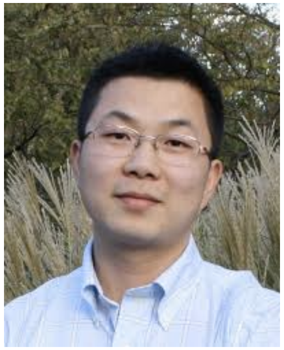
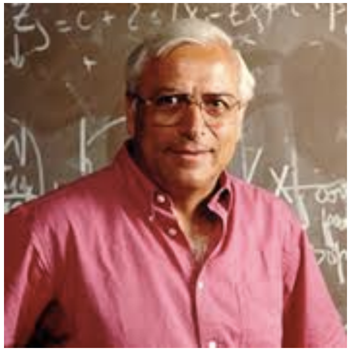
Optimal Gaussian Adaptive Data Analysis

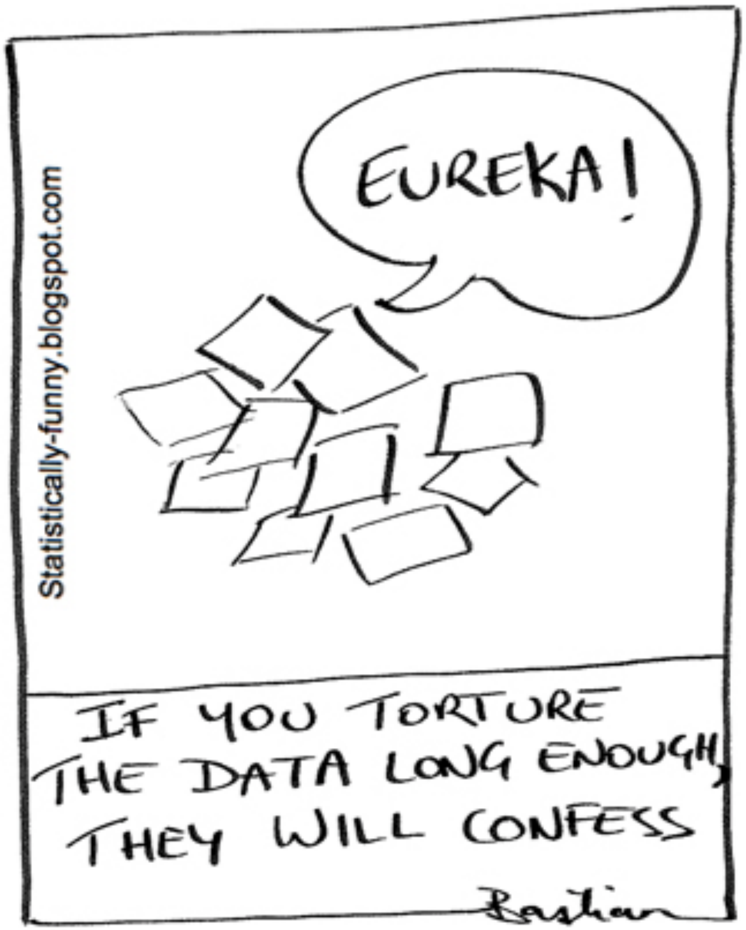
Yu-Xiang Wang

Joint work with Jing Lei and Steve Fienberg





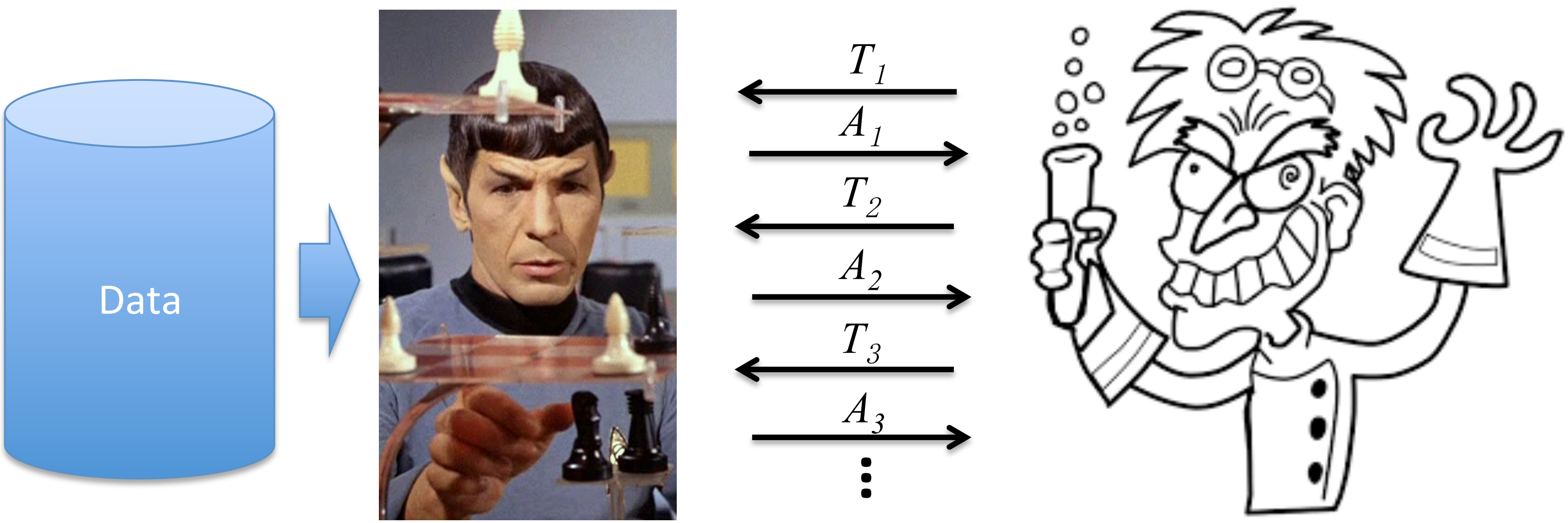
Data analysis is conditional/adaptive



- "All inferences are conditional inferences."
  - Jonathan Taylor (via Ryan)
- "Why most published research findings are false?"
  - John Ioannidis, 2005
- "A garden of forking paths"
  - Gelman and Loken, 2013

A model for adaptive data analysis

$\boxed{\phi_T}$ $\boxed{\sim N(\mu_T, \Sigma)}$ $\boxed{T_1, ..., T_k \in \mathcal{T}}$



Player

I have the data.

I choose how to answer the questions.

Adversary

I have the distribution.

I choose questions T.

Russo and Zou. "Controlling Bias in Adaptive Data Analysis Using Information Theory." AISTATS-2016.

Example: Choosing classifiers

- Adversary:
  - T_1: Give me a risk estimate to the optimal linear classifier using feature 1,5,7
  - T_2: If the answer is greater than 0.5: give me that of feature 2,4,6. Otherwise, give me the risk of a kernel classifier using only feature 1,5,7.
- Player:
  - ϕ_{T_i} empirical estimates of T_i on data. Jointly distributed due to data and T_i (and T_i depends on $T_{1:i-1}$, $A_{1:i-1}$)

Our contribution

- Formulate the minimax problem

- Establish information-theoretic limits
  - Minimax lower bound
  - Per-instance lower bound (for natural estimators)

Minimax setup

- Assuming: $\phi_{\mathcal{T}} \sim N(\mu_{\mathcal{T}}, \Sigma)$ $\Sigma_{tt} \leq \sigma^2$
- No restrictions on adversary.

- How to answer all questions accurately?
  - i.e., how to minimize

$$R(A_{1:k}) = \sup_{T_{1:k}} \left[\max_{i \in [k]} \mathbb{E}(A_i - \mu_{T_i})^2 \right]$$

Known estimators

- Naïve estimator: $A_i = \phi_{T_i}$
  - Achieves rate: $\Theta(k\sigma^2)$

- Noise adding: $A_i \sim \mathcal{N}(\phi_{T_i}, \sqrt{k}\sigma^2)$
  - Achieves rate: $\Theta(\sqrt{k}\sigma^2)$ (Russo and Zou, 2016)

- Can this be improved further?

Lower bound 1 (worst case)

- Assume $|\mathcal{T}| = \Omega(2^k)$

$$\inf_{A_{1:k}} \sup_{\mathcal{D}(\phi_{\mathcal{T}})} \sup_{T_{1:k}} \left(\max_i \mathbb{E}[(A_i - \mu_{T_i})^2] \right) = \Omega(\sqrt{k}\sigma^2)$$

- Any estimators A_i with input

$$\underbrace{\phi_{\mathcal{T}}}_{\text{data}}, \underbrace{T_{1:i-1}, A_{1:i-1}}_{\text{From prev rounds}}, \underbrace{T_i}_{\text{Index}}$$

Lower bound 2 (per-Instance)

- Fix a distribution of $\phi_{\mathcal{T}}$ that's **sufficiently rich**

$$\inf_{Natural\ A_{1:k}} \sup_{T_{1:k}} \left(\max_i \mathbb{E}[(A_i - \mu_{T_i})^2] \right) = \Omega(\sqrt{k}\sigma^2).$$

- Any natural estimators A_i with input

$$\underbrace{\phi_{T_{1:i-1}}}_{\text{Only past data}}, \underbrace{T_{1:i-1}, A_{1:i-1}}_{\text{Shared history}}, \underbrace{\phi_{T_i}}_{\text{Required to avoid triviality.}}$$

*In the previous version of the paper: https://arxiv.org/abs/1602.04287
The estimators are restricted to noise adding ones. New results will be on arxiv soon.

Summary

- Gaussian noise adding is optimal up to constant factors.
- Selection itself is often enough to impose non-trivial lower bound, even for a fixed distribution.

For proof details and open problems

- Talk to me at the poster!
- Thank you!

Supplementary slides

Related work

- ADA via Differential privacy (DFHPRR15, BNSSSU15, etc…)
  - Similar setting. DP is unnecessarily strong for the purpose. Need low-sensitivity.
  - We work with conditional expectations directly.

- Lower bounds via finger printing codes (Hardt, Ullman, Steinke, etc)
  - A different setting. Also, they have a computational lower bound.
  - Suboptimal rate (if we ignore differences in settings).

- Post-selection inference (Taylor, Tibshirani, Fithian, Lee, etc.)
  - The focus is to have correct confidence interval, despite selection bias.
  - Fixed procedure, lasso-like. Not adaptive.
  - We prevent finding significantly biased statistics in the first place.

Sign inference attack

- Choose $T_1 = t_1, \ldots, T_{k-1} = t_{k-1}$
  - Such that $\phi_{t_1} \perp \ldots \perp \phi_{t_{k-1}}$
- Infer the signs of $\phi_{t_1} - \mu_{t_1}, \ldots, \phi_{t_{k-1}} - \mu_{t_{k-1}}$
  - using optimal classifier…
- Construct $T_k = t_k$
  - Such that it's correlation with $\phi_{t_1}, \ldots, \phi_{t_{k-1}}$ are proportional to the inferred signs.

Lower bound idea: Optimal obfuscation of the signs .

Example: linear regression

$$y = X\beta + N(0, \sigma^2 I)$$

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

Hope to discover:
which gene is associated with heart disease?

After looking at a sequence of values:

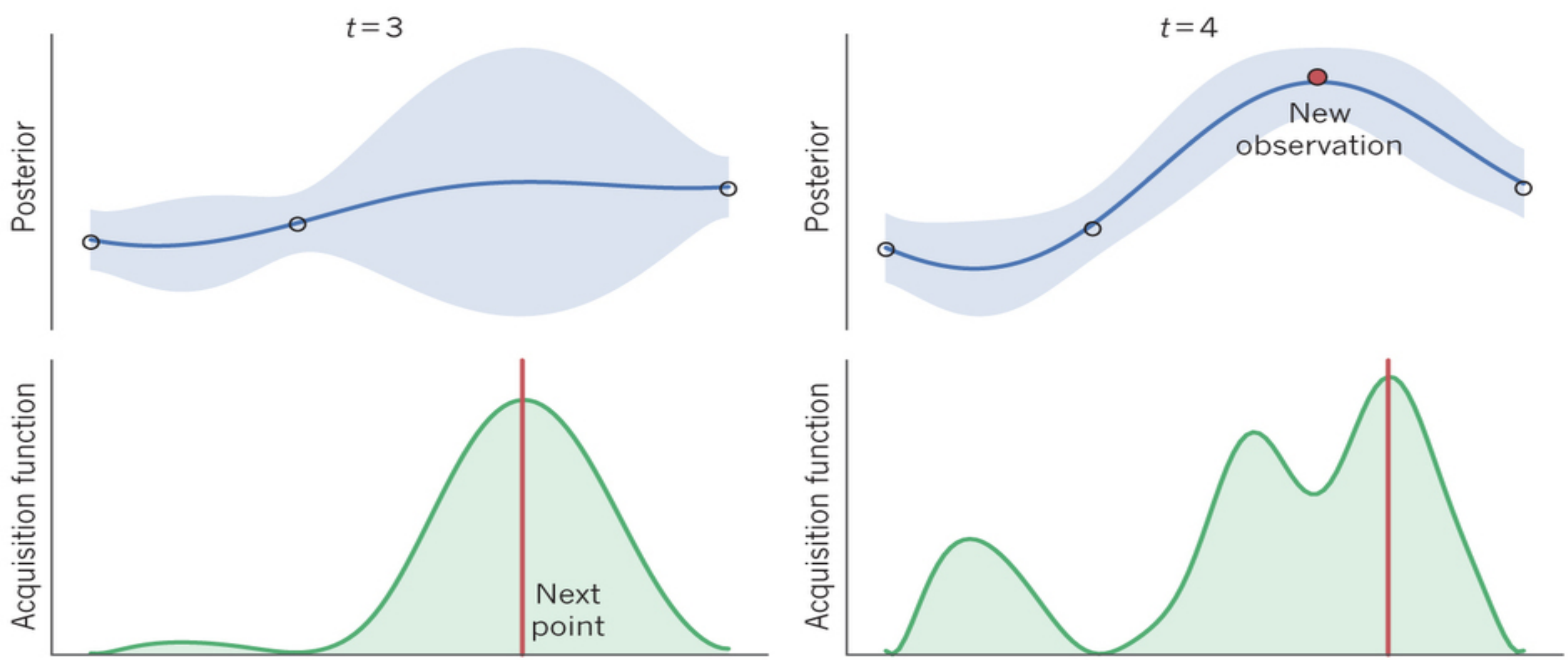
$$\langle t_1, \hat{\beta} \rangle, ..., \langle t_k, \hat{\beta} \rangle$$

We conclude that features indexed by t_k has the a strong association!

* It could also be: choose a feature subset and fit a linear regression.
The fitted parameters will still be jointly Gaussian.

Example: Hyper parameter tuning via Bayesian Optimization

- Set of d hyper parameters $\mathcal{T} = [0,1]^d$
- Grid search is too expensive.
- Often people use sequential adaptive tuning.



What's in common?

- In linear regression:

$$\phi_t = \langle t, \hat{\beta} \rangle \qquad \mathcal{T} = \{t \in \mathbb{R}^d | \|t\|_2 \leq 1\}$$

$$\mu_t = \langle t, \beta \rangle$$

Selection rule: exploratory

- In Bayesian optimization:

$$\phi_t = \mathrm{TestErr}(t) \qquad \mathcal{T} = [0,1]^d$$

$$\mu_t = \mathbb{E}[\mathrm{TestErr}(t)]$$

Selection rule: GP-UCB.

- In both cases:
  - $\phi_{\mathcal{T}}$ is a Gaussian Process.
  - Both sequential, but different selection rules