Online Supplementary Appendix for "Decentralized Nonparametric Multiple Testing"

Subhadeep Mukhopadhyay*

Temple University, Department of Statistical Science

Philadelphia, Pennsylvania, 19122, U.S.A.

* Email correspondence should be directed to deep@temple.edu

This supplementary document contains two Appendices. Appendix A provides several proofs of results in the main paper. Appendix B includes some additional remarks.

A. PROOFS

A1. Proof of Theorem 2

We start by noting the skew-beta model density model:

$$d(u; F_0, F) \;=\; f_{\mathrm{B}}(u;\, \gamma, \beta)\Big\{1 + \sum_j \mathrm{LP}[j; F_{\mathrm{B}}, D]\, T_j(u; F_{\mathrm{B}})\Big\}, \quad \text{for } 0 < u < 1, \tag{4.1}$$

where beta density and cdf with parameters γ and β are denoted by f_{B} and F_{B}, respectively; $T_j(u; F_{\mathrm{B}})$ are called beta-LP polynomials $\mathrm{Leg}_j \circ F_B(u; \gamma, \beta)$. Here the sign '$\circ$' refers to the usual composition of functions. The beta-LP polynomials satisfy the following orthonormality conditions:

$$\mathbb{E}_{F_B}[T_j(U; F_B)] = 0, \;\; \text{and} \;\; \mathbb{E}_{F_B}[T_j(U; F_B) T_k(U; F_B)] = \delta_{jk}.$$

This implies that the LP-Fourier coefficients of (4.1) can now be expressed as

$$\begin{aligned} \mathrm{LP}[j; F_{\mathrm{B}}, D] \;&=\; \int_0^1 \frac{d(u; F_0, F)}{f_{\mathrm{B}}(u;\, \gamma, \beta)} T_j(u; F_B)\, \mathrm{d}F_B(u; \gamma, \beta) \\ &=\; \int_0^1 T_j(u; F_B)\, \mathrm{d}D(u; F_0, F) = \mathbb{E}_D[T_j(u; F_B)]. \end{aligned} \tag{4.2}$$

Complete the proof by replacing the population D in (4.2) by its sample estimator $\widetilde{D}$ to compute $\mathrm{LP}[j; F_{\mathrm{B}}, \widetilde{D}]$. □

A2. Proof of Theorem 3

We begin by recalling the definition of sample comparison density $\widetilde{D}_l \equiv D(u; F_0, \widetilde{F}_l)$ of the l-th partitioned p-values:

$$D(u; F_0, \widetilde{F}_l) \;=\; \widetilde{F}_l(Q(u; F_0)) \;=\; n_l^{-1} \sum_{i=1}^{n_l} \mathbb{I}(u_{li} \le u). \tag{4.3}$$

Theorem 2 implies that the sample LP-Fourier coefficients for the l-th partition is given by

$$\mathrm{LP}[j; F_{\mathrm{B}}, \widetilde{D}_l] = n_l^{-1} \sum_{i=1}^{n_l} \mathrm{Leg}_j \circ F_B(u_{li}; \gamma, \beta), \quad j = 1, \ldots, m. \tag{4.4}$$

This ensures that the full-data sample LP-Fourier coefficients can be expressed as

$$\begin{aligned} \mathrm{LP}[j; F_{\mathrm{B}}, \widetilde{D}] \;&=\; N^{-1} \sum_{l=1}^{K} \sum_{i=1}^{n_l} \mathrm{Leg}_j \circ F_B(u_{li}; \gamma, \beta) \\ &=\; \sum_{l=1}^{K} \Big\{ N^{-1} \sum_{i=1}^{n_l} \mathrm{Leg}_j \circ F_B(u_{li}; \gamma, \beta) \Big\}, \end{aligned}$$

which by virtue of (4.3) and (4.4), can be rewritten as

$$\mathrm{LP}[j; F_{\mathrm{B}}, \widetilde{D}] = \sum_{l=1}^{K} \pi_l \, \mathrm{LP}[j; F_{\mathrm{B}}, \widetilde{D}_l],$$

where $\pi_l = n_l/N$. This proves the claim. $\square$

A3. Proof of Theorem 4

This is immediate from (2.7) and Theorem 3, as noted in (2.6).

A4. Proof of Theorem 5

The chisquare divergence between skew-G comparison density

$$d(u; G, F) = g(u)\big\{1 + \sum_j \mathrm{LP}[j; G, D] T_j(u; G)\big\},$$

and an arbitrary G over the unit interval is given by

$$\chi^2(D||G) = \int_0^1 \left[\frac{d(u)}{g(u)} - 1\right]^2 g(u)\, \mathrm{d}u = \int_0^1 \left\{ \sum_j \mathrm{LP}[j; G, D] T_j(u; G) \right\}^2 g(u)\, \mathrm{d}u. \tag{4.5}$$

Straightforward calculation shows (4.5) has the following analytic form:

$$\sum_j \big| \operatorname{LP}[j;G,D] \big|^2 \int_0^1 T_j^2(u;G)\, \mathrm{d}G + \sum_{j\neq k} \operatorname{LP}[j;G,D]\, \operatorname{LP}[k;G,D] \int_0^1 T_j(u;G) T_k(u;G)\, \mathrm{d}G,$$

which completes the proof. □

B. ADDITIONAL REMARKS

B1. Advantages of LP-skew Density Model. The reason for using LP-skew density model (2.3) instead of classical kernel density estimate (KDE) is threefold:

- *Statistical side*: KDE for compact support $[0,1]$ is known to be a challenging problem due to the "boundary effect," Besides this, difficulty arises to accurately estimate the *highly dynamic* tails near 0 and 1, such as shown in the bottom panel of Fig 2. As noted in Mukhopadhyay (2016), the novelty of our approach lies in its unique ability to "decouple" the density estimation problem into two separate modeling problems: the tail part and the central part of the distribution. Keep in mind that tails (where the signals hide) of $\widehat{d}(u;F_0,F)$ are the most important part for multiple testing.

- *Computational side*: The brute-force application of KDE $\frac{1}{Nh}\sum_{l=1}^{K}\sum_{i=1}^{n_l} K\left(\frac{u-u_{li}}{h}\right)$ requires $O(N^2)$ kernel evaluations and $O(N^2)$ multiplications and additions, making it computationally impractical for large-N problems (even for a fixed-bandwidth case).

- *Compressibility side*: The skew-beta model encodes the shape of the density using few LP-Fourier coefficients†. For example, in the Prostate cancer example, we were able to compress the whole function into three coefficients. This compressive representation is particularly attractive for designing memory-efficient big-data algorithms. Contrast this with classical KDE approach, where storing the density estimate values at each data point could be expensive, if not infeasible.

B2. On The Algorithm. The prescribed embarrassingly parallel inference algorithm:

†Note that, our specially designed LP-basis functions $T_j(u;F_B)$ are: (i) orthonormal basis with respect to the measure F_B, which guarantees parsimony of our density expansion, and (ii) robust in nature (as they are polynomials of rank-transform $F_B(u;\gamma,\beta)$, thus can tackle highly-dynamic tails of the distribution without falling prey to the spurious bumps.

- Upgrades traditional raw-empirical multiple testing procedures to a more stable and smooth-nonparametric versions.
- Performs smooth-BH filtering, by computing $u_{\max} = \sup_u \left\{ \frac{\widehat{D}(u)}{u} \geq \dfrac{\eta}{\alpha} \right\}$, which can be done without any reference to the partitioned-pvalues once we have the $\widehat{D}$. Report the cases with $u_{li} \leq u_{\max}$ as interesting for $l = 1, \ldots, k$. Contrast this with the "naive" $\widetilde{D}(u)$ based BH procedure (2.1), which requires sorting of p-values to count the empirical proportions. Also see Remark 1.
- Along the same line, one can also perform local-fdr analysis by evaluating $\widehat{d}(u_{li}; F_0, F) > \eta/2\alpha$ inside each partition, once we have $\widehat{d}$ (computed in a completely parallelized manner with zero-communication between the nodes).

This again shows the usefulness of comparison-density-based functional reformulation of multiple testing problems.

B3. Functional View of Multiple Testing. As noted in Mukhopadhyay (2016), the notion of comparison distribution allows us to transform the simultaneous hypothesis testing problem into a nonparametric function estimation problem. The transition from discrete analysis and ranking of individual p-values to comparison density function estimation[†] is necessary to develop the decentralized large-scale inference (DSLI) engine.

B4. Model Selection. For constructing skew-beta model it is important to properly select the empirical LP-Fourier coefficients appearing in (4.1). Identify indices j for which $\mathrm{LP}(j; F_B; D)$ are significantly non-zero by using AIC model selection criterion applied to LP means arranged in decreasing magnitude. Choose k to maximize $\mathrm{AIC}(k)$,

$$\mathrm{AIC}(k) \;=\; \text{sum of squares of first } k \text{ sorted LP-means} \;-\; 2k/N$$

This functionality was incorporated as an inbuilt mechanism for our decentralized algorithm. From a theoretical perspective, the proposed AIC-based LP-Fourier coefficient selection criterion can be shown to minimize the mean integrated squared error (Mukhopadhyay, 2017, Sec. 2.4).

[†]This can also be viewed as going from large-N microscopic discrete model to a functional macroscopic model that obeys the superposition principle (see Remark 4 of the main paper).

B5. Same Covariates on Different Machines. Consider the case where we have same covariates on different machines. Define $\bar{x}_0$ and $\bar{x}_1$ to be the global group-specific sample means, which can be computed easily (in a parallelized manner):

$$\bar{x}_0 = \sum_{l=1}^{k} \pi_{l0}\bar{x}_{l0}, \text{ and } \bar{x}_1 = \sum_{l=1}^{k} \pi_{l1}\bar{x}_{l1},$$

where $\pi_{l0} = n_{l0}/N_0$, $\pi_{l1} = n_{l1}/N_1$, $n_l = n_{l0} + n_{l1}$, $N_0 = \sum_{l=1}^{k} n_{l0}$, and $N_1 = \sum_{l=1}^{k} n_{l1}$. Exact similar process is also valid for the sample standard deviations S_1^2 and S_2^2. This implies that we can easily compute the full-data Z or t-statistics $Z_1, \ldots, Z_p$ and can perform multiple-testing without any problem.

On the other hand, this paper addresses the challenging regime where a massive collection of covariates are distributed over the machines, which needs a non-trivial solution and carries more appeal than the 'large-n small-p' case, especially in the context of multiple testing.

References

Efron, B. (2010), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge; New York: Cambridge University Press.

Mukhopadhyay, S. (2016), "Large Scale Signal Detection: A Unifying View," *Biometrics*, 72, 325–334.

Mukhopadhyay, S. (2017), "Large-Scale Mode Identification and Data-Driven Sciences," *Electronic Journal of Statistics*, 11, 215–240.