Precision medicine: predicting best cause of action using omics data

https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf

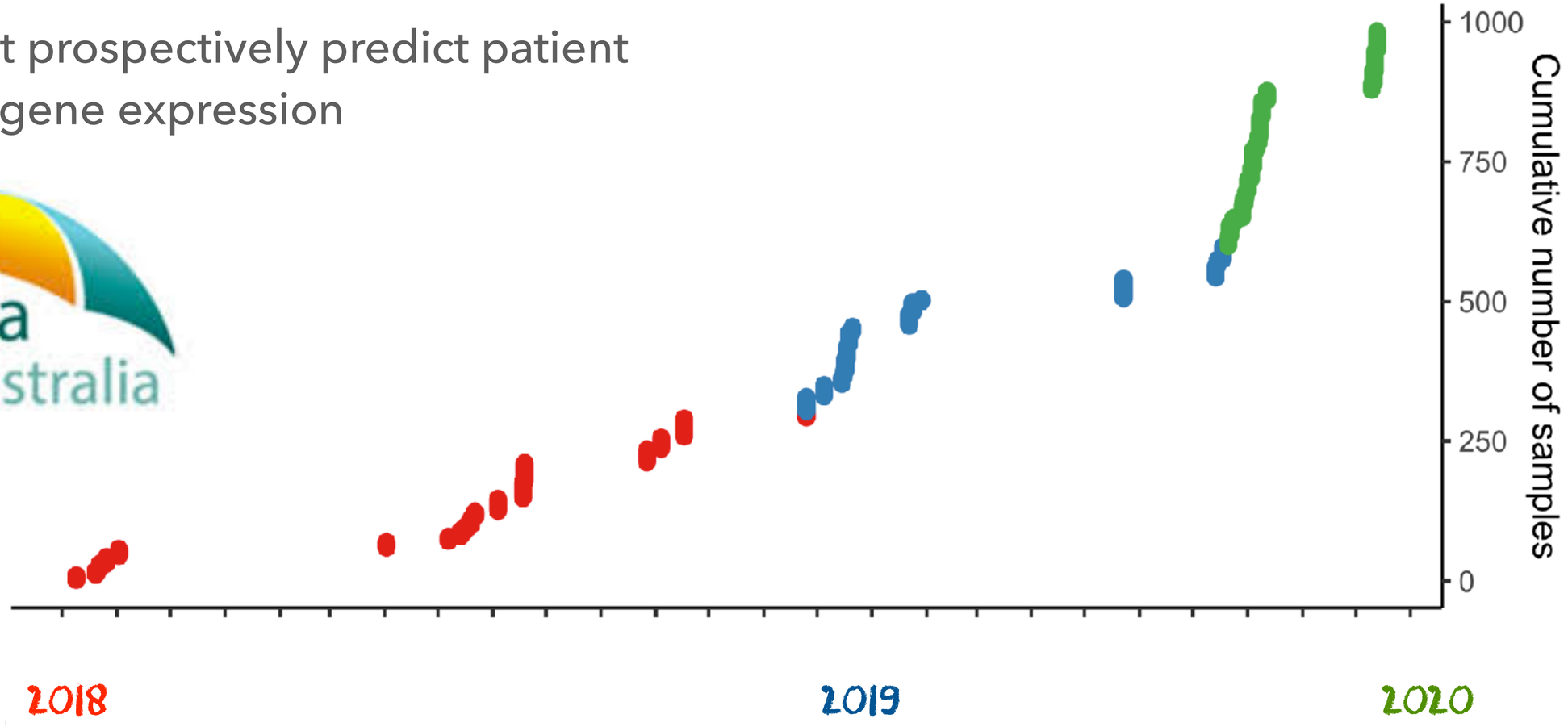
Precision medicine: predicting best cause of action using omics data



https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf

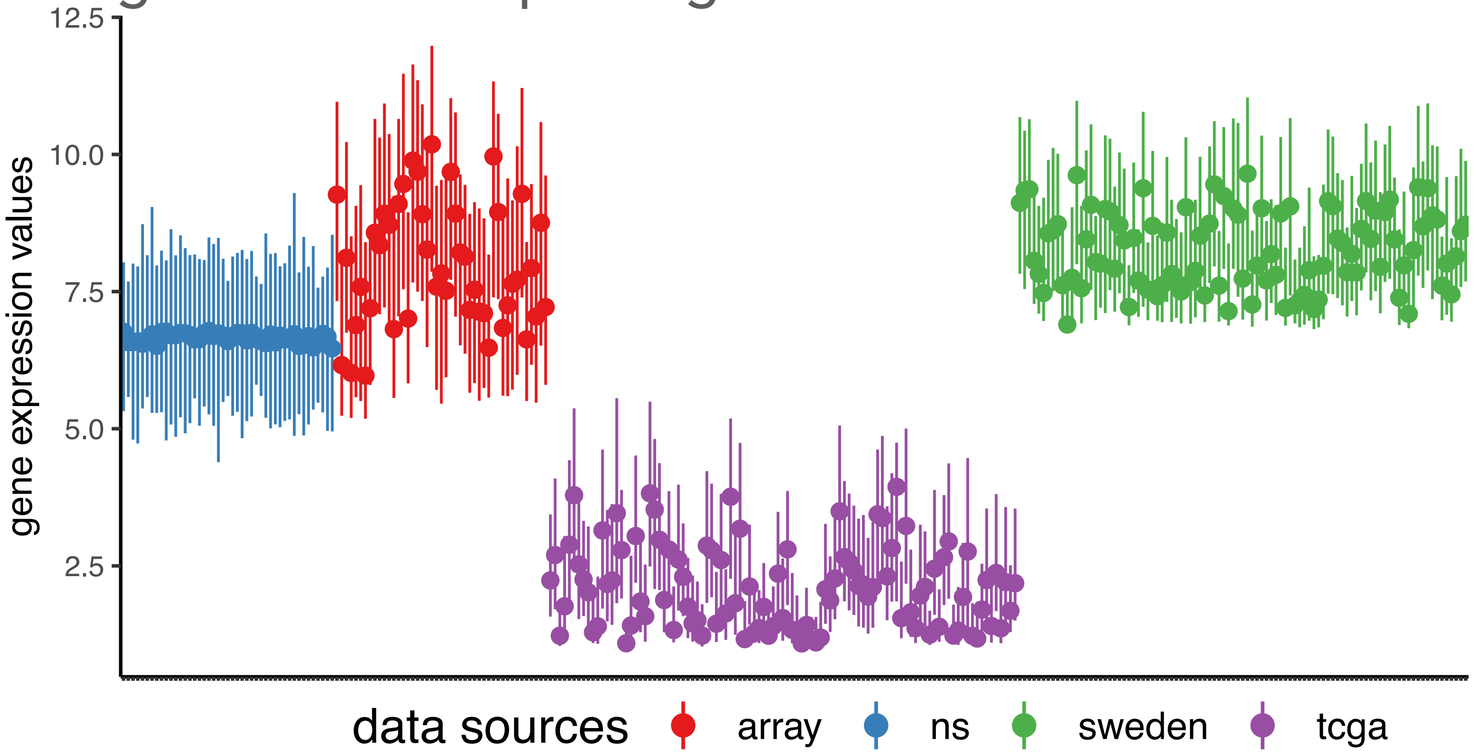
Melanoma Institute Australia

- A framework that prospectively predict patient outcomes using gene expression



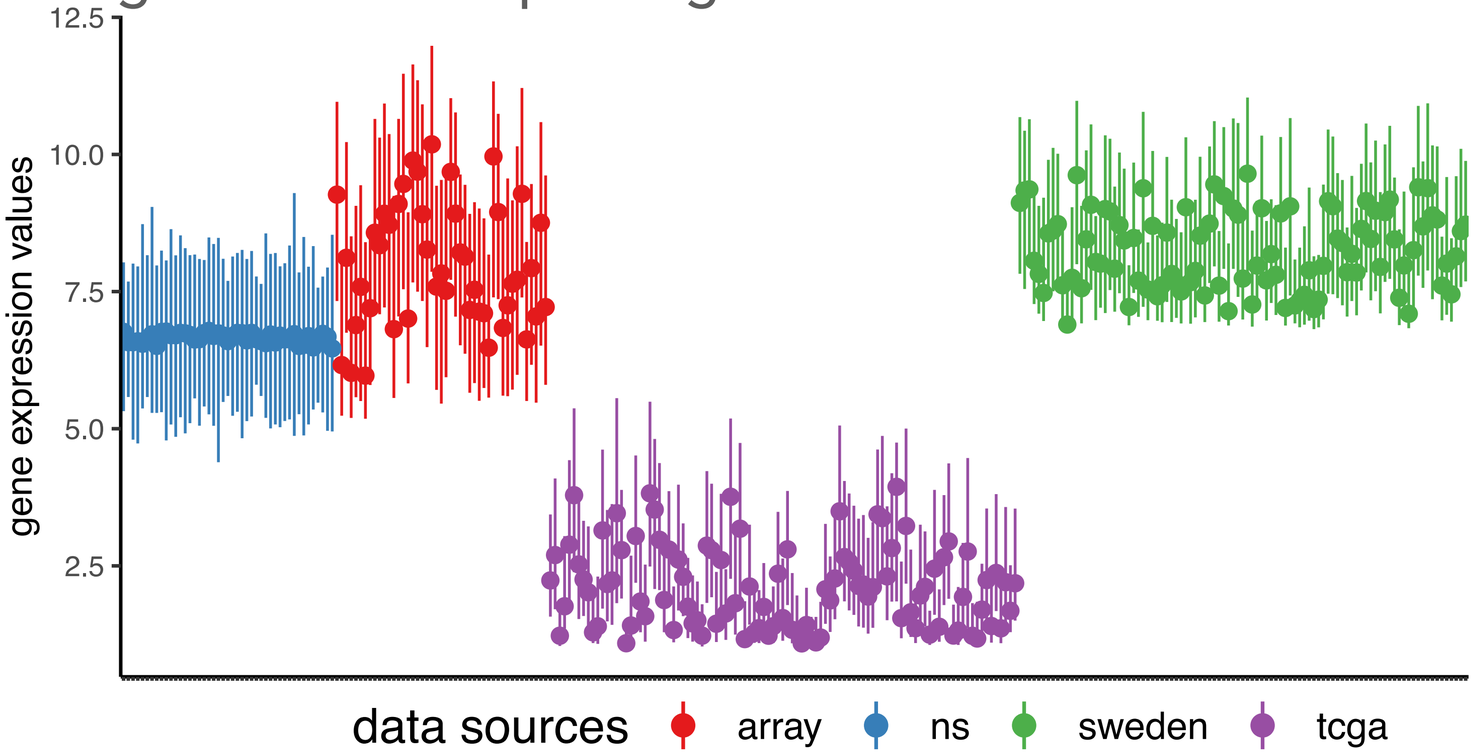
Melanoma Institute Australia

- A framework that prospectively predict patient outcomes using gene expression
- A generalisation: pooling data from various sources



CPOP: Stable prediction for patient survival status using genetic information

- A framework that prospectively predict patient outcomes using gene expression
- A generalisation: pooling data from various sources



1. Traditional statistical methods can't handle high dimensional data
2. High dimensional methods are not always stable
3. Without stability, you can't push this technology out to the clinics ethically

Outline

- Chapter 1: underlying methods
  - Low-dimensional
  - High-dimensional
- Chapter 2: CPOP - Cross Platform Omics Prediction
  - Rationale
  - Results and implications

Chapter 1:
fitting separate models to each data

Section 1:
Regression analysis before the 20th century (low dimensional)

A system of linear equations

y

Tumour shrinkage
5.9
2.1
4.7
3.2

X

Dosage	Age
1.2	32
2.4	21
3.1	46
1.2	31

A system of linear equations

"complex"

y

Tumour shrinkage
5.9
2.1
4.7
3.2

X

Dosage	Age
1.2	32
2.4	21
3.1	46
1.2	31

"simple"

A system of linear equations

$$y = X\beta + \epsilon$$

Tumour shrinkage	Dosage	Age
5.9	1.2	32
2.1	2.4	21
4.7	3.1	46
3.2	1.2	31

β

explains how each variable contribute to the final response

A system of linear equations

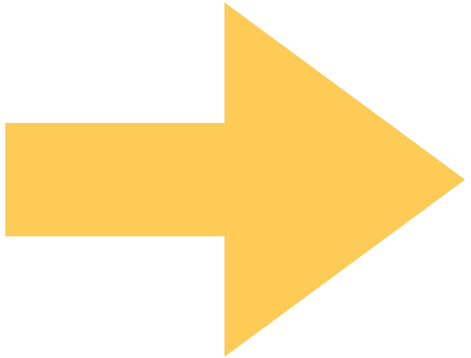
$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

Tumour shrinkage	Dosage	Age
5.9	1.2	32
2.1	2.4	21
4.7	3.1	46
3.2	1.2	31

β explains how each variable contribute to the final response

Framingham heart disease risk score:

- Age (Years)
- Cholesterol (mg/dL)
- If smoker (Yes/No)
- HDL cholesterol (mg/dL)
- Systolic blood pressure (mm Hg)



$$\hat{y} = X\hat{\beta}$$

20 points model

Least squares regression as a minimisation problem

The solution to:

$$\underset{\beta}{\operatorname{argmin}} \qquad \|y - X\beta\|_2^2$$

is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

Problem: when will least squares solution fails

$$\text{If} \quad n > p$$

everything is great!

https://github.com/kevinwang09/shrink_shiny

Problem: when will least squares solution fails

If $n > p$

everything is great!

https://github.com/kevinwang09/shrink_shiny

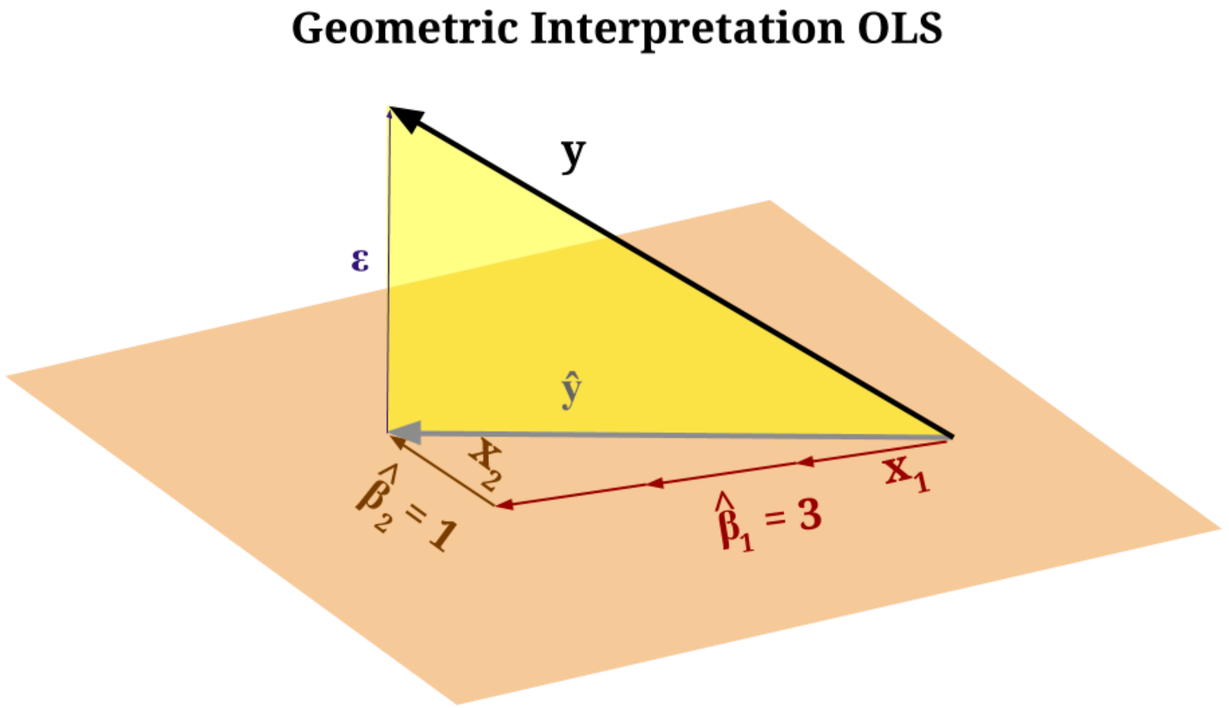
If $n \leq p$

more parameters to be estimated than observations

$\hat{\beta}$ blows up (not estimable)!

Linear regression is also a projection

- A linear regression aims to explain as much complications in the response variable using the predictors as possible.
- L2 projection of y upon a subspace spanned by the column vectors of the predictor matrix.



Section 2:
Linear regression of the 20th century (high dimensional)

Least squares regression (1700's)

Least squares regression solves:

$$\underset{\beta}{\operatorname{argmin}} \qquad \|y - X\beta\|_2^2$$

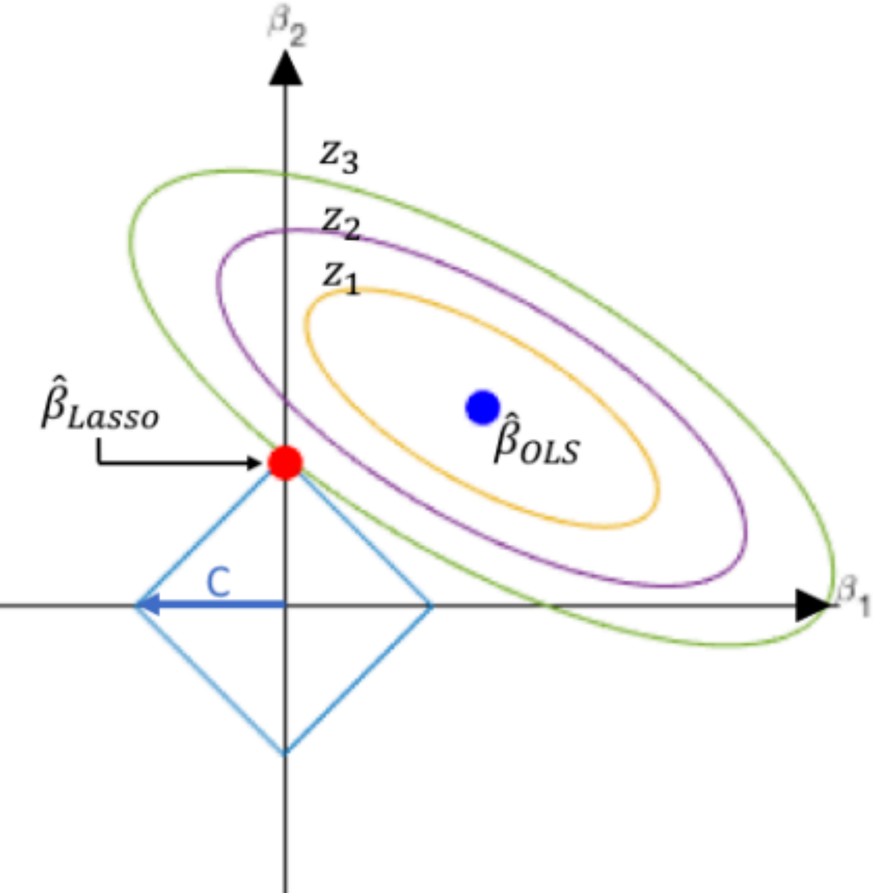
Lasso – Least Absolute Shrinkage and Selection Operator (1996)

Lasso solves:

$$\underset{\beta}{\operatorname{argmin}} \qquad \|y - X\beta\|_2^2 + \lambda \, \|\beta\|_1$$

- Pure optimisation of the first term makes $\hat{\beta}$ blow up, so you penalise on its size at the same time
- λ is usually chosen by refitting the equation using resampled data*
- The original paper, Tibshirani 1996, was cited more than 30,000 times

Why is Lasso so popular?

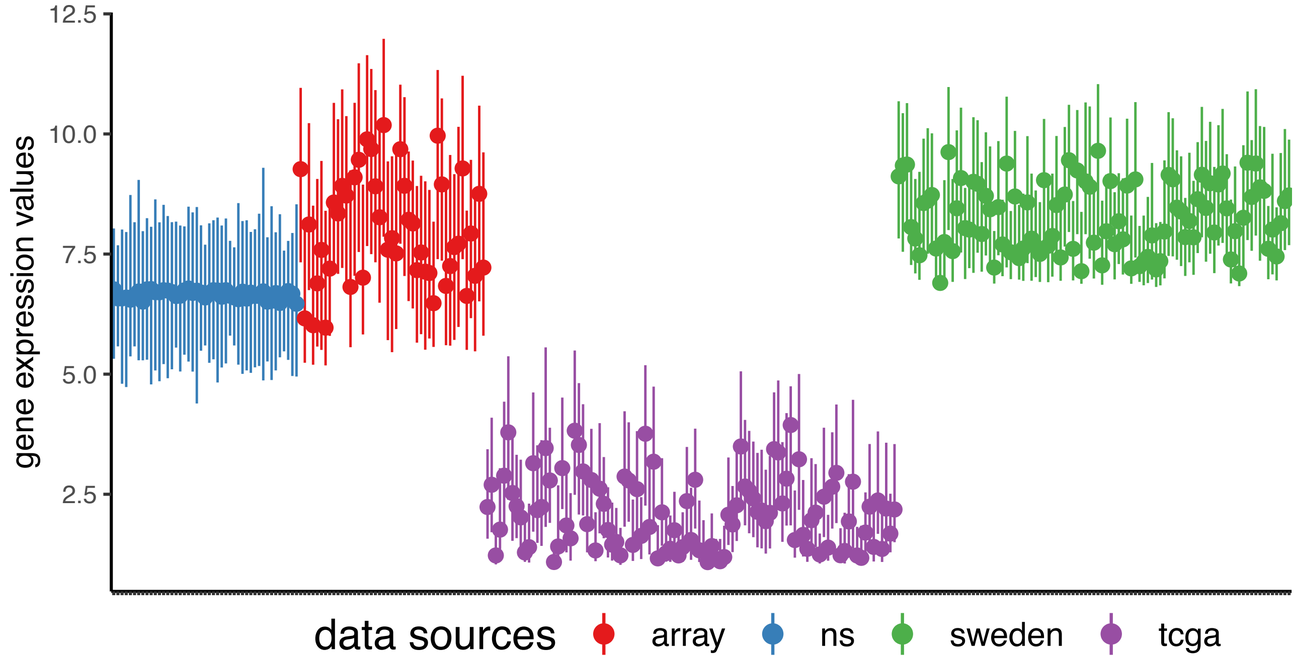


- Shrinks some coefficients to exactly 0
- Every variable introduces variations into model estimation and prediction
- Informative variables: no worries!
- Non-informative variables: a damn nuisance!

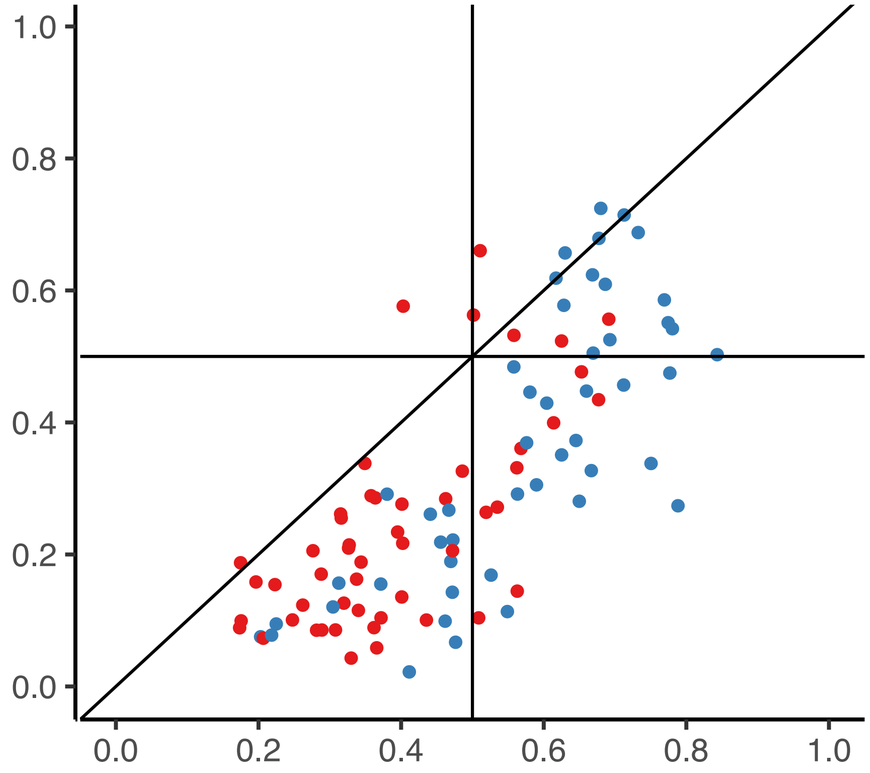
Variable selection aims to keep the informative variables and eliminate the non-informative ones

Lasso on individual dataset

When training data and validation data are not of the same statistical shape, any model would do miserably.



Train: Sweden
Test: TCGA



Train: TCGA
Test: TCGA

But don't cry! Whenever Lasso fails, a small modification is often all you need

- **Logistic Lasso** deals with binary response:

$$\underset{\beta}{\operatorname{argmin}} \quad \left\| y - \frac{1}{1+\exp\left(X\beta\right)} \right\|_2^2 + \lambda \left\| \beta \right\|_1$$

- **Elastic Net** deals with highly correlated variables:

$$\underset{\beta}{\operatorname{argmin}} \quad \|y - X\beta\|_2^2 + \lambda \left[(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right]$$

- **Weighted Lasso** forces some variables into the model

$$\underset{\beta}{\operatorname{argmin}} \quad \|y - X\beta\|_2^2 + \lambda \left\| w \circ \beta \right\|_1$$

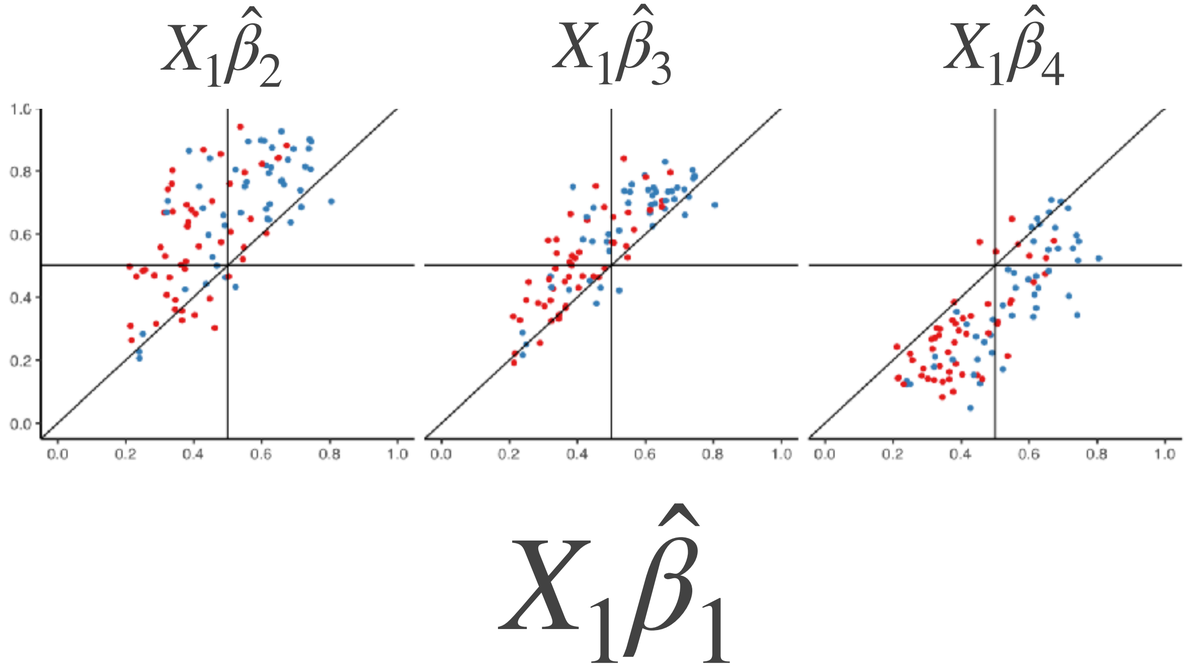
Chapter 2:
CPOP – Cross platform Omics Prediction

Statistical challenges

1. Genes must have similar scaling across datasets
2. Variable selection must be stable across datasets
3. Single-patient prediction

Transferability

For the same samples,
the prediction from one gene expression platform
should be equivalent to another platform



First component of CPOP: feature engineering

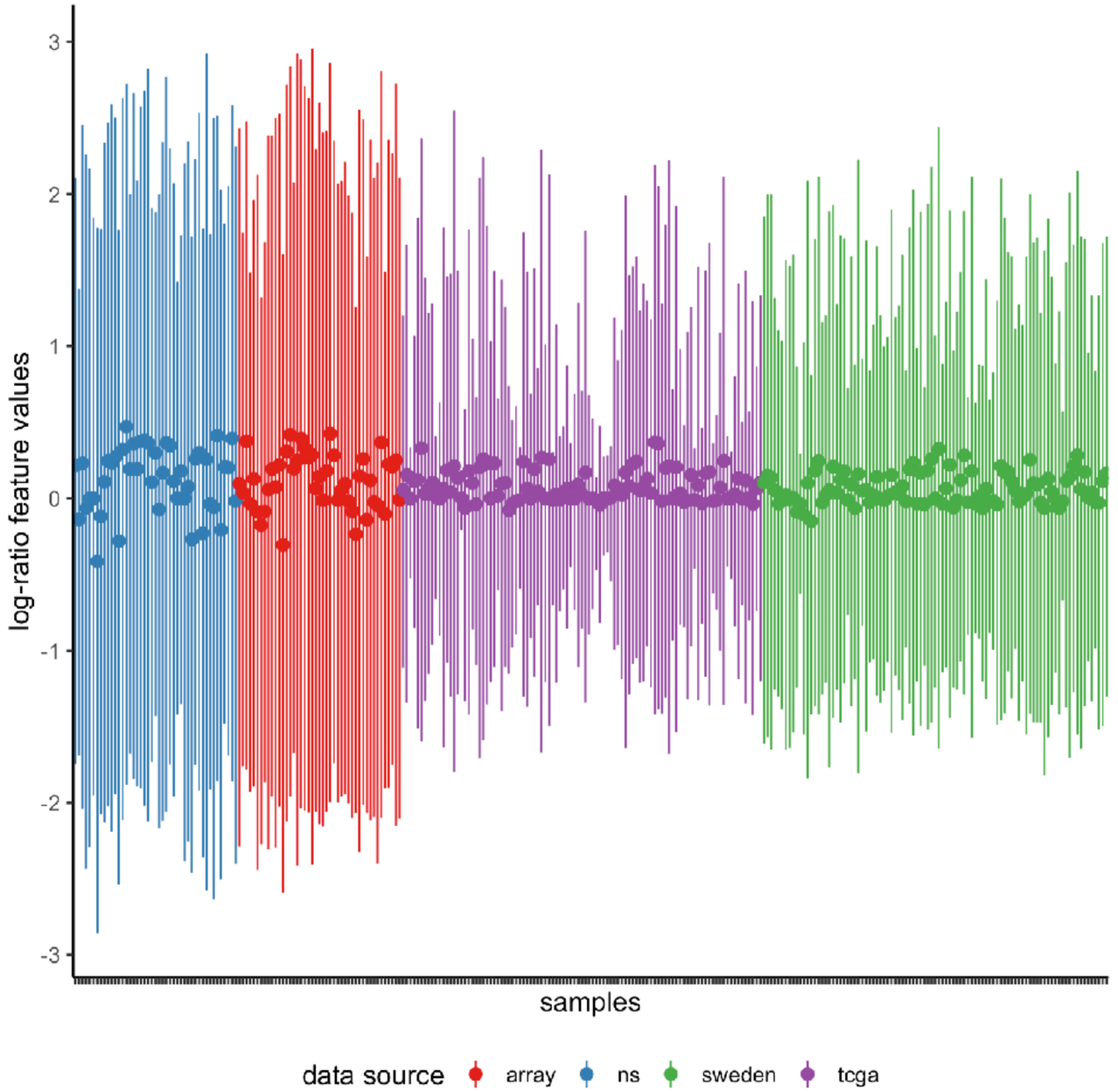
就让我 来次透彻心扉的痛
都拿走 让我再次两手空空
只有奋奋一息过
那个真正的我
他才能够诞生

Within-sample feature standardisation

Single-patient prediction prevents us from calculating any cross-sample statistics, so the natural solution is within-sample standardisation

Log-ratio

log(gene A) – log(gene B)



The solution is trivial?

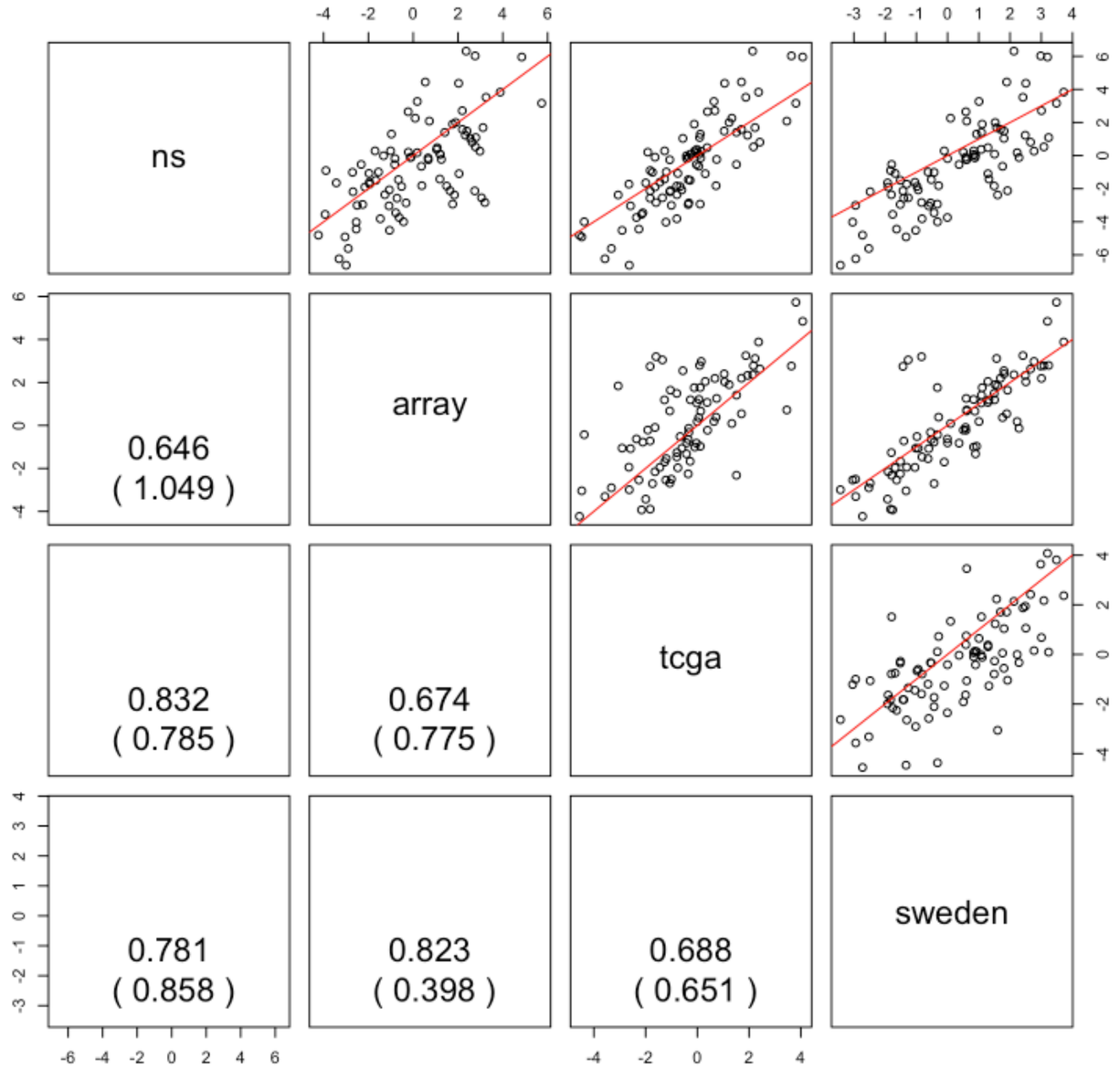
1. Genes must have similar scaling across datasets

2. Variable selection must be stable across datasets

3. Single-patient prediction

The solution is trivial?

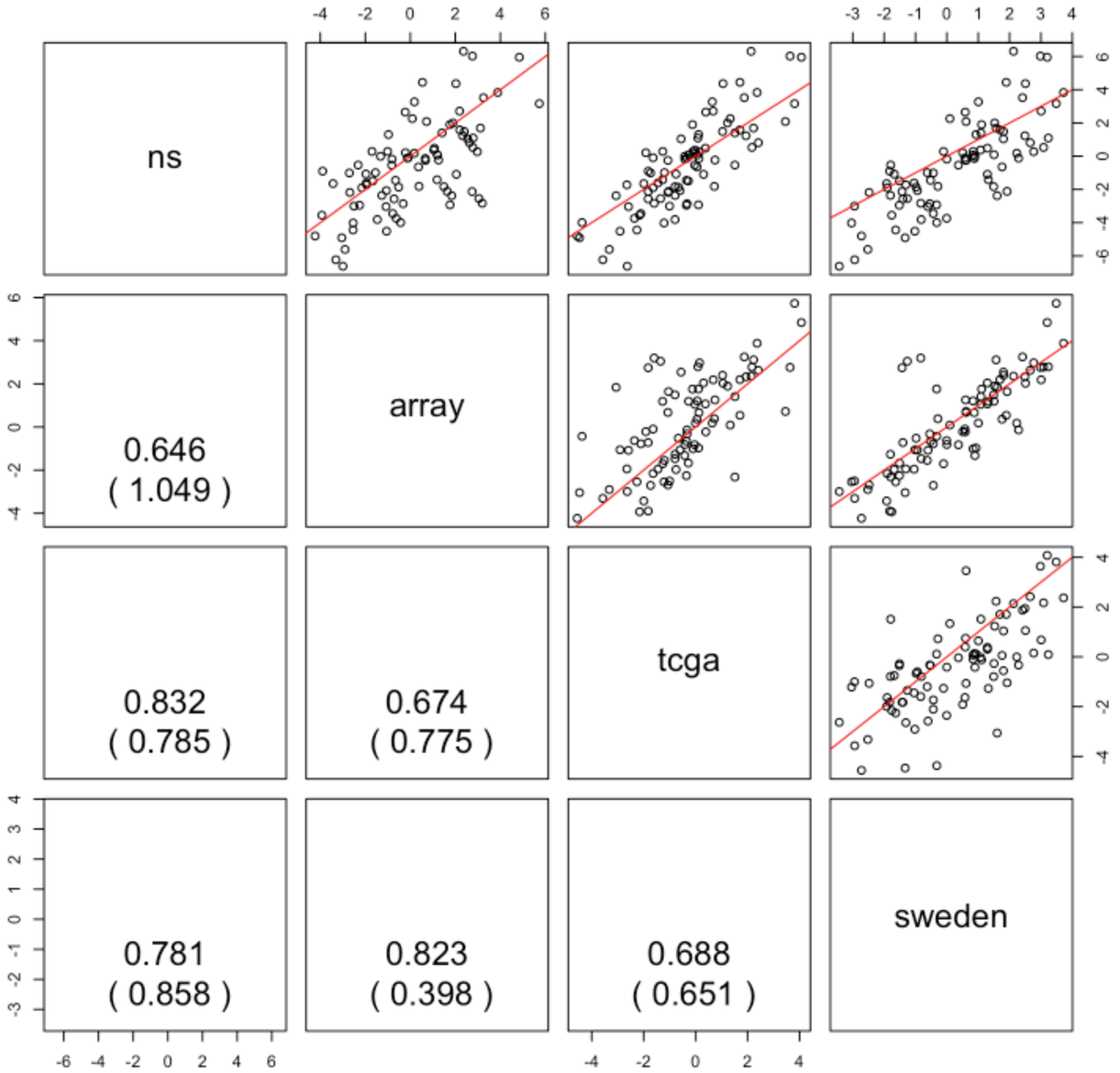
1. Log-ratios must have similar scaling across datasets
2. Variable selection must be stable across datasets
3. ~~Single-patient prediction~~



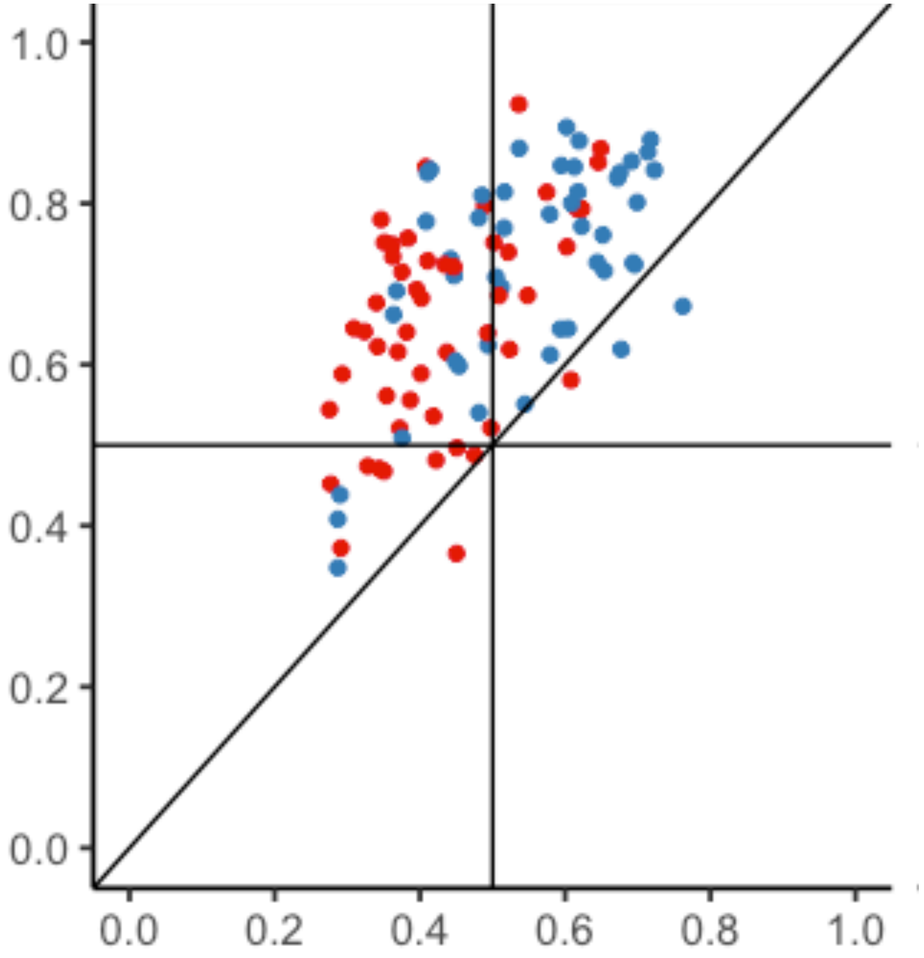
The solution is trivial?

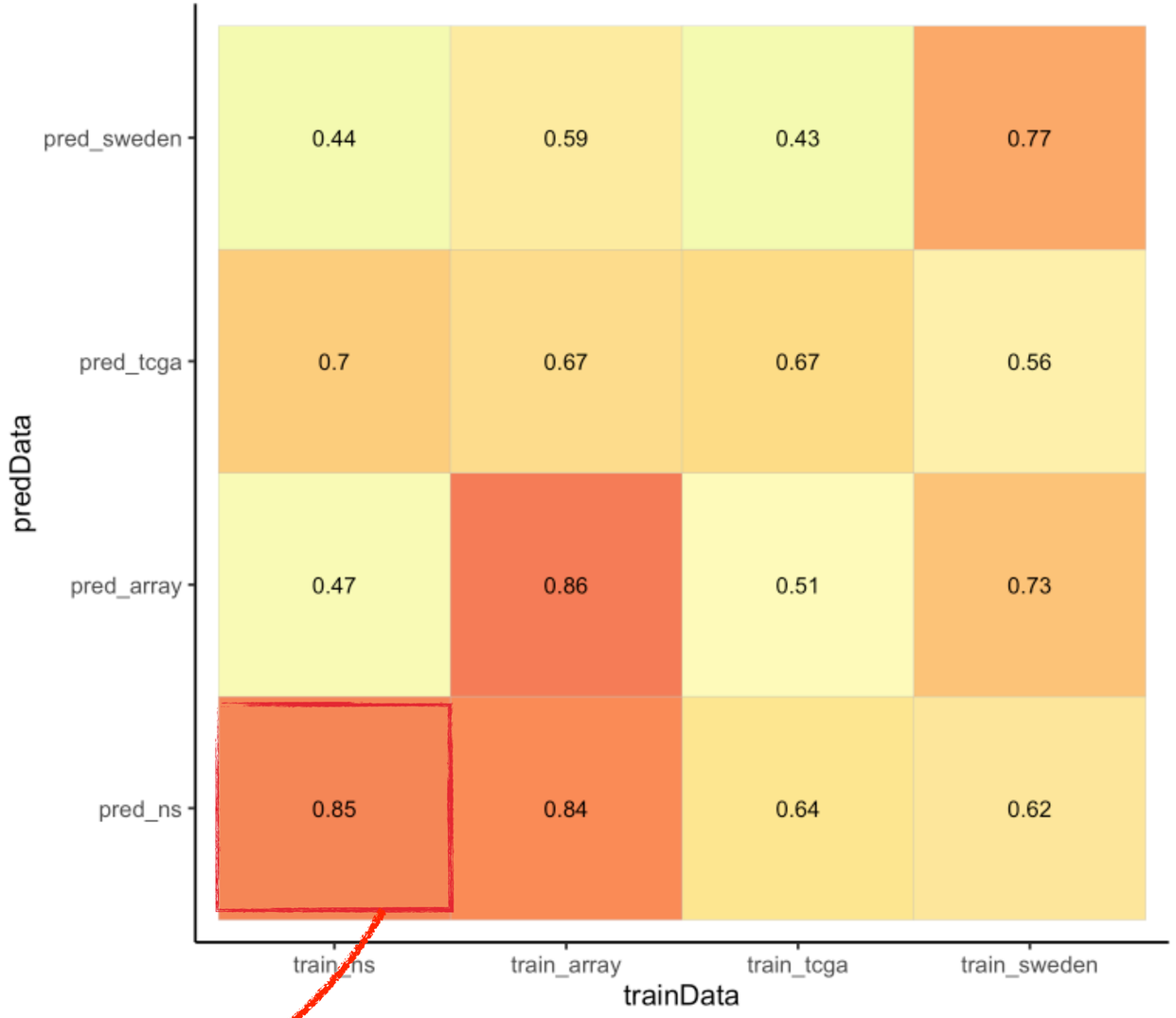
1. Log-ratios must have similar scaling across datasets
2. Variable selection must be stable across datasets
3. ~~Single-patient prediction~~

Lasso variable selection is NOT stable



The solution is not so trivial

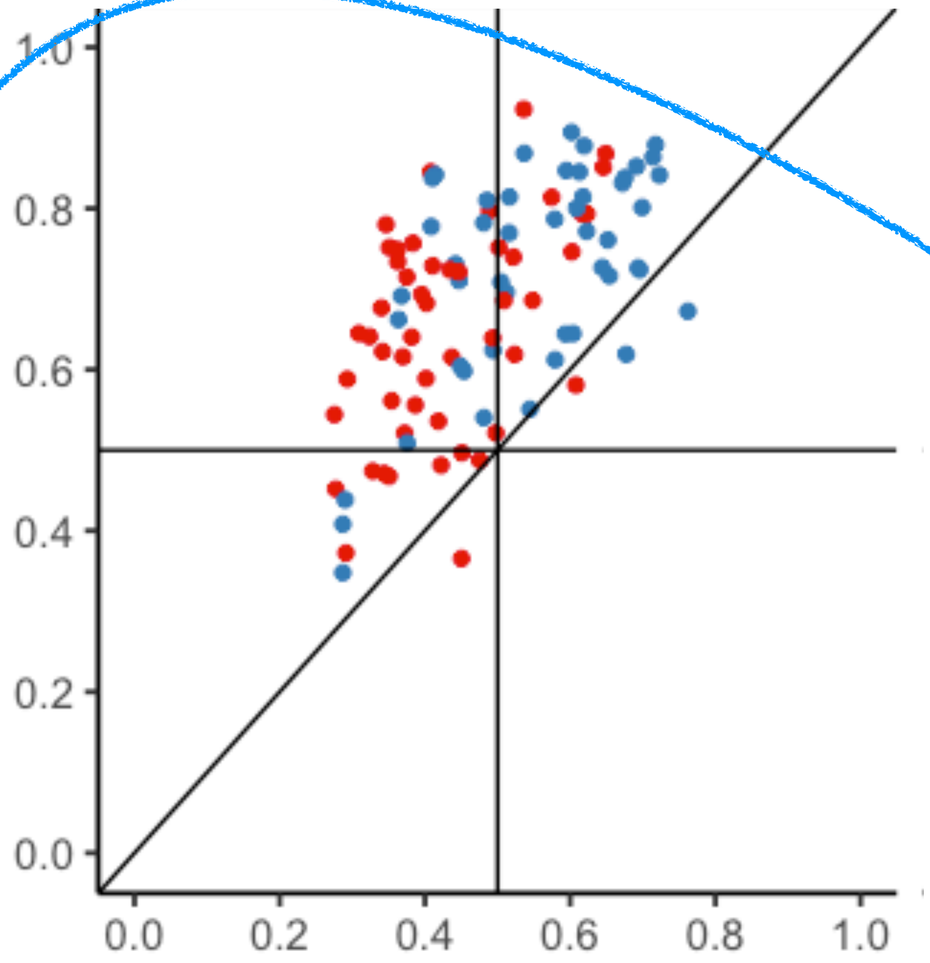




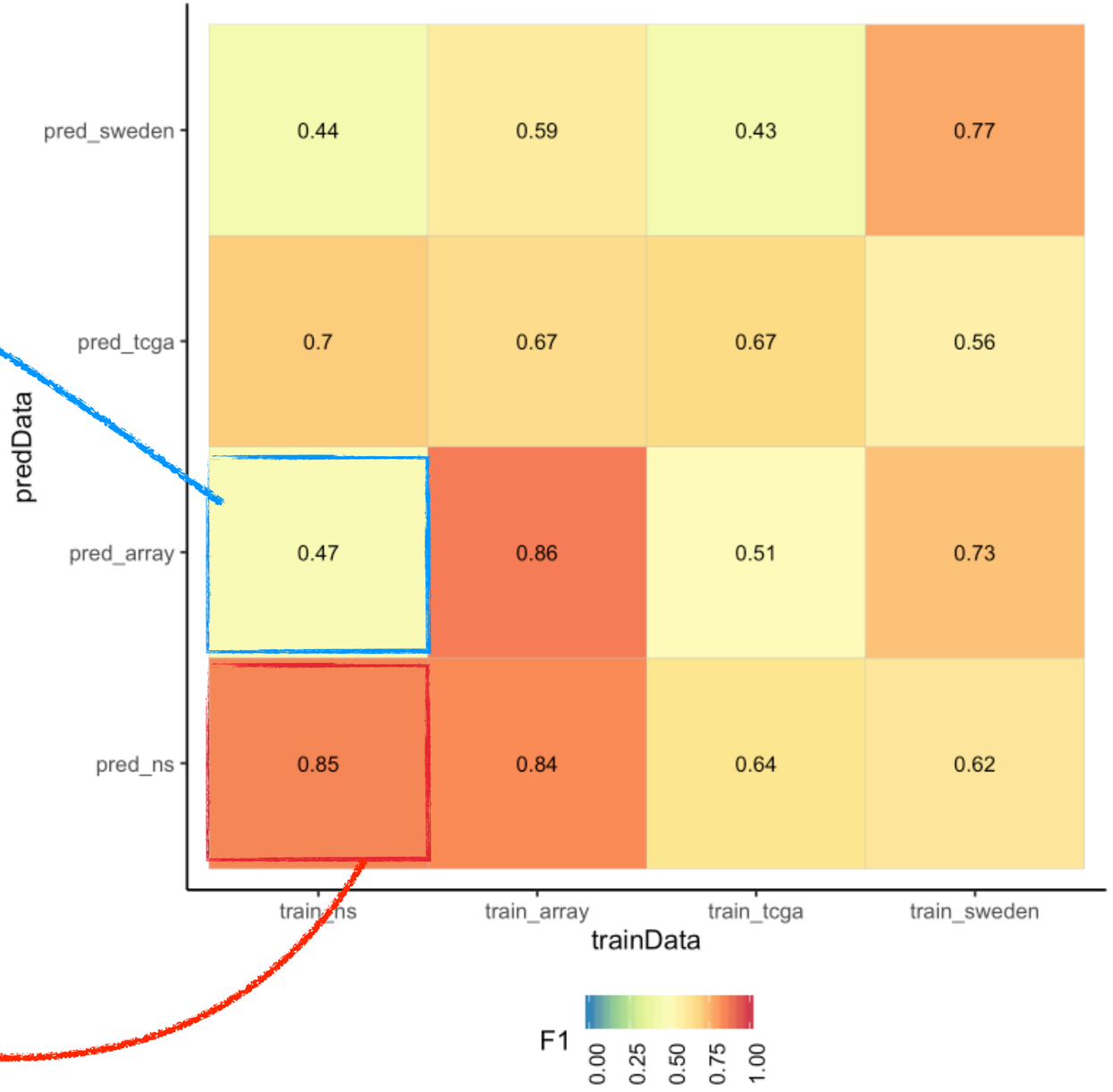
re-substitution probability (good!)

The solution is not so trivial

validation probability (poor!)



re-substitution probability (good!)



The solution is not so trivial

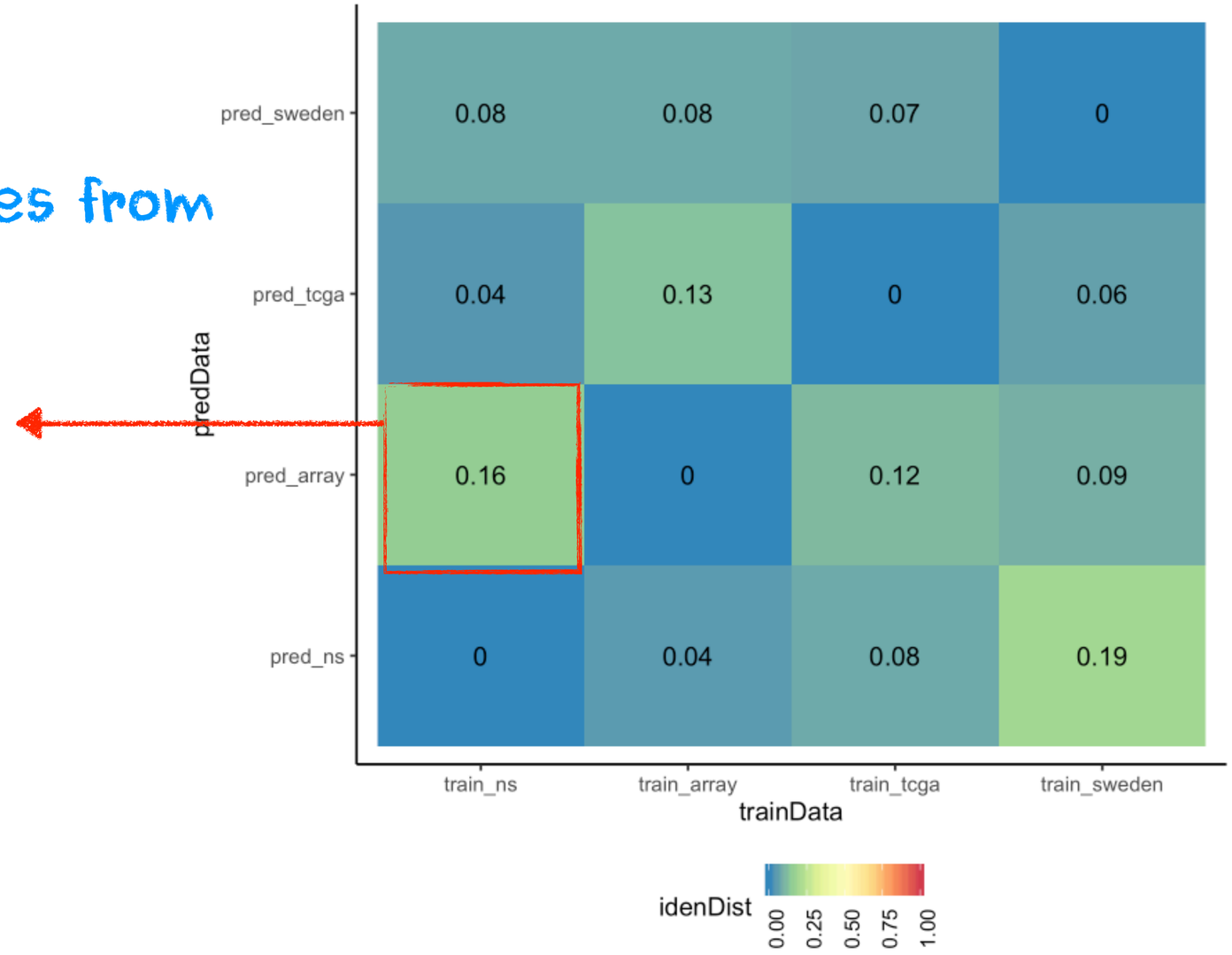
Estimated prognosis probabilities from

training data

vs

validation data

differ by 0.16 on average



Second component of CPOP: stable feature selection

我曾经毁了我的一切
只想永远地离开
我曾经堕入无边黑暗
想挣扎无法自拔
我曾经像你像他像那野草野花
绝望着也渴望着
也哭也笑也平凡着

Motivation for CPOP: one patient cohort, two gene expression data

$$X_1\hat{\beta}_1 \approx X_2\hat{\beta}_2$$

loosely translate to

$$X_1 \approx X_2$$

column-wise

(feature distribution stability)

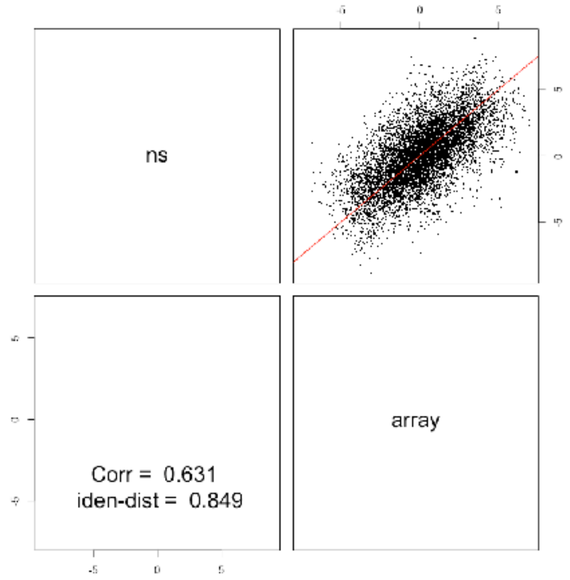
$$\hat{\beta}_1 \approx \hat{\beta}_2$$

element-wise

(mode estimation stability)

CPOP weighted variable selection

1. **Weighted Elastic Net:** higher weights on similar features

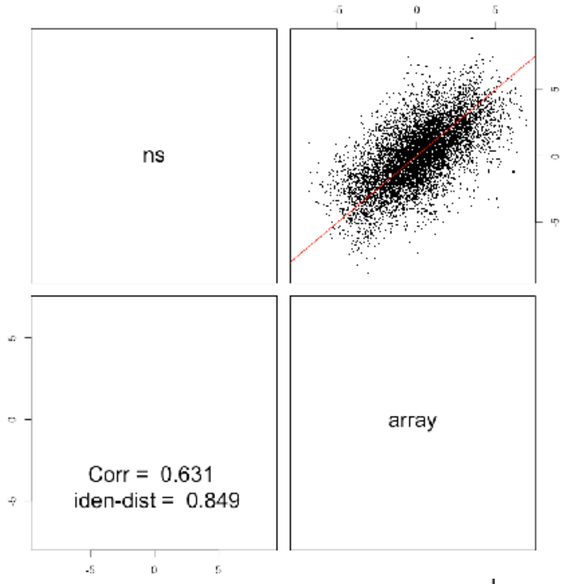


$$X_1 \approx X_2$$

2. Another Elastic Net only retaining stable coefficients

CPOP weighted variable selection

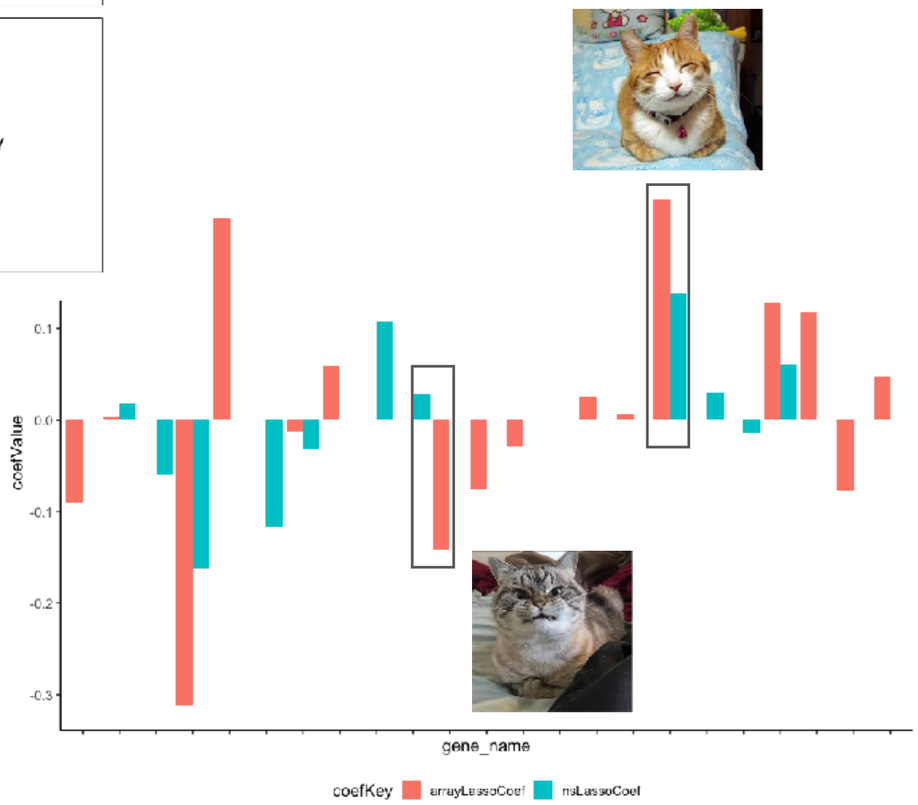
1. **Weighted Elastic Net:** higher weights on similar features



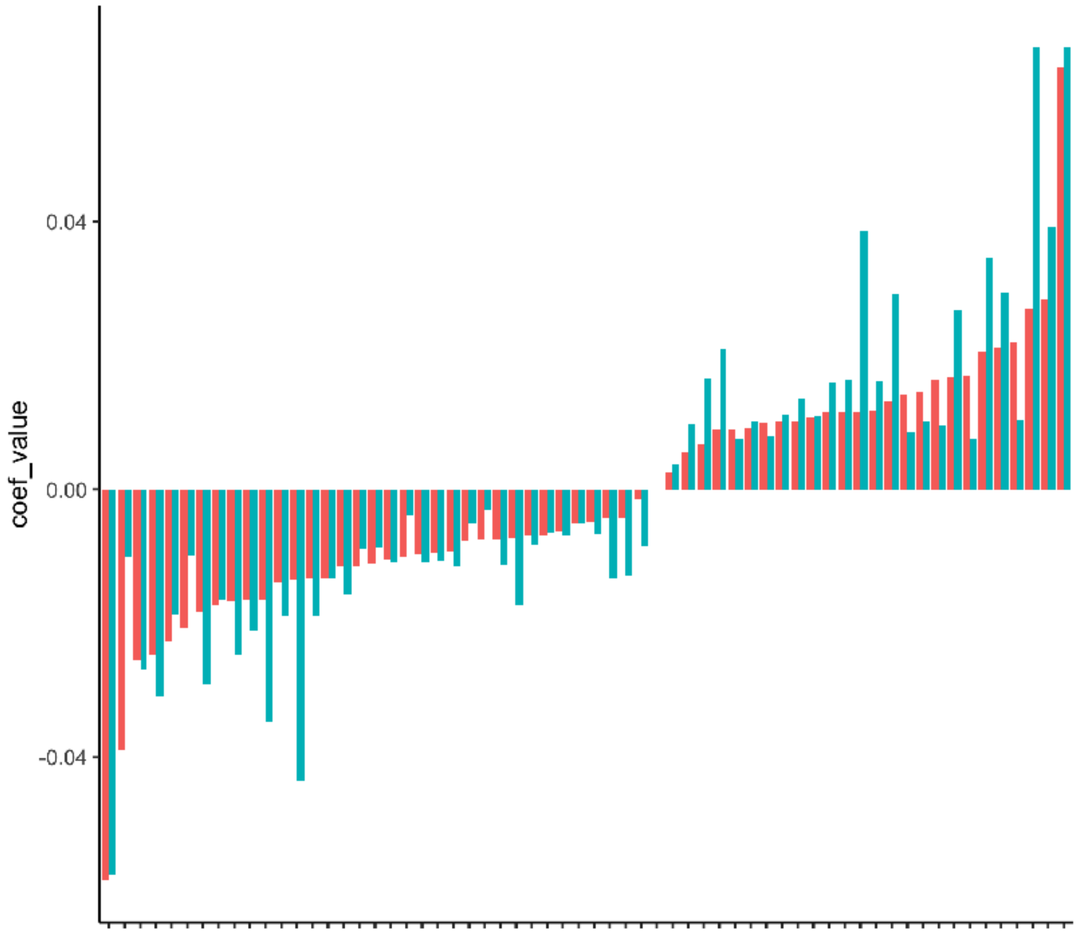
$$X_1 \approx X_2$$

2. Another Elastic Net only retaining stable coefficients

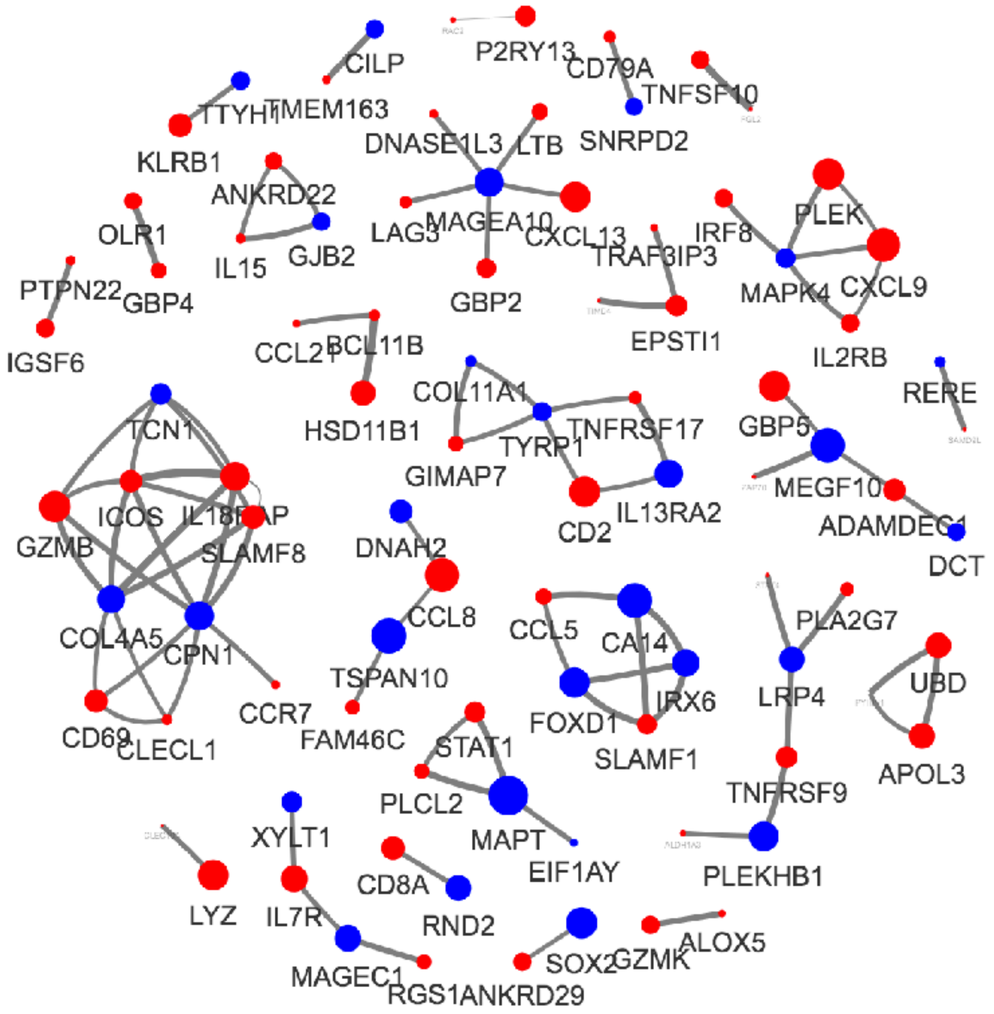
$$\hat{\beta}_1 \approx \hat{\beta}_2$$



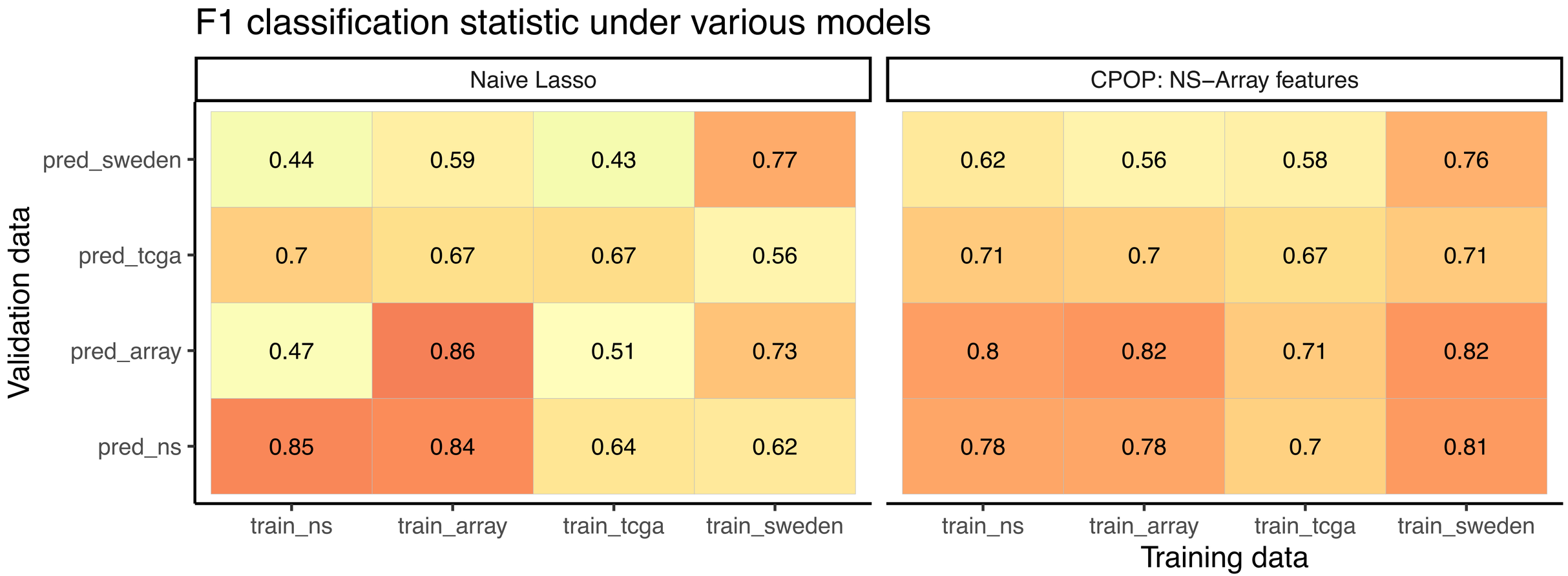
CPOP stable features



Log-ratio features

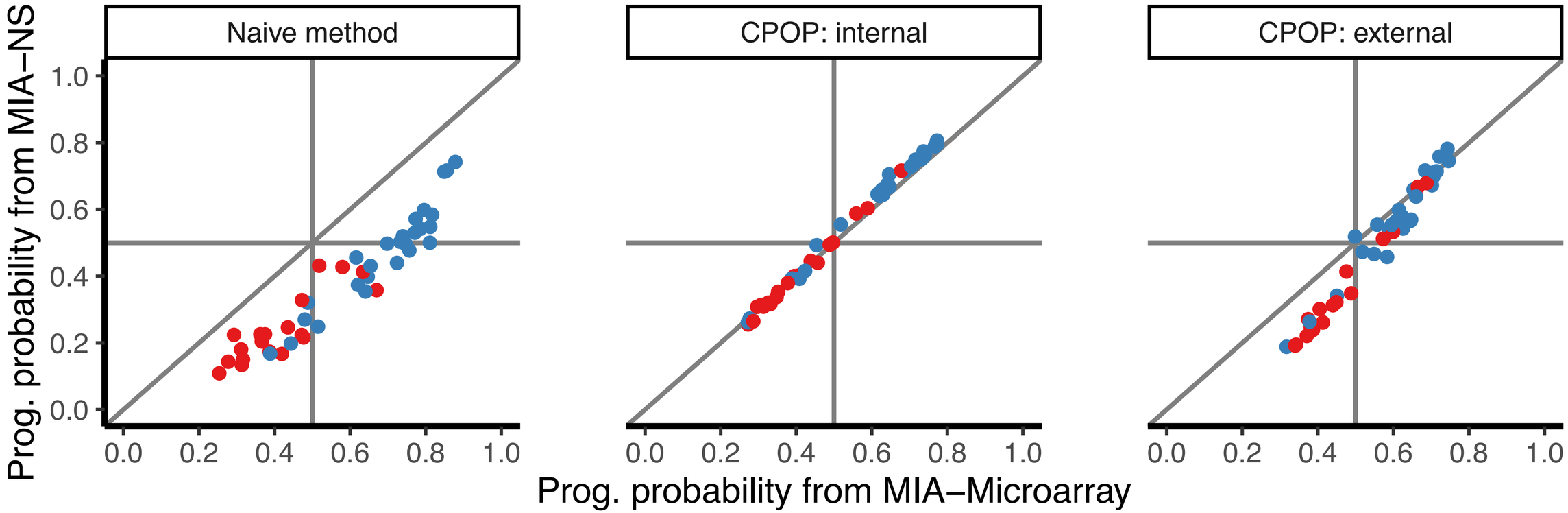


CPOP results 1: four melanoma data



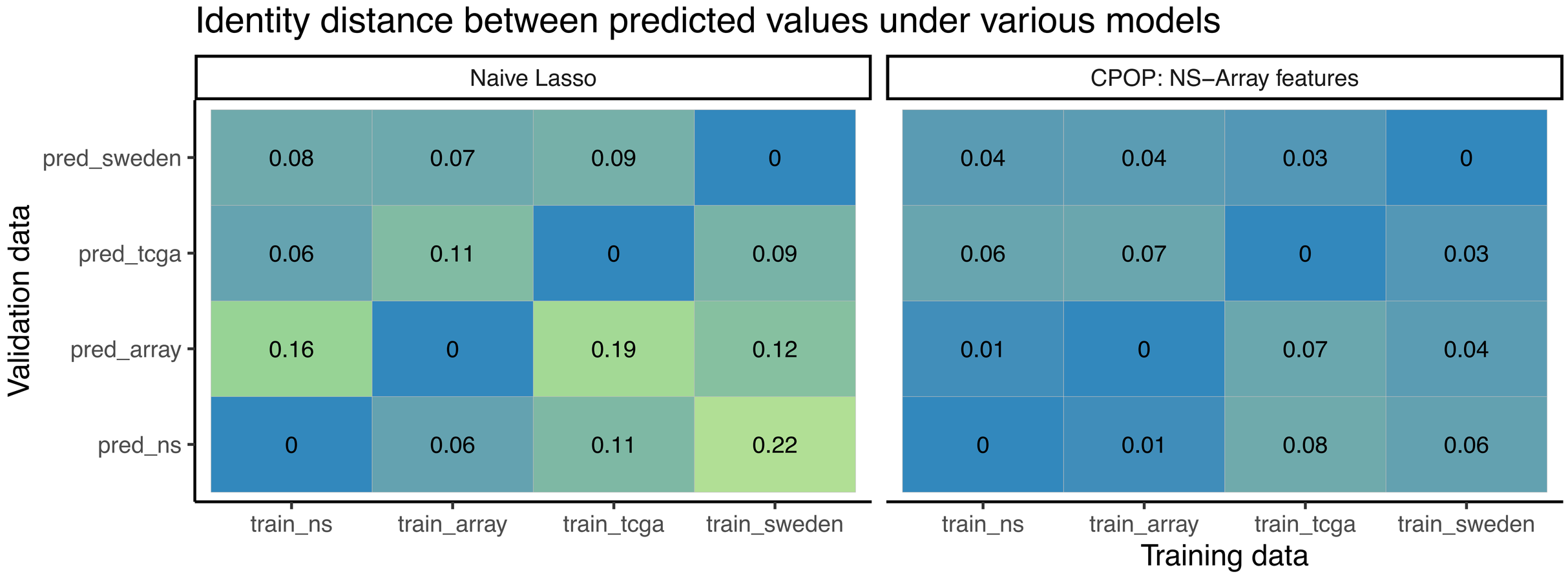
Small deviation in accuracy across datasets

CPOP results 1: four melanoma data



Small deviation in predicted values across datasets

CPOP results 1: four melanoma data

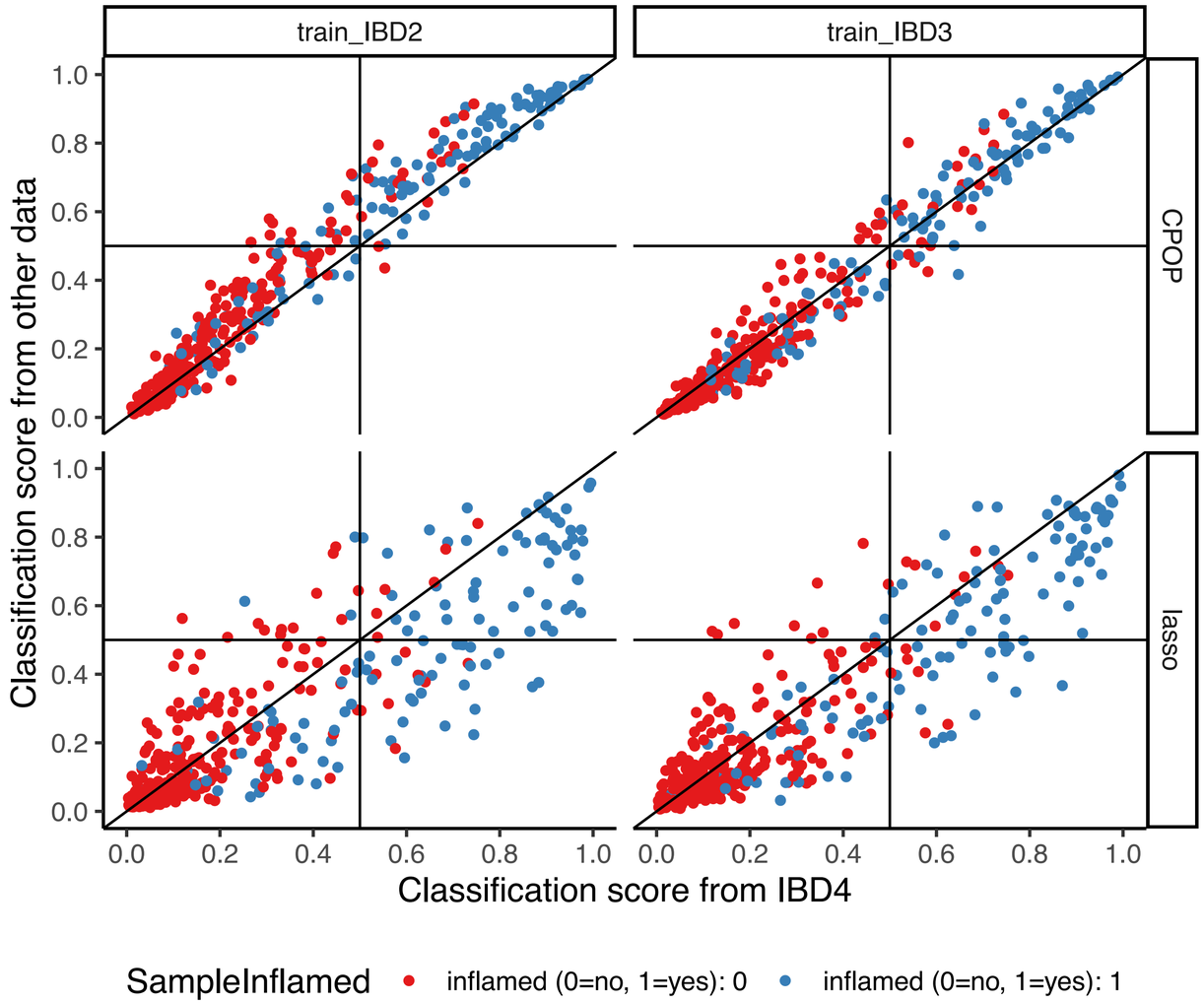


Small deviation in predicted values across datasets

CPOP results 2: prospective prediction

- CPOP on IBD NanoString data demonstrated improvements on stability

- We are planning to exploring other data of higher relevance to precision medicine (e.g. drug sensitivity)



I think that is all.

Is log-ratio really a new innovation?

Here is a list of papers that uses genes as predictors

Here is a list of papers that uses a single ratio for prediction

Our contribution is the advocation using a whole collection of ratios for prediction

This has extra implications in terms of the statistics, but we are happy to tackle these.