Bayesian methods under unknown prior distributions with applications to the analysis of gene expression data

Abbas Rahal

Thesis submitted to the University of Ottawa in partial fulfillment of the
requirements for the degree of
Doctorate in Philosophy Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

The local false discovery rate (LFDR) is one of many existing statistical methods that analyze multiple hypothesis testing. As a Bayesian quantity, the LFDR is based on the prior probability of the null hypothesis and a mixture distribution of null and non-null hypothesis. In practice, the LFDR is unknown and needs to be estimated. The empirical Bayes approach can be used to estimate that mixture distribution.

Empirical Bayes does not require complete information about the prior and hyper prior distributions as in hierarchical Bayes. When we do not have enough information at the prior level, and instead of placing a distribution at the hyper prior level in the hierarchical Bayes model, empirical Bayes estimates the prior parameters using the data via, often, the marginal distribution.

In this research, we developed new Bayesian methods under unknown prior distribution. A set of adequate prior distributions maybe defined using Bayesian model checking by setting a threshold on the posterior predictive p-value, prior predictive p-value, calibrated p-value, Bayes factor, or integrated likelihood. We derive a set of adequate posterior distributions from that set. In order to obtain a single posterior distribution instead of a set of adequate posterior distributions, we used a blended distribution, which minimizes the relative entropy of a set of adequate prior (or posterior) distributions to a "benchmark" prior (or posterior) distribution. We present two approaches to generate a blended posterior distribution, namely, updating-before-blending and blending-before-updating. The blended posterior distribution can be

ABSTRACT

used to estimate the LFDR by considering the nonlocal false discovery rate as a benchmark and the different LFDR estimators as an adequate set.

The likelihood ratio can often be misleading in multiple testing, unless it is supplemented by adjusted p-values or posterior probabilities based on sufficiently strong prior distributions. In case of unknown prior distributions, they can be estimated by empirical Bayes methods or blended distributions. We propose a general framework for applying the laws of likelihood to problems involving multiple hypotheses by bringing together multiple statistical models.

We have applied the proposed framework to data sets from genomics, COVID-19 and other data.

Résumé

Le *local false discovery rate* (LFDR) est l'une de nombreuses méthodes statistiques qui analyse plusieurs tests d'hypothèses ensemble. En tant que quantité bayésienne, le LFDR est basé sur la probabilité a priori de l'hypothèse nulle et sur une loi mixte d'hypothèses nulle et non nulle. Dans la pratique, le LFDR est inconnu et doit être estimé. L'approche empirique de Bayes peut être utilisée pour estimer cette loi mixte.

L'empirique de Bayes ne nécessite pas d'informations complètes sur les lois a priori et hyper-priori comme celle de Bayes hiérarchique. Lorsque nous n'avons pas suffisamment d'informations au niveau de a priori, et au lieu de placer une loi au niveau d'hyper-priori dans le modèle de Bayes hiérarchique, l'approche empirique de Bayes estime les paramètres de la loi a priori en utilisant les données via, souvent, la loi marginale.

Dans ce projet, nous avons développé de nouvelles méthodes bayésiennes sous des lois a priori inconnues. Un ensemble de lois a priori adéquates peut être défini en utilisant un modèle bayésien de vérification et en fixant un seuil sur la valeur-p de la loi prédictive a posteriori, la valeur-p de la loi prédictive a priori, la valeur-p calibrée, le facteur de Bayes, ou la vraisemblance intégrée. Nous dérivons de cet ensemble un autre ensemble de lois a posteriori adéquates. Afin d'obtenir une seule loi a posteriori, nous utilisons la loi mélangée, qui minimise l'entropie relative d'un ensemble de lois a priori (ou a posteriori) adéquates à une loi a priori (ou a posteriori) de référence. Nous présentons deux approches nommées *updating-before-blending* et

RÉSUMÉ

blending-before-updating pour générer une loi a posteriori mélangée. La loi a posteriori mélangée peut être utilisée pour estimer le LFDR en considérant le *nonlocal false discovery rate* comme une référence et les différents estimateurs du LFDR comme un ensemble adéquat.

Le rapport de vraisemblance peut être souvent trompeur dans la comparaison de tests multiples à moins qu'il ne soit complété par des valeurs-p ajustées ou des probabilités a posteriori basées sur des lois a priori suffisamment fortes. En cas de lois a priori inconnues, elles peuvent être estimées par des méthodes empiriques de Bayes ou par de lois mélangées. Nous proposons un cadre général pour appliquer le rapport de vraisemblance aux problèmes impliquant des hypothèses multiples en réunissant plusieurs modèles statistiques.

Nous avons appliqué le cadre proposé à des ensembles de données provenant de la génomique, de la COVID-19 et d'autres données.

Dedications

For my two daughters, Lena and Lyne.

Acknowledgement

I would like to express my special appreciation and thanks to my supervisor, Dr. David R. Bickel, to supervise my Ph.D. study and for his guidance, patience and support. I would also like to thank Dr. Ali Karimnezhad, Dr. Iraj Yadegari and Dr. Marta Padilla for their help, comments and friendship.

I warmly thank the funding sources for this research which was partially supported by the Natural Sciences and Engineering Research Council of Canada, by Agriculture and Agri-Food Canada, by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa.

Many thanks to the Department of Mathematics and Statistics of the University of Ottawa and Carleton University for providing a special environment for studying, learning and research, as well as for financial support.

I would like to thank Dr. Lisa Strug of the University of Toronto, Dr. Shirley Mills of Carleton University, Dr. Mayer Alvo and Dr. David Sankoff of the University of Ottawa for agreeing to examine this thesis, for their time and comments.

Last but not the least, I would like to thank my wife Dania and my family for their encouragement, support and patience.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Preface

This thesis is submitted for the degree of Doctor of Philosophy at the University of Ottawa. The research period for this thesis was between January 2015 and December 2020 under the supervision of David R. Bickel, Adjunct Professor in the Department of Mathematics and Statistics of the University of Ottawa and Associate Professor in the Informatics and Analytics Program of the University of North Carolina at Greensboro.

The multiple hypothesis testing problem is one of the important problems in genomics data analysis, and the LFDR can be used to analyze thousands of genetic features including genes, proteins, etc. together. We used empirical Bayes methods to estimate the LFDR. We are interested in analyzing different LFDR estimators to find out their appropriate characteristics in terms of required specifications such as number of features, sample size of treatment and control groups, etc. LFDR estimators may perform differently for the same test statistic, i.e., one estimator may perform well for a small number of features but not necessarily for another estimator. The question that arises in this context is what to do if there is uncertainty about which estimator to use.

This motivated us to develop new Bayesian methods we call blended distributions. In these methods, instead of estimating the unknown prior distribution as in empirical Bayes approach, we determine a set of prior distributions and then select a single distribution from this set. The selection of a single distribution from the

set of prior distributions is based on minimizing the relative entropy of that set to a benchmark prior distribution. The blended distribution can be a solution to the problem mentioned in the question above.

In addition, we introduce a general framework for applying the laws of likelihood to problems involving multiple hypotheses that may require posterior probabilities to avoid misleading results. This requires strong prior distributions that can be estimated in the case where their parameters are unknown.

Chapter 1

Introduction

This thesis is composed of three papers, each of which is the subject of a chapter. Before presenting them, it is important to show a brief summary of preliminary definitions and to review some statistical concepts.

1.1 Blended distribution

In general, Bayesian inference can be used when we have a complete knowledge of the prior distribution, whereas the blended approach can be used when we have a partial or unknown knowledge of the prior distribution. This approach is based on using a set of prior distributions instead of using a single prior. Many methods can be used to determine a set of adequate prior distributions such as prior predictive p-value, posterior predictive p-value [5], Bayes factor, or integrated likelihood [3].

We define two approaches of blended distribution. The first approach we call updating-before-blending ($\widetilde{P^x}$), which is the posterior distribution, $\dot{P}^x$, in a set of adequate Bayesian posteriors, $\dot{\mathcal{P}}^x(\ddot{P}^x)$, that minimizes the relative entropy of $\dot{\mathcal{P}}^x(\ddot{P}^x)$

to a benchmark posterior, $\ddot{P}^x$. Thus

$$\widetilde{P^x} = \underset{\dot{P}^x \in \dot{\mathcal{P}}^x(\ddot{P}^x)}{\arg\inf} \mathrm{RE}(\dot{P}^x \parallel \ddot{P}^x).$$

The second approach we call blending-before-updating $\left(\widetilde{P}\right)^x$. The first step, we find the blended prior $(\widetilde{P})$ which is one of the prior distribution, $\dot{P}$, within a set of adequate prior distributions, $\dot{\mathcal{P}}(\ddot{P})$, that minimizes the relative entropy to the benchmark prior, $\ddot{P}$, thus

$$\widetilde{P} = \underset{\dot{P} \in \dot{\mathcal{P}}(\ddot{P})}{\arg\inf} \mathrm{RE}(\dot{P} \parallel \ddot{P}).$$

The second step, given the data X we derive from that blended prior its posterior

$$\left(\widetilde{P}\right)^x = \widetilde{P}(\bullet \mid X = x).$$

In Chapter 2, we will apply these two approaches on two well-known models, namely, the normal-normal model and the Poisson-gamma model to see if for each model they produce the same result. These approaches can be used in several applications. In this thesis, we will use them in three applications. Our first application illustrates the number of deaths due to coronavirus in Canada. The second application compares a new treatment to a placebo, where it can be given at home after a myocardial infarction. The third application concerns breast cancer. We will propose a new LFDR estimator which is used to select the proteins that have a high probability of being affected by cancer.

1.2 Local false discovery rate

In genomics applications, we are interested in discovering genes that are differentially expressed by testing differential gene expression levels between case and control

groups.

A number of Bayesian approaches have been used in multiple hypothesis testing. Efron [4] has defined a Bayesian framework for multiple hypothesis testing called LFDR which is an alternative definition to that of Benjamini and Hochberg [2].

Consider the following multiple hypothesis testing problem. Suppose we have two tables of data for the treatment and control groups, as in Figure 1.1. For a feature i of a patient j, we have x_{ij} for the treatment group and y_{ij} for the control group. These x_{ij} and y_{ij} can be gene expression level, log-abundance, etc.

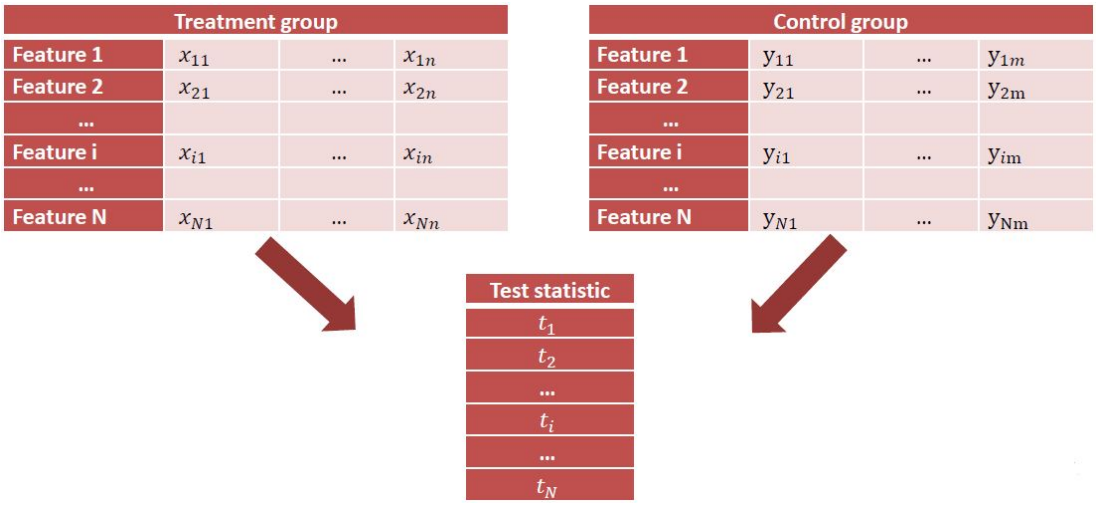

Figure 1.1: Multiple hypothesis testing problem between treatment and control groups.

The null (H_0) and alternative (H_1) hypothesis are defined as follows: H_{0i} : the ith feature is unaffected by a treatment, unassociated with a disease, etc. H_{1i} : the ith feature is affected by a treatment, associated with a disease, etc. H_{0i} (null) means that for the feature i, x_{ij} and y_{ij} have the same distribution. Consider a variable A_i that is equal to 0 if H_{0i} is true, or equal to 1 if H_{1i} is true. The two groups of data in Figure 1.1 can be reduced to some test statistics, say $T_1, T_2..., T_N$ and their realizations are $t_1, t_2..., t_N$. Let $\pi_0 = \mathrm{P}(A_i{=}0)$ be the prior probability of null, $\pi_1 = \mathrm{P}(A_i{=}1)$ be the prior probability of non-null with $\pi_1 = 1 - \pi_0$, $f_0(t)$ be the null

density, and $f_1(t)$ be the non-null density. Let $f(t)$ be the mixture density, such that $f(t) = \pi_0 f_0(t) + \pi_1 f_1(t)$. The LFDR is defined by

$$\psi_i \equiv \mathrm{P}(A_i = 0 | T_i = t_i) = \frac{\pi_0 f_0(t_i)}{f(t_i)}.$$

ψ_i is unknown and needs to be estimated. There are many estimators of the LFDR such as the histogram-based estimator, the maximum likelihood estimator, the binomial-based estimator, etc. Due to the high variance of some LFDR estimators and the lack of availability of software, Chapter 3 defines two new LFDR estimators: the corrected false discovery rate and the re-ranked false discovery rate which are correcting an estimated FDR or the level at which an FDR is controlled.

1.3 Weight of evidence

The p-value is not the only statistical evidence that can be used for reporting outcomes of hypothesis tests. The ASA Statement on Statistical Significance and P-Values [10] included the likelihood ratio as a measure of statistical evidence that might replace the p-value. Some statisticians prefer to supplement or even replace p-values with other approaches such as likelihood ratios or Bayes Factors. Chapter 4 offers a general framework for applying the laws of likelihood to problems involving multiple hypotheses by bringing together multiple statistical models and applying to gene expression data.

1.4 Data

In this thesis, we have used different data for inferences and applications. The main objective of using these data is to apply the theories and approaches we have defined in each paper. In the following, the data we used:

1. Data on the number of deaths due to COVID-19 in Canada. This data was published on the website of Health Canada, Canada.ca/coronavirus at `https://bit.ly/3noHUY5`.

2. Biomedical data of the GREAT trial of early treatment for myocardial infarction. This data contains the number of death after experimental trial for new treatment against a placebo [9, pp. 26, 69].

3. The SELDI-TOF spectra data set that consists of the log-abundance levels of 20 plasma proteins in 35 women with ER/PR positive breast cancer (treatment group) and 64 healthy women (control group), which were measured in Alex Miron's laboratory at the Dana-Farber Cancer Institute [6].

4. The microarray data set that consists of the expression levels of 13,440 genes across 6 biological replicates in tomatoes at three days after the breaker stage of ripening [1].

5. The SAT data contains the means and variances of SAT exam score differences between students participating in a training program and those not participating for each of eight exam sites [8].

1.5 The objectives of the thesis

The main objective of this thesis is to define new Bayesian methods under unknown prior distributions. Each of the following three paragraphs summarizes how that objective is addressed by one of the three papers of the thesis.

We first define two blended distribution approaches which use the Bayesian approach under unknown prior distribution. We compare them for each normal-normal and Poisson-gamma model to check whether they yield the same results, and apply

them to some applications, such as breast cancer, COVID-19 and biomedical treatment.

Second, we define new LFDR estimators under unknown prior distributions of the null and non-null hypothesis. To evaluate the performance of each LFDR estimator, we conduct a simulation study to compare the new LFDR estimators to some existing estimators under different scenarios.

Third, we introduce a general framework for applying the laws of likelihood to problems involving multiple hypotheses that may require posterior probabilities to avoid misleading results. This requires strong prior distributions which can be estimated in the case where their parameters are unknown.

1.6 The objectives of the chapters

This thesis is based on three papers. The paper in Chapter 2 is in preparation; my contributions were the mathematical developments of methods and approaches, the data analysis and interpretation. The paper in Chapter 3 is published; I contributed to the implementation of the simulation studies, the data analysis, the provision of an R package for the methods proposed in the paper, and the discussion of the results. The paper in Chapter 4 is published; my contributions were the data analysis and interpretation, and the discussion of the data analysis results. Overall I had a significant and major contribution, my supervisor was the research leader, he also critically reviewed and approved the versions for publication.

We are working on another paper "Empirical Bayes estimation of gene expression fold change" which we have not included here because there is still a section to complete.

Chapter 2 defines two approaches of blended distributions, updating-before-blending and blending-before-updating. These two approaches will be applied to the normal-normal model and the Poisson-gamma model. To illustrate the use of the

proposed approaches, the chapter presents three applications. The first application illustrates the number of deaths due to coronavirus in Canada. The second application compares a new treatment to a placebo, where it can be given at home after a myocardial infarction. The third application focuses on breast cancer.

Chapter 3 proposes new estimators of LFDR which are correcting false discovery rate, re-ranking false discovery rate, and an estimator of nonlocal false discovery rate. It also presents a stimulation study to compare the performance of the proposed estimators to two existing LFDR estimators, the histogram-based estimator and the maximum likelihood estimator. An application on tomato gene expression data is performed to compare the results of each estimator in terms of the number of genes found that are differentially expressed.

Chapter 4 defines the theory of weight of statistical evidence with respect to a single model. It also defines the fusion model and the weight of evidence with respect to two fused models. For purposes of illustration and application of multiple hypothesis evidence, these theories were applied to exam scores data and to gene expression levels data. At the end, it presents a brief discussion of the practical advantages of the fused model weight of evidence for shrinkage and for evidence supporting a point null hypothesis.

Chapter 2

Commutativity of updating by Bayes's theorem and blending by maximum entropy for the normal-normal and Poisson-gamma models

This chapter defines two approaches of blended distributions, updating-before-blending and blending-before-updating. These two approaches will be applied to the normal-normal model and the Poisson-gamma model. For illustration and application on both approaches, we present three applications. The first application illustrates the number of deaths due to coronavirus in Canada. The second application compares a new treatment to a placebo, where it can be given at home after a myocardial infarction. The third application focuses on breast cancer.

Commutativity of updating by Bayes's theorem and blending by maximum entropy for the normal-normal and Poisson-gamma models

Abbas Rahal* and David R. Bickel†

Keywords: robust Bayesian statistics; imprecise probability; Bayesian model checking; blended inference; posterior predictive p-value.

*Department of Mathematics and Statistics
University of Ottawa
150 Louis-Pasteur Pvt
Ottawa, ON, Canada K1N 6N5
†Informatics and Analytics
University of North Carolina at Greensboro
The Graduate School
241 Mossman Building CAMPUS
Greensboro, NC 27402-6170
dbickel@uncg.edu

Abstract

Given data, a set of adequate prior distributions, and a "benchmark" prior distribution that would have been used were that set unknown, there are two methods of generating a blended posterior distribution, a single posterior distribution for inference and decision making. The methods differ only in the order of applying two transformations: updating according to Bayes's theorem and blending by maximum entropy or, more precisely, minimum relative entropy. The blending transformation results in the adequate distribution of the lowest entropy relative to the benchmark distribution.

The first approach starts by updating all the prior distributions, using Bayes's theorem to generate a set of adequate posterior distributions and a benchmark posterior distribution. That approach then returns the adequate posterior distribution that minimizes the entropy relative to the benchmark posterior distribution.

The second approach instead starts with the adequate prior distribution that minimizes the entropy relative to the benchmark prior distribution. Then that prior distribution is updated to a posterior distribution via Bayes's theorem.

It appears that the order of updating and blending tends to have little effect on the blended posterior distribution in the cases of the normal-normal model and the Poisson-gamma model.

1 Introduction

In data analysis, it is often the case that while the prior distribution is not known with certainty, there is enough partial information about the prior to confine it to some set of distributions that are in some sense adequate. There is a vast literature on such sets of probability distributions, often classified under "robust Bayes" in statistics (e.g., Berger and Berliner, 1986; Sivaganesan, 1999; Bayati Eshkaftaki and Parsian, 2011; Kiapour and Nematollahi, 2011; Jozani et al., 2012) and under "imprecise probability" in many other fields (e.g., Hurwicz, 1951; Moral and De Campos, 1991; Walley, 2000; Weichselberger, 2001;

Tapking, 2004; Gilio, 2005; Harremoës, 2007; de Cooman and Hermans, 2008; Wang and Klir, 2008; Arias-Nicolás et al., 2009; Hable, 2009; Cohen et al., 2010; Shafer, 2011; Benavoli and Ristic, 2011). One way to generate a set of adequate priors is to consider a prior adequate only if it passes a model check (Bickel, 2015b).

When the set of adequate priors is too large for practical use, scientists often turn to a frequentist method of generating confidence intervals. A method of nested confidence intervals can be encoded as a frequentist posterior distribution called a *confidence distribution* (e.g., Schweder and Hjort, 2002; Singh et al., 2005, 2007; Kim and Lindsay, 2011; Bityukov et al., 2011; Tian et al., 2011; Hannig and Xie, 2012; Xie and Singh, 2013; Nadarajah et al., 2015; Schweder and Hjort, 2016; Veronese and Melilli, 2018; Bickel, 2020). In the framework of Bickel (2015a), its use is warranted if it lies in the set of posterior distributions that corresponds to the set of adequate prior distributions.

If not, how can the confidence distribution and the set of adequate posteriors be blended to generate a single posterior distribution for inference in the spirit of Good (1983)? More generally, what if the alternative distribution is a Bayesian prior or posterior distribution rather than a confidence distribution? Whatever such distribution is blended with a set of priors or posteriors is called a *benchmark distribution* (Bickel, 2015a).

Example 1. Due to the COVID-19 coronavirus in Canada, we are interested in knowing $\theta = E(Y_i|\theta)$, the expected value of Y_i, the number of deaths on day i caused by COVID-19 in Canada for the period between March 15, 2020 and August 15, 2020. Based on the values of Y_i published on April 24, 2020 by Public Health Agency of Canada on its website Canada.ca/coronavirus at `https://bit.ly/3noHUY5` (accessed December 22, 2020), the sample mean of deaths number per day until April 23, 2020 was 54 deaths (Figure 1):

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = 54$$

where n is equal to 40, the number of days from March 15 until April 23.

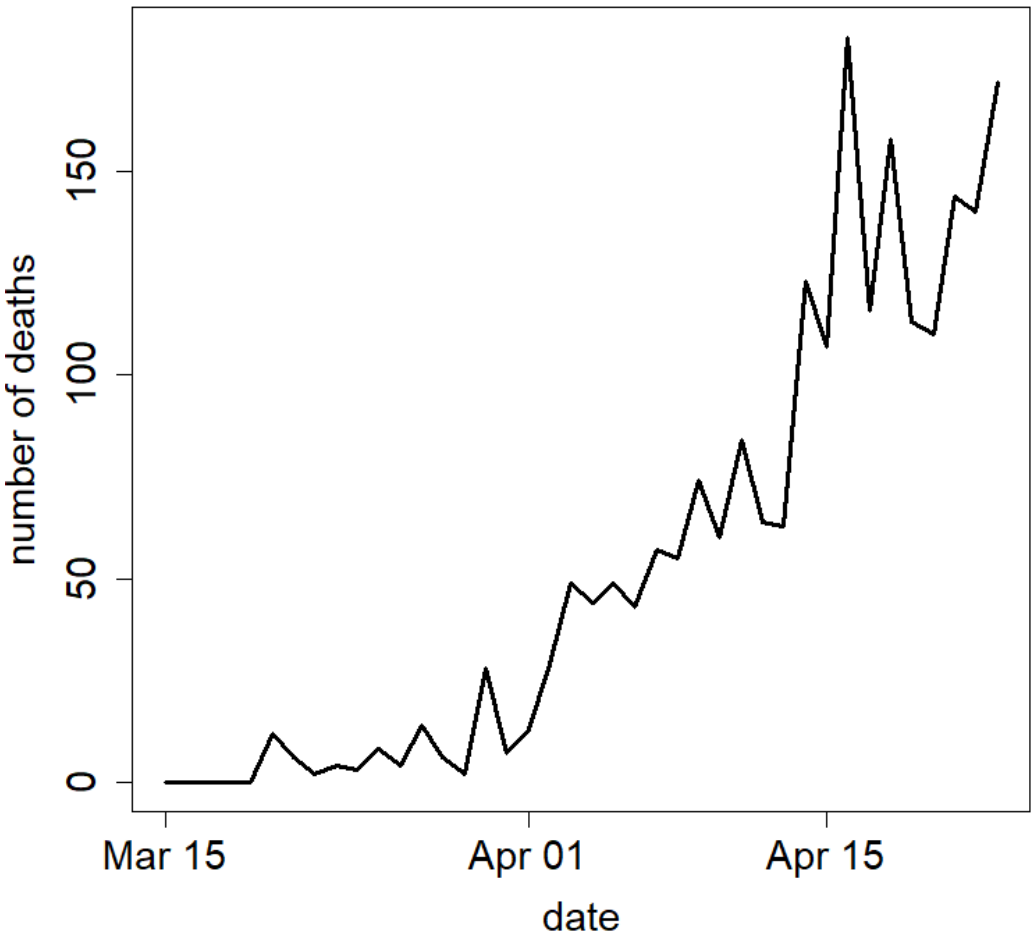

Figure 1: The number of daily deaths caused by COVID-19 in Canada for the period between March 15, 2020 and April 23, 2020.

Although the numbers of deaths are highly correlated and show an upward trend, we make the simplifying assumption for the sake of illustration that the numbers of deaths in each day of that period are mutually independent and drawn from the same Poisson distribution of an unknown parameter value θ. Since we are uncertain about the value of θ, we will use Poisson-gamma Bayesian model to determine the posterior mean expectation value for the period between March 15, 2020 and August 15, 2020. While the hyperparameters α and β of the prior distribution, gamma(α, β), are in general unknown, we arbitrarily used

$\alpha = 35$ for purposes of illustration (For all the values of α between 0 and 120 that we tried, we obtained similar results for this problem. To ensure that the two methods used in this paper had different results, we chose a value for α smaller than 120). We then applied a Bayesian model checking test to determine an interval of adequate values of β. A hyperparameter value or its corresponding prior distribution is considered *adequate* if it is not rejected by such a statistical test. In this example, a hyperparameter value and its prior distribution are considered adequate if a posterior predictive p-value is greater than 0.05. That resulted in $(0\,,\,1.45]$ as the set of adequate values of β and in the corresponding set of adequate prior distributions. Each adequate prior distribution corresponds to an adequate posterior distribution according to Bayes's theorem, resulting in the set of adequate posterior distributions. Section 4.1 explains the method we used to calculate the posterior predictive p-value.

The question that arises here is which adequate posterior distribution should be used for inference and decision making? For that, we use an alternative prior called the "benchmark" that can be determined by using the information of the sample mean number of daily deaths (54 deaths) caused by COVID-19 in Canada until April 23, 2020. Thus, the benchmark prior distribution that we use in this example is gamma$(270, 5)$, calculated as follows: the expectation for a gamma(α, β) distribution is $\frac{\alpha}{\beta}$, we chose α and β in a way that $\frac{\alpha}{\beta} = 54$ in order to ensure that the prior expectation value of θ is equal to the sample mean of the data set:

$$E_{\text{benchmark}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i = 54.$$

(In addition to that constraint, we also constrained α and β in order to obtain different results from different methods used in this paper. That was done for the sake of illustration). This benchmark prior distribution is used to blend with the adequate prior distributions using the two methods of this paper in order to obtain a single posterior distribution for inference and decision making. ▲

The next example similarly involves unknown prior distributions.

Example 2. The GREAT trial of early treatment for myocardial infarction (Spiegelhalter et al., 2004, pp. 26, 69) gave the data of Table 1.

		Treatment		
		New	Control	Total
Event	Death	$a = 13$	$b = 23$	$a + b = 36$
	No death	$c = 150$	$d = 125$	$c + d = 275$
	Total	$a + c = 163$	$b + d = 148$	$N = 311$

Table 1: In this table, we have the numbers of deaths and non-deaths after an experimental trial for a "New" treatment against a placebo ("Control").

The aim of this study is to compare placebo and a new treatment to be given at home as soon as possible after a myocardial infarction. The 30-day mortality rate under each treatment with the benefit of the new treatment is measured by the odds ratio (OR). In this context, the odds ratio for two events with probabilities p_1 and p_2 under two different interventions is defined by

$$\text{OR} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}},$$

where p_1 is the probability of death under the new treatment, and p_2 is the probability of death under the old treatment. An OR less than 1 would then favor the new treatment. The scale of OR is changing between 0 and $+\infty$, the natural logarithm (denoted by log) was used for OR to transform its values between $-\infty$ and $+\infty$. Therefore, the parameter θ is defined by

$$\theta = \log(\text{OR}) = \log(\frac{p_1}{1-p_1}) - \log(\frac{p_2}{1-p_2}).$$

The maximum likelihood estimate of the odds ratio of death is $\frac{\frac{a}{c}}{\frac{b}{d}}$, and θ is estimated by $\log(\frac{\frac{a}{c}}{\frac{b}{d}})$. Spiegelhalter et al. estimated θ by

$$\hat{\theta} = \log(\frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(b+\frac{1}{2})(c+\frac{1}{2})}). \tag{1}$$

And the estimate of the variance of $\hat{\theta}$ is

$$\widehat{V(\hat{\theta})} = \frac{1}{a+\frac{1}{2}} + \frac{1}{b+\frac{1}{2}} + \frac{1}{c+\frac{1}{2}} + \frac{1}{d+\frac{1}{2}}. \tag{2}$$

By using Equations (1) and (2) for the data, Y, in Table 1 we get $\hat{\theta} = \bar{y} = -0.74$ and $\widehat{V(\hat{\theta})} = 0.13$.

Spiegelhalter et al. assumed that when N is not too small it is reasonable to assume a normal distribution for the data, Y, with $Y|\theta \sim N(\theta, \sigma^2 = 4)$. Spiegelhalter et al. (2004, p. 69) used the $N(-0.26, 0.0169)$ as the prior distribution of θ. If the hyperprior mean μ is unknown, the prior distribution would instead be $N(\mu, 0.0169)$. Testing every possible value of μ using a posterior predictive test described in explained in Section 4.2, we found that the only adequate hyperparameters are those between -1.542 and 0.062. This implies that the set of adequate prior distributions is

$$\{N(\mu, 0.0169) : \mu \in [-1.542\,,\, 0.062]\}$$

and that the corresponding set of adequate posterior distributions is

$$\{N(\mu, 0.015) :\ \mu \in [-1.45\,,\, -0.03]\}\,.$$

Which distributions in the set of adequate posterior distributions should be used for inference and decision making? To answer that, we used $N(-0.26\ ,\ 0.5)$ as the benchmark prior distribution in order to blend it with the set of prior distribution using the two methods featured in this paper. We chose the hyperprior mean of -0.26 to follow Spiegelhalter et al. (2004, p. 69). We chose the hyperprior variance of 0.5 in order to ensure that the methods used in this paper would result in different blended posterior distributions. ▲

In this paper, two approaches can be used to find such a blended distribution. While one approach applies an updating transformation before a blending transformation, the other approach applies the blending transformation before the updating transformation. The updating transformation transforms a prior distribution into a posterior distribution according to Bayes's theorem. The blending transformation transforms a set of prior or posterior distributions into a maximum entropy distribution by minimizing the entropy relative to a benchmark distribution such as the gamma distribution at the end of Example 1, the normal

distribution at the end of Example 2, or the Bernoulli distribution in Section 4.3.

The first approach updates first and then blends. Let $\widetilde{P^x}$ be the posterior distribution $(\dot{P}^x)$ in a set Γ^x of adequate posterior distributions that minimizes the relative entropy to a benchmark posterior distribution $(\ddot{P}^x)$ (Bickel, 2015a). Thus

$$\widetilde{P^x} = \underset{\dot{P}^x \in \Gamma^x}{\arg\inf} \mathrm{RE}(\dot{P}^x \parallel \ddot{P}^x).$$

The second approach blends first and then updates. Let $\left(\widetilde{P}\right)^x$ be the resulting posterior distribution. In the first step of this approach, we find the blended prior $(\widetilde{P})$ which is one of the prior distribution $(\dot{P})$ within a set Γ of adequate prior distributions that minimizes the relative entropy to the benchmark prior $(\ddot{P})$, thus

$$\widetilde{P} = \underset{\dot{P} \in \Gamma}{\arg\inf} \mathrm{RE}(\dot{P} \parallel \ddot{P}).$$

In the second step of this approach, given the observation that $X = x$, we apply Bayes's theorem to derive from that blended prior its posterior distribution

$$\left(\widetilde{P}\right)^x = \widetilde{P}(\bullet \mid X = x).$$

These two approaches will be applied to the normal-normal model and the Poisson-gamma model to determine whether for each model they produce the same distribution. If $\widetilde{P^x} = \left(\widetilde{P}\right)^x$, then the blending and updating transformations are said to commute. In general, $\widetilde{P^x} \neq \left(\widetilde{P}\right)^x$, for those transformations do not necessarily commute (Bickel, 2018). In this paper, we want to study the potential impact of that non-commutativity in the special cases of the normal-normal model and the Poisson-gamma model.

This paper is organized as follows. Section 2 defines some concepts and definitions that we will use in this paper such as relative entropy, adequate set, benchmark distribution, updating-before-blending and blending-before-updating. Section 3 defines updating-before-

blending and blending-before-updating approaches for the normal-normal model and the Poisson-gamma model. For illustration and application on both approaches, Section 4 includes three applications. The first application illustrates the number of deaths due to coronavirus in Canada, the second application studies the comparison of placebo and a new treatment to be given at home as soon as possible after a myocardial infarction, and the third application concerns breast cancer.

2 Preliminary concepts

In this section, we define some concepts that we will use in this paper.

2.1 Definition of relative entropy

Let ϑ a random variable with a realization θ in a parameter space Θ. The relative entropy also called Kullback-Leibler divergence of the continuous probability measure P with respect to a measure ν is defined by, $\mathrm{RE}(P||\nu) = \int_\Theta \frac{dP}{d\nu}(\theta) \log \frac{dP}{d\nu}(\theta) d\nu(\theta)$, where $\frac{dP}{d\nu}$ is the Radon-Nikodym derivative (Kullback, 1968, pp. 3-6). We consider two cases:

1. Continuous case: Let ν be the Lebesgue measure, P is absolutely continuous with respect to a probability measure Q ($P \ll Q$) and $Q \ll \nu$, their probability density functions are $p = \frac{dP}{d\nu}$ and $q = \frac{dQ}{d\nu}$ respectively. The relative entropy becomes

$$\mathrm{RE}(p||q) = \int_\Theta p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta = E_P(\log \frac{p(\theta)}{q(\theta)}).$$

2. Discrete case: Assume that $0 \log(0) = 0$. Let p and q be two probability mass functions on a counting measure on a discrete set Θ. The relative entropy becomes

$$\mathrm{RE}(p||q) = \sum_{\theta \in \Theta} p(\theta) \log \frac{p(\theta)}{q(\theta)} = E_P(\log \frac{p(\theta)}{q(\theta)}).$$

The relative entropy has the following properties:

- The relative entropy is not symmetric, therefore it is not a distance.
- $\mathrm{RE}(p||q) \geq 0$ with equality if and only if $p(\theta) = q(\theta)$.

2.2 Definition of set of adequate prior distributions

The definition of set of adequate prior distributions is similar to a set of plausible priors and interval of probability (Weichselberger, 2000; Augustin, 2002, 2004). For a threshold $\alpha \in \mathbb{R}$, the set of adequate prior distributions is defined by

$$\dot{\mathcal{P}} = \{\dot{P} : s(\dot{P}) > \alpha\},$$

where $s(\dot{P})$ is a real-valued function of a prior distribution $\dot{P}$, such as posterior predictive p-value, prior predictive p-value, integrated likelihood, Bayes factor, etc.

The set of adequate hyperparameter values is the set of hyperparameters associated with $\dot{\mathcal{P}}$ as following:

$$H = \{\psi : s_\psi > \alpha\} \text{ such that } \dot{\mathcal{P}} = \{\dot{P}_\psi : \psi \in H\},$$

where $\dot{P}_\psi$ is a prior distribution indexed by a hyperparameter ψ.

2.3 Definition of set of adequate posterior distributions

Let Θ be the set of parameter values on the measurable space $(\Theta, \mathcal{A})$. Let x denote an observed data on the set $\mathcal{X}$ that is endowed with a σ-algebra $\mathfrak{X}$. Denote by, $\dot{\mathcal{P}}$ the set of adequate prior distributions on $(\Theta, \mathcal{A})$.

Let $\dot{P}^x = \dot{P}(\bullet \mid X = x)$ be the Bayesian posterior distribution based on the prior $\dot{P} \in \dot{\mathcal{P}}$ of the parameter and on the observation that $X = x$. The set of adequate posterior distributions on $(\Theta, \mathcal{A})$ is defined by

$$\dot{\mathcal{P}}^x = \left\{ \dot{P}^x \ : \ \dot{P}^x = \dot{P}(\bullet \mid X = x),\ \dot{P} \in \dot{\mathcal{P}} \right\}.$$

Note that $\dot{\mathcal{P}} \subset \mathcal{P}$ and $\dot{\mathcal{P}}^x \subset \mathcal{P}$, where $\mathcal{P}$ is the set of all probability distributions on $(\Theta, \mathcal{A})$.

2.4 Benchmark priors and posteriors

Let $\ddot{P}$ be the benchmark prior distribution, $\ddot{P} \in \mathcal{P}$ but not necessarily in $\dot{\mathcal{P}}$ (defined in section 2.3). The benchmark posterior distribution $\ddot{P}^x = \ddot{P}(\bullet \mid X = x)$ is the posterior distribution resulting from Bayes theorem with $\ddot{P}$ as the prior.

2.5 Updating-before-blending posterior distribution

Let $\dot{\mathcal{P}}^x(\ddot{P}^x)$ be the largest subset of $\dot{\mathcal{P}}^x$ such that the relative entropy of any of its members with respect to $\ddot{P}^x$ is finite. That is,

$$\dot{\mathcal{P}}^x(\ddot{P}^x) = \left\{ \dot{P}^x \in \dot{\mathcal{P}}^x \ : \ \mathrm{RE}(\dot{P}^x \parallel \ddot{P}^x) < \infty \right\}.$$

The updating-before-blending posterior distribution (UB), $\widetilde{P^x}$, is the probability distribution in $\dot{\mathcal{P}}^x$ that minimizes the relative entropy with respect to the benchmark posterior $\ddot{P}^x$ (Bickel, 2015a):

$$\widetilde{P^x} = \underset{\dot{P}^x \in \dot{\mathcal{P}}^x(\ddot{P}^x)}{\arg\inf} \ \mathrm{RE}(\dot{P}^x \parallel \ddot{P}^x). \tag{3}$$

2.6 Blending-before-updating posterior distribution

Let $\dot{\mathcal{P}}(\ddot{P})$ the largest subset of $\dot{\mathcal{P}}$, such that the relative entropy of any of its members with respect to the benchmark $\ddot{P}$ is finite. That is, $\dot{\mathcal{P}}(\ddot{P}) = \left\{ \dot{P} \in \dot{\mathcal{P}} \ : \ \mathrm{RE}(\dot{P} \parallel \ddot{P}) < \infty \right\}$.

The blended prior distribution $\widetilde{P}$ is the probability distribution in $\dot{\mathcal{P}}(\ddot{P})$ that minimizes

the relative entropy with respect to the benchmark prior $\ddot{P}$:

$$\widetilde{P} = \underset{\dot{P} \in \dot{\mathcal{P}}(\ddot{P})}{\arg\inf} \mathrm{RE}(\dot{P} \parallel \ddot{P}).$$

The blending-before-updating posterior distribution (BU), $\left(\widetilde{P}\right)^x$, is the conditional distribution of $\widetilde{P}$ of the parameter given the data $X = x$,

$$\left(\widetilde{P}\right)^x = \widetilde{P}(\bullet \mid X = x).$$

3 Results

In this section, we apply the approaches of sections 2.5 and 2.6 to the normal-normal model (NNM) and the Poisson-gamma model (PGM). We also introduce some theorems to show whether both approaches yield the same or different distributions.

3.1 Normal-Normal model

3.1.1 Background

Let $Y_1, ..., Y_n$ be n independent random variables of $N(\theta, \sigma^2)$, the normal distribution with mean θ and known variance σ^2. The prior distribution of θ is $N(\mu, v)$, where μ is the mean of θ and v its known variance. The hyperparameter μ is unknown and belongs to the following set of adequate hyperparameter values: $[\underline{\mu}, \overline{\mu}]$, where $-\infty < \underline{\mu} < \overline{\mu} < \infty$.

Let $\ddot{P}$ be the benchmark prior distribution of $\theta \sim N(\ddot{\mu}, \ddot{v})$, where $\ddot{\mu}$ and $\ddot{v}$ are the benchmark prior mean and variance, let $\overline{y} = \sum_{i=1}^{n} y_i/n$ be the sample mean. The benchmark posterior distribution $\ddot{P}^y$ of θ is $N(\ddot{\mu}^y, \ddot{v}^y)$, where

$$\ddot{\mu}^y = \frac{\sigma^2}{n\ddot{v} + \sigma^2}\ddot{\mu} + \frac{n\ddot{v}}{n\ddot{v} + \sigma^2}\overline{y} \text{ and } \ddot{v}^y = \frac{\sigma^2 \ddot{v}}{\sigma^2 + n\ddot{v}}.$$

The posterior distribution of θ given the data $y = (y_1, ..., y_n)$ is $N(\mu^y\ ,\ v^y)$, where

$$\mu^y = \frac{\sigma^2}{nv+\sigma^2}\mu + \frac{nv}{nv+\sigma^2}\overline{y} \text{ and } v^y = \frac{\sigma^2 v}{\sigma^2+nv}. \tag{4}$$

Let $[\underline{\mu}^y, \overline{\mu}^y]$ be the set of adequate posterior mean, μ^y. The relation between these two sets $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\mu}^y, \overline{\mu}^y]$ is given by the following equations:

$$\underline{\mu}^y = \frac{\sigma^2}{nv+\sigma^2}\underline{\mu} + \frac{nv\overline{y}}{nv+\sigma^2} \tag{5}$$

$$\overline{\mu}^y = \frac{\sigma^2}{nv+\sigma^2}\overline{\mu} + \frac{nv\overline{y}}{nv+\sigma^2}. \tag{6}$$

Let $\dot{\mathcal{P}} = \left\{N(\mu\ ,\ v)\colon\ \mu \in [\underline{\mu}, \overline{\mu}]\right\}$, and $\dot{\mathcal{P}}(\ddot{P}) = \left\{\dot{P} \in \dot{\mathcal{P}}:\ \mathrm{RE}(\dot{P} \parallel \ddot{P}) < \infty\right\}$ denote the sets of adequate prior distributions. The set of adequate posterior distribution of θ is given by

$$\dot{\mathcal{P}}^y = \left\{N(\mu^y\ ,\ v^y)\colon\ \mu^y \in [\underline{\mu}^y, \overline{\mu}^y]\right\}, \text{ and}$$

$$\dot{\mathcal{P}}^y(\ddot{P}^y) = \left\{\dot{P}^y \in \dot{\mathcal{P}}^y:\ \mathrm{RE}(\dot{P}^y \parallel \ddot{P}^y) < \infty\right\}.$$

3.1.2 Blending-before-updating distribution of NNM

Let $\widetilde{P}$ be the blended prior distribution of NNM with mean $\widetilde{\mu}$, this mean must be in $[\underline{\mu}, \overline{\mu}]$, and variance v.

Lemma 1. *We distinguish three cases of $\widetilde{\mu}$ depending on the value of $\ddot{\mu}$ for BU of NNM.*

Case BU1. If $\ddot{\mu} \in [\underline{\mu}, \overline{\mu}]$ then $\widetilde{\mu} = \ddot{\mu}$ and the BU is $\left(\widetilde{P}\right)^y = N((\widetilde{\mu})^y, v^y)$, where $(\widetilde{\mu})^y = \frac{\sigma^2}{nv+\sigma^2}\ddot{\mu} + \frac{nv}{nv+\sigma^2}\overline{y}$.

Case BU2. If $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$ and $\ddot{\mu} > \overline{\mu}$ then $\widetilde{\mu} = \overline{\mu}$ and the BU is $\left(\widetilde{P}\right)^y = N(\overline{\mu}^y, v^y)$.

Case BU3. If $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$ and $\ddot{\mu} < \underline{\mu}$ then $\widetilde{\mu} = \underline{\mu}$ and the BU is $\left(\widetilde{P}\right)^y = N(\underline{\mu}^y, v^y)$.

Proof. $\widetilde{P}$ is obtained by

$$\begin{aligned} \widetilde{P} &= \underset{\dot{P}\in\dot{\mathcal{P}}(\ddot{P})}{\arg\inf}\, \mathrm{RE}(\dot{P} \parallel \ddot{P}) \\ &= \underset{\dot{P}\in\dot{\mathcal{P}}(\ddot{P})}{\arg\inf} \int_{\Theta} \dot{P}(\theta)\log \frac{\dot{P}(\theta)}{\ddot{P}(\theta)} d\theta \end{aligned} \tag{7}$$

this implies

$$\begin{aligned} \widetilde{\mu} &= \underset{\mu\in[\underline{\mu},\overline{\mu}]}{\arg\inf} \int_{\Theta} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(\theta-\mu)^2}{2v}} (\log(\frac{1}{\sqrt{2\pi v}} e^{-\frac{(\theta-\mu)^2}{2v}}) - \log(\frac{1}{\sqrt{2\pi \ddot{v}}} e^{-\frac{(\theta-\ddot{\mu})^2}{2\ddot{v}}})) d\theta \\ &= \underset{\mu\in[\underline{\mu},\overline{\mu}]}{\arg\inf} \left(\frac{1}{2}\log(\frac{\ddot{v}}{v}) + \frac{1}{2\ddot{v}} \left(v + (\mu - \ddot{\mu})^2 \right) - \frac{1}{2} \right). \end{aligned}$$

Minimizing $\mathrm{RE}(\dot{P} \parallel \ddot{P})$ in Equation (7) is equivalent to minimizing the function

$$g(\mu) = \frac{1}{2}\log(\frac{\ddot{v}}{v}) + \frac{1}{2\ddot{v}} \left(v + (\mu - \ddot{\mu})^2 \right) - \frac{1}{2}$$

over $\mu \in [\underline{\mu}, \overline{\mu}]$. $\widetilde{\mu}$ is the scalar that minimizes $g(\mu)$. The derivative of $g(\mu)$ with respect to μ is $\frac{\partial g(\mu)}{\partial \mu} = \frac{1}{\ddot{v}}(\mu - \ddot{\mu})$, we set $\frac{\partial g(\mu)}{\partial \mu} = 0$, then we get $\widetilde{\mu} = \ddot{\mu}$. Since the second derivative is positive, $\frac{\partial^2 g(\mu)}{\partial \mu^2} = \frac{1}{\ddot{v}} > 0$, therefore $g(\widetilde{\mu})$ is the minimum of $g(\mu)$.

Case BU1. If $\ddot{\mu} \in [\underline{\mu}, \overline{\mu}]$ then $\widetilde{\mu} = \ddot{\mu}$. In this case the blended prior is $\widetilde{P} = N(\ddot{\mu}, v)$, therefore, by computing the posterior distribution of $\widetilde{P}$ we get the BU, $\left(\widetilde{P}\right)^y = N((\widetilde{\mu})^y, v^y)$, where $(\widetilde{\mu})^y = \frac{\sigma^2}{nv+\sigma^2}\ddot{\mu} + \frac{nv}{nv+\sigma^2}\overline{y}$.

Case BU2. If $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$ and $\ddot{\mu} > \overline{\mu}$ then $\widetilde{\mu} = \overline{\mu}$. Since $\widetilde{\mu} \in [\underline{\mu}, \overline{\mu}]$ and $\widetilde{\mu}$ should be equal to $\ddot{\mu}$ but $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$, in this case $\widetilde{\mu}$ equals the value closest to $\ddot{\mu}$ which is $\overline{\mu}$. Then the blended prior is $\widetilde{P} = N(\overline{\mu}, v)$, therefore, the BU is $\left(\widetilde{P}\right)^y = N(\overline{\mu}^y, v^y)$.

Case BU3. If $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$ and $\ddot{\mu} < \underline{\mu}$ then $\widetilde{\mu} = \underline{\mu}$, since $\widetilde{\mu} \in [\underline{\mu}, \overline{\mu}]$ and $\widetilde{\mu}$ should be equal to $\ddot{\mu}$ but $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$, in this case $\widetilde{\mu}$ equals the value closest to $\ddot{\mu}$ which is $\underline{\mu}$. Then the blended prior is $\widetilde{P} = N(\underline{\mu}, v)$, therefore, the BU is $\left(\widetilde{P}\right)^y = N(\underline{\mu}^y, v^y)$. □

3.1.3 Updating-before-blending distribution of NNM

Let $\widetilde{P^y}$ be the UB of NNM with mean $\widetilde{\mu^y}$, such that $\widetilde{\mu^y} \in [\underline{\mu}^y, \overline{\mu}^y]$, and variance v^y.

Lemma 2. *We distinguish three cases of $\widetilde{\mu^y}$ depending on $\ddot{\mu}^y$ for UB of NNM.*

Case UB1. If $\ddot{\mu}^y \in [\underline{\mu}^y, \overline{\mu}^y]$ then $\widetilde{\mu^y} = \ddot{\mu}^y$ and the UB is $\widetilde{P^y} = N(\ddot{\mu}^y, v^y)$.

Case UB2. If $\ddot{\mu}^y \notin [\underline{\mu}^y, \overline{\mu}^y]$ and $\ddot{\mu}^y > \overline{\mu}^y$ then $\widetilde{\mu^y} = \overline{\mu}^y$ and the UB is $\widetilde{P^y} = N(\overline{\mu}^y, v^y)$.

Case UB3. If $\ddot{\mu}^y \notin [\underline{\mu}^y, \overline{\mu}^y]$ and $\ddot{\mu}^y < \underline{\mu}^y$ then $\widetilde{\mu^y} = \underline{\mu}^y$ and the UB is $\widetilde{P^y} = N(\underline{\mu}^y, v^y)$.

Proof. The UB, $\widetilde{P^y}$, is obtained by

$$\begin{aligned} \widetilde{P^y} &= \underset{\dot{P}^y \in \dot{\mathcal{P}}^y(\ddot{P}^y)}{\arg\inf} \ \mathrm{RE}(\dot{P}^y \parallel \ddot{P}^y) \\ &= \underset{\dot{P}^y \in \dot{\mathcal{P}}^y(\ddot{P}^y)}{\arg\inf} \int_{\Theta} \dot{P}^y(\theta) \log \frac{\dot{P}^y(\theta)}{\ddot{P}^y(\theta)} d\theta \end{aligned} \tag{8}$$

this implies

$$\begin{aligned} \widetilde{\mu^y} &= \underset{\mu^y \in [\underline{\mu}^y, \overline{\mu}^y]}{\arg\inf} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi v^y}} e^{-\frac{(\theta-\mu^y)^2}{2v^y}} \left(\log\left(\frac{1}{\sqrt{2\pi v^y}} e^{-\frac{(\theta-\mu^y)^2}{2v^y}}\right) - \log\left(\frac{1}{\sqrt{2\pi \ddot{v}^y}} e^{-\frac{(\theta-\ddot{\mu}^y)^2}{2\ddot{v}^y}}\right)\right) d\theta \\ &= \underset{\mu^y \in [\underline{\mu}^y, \overline{\mu}^y]}{\arg\inf} \left(\frac{1}{2} \log\left(\frac{\ddot{v}^y}{v^y}\right) + \frac{1}{2\ddot{v}^y} \left(v^y + (\mu^y - \ddot{\mu}^y)^2\right) - \frac{1}{2} \right). \end{aligned}$$

To minimize the $\mathrm{RE}(\dot{P}^y \parallel \ddot{P}^y)$ in Equation (8) we have to minimize the function

$$h(\mu^y) = \frac{1}{2} \log\left(\frac{\ddot{v}^y}{v^y}\right) + \frac{1}{2\ddot{v}^y} \left(v^y + (\mu^y - \ddot{\mu}^y)^2\right) - \frac{1}{2}$$

over $\mu^y \in [\underline{\mu}^y, \overline{\mu}^y]$. The UB mean, $\widetilde{\mu}^y$, is the scalar that minimizes the function $h(\mu^y)$. To find $\widetilde{\mu}^y$, we compute the first derivative of $h(\mu^y)$ with respect to μ^y, $\frac{\partial h(\mu^y)}{\partial \mu^y} = \frac{1}{\ddot{v}^y}(\mu^y - \ddot{\mu}^y)$, we set $\frac{\partial h(\mu^y)}{\partial \mu^y} = 0$, this implies $\widetilde{\mu^y} = \ddot{\mu}^y$. The second derivative of $h(\mu^y)$ with respect to μ^y is positive, $\frac{\partial^2 h(\mu^y)}{\partial (\mu^y)^2} = \frac{1}{\ddot{v}^y} > 0$, this is confirm that $h(\ddot{\mu}^y)$ is the minimum of $h(\mu^y)$. Moreover, we have to study whether $\ddot{\mu}^y$ in $[\underline{\mu}^y, \overline{\mu}^y]$ or not, since $\widetilde{\mu^y}$ must be in this set of adequate posterior mean.

Case UB1. If $\ddot{\mu}^y \in [\underline{\mu}^y, \overline{\mu}^y]$ then $\widetilde{\mu^y} = \ddot{\mu}^y$, therefore the UB is $\widetilde{P^y} = N(\ddot{\mu}^y, v^y)$.

Case UB2. If $\ddot{\mu}^y \notin [\underline{\mu}^y, \overline{\mu}^y]$ and $\ddot{\mu}^y > \overline{\mu}^y$ then $\widetilde{\mu^y} = \overline{\mu}^y$, therefore the UB is $\widetilde{P^y} = N(\overline{\mu}^y, v^y)$.

Case UB3. If $\ddot{\mu}^y \notin [\underline{\mu}^y, \overline{\mu}^y]$ and $\ddot{\mu}^y < \underline{\mu}^y$ then $\widetilde{\mu^y} = \underline{\mu}^y$, therefore the UB is $\widetilde{P^y} = N(\underline{\mu}^y, v^y)$. □

Theorem 1. *If $v \neq \ddot{v}$, BU and UB distributions of the normal-normal model are not necessarily the same distribution.*

Proof. To prove that this is true, we need to find under the same assumptions at least two different BU and UB. For this end, we use the following counterexample:

Let $n = 8$, $\sigma^2 = 2$, $\bar{y} = -2.6$, $[\underline{\mu}, \overline{\mu}] = [\frac{1}{6}, \frac{1}{4}]$, $v = \frac{1}{4}$, $\ddot{\mu} = \frac{1}{3}$, $\ddot{v} = \frac{7}{4}$. We can compute $\underline{\mu}^y = \frac{\sigma^2}{nv+\sigma^2}\underline{\mu} + \frac{nv\bar{y}}{nv+\sigma^2} = -1.216667$, $\overline{\mu}^y = \frac{\sigma^2}{nv+\sigma^2}\overline{\mu} + \frac{nv\bar{y}}{nv+\sigma^2} = -1.176667$ and then $[\underline{\mu}^y, \overline{\mu}^y] = [-1.216667, -1.176667]$. We also compute $\ddot{\mu}^y = \frac{\sigma^2}{n\ddot{v}+\sigma^2}\ddot{\mu} + \frac{n\ddot{v}}{n\ddot{v}+\sigma^2}\overline{y} = -2.233333$, $\ddot{v}^y = \frac{\sigma^2\ddot{v}}{\sigma^2+n\ddot{v}} = 0.21875$ and $v^y = \frac{\sigma^2 v}{\sigma^2+nv} = 0.125$.

Since $\ddot{\mu} > \overline{\mu}$, we use case BU2 of lemma 1 we get the blended prior $\widetilde{P} = N(\overline{\mu}, v)$ and then the BU is $\left(\widetilde{P}\right)^y = N(\overline{\mu}^y, v^y)$. In other hand, we have $\ddot{\mu}^y < \underline{\mu}^y$, we use case UB3 of lemma 2 and then the UB is $\widetilde{P^y} = N(\underline{\mu}^y, v^y)$. Hence, BU and UB are different. □

Theorem 2. *If $v = \ddot{v}$, BU and UB of the normal-normal model are the same distribution.*

Proof. Let $a = \frac{\sigma^2}{nv+\sigma^2}$, $b = \frac{nv\bar{y}}{nv+\sigma^2}$, $c = \frac{\sigma^2}{n\ddot{v}+\sigma^2}$ and $d = \frac{n\ddot{v}\bar{y}}{n\ddot{v}+\sigma^2}$. We have by hypothesis $v = \ddot{v}$ this implies $a = c$ and $b = d$. As the variance and the sample size are positive then $a > 0$. We check the three cases according to $\ddot{\mu}$.

If $\ddot{\mu} \in [\underline{\mu}, \overline{\mu}]$ which is equivalent to $\underline{\mu} \leq \ddot{\mu} \leq \overline{\mu}$, since $a > 0$ this implies $a\underline{\mu} + b \leq a\ddot{\mu} + b \leq a\overline{\mu} + b$ thus $\underline{\mu}^y \leq \ddot{\mu}^y \leq \overline{\mu}^y$. By using first case of lemma 1 and lemma 2 the BU and UB are the same distribution, $\left(\widetilde{P}\right)^y = \widetilde{P^y}$, and has $N(\ddot{\mu}^y, v^y)$.

If $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$ and $\ddot{\mu} > \overline{\mu}$ that means $\underline{\mu} < \overline{\mu} < \ddot{\mu}$, since $a > 0$ this implies $a\underline{\mu} + b < a\overline{\mu} + b < a\ddot{\mu} + b$ thus $\underline{\mu}^y < \overline{\mu}^y < \ddot{\mu}^y$. By using second case of lemma 1 and lemma 2 the BU and UB are the same distribution, $\left(\widetilde{P}\right)^y = \widetilde{P^y}$, and has $N(\overline{\mu}^y, v^y)$.

If $\ddot{\mu} \notin [\underline{\mu}, \overline{\mu}]$ and $\ddot{\mu} < \underline{\mu}$ that means $\ddot{\mu} < \underline{\mu} < \overline{\mu}$, since $a > 0$ this implies $a\ddot{\mu} + b < a\underline{\mu} + b < a\overline{\mu} + b$ thus $\ddot{\mu}^y < \underline{\mu}^y < \overline{\mu}^y$. By using third case of lemma 1 and lemma 2 the BU and UB are the same distribution, $\left(\widetilde{P}\right)^y = \widetilde{P^y}$, and has $N(\underline{\mu}^y, v^y)$. □

3.2 Poisson-gamma model

3.2.1 Background

Let $Y_1, ..., Y_n$ are n independent Poisson random variables with mean θ such that $\theta > 0$. The prior of θ has gamma distribution gamma(α, β), where α is the shape and β is the rate. The mean of θ is equal to $\mu = \frac{\alpha}{\beta}$ and the variance of θ is equal to $v = \frac{\alpha}{\beta^2}$, such that, α is positive and known. Let $[\underline{\beta}, \overline{\beta}]$ be the set of adequate hyperparameter values of β, where $0 < \underline{\beta} < \overline{\beta} < \infty$. Let $\ddot{P} = \text{gamma}(\ddot{\alpha}, \ddot{\beta})$ be the benchmark prior distribution of θ, both $\ddot{\alpha}$ and $\ddot{\beta}$ are positive and known. The posterior distribution of θ given the data $y = (y_1, ..., y_n)$ is gamma$(\alpha^y\ ,\ \beta^y)$, where $\alpha^y = \alpha + \sum_{i=1}^{n} y_i$ and $\beta^y = n + \beta$ (Lynch, 2007, section 3.3.2). While the benchmark posterior distribution of θ is equal to $\ddot{P}^y = \text{gamma}(\ddot{\alpha}^y\ ,\ \ddot{\beta}^y)$, where $\ddot{\alpha}^y = \ddot{\alpha} + \sum_{i=1}^{n} y_i$ and $\ddot{\beta}^y = n + \ddot{\beta}$. Let $[\underline{\beta}^y, \overline{\beta}^y]$ be the set of adequate values of β^y. The relation between the sets $[\underline{\beta}, \overline{\beta}]$ and $[\underline{\beta}^y, \overline{\beta}^y]$ is given by the following equations: $\underline{\beta}^y = \underline{\beta} + n$ and $\overline{\beta}^y = \overline{\beta} + n$.

Let $\dot{\mathcal{P}} = \left\{\text{gamma}(\alpha\ ,\ \beta) :\ \beta \in [\underline{\beta}, \overline{\beta}]\right\}$, and $\dot{\mathcal{P}}(\ddot{P}) = \left\{\dot{P} \in \dot{\mathcal{P}} :\ \text{RE}(\dot{P} \parallel \ddot{P}) < \infty\right\}$ denote the sets of adequate prior distributions. The sets of adequate posterior distributions of θ are given by

$$\dot{\mathcal{P}}^y = \left\{\text{gamma}(\alpha^y\ ,\ \beta^y) :\ \beta^y \in [\underline{\beta}^y, \overline{\beta}^y]\right\}, \text{ and}$$

$$\dot{\mathcal{P}}^y(\ddot{P}^y) = \left\{\dot{P}^y \in \dot{\mathcal{P}}^y\text{:}\ \text{RE}(\dot{P}^y \parallel \ddot{P}^y) < \infty\right\}.$$

3.2.2 Blending-before-updating distribution of PGM

Let $\widetilde{P}$ be the blended prior distribution of PGM, $\widetilde{P} = \text{gamma}(\alpha, \widetilde{\beta})$.

Lemma 3. *If $\alpha = \ddot{\alpha}$, we distinguish three cases of $\widetilde{\beta}$ depending on the value of $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$ for BU of PGM.*

Case BU1. If $\ddot{\beta} \in [\underline{\beta}, \overline{\beta}]$ then $\widetilde{\beta} = \ddot{\beta}$ and the BU is $\left(\widetilde{P}\right)^y = gamma(\alpha^y, \ddot{\beta}^y)$.

Case BU2. If $\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\ddot{\beta} > \overline{\beta}$ then $\widetilde{\beta} = \overline{\beta}$ and the BU is $\left(\widetilde{P}\right)^y = gamma(\alpha^y, \overline{\beta}^y)$.

Case BU3. If $\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\ddot{\beta} < \underline{\beta}$ then $\widetilde{\beta} = \underline{\beta}$ and the BU is $\left(\widetilde{P}\right)^y = gamma(\alpha^y, \underline{\beta}^y)$.

Proof. $\widetilde{P}$ is obtained by

$$\begin{aligned} \widetilde{P} &= \underset{\dot{P} \in \dot{\mathcal{P}}(\ddot{P})}{\arg\inf} \mathrm{RE}(\dot{P} \parallel \ddot{P}) \\ &= \underset{\dot{P} \in \dot{\mathcal{P}}(\ddot{P})}{\arg\inf} \int_{\Theta} \dot{P}(\theta) \log \frac{\dot{P}(\theta)}{\ddot{P}(\theta)} d\theta \end{aligned} \tag{9}$$

this implies

$$\begin{aligned} \widetilde{\beta} &= \underset{\beta \in [\underline{\beta}, \overline{\beta}]}{\arg\inf} \int_{\Theta} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} (\log(\frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}) - \log(\frac{\ddot{\beta}^{\ddot{\alpha}}}{\Gamma(\ddot{\alpha})} \theta^{\ddot{\alpha}-1} e^{-\ddot{\beta}\theta})) d\theta \\ &= \underset{\beta \in [\underline{\beta}, \overline{\beta}]}{\arg\inf} \left(\log(\frac{\beta^{\alpha}}{\ddot{\beta}^{\ddot{\alpha}}}) + \log(\frac{\Gamma(\ddot{\alpha})}{\Gamma(\alpha)}) + \ddot{\beta}\frac{\alpha}{\beta} - \alpha + (\alpha - \ddot{\alpha})(\psi(\alpha) - \log(\beta)) \right), \end{aligned}$$

where $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digamma function.

Minimizing $\mathrm{RE}(\dot{P} \parallel \ddot{P})$ in Equation (9) is equivalent to minimizing the function

$$g(\beta) = \log(\frac{\beta^{\alpha}}{\ddot{\beta}^{\ddot{\alpha}}}) + \log(\frac{\Gamma(\ddot{\alpha})}{\Gamma(\alpha)}) + \ddot{\beta}\frac{\alpha}{\beta} - \alpha + (\alpha - \ddot{\alpha})(\psi(\alpha) - \log(\beta))$$

over $\beta \in [\underline{\beta}, \overline{\beta}]$. Let $\widetilde{\beta}$ be the scalar that minimizes $g(\beta)$. The derivative of $g(\beta)$ with respect to β is $\frac{\partial g(\beta)}{\partial \beta} = \frac{\alpha}{\beta} - \frac{\ddot{\beta}\alpha}{\beta^2} - \frac{(\alpha - \ddot{\alpha})}{\beta}$, we set $\frac{\partial g(\beta)}{\partial \beta} = 0$, then $\widetilde{\beta} = \frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$. The second derivative with respect to β is $\frac{\partial^2 g(\beta)}{\partial \beta^2} = \frac{2\alpha\ddot{\beta} - \ddot{\alpha}\beta}{\beta^3}$. $g(\widetilde{\beta})$ is minimum of $g(\beta)$ if $\frac{\partial^2 g(\beta)}{\partial \beta^2} > 0$, this implies $2\alpha\ddot{\beta} - \ddot{\alpha}\beta > 0$ and then $\widetilde{\beta} < 2\frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$ which is always true since $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} > 0$ and $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} < 2\frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$. Moreover, we have to study whether $\ddot{\beta}$ in $[\underline{\beta}, \overline{\beta}]$ or not, since $\widetilde{\beta}$ must be in that set.

Case BU1. If $\ddot{\beta} \in [\underline{\beta}, \overline{\beta}]$ then $\widetilde{\beta} = \ddot{\beta}$. In this case the blended prior is $\widetilde{P} = \mathrm{gamma}(\alpha, \ddot{\beta})$,

therefore, by computing the posterior distribution of $\widetilde{P}$ we get the BU $\left(\widetilde{P}\right)^y = \text{gamma}(\alpha^y, \ddot{\beta}^y)$.

Case BU2. If $\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\ddot{\beta} > \overline{\beta}$ then $\widetilde{\beta} = \overline{\beta}$. The blended prior is $\widetilde{P} = \text{gamma}(\alpha, \overline{\beta})$, therefore, the BU is $\left(\widetilde{P}\right)^y = \text{gamma}(\alpha^y, \overline{\beta}^y)$.

Case BU3. If $\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\ddot{\beta} < \underline{\beta}$ then $\widetilde{\beta} = \underline{\beta}$. The blended prior is gamma$(\alpha, \underline{\beta})$, therefore, the BU is $\left(\widetilde{P}\right)^y = \text{gamma}(\alpha^y, \underline{\beta}^y)$. □

Lemma 4. *If $\alpha \neq \ddot{\alpha}$, it may $\ddot{\beta} \in [\underline{\beta}, \overline{\beta}]$ but $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and vice versa. We know that $\widetilde{\beta}$ belongs to $[\underline{\beta}, \overline{\beta}]$, its value will be determined according to the position of $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$. Under this assumption, we also have three cases for BU of PGM:*

Case BU1. If $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} \in [\underline{\beta}, \overline{\beta}]$ then the BU is $\left(\widetilde{P}\right)^y = gamma(\alpha^y, \frac{\alpha}{\ddot{\alpha}}\ddot{\beta} + n)$.

Case BU2. If $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} > \overline{\beta}$ then the BU is $\left(\widetilde{P}\right)^y = gamma(\alpha^y, \overline{\beta}^y)$.

Case BU3. If $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} < \underline{\beta}$ then the BU is $\left(\widetilde{P}\right)^y = gamma(\alpha^y, \underline{\beta}^y)$.

Proof. From the proof of lemma 3, we found that $g(\beta)$ is minimized by $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$. We have $\alpha \neq \ddot{\alpha}$, this implies $\ddot{\beta}$ and $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$ are different.

CaseBU1. If $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} \in [\underline{\beta}, \overline{\beta}]$ then $\widetilde{\beta} = \frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$. In this case the blended prior is $\widetilde{P} = \text{gamma}(\alpha, \frac{\alpha}{\ddot{\alpha}}\ddot{\beta})$, therefore, by computing the posterior distribution of $\widetilde{P}$ we get the BU $\left(\widetilde{P}\right)^y = \text{gamma}(\alpha^y, \frac{\alpha}{\ddot{\alpha}}\ddot{\beta} + n)$.

The proof of the other two cases is similar to that of lemma 3 by substituting $\ddot{\beta}$ by $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta}$. □

3.2.3 Updating-before-blending distribution of PGM

Let $\widetilde{P^y}$ be the UB of PGM, $\widetilde{P^y} = \text{gamma}(\alpha^y, \widetilde{\beta^y})$.

Lemma 5. *If $\alpha^y = \ddot{\alpha}^y$, we distinguish three cases of $\widetilde{\beta^y}$ depending on the value of $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y$ for UB of PGM.*

Case UB1. If $\ddot{\beta}^y \in [\underline{\beta}^y, \overline{\beta}^y]$ then $\widetilde{\beta^y} = \ddot{\beta}^y$ and the UB is $\widetilde{P^y} = gamma(\alpha^y, \ddot{\beta}^y)$.

Case UB2. If $\ddot{\beta}^y \notin [\underline{\beta}^y, \overline{\beta}^y]$ and $\ddot{\beta}^y > \overline{\beta}^y$ then $\widetilde{\beta^y} = \overline{\beta}^y$ and the UB is $\widetilde{P^y} = gamma(\alpha^y, \overline{\beta}^y)$.

Case UB3. If $\ddot{\beta}^y \notin [\underline{\beta}^y, \overline{\beta}^y]$ *and* $\ddot{\beta}^y < \underline{\beta}^y$ *then* $\widetilde{\beta^y} = \underline{\beta}^y$ *and the UB is* $\widetilde{P^y} = gamma(\alpha^y, \underline{\beta}^y)$.

Proof. $\widetilde{P^y}$ is obtained by

$$\begin{aligned} \widetilde{P^y} &= \underset{\dot{P}^y \in \dot{\mathcal{P}}^y(\ddot{P}^y)}{\text{Arg inf}} \ \text{RE}(\dot{P}^y \parallel \ddot{P}^y) \\ &= \underset{\dot{P}^y \in \dot{\mathcal{P}}^y(\ddot{P}^y)}{\text{Arg inf}} \int_\Theta \dot{P}^y(\theta) \log \frac{\dot{P}^y(\theta)}{\ddot{P}^y(\theta)} d\theta \end{aligned} \tag{10}$$

this implies

$$\begin{aligned} \widetilde{\beta^y} &= \underset{\beta^y \in [\underline{\beta}^y, \overline{\beta}^y]}{\arg\inf} \int_\Theta \frac{(\beta^y)^{\alpha^y}}{\Gamma(\alpha^y)} \theta^{\alpha^y - 1} e^{-\beta^y \theta} (\log(\frac{(\beta^y)^{\alpha^y}}{\Gamma(\alpha^y)} \theta^{\alpha^y - 1} e^{-\beta^y \theta}) - \log(\frac{(\ddot{\beta}^y)^{\ddot{\alpha}^y}}{\Gamma(\ddot{\alpha}^y)} \theta^{\ddot{\alpha}^y - 1} e^{-\ddot{\beta}^y \theta})) d\theta \\ &= \underset{\beta^y \in [\underline{\beta}^y, \overline{\beta}^y]}{\arg\inf} \left(\alpha^y \log(\beta^y) - \ddot{\alpha}^y \log(\ddot{\beta}^y) + \log(\frac{\Gamma(\ddot{\alpha}^y)}{\Gamma(\alpha^y)}) + \ddot{\beta}^y \frac{\alpha^y}{\beta^y} - \alpha^y + (\alpha^y - \ddot{\alpha}^y)(\psi(\alpha^y) - \log(\beta^y)) \right), \end{aligned}$$

where $\psi(\alpha^y) = \frac{\Gamma'(\alpha^y)}{\Gamma(\alpha^y)}$ is the digamma function.

The $\text{RE}(\dot{P}^y \parallel \ddot{P}^y)$ in Equation (10) can be minimized by minimizing the function

$$h(\beta^y) = \alpha^y \log(\beta^y) - \ddot{\alpha}^y \log(\ddot{\beta}^y) + \log(\frac{\Gamma(\ddot{\alpha}^y)}{\Gamma(\alpha^y)}) + \ddot{\beta}^y \frac{\alpha^y}{\beta^y} - \alpha^y + (\alpha^y - \ddot{\alpha}^y)(\psi(\alpha^y) - \log(\beta^y))$$

over $\beta^y \in [\underline{\beta}^y, \overline{\beta}^y]$. Let $\widetilde{\beta}^y$ be the scalar that minimizes the function $h(\beta^y)$. The derivative of $h(\beta^y)$ with respect to β^y is $\frac{\partial h(\beta^y)}{\partial \beta^y} = \frac{\alpha^y}{\beta^y} - \frac{\ddot{\beta}^y \alpha^y}{(\beta^y)^2} - \frac{(\alpha^y - \ddot{\alpha}^y)}{\beta^y}$, we set $\frac{\partial h(\beta^y)}{\partial \beta^y} = 0$, then $\widetilde{\beta^y} = \frac{\alpha^y}{\ddot{\alpha}^y} \ddot{\beta}^y$. To check if $h(\widetilde{\beta^y})$ is the minimum of $h(\beta^y)$ we have to calculate the second derivative. $\frac{\partial^2 h(\beta^y)}{\partial (\beta^y)^2} = \frac{2\alpha^y \ddot{\beta}^y - \ddot{\alpha}^y \beta^y}{(\beta^y)^3}$, the $\frac{\partial^2 h(\beta^y)}{\partial (\beta^y)^2}$ is positive when $2\alpha^y \ddot{\beta}^y - \ddot{\alpha}^y \beta^y > 0$ and then $\widetilde{\beta^y} < 2\frac{\alpha^y}{\ddot{\alpha}^y} \ddot{\beta}^y$ which is always true since $\frac{\alpha^y}{\ddot{\alpha}^y} \ddot{\beta}^y > 0$ and $\frac{\alpha^y}{\ddot{\alpha}^y} \ddot{\beta}^y < 2\frac{\alpha^y}{\ddot{\alpha}^y} \ddot{\beta}^y$. By assumption, we have $\alpha^y = \ddot{\alpha}^y$ this implies $\widetilde{\beta^y} = \frac{\alpha^y}{\ddot{\alpha}^y} \ddot{\beta}^y = \ddot{\beta}^y$. Note that $\widetilde{\beta^y}$ must be in $[\underline{\beta}^y, \overline{\beta}^y]$ but not necessarily the case for $\ddot{\beta}^y$.

Case UB1. If $\ddot{\beta}^y \in [\underline{\beta}^y, \overline{\beta}^y]$ then $\widetilde{\beta^y} = \ddot{\beta}^y$. Therefore, the UB is $\widetilde{P^y} = \text{gamma}(\alpha^y, \ddot{\beta}^y)$.

Case UB2. If $\ddot{\beta}^y \notin [\underline{\beta}^y, \overline{\beta}^y]$ and $\ddot{\beta}^y > \overline{\beta}^y$ then $\widetilde{\beta^y} = \overline{\beta}^y$. Therefore, the UB is $\widetilde{P^y} = \text{gamma}(\alpha^y, \overline{\beta}^y)$.

Case UB3. If $\ddot{\beta}^y \notin [\underline{\beta}^y, \overline{\beta}^y]$ and $\ddot{\beta}^y < \underline{\beta}^y$ then $\widetilde{\beta^y} = \underline{\beta}^y$. Therefore, the UB is $\widetilde{P^y} = \text{gamma}(\alpha^y, \underline{\beta}^y)$. □

Lemma 6. *If $\alpha^y \neq \ddot{\alpha}^y$, it may $\ddot{\beta}^y \in [\underline{\beta}^y, \overline{\beta}^y]$ but $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y \notin [\underline{\beta}^y, \overline{\beta}^y]$ and vice versa. We know that $\widetilde{\beta^y}$ belongs to $[\underline{\beta}^y, \overline{\beta}^y]$, its value will be determined according to the position of $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y$. We can distinguish three cases for UB of PGM:*

Case UB1. If $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y \in [\underline{\beta}^y, \overline{\beta}^y]$ then the UB is $\widetilde{P^y} = gamma(\alpha^y, \frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y)$.

Case UB2. If $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y \notin [\underline{\beta}^y, \overline{\beta}^y]$ and $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y > \overline{\beta}^y$ then the UB is $\widetilde{P^y} = gamma(\alpha^y, \overline{\beta}^y)$.

Case UB3. If $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y \notin [\underline{\beta}^y, \overline{\beta}^y]$ and $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y < \underline{\beta}^y$ then the UB is $\widetilde{P^y} = gamma(\alpha^y, \underline{\beta}^y)$.

Proof. From the proof of lemma 5, we found that $h(\beta^y)$ is minimized by $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y$. By assumption, we have $\alpha^y \neq \ddot{\alpha}^y$, this implies $\ddot{\beta}^y \neq \frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y$.

For case UB1. If $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y \in [\underline{\beta}^y, \overline{\beta}^y]$ then $\widetilde{\beta^y} = \frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y$. Therefore, the UB is $\widetilde{P^y} = \text{gamma}(\alpha^y, \frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y)$.

The proof of the other two cases is similar to that of lemma 5 by substituting $\ddot{\beta}^y$ by $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y$. □

Theorem 3. *If $\alpha \neq \ddot{\alpha}$, BU and UB distributions of the Poisson-gamma model are not necessarily the same distribution.*

Proof. To prove that, we need to find under the same assumptions two BU and UB are different. We use the following counterexample:

Let $n = 5$, $\sum_{i=1}^{n} y_i = 30$, $\alpha = 1$, $[\underline{\beta}, \overline{\beta}] = [1, 3]$, $\ddot{\alpha} = 4$ and $\ddot{\beta} = \frac{5}{2}$. We compute $\alpha^y = \alpha + \sum_{i=1}^{n} y_i = 31$, $\underline{\beta}^y = \underline{\beta} + n = 6$, $\overline{\beta}^y = \overline{\beta} + n = 8$ and then $[\underline{\beta}^y, \overline{\beta}^y] = [6, 8]$. We also compute $\ddot{\alpha}^y = \ddot{\alpha} + \sum_{i=1}^{n} y_i = 34$ and $\ddot{\beta}^y = n + \ddot{\beta} = 7.5$.

Since $\alpha \neq \ddot{\alpha}$ and $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} < \underline{\beta}$, we use case BU3 of lemma 4 we get the blended prior $\widetilde{P} = \text{gamma}(\alpha, \underline{\beta})$ and then the BU is $\left(\widetilde{P}\right)^y = \text{gamma}(\alpha^y, \underline{\beta}^y)$. In other hand, we have $\alpha^y \neq \ddot{\alpha}^y$ and $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y \simeq 6.84 \in [\underline{\beta}^y, \overline{\beta}^y] = [6, 8]$, we use case UB1 of lemma 6 and then the UB is $\widetilde{P^y} = \text{gamma}(\alpha^y, \frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y)$. Hence, BU and UB are different. □

Theorem 4. *If $\alpha = \ddot{\alpha}$, BU and UB distributions of the Poisson-gamma model are the same distribution.*

Proof. We have by assumption $\alpha = \ddot{\alpha}$ this implies $\alpha^y = \ddot{\alpha}^y$, thus $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} = \ddot{\beta}$ and $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y = \ddot{\beta}^y$. We have three cases according to the position of $\ddot{\beta}$ and $\ddot{\beta}^y$.

If $\ddot{\beta} \in [\underline{\beta}, \overline{\beta}]$ which is equivalent to $\underline{\beta} \leq \ddot{\beta} \leq \overline{\beta}$, this implies $\underline{\beta} + n \leq \ddot{\beta} + n \leq \overline{\beta} + n$ thus $\underline{\beta}^y \leq \ddot{\beta}^y \leq \overline{\beta}^y$. By using first case of lemma 3 and lemma 5 the BU and UB are the same distribution, $\left(\widetilde{P}\right)^y = \widetilde{P^y}$, and has gamma$(\ddot{\alpha}^y, \ddot{\beta}^y)$.

If $\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\ddot{\beta} > \overline{\beta}$ that means $\underline{\beta} < \overline{\beta} < \ddot{\beta}$, this implies $\underline{\beta} + n < \overline{\beta} + n < \ddot{\beta} + n$ thus $\underline{\beta}^y < \overline{\beta}^y < \ddot{\beta}^y$. By using second case of lemma 3 and lemma 5 the BU and UB are the same distribution, $\left(\widetilde{P}\right)^y = \widetilde{P^y}$, and has gamma$(\alpha^y, \overline{\beta}^y)$.

If $\ddot{\beta} \notin [\underline{\beta}, \overline{\beta}]$ and $\ddot{\beta} < \underline{\beta}$ that means $\ddot{\beta} < \underline{\beta} < \overline{\beta}$, this implies $\ddot{\beta} + n < \underline{\beta} + n < \overline{\beta} + n$ thus $\ddot{\beta}^y < \underline{\beta}^y < \overline{\beta}^y$. By using third case of lemma 3 and lemma 5 the BU and UB are the same distribution, $\left(\widetilde{P}\right)^y = \widetilde{P^y}$, and has gamma$(\alpha^y, \underline{\beta}^y)$. □

4 Case studies

In this section, we show three case studies:

- Two case studies illustrating the difference between UB and BU:
  - Section 4.1 features a case study on the coronavirus in Canada using the Poisson-gamma model (Example 1 of the introduction section).
  - Section 4.2 features a case study on a new biomedical treatment for the early treatment of myocardial infarction using the normal-normal model (Example 2 of the introduction section).
- Section 4.3 has a breast cancer case study in which UB but not BU is available.

4.1 Application of BU and UB of the Poisson-gamma model on the coronavirus

4.1.1 Background

This section illustrates the methods of this paper by applying them to the problem of Example 1. We are interested in knowing the expected value of the daily number of deaths caused by COVID-19 in Canada for the period between March 15, 2020 and August 15, 2020. In our model, we will use posterior predictive p-value method to determine the set of adequate prior distributions, and then derive from that set the set of adequate posterior distributions. Both the BU and UB approaches will be used to make inference for this problem.

4.1.2 Determine the adequate set

Let Y_i be the number of deaths on day i due to COVID-19 in Canada for the period between March 15, 2020 and August 15, 2020. We assumed they are independent for simplicity. Let y_i be the observed value of Y_i, and θ be the expectation of number of deaths per day for that period, $E(Y_i|\theta) = \theta$. Thus, the Poisson-gamma model is defined by the following

$$\begin{cases} Y_i|\theta \sim Poisson(\theta) & Y_1, Y_2, ..., Y_n \text{ are i.i.d} \\ \theta \sim \text{gamma}(\alpha, \beta) & \\ E(\theta) = \frac{\alpha}{\beta} \text{ is the expectation of } \theta & \\ V(\theta) = \frac{\alpha}{\beta^2} \text{ is the variance of } \theta. & \end{cases}$$

To build the prior distribution, gamma(α, β), we have to choose values for the hyperparameters α and β. The hyperparameter β belongs to set of adequate hyperparameter values $H = [\underline{\beta}, \overline{\beta}]$ which can be determined using the posterior predictive p-value method. While we fix the hyperparameter $\alpha = 35$, later we will discuss the choice of this value.

We drew a random sample of $n = 30$ days with replacement from the COVID-19 database in Canada for a total of 64 days. Let $y = (y_1, y_2, ..., y_n)$ represent that sample. The number

of deaths in the n days is $\sum_{i=1}^{n} y_i = 2510$. Let $T(y)$ be a test statistic, in our model the test statistic is equal to the sample mean, $T(y) = \bar{y}$. Figure 2 shows the histogram of the sample data that we analyzed.

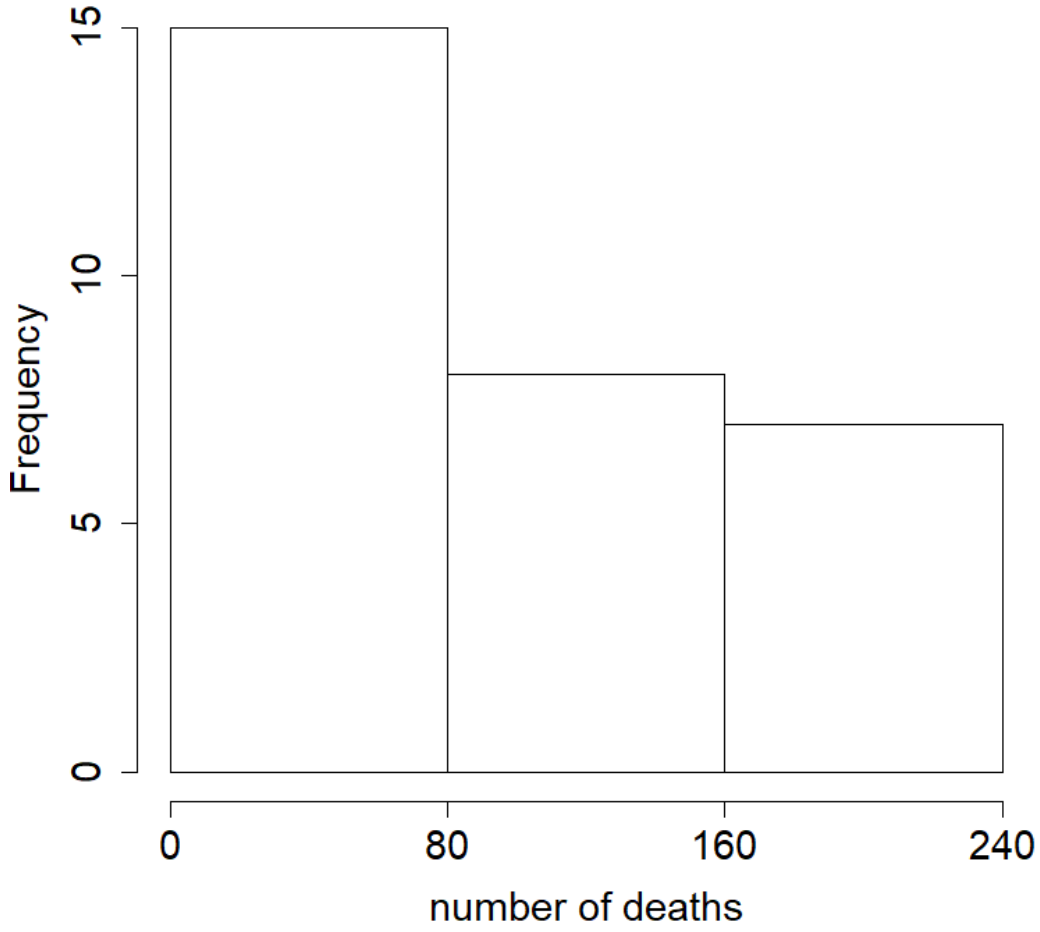

Figure 2: The histogram of a sample of 30 days of data from the COVID-19 database in Canada.

The posterior distribution of θ is given by

$$\theta|y \sim \text{gamma}(\sum_{i=1}^{n} y_i + \alpha\,,\, n + \beta).$$

The posterior predictive distribution of a future observation y^{rep} is negative binomial

(NB), and its probability mass function (pmf) is given by

$$\begin{aligned} p(y^{rep}|y) &= \int_{\Theta} p(y^{rep}|\theta)p(\theta|y)d\theta \\ &= \frac{\Gamma(y^{rep}+\sum_{i=1}^{n}y_i+\alpha)}{\Gamma(\sum_{i=1}^{n}y_i+\alpha)\Gamma(y^{rep}+1)}\left(\frac{n+\beta}{n+\beta+1}\right)^{\sum_{i=1}^{n}y_i+\alpha}\left(\frac{1}{n+\beta+1}\right)^{y^{rep}} \end{aligned}$$

$$Y^{rep}|y \sim \text{NB}(r\,,\,p),$$

where $p=\frac{n+\beta}{n+\beta+1}$, $q=1-p=\frac{1}{n+\beta+1}$ and $r=\sum_{i=1}^{n}y_i+\alpha$ (Hyvönen and Tolonen, 2019, section 2.1.2).

The expectation of negative binomial distribution is given by

$$E(Y^{rep}|y)=r\frac{q}{p}=\frac{\sum_{i=1}^{n}y_i+\alpha}{n+\beta}$$

and the variance is given by

$$V(Y^{rep}|y)=r\frac{q}{p^2}=\frac{\sum_{i=1}^{n}y_i+\alpha}{(n+\beta)^2}(n+\beta+1).$$

The posterior predictive p-value (Rubin, 1984; Gelman et al., 1996) is defined by

$$\begin{aligned} \text{p-value} &= P[T(Y^{rep})>T(y)\,|y] \\ &= 1-P[\bar{Y}^{rep}\leq\bar{y}\,|y] \\ &= 1-P[\sum_{i=1}^{n}Y_i^{rep}\leq\sum_{i=1}^{n}y_i\,|y] \\ &= 1-F(n\bar{y}), \end{aligned}$$

where $F(n\bar{y})$ is the cumulative distribution function of the posterior predictive distribution $\text{NB}(nr\,,\,p)$.

For a threshold δ, the set of adequate hyperparameter values of β is given by

$$H = \{\beta\,:\,\text{p-value}=1-F_{\beta}(n\bar{y})>\delta\}.$$

For a threshold $\delta=0.05$, the set of adequate hyperparameter values of β becomes

$$
\begin{aligned}
H &= \{\beta : 1 - F_\beta(n\bar{y}) > 0.05\} \\
H &= (0\,,\, 1.45].
\end{aligned}
$$

4.1.3 Determine the BU and UB

The average number of daily deaths caused by COVID-19 in Canada until April 23, 2020 was 54, as we showed in Example 1, the benchmark prior distribution of θ is $\ddot{P} = \text{gamma}(270, 5)$ its mean is equal to 54, the corresponding benchmark posterior distribution is $\ddot{P}^y = \text{gamma}(2780\ ,\ 35)$. The sets of adequate prior distributions are

$$
\dot{\mathcal{P}} = \{\text{gamma}(35\ ,\ \beta) : \beta \in (0\,,\, 1.45]\}\,, \text{ and } \dot{\mathcal{P}}(\ddot{P}) = \left\{\dot{P} \in \dot{\mathcal{P}} : \text{RE}(\dot{P} \parallel \ddot{P}) < \infty\right\}.
$$

The corresponding sets of adequate posterior distributions of θ are given by

$$
\dot{\mathcal{P}}^y = \{\text{gamma}(2545\ ,\ \beta^y) : \beta^y \in (30\,,\, 31.45]\}\,, \text{ and } \dot{\mathcal{P}}^y(\ddot{P}^y) = \left\{\dot{P}^y \in \dot{\mathcal{P}}^y : \text{RE}(\dot{P}^y \parallel \ddot{P}^y) < \infty\right\}.
$$

We have $\alpha \neq \ddot{\alpha}$ and $\frac{\alpha}{\ddot{\alpha}}\ddot{\beta} \simeq 0.65 \in H$, according to case BU1 of lemma 4 the BU is $\left(\widetilde{P}\right)^y = \text{gamma}(\alpha^y, \frac{\alpha}{\ddot{\alpha}}\ddot{\beta} + n)$ which is equal to $\text{gamma}(2545, 30.65)$. On the other hand we have $\alpha^y \neq \ddot{\alpha}^y$ and $\frac{\alpha^y}{\ddot{\alpha}^y}\ddot{\beta}^y \simeq 32.04 \notin (30\,,\, 31.45]$, then according to case UB2 of lemma 6 the UB is $\widetilde{P^y} = \text{gamma}(\alpha^y, \overline{\beta}^y)$ which is $\text{gamma}(2545, 31.45)$. Hence, BU and UB are different, this is what we can see in Figure 3 the probability density functions of BU, UB and benchmark posterior are different.

We have chosen $\alpha = 35$ for $\text{gamma}(\alpha\ ,\ \beta)$, the prior distribution, in the set of adequate prior distributions $\dot{\mathcal{P}}$. For any $\alpha < 120$, we will almost get similar results, which in turn show that BU is different from UB. Otherwise, we will get BU and UB are the same.

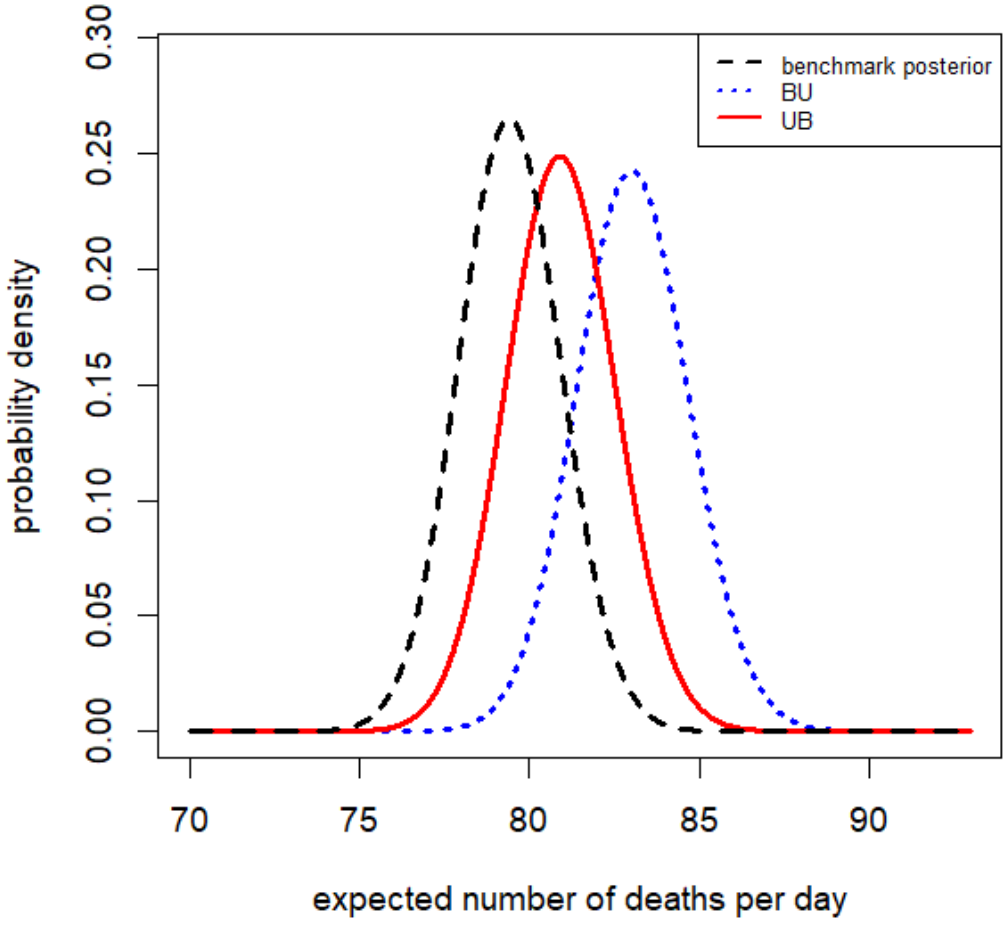

Figure 3: Comparison of the probability density function of the benchmark posterior (dashed), BU (dotted) and UB (solid) of average daily deaths caused by COVID-19 in Canada.

4.1.4 Interpretation

In Table 2, for each approach, we have calculated the posterior mean of θ that we defined in Section 4.1.2, the posterior variance and the 95% credible interval of θ. The benchmark posterior mean of θ is 79 deaths, whereas the posterior mean of θ using BU is 83 deaths, and it is 81 using UB. The mean values of both approaches are close to each other, this is due to a small range, 1.45, of the set of adequate hyperparameter values of β, $(30\,,\,31.45]$. The

Approach	distribution	mean	variance	95% Credible interval of θ
BU	$\left(\widetilde{P}\right)^y = \text{gamma}(2545, 30.65)$	83.03	2.71	[79.8 , 86.3]
UB	$\widetilde{P^y} = \text{gamma}(2545, 31.45)$	80.92	2.57	[77.8 , 84.1]

Table 2: The mean, the variance and the 95% credible interval of θ for each approach BU and UB.

95% credible interval of θ based on the BU is $[79.8, 86.3]$ and for that of UB is $[77.8, 84.1]$. We can conclude that the two approaches for this application give very close results with a small difference.

4.2 Application of BU and UB of the normal-normal model on new biomedical treatment

4.2.1 Background

This section illustrates the methods of this paper by applying them to the problem of Example 2. As we have seen in the Example 2, the aim of this study is to compare placebo and a new treatment to be given at home as soon as possible after a myocardial infarction. We have the information that the hyperparameter μ of our prior distribution belongs to a set of adequate hyperparameter values $[\underline{\mu}, \overline{\mu}]$. We will determine the adequate set of the prior distribution in the next section. Our model becomes

$$\begin{cases} Y|\theta \sim N(\theta, 4) \\ \theta \sim N(\mu, 0.0169). \end{cases}$$

The posterior distribution of θ is normal, and using Equation (4) we get

$$\theta|y \sim N(\mu^y = 0.886\mu - 0.084, v^y = 0.015). \tag{11}$$

4.2.2 Determine the adequate set

The posterior predictive distribution of a future observation y^{rep} is normal distribution, and its probability density function (pdf) is given by

$$
\begin{aligned}
p(y^{rep}|y) &= \int_{\Theta} p(y^{rep}|\theta)p(\theta|y)d\theta \\
&= \int_{\Theta} \phi_{\theta,\sigma^2}(y^{rep})\phi_{\mu^y,v^y}(\theta)d\theta \\
&= \frac{1}{\sqrt{2\pi(\sigma_n^2+\sigma^2)}} e^{-\frac{(y^{rep}-\mu^y)^2}{2(v^y+\sigma^2)}}, \qquad (12)
\end{aligned}
$$

where $\phi_{\theta,\sigma^2}(y^{rep})$ is the normal pdf of y^{rep}, with mean θ and variance σ^2, and $\phi_{\mu^y,v^y}(\theta)$ is the normal pdf of θ with mean μ^y and variance v^y defined in Equation (11).

For the proof of Equation (12) see Equation 2.115 of Bishop (2006). The posterior predictive distribution of y^{rep} has the same mean as the posterior distribution of θ in Equation (11), $\mu^y = 0.886\mu - 0.084$, and its variance is $v^y + \sigma^2$, thus

$$Y^{rep}|y \sim N(0.886\mu - 0.084, 4.015).$$

Let $T(Y)$ be a test statistic, in our model the test statistic is equal to the sample mean, $T(Y) = \bar{Y}$. The one-sided posterior predictive p-value is

$$
\begin{aligned}
\text{pp} &= P[T(Y^{rep}) > T(y)\,|y] \\
&= 1 - P[\bar{Y}^{rep} \leq \bar{y}\,|y] \\
&= 1 - P[\frac{\bar{Y}^{rep} - E(\bar{Y}^{rep}|y)}{sd(\bar{Y}^{rep}|y)} < \frac{\bar{y} - E(\bar{Y}^{rep}|y)}{sd(\bar{Y}^{rep}|y)}] \\
&= 1 - \Phi(\frac{\bar{y} - E(\bar{Y}^{rep}|y)}{sd(\bar{Y}^{rep}|y)}) \\
&= 1 - \Phi(\frac{-0.656 - 0.886\mu}{0.363}),
\end{aligned}
$$

where Φ is the cumulative distribution function of the standard normal distribution. To convert the one-sided p-value to two-sided p-value we use one of these two formulas: $2\min(\Phi(\frac{-0.656-0.886\mu}{0.363}), 1-\Phi(\frac{-0.656-0.886\mu}{0.363}))$ or $2(1-\Phi(|\frac{-0.656-0.886\mu}{0.363}|))$. Therefore, the two-sided posterior predictive p-value becomes

$$\text{p-value} = 2(1 - \Phi(|\frac{-0.656 - 0.886\mu}{0.363}|)).$$

For a threshold δ, the set of adequate hyperparameter values of μ is given by

$$H = \{\mu : \text{p-value} > \delta\}.$$

For a threshold $\delta = 0.05$, the set of adequate hyperparameter values of μ becomes

$$
\begin{aligned}
H &= \{\mu : 2(1-\Phi(|\frac{-0.656-0.886\mu}{0.363}|)) > 0.05\} \\
H &= [-1.542\,,\, 0.062].
\end{aligned}
$$

Let H^y be the set of adequate posterior mean μ^y. By using Equations (5) and (6) we get

$$H^y = [-1.45\,,\, -0.03].$$

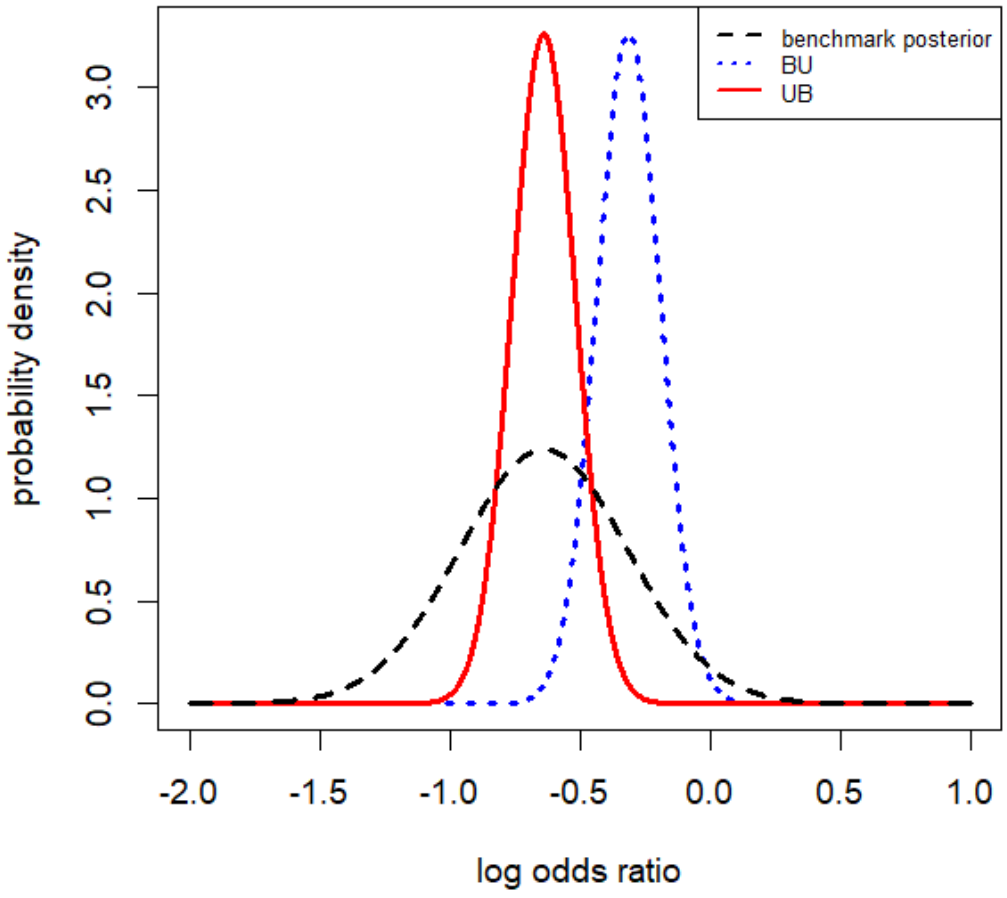

Figure 4: Comparison of the probability density function of the benchmark posterior (dashed), BU (dotted) and UB (solid) of $\theta = \log(\text{OR})$.

4.2.3 Determine the BU and UB

In this application we will use the prior of Spiegelhalter et al. (2004, pp. 26, 69) as a benchmark prior ($\ddot{P}$) with different variance, thus $\theta \sim N(\ddot{\mu} = -0.26\ ,\ \ddot{v} = 0.5)$. The corresponding benchmark posterior distribution ($\ddot{P}^y$) is $N(\ddot{\mu}^y = -0.64\ ,\ \ddot{v}^y = 0.1)$. The set of adequate prior distribution is $\dot{\mathcal{P}} = \{N(\mu\ ,\ v = 0.0169) : \mu \in [-1.542\,,\ 0.062]\}$, and $\dot{\mathcal{P}}(\ddot{P}) = \left\{\dot{P} \in \dot{\mathcal{P}} : \mathrm{RE}(\dot{P} \parallel \ddot{P}) < \infty\right\}$. The set of adequate posterior distributions of θ is given by $\dot{\mathcal{P}}^y = \{N(\mu^y\ ,\ v^y = 0.015) : \mu^y \in [-1.45\,,\ -0.03]\}$, and $\dot{\mathcal{P}}^y(\ddot{P}^y) = \left\{\dot{P} \in \dot{\mathcal{P}}^y : \mathrm{RE}(\dot{P}^y \parallel \ddot{P}^y) < \infty\right\}$.

We have $\ddot{\mu} = -0.26 \in H$, according to case BU1 of lemma 1 the BU is $\left(\widetilde{P}\right)^y = N(-0.31\ ,\ 0.015)$. On the other hand we have $\ddot{\mu}^y \in [-1.45\,,\ -0.03]$, then according to case UB1 of lemma 2 the UB is $\widetilde{P^y} = N(-0.64\ ,\ 0.015)$. Figure 4 shows that the pdfs of BU, UB and benchmark posterior are different. However, the UB and benchmark posterior have the same mean, $\widetilde{\mu^y} = \ddot{\mu}^y = -0.64$.

4.2.4 Interpretation

Remember that, for a $\theta < 0$ this is equivalent to OR < 1, it is to favor the new treatment, because mortality with the new intervention is low. The values of this analysis are summarized in Table 3. The estimated OR for both approaches is smaller than 1, for the UB approach is 0.53 with a corresponding risk reduction 47%, but for that of the BU is equal to 0.73 with a corresponding risk reduction 27%. Moreover, the 95% credible interval of OR corresponding to 95% credible interval of θ is $[0.41, 0.67]$ which is more confident (compared to its distance from OR $= 1$) than of that of BU which is equal to $[0.58, 0.93]$. The posterior probability of OR < 0.5 (equivalent to $P(\theta < -0.69|y)$) corresponding to risk reduction at least 50% is equal to 0.001 for BU, while, it is more important for UB which is equal to 0.338. Also, The posterior probability of OR < 1 (equivalent to $P(\theta < 0|y)$) is equal to 99.5% for BU and 100% for UB.

We can conclude from the above analysis that the benefit is for the new treatment, with strong confirmation results obtained by the UB approach.

Distribution	Estimated OR	95% Credible interval of θ	$P(\text{OR} < 0.5 \mid y)$	$P(\text{OR} < 1 \mid y)$
$\left(\widetilde{P}\right)^y = N(-0.31\ ,\ 0.015)$	$e^{-0.31} = 0.73$	$[-0.55, -0.07]$	0.001	0.995
$\widetilde{P^y} = N(-0.64\ ,\ 0.015)$	$e^{-0.64} = 0.53$	$[-0.88, -0.40]$	0.338	1

Table 3: The estimated OR, the 95% credible interval of θ, the probability of OR smaller than 0.5 and the probability of OR smaller than 1 for each approach.

4.3 Application of breast cancer

4.3.1 Background

The SELDI-TOF spectra data set (ProData package, available at `https://rb.gy/yjxrej`) from Alex Miron's laboratory at the Dana-Farber Cancer Institute (Li, 2009) consists of the log-abundance levels of 20 plasma proteins in 35 women with ER/PR positive breast cancer (disease group) and 64 healthy women (control group). This is a multiple hypothesis testing problem, with one test per protein of the null hypothesis that the mean protein log-abundance level does not differ between the disease and control groups. The data can be reduced to a vector of t-test statistics between the disease and control groups. The null hypothesis of the t-test considers the ith protein unaffected by the cancer, whereas the alternative hypothesis considers it affected. Many estimators of the local false discovery rate can be computed for each protein, such as the corrected false discovery rate (CFDR) (Bickel and Rahal, 2019), the re-ranked false discovery rate (RFDR) (Bickel and Rahal, 2019), the histogram-based estimator (HBE) (Efron, 2004), the binomial-based estimator (BBE) (Bickel, 2013) or the maximum likelihood estimator (MLE) (Padilla and Bickel, 2012, and references). What if there is uncertainty about which estimator to use?

4.3.2 Determine the UB

Among the 20 proteins in ProData package, we are interested in finding proteins affected by cancer. As an example of data visualization, Figure 5 shows the histogram of protein 5 for the cancer group (left) and the control group (right). Let's remember the definition of

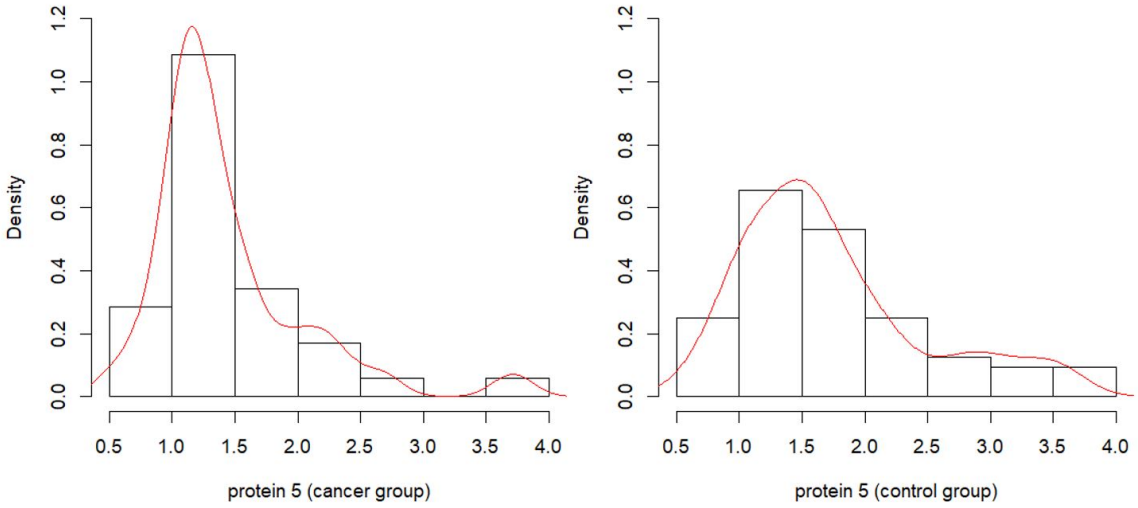

Figure 5: Histograms of protein 5 for both cancer group (left) and control group (right).

LFDR which can be used to find the affected proteins by the cancer.

The ProData can be reduced to some test statistics, say $X_1, X_2, ..., X_N$, where N is the total number of protein, $N = 20$. The null hypothesis of the ith test considers the protein is unaffected by the cancer against the alternative which considers it affected. Denote by $p(X_i)$ the p-value for the ith test, such that $i = 1, 2, ..., N$. Let A_i be the random quantity that defined by

$$A_i = \begin{cases} 0 & \text{if the } i\text{th null hypothesis is true} \\ 1 & \text{if the } i\text{th alternative hypothesis is true.} \end{cases}$$

The LFDR is the posterior probability that null hypothesis i is true given the p-value (Bickel and Rahal, 2019), thus

$$\text{LFDR}(p(x_i)) = P[A_i = 0 | p(X_i) = p(x_i)]. \tag{13}$$

$A_i | p(X_i) = p(x_i)$ is the random variable drawn from the Bernoulli distribution with parameter $1-\text{LFDR}(p(x_i))$.

The nonlocal false discovery rate (NFDR) (Efron and Tibshirani (2002); Bickel (2013)) at a significance level of α is

$$\text{NFDR}(\alpha) = P[A_i = 0 | p(X_i) \leq \alpha].$$

There are several estimators of the LFDR defined in Equation (13). In this application we use four of them, CFDR, RFDR, HBE and MLE. As all these estimators do not return the same estimation of LFDR, we are interested in defining a new estimator, UB, of the LFDR using the NFDR estimate to generate the benchmark distribution $\ddot{P}^y = \text{Bern}(1-\widehat{\text{NFDR}})$, where, following Bickel and Rahal (2019),

$$\widehat{\text{NFDR}}(p(x_j)) = \frac{p(x_j)N}{\#(p(x_i) \leq p(x_j))} \text{ if } \frac{p(x_j)N}{\#(p(x_i) \leq p(x_j))} \leq 1 \text{ else } \widehat{\text{NFDR}}(p(x_j)) = 1.$$

Let $\dot{\mathcal{P}}^y$ denote this set of adequate posterior distributions:

$$\dot{\mathcal{P}}^y = \left\{ \text{Bern}(1-\widehat{\text{LFDR}(p(x))}) : \text{LFDR} \in \{\text{CFDR, RFDR, HBE, MLE}\} \right\}, \text{ and}$$

$$\dot{\mathcal{P}}^y(\ddot{P}^y) = \left\{ \dot{P}^y \in \dot{\mathcal{P}}^y : \text{RE}(\dot{P}^y \parallel \ddot{P}^y) < \infty \right\}.$$

For each protein i, the UB can be used to estimate a single LFDR from the set of estimates, thus

$$\begin{aligned}
\widetilde{\text{LFDR}}_i &= \underset{\dot{P}_i^y \in \dot{\mathcal{P}}_i^y(\ddot{P}_i^y)}{\arg\min} \text{RE}(\dot{P}_i^y \parallel \ddot{P}_i^y) \\
&= \underset{\text{LFDR} \in \{\text{CFDR, RFDR, HBE, MLE}\}}{\arg\min} \text{RE}\left(\text{Bern}(1-\widehat{\text{LFDR}(p(x_i))}) || \text{Bern}(1-\widehat{\text{NFDR}(p(x_i))})\right) \\
&= \underset{\text{LFDR} \in \{\text{CFDR, RFDR, HBE, MLE}\}}{\arg\min} (1-\widehat{\text{LFDR}(p(x_i))}) \log \frac{(1-\widehat{\text{LFDR}(p(x_i))})}{(1-\widehat{\text{NFDR}(p(x_i))})} + \widehat{\text{LFDR}(p(x_i))} \log \frac{\widehat{\text{LFDR}(p(x_i))}}{\widehat{\text{NFDR}(p(x_i))}}.
\end{aligned}$$

Note that there is no BU available for this application, since by definition the LFDR is a posterior probability and the BU is mainly based on a set of adequate prior distributions which is not the case here.

4.3.3 Interpretation

We used three R packages in our computations of LFDR estimators. The package *locfdr* (Efron et al., 2015) used to compute the HBE, we use the standard normal quantile of the one-sided p-value as the z-score, we set the number of histogram breaks in the discretization of the z-score equal to the minimum of 120 and number of row of cancer group, and we set the degrees of freedom equal to 7 for fitting the mixture estimated density. The package *LFDR.MLE* (Padilla et al., 2015) used to compute the MLE, we use the statistic of the t test between cancer and control groups with its degrees of freedom 97. The package *CorrectedFDR* (Rahal et al., 2018) used to compute the NFDR, CFDR, RFDR and UB, we use the two-sided p-value of the t test between cancer and control groups.

Figure 6 shows that some estimators are below the threshold 0.2, all LFDR estimators with a probability greater than 80% for a protein affected by cancer can be found below this threshold. We summary those proteins (3, 4 and 14) in Table 4, it is clear that the UB estimator chooses from the LFDR estimators the values closest to those of the NFDR. The advantage of UB is to provide mixture values from the different estimators, the $3rd$ protein was chosen from CFDR, while the $4th$ protein was chosen from RFDR and the $14th$ protein was chosen from HBE.

Protein	NFDR	CFDR	RFDR	MLE	HBE	UB
3	0.00003	**0.00003**	0.00712	0.00024	0.00163	**0.00003**
4	0.15497	0.28412	**0.13055**	0.65316	1.00000	**0.13055**
14	0.00712	0.01068	0.15497	0.02159	**0.00442**	**0.00442**

Table 4: Summary of proteins that have a UB less than 0.2 as well as other LFDR estimators for these proteins.

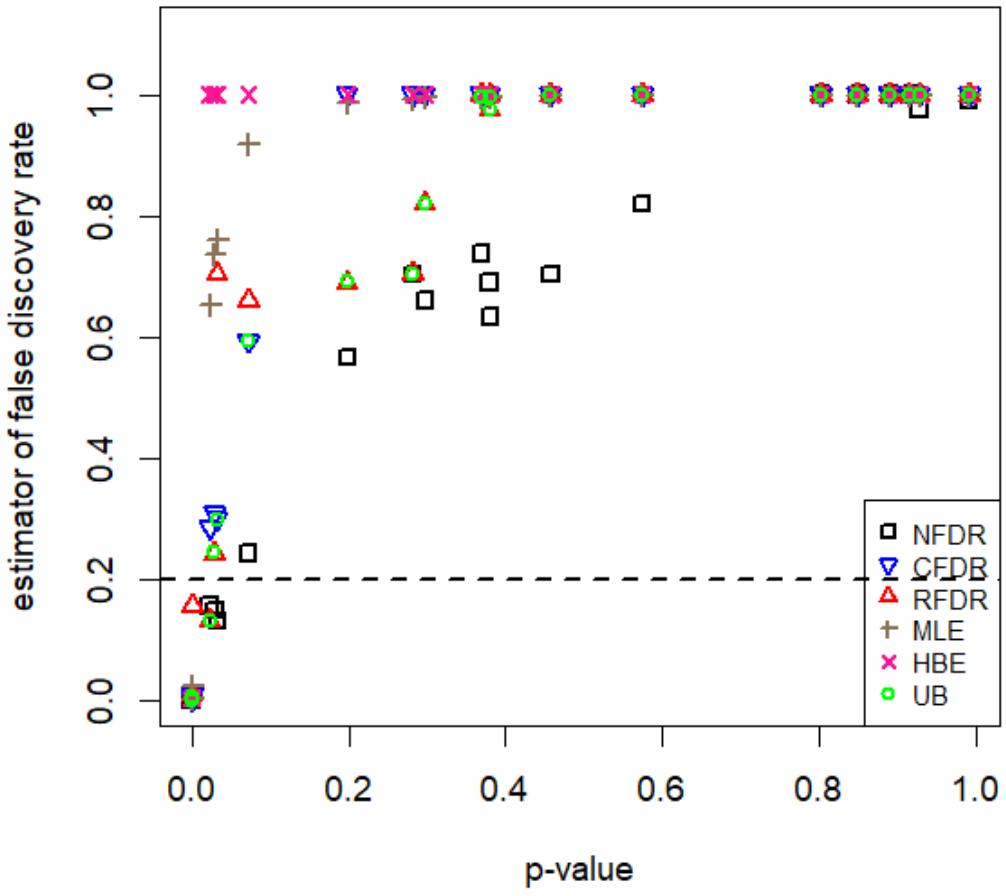

Figure 6: In x-axis the p-value of t test statistics, in y-axis NFDR, CFDR, RFDR, MLE, HBE and UB estimators for each protein of breast cancer data.

5 Conclusion

This paper proposes the BU, new approach of blended distribution that select a prior distribution from a set of adequate prior distributions and derive from that prior a posterior distribution. For the normal-normal model and the Poisson-gamma model, the BU and UB do not necessarily yield the same distribution, they can be the same under some assumptions. For inference, we presented three applications in this paper. The first application, in

Section 4.1, estimated the expected number of deaths caused by coronavirus in Canada. The second application, in Section 4.2, studied a new biological treatment compared to a placebo administrated at home after a myocardial infarction. The third application, in section 4.3, defined a new estimator of the local false discovery rate to find the proteins most likely affected by breast cancer.

In the first two applications (Sections 4.1 and 4.2), we found the results of BU and UB are close, but that may be attributed to the small range of the set of adequate hyperparameter values. Had the range been larger, we might have seen a much larger difference between BU and UB. On the other hand, we chose some of the hyperparameters to accentuate the differences between BU and UB, as specified in Examples 1 and 2.

That suggests that whether BU or UB is used may not have any important effect on conclusions. If so, the choice could be left to personal preference or computational convenience. For example, BU would be computationally faster than UB if each use of a posterior distribution requires many Markov chain Monte Carlo iterations (Mayer Alvo, personal communication).

Further work

- In this paper, we have shown the blended distribution based on the set of adequate prior or posterior distributions for the normal-normal and Poisson-gamma models. For these two models we found that the BU and UB do not necessarily yield the same distribution. For future work, it would be interesting to check if this result can be extended to all distributions for the exponential family.

- In the set of adequate prior or posterior distributions that we showed in this paper, it was only one parameter unknown. It would be interesting to consider two or more unknown parameters. In this case the gradient descent can be used to solve the minimization problem for two dimensions or more.

- We used relative entropy to minimize the set of adequate prior or posterior distributions to the benchmark distribution. It would be interesting to check if the reverse relative entropy or Euclidean distance yield the same results as the relative entropy.

Acknowledgments

We gratefully acknowledge Lisa Strug and Mayer Alvo for valuable feedback. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009) and by the Faculty of Medicine of the University of Ottawa.

Conflict of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Arias-Nicolás, J., Martín, J., Ruggeri, F., Suárez-Llorens, A., 2009. Optimal actions in problems with convex loss functions. International Journal of Approximate Reasoning 50, 303 – 314.

Augustin, T., 2002. Expected utility within a generalized concept of probability - a comprehensive framework for decision making under ambiguity. Statistical Papers 43, 5–22.

Augustin, T., 2004. Optimal decisions under complex uncertainty - basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. Zeitschrift fur Angewandte Mathematik und Mechanik 84, 678–687. doi:{10.1002/zamm.20410151}.

Bayati Eshkaftaki, A., Parsian, A., 2011. Robust Bayes Estimation. Communications in Statistics - Theory and Methods 40, 929–941.

Benavoli, A., Ristic, B., 2011. Classification with imprecise likelihoods: A comparison of tbm, random set and imprecise probability approach.

Berger, J., Berliner, L.M., 1986. Robust Bayes and Empirical Bayes Analysis with ϵ-Contaminated Priors. The Annals of Statistics 14, pp. 461–486.

Bickel, D.R., 2013. Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Statistical Applications in Genetics and Molecular Biology 12, 529–543.

Bickel, D.R., 2015a. Blending Bayesian and frequentist methods according to the precision of prior information with applications to hypothesis testing. Statistical Methods & Applications 24, 523–546.

Bickel, D.R., 2015b. Inference after checking multiple Bayesian models for data conflict

and applications to mitigating the influence of rejected priors. International Journal of Approximate Reasoning 66, 53–72.

Bickel, D.R., 2018. Bayesian revision of a prior given prior-data conflict, expert opinion, or a similar insight: A large-deviation approach. Statistics 52, 552–570.

Bickel, D.R., 2020. Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors. Communications in Statistics - Theory and Methods 49, 2703–2712.

Bickel, D.R., Rahal, A., 2019. Correcting false discovery rates for their bias toward false positives. Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1630432 .

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Bityukov, S., Krasnikov, N., Nadarajah, S., Smirnova, V., 2011. Confidence distributions in statistical inference. AIP Conference Proceedings 1305, 346–353. doi:10.1063/1.3573637.

Cohen, M., Chateauneuf, A., Danan, E., Gajdos, T., Giraud, R., Jeleva, M., Philippe, F., Tallon, J.M., Vergnaud, J.C., 2010. Tribute to Jean-Yves Jaffray. Theory and Decision 71, 1–10. doi:10.1007/s11238-010-9210-y.

de Cooman, G., Hermans, F., 2008. Imprecise probability trees: Bridging two theories of imprecise probability. Artificial Intelligence 172, 1400–1427.

Efron, B., 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99, 96–104.

Efron, B., Tibshirani, R., 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, 70–86.

Efron, B., Turnbull, B., Narasimhan, B., Strimmer, K., 2015. locfdr: Computes Local False Discovery Rates. R package version 1.1-8, https://goo.gl/1e284V.

Gelman, A., Meng, X.L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica 6, 733–760.

Gilio, A., 2005. Probabilistic logic under coherence, conditional interpretations, and default reasoning. Synthese 146, 139–152.

Good, I., 1983. Good Thinking: the Foundations of Probability and Its Applications. G - Reference, Information and Interdisciplinary Subjects Series, University of Minnesota Press.

Hable, R., 2009. Data-based decisions under imprecise probability and least favorable models. International Journal of Approximate Reasoning 50, 642–654. doi:{10.1016/j.ijar.2008.03.009}.

Hannig, J., Xie, M., 2012. A note on Dempster-Shafer recombination of confidence distributions. Electronic Journal of Statistics 6, 1943–1966. doi:10.1214/12-EJS734.

Harremoës, P., 2007. Information topologies with applications, in: Csiszár, I., Katona, G.O.H., Tardos, G., Wiener, G. (Eds.), Entropy, Search, Complexity. Springer Berlin Heidelberg, Berlin, Heidelberg. volume 16 of *Bolyai Society Mathematical Studies*, pp. 113–150. doi:10.1007/978-3-540-32777-6.

Hurwicz, L., 1951. Optimality criteria for decision making under ignorance. Cowles Commission Discussion Paper 370.

Hyvönen, V., Tolonen, T., 2019. Bayesian inference 2019 .

Jozani, M., Marchand, É., Parsian, A., 2012. Bayesian and robust bayesian analysis under a general class of balanced loss function. Statistical Papers 53, 51–60.

Kiapour, A., Nematollahi, N., 2011. Robust bayesian prediction and estimation under a squared log error loss function. Statistics and Probability Letters 81, 1717–1724.

Kim, D., Lindsay, B.G., 2011. Using confidence distribution sampling to visualize confidence sets. Statistica Sinica 21, 923–948.

Kullback, S., 1968. Information Theory and Statistics. Dover, New York.

Li, X., 2009. ProData. Bioconductor.org documentation for the ProData package .

Lynch, S.M., 2007. Introduction to applied Bayesian statistics and estimation for social scientists. Springer.

Moral, S., De Campos, L., 1991. Updating uncertain information, in: Bouchon-Meunier, B., Yager, R., Zadeh, L. (Eds.), Uncertainty in Knowledge Bases. Springer Berlin / Heidelberg. volume 521 of *Lecture Notes in Computer Science*, pp. 58–67.

Nadarajah, S., Bityukov, S., Krasnikov, N., 2015. Confidence distributions: A review. Statistical Methodology 22, 23–46.

Padilla, M., Bickel, D.R., 2012. Estimators of the local false discovery rate designed for small numbers of tests. Statistical Applications in Genetics and Molecular Biology 11, art. 4.

Padilla, M., Yang, Y., Ali, A., Leckett, K., Yang, Z., Li, Z., Yanofsky, C.M., Bickel, D.R., 2015. LFDR.MLE: Estimation of the Local False Discovery Rates by type II Maximum Likelihood Estimation. R package version 1.0, https://goo.gl/d9MTNE.

Rahal, A., Akpawu, A., Chitpin, J., Bickel, D.R., 2018. CorrectedFDR: Correcting False Discovery Rates. R package version 1.0, https://goo.gl/qMQsXa.

Rubin, D.B., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann.Statist. 12, 1151–1172.

Schweder, T., Hjort, N., 2016. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Schweder, T., Hjort, N.L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29, 309–332.

Shafer, G., 2011. A betting interpretation for probabilities and Dempster-Shafer degrees of belief. International Journal of Approximate Reasoning 52, 127–136. doi:10.1016/j.ijar.2009.05.012.

Singh, K., Xie, M., Strawderman, W.E., 2005. Combining information from independent sources through confidence distributions. Annals of Statistics 33, 159–183.

Singh, K., Xie, M., Strawderman, W.E., 2007. Confidence distribution (CD) – distribution estimator of a parameter. IMS Lecture Notes Monograph Series 2007 54, 132–150.

Sivaganesan, S., 1999. A likelihood based robust Bayesian summary. Statistics & Probability Letters 43, 5–12. doi:{10.1016/S0167-7152(98)00206-5}.

Spiegelhalter, D.J., Abrams, K.R., Myles, J.P., 2004. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Statistics in Practice, John Wiley & Sons, Ltd., Chichester.

Tapking, J., 2004. Axioms for preferences revealing subjective uncertainty and uncertainty aversion. Journal of Mathematical Economics 40, 771–797. doi:10.1016/j.jmateco.2003.07.004.

Tian, L., Wang, R., Cai, T., Wei, L.J., 2011. The highest confidence density region and its usage for joint inferences about constrained parameters. Biometrics 67, 604–10. doi:10.1111/j.1541-0420.2010.01486.x.

Veronese, P., Melilli, E., 2018. Some asymptotic results for fiducial and confidence distributions. Statistics & Probability Letters 134, 98–105.

Walley, P., 2000. Towards a unified theory of imprecise probability. International Journal of Approximate Reasoning 24, 125–148.

Wang, Z., Klir, G.J., 2008. Generalized Measure Theory (IFSR International Series on Systems Science and Engineering). Springer, New York.

Weichselberger, K., 2000. The theory of interval-probability as a unifying concept for uncertainty. International Journal of Approximate Reasoning 24, 149–170.

Weichselberger, K., 2001. Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept. Physica-Verlag, Heidelberg.

Xie, M.G., Singh, K., 2013. Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review 81, 3–39.

Chapter 3

Correcting false discovery rates for their bias toward false positives

This chapter proposes two new estimators of the local false discovery rate, the correcting false discovery rate, the re-ranking false discovery rate, and an estimator of the nonlocal false discovery rate. It also presents a stimulation study to compare the performance of the proposed estimators to two existing estimators of the local false discovery rate, the histogram-based estimator and the maximum likelihood estimator.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1630432

Correcting false discovery rates for their bias toward false positives

David R. Bickel and Abbas Rahal

Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

ABSTRACT

The way false discovery rates (FDRs) are used in the analysis of genomics data leads to excessive false positive rates. In this sense, FDRs overcorrect for the excessive conservatism (bias toward false negatives) of methods of adjusting p values that control a family-wise error rate. Estimators of the local FDR (LFDR) are much less biased but have not been widely adopted due to their high variance and lack of availability in software. To address both issues, we propose estimating the LFDR by correcting an estimated FDR or the level at which an FDR is controlled.

ARTICLE HISTORY
Received 20 July 2018
Accepted 5 June 2019

KEYWORDS
Bayesian false discovery rate; Empirical Bayes; Local false discovery rate; Principle of maximum entropy; Multiple comparison procedure; Multiple testing

1. Introduction

Your job is to provide the selection committee with a list of applicants predicted to graduate with honors if selected. While you are trying to decide on whether to report those with less than a 5%, 20%, or 50% probability of failure, a colleague informs you of a popular new procedure. It replaces each applicant's failure probability with its average with the failure probabilities of all applicants with lower failure probabilities. Having decided to predict the graduation of each applicant with less than 20% failure probability, you now wonder how to calibrate the accurately estimated average failure probability among the better candidates in order to make it reasonable as an estimate of the applicant's failure probability.

Given a gene expression data set, you need to determine which genes to consider differentially expressed. If you decide that only genes with more than 80% estimated probability of differential expression belong on the list, then every gene on the list will have less than 20% estimated probability that its null hypothesis of equivalent expression is true. That probability is the local false discovery rate (LFDR). Since that probability is difficult to estimate accurately, the false discovery rate (FDR) looks like a viable alternative. In fact, software is readily available to report the FDR achieved by each gene. However, since the FDR achieved by a gene is for practical purposes an average of the LFDRs over all genes with lower LFDRs of that gene, the achieved FDR is by

CONTACT David R. Bickel dbickel@uottawa.ca Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Department of Mathematics and Statistics, University of Ottawa, 451 Smyth Road, Ottawa, Ontario K1H 8M5, Canada.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

construction too low to use as an estimate of the null hypothesis probability. That bias requires a correction in order to construct the desired list of genes with over 80% estimated probability of differential expression.

The two bias corrections introduced in this note transform achieved FDRs into LFDR estimates. As calibrated FDR values, those LFDR estimates in effect determine reasonable FDR thresholds to apply to the output of genomics software packages. This applicability of the bias corrections to popular FDR methods achieves the primary goal of this note, to make the benefits of LFDR estimation more accessible.

Suppose all null hypotheses are rejected that have p values below α, a significance level such as 0.01. The simplest FWER is $\mathrm{FWER}(\alpha)$, the frequentist probability that at least one null hypothesis is rejected at level α assuming that all null hypotheses are true (Dudoit and van der Laan 2008). The simplest FDR, called the nonlocal FDR or $\mathrm{NFDR}(\alpha)$ (Bickel 2013), is the empirical Bayes probability that a null hypothesis rejected at level α is true given the data (Efron 2010). Let p_i denote the p value of the ith hypothesis test. This p_i depends on x_i, the observed data set for the ith biological feature. In the case that the ith null hypothesis says the ith gene has no differential expression, x_i is one or more vectors of measurements of the expression of the ith gene. The adjusted p value or *achieved FWER* for the ith test is $\mathrm{FWER}(p_i)$, the FWER at the significance level at which the ith null hypothesis is barely rejected ($\alpha = p_i$). Likewise, the *achieved NFDR* for the ith test is $\mathrm{NFDR}(p_i)$, the probability that a randomly selected p value less than p_i corresponds to a true null hypothesis. That is misleading as a measure of the significance of the ith hypothesis test since it can be substantially lower than the posterior probability that the null hypothesis is true. That posterior probability is the local false discovery rate (LFDR). In fact, the achieved NFDR of the ith null hypothesis is the expectation value of all LFDRs corresponding to p values less than p_i (Efron 2010). That is why estimates of NFDRs are often much less than estimates of LFDRs (e.g. Hong, Tibshirani, and Chu 2009; Bickel 2012, Figure 3; cf. Genovese and Wasserman 2002), leading to unwarranted reports of significance and potentially many more false positives.

This anti-conservative bias of estimators of $\mathrm{NFDR}(p_i)$, the achieved NFDR of the ith null hypothesis, undermines the most common use of the FDR. A researcher seeking to apply a multiple-comparison procedure with a list of p values typically selects either a method of FWER control or a method FDR control (Benjamini and Hochberg 1995). With the former now known for its excessive bias toward false negatives in genomics-scale applications (Dudoit and van der Laan 2008, p. 145), the latter is now seen as a viable alternative, as indicated both by its availability in data analysis software and by the literature (e.g. Van Den Oord 2008; Glickman, Rao, and Schultz 2014). The original method of FDR control (Benjamini and Hochberg 1995) remains the most commonly used. At significance level α, $\mathrm{FDR}(\alpha)$, the false discovery rate as defined by Benjamini and Hochberg (1995), tends to be similar in value to $\mathrm{NFDR}(\alpha)$. In fact, with $\widehat{\mathrm{NFDR}}(\alpha)$ as the estimator of $\mathrm{NFDR}(\alpha)$ to be defined in Sec. 2, the practice of setting α to the smallest value such that $\widehat{\mathrm{NFDR}}(\alpha) \leq q$ for some specified value q guarantees that $\mathrm{FDR}(\alpha) \leq q$ under the Benjamini and Hochberg (1995) independence assumption (Efron 2010, pp. 53-54). Unfortunately, that means the rejected hypothesis with $p_{\max,q}$, the highest p value, is rejected merely because $\widehat{\mathrm{NFDR}}(p_{\max,q}) \leq q$ even though

$\mathrm{NFDR}(p_{\max,q})$ is the probability of the truth of a random null hypothesis with a p value *less than* $p_{\max,q}$ rather than *equal to* $p_{\max,q}$.

As the flaw in the procedure is practical, it may be most clearly seen in terms of particular applications to genomics data. Suppose the null hypothesis is that a gene is not differentially expressed. To the extent that $\widehat{\mathrm{NFDR}}(p_{\max,q})$ accurately estimates $\mathrm{NFDR}(p_{\max,q})$, selecting a gene for further consideration on the basis of FDR control in effect selects a gene on the basis of the probability that a more significant gene is differentially expressed. That is equivalent to selecting an applicant for admission on the basis of an average of his or her probability of failure with the failure probabilities of all the better applicants.

Later methods of FDR control do not necessarily fare any better. The opposite is often the case since many such methods are designed to be even less conservative than the Benjamini and Hochberg (1995) procedure (e.g. Benjamini and Liu 1999), resulting in even more unwarranted discoveries and the accompanying bias toward false positives. In short, that bias is not a special feature of a particular FDR algorithm but rather of every successful algorithm, for the way the FDR is defined automatically leads to that bias.

The strategy in Section 3 is not to replace FDR methods but rather to take advantage of their stability (low variance compared to LFDR estimators (cf. Dudoit and van der Laan 2008, p. 316)) while correcting their bias toward false positives. With the principle of maximum entropy, the strategy results in two new estimators of the LFDR.

For example, the achieved NFDR estimate corresponding to $p(x_{(i)})$, the ith smallest p value, is multiplied by a simple sum to become $\mathrm{CFDR}(x_{(i)})$, the corrected FDR defined by

$$\mathrm{CFDR}(x_{(i)}) = \left(\sum_{k=1}^{i} \frac{1}{i-k+1}\right) \widehat{\mathrm{NFDR}}(p(x_{(i)})), \tag{1}$$

a generalization of which is derived in Sec. 3. Wide discrepancies between the new estimators and $\widehat{\mathrm{NFDR}}(p(x_{(i)}))$ are revealed in applications to a biomedical data set and a gene expression data set, illustrating the extent to which FDR methods require bias correction.

While it seems clear that FDR methods benefit from correcting their bias, the competitiveness of the resulting LFDR estimators compared to previous LFDR estimations is more debatable. Thus, Sec. 4 presents a simulation study comparing the performance of LFDR estimators based on NFDR bias correction to that of previous estimators. The CFDR performs well not only compared to NFDR estimation but also compared to previous methods of LFDR estimation, as is discussed in Sec. 5. The R packages behind the computations are listed in Sec. 6, which also cites interactive software intended to explain the need for bias correction.

2. False discovery rates

Let A_i stand for the random quantity that is equal to 1 if the ith feature is affected by a treatment, associated with a disease, etc. If the feature is unaffected (or unassociated),

then A_i is equal to 0. In general, $A_i = 1$ if the ith alternative hypothesis is true, but $A_i = 0$ if the ith null hypothesis is true.

The *local false discovery rate* (LFDR) is the posterior probability that null hypothesis i is true given the p value:

$$\text{LFDR}(p(x_i)) = P\big(A_i = 0 | p(X_i) = p(x_i)\big), \tag{2}$$

where X_i is a random variable of observed value x_i. In the gene expression case, the LFDR is a conditional probability of the hypothesis that gene i is not differentially expressed given the p value. Conversely, $1-\text{LFDR}(p(x_i))$ is $P(A_i = 1 | p(X_i) = p(x_i))$, the posterior probability of the hypothesis that feature i is differentially expressed.

Another "Bayesian false discovery rate" (Efron and Tibshirani 2002) is the *nonlocal false discovery rate* (NFDR) (Bickel 2013) at a significance level of α: $\text{NFDR}(\alpha) = P(A_i = 0 | p(X_i) \leq \alpha)$. The NFDR evaluated at significance level α is equal to the average LFDR over all genes with p values less than α. That is to say, $\text{NFDR}(\alpha) = P(A_i = 0 | p(X_i) \leq \alpha)$

$$= E\big(P\big(A_i = 0 | p(X_i)\big) | p(X_i) \leq \alpha\big) = E\big(\text{LFDR}(p(X_i)) | p(X_i) \leq \alpha\big), \tag{3}$$

by the law of total probability (Efron 2010).The NFDR may be estimated by this ratio of the estimated average number of false discoveries to the number of null hypotheses rejected at significance level α:

$$\widehat{\text{NFDR}}(\alpha) = \frac{\alpha d}{\#(p(x_i) \leq \alpha)} \text{ if } \frac{\alpha d}{\#(p(x_i) \leq \alpha)} \leq 1 \text{ else } \widehat{\text{NFDR}}(\alpha) = 1, \tag{4}$$

where the dimension d is number of tests performed.Many software packages substitute each p value for α in Eqs. (3) and (4), yielding the achieved FDR and its estimate for test j:

$$\begin{aligned} \text{NFDR}(p(x_j)) &= P\Big(A_i = 0 | p(X_i) \leq p(x_j)\Big) \\ &= E\Big(P\big(A_i = 0 | p(X_i)\big) | p(X_i) \leq p(x_j)\Big) = E\Big(\text{LFDR}(p(X_i)) | p(X_i) \leq p(x_j)\Big); \end{aligned} \tag{5}$$

$$\widehat{\text{NFDR}}(p(x_j)) = \widehat{\text{NFDR}}(\alpha) \text{ with } \alpha = p(x_j). \tag{6}$$

However, that can be misleadingly low, for Eq. (5) says the achieved FDR of the test corresponding to the p value $p(x_j)$ is the average posterior probability that the null hypothesis is true given that the p value is less than $p(x_j)$. That probability is often much less than $\text{LFDR}(p(x_j))$, the posterior probability that the null hypothesis is true given that the p value is equal to $p(x_j)$; in short, $\text{NFDR}(p(x_j)) \ll \text{LFDR}(p(x_j))$.

As a result, to the extent that $\widehat{\text{NFDR}}(p(X_j))$ is an unbiased estimate of $\text{NFDR}(p(X_j))$, its negative bias tends to be excessive as an estimator of $\text{LFDR}(p(X_j))$, the quantity most relevant to deciding whether to reject the jth null hypothesis given the data. Under the assumption that $\widehat{\text{NFDR}}(p(X_j))$ and $\widehat{\text{LFDR}}(p(X_j))$ are unbiased estimators of $\text{NFDR}(p(X_j))$ and $\text{LFDR}(p(X_j))$, respectively, the bias in $\widehat{\text{NFDR}}(p(X_j))$ as an estimate of $\text{LFDR}(p(X_j))$ is $E(\widehat{\text{NFDR}}(p(X_j)) - \text{LFDR}(p(X_j))) = E(\widehat{\text{NFDR}}(p(X_j)) - \text{NFDR}(p(X_j))) + E(\text{NFDR}(p(X_j)) - \text{LFDR}(p(X_j))) = 0 + E(\text{NFDR}(p(X_j)) - \text{LFDR}(p(X_j))) \ll 0$.

The practical consequence of this negative bias in the case of gene expression data is that more genes will be considered differentially expressed on the basis of a low $\mathrm{NFDR}(p(x_j))$ than is warranted since the probability that the null hypothesis is true tends to be higher than $\mathrm{NFDR}(p(x_j))$.

That bias toward false positives occurs not only for estimates of achieved NFDRs but also for various procedures that control the FDR (§1). Can the bias can be corrected in the same way as a systematically erring instrument can be calibrated?

3. Salvaging false discovery rates

3.1. Inferring local false discovery rates from a false discovery rate

Equation (3) constrains the conditional expectation value of the LFDR to be the NFDR: $E(\mathrm{LFDR}(p(X_i))|p(X_i) \leq \alpha) = \mathrm{NFDR}(\alpha)$. With the additional constraint that $\mathrm{LFDR}(p(X_i)) \geq 0$, the maximum entropy principle yields $\mathrm{Expon}((\mathrm{NFDR}(\alpha))^{-1})$, the $(\mathrm{NFDR}(\alpha))^{-1}$-rate exponential distribution of the LFDR given $p(X_i) \leq \alpha$:

$$\mathrm{LFDR}(p(X_i))|p(X_i) \leq \alpha \sim f_\alpha^\star(\bullet) := \arg\max_{f(\bullet)} \left(-\int_0^\infty f(L) \ln f(L) dL \right) = \frac{e^{-\frac{\bullet}{\mathrm{NFDR}(\alpha)}}}{\mathrm{NFDR}(\alpha)}, \quad (7)$$

where L is a dummy variable for the LFDR, and each $f(\bullet)$ is a probability density function satisfying $\int_0^\infty f(L)dL = 1$ and $\int_0^\infty Lf(L)dL = \mathrm{NFDR}(\alpha)$. Thus, if $p(x_i) \leq \alpha$, then the conditional probability density of $\mathrm{LFDR}(p(x_i))$ given $p(X_i) \leq \alpha$ is $f_\alpha^\star(\mathrm{LFDR}(p(x_i))) = e^{-\frac{\mathrm{LFDR}(p(x_i))}{\mathrm{NFDR}(\alpha)}}/\mathrm{NFDR}(\alpha)$.

Under the additional constraint that $\mathrm{LFDR}(p(X_i)) \leq 1$, the distribution that maximizes the entropy becomes more complicated (Conrad 2014). Since that distribution is very closely approximated by Eq. (7) if $\alpha < 1/2$, it does not provide a practical benefit for tests at any significance level much less than 0.5.

Consider $\mathrm{LFDR}_{(1)}, \ldots, \mathrm{LFDR}_{(m)}$, the order statistics of the m independent draws $\mathrm{LFDR}(p(X_1))$, ..., $\mathrm{LFDR}(p(X_m))$ from $\mathrm{Expon}((\mathrm{NFDR}(\alpha))^{-1})$. The ith expected order statistic is known for the exponential distribution (e.g. Cox and Hinkley 1979) to be

$$E\left(\mathrm{LFDR}_{(i)}|p(X_1), \ldots, p(X_m) \leq \alpha\right) = \left(\sum_{k=1}^{m} \frac{1}{m-k+1} \right) \mathrm{NFDR}(\alpha). \quad (8)$$

Let $m(\alpha)$ denote the number of hypotheses rejected at level α:

$$m(\alpha) = \left| \left\{ p(x_i) : i = 1, \ldots, d; p(x_i) \leq \alpha \right\} \right|, \quad (9)$$

assuming there are no ties. With $p(x_{(1)}), \ldots, p(x_{(m(\alpha))})$ as the $m(\alpha)$ observed order statistics of the p values corresponding to rejected null hypotheses, Eq. (8) suggests estimating $\mathrm{LFDR}(p(x_{(i)}))$, the LFDR corresponding to ith-highest such p value, by

$$\widehat{\mathrm{LFDR}}^\star(x_{(i)}; q) := \left(\sum_{k=1}^{i} \frac{1}{m-k+1} \right) q \quad (10)$$

for $i = 1, \ldots, m(\alpha)$, where, for now, $q = \widehat{\mathrm{NFDR}}(\alpha)$ and $m = m(\alpha)$. The $\star$ indicates dependence on the maximum entropy solution (7).

As FDR and NFDR methods analyze data much more similarly to each other than to LFDR estimators, the proposed method is not limited to the NFDR. Rather than setting $q = \widehat{\text{NFDR}}(\alpha)$, the q in Eq. (10) may instead be a level at which the FDR is controlled according to a method that determines which null hypotheses to reject, as in Sec. 1. The number of null hypotheses rejected by the method is the m in Eq. (10). In some cases, the level of FDR control is identical to an achieved NFDR estimate (§1).

Example 1. Of the $d = 15$ hypotheses on real-data thrombolytic-treatment outcomes tested by Neuhaus et al. (1992), 4 are rejected when controlling the FDR at the 0.05 level (Benjamini and Hochberg 1995). However, their expected order statistics of the LFDR according to $f^\star_{0.05}$ are $\widehat{\text{LFDR}}^\star(x_{(1)}; 0.05) = 0.012$, $\widehat{\text{LFDR}}^\star(x_{(2)}; 0.05) = 0.029$, $\widehat{\text{LFDR}}^\star(x_{(3)}; 0.05) = 0.054$, and $\widehat{\text{LFDR}}^\star(x_{(4)}; 0.05) = 0.104$, only 2 of which are below 0.05. That the hypotheses with estimated LFDRs of 0.054 and 0.104 would not be rejected merely because the mean of 0.012, 0.029, 0.054, and 0.104 is 0.05 highlights the problem identified above (§1, §2). ▲

3.2. Correcting and re-ranking false discovery rates

Let $q(\alpha)$ denote the smallest value of q such that all hypothesis with p values in $[0, \alpha]$ are rejected according to some procedure that guarantees that the FDR, NFDR, or an estimate of either is no higher than q. The substitutions $q = q(\alpha)$ and $m = m(\alpha)$ (9) turn Eq. (10) into

$$\widehat{\text{LFDR}}^\star\left(x_{(i)}; q(\alpha), \alpha\right) := \left(\sum_{k=1}^{m(\alpha)} \frac{1}{m(\alpha) - k + 1}\right) q(\alpha). \tag{11}$$

As explained in Sections 1 and 2, $q(p(x_{(i)}))$ would be overly optimistic as an estimator of $\text{LFDR}(x_{(i)})$. It does not follow that $q(p(x_{(i)}))$ is useless, for its bias may be corrected by Eq. (11):

$$\begin{aligned}\text{CFDR}(x_{(i)}) := \widehat{\text{LFDR}}^\star(x_{(i)}; q(p(x_{(i)})), p(x_{(i)})) &= \left(\sum_{k=1}^{m(p(x_{(i)}))} \frac{1}{m(p(x_{(i)})) - k + 1}\right) q(p(x_{(i)})) \\ &= \left(\sum_{k=1}^{i} \frac{1}{i - k + 1}\right) q(p(x_{(i)}))\end{aligned} \tag{12}$$

since $m(p(x_{(i)})) = i$ according to Eq. (9) as, under the assumption that there are no ties among the p values, $p(x_{(i)})$ is the p value of rank i.

$\widehat{\text{LFDR}}^\star(x_{(i)}; q(p(x_{(i)})), p(x_{(i)}))$ is called the *corrected false discovery rate* (CFDR) of the ith smallest p value and is abbreviated by $\text{CFDR}(x_{(i)})$. The special case with $q(p(x_{(i)})) = \widehat{\text{NFDR}}(p(x_{(i)}))$ appears in Eq. (1). Like other false discovery rate estimates (Efron 2010), CFDR values greater than 1 are reset to 1.

A different estimator of the LFDR arises from the simplifying assumption that the p values and LFDRs are similarly ordered in the sense that they have the same ranks. Let

$F^\star_\alpha$ denote the cumulative distribution function of the maximum-entropy LFDR: $F^\star_\alpha(L^\star) = \int_0^{L^\star} f^\star_\alpha(L)dL$ for $0 \le L^\star \le 1$. Then, for any α, the quantile rank of the NFDR is $F^\star_\alpha(\text{NFDR}(\alpha)) = 1-e^{-1}$ since $F^\star_\alpha$ is exponential with rate $(\text{NFDR}(\alpha))^{-1}$. As $1-e^{-1}$ is constant in α, $F^\star_\alpha(\text{NFDR}(\alpha))$ will be abbreviated by $F^\star(\text{NFDR})$.

Let $[i/F^\star(\text{NFDR})]$ denote the closest integer to $i/F^\star(\text{NFDR})$, with $i = 1, 2, \ldots$ small enough that $[i/F^\star(\text{NFDR})] \le d$. For large i, the expected value of a single LFDR drawn from $F^\star_{p(x_{([i/F^\star(\text{NFDR})])})}$ is $\text{NFDR}(p(x_{([i/F^\star(\text{NFDR})])}))$, which, with high probability, is approximately $\text{LFDR}(p(x_{(i)}))$ since $\text{LFDR}_{(i)} = \text{LFDR}(p(x_{(i)}))$ by the assumed equality of ranks and since $\text{LFDR}_{(i)}$ is approximately equal to the $F^\star(\text{NFDR})$-quantile of $F^\star_{p(x_{([i/F^\star(\text{NFDR})])})}$ with high probability. The *re-ranked false discovery rate* (RFDR) is accordingly defined by

$$\text{RFDR}(x_{(i)}) = q\left(p\left(x_{\left(\left[i/F^\star(\text{NFDR})\right]\right)}\right)\right) \text{ if } \left[i/F^\star(\text{NFDR})\right] \le d \text{ else } \text{RFDR}(x_{(i)}) = 1. \quad (13)$$

This is identical to the LFDR estimator $\widehat{\psi}(x_{(i)})$ proposed by Bickel (2013, p. 537) except that $\widehat{\psi}(x_{(i)})$ uses 1/2 in place of $F^\star(\text{NFDR})$ (Bickel 2015). Thus, it is called the *double-ranked false discovery rate* (DFDR). The fact that $F^\star(\text{NFDR})>1/2$ means $\widehat{\psi}(x_{(i)})$ is based on a p value of higher rank than that of $\text{RFDR}(x_{(i)})$, suggesting that the DFDR has a positive bias. To avoid possible confusion, it is emphasized that the CFDR, RFDR, and DFDR are not quantities to be estimated but are estimates of the LFDR.

Example 2. The p values from Example 1 yield the achieved NFDR estimates (6) and the corresponding LFDR estimates (12)–(13) displayed in the left panel of Figure 1. Using a significance threshold of 0.05 results in the rejection of 4 hypotheses $(\widehat{\text{NFDR}}(p(x_{(i)})) \le 0.05)$, 3 hypotheses $(\text{CFDR}(x_{(i)}) \le 0.05)$, or 2 hypotheses $(\text{RFDR}(x_{(i)}) \le 0.05)$. ▲

Example 3. Using microarray technology and $n = 6$ biological replicates, Alba et al. (2005) recorded microarray measurements of the expression levels of 13,440 genes in tomatoes at three days after the breaker stage of ripening. Following Bickel (2012), only the $d = 6103$ genes with data available for all 6 replicates are used for multiple applications of the paired t-test, with 6 (mutant, wild type) pairs for each gene.

The right panel of Figure 1 reveals that $\widehat{\text{NFDR}}(p(x_{(i)}))$ is much less than $\text{CFDR}(x_{(i)})$ and $\text{RFDR}(x_{(i)})$ for the 545 genes with $\widehat{\text{NFDR}}(p(x_{(i)})) \le 0.2$. By comparison, only 69 or 344 of which satisfy $\text{CFDR}(x_{(i)}) \le 0.2$ or $\text{RFDR}(x_{(i)}) \le 0.2$, respectively. Lowering the threshold defining what is meant by a "discovery" has no effect on the relative magnitudes of the discrepancies (Figure 2). This marked discrepancy between LFDR and NFDR estimators corroborates previous findings for this data set (Bickel, 2012, Figure 3) and for other gene expression data (Hong, Tibshirani, and Chu 2009). Since the number of discoveries in Figure 2 is the sum of the number of true discoveries and the number of false discoveries, higher numbers of discoveries tend to coincide with higher numbers of false discoveries. In this example, $q(p(x_{(i)})) = \widehat{\text{NFDR}}(p(x_{(i)}))$. ▲

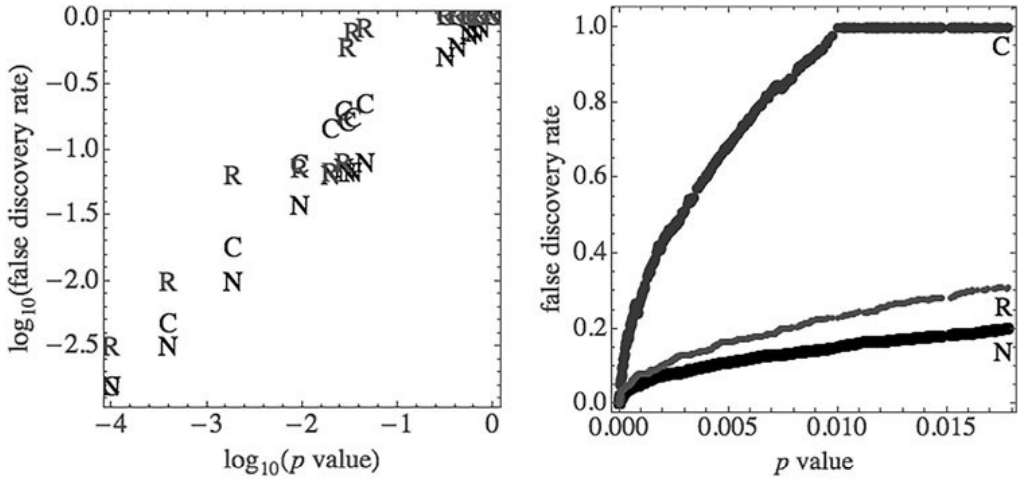

Figure 1. Local ("C"; "R") and nonlocal ("N") false discovery rate estimates as a function of $p(x_{(i)})$ for biomedical data (left) and gene expression data (right). "N" $= \widehat{\text{NFDR}}(p(x_{(i)}))$; "C" $= \text{CFDR}(x_{(i)})$; "R" $= \text{RFDR}(x_{(i)})$.

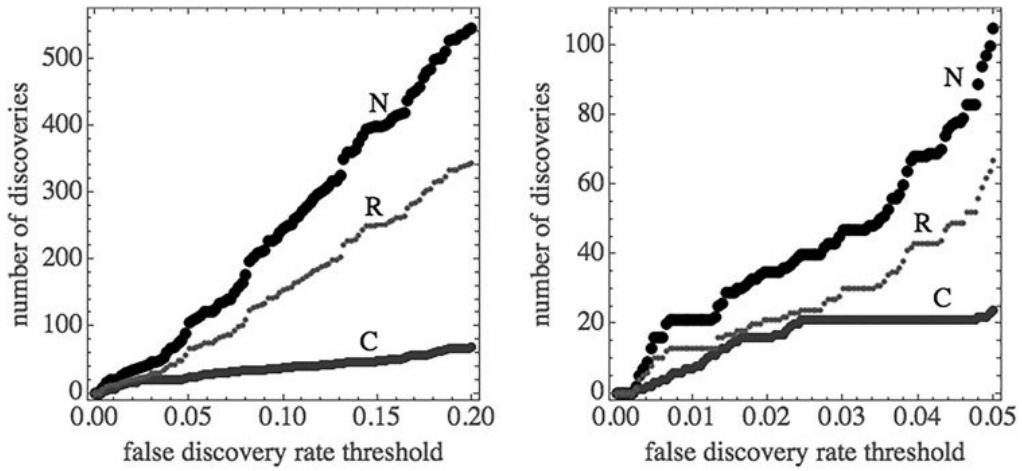

Figure 2. Number of genes discovered to be differentially expressed as a function of the threshold of the local ("C"; "R") and nonlocal ("N") false discovery rate estimates. The right plot magnifies the lower-left sixteenth of the left plot.

4. Performance of the proposed estimators of the local false discovery rate

This section presents a simulation study comparing the performance of the proposed estimators to that of previous estimators of the LFDR. In Sec. 4.1, we introduce a brief description of the previous estimators considered. In Sec. 4.2, we explain how we simulate the data and how we compute the root mean square error for each LFDR estimator. In Sec. 4.3, we compare the simulation results.

4.1. Previous local false discovery rate estimators

The previous methods of estimating the LFDR are explained in this subsection in terms of f_0 and f_1, the null-hypothesis and alternative-hypothesis probability density functions

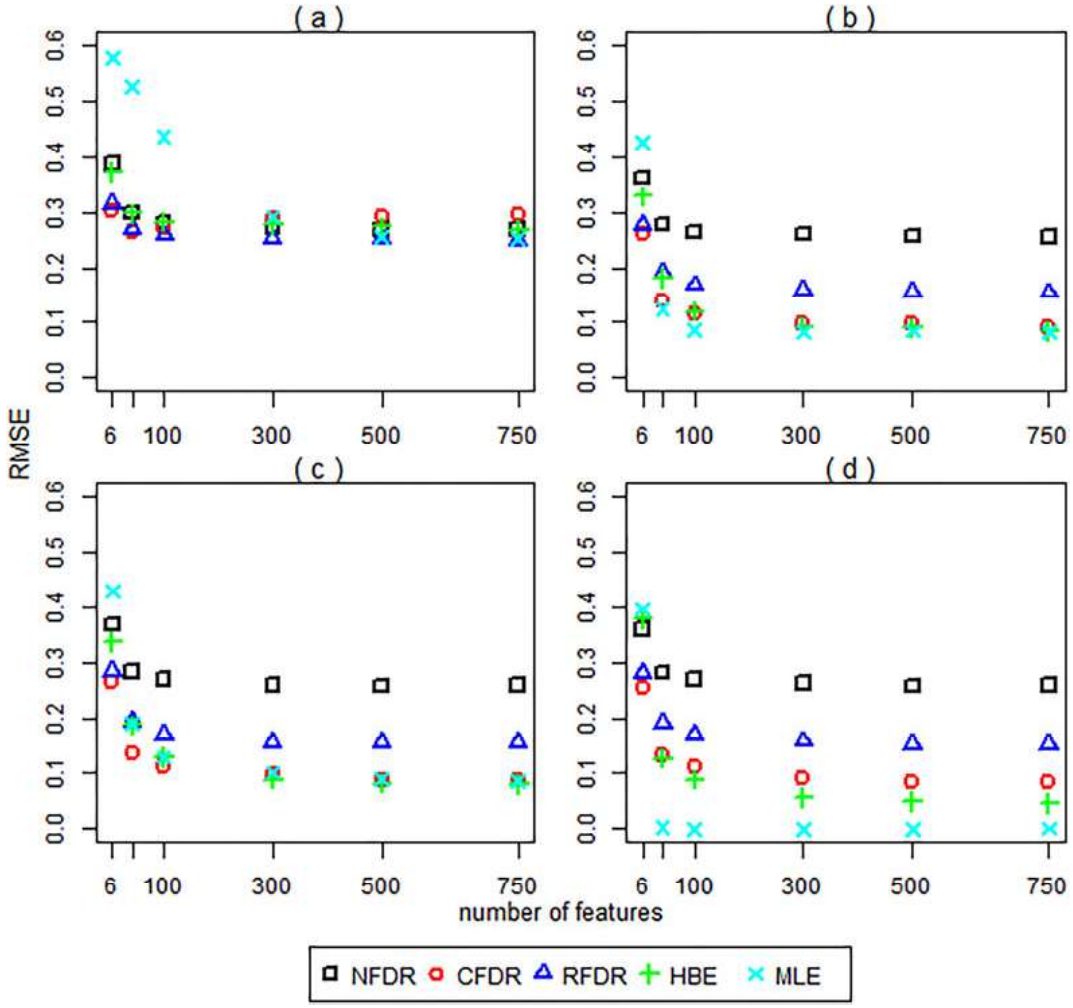

Figure 3. RMSE of $\widehat{\widehat{A}}$ as a function of d, the number of features, based on different values of the parameter of interest under the alternative hypothesis θ_{alt} and the sample sizes n and m. $\theta_{\text{alt}} = 1.5$ in the top two panels ((a) and (b)), and $\theta_{\text{alt}} = 4$ in the bottom two panels ((c) and (d)); $n = m = 5$ in the left two panels ((a) and (c)), and $n = m = 20$ in the right two panels ((b) and (d)).

of a p value $p(X_i)$ or other statistic $u(X_i)$ that summarizes the data. For example, if the gene expression data vector x_i for the ith gene may be summarized by a scalar $u(x_i)$, then the probability density would be $f_0(u(x_i))$ if that gene is equivalently expressed but would be $f_1(u(x_i))$ if it is differentially expressed. Let $\pi_0 = P(A_i = 0)$ denote the prior probability that the ith null hypothesis is true. By the definition of the LFDR as the posterior probability that the ith null hypothesis is true and by Bayes's theorem, we have

$$\text{LFDR}(u(x_i)) = \frac{\pi_0 f_0(u(x_i))}{\pi_0 f_0(u(x_i)) + (1-\pi_0) f_1(u(x_i))}, \tag{14}$$

which is consistent with the previous notation when $u(x_i) = p(x_i)$.

4.1.1. Histogram-based estimator of the local false discovery rate

The histogram-based estimator (HBE) of the local false discovery rate was introduced by Efron (2004). If the null hypothesis H_i is true, then $p(X_i)$ is uniformly distributed

between 0 and 1 for $i = 1, \ldots, d$. Both π_0 and the probability density of $p(X_i)$ can be estimated by using the natural spline fit to the histogram counts by Poisson regression (Efron 2004). Using the former estimate as π_0 and the latter estimate as the denominator of Eq. (14) with $u(x_i) = p(x_i)$ yields the HBE.

4.1.2. Maximum likelihood estimator of the local false discovery rate

In this subsection, we explain a simple method following the maximum likelihood strategy of estimating the LFDR. This strategy is commonly used; for example, Muralidharan (2010) estimated the LFDR by maximizing the likelihood assuming exponential families.

Let $f(u(x_i); \delta_{\text{alt}})$ denote the probability density of $u(X_i)$ given the value δ_{alt} of a noncentrality parameter such that $f(u(x_i); 0) = f_0(u(x_i))$. The true LFDR is written as Eq. (14) with $f_1(u(x_i)) = f(u(x_i); \delta_{\text{alt}})$. A maximum likelihood estimator (MLE) of the LFDR estimates the unknown parameters π_0 and δ_{alt} by $\widehat{\delta}_{\text{alt}}$ and $\widehat{\pi}_0$, the values that maximize the likelihood

$$\prod_{i=1}^{d} \left(\pi_0 f_0(u(x_i)) + (1 - \pi_0) f_{\delta_{\text{alt}}}(u(x_i))\right)$$

when allowing $0 \leq \pi_0 \leq 1$ and $\delta_{\text{alt}} > 0$. Substituting those estimates into Equation (14) yields

$$\frac{\widehat{\pi}_0 f_0(u(x_i))}{\widehat{\pi}_0 f_0(u(x_i)) + (1 - \widehat{\pi}_0) f_{\widehat{\delta}_{\text{alt}}}(u(x_i))},$$

which is called the *MLE of the LFDR* (Padilla and Bickel 2012; Bickel 2013; Yang, Aghababazadeh, and Bickel 2013).

4.2. Method of simulation and performance quantification

This section compares several LFDR estimators by simulating the log-abundance of proteins. To do this comparison, we build treatment and control groups. We call X the data matrix of the treatment group, and Y that of control group. Let n and m denote the sample sizes of the treatment and control groups, respectively. The dimension of X is $d \times n$ and that of Y is $d \times m$, and the vectors of the ith feature of X and Y are $X_i = (X_{i,1}, \ldots, X_{i,j}, .., X_{i,n}) \in \mathbb{R}^n$, $i = 1, \ldots, d$ and $j = 1, \ldots, n$, and $Y_i = (Y_{i,1}, \ldots, Y_{i,j}, .., Y_{i,m}) \in \mathbb{R}^m$, $i = 1, \ldots, d$ and $j = 1, \ldots, m$. We use values of d equal to 6, 42, 100, 300, 500 and 750, and two cases of sample sizes: $n = m = 5$ and $n = m = 20$.

The probability that the ith feature is unaffected by the treatment is $\pi_0 = 0.9$. In Sec. 2, we defined the random variables A_i, $i = 1, \ldots, d$. We now add an assumption that the A_i are mutually independent.

Denote by $\overline{A_i}$ the null-hypothesis indicator variable defined by $\overline{A_i} = 1 - A_i$:

$$\overline{A_i} = \begin{cases} 0 & \text{if the } i\text{th feature is affected by the disease} \\ 1 & \text{if the } i\text{th feature is unaffected by the disease.} \end{cases}$$

We draw $\overline{A_i}$ from Bernoulli(π_0), the Bernoulli distribution with parameter π_0. If $\overline{A_i} = 1$, we draw X_i from $N(0,1)$, the standard normal distribution, but if $\overline{A_i} = 0$, we draw X_i from $N(\mu,1)$, where the mean μ is 1.5 or 4, with the choice of these two values following Padilla and Bickel (2015). Regardless of the value of $\overline{A_i}$, we draw Y_i from $N(0,1)$. Our parameter of interest for the feature i is θ_i, the difference in the logarithms of protein abundance between the sample mean of treatment and control groups. Under null hypothesis $\overline{A_i} = 1$, the parameter θ_i is equal to zero ($\theta_i = 0$), and under alternative hypothesis $\overline{A_i} = 0$, the parameter θ_i is equal to μ ($\theta_i = \mu$). Algorithm 1 is used to this end.

The p values for the NFDR, CFDR, and RFDR come from the two-sided, equal-variance t test. We also compute two previous estimators, HBE and MLE, explained in Sections 4.1.1 and 4.1.2. For the HBE software (Efron et al. 2015), we use the standard normal quantile of the one-sided p value as the z-score, we set the number of histogram breaks in the discretization of the z-score equal to the minimum of 120 and $1.5d$, and we set the degrees of freedom equal to 7 for fitting the mixture estimated density. For the MLE, $u(x_i) = |t(x_i)|$, where the Student t statistic $t(x_i)$ is modeled a realization of $t(X_i)$, the statistic of the t distribution f_0 under the null hypothesis with $\nu = n + m - 2$ degrees of freedom or the noncentral t distribution $f(\bullet; \delta_{\text{alt}})$ under the alternative hypothesis with ν degrees of freedom and noncentrality parameter $\delta_{\text{alt}} > 0$ (Bickel 2014).

We assess the performance of both the NFDR and of the LFDR estimators as estimators of the null-hypothesis indicator $\overline{A_i}$, following Yang, Aghababazadeh, and Bickel (2013) and Bickel (2014). For B data sets (in our simulation we set $B = 500$), we compute for each feature the root mean square error (RMSE) between each LFDR estimator and the indicator variable $\overline{A_i}$, and then we compute the average over all the features. Algorithm 2 explains how we compute the RMSE for each of the LFDR estimators. We set $q(p(x_{(i)})) = \widehat{\text{NFDR}}(p(x_{(i)}))$ for the simulations.

Algorithm 1. Algorithm to simulate a data set.

The input parameters are d, π_0, n, m, μ.
For each feature $i = 1, 2, \ldots, d$, draw $\overline{A_i}$ from Bernoulli(π_0).

- If $\overline{A_i} = 1$ then draw a vector of length n, $x_i = [x_{i,1}, \ldots, x_{i,j}, \ldots, x_{i,n}]$ from $N(0, 1)$ and a vector of length m, $y_i = [y_{i,1}, \ldots, y_{i,j}, \ldots, y_{i,m}]$ from $N(0, 1)$.
- If $\overline{A_i} = 0$ then draw a vector of length n, $x_i = [x_{i,1}, \ldots, x_{i,j}, \ldots, x_{i,n}]$ from $N(\mu, 1)$ and a vector of length m, $y_i = [y_{i,1}, \ldots, y_{i,j}, \ldots, y_{i,m}]$ from $N(0, 1)$.
- Thus, we get the vectors $x_1, \ldots, x_d$ and $y_1, \ldots, y_d$, and a vector $\overline{A} = [\overline{A_1}, \ldots, \overline{A_d}]$.

Algorithm 2. Algorithm to compute RMSE.

Notation: $\widehat{\overline{A}}([x_1, \ldots, x_i, \ldots, x_d], [y_1, \ldots, y_i, \ldots, y_d])$ is a function to compute the estimator of LFDR or NFDR for the data set $([x_1, \ldots, x_i, \ldots, x_d], [y_1, \ldots, y_i, \ldots, y_d])$. $\widehat{\overline{A}}$ can be CFDR, RFDR, HBE, MLE or NFDR.
The inputs are: The parameters d, π_0, n, m, μ, $B$and the function $\widehat{\overline{A}}$.
Step 1. Repeat Algorithm 1 B times to get B data sets.

Step 2. For each data set $([x_1^b, \ldots, x_i^b, \ldots, x_d^b], [y_1^b, \ldots, y_i^b, \ldots, y_d^b])$, $b = 1, 2, \ldots, B$, compute the estimator $\widehat{\overline{A^b}} = \widehat{\bar{A}}([x_1^b, \ldots, x_i^b, \ldots, x_d^b], [y_1^b, \ldots, y_i^b, \ldots, y_d^b]) \in [0, 1]^{\mathrm{d}}$, $\widehat{\overline{A^b}}$ is a vector of d elements corresponding to d features $\widehat{\overline{A^b}} = [\widehat{\overline{A_1^b}}, \ldots, \widehat{\overline{A_i^b}}, \ldots, \widehat{\overline{A_d^b}}]$.

Step 3. Compute RMSE for each estimator over all B data sets,

$$\mathrm{RMSE}\left(\widehat{\bar{A}}\right) = \sqrt{\frac{1}{d}\frac{1}{B}\sum_{i=1}^{d}\sum_{b=1}^{B}\left(\overline{A_i^b} - \widehat{\overline{A_i^b}}\right)^2}.$$

Thus, we get: RMSE(CFDR), RMSE(RFDR), RMSE(HBE), RMSE(MLE) and RMSE (NFDR).

4.3. Results of the simulation study

Figure 3 presents the RMSE for different cases of θ_{alt} and for the sample sizes of the treatment and control groups. Panel (a), corresponding to $\theta_{\mathrm{alt}} = 1.5$ and $n = m = 5$, shows that the CFDR performs better than other estimators for a small number of features $(6 \leq d \leq 42)$, and the MLE looks the worst for $d \leq 200$. The RFDR performs better than other estimators for $d \geqslant 60$. When $d \geqslant 750$, the HBE, MLE, and RFDR performed better than the CFDR.

For the same $\theta_{\mathrm{alt}} = 1.5$ but for a sample size $n = m = 20$, Panel (b) of Figure 3 shows that the CFDR for small and large features performed better than the RFDR and NFDR. For small number of features $(d \approx 6)$, the CFDR has lower values of RMSE, whereas MLE becomes better for $d \geq 42$. When $d \geq 750$, the MLE, CFDR and HBE have close values of RMSE.

Panel (c) corresponds to $\theta_{\mathrm{alt}} = 4$ and $n = m = 5$. It is clear that the CFDR performs better for both small and large features, while for $d \geq 300$, MLE and HBE start to performs like CFDR despite MLE being the worst for $d = 6$. However, NFDR appears as the worst for $d \geq 18$.

Panel (d) of Figure 3, corresponding to $\theta_{\mathrm{alt}} = 4$ and $n = m = 20$, shows that the RMSE of the CFDR is small for $d \leq 18$. However, for $d \geq 24$, the RMSE of the MLE becomes smaller. This is due to large sample size of treatment and control groups $n = m = 20$ and the large number of features. We can also see that the HBE, designed for large-scale inference, is worst for small numbers of features $(d \leq 12)$, while the NFDR becomes the worst for $d \geq 18$.

The anonymous referee pointed out that the estimator that has a value of 1 regardless of the data has an RMSE of $\sqrt{1/10} \approx 0.32$, which is only slightly worse than the RMSE of the NFDR estimate and of the LFDR estimators displayed in panel (a).

5. Discussion

This paper proposes the CFDR and the RFDR, two bias-correction methods that transform the output of FDR software packages to LFDR estimates. Their advantage over uncorrected FDR methods is the reduction in bias toward false positives. An advantage of the CFDR and RFDR over previous LFDR methods is its applicability to the output

of popular FDR software. Thus, they provide simple ways to improve the interpretation of the output of readily available FDR software packages.

Unfortunately, the reduction in the bias compared to the uncorrected FDR methods may mean some increase in variance. This may be tolerated since it purchases the much-needed reduction of an anti-conservative bias, a bias resulting in reporting more false positives than the nominal thresholds suggest. Equation (1) indicates that the CFDR is most advantageous compared to the NFDR for genes with larger p values and that the CFDR is most similar to the NFDR for the genes that have the lowest p values.

On the other hand, the potentially lower variance of the CFDR and RFDR compared to previous LFDR estimators suggests that the bias of the CFDR and RFDR is higher. That is why the overall performance of the CFDR and RFDR compared to that of previous LFDR estimators was assessed in the simulation study in Sec. 4. The CFDR performs well for small numbers of features. While the MLE performs better than the other estimators of LFDR for large numbers of features, that is because it leans heavily on the parametric model used to generate the data. The CFDR and HBE perform well for larger numbers of features with less dependence on that model.

6. Software

The methods introduced in this paper are provided in the R package *CorrectedFDR* (Rahal et al. 2018). Other R packages used in our computations include *locfdr* (Efron et al. 2015) *and LFDR.MLE* (Padilla et al. 2015). The simulated data were generated by the R functions in the *SimLFDR* suite (Rahal, 2019). Interactive software to illustrate the need for bias correction is also available (Bickel, 2019b); see Bickel (2019a) for a complementary exercise.

Acknowledgments

We thank the anonymous referee for many comments that led to improvements in clarity.

Funding

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009), by Agriculture and Agri-Food Canada (ABIP 000159B), and by the Faculty of Medicine of the University of Ottawa.

ORCID

David R. Bickel http://orcid.org/0000-0002-9623-3946

References

Alba, R., P. Payton, Z. Fei, R. McQuinn, P. Debbie, G. B. Martin, S. D. Tanksley, and J. J. Giovannoni. 2005. Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell* 17 (11):2954–65. doi:10.1105/tpc.105.036053.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57:289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Benjamini, Y., and W. Liu. 1999. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 82 (1–2):163–70. doi:10.1016/S0378-3758(99)00040-3.

Bickel, D. R. 2012. Game-theoretic probability combination with applications to resolving conflicts between statistical methods. *International Journal of Approximate Reasoning* 53 (6): 880–91. doi:10.1016/j.ijar.2012.04.002.

Bickel, D. R. 2013. Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. *Statistical Applications in Genetics and Molecular Biology* 12 (4):529–43. doi:10.1515/sagmb-2013-0003.

Bickel, D. R. 2014. Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison. *International Statistical Review* 82 (3):457–76. doi:10.1111/insr.12064.

Bickel, D. R. 2015. Corrigendum to: Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. *Statistical Applications in Genetics and Molecular Biology* 14 (2):225. doi:10.1515/sagmb-2014-0100.

Bickel, D. R. 2019a. *Genomics data analysis: False discovery rates and empirical Bayes methods*. London: Chapman and Hall/CRC.

Bickel, D. R. 2019b. Interactive comparison of false discovery rates and local false discovery rates. *Open Access Software*. doi:10.5281/zenodo.3234944.

Conrad, K. 2014. Probability distributions and maximum entropy. Technical Report, University of Connecticut. Accessed October 19, 2015. http://www.math.uconn.edu/kconrad/blurbs/analysis/entropypost.pdf.

Cox, D., and D. Hinkley. 1979. *Theoretical statistics*. Boca Raton, FL: Taylor & Francis.

Dudoit, S., and M. J. van der Laan. 2008. *Multiple testing procedures with applications to genomics*. New York: Springer.

Efron, B. 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* 99 (465):96–104. doi:10.1198/016214504000000089.

Efron, B. 2010. *Large-Scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge: Cambridge University Press.

Efron, B., and R. Tibshirani. 2002. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* 23 (1):70–86. doi:10.1002/gepi.1124.

Efron, B., B. Turnbull, B. Narasimhan, and K. Strimmer. 2015. Locfdr: Computes local false discovery rates. R package version 1, 1–8. https://goo.gl/1e284V.

Genovese, C., and L. Wasserman. 2002. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3):499–517. doi:10.1111/1467-9868.00347.

Glickman, M., S. Rao, and M. Schultz. 2014. False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology* 67 (8):850–7. doi:10.1016/j.jclinepi.2014.03.012.

Hong, W.-J., R. Tibshirani, and G. Chu. 2009. Local false discovery rate facilitates comparison of different microarray experiments. *Nucleic Acids Research* 37 (22):7483–97. doi:10.1093/nar/gkp813.

Muralidharan, O. 2010. An empirical Bayes mixture method for effect size and false discovery rate estimation. The *Annals of Applied Statistics* 4 (1):422–38. doi:10.1214/09-AOAS276.

Neuhaus, K.-L., R. Von Essen, U. Tebbe, A. Vogt, M. Roth, M. Riess, W. Niederer, F. Forycki, A. Wirtzfeld, W. Maeurer, et al. 1992. Improved thrombolysis in acute myocardial infarction with front-loaded administration of alteplase: Results of the rt-PA-APSAC patency study (TAPS). *Journal of the American College of Cardiology* 19 (5):885–91. doi:10.1016/0735-1097(92)90265-O.

Padilla, M., and D. R. Bickel. 2012. Estimators of the local false discovery rate designed for small numbers of tests. *Statistical Applications in Genetics and Molecular Biology* 11 (5):4.

Padilla, M., Bickel, D. R., 2015. Empirical Bayes and fiducial effect-size estimation for small numbers of tests. Working Paper, University of Ottawa, deposited in uO Research at: http://hdl.handle.net/10393/32151.

Padilla, M., Y. Yang, A. Ali, K. Leckett, Z. Yang, Z. Li, C. M. Yanofsky, and D. R. Bickel. 2015. LFDR.MLE: Estimation of the Local False Discovery Rates by type II Maximum Likelihood Estimation. R package version 1.0. https://goo.gl/d9MTNE.

Rahal, A. 2019. SimLFDR: A suite of R functions to simulate data in order to compare the performance of LFDR estimators. http://doi.org/10.5281/zenodo.2695970

Rahal, A., A. Akpawu, J. Chitpin, and D. R. Bickel. 2018. CorrectedFDR: Correcting false discovery rates. R package version 1.0. https://goo.gl/qMQsXa.

Van Den Oord, E. 2008. Review article: Controlling false discoveries in genetic studies. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 147 (5):637–44. doi:10.1002/ajmg.b.30650.

Yang, Y., F. A. Aghababazadeh, and D. R. Bickel. 2013. Parametric estimation of the local false discovery rate for identifying genetic associations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10:98–108.

Yang, Z., Z. Li, and D. R. Bickel. 2013. Empirical Bayes estimation of posterior probabilities of enrichment: A comparative study of five estimators of the local false discovery rate. *BMC Bioinformatics* 14 (1):87. doi:10.1186/1471-2105-14-87.

Chapter 4

Model fusion and multiple testing in the likelihood paradigm: shrinkage and evidence supporting a point null hypothesis

This chapter defines the theory of weight of statistical evidence with respect to a single model. It also defines the fusion model and the weight of evidence with respect to two fused models. For purposes of illustration and application of multiple hypothesis evidence, these theories were applied to exam scores data and to gene expression levels data. At the end, it presents a brief discussion of the practical advantages of the fused model weight of evidence for shrinkage and for evidence supporting a point null hypothesis.

STATISTICS
https://doi.org/10.1080/02331888.2019.1660342

Taylor & Francis
Taylor & Francis Group

Model fusion and multiple testing in the likelihood paradigm: shrinkage and evidence supporting a point null hypothesis

David R. Bickel [a,b] and Abbas Rahal[b]

[a]Department of Biochemistry, Microbiology, and Immunology, Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Canada; [b]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada

ABSTRACT

According to the general law of likelihood, the strength of statistical evidence for a hypothesis as opposed to its alternative is the ratio of their likelihoods, each maximized over the parameter of interest. Consider the problem of assessing the weight of evidence for each of several hypotheses. Under a realistic model with a free parameter for each alternative hypothesis, this leads to weighing evidence without any shrinkage toward a presumption of the truth of each null hypothesis. That lack of shrinkage can lead to many false positives in settings with large numbers of hypotheses. A related problem is that point hypotheses cannot have more support than their alternatives. Both problems may be solved by fusing the realistic model with a model of a more restricted parameter space for use with the general law of likelihood. Applying the proposed framework of model fusion to data sets from genomics and education yields intuitively reasonable weights of evidence.

ARTICLE HISTORY
Received 28 January 2018
Accepted 5 August 2019

KEYWORDS
Direct likelihood inference; likelihoodism; measure of evidence; multiple testing; pure likelihood methods

1. Introduction

The *ASA Statement on Statistical Significance and P-Values* [1] included the likelihood ratio as a measure of statistical evidence that might replace the p value:

> In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include ⋯ alternative measures of evidence, such as likelihood ratios or Bayes Factors ⋯

The *(special) law of likelihood* [2] uses the likelihood ratio $f_1(x)/f_2(x)$ as the measure of the strength of evidence in favour of a distribution f_1 over another distribution f_2 given x, the observed sample. The special law has been extended to the *general law of likelihood* [3, Section 2.2.3], which defines

$$W(\mathcal{H}_1;\mathcal{H}_2) = \frac{\sup_{\theta\in\mathcal{H}_1} f_\theta(x)}{\sup_{\theta\in\mathcal{H}_2} f_\theta(x)} \tag{1}$$

CONTACT David R. Bickel dbickel@uottawa.ca Department of Biochemistry, Microbiology, and Immunology, Ottawa Institute of Systems Biology, University of Ottawa, 451 Smyth Road, Ottawa, Ontario K1H 8M5, Canada Department of Mathematics and Statistics, University of Ottawa, 451 Smyth Road, Ottawa, Ontario K1H 8M5, Canada

as the strength or *weight of the evidence* in favour of the hypothesis that $\theta \in \mathcal{H}_1$ over the hypothesis that $\theta \in \mathcal{H}_2$ for the competing subsets $\mathcal{H}_1$ and $\mathcal{H}_2$ of the parameter space Θ [cf. 4], where $f_\theta(x)$ is the probability density of x given θ. (The dependence of $W(\mathcal{H}_1; \mathcal{H}_2)$ on x is suppressed to avoid complex notation.)

While the ASA Statement is relevant to multiple comparisons, the general law of likelihood does not in itself take them into account. Multiple testing is currently addressed by correcting p values for testing multiple null hypotheses.

Example 1.1: Consider the microarray data set that consists of the expression levels of 13,440 genes across $n = 6$ biological replicates in tomatoes at three days after the breaker stage of ripening [5]. As complete data are only available for $N = 6103$ of the genes, Bickel [6], Bickel and Rahal [7] used them for applications of the paired t-test of the null or alternative hypothesis that the ith gene is equivalently expressed or differentially expressed, respectively, between mutants and wild types.

Genomics p values like those are typically adjusted to control a family-wise error rate or a false discovery rate [e.g., 8]. However, since an adjusted p value reduces to an unadjusted p value in the degenerate case of a single null hypothesis, the criticisms that the *ASA Statement* [1] levels against the misuse of p values apply with equal force against misusing adjusted p values [see 9, Section 6.2].

In response to claims that the special law of likelihood rendered adjustments for multiple testing in problems like Example 1.1 unnecessary, Korn and Freidlin [10] supplied clinical examples to point out that although the likelihood ratio measures the strength of evidence in simple situations, it often can be misleading in multiple comparison situations unless supplemented with adjusted p values or posterior probabilities based on sufficiently strong prior distributions, perhaps estimated by empirical Bayes methods. In addressing the problem, Strug and Hodge [11] emphasized the role of frequentist considerations when planning an experiment, whereas Bickel [3] argued against comparing simple null hypotheses to composite alternative hypotheses. Nonetheless, exactly that comparison remains important in many scientific applications, especially those involving multiple comparisons (e.g., Example 1.1).The primary motivation of this paper is to offer a general framework for applying the laws of likelihood to problems involving multiple hypotheses by bringing together multiple statistical models. As with all likelihood-based inference [12, Section 6.5], successful application of the weight of evidence hinges on the selection of models at appropriate levels of abstraction, keeping in mind not only model realism but also operational characteristics. Against this background, different models may be brought together in such a way that the weight of evidence is reliable across a spectrum of applications broad enough to include those involving multiple comparisons.

Advantages of considering multiple models together motivate Bayesian, frequentist, and fiducial model averaging [e.g., 13–17]. This paper proposes a method of evidential model averaging called 'model fusion' as an alternative. Extending the laws of likelihood to model averaging may make the likelihood paradigm [2] less dependent on its frequentist and Bayesian parents.

Section 2 defines the theory of the weight of statistical evidence with respect to a single model. Section 3 defines model fusion and the weight of evidence with respect to two fused models. To address the multiple comparison problem, a special case of the theory of model

fusion is formulated in Section 4. Section 5 illustrates that case by applying it to data on exam scores and to data on gene expression and briefly discusses practical advantages of the fused-model weight of evidence for shrinkage and for evidence supporting a point null hypothesis. Proofs appear in Appendix 1.

2. Weight of evidence under a single model

The *marginal likeliness* that θ is in a set $\mathcal{H}$ of parameter values of a hypothesis is

$$L(\mathcal{H}) = W(\mathcal{H}; \Theta), \tag{2}$$

where $\mathcal{H}$ is a subset of the parameter space Θ. The *conditional likeliness* of the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$, restricting the parameter to values in a set $\mathcal{R}$, is

$$L(\mathcal{H}|\mathcal{R}) = \frac{L(\mathcal{H} \cap \mathcal{R})}{L(\mathcal{R})}. \tag{3}$$

The term 'likeliness' replaces the term 'extended likelihood' of Giang and Shenoy [18] for brevity and to avoid confusion with previous usages of the latter term in the statistics literature [19–21]. The *weight of evidence* for the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$ is

$$W(\mathcal{H}|\mathcal{R}) = W(\mathcal{H} \cap \mathcal{R}; \mathcal{R} \backslash \mathcal{H}) = \frac{L(\mathcal{H} \cap \mathcal{R})}{L(\mathcal{R} \backslash \mathcal{H})}.$$

These equations follow immediately from Equation (1) and the definitions:

$$L(\mathcal{H}) = \frac{\sup_{\theta \in \mathcal{H}} f_\theta(x)}{\sup_{\theta \in \Theta} f_\theta(x)};$$

$$L(\mathcal{H}|\mathcal{R}) = \frac{\sup_{\theta \in \mathcal{H} \cap \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{R}} f_\theta(x)}; \tag{4}$$

$$W(\mathcal{H}|\mathcal{R}) = \frac{\sup_{\theta \in \mathcal{H} \cap \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{R} \backslash \mathcal{H}} f_\theta(x)}. \tag{5}$$

The *unlikeliness* of the hypothesis that $\theta \in \mathcal{H}$ is $U(\mathcal{H}|\mathcal{R}) = L(\overline{\mathcal{H}}|\mathcal{R})$, where $\overline{\mathcal{H}} = \Theta \backslash \mathcal{H}$ is the complement of $\mathcal{H}$.

A generalization of the following result is known from both possibility theory and ranking theory. A set function Π is a *possibility measure* if $\Pi(\emptyset) = 0$, $\Pi(\Theta) = 1$, and $\Pi(\mathcal{H}) = \sup_{\theta \in \mathcal{H}} \Pi(\{\theta\})$ for any suitable $\mathcal{H} \subset \Theta$; a set function N is a *necessity measure* if $\mathrm{N}(\mathcal{H}) = 1 - \Pi(\overline{\mathcal{H}})$ for any suitable $\mathcal{H} \subset \Theta$ [22]. Thus, $L(\bullet|\mathcal{R})$ is a possibility measure, and $1 - U(\bullet|\mathcal{R})$ is a necessity measure. In terms of the theory of ranking functions [23, Section 5.2], $\log W(\bullet|\mathcal{R})$ is a two-sided ranking function, where the base of log is greater than 1.

Lemma 2.1: *If $\mathfrak{H}$ is a set of measurable subsets of Θ and $\mathcal{H}, \mathcal{R} \in \mathfrak{H}$ such that $L(\mathcal{R}) > 0$, then $L(\mathcal{H}|\mathcal{R}) = 1$ and $U(\mathcal{H}|\mathcal{R}) = 1/W(\mathcal{H}|\mathcal{R})$ if $W(\mathcal{H}|\mathcal{R}) \geq 1$ but $L(\mathcal{H}|\mathcal{R}) = W(\mathcal{H}|\mathcal{R})$ and $U(\mathcal{H}|\mathcal{R}) = 1$ if $W(\mathcal{H}|\mathcal{R}) < 1$.*

In addition, for a partition $\mathfrak{P}$ of the set of subsets of the parameter space Θ,

$$L(\mathcal{H}) = \sup_{\mathcal{R} \in \mathfrak{P}} L(\mathcal{R})\, L(\mathcal{H}|\mathcal{R}) \tag{6}$$

holds according to the theory of idempotent probability [24] since $L(\mathcal{H})$ and $L(\mathcal{H}|\mathcal{R})$ are marginal and conditional idempotent probabilities as well as possibility values from possibility theory. Bickel [25] shows how to derive it. (Recall that a partition of Θ is by definition a set of non-intersecting subsets of Θ such that their union is Θ.) For mathematically rigorous restrictions on $\mathcal{H}$ and $\mathcal{R}$, see Puhalskii [24].

Example 2.1: For the problem of Example 1.1, Equation (6) will be used to obtain the likeliness that the ith gene is differentially expressed by possibilistically marginalizing over the number of genes that are not differentially expressed:

$$\begin{aligned} &L\left(i\text{th gene is differentially expressed}\right) \\ &\quad = \max_{N_0 = 0,1,\ldots,6103} L\left(N_0 \text{ of the 6103 genes are not differentially expressed}\right) \\ &\qquad \times L\left(i\text{th gene is differentially expressed} \,|N_0 \text{ of the 6103 genes are not differentially expressed}\right). \end{aligned}$$

Sections 3 and 4 add technical concepts for stating that more precisely.

3. Fusing models from different levels

Likelihood-based frequentist inference may involve eliminating nuisance parameters by using an approximate marginal likelihood function or another pseudolikelihood function of the parameter of interest rather than the likelihood function of the full parameter [e.g., 26, chapters 8–9]. Which pseudolikelihood function is chosen for eliminating nuisance parameters may improve applications of the general law of likelihood by depending on criteria outside the statistical model [3,27].

More generally, paradoxes arise from likelihood inference under inappropriate models of the system of interest [12, Section 6.5]. The level of abstraction in a model must be selected with the use of the model in mind. In complex problems such as large-scale inference (e.g., Example 1.1), multiple levels of abstraction may be used together to weigh the evidence. This section applies two levels of abstraction to the problem of testing multiple hypotheses.

Considering some function φ and a parameter set $\Phi = \{\varphi(\theta) : \theta \in \Theta\}$, let $\mathfrak{I}$ denote the smallest set of subsets of Φ that has $\{\{\varphi(\theta) : \theta \in \mathcal{H}\} : \mathcal{H} \in \mathfrak{H}\}$ as a subset. For every $\phi \in \Phi$, let g_ϕ be a probability density function on a set $\mathcal{Z}$ of possible values of an observable random element Z of observed value z.

Definition 3.1: The triple $(\{f_\theta : \theta \in \Theta\}, \{g_\phi : \phi \in \Phi\}, \mathfrak{Omega})$ is called a *fusion* of the two parametric models, the density families $\{f_\theta : \theta \in \Theta\}$ and $\{g_\phi : \phi \in \Phi\}$, with respect to a partition $\mathfrak{Omega} \subset \mathfrak{I}$ of Φ. A model used to quantify the weight of evidence without any fusion (Section 2) is a *pure model*.

The function $L^f : \mathfrak{H} \times \mathfrak{H} \to [0, \infty]$ is defined in accordance with Equations (2) and (3) such that $L^f(\mathcal{H}|\mathcal{R}) = L(\mathcal{H}|\mathcal{R})$ for all $\mathcal{H}, \mathcal{R} \in \mathfrak{H}$. In analogy with Equation (2), the function $L^g : \mathfrak{I} \times \mathfrak{I} \to [0, \infty]$ is defined such that

$$L^g(\mathcal{I}|\mathcal{S}) = \frac{\sup_{\phi \in \mathcal{I} \cap \mathcal{S}} g_\phi(z)}{\sup_{\phi \in \mathcal{S}} g_\phi(z)}$$

for all $\mathcal{I}, \mathcal{S} \in \mathfrak{I}$. For simplicity of notation, we suppress the dependence of $L^g(\mathcal{I}|\mathcal{S})$ and related values on z. For any $\mathcal{I} \in \mathfrak{I}$, define the function φ^{-1} such that

$$\varphi^{-1}(\mathcal{I}) = \{\theta \in \Theta : \varphi(\theta) \in \mathcal{I}\}.$$

Let $L^{fg} : \mathfrak{H} \times \mathfrak{I} \times \mathfrak{I} \to [0, \infty]$ denote the function satisfying

$$L^{fg}(\mathcal{H}, \mathcal{I}|\mathcal{S}) = L^f\left(\mathcal{H}|\varphi^{-1}(\mathcal{I} \cap \mathcal{S})\right) L^g(\mathcal{I}|\mathcal{S}), \tag{7}$$

which reduces to $L^f(\mathcal{H} \cap \mathcal{I}|\mathcal{S})$ according to Equation (3) whenever $\varphi(\theta) = \theta$ and $g_\theta = f_\theta$ for all $\theta \in \Theta$. With $L^{fg}(\bullet, \bullet) = L^{fg}(\bullet, \bullet|\Phi)$, $L^f(\bullet) = L^f(\bullet|\Theta)$, and $L^g(\bullet) = L^g(\bullet|\Phi)$, Equation (7) degenerates to

$$L^{fg}(\mathcal{H}, \mathcal{I}) = L^f\left(\mathcal{H}|\varphi^{-1}(\mathcal{I})\right) L^g(\mathcal{I}). \tag{8}$$

Analogously generalizing Equation (6), let

$$L^{fg}_{\mathfrak{Q}}(\mathcal{H}|\mathcal{S}) = \sup_{\mathcal{I} \in \mathfrak{Q}} L^{fg}(\mathcal{H}, \mathcal{I}|\mathcal{S}), \tag{9}$$

and let $L^{fg}_{\mathfrak{Q}}(\mathcal{H}) = L^{fg}_{\mathfrak{Q}}(\mathcal{H}|\Phi)$.

The corresponding conditional weight of evidence in the observation that $X = x$ and $Z = z$ substantiating the hypothesis that $\theta \in \mathcal{H}_1$ as opposed to the hypothesis that $\theta \in \mathcal{H}_2$ is generalized to

$$W^{fg}_{\mathfrak{Q}}(\mathcal{H}_1; \mathcal{H}_2|\mathcal{S}) = \frac{L^{fg}_{\mathfrak{Q}}(\mathcal{H}_1|\mathcal{S})}{L^{fg}_{\mathfrak{Q}}(\mathcal{H}_2|\mathcal{S})}, \tag{10}$$

and that substantiating the hypothesis that $\theta \in \mathcal{H}$ given $\varphi(\theta) \in \mathcal{S}$ to

$$W^{fg}_{\mathfrak{Q}}(\mathcal{H}|\mathcal{S}) = W^{fg}_{\mathfrak{Q}}\left(\mathcal{H}; \overline{\mathcal{H}}|\mathcal{S}\right) \tag{11}$$

for all $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2 \in \mathfrak{H}$ and $\mathcal{S} \in \mathfrak{I}$, where $\overline{\mathcal{H}} = \Theta \backslash \mathcal{H}$. The marginal counterparts are $W^{fg}_{\mathfrak{Q}}(\bullet, \bullet) = W^{fg}_{\mathfrak{Q}}(\bullet, \bullet|\Phi)$ and $W^{fg}_{\mathfrak{Q}}(\bullet) = W^{fg}_{\mathfrak{Q}}(\bullet|\Phi)$. Analogously, W^f is identified with the functions denoted by W in Section 2.

Theorem 3.1: *For a partition $\mathfrak{Q} \subset \mathfrak{I}$ of Φ and any $\mathcal{H} \in \mathfrak{H}$,*

$$L^{fg}_{\mathfrak{Q}}(\mathcal{H}) = \sup_{\mathcal{I} \in \mathfrak{Q}} L^f\left(\mathcal{H}|\varphi^{-1}(\mathcal{I})\right) L^g(\mathcal{I}); \tag{12}$$

$$W_{\Omega}^{fg}(\mathcal{H}) = \frac{\sup_{\mathcal{I}\in\Omega} L^f\left(\mathcal{H}|\varphi^{-1}(\mathcal{I})\right) L^g(\mathcal{I})}{\sup_{\mathcal{I}\in\Omega} L^f\left(\overline{\mathcal{H}}|\varphi^{-1}(\mathcal{I})\right) L^g(\mathcal{I})}. \tag{13}$$

Properties analogous to those of Theorem 3.1 may be proven conditional on $\varphi(\theta) \in \mathcal{S}$.

Example 3.1: Equation (12) enables the display equation in Example 2.1 to be stated in terms of two different models, f and g:

$$\begin{aligned} &L^{fg}_{\{\{0\},\{1\},\ldots,\{6103\}\}}\left(i\text{th gene is differentially expressed}\right) \\ &\quad = \max_{N_0=0,1,\ldots,6103} L^g\left(N_0 \text{ of the 6103 genes are not differentially expressed}\right) \\ &\qquad \times L^f\left(i\text{th gene is differentially expressed } | N_0 \text{ of the 6103 genes are} \right. \\ &\qquad\quad \left. \text{not differentially expressed}\right). \end{aligned}$$

More specific versions of this example will be stated in the next section.

Theorem 3.1 is the foundation of the class of models proposed in Section 4.3.1. The models of that class that are specified in Sections 4.3.2–4.3.4 are applied to real data in Section 5.

4. Multiple hypothesis weights of evidence

4.1. Notation for multiple hypothesis evidence

A typical problem of testing N null hypotheses involves data consisting of N statistics represented by the N-tuple $x = (x_1, \ldots, x_N)$, where $x_i \in \mathcal{X}$ for all $i = 1, \ldots, N$ with $\mathcal{X} \subseteq \mathbb{R}$. Such data may be a function of the original observations, as when each x_i is reduced to a test statistic or, as in Example 1.1, to a p value. They are modelled as realizations of random elements, the N-tuple of which is $X = (X_1, \ldots, X_N)$, with X_i distributed with density f_{θ_i} given $\theta_i \in \Theta$ for all $i = 1, \ldots, N$. For a parameter space Θ and N unknown parameters in Θ, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \in \Theta^N$. ($\boldsymbol{\theta}$ here is in boldface since its components $\theta_1, \ldots, \theta_N$ may be vectors.) The probability density function on $\mathbb{R}^N$ corresponding to the joint distribution of $X_1, \ldots, X_N$ is denoted by $f_{(\theta_1,\ldots,\theta_N)}$. Thus, the likelihood function of $\boldsymbol{\theta}$ is the function $f_\bullet(x_1, \ldots, x_N) : \Theta^N \to [0, \infty[$. Hypotheses about $\boldsymbol{\theta}$ correspond to members of $\mathfrak{H}$, the set of subsets of Θ^N, with the boldface again emphasizing the multidimensionality of $\boldsymbol{\theta}$.

For $i = 1, \ldots, N$ and for some $\theta_0 \in \Theta$, the ith null hypothesis is that $\theta_i = \theta_0$, assuming the idea of no effect is the same for all hypotheses. In Example 1.1, $\theta_0 = 0$, representing no differential expression, since θ_i is the expectation value of the logarithm of a ratio of expression between mutant and wild type. A referee pointed out that the assumption may be relaxed.

Given some *target set*, a nonempty set $\mathcal{T} \subseteq \{1, \ldots, N\}$ of cardinality M, consider the M null hypotheses in $\{\theta_j = \theta_0 : j \in \mathcal{T}\}$. The maximum likelihood estimate of θ_i is $\widehat{\theta}(x_i) = \arg\sup_{\theta\in\Theta} f_\theta(x_i)$, unique by assumption, leading to the abbreviation $L_i^{\max} = f_{\widehat{\theta}(x_i)}(x_i)/f_{\theta_0}(x_i)$ for every $i = 1, \ldots, N$.

4.2. Pure model for multiple hypothesis evidence

The following model is that of Bickel [3, Section 2.4.1]. Again consider some restriction set $\boldsymbol{\mathcal{R}} \in \mathfrak{H}$. According to Equation (5), the joint weight of evidence substantiating the alternative hypotheses that $\theta_j \neq \theta_0$ for all $j \in \mathcal{T}$, given that $\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}}$, is

$$W\left(\left\{\theta_j \neq \theta_0 : j \in \mathcal{T}\right\} | \boldsymbol{\mathcal{R}}\right) \equiv W\left(\boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}} | \boldsymbol{\mathcal{R}}\right) = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}} \cap \boldsymbol{\mathcal{R}}} f_{\boldsymbol{\theta}}(x)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}} \setminus \boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}}} f_{\boldsymbol{\theta}}(x)}; \tag{14}$$

$$\boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}} = \left\{(\theta_1, \ldots, \theta_N) \in \Theta^N : \forall j \in \mathcal{T}\ \theta_j \neq \theta_0\right\}. \tag{15}$$

Similarly, the marginal weight of evidence substantiating the alternative hypotheses that $\theta_j \neq \theta_0$ for some $j \in \mathcal{T}$, given that $\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}}$, is

$$W\left(\left\{\exists j \in \mathcal{T}\ \theta_j \neq \theta_0\right\} | \boldsymbol{\mathcal{R}}\right) \equiv W\left(\boldsymbol{\mathcal{H}}_{\bigvee \mathcal{T}} | \boldsymbol{\mathcal{R}}\right) = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\bigvee \mathcal{T}} \cap \boldsymbol{\mathcal{R}}} f_{\boldsymbol{\theta}}(x)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}} \setminus \boldsymbol{\mathcal{H}}_{\bigvee \mathcal{T}}} f_{\boldsymbol{\theta}}(x)};$$

$$\boldsymbol{\mathcal{H}}_{\bigvee \mathcal{T}} = \left\{(\theta_1, \ldots, \theta_N) \in \Theta^N : \exists j \in \mathcal{T}\ \theta_j \neq \theta_0\right\}.$$

Theorem 4.1: *Suppose that the X_is are mutually independent, that $L_i^{\max} > 1$ for all $i = 1, \ldots, N$, and that $J(\mathcal{T}) = \arg\min_{j \in \mathcal{T}} L_j^{\max}$ is unique. It follows that, with probability 1, the joint weight of evidence substantiating the joint alternative hypotheses that $\theta_j \neq \theta_0$ for all $j \in \mathcal{T}$ is*

$$W\left(\boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}}\right) = \min_{j \in \mathcal{T}} L_j^{\max} \tag{16}$$

and that the marginal weight of evidence substantiating the alternative hypotheses that $\theta_j \neq \theta_0$ for some $j \in \mathcal{T}$ is

$$W\left(\boldsymbol{\mathcal{H}}_{\bigvee \mathcal{T}}\right) = \prod_{j \in \mathcal{T}} L_j^{\max}. \tag{17}$$

Since Theorem 4.1 holds for $N = 1$ as well as $N \geq 2$, there is continuity in the evaluation of evidence from a single comparison to as many comparisons as needed. Similarly, the p-value adjustments that control family-wise error rates and non-local false discovery rates are continuous with single, unadjusted p values, which is why those adjustments do not overcome alleged problems with null hypothesis significance testing (Example 1.1).

Nonetheless, that continuity contrasts with the sharp discontinuity seen in applications of completely different frequentist methods to problems involving very different numbers of comparisons [28, Section 6]. When one method is recommended for small numbers of comparisons and another for large numbers of comparisons, there is an undesirable discontinuity in the jump from one method to another as the number of comparisons gradually increases.

Example 4.1: The discontinuity appears in the practice of applying p values to single comparisons but estimates of local false discovery rates to large numbers of comparisons [29]. The problem is that an estimate of a local false discovery rate is an estimated posterior probability, which differs markedly from a p value not only in philosophical interpretation but also in numerical value, at least in the case of two-sided p values [30]. What should be

done when the number of comparisons is too large for a p value but too small for a reliable estimate of the local false discovery rate? (Efron [31] suggested controlling non-local false discovery rates for that case, but their use in practice suffers from an anti-conservative bias [7,32, Chapter 6].) Why should the way evidence is quantified depend on the number of comparisons? At exactly how many comparisons should one suddenly switch methods? Is there any argument to support that practice?

Like subjective Bayesian statistics, the general law of likelihood solves the problem in principle by its applicability to any number of comparisons, ensuring a continuity in statistical inference as the number of comparisons increases. For other solutions to this problem, see Bickel [9,28,33,34].

4.3. Model fusion for multiple hypothesis evidence

The methods of this section are broadly applicable to problems of multiple testing, including that of Example 1.1. By generalizing Example 3.1, Section 4.3.1 presents a general method for determining two strengths of evidence:

(1) The strength of evidence for a hypothesis given the weight of evidence for each possible number of true null hypotheses
(2) The strength of evidence that a particular null alternative hypothesis is true given that number

Specific models for the former strength of evidence then appear in Sections 4.3.2–4.3.4. Each of those models has its own motivation and applicability (Table 1).

Further research is needed before one of the models can be recommended as a generally applicable default. The likelihood paradigm remains in its adolescence.

4.3.1. Marginalization over how many null hypotheses are true

This section presents a method of combining the strength of evidence for a number of true null hypotheses with the strength of evidence that a particular null alternative hypothesis is true given that number. That is accomplished by applying the formalism of Section 3 to marginalization over how many null hypotheses are true.

The number of true null hypotheses is $\sum_{i=1}^{N} 1_0(\theta_i)$, where $1_0(\theta_i) = 1$ if $\theta_i = \theta_0$ and $1_0(\theta_i) = 0$ if $\theta_i \neq \theta_0$. The joint weight of evidence substantiating the alternative hypotheses given by $\theta_j \neq \theta_0$ for all $j \in \mathcal{T}$, conditional on the truth of exactly N_0 null

Table 1. Some models for determining the strength of evidence for each possible N_0, the number of true null hypotheses.

Model	Section	Motivation	Applicability
Non-mixture model	4.3.2	No random parameters	No information on N_0
Mixture model	4.3.3	Random parameters	Many hypotheses
Model of p values	4.3.4	Given p values	$\underline{N}_0$ is available

Note: $\underline{N}_0$ denotes an informative lower bound on N_0.

hypotheses, is

$$W\left(\left\{\theta_j \neq \theta_0 : j \in \mathcal{T}\right\} \mid \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right) \equiv W\left(\boldsymbol{\mathcal{H}}_{\wedge \mathcal{T}} | \boldsymbol{\mathcal{R}}_{N_0}\right) = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\wedge \mathcal{T}} \cap \boldsymbol{\mathcal{R}}_{N_0}} f_{\boldsymbol{\theta}}\left(x\right)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}}_{N_0} \setminus \boldsymbol{\mathcal{H}}_{\wedge \mathcal{T}}} f_{\boldsymbol{\theta}}\left(x\right)};$$

$$\boldsymbol{\mathcal{R}}_{N_0} = \left\{\left(\theta_1, \ldots, \theta_N\right) \in \Theta^N : \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right\}. \tag{18}$$

Equation (18) may be simplified with some additional notation. Let $L_{(k)}^{\max}$ denote the kth of the $N-M$ order statistics of $L_1^{\max}, \ldots, L_N^{\max}$ that exclude $L_j^{\max}$ for every $j \in \mathcal{T}$.

Theorem 4.2: *Suppose $f_{\boldsymbol{\theta}}$ is the probability density function of the distribution of $(X_1, \ldots, X_N)$ that corresponds to $\boldsymbol{\theta}$ and that the X_is are mutually independent. It follows that, with probability 1, the joint weight of evidence substantiating the joint alternative hypotheses that $\theta_j \neq \theta_0$ for all $j \in \mathcal{T}$, conditional on the truth of exactly N_0 null hypotheses, is*

$$W^f\left(\boldsymbol{\mathcal{H}}_{\wedge \mathcal{T}} \left| \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right.\right) = \begin{cases} \dfrac{\min_{j \in \mathcal{T}} L_j^{\max}}{L_{(N_0)}^{\max}} & \text{if } 1 \leq N_0 \leq N - M \\ 0 & \text{if } N_0 > N - M \\ \infty & \text{if } N_0 = 0 \end{cases}. \tag{19}$$

With the weight of evidence from Equation (19), Lemma 2.1 gives

$$L^f\left(\boldsymbol{\mathcal{H}}_{\wedge \mathcal{T}} \left| \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right.\right) = \begin{cases} 1 & \text{if } N_0 \leq N - M \\ 0 & \text{if } N_0 > N - M \end{cases}. \tag{20}$$

While the expression for $W^f(\boldsymbol{\mathcal{H}}_{\vee \mathcal{T}} | \sum_{i=1}^{N} 1_0(\theta_i) = N_0)$ cannot be written as concisely as that of $W^f(\boldsymbol{\mathcal{H}}_{\wedge \mathcal{T}} | \sum_{i=1}^{N} 1_0(\theta_i) = N_0)$, it is easily implemented numerically, providing $L^f(\boldsymbol{\mathcal{H}}_{\vee \mathcal{T}} | \sum_{i=1}^{N} 1_0(\theta_i) = N_0)$ via Lemma 2.1.

The model $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta^N\}$ is used with another model, $\{g_{\phi} : \phi \in \boldsymbol{\Phi}\}$, to obtain $L_{\mathfrak{Q}}^{fg}(\boldsymbol{\mathcal{H}})$ and $W_{\mathfrak{Q}}^{fg}(\boldsymbol{\mathcal{H}})$ for some hypothesis that $\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}$, such as $\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}}$ or $\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\bigvee \mathcal{T}}$, according to Section 3. The parameters are related by a function $\varphi : \Theta^N \to \boldsymbol{\Phi}$ that transforms $\boldsymbol{\theta}$ to $\varphi(\boldsymbol{\theta}) = \varphi((\theta_1, \ldots, \theta_N))$ for all $\boldsymbol{\theta} \in \Theta^N$. Under the second model, the number of true null hypotheses is $\nu_0(\phi)$, where ν_0 is the function such that $\nu_0(\varphi((\theta_1, \ldots, \theta_N))) = \sum_{i=1}^{N} 1_0(\theta_i)$ for all $\boldsymbol{\theta} \in \Theta^N$. Hereafter, ν_0 is assumed to exist.

Corollary 4.1: *Consider the model fusion $(\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta^N\}, \{g_{\phi} : \phi \in \boldsymbol{\Phi}\}, \mathfrak{Q})$ with respect to the partition*

$$\mathfrak{Q} = \left\{\left\{\phi \in \boldsymbol{\Phi} : \nu_0\left(\phi\right) = 0\right\}, \left\{\phi \in \boldsymbol{\Phi} : \nu_0\left(\phi\right) = 1\right\}, \ldots, \left\{\phi \in \boldsymbol{\Phi} : \nu_0\left(\phi\right) = N\right\}\right\}.$$

For any $\boldsymbol{\mathcal{H}} \in \mathfrak{H}$,

$$L_{\mathfrak{Q}}^{fg}\left(\boldsymbol{\mathcal{H}}\right) = \max_{N_0 = 0, 1, \ldots, N} L^f\left(\boldsymbol{\mathcal{H}} \left| \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right.\right) L^g\left(\left\{\phi \in \boldsymbol{\Phi} : \nu_0\left(\phi\right) = N_0\right\}\right). \tag{21}$$

The density family $\{g_\phi : \phi \in \mathbf{\Phi}\}$ could be, for example, the model of Section 4.3.2. The model of Section 4.3.3 may be similarly applied via an approximation. The model of Section 4.3.4 is applicable given p values and an informative lower bound on N_0. Table 1 compares those three models.

Substituting Equation (20) into Equation (21) gives

$$\begin{aligned} L^{fg}_{\Omega}\left(\mathcal{H}_{\wedge\mathcal{T}}\right) &= \max_{N_0=0,1,\ldots,N} L^f\left(\mathcal{H}_{\wedge\mathcal{T}} \middle| \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right) L^g\left(\{\phi \in \mathbf{\Phi} : \nu_0\left(\phi\right) = N_0\}\right) \\ &= \max_{N_0=0,1,\ldots,N-M} L^g\left(\{\phi \in \mathbf{\Phi} : \nu_0\left(\phi\right) = N_0\}\right) \end{aligned} \tag{22}$$

as a generalization of the display equation in Example 3.1. The key quantity is $L^g(\{\phi \in \mathbf{\Phi} : \nu_0(\phi) = N_0\})$, the likeliness of N_0. Sections 4.3.2–4.3.4 derive it for fusion with different models.

4.3.2. Fusion with a non-mixture model

A K-component non-mixture model will be built from a density family $\{g_{\phi'} : \phi' \in \Phi\}$ given some $\Phi \subseteq \mathbb{R}$ and $K \in \{2, 3, \ldots\}$. Let $\mathbf{\Phi}$ denote the parameter set of ϕ constrained such that its components have K values, including the null hypothesis value $\phi(0)$:

$$\begin{aligned} \mathbf{\Phi} = \Big\{ &\phi \in \{\phi\left(0\right), \phi\left(1\right), \ldots, \phi\left(K-1\right)\}^N : \phi\left(j\right), \phi\left(k\right) \in \Phi \backslash \phi\left(0\right); \phi\left(j\right) \\ &\neq \phi\left(k\right); j, k = 1, \ldots K-1 \Big\}. \end{aligned}$$

With the independence assumption,

$$g_\phi\left(z\right) = g_{(\phi_1,\phi_2,\ldots,\phi_N)}\left(z\right) = \prod_{i=1}^{N} g_{\phi_i}\left(z_i\right)$$

for all $(\phi_1, \phi_2, \ldots, \phi_N) \in \mathbf{\Phi}$, where z_i is a p value or test statistic for the ith null hypothesis. Padilla and Bickel [35] and Yang et al. [36] used this without fusion to quantify the weight of evidence for the hypothesis that $\phi_i \in \mathcal{I}_1$ relative to that of $\phi_i \in \mathcal{I}_2$ for some $\mathcal{I}_1, \mathcal{I}_2 \in \mathfrak{I}$ by $\sup_{\phi\in\mathcal{I}_1} \prod_{i=1}^N g_\phi(z_i) / \sup_{\phi\in\mathcal{I}_2} \prod_{i=1}^N g_\phi(z_i)$. That approach fails when the $\{g_\phi : \phi \in \mathbf{\Phi}\}$ model is misspecified in the sense that K is not large enough, for in that case, the weight of evidence for some alternative hypotheses can be spuriously low. On the other hand, making K too large leads to insufficient evidence for the null hypothesis.

A more robust approach is available in the fusion of the two models, enabling the use of Equation (21) with the next result.

Lemma 4.1: *For all* $N_0 = 0, 1, \ldots, N$,

$$L^g\left(\{\phi \in \mathbf{\Phi} : \nu_0\left(\phi\right) = N_0\}\right) = \frac{\sup_{(\phi_1,\phi_2,\ldots,\phi_N)\in\mathbf{\Phi}:\nu_0((\phi_1,\phi_2,\ldots,\phi_N))=N_0} \prod_{i=1}^N g_{\phi_i}\left(z_i\right)}{\sup_{(\phi_1,\phi_2,\ldots,\phi_N)\in\mathbf{\Phi}} \prod_{i=1}^N g_{\phi_i}\left(z_i\right)}$$

Example 4.2: Following a general approach in Padilla and Bickel [35], Yang et al. [36] considered $K = 2$, $\Phi = \{0, \phi_{\text{alt}}\}$, g_0 as the central χ^2 density function with 1 degree of freedom, and $g_{\phi_{\text{alt}}}$ as the noncentral χ^2 density function with unknown noncentrality parameter ϕ_{alt}

and 1 degree of freedom. Let $\nu_0((\phi_1, \phi_2, \ldots, \phi_N)) = \sum_{i=1}^{N} 1_{\{0\}}(\phi_i)$, where $1_{\{0\}}(\phi_i) = 1$ if $\phi_i = 0$ or $1_{\{0\}}(\phi_i) = 1$ if $\phi_i = \phi_{\text{alt}}$. According to Lemma 4.1,

$$L^g\left(\left\{\phi \in \{0, \phi_{\text{alt}}\}^N : \nu_0(\phi) = N_0\right\}\right) = \frac{\sup_{\phi_{\text{alt}}>0,(\phi_1,\phi_2,\ldots,\phi_N)\in\{0,\phi_{\text{alt}}\}^N:\nu_0((\phi_1,\phi_2,\ldots,\phi_N))=N_0} \prod_{i=1}^{N} g_{\phi_i}(z_i)}{\sup_{\phi_{\text{alt}}>0} \prod_{i=1}^{N} \left(g_0(z_i) \vee g_{\phi_{\text{alt}}}(z_i)\right)},$$

with $\vee$ denoting the maximum.

4.3.3. Fusion with a mixture model

In a K-component mixture model formulated to obtain maximum likelihood estimates of false discovery rates [37,38],

$$g_\phi = \sum_{k=0}^{K-1} \pi_k g_{\phi(k)},$$

where $\pi_k \in [0, 1]$ for each $k \in 0, 1, \ldots, K-1$ such that $\sum_{k=0}^{K-1} \pi_k = 1$ and where each $g_{\phi(k)}$ with $\phi(k) \in \Phi$ is a probability density function according to the $\{g_{\phi'} : \phi' \in \Phi\}$ family. Thus, $(\pi_k, \phi(k)) \in [0, 1] \times \Phi$ for each k, and $\mathbf{\Phi} = ([0, 1] \times \Phi)^K$, unlike the parameter set in Section 4.3.2. For all matrices $\phi = ((\pi_0, \phi(0)), (\pi_1, \phi(1)), \ldots, (\pi_{K-1}, \phi(K-1))) \in ([0, 1] \times \Phi)^K$,

$$g_\phi(z) = \prod_{i=1}^{N} g_\phi(z_i).$$

Without model fusion, the weight of evidence for the ith alternative hypothesis has been quantified as $(\sum_{k=1}^{K-1} \widehat{\pi}_k g_{\widehat{\phi}(k)}(z_i) / \sum_{k=1}^{K-1} \widehat{\pi}_k) / g_{\widehat{\phi}(0)}(z_i)$, where $(\widehat{\pi}_k, \widehat{\phi}(k))$ is the maximum likelihood estimate of $(\pi_k, \phi(k))$ for each $k = 0, \ldots, K-1$ [35,36]. As this fails under marked misspecification in the same way as does using the non-mixture model of Section 4.3.2, the $\{g_\phi : \phi \in \mathbf{\Phi}\}$ model may be more widely applicable when fused with a less restrictive model.

Such model fusion may be implemented by letting $\nu_0(\phi) = \pi_0 N$ for all

$$((\pi_0, \phi(0)), (\pi_1, \phi(1)), \ldots, (\pi_{K-1}, \phi(K-1))) \in ([0, 1] \times \Phi)^K$$

and by the constraint $\pi_k \in \mathcal{P}_N = \{0, 1/N, 2/N, \ldots, N\}$ for each $k \in 0, 1, \ldots, K$. (Although $\nu_0(\phi)$ in general is a function of ϕ, it is constant in this case.) The equation $\nu_0(\phi) = \pi_0 N$ treats π_0 as if it were the proportion of true null hypotheses, which, while not strictly correct, is an adequate approximation for sufficiently large N according to the law of large numbers. (The consequences for small N have not been studied.) The fusion of the two models provides marginalization according to $\mathbf{\Phi} = (\mathcal{P}_N \times \Phi)^K$ and Equation (21), as follows, with the proof analogous to that of Lemma 4.1.

Lemma 4.2: *For all $N_0 = 0, 1, \ldots, N$ and any $K = 2, 3, \ldots$,*

$$L^g(\{\phi \in \Phi : \nu_0(\phi) = N_0\}) = \frac{\sup_{\phi\in(\mathcal{P}_N\times\Phi)^K:\pi_0=N_0/N,\sum_{k=0}^{K-1}\pi_k=1} \prod_{i=1}^{N} g_\phi(z_i)}{\sup_{\phi\in(\mathcal{P}_N\times\Phi)^K:\sum_{k=0}^{K-1}\pi_k=1} \prod_{i=1}^{N} g_\phi(z_i)},$$

where z_i is a p value or test statistic for the ith null hypothesis.

Example 4.3: In the mixture-model equivalent of Example 4.2, $K = 2$, $\Phi = \{0, \phi_{\text{alt}}\}$, and g_0 and $g_{\phi_{\text{alt}}}$ are the same as before [36]. By Lemma 4.2,

$$L^g\left(\{\phi \in \{0, \phi_{\text{alt}}\} : \nu_0(\phi) = N_0, \phi_{\text{alt}} > 0\}\right) = \frac{\sup_{\phi_{\text{alt}}>0} \prod_{i=1}^N \left(\frac{N_0}{N} g_0(z_i) + \left(1 - \frac{N_0}{N}\right) g_{\phi_{\text{alt}}}(z_i)\right)}{\sup_{\phi_{\text{alt}}>0, \pi_0 \in \mathcal{P}_N} \prod_{i=1}^N \left(\pi_0 g_0(z_i) + (1-\pi_0) g_{\phi_{\text{alt}}}(z_i)\right)}. \quad (23)$$

4.3.4. Fusion with a model of p values

This subsection illustrates how to weigh statistical evidence in a genomics data analysis problem that requires reporting a list of 'significant features' from a larger list of N biological features such as protein, lipid, or metabolite species. That is usually accomplished using p values, with one p value per biological feature. Typical examples are a list of genes with evidence of differential expression (Example 1.1) and a list of genetic variants with evidence of being associated with a disease or other trait [e.g., 39].

In this context, the ith feature corresponds to the null hypothesis that $Z_i \sim g_1$ and to the alternative hypothesis that $Z_i \sim g_{\phi_i}$, where $\phi_i \neq 1$ and, for any $\phi > 0$, g_ϕ is the normal density with mean 0 and standard deviation ϕ and Z_i is the random variable of observed statistic z_i [40]. The z_i is often the standard normal quantile of a one-sided p value.

For some list size M, let $\mathcal{T}_M$, called the *M-feature list*, denote the set of M indices of the alternative hypotheses that have the highest maximum likelihood ratios, that is, such that, for every $i \in \mathcal{T}_M$ and $j \notin \mathcal{T}_M$, $\ell_i^{\max} > \ell_j^{\max}$, assuming no ties, where $\ell_i^{\max} = \sup_{\phi>0} g_\phi(z_i)/g_1(z_i)$. Bickel [40] noted that

$$\ell_i^{\max} = \frac{\sup_{\phi>0} \phi^{-1} e^{-z_i^2/(2\phi^2)}}{e^{-z_i^2/2}} = \frac{1}{|z_i|} e^{\frac{z_i^2-1}{2}}, \quad (24)$$

which is an upper bound of a Bayes factor if $|z_i| \geq 1$ [41]. Let $\ell_{(k)}^{\max}$ denote the kth order statistic of the $\ell_i^{\max}$s. The rationale is that the list of significant features should consist of the features with sufficient evidence that $Z_i \sim g_{\phi_i}$, with M to be determined as follows.

The $\Phi \in \mathbf{\Phi}$ restriction of Section 4.3.2 is replaced with the $\nu_0(\phi) \geq \underline{N}_0$ restriction, where $\phi = (\phi_1, \ldots, \phi_N)$ and $\underline{N}_0$ is the lowest plausible number of true null hypotheses [cf. 9].

Lemma 4.3: *For all* $N_0 = 0, 1, \ldots, N$,

$$L^g\left(\{\phi \in \mathbf{\Phi} : \nu_0(\phi) = N_0\}\right) = \begin{cases} \frac{\prod_{k=N-N_0+1}^N \ell_{(k)}^{\max}}{\prod_{k=N-\underline{N}_0+1}^N \ell_{(k)}^{\max}} & \text{if } N_0 \geq \underline{N}_0 \\ 0 & \text{if } N_0 < \underline{N}_0 \end{cases}$$

if $N_0 \geq \underline{N}_0$ *but* $L^g(\{\phi_i \neq 1 : i \in \{1, \ldots, N\}, \nu_0(\phi) = N_0\} | \nu_0(\phi) \geq \underline{N}_0) = 0$ *otherwise.*

Plugging $L^g(\{\phi \in \mathbf{\Phi} : \nu_0(\phi) = N_0\})$ from Lemma 4.3 into Equation (22) gives

$$L_{\Omega}^{fg}(\mathcal{T}_M) = \max_{N_0 \geq \underline{N}_0} \frac{\prod_{k=N-N_0+1}^N \ell_{(k)}^{\max}}{\prod_{k=N-\underline{N}_0+1}^N \ell_{(k)}^{\max}} = 1$$

for $M \leq N - \underline{N}_0$ and $L_{\Omega}^{fg}(\boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}}) = 0$ for $M > N - \underline{N}_0$. The resulting weight of evidence is

$$W^f\left(\mathcal{T}_M | \nu_0\left(\phi\right) \geq \underline{N}_0\right)$$
$$= W^f\left(\mathcal{T}_M | \nu_0\left(\phi\right) = \underline{N}_0\right) = \begin{cases} \dfrac{\min_{j \in \mathcal{T}_M} \ell_j^{\max}}{\ell_{(\underline{N}_0)}^{\max}} & \text{if } 1 \leq \underline{N}_0 \leq N - M \\ 0 & \text{if } \underline{N}_0 > N - M \\ \infty & \text{if } \underline{N}_0 = 0 \end{cases} \tag{25}$$

according to Equation (19).

Let $M_{N_0}(\Lambda)$ denote the size of the largest M-feature list such that its weight of evidence conditional on at least $\underline{N}_0$ true null hypotheses is at least Λ for some $\Lambda \geq 1$. Then for at least negligible evidence ($\Lambda = 1$) according to Table 2,

$$\begin{aligned} M_{N_0}\left(1\right) &= \arg \max_{M=0,1,\ldots,N} W^f\left(\mathcal{T}_M | \nu_0\left(\phi\right) \geq \underline{N}_0\right) \\ &= \arg \max_{M \leq N - \underline{N}_0} W^f\left(\mathcal{T}_M | \nu_0\left(\phi\right) \geq \underline{N}_0\right) \\ &= N - \underline{N}_0. \end{aligned}$$

For higher levels of evidence ($\Lambda > 1$), perhaps corresponding to another grade of evidence in Table 2, $M_{\underline{N}_0}(\Lambda) = 0$ if $\ell_{(N)}^{\max} / \ell_{(\underline{N}_0)}^{\max} < \Lambda$ and

$$\begin{aligned} M_{\underline{N}_0}\left(\Lambda\right) &= \arg \max_{M=0,1,\ldots,N: W^f\left(\mathcal{T}_M | \nu_0(\phi) \geq \underline{N}_0\right) \geq \Lambda} W^f\left(\mathcal{T}_M | \nu_0\left(\phi\right) \geq \underline{N}_0\right) \\ &= \arg \max_{M \leq N - \underline{N}_0 : \min_{j \in \mathcal{T}_M} \ell_j^{\max} \geq \Lambda \ell_{(\underline{N}_0)}^{\max}} \frac{\min_{j \in \mathcal{T}_M} \ell_j^{\max}}{\ell_{(\underline{N}_0)}^{\max}} \\ &= \arg \max_{M \leq N - \underline{N}_0 : \ell_{(N-M+1)}^{\max} \geq \Lambda \ell_{(\underline{N}_0)}^{\max}} \frac{\min_{j \in \mathcal{T}_M} \ell_{(N-M+1)}^{\max}}{\ell_{(\underline{N}_0)}^{\max}} \end{aligned}$$

if $\ell_{(N)}^{\max} / \ell_{(\underline{N}_0)}^{\max} \geq \Lambda$. Thus, $\mathcal{T}_{\underline{N}_0}(\Lambda)$, the $M_{\underline{N}_0}(\Lambda)$-feature list, is the Λ-*joint-evidence list*, the largest list of features such that its weight of evidence, conditional on at least $\underline{N}_0$ true null hypotheses, is at least Λ. That is the set of indices corresponding to the $M_{\underline{N}_0}(\Lambda)$-largest

Table 2. Intervals for general likelihood ratios as the weight of statistical evidence [2,14,28].

Negligible	Weak	Moderate	Strong	Very strong	Overwhelming
$[2^0, 2^1[$	$[2^1, 2^2[$	$[2^2, 2^3[$	$[2^3, 2^5[$	$[2^5, 2^7[$	$[2^7, \infty]$

Notes: While the thresholds are largely set by convention [cf. 42], they to this day motivate arguments about how to improve the application of statistics to scientific reporting [cf. 43].

order statistics:

$$\begin{aligned}
&\mathcal{T}_{\underline{N}_0}(\Lambda) \coloneqq \mathcal{T}_{M_{\underline{N}_0}(\Lambda)} \\
&= \begin{cases} \left\{ i \in \{1,\ldots,N\} : \ell_i^{\max} = \ell_{(k)}^{\max}, k \in \left\{ N - M_{\underline{N}_0}(\Lambda) + 1, \ldots, N \right\} \right\} & \text{if } M_{\underline{N}_0}(\Lambda) \geq 1 \\ \emptyset & \text{if } M_{\underline{N}_0}(\Lambda) = 0 \end{cases} \\
&= \begin{cases} \left\{ i \in \{1,\ldots,N\} : \ell_i^{\max} \geq \ell_{(N - M_{\underline{N}_0}(\Lambda)+1)}^{\max} \right\} & \text{if } \ell_{(N)}^{\max} / \ell_{(\underline{N}_0)}^{\max} \geq \Lambda \\ \emptyset & \text{if } \ell_{(N)}^{\max} / \ell_{(\underline{N}_0)}^{\max} < \Lambda. \end{cases}
\end{aligned}$$

On the other hand, the Λ-*individual-evidence list*, the largest list of features corresponding to alternative hypotheses of at least Λ when considering each feature individually, is

$$\mathcal{E}_{\underline{N}_0}(\Lambda) \coloneqq \left\{ i \in \{1,\ldots,N\} : W^f\left(\{i\} \,|\nu_0(\phi) \geq \underline{N}_0\right) \geq \Lambda \right\},$$

where, as per Equation (25),

$$W^f\left(\{i\} \,|\nu_0(\phi) \geq \underline{N}_0\right) = W^f\left(\{i\} \,|\nu_0(\phi) = \underline{N}_0\right) = \begin{cases} \dfrac{\ell_i^{\max}}{\ell_{(\underline{N}_0)}^{\max}} & \text{if } 1 \leq \underline{N}_0 \leq N-1 \\ 0 & \text{if } \underline{N}_0 = N \\ \infty & \text{if } \underline{N}_0 = 0 \end{cases}.$$

$$\therefore \mathcal{E}_{\underline{N}_0}(\Lambda) \coloneqq \begin{cases} \left\{ i \in \{1,\ldots,N\} : \dfrac{\ell_i^{\max}}{\ell_{(\underline{N}_0)}^{\max}} \geq \Lambda \right\} & \text{if } 1 \leq \underline{N}_0 \leq N-1 \\ \emptyset & \text{if } \underline{N}_0 = N \\ \{1,\ldots,N\} & \text{if } \underline{N}_0 = 0 \end{cases}. \tag{26}$$

Comparing the above equations for $\mathcal{T}_{\underline{N}_0}(\Lambda)$ and $\mathcal{E}_{\underline{N}_0}(\Lambda)$ reveals that the two sets are identical. That highlights the practical utility of the deductive cogency property of possibility measures [see 44].

The approach of this section is used to analyze gene expression data in Section 5.1.

5. Applications of multiple hypothesis evidence

5.1. Gene expression data

The gene expression data analysis of this section illustrates the general method of Section 4.3.4. The $N = 6103$ one-sided p values of Example 1.1 are transformed to z_i statistics, as specified in Section 4.3.4, for $i = 1, \ldots, N$.

As seen in Equation (26), the weight of evidence in favour of each gene's differential expression between the mutant and wild type requires $\underline{N}_0$, which is set to the largest integer that is less than or equal to $\pi_0 N$, where π_0 is a lower bound of the prior probability that the ith gene is equivalently expressed for the mutant and wild type. Figure 1 shows how the weight of evidence decreases as the p value increases. For any p value smaller than 10^{-5},

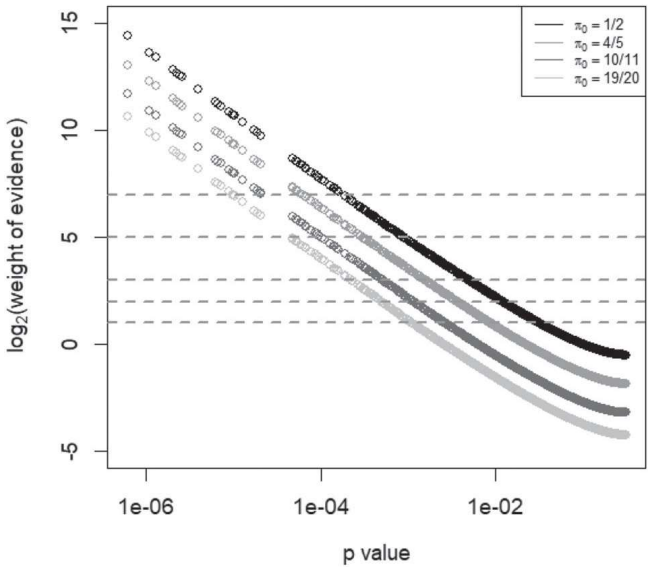

Figure 1. The weight of evidence in favour of each gene's differential expression according to equations (24) and (26) for the data set introduced in Example 1.1 at $\pi_0 = 1/2, 4/5, 10/11, 19/20$. The gray dashed lines correspond to the thresholds in Table 2. Note that since the weight of evidence is the strength of the evidence against the null hypothesis, it increases as the p value decreases. While the strength of evidence is a decreasing function of the p value in this example, the exact relation of the weight of evidence to the p value differs from one data set to another. As a result, the p value cannot be transformed to the weight of evidence using a simple calibration formula.

the evidence for differential gene expression is overwhelming according to Table 1. At the other extreme, for any p value greater than 0.035, the weight of evidence for differential expression is negligible. In less extreme cases, when the p value lees between 10^{-5} and 0.035, the weight of evidence depends on the value of π_0. For a p value equal to 0.0006, the weight of evidence based on $\pi_0 = 1/2$, 4/5, 10/11, and 19/20 would yield very strong evidence, strong evidence, moderate evidence, and weak evidence, respectively.

With π_0 as a prior probability rather than a lower bound, the most common default is $\pi_0 = 1/2$. The importance of $\pi_0 = 10/11$ is that Benjamin et al. [43] recommended 10:1 as the prior odds $\pi_0/(1 - \pi_0)$ on the basis of meta-analyses, but the value of π_0 depends on the domain.

As used in Figure 1 and below, π_0 is a known lower bound of the prior probability that the ith gene is equivalently expressed. The value of N_0 is constrained to be at least the corresponding value of $\underline{N}_0$. In contrast with published methods of estimating π_0, the value of N_0 is not estimated since it is not of interest in itself. Rather, the weight of evidence for differential expression is determined by marginalizing over N_0 to determine the likeliness of differential expression according to Example 3.1. If no lower bound can be assumed, then the model of Section 4.3.4 cannot be applied. In that case, another model mentioned in Table 1 may prove more useful. See Section 5.2 for applications of models not requiring the lower bound.

The $\ell_i^{\max}$ of Equations (24) and (26) is the bridge linking the weight of evidence to the ith gene's *local false discovery rate* (LFDR), the posterior probability that ith gene is

equivalently expressed [cf 3]. A simple estimator of that LFDR is

$$\widehat{\text{LFDR}}_i(\pi_0) = \begin{cases} \left(1 + \ell_i^{\max}(1-\pi_0)/\pi_0\right)^{-1} & \text{if } |z_i| \geq 1 \\ 1 & \text{if } |z_i| < 1 \end{cases}, \tag{27}$$

which is derived from Bayes's theorem:

$$\begin{aligned} \text{LFDR}_i(\pi_0) = \text{Prob}(\theta_i = \theta_0 | Z_i = z_i) &= \frac{\pi_0 g_0(z_i)}{\pi_0 g_0(z_i) + (1-\pi_0) g_{\phi_{\text{alt}}}(z_i)} \\ &= \left(1 + \frac{g_{\phi_{\text{alt}}}(z_i)}{g_0(z_i)}(1-\pi_0)/\pi_0\right)^{-1}. \end{aligned}$$

See Bickel [32] for an exposition.

Figure 2 compares $\widehat{\text{LFDR}}_i(1/2)$ and $\widehat{\text{LFDR}}_i(10/11)$ to two previous estimators of the LFDR, a histogram-based estimator (HBE) using Poisson regression with the 'theoretical null' assumption [45] and the corrected false discovery rate (CFDR) [7]. There, a gene is considered statistically *significant* if its LFDR estimate is below some threshold. Figure 2 shows the number of significant genes according to each estimator of the LFDR as a function of that threshold. It is clear that the CFDR yields the lowest numbers of significant genes. $\widehat{\text{LFDR}}_i(10/11)$ has values close to those of the CFDR, whereas $\widehat{\text{LFDR}}_i(1/2)$ yields the largest numbers of significant genes. For a threshold equal to the conventional value 0.2 [46], the numbers of genes considered differentially expressed are as follows: CFDR, 69 genes; $\widehat{\text{LFDR}}_i(10/11)$, 115 genes; HBE, 344 genes; $\widehat{\text{LFDR}}_i(1/2)$, 573 genes. The CFDR and HBE differ from each other due to very different methods of estimating the LFDR: the CFDR calibrates an estimated non-local false discovery rate, whereas the HBE instead

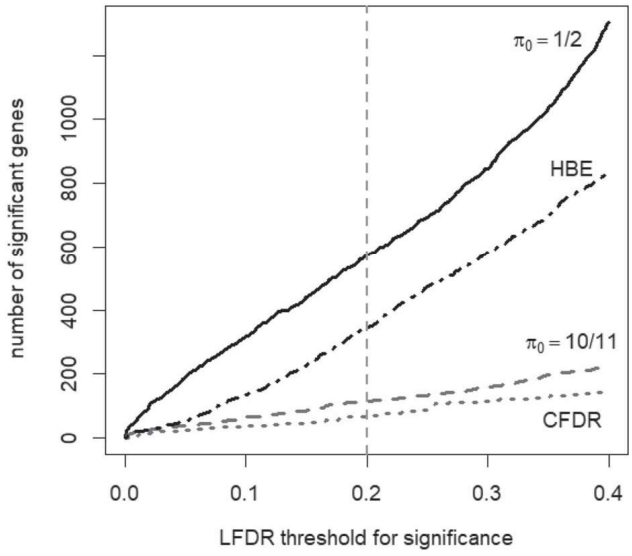

Figure 2. Number of significant genes for the data set introduced in Example 1.1 for different estimators of LFDR, corrected false discovery rate ('CFDR'), histogram based estimator ('HBE'), $\widehat{\text{LFDR}}_i(1/2)$ ('$\pi_0 = 1/2$'), and $\widehat{\text{LFDR}}_i(10/11)$ ('$\pi_0 = 10/11$').

relies on a histogram. An anonymous reviewer disagrees, instead attributing the difference between the CFDR and the HBE to differing tacit assumptions about the value of π_0.

The independence assumption of Section 4.3.1, while not strictly true in applications to genomics data, provides a practical approximation or idealization in many cases, as seen in its ubiquity in the bioinformatics literature. Much of the research in systems biology that replaces that simple assumption with more complex structures may provide further opportunities for model fusion according to the general framework of Section 3.

5.2. Exam score data

Rubin [47] reports means and variances of SAT exam score differences between students participating in a training programme and those not participating for each of eight exam sites. The standardized mean differences are denoted by $x = (x_1, \ldots, x_8)$ and modelled as independent variates from $\mathrm{N}(\theta_1, 1), \ldots, \mathrm{N}(\theta_8, 1)$, respectively. Thus, $f_{\boldsymbol{\theta}}(x) = \prod_{i=1}^{8} f_{\theta_i}(x_i)$, where f_{θ_i} is the normal density function of mean θ_i and unit variance for all $i = 1, \ldots, 8$. The ith of the $N = 8$ null hypotheses is that $\theta_i = 0$, and the ith alternative hypothesis that $\theta_i \neq 0$.

Three approaches to modelling are employed to highlight advantages of model fusion (Section 3). The approaches represent all exam sites together with a pure non-mixture model (Section 4.2), with a fusion of two non-mixture models (Section 4.3.2), and with a fusion of a non-mixture model with a mixture model (Section 4.3.3). In the first approach, the weights of evidence considered are simply given by Theorem 4.1.

Each of the two fusion approaches needs its own model to fuse with $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^8\}$. Both fusion approaches use $z = (z_1, \ldots, z_8) = (x_1, \ldots, x_8)$, $K = 2$, $\Phi = \{0, \phi_{\mathrm{alt}}\}$, g_0 as the standard normal density function, and $g_{\phi_{\mathrm{alt}}}$ as the $\mathrm{N}(\phi_{\mathrm{alt}}, 1)$ density function with known mean $\phi_{\mathrm{alt}} = 2$. The family $\{g_\phi : \phi \in \{0, \phi_{\mathrm{alt}}\}^8\}$ has a non-mixture model version (Section 4.3.2) and a mixture model version (Section 4.3.3). For the former, Lemma 4.1 gives Equation (4.2) as the likeliness function of N_0 with the density functions of this section in place of the χ^2 density functions of Example 4.2. For the latter, Lemma 4.2 yields

$$L^g\left(\{\phi \in \{0, \phi_{\mathrm{alt}}\} : \nu_0(\phi) = N_0\}\right) = \frac{\prod_{i=1}^{N}\left(\frac{N_0}{N} g_0(z_i) + \left(1 - \frac{N_0}{N}\right) g_{\phi_{\mathrm{alt}}}(z_i)\right)}{\sup_{\pi_0 = 0, \frac{1}{8}, \ldots, \frac{7}{8}, 1} \prod_{i=1}^{N}\left(\pi_0 g_0(z_i) + (1 - \pi_0) g_{\phi_{\mathrm{alt}}}(z_i)\right)}$$

as the likeliness function of N_0 in contrast with that of Equation (23), in which ϕ_{alt} is unknown.

Those two likeliness functions of N_0 are plotted in Figure 3. The three modelling approaches are compared in Figures 4–6.

In Figure 3, the non-mixture model is seen to have a sharper likeliness function than the mixture model. Since the likeliness under the non-mixture model is concentrated around high values of N_0, the evidence supports the null hypothesis more than under the mixture model. Figures 4–6 reflect that in the low weights of evidence assigned by the fusion with the non-mixture model. The mixture model may perform better than the non-mixture model if N is sufficiently large and if π_0 is close to 1 [36].

Figures 4–6 also indicate that under the pure model approach, the evidence is weighed without any shrinkage toward the null hypotheses. That is because the pure model reports

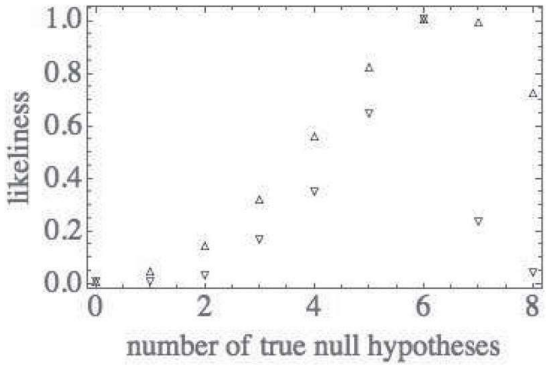

Figure 3. Likeliness as a function of N_0 under a non-mixture model (∇; Section 4.3.2) and a mixture model (Δ; Section 4.3.3).

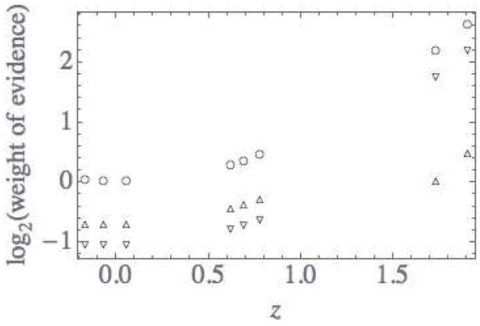

Figure 4. The weight of the evidence favouring the hypothesis that $\theta_i \neq 0$ for each $i = 1, \ldots, 8$ under a pure non-mixture model ($\bigcirc$; Section 4.2), under a fusion of two non-mixture models (∇; Section 4.3.2), and under a fusion of a non-mixture model with a mixture model (Δ; Section 4.3.3). Each fusion includes the model of Figure 3 with the corresponding symbol.

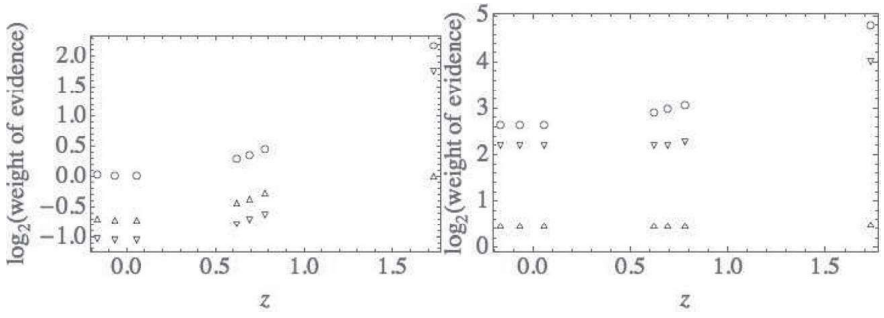

Figure 5. The weight of the evidence favouring the conjunctive hypothesis that $\theta_i \neq 0$ and $\theta_1 \neq 0$ (left) or favouring the disjunctive hypothesis that $\theta_i \neq 0$ or $\theta_1 \neq 0$ (right) for each $i = 2, \ldots, 8$. $\bigcirc < \nabla < \Delta$: caption of Figure 4.

the strength of evidence as if all likeliness were concentrated at $N_0 = 0$, as would be the case were it known that no null hypothesis is true.

An undesirable manifestation of the lack of shrinkage under the pure model is that all the alternative hypotheses are seen in the plots to be favoured over the null hypotheses

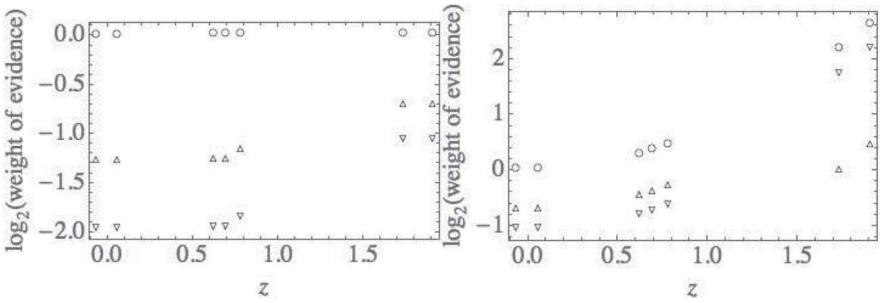

Figure 6. The weight of the evidence favouring the conjunctive hypothesis that $\theta_i \neq 0$ and $\theta_3 \neq 0$ (left) or favouring the disjunctive hypothesis that $\theta_i \neq 0$ or $\theta_3 \neq 0$ (right) for each $i = 1, 2, 4, \ldots, 8$. $\bigcirc < \nabla < \Delta$: caption of Figure 4.

($\log_2 W(\mathcal{H}) > 0$), at least to a negligible extent (Table 2). In fact, this occurs more generally [3]. By contrast, both fused-model approaches lead to the support of most of the null hypotheses by more evidence than their alternative hypotheses, as indicated by the negative log weights of evidence in Figures 4–6.

Thus, the fusion between the two non-mixture models in this application strikes an intuitively reasonable balance between the extremes of excessive weights of evidence according to the pure model and overly conservative weights of evidence according to the fusion with the mixture model.

Acknowledgments

We thank the anonymous reviewers and the Associate Editor for several comments leading to a clearer presentation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was partially supported by the Canada Foundation for Innovation (CFI16604), by the Ministry of Research and Innovation of Ontario (MRI16604), and by the Faculty of Medicine of the University of Ottawa.

ORCID

David R. Bickel http://orcid.org/0000-0002-9623-3946

References

[1] Wasserstein RL, Lazar NA. The ASA's statement on p-values:context, process, and purpose. Am Stat. 2016;70(2):129–133.
[2] Royall R. Statistical evidence: a likelihood paradigm. New York: CRC Press; 1997.
[3] Bickel DR. The strength of statistical evidence for composite hypotheses:inference to the best explanation. Stat Sin. 2012;22:1147–1198.
[4] Zhang Z, Zhang B. A likelihood paradigm for clinical trials (with discussion). J Stat Theory Pract. 2013;7:157–203.

[5] Alba R, Payton P, Fei Z, et al. Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. Plant Cell. 2005;17:2954–2965.
[6] Bickel DR. Game-theoretic probability combination with applications to resolving conflicts between statistical methods. Int J Approx Reason. 2012;53:880–891.
[7] Bickel DR, Rahal A. Correcting false discovery rates for their bias toward false positives. Commun Stat Simul Comput. 2019. doi:10.1080/03610918.2019.1630432.
[8] Dudoit S, van der Laan MJ. Multiple testing procedures with applications to genomics. New York: Springer; 2008.
[9] Bickel DR. Small-scale inference: empirical Bayes and confidence methods for as few as a single comparison. Int Stat Rev. 2014;82:457–476.
[10] Korn EL, Freidlin B. The likelihood as statistical evidence in multiple comparisons in clinical trials: no free lunch. Biometrical J. 2006;48:346–355.
[11] Strug LJ, Hodge SE. An alternative foundation for the planning and evaluation of linkage analysis I. Decoupling 'error probabilities' from 'measures of evidence'. Hum Hered. 2006;61:166–188.
[12] Lindsey J. Parametric statistical inference. Oxford: Oxford Science Publications, Clarendon Press; 1996.
[13] Ando T. Bayesian model selection and statistical modeling. Statistics: a series of textbooks and monographs. Taylor & Francis; 2010
[14] Bickel DR. Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors. Int J Approx Reason. 2015;66:53–72.
[15] Bickel DR. A note on fiducial model averaging as an alternative to checking Bayesian and frequentist models. Commun Stat Theory Methods. 2018;47:3125–3137.
[16] Claeskens G, Hjort NL. Model selection and model averaging. Cambridge: Cambridge University Press; 2008.
[17] Wasserman L. Bayesian model selection and model averaging. J Math Psychol. 2000;44(1):92–107.
[18] Giang PH, Shenoy PP. Decision making on the sole basis of statistical likelihood. Artif Intell. 2005;165:137–163.
[19] Barndorff-Nielsen OE. Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. J R Stat Soc B. 1994;56:125–140.
[20] Bjørnstad JF. On the generalization of the likelihood function and the likelihood principle. J Am Stat Assoc. 1996;91:791–806.
[21] Pawitan Y. In all likelihood: statistical modeling and inference using likelihood. Oxford: Clarendon Press; 2001.
[22] Wang Z, Klir GJ. Generalized measure theory (IFSR international series on systems science and engineering). New York: Springer; 2008.
[23] Spohn W. The laws of belief: ranking theory and its philosophical applications. Oxford: Oxford University Press; 2012.
[24] Puhalskii A. Large deviations and idempotent probability. Monographs and surveys in pure and applied mathematics. New York: CRC Press; 2001.
[25] Bickel DR. The sufficiency of the evidence, the relevancy of the evidence, and quantifying both with a single number; 2019. Working paper, doi:10.5281/zenodo.2538412.
[26] Severini T. Likelihood methods in statistics. Oxford: Oxford University Press; 2000.
[27] Bickel DR. Pseudo-likelihood, explanatory power, and Bayes's theorem [comment on "A likelihood paradigm for clinical trials"]. J Stat Theory Pract. 2013;7:178–182.
[28] Bickel DR. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. Can J Stat. 2011;39:610–631.
[29] Westfall PH. Comment on B. Efron, "Correlated z-values and the accuracy of large-scale statistical estimates". J Am Stat Assoc. 2010;105:1063–1066.
[30] Lindley DV. A statistical paradox. Biometrika. 1957;44:187–192.
[31] Efron B. Rejoinder to comments on B. Efron, "Correlated z-values and the accuracy of large-scale statistical estimates". J Am Stat Assoc. 2010;105:1067–1069.

[32] Bickel DR. Genomics data analysis: false discovery rates and empirical Bayes methods. New York: Chapman and Hall/CRC; 2020. Available from: https://davidbickel.com/genomics/.
[33] Bickel DR. Blending Bayesian and frequentist methods according to the precision of prior information with applications to hypothesis testing. Stat Methods Appt. 2015;24:523–546.
[34] Bickel DR. Confidence distributions applied to propagating uncertainty to inference based on estimating the local false discovery rate: a fiducial continuum from confidence sets to empirical Bayes set estimates as the number of comparisons increases. Commun Stat Theory Methods. 2017;46:10788–10799.
[35] Padilla M, Bickel DR. Estimators of the local false discovery rate designed for small numbers of tests. Stat Appl Genet Mol Biol. 2012;11(5):art. 4.
[36] Yang Y, Aghababazadeh FA, Bickel DR. Parametric estimation of the local false discovery rate for identifying genetic associations. IEEE/ACM Trans Comput Biol Bioinf. 2013;10:98–108.
[37] Muralidharan O. An empirical Bayes mixture method for effect size and false discovery rate estimation. Ann Appl Stat. 2010;4:422–438.
[38] Pawitan Y, Murthy K, Michiels S, Ploner A. Bias in the estimation of false discovery rate in microarray studies. Bioinformatics. 2005;21:3865–3872.
[39] Karimnezhad A, Bickel DR. Incorporating prior knowledge about genetic variants into the analysis of genetic association data: an empirical Bayes approach. IEEE/ACM Trans Comput Biol Bioinf. 2018. doi:10.1109/TCBB.2018.2865420. Available from: https://ieeexplore.ieee.org/document/8436435/.
[40] Bickel DR. Sharpen statistical significance: evidence thresholds and Bayes factors sharpened into Occam's razor. Stat. 2019;8(1):e215.
[41] Held L, Ott M. How the maximal evidence of p-values against point null hypotheses depends on sample size. Am Stat. 2016;70(4):335–341.
[42] Jeffreys H. Theory of probability. London: Oxford University Press; 1948.
[43] Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav. 2017;1:0–0.
[44] Bickel DR, Patriota AG. Self-consistent confidence sets and tests of composite hypotheses applicable to restricted parameters. Bernoulli. 2019;25(1):47–74.
[45] Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc. 2004;99:96–104.
[46] Efron B. Large-Scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge: Cambridge University Press; 2010.
[47] Rubin DB. Estimation in parallel randomized experiments. J Educ Stat. 1981;6:377–401.

Appendix

Appendix 1. Proofs of claims

A.1 Proofs of claims made in Sections 2 and 3

A.1.1 Proof of Lemma 2.1

In the $W(\mathcal{H}|\mathcal{R}) \geq 1$ case, Equation (5) implies that $\sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x) \geq \sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)$ and thus that $\sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x) = \sup_{\theta\in\mathcal{R}} f_\theta(x)$. By Equation (4), $L(\mathcal{H}|\mathcal{R}) = 1$. Therefore, Equation (5) says $U(\mathcal{H}|\mathcal{R}) = 1/W(\mathcal{H}|\mathcal{R})$. The analogous argument for the $W(\mathcal{H}|\mathcal{R}) < 1$ case yields this chain of results: $\sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x) < \sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)$, $\sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x) = \sup_{\theta\in\mathcal{R}} f_\theta(x)$, $U(\mathcal{H}|\mathcal{R}) = 1$, and $L(\mathcal{H}|\mathcal{R}) = W(\mathcal{H}|\mathcal{R})$, in that order.

A.1.2 Proof of Theorem 3.1

Equation (9) yields

$$L^{fg}_{\mathfrak{Q}}(\mathcal{H}) = \sup_{\mathcal{I}\in\mathfrak{Q}} L^{fg}_{\mathfrak{Q}}(\mathcal{H},\mathcal{I}),$$

which, with Equation (8), in turn yields Equation (12). Equation (13) follows immediately from Equations (10) and (11).

A.2 Proofs of claims made in Section 4

A.2.1 Proof of Theorem 4.1

The following statements hold with probability 1. By Equation (14) and the fact that $L_i^{\max} > 1$,

$$W\left(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}\right) = W\left(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}|\Theta^N\right) = \frac{\sup_{(\theta_1,\ldots,\theta_N)\in\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}}\prod_{i=1}^N f_{\theta_i}(x_i)}{\sup_{(\theta_1,\ldots,\theta_N)\in\Theta^N\setminus\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}}\prod_{i=1}^N f_{\theta_i}(x_i)}$$

$$= \frac{\prod_{i\in\{1,\ldots,N\}} f_{\widehat{\theta}(x_i)}(x_i)}{f_{\theta_0}\left(x_{J(\mathcal{T})}\right)\prod_{i\in\{1,\ldots,N\}\setminus\{J(\mathcal{T})\}} f_{\widehat{\theta}(x_i)}(x_i)} = \frac{f_{\widehat{\theta}(x_{J(\mathcal{T})})}\left(x_{J(\mathcal{T})}\right)\prod_{i\in\{1,\ldots,N\}\setminus\{J(\mathcal{T})\}} f_{\widehat{\theta}(x_i)}(x_i)}{f_{\theta_0}\left(x_{J(\mathcal{T})}\right)\prod_{i\in\{1,\ldots,N\}\setminus\{J(\mathcal{T})\}} f_{\widehat{\theta}(x_i)}(x_i)},$$

and cancellation of the products yields $W(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}) = f_{\widehat{\theta}(x_{J(\mathcal{T})})}(x_{J(\mathcal{T})})/f_{\theta_0}(x_{J(\mathcal{T})})$ and thus Equation (16). Similarly,

$$W\left(\boldsymbol{\mathcal{H}}_{\vee\mathcal{T}}\right) = \frac{\sup_{(\theta_1,\ldots,\theta_N)\in\boldsymbol{\mathcal{H}}_{\vee\mathcal{T}}}\prod_{i=1}^N f_{\theta_i}(x_i)}{\sup_{(\theta_1,\ldots,\theta_N)\in\Theta^N\setminus\boldsymbol{\mathcal{H}}_{\vee\mathcal{T}}}\prod_{i=1}^N f_{\theta_i}(x_i)} = \frac{\prod_{i\in\{1,\ldots,N\}} f_{\widehat{\theta}(x_i)}(x_i)}{\prod_{i\in\mathcal{T}} f_{\theta_0}(x_i)\prod_{i\in\{1,\ldots,N\}\setminus\mathcal{T}} f_{\widehat{\theta}(x_i)}(x_i)}$$

$$= \frac{\prod_{i\in\mathcal{T}} f_{\widehat{\theta}(x_i)}(x_i)\prod_{i\in\{1,\ldots,N\}\setminus\mathcal{T}} f_{\widehat{\theta}(x_i)}(x_i)}{\prod_{i\in\mathcal{T}} f_{\theta_0}(x_i)\prod_{i\in\{1,\ldots,N\}\setminus\mathcal{T}} f_{\widehat{\theta}(x_i)}(x_i)},$$

leading to Equation (17) by cancellation.

A.2.2 Proof of Theorem 4.2

The order statistics are unambiguously defined since the absolute continuity of the distribution of each X_i with respect to the Lebesgue measure guarantees that there are almost surely no ties. With $J(\mathcal{T}) = \arg\min_{j\in\mathcal{T}} L_j^{\max}$, it is seen that $L_{J(\mathcal{T})}^{\max}$ is the lowest of the maximized likelihood ratios that correspond to the target parameters. Define each $x_{(k)}$ to be the value in $\{x_i : i = 1,\ldots,N\}$ such that $f_{\widehat{\theta}(x_{(k)})}(x_{(k)})/f_{\theta_0}(x_{(k)}) = L_{(k)}^{\max}$. In the case that $1 \le N_0 = N - M$, Equation (18) simplifies to

$$W^f\left(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}|\sum_{i=1}^N 1_0(\theta_i) = N_0\right)$$

$$= \left(\frac{\prod_{k=1}^{N_0} f_{\theta_0}\left(x_{(k)}\right)}{f_{\theta_0}\left(x_{J(\mathcal{T})}\right) f_{\widehat{\theta}\left(x_{(N_0)}\right)}\left(x_{(N_0)}\right)\prod_{k=1}^{N_0-1} f_{\theta_0}\left(x_{(k)}\right)}\right)\left(\frac{\prod_{j\in\mathcal{T}} f_{\widehat{\theta}(x_j)}\left(x_j\right)}{\prod_{j\in\mathcal{T}\setminus\{J(\mathcal{T})\}} f_{\widehat{\theta}(x_j)}\left(x_j\right)}\right),$$

and cancelation wherever possible yields

$$W^f\left(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}|\sum_{i=1}^N 1_0(\theta_i) = N_0\right) = \left(\frac{f_{\theta_0}\left(x_{(N_0)}\right)}{f_{\theta_0}\left(x_{J(\mathcal{T})}\right) f_{\widehat{\theta}\left(x_{(N_0)}\right)}\left(x_{(N_0)}\right)\times 1}\right)\left(\frac{f_{\widehat{\theta}(x_{J(\mathcal{T})})}\left(x_{J(\mathcal{T})}\right)}{1}\right).$$

In the case that $1 \le N_0 < N - M$, Equation (18) simplifies with cancellation to

$$W^f\left(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}|\sum_{i=1}^N 1_0(\theta_i) = N_0\right) = \left(\frac{\prod_{k=1}^{N_0} f_{\theta_0}\left(x_{(k)}\right)\prod_{k=N_0+1}^{N-M} f_{\widehat{\theta}(x_{(k)})}\left(x_{(k)}\right)}{f_{\theta_0}\left(x_{J(\mathcal{T})}\right)\prod_{k=1}^{N_0-1} f_{\theta_0}\left(x_{(k)}\right)\prod_{k=N_0}^{N-M} f_{\widehat{\theta}(x_{(k)})}\left(x_{(k)}\right)}\right)$$

$$\times\left(\frac{\prod_{j\in\mathcal{T}} f_{\widehat{\theta}(x_j)}\left(x_j\right)}{\prod_{j\in\mathcal{T}\setminus\{J(\mathcal{T})\}} f_{\widehat{\theta}(x_j)}\left(x_j\right)}\right) = \left(\frac{f_{\theta_0}\left(x_{(N_0)}\right)\times 1}{f_{\theta_0}\left(x_{J(\mathcal{T})}\right)\times f_{\widehat{\theta}\left(x_{(N_0)}\right)}\left(x_{(N_0)}\right)}\right)\left(\frac{f_{\widehat{\theta}(x_{J(\mathcal{T})})}\left(x_{J(\mathcal{T})}\right)}{1}\right).$$

Both cases together establish the result for $1 \le N_0 \le N - M$. In the case that $N_0 > N - M$, the numerator in Equation (18) is 0 since $\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}} \cap \boldsymbol{\mathcal{R}}_{N_0} = \emptyset$. In the case that $N_0 = 0$, the denominator in Equation (18) is 0 since $\boldsymbol{\mathcal{R}}_0 \setminus \boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}} = \emptyset$.

A.2.3 Proof of Corollary 4.1

With $\theta = \boldsymbol{\theta}$ and $\mathcal{H} = \boldsymbol{\mathcal{H}}$, the claim results from Theorem 3.1 since $\mathfrak{Q}$ is a member of $\mathfrak{I}$, a partition of $\boldsymbol{\Phi}$, and since

$$\varphi^{-1}\left(\{N_0\}\right) = \left\{(\theta_1, \ldots, \theta_N) \in \Theta^N : \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right\} = \boldsymbol{\mathcal{R}}_{N_0}$$

for all $N_0 = 0, 1, \ldots, N$.

A.2.4 Proof of Lemma 4.1 and, by analogy, Lemma 4.2

By Equation (4),

$$L^g\left(\{\phi \in \boldsymbol{\Phi} : \nu_0\left(\phi\right) = N_0\}\right) = \frac{\sup_{\phi \in \boldsymbol{\Phi}: \nu_0(\phi) = N_0} g_\phi\left(z_i\right)}{\sup_{\phi \in \boldsymbol{\Phi}} g_\phi\left(z_i\right)}.$$

A.2.5 Proof of Lemma 4.3

By Equation (4) and the definition of conditional likeliness,

$$\begin{aligned} L^g\left(\left\{\phi_i \neq 1 : i \in \{1, \ldots, N\}, \nu_0\left(\phi\right) = N_0\right\} \middle| \nu_0\left(\phi\right) \geq \underline{N}_0\right) &= \frac{\sup_{\phi_i \neq 1: i \in \{1, \ldots, N\}, \nu_0(\phi) = N_0} \prod_{i=1}^{N} g_{\phi_i}\left(z_i\right)}{\sup_{\phi_i \neq 1: i \in \{1, \ldots, N\}, \nu_0(\phi) \geq \underline{N}_0} \prod_{i=1}^{N} g_{\phi_i}\left(z_i\right)} \\ &= \frac{\prod_{k=N-N_0+1}^{N} \ell_{(k)}^{\max}}{\prod_{k=N-\underline{N}_0+1}^{N} \ell_{(k)}^{\max}}. \end{aligned}$$

Chapter 5

Concluding summary

In this thesis, we defined two approaches of blended distribution we call blending-before-updating (BU) and updating-before-blending (UB). These BU and UB select from a set of adequate posterior distributions a single posterior distribution in order to use it for inference and decision making. These approaches can be used in many Bayesian statistical models. Nevertheless, in this thesis we worked on two models: the normal-normal model and the Poisson-gamma model. We have shown that the BU and UB approaches for these two models do not necessarily yield the same distribution. However, under certain assumptions, they can give the same distribution. As applications on these two approaches, we presented in Chapter 2 three applications. The first application estimated the expected number of deaths caused by coronavirus in Canada, the second application compared a new biological treatment to a placebo administrated at home after a myocardial infarction, and the third application defined a new LFDR estimator that was used to find the proteins most likely affected by breast cancer.

Equally important, this research opens the door to verify if the results can be extended to all distributions for the exponential family. Also, we had one unknown parameter in the set of adequate prior or posterior distributions, and it would be

interesting to consider two or more unknown parameters. In this case the gradient descent can be used to solve the minimization problem for two dimensions or more. In addition, we used relative entropy to minimize the set of adequate prior or posterior distributions to the benchmark distribution. It would be interesting to check if the reverse relative entropy or Euclidean distance yield the same results. Furthermore, it would be interesting to check in which cases the BU and UB should be used and which one is the most accurate.

Regarding the Bayesian approach of LFDR, we presented in Chapter 3 two new estimators of the LFDR, namely, the CFDR and the RFDR, and an estimator of the NFDR. We also presented a simulation study to compare the performance of the proposed estimators to two existing LFDR estimators, the HBE and the MLE. The results of this study showed that the CFDR performs well for the small number of features. While the MLE outperforms the other estimators for the large number of features, it is highly dependent on the parametric model used to generate the data. The CFDR and the HBE perform well for the large number of features with less dependence on that model. An application on tomato gene expression data was performed to compare the results of each estimator in terms of the number of genes found that are differentially expressed. We developed an R package *CorrectedFDR* [7] to compute these estimators.

In terms of model fusion and multiple testing, we defined in Chapter 4 the theory of weight of evidence under a single model, also fusing models from different levels. To address the multiple hypothesis weights of evidence problem, we developed two models, pure and fusion models. More research is needed before recommending an applicable default model of fusion with a non-mixture model, fusion with a mixture model, and fusion with a model of p-values, which bridge the link of the weight of evidence to the LFDR. Hence, it defines a new LFDR estimator. For purposes of illustration and application of multiple hypothesis evidence, we applied these theories to gene expression data. In this application, we found that the weight of evidence

decreases as the p-value increases. We also performed a comparison between the different LFDR estimators to show for each estimator the number of genes that are differentially expressed.

Bibliography

[1] R. Alba, P. Payton, Z. Fei, R. McQuinn, P. Debbie, G. B. Martin, S. D. Tanksley, and J. J. Giovannoni. Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell*, 17:2954–2965, 2005.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[3] D. R. Bickel. Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors. *International Journal of Approximate Reasoning*, 66:53–72, 2015.

[4] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge, 2010.

[5] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 6(4):733–760, 1996.

[6] Xiaochun Li. ProData. *Bioconductor.org documentation for the ProData package*, 2009.

BIBLIOGRAPHY

[7] Abbas Rahal, Anna Akpawu, Justin Chitpin, and David R. Bickel. *CorrectedFDR: Correcting False Discovery Rates*, 2018. R package version 1.0, https://goo.gl/qMQsXa.

[8] Donald B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6:pp. 377–401, 1981.

[9] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Statistics in Practice. John Wiley & Sons, Ltd., Chichester, 2004.

[10] Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.