Hypothesis testing with e-values

ISI WSC IPS Discussion

Wouter Koolen

July 13th, 2021

Centrum Wiskunde & Informatica

Conclusion

E-values are an exciting way to measure evidence against a (composite) null hypothesis.

Conclusion

E-values are an exciting way to measure evidence against a (composite) null hypothesis.

We (i.e. you!) are developing a better understanding of the tools, techniques, trade-offs ...

Conclusion

E-values are an exciting way to measure evidence against a (composite) null hypothesis.

We (i.e. you!) are developing a better understanding of the tools, techniques, trade-offs . . .

- Practical, intuitive toolbox for designing things that work.
- Beautiful open questions in the theory.

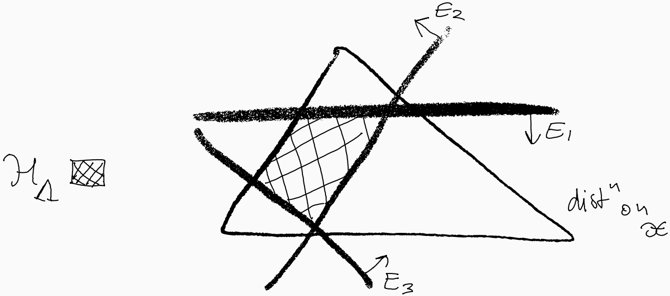
GRO Non-parametrics

Larsson & De Heide

Larsson constructs hypotheses by requiring a family $(E_\lambda)_{\lambda \in \Lambda}$ to be e-values:

e-value hypothesis class

$$\mathcal{H}_\Lambda := \left\{ P \text{ on } \mathcal{X} \;\middle|\; \forall \lambda \in \Lambda : \mathbb{E}_{X \sim P}[E_\lambda(X)] \le 1 \right\}$$



Larsson & De Heide

Larsson constructs hypotheses by requiring a family $(E_\lambda)_{\lambda \in \Lambda}$ to be e-values:

e-value hypothesis class

$$\mathcal{H}_\Lambda \ := \ \left\{ P \text{ on } \mathcal{X} \;\middle|\; \forall \lambda \in \Lambda : \mathbb{E}_{X \sim P}\left[E_\lambda(X)\right] \leq 1 \right\}$$

Example

Interesting example: bounded $(1+\epsilon)$-th moment for $\epsilon > 0$:

$$\mathbb{E}_{X \sim P}\left[\left|X\right|^{1+\epsilon}\right] \ \leq \ B$$

Heavy-tailed distributions including Pareto, Fisher-Tippett, ...

De Heide proposes the GRO criterion.

GRO

$$S^* := \operatorname*{arg\,max}_{S \text{ an e-value for } \mathcal{H}_0} \mathbb{E}_{X \sim Q} [\ln S(X)]$$

Larsson & De Heide

De Heide proposes the GRO criterion.

GRO

$$S^* \;:=\; \underset{S \text{ an e-value for } \mathcal{H}_0}{\arg\max} \;\; \mathbb{E}_{X \sim Q}\left[\ln S(X)\right]$$

Question

What does the non-parametric GRO look like?

GRO e-value Hypotheses

Larsson's theorem (finite Λ version)

Theorem

S is an e-value for $\mathcal{H}_\Lambda$ if there are non-negative $\pi \geq 0$ such that

$$S(x) \;\leq\; 1 + \sum_{\lambda \in \Lambda} \pi_\lambda (E_\lambda(x) - 1) \qquad \text{for all } x \in \mathcal{X}.$$

GRO e-value Hypotheses

Larsson's theorem (finite Λ version)

Theorem

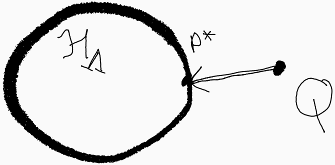
S is an e-value for $\mathcal{H}_\Lambda$ if there are non-negative $\pi \geq 0$ such that

$$S(x) \;\leq\; 1 + \sum_{\lambda \in \Lambda} \pi_\lambda (E_\lambda(x) - 1) \qquad \textit{for all } x \in \mathcal{X}.$$

The GRO e-value for $\mathcal{H}_\Lambda$ is found by solving

$$\max_{S \text{ an e-value for } \mathcal{H}_0} \mathbb{E}_{X \sim Q}[\ln S(X)] \;=\; \max_{\pi \geq 0} \mathbb{E}_{X \sim Q}\left[\ln \left(1 + \sum_{\lambda \in \Lambda} \pi_\lambda \left(E_\lambda(X) - 1\right)\right)\right]$$

RIPr version



We may also write the following

$$
\begin{aligned}
&\min_{P \in \mathcal{H}_\Lambda} \ \mathrm{KL}(Q \| P) \\
&= \max_{\substack{\nu \in \mathbb{R} \\ \pi \geq 0}} \min_{P \geq 0} \ \mathbb{E}_Q \left[\ln \frac{Q(X)}{P(X)} \right] + \sum_{\lambda \in \Lambda} \pi_\lambda \left(\mathbb{E}_{X \sim P} \left[E_\lambda(X) - 1 \right] \right) + \nu \left(\mathbb{E}_{X \sim P}[1] - 1 \right)
\end{aligned}
$$

So that

$$
P(x) = \frac{Q(x)}{\nu + \sum_{\lambda \in \Lambda} \pi_\lambda E_\lambda(x)}
$$

All in all, the GRO e-value is a likelihood ratio

$$
S^*(x) = \frac{Q(x)}{P_{\nu^*, \lambda^*}(x)}
$$

And we are back in the **Turner 2021** case.

Application and Extensions

Techniques partially developed/exploited in Bandit literature under the names **KLInf** and **empirical likelihood**

(Honda and Takemura, 2010; Cappé et al., 2013; Agrawal, Juneja, and Glynn, 2020; Agrawal, Juneja, and Koolen, 2021; Agrawal, Koolen, and Juneja, 2020)

Application and Extensions

Techniques partially developed/exploited in Bandit literature under the names **KLInf** and **empirical likelihood**

(Honda and Takemura, 2010; Cappé et al., 2013; Agrawal, Juneja, and Glynn, 2020; Agrawal, Juneja, and Koolen, 2021; Agrawal, Koolen, and Juneja, 2020)

Optimality both for regret and PAC learning objectives.

Application and Extensions

Techniques partially developed/exploited in Bandit literature under the names **KLInf** and **empirical likelihood**

(Honda and Takemura, 2010; Cappé et al., 2013; Agrawal, Juneja, and Glynn, 2020; Agrawal, Juneja, and Koolen, 2021; Agrawal, Koolen, and Juneja, 2020)

Optimality both for regret and PAC learning objectives.

Connections to **worst-case regret bounds** for exp-concave losses (yield anytime-valid confidence intervals with $|\Lambda|$ as the notion of capacity).

Application and Extensions

Techniques partially developed/exploited in Bandit literature under the names **KLInf** and **empirical likelihood**

(Honda and Takemura, 2010; Cappé et al., 2013; Agrawal, Juneja, and Glynn, 2020; Agrawal, Juneja, and Koolen, 2021; Agrawal, Koolen, and Juneja, 2020)

Optimality both for regret and PAC learning objectives.

Connections to **worst-case regret bounds** for exp-concave losses (yield anytime-valid confidence intervals with $|\Lambda|$ as the notion of capacity).

Question

Non-linear conditions? Variance? CVaR? Centered moment constraints?

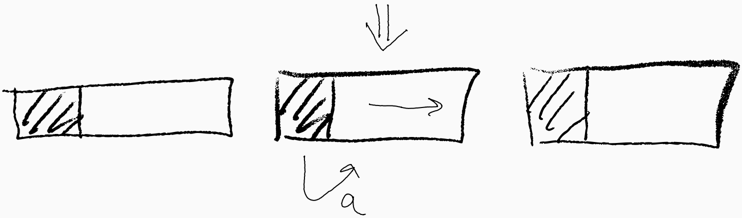
Choosing the Alternative

Vovk

Cox' problem: test all means 0 vs one mean $\mu \neq 0$.

Vovk proposes

- Use mean on first blocks to pick a population.
- Compute the mean (ML) a of the first block.
- Use i.i.d. P_a as the alternative model for the second block.



Vovk

Cox' problem: test all means 0 vs one mean $\mu \neq 0$.

Vovk proposes

- Use mean on first blocks to pick a population.
- Compute the mean (ML) a of the first block.
- Use i.i.d. P_a as the alternative model for the second block.

General(!) Simple. Elegant. Beautiful.

Vovk

Cox' problem: test all means 0 vs one mean $\mu \neq 0$.

Vovk proposes

- Use mean on first blocks to pick a population.
- Compute the mean (ML) a of the first block.
- Use i.i.d. P_a as the alternative model for the second block.

General(!) Simple. Elegant. Beautiful.

Consideration:

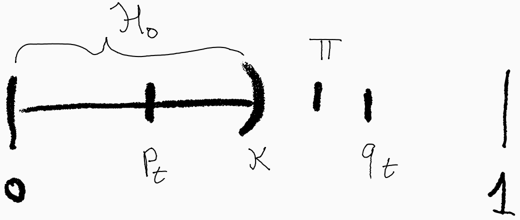
- The first blocks are used to select a block: a will overfit (slightly).
- If $a > \mu$ then $E = P_a/P_0$ is only expected to win if $a < 2\mu$.
- Are there ways to dampen a?
- Curiously: problem gets **worse** if there are *multiple populations* with $\mu \neq 0$.

Henzi

At each time-point t, the hypothesis is that p_t is a better prediction that q_t (as measured by score function S).

Henzi proposes (say $p_t < q_t$)

- Compute the mid-point $\kappa \in (p_t, q_t)$ (halfway for Brier).
- Use null $\mathcal{H}_0 = \{\mathrm{Ber}(\theta) \mid \theta \in [0, \kappa]\}$.
- Use alternative $\pi = \frac{3}{4} q_t + \frac{1}{4} p_t > \kappa$.
- Multiply evidence by e-value $P_\pi(Y)/P_\kappa(Y)$.



Henzi

At each time-point t, the hypothesis is that p_t is a better prediction that q_t (as measured by score function S).

Henzi proposes (say $p_t < q_t$)

- Compute the mid-point $\kappa \in (p_t, q_t)$ (halfway for Brier).
- Use null $\mathcal{H}_0 = \{\mathrm{Ber}(\theta) \mid \theta \in [0, \kappa]\}$.
- Use alternative $\pi = \frac{3}{4} q_t + \frac{1}{4} p_t > \kappa$.
- Multiply evidence by e-value $P_\pi(Y)/P_\kappa(Y)$.

Ingenious. Pretty. (A Larsson e-value hypothesis!)

Henzi

At each time-point t, the hypothesis is that p_t is a better prediction that q_t (as measured by score function S).

Henzi proposes (say $p_t < q_t$)

- Compute the mid-point $\kappa \in (p_t, q_t)$ (halfway for Brier).
- Use null $\mathcal{H}_0 = \{\mathrm{Ber}(\theta) \mid \theta \in [0, \kappa]\}$.
- Use alternative $\pi = \frac{3}{4} q_t + \frac{1}{4} p_t > \kappa$.
- Multiply evidence by e-value $P_\pi(Y)/P_\kappa(Y)$.

Ingenious. Pretty. (A Larsson e-value hypothesis!)

Consideration:

- π is expected to gain evidence for true parameter μ between κ and π. But what if $\mu = \kappa + \epsilon$?
- Are there (effect size?) guidelines for setting ϵ a priori?
- Universal modelling (mixture) over ϵ?

Conclusion

Safety and Power

Joy and elegance

Sequential story only partially understood.

- Product of GRO may not be GRO itself.
  Order of quantifiers matters!
- GRO+invariance sometimes leads to test-super-martingale in reduced filtration.
  When? Approximately?

Exciting area!

Thanks!

References

Agrawal, S., S. Juneja, and P. Glynn (Aug. 2020). “Optimal δ-Correct Best-Arm Selection for Heavy-Tailed Distributions”. In: *Proceedings of the 31st International Conference on Algorithmic Learning Theory*. Vol. 117. Proceedings of Machine Learning Research, pp. 61–110.

Agrawal, S., S. Juneja, and W. M. Koolen (Feb. 2021). “Regret Minimization in Heavy-Tailed Bandits”. In: *ArXiv*. Accepted to Annual Conference on Learning Theory (COLT).

Agrawal, S., W. M. Koolen, and S. Juneja (Aug. 2020). “Optimal Best-Arm Identification Methods for Tail-Risk Measures”. In: *ArXiv*. Submitted to Advances in Neural Information Processing Systems (NeurIPS).

Cappé, O., A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz, et al. (2013). "Kullback–Leibler upper confidence bounds for optimal sequential allocation". In: *The Annals of Statistics* 41.3, pp. 1516–1541.

Honda, J. and A. Takemura (2010). "An asymptotically optimal bandit algorithm for bounded support models". In: *In Proceedings of the Twenty-third Conference on Learning Theory (COLT 2010*, pp. 67–79.