Statistica Sinica **34** (2024), 2237-2263
doi:https://doi.org/10.5705/ss.202022.0261

ON p-VALUE COMBINATION OF INDEPENDENT ANDNONSPARSE SIGNALS: ASYMPTOTIC EFFICIENCY AND FISHER ENSEMBLE

Yusi Fang[1], Chung Chang*[2] and George C. Tseng*[1]

[1] *University of Pittsburgh and* [2] *National Sun Yat-sen University*

Abstract: Combining p-values to integrate multiple effects is of long-standing interest in social science and biomedical research. In this study, we revisit a classical scenario closely related to a meta-analysis with unknown heterogeneity that combines a finite and fixed number of p-values, while allowing the sample size for generating each p-value to go to infinity. Although many modified Fisher's methods have been developed for this purpose, their asymptotic properties and finite-sample numerical performance have not been examined, and so is the motivation for our study. Our results show that Fisher and adaptive rank truncated product methods have top performance and complementary advantages across different proportions of true signals. Consequently, we propose an ensemble method, called the Fisher ensemble, that combines the two top-performing Fisher-related methods using a robust harmonic mean ensemble approach. We show that the Fisher ensemble achieves asymptotic Bahadur optimality and integrates advantages of the two methods in simulations. We subsequently extend the Fisher ensemble to a variation that is particularly powerful for concordant effect size directions. A transcriptomic meta-analysis application confirms the theoretical and simulation conclusions, generates intriguing biomarker and pathway findings, and demonstrates the strengths and strategy of using the proposed Fisher ensemble methods.

Key words and phrases: Ensemble method, global hypothesis testing, p-value combination, omnibus test.

1. Introduction

Methods for combining p-values are of substantial interest in statistics and scientific fields for aggregating homogeneous or possibly heterogeneous information from multiple sources. Consider the problem of combining K p-values, $\vec{p} = (p_1, \ldots, p_K)$, where p_i is the p-value of testing $H_0^{(i)} : \theta_i \in \Theta_0^{(i)}$ versus $H_1^{(i)} : \theta_i \in \Theta^{(i)} - \Theta_0^{(i)}$. Here, θ_i denotes the parameter of interest, and $\Theta^{(i)}$ and $\Theta_0^{(i)}$ denote the total possible parameter space and null parameter space of θ_i, respectively. For example, $\theta_i = \mu_i$ for $N(\mu_i, 1)$, $\Theta^{(i)} = \mathbb{R}$, and $\Theta_0^{(i)} = \{\mu_i = 0\}$ for a simple Z-test. The global union-intersection test for detecting any signal in the K p-values is $H_0 : \cap_{1 \leqslant i \leqslant K} \{\theta_i \in \Theta_0^{(i)}\}$ versus $H_1 : \cup_{1 \leqslant i \leqslant K} \{\theta_i \in \Theta^{(i)} - \Theta_0^{(i)}\}$. A general strategy is to combine the input p-values to form a test statistic

*Corresponding author.

for globally testing the existence of any signal. In the literature, three major categories of methods have been developed, depending on the types of input data and signal. The first category considers the combination of independent p-values, where K is small and fixed (e.g., K between 5 and 30). The sample size n_i $(1 \leqslant i \leqslant K)$ for deriving p_i is large, and can asymptotically go to infinity. This first classical scenario is closely related to meta-analysis applications that integrate multiple small effects to increase statistical power. Traditional methods include Fisher's method $T_{\text{Fisher}} = \sum_{i=1}^{K} -2\log p_i$ (Fisher (1934)) and Stouffer's method $T_{\text{Stouffer}} = \sum_{i=1}^{K} \Phi^{-1}(1-p_i)$(Stouffer et al. (1949)), as well as many other transformation selections. The second category combines independent, sparse, and weak signals. Here, a large number of p-values are combined $(K \to \infty)$, but only a small number ℓ of the K p-values ($\ell = K^{\beta}$ with $0 < \beta < 1/2$) have weak signals, and all remaining p-values have no signal. The higher criticism test (the HC test, hereafter; Donoho and Jin (2004)) and the Berk-Jones test (the BJ test, hereafter; Berk and Jones (1979); Li and Siegmund (2015)) are two representative methods, and are shown to be asymptotically optimal in terms of the detection boundary across varying levels of signal sparsity $(0 < \beta < 1/2)$ as $K \to \infty$. In the third category, the K p-values are integrated with an unknown correlation structure and with sparse and weak signals. Liu and Xie (2020) and Wilson (2019) proposed a Cauchy test (CA test) and a harmonic mean test (HM test), respectively. These methods provide robustness under an unknown dependency structure when inference is established under an independence assumption. They also attain the optimal detection boundary for detecting highly sparse signals (with $s = K^{\beta}$, $0 < \beta < 1/4$, but not for $1/4 < \beta < 1/2$) as $K \to \infty$ (Liu and Xie (2020); Fang, Tseng and Chang (2023)).

In this study, we revisit the methods of the first category, evaluate their asymptotic efficiencies, assess their finite-sample numerical performance, and propose an ensemble method that combines two complementary top performers for general applications. To differentiate between the first category and the second and third categories, we focus on detecting independent and nonsparse signals inside a small and fixed number of p-values for scenarios of the first category. Here, the nonsparse signals differ from the sparse signals in the second and third categories in the sense that the proportion of true signals varies from $1/K$ to 1 and is unknown, whereas the proportions in the second and third categories vanish to zero as $K \to \infty$. Methods for the first "meta-analytic scenario with unknown heterogeneity" remain in high demand, and present new challenges in applications such as transcriptomic, GWAS, CNV, and methylation meta-analyses (Li and Tseng (2011); Tseng, Ghosh and Feingold (2012); Begum et al. (2012); Guerra and Goldstein (2016)).

Prior to the 1970s and 1980s, methods for the first category focused on aggregating transformed scores from the p-values: $T = \sum_{i=1}^{K} g(p_i) = \sum_{i=1}^{K} F_U^{-1}(p_i)$, where $F_U^{-1}(\cdot)$ is the inverse CDF of U. For example, U is a chi-squared distribu-

tion for the Fisher test, and a standard normal distribution for the Stouffer test. Littell and Folks (1973) showed that Fisher's method is asymptotically optimal in terms of the Bahadur relative efficiency, providing a theoretical justification for the log-transformation over the other types of transformations (see Section 2 for more details). Despite the optimal asymptotic efficiency of the Fisher test, its finite-sample performance in terms of statistical power is often poor if only part of the K p-values have signals. For this commonly encountered situation, with unknown heterogeneous signals, many modified Fisher methods have been developed to improve the original method. Dudbridge and Koeleman (2003) proposed the rank truncated product (RTP) method to aggregate signals only for the top ordered (i.e., the most significant) p-values: $T_m = -2\sum_{i=1}^{m} \log p_{(i)}$, where $p_{(i)}$ is the ith ordered p-value and $1 \leqslant m \leqslant K$ is a user-specified truncation point on the ranks of the input p-values. However, the choice of m is subjective, and the RTP method can suffer substantial power loss with a misspecified m. To address this challenge, several works have focused on improving the RTP method by adaptively determining m from an optimization criterion. For example, Song, Min and Zhang (2016) developed an adaptive Fisher procedure using a partial sum optimized by $\underline{z}$-standardization, similar to higher criticism (AFz, hereafter): $T_{\mathrm{AFz}} = \max_{1\leqslant j\leqslant K}\{-\sum_{i=1}^{j} \log p_{(i)} - \sum_{i=1}^{n} w(j,i)\}/\sqrt{\sum_{i=1}^{n} w^2(j,i)}$, where $w(j,i) = \min\{1, j/i\}$. Let $\bar{F}_{\chi^2_{2j}}(t) = 1 - F_{\chi^2_{2j}}(t)$, where $F_{\chi^2_{2j}}(t)$ denotes the CDF of a chi-squared random variable with degrees of freedom $2j$. Li and Tseng (2011) proposed an adaptive Fisher procedure using a partial sum optimized by the corresponding pseudo/surrogate "p-values"(AFs, hereafter): $T_{\mathrm{AFs}} = \max_{1\leqslant j\leqslant K} -\log\{h(\vec{p}, j)\}$. Here, $h(\vec{p}, j) = \bar{F}_{\chi^2_{2j}}(-2\sum_{i=1}^{j} \log p_{(i)})$ is the corresponding surrogate "p-value" of the partial sum. This is not a true and valid p-value, but rather a surrogate for fast computation by importance sampling (Huo et al. (2020)). Instead of using the surrogate p-values in AFs, Yu et al. (2009) proposed the adaptive rank truncated product (ARTP) method, which is based on the exact $\underline{p}$-values of the partial sum (AFp, hereafter): $T_{\mathrm{AFp}} = \max_{1\leq j\leq K} -\log\{h_j(\vec{p}, j)\}$, where $h_j(\vec{p}, j) = 1 - G_j(-2\sum_{i=1}^{j} \log p_{(i)})$, in which $G_j(t)$ denotes the CDF function of $-2\sum_{i=1}^{j} \log p_{(i)}$ under the null. For computation, Yu et al. (2009) proposed an algorithm that requires large storage memory to achieve manageable computing.

Another related strategy in the literature is to directly filter out p-values greater than a user-specified threshold $\tau \in (0, 1]$. For example, the truncated Fisher with hard-thresholding (TFhard) $T_{\mathrm{TFhard}}(\tau) = \sum_{i=1}^{K} -\log(p_i)\mathrm{I}_{\{p_i\leq\tau\}}$ (Zaykin et al. (2002)), where $\mathrm{I}_{\{\cdot\}}$ denotes the indicator function. Zhang et al. (2020) proposed the truncated Fisher with soft-thresholding (TFsoft) to improve on TFhard, arguing that the continuous soft-thresholding scheme leads to more stable performance with varying input p-values. In both TFsoft and TFhard, the choice of τ is not straightforward. Zhang et al. (2020) investigated the optimal choice of τ for TFhard under a theoretical setting of a Gaussian mixture, where

the mixture probability and the mean of the signals are known and $K \to \infty$. However, such prior information is usually unknown in practice. As such, they replaced the single user-specified τ with a user-specified set of thresholds $\mathcal{T}$, and proposed two omnibus tests for TFhard and TFsoft that alleviate the problem of choosing τ to some extent, but the selection of $\mathcal{T}$ is still prespecified and ad hoc.

Another line of research in p-value combination incorporates weighting in the procedure. For example, Xu et al. (2016) proposed an adaptive two-sample test for high-dimensional means, which can be regarded as a weighted test. Liptak's test (Lipták (1958)) can be considered as Stouffer's method with weights, and is commonly referred to as the weighted z-test. Won et al. (2009) estimated the best weights for Liptak's method from a simple alternative hypothesis, assuming an expected effect size. Other choices of weights for the z-test have been suggested by other researchers, including Mosteller and Bush (1954) and Zaykin (2011). In addition to the weighted z-test (Liptak's test), many tests constructed using the sum of transformed p-values also have a weighted version. For example, the CA and HM tests were originally proposed with weights, and in this study, we use the version with equal weights (Wilson (2019); Liu et al. (2019); Liu and Xie (2020)). Chen et al. (2014) proposed a test combining p-values based on the sum of an inverse gamma distribution, which can also be regarded as a weighted test, in the sense that it gives larger "weights" to smaller p-values. In fact, AFs and AFp can be considered as adaptively weighted methods using binary weights, and TFsoft (Zhang et al. (2020)) can be viewed as thresholding and weighting of the Fisher method.

Notwithstanding the active development of modified Fisher methods, there is no comprehensive and systematic evaluation of the asymptotic properties and finite-sample numerical performance of the methods in the first category. Our study sets out to fill this gap. In Section 2, we examine the Asymptotic Bahadur optimality (ABO) of seven methods in the first category: Fisher, Stouffer, AFs, AFz, AFp, TFhard, and TFsoft. The two adaptive Fisher methods, AFs and AFz, provide estimates of the subset of p-values contributing to the signal. Therefore, we also investigate whether the estimates in these two methods consistently select the subset of p-values containing the true signals (signal selection consistency). For completeness, we also examine the asymptotic efficiency of methods developed for sparse signals, including the CA, Pareto family, minimum p-value (minP), BJ, and HC methods. In Section 3, we perform finite-sample numerical evaluations to compare the statistical power of these methods under different K, signal strength, and proportions of true signals. The results reported in of Sections 2 and 3 show complementary advantages of two top performers, namely, Fisher and AFp, especially for varying proportions of true signals. Consequently, we develop a Fisher ensemble (FE) method in Section 4 that applies an HM ensemble approach to combine the Fisher and AFp methods. We prove the ABO of the FE (Section 4.2) and demonstrate its consistently high performance in various simulation

scenarios (Section 4.3). Section 5 develops an extension of the FE method, called FE_{CS}, for enhanced statistical power in terms of detecting signals with concordant effect size directions. Section 6 applies FE, FE_{CS}, and several existing methods to a transcriptomic meta-analysis on biomarker and pathway detection for aging (Zahn et al. (2007)). Section 7 concludes the paper.

2. Asymptotic Efficiencies of Existing Methods

This section investigates the asymptotic efficiencies of existing p-value combination methods. Because we focus on the scenarios with independent and nonsparse signals inside a finite number of p-values, we slightly generalize the setup proposed in Littell and Folks (1973) (differences are discussed in Remark 1), which uses the criterion of exact Bahadur relative efficiency (Bahadur (1967b)). Under this setting, Fisher's method is ABO (Littell and Folks (1973)) and shows theoretical advantages in terms of log-transformation. Numerous modified Fisher methods (e.g., AFs, AFp, AFz, TFhard, and TFsoft) have been developed to improve the finite-sample statistical power, but their asymptotic efficiency has not been investigated. Section 2.1 introduces the problem setting and defines the exact slope of a hypothesis test, which is a natural concept derived from the exact Bahadur relative efficiency. Section 2.2 presents the ABO results of the five modified Fisher methods.

2.1. Bahadur relative efficiency and exact slope

We first introduce the concept of an exact slope of a hypothesis test (Bahadur (1967b); Littell and Folks (1973)). Consider $(x_1, x_2, \ldots)$ as an infinite sequence of independent observations of a random variable X from the probability distribution P_θ with parameter $\theta \in \Theta$. Let T_n be a real-valued and continuous test statistic depending on the first n observations $(x_1, \ldots, x_n)$, where large values of T_n result in rejecting the null hypothesis. Assume that the probability distribution of T_n is the same for $\forall\theta \in \Theta_0$, which leads to $\mathbb{P}_\theta(T_n < t) = \mathbb{P}_0(T_n < t)$, for all $\theta \in \Theta_0$, and assume $1-\mathbb{P}_0(T_n < t)$ is uniformly distributed on $[0, 1]$ (Littell and Folks (1973)). Further, denote $p^{(n)} = 1-F_n(t_n)$ as the p-value of the observed $T_n = t_n$, where $F_n(t) = \mathbb{P}_0(T_n < t)$. We then define the exact slope of T_n as follows.

Definition 1. For the test statistic T_n with p-value $p^{(n)}$, if there is a positive-valued function $c(\theta)$, such that for any $\theta \in \Theta - \Theta_0$, $-(2/n)\log p^{(n)} \to c(\theta)$ as $n \to \infty$ with probability one, then $c(\theta)$ is called the `exact slope` of T_n.

As a simple example, consider testing for a zero mean ($\mu = 0$) with known variance under a univariate Gaussian distribution and T_n is the conventional z-test. It is easily seen that $c(\mu) = \mu^2$ is the exact slope of the z-test. For more examples, see Abrahamson (1967) and Bahadur (1967a). The exact slope of a test naturally connects to the exact Bahadur efficiency between the test

statistics. Consider two sequences of test statistics $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$ testing the same null hypothesis, with exact slopes $c_1(\theta)$ and $c_2(\theta)$, respectively. We define the ratio $\phi_{12}(\theta) = c_1(\theta)/c_2(\theta)$ as the *exact Bahadur relative efficiency* of $\{T_n^{(1)}\}$ relative to $\{T_n^{(2)}\}$, comparing the relative asymptotic efficiency between two test statistics. Indeed, considering any significance level $\alpha > 0$, for $i = 1, 2$, denote $N^{(i)}(\alpha)$ as the smallest sample size such that, for any $n \geqslant N^{(i)}(\alpha)$, the p-value of $T_n^{(i)}$ is smaller than α. Then, we can show with probability one that $\lim_{\alpha\to 0} N^{(2)}(\alpha)/N^{(1)}(\alpha) = \phi_{12}(\theta)$, which asymptotically characterizes the ratio of the smallest sample sizes of the two test statistics required to attain the same sufficiently small significance level α (Littell and Folks (1973)).

For $\theta \in \Theta_0$, the p-value $p^{(n)}$ follows a uniform distribution Unif$(0, 1)$. Lemma 1 shows that the analogous "exact slope" $-(2/n)\log p^{(n)}$ converges to zero with probability one.

Lemma 1. *For $\theta \in \Theta_0$, as n diverges, $-(2/n)\log p^{(n)} \to 0$ with probability one.*

The proof of Lemma 1 can be found in the Supplementary Material, Section S2.1. Here, we extend the definition of an exact slope to the null parameter space, where $c(\theta) = 0$, for $\theta \in \Theta_0$.

To benchmark the asymptotic efficiency of a p-value combination method, we introduce the theoretical setup adopted from the framework in Littell and Folks (1973). Suppose we have $K < \infty$ sequences of test statistics $\{T_{n_1}^{(1)}\}, \ldots, \{T_{n_K}^{(K)}\}$ for testing $\theta_i \in \Theta_0^{(i)}$, for $1 \leqslant i \leqslant K$. Assume that for all sample sizes $n_1, \ldots, n_K$, and when $\theta_i \in \Theta_0^{(i)}$, for $1 \leqslant i \leqslant K$, $\{T_{n_1}^{(1)}\}, \ldots, \{T_{n_K}^{(K)}\}$ are independently distributed. Denote $p_i^{(n_i)}$ as the p-value of the ith test statistic $T_{n_i}^{(i)}$. For each $1 \leqslant i \leqslant K$, assume that the sequence $\{T_{n_i}^{(i)}\}$ has the exact slope $c_i(\theta_i)$ as $-(2/n_i)\log p_i^{(n_i)} \to c_i(\theta_i) \geqslant 0$ with probability one as $n_i \to \infty$. We further assume the sample sizes $n_1, \ldots, n_K$ satisfy $n = (1/K)\sum_i^K n_i$ and $\lim_{n\to\infty}(n_i/n) = \lambda_i$, where $\lambda_i > 0$ and $\sum_i^K \lambda_i = K$. Under the above setup, the goal of any p-value combination method is to test

$$H_0 : \cap_{i=1}^K \{\theta_i \in \Theta_0^{(i)}\} \text{ versus } H_1 : \cup_{i=1}^K \{\theta_i \in \Theta^{(i)} - \Theta_0^{(i)}\}. \tag{2.1}$$

For simplicity, we assume under the null

$$\lambda_1 c_1(\theta_1) \geqslant \lambda_2 c_2(\theta_2) \geqslant \cdots \geqslant \lambda_K c_K(\theta_K) \geqslant 0,$$

where the first ℓ p-values have true signals (i.e., θ_i belong to $\Theta^{(i)} - \Theta_0^{(i)}$, for $1 \leqslant i \leqslant \ell$) with exact slopes $c_i(\theta_i) > 0$, and $c_i(\theta_i) = 0$ for the remaining $\theta_i \in \Theta_0^{(i)}$, for $\ell + 1 \leqslant i \leqslant K$.

Remark 1. There are two main differences between the original setup in Littell and Folks (1973) and ours. First, Littell and Folks (1973) assume that all studies have strictly positive exact slopes, whereas we allow some studies to have zero-valued exact slopes. Second, Littell and Folks (1973) consider a general parameter

Table 1. Results of asymptotic properties of 12 p-value combination methods: Fisher, Stouffer, five modified Fisher (AFs, AFp, AFz, TFhard, and TFsoft) and five methods designed for sparse and weak signals (Cauchy, Pareto, minP, BJ, and HC).

Methods	ABO	Exact slopes	Signal selection consistency	Proofs
Fisher	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	–	Theorem S1
Stouffer	No	$\frac{1}{K}[\sum_{i=1}^{\ell}(\lambda_i c_i(\theta_i))^{1/2}]^2$	–	Theorem S1
AFs	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	Yes	Theorems 1 & 4
AFp	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	Yes	Theorems 2 & 5
AFz	No	$\leqslant \max_{j\leq \ell} \frac{\sqrt{\sum_{i=1}^{K} \min^2\{1,1/i\}} \sum_{i=1}^{K} \lambda_i c_i(\theta_i)}{\sqrt{\sum_{i=1}^{K} \min^2\{1,j/i\}}}$	No	Theorem 3
TFhard	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	–	Theorem 6
TFsoft	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	–	Theorem 6
Pareto	No	$\max_i \lambda_i c_i(\theta_i)$	–	Theorem S2
Cauchy	No	$\max_i \lambda_i c_i(\theta_i)$	–	Theorem S3
minP	No	$\max_i \lambda_i c_i(\theta_i)$	–	Littell and Folks (1973)
BJ	No	$\max_i i\lambda_i c_i(\theta_i)$	No	Theorem S4
HC	No	–	No	Proposition S1

space Θ, whereas we consider a product of parameter spaces $\Theta^{(1)} \times \Theta^{(2)} \times \cdots \times \Theta^{(K)}$. Although differences exist, one can still establish the results in Littell and Folks (1973) by combining their arguments with Lemma 1.

Following Theorem 2 and the arguments in Section 4 in Littell and Folks (1973), under the alternatives, the maximum attainable exact slope for any p-value combination method is $\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$. Hence, we define the Asymptotic Bahadur Optimality (ABO) of a p-value combination method as follows.

Definition 2. Denote $\vec{\theta} = (\theta_1, \ldots, \theta_K)$. Under the above setup, a p-value combination test $H(p_1, \ldots, p_K)$ is Asymptotic Bahadur Optimal (ABO) if its exact slope $C_H(\vec{\theta})$ satisfies $C_H(\vec{\theta}) = \sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$.

2.2. The ABO property of p-value combination methods

Littell and Folks (1973) showed that the Fisher test is ABO, whereas Stouffer and minP tests are not. Except for these methods, there is a lack of asymptotic efficiency analysis for such methods. This subsection discusses five modified Fisher methods: AFs, AFp, AFz, TFhard, and TFsoft. We additionally analyze four methods designed for combining sparse and weak signals: Cauchy, Pareto, BJ, and HC. As expected, the latter four tests do not enjoy ABO property; the proofs are provided in the Supplementary Material. The theoretical results for the ABO, exact slope, and signal selection consistency (discussed in Theorems 4 and 5 and Remarks S3, S5, and S6) are summarized in Table 1.

Recall that the Fisher and the five modified Fisher methods combine p-values using the following test statistics:

$$T_{\text{Fisher}} = \sum_{i=1}^{K} -2\log p_{(i)}; \quad T_{\text{AFz}} = \max_{1\leqslant j\leqslant K} \frac{-\sum_{i=1}^{j}\log p_{(i)} - \sum_{i=1}^{K} w(i,j)}{\sqrt{\sum_{i=1}^{K} w^2(i,j)}};$$

$$T_{\text{AFs}} = \max_{1\leqslant j\leqslant K} -\log\left\{\bar{F}_{\chi^2_{2j}}\left(-2\sum_{i=1}^{j}\log p_{(i)}\right)\right\}; \quad T_{\text{AFp}} = \max_{1\leqslant j\leqslant K} -\log\{h_j(\vec{p},j)\};$$

$$T_{\text{TFhard}}(\tau) = \sum_{i=1}^{K}(-2\log p_i)\mathrm{I}_{\{p_i\leqslant\tau\}}; \quad T_{\text{TFsoft}}(\tau) = \sum_{i=1}^{K}(-2\log p_i + 2\log\tau)_+.$$

Here, $w(i,j) = \min\{1, j/i\}$. In addition, $\tau \in (0,1]$ is a user-specified constant for the two truncated Fisher methods, and $(x)_+$ denotes $\max(x,0)$.

All six methods can be characterized in the form of $H(-\log p_1, \ldots, -\log p_K)$ by some function $H(\cdot)$. With the log-transform on p-values as a key ingredient, the above methods can potentially achieve high asymptotic efficiency. Indeed, together with Lemma 1, by using almost the same arguments as those in Littell and Folks (1973), one can show that the Fisher test is ABO; see Theorem S1.

Although achieving high asymptotic efficiency, the Fisher test has been shown to have poor performance empirically when only few of the part of p-values contain signals (e.g., 2 out of 10 p-values have signals); see Song, Min and Zhang (2016) and Li and Tseng (2011) for more discussions. Many modified Fisher methods have been proposed to address this problem (Zaykin et al. (2002); Yu et al. (2009); Kuo and Zaykin (2011); Zhang et al. (2020); Li and Tseng (2011); Song, Min and Zhang (2016)). The idea is to filter out large p-values that are less likely to carry signals, and reduce the impact of noise, while still using the log-transformation of the p-values to achieve high efficiency. In Particular, AFs, AFp, and AFz combine the first m smallest ordered p-values. All three methods use some optimization criterion that adaptively selects m to achieve superior finite-sample power in varying proportions of signals. Whether AFs, AFp, and AFz retain the ABO property of Fisher is an intriguing question, and is investigated below. In fact, as surprisingly found in the following three theorems, that AFs and AFp are ABO, but AFz is not.

Theorem 1 (AFs is ABO). *Under the setup in Section* 2.1*, T_{AFs} is similar to the Fisher test in terms of being ABO, with exact slope $C_{AFs}(\vec{\theta}) = \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$.*

Theorem 2 (AFp is ABO). *Under the setup in Section* 2.1*, T_{AFp} is similar to the Fisher test in terms of being ABO, with exact slope $C_{AFp}(\vec{\theta}) = \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$.*

Theorem 3 (AFz is not ABO). *Under the setup in Section* 2.1*, consider the test statistic $T_A = \max_{1\leqslant j\leqslant K}\{-2\sum_{i=1}^{j}\log p_{(i)} - A_j\}/B_j$, where $B_j > 0$ and A_j are some finite constants that depend only on j and K. Assume there is no tie*

for $\sum_{i=1}^{j}\lambda_i c_i(\theta_i)/B_j$*,for* $j=1,\ldots,K$*, and* B_j *is monotonic increasing. Then,* T_A *is not ABO, with exact slope*

$$C_A(\vec{\theta}) \leqslant \max_{1\leqslant j\leqslant \ell}\frac{B_1}{B_j}\sum_{i=1}^{j}\lambda_i c_i(\theta_i).$$

The equality holds if and only if $\ell=1$ *(i.e., there is only one signal inside the* K *p-values).*

By taking $A_j=2\sum_{i=1}^{K}w(j,i)$ and $B_j=2\{\sum_{i=1}^{K}w^2(j,i)\}^{1/2}$, T_A reduces to T_{AFz}, indicating that AFz is not ABO, in general (e.g., a special case that AFz is ABO is when $\ell=1$).

The better asymptotic efficiency of AFp and AFs compared with that of AFz may be because the latter tries to estimate the subset of p-values with true signals. Consider the equivalent form of AFs for combining independent p-values:

$$T'_{\mathrm{AFs}}=\min_{\vec{w}}\bar{F}_{\chi^2_{2(\sum_{i=1}^{K}w_i)}}\left(-2\sum_{i=1}^{K}w_i\log p_i\right),$$

where $\vec{w}=(w_1,\ldots,w_K)\in\{0,1\}^K$ is the vector of binary weights that identify the candidate subset of p-values with true signals. Note that T'_{AFs} is the original form proposed in Li and Tseng (2011). Denoted by $\hat{\vec{w}}=\mathrm{argmin}_{\vec{w}}\ \bar{F}_{\chi^2_{2(\sum_{i=1}^{K}w_i)}}(-2\sum_{i=1}^{K}w_i\log p_i)$, and let $\vec{w}^*=\{(w_1^*,\ldots,w_K^*):w_k^*=1\text{ if }\theta_i\in\Theta-\Theta_0\text{ and }w_k^*=0\text{ if }\theta_i\in\Theta_0\}$ be the indicators of the true signals. We show the signal selection consistency of AFs in the following theorem.

Theorem 4 (Signal selection by AFs is consistent). *Under the setup in Section* 2.1*,* $\hat{\vec{w}}\to\vec{w}^*$ *as* $n\to\infty$ *in probability for the AFs test.*

Theorem 5 (Signal selection by AFp is consistent). *Under the setup in Section* 2.1*, AFp selects the true subset of p-values by selecting* $p_{(1)},\ldots,p_{(\hat{j})}$*, where*

$$\hat{j}=\underset{1\leqslant j\leqslant K}{\mathrm{argmax}}-\log\{h_j(\vec{p},j)\}.$$

Theorem 6 states that for any given value of $\tau\in(0,1]$, TFhard and TFsoft are ABO.

Theorem 6 (TFhard and TFsoft are ABO). *Under the setup in Section* 2.1*, TFhard and TFsoft are ABO, with exact slopes* $C_{TFhard}(\vec{\theta})=C_{TFsoft}(\vec{\theta})=\sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$.

Although TFhard and TFsoft are ABO, the choice of τ may significantly affect their finite-sample performance (Zhang et al. (2020)). To address this issue, Zhang et al. (2020) proposed the following omnibus tests for both methods (denoted by oTFhard and oTFsoft, respectively):

$$T_{\text{oTFhard}} = \min_{\tau\in\mathcal{T}}[1 - F_{U_{\text{TFhard}}(\tau)}\{T_{\text{TFhard}}(\tau)\}]$$
$$T_{\text{oTFsoft}} = \min_{\tau\in\mathcal{T}}[1 - F_{U_{\text{TFsoft}}(\tau)}\{T_{\text{TFsoft}}(\tau)\}],$$

where $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ is a user-specified set of the candidates of τ. Here, $U_{\text{TFhard}}(\tau)$ and $U_{\text{TFsoft}}(\tau)$ denote the random variables that follow the null distributions of $T_{\text{TFhard}}(\tau)$ and $T_{\text{TFsoft}}(\tau)$, respectively. Although the omnibus tests alleviate the sensitivity of the choice of τ for both TFhard and TFsoft to some extent, $\mathcal{T}$ is still user specified and subjective. In addition, Zhang et al. (2020) derive the null distributions of both omnibus tests in an asymptotic sense as $K \to \infty$, which may not be accurate for small K with small p-value thresholds, which are commonly used in applications, such as genomics studies, to handle multiplicity.

Proofs of the theorems for the Fisher and modified Fisher methods in this subsection can be found in the Supplementary Material, Section S2.1. For completeness, we also show that methods that combine sparse and weak signals, such as the CA, Pareto, BJ, and HC methods, are not ABO (Supplementary Material, Section S1); the proofs are available in Supplementary Material, Section S2.3. In conclusion, the Fisher, AFs, AFp, TFhard, and TFsoft methods are the only ones with the ABO property. AFs, AFp, and AFz provide signal selection (i.e., subset estimation of the true signal), and AFs and AFp are the only two methods that exhibit consistency in the signal identification.

3. Power Comparison in Finite-Sample Simulations

Although Section 2 evaluates the asymptotic efficiency of p-value combination methods, the finite-sample statistical power of such methods under different proportions of signals has not been assessed. In this section, we evaluate the seven methods designed for nonsparse signal setting described in Section 2: Fisher, Stouffer, AFs, AFp, AFz, TFhard, and TFsoft. Additionally, we evaluate the following methods designed for combining sparse and weak signals for completeness: minP, CA, HM, BJ, and HC. Because TFhard and TFsoft are sensitive to the choice of the tuning parameter τ, for a fair comparison, we use the corresponding omnibus tests oTFhard and oTFsoft, instead. The tuning candidate set $\mathcal{T}$ is $\{0.01, 0.05, 0.5, 1\}$, which is used in the original paper (Zhang et al. (2020)).

For better illustration, we first present the results of the seven methods that combine nonsparse signals in Figures 1 and S1. Results comparing all 12 methods can be found in the Supplementary Material, Figures S2 and S3, where the modified Fisher methods dominate other methods designed for sparse and weak signals, in general, unless the signals are indeed sparse and weak (e.g., cases of $\ell/K \leqslant 0.1$ in Figure S3). However, in such cases, methods such as AFp and AFz still have comparable power with the top-performing methods, such as minP.

We simulate $X = (X_1, \ldots, X_K) \overset{D}{\sim} N(\vec{\mu}, I_K)$, where $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$ contains ℓ nonzero signals $\mu_1 = \cdots = \mu_\ell = \mu_0$, and $K - \ell$ with no signal ($\mu_{\ell+1} = \cdots = \mu_K = 0$). We evaluate for a wide range of $K = 10, 20, 40, 80$. For each K, we vary the proportions of the true signals ℓ/K: $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$. We also vary $\mu_0 = 0.5, 0.65, \ldots, 6$ for a broad range of signal strengths. The p-values are calculated using the two-sided test $p_i = 2(1 - \Phi(|X_i|))$, for $i = 1, \ldots, K$. For each combination of parameter values, we draw 10^6 Monte Carlo samples to calculate the critical values for all the methods at a given significance level α, because the p-value calculation algorithms for some methods, such as oTFsoft and oTFhard, are not accurate for small K.

Figure 1 shows the empirical power of the Fisher, Stouffer, and five modified Fisher methods with varying proportions of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ at significance level $\alpha = 0.01$. For a given K and proportion of signals ℓ/K, we choose the smallest μ_0 such that the best method has at least 0.5 statistical power, which allows optimized visualization and a comparison of the methods in different signal settings. We first note that AFz is inferior to the other modified Fisher methods, consistent with our theoretical result that AFz is not ABO. We further note that AFs, AFp, oTFhard, and oTFsoft exhibit comparable performance across varying proportions of signals. Fisher outperforms all other methods in terms of detecting frequent signals (e.g., when the proportion of true signals is greater than 0.3).

Although combining a small number of strong signals is not our primary focus, out of curiosity, and for a more comprehensive evaluation of existing methods, we simulate the alternatives with fixed numbers of true signals $\ell = 1, 2, \ldots, 6$, for $K = 20, 40, 80$, following the above simulation scheme. Figure S1 shows the empirical power of the Fisher, Stouffer, and five modified Fisher methods with varying numbers of signals $\ell = 1, 2, \ldots, 6$ at $\alpha = 0.05$. Similarly, for a given K and ℓ, we choose the smallest μ_0 such that the best method has at least 0.9 statistical power. Clearly, this simulation setting focuses more on the performance of combining less frequent, but relatively strong signals. We note that AFz, AFp, and oTFsoft have comparable statistical power across varying numbers of true signals, followed by AFs and oTFhard. However, the Fisher method, is significantly inferior to the modified Fisher methods when ℓ is much smaller than K (e.g., $\ell \leqslant 3$ for $K = 20, 40, 80$).

In many real applications (e.g., the transcriptomic meta-analysis in Section 6), the p-value combination test is repeated many times (i.e., for each gene). It is expected that some true biomarkers are more homogeneous with frequent true signals and some with less-frequent signals. The results in Figures 1 and S1 show the need to develop an ensemble method to integrate the advantages of Fisher and one of the top-performing modified Fisher methods, which is presented in the next section.

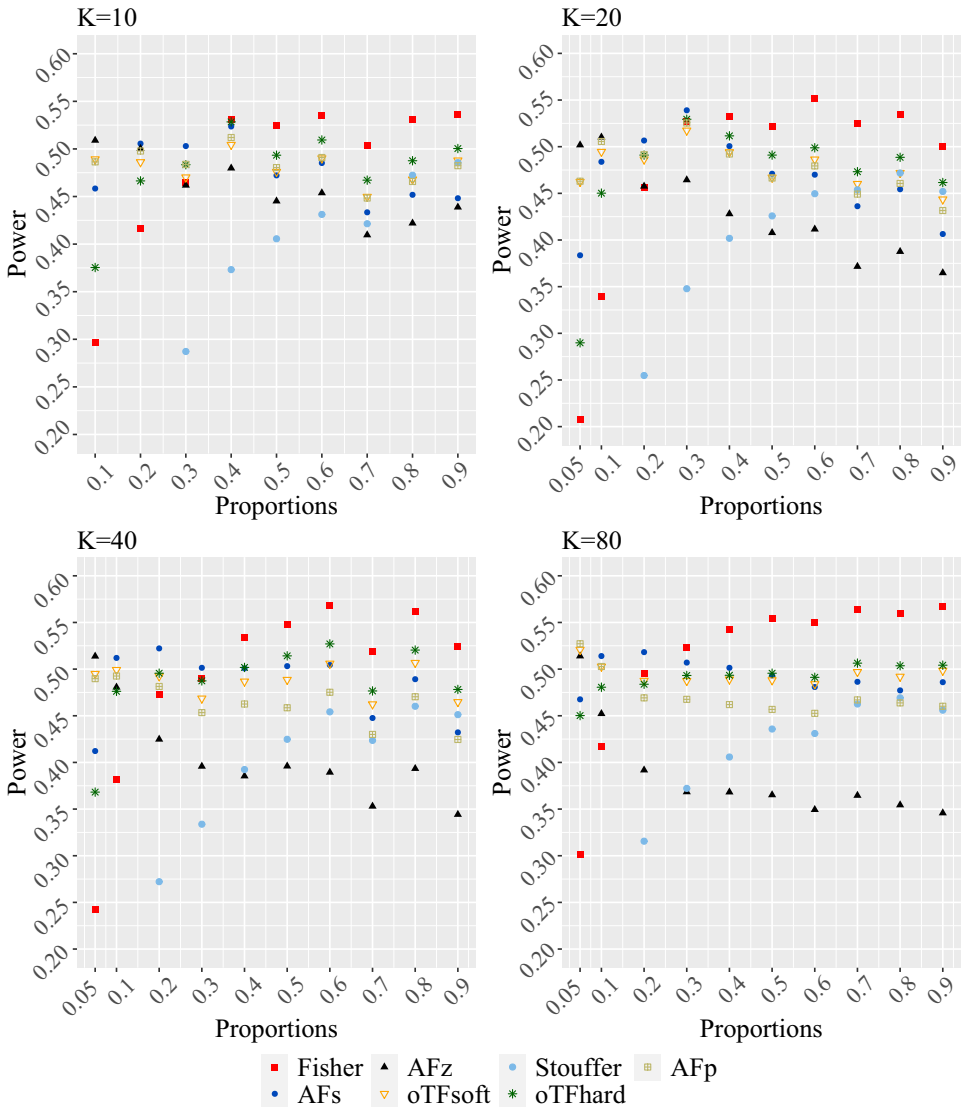

Figure 1. Statistical power of the Fisher, Stouffer, and five modified Fisher's methods at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. The standard errors are negligible compared with the scale of the mean power (smaller than 0.1% of the power), and hence are omitted. The results of the Stouffer test with power smaller than 0.25 are omitted.

4. Fisher Ensemble to Combine Fisher and AFp

As shown in Sections 2 and 3, the Fisher and four modified Fisher methods (AFs, AFp, TFhard, and TFsoft) are ABO, and have complementary strength in finite-sample evaluation of varying proportions and numbers of true signals. A natural idea is to ensemble Fisher and one of the four modified Fisher methods

for more stable and universally competitive performance. Because oTFhard and oTFsoft methods require an ad hoc decision of a user-specified set $\mathcal{T}$, and their existing computing algorithms are not accurate for small K, we develop an ensemble method to combine Fisher and AFp methods. In Section 4.1, we propose an ensemble approach, namely the FE, using the HM method (Wilson (2019); Fang, Tseng and Chang (2023)) to combine Fisher and AFp. In Section 4.2, we provide the theoretical support of the FE and show that the FE is ABO. Section 4.3 presents simulation results similar to those in Section 3 to demonstrate the balanced and superior performance of FE across varying proportions of true signals.

4.1. FE by HM integration

Denoted by p^{Fisher} and p^{AFp} the p-values derived from the Fisher and AFp combination tests, respectively. We propose combining the p-values of the two methods using $T_h = \{h(p^{\text{Fisher}}) + h(p^{\text{AFp}})\}/2$, with function h. Because the p^{Fisher} and p^{AFp} can be highly dependent, one option is to use the Cauchy combination test with $h(p) = \tan\{\pi(1/2 - p)\}$, because the theorems and simulations in Liu and Xie (2020) and Liu et al. (2019) show that the Cauchy combination test is robust to dependency of the combined p-values, and results in a fast algorithm with a Cauchy distribution under the null hypothesis (i.e., the null distribution is standard Cauchy). However, this Cauchy ensemble approach is problematic when either p^{Fisher} or p^{AFp} is close to one. In this case, the Cauchy transformation generates a $-\infty$ score, and the power is greatly reduced. We propose using the HM method (Wilson (2019)), $h(p) = 1/p$, in our FE, as follows:

$$T_{\text{FE}} = \frac{1}{2}\left(\frac{1}{p^{\text{Fisher}}} + \frac{1}{p^{\text{AFp}}}\right), \tag{4.1}$$

where the HM method is shown to be approximately equivalent to Cauchy in (Fang, Tseng and Chang (2023)). When the p-value p follows Unif$(0, 1)$, the reciprocal of p follows the Pareto distribution Pareto$(1, 1)$ with both the scale and shape parameters equal to one. We use the reciprocal of p-values instead of the Cauchy transformation to avoid the large negative score issue described above; also see Fang, Tseng and Chang (2023) for more details. Other than avoiding large negative score issue, using the HM with the reciprocal of the p-value $1/p$ performs almost identically to Cauchy $h(p)$. The Supplementary Material, Section S3.7, provides numeric results in which the ensemble method using the HM outperforms the Cauchy combination.

In the implementation, FE is fully data-driven with fast algorithms. Indeed, for $p_1, \ldots, p_K \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$, the null distribution of the Fisher test follows a chi-squared distribution with degrees of freedom $2K$. For the p-value calculation for AFp, Yu et al. (2009) proposed an empirical approach to avoid cumbersome two-

layer permutation. Finally, Theorems 1 and 2 in Fang, Tseng and Chang (2023) show that the harmonic approach using the reciprocal of the p-values can have robust type I error control if we naively use the Pareto distribution Pareto$(1,1)$ as the null distribution (see the Supplementary Material, Section S1.2).

As a result, the fast p-value computation for the FE T_{FE} is warranted. Table S1 in Section S3.1 justifies the above procedure, where we show that the type-I error control for FE is accurate for $\alpha \leqslant 0.05$ across a broad range of $5 \leqslant K \leqslant 100$.

4.2. Asymptotic efficiency of the FE

In this subsection, we show that the FE is ABO. We first introduce a heavy-tailed distribution family, namely, the regularly varying distribution R (Mikosch (1999)), where Cauchy and Pareto distributions are special cases of the family. Consider an ensemble method induced by a regularly varying distribution (e.g., Pareto$(1,1)$ for $1/p$, in our case) to combine multiple p-value combination methods (e.g., Fisher and AFp, in our case). The ensemble method is ABO if at least one of the p-value combination methods is ABO. Because both Fisher and AFp are ABO, and Pareto$(1,1)$ (corresponding to $1/p$) is a regularly varying distribution, we conclude that the FE is also ABO. Below, we outline the definition of the regularly-varying distributions and the theorem. The detailed proof is available in the Supplementary Material, Section S2.2.

Definition 3. A distribution F is said to belong to the regularly varying tailed family with index γ (denoted by $F \in R_{-\gamma}$) if $\lim_{x\to\infty} \bar{F}(xy)/\bar{F}(x) = y^{-\gamma}$, for some $\gamma > 0$ and all $y > 0$.

We denote the whole family of regularly varying tailed distributions by R. For two positive functions $u(\cdot)$ and $v(\cdot)$, we write $u(t) \sim v(t)$ if $\lim_{t\to\infty} u(t)/v(t) = 1$. It can be shown that every distribution F belonging to $R_{-\gamma}$ can be characterized by $\bar{F}(t) \sim L(t)t^{-\gamma}$, where $\bar{F}(t) = 1 - F(t)$ and $L(t)$ is a slowly varying function. A function L is called slowly varying if $\lim_{y\to\infty} L(ty)/L(y) = 1$ for any $t > 0$. Regularly varying distributions represent a wide class of heavy-tailed distributions, including the Cauchy, Pareto$(1,1)$ (HM), and general Pareto distributions.

Consider $L < \infty$ p-value combination test statistics $T_1, \ldots, T_L$. Denoted by $p_{T_1}, \ldots, p_{T_L}$ the resulting p-values of $T_1, \ldots, T_L$. In the FE, we have $L = 2$, and (T_1, T_2) are Fisher and AFp. Under Definition 3, consider the following ensemble method using a regularly varying tailed distribution:

$$T_{\mathrm{RV}}(\gamma) = \sum_{i=1}^{L} g_\gamma(p_{T_i}) = \sum_{i=1}^{L} F^{-1}_{U(\gamma)}(1 - p_{T_i}),$$

where $F_{U(\gamma)}$ is the CDF of $U(\gamma)$ and $U(\gamma) \in R_{-\gamma}$. Under the null hypothesis, the test statistic transforms all p_{T_i} into regularly varying tailed random variables

with index γ. The following theorem suggests that under mild conditions, the ensemble method with a regularly varying tailed distribution exhibits the ABO property.

Theorem 7. *For each $i = 1, \ldots, L$, let $C_i(\vec{\theta})$ be the exact slope of T_i, and assume $\max_{1 \leqslant i \leqslant L} C_i(\vec{\theta}) > 0$. Let $C_{RV}^{(\gamma)}(\vec{\theta})$ be the exact slope of $T_{RV}(\gamma)$. If one of the following two conditions holds:* (C1) *$F_{U(\gamma)}^{-1}(1-p)$ is bounded below: $F_{U(\gamma)}^{-1}(1-p) \geqslant \nu$, for some constant ν and $\forall p \in [0,1]$, and* (C2) *all T_i have nonzero exact slopes: $\min_{1 \leqslant i \leqslant L} C_i(\vec{\theta}) > 0$, then we have $C_{RV}^{(\gamma)}(\vec{\theta}) = \max_{1 \leqslant i \leqslant L} C_i(\vec{\theta})$.*

Remark 2. Because $1/p$ (reciprocal of p-value) is bounded below and $h(p)$ (Cauchy) is not, using $1/p$ rather than $h(p)$ can satisfy Condition (C1) in Theorem 7. In general, if Condition (C1) is not satisfied, Condition (C2) is a mild condition (meaning all tests T_i are at least minimally effective and have a nonzero slope), but not always easy to check or satisfy in practice. For example, when we aggregate the methods combining left one-sided p-values and right one-sided p-values in Section 4, methods that only combining left one-sided p-values will produce p-values converging to one when only positive effects exit; see Section 5 and the Supplementary Material, Section S3.7 for more details.

Theorem 7 suggests that $T_{\mathrm{RV}}(\gamma)$ is ABO as long as at least one of $T_1, \ldots, T_L$ is ABO. Consequently, the FE is ABO, because Pareto$(1,1)$ (corresponding to $1/p$) belongs to a regularly varying tailed distribution, and both Fisher and AFp are ABO.

4.3. Finite-sample power comparison for the FE

In this subsection, we evaluate the finite-sample power of FE. To illustrate that FE can take advantage of integrated methods, we also include AFs and Fisher as baseline methods. We use the same simulation scheme as that in Section 3 to generate the simulated data. Figure 2 shows the statistical power of FE, AFp, and Fisher, with varying proportions of true signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ at $\alpha = 0.01$. Similarly to Figure 1, for a given proportion of signals ℓ/K and number of combined p-values K, we choose the smallest μ_0 that allows the best method to have power larger than 0.5 in Figure 2. Figure S4 shows the statistical power of FE, AFp, and Fisher when combining $K = 20, 40, 80$ p-values with varying numbers of true signals $\ell = 1, 2, \ldots, 6$ at $\alpha = 0.05$. Similarly to Figure S1, for a given ℓ and K, we choose the smallest μ_0 that allows the best method to have power larger than 0.9 in Figure S4, which is supposed to focus on combining less frequent, but strong signals. As expected,the FE has stable statistical power that is comparable to the better of Fisher and AFp in settings with either dense but weak signals, or less frequent but strong signals. Specifically, when the proportion of signals is high, FE performs close to Fisher and is superior to AFp. When the number of true signals is small, FE performs close to AFp and outperforms Fisher. In the Supplementary Material, Figures S5 and S6, we implement another Fisher

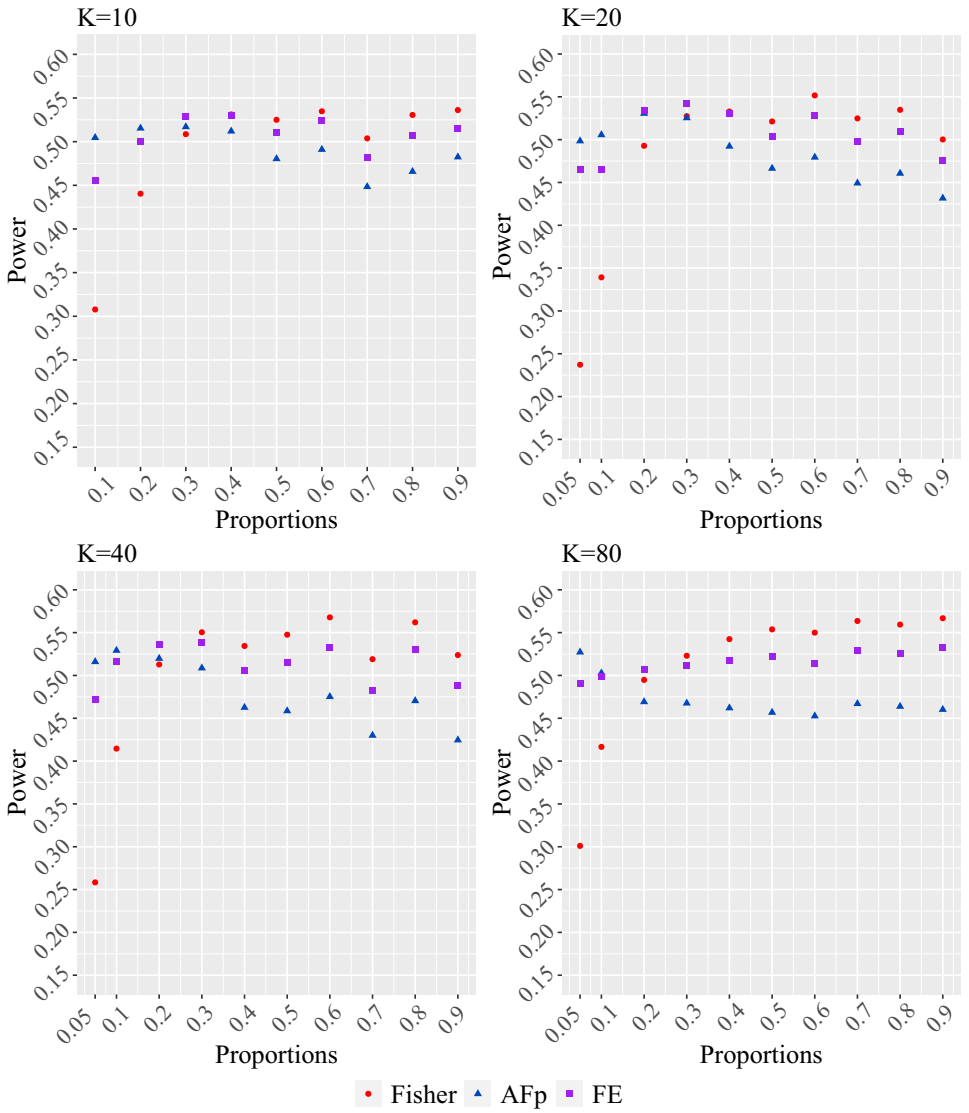

Figure 2. Statistical power of FE, Fisher, and AFp at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. The standard errors are negligible, and hence are omitted.

ensemble method (FE2) that combines Fisher, AFp, and minP. As expected, its power for only a small number of signals is slightly improved over that of FE, but at the expense of a large reduction of power when signals are frequent. From the asymptotic efficiency in Section 4.2 and the simulations above, we recommend using the FE method that combines Fisher and AFp for general applications.

5. Detection of Signals with Concordant Directions

5.1. FE focused on concordant signals ($\mathrm{FE_{CS}}$)

For all methods we have discussed so far, the global hypothesis setting is designed for two-sided tests, regardless of the directions of the effects. Recall from Equation 2.1 that the hypothesis testing considered is $H_0 : \cap_{i=1}^{K}\{\theta_i = 0\}$ versus $H_1 : \cup_{i=1}^{K}\{\theta_i \neq 0\}$. Consider the alternative hypothesis that only the first ℓ p-values have true signals (i.e., $\theta_i \neq 0$ for $1 \leqslant i \leqslant \ell$, and $\theta_{\ell+1} = \cdots = \theta_K = 0$). The two-sided tests to obtain p_i $(1 \leqslant i \leqslant K)$ cannot guarantee signals with concordant directions ($\mathrm{sgn}(\theta_1) = \cdots = \mathrm{sgn}(\theta_\ell)$, denoted by $\mathrm{sgn}(\cdot)$, the sign function), which is desirable in most applications. For example, when conducting a meta-analysis of K transcriptomic studies believed to be relatively homogeneous, we are interested in identifying biomarkers concordantly up-regulated or down-regulated. For this problem, Owen (2009) revisited the Pearson test statistic, and proposed using $T_{\mathrm{Pearson}} = \min\{\tilde{p}^{\mathrm{Fisher},L}, \tilde{p}^{\mathrm{Fisher},R}\}$, where $\tilde{p}^{\mathrm{Fisher},L}$ and $\tilde{p}^{\mathrm{Fisher},R}$ use Fisher to combine the left and right one-sided p-values respectively, and the Pearson test takes the more significant one as the test statistic. In this subsection, we similarly extend the FE method to use the HM approach to combine the two left and right one-sided p-values of Fisher and AFs (denoted by $\mathrm{FE_{CS}}$; Fisher ensemble for concordant signal):

$$T_{\mathrm{FE_{CS}}} = \frac{1}{4}\left(\frac{1}{\tilde{p}^{\mathrm{Fisher},L}} + \frac{1}{\tilde{p}^{\mathrm{Fisher},R}} + \frac{1}{\tilde{p}^{\mathrm{AFp},L}} + \frac{1}{\tilde{p}^{\mathrm{AFp},R}}\right).$$

Remark 3. When combining one-sided p-values, it is common to observe p-values close to one and it is critical to use the HM rather than Cauchy, to avoid $-\infty$ scores.

Remark 4. Let $C^L(\vec{\theta})$ be the maximum attainable exact slope for any p-value combination method combining left one-sided p-values, and define $C^R(\vec{\theta})$ in a similar manner for right one-sided p-values. By Theorem 7, the exact slope of $\mathrm{FE_{CS}}$ is $\max\{C^L(\vec{\theta}), C^R(\vec{\theta})\}$, indicating high asymptotic efficiency, because even if one has prior knowledge of the effect size direction, it is not possible to design a p-value combination method with a larger exact slope for detecting concordant signals.

For the computation, similarly to FE, one can use the p-value calculation of $\mathrm{Pareto}(1,1)$ to calculate the p-value for $\mathrm{FE_{CS}}$. This approximation procedure is justified by simulation results in Table S1 in Section S3.1 for a broad range of significance levels α and numbers of input p-values K.

5.2. Finite-sample power comparison for the FE for concordant signals

In this subsection, we evaluate the finite-sample power of $\mathrm{FE_{CS}}$. To demonstrate the advantages of $\mathrm{FE_{CS}}$, we also include the regular FE and Pearson as

baseline methods. We use the same simulation scheme as that in Section 3 to generate the simulated data. For FE_{CS} and Pearson, the one-sided p-values are generated by $\tilde{p}_i^{(L)} = 1 - \Phi(X_i)$ and $\tilde{p}_i^{(R)} = \Phi(X_i)$ $(i = 1, \ldots, K)$, respectively. For the regular FE, we combine the two-sided p-values $p_i = 2\{1 - \Phi(|X_i|)\}$, for $i = 1, \ldots, K$.

Figures 3 and S7 show the empirical power of FE_{CS}, Pearson, and the regular FE. For Figure 3, we choose the smallest μ_0 that allows the best method to have power larger than 0.5 for a given proportion of signals ℓ/K and a number of combined p-values K. Both FE_{CS} and Pearson dominate the regular FE, indicating the former two methods perform better for the alternatives with one-sided direction consistent effects (because $\mu_1 = \cdots = \mu_s = \mu_0 > 0$ under the alternatives). In addition, FE_{CS} has comparative performance with that of Pearson for $\ell/K \geqslant 0.2$, and outperforms Pearson when $\ell/K < 0.2$. For Figure S7, we choose the smallest μ_0 that allows the best method to have power larger than 0.9 for a given number of signals ℓ and number of combined p-values K. This setting focuses on less frequent but strong signals. Note that FE_{CS} outperforms Pearson when the number of signals is low (e.g., $\ell \leqslant 4$).

6. Real Application to AGEMAP Data

In this section, we apply different p-value combination methods to analyze data from the AGEMAP study (Zahn et al. (2007)). The data set contains microarray expressions of 8,932 genes in 16 tissues and age and sex variables of 618 mice subjects. We are interested in identifying age-associated marker genes. Following the original paper, we fit the following regression model to detect age-associated genes in each tissue:

$$Y_{ijk} = \beta_{0jk} + \beta_{\text{age},jk}\text{Age}_{ijk} + \beta_{\text{sex},jk}\text{Sex}_{ijk} + \varepsilon_{ijk} \quad \text{for} \quad i = 1, \ldots, m_{jk},$$

where Y_{ijk} is the expression level of the ith subject for the jth gene and kth tissue. For each gene j, we consider designs of both two-sided and one-sided tests when combining p-values across tissues. In the two-sided test design, the two-sided p-values $(p_{j1}, \ldots, p_{jK})$ for their corresponding $\beta_{\text{age},jk}$ coefficients are combined using the Fisher, AFp, and FE methods. In this case, the association directions (positive or negative associations) are not considered. In contrast, the one-sided test design combines left-tailed p-values $(\tilde{p}_{j1}^L, \ldots, \tilde{p}_{jK}^L)$ or right-tailed p-values $(\tilde{p}_{j1}^R, \ldots, \tilde{p}_{jK}^R)$ using FE_{CS}. Figure 4 shows the general workflows of transcriptomic meta-analysis for the jth gene with two-sided and one-sided designs. Compared with FE, FE_{CS} is expected to have increased power in terms of detecting age-related biomarkers with concordant signals (all positive associated or all negative associated) across tissues, but have reduced power for markers with heterogeneous signals (i.e., positive associations in some tissues and negative associations in some others). In this application, both concordant and heterogeneous age-related

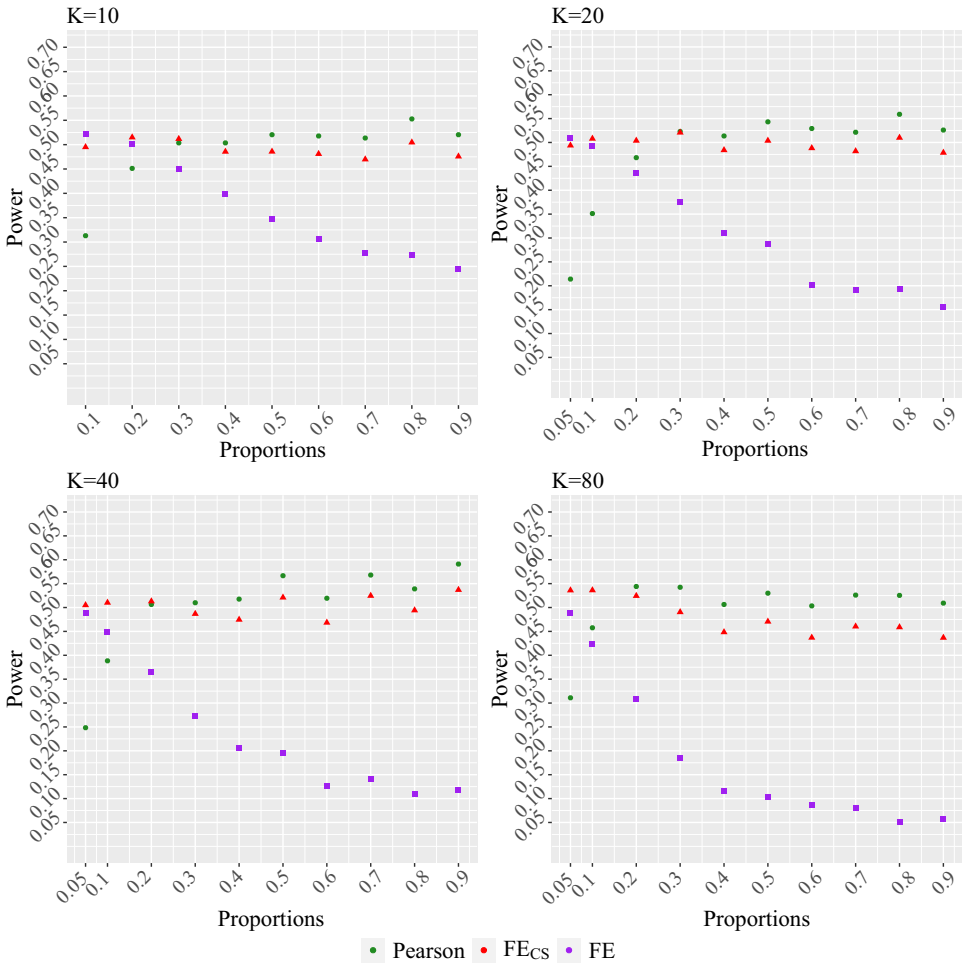

Figure 3. Statistical power of FE, FE_{CS}, and Pearson at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. The standard errors are negligible, and hence are omitted.

biomarkers are of interest. Heterogeneous biomarkers detected by FE can have different age-associations (positive, negative, or non-association) across tissues, whereas concordant biomarkers detected by FE_{CS} are tissue-invariant. FE and FE_{CS} serve as complementary tools for different biological objectives.

Figure 5(a) shows the Fisher, AFp, and FE p-value combination results in the two-sided test design. Under the significance level of q-value $\leqslant 0.05$, Fisher detects 576 genes (yellow color) and AFp detects 473 genes (green color), where Category II (392 genes) represents genes detected by Fisher and AFp and Categories I (184 genes) and III (81 genes) represent biomarkers uniquely detected by Fisher and by AFp, respectively. The heatmap shows an age-association measure defined as: $E_{jk} = -\text{sign}(\beta_{\text{age},jk}) \log(\min\{\tilde{p}^L_{jk}, \tilde{p}^R_{jk}\})$ for gene

j on the rows and tissue k on the columns, that is, the signed log-transformed (base 10) one-sided p-values. Consequently, a red color of E_{jk} represents a strong positive association with age, and blue denotes a strong negative association. As expected, FE combines the strengths of Fisher and AFp to detect 593 genes (purple color) that contain all genes in Category II and most genes in Categories I and III. By counting the number of tissues with p-values $p_{jk} \leqslant 0.05$, Figure S10 in the Supplementary Material shows that Category I genes (detected by Fisher, but not by AFp) are age-associated in more tissues, while Category III (detected by AFp, but not by Fisher) are age-associated in fewer tissues, which is consistent with the theoretical insight and simulation result that Fisher is more powerful for detecting frequent signals, and AFp is more powerful for relatively less frequent signals.

We next perform hierarchical clustering (using 1-correlation between tissues as the dissimilarity measure and complete linkage) for the 16 tissues based on the E_{jk} values in the 593 age-related genes detected by FE; the dendrogram is shown in Figure 5(a). By cutting the dendrogram, we identify five clear tissue modules with similar age-association patterns: (1) thymus and gonads; (2) spleen and lung; (3) eye, kidney, and heart; (4) hippocampus, adrenal glands, and muscle; and (5) cerebrum and spinal cord (also see Figure 5(b) for the heatmap of the pair-wise correlations). For the first module, the thymus has long been regarded as an endocrine organ that is closely related to gonads and sexual physiology, such as sexual maturity and reproduction (Grossman (1985); Leposavić and Pilipović (2018)). The spleen lung module is consistent with the finding in Zahn et al. (2007), and many reports suggest that the spleen and lung share a similar aging pattern (e.g., Schumacher et al. (2008)). For the third module, literature shows that the kidney and eye share structural, developmental, physiological, and pathogenic similarities and pathways. The relationships between age-related eye, kidney, and cardiovascular diseases have been widely reported (e.g., Farrah et al. (2020)). For the fourth module, numerous studies have reported a relationship between adrenal glands and hippocampal aging (e.g., Landfield, Waymire and Lynch (1978)). For the last module, few existing studies have investigated the aging process of the spinal cord (Knight and Nigam (2017)). However, it is reasonable that the cerebrum and spinal cord might share a similar aging pattern, because they both belong to the central nervous system. On the other hand, the liver has intriguingly negative correlations of aging effects with muscle, adrenal glands, and several brain regions, such as the hippocampus, cerebellum, and cerebrum (also see Figure 5(b)).

Next, we evaluate $\mathrm{FE}_{\mathrm{CS}}$ for the one-sided test design and compare it with FE. We calculate $S_{\mathrm{sign},j} = \sum_{k=1}^{16} \mathrm{sign}(\beta_{\mathrm{age},jk}) \mathrm{I}_{\{\min\{\tilde{p}_{jk}^L, \tilde{p}_{jk}^R\} \leqslant 0.05\}}$ to determine whether the detected concordant aging marker j is positively associated ($S_{\mathrm{sign},j} > 0$) or negatively associated ($S_{\mathrm{sign},j} \leqslant 0$), and use it to determine whether a detected marker is dominant with the positive association or negative association.

Similarly to the previous analysis, Figure 6 shows the age-associated genes detected by FE (593 genes, Categories II(A), II(B), and III) and FE_{CS} (398 genes, Categories I(A), I(B), II(A), and II(B)), where Categories II(A) and II(B) are genes detected by FE and FE_{CS}, Category III are detected only by FE, and Categories I(A) and I(B) are detected only by FE_{CS}. For genes detected by FE_{CS}, Categories I(A) and II(A) are concordant aging markers with a positive association (mostly red), and Categories I(B) and II(B) are negatively associated (mostly blue), which are visually consistent with the heatmap. In contrast, genes in Category III mostly have discordant association directions (partial red and partial blue). The Supplementary Material, Figure S11, shows the distributions of $S_{\text{sign},j}$ in the gene categories.

At significance level $q \leqslant 0.05$, FE_{CS} identifies 184 positively associated genes (Categories I(A) and II(A)) and 214 negatively associated genes (Categories I(B) and II(B)). We perform an Ingenuity Pathway Analysis (IPA) on these two concordant age-associated gene lists. The result identifies 11 enriched pathways from the 184 positively associated genes, and four enriched pathways from the 214 negatively associated genes (enrichment $p \leqslant 0.01$). Table S2 shows these enriched pathways with pathway names, enrichment p-values, and abundant supporting literature of the pathways related to aging/early development processes (see complete references in the Supplementary Material, References II). The result shows that FE_{CS} has an advantage in terms of identify age-associated markers concordant across tissues and delivering interpretable biological insights.

7. Conclusion

Combining p-values is a common and effective tool in many scientific applications. We focus on the scenario of meta-analysis with unknown heterogeneity, in which the number of combined p-values K is finite and fixed but the sample size for generating each p-value can increase to infinity (i.e., the first category described in the Introduction). The goal of this category is to aggregate heterogeneous independent signals, where the proportion of true signals is unknown and can range from $1/K$ to 1. Note that our goal is to combine independent and nonspare signals and to distinguish it from combining sparse signals in the asymptotic rare and weak (ARW) model when $K \to \infty$, which is commonly considered in the second and third categories described in the Introduction.

Our results contribute to the literature in three ways. First, this is the first study to comprehensively evaluate p-value combination methods for their asymptotic efficiency in terms of asymptotic Bahadur optimality (ABO). We investigate classical methods (Fisher and Stouffer) and modified Fisher methods (AFs, AFp, AFz, TFhard, and TFsoft). The result shows that Fisher, AFs, AFp, TFhard, and TFsoft are ABO, but Stouffer and AFz are not. We also find interesting consistency properties when estimating the subset of contributing

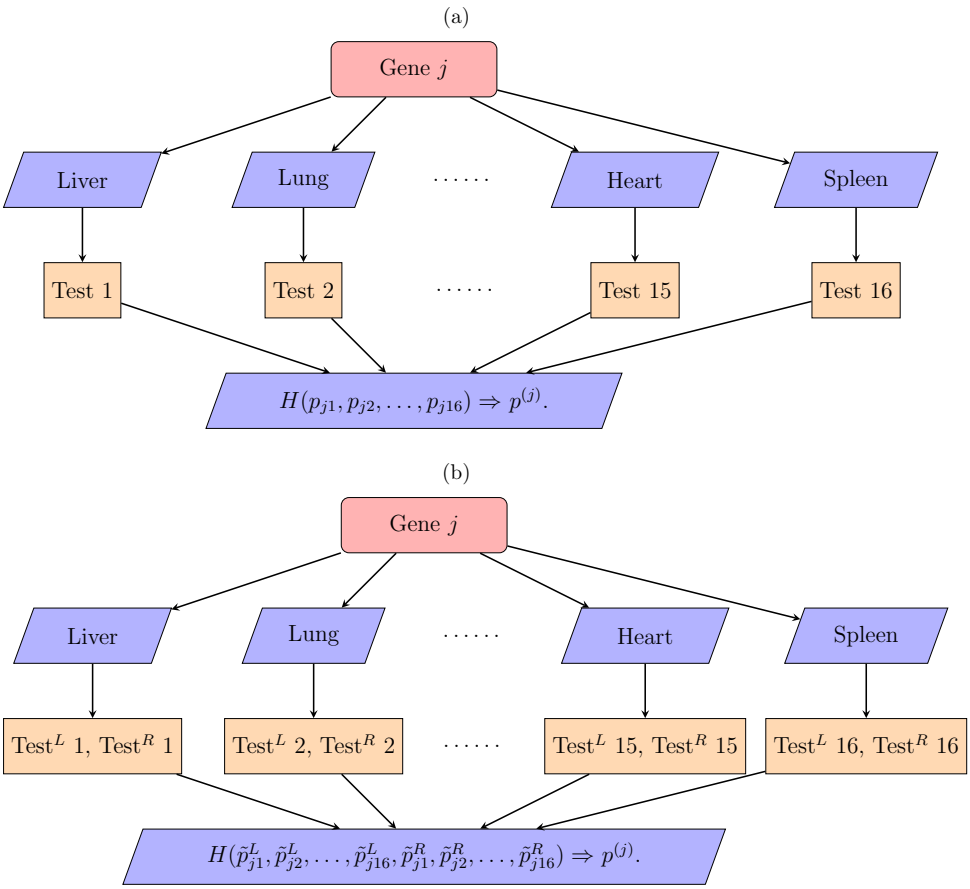

Figure 4. Procedures of transcriptomic meta-analysis on AGEMAP data set (two-sided design (Figure 4(a)) and one-sided design (Figure 4(b)), where $H(\cdot)$ denotes a chosen p-value combination method and $p^{(j)}$ denotes the corresponding p-value of H with input p-values. Here, p_{jk} is the two-sided p-value for the jth gene on the kth tissue, and $\tilde{p}^L_{jk}$ and $\tilde{p}^R_{jk}$ are the left-tailed and right-tailed p-values for the jth gene on the kth tissue, respectively.

signals in AFs and AFp (Theorems 4 and 5). Second, we perform an extensive finite-sample power comparison and conclude that Fisher and AFp are the two top performers, with complementary advantages, where Fisher is more powerful with frequent signals and AFp is more powerful in relatively sparse settings. Third, we propose a Fisher ensemble (FE) method that combines Fisher and AFp. A one-sided test modification, FE_{CS}, is further developed for detecting concordant signals. Here, FE and FE_{CS} offer several advantages: First, both methods have high asymptotic efficiency (FE is ABO). Second, the HM combination avoids the $-\infty$ score in the Cauchy. Third, we numerically demonstrate their constantly high performance across varying proportions of signals. Fourth, both methods

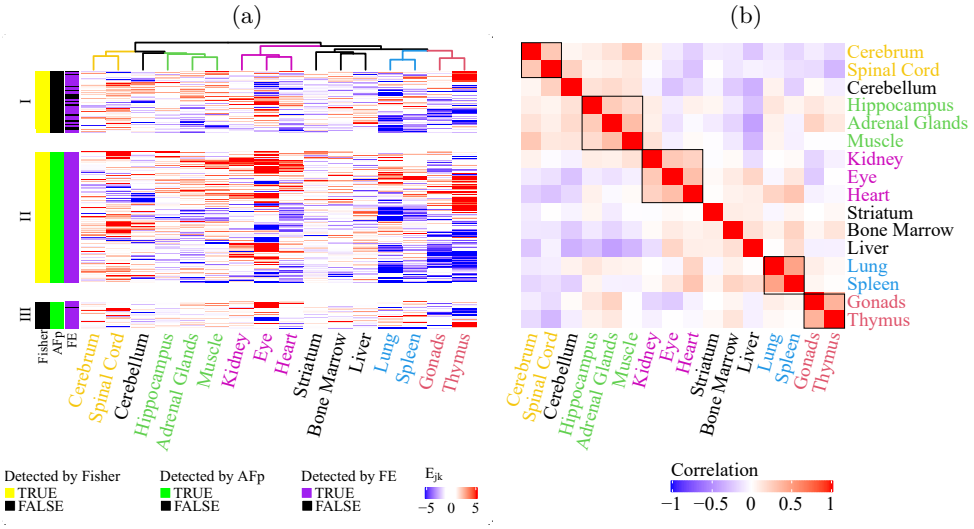

Figure 5. (a) Heatmaps of age-association measure E_{jk} of significant genes ($q <= 0.05$) detected in the two-sided test design. Category I: genes detected by Fisher, but not by AFp; II: genes detected by both Fisher and AFp; III: genes detected by AFp, but not by Fisher. (b) Heatmap of pair-wise correlations between tissues based on the detected genes by FE ($q \leqslant 0.05.$) in (a).

have fast procedures. Finally, an application to AGEMAP transcriptomic data verifies our theoretical conclusions, demonstrates the superior performance of FE and FE_{CS}, and discovers intriguing biological findings in age-associated biomarkers and pathways.

Modern data science faces challenges from data heterogeneity, increasingly complex data structures, and the need for effective methods for new scientific hypotheses. The ensemble methods proposed in this paper, FE and FE_{CS}, have solid theoretical and numerical support for their superior performance in a wide range of signal settings. Therefore, we believe these methods will be useful in many other scientific problems.

Supplementary Material

The online Supplementary Material includes proofs of Lemma 1, Theorems 1–7, and all technical lemmas, as well as additional theoretical results (Theorems S1-S4 and Proposition S1 and their proofs) and additional simulation results.

Acknowledgments

YF and GCT were funded by NIH R01LM014142 and R21LM012752; CC was funded by the Ministry of Science and Technology of ROC 109-2118-M-110-

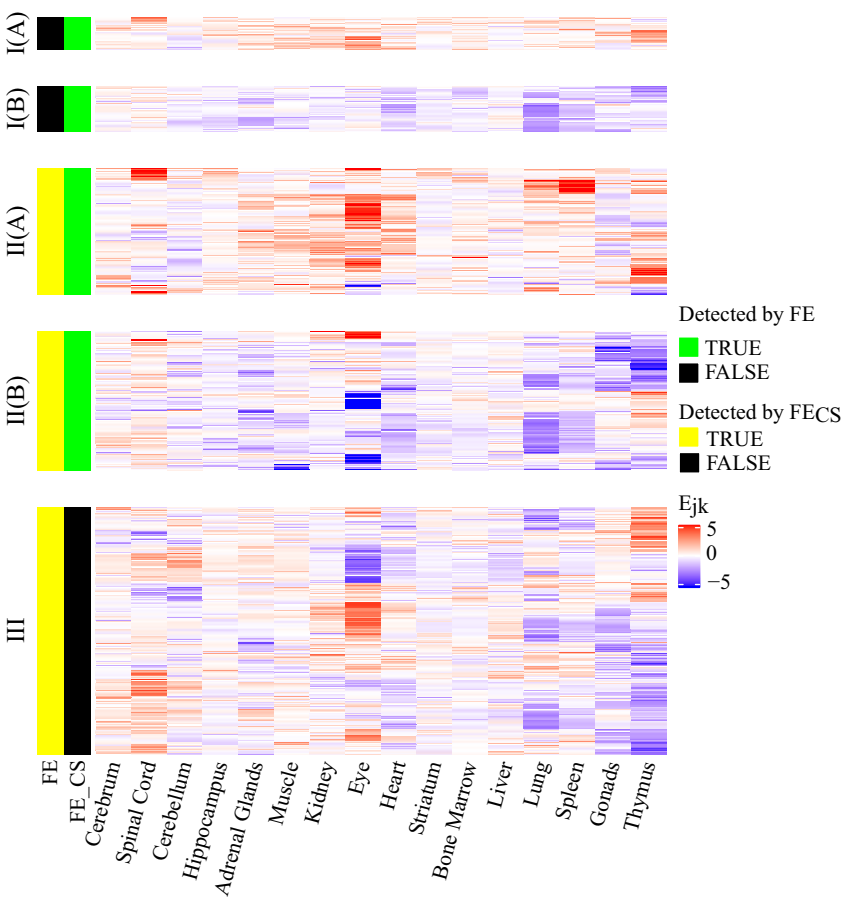

Figure 6. Heatmaps of age-association measure E_{jk} of genes detected by $\mathrm{FE_{CS}}$ or by FE ($q <= 0.05$). Heatmap I(A) represents up-regulated genes detected only by $\mathrm{FE_{CS}}$ (38 genes); heatmap I(B) represents down-regulated genes detected only by $\mathrm{FE_{CS}}$ (53 genes); heatmap II(A) represents up-regulated genes detected by both $\mathrm{FE_{CS}}$ and FE (146 genes); heatmap II(B) represents down-regulated genes detected by both $\mathrm{FE_{CS}}$ and FE (161 genes); heatmap III represents genes detected only by FE (286 genes).

002 and 110-2118-M-110-001. The authors thank Yongseok Park for the helpful discussions and the editors and anonymous reviewers for their careful reading of our manuscript and many insightful comments and suggestions.

References

Abrahamson, I. G. (1967). Exact Bahadur efficiencies for the Kolmogorov-Smirnov and Kuiper one-and two-sample statistics. *The Annals of Mathematical Statistics* **38**, 1475–1490.

Bahadur, R. R. (1967a). An optimal property of the likelihood ratio statistic. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Edited by L. M. Le Cam and J. Neyman), 13–26.

Bahadur, R. R. (1967b). Rates of convergence of estimates and test statistics. *The Annals of Mathematical Statistics* **38**, 303–324.

Begum, F., Ghosh, D., Tseng, G. C. and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research* **40**, 3777–3784.

Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Probability Theory and Related Fields* **47**, 47–59.

Chen, Z., Yang, W., Liu, Q., Yang, J. Y., Li, J. and Yang, M. Q. (2014). A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. *BMC Bioinformatics* **15**, S3.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32**, 962–994.

Dudbridge, F. and Koeleman, B. P. (2003). Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **25**, 360–366.

Fang, Y., Tseng, G. C. and Chang, C. (2023). Heavy-tailed distribution for combining dependent p-values with asymptotic robustness. *Statistica Sinica* **33**, 1115–1142.

Farrah, T. E., Dhillon, B., Keane, P. A., Webb, D. J. and Dhaun, N. (2020). The eye, the kidney, and cardiovascular disease: Old concepts, better tools, and new horizons. *Kidney International* **98**, 323–342.

Fisher, R. (1934). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh and London.

Grossman, C. J. (1985). Interactions between the gonadal steroids and the immune system. *Science* **227**, 257–261.

Guerra, R. and Goldstein, D. R. (2016). *Meta-Analysis and Combining Information in Genetics and Genomics*. Chapman and Hall/CRC.

Huo, Z., Tang, S., Park, Y. and Tseng, G. (2020). P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. *Bioinformatics* **36**, 524–532.

Knight, J. and Nigam, Y. (2017). Anatomy and physiology of ageing 5: The nervous system. *Nursing Times* **113**, 55–58.

Kuo, C.-L. and Zaykin, D. V. (2011). Novel rank-based approaches for discovery and replication in genome-wide association studies. *Genetics* **189**, 329–340.

Landfield, P. W., Waymire, J. and Lynch, G. (1978). Hippocampal aging and adrenocorticoids: Quantitative correlations. *Science* **202**, 1098–1102.

Leposavić, G. M. and Pilipović, I. M. (2018). Intrinsic and extrinsic thymic adrenergic networks: Sex steroid-dependent plasticity. *Frontiers in Endocrinology* **9**. Web: `https://doi.org/10.3389/fendo.2018.00013`.

Li, J. and Siegmund, D. (2015). Higher criticism: p-values and criticism. *The Annals of Statistics* **43**, 1323–1350.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5**, 994–1019.

Lipták, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl* **3**, 171–197.

Littell, R. C. and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *Journal of the American Statistical Association* **68**, 193–194.

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E. and Lin, X. (2019). ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics* **104**, 410–421.

Liu, Y. and Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **115**, 393–402.

Mikosch, T. (1999). *Regular Variation, Subexponentiality and their Applications in Probability Theory*. Eindhoven University of Technology, Eindhoven.

Mosteller, F. and Bush, R. R. (1954). *Selected Quantitative Techniques*. Addison-Wesley.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *The Annals of Statistics* **37**, 3867–3892.

Schumacher, B., van der Pluijm, I., Moorhouse, M. J., Kosteas, T., Robinson, A. R., Suh, Y. et al. (2008). Delayed and accelerated aging share common longevity assurance mechanisms. *PLoS Genetics* **4**, e1000161.

Song, C., Min, X. and Zhang, H. (2016). The screening and ranking algorithm for change-points detection in multiple samples. *The Annals of Applied Statistics* **10**, 2102–2129.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. and Williams, R. M., Jr. (1949). The American soldier: Adjustment during army life. *Studies in Social Psychology in World War II* **5**.

Tseng, G. C., Ghosh, D. and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* **40**, 3785–3799.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* **116**, 1195–1200.

Won, S., Morris, N., Lu, Q. and Elston, R. C. (2009). Choosing an optimal method to combine p-values. *Statistics in Medicine* **28**, 1537–1553.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika* **103**, 609–624.

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N. et al. (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **33**, 700–709.

Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A. et al. (2007). AGEMAP: A gene expression database for aging in mice. *PLoS Genetics* **3**, e201.

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology* **24**, 1836–1841.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **22**, 170–185.

Zhang, H., Tong, T., Landers, J. and Wu, Z. (2020). TFisher: A powerful truncation and weighting procedure for combining p-values. *The Annals of Applied Statistics* **14**, 178–201.

Yusi Fang

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260, USA.

E-mail: yuf31@pitt.edu

Chung Chang
Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 80424, Taiwan.
E-mail: cchang@math.nsysu.edu.tw
George C. Tseng
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260, USA.
E-mail: ctseng@pitt.edu

(Received July 2022; accepted March 2023)