The Reusable Holdout: Preserving Statistical Validity in Adaptive Data Analysis

Moritz Hardt
IBM Research Almaden

Joint work with Cynthia Dwork, Vitaly Feldman, Toni Pitassi, Omer Reingold, Aaron Roth
False discovery — a growing concern

“Trouble at the Lab” – The Economist
Most published research findings are probably false. – John Ioannidis

P-hacking is trying multiple things until you get the desired result. – Uri Simonsohn

She is a p-hacker, she always monitors data while it is being collected. – Urban Dictionary

The p value was never meant to be used the way it's used today. – Steven Goodman
Preventing false discovery

Decade old subject in Statistics

Powerful results such as Benjamini-Hochberg work on controlling False Discovery Rate

Lots of tools:
Cross-validation, bootstrapping, holdout sets

Theory focuses on non-adaptive data analysis
Non-adaptive data analysis

- Specify exact experimental setup
  - e.g., hypotheses to test
- Collect data
- Run experiment
- Observe outcome

Data analyst

Can’t reuse data after observing outcome.
Adaptive data analysis

- Specify exact experimental setup
  - e.g., hypotheses to test
- Collect data
- Run experiment
- Observe outcome
- Revise experiment
Adaptivity

Data dredging, data snooping, fishing, p-hacking, post-hoc analysis, garden of the forking paths

Some caution strongly against it:
“Pre-registration” — specify entire experimental setup ahead of time

Humphreys, Sanchez, Windt (2013), Monogan (2013)
Adaptivity

“Garden of Forking Paths”

The most valuable statistical analyses often arise only after an iterative process involving the data — Gelman, Loken (2013)
From art to science

Can we guarantee statistical validity in adaptive data analysis?

Our results: To a surprising extent, yes.

Our hope: To inform discourse on false discovery.
Main result:
The outcome of any differentially private analysis generalizes*.

* If we sample fresh data, we will observe roughly the same outcome.

Moreover, there are powerful differentially private algorithms for adaptive data analysis.
Intuition

Differential privacy is a *stability* guarantee:

- Changing one data point doesn’t affect the outcome much

Stability implies generalization

- “Overfitting is not stable”
Does this mean I have to learn how to use differential privacy?

Resoundingly, no!

Thanks to our reusable holdout method
Standard holdout method

Data

Training data

Holdout

Unrestricted access

Good for one validation

Non-reusable: Can’t use information from holdout in training stage adaptively
One corollary: a reusable holdout

Data

training data

unrestricted access

reusable holdout

can be used many times adaptively

essentially as good as using fresh data each time!
More formally

Domain $X$. Unknown distribution $D$ over $X$

Data set $S$ of size $n$ sampled i.i.d. from $D$

What the holdout will do:

Given a function $q : X \rightarrow [0, 1]$, estimate the expectation $\mathbb{E}_D[q]$ from sample $S$

Definition: An estimate $a$ is valid if $|a - \mathbb{E}_D[q]| < 0.01$

Enough for many statistical purposes, e.g., estimating quality of a model on distribution $D$
Example: Model Validation

We trained predictive model $f: Z \rightarrow Y$
and want to know its accuracy

Put $X = Z \times Y$.
Joint distribution $D$ over data x labels

Estimate accuracy of classifier
using the function $q(z,y) = 1\{ f(z) = y \}$

$\mathbb{E}_S[q] = $ accuracy with respect to sample $S$
$\mathbb{E}_D[q] = $ true accuracy with respect to unknown $D$
A reusable holdout: *Thresholdhound*

**Theorem.** Thresholdout gives valid estimates for any sequence of adaptively chosen functions until $n^2$ overfitting* functions occurred.

* Function $q$ overfits if $|\mathbb{E}_S[q] - \mathbb{E}_D[q]| > 0.01$.

Example: Model is good on $S$, bad on $D$. 
Thresholdout

Input:
Data $S$, holdout $H$, threshold $T > 0$, tolerance $\sigma > 0$

Given function $q$:

Sample $\eta, \eta'$ from $N(0, \sigma^2)$

If $|\text{avg}_H[q] - \text{avg}_S[q]| > T + \eta$:
  output $\text{avg}_H[q] + \eta'$

Otherwise:
  output $\text{avg}_S[q]$
An illustrative experiment

- Data set with $2n = 20,000$ rows and $d = 10,000$ variables. Class labels in $\{-1,1\}$
- Analyst performs **stepwise variable selection**:
  1. Split data into training/holdout of size $n$
  2. Select “best” $k$ variables on training data
  3. Only use variables also good on holdout
  4. Build linear predictor out of $k$ variables
  5. Find best $k = 10,20,30,\ldots$
No correlation between data and labels

data are random gaussians
labels are drawn *independently* at random from \{-1,1\}

Thresholdout correctly detects overfitting!
High correlation

20 attributes are highly correlated with target remaining attributes are uncorrelated

Thresholdout correctly detects right model size!
Conclusion

Powerful new approach for achieving statistical validity in adaptive data analysis building on differential privacy!

• Reusable holdout:
  • Broadly applicable
  • Complete freedom on training data
  • Guaranteed accuracy on the holdout
  • No need to understand Differential Privacy
  • Computationally fast and easy to apply
Go read this paper for a proof:

On the Generalization Properties of Differential Privacy

Kobbi Nissim, Uri Stemmer

A new line of work, started with Dwork et al., studies the task of answering statistical queries using a sample and relates it to the notion of differential privacy. By the Hoeffding bound, a sample of size $O(\log k/\epsilon^2)$ suffices to answer $k$ non-adaptive queries when the answers are computed by evaluating the statistical queries on the sample. This argument fails when the queries are chosen adaptively (they depend on the sample). Dwork et al. showed that if the answers are computed with $(\epsilon, \delta)$-differential privacy then $O(\epsilon)$ holds but for $\delta = O(1/k)$ the probability $1 - O(\delta^c)$. Using the Private Multiplicative Weights mechanism, they concluded that the sample size can still be at least $\log k$.

Very recently, Bassily et al. presented an improved bound and showed that (a variant of) the private multiplicative weights algorithm can be adapted to the case of adaptively chosen statistical queries using sample complexity that grows logarithmically in $k$. However, their results no longer hold when using a differentially private algorithm, and require modifying the private multiplicative weights algorithm in order to obtain their desired result. We greatly simplify the results of Dwork et al. and improve on the bound by showing that differential privacy guarantees a bound $\epsilon$ with probability $1 - O(\delta \log(1/\epsilon)/\epsilon)$. It would be tempting to guess that an $(\epsilon, \delta)$-differentially private computation should guarantee a bound $1 - O(\delta)$. However, we show that this is not the case, and that our bound is tight (up to logarithmic factors).
Thank you.