Advances in random forest tuning and improvements in false discovery rate controlling procedures via test-specific covariate adjustments

by

Hyeongseon Jeon

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Dan Nettleton, Major Professor
Peng Liu
Dan Nordman
Lily Wang
Huaiqing Wu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

DEDICATION

This dissertation is dedicated to my one and only friend Yanghyeon Cho and my parents, Bunsaeng Kwon and Kwangjun Jeon, for their unrepayable love and support. Through the difficulties of my doctoral studies, Yanghyeon Cho constantly listened to my research concerns and encouraged me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I want to express my gratitude to everyone who positively influenced my life during my doctoral studies. Before anything else, I would like to thank my advisor, Dan Nettleton. He gave me ownership over the topic and direction of my research. Combined with my stubborn nature, I have experienced many obstacles reaching this milestone. However, through such experiences, I learned the research direction I should not go, and I could have much better statistical intuition. I believe that he has nourished my growth as an independent researcher the most.

The following individuals made indirect, but significant, contributions to my doctoral studies. Huaiqing Wu, one of my committee members, made me realize that humanity is also important as a researcher. His smiles and greetings literally kept me alive. The most valuable piece of advice that changed my perspective was from Junho Lee, who advised me to prioritize my studies over interpersonal relationships. After receiving his advice, I could finally live my own life. Through weekly discussions with Kyusang Lim, I built the capacity for critical statistical reasoning. As a result of vigorous discussions with him, I obtained intuitions for methods related to gene expression data analysis.

I want to thank all of the committee members, Peng Liu, Dan Nordman, Lily Wang, and Huaiqing Wu, for their insightful questions during my oral preliminary examination. I would also like to emphasize that Jack Dekker's thoughts and intuitions have substantially impacted my study. Like all the researchers I admired during my doctoral program, I will strive to serve as a role model for future researchers.

Before concluding this acknowledgment, I would like to note that the Laurence H. Baker Endowment and the Plant Science Institute at Iowa State University have provided me with consistent financial support. Without the support and care, I would not have recovered from various hardships and completed my studies. Iowa State University provided me with a large

amount of support and endless opportunity, for which I am grateful and hope to give back one day. Ames, Iowa will remain one of my fond memories to remember at the end of my life's journey.

ABSTRACT

Each chapter of this dissertation is devoted to one of three topics. The first two are novel false discovery rate (FDR) controlling methods in different situations, and the third deals with a new tuning parameter selection approach for the random forest method in regression problems.

The primary research of this dissertation is to develop methods for controlling FDR while conducting multiple hypothesis tests with gene expression data. The first topic of this dissertation is a gene-specific covariate-based FDR-controlling method. We propose gene length as a potential gene-specific covariate. We develop a method based on covariate-specific conditional null probability for promising hypotheses with low p-values. We prove that the method controls positive FDR (pFDR) and provide an equivalent statement producing the method's rejection rule. Simulations demonstrate our method controls over pFDR, and the suggested method is better than existing methods in terms of true positive rate and summary statistics for the receiver operating characteristic (ROC) curve. Using data provided by Dr. Lim, we observe that our method rejects more null hypotheses at most target levels than existing methods.

Another topic of this dissertation is developing an FDR-controlling method for circumstances where data are obtained from the pilot and main studies. We assume each study has unique properties such as sample size and error variance. Our method's rejection rule permits a higher p-value rejection threshold for the main study when the p-value for the pilot study is relatively low. This relationship enables us to evaluate fewer rejection rules than a competing method, resulting in more inference power. Our simulation study demonstrates our approach for combining results from two studies is superior to existing methods and controls FDR to a predetermined level. The number of rejected null hypotheses in the data analysis was greater than that of competing methods.

The last topic of the dissertation is a unique tuning approach for random forest (RF) regression. We propose a case-specific tuning strategy for selecting the RF tuning parameter values of mtry and nodesize. We provide an example showing case-specific tuning parameters can be useful by demonstrating that the best choice for tuning parameter values varies across the predictor space. The tuning algorithm is then outlined mathematically. In a simulation study, our approach outperforms the conventional algorithms implemented in various `R` packages to minimize mean squared prediction error. Moreover, this method outperforms competing methods for the majority of the datasets we examined.

CHAPTER 1. GENERAL INTRODUCTION

1.1 RNA-seq differential expression analysis adjusting for relevant gene-specific covariates

Gene expression, which is an abundance of mRNA transcript, is quantified by RNA profiling techniques. The profiling technologies such as RNA-seq and microarray are detailed in Mantione et al. (2014) and Lowe et al. (2017). A typical research question concerning gene expression data is identifying differentially expressed (DE) genes with intriguing features. Genes with opposite characteristics are called equivalently expressed (EE) genes. In general, DE genes are identified by conducting hypothesis tests for thousands of genes. In this circumstance involving multiple tests, statisticians have developed approaches based on different error quantities to boost the inference power. The error quantities include familywise type 1 error (FWER), defined as the probability of producing one or more false positives. Bonferroni correction and Holm's (1979) method are used to control the FWER. Regarding gene expression data, the false discovery rate (FDR) introduced by Benjamini and Hochberg (1995) is the often applied error measure, and Storey's (2002) q-value method is the most commonly employed approach to control the FDR. Modern FDR-controlling approaches have attempted to increase testing power by using appropriate covariate variables [Korthauer et al. (2019)]. The modern approaches include methods such as Cai and Sun (2009), Scott et al. (2015), Ignatiadis et al. (2016), Boca and Leek (2018), and Lei and Fithian (2018). Under this contemporary tendency, two novel approaches adjusting for appropriate covariates are proposed in Chapters 2 and 3.

A biological discovery in Lopes et al. (2021) about the relationship between biological timing and gene length inspired the development of the positive FDR-controlling method presented in Chapter 2. Shorter genes tend to regulate immediate biological processes such as skin recovery, whereas longer genes tend to regulate long-term biological processes such as muscle development

[Lopes et al. (2021)]. Depending on the treatment factor, the fraction of DE or EE genes may vary by gene length. In the Bayesian perspective, EE gene probability (i.e., null probability) may vary with gene length. We suggest a rejection rule that accounts for heterogeneity among tests resulting from their distinct null probabilities. Our approach gives a higher p-value rejection threshold for genes with a low conditional null probability, allowing us to focus more on promising hypotheses. Lei and Fithian's (2018) method enhances inference power by focusing on promising hypotheses using adaptively established p-value rejection thresholds. In addition, we provide a positive FDR estimator and identify a condition equivalently determining our rejection rule under a reasonable model assumption.

The second FDR-controlling method was inspired by the independent hypothesis weighting (IHW) method [Ignatiadis et al. (2016)]. IHW considers all possible rejection rules defined by groups specified by a covariable variable. If the considered rejection rules can be reduced reasonably, the inference power may be increased. From such intuition, we developed a method to infer DE genes, using the p-values obtained from a pilot study and the main study. Our method allows a higher p-value rejection threshold for the main study when the p-value for the pilot study is relatively small. In this context, we consider the pilot study's p-value as a covariate. Compared to the IHW method, we can considerably reduce the number of examined rejection rules from the negative relationship between the pilot study p-value and the main study p-value rejection threshold used by out method. However, since many rejection rules are still evaluated, calibration is necessary for computing FDR, and we provided a calibration method.

1.2 Random forest regression tuning algorithm

The random forest (RF) method developed by Breiman (2001) is a popular machine learning algorithm for predictions. Lin and Jeon (2006) established the perspective of viewing the RF method as an adaptive nearest-neighbors algorithm. The forest weight, a byproduct of the RF method, has been used in numerous purposes. For example, see Meinshausen (2006). Xu et al. (2016), Zhang et al. (2019), and Friedbergå et al. (2020). Cross-validation is frequently applied to

assess prediction errors in machine learning literature. Instead, out-of-bag (OOB) prediction error can be used to evaluate prediction error when utilizing bootstrap aggregation methods such as RF. OOB prediction is the mean prediction on each training sample using just the trees that did not include the sample in their bootstrap sample, and the OOB prediction error is the error associated with the OOB prediction. The standard tuning parameter selection strategy of the RF method uses the parameter values that minimize the average squared OOB prediction error.

In the third topic of this dissertation, which is presented in Chapter 4, we argue that the best choice for values of the tuning parameters may depend on the target value of the predictor vector at which a prediction is desired. According to this perspective, using a case-specific tuning parameter for each predictor vector value may lower the overall prediction error. Using the proximity weight, we can evaluate the predictor vectors close to the target predictor vector. At the same time, each training case's prediction error can be evaluated via its OOB prediction error. In this regard, we suggest a case-specific tuning algorithm based on weighted average squared OOB prediction error using proximity weight.

1.3 Dissertation Structure

Following this general introduction, the structure of this dissertation is as follows. We adhere to the journal format through Chapters 2, 3, and 4. Each of the three chapters contains the introduction, method proposal, simulation study, data analysis, and discussion sections. Chapters 2 and 3 present new methods for analyzing RNA-seq data. The second chapter discusses a rejection rule employing covariate-specific conditional null probability. The third chapter discusses a novel rejection rule that uses the p-value as a covariate. The fourth chapter discusses a tuning parameter selection strategy for the RF regression method. The concluding chapter provides a concise overview of the entire dissertation and future research.

1.4 References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Boca, S. M. and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6:e6035.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481.

Friedbergå, R., Tibshirani, J., Athey, S., and Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):1–21.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Lopes, I., Altab, G., Raina, P., and De Magalhaes, J. P. (2021). Gene size matters: An analysis of gene length in the human genome. *Frontiers in Genetics*, 12:30.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS computational biology*, 13(5):e1005457.

Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., and Stefano, G. B. (2014). Comparing bioinformatic gene expression profiling methods: microarray and rna-seq. *Medical science monitor basic research*, 20:138.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Xu, R., Nettleton, D., and Nordman, D. J. (2016). Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25(1):49–65.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*.

CHAPTER 2. ADJUSTING FOR GENE-SPECIFIC COVARIATES TO IMPROVE REJECTION RULES IN RNA-SEQ ANALYSIS

Hyeongseon Jeon[1], Dan Nettleton[1], and Kyu-Sang Lim[2]

[1]Department of Statistics, Iowa State University, Ames, IA 50011, USA

[2]Department of Animal Resources Science, Kongju National University, Yesan-gun, Chungnam 32439, Republic of Korea

Modified from a manuscript to be submitted to *Bioinformatics*

Abstract

This paper suggests a novel positive false discovery rate (pFDR) controlling method using a gene-specific covariate variable, such as gene length. We suppose the null probability depends on the covariate variable. In this context, we propose a rejection rule that accounts for heterogeneity among promising tests with low p-values, while accounting for different null probabilities. We establish a pFDR estimator for a given rejection rule by following Storey's q-value framework. A condition on a type 1 error posterior probability is provided that equivalently characterizes our rejection rule. We also present a suitable procedure for selecting a tuning parameter through cross-validation that maximizes the expected number of hypotheses declared significant. A simulation study demonstrates that our method is comparable to or better than existing methods across a variety of realistic scenarios. In data analysis, we find support for our method's premise that the null probability varies with a gene-specific covariate variable.

2.1 Introduction

Gene expression refers to an abundance of messenger RNA transcripts quantified by RNA profiling techniques like microarray and RNA-seq. The invention of the sequencing technique

RNA-seq enables researchers to profile nearly all genes in an organism simultaneously. At the same time, cost decreases in RNA profiling techniques have led to an increase in the volume of gene expression data. The accumulated knowledge on gene expression data has been made available in public databases, and open-source software or computer packages enable scientists to quickly access the information. As knowledge has increased, statistical methods that efficiently select genes with interesting expression patterns have received more attention. The research question involving gene selection often focuses on identifying genes differentially expressed (DE) across different experimental conditions. Genes other than DE genes are called equally or equivalently expressed (EE) genes.

DE genes are typically identified through hypothesis testing on each gene in a statistical framework, viewed as a multiple testing problem. Under the multiple testing framework, the classic rejection rule $P \leq 0.05$ generally generates too many false positives (type I errors). Therefore, statisticians have suppressed false positives by controlling different error quantities. When dealing with gene expression data, the most useful error quantity is typically the false discovery rate (FDR), introduced by Benjamini and Hochberg (1995). FDR refers to the expected proportion of false positives among all tests whose null hypotheses have been rejected. The most widely used procedure is Storey's (2002) q-value method. Storey's (2002) standard q-value rejection rule is based on null probability π_0, the probability that a tested null hypothesis is true. Typically, π_0 is unknown and needs to be estimated [Liang and Nettleton (2012), Nettleton et al. (2006), Storey (2002), Storey et al. (2004)].

Contemporary methods for FDR control are based on gene-specific covariate variables such as mean non-zero expression and the proportion of samples with the non-zero expression [Korthauer et al. (2019)]. The covariate variables are used for grouping or modeling purposes. As circumstances vary across hypothesis tests, it is vital to consider each test separately. Efron et al. (2001) proposed local FDR adjusting for different features by employing true null probability and density functions under null and alternatives. Sun and Cai (2007) developed a method for estimating global FDR through the local FDR. Subsequently, Cai and Sun (2009) developed an

FDR-controlling method using external grouping information. An FDR regression method proposed by Scott et al. (2015) regulates FDR by utilizing the local FDR and treating the null probability as a function of covariate variables. Boca and Leek (2018) also proposed a method (BL), considering the FDR and null probability as functions of a covariate variable. Ignatiadis et al. (2016) proposed an independent hypothesis weighting method (IHW) which maximizes the number of rejected null hypotheses, based on covariate-variable-based groups. Recently, Lei and Fithian (2018) developed a covariate-specific p-value thresholding method (AdaPT), based on adaptively determined significance thresholds and the local FDR.

The AdaPT method has developed into a powerful approach that is expected to yield more discoveries by focusing on promising hypotheses and utilizing adaptively defined p-value rejection thresholds. Initially, the method establishes a constant threshold across all covariate values. The initial threshold is updated continuously to gradually increase rejection power. As a result of considering multiple thresholds, we predict that the method's average ability to classify the true positives across all nominal FDR levels may deteriorate. Simultaneously, adaptively determined thresholds complicate FDR estimation.

This paper presents a novel and more straightforward rejection rule that accounts for the heterogeneity between promising hypotheses with low p-values determined by the classic rejection rule $P \leq \alpha$. To be more precise, our rejection rule is based on the product of the p-value and covariate-specific conditional null probability, given the p-value is no larger than α. Due to the simplicity of the rejection rule, Storey's (2002) positive FDR (pFDR) is naturally estimated. Because pFDR provides an upper bound on FDR, pFDR control implies FDR control. We demonstrate that the rejection rule is uniquely determined by a property of equalizing type 1 error posterior probabilities among all tests with p-values no larger than α and any given covariate value.

New biological discoveries are good motivating sources to develop new statistical methods. Recently, it was discovered that there exist relationships between biological timing and gene length: shorter genes tend to regulate immediate physical processes such as skin recovery, whereas

longer genes tend to regulate long-term physical processes such as muscle development [Lopes et al. (2021)]. Thus, the fraction of DE or EE genes may vary by gene length depending on the experimental conditions studied. From a Bayesian perspective, the null probability may vary by gene length. Because of this heterogeneity, we consider gene length as a covariate variable potentially important to consider when identifying DE genes. Though we focus exclusively on gene length in this paper, our approach is applicable for any gene-specific covariate.

The remainder of this paper is organized as follows. In Section 2.2, we define our method in detail and argue its mathematical implications in terms of posterior probability. In Section 2.3, we demonstrate the effectiveness of the method through simulation studies. In Section 2.4, we illustrate our method's efficacy through data analysis. Lastly, Section 2.5 evaluates the proposed method's potential for further development.

2.2 Method Proposal

Our research objective is to declare genes to be DE while controlling pFDR in the multiple testing framework. Our method is inspired by Storey's (2002) q-value method based on the Bayesian perspective. From the Bayesian perspective, two types of conditional prior probabilities of being an EE gene, also referred to as conditional null probabilities, are considered. In our work, both conditional null probabilities are considered as functions of a covariate variable. Section 2.2.1 presents a rejection rule based on a conditional null probability. By inverting the rejection rule, its rejection region is naturally determined in Section 2.2.2. This rejection region is needed for estimating the pFDR associated with the rejection rule. In Section 2.2.3, we establish the pFDR estimator based on another conditional null probability through mathematical reasoning. At the same time, the q-value estimator is obtained. Section 2.2.4 describes a procedure for estimating the conditional null probabilities, which serves as the foundation for our method. Section 2.2.5 delves into the rejection rule's intrinsic meaning regarding posterior probability.

2.2.1 Rejection Rule

Our rejection rule is based on the premise that a p-value rejection threshold should be negatively associated with null probability. Furthermore, we assume that null probability may be associated with a gene-specific covariate. This assumption is reasonable given the change in the fraction of DE genes with gene length discussed in the previous section. We also believe that the association between null probability and the covariate may be most relevant among tests with low p-values. This paper, therefore, presents a rejection rule based on the conditional null probability, given the covariate and a low p-value event.

Consider hypothesis testing for each of m genes. For gene $j \in \{1, \ldots, m\}$, let X_j and P_j denote the value of a covariate and the p-value, respectively. Let H_{0j} denote the event that gene j is an EE gene. Let

$$\pi_0(X_j) = \mathbb{P}(H_{0j} \mid X_j), \text{ and} \tag{2.1}$$

$$\pi_{0|\alpha}(X_j) = \mathbb{P}(H_{0j} \mid P_j \leq \alpha, X_j). \tag{2.2}$$

Expressions (2.1) and (2.2) are conditional probabilities of gene j being an EE gene. These conditional null probabilities, as stated previously, are functions of the covariate value X_j. Furthermore, (2.2) is the conditional null probability conditioning on the classic rejection rule $P \leq \alpha$. It is worth noting that α can either be specified as a value or selected via a procedure, as described in Section 2.3. Define the jth $\tilde{p}$-value as $\tilde{P}_j = P_j \cdot \pi_{0|\alpha}(X_j)$. Based on the configurations stated so far, the following rejection rule is proposed:

Rejection Rule 2.2.1. Reject all null hypotheses whose $\tilde{p}$-value is less than or equal to t, for some $t > 0$.

The genes declared to be DE (DDE) following Rejection Rule 2.2.1 are naturally determined by $\{j : \tilde{P}_j \leq t\}$. Under the rejection rule, both the p-value and the conditional null probability in (2.2) affect the rejection decision for each hypothesis test. Note that we initially assume that $\pi_0(\cdot)$ and $\pi_{0|\alpha}(\cdot)$ are known and then replace these functions with estimates discussed in Section 2.2.4. Section 2.2.5 discusses the rejection rule's intrinsic meaning.

2.2.2 Rejection Region

Considering the rejection region associated with a rejection rule is useful for estimating the pFDR and for gaining a better understanding of the rule. By inverting the rejection rule, the rejection region for the p-value of the jth gene can be obtained as follows:

$$\begin{aligned}\Gamma_{X_j}(t) &= \left\{p \in [0,1] : p \cdot \pi_{0|\alpha}(X_j) \leq t\right\} \\ &= [0, u_t(X_j)],\end{aligned} \tag{2.3}$$

where $u_t(X_j) = 1$ if $\pi_{0|\alpha}(X_j) \leq t$ and $u_t(X_j) = \frac{t}{\pi_{0|\alpha}(X_j)}$ otherwise. Note that

$$\tilde{P}_j \leq t \iff P_j \in \Gamma_{X_j}(t) \iff P_j \leq u_t(X_j). \tag{2.4}$$

Figure 2.1.B illustrates how the rejection region's upper bound varies as a function of x for various t-values for the arbitrarily chosen $\pi_{0|\alpha}(x)$ in Figure 2.1.A. Additionally, Figure 2.1.B demonstrates that genes with relatively high p-values may, nonetheless, be declared to be DE genes when their conditional null probabilities are low. The phenomenon is noticeable when x is between 2 and 3.

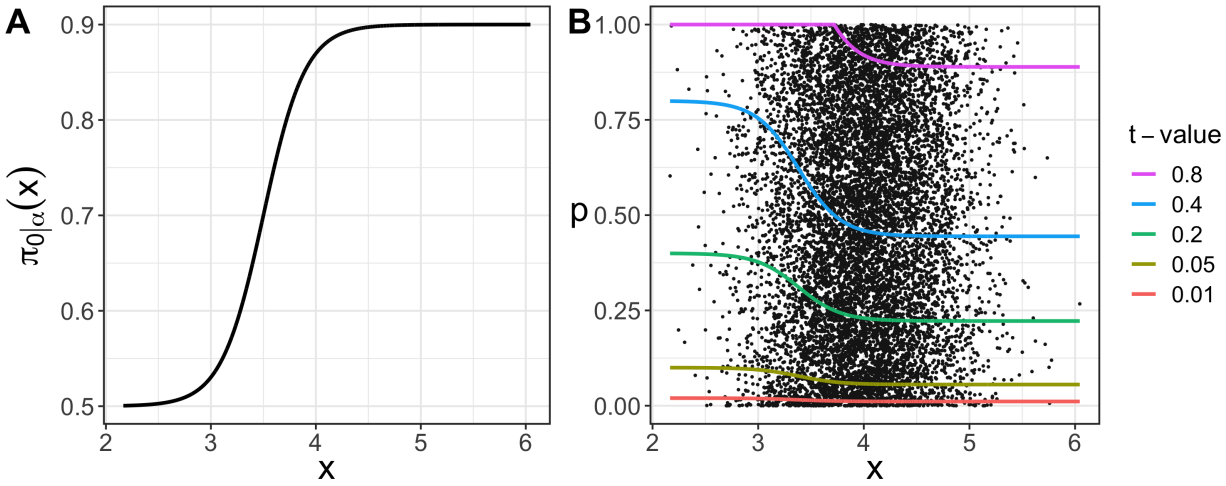


Figure 2.1: An example function $\pi_{0|\alpha}(x)$ is depicted in Figure A, and the rejection regions' upper bounds created by five distinct t-values are illustrated in Figure B.

2.2.3 False Discovery Rate Estimator

For a given $\tilde{p}$-value significance threshold t, the number of genes declared to be DE is

$$R(t) = \sum_{j=1}^{m} 1\big(\tilde{P}_j \le t\big). \tag{2.5}$$

The number of false positives among the $R(t)$ genes can be expressed by

$$V(t) = \sum_{j=1}^{m} V_j(t), \text{ where } V_j(t) = 1\big(\tilde{P}_j \le t,\, H_{0j}\big). \tag{2.6}$$

Clearly, $V(t) \le R(t)$. From the equivalence (2.4), $V_j(t)$ has another expression:

$$V_j(t) = 1\Big\{P_j \le u_t(X_j),\, H_{0j}\Big\}. \tag{2.7}$$

FDR can be expressed as $\text{FDR}(t) = \mathbb{E}\Big\{\frac{V(t)}{R(t)\vee 1}\Big\}$. Our proposed method is based on pFDR defined as $\text{pFDR}(t) = \mathbb{E}\Big\{\frac{V(t)}{R(t)} \,\Big|\, R(t) > 0\Big\}$. When $\mathbb{P}(R(t) > 0)$ is close to 1, $\text{FDR}(t)$ and $\text{pFDR}(t)$ are nearly identical. For a generalized significance region $\tilde{\Gamma}$ of $\tilde{P}_j$, $V(\tilde{\Gamma})$ and $R(\tilde{\Gamma})$ can be naturally defined by replacing $\tilde{P}_j \le t$ with $\tilde{P}_j \in \tilde{\Gamma}$ in the definitions (2.5) and (2.6). By using the notations, the positive FDR is defined by $\text{pFDR}(\tilde{\Gamma}) = \mathbb{E}\Big\{\frac{V(\tilde{\Gamma})}{R(\tilde{\Gamma})} \,\Big|\, R(\tilde{\Gamma}) > 0\Big\}$. The following theorem is based on the generalized significance region $\tilde{\Gamma}$.

Theorem 2.2.1. Suppose m identical hypothesis tests are performed with $\tilde{P}_1, \cdots, \tilde{P}_m$ and significance region $\tilde{\Gamma}$. Let $\pi_A(\cdot) = 1 - \pi_0(\cdot)$. Assume that $(P_1, H_1, X_1), \ldots, (P_m, H_m, X_m)$ are i.i.d. random vectors, where $\tilde{P}_j = P_j \cdot \pi_{0|\alpha}(X_j)$, $P_j \mid H_j, X_j \sim (1 - H_j) \cdot F_0 + H_j \cdot F_1$ for some null distribution F_0 and alternative distribution F_1, and $H_j \mid X_j \sim \text{Bern}(\pi_A(X_j))$, $X_j \sim F_X$ for $j = 1, \cdots, m$. Then,

$$\text{pFDR}(\tilde{\Gamma}) = \mathbb{P}(H_j = 0 \mid \tilde{P}_j \in \tilde{\Gamma}) = \frac{\mathbb{E}V(\tilde{\Gamma})}{\mathbb{E}R(\tilde{\Gamma})},\ \forall\, j = 1, \ldots, m. \tag{2.8}$$

Appendix I contains the proof of Theorem 2.2.1, which closely follows an analogous proof in Storey (2003).

Remark 2.2.1. The marginal distribution of H_j in Theorem 2.2.1 is $\text{Bern}(\pi_A)$, where $\pi_A = 1 - \pi_0$ and $\pi_0 = \mathbb{P}(H_j = 0)\ \forall\, j = 1, \ldots, m$. In this perspective, (P_j, H_j) are i.i.d. random

variables, where $P_j \mid H_j \sim (1-H_j)\cdot F_0 + H_j \cdot F_1$ and $H_j \sim \text{Bern}(\pi_A)$. The standard q-value method is established on this modeling setup. Therefore, we can still apply the standard q-value method, while controlling pFDR, to the p-values generated from the model in Theorem 2.2.1.

Theorem 2.2.1 establishes that $\text{pFDR}(t) = \frac{\mathbb{E}V(t)}{\mathbb{E}R(t)}$. Our estimator is obtained by estimating $\mathbb{E}V(t)$ and $\mathbb{E}R(t)$. The denominator $\mathbb{E}R(t)$ can be easily estimated as $R(t)$. However, the number of false positives $V(t)$ is unknown. To estimate the numerator $\mathbb{E}V(t)$, we might consider estimating F_1. Because estimating F_1 is not straightforward, we propose to estimate $\mathbb{E}V(t)$ using $\mathbb{E}\big\{V(t) \mid \vec{X} = (X_1, \cdots, X_m)\big\}$, which is both the best predictor of $V(t)$ under a squared error loss function and an unbiased estimator of $\mathbb{E}V(t)$. Some conditions are required to derive $\mathbb{E}\big\{V(t) \mid \vec{X}\big\}$. Let $\vec{X}_{-j}$ denote a vector $\vec{X}$ without the jth element. Under the model assumption described in Theorem 2.2.1, the following properties are obtained:

$$(P_j, H_j, X_j) \perp \vec{X}_{-j} \quad \rightarrow \quad P_j \mid \vec{X} \overset{d}{=} P_j \mid X_j \tag{2.9}$$

$$(P_j, H_j, X_j) \perp \vec{X}_{-j} \quad \rightarrow \quad H_j \mid \vec{X} \overset{d}{=} H_j \mid X_j \tag{2.10}$$

$$(P_j, H_j, X_j) \perp \vec{X}_{-j} \quad \rightarrow \quad P_j \mid H_j, \vec{X} \overset{d}{=} P_j \mid H_j, X_j \tag{2.11}$$

$$X_j \perp P_j \mid H_j \quad \rightarrow \quad P_j \mid H_j, X_j \overset{d}{=} P_j \mid H_j. \tag{2.12}$$

When the simple null hypothesis is true, and the test statistic is continuous, the p-value follows a uniform distribution between 0 and 1. Motivated by this fact, the following assumption is made:

Assumption 2.2.1. $\text{P}_j \mid H_j = 0 \sim \text{Unif}(0,1)$.

Under properties (2.10) to (2.12) and Assumption 2.2.1, $\mathbb{E}\big\{V(t) \mid \vec{X}\big\}$ can be derived as follows:

$$\begin{aligned}
\mathbb{E}\{V(t) \mid \vec{X}\} &= \sum_{j=1}^{m} \mathbb{E}\{V_j(t) \mid \vec{X}\} \ \because V(t) = \sum_{j=1}^{m} V_j(t) \text{ and linearity} \\
&= \sum_{j=1}^{m} \mathbb{P}\Big\{P_j \le u_t(X_j),\, H_{0j} \,\Big|\, \vec{X}\Big\} \ \because (2.7) \\
&= \sum_{j=1}^{m} \mathbb{P}\Big\{P_j \le u_t(X_j) \,\Big|\, H_{0j},\, \vec{X}\Big\} \cdot \mathbb{P}(H_{0j} \mid \vec{X}) \\
&= \sum_{j=1}^{m} \mathbb{P}\Big\{P_j \le u_t(X_j) \,\Big|\, H_{0j},\, X_j\Big\} \cdot \mathbb{P}(H_{0j} \mid X_j) \ \because (2.10, 2.11) \\
&= \sum_{j=1}^{m} u_t(X_j) \cdot \pi_0(X_j) \because (2.12, \text{Assumption 2.2.1}).
\end{aligned} \tag{2.13}$$

By combining the predetermined form of pFDR(t) and (2.13), the pFDR estimator is established:

$$\widehat{\text{pFDR}}(t) = \frac{\sum_{j=1}^{m} u_t(X_j) \cdot \pi_0(X_j)}{R(t)} \tag{2.14}$$

$$\leq \frac{t}{R(t)} \cdot \sum_{j=1}^{m} \frac{\pi_0(X_j)}{\pi_{0|\alpha}(X_j)}, \tag{2.15}$$

where $\pi_0(\cdot)$ and $\pi_{0|\alpha}(\cdot)$ are considered known. The pFDR estimator (2.15) serves as an upper bound for (2.14), where the equality holds when $\pi_{0|\alpha}(X_j) \geq t$ for all j. We adopt the simpler version (2.15) as our pFDR estimator, which is also used in the simulation study and data analysis, where our results using either (2.14) or (2.15) were nearly identical. Furthermore, we may define q-values that can be utilized easily to declare genes to be DE:

$$Q_j = \min_{t:\, t \geq \tilde{P}_j} \text{pFDR}(t). \tag{2.16}$$

By declaring the genes with the q-value less than or equal to a nominal level γ as DE genes, one can acquire the DDE genes list that controls the pFDR at less than or equal to γ. In addition, the q-value estimator of (2.16) can be obtained by inserting the pFDR estimator (2.15):

$$\hat{Q}_j = \min_{t:\, t \geq \tilde{P}_j} \widehat{pFDR}(t). \tag{2.17}$$

Up to this point, $\pi_0(\cdot)$ and $\pi_{0|\alpha}(\cdot)$ have been treated as given. In practice, we must estimate both conditional null probabilities to apply our method. The following section discusses an estimating procedure.

2.2.4 Estimation of $\pi_0(\cdot)$ and $\pi_{0|\alpha}(\cdot)$

To simplify the problem of estimating $\pi_0(\cdot)$ and $\pi_{0|\alpha}(\cdot)$, we first derive a useful property. Under the model described in Theorem 2.2.1 and Assumption 2.2.1, $\pi_{0|\alpha}(\cdot)$ satisfies

$$\pi_{0|\alpha}(X_j) = \mathbb{P}(H_{0j} \mid P_j \leq \alpha, X_j) = \frac{\mathbb{P}(P_j \leq \alpha \mid H_{0j},\, X_j) \cdot \mathbb{P}(H_{0j} \mid X_j)}{\mathbb{P}(P_j \leq \alpha \mid X_j)}$$

$$= \alpha \cdot \frac{\pi_0(X_j)}{\mathbb{P}(P_j \leq \alpha \mid X_j)} \quad \because (2.12, \text{Assumption } 2.2.1). \tag{2.18}$$

According to equality (2.18), when both $\pi_0(X_j)$ and $\mathbb{P}(P_j \leq \alpha \mid X_j)$ are known, $\pi_{0|\alpha}(X_j)$ can be obtained. Thus, we now discuss how to estimate $\pi_0(X_j)$ and $\mathbb{P}(P_j \leq \alpha \mid X_j)$.

The estimation procedure is based on the idea that the values of a function at two close points may be similar. We suggest using the Euclidean distance between covariate variables to quantify their closeness. Let N_{nh} be a user-selected neighborhood size. Let $N_j \subseteq \{1, \cdots, m\}$ contain the N_{nh} indices corresponding to the N_{nh} genes whose covariate values are closest to X_j. Both probabilities are estimated using only the neighborhood p-values $\{P_i : i \in N_j\}$. First, $\pi_0(X_j)$ is estimated using the method of Nettleton et al. (2006) applied to $\{P_i : i \in N_j\}$, which gives

$$\hat{\pi}_0(X_j) = \frac{\sum_{i \in N_j} 1\big(P_i \geq P_{cut,j}\big)}{N_{nh}} \cdot \frac{1}{1 - P_{cut,j}}, \tag{2.19}$$

where $P_{cut,j}$ is a threshold determined by Nettleton et al. (2006) such that the empirial distribution of $\{P_i : i \in N_j,\, P_i \geq P_{cut,j}\}$ is approximately uniform. See Nettleton et al. (2006) for the details.

Next, $\mathbb{P}(P_j \leq \alpha \mid X_j)$ can be easily estimated as the proportion of the p-values in $\{P_i : i \in N_j\}$ less than or equal to α:

$$\hat{\mathbb{P}}(P_j \leq \alpha \mid X_j) = \frac{\sum_{i \in N_j} 1\big(P_i \leq \alpha\big)}{N_{nh}}. \tag{2.20}$$

By (2.18), a natural estimator of $\pi_{0|\alpha}(X_j)$ is $\hat{\pi}_{0|\alpha}(X_j) = 1 \wedge \left\{\alpha \cdot \frac{\hat{\pi}_0(X_j)}{\hat{\mathbb{P}}(P_j \leq \alpha | X_j)}\right\}$. As a result, all necessary components for our method are obtained. The following Section 2.2.5 provides an in-depth discussion of the rejection rule.

2.2.5 Implications of the Rejection Rule

To better understand our rejection rule, we derive an equivalent condition characterizing the rejection rule in terms of a conditional type 1 error posterior probability, as specified in the following theorem.

Theorem 2.2.2. Consider the same inference setup described in Theorem 2.2.1 with a rejection rule $P_j \leq u(X_j)$, for a given non-negative function $u(\cdot)$. Assume that Assumption 2.2.1 holds. Let

T_{1j} be the event that a type 1 error occurs for test j. If the rejection rule is more conservative than the classic rejection rule $P_j \leq \alpha$, that is, $\max_j u(X_j) \leq \alpha$, then,

$$\mathbb{P}(\mathrm{T}_{1j} \mid P_j \leq \alpha, \vec{X}) \text{ is the same for all } j = 1, \ldots, m \tag{2.21}$$

$$\iff u(X_j) = \frac{t}{\pi_{0|\alpha}(X_j)} \text{ for all } j = 1, \ldots, m \text{ and some } t > 0. \tag{2.22}$$

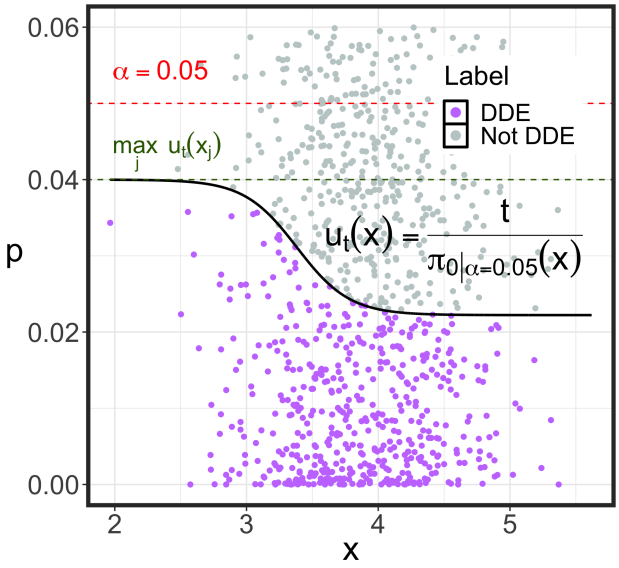


Figure 2.2: A scatterplot illustrating upper bounds satisfying the conservativeness condition stated in Theorem 2.2.2 for a given $\alpha = 0.05$. The green horizontal line is below the red horizontal line indicating that the condition is satisfied. Because $\max_j u_t(X_j) \leq 0.05$, $u_t(x)$ is simplified to $\frac{t}{\pi_{0|\alpha}(x)}$.

The proof is included in Appendix II. According to Theorem 2.2.2, among more conservative rejection rules than the classic rejection rule, referred to as a conservativeness condition, the rejection rule that preserves constant type 1 error posterior probability given the low p-value condition and covariate variables $\vec{X}$ is uniquely determined by $u(X_j) = \frac{t}{\pi_{0|\alpha}(X_j)}$ for some t. Under the conservativeness condition $u(X_j) \leq \alpha$, $u(X_j)$ is identical to our proposed rejection rule's upper bound $u_t(X_j)$. In other words, under the conservativeness condition, our proposed rejection rule is the only one that equalizes the conditional type 1 error posterior probability across all tests. For clarity, $u_t(X_j)$ now refers to $u(X_j)$.

According to the model assumed in Theorem 2.2.2, rejection situations vary by covariate variables. A rejection rule ignoring the distinct situations is incapable of equalizing error control as described in Theorem 2.2.2. According to Theorem 2.2.2, however, our rejection rule ensures constant conditional type 1 error posterior probabilities across all tests. This is in contrast to traditional rejection rules, which provide conditional type 1 error probabilities that vary across tests.

Lastly, let us discuss the conservativeness condition $\max_j u_t(X_j) \leq \alpha$ in Theorem 2.2.2. The condition indicates that the rejection region's upper bound is less than or equal to α. In other words, rejections of null hypotheses occur only when the p-values are less than or equal to α. When α is set to 0.05, Figure 2.2 illustrates the condition visually. If α is not chosen too small, the classic rejection rule is fairly liberal, and the conservativeness condition of Theorem 2.2.2 is easily satisfied in practice. Therefore, we can conclude that Theorem 2.2.2 is stated for a reasonably confined condition on t.

2.3 Simulation Study

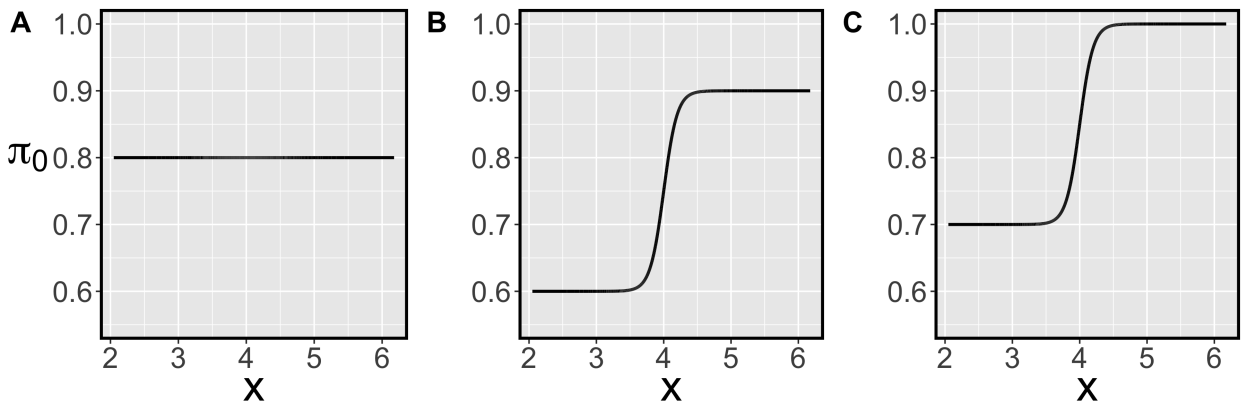


Figure 2.3: Functions from A to C illustrate three $\pi_0(\cdot)$ functions used in the simulation, where $\pi_0^A(x) = 0.8$, $\pi_0^B(x) = 0.6 + \frac{0.3}{1+\exp\{-10\cdot(x-4)\}}$, and $\pi_0^C(x) = 0.7 + \frac{0.3}{1+\exp\{-10\cdot(x-4)\}}$.

2.3.1 Model Description

We conduct a simulation study to assess our method's performance. The simulation's model is inspired by the model in Theorem 2.2.1. Essentially, we consider gene expression data sets with $m = 10,000$ genes generated independently from normal distributions with gene-specific standard deviation from an inverse chi-square distribution. The log-transformed gene length denoted by X, which affects the probability of being an EE gene or equivalently DE gene, is assumed to be normally distributed. If a gene is chosen as a DE gene, determined by $\pi_0(\cdot)$, the treatment effect is randomly generated from a normal distribution. Let j and k be the indices for the gene and treatment group, respectively. Let s denote a sample index within a treatment group. The sample size within a treatment group n is set to 10. For jth gene, the data model with Y_{jks} as the response variable is described as follows. Note that independence holds unless otherwise specified.

$$\begin{aligned}
&Y_{jks} \sim N(\delta_{jk},\ \sigma_j^2),\\
&\delta_{j0} = 0 \text{ and } \delta_{j1} \mid H_j = (1 - H_j) \cdot 0 + H_j \cdot N(\mu_\delta,\ \sigma_\delta^2 = 0.02^2),\\
&H_j \mid X_j \sim \text{Bern}(1 - \pi_0(X_j)),\\
&X_j \sim N(\mu_X = 4,\ \sigma_X^2 = 0.5^2), \text{ and } \sigma_j \sim \text{Inv-}\chi_5^2.
\end{aligned} \tag{2.23}$$

After generating the data set from (2.23), a two-sample t-test is used to obtain a p-value for testing each gene's treatment effect.

The simulation is conducted with different combinations of μ_δ and $\pi_0(\cdot)$. μ_δ is chosen from a set of four equally spaced values. As illustrated in Figure 2.3, three $\pi_0(\cdot)$ functions are considered. The function $\pi_0^A(\cdot)$ is a constant function, whereas $\pi_0^B(\cdot)$ and $\pi_0^C(\cdot)$ are increasing sigmoid functions. Using $\pi_0^A(\cdot)$, we determine whether the proposed method works well when the true model does not follow the working model in which the probability of being an EE gene varies with gene length. Using $\pi_0^B(\cdot)$ and $\pi_0^C(\cdot)$, we determine whether the proposed method performs better than other methods when the true model follows the working model. $\pi_0^C(\cdot)$ has a more extreme characteristic than $\pi_0^B(\cdot)$ due to a gene-length region with a null probability of one.

Under a target FDR level of 0.05, the proposed method is compared to the standard q-value, IHW, BL, and AdaPT methods. These methods are chosen because they enable control of the FDR in the simulation study of Korthauer et al. (2019). The tuning parameters of the proposed method, N_{nh} and α are specified as follows. N_{nh} is set to $2,000$. The value of α is chosen arbitrarily or through cross-validation. First, we choose α values of 0.05 and 1 to better understand the proposed method's properties. Additionally, when α equals 1, we include the proposed method with true null probability $\pi_0(\cdot)$ for a reference. We also employ an α selection procedure based on repeated 10-fold cross-validation that maximizes the expected number of DDE genes. More precisely, we partition the observations $\{(X_j, P_j) : j = 1, \ldots, m\}$ completely at random into 10 parts. Holding each part out as a test set in turn, the other 9 parts are used as a training set. For each of 100 equally spaced α values between 0.001 and 0.2, and an α value 1, the training data are used to estimate $\pi_{0|\alpha}(\cdot)$ and our rejection rule for controlling pFDR at the target level 0.05. The number of DDE genes is determined based on applying the estimated rejection rule to the test data. This entire 10-fold cross-validation process is repeated M times, and the average number of DDE genes across the $10 \times$ M test sets is determined for each value of α. The value of α with the highest average number of DDE genes is selected and used with our proposed procedure on the entire data set to identify differentially expressed genes. In the simulation study, we use M $= 1$. M $= 100$ is used in the following data analysis section. Depending on whether the true $\pi_0(\cdot)$ is used or not, and on the value of α, the proposed method's procedures are referred to as prop.q(true, $\alpha = 1$), prop.q(est, $\alpha = 1$), prop.q(est, $\alpha = 0.05$), and prop.q(est, $\alpha =$ cv).

As discussed in Remark 2.2.1, the standard q-value method is still applicable in our simulation setup and is guaranteed to control pFDR. To estimate $\pi_0 = \mathbb{P}(H = 0)$, the histogram-based method of Nettleton et al. (2006) is used. Moreover, π_0 can be easily approximated from a property as follows:

$$\mathbb{P}(H_1 = 0) = \mathbb{E}_{X_1}\mathbb{P}(H_1 = 0 \mid X_1) \approx \frac{\sum_{j=1}^{m} \mathbb{P}(H_j = 0 \mid X_j)}{m} = \frac{\sum_{j=1}^{m} \pi_0(X_j)}{m}. \tag{2.24}$$

Because m is large, the Monte Carlo approximation (2.24) is accurate. For a reference, the standard q-value method with true null probability $\pi_0 \approx \frac{\sum_{j=1}^{m} \pi_0(X_j)}{m}$ is also included in the

comparison. Again, depending on whether the true parameter is used or not, the standard q-value method's procedures are referred to as std.q(true) and std.q(est). For simplicity, the omission of the estimator and true parameter symbols indicates the version of the procedure with parameters estimated from data (i.e., the version of the procedure that can be used in practice). For example, std.q = std.q(est).

Lastly, we turn to the IHW, BL, and AdaPT methods implemented in `R` packages `IHW`, `swfdr`, and `adaptMT`. `IHW` and `swfdr` are `Bioconductor R` packages, and `adaptMT` is a `CRAN R` package. Essentially, we follow the default configuration of the packages. For the AdaPT method, inspired by the simulation results in Korthauer et al. (2019), we use the *adapt_glm* function with the settings specified in the paper. Moreover, in the case of the IHW and AdaPT methods, the target FDR is set to 0.05. The procedures associated with the three methods are denoted by their respective names. In total, nine procedures are compared. The simulation results are analyzed mostly without employing the procedures that use true parameter values because these methods cannot be used in practice.

2.3.2 Simulation Results

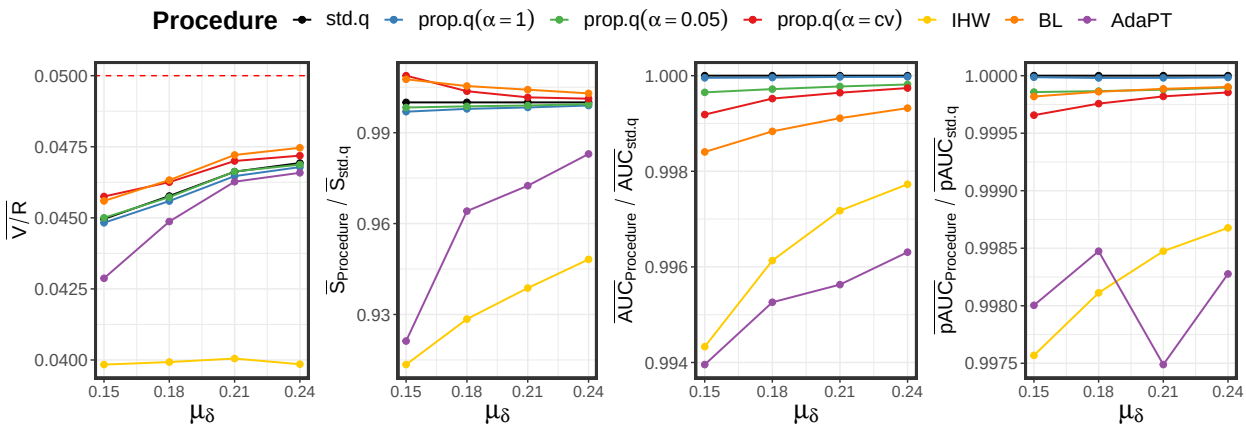


Figure 2.4: Four graphs of summary statistics of $\overline{\mathrm{V/R}}$, $\overline{\mathrm{S}}$, $\overline{\mathrm{AUC}}$, and $\overline{\mathrm{pAUC}}$ derived from the scenarios of $\pi_0^A(\cdot)$.

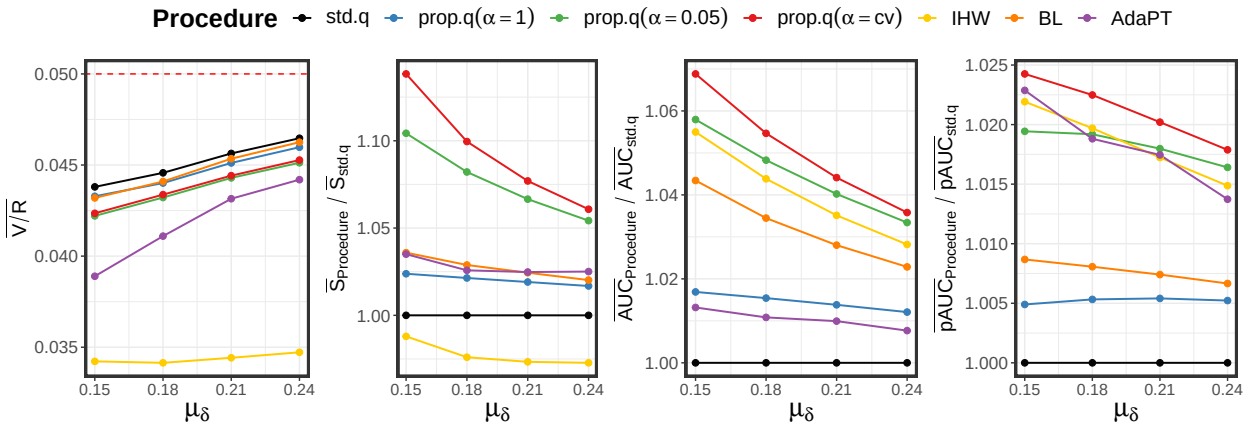


Figure 2.5: Four graphs of summary statistics of $\overline{\mathrm{V/R}}$, $\overline{\mathrm{S}}$, $\overline{\mathrm{AUC}}$, and $\overline{\mathrm{pAUC}}$ derived from the scenarios of $\pi_0^B(\cdot)$.

The nine procedures are compared in terms of mean false discovery proportion, mean true positive number, mean area under the receiver-operating characteristic (ROC) curve (AUC), and mean partial area under the ROC curve (pAUC). The ROC curve displays the trade-off between true-positive rate and false-positive rate. AUC and pAUC are the ROC curve's summary statistics, calculated based on each procedure's adjusted p-values or q-values. High AUC and pAUC values indicate that the procedure generally prioritizes true positives over false positives. The pAUC value is calculated by the standardized area under the ROC curve with a false-positive rate less than or equal to 0.1, regarded as a relevant region in our situation.

For each scenario composed of μ_δ and $\pi_0(\cdot)$, we generated 5000 data sets, which were used to approximate the four mean values: mean false discovery proportion, mean true positive number, mean AUC, and mean pAUC, denoted by $\overline{\mathrm{V/R}}$, $\overline{\mathrm{S}}$, $\overline{\mathrm{AUC}}$ and $\overline{\mathrm{pAUC}}$. When a procedure declares no significant hypotheses, the false discovery proportion is set to zero, which means $\overline{\mathrm{V/R}}$ is an empirical estimate of FDR rather than pFDR. However, in all our simulation scenarios, the probabilities of discovery corresponding to our proposed procedures are approximately 1. Therefore, for our proposed procedures, FDR $\approx$ pFDR in our simulation.

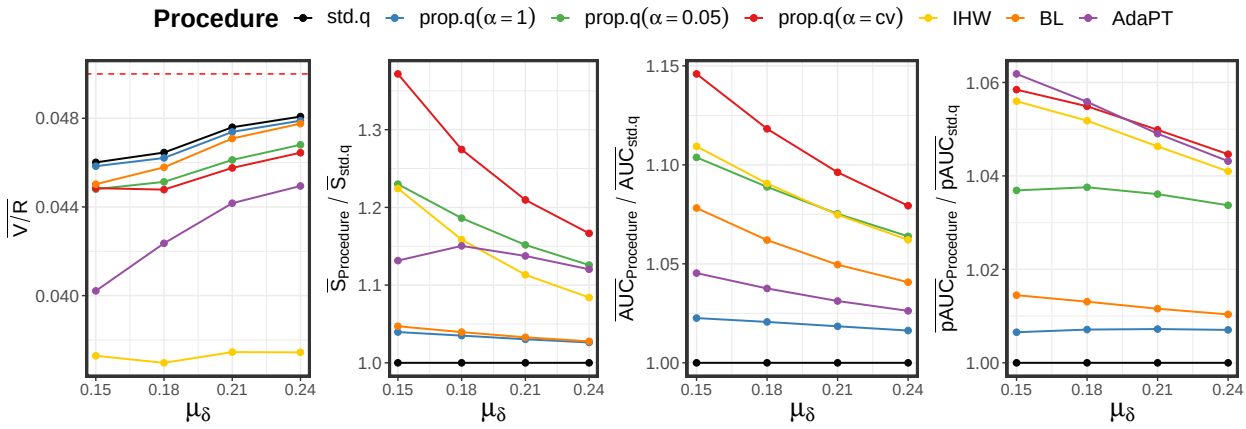


Figure 2.6: Four graphs of summary statistics of $\overline{\mathrm{V/R}}$, $\overline{\mathrm{S}}$, $\overline{\mathrm{AUC}}$, and $\overline{\mathrm{pAUC}}$ derived from the scenarios of $\pi_0^C(\cdot)$.

Table 2.3 summarizes all the simulation results. Figures 2.4 to 2.6 illustrate the results associated with the functions $\pi_0^A(\cdot)$, $\pi_0^B(\cdot)$ and $\pi_0^C(\cdot)$, respectively. In the figures, except for $\overline{\mathrm{V/R}}$, the ratio to std.q is calculated to illustrate the relative performance. Above all, all procedures under consideration control FDR in all scenarios.

Let us discuss the $\pi_0^A(\cdot)$ results. As illustrated in Figure 2.4, all procedures have nearly identical $\overline{\mathrm{AUC}}$ and $\overline{\mathrm{pAUC}}$ across all scenarios, showing that they perform similarly in terms of prioritizing true discoveries. In terms of true positive number $\overline{\mathrm{S}}$, when μ_δ is small, the std.q outperforms the IHW and AdaPT. On the other hand, all procedures associated with the proposed method perform nearly identically to the std.q, which is understandable as the proposed method generalizes the standard q-value method. Given the constant $\pi_0(\cdot)$, the rejection rule and FDR estimator for both methods are equivalent. In other words, if we know the constant null probability, both methods are equivalent. The results of $\pi_0^A(\cdot)$ suggest that the proposed method performs as well as std.q even when the covariate is irrelevant.

We now turn to the $\pi_0^B(\cdot)$ and $\pi_0^C(\cdot)$ results. First, we explore the proposed method's properties by comparing the related procedures to std.q. The summary statistics $\overline{\mathrm{S}}$, $\overline{\mathrm{AUC}}$, and $\overline{\mathrm{pAUC}}$ all indicate the same conclusion. The procedure performs best in the order of prop(α=cv),

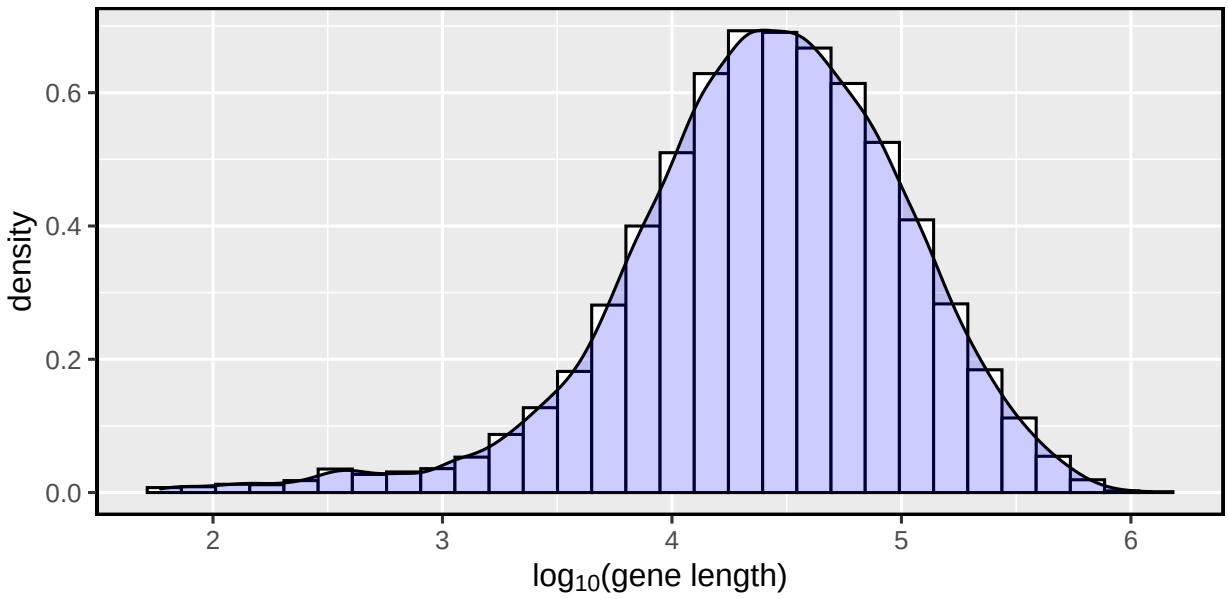


Figure 2.7: The histogram of log10 transformed gene length for 10,858 genes. The log10 transformed gene lengths have a mean of 4.4 and a standard deviation of 0.61.

prop(α=0.05), prop(α=1), then std.q. The order is well illustrated in Figures 2.5 and 2.6. Since prop.q($\alpha = 1$) is better than std.q, we can conclude that there is an improvement by considering covariate-specific null probability. It is noteworthy that the BL method consistently outperforms prop(α=1), even though both methods employ the covariate-specific null probability. By comparing prop(α=0.05) and prop(α=1), we can conclude that incorporating the classic rejection rule improves the proposed method. From the comparison between prop(α=cv) and prop(α=0.05), it can be concluded that cross-validation is beneficial for α selection to improve all evaluation criteria. As a result, we recommend using cross-validation to determine the value of α and setting the default value to 0.05.

The prop(α=cv) method is now compared to IHW, BL, and AdaPT. In terms of $\overline{\text{S}}$ and $\overline{\text{AUC}}$, prop(α=cv) surpasses other procedures in all scenarios. In the case of the AdaPT method, we can see that the performance is weakened in terms of $\overline{\text{AUC}}$, which is likely due to the vulnerability stated in Section 2.1. As seen in Figure 2.6, there are scenarios where AdaPT method

outperforms prop(α=cv) regarding $\overline{\text{pAUC}}$. For the corresponding scenarios, however, prop(α=cv) consistently generates more true positives $\overline{\text{S}}$ than AdaPT, which may be attributed to the differing FDR estimators. Based on our simulation setup, we can conclude that the proposed method employing cross-validation to select α outperforms the competing FDR-controlling methods in most scenarios and evaluation criteria that we considered.

2.4 Data Analysis

We tested our proposed method using RNA-seq data regarding disease resilience in young, healthy pigs [Lim et al. (2021)], and additional data on gene lengths. A comprehensive description of the study's design and hypotheses testing is described in Lim et al. (2021), which is summarized as follows. The study enrolled 912 F1 barrows at ~ 27 days of age in 15 batches. After three weeks in a healthy quarantine nursery, the piglets were exposed to natural polymicrobial diseases found on commercial farms. Not only were gene expression levels of the piglets' blood samples quantified, but also disease resilience phenotypes such as subjective health score, treatment rate, mortality, and growth rate. Although the paper [Lim et al. (2021)] tested numerous hypotheses, our current paper focuses on the association between gene expression and concurrent growth rate using blood samples taken during quarantine nursery periods before disease exposure. We anticipated that the disease-independent growth rate would be a long-term physical process, which is expected to be associated with the expression of longer genes. This expectation motivated us to concentrate on the association involving growth rate before disease exposure.

The following is the analysis we conducted. The gene expression in blood samples acquired during quarantine nursery was quantified using 3'mRNA sequencing with a globin block. Focusing on the data in Lim et al. (2021) from profiling period 2 and genes in the Ensembl database, we analyzed $10,858$ genes with a non-zero read count for at least 80% of the samples. The growth rate of a pig was used as a common dependent variable. We used log-scale read counts normalized and adjusted for nuisance factors as described by Lim et al. (2021). A p-value was calculated for each gene, testing whether the adjusted log2 transformed read count has a zero

slope coefficient. In total, we generated $10,858$ p-values. Figure 2.7 illustrates the histogram of log10-transformed gene lengths utilized to determine the covariate distribution in the simulation discussed in Section 2.3.

Table 2.1: Summary of the number of tests declared to be significant by the seven procedures at four nominal FDR levels 0.01, 0.05, 0.1, and 0.2.

nominal FDR	**std.q**	**prop.q(α=1)**	**prop.q(α=0.05)**	**prop.q(α=cv)**	**IHW**	**BL**	**AdaPT**
0.01	181	182	182	184	185	184	0
0.05	298	298	305	306	290	299	291
0.10	419	425	442	443	385	426	455
0.20	707	753	774	774	608	736	725

Table 2.2: Summary of the number of tests declared to be significant by multiple procedures at a nominal FDR level of 0.2 for four gene length-based groups. The grouping condition is expressed as a gene-length interval, and n denotes the number of genes contained within a group.

Group	**Grouping Criteria**	**n**	**std.q**	**prop.q(α=1)**	**prop.q(α=0.05)**	**prop.q(α=cv)**	**IHW**	**BL**	**AdaPT**
1	$[0, 11365)$	2715	178	177	174	173	123	161	136
2	$[11365, 27566)$	2714	167	167	165	162	114	165	159
3	$[27566, 66235)$	2714	177	199	202	203	159	188	180
4	$[66235, \infty)$	2715	185	210	233	236	212	222	250

We applied the seven procedures, described in Section 2.3, to the p-values and their associated gene lengths. For prop.q(α=cv), we employed repeated 10-fold cross-validation 100 times to reduce the sampling variation associated with cross-validation. The number of significant tests at various nominal FDR levels are summarized in Table 2.1. For each nominal FDR level, we varied the target FDR level when applying the IHW, AdaPT, and prop.q(α=cv). Regarding the proposed method, decreasing α from 1 to 0.05 or using cross-validation to select α tended to increase the number of significant tests, consistent with the simulation outcome. Furthermore, prop.q(α=cv) consistently declared a greater or similar number of tests significant than the std.q, IHW, and BL methods. Except for the nominal level of 0.1, the prop.q(α=cv) generated more significant results than AdaPT. When the nominal level is set to 0.01, the AdaPT method declared no tests significant. According to Figure 2.8, the null probability estimates tend to

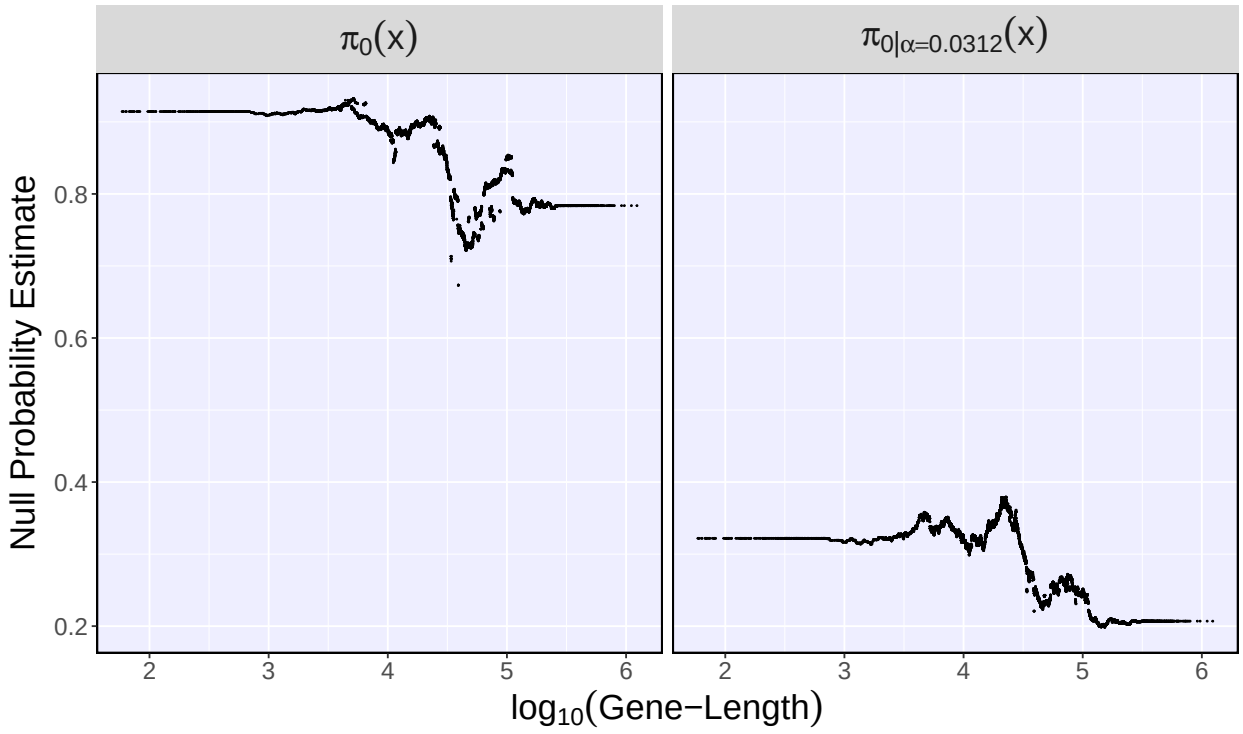


Figure 2.8: Null probability estimates of $\pi_0(x)$ and $\pi_{0|\alpha}(x)$ for 10,858 covariate values, following the procedure explained in Section 2.2.4. The nominal FDR level is set to 0.2. An α value of 0.0312 was chosen through the cross-validation approach.

decrease as the gene length grows. This tendency supports our assumption that the null probability varies with covariates, which is also consistent with our expectation that longer genes are more likely to be DE genes.

We classified genes according to their lengths into four groups with almost equal numbers. The grouping criteria are summarized in Table 2.2. The table shows the number of significant tests at a nominal FDR level of 0.2 for each group. Across all procedures, some interesting features are observed. As illustrated in Figure 2.9, the significant tests are observed in greater abundance in the 4th quantile group than in all other quantile groups. The phenomenon is noticeable when the AdaPT method is used. Additionally, the number of significant tests increases gradually from the second quantile group. This finding supports our intuition that the disease-independent growth rate is a long-term physical process that tends to involve longer genes. The standard q-value

method, which does not require knowledge of gene length, produced the same trend, supporting the validity of our method's assumption that the null probability varies with gene length.

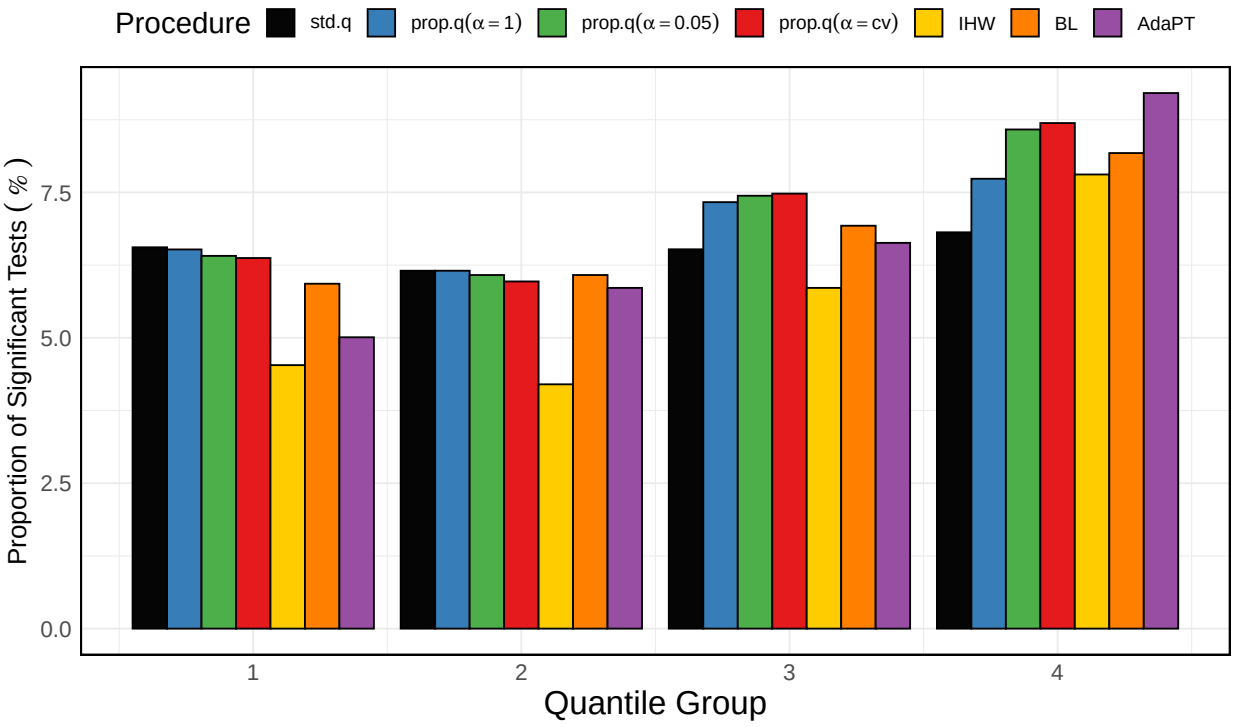


Figure 2.9: Barplot depiction of the proportion of tests declared to be significant by the three procedures at a nominal FDR level of 0.2 for four gene length-based groups. The grouping criteria are explained in Table 2.2.

2.5 Discussion

While the proposed method demonstrates significant gains over existing methods, there are still areas for improvement. First, the modeling framework upon which our method is developed is generalizable. One may consider a method in which the alternative distribution F_1 varies with the covariate variable. Second, the estimation procedure for estimating the null probabilities can be improved. The simulation results indicate that the BL method consistently beats our method with $\alpha = 1$, indicating a promising direction for further development of the estimation procedure.

Finally, different rejection rules can be defined using different posterior probability types. Performance is predicted to vary according to the target posterior probability. We anticipate that subsequent studies will examine our method from various perspectives. Simultaneously, we hope that our paper will inspire other scholars and be used in various fields.

2.6 Acknowledgment

The data used in this study were generated with funding from USDA National Institute of Food and Agriculture grant 2017-67007-26144, Genome Canada, Genome Alberta, and PigGen Canada.

2.7 References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Boca, S. M. and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6:e6035.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):1–21.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182.

Lim, K.-S., Cheng, J., Putz, A., Dong, Q., Bai, X., Beiki, H., Tuggle, C. K., Dyck, M. K., Fortin, F., Harding, J., et al. (2021). Quantitative analysis of the blood transcriptome of young healthy pigs and its relationship with subsequent disease resilience. *BMC Genomics*, 22(1):1–18.

Lopes, I., Altab, G., Raina, P., and De Magalhaes, J. P. (2021). Gene size matters: An analysis of gene length in the human genome. *Frontiers in Genetics*, 12:30.

Nettleton, D., Hwang, J. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337–356.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.

2.8 Appendix I: Proof of Theorem 2.2.1

Proof. The equality is derived as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\frac{V(\tilde{\Gamma})}{R(\tilde{\Gamma})}\Big|R(\tilde{\Gamma})>0\Big]=\sum_{k=1}^{m}\mathbb{E}\Big[\frac{V(\tilde{\Gamma})}{R(\tilde{\Gamma})}\Big|_{R(\tilde{\Gamma})=k,\,R(\tilde{\Gamma})>0}\Big]\cdot\mathbb{P}(R(\tilde{\Gamma})=k\mid R(\tilde{\Gamma})>0)\\
&=\sum_{k=1}^{m}\mathbb{E}\Big[\frac{V(\tilde{\Gamma})}{k}\Big|R(\tilde{\Gamma})=k\Big]\cdot\mathbb{P}(R(\tilde{\Gamma})=k(\tilde{\Gamma})>0)\\
&\overset{(\star)}{=}\mathbb{P}(H_j=0\mid\tilde{P}_j\in\tilde{\Gamma})\cdot\mathbb{P}(R(\tilde{\Gamma})>0\mid R(\tilde{\Gamma})>0)=\mathbb{P}(H_j=0\mid\tilde{P}_j\in\tilde{\Gamma})\\
&=\frac{m\cdot\mathbb{P}(H_j=0,\tilde{P}_j\in\tilde{\Gamma})}{m\cdot\mathbb{P}(\tilde{P}_j\in\tilde{\Gamma})}=\frac{\mathbb{E}V(\tilde{\Gamma})}{\mathbb{E}R(\tilde{\Gamma})}
\end{aligned}$$

□

$$\because\ \mathbb{E}[V(\tilde{\Gamma}) \mid R(\tilde{\Gamma}) = k] = \mathbb{E}\Big[\sum_{i=1}^{m} 1(H_j = 0) 1(\tilde{P}_i \in \tilde{\Gamma}) \mid \tilde{P}_1,\ldots,\tilde{P}_k \in \tilde{\Gamma}, \tilde{P}_{k+1},\ldots,\tilde{P}_m \notin \tilde{\Gamma}\Big]$$

$$= \mathbb{E}\Big[\sum_{i=1}^{k} 1(H_i = 0) \mid \tilde{P}_1,\ldots \tilde{P}_k \in \tilde{\Gamma}, \tilde{P}_{k+1},\ldots \tilde{P}_m \notin \tilde{\Gamma}\Big] = \sum_{i=1}^{k} \mathbb{P}\Big[H_i = 0 \mid \tilde{P}_1,\ldots \tilde{P}_k \in \tilde{\Gamma}, \tilde{P}_{k+1},\ldots \tilde{P}_m \notin \tilde{\Gamma}\Big]$$

$$= k \cdot \mathbb{P}(H_j = 0 \mid \tilde{P}_j \in \tilde{\Gamma})\ \ \because \text{i.i.d.}\ -(\star)$$

□

2.9 Appendix II: Proof of Theorem 2.2.2

Proof. Let H_{0j} denote $H_j = 0$.

$$\begin{aligned}
\mathbb{P}(T_{1j} \mid P_j \le \alpha, \vec{X}) &= \mathbb{P}\big(P_j \le u(X_j), H_{0j} \mid P_j \le \alpha, \vec{X}\big) \quad \because (2.7) \\
&= \mathbb{P}\big(P_j \le u(X_j) \mid H_{0j}, P_j \le \alpha, \vec{X}\big) \cdot \mathbb{P}\big(H_{0j} \mid P_j \le \alpha, \vec{X}\big) \\
&= \frac{\mathbb{P}\big(P_j \le u(X_j), P_j \le \alpha \mid H_{0j}, \vec{X}\big)}{\mathbb{P}\big(P_j \le \alpha \mid H_{0j}, \vec{X}\big)} \cdot \frac{\mathbb{P}\big(H_{0j}, P_j \le \alpha \mid \vec{X}\big)}{\mathbb{P}\big(P_j \le \alpha \mid \vec{X}\big)} \\
&= \frac{\mathbb{P}\big(P_j \le u(X_j) \wedge \alpha \mid H_{0j}, \vec{X}\big)}{\mathbb{P}\big(P_j \le \alpha \mid H_{0j}, \vec{X}\big)} \cdot \frac{\mathbb{P}\big(P_j \le \alpha \mid H_{0j}, \vec{X}\big) \cdot \mathbb{P}\big(H_{0j} \mid \vec{X}\big)}{\mathbb{P}\big(P_j \le \alpha \mid \vec{X}\big)} \\
&= \frac{\mathbb{P}\big(P_j \le u(X_j) \wedge \alpha \mid H_{0j}, \vec{X}\big)}{\mathbb{P}\big(P_j \le \alpha \mid H_{0j}, \vec{X}\big)} \cdot \frac{\mathbb{P}\big(P_j \le \alpha \mid H_{0j}, X_j\big) \cdot \mathbb{P}\big(H_{0j} \mid X_j\big)}{\mathbb{P}\big(P_j \le \alpha \mid X_j\big)} \quad \because (2.9, 2.10, 2.11) \\
&= \frac{u(X_j) \wedge \alpha}{\alpha} \cdot \mathbb{P}\big(H_{0j} \mid P_j \le \alpha, X_j\big) \quad \because (2.11, 2.12, \text{Assumption } 2.2.1) \\
&= \frac{u(X_j)}{\alpha} \cdot \pi_{0|\alpha}(X_j) \quad \because \max_j u(X_j) \le \alpha \ -(\star\star)
\end{aligned}$$

From the equality $(\star\star)$, the property that $\mathbb{P}(T_{1j} \mid P_j \le \alpha, \vec{X})$ is constant is equivalent to $u(X_j) \propto \frac{1}{\pi_{0|\alpha}(X_j)}$, with respect to j. □

Table 2.3: Summary of statistics of $\overline{\text{V/R}}$, $\overline{\text{S}}$, $\overline{\text{AUC}}$, and $\overline{\text{pAUC}}$ derived for different values of μ_δ and different functions $\pi_0(\cdot)$.

$\pi_0(\cdot)$	μ_δ	Label	std.q(true)	std.q	prop.q(true,α=1)	prop.q(α=1)	prop.q(α=0.05)	prop.q(α=cv)	IHW	BL	AdaPT
$\pi_0^A(\cdot)$	0.15	$\overline{V/R}$	0.050	0.045	0.050	0.045	0.045	0.046	0.040	0.046	0.043
·	·	$\overline{S}$	200.812	187.972	200.812	187.395	187.654	189.628	171.705	189.404	173.166
·	·	$\overline{AUC}$	0.744	0.744	0.744	0.744	0.744	0.743	0.740	0.743	0.740
·	·	$\overline{pAUC}$	0.657	0.657	0.657	0.657	0.656	0.656	0.655	0.656	0.655
·	0.18	$\overline{V/R}$	0.050	0.046	0.050	0.046	0.046	0.046	0.040	0.046	0.045
·	·	$\overline{S}$	369.905	355.347	369.905	354.587	354.882	356.657	329.934	357.264	342.596
·	·	$\overline{AUC}$	0.785	0.785	0.785	0.785	0.785	0.784	0.782	0.784	0.781
·	·	$\overline{pAUC}$	0.697	0.697	0.697	0.697	0.697	0.697	0.696	0.697	0.696
·	0.21	$\overline{V/R}$	0.050	0.047	0.050	0.046	0.047	0.047	0.040	0.047	0.046
·	·	$\overline{S}$	547.103	533.164	547.103	532.271	532.616	534.052	500.493	535.400	518.499
·	·	$\overline{AUC}$	0.818	0.818	0.818	0.818	0.818	0.818	0.816	0.817	0.815
·	·	$\overline{pAUC}$	0.734	0.734	0.734	0.734	0.734	0.734	0.733	0.734	0.732
·	0.24	$\overline{V/R}$	0.050	0.047	0.050	0.047	0.047	0.047	0.040	0.047	0.047
·	·	$\overline{S}$	715.256	702.926	715.256	702.202	702.527	703.779	666.500	705.014	690.958
·	·	$\overline{AUC}$	0.845	0.845	0.845	0.845	0.845	0.845	0.843	0.845	0.842
·	·	$\overline{pAUC}$	0.766	0.766	0.766	0.766	0.766	0.766	0.765	0.766	0.765
$\pi_0^B(\cdot)$	0.15	$\overline{V/R}$	0.050	0.044	0.050	0.043	0.042	0.042	0.034	0.043	0.039
·	·	$\overline{S}$	297.371	273.561	316.116	280.075	302.094	311.394	270.266	283.380	283.147
·	·	$\overline{AUC}$	0.744	0.744	0.768	0.756	0.787	0.795	0.785	0.776	0.754
·	·	$\overline{pAUC}$	0.657	0.657	0.663	0.660	0.669	0.672	0.671	0.662	0.672
·	0.18	$\overline{V/R}$	0.050	0.045	0.050	0.044	0.043	0.043	0.034	0.044	0.041
·	·	$\overline{S}$	525.801	499.692	549.915	510.387	540.725	549.442	487.702	514.103	512.576
·	·	$\overline{AUC}$	0.785	0.785	0.805	0.797	0.823	0.828	0.819	0.812	0.793
·	·	$\overline{pAUC}$	0.697	0.697	0.704	0.701	0.711	0.713	0.711	0.703	0.710
·	0.21	$\overline{V/R}$	0.050	0.046	0.050	0.045	0.044	0.044	0.034	0.045	0.043
·	·	$\overline{S}$	755.219	730.937	781.826	744.888	779.586	787.204	711.465	748.779	749.066
·	·	$\overline{AUC}$	0.818	0.818	0.835	0.830	0.851	0.854	0.847	0.841	0.826
·	·	$\overline{pAUC}$	0.734	0.734	0.740	0.738	0.747	0.749	0.746	0.739	0.747
·	0.24	$\overline{V/R}$	0.050	0.046	0.050	0.046	0.045	0.045	0.035	0.046	0.044
·	·	$\overline{S}$	968.460	947.227	995.371	963.159	998.606	1004.795	921.456	966.355	970.972
·	·	$\overline{AUC}$	0.845	0.845	0.860	0.856	0.874	0.876	0.869	0.865	0.852
·	·	$\overline{pAUC}$	0.766	0.766	0.772	0.770	0.779	0.780	0.777	0.771	0.776
$\pi_0^C(\cdot)$	0.15	$\overline{V/R}$	0.050	0.046	0.050	0.046	0.045	0.045	0.037	0.045	0.040
·	·	$\overline{S}$	121.737	115.952	132.373	120.544	142.610	159.057	141.967	121.412	131.207
·	·	$\overline{AUC}$	0.744	0.744	0.776	0.761	0.821	0.852	0.825	0.802	0.778
·	·	$\overline{pAUC}$	0.656	0.656	0.665	0.661	0.681	0.695	0.693	0.666	0.697
·	0.18	$\overline{V/R}$	0.050	0.046	0.049	0.046	0.045	0.045	0.037	0.046	0.042
·	·	$\overline{S}$	236.360	229.391	251.640	237.423	272.098	292.350	265.758	238.449	263.901
·	·	$\overline{AUC}$	0.785	0.785	0.812	0.801	0.854	0.878	0.856	0.833	0.814
·	·	$\overline{pAUC}$	0.697	0.697	0.706	0.702	0.723	0.735	0.733	0.706	0.736
·	0.21	$\overline{V/R}$	0.050	0.048	0.050	0.047	0.046	0.046	0.037	0.047	0.044
·	·	$\overline{S}$	363.087	356.173	381.231	366.962	410.210	430.825	396.529	367.860	405.154
·	·	$\overline{AUC}$	0.818	0.818	0.841	0.833	0.880	0.897	0.880	0.859	0.844
·	·	$\overline{pAUC}$	0.734	0.734	0.742	0.739	0.760	0.771	0.768	0.742	0.770
·	0.24	$\overline{V/R}$	0.050	0.048	0.050	0.048	0.047	0.046	0.037	0.048	0.045
·	·	$\overline{S}$	486.493	480.227	505.832	492.826	540.639	560.260	520.587	493.493	538.035
·	·	$\overline{AUC}$	0.845	0.845	0.865	0.859	0.899	0.913	0.898	0.880	0.868
·	·	$\overline{pAUC}$	0.766	0.766	0.774	0.771	0.792	0.800	0.797	0.774	0.799

CHAPTER 3. DETECTING DIFFERENTIALLY EXPRESSED GENES BY COMBINING INFORMATION FROM A PILOT AND A MAIN STUDY

Hyeongseon Jeon[1], Dan Nettleton[1], and Kyu-Sang Lim[2]

[1]Department of Statistics, Iowa State University, Ames, IA 50011, USA

[2]Department of Animal Resources Science, Kongju National University, Yesan-gun, Chungnam 32439, Republic of Korea

Modified from a manuscript to be submitted to *Bioinformatics*

Abstract

This paper presents a novel false discovery rate (FDR) controlling method for incorporating gene expression data sets collected from a pilot and main study. Our approach allows a higher p-value rejection threshold for the main study when the p-value for the pilot study is relatively low, as determined by what we call an L-shaped rejection region. We search over a smaller set of rejection rules than a competing independent hypothesis weighting method, which leads to an increase in testing power for our approach. Nevertheless, a bias correction is still required when estimating FDR due to the many rejection rules we consider. Accordingly, we propose a bias correction method using the relationship between two types of FDR estimators. A simulation study demonstrates that our method effectively regulates FDR and surpasses existing methods in terms of true positive rate. At all nominal FDR levels, the proposed procedure declares more tests to be significant than the existing methods throughout data analysis.

3.1 Introduction

Researchers conduct gene expression studies to identify genes with potentially interesting expression patterns. Such genes, often referred to as differentially expressed (DE), may change

mRNA abundance levels across treatments or exhibit non-null associations with a continuous variable of interest. Typically, DE genes are identified by testing a hypothesis for each gene, which results in a multiple testing problem. When confronted with the multiple testing problem, researchers aim to maximize the number of true positives while controlling the number of false positives. Researchers consider various error quantities to reach this goal. When analyzing expression data, the most commonly used error quantity is the false discovery rate (FDR), proposed by Benjamini and Hochberg (1995) and closely related to the expected proportion of false positives among tests whose null hypothesis is rejected.

The most widely used method for FDR control is Storey's (2002) q-value method, based on the number of null or equivalently expressed (EE) genes m_0, among all m genes tested for differential expression. Modern FDR-controlling methods are based on gene-specific covariates such as non-zero mean expression and the proportion of samples having non-zero expression, which are described in Korthauer et al. (2019). The methods include Cai and Sun (2009), Scott et al. (2015), Ignatiadis et al. (2016), Boca and Leek (2018), and Lei and Fithian (2018). Among the contemporary methods, the independent hypothesis weighting (IHW) method [Ignatiadis et al. (2016)] increases detection power by maximizing the number of declared differentially expressed (DDE) genes when the corresponding covariate values are used to categorize the genes. The IHW method considers nearly all possible group-specific p-value rejection thresholds. We argue that the inference power may be increased by reducing the number of rejection threshold combinations. In this perspective, we present a gene expression data analysis procedure involving a pilot and main study.

We use a pilot study's findings to refine a larger-scale main study. After collecting data from both studies, researchers often decide whether to undertake further analysis using both data sets or simply the data from the main study. When researchers focus on only the main study, critical information gathered during the pilot study is lost. Inspired by these concerns and the IHW method, we provide an approach based on two independently generated p-value vectors, P_{pilot} and P_{main}, obtained from the two studies. We present an FDR-controlling method in which P_{main}

serves as the main source p-value vector while P_{pilot} serves as a covariate variable. If the pilot study is informative for DE genes, we expect to detect more DE genes among genes with relatively low P_{pilot}. From this perspective, we propose a novel rejection rule with a negative relationship between the main study's p-value rejection threshold and the pilot study's p-value.

The remainder of this paper is organized as follows. Section 3.2 details our method, including a calibration procedure for achieving a target FDR level. In Section 3.3, we demonstrate our method's efficiency through simulation studies, and in Section 3.4 we illustrate our method through data analysis. Finally, Section 3.5 assesses the proposed approach's potential for further development.

3.2 Method Proposal

This section describes our method and is divided into three subsections. Section 3.2.1 characterizes the gene expression data that we analyze. Although we focus on analyzing a pilot and corresponding main study, our method is more generally applicable to gene expression data analysis of multiple studies satisfying the outlined characteristics. Sections 3.2.2 and 3.2.3 discuss approaches for estimating m_0 and detecting DE genes, respectively. Both approaches use independently generated p-value vectors from the two studies, denoted by P_{pilot} and P_{main}.

3.2.1 Data Description

We present three characteristics of the gene expression data derived from pilot and main studies. First, both data sets are independently collected. This is a natural assumption when the studies are conducted separately with independent experimental or observational units. The second characteristic is that the DE genes in both studies are identical, which is reasonable when treatment factors are identical and experimental or observational units come from the same population for both pilot and main studies. Let H_0^i and H_A^i denote the collection of EE and DE genes, respectively, in study i. For example, H_0^{pilot} represents EE genes for the pilot study. The

second characteristic can be expressed as

$$H_A^{pilot} = H_A^{main} \iff H_0^{pilot} = H_0^{main}. \tag{3.1}$$

From (3.1), we define $H_0 = H_0^{pilot} = H_0^{main}$ and $H_A = H_A^{pilot} = H_A^{main}$. When establishing the m_0 estimator in Section 3.2.2, the first and second characteristics are crucial.

The third characteristic is that the testing power of the main study to detect a DE gene is greater than that of the pilot study for the majority of genes. This is a reasonable characteristic because the main study will naturally use more experimental or observational units than the pilot study. Furthermore, the pilot study may give an investigator practice with laboratory protocols that can lead to the use of improved or refined techniques in the main study, thereby reducing error variance in the main study relative to the pilot. This third characteristic helps us decide our rejection rule in Section 3.2.3.1.

3.2.2 Estimating m_0

The total number of EE genes m_0 is required when estimating FDR for a list of DDE genes. To estimate m_0, we employ a method suggested by Orr et al. (2012). The method provides a conservative estimator of $m_{00} = \left|H_0^A \cap H_0^B\right|$ for any two independent studies A and B. The method estimates m_{00} using two independently generated p-value vectors from the studies. Let $P_j = (P_{Aj},\ P_{Bj})$ denote a p-value pair for the jth gene. Orr et al.'s (2012) m_{00} estimator is defined as

$$\hat{m}_{00} = \frac{\sum_{j=1}^{m} \mathbf{1}\{P_j \in [\lambda_A, 1] \times [\lambda_B, 1])\}}{(1-\lambda_A)(1-\lambda_B)}, \tag{3.2}$$

where the cutoff points λ_A and λ_B are chosen by the histogram-based method of Nettleton et al. (2006). Each cutoff point is chosen so that the histogram of p-values is approximately uniform to the right of the cutoff point. As illustrated in Figure 3.1, Orr et al.'s (2012) estimator (3.2) can be interpreted as a normalized count of the p-value pairs within a rectangle N with the size of the rectangle $(1-\lambda_A)(1-\lambda_B)$ serving as the normalizing constant. Orr et al.'s (2012) estimator is well-matched to our data as our data sets are independently collected which is the first

characteristic described in Section 3.2.1. Furthermore, the second characteristic, $H_0^{pilot} = H_0^{main}$, allows us to use Orr et al.'s (2012) estimator when estimating $m_0 = |H_0|$. Throughout this paper, we use $\hat{m}_0$ to refer to the estimator of m_0 obtained using Orr et al. (2012).

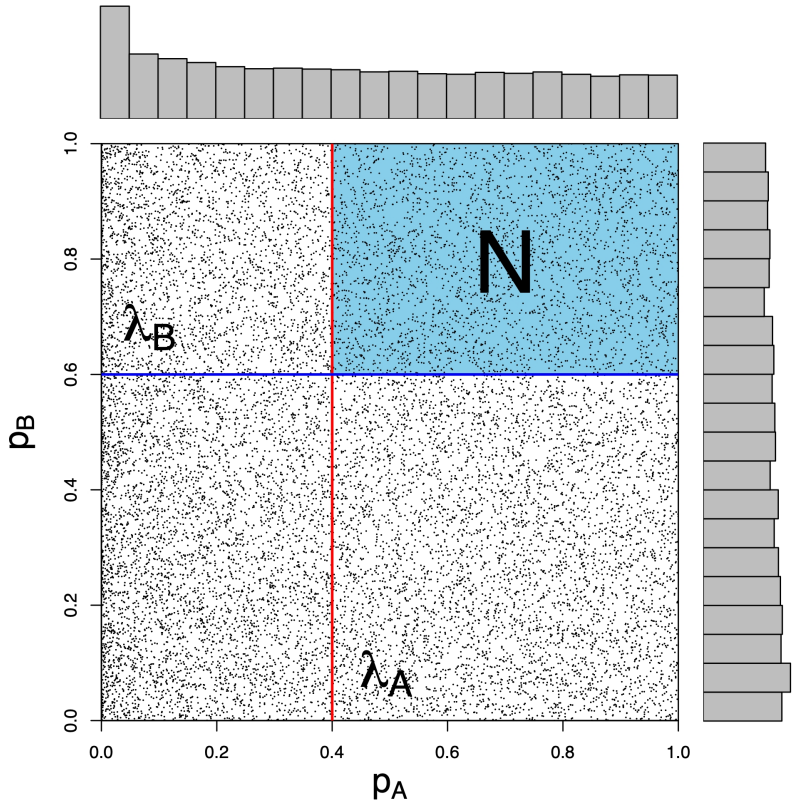


Figure 3.1: Scatter plot of p_B versus p_A with marginal histograms in each margin and a rectangle N defined by cutoff points λ_A and λ_B.

3.2.3 Detecting DE genes

This section describes an approach for detecting DE genes while controlling FDR at level α. Section 3.2.3.1 begins by defining a collection of rejection rules we consider. Each rejection rule has a one-to-one correspondence with a rejection region (RR). The rejection region is used to define FDR estimators and aids in the comprehension of our rejection rule. In Section 3.2.3.2, two FDR estimators are defined. Section 3.2.3.3 employs an FDR estimator to construct an optimal

rejection region at FDR level α. However, the FDR estimator for the optimal rejection region can be biased. Therefore, Section 3.2.3.4 presents a calibration method for correcting the bias using the relationship between the two FDR estimators defined in Section 3.2.3.2.

3.2.3.1 L-Shape Rejection Region

To discuss our method for detecting DE genes, we need to specify the collection of rejection rules that we consider. The following three types define all the rejection rules:

1. For a given $g \in \{1, \ldots, m\}$, reject all genes corresponding to indices in $DDE^1_g = \{j \in \{1, \ldots, m\} : P_{main,\, j} \leq P_{main,\, g}\}$.

2. For a given $g \in \{1, \ldots, m\}$, reject all genes corresponding to indices in $DDE^2_g = \{j \in \{1, \ldots, m\} : P_{pilot,\, j} \leq P_{pilot,\, g},\, P_{main,\, j} \leq P_{main,\, g}\}$.

3. For given $g_1 \neq g_2 \in \{1, \ldots, m\}$ $(P_{main,\, g1} < P_{main,\, g2})$, reject all genes corresponding to indices in $DDE^3_{g_1,\, g_2} = DDE^1_{g_1} \cup DDE^2_{g_2}$.

The above definition yields $\frac{m(m+3)}{2} = O(m^2)$ number of distinct rejection rules almost surely, while the IHW method considers $O(m^p)$ for p equal to the number of covariate-specific groups. Each rejection rule provides a DDE gene list by DDE^1_g, DDE^2_g, and $DDE^3_{g_1,\, g_2}$, respectively. The type 1 rejection rules are determined only by P_{main}. The type 2 rejection rules use both p-value vectors P_{pilot} and P_{main}. Because type 3 rejection rules incorporate type 1 and 2 rejection rules, they are constructed using data from both studies. The rejection rules established solely with P_{pilot} are excluded due to the third characteristic, implying that P_{main} is required to establish the rejection rules.

We now specify rejection regions corresponding to the previously defined rejection rules. Each rejection region enables us to estimate the FDR for a rejection rule. The rejection regions corresponding to the previously defined rejection rules are as follows:

1. For a given $g \in \{1, \ldots, m\}$, the rejection region is $RR^1_g = \{(p_1, p_2) \in [0, 1]^2 : p_2 \leq P_{main,\, g}\}$.

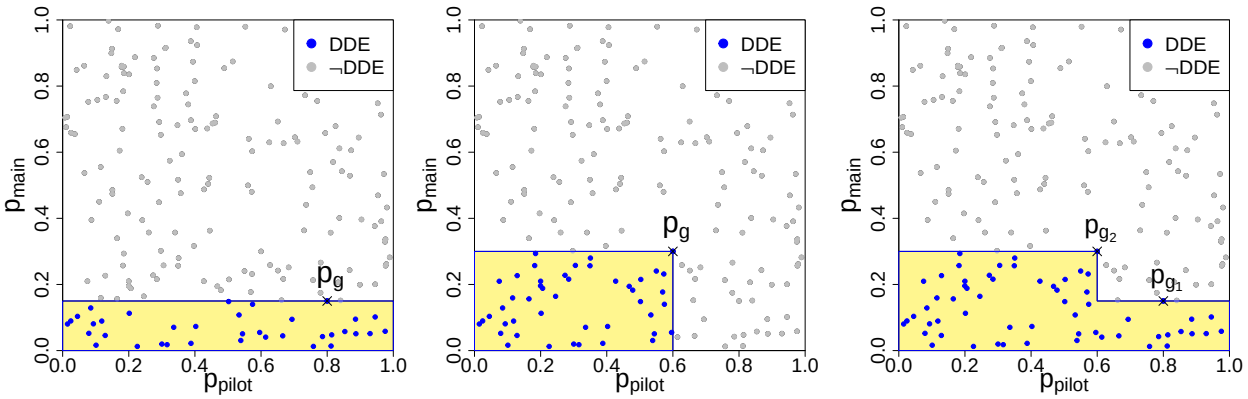
2. For a given $g \in \{1, \ldots, m\}$, the rejection region is

   $RR^2_g = \{(p_1, p_2) \in [0,1]^2 : p_1 \leq P_{pilot,\, g},\ p_2 \leq P_{main,\, g}\}$.

3. For given $g_1 \neq g_2 \in \{1, \ldots, m\}$ ($P_{main,\, g1} < P_{main,\, g2}$), the rejection region is

   $RR^3_{g_1,g_2} = RR^1_{g_1} \cup RR^2_{g_2}$.

Due to the constraint in 3 ($P_{main,\, g1} < P_{main,\, g2}$), no type 3 rejection region is a type 1 rejection region and vice versus. Furthermore, the rejection rule and rejection region of all types have one-to-one correspondence almost surely. In other words, a rejection region and a rejection rule uniquely identify each other. All procedures are now described in terms of rejection regions due to the equivalence. Figure 3.3 illustrates the rejection regions of all types. The p-value pairs for DDE genes in each rejection region are colored blue. Both type 1 and 2 rejection regions have rectangular forms. Each type 3 rejection region is formed by the union of two rectangles and has an "L"-shape. Because most rejection regions are of type 3, we refer to the entire collection of type 1, 2, and 3 rejection regions as L-shaped and use $\mathcal{L}$ to denote the entire collection.



(a) Type 1 RR **(b)** Type 2 RR **(c)** Type 3 RR

Figure 3.3: Scatter plots of P_{main} versus P_{pilot} with rejection regions (yellow-colored areas) for all three types.

3.2.3.2 False Discovery Rate Estimators

We present two FDR estimators for an L-shaped rejection region L. Note that both estimators are defined assuming m_0 is known, but in practice $\hat{m}_0$ is substituted in place of m_0. Let $R(L)$ denote the number of DDE genes corresponding to L, i.e., the number of p-value pairs in the region L. Let $V(L)$ denote the number of false positives associated with L. Let $P_j = (P_{pilot,\,j}, P_{main,\,j})$ denote the p-value pair for the jth gene. $R(L)$ and $V(L)$ then can be expressed as $\sum_{j\in\{1,\ldots,m\}} \mathbf{1}(P_j \in L)$ and $\sum_{j\in H_0} \mathbf{1}(P_j \in L)$, respectively. Let $\text{Area}(L)$ denote the area of L. Let $\text{FDP}(L|H_0) = \frac{V(L)}{R(L)\vee 1}$. FDR corresponding to the L can be expressed as $\text{FDR}(L) = \mathbb{E}\{\text{FDP}(L|H_0)\}$. Moreover, the positive FDR (pFDR), which was introduced by Storey (2003), can be expressed as $\text{pFDR}(L) = \mathbb{E}\big\{\frac{V(L)}{R(L)} \mid R(L) > 0\big\}$.

When the jth gene's simple null hypothesis is true, expressed by $j \in H_0$, and the test statistic is continuous, the p-value follows a uniform distribution between 0 and 1. From this fact and the first data characteristic of data independence, the following assumption is established:

Assumption 3.2.1. $\text{P}_j \mid j \in H_0 \overset{\text{iid}}{\sim} \text{Unif}(0,1)^2$.

Our initial estimator of FDR is based on Storey's (2003) pFDR estimator. Because pFDR is an upper bound of the FDR, we can use pFDR estimator as a conservative estimator of FDR. By using Assumption 2.1 and the reasonings in Storey (2003), we define our initial FDR estimator as

$$\widehat{\text{FDR}}_1(L \mid m_0) = \frac{m_0 \cdot \text{Area}(L)}{R(L) \vee 1}. \tag{3.3}$$

This estimator's properties are not described in detail because procedures maximizing the number of DDE genes over rejection regions L tend to underestimate FDR, as discussed by Ignatiadis et al. (2016). To correct this potential bias, we introduce a second FDR estimator.

The second estimator is developed under the following assumption.

Assumption 3.2.2. $\forall\, j \neq j' \in H_0,\ \mathbb{E}\Big\{\frac{\mathbf{1}(P_j\in L)}{R(L)\vee 1}\,\Big|\,H_0\Big\} = \mathbb{E}\Big\{\frac{\mathbf{1}(P_{j'}\in L)}{R(L)\vee 1}\,\Big|\,H_0\Big\}$.

Suppose I_0 is a random sample of H_0. From the Assumption 3.2.2, we can derive the following property.

$$\mathbb{E}\Big\{\frac{\sum_{j\in H_0}\mathbf{1}(P_j\in L)}{(R(L)\vee 1)\cdot|H_0|}\,\Big|\,H_0\Big\}=\mathbb{E}\Big\{\frac{\sum_{j\in I_0}\mathbf{1}(P_j\in L)}{(R(L)\vee 1)\cdot|I_0|}\,\Big|\,H_0\Big\}$$
$$\iff\quad \mathbb{E}\Big\{\frac{V(L)}{R(L)\vee 1}\,\Big|\,H_0\Big\}=\mathbb{E}\Big\{\frac{\sum_{j\in I_0}\mathbf{1}(P_j\in L)}{R(L)\vee 1}\cdot\frac{|H_0|}{|I_0|}\,\Big|\,H_0\Big\}. \tag{3.4}$$

According to the property (3.4), when $|H_0|$ and $|I_0|$ are known $\frac{\sum_{j\in I_0}\mathbf{1}(P_j\in L)}{(R(L)\vee 1)}\cdot\frac{|H_0|}{|I_0|}$ is an unbiased predictor of $\mathrm{FDP}(L|H_0)$, and the second FDR estimator is chosen as follows:

$$\widehat{\mathrm{FDR}}_2(L\mid m_0,I_0)=\frac{\sum_{j\in I_0}\mathbf{1}(P_j\in L)}{R(L)\vee 1}\cdot\frac{m_0}{|I_0|}. \tag{3.5}$$

The second FDR estimator (3.5) has the limitation of requiring unrealistic knowledge of I_0. Nonetheless, we demonstrate in Section 3.3 that $\widehat{\mathrm{FDR}}_2(L\mid m_0,I_0)$ is useful for correcting the bias of the first FDR estimator (3.3).

3.2.3.3 Optimal Rejection Region

From the L-shaped rejection regions defined in Section 3.2.3.1, we seek a region that is optimal in the sense that the number of true positive results is maximized while FDR is bounded above at the desired level. Initially, we consider rejection regions resulting in r DDE genes for some fixed $r\in\{1,\ldots,m\}$. Because null p-value pairs are assumed to be uniformly distributed, the larger the region, the greater the expected number of null p-value pairs falling in the region. Therefore, to produce the fewest false positives, we seek the smallest rejection region among those containing r p-value pairs. Because we consider rejection regions with the number of DDE genes fixed at r, the lowest number of false positives indicates the greatest number of true positives. From this perspective,

$$\arg\min\big\{\mathrm{Area}(L):L\in\mathcal{L},\,R(L)=r\big\} \tag{3.6}$$

is a natural choice for our rejection region when the number of DDE genes is fixed at r. We refer to this area minimizing region as a locally optimal rejection region. Figure 3.4 illustrates how the

locally optimal rejection region is obtained. Notably, regardless of the number of DDE genes between 1 and m, there is always a type 1 L-shaped rejection region (a blue-lined region in Figure 3.4), indicating that all locally optimal rejection regions are well-defined. Let $L^*(r) = \arg\min \{\text{Area}(L) : L \in \mathcal{L},\ R(L) = r\}$ for $\forall\, r \in \{1, \ldots, m\}$, and let $L^*(0) = \emptyset$. Then, define a collection of all locally optimal rejection regions for specified DDE genes numbering 0 to m by

$$\mathcal{L}^* = \{L^*(r) : r = 0, \ldots, m\}. \tag{3.7}$$

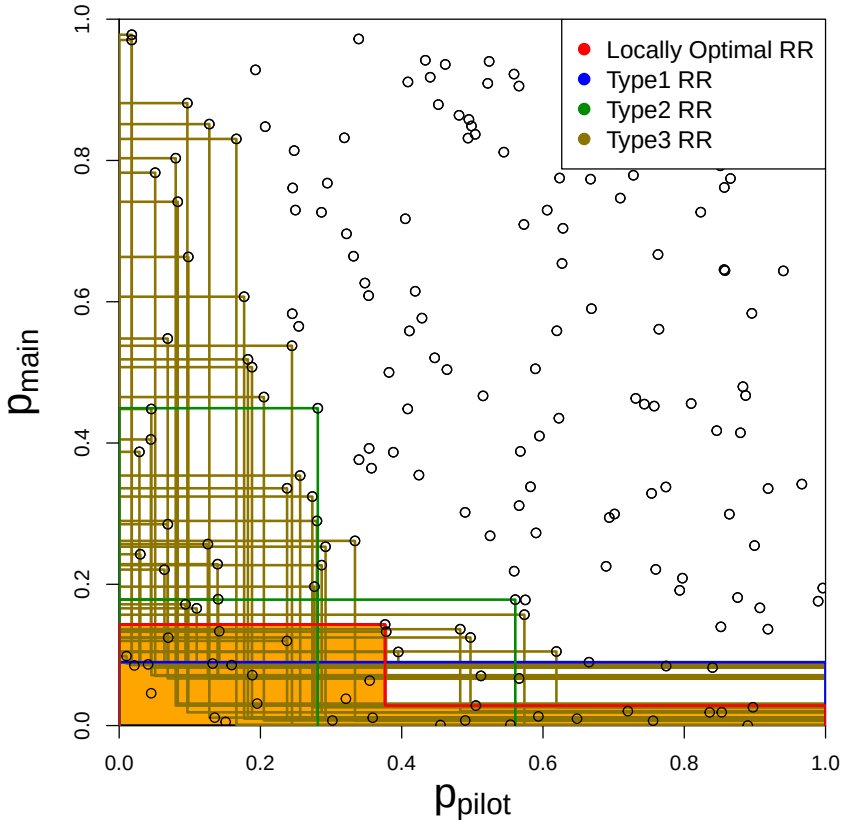


Figure 3.4: Scatter plot of P_{main} versus P_{pilot} with a locally optimal rejection region (orange colored region) and all types of L-shape rejection regions containing 30 p-value pairs ($r = 30$).

Now, let us identify the optimal rejection region among locally optimal rejection regions in $\mathcal{L}^*$. Practically, we seek a rejection region where the number of DDE genes is maximized, and the FDR is kept below a target level a. Once $\widehat{\text{FDR}}_1(\cdot \mid \hat{m}_0)$ is substituted for the FDR, we

approximate the optimal rejection region as

$$L_a^* \stackrel{\text{def}}{=} \arg\max \left\{ R(L) : L \in \mathcal{L}^*, \widehat{\text{FDR}}_1(L \mid \hat{m}_0) \leq a \right\} \tag{3.8}$$

Note that L_a^* is well-defined for every $a \in [0,1]$ as $\mathcal{L}^*$ contains the empty set ($\widehat{\text{FDR}}_1(\emptyset \mid m_0) = 0$ $\because$ Area$(\emptyset) = 0$). Ideally, using L_α^* as our rejetion region would result in FDR control at level α. However, $\widehat{\text{FDR}}_1(L^*(r) \mid \hat{m}_0)$ tends to underestimate the FDR due to the area-minimizing scheme of $L^*(r)$. Due to the underestimation bias, we calibrate a of L_a^* to control FDR at α. The following section describes the calibration method.

3.2.3.4 Calibration on a of L_a^* to control FDR at α

For the calibration method, a theorem is established using the relationship between $\widehat{\text{FDR}}_1(L_a^* \mid \hat{m}_0)$ and $\text{FDP}(L_a^* \mid H_0)$, $\forall\, a \in [0,1]$. The theorem is based on an assumption:

Assumption 3.2.3. $\forall\, L_1 \neq L_2 \in \mathcal{L}$, $\widehat{\text{FDR}}_1(L_1|\hat{m}_0) \neq \widehat{\text{FDR}}_1(L_2|\hat{m}_0)$ almost surely.

Each L-shape rejection region is defined by one or two p-value pairs. Additionally, because p-value pairs are continuous random variables, they are almost surely distinct. As a result, the $\widehat{\text{FDR}}_1(\cdot \mid \hat{m}_0)$ values determined by the distinct p-value pairs are also almost surely distinct for all L-shaped rejection regions. The assumption is formed from this perspective. Under the assumption, we provide the following theorem:

Theorem 3.2.1. Let $\tilde{\alpha} = \max\left\{ \widehat{\text{FDR}}_1(L_a^* \mid \hat{m}_0) : a \in [0,1], \text{FDP}(L_a^* \mid H_0) \leq \alpha \right\}$. Then, $\text{FDP}(L_{\tilde{\alpha}}^* \mid H_0) \leq \alpha$ almost surely..

According to the Theorem 3.2.1 which we prove in the Appendix, $L_{\tilde{\alpha}}^*$ has an FDR that is less than or equal to α. However, the $\tilde{\alpha}$ cannot be determined because the relationship between $\widehat{\text{FDR}}_1(L_a^* \mid \hat{m}_0)$ and $\text{FDP}(L_a^* \mid H_0)$ is unknown. Instead, we suggest using the relationship between $\mathbb{E}_{rep}\{\widehat{\text{FDR}}_1(L_a^* \mid \hat{m}_{0,rep})\}$ and $\mathbb{E}_{rep}\{\widehat{\text{FDR}}_2(L_a^* \mid \hat{m}_{0,rep}, I_{0,rep})\}$. The expectation $\mathbb{E}_{rep}$ is approximated by simulation. We repeatedly replace a randomly selected 5% of the p-value pairs with independently generated $\text{Unif}(0,1)^2$ draws following the Assumption 3.2.1. Substituting a

small fraction (5%) does not significantly alter the relationship between $\widehat{\text{FDR}}_1(L_a^* \mid \hat{m}_0)$ and $\text{FDP}(L_a^* \mid H_0)$. In addition, under Assumption 3.2.2, $\widehat{\text{FDR}}_2$ is an unbiased predictor of FDP, allowing us to substitue $\widehat{\text{FDR}}_2$ for FDP. Note that m_0 is repeatedly estimated in each simulation run and denoted by $\hat{m}_{0,rep}$. Then, the proposed adjusted value of a is as follows:

$$\alpha^* = \max\Big[\mathbb{E}_{rep}\big\{\widehat{\text{FDR}}_1(L_a^* \mid \hat{m}_{0,rep})\big\} : a \in [0,1],\, \mathbb{E}_{rep}\big\{\widehat{\text{FDR}}_2(L_a^* \mid \hat{m}_{0,rep}, I_{0,rep})\big\} \leq \alpha\Big]. \tag{3.9}$$

Using the value in (3.9), our DE genes detection method is established. Following the definition of L_a^* in (3.8), all genes corresponding to the rejection region $L_{\alpha^*}^*$ are declared as DE genes. In the following section, our simulations demonstrate effective FDR control in practice.

3.3 Simulation Study

3.3.1 Model Description

We conduct a simulation study to assess our method's performance, investigating independent and normally distributed gene expression data for $m = 10,000$ genes. When creating the gene expression data, two factors, *study* and *treatment*, are considered. The *study* factor has levels of pilot and main, and the *treatment* factor has levels of cntrl and trt. For *study* i and *treatment* j, suppose there are n_i observations. For a given π_0, $m_1 = m \times (1 - \pi_0)$ genes are randomly selected as DE genes. Without loss of generality, the first m_1 genes are considered DE genes, and the treatment effects are generated from a normal distribution. In addition, gene and study-specific variance is assumed. For each gene, the two studies' variances are determined by the main study's standard deviation and the two studies' variance ratio, generated from an inverse chi-square distribution and a log-normal distribution. A fixed study effect of 1 is assumed across all genes. For gene j, *study* s, *treatment* t, and observation k, the response variable is generated following

the model below. Note that independence holds unless otherwise specified.

$$
\begin{aligned}
&Y_{jstk} \sim N(\mu_{jst},\ \sigma^2_{js}), \text{ where } \mu_{jst} = \mathbf{1}(s = \text{ main}) + \delta_{jt}, \\
&j \in \{1, \cdots m\},\ s \in \{\text{pilot, main}\},\ t \in \{\text{cntrl, trt}\},\ \text{and } k \in \{1, \cdots n_s\}, \\
&\delta_{j,\text{cntrl}} = 0 \text{ and } \delta_{j,\text{trt}} = \mathbf{1}(j \le m_1) \cdot N(\mu_\delta,\ \sigma^2_\delta = 0.02^2), \\
&\sigma^2_{j,\text{pilot}} = R_{\sigma^2 j} \cdot \sigma^2_{j,\text{main}}, \\
&\sigma_{j,\text{main}} \sim \text{Inv-}\chi^2_5 \text{ and } R_{\sigma^2 j} \sim \text{Lognormal}(\mu_R,\ \sigma^2_R = 0.1^2).
\end{aligned} \tag{3.10}
$$

Suppose that a data set following the model (3.10) is generated. For each gene, to infer the *treatment* effect, we can apply a cell-means model with the two factors assuming heterogeneous variance between the *studies*. Then, we can calculate the p-value testing the null hypothesis of $H_0^j : \frac{\bar{\mu}_{j\cdot\text{trt}} - \bar{\mu}_{j\cdot\text{cntrl}}}{2} = 0$, where $\bar{\mu}_{j\cdot t}$ denotes a marginal mean across the *study*-factor levels, using the t-test with degrees of freedom approximated by the Cochran Satterthwaite method. The p-value vector is denoted by $P_{combined}$. Additionally, for each *study* data, the p-value testing the *treatment* effect is calculated using a two-sample t-test, which is the basis for our method. Each *study*'s p-value vector is denoted by P_{pilot} and P_{main}, respectively.

Table 3.1: Summary of the simulation's model parameters.

Parameter Set	π_0	μ_δ	$\mathbb{E}R_{\sigma^2}$	n_{pilot}	n_{main}
1	0.9	0.05, 0.07, 0.09, 0.11, 0.13	1, 1.5, 2, 2.5	20	50
2	0.9	0.05, 0.07, 0.09, 0.11, 0.13	2	10, 20, 30, 40	50

In the model (3.10), there are five model parameters π_0, μ_δ, μ_R, n_{pilot}, and n_{main}. We set π_0 and n_{main} to 0.9 and 50, respectively. The variables μ_δ, μ_R, and n_{pilot} are combined differently. Increases in δ_{jt} imply an increase in testing power for both *studies*, and its distribution is determined by μ_δ. The parameters μ_R and n_{pilot} are chosen to reflect the third data characteristic, that the main *study*'s testing power is greater than the pilot *study*'s testing power for most genes. As $R_{\sigma^2 j}$ increases or n_{pilot} decreases, the testing power of the pilot *study* declines, increasing the difference in testing power between the two *studies*. The value of μ_R is determined

via a more intuitive parameter $\mathbb{E}R_{\sigma^2}$ (because $\mathbb{E}R_{\sigma^2 j}$ is constant regardless of gene, j is omitted). Table 3.1 summarizes the five model parameters used in the simulation, which are generated with two constraints of $n_{pilot} \leq n_{main}$ and $\mathbb{E}R_{\sigma^2} \geq 1$. There are two distinct sets of scenarios in which one is fixed to examine the effect of μ_R and n_{pilot} separately.

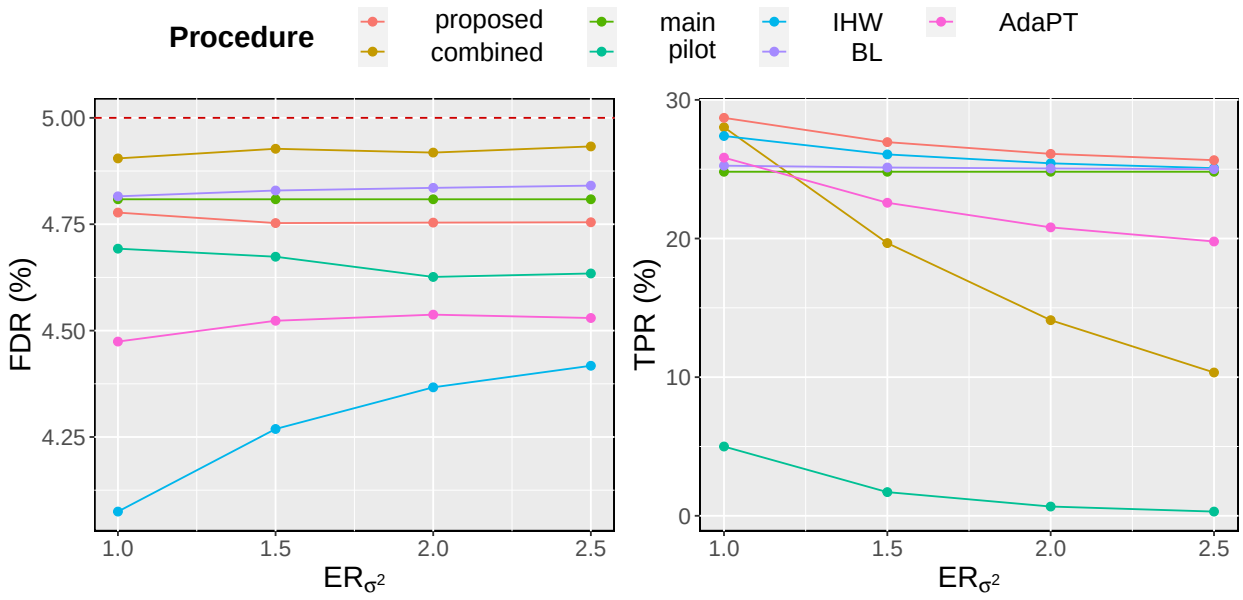


Figure 3.5: The FDR and TPR (%) graphs of the four procedures for scenarios of $n_{pilot} = 20$ and $\mu_\delta = 0.09$ in the set 1.

When making inferences about DE genes, the target FDR level is set to 0.05. The pre-specified conditions when applying our procedure are as follows: α^* is selected from a set of 1000 evenly spaced values between 0 and 1, and $\mathbb{E}_{rep}$ is calculated by 1000 times, replacing p-value pairs. The procedure is denoted by *proposed*. We compare the *proposed* procedure to six distinct procedures. First, the procedures that apply Storey's (2002) q-value method with the histogram-based π_0 estimator by Nettleton et al. (2006) to $P_{combined}$, P_{pilot}, and P_{main} are considered. The procedures are denoted by *combined*, *pilot*, and *main*, respectively, depending on the considered data type.

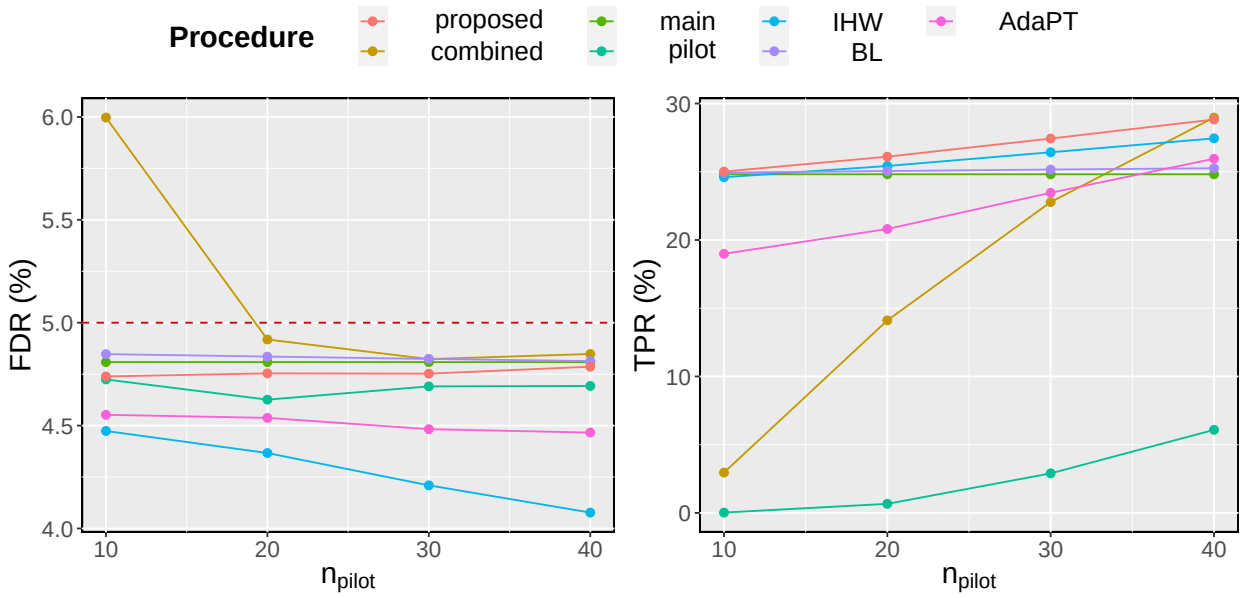


Figure 3.6: The FDR and TPR (%) graphs of the four procedures for scenarios of $ER_{\sigma^2} = 2$ and $\mu_\delta = 0.09$ in the set 2.

We also consider modern FDR-controlling methods utilizing covariate variables. Because P_{pilot} can be considered a covariate variable, the modern methods are also evaluated, with P_{main} serving as the primary source p-value and P_{pilot} serving as a covariate variable. We choose the methods of IHW, Boca and Leek (2018) (BL), and Lei and Fithian (2018) (AdaPT) based on the simulation results in Korthauer et al. (2019). The methods are implemented in the `R` packages `IHW`, `swfdr`, and `adaptMT`, respectively. Essentially, we adhere to the default package settings. For the AdaPT method, we employ the *adapt_glm* function with the settings described in the paper Korthauer et al. (2019). Moreover, the target FDR for the IHW and AdaPT methods is set to the nominal FDR level, which is 0.05 in the simulation study.

3.3.2 Simulation Results

The seven procedures are compared in terms of FDR and mean true positive rate (TPR). The TPR indicates the proportion of true positives among all DDE genes. For each scenario, $10,000$

data sets were generated for analysis purposes to approximate the mean values, referred to as empirical values. Note that when a procedure declares no significant hypotheses, the false discovery proportion is zero. Table 3.2 summarizes the simulation results. The FDR is effectively controlled at or below 0.05 for all considered scenarios and procedures, except the *combined* procedure. In particular, the *proposed* procedure maintains the FDR level, indicating that the calibration method is effective.

The TPR grows in all procedures as μ_δ increases, given that other model parameters remain constant. As illustrated in Figures 3.5 and 3.6, when n_{pilot} is small or ER_{σ^2} is large, the *main* procedure typically has a higher TPR value than the *combined* procedure. This phenomenon supports the practice of using only the main study data. Additionally, n_{pilot} and $\mathbb{E}R_{\sigma^2}$ have the same influence on the TPR in opposite directions, as seen in Figures 3.5 and 3.6. Given the t-test statistic's reciprocal relationship between variance and sample size, the phenomena are understandable.

In most scenarios, the *proposed* procedure has a higher TPR than all other procedures. As seen in Figures 3.5 and 3.6, our procedure's TPR line serves as the upper limit for other procedures' TPR lines. The proposed procedure outperforms all considered procedures, particularly *combined* and *main* procedures. Therefore, we may conclude that the *proposed* procedure combines the strengths of the *combined* and *main* procedures. In addition, all procedures including the *proposed* procedure that include P_{pilot} as a covariate variable demonstrate that the TPR increases as n_{pilot} increases or $\mathbb{E}R_{\sigma^2}$ decreases, indicating that P_{pilot} can be considered a useful covariate variable. In particular, the IHW procedure is the second best, understandable in that both the proposed and IHW methods are based on maximization of DDE genes.

3.4 Data Analysis

We evaluated the proposed method using RNA-seq data on the disease resistance of young, healthy pigs. Lim et al. (2021) provides a comprehensive description of the study's design and

testing of hypotheses, which is summarized as follows. At ~ 27 days of age, 912 F1 barrows from 15 batches were enrolled in the study. After three weeks in a healthy quarantine nursery, the piglets were exposed to natural polymicrobial diseases found on commercial farms. In addition to quantifying gene expression levels in blood samples from piglets, disease resilience phenotypes such as subjective health score, treatment rate, mortality, and growth rate were also measured. Although the paper Lim et al. (2021) tested numerous hypotheses, we focused on the association between gene expression and concurrent growth rate using blood samples collected during periods of quarantine nursery prior to disease exposure.

The analysis we conducted is summarized below. The gene expression in blood samples collected during quarantine nursery was quantified using 3'mRNA sequencing with a globin block. Using the data in Lim et al. (2021) and genes in the Ensembl database, we analyzed $10,858$ genes with a non-zero read count for at least 80% of the samples. In addition, we employed log-scale read counts normalized and adjusted for nuisance factors according to Lim et al. (2021). We followed the inference procedure outlined in Lim et al. (2021) except for separating the data into two profiling periods. The same inference procedure was conducted for each profiling period, summarized as follows. The growth rate of a pig was used as a common dependent variable. A p-value was calculated for each gene to determine whether the adjusted log2 transformed read count has a slope coefficient of zero. In total, we generated $10,858$ p-value pairs.

Figure 3.7 illustrates the density of the p-value pairs. A density greater than 1 indicates that the region contains more p-value pairs than expected. When the p-value of profiling period 2 is low, the observed density exceeds the expected value, indicating that profiling period 2 has many low p-values. In contrast, the p-values for profiling period 1 contain fewer low values. When the p-value of the first profiling period is less than 0.25, there are more low p-values in the second period, exhibiting the L-shaped rejection region. Numerous low p-values may be indicative of a strong testing power. According to the third data characteristic, we determine that the p-value of profiling period 2 has the main study's feature. As a result, we apply our approach to the data by

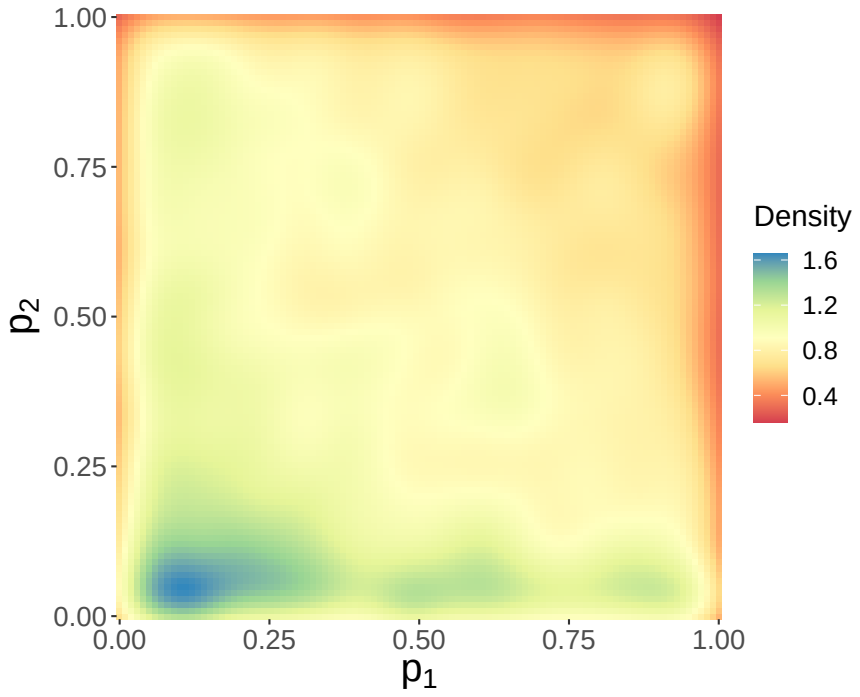


Figure 3.7: Density plot of the p-value of the second profiling period p_2 against the p-value of the first profiling period p_1.

considering the first profiling period as the pilot study and the second profiling period as the main study. In other words, we consider the p-value from the second profiling period as a covariate.

The seven procedures described in Section 3.3 are applied to the p-value pairs. For the proposed procedure, $\mathbb{E}_{rep}$ is calculated by 10,000 times, replacing p-value pairs. We considered the p-values obtained in Lim et al. (2021) as the combined procedure. Table 3.1 summarizes the number of significant tests at various nominal FDR levels. For each nominal FDR level, the target FDR level is modified when implementing the IHW, AdaPT, and proposed procedure. The pilot procedure declared only a few tests significant across all nominal FDR levels. Once more, it can be stated that the testing power of the first profiling period is lower than that of the second profiling period. The proposed procedure consistently declares more tests as significant than other procedures, including combined and main procedures. As illustrated in Figure 3.8, the proposed procedure provides a larger L-shaped rejection region as the target FDR level increases. Regardless of the target FDR level, the main p-value rejection threshold changes when the

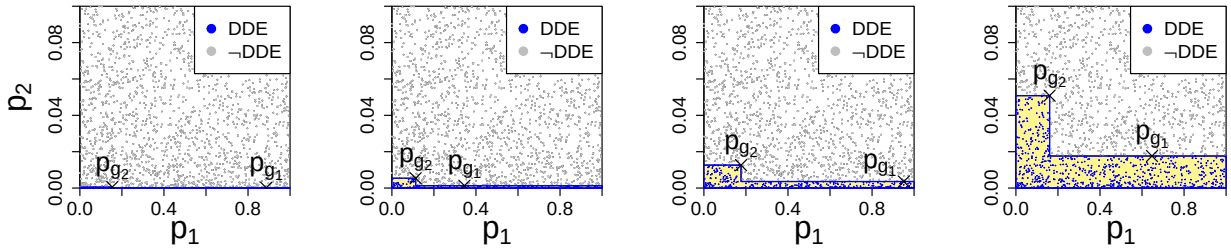


Figure 3.8: The proposed method is applied to the p-value pairs using different nominal FDR levels, and the chosen L-shaped rejection regions are visualized in the four scatterplots of the p-value pairs with a vertical axis limit of 0.1. The rejection regions correspond to nominal FDR levels of 0.01, 0.05, 0.1, and 0.2, from left to right.

covariate p-value is approximately 0.2. By examing Figure 3.7, we can see that our approach can be understood to increase the inference power by focusing our rejection region on the region where the density of the p-value pair is high. In contrast, in most instances, modern FDR-controlling procedures, like IHW, BL, and AdaPT, declare fewer tests as significant than the main procedure.

3.5 Discussion

Even though the proposed method offers substantial improvements over existing methods, there is still potential for improvement. First, our method provides a DDE gene list for a specified FDR level, which is a disadvantage compared to the existing methods that provide adjusted p-values or q-values. This may not be a major concern if we only want the DDE gene list for a given FDR level, but there is space for improvement in this area. Next, the proposed rejection region can be further generalized, and we can consider a sigmoid function to define the region. Additional research is required for the type of rejection region. At the same time, we should explore the covariate suited for such rejection regions. We hope our paper may inspire other scholars and be implemented in various fields.

3.6 Acknowledgment

The data used in this study were generated with funding from USDA National Institute of Food and Agriculture grant 2017-67007-26144, Genome Canada, Genome Alberta, and PigGen Canada.

3.7 References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Boca, S. M. and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6:e6035.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):1–21.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Lim, K.-S., Cheng, J., Putz, A., Dong, Q., Bai, X., Beiki, H., Tuggle, C. K., Dyck, M. K., Fortin, F., Harding, J., et al. (2021). Quantitative analysis of the blood transcriptome of young healthy pigs and its relationship with subsequent disease resilience. *BMC Genomics*, 22(1):1–18.

Nettleton, D., Hwang, J. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337–356.

Orr, M., Liu, P., and Nettleton, D. (2012). Estimating the number of genes that are differentially expressed in both of two independent experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(4):583–600.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035.

3.8 Appendix

Lemma : $L^*_{\widehat{\mathrm{FDR}}_1(L^*_a|\hat{m}_0)} = L^*_a$ almost surely.

Proof. Let $\tilde{\mathcal{L}}_a$ indicate a collection of all L-shape rejection regions holding a constraint of $\widehat{\mathrm{FDR}}_1(\cdot \mid \hat{m}_0) \leq a$. By definition of L^*_a, the best L-shape rejection region L^*_a also holds the constraint $\widehat{\mathrm{FDR}}_1(L^*_a \mid \hat{m}_0) \leq a$. As a result, we can show that $\tilde{\mathcal{L}}_{\widehat{\mathrm{FDR}}_1(L^*_a|\hat{m}_0)} \subseteq \tilde{\mathcal{L}}_a$. From the definition of $\tilde{\mathcal{L}}_{\widehat{\mathrm{FDR}}_1(L^*_a|\hat{m}_0)}$ and an obvious fact of $\widehat{\mathrm{FDR}}_1(L^*_a \mid \hat{m}_0) \leq \widehat{\mathrm{FDR}}_1(L^*_a \mid \hat{m}_0)$, we can show that $L^*_a \in \tilde{\mathcal{L}}_{\widehat{\mathrm{FDR}}_1(L^*_a|\hat{m}_0)}$. Because the best rejection region L^*_a for a larger set $\tilde{\mathcal{L}}_a$ is also included in the smaller set $\tilde{\mathcal{L}}_{\widehat{\mathrm{FDR}}_1(L^*_a|\hat{m}_0)}$, L^*_a is still the best rejection region in $\tilde{\mathcal{L}}_{\widehat{\mathrm{FDR}}_1(L^*_a|\hat{m}_0)}$, in perspective of DDE genes number and $\widehat{\mathrm{FDR}}_1$. By combining the previous result with the Assumption 3.2.3, we can conclude that $L^*_{\widehat{\mathrm{FDR}}_1(L^*_a|\hat{m}_0)} = L^*_a$ almost surely.

Theorem : Let $\tilde{\alpha} = \max\left\{\widehat{\mathrm{FDR}}_1(L^*_a \mid \hat{m}_0) : a \in [0,1],\ \mathrm{FDP}(L^*_a \mid H_0) \leq \alpha\right\}$. Then, $\mathrm{FDP}(L^*_{\tilde{\alpha}} \mid H_0) \leq \alpha$ almost surely..

Proof. Define $\mathcal{F}_\alpha = \left\{\widehat{\mathrm{FDR}}_1(L^*_a \mid \hat{m}_0) : a \in [0,1],\ \mathrm{FDP}(L^*_a \mid H_0) \leq \alpha\right\}$. From the definition of $\tilde{\alpha}$, $\tilde{\alpha}$ can be expressed as $\max(\mathcal{F}_\alpha)$. Because $\mathcal{F}_\alpha$ is not an empty set ($0 \in \mathcal{F}_\alpha \because L^*_0 = \emptyset$), $\tilde{\alpha} = \widehat{\mathrm{FDR}}_1(L^*_{\tilde{a}} \mid \hat{m}_0)$ for some $\tilde{a} \in [0,1]$ such that $\mathrm{FDP}(L^*_{\tilde{a}} \mid H_0) \leq \alpha$. From the Assumption 3.2.3, $L^*_{\tilde{a}}$ satisfying $\widehat{\mathrm{FDR}}_1(\cdot \mid \hat{m}_0) = \tilde{\alpha}$ is uniquely determined almost surely. Without loss of generality, we can select a $\tilde{a}$ such that $\tilde{\alpha} = \widehat{\mathrm{FDR}}_1(L^*_{\tilde{a}} \mid \hat{m}_0)$. From the lemma and the fact that $\tilde{\alpha} = \widehat{\mathrm{FDR}}_1(L^*_{\tilde{a}} \mid \hat{m}_0)$, we can show that $L^*_{\tilde{\alpha}} = L^*_{\tilde{a}}$ almost surely. Because $\mathrm{FDP}(L^*_{\tilde{a}} \mid H_0) \leq \alpha$ and $L^*_{\tilde{\alpha}} = L^*_{\tilde{a}}$ almost surely, we can conclude that $\mathrm{FDP}(L^*_{\tilde{\alpha}} \mid H_0) \leq \alpha$ almost surely.

Table 3.2: The empirical TPR (%) with corresponding empirical FDR (%) in parentheses of the seven procedures for all scenarios of set 1 and 2.

μ_δ	$\mathbb{E}R_{\sigma^2}$	n_{pilot}	**proposed**	**combined**	**main**	**pilot**	**IHW**	**BL**	**AdaPT**
0.05	1.00	20	6.4 (4.76)	6.04 (4.89)	4.83 (4.68)	0.29 (4.44)	5.79 (4.28)	4.91 (4.7)	5.19 (3.75)
·	1.50	20	5.66 (4.73)	3.11 (4.98)	4.83 (4.68)	0.07 (4.25)	5.27 (4.4)	4.89 (4.72)	3.68 (3.6)
·	2.00	20	5.31 (4.69)	1.69 (4.99)	4.83 (4.68)	0.03 (4.46)	5.03 (4.43)	4.88 (4.73)	2.81 (3.38)
·	2.50	20	5.13 (4.71)	0.98 (5.05)	4.83 (4.68)	0.02 (4.48)	4.91 (4.46)	4.87 (4.73)	2.28 (3.22)
0.07	1.00	20	16.33 (4.76)	15.78 (4.88)	13.4 (4.77)	1.53 (4.61)	15.35 (4.2)	13.64 (4.78)	14.12 (4.23)
·	1.50	20	14.97 (4.71)	9.77 (4.92)	13.4 (4.77)	0.42 (4.43)	14.32 (4.3)	13.57 (4.8)	11.57 (4.21)
·	2.00	20	14.35 (4.74)	6.25 (4.99)	13.4 (4.77)	0.15 (4.52)	13.83 (4.39)	13.53 (4.8)	10.14 (4.18)
·	2.50	20	13.98 (4.71)	4.11 (5)	13.4 (4.77)	0.07 (4.59)	13.56 (4.4)	13.51 (4.8)	9.32 (4.19)
0.09	1.00	20	28.71 (4.78)	28.02 (4.91)	24.81 (4.79)	5 (4.67)	27.4 (4.06)	25.25 (4.8)	25.83 (4.46)
·	1.50	20	26.95 (4.75)	19.66 (4.92)	24.81 (4.79)	1.7 (4.73)	26.06 (4.27)	25.12 (4.81)	22.58 (4.53)
·	2.00	20	26.11 (4.75)	14.11 (4.93)	24.81 (4.79)	0.66 (4.64)	25.42 (4.35)	25.05 (4.82)	20.8 (4.52)
·	2.50	20	25.66 (4.75)	10.33 (4.93)	24.81 (4.79)	0.3 (4.59)	25.07 (4.41)	25 (4.83)	19.78 (4.52)
0.11	1.00	20	40.64 (4.83)	39.91 (4.93)	36.4 (4.86)	10.74 (4.78)	39.1 (4.03)	37 (4.88)	37.47 (4.59)
·	1.50	20	38.75 (4.82)	30.51 (4.92)	36.4 (4.86)	4.68 (4.67)	37.65 (4.22)	36.82 (4.88)	33.93 (4.64)
·	2.00	20	37.85 (4.81)	23.71 (4.94)	36.4 (4.86)	2.17 (4.71)	36.97 (4.33)	36.73 (4.89)	32.13 (4.68)
·	2.50	20	37.34 (4.8)	18.62 (4.94)	36.4 (4.86)	1.07 (4.64)	36.57 (4.38)	36.66 (4.9)	31.17 (4.67)
0.13	1.00	20	50.83 (4.86)	50.13 (4.95)	46.65 (4.89)	17.93 (4.8)	49.11 (3.96)	47.34 (4.91)	47.58 (4.68)
·	1.50	20	48.97 (4.85)	40.62 (4.97)	46.65 (4.89)	9.28 (4.77)	47.72 (4.19)	47.15 (4.92)	44.05 (4.72)
·	2.00	20	48.09 (4.83)	33.32 (4.94)	46.65 (4.89)	5.04 (4.7)	47.07 (4.29)	47.03 (4.93)	42.21 (4.74)
·	2.50	20	47.6 (4.84)	27.6 (4.95)	46.65 (4.89)	2.85 (4.66)	46.7 (4.35)	46.96 (4.93)	41.44 (4.77)
0.05	2.00	10	4.87 (4.7)	0.15 (7.95)	4.83 (4.68)	0 (4.59)	4.76 (4.49)	4.87 (4.74)	1.48 (2.83)
·	2.00	20	5.31 (4.69)	1.69 (4.99)	4.83 (4.68)	0.03 (4.46)	5.03 (4.43)	4.88 (4.73)	2.81 (3.38)
·	2.00	30	5.86 (4.7)	4.15 (4.79)	4.83 (4.68)	0.15 (4.5)	5.41 (4.4)	4.9 (4.72)	4.14 (3.64)
·	2.00	40	6.45 (4.78)	6.52 (4.75)	4.83 (4.68)	0.42 (4.5)	5.82 (4.3)	4.91 (4.71)	5.27 (3.75)
0.07	2.00	10	13.51 (4.7)	0.78 (6.7)	13.4 (4.77)	0.01 (4.52)	13.23 (4.48)	13.47 (4.82)	8.3 (4.13)
·	2.00	20	14.35 (4.74)	6.25 (4.99)	13.4 (4.77)	0.15 (4.52)	13.83 (4.39)	13.53 (4.8)	10.14 (4.18)
·	2.00	30	15.36 (4.74)	11.93 (4.84)	13.4 (4.77)	0.8 (4.63)	14.59 (4.31)	13.59 (4.79)	12.26 (4.22)
·	2.00	40	16.43 (4.75)	16.58 (4.78)	13.4 (4.77)	2.07 (4.71)	15.38 (4.19)	13.64 (4.78)	14.28 (4.24)
0.09	2.00	10	25.02 (4.73)	2.95 (5.98)	24.81 (4.79)	0.03 (4.67)	24.61 (4.46)	24.94 (4.83)	18.98 (4.55)
·	2.00	20	26.11 (4.75)	14.11 (4.93)	24.81 (4.79)	0.66 (4.64)	25.42 (4.35)	25.05 (4.82)	20.8 (4.52)
·	2.00	30	27.43 (4.75)	22.79 (4.83)	24.81 (4.79)	2.9 (4.68)	26.43 (4.2)	25.16 (4.81)	23.47 (4.49)
·	2.00	40	28.83 (4.79)	28.98 (4.85)	24.81 (4.79)	6.09 (4.7)	27.45 (4.07)	25.26 (4.8)	25.96 (4.47)
0.11	2.00	10	36.63 (4.79)	7.23 (5.68)	36.4 (4.86)	0.08 (4.86)	36.05 (4.45)	36.56 (4.9)	30.95 (4.73)
·	2.00	20	37.85 (4.81)	23.71 (4.94)	36.4 (4.86)	2.17 (4.71)	36.97 (4.33)	36.73 (4.89)	32.13 (4.68)
·	2.00	30	39.27 (4.82)	34.08 (4.89)	36.4 (4.86)	6.93 (4.72)	38.05 (4.17)	36.87 (4.88)	34.86 (4.63)
·	2.00	40	40.78 (4.84)	40.87 (4.88)	36.4 (4.86)	12.32 (4.75)	39.15 (4.01)	37.01 (4.88)	37.57 (4.59)
0.13	2.00	10	46.89 (4.83)	13.24 (5.45)	46.65 (4.89)	0.23 (4.7)	46.21 (4.42)	46.84 (4.93)	41.67 (4.79)
·	2.00	20	48.09 (4.83)	33.32 (4.94)	46.65 (4.89)	5.04 (4.7)	47.07 (4.29)	47.03 (4.93)	42.21 (4.74)
·	2.00	30	49.49 (4.84)	44.26 (4.91)	46.65 (4.89)	12.58 (4.75)	48.09 (4.12)	47.21 (4.92)	45.01 (4.71)
·	2.00	40	50.93 (4.84)	51.02 (4.91)	46.65 (4.89)	19.79 (4.82)	49.16 (3.96)	47.35 (4.92)	47.7 (4.68)

Table 3.3: The number of tests declared to be significant for different nominal FDR levels and procedures.

nominal FDR	**proposed**	**combined**	**main**	**pilot**	**IHW**	**BL**	**AdaPT**
0.01	188	126	181	0	177	182	0
0.05	314	254	298	0	290	297	308
0.10	450	429	419	0	391	424	413
0.20	894	852	707	3	565	704	648

CHAPTER 4. CASE-SPECIFIC TUNING FOR RANDOM FOREST REGRESSION

Hyeongseon Jeon[1] and Dan Nettleton[1]

[1]Department of Statistics, Iowa State University, Ames, IA 50011, USA

Modified from a manuscript to be submitted to *Statistical Analysis and Data Mining*

Abstract

The Random Forest (RF) method is a well-known machine learning algorithm for prediction. The performance of the method depends on the values of a few tuning parameters. This paper suggests a unique case-specific tuning algorithm for RF regression. We provide a simple example to demonstrate that the values of the tuning parameters that lead to the best of the predictor domain. Thus, case-specific tuning can be advantageous. A simulation study reveals that our strategy for case-specific tuning may be a good alternative to existing approaches for selecting tuning parameter values, especially when the signal-to-noise ratio is relatively high. In addition, in real data analysis, our approach provides competitive results compared to existing methods.

4.1 Introduction

Since the RF method was proposed by Breiman (2001), it has been heavily used in practice to make predictions and extensively studied in statistics and machine learning literature. Several articles, such as Boulesteix et al. (2012), Biau and Scornet (2016), and Probst and Boulesteix (2017), have reviewed the RF method.

According to Lin and Jeon (2006), the RF method can be viewed as an adaptive nearest-neighbors method. The RF prediction of the response value as a function of a predictor vector is simply a weighted average of the response values in the training data set. The weights

on the training response values are determined by the proximity of each predictor vector's value in the training data set to the value of the predictor vector targeted for prediction. The biggest weights are assigned to training response values whose predictor vector values are in close proximity to the target value. The RF algorithm automatically accounts for the underlying relationship between the predictor vector and the response while determining the proximity weights. Predictor variables that are only weakly associated or unassociated with the response play little to no role in determining the proximity weights, while predictor variables strongly associated with the response have a large impact on the RF assessment of proximity. Several authors have described RF proximity weighting and used it for various purposes. For example, see Meinshausen (2006). Xu et al. (2016), Zhang et al. (2019), and Friedbergå et al. (2020)

To implement the RF method in actual data analysis, specifying the values of several tuning parameters is necessary. From the statistical point of view, the two most interesting tuning parameters are, arguably, mtry and nodesize. These tuning parameters are interesting because the best values to choose, in terms of minimizing mean squared prediction error (MSPE), depend on unknowns specific to a given data set. The values of the tuning parameters mtry and nodesize tradeoff variance and bias of the RF predictor by their impact on RF proximity weights. Large values of mtry and small values of nodesize tend to concentrate weights on the response values of a few nearby cases, thereby reducing bias at the cost of larger predictor variance. Small values of mtry and large values of nodesize, on the other hand, tend to place positive weights more widely and uniformly over the training response values, potentially introducing bias in exchange for a reduction in predictor variance.

Cross-validation is perhaps one of the most common strategies for selecting values of mtry and nodesize to minimize MSPE. For random forests and other methods relying on bootstrap aggregation (i.e., bagging) cross-validation is carried out conveniently and efficiently by minimizing the sum of the squared out-of-bag (OOB) prediction errors over a grid of tuning parameter value choices. For readers unfamiliar with OOB prediction, the idea is as follows. The RF algorithm starts by selecting B bootstrap samples from the training data set. The RF

prediction for a target predictor vector is obtained by averaging the predictions of B random trees, each constructed using one of the B bootstrap samples. According to Breiman (2001), the OOB prediction of the ith response value is obtained by averaging the predictions from the random trees built from the bootstrap samples that exclude the ith training case. Because the i training case is excluded from a bootstrap sample with probability $(\frac{n-1}{n})^n \approx e^{-1} \approx 0.368$, there will be, with high probability, a subforest of trees built without knowledge of ith training case from which the OOB prediction is obtained. The OOB prediction error is the difference between the ith response value and its OOB prediction. It is straightforward to obtain each OOB prediction error for a variety of tuning parameter values.

In this paper, we point out that the best values for mtry and nodesize may depend on the target value of the predictor vector. Thus, it may be advantageous to choose the values of mtry and nodesize in a case-specific way. When the target for prediction lies in a region of the predictor vector domain where the true underlying mean function is relatively flat, assigning positive weights to many nearby observations is likely to reduce MSPE by diminishing predictor variance without introducing substantial bias. On the other hand, if the target value lies in a region of the predictor domain where the true underlying mean function changes relatively quickly, weights may need to be more concentrated on a small number of nearby cases to avoid introducing substantial bias at the cost of higher prediction variance. One illustrative example is provided in Figure 4.1. When X_1 is less than 0.75, the mean function is flat, whereas when X_1 is greater than 0.75, the mean function changes relatively quickly. Therefore, when X_1 is relatively small, selecting the tuning parameter in the direction of reducing the predictor's variance will help reduce MSPE. When considering the bias-variance tradeoff associated with nodesize, the predictor's variance can be diminished when nodesize is large. Consequently, when X_1 belongs to the first four interval groups (defined on the horizontal axis of the second panel of Figure 4.1), MSPE is smaller when nodesize is 20 than when nodesize is 1. In contrast, the converse occurs when X_1 belongs to the fifth interval group.

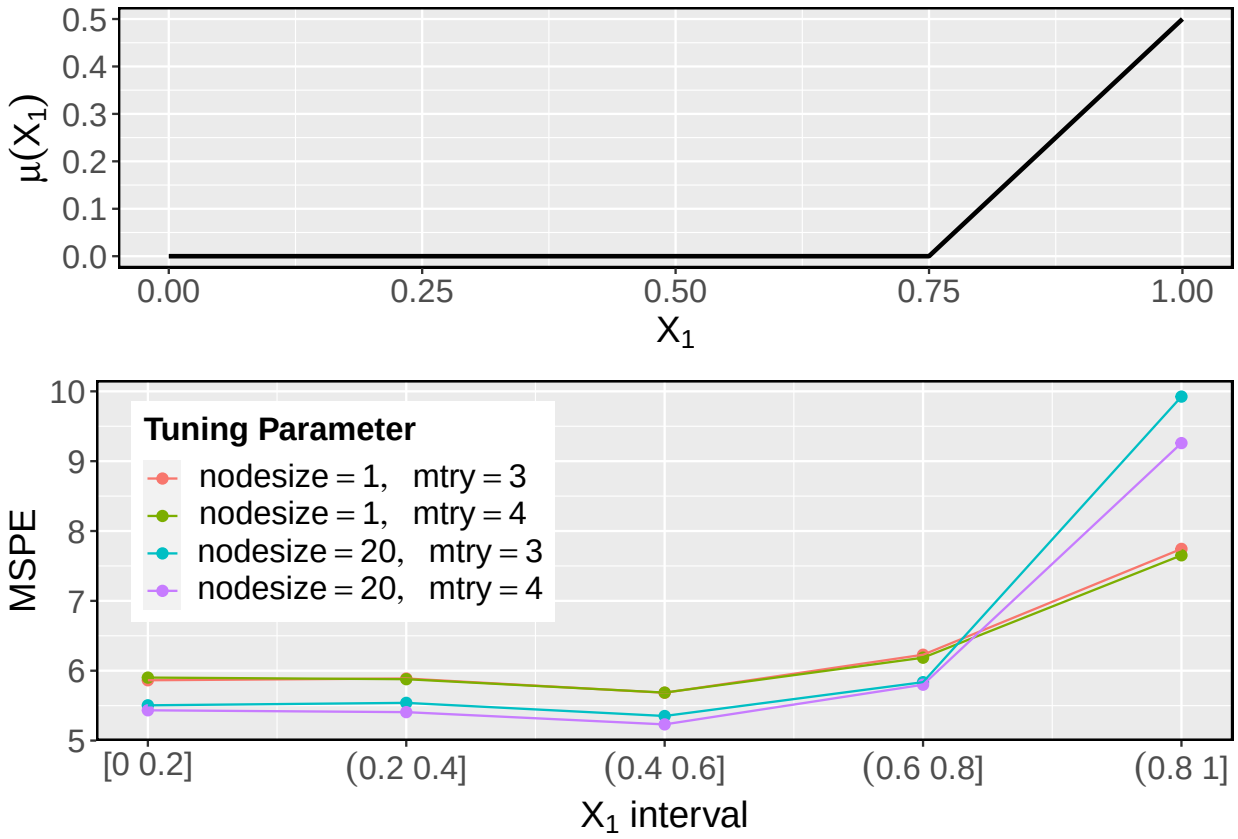


Figure 4.1: The data generating model is $Y = \mu(X_1) + \epsilon$, with five i.i.d. predictor variables from U(0,1), where $\mu(X_1)$ is illustrated in the graph above. RF is implemented using the specified values of nodesize and mtry, and MSPE is computed for five intervals defined by X_1 values. The MSPE values are depicted in the graph below.

Rather than finding the values of mtry and nodesize that minimize the sum of all squared OOB prediction errors, our idea for case-specific tuning of random forests relies on minimizing a weighted sum of squared OOB prediction errors. We use RF proximity weights to pay the most attention to the OOB prediction errors of cases that are in closest proximity to the target value of the predictor vector. In this way, we use the observed data to infer whether the target lies in a region of the predictor domain where averaging over response values of many nearest cases is advantageous or whether the complexity of the mean function in the target region requires a more refined and narrow focus on only a relatively few of the nearest cases to improve MSPE.

The remainder of this paper is organized as follows. Section 4.2 describes the RF algorithm, its tuning parameters, and our novel tuning algorithm. Section 4.3 and 4.4 demonstrate the effectiveness of our approach through a simulation study and data analysis, respectively. Section 4.5 examines our approach's development potential.

4.2 Method Proposal

4.2.1 Random Forest Prediction

The following describes the classic RF algorithm, proposed by Breiman (2001). Let $\mathbb{Z} = \{\mathbf{Z}_i = (\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$ represent training data set, where $\mathbf{X}_i$ and Y_i are ith p-dimensional predictor vector value and response variable value, respectively. RF generates B bootstrap samples from $\mathbb{Z}$. For a tuning parameter vector $\boldsymbol{\lambda}$, a random regression tree is constructed for the bth bootstrap sample $\mathbb{Z}_b$. A set of split rules outlined below determines the regression tree. Each tree node is specified by a hyperrectangle sequentially defined by the split rules, and the initial node's hyperrectangle is set to p-dimensional Euclidean space. For a given node, a split rule is applied to the subset of the observations in $\mathbb{Z}_b$ whose predictor vector observations fall within the hyperrectangle if the number of observations in the subset exceeds the nodesize. If the number of observations in the subset is less than or equal to the nodesize, the node is considered a terminal node and is not split further. The split rule is described as follows. Initially, mtry candidate variables are selected at random. For each candidate variable, the split rule is selected that minimizes the sum of squared deviations of each response variable value from its subnode mean response. As the node's split rule, the rule with the minimum sum of squared deviations across candidate variables is selected. The hyperrectangle for the next two nodes is created by dividing the hyperrectangle of the current node using the split rule.

Following the construction of the random regression tree using the bth bootstrap sample, the tree is used to predict the response at a target point $\mathbf{X}_0$. A terminal node is first identified relative to the target point following the set of split rules. Then, the tree prediction is obtained by averaging the response values of training observations in the terminal node. Let $T(\mathbf{X}_0 : \mathbb{Z}_b, \boldsymbol{\lambda})$

denote the tree prediction using bth bootstrap sample at a target point $\mathbf{X}_0$ with tuning parameter value $\boldsymbol{\lambda}$. The RF prediction at $\mathbf{X}_0$ is

$$\hat{Y}^{\boldsymbol{\lambda}}(\mathbf{X}_0) = \frac{\sum_{b=1}^{B} T(\mathbf{X}_0 : \mathbb{Z}_b, \boldsymbol{\lambda})}{B} \tag{4.1}$$

$$= \frac{\sum_{i=1}^{n} W_i(\mathbf{X}_0 : \mathbb{Z}_1, \ldots, \mathbb{Z}_B, \boldsymbol{\lambda}) \times Y_i}{n}, \tag{4.2}$$

where $\left\{W_i(\mathbf{X}_0 : \mathbb{Z}_1, \ldots, \mathbb{Z}_B, \boldsymbol{\lambda})\right\}_{i=1}^{n}$ are nonnegative proximity weights that sum to 1. As described in Lin and Jeon (2006), the proximity weight $W_i(\mathbf{X}_0 : \mathbb{Z}_1, \ldots, \mathbb{Z}_B, \boldsymbol{\lambda})$ in (4.2) can be employed as a measure of the proximity between $\mathbf{X}_0$ and $\mathbf{X}_i$.

4.2.2 Tuning Parameters

According to Probst et al. (2019), RF method contains several hyperparameters that can be regarded as tuning parameters. Most hyperparameters are utilized when generating random trees. The hyperparameters include the number of randomly selected candidate variables (mtry), the maximum number of observations per a terminal node (nodesize), the sampling procedure including sample size for producing bootstrap samples, and the number of random trees. Xu et al. (2016) successfully improves prediction by adjusting the sampling procedure for each target point. Even though nodesize and mtry are considered tuning parameters in this paper, our method is still valid when a different combination of hyperparameters determines the tuning parameter.

4.2.3 Standard Tuning Algorithm

To assess RF's predictive performance, cross-validation is a suitable method, which is a time-consuming process, especially when dealing with large data sets. As an alternative, to assess RF performance, OOB predictions are typically employed. The OOB prediction $\hat{Y}_{(i)}^{\boldsymbol{\lambda}}$ for ith training data $\mathbf{Z}_i$, is the prediction only of the trees built without $\mathbf{Z}_i$ in their corresponding bootstrap sample. Let O_{bi} be 1 if the ith training observation is left out of the bth bootstrap sample and 0 otherwise. The OOB prediction is expressed as

$$\hat{Y}_{(i)}^{\boldsymbol{\lambda}} = \frac{\sum_{b=1}^{B} O_{bi} \times T(\mathbf{X}_i : \mathbb{Z}_b, \boldsymbol{\lambda})}{\sum_{b=1}^{B} O_{bi}}. \tag{4.3}$$

and the OOB prediction error is defined as $Y_i - \hat{Y}^{\boldsymbol{\lambda}}_{(i)}$. Average squared OOB prediction error (ASOOBPE) is defined as $\frac{\sum_{i=1}^{n}\{Y_i - \hat{Y}^{\boldsymbol{\lambda}}_{(i)}\}^2}{n}$. The standard RF tuning algorithm selects tuning parameter values $\boldsymbol{\lambda}$ that minimizes ASOOBPE among a set of candidate choices for $\boldsymbol{\lambda}$.

4.2.4 Case-Specific Tuning Algorithm

The standard RF tuning algorithm chooses a vector of tuning parameter values $\boldsymbol{\lambda}^*$ that minimizes OOB ASPE and uses the $\boldsymbol{\lambda}^*$ when predicting responses at any target point. Our case-specific tuning (CST) algorithm recommends a case-specific tuning parameter $\boldsymbol{\lambda}^*(\mathbf{X}_0)$ when predicting the response at a target point $\mathbf{X}_0$, employing weighted average squared OOB prediction error (WASOOBPE) with proximity weights. The following describes the CST algorithm.

Case-Specific Tuning Algorithm

1. For a target point $\mathbf{X}_0$, generate proximity weights $\{W_i(\mathbf{X}_0 : \mathbb{Z}_1, \ldots, \mathbb{Z}_B, \boldsymbol{\lambda}_1)\}_{i=1}^n$, using a predetermined tuning parameter $\boldsymbol{\lambda}_1$.

2. Choose a tuning parameter $\boldsymbol{\lambda}^*(\mathbf{X}_0)$ minimizing a weighted average squared OOB prediction error with the weights generated in step 1:
   $\boldsymbol{\lambda}^*(\mathbf{X}_0) = \arg\min_{\boldsymbol{\lambda}_2} \sum_{i=1}^{n} W_i(\mathbf{X}_0 : \mathbb{Z}_1, \ldots, \mathbb{Z}_B, \boldsymbol{\lambda}_1) \times \{Y_i - \hat{Y}^{\boldsymbol{\lambda}_2}_{(i)}\}^2$.

3. Use the tuning parameter $\boldsymbol{\lambda}^*(\mathbf{X}_0)$ for the RF prediction at $\mathbf{X}_0$.

The tuning parameter $\boldsymbol{\lambda}_1$ in step 1 can either be set to default value or be selected by cross-validation, minimizing test error. Typical default values of mtry and nodesize in the RF regression are $\max\{1, \lfloor p/3 \rfloor\}$ and 5, respectively.

Alternatively, the CST algorithm can be implemented by replacing the proximity weights with typical kernel weights. However, we present the CST algorithm with proximity weights for the following two reasons. First, while implementing RF, we normally include a large number of covariate variables, and therefore typical kernel weights may suffer from high-dimensionality problems. According to Friedbergå et al. (2020), proximity weights can alleviate problems

regarding high dimensionality, compared to typical kernel weights. In addition, the classic tuning algorithm also generates random trees to evaluate the prediction error for different tuning parameter combinations. Since the generated random trees can be used to calculate proximity weights, our method can be implemented efficiently.

4.3 Simulation Study

Table 4.1: A description of the five simulation scenario-building mean functions. Each function is made up of a combination of linear and nonlinear components. In particular, Friedman (1991) introduced the third model for the $p = 10$ case, which has been utilized in multiple publications [e.g., Friedbergå et al. (2020), Xu et al. (2016).]

Model ID (j)	$\mu_j(\cdot)$
1	$10\,X_{1i} + 10\,X_{2i} + 5\,X_{3i}X_{4i}$
2	$8\sin(2\pi X_{1i}) + 6\,X_{2i} + 3\,X_{3i} + 2\,X_{2i}X_{3i}$
3	$10\sin(\pi X_{1i}X_{2i}) + 20\,(X_{3i} - 0.5)^2 + 10\,X_{4i} + 5\,X_{5i}$
4	$\log\{1 + e^{20(X_{1i}-0.5)}\} + \frac{1}{X_{2i}X_{3i}+0.1} + 5\,X_{4i}$
5	$e^{2X_{1i}X_{2i}+1.5X_{3i}}$

To evaluate our method's performance, we conduct a simulation study. The data model is based on the additive error model $Y_i = \mu(\mathbf{X}_i) + \epsilon_i$. $\mathbf{X}_i$ and ϵ_i are independently generated from $\mathrm{U}(0,1)^p$ and $\mathrm{N}(0, \sigma_\epsilon^2)$, respectively, for each observation i. We set p equal to 10 and select σ_ϵ from $\{0.1, 1, 10\}$. We consider five mean functions $\mu(\cdot)$, provided in Table 4.1. Each simulation scenario is defined by a combination of $\mu(\cdot)$ and σ_ϵ. For each scenario, 1000 simulation runs were conducted. The model generates 1000 training observations and 100 test observations in each run independently. The test data are utilized to assess a procedure's prediction error. In detail, MSPE is computed using the test data and then averaged over the 1000 simulation runs for a given procedure, referred to as $\overline{\mathrm{MSPE}}$.

We compare seven comparable tuning procedures for RF, including two from our method. When implementing all procedures, the number of random trees is fixed to 500. Regarding our method, we employ a grid search strategy while selecting tuning parameter values for $\boldsymbol{\lambda}_1$ in step 1 and $\boldsymbol{\lambda}_2$ in step 2. Inspired by `randomForestSRC` `R` package, mtry and nodesize values are selected from $\{1, 2, \ldots, p\}$ and $\{1, 2, \ldots, 9, 10, 20, \ldots, 100\}$, respectively. $\boldsymbol{\lambda}_1$ is determined through 10-fold cross-validation or its default value stated in Section 4.2.4. Depending on whether cross-validation is employed, the two procedures are referred to as CST.RF and cv.CST.RF. The descriptions and abbreviations of the remaining five procedures are provided in Table 4.2. Essentially, the procedures are implemented according to each package's default configuration. When implementing case-specific RF method (CS.RF), nodesize is set to 10 in the RF to create bootstrap sampling probabilities, while other tuning parameter settings use the RF's default values. See Xu et al. (2016) for details.

Table 4.2: This table contains the seven procedures' abbreviations, used `R` packages, used functions in the package, and brief descriptions. The tuneRanger method combines all useful features of the ranger, mlrMBO, and mlr `R` packages, where mlrMBO package is built on sequential model-based optimization. Details can be found at Probst et al. (2019).

Name	`R` Package	Used Function	Brief Description
default	`randomForest`	randomForest	*node size* $= 5$ & *mtry* $= \max\{1, \lfloor p/3 \rfloor\}$
CS.RF	`randomForest`	·	For each data point, case-specific RF is applied.
tuneRF	`randomForest`	tuneRF & randomForest	*node size* $= 5$ & *mtry* minimizing ASOOBPE
tune.rfsrc	`randomForestSRC`	tune & rfsrc	*node size* & *mtry* minimizing ASOOBPE
tuneRanger	`tuneRanger`	makeRegrTask & tuneRanger	All useful features of several `R` packages are implemented.
(cv.)CST.RF	`randomForestSRC`	·	For each data point, CST-algorithm is applied.

The simulation results are summarized in Table 4.5. The following analysis is focused on the tuning algorithms for mtry and nodesize. In terms of $\overline{\text{MSPE}}$, when the error variance is relatively small ($\sigma_\epsilon = 0.1$ or 1), alternative approaches are usually superior to the default procedure. Therefore, we may conclude that tuning nodesize and mtry is advantageous when the error variance appears small. At the same time, in the scenarios ($\sigma_\epsilon = 0.1$ or 1), our methods outperform other procedures. This effect appears to result from the fact that when the variance is

smaller, the proximity weight identifies the nearest data point more precisely. As illustrated in Figure 4.2, cv.CST.RF procedure consistently provides the minimum prediction error in the least error variance scenarios. Almost every value on the x- and y-axes of the scatterplots is negative, leading us to conclude that the tuning algorithms outperform the default approach in most simulation runs. In addition, the greater number of dots in the top left corner suggests that the cv.CST.RF procedure consistently outperforms alternative tuning approaches. When the error variance is rather large ($\sigma_\epsilon = 10$), tuneRanger has the minimum prediction error compared to all other procedures. Regardless of the scenarios, tuneRanger and our method work well. Additionally, cv.CST.RF consistently outperforms CST.RF, although the difference in $\overline{\text{MSPE}}$ is not as substantial. Thus, our method with a default setting is also a good alternative.

4.4 Data Analysis

Table 4.3: Summary of preprocessed data sets that we analyzed. All data sets were obtained from the UCI machine learning repository (https://archive.ics.uci.edu). For each data set, observations with missing values and predictor variables with many levels were excluded from our analysis. The response variable was chosen based on the repository's descriptions. We analyzed the energy efficiency data using two distinct response variables and treat these as separate data sets.

Data set	**Data name**	**# of observations**	**# of predictors**	**Response variable**
1	Airfoil self-noise	1503	5	Scaled sound pressure level
2	Auto MPG	392	7	mpg
3	Concrete compressive strength	1030	8	Concrete compressive strength
4	Energy efficiency	768	8	Y_1
5	Energy efficiency	768	8	Y_2
6	Forest fires	517	11	log(area+1)
7	QSAR aquatic toxicity	546	8	LC50
8	QSAR bioconcentration	779	9	Bioconcentration Factor in log units
9	QSAR fish toxicity	908	6	LC50
10	Yacht hydrodynamics	308	6	Residuary resistance per unit weight

We conduct real data analysis on ten data sets from the UCI repository, as described in Table 4.3. To preserve generality, most data sets in the repository with more than five predictor variables and approximately 1500 observations or fewer were considered. The following explains

how we analyzed each data set. Each data set is randomly separated into a training set and a test set to assess prediction error, with the training set approximately four times bigger than the test set. The seven procedures described in Section 4.3 are implemented using the training set. The test set is subsequently used to evaluate prediction error. Consequently, $\overline{\text{MSPE}}$ values are computed to evaluate each procedure's performance, where $\overline{\text{MSPE}}$ is defined as the mean squared error in prediction for an average of over 1000 random training and test set partitions.

According to the data analysis results in Table 4.4, we can conclude that the tuneRanger method and our methods are superior to all other procedures. Additionally, the $\overline{\text{MSPE}}$ and average rank performance of tuneRanger and our methods are comparable. Upon closer inspection, the results vary depending on the data type. Therefore, we believe that both methods are suitable for use in practice. According to average rank, cv.CST.RF beats CST.RF, while the difference across all data sets seems insignificant. If there is a time constraint, CST.RF is a suitable alternative to cv.CST.RF. Lastly, the superior performance of the default approach in data sets 8 and 9 demonstrates that there is no single optimal tuning strategy.

Table 4.4: Summary of the $\overline{\text{MSPE}}$ values derived from the analysis of data sets 1 through 10. Additionally, the rank is computed using the $\overline{\text{MSPE}}$ value for each data set, displayed between parentheses next to the $\overline{\text{MSPE}}$ value. The procedure with the lowest $\overline{\text{MSPE}}$ is the one with the lowest rank. The average rank is displayed between parentheses next to each procedure's name.

Data set	**default (5)**	**CS.RF (5.8)**	**tuneRF (4)**	**tune.rfsrc (4.9)**	**tuneRanger (3)**	**CST.RF (2.8)**	**cv.CST.RF (2.5)**
1	12.96663 (7)	3.87302 (6)	3.57957 (5)	3.41309 (4)	3.25631 (2)	3.24823 (1)	3.25875 (3)
2	7.71837 (5)	7.93895 (7)	7.66451 (3)	7.69696 (4)	7.74673 (6)	7.59487 (1)	7.60856 (2)
3	30.05015 (7)	25.5282 (5)	24.65954 (4)	26.38998 (6)	22.23771 (1)	23.16951 (3)	22.96185 (2)
4	1.24491 (7)	0.25351 (5)	0.27688 (6)	0.25125 (4)	0.23019 (3)	0.22606 (1)	0.22824 (2)
5	3.40837 (7)	2.88359 (3)	3.03665 (5)	2.93579 (4)	3.03731 (6)	2.83769 (2)	2.74709 (1)
6	2.11759 (6)	2.34238 (7)	2.09236 (5)	2.02909 (3)	1.98835 (2)	2.07024 (4)	1.9877 (1)
7	1.21683 (2)	1.26301 (6)	1.21928 (3)	1.27071 (7)	1.21662 (1)	1.25141 (5)	1.2321 (4)
8	0.47946 (1)	0.50923 (6)	0.48185 (2)	0.51784 (7)	0.48337 (3)	0.49994 (5)	0.49326 (4)
9	0.77004 (1)	0.83569 (7)	0.77007 (2)	0.79886 (6)	0.77304 (3)	0.79086 (5)	0.78749 (4)
10	16.39026 (7)	1.22974 (6)	1.20511 (5)	1.19234 (4)	1.18749 (3)	1.14365 (1)	1.15099 (2)

Table 4.5: $\overline{\text{MSPE}}$ values derived for different models and values of σ_ϵ. For each scenario, the smallest $\overline{\text{MSPE}}$ is in bold font.

Model ID	σ_ϵ	default	CS.RF	tune.rfsrc	tuneRF	tuneRanger	CST.RF	cv.CST.RF
1	0.1	0.98102	0.74114	0.57640	0.50929	0.46241	0.45886	**0.43631**
·	1	1.98977	1.91859	1.65325	1.55738	1.55673	1.52995	**1.52354**
·	10	104.43360	112.00949	106.87260	104.31740	**102.55088**	104.96307	103.89444
2	0.1	2.07559	0.95094	0.56988	0.55675	0.51447	0.50965	**0.47081**
·	1	3.08743	2.18807	1.70351	1.67047	1.65557	1.64016	**1.62609**
·	10	105.77790	113.65692	108.49098	106.08697	**103.98705**	106.49796	105.32098
3	0.1	3.11834	3.29546	2.85259	2.71925	2.61802	2.56033	**2.52681**
·	1	4.12401	4.41952	3.89942	3.74910	3.68509	3.60569	**3.58591**
·	10	106.60905	114.53857	109.25835	106.70038	**105.48454**	107.51734	106.80550
4	0.1	1.13538	1.06260	0.64415	0.63631	0.61954	0.58357	**0.55720**
·	1	2.16040	2.27861	1.74242	1.71388	1.72399	1.67577	**1.67032**
·	10	105.58647	113.17093	108.24301	105.49766	**103.84140**	106.11952	105.11590
5	0.1	1.08306	0.75013	0.32088	0.35890	0.36625	0.30934	**0.30315**
·	1	2.11382	1.98602	1.43523	1.45624	1.48277	1.41810	**1.42018**
·	10	105.19142	112.14874	106.55931	105.14231	**104.01971**	104.87568	104.43176

4.5 Discussion

This paper presents the CST algorithm as a viable alternative to the existing RF tuning approaches. However, the approach has significant room for improvement. The parameter domain of tuning parameter values should be reduced for two reasons. First, because grid search strategies are typically slow, restricting the parameter domain will accelerate the algorithm. The second reason is depicted in Figure 4.1. The figure demonstrates that it is advantageous to use unique tuning parameter values for each target point. There is a further noteworthy observation. Only two of the four-parameter vectors minimize the sum of squared prediction errors in at least one predictor domain. In other words, when implementing the CST algorithm, we may not require all tuning parameter vectors. Therefore, a moderately reduced parameter domain may improve the CST algorithm in terms of minimizing prediction error.

It would be intriguing to evaluate suitable prediction intervals for our method. We may employ Zhang et al.'s (2019) prediction interval utilizing the distribution of OOB prediction errors to our method. We also want to investigate whether the case-specific tuning algorithm is useful for classification problems. In addition, it could be interesting to determine whether our strategy is useful for bias correction of other nonparametric methods.

4.6 References

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Friedbergå, R., Tibshirani, J., Athey, S., and Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Probst, P. and Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(1):6673–6690.

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301.

Xu, R., Nettleton, D., and Nordman, D. J. (2016). Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25(1):49–65.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*.

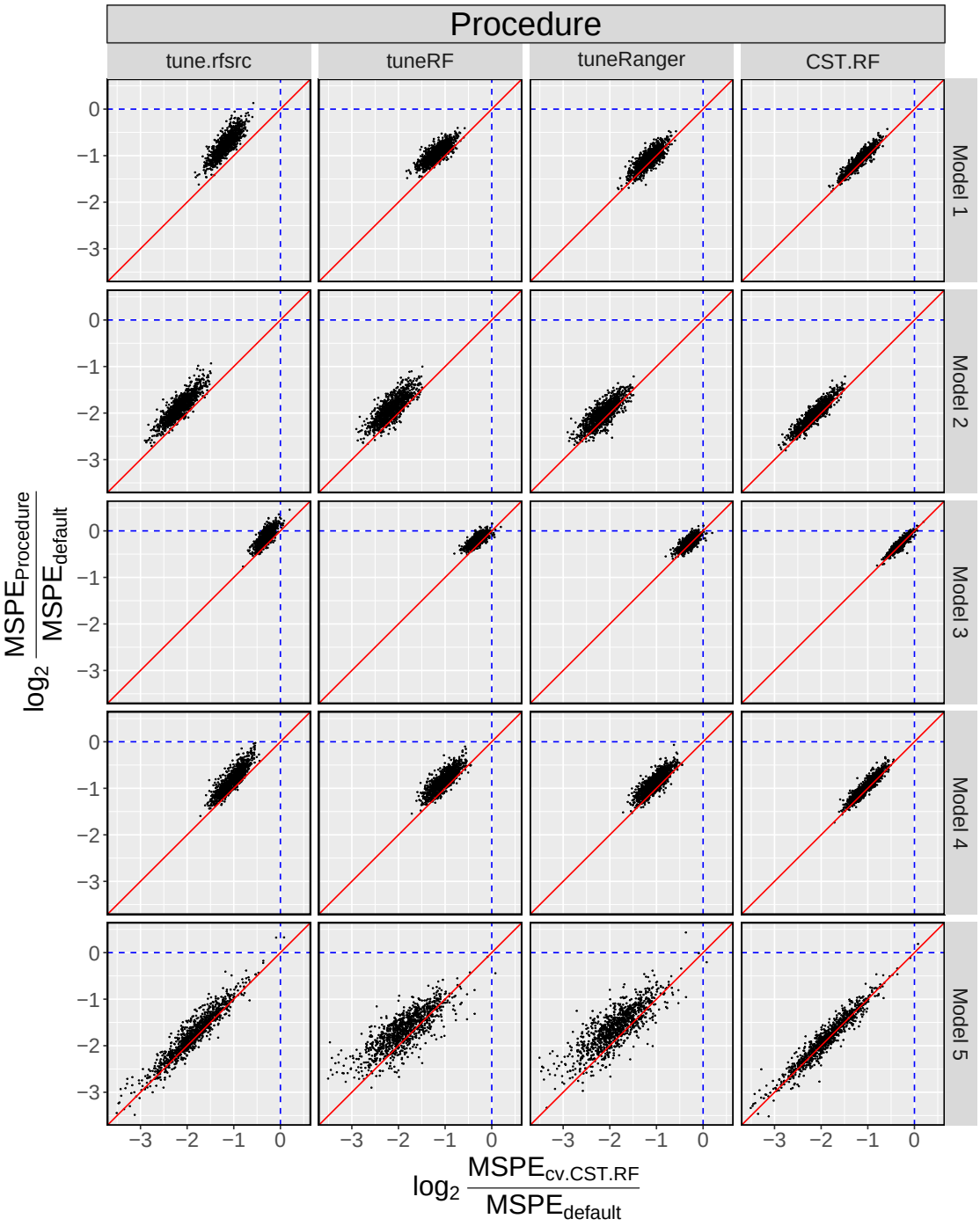


Figure 4.2: Scatterplot showing log2-transformed MSPE ratios for a given procedure and cv.CST.RF procedure, in which the scenario of $\sigma_\epsilon = 0.1$. Each point represents a simulation run, and the MSPE ratio is calculated by dividing the MSPE of a procedure by the MSPE of the default procedure. The red line indicates that the y and x axis values are identical.

CHAPTER 5. GENERAL CONCLUSIONS

5.1 Summary

In the dissertation, three topics are covered. The first two topics are RNA-seq analysis methods incorporating covariates in distinct inference circumstances. The method described in Chapter 2, which employs a gene-specific covariate such as gene length, is a generally applicable approach. We have demonstrated through simulation that the rejection rule that includes the simple information of a low p-value can effectively boost the inference power. The method in Chapter 3 is another RNA-seq analysis method for a more particular circumstance compared to the first method. Simulations indicated that the inference power could be improved by effectively reducing the number of rejection rules when the data include both a pilot study and the main study. Chapter 4 covered the final topic of the RF's case-specific tuning algorithm. Despite the simplicity of the approach, we demonstrated that the prediction error could be lowered practically through simulations and data analysis.

5.2 Future Work

I am planning to conduct more research on topics related to this dissertation. Regarding the first topic, I wish to study rejection rules based on various conditional null probabilities. Incorporating informative conditions of promising hypotheses into the conditional null probability may increase the power of inference. We can also consider a method based on a more generalized mixture model. Next, multiple research topics can be derived from the second topic of this dissertation. By incorporating p-value as a covariate, we developed a novel method, while its applicability is limited. I want to investigate additional covariates having a unique functional relationship with p-value thresholds that may apply to a wider range of circumstances. Then, methods that accommodate the specific relationship can be devised. Alternately, the covariate

required by our method can be obtained from p-value vectors derived from experiments with comparable experimental designs. It may be essential to appropriately handle missing p-values, which is a potentially intriguing statistical question. Lastly, I would like to apply the case-specific tuning philosophy to various machine learning methods.