A selective survey of selective inference

J. Taylor (Stanford)

ICM 2018
Replicability crisis in science

Caveat: embarrassingly untechnical talk

Begin with focus on statisticians’ efforts addressing replicability in science.
Common mathematical theme

**Normal cycle** $N(K)$

\[ N(K) = \{(u, \beta) : u \in K, \beta \in N_u K\} \]
Replicability crisis in science

Scientists collect data first and ask questions later. (Candes)

The idea of a scientist, struck, as if by lightning with a question, is far from the truth. (Tukey)
Replicability crisis in science

Exploratory data analysis

- Tukey: scientists have always used data to form new questions – this is classical!

Confirmatory data analysis

- The standards of science require some confirmation (often statistical).

Conflict identified by Candes (and certainly others)

- Misleading to (naively) use the same data for exploration and confirmation.
Modern science

A (typical?) data scientist $U$’s workflow...

- Query $Q^U_1$ might be choice of a tuning parameter
- Query $Q^U_2$ might be a feature selection step
- Data $T'$ might be validation data
A (typical?) data scientist $U$’s workflow...

- Could have more data nodes…
- Could have more queries…
Example: predicting drug resistance

- $T$ denotes mutation patterns of HIV viruses and in vitro response to 3TC
- $Q_1^U$ asks for important main effects, $Q_2^U$ asks for important interaction; $T'$ is empty – no new data.
Modern science

Simple example: Drop the losers (Sampson and Sill)

- $T$ denotes $K$ different treatments in a clinical trial.
- $Q_1^U$ asks which treatment (apparently) works best.
Modern science

Simple example: Drop the losers (Sampson and Sill)

- $Q^U_2$ asks which is second best (variant of Sampson and Sill).
- Data $T'$ is confirmatory sample: **gold standard** reports an estimate of best treatment effect based on $T'$ alone.
Modern science

Simple example: Drop the losers (Sampson and Sill)

- Wasteful to only use $T$ for selection?
- Can we reuse earlier data?
Modern science

**Conflict between confirmatory and exploratory**

- **Unadjusted estimator** \( (T_k + T'_k)/2 \) is biased, **gold standard estimator** \( T'_k \) is unbiased but more variable.

- Simple manifestation of Candes’ observation, **researcher degrees of freedom**.

- Difference can easily be bigger with larger \( K \), different sample sizes, etc.
Modern science

Reproducibility and replicability

- Great efforts have been made to make \( U \)'s results **reproducible**.
- For **replicability** we need to (statistically) understand this collection of random variables.
Modern science

Selective inference

- Valid inference in the presence of selection effects determined by $Q^U_1, Q^U_2$. 
Modern science

Classical inference

- \( U \)'s exploratory interaction with the data is limited to \( T \).
- Earlier data is “wasted”, confirmatory focus on \( T' \).
- No selection effect.
Modern science

Challenge for selective inference

- A scientist’s question does not always translate easily into statistical objects, a necessary step to model \( U \)’s workflow.

Statistical objects

- Statistical model: e.g. for gold standard \( T' | T \sim F \in \mathcal{M} \)
- Parameter: \( \theta : \mathcal{M} \rightarrow \mathbb{R} \), e.g. treatment effect for “best” treatment.
- Language of statistics: for parameter \( \theta \) we have
  1. point estimators
  2. confidence intervals
  3. posterior distributions

Scientists, when struck by anything, are not struck with null hypotheses…
Modern science

A prototype: simultaneous (large scale) inference

(Wikipedia)

- Measure disease status $D$ and a genomic signature for each of $N$ markers, $(M_i)_{1 \leq i \leq N}$.
- Natural choice of parameters: $\theta_i$ be the association of marker $M_i$ with disease $D$.
- Data: $T' = (D, (M_i)_{1 \leq i \leq N})$. 
Modern science

Large scale inference and multiple comparisons

- Within each marker, estimate association $\hat{\theta}_i$ and consider testing no association between $D$ and marker $M_i$

$$H_i : \theta_i(F) = 0, \text{ i.e. } F \ni \{ G \in \mathcal{M} : \theta_i(G) = 0 \}?$$

False Discovery Rate (FDR)

- Benjamini & Hochberg (1995) hugely influential in multiple comparisons over past 20 years, particularly in large scale inference.
Modern science

**Family Wise Error Rate (FWER)**

- Tests based on maximum association

\[
\max_{1 \leq i \leq N} \left| \frac{\hat{\theta}_i - \theta_i}{SD(\hat{\theta}_i)} \right| = \max_{1 \leq i \leq N} |Z_i|
\]

**Bonferroni and volume of tubes**

- Embedding sampling of genome (or other measurements) into some continuous space

\[
P\left( \sup_{x \in M} |Z_x| \geq u \right) \approx \sum_j L_j(M) \rho_j(u)
\]
Volume of tubes

\[ \lambda (\text{Tube}(M, r)) = \int_{N(M)} J(u, \beta) \mathcal{H}(du \, d\beta) \]
Pause: does large scale inference address $U$’s workflow?

- Arguments for: Bonferroni can be used for a confidence interval in drop the losers.
- Arguments against: questions are determined entirely by structure of $T’$. 
Conditional inference

Classical inference

- Required to collect data $T'$.
- Throwing away $T$ is conditioning on $T$. 
Conditional inference

Drop the losers

- Instead of throwing out all of $T$, condition only on which treatment is apparently best:

- Rao-Blackwell (Cohen + Sacrowicz)

$$\hat{\theta}_{\hat{K}} = E[T_k' | (T' + T)_k, (T_j)_{j\neq k}, \hat{K} = k]$$
Conditional inference

The (classical) scientific method is inadmissible!

- Tests and confidence intervals also available.
- Similar technique can be used when looking at best 2 treatments, rather than just single best treatment.
Conditional inference

The (classical) scientific method is inadmissible!

- Tests and confidence intervals also available.
- Similar technique can be used when looking at best 2 treatments, rather than just single best treatment.
Conditional inference

Classical scientific method

- Specifically allows $U$’s intervention – even model $\mathcal{M}$ is chosen after observing $(Q_1^U, Q_2^U)$!
- $U$ specifies model $\mathcal{M}$ for the law $T'|T$. 
Conditional inference

General approach

- $U$ specifies model $M$ for the law $(T', T)$.
- Do statistics on all data $(T, T')$. 
Conditional inference

General approach

- Is this improvement limited to *drop the losers*? **No.**
- Do we need confirmatory sample $T'$? **No. We can even have** $T = T'$.  
- Can we allow arbitrary queries? **Probably not.**
Conditional inference

General approach

- Inference is carried out in **selective model**

\[ \mathcal{M}^* = \left\{ F^* : \frac{dF^*}{dF} \propto \zeta^* \right\} \]

- The function \( \zeta^* \) can be “read off” the dependency graph knowing the observed values of \( Q_1^U \) and \( Q_2^U \).
Conditional inference

General approach

- Theoretical crux becomes transferring what we know about $\mathcal{M}$ to $\mathcal{M}^*$ (i.e. consistency, CLT, etc.)
- Computational crux becomes describing $\zeta^*$ in silico.
Conditional inference

Randomized convex programs

- In drop the losers, for $Q_1^U$ we solve

  \[
  \max_{\alpha \in S_K} \langle \alpha, T \rangle
  \]

  with

  \[
  S_K = \left\{ \alpha \in \mathbb{R}^K : \alpha_i \geq 0, \sum_{i=1}^K \alpha_i = 1 \right\}.
  \]

- With $\omega = T' - T$, this is (essentially) equivalent to

  \[
  \max_{\alpha \in S_K} \langle \alpha, T + T' \rangle - \langle \alpha, \omega \rangle
  \]

- A perturbed version of

  \[
  \max_{\alpha \in S_K} \langle \alpha, T + T' \rangle
  \]
Conditional inference

Randomized convex programs

\[ \zeta^*(T + T') = \int_{N_k S_K} g_\omega(T + T' - \eta) \cdot J(T + T', \eta) \, d\eta \]
Conditional inference

Randomized convex programs

\[
\min_{\beta} \ell(\beta; T) + P(\beta) - \omega^T \beta, \quad \omega \sim G
\]

Structure inducing penalties

\[
P(\beta) = \sup_{u \in K} \langle u, \beta \rangle, \quad \text{e.g. } P(\beta) = \lambda \| \beta \|_1
\]

Subgradient equations

\[
\hat{u} = -\nabla \ell(\hat{\beta}; T), \quad \hat{u} \in \partial P(\hat{\beta})
\]

A model for queries: \( Q^U = Q^U(\hat{\beta}, \hat{u}) \)
Conditional inference

Structure inducing penalties

\[ u \in \partial P(\beta) \iff \beta \in N_u K. \]
Conditional inference

Adjustment factor

\[ \zeta^*(T) = \int_{\{u, \beta\} : Q^v(\beta, u) = q} g_\omega (\nabla \ell(\beta; T) + u) J(T, \beta, u) \mathcal{H}(du \, d\beta) \]
Conditional inference

What is the payoff?

- Many structure-detection algorithms in modern applied statistics can be cast as convex problems.

Canonical example

- LASSO (noisy version of compressed sensing)

  $$\minimize_{\beta} \ell(\beta; T) + \lambda \|\beta\|_1$$

- Solution is sparse for large values of $\lambda$.

- Also **hugely influential** over last 20 years in statistics.
Conditional inference

Randomized LASSO (Tian-Harris et al. 2016, arxiv/1609.05609)

\[
\text{minimize}_{\beta} \ell(\beta; T) + \lambda \| \beta \|_1 - \omega^T \beta + \frac{\epsilon}{2} \| \beta \|_2^2
\]

- Query:

\[Q^U(T, \omega) = \hat{u}(T, \omega) = u_{\text{obs}}\]

- Adjustment:

\[\zeta^*(T) = \int_{\mathbb{R}^p} g_{\omega} (\nabla \ell(T; \beta) + \epsilon \beta + u_{\text{obs}}) \mathcal{H}(d\beta)\]
Conditional inference

What is the payoff?

- Many queries can be combined.
- Fairly flexible set of analysis pipelines can be subsumed in this model.
Conditional inference

What is the payoff?

\[ \zeta^* (T, T', T'') = \zeta^*_{1,q_1} (T) \cdot \zeta^*_{2,(q_1,q_2)} (T) \cdot \zeta^*_{3,(q_1,q_2.q_3)} (T, T') \]
Conditional inference

What is the payoff?

- Revisit our HIV resistance data: \( n, p = 633, 91 \).
- Goal: predict in-vitro resistance from mutation pattern.

Workflow

1. Search for important main effects using (randomized) marginal screening at some threshold.

2. \( U \) decides that even though mutation K65R was not discovered 1., it should be included.

3. Interaction effects for these first stage mutations are discovered using a (randomized) LASSO.

4. Report desired \( p \)-values, point estimates, intervals.
Conditional inference

What is the payoff?

- Not clear how to do valid inference in other ways besides data splitting (or collecting new data).
Conditional inference

Cost of selection
Conditional inference

Does it work?

![Graph showing coverage vs. rho with signal = 4.5](image)
Conditional inference

Does it work?
Conditional inference

Challenges

- Practical concerns:
  1. Tradeoff between selection quality and inferential power.

- Theoretical properties (some preliminary results e.g. Tian and Taylor (2018)):
  1. consistency
  2. CLT
  3. High dimensions

- Computational properties:
  1. Evaluation of $\zeta^*$
  2. Quality of MCMC (also a theoretical question)
Conclusion

Takeaways

- Modern science requires statistics to adjust to how data is used.
- Simultaneous and conditional inference: we need both!
- Interesting practical, theoretical and computation questions.
Image credits

- Wikipedia for Tukey and Manhattan plot.