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

December 31, 2014

David R. Bickel

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

Abstract

According to the general law of likelihood, the weight of 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, which can be undesirable in settings with a large number 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 a familiar data set yields intuitively reasonable weights of evidence.

Keywords: direct likelihood inference; foundations of statistics; hypothesis testing; law of likelihood; likelihood principle; likelihood paradigm; multiple testing; pure likelihood methods; regularization; robust statistics; strength of statistical evidence

1 Introduction

As a rule, scientists seek to publish observations that constitute sufficient evidence to accept a hypothesis as a contribution to scientific knowledge. For measuring the strength of evidence, the likelihood paradigm has substantial advantages over the frequentist and Bayesian paradigms (Edwards, 1992; Royall, 1997; Blume, 2011; Bickel, 2012; Rohde, 2014).

The paradigm has roots in the likelihood intervals of R. A. Fisher. In a certain sense, a scalar parameter value θ is "consistent with the observations" at some level Λ if and only if $\theta^{-}(\Lambda) \leq \theta \leq \theta^{+}(\Lambda)$, where $[\theta^{-}(\Lambda), \theta^{+}(\Lambda)]$ is the interval of parameter values with likelihood within a factor of Λ of the maximum likelihood, provided that $\Lambda > 1$ (Royall, 1997, p. 26). For example, Fisher (1973, pp. 75-76) considered $\Lambda = 2, 5, 15$, flagging parameter values outside the $[\theta^{-}(\Lambda), \theta^{+}(\Lambda)]$ intervals as "implausible" and those outside even the $[\theta^{-}(15), \theta^{+}(15)]$ as "obviously open to grave suspicion" (cf. Barnard, 1967; Hoch and Blume, 2008). For vector parameters, the Λ-level *likelihood set* is the set of parameter values with likelihood within a factor of Λ of the maximum likelihood.

Just as nested confidence sets may be inverted to define a p-value for each parameter value, likelihood sets may be inverted to obtain the likelihood ratio of each parameter value relative to the maximum likelihood. Edwards (1992) and Royall (1997) interpreted the likelihood ratio as a strength of evidence, carefully limiting the scope to comparisons between simple (point) hypotheses, in which case the Bayes factor is the likelihood ratio. According to the *(special) law of likelihood* attributed to Ian Hacking, the likelihood ratio between two simple hypotheses is sufficient to quantify the strength of evidence of one hypothesis over the other, apart from prior distributions, loss functions, and the sample size (Edwards, 1992; Royall, 1997). This contrasts with the more generally applicable practice of measuring

statistical evidence for general hypotheses with the Bayes factor (cf. Jeffreys, 1948).

The primary motivation for the limitation to simple hypotheses was to avoid the subjectivity involved in specifying the prior distributions needed to define a Bayes factor for composite hypotheses, for the Bayes factor is the ratio of the likelihood means with respect to a prior distribution conditional on each hypothesis compared. To achieve both objectivity and applicability to composite hypotheses, the average likelihood over each composite hypothesis was replaced with the maximum likelihood over each composite hypothesis. The resulting ratio of maximum likelihoods is interpreted as the weight of the statistical evidence that supports one composite hypothesis over another under the *general law of likelihood* (Bickel, 2012, §2.2.3), with grades similar to those of Table 1. For example, the weight of the evidence substantiating the hypothesis that $\theta^{-}(\Lambda) \leq \theta \leq \theta^{+}(\Lambda)$ over the hypothesis that $\theta \notin [\theta^{-}(\Lambda), \theta^{+}(\Lambda)]$ is Λ. Zhang and Zhang (2013a) recommended essentially the same measure of the strength of statistical evidence for regular models and sufficiently large samples. Motivated by different concerns, Dubois et al. (1997), Walley and Moral (1999), Giang and Shenoy (2005), and Coletti et al. (2009) had previously considered the ratio of maximum likelihoods within possibility theory and upper probability theory.

The idea remains controversial for a variety of reasons. For composite hypotheses, that quantity can differ markedly from the degree of *support* defined by the degree of change from prior odds to posterior odds (Bickel, 2013a,b), the sense supported by Edwards (1992) for compatibility with analyses in the presence of priors and by Royall (2000) to interpret the likelihood ratio. Similarly, Blume (2013) does not recognize a need for assigning a strength of evidence to a composite hypothesis, maintaining that the level-Λ likelihood set simply indicates which distributions are better supported than the others by the data (cf. Zhang and Zhang, 2013b).

Further, it is often thought that the likelihood ratio cannot be directly compared to a fixed threshold Λ but that it requires calibration (Severini, 2000; Morgenthaler and Staudte, 2012; Spanos, 2013). For example, Sprott (2000, §5.3) recommends the use of fixed-confidence likelihood intervals, and Patriota (2013) proposed a quantity based on the likelihood ratio test. This type of calibration would indeed be needed to achieve specified repeated-sampling coverage rates since a level-Λ likelihood set can cover the true value of the parameter with much less than 95% frequentist probability. Likewise, from a Bayesian perspective, a level-Λ likelihood set can have a very low posterior probability.

In response to claims that the special law of likelihood rendered adjustments for multiple testing unnecessary, Korn and Freidlin (2006) 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 (2006) emphasized the role of frequentist considerations when planning an experiment, whereas Bickel (2012) argued against comparing simple null hypotheses to composite alternative hypotheses. Nonetheless, exactly that comparison remains important in scientific applications, especially those involving multiple comparisons.

In short, since the practice of rejecting parameter values outside $[\theta^{-}(\Lambda), \theta^{+}(\Lambda)]$ or outside a level-Λ likelihood set for a vector parameter is known to yield counterintuitive results under some models, neither it nor the equivalent idea of accepting a hypothesis of sufficiently high weight of evidence has been widely adopted. As with all likelihood-based inference (Lindsey, 1996, §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

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

Table 1: Intervals for general likelihood ratios as the weight of statistical evidence (Royall, 1997; Bickel, 2011, 2014a; cf. Jeffreys (1948)).

also operational characteristics. With this background, the main goal of the present paper is to bring formally contradictory models 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.

Section 2 defines the theory of the weight of statistical evidence with respect to a single model. In Section 3, it will be seen that the weight of evidence uniquely measures the strength of conclusive evidence as defined mathematically to have certain properties appropriate for accepting a hypothesis of sufficiently high evidence. This new justification for calling $W(\mathcal{H}|\mathcal{R})$ the "weight of evidence" is simpler than that of Bickel (2012). Section 4 defines model fusion and the weight of evidence with respect to two fused models. To address the multiple comparisons problem, a special case of the theory of model fusion is formulated in Section 5. Section 6 illustrates that special case by applying it to data on exam scores. Referring to that data analysis, Section 7 briefly discusses practical advantages of the fused-model weight of evidence for shrinkage and for evidence supporting a point null hypothesis.

2 Weight of evidence

2.1 Preliminary notation and definitions

Let x denote an observed scalar, vector, or matrix in some set $\mathfrak{X}$ of possible observations. This x, a realization of a random variable X, may be a statistic that depends on other observations.

Consider a set Θ and a family of exact or approximate density functions $\{f_{\theta_0} : \theta_0 \in \Theta\}$ such that $f_{\theta_0} \neq f_\theta$ for all $\theta_0 \neq \theta$. If the interest parameter value in a parameter space Φ were equal to θ, then $f_\theta(x)$ would be the probability density or probability mass of the observation that $X = x$. The *likelihood function* is the function $f_\bullet(x)$, that is, $f_\theta(x)$ as a function of θ for all $\theta \in \Theta$, and its maximum likelihood estimate (MLE) is $\widehat{\theta} = \arg\sup_\theta f_\theta(x)$.

The function $f_\bullet(x)$ may be any pseudo-likelihood function such that $f_\theta(x)$ approximates a probability density for every $\theta \in \Theta$. Thus, $f_\bullet(x)$ may be a marginal, conditional, estimated, or integrated likelihood, eliminating a nuisance parameter. If the profile likelihood does not approximate a density for a particular model, it may nevertheless be corrected to approximate a conditional or marginal likelihood in certain cases (Severini, 2000, pp. 310-312, 323). The prefix "pseudo" is somewhat misleading: even the "true" likelihood function might be considered a pseudo-likelihood function since a statistical model cannot completely capture the data-generation process (Lindsey, 1996, §6.5).

All possible hypotheses about θ correspond to members of $\mathfrak{H}$, a σ-field of subsets of Θ. For example, if $\Theta = \mathbb{R}$ and $\mathfrak{H}$ is the set of Borel subsets of Θ, then the hypothesis that $\theta \neq 0$ is the hypothesis that $\theta \in \Theta \setminus \{0\}$, corresponding to the subset $\Theta \setminus \{0\}$, which is a member of $\mathfrak{H}$.

2.2 Likeliness and unlikeliness

For any $\mathcal{H}, \mathcal{R} \in \mathfrak{H}$, let

$$L(\mathcal{H}) = \frac{\sup_{\theta \in \mathcal{H}} f_\theta(x)}{\sup_{\theta \in \Theta} f_\theta(x)} \tag{1}$$

be called the *marginal likeliness* of the hypothesis that $\theta \in \mathcal{H}$ and

$$L(\mathcal{H}|\mathcal{R}) = \frac{L(\mathcal{H} \cap \mathcal{R})}{L(\mathcal{R})} \tag{2}$$

the *conditional likeliness* of the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$. Here, the supremum is the least upper bound in $[0, \infty[$, and $\sup \emptyset \equiv 0$. To accommodate $1/0 = \infty$, $L(\mathcal{H}|\mathcal{R})$ is an extended real number in the closed interval $[0, \infty]$. The term "likeliness" replaces the "extended likelihood" of Giang and Shenoy (2005) for brevity and to avoid confusion with previous usages of the latter term in the statistics literature (Barndorff-Nielsen, 1994; Bjørnstad, 1996; Pawitan, 2001).

The likeliness of a hypothesis is insufficient as a measure of its strength of evidence since the likeliness of the hypothesis's alternative must also be considered. For that reason, it is convenient to define the *marginal unlikeliness* of the hypothesis that $\theta \in \mathcal{H}$ as $U(\mathcal{H}) = L\left(\overline{\mathcal{H}}\right)$ and the *conditional unlikeliness* of the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$ as $U(\mathcal{H}|\mathcal{R}) = L\left(\overline{\mathcal{H}}|\mathcal{R}\right)$, where $\overline{\mathcal{H}}$ is the complement of $\mathcal{H}$. The likeliness and unlikeliness of a hypothesis are combined into a single measure of evidence in Section 2.4.

2.3 Marginal and conditional weight of evidence

Suppose $\mathcal{H}_1, \mathcal{H}_2 \in \mathfrak{H}$. According to the general law of likelihood Bickel (2012), the *weight of evidence* in the observation that $X = x$ substantiating the hypothesis that $\theta \in \mathcal{H}_1$ as

opposed to the hypothesis that $\theta \in \mathcal{H}_2$ is

$$W\left(\mathcal{H}_1; \mathcal{H}_2\right) = \frac{\sup_{\theta \in \mathcal{H}_1} f_\theta\left(x\right)}{\sup_{\theta \in \mathcal{H}_2} f_\theta\left(x\right)}. \tag{3}$$

That will be called the *marginal weight of evidence* to distinguish it from the conditional weight of evidence, defined below. Replacing $f_\bullet\left(x\right)$ with a profile likelihood function yields the quantity considered by Zhang and Zhang (2013a), as discussed in Bickel (2013b).

The *conditional weight of evidence* in the observation that $X = x$ substantiating the hypothesis that $\theta \in \mathcal{H}_1$ as opposed to the hypothesis that $\theta \in \mathcal{H}_2$ given $\theta \in \mathcal{R}$ is

$$W\left(\mathcal{H}_1; \mathcal{H}_2 | \mathcal{R}\right) = W\left(\mathcal{H}_1 \cap \mathcal{R}; \mathcal{H}_2 \cap \mathcal{R}\right) \tag{4}$$

for all $\mathcal{H}_1, \mathcal{H}_2, \mathcal{R} \in \mathfrak{H}$ such that $L\left(\mathcal{R}\right) > 0$. This is connected to the likeliness of Section 2.2 as follows.

Theorem 1. *For any* $\mathcal{H}_1, \mathcal{H}_2, \mathcal{R} \in \mathfrak{H}$*,*

$$W\left(\mathcal{H}_1; \mathcal{H}_2 | \mathcal{R}\right) = \frac{L\left(\mathcal{H}_1 | \mathcal{R}\right)}{L\left(\mathcal{H}_2 | \mathcal{R}\right)}. \tag{5}$$

For any set $\mathfrak{H}_0 \subset \mathfrak{H}$ *such that* $\bigcup_{\mathcal{H}_0 \in \mathfrak{H}_0} \mathcal{H}_0 = \mathcal{H}$*,*

$$L\left(\mathcal{H} | \mathcal{R}\right) = \frac{\sup_{\theta \in \mathcal{H} \cap \mathcal{R}} f_\theta\left(x\right)}{\sup_{\theta \in \mathcal{R}} f_\theta\left(x\right)} = \sup_{\mathcal{H}_0 \in \mathfrak{H}_0} L\left(\mathcal{H}_0 | \mathcal{R}\right). \tag{6}$$

For any partition $\mathfrak{P} \subset \mathfrak{H}$ *of* Θ*,*

$$L\left(\mathcal{H}\right) = \sup_{\mathcal{R} \in \mathfrak{P}} L\left(\mathcal{R}\right) L\left(\mathcal{H} | \mathcal{R}\right) \tag{7}$$

Proof. By equations (1), (4), and (3),

$$W\left(\mathcal{H}_1;\mathcal{H}_2|\mathcal{R}\right)=\frac{\sup_{\theta\in\mathcal{H}_1\cap\mathcal{R}}f_\theta\left(x\right)}{\sup_{\theta\in\mathcal{H}_2\cap\mathcal{R}}f_\theta\left(x\right)}=\frac{\sup_{\theta\in\mathcal{H}_1\cap\mathcal{R}}f_\theta\left(x\right)/\sup_{\theta\in\mathcal{R}}f_\theta\left(x\right)}{\sup_{\theta\in\mathcal{H}_2\cap\mathcal{R}}f_\theta\left(x\right)/\sup_{\theta\in\mathcal{R}}f_\theta\left(x\right)},$$

which is the right-hand side of equation (5) according to equation (2). Equation (1) and (2) imply that

$$L\left(\mathcal{H}|\mathcal{R}\right)=L\left(\bigcup_{\mathcal{H}_0\in\mathfrak{H}_0}\mathcal{H}_0|\mathcal{R}\right)=\frac{\sup_{\mathcal{H}_0\in\mathfrak{H}_0}\sup_{\theta\in\mathcal{H}_0\cap\mathcal{R}}f_\theta\left(x\right)}{\sup_{\theta\in\mathcal{R}}f_\theta\left(x\right)}=\sup_{\mathcal{H}_0\in\mathfrak{H}_0}\frac{\sup_{\theta\in\mathcal{H}_0\cap\mathcal{R}}f_\theta\left(x\right)}{\sup_{\theta\in\mathcal{R}}f_\theta\left(x\right)},$$

yielding $L\left(\mathcal{H}|\mathcal{R}\right)=\sup_{\mathcal{H}_0\in\mathfrak{H}_0}L\left(\mathcal{H}_0|\mathcal{R}\right)$. The other portion of formula (6) is established by substituting $\{\{\theta\}:\theta\in\mathcal{H}\}$ for $\mathfrak{H}_0$. Since $\mathfrak{P}\subset\mathfrak{H}$ is a partition,

$$L\left(\mathcal{H}\right)=L\left(\mathcal{H}\cap\Theta\right)=L\left(\mathcal{H}\cap\bigcup_{\mathcal{R}\in\mathfrak{P}}\mathcal{R}\right)=L\left(\bigcup_{\mathcal{R}\in\mathfrak{P}}\left(\mathcal{H}\cap\mathcal{R}\right)\right)=L\left(\bigcup_{\mathcal{R}_0\in\mathfrak{P}\left(\mathcal{H}\right)}\mathcal{R}_0\right),$$

where $\mathfrak{P}\left(\mathcal{H}\right)=\{\mathcal{R}\in\mathfrak{P}:\mathcal{R}\subseteq\mathcal{H}\}$. Thus, using equation (6),

$$L\left(\mathcal{H}\right)=\sup_{\mathcal{R}_0\in\mathfrak{P}\left(\mathcal{H}\right)}L\left(\mathcal{R}_0\right)=\sup_{\mathcal{R}\in\mathfrak{P}}L\left(\mathcal{H}\cap\mathcal{R}\right)=\sup_{\mathcal{R}\in\mathfrak{P}}L\left(\mathcal{R}\right)L\left(\mathcal{H}|\mathcal{R}\right),$$

with the last equality following from equation (2). □

2.4 Absolute weight of evidence

The strength of evidence favoring the hypothesis that $\theta\in\mathcal{H}$ can also be quantified without explicit reference to a second hypothesis by taking that second hypothesis to be its negation, $\theta\notin\mathcal{H}$. The *conditional weight of evidence* in the observation that $X=x$ substantiating the hypothesis that $\theta\in\mathcal{H}$ given $\theta\in\mathcal{R}$ is $W\left(\mathcal{H}|\mathcal{R}\right)=W\left(\mathcal{H};\overline{\mathcal{H}}|\mathcal{R}\right)$. Likewise, the *marginal*

weight of evidence in the observation that $X = x$ substantiating the hypothesis that $\theta \in \mathcal{H}$ is $W(\mathcal{H}) = W(\mathcal{H}|\Theta)$.

Corollary 1. *For any* $\mathcal{H}, \mathcal{R} \in \mathfrak{H}$,

$$W(\mathcal{H}|\mathcal{R}) = \frac{L(\mathcal{H}|\mathcal{R})}{U(\mathcal{H}|\mathcal{R})} = \frac{L(\mathcal{H} \cap \mathcal{R})}{L(\mathcal{R} \backslash \mathcal{H})} = \frac{\sup_{\theta \in \mathcal{H} \cap \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{R} \backslash \mathcal{H}} f_\theta(x)} \tag{8}$$

$$W(\mathcal{H}|\mathcal{R}) = \frac{\sup_{\mathcal{R} \in \mathfrak{P}} L(\mathcal{R})\, L(\mathcal{H}|\mathcal{R})}{\sup_{\mathcal{R} \in \mathfrak{P}} L(\mathcal{R})\, L\left(\overline{\mathcal{H}}|\mathcal{R}\right)}. \tag{9}$$

Proof. The claims follow directly from $U(\mathcal{H}|\mathcal{R}) = L\left(\overline{\mathcal{H}}|\mathcal{R}\right)$ and from equations (1), (5), and (7). □

Equation (8) indicates that $W(\mathcal{H}|\mathcal{R})$ is a coherent measure of support in the sense to be defined in Section 3. As will be seen, that property helps explain why it is appropriate to call $W(\mathcal{H}|\mathcal{R})$ the weight of evidence.

2.5 Likeliness and unlikeliness from the weight of evidence

While the weight of evidence is the ratio of likeliness to the unlikeliness (8), it is convenient in some applications to derive the likeliness and unlikeliness from the weight of evidence.

Lemma 1. *Given* $\mathcal{H}, \mathcal{R} \in \mathfrak{H}$, *it follows that* $L(\mathcal{H}|\mathcal{R}) = 1$ *and* $U(\mathcal{H}|\mathcal{R}) = 1/W(\mathcal{H}|\mathcal{R})$ *if* $W(\mathcal{H}|\mathcal{R}) \geq 1$ *but that* $L(\mathcal{H}|\mathcal{R}) = W(\mathcal{H}|\mathcal{R})$ *and* $U(\mathcal{H}|\mathcal{R}) = 1$ *if* $W(\mathcal{H}|\mathcal{R}) < 1$.

Proof. In the $W(\mathcal{H}|\mathcal{R}) \geq 1$ case, equation (8) implies that $\sup_{\theta \in \mathcal{H} \cap \mathcal{R}} f_\theta(x) \geq \sup_{\theta \in \mathcal{R} \backslash \mathcal{H}} f_\theta(x)$ and thus that $\sup_{\theta \in \mathcal{H} \cap \mathcal{R}} f_\theta(x) = \sup_{\theta \in \mathcal{R}} f_\theta(x)$. By equation (6), $L(\mathcal{H}|\mathcal{R}) = 1$. Therefore, equation (8) 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 generalization of this result is known both from possibility theory, in terms of which $L(\bullet|\mathcal{R})$ is a possibility measure and $1 - U(\bullet|\mathcal{R})$ is a necessity measure, and from the theory of ranking functions (Spohn, 2012, §5.2), in terms of which $\log W(\bullet|\mathcal{R})$ is a two-sided ranking function, where log is of a base greater than 1.

3 Derivation from coherence and Bayes compatibility

Let P stand for a probability measure on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$ and a random parameter ϑ of prior distribution $P_0 = P(\bullet \times \mathcal{X})$ on $(\Theta, \mathfrak{H})$ such that the posterior probability that $\vartheta \in \mathcal{H}$ is

$$P(\vartheta \in \mathcal{H}|x) = \frac{P_0(\vartheta \in \mathcal{H}) \int_{\mathcal{H}} f_\theta(x)\, dP_0(\theta|\mathcal{H})}{\int f_\theta(x)\, dP_0(\theta)}.$$

This is considered a function of P such that, if Q were the joint distribution on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$, the posterior distribution would be $Q(\vartheta \in \mathcal{H}|x)$ with Q in place of P and $Q_0 = Q(\bullet \times \mathcal{X})$ in place of P_0. The *increase in the odds ratio* due to the observation that $X = x$ in favor of the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$ is the ratio of the conditional posterior odds to the conditional prior odds:

$$\Delta(\mathcal{H}; P|\mathcal{R}) = \frac{P(\vartheta \in \mathcal{H}|x, \mathcal{R}) / P(\vartheta \notin \mathcal{H}|x, \mathcal{R})}{P_0(\vartheta \in \mathcal{H}|\mathcal{R}) / P_0(\vartheta \notin \mathcal{H}|\mathcal{R})} = \frac{\int_{\mathcal{H}\cap\mathcal{R}} f_\theta(x)\, dP_0(\theta|\mathcal{H}, \mathcal{R})}{\int_{\mathcal{R}\setminus\mathcal{H}} f_\theta(x)\, dP_0\left(\theta|\overline{\mathcal{H}}, \mathcal{R}\right)}. \tag{10}$$

The conditional Bayes factor $B(\mathcal{H}|\mathcal{R})$ as a function of $\mathcal{H}$ and $\mathcal{R}$, is defined such that $B(\mathcal{H}|\mathcal{R}) = \Delta(\mathcal{H}; P|\mathcal{R})$ for some fixed probability measure P on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$.

The requirement of Edwards (1992) that a measure of support for one hypothesis over

another be compatible with Bayes's theorem is generalized to composite hypotheses by the following definition, differing from the generalization that often forbids accepting a hypothesis of sufficiently high weight of evidence (Bickel, 2013a,b). Any function $u : \mathfrak{H}^2 \to [0, \infty]$ measures the *odds ratio increase* due to the observation that $X = x$ if there is a probability measure $P_{\mathcal{H}}$ on $(\Theta, \mathfrak{H})$ such that

$$u\left(\mathcal{H}|\mathcal{R}\right) = \Delta\left(\mathcal{H}; P_{\mathcal{H}}|\mathcal{R}\right) \tag{11}$$

for all $\mathcal{H} \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$ satisfying $L\left(\mathcal{R}\right) > 0$.

Another desirable property of a measure of evidence is the avoidance of asserting that contradictory statements are individually supported by the evidence (Schervish, 1996; Lavine and Schervish, 1999; Bickel, 2012; Zhang and Zhang, 2013a). More formally, a function $v : \mathfrak{H}^2 \to [0, \infty]$ is *logically coherent* if

$$v\left(\mathcal{H}_0|\mathcal{R}\right) \leq v\left(\mathcal{H}_1|\mathcal{R}\right) \iff \left(\theta \in \mathcal{H}_0 \cap \mathcal{R} \implies \theta \in \mathcal{H}_1 \cap \mathcal{R}\right) \tag{12}$$

for all $\mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$. The fact that conclusions may be drawn on the basis of logically coherent evidence leads to the following definition and theorem.

Definition 1. A function $w : \mathfrak{H}^2 \to [0, \infty]$ *measures the strength of conclusive evidence* if it both measures the odds ratio increase and is logically coherent.

Theorem 2. *A function $w : \mathfrak{H}^2 \to [0, \infty]$ measures the strength of conclusive evidence if and only if it is the weight of evidence function $W\left(\bullet|\bullet\right)$.*

Proof. ($\Longleftarrow$). The following statements apply for all $\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$. Let $\delta\left(\bullet; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right)$ and $\delta\left(\bullet; \widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}}\right)$ denote the Dirac probability measures on $(\Theta, \mathfrak{H})$ with mass at

$\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}} = \arg\sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x)$ and $\widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}} = \arg\sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)$, respectively. By equation (8),

$$W(\mathcal{H}|\mathcal{R}) = \frac{\int f_\theta(x)\, d\delta\left(\theta; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right)}{\int f_\theta(x)\, d\delta\left(\theta; \widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}}\right)}. \tag{13}$$

There is a probability measure P on $(\Theta\times\mathcal{X}, \mathfrak{H}\otimes\mathfrak{X})$ such that $P_0(\bullet|\mathcal{H},\mathcal{R}) = \delta\left(\bullet; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right)$ and $P_0\left(\bullet|\overline{\mathcal{H}},\mathcal{R}\right) = \delta\left(\bullet; \widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}}\right)$, in which case equations (13) and (10) imply that $W(\mathcal{H}|\mathcal{R}) = \Delta(\mathcal{H}; P|\mathcal{R})$. Thus, $W(\bullet|\bullet)$ measures the odds ratio increase. The fact that $W(\mathcal{H}_0,\mathcal{R}) \leq W(\mathcal{H}_1,\mathcal{R})$ if and only if $\mathcal{H}_0\subseteq\mathcal{H}_1$ demonstrates equation (12). Therefore, $W(\bullet|\bullet)$ is logically coherent. Thus, both criteria of Definition 1 are satisfied.($\Longrightarrow$). Let $w : \mathfrak{H}^2 \to [0,\infty]$ denote a function that measures the strength of conclusive evidence. By Definition 1, w both measures the odds ratio increase and is logically coherent. Assume that there are $\mathcal{H}\in\mathfrak{H}$ and $\mathcal{R}\in\mathfrak{H}$ and, contrary to the $w = W$ claim and equation (8), such that

$$w(\mathcal{H}|\mathcal{R}) \neq \frac{\sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x)}{\sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)} \tag{14}$$

in order to prove the claim by contradiction. Since $w = v$, equations (10), (11), and (12) yield

$$\int_{\mathcal{H}_0\cap\mathcal{R}} f_\theta(x)\, dP_0(\theta|\mathcal{H}_0,\mathcal{R}) \leq \int_{\mathcal{H}_1\cap\mathcal{R}} f_\theta(x)\, dP_0(\theta|\mathcal{H}_1,\mathcal{R}) \iff \mathcal{H}_0\subseteq\mathcal{H}_1.$$

Since $\left\{\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right\} \subseteq \mathcal{H}\cap\mathcal{R}$,

$$\begin{aligned}\int_{\mathcal{H}\cap\mathcal{R}} f_\theta(x)\, dP_0(\theta|\mathcal{H},\mathcal{R}) &\geq \int_{\left\{\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right\}} f_\theta(x)\, dP_0\left(\theta|\left\{\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right\},\mathcal{R}\right)\\ &= \int f_\theta(x)\, d\delta\left(\theta; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right) = \sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x),\end{aligned}$$

but that requires that $\int_{\mathcal{H}\cap\mathcal{R}} f_\theta(x)\, dP_0(\theta|\mathcal{H},\mathcal{R}) = \sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x)$ (cf. Coletti et al., 2009). Analogous reasoning leads to $\int_{\mathcal{R}\setminus\mathcal{H}} f_\theta(x)\, dP_0(\theta|\mathcal{H},\mathcal{R}) = \sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)$. Thus, equations (10), (11), and (12) establish equation (8), contradicting equation (14), thereby proving the $w = W$ claim. □

Theorem 2 says the weight of evidence uniquely measures the strength of conclusive evidence.

That raises questions about the senses in which the posterior probability and the Bayes factor fall short as a measures of the strength of conclusive evidence. Lavine and Schervish (1999) demonstrated that the posterior probability but not the Bayes factor is coherent as a measure of evidence. This is restated in the following corollaries in addition to whether each measures the odds ratio increase.

Corollary 2. *Given any probability measure P on $(\Theta\times\mathcal{X}, \mathfrak{H}\otimes\mathfrak{X})$ satisfying the above conditions, the conditional Bayes factor function B measures the odds ratio increase but is not necessarily logically coherent.*

Proof. By the definition of the conditional Bayes factor $B(\mathcal{H}|\mathcal{R}) = \Delta(\mathcal{H}; P|\mathcal{R})$. Thus, $u = B$ yields equation (11), establishing the first claim. The second claim is established by noting that, according to Theorem 2, B is only logically coherent in the special case that $B(\mathcal{H}|\mathcal{R}) = W(\mathcal{H}|\mathcal{R})$ for all $\mathcal{H}\in\mathfrak{H}$ and all $\mathcal{R}\in\mathfrak{H}$. □

Corollary 3. *Given any probability measure P on $(\Theta\times\mathcal{X}, \mathfrak{H}\otimes\mathfrak{X})$ satisfying the above conditions, the posterior probability function $P(\vartheta\in\bullet|x,\bullet)$ on $\mathfrak{H}^2$ is logically coherent but does not necessarily measure the odds ratio increase.*

Proof. Consider an $\mathcal{R}\in\mathfrak{H}$ that satisfies $L(\mathcal{R}) > 0$ and $\mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ such that $\mathcal{H}_0 \subseteq \mathcal{H}_1$. By the additivity of probability measures, $P(\vartheta\in\mathcal{H}_0|x,\mathcal{R}) \leq P(\vartheta\in\mathcal{H}_1|x,\mathcal{R})$. Likewise, any

$\mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ such that $P(\vartheta \in \mathcal{H}_0|x, \mathcal{R}) \leq P(\vartheta \in \mathcal{H}_1|x, \mathcal{R})$ are related by $\mathcal{H}_0 \subseteq \mathcal{H}_1$. Thus, $v = P(\vartheta \in \bullet|x, \bullet)$ yields equation (12), establishing the first claim. The second claim is established by noting that, according to Theorem 2, $P(\vartheta \in \bullet|x, \bullet)$ only measures the odds ratio increase in the special case that $P(\vartheta \in \mathcal{H}|x, \mathcal{R}) = W(\mathcal{H}|\mathcal{R})$ for all $\mathcal{H} \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$ satisfying $L(\mathcal{R}) > 0$. □

Lavine and Schervish (1999) likewise argued that the posterior probability is coherent as a measure of evidence. In short, while the Bayes factor measures the odds ratio increase and the posterior probability is logically coherent, the weight of evidence is the only quantity with both properties.

4 Fusing models from different levels

The procedure chosen for eliminating nuisance parameters may depend on criteria outside the statistical model (Bickel, 2012). Indeed, paradoxes arise from likelihood inference under inappropriate models of the system of interest (Lindsey, 1996, §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, multiple levels of abstraction may merit simultaneous consideration. 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 σ-field of subsets of Φ satisfying $\mathfrak{I} = \{\{\varphi(\theta) : \theta \in \mathcal{H}\} : \mathcal{H} \in \mathfrak{H}\}$. 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 variable Z of observed value z. The pair $(\{f_\theta : \theta \in \Theta\}, \{g_\phi : \phi \in \Phi\})$ is called a *fusion* of the two parametric models, the density families $\{f_\theta : \theta \in \Theta\}$ and $\{g_\phi : \phi \in \Phi\}$. A model used to quantify the weight of evidence without any fusion (§2) is a *pure model*.

The function $L^f : \mathfrak{H} \times \mathfrak{H} \to [0, \infty]$ is defined in accordance with equations (1)-(2) 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 (1), 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 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

$$L^{fg}(\mathcal{H}, \mathcal{I}|\mathcal{S}) = L^f\left(\mathcal{H}|\varphi^{-1}(\mathcal{I} \cap \mathcal{S})\right) L^g(\mathcal{I}|\mathcal{S}), \tag{15}$$

which reduces to $L^f(\mathcal{H} \cap \mathcal{I}|\mathcal{S})$ according to equation (2) 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 (15) degenerates to

$$L^{fg}(\mathcal{H}, \mathcal{I}) = L^f\left(\mathcal{H}|\varphi^{-1}(\mathcal{I})\right) L^g(\mathcal{I}). \tag{16}$$

Likewise generalizing equation (7), let

$$L^{fg}(\mathcal{H}|\mathcal{S}) = \sup_{\mathcal{I} \in \mathfrak{Q}} L^{fg}(\mathcal{H}, \mathcal{I}|\mathcal{S}) \tag{17}$$

for any partition $\mathfrak{Q} \subset \mathfrak{I}$ of Φ, and let $L^{fg}(\mathcal{H}) = L^{fg}(\mathcal{H}|\Phi)$.

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

$$W^{fg}\left(\mathcal{H}_1;\mathcal{H}_2|\mathcal{S}\right)=\frac{L^{fg}\left(\mathcal{H}_1|\mathcal{S}\right)}{L^{fg}\left(\mathcal{H}_2|\mathcal{S}\right)}, \tag{18}$$

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

$$W^{fg}\left(\mathcal{H}|\mathcal{S}\right)=W^{fg}\left(\mathcal{H};\overline{\mathcal{H}}|\mathcal{S}\right) \tag{19}$$

for all $\mathcal{H},\mathcal{H}_1,\mathcal{H}_2\in\mathfrak{H}$ and $\mathcal{S}\in\mathfrak{I}$. The marginal counterparts are $W^{fg}\left(\bullet,\bullet\right)=W^{fg}\left(\bullet,\bullet|\Phi\right)$ and $W^{fg}\left(\bullet\right)=W^{fg}\left(\bullet|\Phi\right)$. Analogously, W^{f} is identified with the functions denoted by W in Sections 2. Thus, $W^{fg}\left(\bullet,\bullet|\mathcal{S}\right)=W^{f}\left(\bullet,\bullet|\varphi^{-1}\left(\mathcal{S}\right)\right)$, $W^{fg}\left(\bullet,\bullet\right)=W^{f}\left(\bullet,\bullet\right)$, etc.

Theorem 3. *For any partition $\mathfrak{Q}\subset\mathfrak{I}$ of Φ and any $\mathcal{H}\in\mathfrak{H}$,*

$$L^{fg}\left(\mathcal{H}\right)=\sup_{\mathcal{I}\in\mathfrak{Q}}L^{f}\left(\mathcal{H}|\varphi^{-1}\left(\mathcal{I}\right)\right)L^{g}\left(\mathcal{I}\right); \tag{20}$$

$$W^{fg}\left(\mathcal{H}\right)=\frac{\sup_{\mathcal{I}\in\mathfrak{Q}}L^{f}\left(\mathcal{H}|\varphi^{-1}\left(\mathcal{I}\right)\right)L^{g}\left(\mathcal{I}\right)}{\sup_{\mathcal{I}\in\mathfrak{Q}}L^{f}\left(\overline{\mathcal{H}}|\varphi^{-1}\left(\mathcal{I}\right)\right)L^{g}\left(\mathcal{I}\right)}. \tag{21}$$

Proof. Equation (17) yields

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

which, with equation (16), in turn yields equation (20). Equation (21) follows immediately from equations (18)-(19).

□

Analogous properties may be proven conditional on $\varphi\left(\theta\right)\in\mathcal{S}$.

5 Multiple-hypothesis weights of evidence

5.1 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 $i = 1, \ldots, N$. Such data may be a function of the original observations, as when each x_i is reduced to a test statistic. They are modeled as realizations of random variables, 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 set Θ and N unknown parameters in Θ, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \in \Theta^N$. The generalized 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 $f_\bullet(x_1, \ldots, x_N) : \Theta^N \to [0, \infty[$. With that slight change of the notation in Section 2.1, hypotheses about $\boldsymbol{\theta}$ correspond to members of $\mathfrak{H}$, the σ-field of subsets of Θ^N.

For $i = 1, \ldots, N$ and for some $\theta_0 \in \Theta$, the ith null hypothesis is that $\theta_i = \theta_0$. 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 $\widehat{\theta}(x_i) = \arg\sup_{\theta\in\Theta} f_\theta(x_i)$, leading to the abbreviation $L_i^{\max} = f_{\widehat{\theta}(x_i)}(x_i) / f_{\theta_0}(x_i)$ for every $i = 1, \ldots, N$. It is assumed for simplicity that $L_i^{\max} > 1$, as occurs with probability 1 for continuous models.

5.2 Pure model for multiple-hypothesis evidence

Again consider some restriction set $\boldsymbol{\mathcal{R}} \in \mathfrak{H}$. According to equation (8), 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

$$W\left(\left\{\theta_j \neq \theta_0 : j \in \mathcal{T}\right\} | \boldsymbol{\mathcal{R}}\right) \equiv W\left(\boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}} | \boldsymbol{\mathcal{R}}\right) = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}} \cap \boldsymbol{\mathcal{R}}} f_{\boldsymbol{\theta}}(x)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}} \setminus \boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}}} f_{\boldsymbol{\theta}}(x)}; \tag{22}$$

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

(cf. Bickel (2012, §2.4.1)). 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. *Suppose that T_i is independent of T_j for all $i \neq j$. It follows that the joint weight of evidence substantiating the joint alternative hypotheses that $\theta_j \neq \theta_0$ for all $j \in \mathcal{T}$, conditional on $\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}}$, is*

$$W\left(\boldsymbol{\mathcal{H}}_{\bigwedge \mathcal{T}} | \boldsymbol{\mathcal{R}}\right) = \min_{j \in \mathcal{T}} L_j^{\max} \tag{23}$$

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

$$W\left(\boldsymbol{\mathcal{H}}_{\bigvee \mathcal{T}} | \boldsymbol{\mathcal{R}}\right) = \prod_{j \in \mathcal{T}} L_j^{\max}. \tag{24}$$

Proof. Let $J(\mathcal{T}) = \arg\min_{j \in \mathcal{T}} L_j^{\max}$. By equation (22),

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

and cancellation of the products yields $W\left(\boldsymbol{\mathcal{H}}_{\bigwedge\mathcal{T}}|\boldsymbol{\mathcal{R}}\right) = f_{\widehat{\theta}\left(x_{J(\mathcal{T})}\right)}\left(x_{J(\mathcal{T})}\right)/f_{\theta_0}\left(x_{J(\mathcal{T})}\right)$ and thus equation (23). Similarly,

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

leading to equation (24) by cancellation. □

5.3 Fusing models for multiple-hypothesis evidence

5.3.1 Marginalization over how many null hypotheses are true

A practical application of the formalism of Section 4 is marginalization over how many null hypotheses are true according to two models, which correspond to different levels of abstraction. The number of true null hypotheses is $\sum_{i=1}^{N} 1_0\left(\theta_i\right)$, where $1_0\left(\theta_i\right) = 1$ if $\theta_i = \theta_0$ and $1_0\left(\theta_i\right) = 0$ 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 hypotheses, is

$$W\left(\left\{\theta_j \neq \theta_0 : j \in \mathcal{T}\right\} \mid \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right) \equiv W\left(\boldsymbol{\mathcal{H}}_{\bigwedge\mathcal{T}}|\boldsymbol{\mathcal{R}}_{N_0}\right) = \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\mathcal{H}}_{\bigwedge\mathcal{T}}\cap\boldsymbol{\mathcal{R}}_{N_0}} f_{\boldsymbol{\theta}}\left(x\right)}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\mathcal{R}}_{N_0}\setminus\boldsymbol{\mathcal{H}}_{\bigwedge\mathcal{T}}} f_{\boldsymbol{\theta}}\left(x\right)}; \quad (25)$$

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

In the case of continuous density functions, equation (25) 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 5. *Suppose $f_{\boldsymbol{\theta}}$ is the Radon-Nikodym derivative of the distribution of $(T_1, \ldots, T_N)$ that corresponds to $\boldsymbol{\theta}$, with respect to the Lebesgue measure, and that T_i is independent of T_j for all $i \neq j$. 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 true null hypotheses, is*

$$W^f\left(\boldsymbol{\mathcal{H}}_{\bigwedge\mathcal{T}}|\sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right) = \begin{cases} \frac{\min_{j\in\mathcal{T}} L_j^{\max}}{L_{(N_0)}^{\max}} & \text{if } 1 \leq N_0 \leq N - M \\ 0 & \text{if } N_0 > N - M \\ \infty & \text{if } N_0 = 0 \end{cases}. \tag{26}$$

Proof. The order statistics are unambiguously defined since the absolute continuity of the distribution of each T_i with respect to the Lebesgue measure guarantees that there are almost surely no ties. With $J\left(\mathcal{T}\right) = \arg\min_{j\in\mathcal{T}} L_j^{\max}$, 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 $f_{\widehat{\theta}\left(x_{(k)}\right)}\left(x_{(k)}\right) / f_{\theta_0}\left(x_{(k)}\right) = L_{(k)}^{\max}$. In the case that $1 \leq N_0 = N - M$, equation (25) simplifies to

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

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

and cancelation wherever possible yields

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

In the case that $1 \leq N_0 < N - M$, equation (25) simplifies with cancellation to

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

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

Both cases together establish the result for $1 \leq N_0 \leq N - M$. In the case that $N_0 > N - M$, the numerator in equation (25) 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 (25) is 0 since $\boldsymbol{\mathcal{R}}_0 \backslash \boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}} = \emptyset$. □

With the weight of evidence from equation (26), Lemma 1 gives $L^f\left(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}} | \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right)$. While the expression for $W^f\left(\boldsymbol{\mathcal{H}}_{\vee\mathcal{T}} | \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right)$ cannot be written as concisely as that of $W^f\left(\boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}} | \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right)$, it is easily implemented numerically, providing

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

via Lemma 1.

The model $\left\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta^N\right\}$ is used with another model, $\{g_\phi : \phi \in \boldsymbol{\Phi}\}$, to obtain $L^{fg}\left(\boldsymbol{\mathcal{H}}\right)$ and $W^{fg}\left(\boldsymbol{\mathcal{H}}\right)$ for some hypothesis that $\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}$, such as $\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\wedge\mathcal{T}}$ or $\boldsymbol{\theta} \in \boldsymbol{\mathcal{H}}_{\vee\mathcal{T}}$, according to Section 4. The parameters are related by a function $\varphi : \Theta^N \to \boldsymbol{\Phi}$ that transforms $\boldsymbol{\theta}$ to $\varphi\left(\boldsymbol{\theta}\right) = \varphi\left(\left(\theta_1, \ldots, \theta_N\right)\right)$ for all $\boldsymbol{\theta} \in \Theta^N$. Under the second model, the number of true null hypotheses is $\nu_0\left(\phi\right)$, where ν_0 is the function such that $\nu_0\left(\varphi\left(\left(\theta_1, \ldots, \theta_N\right)\right)\right) = \sum_{i=1}^{N} 1_0\left(\theta_i\right)$ for all $\boldsymbol{\theta} \in \Theta^N$.

Corollary 4. *For any* $\boldsymbol{\mathcal{H}} \in \mathfrak{H}$ *and* $N_0 = 0, 1, \ldots, N$,

$$L^{fg}\left(\boldsymbol{\mathcal{H}}\right) = \max_{N_0=0,1,\ldots,N} L^f\left(\boldsymbol{\mathcal{H}} | \sum_{i=1}^{N} 1_0\left(\theta_i\right) = N_0\right) L^g\left(\left\{\phi \in \boldsymbol{\Phi} : \nu_0\left(\phi\right) = N_0\right\}\right). \tag{27}$$

Proof. With $\theta = \boldsymbol{\theta}$, $\mathcal{H} = \boldsymbol{\mathcal{H}}$, and

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

the claim results from Theorem 3 since $\mathfrak{Q}$ is a member of $\mathfrak{I}$, a partition of $\boldsymbol{\Phi}$, and since

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

for all $N_0 = 0, 1, \ldots, N$. □

The density family $\{g_\phi : \phi \in \boldsymbol{\Phi}\}$ could be, for example, the model of Section 5.3.2. The model of Section 5.3.3 may be similarly applied via an approximation.

5.3.2 Fusion with a non-mixture model

A non-mixture model will be built from a density family $\{g_{\phi'} : \phi' \in \Phi\}$. Suppose that $\Phi = \{\phi\left(0\right), \phi\left(1\right), \ldots, \phi\left(K-1\right)\}$, that $\boldsymbol{\Phi} = \Phi^N$ is a parameter set of some finite cardinality $K \geq 2$, that $\boldsymbol{\mathfrak{I}} = 2^{\boldsymbol{\Phi}}$, and that

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

for all $(\phi_1, \phi_2, \ldots, \phi_N) \in \Phi^N$, with $\phi_i = \phi\left(0\right)$ corresponding to the ith null hypothesis. Padilla and Bickel (2012) and Yang et al. (2013a) 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 $\left\{g_\phi : \phi \in \Phi^N\right\}$ model is misspecified in the sense that either K or $\max_{k=0,1,\ldots,K-1} \phi(k)$ is not large enough, for in either 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 (see Yang et al., 2013b).

A more robust approach is available in the fusion of the two models, enabling the use of equation (27) with the next result.

Lemma 2. *For all* $N_0 = 0, 1, \ldots, N$,

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

Proof. By Theorem 1, □

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

Example 1. Following a general approach in Padilla and Bickel (2012), Yang et al. (2013a) 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 if $1_{\{0\}}(\phi_i) = 1$ if $\phi_i = \phi_{\text{alt}}$. According to Lemma 2,

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

with $\vee$ denoting the maximum.

5.3.3 Fusion with a mixture model

In a K-component mixture model formulated to obtain maximum likelihood estimates of false discovery rates (Pawitan et al., 2005; Muralidharan, 2010; Bickel, 2014b),

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

where $\pi_k \in [0,1]$ for each $k \in 0,1,\ldots,K$ such that $\sum_{k=0}^{K-1} \pi_k = 1$ and where each $g_{\phi(k)}$ with $\phi(k) \in \{\phi(0), \phi(1), \ldots, \phi(K-1)\}$ is a probability density function according to the family in Section 5.3.2. Thus, $(\pi_k, \phi(k)) \in [0,1] \times \Phi = [0,1] \times \{\phi(0), \phi(1), \ldots, \phi(K-1)\}$ for each k, and $\mathbf{\Phi} = ([0,1] \times \Phi)^K$, unlike the parameter space of Section 5.3.2. For all matrices $\phi = ((\pi_0, \phi(0)), (\pi_1, \phi(1)), \ldots, (\pi_{K-1}, \phi(K-1))) \in ([0,1] \times \Phi)^K$,

$$g_{\phi \in ([0,1] \times \Phi)^K}(z) = \prod_{i=1}^{N} g_\phi(z_i).$$

Without model fusion, the weight of evidence for the ith alternative hypothesis has been quantified as $\left(\sum_{k=1}^{K-1} \widehat{\pi}_k g_{\widehat{\phi}(k)}(z_i) / \sum_{k=1}^{K-1} \widehat{\pi}_k\right) / g_{\widehat{\phi}(0)}(z_i)$, where $\left(\widehat{\pi}_k, \widehat{\phi}(k)\right)$ is the maximum likelihood estimate of $(\pi_k, \phi(k))$ for each $k = 0, \ldots, K-1$ (Padilla and Bickel, 2012; Yang et al., 2013a). As this fails under marked misspecification in the same way as does using the non-mixture model of Section 5.3.2, 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 $k \in 0,1,\ldots,K$, as if π_0 were

the proportion of true null hypotheses. While that is not strictly correct, it is an adequate approximation for sufficiently large N (Bickel, 2014b). The fusion of the two models provides marginalization according to $\mathbf{\Phi} = (\mathcal{P}_N \times \Phi)^K$ and equation (27), as follows, with the proof analogous to that of Lemma 2.

Lemma 3. *For all* $N_0 = 0, 1, \ldots, N$,

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

Example 2. In the mixture-model equivalent of Example 1, $K = 2$, $\Phi = \{0, \phi_{\text{alt}}\}$, and g_0 and $g_{\phi_{\text{alt}}}$ are the same as before (Yang et al., 2013a). By Lemma 3,

$$L^g\left(\{\phi \in \{0, \phi_{\text{alt}}\} : \nu_0(\phi) = N_0\}\right) = \frac{\sup_{\phi_{\text{alt}} > 0} \prod_{i=1}^N \left(\frac{N_0}{N} g_0(z_i) + \left(1 - \frac{N_0}{N}\right) g_{\phi_{\text{alt}}}(z_i)\right)}{\sup_{\phi_{\text{alt}} > 0, \pi_0 \in \mathcal{P}_N} \prod_{i=1}^N \left(\pi_0 g_0(z_i) + (1 - \pi_0) g_{\phi_{\text{alt}}}(z_i)\right)}. \tag{28}$$

6 Application of multiple-hypothesis evidence

Rubin (1981) reports 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. The standardized mean differences are denoted by $x = (x_1, \ldots, x_8)$ and modeled 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 modeling are employed to highlight advantages of model fusion (§4). The approaches represent all exam sites together with a pure non-mixture model (§5.2), with

a fusion of two non-mixture models (§5.3.2), and with a fusion of a non-mixture model with a mixture model (§5.3.3). In the first approach, the weights of evidence considered are simply given by Theorem 4.

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 $z = (z_1, \ldots, z_8) = (x_1, \ldots, x_8)$, $K = 2$, $\Phi = \{0, \phi_{\text{alt}}\}$, g_0 as the standard normal density function, and $g_{\phi_{\text{alt}}}$ as the $\text{N}(\phi_{\text{alt}}, 1)$ density function with known mean $\phi_{\text{alt}} = 2$. The family $\left\{g_\phi : \phi \in \{0, \phi_{\text{alt}}\}^8\right\}$ has a non-mixture model version (§5.3.2) and a mixture model version (§5.3.3). For the former, Lemma 2 gives equation (1) as the likeliness function of N_0 with the density functions of this section in place of the χ^2 density functions of Example 1. For the latter, Lemma 3 yields

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

as the likeliness function of N_0 in contrast with that of equation (28), in which ϕ_{alt} is unknown.

Those two likeliness functions of N_0 are plotted in Figure 1. The three modeling approaches are compared in Figures 2, 3, and 4. Implications are discussed in Section 7.

7 Discussion

In Figure 1, the non-mixture model is seen to be more informative than the mixture model. Figures 2, 3, and 4 reflect that in the low weights of evidence assigned by the fusion with the mixture model. The overly conservative nature of the mixture model may be due in part to the fact that π_0 is a poor approximation to the proportion of null hypotheses that are

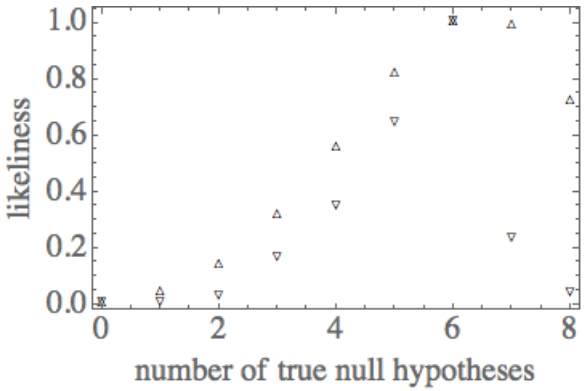

Figure 1: Likeliness as a function of N_0 under a non-mixture model ($\bigtriangledown$; §5.3.2) and a mixture model ($\bigtriangleup$; §5.3.3).

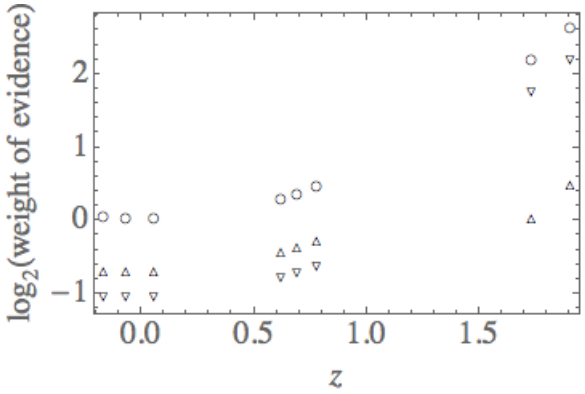

Figure 2: The weight of the evidence favoring the hypothesis that $\theta_i \neq 0$ for each $i = 1, \ldots, 8$ under a pure non-mixture model ($\bigcirc$; §5.2), under a fusion of two non-mixture models ($\bigtriangledown$; §5.3.2), and under a fusion of non-mixture model with a mixture model ($\bigtriangleup$; §5.3.3). Each fusion includes the model of Figure 1 with the corresponding symbol.

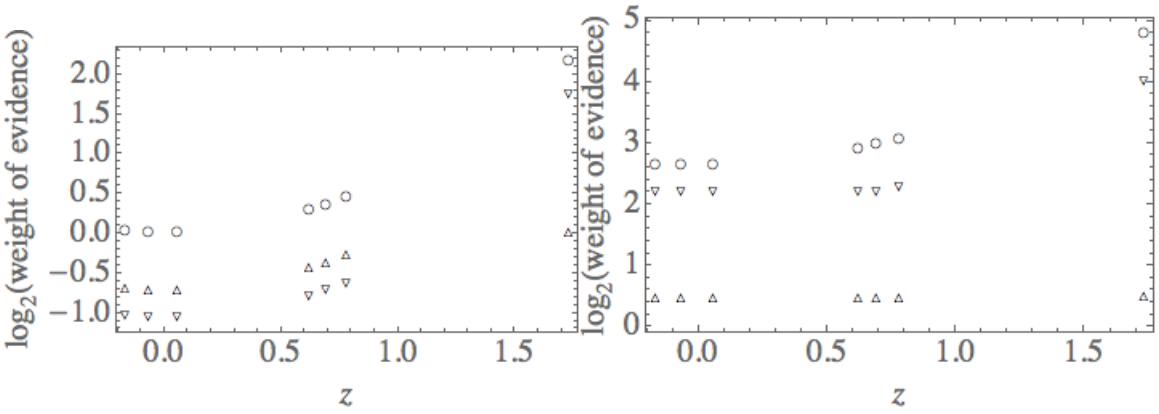

Figure 3: The weight of the evidence favoring the conjunctive hypothesis that $\theta_i \neq 0$ and $\theta_1 \neq 0$ (left) or favoring the disjunctive hypothesis that $\theta_i \neq 0$ or $\theta_1 \neq 0$ (right) for each $i = 2, \ldots, 8$. $\bigcirc\bigtriangledown\bigtriangleup$: caption of Figure 2.

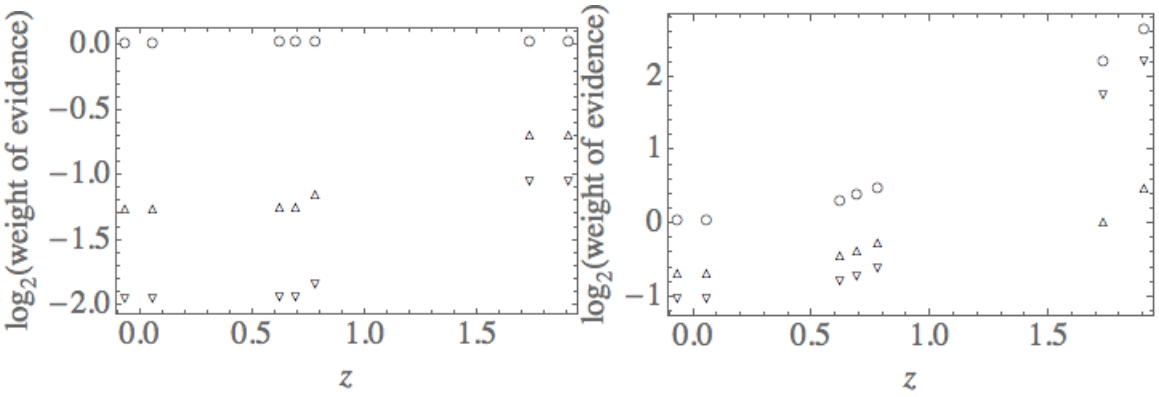

Figure 4: The weight of the evidence favoring the conjunctive hypothesis that $\theta_i \neq 0$ and $\theta_3 \neq 0$ (left) or favoring the disjunctive hypothesis that $\theta_i \neq 0$ or $\theta_3 \neq 0$ (right) for each $i = 1, 2, 4, \ldots, 8$. $\bigcirc\bigtriangledown\bigtriangleup$: caption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

true (§5.3.3). The mixture model may perform better than the non-mixture model if N is sufficiently large and if π_0 is close to 1 (Yang et al., 2013a).

Figures 2-4 also indicate that under the pure model approach, the evidence is weighed without any shrinkage toward the null hypotheses. Thus, the level-Λ *fused likeliness set* defined by $\left\{\theta \in \Theta : L^{fg}\left(\{\theta\}\,|\mathcal{S}\right) \geq 1/\Lambda\right\}$ exhibits shrinkage in set estimation not possible with the level-Λ likelihood set consided in Section 1.

An undesirable manifestation of pure-model's lack of shrinkage is that all the alternative hypotheses are seen in the plots to be favored over the null hypotheses ($\log_2 W\left(\mathcal{H}\right) > 0$), at least to a negligible extent (Table 1). In fact, this occurs more generally with probability 1 (Bickel, 2012). 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 2-4.

In conclusion, the fusion between two non-mixture models 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

This research was partially supported 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.

References

Barnard, G. A., 1967. The use of the likelihood function in statistical practice. Proc. 5th Berkeley Symp. on Math. Stat. Prob. Vol. I, 27–40.

Barndorff-Nielsen, O. E., 1994. Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. Journal of the Royal Statistical Society B 56, 125–140.

Bickel, D. R., 2011. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. Canadian Journal of Statistics 39, 610–631.

Bickel, D. R., 2012. The strength of statistical evidence for composite hypotheses: Inference to the best explanation. Statistica Sinica 22, 1147–1198.

Bickel, D. R., 2013a. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. International Statistical Review 81, 188–206.

Bickel, D. R., 2013b. Pseudo-likelihood, explanatory power, and Bayes's theorem [comment on "A likelihood paradigm for clinical trials"]. Journal of Statistical Theory and Practice 7, 178–182.

Bickel, D. R., 2014a. Inference after checking multiple Bayesian models for data conflict. Working Paper, University of Ottawa, deposited in uO Research at http://hdl.handle.net/10393/31135.

Bickel, D. R., 2014b. Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison. International Statistical Review 82, 457–476.

Bjørnstad, J. F., 1996. On the generalization of the likelihood function and the likelihood principle. Journal of the American Statistical Association 91, 791–806.

Blume, J., 2013. Likelihood and composite hypotheses [comment on "A likelihood paradigm for clinical trials"]. Journal of Statistical Theory and Practice 7 (2), 183–186.

Blume, J. D., 2011. Likelihood and its evidential framework. In: Bandyopadhyay, P. S., Forster, M. R. (Eds.), Philosophy of Statistics. North Holland, Amsterdam, pp. 493–512.

Coletti, G., Scozzafava, R., Vantaggi, B., 2009. Integrated likelihood in a finitely additive setting. In: Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Vol. 5590 of Lecture Notes in Comput. Sci. Springer, Berlin, pp. 554–565.

Dubois, D., Moral, S., Prade, H., Jan. 1997. A semantics for possibility theory based on likelihoods. Journal of Mathematical Analysis and Applications 205 (2), 359–380.

Edwards, A. W. F., 1992. Likelihood. Johns Hopkins Press, Baltimore.

Fisher, R. A., 1973. Statistical Methods and Scientific Inference. Hafner Press, New York.

Giang, P. H., Shenoy, P. P., 2005. Decision making on the sole basis of statistical likelihood. Artificial Intelligence 165, 137–163.

Hoch, J. S., Blume, J. D., 2008. Measuring and illustrating statistical evidence in a cost-effectiveness analysis. Journal of Health Economics 27, 476–495.

Jeffreys, H., 1948. Theory of Probability. Oxford University Press, London.

Korn, E. L., Freidlin, B., 2006. The likelihood as statistical evidence in multiple comparisons in clinical trials: No free lunch. Biometrical Journal 48, 346–355.

Lavine, M., Schervish, M. J., 1999. Bayes factors: What they are and what they are not. American Statistician 53, 119–122.

Lindsey, J., 1996. Parametric Statistical Inference. Oxford Science Publications. Clarendon Press, Oxford.

Morgenthaler, S., Staudte, R. G., 2012. Advantages of Variance Stabilization. Scandinavian Journal of Statistics 39 (4), 714–728.

Muralidharan, O., 2010. An empirical Bayes mixture method for effect size and false discovery rate estimation. Annals of Applied Statistics 4, 422–438.

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 (5), art. 4.

Patriota, A. G., 2013. A classical measure of evidence for general null hypotheses. Fuzzy Sets and Systems 233, 74 – 88.

Pawitan, Y., 2001. In All Likelihood: Statistical Modeling and Inference Using Likelihood. Clarendon Press, Oxford.

Pawitan, Y., Murthy, K., Michiels, S., Ploner, A., 2005. Bias in the estimation of false discovery rate in microarray studies. Bioinformatics 21, 3865–3872.

Rohde, C., 2014. Pure likelihood methods. In: Introductory Statistical Inference with the Likelihood Function. Springer International Publishing, pp. 197–209.

Royall, R., 1997. Statistical Evidence: A Likelihood Paradigm. CRC Press, New York.

Royall, R., 2000. On the probability of observing misleading statistical evidence (with discussion). Journal of the American Statistical Association 95, 760–780.

Rubin, D. B., 1981. Estimation in parallel randomized experiments. Journal of Educational Statistics 6, pp. 377–401.

Schervish, M. J., 1996. P values: What they are and what they are not. American Statistician 50, 203–206.

Severini, T., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.

Spanos, A., 2013. Revisiting the likelihoodist evidential account [comment on "A likelihood paradigm for clinical trials"]. Journal of Statistical Theory and Practice 7 (2), 187–195.

Spohn, W., 2012. The Laws of Belief: Ranking Theory and Its Philosophical Applications. Oxford University Press.

Sprott, D. A., 2000. Statistical Inference in Science. Springer, New York.

Strug, L. J., Hodge, S. E., 2006. An alternative foundation for the planning and evaluation of linkage analysis i. decoupling 'error probabilities' from 'measures of evidence'. Human Heredity 61, 166–188.

Walley, P., Moral, S., 1999. Upper probabilities based only on the likelihood function. Journal of the Royal Statistical Society.Series B: Statistical Methodology 61, 831–847.

Yang, Y., Aghababazadeh, F. A., Bickel, D. R., 2013a. 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., Li, Z., Bickel, D. R., 2013b. Empirical Bayes estimation of posterior probabilities of enrichment: A comparative study of five estimators of the local false discovery rate. BMC Bioinformatics 14, art. 87.

Zhang, Z., Zhang, B., 2013a. A likelihood paradigm for clinical trials. Journal of Statistical Theory and Practice 7, 157–177.

Zhang, Z., Zhang, B., 2013b. Rejoinder [on "A likelihood paradigm for clinical trials"]. Journal of Statistical Theory and Practice 7, 196–203.