STATS 300C: Theory of Statistics **Spring 2022**

Lecture 15 — May 17, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Timothy Sudijono*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and 2021 scribe notes written by Cheuk To Tsui and Yuchen Hu.*

Reading: Material for this lecture is based on [1, 2, 3].

15.1 A recap on e-values

In the last lecture, we introduced the notion of an e-value, a quantity similar to the p-value, which solves the optional continuation problem. The important aspects of e-values include:

- Validity for arbitrary dependence.
- They are easy to combine.[1]
- Flexible with regards to stop / continue procedures.
- Non-asymptotic and model free.

Many of these properties come from the fact that e-variables control the *expectation* under the null. In general, expectations are much easier to work with: the linearity of expectation is convenient, and martingale theory plays well with the expectation too (as in the Optional Stopping Theorem).

15.2 Reverse Information Projection and an analog of power

How do we generate e-values? As we saw in Lecture 14, a source of e-values comes from the notion of a Bayes Factor, introduced by Jeffreys in Bayesian statistics as a form of model selection. The setup involves two models of observed data X, whose averaged likelihoods are given by $p_{W_1}(X)$ and $p_{W_0}(X)$ respectively. Recall the notation p_W:

$$p_W(X) := \int p_\theta(X) dW(\theta).$$

[1] The average of e-values is also an e-value. The same statement for p-values is not true, although twice the average of p-values is a p-value.

Then the Bayes factor may be defined as

$$M(X) := \frac{p_{W_1}(X)}{p_{W_0}(X)}.$$

In the case of a point null, the Bayes factor does define an e-value. In the general composite case, this quantity does not define an e-value, since we require $\mathbb{E}_{X\sim P_0}[M(X)] \leq 1$, and for a Bayes factor, we can only guarantee that $\mathbb{E}_{X\sim P_{W_0}}[M(X)] \leq 1$. However, we have some recourse; another procedure known as **Reverse Information Projection** also creates e-values. In fact, the resulting e-values are optimal in a sense we shall describe. Before we explain this idea, let us quickly recall the notion of KL divergence. Let P, Q be two distributions with densities p, q. The KL divergence is defined as the quantity

$$D(P||Q) = \mathbb{E}_{X\sim P} \log \frac{p(X)}{q(X)}$$

The KL divergence is not exactly a metric, as it is not symmetric, but it does satisfy $D(P||Q) \geq 0$ with equality only if $P = Q$. This can be shown by Jensen's inequality.

If the Bayes factor for a point null defines a p-value, perhaps we can summarize the Bayes factor for a composite null using a single distribution. This is the idea behind reverse information projection.

Definition 1 (Reverse Information Projection)**.** Define

$$W_0^* = \operatorname*{argmin}_{W_0} D(P_{W_1}||P_{W_0})$$

where W_0 ranges over all priors on Θ_0. Then $P_{W_0^*}$ is the **reverse information projection** of P_{W_1} on the set $\overline{H}_0 = \{P_W : W \text{ a prior on } \Theta_0\}$.

Theorem 1 (Li '99, Barron & Li '99, Grunwald et al. '19)**.** If W_0^* exists, then

$$P_{W_1}(X)/P_{W_0^*}(X)$$

is an e-variable. It is the Bayes **GROW** e-variable relative to W_1, achieving

$$\max \mathbb{E}_{X\sim P_{W_1}}[\log E]$$

where the maximum is over e-variables for H_0.

We have just introduced an optimality criterion for e-variables, GROW[2], that can be seen as an analog of power for p-values. Why is this the right optimality criterion to consider? Intuitively, we would like to minimize the chance that E is small, or even zero. Revisiting the optional continuation examples from the last lecture, the right way to define an e-value for this setting is the product $\prod_i E_i$. If any of the E_i are zero, then the product of these quantities is zero, which is useless for inference. But why should we look to maximize the logarithm of E rather than the expectation of E itself? One answer comes from an interpretation of e-values via betting.

[2]In [3], the authors use the acronym GRO, which stands for growth-rate optimality.

15.2.1 A betting perspective on e-values

Suppose we turn the hypothesis testing problem between two hypotheses H_0, H_1 into a betting game. At each turn, you may choose the amount to stake, and for each dollar bet, the payoff is equal to the e-value $E(x)$, dependent on the outcome x. For simplicity, under the null, suppose that $\mathbb{E}_0 E(x) = 1$, so you won't make money on average. Now if the alternative is true, suppose that $\mathbb{E}_1 E(x) > 1$. This situation corresponds to a game in which you think the alternative hypothesis is true, and under this regime, the payoff $E(x)$ is positive. What is the long term wealth you should expect to have at time T? In general, the wealth should be exponential in T. Taking the logarithm, we should expect the total wealth to be like $T \log E(x)$. In light of this it is not unreasonable to maximize $\mathbb{E}_1 \log E(x)$. This strategy is known as *Kelly Gambling,* or the *Kelly Criterion*.

To complete this connection to betting, we show that if $E(x)$ is the likelihood ratio $\frac{p_1(X)}{p_0(X)}$ then it maximizes the logarithm. Because the expectation of $E(x)$ under H_0 is one, we may write

$$E(x) = \frac{q(x)}{p_0(x)}$$

for q some probability density. But then

$$\begin{aligned}\mathbb{E}_1\left[\log \frac{p_1(x)}{p_0(x)}\right] - \mathbb{E}_1[\log E(x)] &= \mathbb{E}_1\left[\log \frac{p_1(x)}{p_0(x)}\right] - \mathbb{E}_1\left[\log \frac{q(x)}{p_0(x)}\right] \\ &= \mathbb{E}_1\left[\log \frac{p_1(x)}{q(x)}\right].\end{aligned}$$

But this last term is the KL divergence of p_1 with respect to q which is always positive.

15.3 FDR control via e-values

We have seen safe testing and continued testing as applications of e-values. Can we utilize this idea to do multiple hypothesis testing, ideally with some control on FWER or FDR? In this setup we observe n realized e-values $e_1, \dots, e_n$ associated to null hypotheses $H_1, \dots, H_n$. Because inverse e-values e_i^{-1} are p-values, a natural idea is to run Benjamini Hochberg using these quantities as p-values. This defines a procedure which we'll call the e-BH procedure.

The e-BH procedure fixes an FDR level $\alpha \in (0, 1)$, and orders the realized e-values as $e_{(1)} \geq \cdots \geq e_{(n)}$. The procedure then rejects hypotheses with the largest $\hat{k}$ e-values, where

$$\hat{k} := \max\left\{i : \frac{i e_{(i)}}{n} \geq \frac{1}{\alpha}\right\}. \tag{15.1}$$

If we define the p-values $p_{(i)} = e_{(i)}^{-1}$, the stopping condition of Equation 15.1 can be rewritten as

$$\max\left\{i : p_{(i)} \leq \frac{i\alpha}{n}\right\},$$

which is exactly the stopping condition for Benjamini Hochberg. Given this relation to BH, it is not surprising that the false discovery rate is controlled:

Theorem 2 (Wang & Ramdas '20 [1])**.** The e-BH procedure has FDR at most $n_0\alpha/n$, where n_0 is the number of nulls among the hypothesis H_i.

The proof shares many similarities with proof of the BH procedure from previous lectures.

Proof. As usual, write

$$FDP = \sum_{i \in H_0} \frac{V}{1 \wedge R} = \sum_{i \in H_0} \frac{V}{1 \wedge \hat{k}}.$$

By definition, $i \leq \hat{k}$ for every $i \in H_0$. Thus,

$$\begin{aligned} FDP \leq \sum_{i \in H_0} \frac{V}{1 \wedge \hat{k}} &\leq \sum_{i \in H_0} \frac{V_{(i)}}{i} \\ &\leq \sum_{i \in H_0} V_{(i)} e_{(i)} \frac{\alpha}{n} \\ &\leq \sum_{i \in H_0} e_{(i)} \frac{\alpha}{n}. \end{aligned}$$

Taking an expectation on the last line gives

$$\sum_{i \in H_0} \mathbb{E}[e_{(i)}] \frac{\alpha}{n} \leq \frac{\alpha n_0}{n}$$

since the e_i are e-values. □

At first glance, this guarantee is similar to the FDR control of Benjamini Hochberg (BH) using p-values. The key difference however, is that Theorem 2 controls FDR under **arbitrary** dependence. In an earlier lecture, we proved a result on FDR control up to logarithmic factors, under arbitrary dependence. This suggests that the e-BH procedure somehow accounts for this, so we should expect the procedure to have low power.

15.4 Confidence Intervals with e-values

In recent work [2], Xu et al. look into the problem of designing confidence intervals with proper coverage using e-values, in the setting of post-selection inference. In this problem, we are interested in concurrently estimating many parameters $\{1, \ldots, n\}$; to quantify uncertainty about these parameters, we create $1 - \alpha$ confidence intervals for each of these parameters. In modern science however, there is usually some form of data snooping in which the scientist selects some subset $S \subseteq \{1, \ldots, n\}$ of these parameters to investigate, based on ad-hoc decisions or the data that has been observed.

Due to the snooping procedure, we should expect much fewer than $1 - \alpha$ of these confidence intervals to cover their true parameters. How can we adapt the given confidence intervals to retain proper coverage?

To see how e-values can help here, let us first try to construct a confidence interval for a parameter $\theta \in H_0$ using an e-value. Suppose we have an e-value $E(\theta)$ so that $\mathbb{E}_\theta E(\theta) \leq 1$ for all $\theta \in H_0$. Define the confidence interval

$$C(\alpha) = \left\{\theta : E(\theta) \leq \frac{1}{\alpha}\right\},$$

which is called an **e-CI**. This is valid by the following calculation. Letting θ^* be the true value of θ,

$$\mathbb{P}_{\theta^*}(\theta^* \notin C(\alpha)) = \mathbb{P}_{\theta^*}\left(E \leq \frac{1}{\alpha}\right) \leq \alpha \mathbb{E}_\theta E \leq \alpha,$$

where we have used Markov's inequality.

Given these e-CIs, [2] introduces an adapted version of the Benjamini-Yekutieli procedure from Lecture 12, called the e-BY procedure. One of the main results of the work is that this procedure controls the FCR.

Definition 2 (The e-BY procedure [2])**.** Suppose we have a method $C_i(\alpha)$ for creating e-CIs at any confidence level $1-\alpha$ for each parameter $i \in \{1, \ldots, n\}$. After data snooping, select a subset of parameters S for which we will construct confidence intervals.

Return the widened confidence intervals $C_i(\alpha_i)$ where

$$\alpha_i := \frac{\delta |S|}{n},$$

for δ a pre-specified level of control on the False Coverage Rate.

Theorem 3 (e-BY controls FCR [2])**.** The e-BY procedure produces confidence intervals which control the False Coverage Rate at level δ.

There are a few things to like about this result. Firstly, there is no independence assumption. Comparing Theorem 3 to the result that Benjamini-Yekutieli controls FCR, we also see that we remove the complex $R_{(i)}$ terms.

Proof. We first begin by writing the definition of the False Coverage Proportion (FCP), and then proceed by some loose approximations:

$$\begin{aligned}
FCP &= \sum_{i=1}^{n} \frac{\mathbf{1}\{i \in S, \theta_i \notin C_i(\frac{\delta}{n}|S|)\}}{|S|} \\
&= \sum_{i=1}^{n} \frac{\mathbf{1}\{i \in S, E(\theta_i) \geq \frac{n}{\delta|S|}\}}{|S|} \\
&= \sum_{i=1}^{n} \mathbf{1}\{i \in S, E(\theta_i) \geq \frac{n}{\delta|S|}\} E(\theta_i) \frac{\delta}{n} \\
&\leq \sum E(\theta_i) \frac{\delta}{n}.
\end{aligned}$$

Thus

$$FCR = \mathbb{E}[FCP] \leq \sum \mathbb{E} E(\theta_i) \frac{\delta}{n} \leq \delta,$$

as desired. □

15.5 Summary

Overall, e-values are useful in that they allow for safe Type I error testing under complicated optional continuation situations. The major pitfall of these methods is low power; tests based on e-values are less likely to reject the null. This might be a issue when proposing new e-values, as in the following problem:

Problem 1. How do you use e-values to test for nonzero coefficients in the logistic regression model?[3]

[3]For one approach, see universal inference.

Bibliography

[1] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *arXiv preprint arXiv:2009.02824*, 2020.

[2] Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. Post-selection inference for e-value based confidence intervals. *arXiv preprint arXiv:2203.12572*, 2022.

[3] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.