Decentralized False Discovery Rate (FDR) Control in Wireless Sensor Networks

Final Project Report, ECE 6960: Large-Scale Inference

Xiang Zhang
Department of Electrical and Computer Engineering, University of Utah
Email: xiang.zhang@utah.edu

Abstract

In this report we investigate some of the recent results in distributed/decentralized False Discovery Rate (FDR) control for multiple testing with applications in wireless sensor networks. The work is mainly based on [1]. In this paper, an algorithm called QuTE is proposed to control FDR in a wireless sensor network, in a distributed manner. The QuTE algorithm works in three stages, i.e., Quest, Test and Exchange. Theoretical analysis shows that this algorithm controls the overall FDR at a certain level. Simulations are provided to demonstrate the performance of the proposed algorithm. We focus on the proof of the convergence of FDR and interpretation of the simulation result. Several possible future directions are discussed.

I. Introduction

This work focuses on multiple testing of hypotheses in a distributed manner, based on the well-known Benjamini-Hochberg (BH) procedure. This work is also among the very first attempts to address the FDR control in a distributed manner. The problem setting is as follows. The network is represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and each node a is responsible for set of local hypotheses denoted as $\mathcal{H}_a = \{H_{a,i}\}_{i=1}^{n_a}$, among which $\mathcal{H}_a^0$ denotes the set of true nulls. The overall set of true nulls is denoted as $\mathcal{H}^0$ and there are $N = \sum_{a\in\mathcal{V}} n_a$ hypotheses to test. Assume that all nodes will test different hypotheses, i.e., for any two nodes $a, b \in \mathcal{V}$, we have $\mathcal{H}_a \cap \mathcal{H}_b = \emptyset$. Prior to applying the algorithm, each node computes a p-vector $\mathbf{P} = (P_{a,1}, P_{a,2}, \cdots, P_{a,n_a})$, whose entries correspond to the p-value for the local hypotheses.

The proposed QuTE algorithm operates in three different stages, i.e., Query, Test and Exchange. More specially, in the Query stage, each node queries its neighboring nodes for the p-vectors corresponding to their local hypotheses. Then in the Test stage, each node performs the BH procedure on the set of p-values after the query (denoted as $\mathcal{S}_a$) at level $\alpha_a := \alpha\frac{|S_a|}{N}$. Denote $\widehat{k}^a$ as the number of rejections made in this stage by node a. In the last Exchange stage, all nodes exchange their own rejection decisions with the neighboring nodes. Hypotheses rejected by any node is a valid rejection of the QuTE algorithm. As a result, the overall FDR is controlled as level $\alpha\frac{|\mathcal{H}_0|}{N}$. The extended version of the algorithm includes multiple rounds of communication ($c \geq 1$ rounds), which is indeed the same as the single-round version except for the enlarged communication range, i.e., each node can communicate with all other nodes within distance $d = c$. We use the notation d-Neighbor(a) to denote the set of nodes which distance d of node a.

In general, each node can only reach its neighboring nodes and therefore obtain partial information about the overall system.Sort of counter-intuitive, under the setting that each node can only perform local BH procedure, the overall FDR can be controlled in a similar way as the centralized FDR control. The connectivity (the number of edges on a given vertex set) of the graph has a significant impact on the QuTE algorithm. The QuTE algorithm reduces to the very conservative Bonferroni test when the graph is empty (no edge) given that each nodes tests exactly one hypothesis. This is because no nodes are able to communicate with other nodes and hence the only information available is their local p-values. For a complete graph, the QuTE algorithm becomes the centralized FDR control since all nodes can reach any other node in the network and thus obtain all the p-values of the system.

II. Main Result

The main result of this paper is stated as follows.

Theorem 1: *Suppose that p-values are independent, or positively dependent. Then for any graph topology, the QuTE algorithm achieves FDR control at level $\alpha\frac{|\mathcal{H}^0|}{N}$.*

Proof: First we denote $\widehat{k}^{(a)} = \max\limits_{s\in\text{Neighbor}(a)\cup\{a\}} \widehat{k}^s$, i.e., the maximum of of rejections made among a and its neighbors.

Following the definition of FDR, we have

$$
\begin{aligned}
FDR &= \mathbb{E}\left[\frac{V}{R}\right] = \mathbb{E}\left[\frac{\sum_{a\in\mathcal{V}}\sum_{j\in\mathcal{H}_a^0}\mathbb{1}\left\{P_{a,j}\le\alpha\frac{\widehat{k}^{(a)}}{N}\right\}}{\widehat{k}^{qute}}\right]\\
&\overset{(1)}{=} \sum_{a\in\mathcal{V}}\sum_{j\in\mathcal{H}_a^0}\mathbb{E}\left[\frac{\mathbb{1}\left\{P_{a,j}\le\alpha\frac{\widehat{k}^{(a)}}{N}\right\}}{\widehat{k}^{qute}}\right]\\
&\overset{(2)}{\le} \sum_{a\in\mathcal{V}}\sum_{j\in\mathcal{H}_a^0}\mathbb{E}\left[\frac{\mathbb{1}\left\{P_{a,j}\le\alpha\frac{\widehat{k}^{(a)}}{N}\right\}}{\widehat{k}^{(a)}}\right]\\
&\overset{(3)}{=} \sum_{a\in\mathcal{V}}\sum_{j\in\mathcal{H}_a^0}\frac{\alpha}{N}\mathbb{E}\left[\frac{\mathbb{1}\left\{P_{a,j}\le\alpha\frac{\widehat{k}^{(a)}}{N}\right\}}{\alpha\frac{\widehat{k}^{(a)}}{N}}\right]\\
&\overset{(4)}{\le} \sum_{a\in\mathcal{V}}\sum_{j\in\mathcal{H}_a^0}\frac{\alpha}{N}.1\\
&= \sum_{a\in\mathcal{V}}\alpha\frac{|\mathcal{H}_a^0|}{N}\\
&\overset{(5)}{=} \alpha\frac{|\mathcal{H}^0|}{N}
\end{aligned}
$$

where in step (2), we used the fact that $\widehat{k}^{(a)}\le\widehat{k}^{qute},\forall a\in\mathcal{V}$, which is straightforward since the total number of discoveries of the whole system can never be smaller than the number of rejections made by any single node. From (3) to (4), we used the following corollary in [2].

Corollary: *Let P_i be the true null, satisfying the superunifromity assumption (including the uniform case), and assume that the p-vector P is PRDS (also including the independent case) w.r.t. P_i. Then, for any non-increasing function $f:[0,1]^n\to[0,\infty)$, (w.r.t to the orthant ordering), we have*

$$\mathbb{E}\left[\frac{P_i\le f(P)}{f(P)}\right]\le 1 \tag{1}$$

Proof: Refer to [2] for the details of the proof.

Define the function $f(P)=\alpha\frac{\widehat{k}^{(a)}}{N}$ for node a. $f(P)$ is non-increasing since reducing the p-vector can only possibly increase $\widehat{k}^{(a)}$, the number of rejections made by the BH procedure at node a. Then combined with the above corollary, we justify the inequality from (3) to (4). Hence, the proof of Theorem 1 is complete. Since $|\mathcal{H}^0|\le N$, we see that $\alpha\frac{|\mathcal{H}^0|}{N}\le\alpha$, which means that the overall FDR is controlled at least at the predefined level α.

Theorem 2: *If the p-values are independent, or positively dependent, then the multi-step QuTE algorithm with $c\ge1$ rounds of communication guarantees that* $\mathrm{FDR}\le\alpha\frac{|\mathcal{H}^0|}{N}$.

Proof: The proof of this theorem is very similar to Theorem 1, except that we we redefine $\widehat{k}^{(a)}=\max\limits_{s\in c-\text{Neighbor}(a)\cup\{a\}}\widehat{k}^{s}$, i.e., the maximum of of rejections made among a and nodes within distance c from node a.

III. Discussion

This paper provides a framework for implementing distributed FDR control by performing local BH procedure. However, some of the assumptions involved is not practical, which put a limit on the application of the QuTE algorithm. The main practical issue is that the sensors often have a limited computation ability and repetition of computations are not acceptable. Here is a list of possible directions where potential improvements can be made.

A. Imperfect p-value

It is assumed in the paper that the p-values are real numbers with unlimited accuracy, which is not feasible for practical settings. For example, in most wireless sensor networks, sensors are equipped with limited power supply, which puts a threshold on the computation ability and communication range of these sensors. Hence, p-values must be quantized. *How many bits are necessary to guarantee overall FDR control?* It is an interesting topic to characterize the effect of quantization error on the FDR control performance. It is anticipated that there will be an optimal tradeoff between quantization level (accuracy) and FDR control. With small enough quantization level, the overall FDR can possibly be controlled using the QuTE algorithm with high probability. However, for limited quantization accuracy, whether the FDR can be controlled is still not clear.

B. Computation Redundancy

An easy observation is that in a complete graph, all the agents are doing BH procedure over the same set of p-values. That is, the same computation is performed multiple times, which is highly redundant and will result in unnecessary power consumption. One possible way to resolve this issue is to do *sampling*. More specifically, we can select some certain agents (not all) to perform local BH procedure and other agents will not. The sampling method may depends on the *topology of the graph* ans also the *distribution of the hypotheses* n_a such that $\sum_{a \in \mathcal{V}} n_a = N$. The general principle is to let nodes which are able to gather rich information from other nodes and also nodes with large n_a to perform the local BH procedure. For example, in the star-shaped graph, it is sufficient to only let the center node perform BH procedure since it has all the N p-values of the system.

C. Robustness And Temporal Effect

We may encounter cases where some of the nodes in the system fail and are not able to perform the local BH procedure, or the communication between nodes, either in the Query stage or the Exchange stage, fails. In this case, some of the requested p-values are are not available to the desired nodes. These node/communication failure will make the FDR control more conservative since nodes can access less information about the system. It will be an interesting topic to address the effect of stragglers and try to avoid it. Recall that in the paper it is assumed all nodes are responsible for disjoint sets of hypotheses, which means no two nodes will test the same hypothesis. *However, we can add some redundancy to mitigate the effect of failing nodes*. If we treat the assignment of the set of N hypotheses as design variables, i.e., $\mathcal{H} = \cup_{a \in \mathcal{V}} \mathcal{H}_a$, and allow each hypothesis to be tested multiple times at distinct nodes, then it is possible to mitigate the effect of failing nodes at the cost of increased computation. The *computation redundancy metric* may be defined as $r \triangleq \frac{\sum_{a \in \mathcal{V}} |\mathcal{H}_a|}{N} \in \mathbb{Z}^+$, meaning that each hypothesis is on average tested at r different nodes. Then we may find the tradeoff between computation redundancy r and the maximum number of failing nodes such that the overall FDR is controlled at a predefined level. A lot of results in *error control coding* can be potentially introduced and applied.

Also, in the studied paper it is assumed that the system is static, which means that the network topology, p-values do not change with time. In practical settings, FDR may need to be controlled over time. The topology of the network, wireless channel quality, and even the assignment of hypotheses can vary with time. Hence, it is relevant to study the distributed FDR control under such circumstances.

References

[1] Ramdas, Aaditya, et al. ”QuTE: Decentralized multiple testing on sensor networks with false discovery rate control.” Decision and Control (CDC), 2017 IEEE 56th Annual Conference on. IEEE, 2017.

[2] Barber, Rina Foygel, and Aaditya Ramdas. ”The p-filter: multilayer false discovery rate control for grouped hypotheses.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79.4 (2017): 1247-1268.

[3] Blanchard, Gilles, and Etienne Roquain. ”Two simple sufficient conditions for FDR control.” Electronic journal of Statistics 2 (2008): 963-992.