Estimator selection and adaptive learning

Christophe Giraud

Université Paris-Sud

M2 DS

Plan for today

- Estimator selection: classics
- Overfit in practice: an example
- Adaptive data analysis: issue
- Paper: reusable holdout with privacy

What shall I do with these data ?

Classical steps

1. Elucidate the question(s) you want to answer to, and check your data
   This requires some
   - deep discussions with experts (data collector, biologists, physicians, etc),
   - low level analyses (PCA, scatterplots, etc) to detect key features, outliers, etc
   - and ... experience !
2. Choose and apply an estimation procedure
3. Check your results (possible bias in residues, stability, etc)

Estimator selection

Example

Regression with unknown variance:

- $Y_i = f_i^* + \varepsilon_i$ with ε_i i.i.d. with variance σ^2
- $f^* = (f_1^*, \ldots, f_n^*)^T$ and σ^2 are unknown
- we want to estimate f^*

Ex 1 : sparse linear regression

- $f^* = \mathbf{X}\beta^*$ with β^* "sparse" in some sense and $\mathbf{X} \in \mathbb{R}^{n \times p}$ with possibly $p > n$

Ex 2 : non-parametric regression

- $f_i^* = F^*(x_i)$ with $F^* : \mathcal{X} \to \mathbb{R}$

A plethora of estimators

Sparse linear regression

- **Coordinate sparsity:** Lasso, Dantzig, Elastic-Net, Exponential-Weighting, Projection on subspaces $\{V_\lambda : \lambda \in \Lambda\}$ given by PCA, Random Forest, PLS, etc.
- **Structured sparsity:** Group-lasso, Fused-Lasso, Bayesian estimators, etc

Non-parametric regression

- Spline smoothing, Nadaraya kernel smoothing, kernel ridge estimators, nearest neighbors, L^2-basis projection, Sparse Additive Models, Neural Networks, etc

Important practical issues

Which estimator shall I use?

- **Sparse regression :** Lasso? Group-Lasso? Random-Forest? Exponential-Weighting? Forward–Backward?
- **Non-parametric regression :** Kernel regression? (which kernel?) Spline smoothing?

With which tuning parameter?

- which penalty level λ for the lasso?
- which bandwith h for kernel regression?
- etc

Difficulties

- No procedure is universally better than the others
- A sensible choice of the tuning parameters depends on
  - some unknown characteristics of f (sparsity, smoothness, etc)
  - the unknown variance σ^2.

Even if you are a pure Lasso-enthusiast, you miss some key informations in order to apply properly the lasso procedure !

The objective

Formalization

We have a collection of estimation schemes (lasso, group-lasso, etc) and for each scheme we have a grid of different values for the tuning parameters.

At the end, putting all the estimators together we have a collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ of estimators.

Ideal objective

Select the "best" estimator among the collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$.

(alternative objective: aggregate at best the estimators)

Cross-Validation

The most popular technique for choosing tuning parameters

Principle

split the data into a training set and a validation set: the estimators are built on the *training* set and the *validation* set is used for estimating their prediction risk.

Most popular cross-validation scheme

- Hold-out : a single split of the data for *training* and *validation*.
- V-fold CV : the data is split into V subsamples. Each subsample is successively removed for *validation*, the remaining data being used for *training*.
- Leave-one-out : corresponds to n-fold CV.
- Leave-q-out : every possible subset of cardinality q of the data is removed for *validation*, the remaining data being used for *training*.

Classical choice of V : between 5 and 10 (remains tractable).

V-fold CV

train	train	train	train	test
train	train	train	test	train
train	train	test	train	train
train	test	train	train	train
test	train	train	train	train

Recursive data splitting for 5-fold Cross-Validation

Pros and Cons

- **Universality:** Cross-Validation can be implemented in most statistical frameworks and for most estimation procedures.
- Usually (but not always!) give good results in practice.
- But **limited theoretical garanties** in large dimensional settings.

Scaled-Lasso

Automatic tuning of the Lasso

Scale invariance

The estimator $\widehat{\beta}(Y, \mathbf{X})$ of β^* is scale-invariant if $\widehat{\beta}(sY, \mathbf{X}) = s\widehat{\beta}(Y, \mathbf{X})$ for any $s > 0$.

Example: the estimator

$$\widehat{\beta}(Y, \mathbf{X}) \in \operatorname*{argmin}_{\beta} \|Y - \mathbf{X}\beta\|^2 + \lambda\Omega(\beta),$$

where Ω is homogeneous with degree 1 is not scale-invariant unless λ is proportional to σ.

In particular the Lasso estimator is not scale-invariant when λ is not proportional to σ.

Rescaling

Idea:

- estimate σ with $\widehat{\sigma} = \|Y - \mathbf{X}\beta\|/\sqrt{n}$.
- set $\lambda = \mu\widehat{\sigma}$
- divide the criterion by $\widehat{\sigma}$ to get a convex problem

Scale-invariant criterion

$$\widehat{\beta}(Y, \mathbf{X}) \in \operatorname*{argmin}_{\beta} \sqrt{n}\|Y - \mathbf{X}\beta\| + \mu\Omega(\beta).$$

Example: scaled-Lasso

$$\widehat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \sqrt{n}\|Y - \mathbf{X}\beta\| + \mu|\beta|_1 \right\}.$$

Pros and Cons

- Universal choice $\mu = 5\sqrt{\log(p)}$
- strong theoretical guaranties
- computationally feasible
- but poor performances in practice (disappointing!)

Numerical experiments (1/2)

Tuning the Lasso

- 165 examples extracted from the literature
- each example e is evaluated on the basis of 400 runs

Comparison to the oracle $\widehat{\beta}_{\lambda^*}$

procedure	quantiles			
	0%	50%	75%	90%
Lasso 10-fold CV	1.03	1.11	1.15	1.19
Lasso LinSelect	0.97	1.03	1.06	1.19
Scaled Lasso	1.32	2.61	3.37	11.2

For each procedure ℓ, quantiles of $\mathcal{R}\left[\widehat{\beta}_{\hat{\lambda}_\ell}; \beta_0\right] / \mathcal{R}\left[\widehat{\beta}_{\lambda^*}; \beta_0\right]$, for $e = 1, \ldots, 165$.

Numerical experiments (2/2)

Computation time

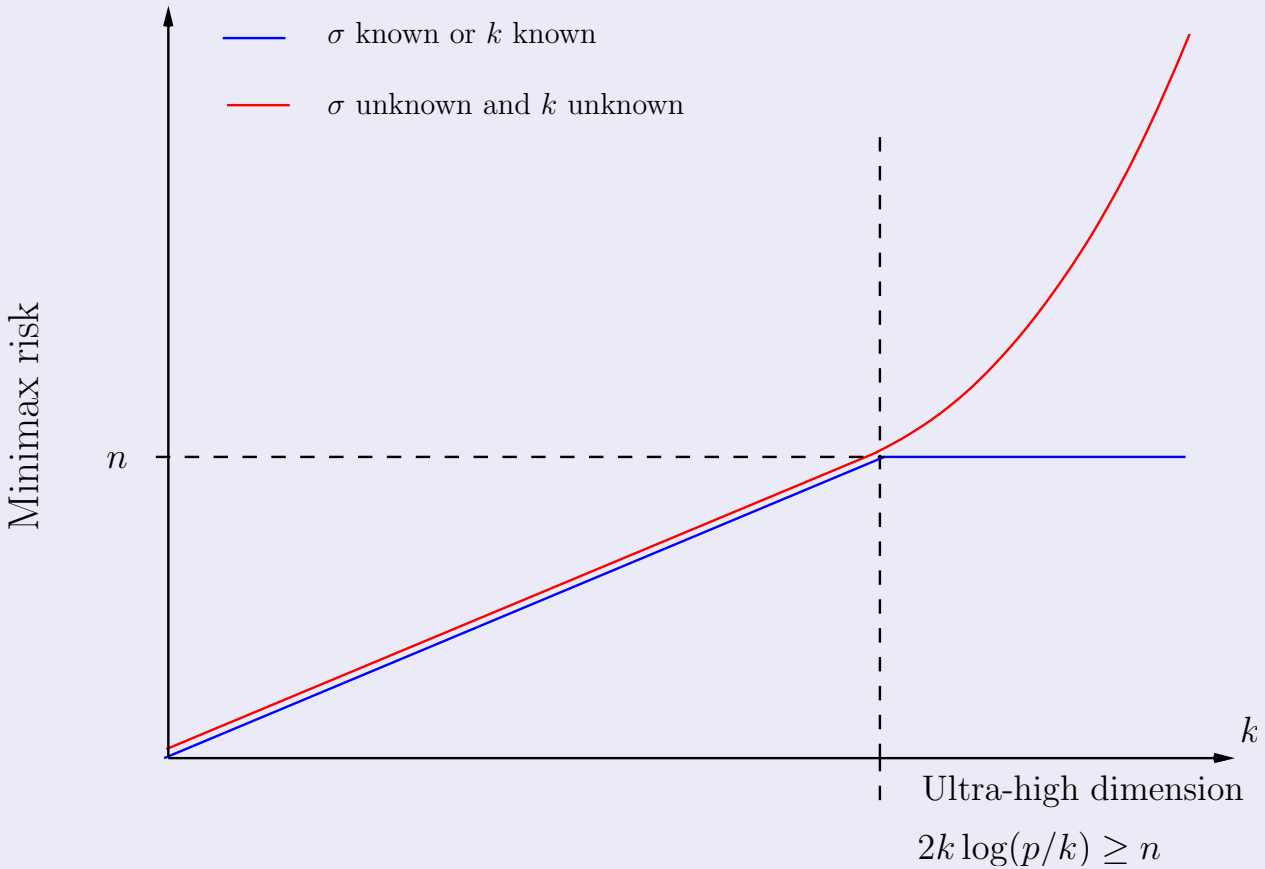
n	p	10-fold CV	LinSelect	Scaled-Root
100	100	4 s	0.21 s	0.18 s
100	500	4.8 s	0.43 s	0.4 s
500	500	300 s	11 s	6.3 s

Packages:

- `enet` for 10-fold CV and LinSelect
- `lars` for Scaled Lasso (procedure of Sun & Zhang)

Impact of the unknown variance?

Case of coordinate-sparse linear regression



Minimax prediction risk over k-sparse signal as a function of k

The danger of overfitting

- selection methods are widely used by data scientists
- V-fold CV is the most popular selection scheme
- yet, despite this step, data scientist can face severe overfitting...

A Kaggle Post-morterm

The curse of overfitting

A Kaggle postmorterm

A blog-post by Greg Park:
"The dangers of overfitting: a Kaggle postmortem"
`http://gregorypark.org/blog/Kaggle-Psychopathy-Postmortem/`

Psychopathy Kaggle contest

- **Goal:** predict the psychopathy levels of Twitter users
- **Competition:**
  - Competitors can submit two sets of predictions each day;
  - Each submission is instantly scored from 0 (worst) to 1 (best) and competitors are ranked on a public leaderboard;
  - When the competition closes, the private leaderboard is revealed (competitors scores are computed on a mostly independent data set).

Listen to Greg Park's experience...

The competition

"I made 42 submissions [...]. By the end of the contest, I had slowly worked my way up to 2nd place on the public leaderboard, shown below."

#	Δ1w	Team Name	MCAP	Entries	Last Submission UTC (Best Submission - Last)
1	-	willkurt *	0.85295	60	Tue, 26 Jun 2012 03:41:54 (-21.9d)
2	-	Greg Park	0.85195	42	Fri, 29 Jun 2012 01:08:38 (-14.1d)
3	-	Luca Massaron	0.85163	50	Thu, 28 Jun 2012 14:51:04 (-25.6d)
4	-	Stein	0.85093	28	Thu, 28 Jun 2012 08:15:16 (-9.8d)
5	-	NoisyServer	0.85082	36	Fri, 29 Jun 2012 21:21:52 (-21.9d)

"I felt confident that I would maintain a decent spot on the private leaderboard"

Competition outcome

"Soon after the competition closed, the private leaderboard was revealed. Here's what I saw at the top :"

#	Δ1w	Team Name	MCAP	Entries	Last Submission UTC (Best Submission - Last)
1	-	y_tag *	0.86997	12	Tue, 26 Jun 2012 12:46:19
2	↑1	Bruce Cragin	0.86745	10	Fri, 29 Jun 2012 22:28:17 (-47.6h)
3	new	Indy Actuaries	0.86700	6	Fri, 29 Jun 2012 03:40:38 (-3.4d)
4	↓2	jontix	0.86697	7	Thu, 31 May 2012 11:05:41 (-9.9d)
5	-	JKARP	0.86683	35	Fri, 29 Jun 2012 22:02:34 (-3.3d)

Competition outcome

"Where'd I go? I scrolled down the leaderboard... further... and further... and finally found my name:"

	↓8	Random Forest Benchmark	0.86141		
45	↓8	dickoa	0.86141	1	Tue, 22 May 2012 12:07:36
45	↓8	Rohit	0.86141	2	Fri, 25 May 2012 21:00:14
45	↓8	squawkboxed	0.86141	1	Fri, 08 Jun 2012 14:57:28
45	new	BLetson	0.86141	3	Fri, 29 Jun 2012 14:49:38
50	↓9	testing	0.86135	4	Sat, 16 Jun 2012 05:18:44 (-26.1h)
51	↓9	schappi	0.86130	7	Sat, 16 Jun 2012 12:53:13
52	↓8	Greg Park	0.86116	42	Fri, 29 Jun 2012 01:08:38 (-14.1d)
53	↓8	Glen	0.86111	35	Tue, 05 Jun 2012 23:44:06 (-3.3d)

What happened?

"Somehow I managed to fall from 2nd all the way down to 52nd!

I wasn't the only one who took a big fall: the top five users on the public leaderboard ended up in 64th, 52nd, 58th, 16th, and 57th on the private leaderboard, respectively.

I even placed below the random forest benchmark, a solution publicly available from the start of the competition."

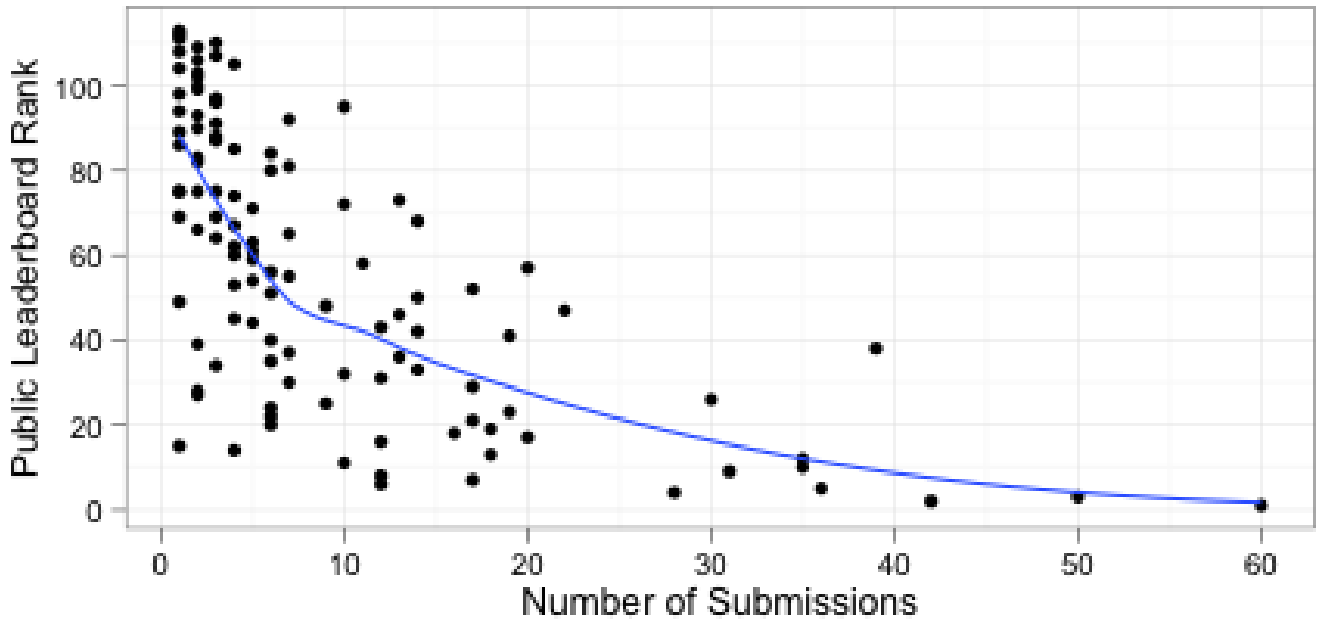
What happened?

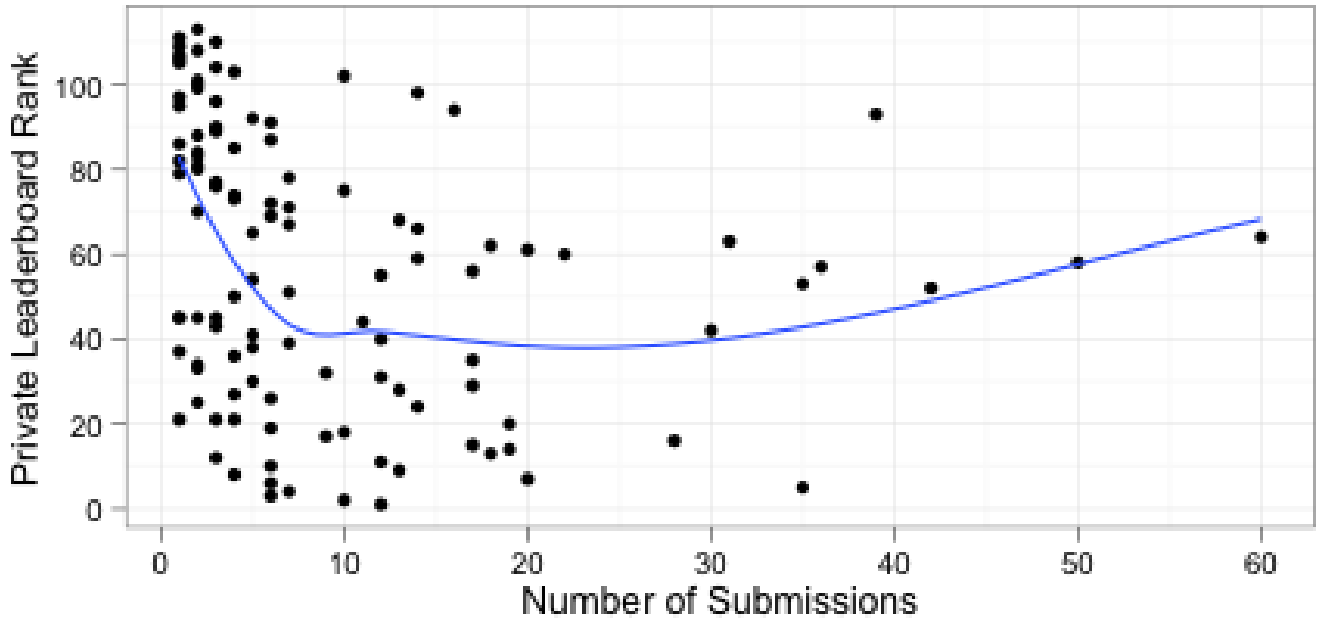
An overfitting problem

Rank versus number of submissions

- the top five in the public leaderboard have from 28 to 60 submissions
- the top four in the private leaderboard have less than 12 submissions

The harder you work, the worse is your job?


Public Leaderboard Rank
100
80
60
40
20
0
0
10
20
30
40
50
60
Number of Submissions


Private Leaderboard Rank
100
80
60
40
20
0
0
10
20
30
40
50
60
Number of Submissions

Observations

Observations

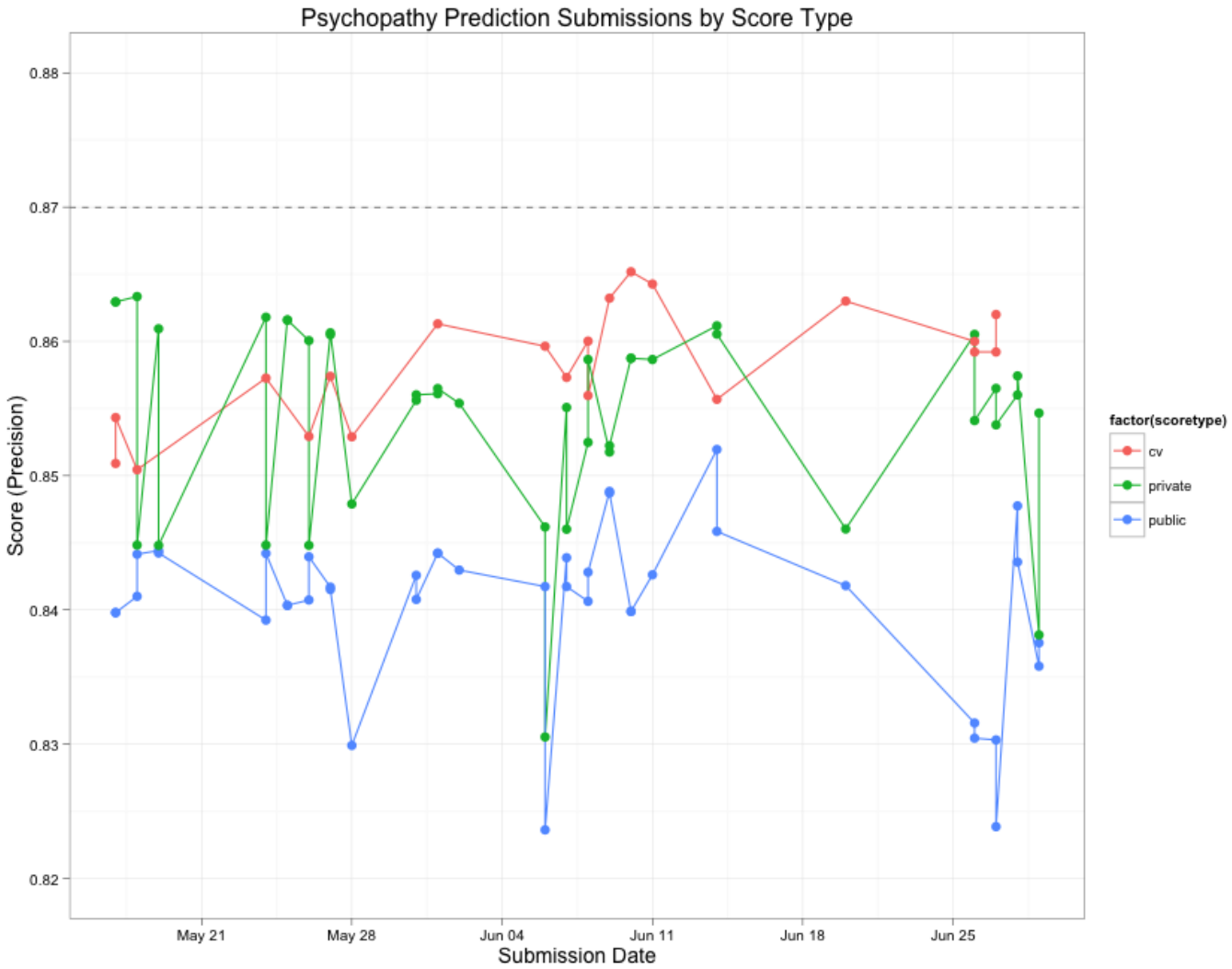
- On the public leaderboard, more submissions are consistently related to a better standing.

  "It could be that the public leaderboard actually reflects the amount of brute force from a competitor rather than predictive accuracy."

- The private leaderboard has a U-shaped curve. After about 25 submissions or so, private leaderboard standings get worse with the number of submissions.

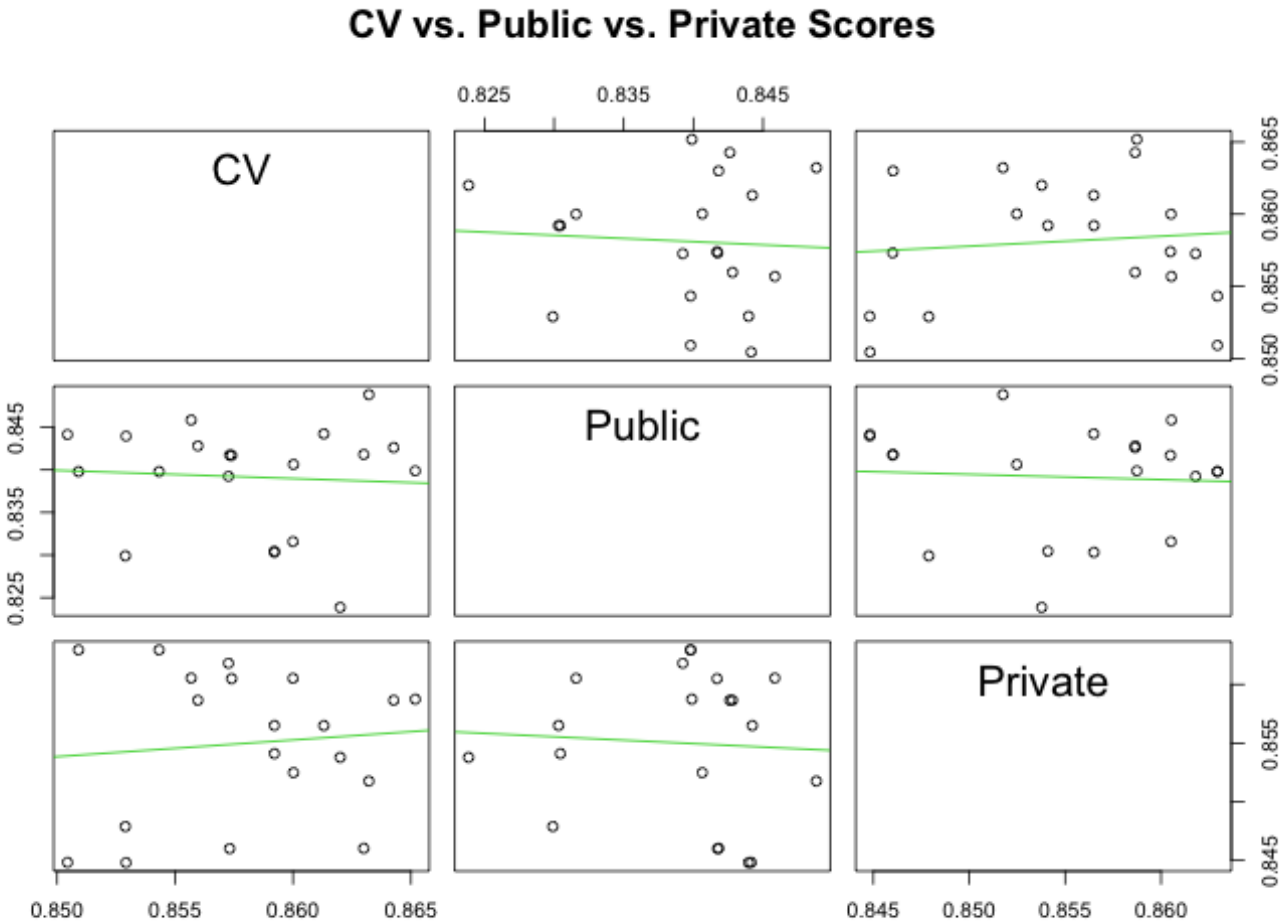
"I knew not to trust the public leaderboard, but when I started to edge towards to the top, I began to trust it again!"

Temporal evolution



Orange: CV criterion. **Blue:** public score. **Green:** private score.

Scatterplots



Greg Park's lessons

Lessons

- "It is easier to overfit the public leaderboard than previously thought. Be more selective with submissions."
- "On a related note, perform cross-validation the right way: include all training (feature selection, preprocessing, etc.) in each fold."
- "Try to ignore the public leaderboard, even when it is telling you nice things about yourself."

Conclusions

Take home message

- Take care of overfitting even if you use CV or FDR control!
- Be careful: you usually work sequentially, taking decisions based on previous outcomes.

Can we design statistical strategies to overcome this curse?

Adaptive Data Analysis

Introduction of *Preserving Statist. Validity...* Dwork et. al.

Throughout the scientific community there is a growing recognition that claims of statistical significance in published research are frequently invalid. The past few decades have seen a great deal of effort to understand and propose mitigations for this problem. These efforts range from the use of sophisticated validation techniques and deep statistical methods for controlling the false discovery rate in multiple hypothesis testing to proposals for preregistration (that is, defining the entire data-collection and data-analysis protocol ahead of time). The statistical inference theory surrounding this body of work assumes a fixed procedure to be performed, selected before the data are gathered. In contrast, the practice of data analysis in scientific research is by its nature an adaptive process, in which new hypotheses are generated and new analyses are performed on the basis of data exploration and observed outcomes on the same data. This disconnect is only exacerbated in an era of increased amounts of open access data, in which multiple, mutually dependent, studies are based on the same datasets.

From Theory to Practice

Statistical theory

Classical statistical theory assumes that a fixed procedure is performed, selected before the data are gathered.

Practice

In contrast, the practice of data analysis in scientific research is, by nature, an adaptive process in which new analyses are chosen on the basis of data exploration and previous analyses of the same data.

Adaptive data analysis

Adaptive learning

We apply sequentially m algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_m$ on the data with

$$\mathcal{A}_j = \mathcal{A}_j(X, \mathcal{A}_1(X), \ldots, \mathcal{A}_{j-1}(X)).$$

A simple but expansive scheme

Multiple hold-out

For each procedure, use a new hold-out sample!

$\longrightarrow$ requires a huge amount of data!!

Reusable hold-out?

Could we reuse the holdout?

Idea:

- access the hold out sample via differentially private mechanism,
- hence you can learn about your query, but you learn little about the data.

References

Estimator selection: Introduction to high-dimensional statistics, chap. 5

Differential Privacy: Duchi, Jordan and Wainwright. Privacy Aware learning.
https://people.eecs.berkeley.edu/~wainwrig/Papers/DucJorWai14.pdf

Reusable Holdout:
Dwork et al. Preserving Statistical Validity in Adaptive Data Analysis
https://arxiv.org/pdf/1411.2664.pdf

Dwork et al. The reusable holdout: Preserving validity in adaptive data analysis. Science (2015)
https://pdfs.semanticscholar.org/25fe/96591144f4af3d8f8f79c95b37f415e5bb75.pdf