Smaller $p$-Values via Indirect Information

Peter Hoff

Department of Statistical Science, Duke University, Durham, NC

**ABSTRACT**

This article develops $p$-values for evaluating means of normal populations that make use of indirect or prior information. A $p$-value of this type is based on a biased frequentist hypothesis test that has optimal average power with respect to a probability distribution that encodes indirect information about the mean parameter, resulting in smaller $p$-value if the indirect information is accurate. In a variety of multiparameter settings, we show how to adaptively estimate the indirect information for each mean parameter while still maintaining uniformity of the $p$-values under their null hypotheses. This is done using a linking model through which indirect information about the mean of one population may be obtained from the data of other populations. Importantly, the linking model does not need to be correct to maintain the uniformity of the $p$-values under their null hypotheses. This methodology is illustrated in several data analysis scenarios, including small area inference, spatially arranged populations, interactions in linear regression, and generalized linear models. Supplementary materials for this article are available online.

1. Introduction

Many statistical procedures are built upon the evaluation of evidence that one or more unknown scalar quantities are not equal to some null value, say zero. A controversial yet ubiquitous measure of evidence that a scalar $\theta$ is not zero is the $p$-value, a function of the data that is uniformly distributed on $(0, 1)$ if $\theta = 0$. One way to construct a $p$-value is with a collection of tests that, for each value of $\alpha \in (0, 1)$, includes a single test with Type I error rate equal to $\alpha$. Given such a collection, the $p$-value can be defined as the smallest value of $\alpha$ for which the corresponding test rejects $H$, assuming the tests satisfy a monotonicity property (Dickhaus 2014). As such, a $p$-value will likely be small if the true value of $\theta$ is one for which the corresponding tests have a high probability of rejecting $H$.

If indirect or prior information about $\theta$ is available, then $p$-values derived from such tests will generally be smaller than a standard $p$-value that does not make use of the information. In this article, $p$-values are developed that correspond to optimal tests when the indirect information is summarized with a probability distribution over $\theta$. Specifically, it is shown that if the direct information about $\theta$ is from data $Y \sim N(\theta, \sigma^2)$, and the indirect information about $\theta$ is encoded with a normal distribution $\pi$, then the test statistic with optimal power on average with respect to $\pi$ may be derived analytically, and the corresponding $p$-value may be expressed very simply as $1 - |\Phi(Y/\sigma + 2\mu/\tau^2) - \Phi(-Y/\sigma)|$, where $\mu$ and $\tau^2$ are the mean and variance of $\pi$, and $\Phi$ is the standard normal cumulative distribution function (CDF). A similar $p$-value for the case that $\sigma^2$ is estimated from data is also derived. These $p$-values are uniformly distributed if $\theta = 0$, regardless of the values of $\mu$ and $\tau^2$.

We refer to such a $p$-value as being “frequentist, assisted by Bayes,” or FAB. It is frequentist in the sense that it has guaranteed sampling properties (uniformity under the null distribution), and it is Bayesian in the sense that the corresponding tests maximize expected power, on average across $\theta$-values with respect to a probability distribution. The basic idea of using a Bayesian criteria to select a frequentist procedure goes back at least to Pratt (1963), who constructed a confidence interval for a normal mean that has minimum expected width with respect to a distribution over the possible values of the mean. In fact, the FAB $p$-value may alternatively be derived by finding the smallest $\alpha$ for which Pratt’s $1 - \alpha$ confidence interval does not contain zero. Related to this, Good and Crook (1974) compared Bayes factors to a variety of other statistics for evaluating equiprobability of multinomial probabilities, and referred to the use of such statistics and consideration of average power as a “Bayes/non-Bayes compromise” (see also Good (1992) for other such compromises). More recently, Servin and Stephens (2007) used a Bayes factor to evaluate a global test of association between a phenotypic outcome and several genetic markers. A null distribution for their global test is obtained via permutation, as in this case the null hypothesis corresponds to an exchangeability assumption. Going the other direction, Wakefield (2009) and Benjamin and Berger (2019) studied how standard $p$-value criteria should be modified to produce inferences that resemble those that would be obtained using Bayes factors.

**CONTACT** Peter Hoff peter.hoff@duke.edu Department of Statistical Science, Duke University, Durham, NC 27708.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/RJASA.

© 2020 American Statistical Association
In addition to their use for measuring evidence, \( p \)-values are also used as inputs into statistical decision procedures that control various frequentist error rates. For example, rejecting the hypothesis that \( \theta = 0 \) whenever the \( p \)-value is below some threshold \( \alpha \) is a procedure that of course maintains a false rejection rate of \( \alpha \). In multiparameter analyses, \( p \)-values are often combined in ways to evaluate a global hypothesis (Birnbaum 1954; Heard and Rubin-Delanchy 2018), to adaptively estimate a threshold that maintains control of the false discovery rate (Benjamini and Hochberg 1995), or to maintain such a false discovery rate among multiple groups of hypotheses (Barber and Ramdas 2017). Frequentist multiparameter inference procedures such as these take as their input a list of \( p \)-values, typically without specifying how the \( p \)-values are constructed. In contrast, the focus of this article is on how, in multiparameter scenarios, adaptive FAB \( p \)-values that are smaller on average than standard \( p \)-values may be constructed by sharing information across an entire dataset. Specifically, while the FAB \( p \)-value can be as small as half the usual \( p \)-value derived from a uniformly most powerful unbiased (UMPU) test, to realize this gain the indirect information needs to be in accord with the actual value of \( \theta \). Scenarios where such accordance can be obtained statistically include multiparameter problems where the parameters are believed to be similar to one another in some way. For example, if the data consist of independent samples from each of several groups, then the indirect information for the parameter of one group may be inferred from the data of the others using a model for across-group heterogeneity of the parameters, which we refer to as a linking model. If the linking model is precise then the FAB \( p \)-values will be smaller than the UMPU \( p \)-values, on average across groups. If the linking model is diffuse, then the FAB \( p \)-values will be similar to the UMPU \( p \)-values, as the latter are a special case of the former. Importantly, the resulting FAB \( p \)-values are adaptive in that the parameters of the linking model are estimated from the data, and robust in the sense that the uniformity of the \( p \)-values under their null hypotheses does not depend on the linking model being correct.

The utility of indirect information for multiparameter inference problems has been well-recognized (Ghosh and Rao 1994; Efron 2010). While hierarchical models and empirical Bayes procedures provide a powerful and flexible class of methods for information-sharing in such problems, their performance guarantees typically hold globally, on average across parameter values rather than for each parameter individually. For parameter-specific inference, the basic strategy of using indirect information and a linking model to choose statistical procedures with parameter-specific frequentist guarantees was used previously in Yu and Hoff (2018), Hoff and Yu (2019), and Burris and Hoff (2019) to obtain adaptive versions of Pratt’s confidence interval for multiparameter settings. In the present article, the use of adaptive tests to construct a FAB \( p \)-value for each parameter is similar to the adaptive empirical likelihood ratio used by Storey (2007), which combines data from multiple experiments to form an empirically estimated significance threshold, and is then individually applied to each experiment. Storey’s empirical likelihood ratio corresponds in some sense to a semiparametric exchangeable linking model, while the parametric approach proposed here permits other types of information to be shared, such as spatial information or parameter-specific explanatory variables.

Other recently developed evidence measures that incorporate auxiliary information include covariate-assisted ranking and screening (Cai, Sun, and Wang 2019), the “skeptical \( p \)-value” (Held 2020), and the “\( s \)-value” (Grünwald, de Heide, and Koolen 2019). The first of these proposes a method for two-sample comparisons of multiple populations that estimates a relationship between effect size and an auxiliary population-level statistic, and uses any such relationship to obtain a potentially improved ranking of hypotheses by likely effect size, before applying a thresholding procedure. While similar to the approach presented here in that it infers relationships among parameter values, it differs in a variety of ways. In particular, the FAB approach presented here provides population-specific \( p \)-values that are individually uniformly distributed under each population-specific null hypothesis.

The skeptical \( p \)-value and the \( s \)-value are similar in that they are both related to sequential decision making, and both tend to be more conservative than a regular \( p \)-value. The skeptical \( p \)-value is obtained by comparing a test statistic from a replication study not to its null distribution, but to a conservative prior predictive distribution partly obtained from the original study. In contrast, the objective of the FAB \( p \)-value is to increase statistical power rather than reduce it. However, when used for evaluating a hypothesis in a replication study, the FAB \( p \)-value shares some properties of the one-sided skeptical \( p \)-value, in that it has more power when the true effect is in accord with the estimated effect from the original study. This is discussed in more detail at the end of Section 2.2. The \( s \)-value is a measure of evidence that facilitates sequential information accumulation, and is in some sense more conservative than a \( p \)-value. However, the objective of a sequential procedure based on \( s \)-values is on the accumulation of evidence against a hypothesis about a single population or parameter, whereas the utility of FAB \( p \)-values is in their ability to share information across multiple populations. However, FAB \( p \)-values are well-suited to situations where multiple populations are observed sequentially, as is described more fully in Section 5.

The FAB \( p \)-value is derived and studied in Section 2. Section 3 illustrates how adaptive FAB \( p \)-values may be constructed in two data analysis scenarios in which there is independent direct data for each parameter. These include small area inference using the Fay–Herriot model (Fay and Herriot 1979), and inference for a sequence of means using data from a hidden Markov process. The methodology is extended in Section 4 to situations where the direct data for each parameter may be dependent, such as estimates of linear regression coefficients.

This article primarily focuses on scenarios involving normally (or \( t \)) distributed test statistics and normal prior distributions, as in these cases the FAB tests and \( p \)-values have surprisingly simple forms. To accommodate more general data-analysis scenarios, approximate FAB \( p \)-values for asymptotically normal estimators are developed in Section 4, where it is proven that these \( p \)-values are asymptotically uniform under the null distribution. These approximate FAB \( p \)-values can be used, for example, to evaluate coefficients in widely used generalized
linear models such as logistic or Poisson regression. A discussion follows in Section 5.

2. FAB p-Values

2.1. Bayes-Optimal Tests and p-Values

Consider performing a level-α test of $H : \theta = 0$ based on the observation of $Y \sim N(\theta, \sigma^2)$, with $\sigma^2$ known. Suppose additionally that indirect information about $\theta$ is available, via prior knowledge or data independent of $Y$, that is encoded with a distribution having density $\pi$. The average power of a test function $f : \mathbb{R} \rightarrow [0, 1]$ with respect to $\pi$ is given by

$$
\mathbb{E}[f(Y)|\theta] = \int \int f(y)p(y|\theta)\pi(\theta)dyd\theta = \int f(y)\left(\int p(y|\theta)\pi(\theta)d\theta\right)dy = \int f(y)p_\pi(y)dy,
$$

where $p(y|\theta)$ is the $N(\theta, \sigma^2)$ density and $p_\pi$ is the marginal density of $Y$ under $\pi$.

According to our indirect information, the level-α test that has the highest probability of rejecting $H$ is the function $f_\alpha$ that maximizes (1) among all functions $f : \mathbb{R} \rightarrow [0, 1]$ that satisfy $\mathbb{E}[f(Y)|\theta] = \alpha$ when $\theta = 0$. The corresponding test has a frequentist Type I error rate of $\alpha$, but it is also the Bayes-optimal level-α test if $\pi$ is viewed as a prior density. As such, we describe this test as being “frequentist, assisted by Bayes,” or FAB.

By the Neyman–Pearson lemma, the FAB test is to reject $H$ if $p_\pi(y)/p_0(y)$ exceeds its upper $\alpha$ quantile under $H$, where $p_0(y)$ is the $N(0, \sigma^2)$ null density. If $\pi$ is a normal density, then the test has a simple analytic form.

**Theorem 1.** The level-α FAB test of $H$ corresponding to a $N(\mu, \tau^2)$ distribution for $\theta$ rejects when $|Y + \mu\sigma^2/\tau^2| > c$, where $c$ is the solution to $[\Phi(c/\sigma) + \Phi(\mu/\tau^2)]/2 = 1 - \alpha/2$ and $\Phi$ is the standard normal CDF.

**Proof.** If $Y|\theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$ then the marginal distribution of $Y$ is $N(\mu, \sigma^2 + \tau^2)$. The likelihood ratio is then

$$
\frac{p_\pi(y)}{p_0(y)} = \exp(-[(y^2 - 2\mu y + \mu^2)/(\sigma^2 + \tau^2) - y^2/\sigma^2]/2).
$$

Some algebra shows that this is monotonically increasing in $|y + \mu\sigma^2/\tau^2|$.

Different prior distributions for $\theta$ typically lead to different Bayes-optimal test statistics. However, one particular prior distribution, used in an example in Section 4, leads to the same test statistic as a normal prior distribution.

**Corollary 1.** The level-α FAB test of $H$ when the prior distribution for $\theta$ is a mixture of a $N(\mu, \tau^2)$ distribution with a point-mass at zero is the same as the test when the prior distribution is $N(\mu, \tau^2)$.

**Proof.** Let $p_\mu(y)$ be the marginal density of $Y$ under the mixture prior distribution with weight $w$ on the normal component. Then $p_\mu(y) = w p_\pi(y) + (1-w)p_0(y)$, where $p_\pi$ and $p_0$ are as in the proof of Theorem 1. The likelihood ratio is $|w p_\pi(y) + (1-w)p_0(y)|/p_0(y)$, which simplifies to $w p_\pi(y)/p_0(y) + 1-w$. This is monotonically increasing in $p_\pi(y)/p_0(y)$, the likelihood ratio when $w = 1$, and so the tests are the same.

The FAB $p$-value function is the function that maps potential values $y$ of $Y$ to the smallest value of $\alpha$ such that the FAB level-α test rejects when $Y = y$. It is also the probability that $|Y + \mu\sigma^2/\tau^2| > |y + \mu\sigma^2/\tau^2|$ when $Y \sim N(0, \sigma^2)$ (see, e.g., Dickhaus 2014, chap. 2). The functional form of the FAB $p$-value is quite simple, and is easily derived using this latter characterization.

**Theorem 2.** For evaluating $H : \theta = 0$ from $Y|\theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$, the FAB $p$-value function may be written as $1 - |\Phi((y + 2\mu)/\sigma) - \Phi((-y)/\sigma)|$ where $a = \mu\sigma^2/\tau^2$, or as $p(z, b) = 1 - |\Phi(z - b) - \Phi(-z)|$, where $z = y/\sigma$ and $b = 2\mu\sigma^2/\tau^2$.

**Proof.** The $p$-value is

$$
\text{Pr}(|Y + \mu\sigma^2/\tau^2| > |y + \mu\sigma^2/\tau^2|)
\begin{align*}
\text{Pr}(Z + b/2 > |z + b/2|) &= \text{Pr}(Z + b/2 < |z + b/2|) + \text{Pr}(Z + b/2 > |z + b/2|)
\end{align*}
\begin{align*}
&= \Phi(-|z + b/2| - b/2) + 1 - \Phi(|z + b/2| - b/2).
\end{align*}
$$

First suppose $z + b/2 > 0$. In this case, the $p$-value is

$$
\Phi(-z - b) + 1 - \Phi(z) = 1 - \Phi(z + b) + \Phi(-z).
$$

Note that $z + b/2 > 0$ implies $z + b > -z$, and so the $p$-value can be written $1 - |\Phi(z + b) - \Phi(-z)|$. Now suppose $z + b/2 < 0$, so that the $p$-value is

$$
\Phi(z) + 1 - \Phi(-z - b) = 1 - \Phi(-z) + \Phi(z + b).
$$

In this case, $z + b < -z$ so the $p$-value may be written as $1 - |\Phi(z + b) - \Phi(-z)|$, which is the same as in the case $z + b/2 > 0$. So in either case, the $p$-value is $1 - |\Phi(z + b) - \Phi(-z)|$.

Note that the standard $p$-value based on the UM PU test statistic $|Z|$ can be written as $p(Z, 0) = 1 - |\Phi(Z) - \Phi(-Z)|$, and so is a special case of the FAB $p$-value with $\mu = 0$. For nonzero $\mu$, if the indirect information is accurate in the sense that $\theta$ and $\mu$ are of the same sign, then $Y$ and $2a$ (or $Z$ and $b$) will likely have the same sign, making it probable that the absolute difference $|\Phi(Z + b) - \Phi(-Z)|$ is larger than $|\Phi(Z - (b))|$, thereby making the FAB $p$-value smaller than the UM PU $p$-value. However, if the sign of $\mu$ does not match that of $\theta$, then this absolute difference will likely be smaller than $|\Phi(Z) - \Phi(-Z)|$, making the FAB $p$-value larger than the UM PU $p$-value. This can be made more precise as follows:

**Theorem 3.** Let $Z \sim N(\theta, 1)$. Then

$$
\text{Pr}(p(Z, b) < p(Z, 0)/\theta) > \Phi(\text{sign}(b) \times \theta).
$$

**Proof.** If $b$ is positive then $Z > 0$ implies $p(Z, b) < p(Z, 0)$, so $\text{Pr}(p(Z, b) < p(Z, 0)/\theta) > \text{Pr}(Z > 0/\theta) = \Phi(\theta)$. Similarly, if $b$ is negative then $\text{Pr}(p(Z, b) < p(Z, 0)/\theta) > \Phi(-\theta)$. Combining these gives the result.
Figure 1 plots the UMPU and FAB p-value functions for a variety of values of $b$. The FAB p-value functions are symmetric around $-b/2 = -\mu \sigma / \tau^2$, and so if $\mu$ is positive these p-values will be lower than the UMPU p-value if $z$ is positive, and can even be lower for a range of negative $z$-values. This is to be expected—the FAB p-value will be lower if $z$ and $\mu$ match, but can be higher otherwise. Like the UMPU p-value, for each finite value of $b$ the FAB p-value decreases to zero as $|z|$ increases. As shown in the right-side panel of Figure 1, for large $|z|$ the FAB p-value can be as small as half of the UMPU p-value, if $z\mu > 0$. For example, for the nonzero values of $b$ shown in the figure, when the UMPU p-value is 0.10, the FAB p-values are close to 0.05. However, the ratio of the FAB p-value to the UMPU p-value can be unboundedly large if there is a mismatch between the data and the indirect information (that is, if $z\mu < 0$). These comments are summarized as follows:

**Theorem 4.** Let $p(z, b) = 1 - |\Phi(z + b) - \Phi(-z)|$. Then $p(z, 0)$ is the UMPU p-value function, and for $b > 0$,
1. $\lim_{z \to \infty} p(z, b)/p(z, 0) = 1/2$;
2. $\lim_{z \to -\infty} p(z, b)/p(z, 0) = \infty$.

**Proof.** For positive $z$ and $b$ the ratio of the p-values is
\[
p(z, b)/p(z, 0) = \frac{1 - \Phi(z + b) + \Phi(-z)}{1 - \Phi(z) + \Phi(-z)} = \frac{2\Phi(-z)}{1 - \Phi(z) + \Phi(-z)} = \frac{1}{2} \left( 1 + \frac{1 - \Phi(z + b) + \Phi(-z)}{1 - \Phi(z) + \Phi(-z)} \right),
\]
which converges to 1/2 as $z \to \infty$ by L'Hôpital's rule. For $z < -b/2 < 0$ the ratio is
\[
\frac{1 + \Phi(z + b) - \Phi(-z)}{1 + \Phi(z) - \Phi(-z)} = \frac{1}{2} \frac{\Phi(z + b) + \Phi(z)}{\Phi(z)} = \frac{1}{2} \left( 1 + \frac{\Phi(z + b)}{\Phi(z)} \right),
\]
which diverges to infinity as $z \to -\infty$.

The FAB p-value is based on a test that has better performance on one side of the real line than the other, suggesting a relationship to a one-sided test. Indeed, one-sided p-values are limiting cases of FAB p-values.

**Theorem 5.**
1. $\lim_{b \to \infty} p(z, b) = 1 - \Phi(z)$;
2. $\lim_{b \to -\infty} p(z, b) = \Phi(z)$.

In other words, for large positive $b$ the FAB p-value approximates the p-value from the uniformly most powerful tests of $H : \theta \leq 0$ versus $K : \theta > 0$, and vice versa for large negative $b$.

### 2.2. Null and Alternative Distributions of FAB p-Values

Like UMPU p-values, FAB p-values are uniformly distributed under the null hypothesis.

**Theorem 6.** Let $Z \sim N(0, 1)$. Then for any $b \in \mathbb{R}$, $\Pr(1 - |\Phi(Z + b) - \Phi(-Z)| \leq u) = u$.

**Proof.** This is guaranteed by the fact that the p-value is the probability of obtaining a value of $|Z + b/2|$ as or more extreme than the one observed, and that the null distribution of the test statistic is continuous (Dickhaus 2014, chap. 2). Alternatively, the result can be shown directly as follows: Since the FAB p-value function $p(z, b)$ is symmetrically decreasing from $z = -b/2$, the probability that $p(Z, b)$ is less than $u$ is the probability that $Z$ lies outside the interval $(-b/2 - c, -b/2 + c)$, where $c > 0$ satisfies $p(-b/2 + c, b) = u$. Plugging this into the p-value function gives
\[
u = 1 - |\Phi(b/2 + c) - \Phi(b/2 - c)| = 1 - \Phi(b/2 + c) + \Phi(b/2 - c) = \Phi(-b/2 - c) + [1 - \Phi(-b/2 + c)],
\]
so the probability that $Z$ lies outside of $(-b/2 - c, -b/2 + c)$, and that $p(Z, b) < u$, is $u$. 

The CDF and density of the FAB p-value under nonzero values of $\theta$ may be of use for power calculations and theoretical study. To find these quantities, recall that the FAB p-value function is symmetric around $-b/2$, and so each possible value $u$ of the p-value other than 1 is achieved by two values of $z$, say $z_l$ and $z_h$, which satisfy $z_l < -b/2 < z_h$. These values are the solutions to the equations
\[
\Phi(z_h + b) - \Phi(-z_h) = 1 - u, \quad (2)
\]
\[
\Phi(-z_l) - \Phi(z_l + b) = 1 - u. \quad (3)
\]
Since the p-value is monotonically increasing on $z < -b/2$ and decreasing on $z > -b/2$, the p-value will be less than $u$ if $z < z_l$ or $z > z_h$. Now let $U = p(Z, b)$ where $Z \sim N(\theta, 1)$. Then the probability that $U \leq u$ is the probability that $Z < z_l$ or $Z > z_h$, which is

$$F_U(u) = \Pr(U \leq u) = \Phi(z_l - \theta) + 1 - \Phi(z_h - \theta).$$

Note that $z_l$ and $z_h$ are functions of $u$, and can be computed with a zero-finding algorithm. Taking the derivative of $F_U$ gives the density of $U$,

$$f_U(u) = \frac{dF_U(u)}{du} = \phi(z_l - \theta) \frac{dz_l}{du} - \phi(z_h - \theta) \frac{dz_h}{du},$$

where $\phi$ is the standard normal density. The derivatives on the right side of the equation can be found by implicit differentiation of (2) and (3), giving

$$f_U(u) = \frac{\phi(z_l - \theta)}{\phi(z_l + b)} + \frac{\phi(z_h - \theta)}{\phi(z_h + b)}.$$

Functions for computing the CDF and density of $U = p(Z, b)$ are available in the supplementary materials for this article.

Figure 2 displays the density of $p(Z, b)$ for various values of $b$ and under various alternative values of $\theta$. If $b$ and $\theta$ are both positive (or are both negative) then the FAB p-value density is more concentrated on small values than the density of the UMPU p-value. However, if there is a mismatch between $\theta$ and $b$ then the FAB p-value is expected to be larger. For example, if $\theta$ is negative then as $b$ increases, the distribution of the p-value converges to a uniform distribution. This property is similar to that of the "one-sided skeptical p-value" proposed by Held (2020), whereby evidence against a zero value of $\theta$ in a replication study is only assessed in the same direction as the estimate from the prior original study, and so there is no power to detect deviations from the null hypothesis that are in conflict with the prior information. The FAB p-value is less extreme in this regard, in that the corresponding test will still have power to detect nonzero $\theta$ values that are in conflict with the prior information $b$, as long as they are sufficiently large. In particular, the FAB p-value converges to zero as $|z| \to \infty$ for any fixed $b$.

Since a p-value can be used to perform a level-$\alpha$ test (by rejecting when the p-value is below $\alpha$), it is of interest to evaluate a p-value in terms of the power of the corresponding test. The right most panel of Figure 2 plots this power as a function of $\theta$ and $b$ for $\alpha = 0.05$ and $\alpha = 0.10$. Additionally, a simple way to summarize the power over a range of $\alpha$-values is by computing the expected power, on average with respect to the uniform distribution on $\alpha$, which turns out to be one minus the expectation of the p-value, or EPV (Sackrowitz and Samuel-Cahn 1999; Vexler 2020). This average power is given by the top set of lines in the right-most panel of Figure 2. As can be seen from the figure, these power functions are all eventually increasing as $|\theta|$ increases for all values of $b$, but they increase most rapidly as $\theta$ increases in the direction of $b$.

### 2.3. FAB p-Values for Nonnormal Populations and Statistics

An FAB p-value for testing hypotheses about nonnormal families may be constructed analogously. For example, consider a family of distributions $P = \{P_\theta : \theta \in \Theta\}$, for which prior information about $\theta$ is available in the form of a prior density $\pi$. By the same argument as that preceding Theorem 1, the Bayes-optimal test of $H : Y \sim P_0$ is to reject when $p_\pi(Y)/p_\pi(\tilde{Y})$ is large, where $P_0$ is the density of $P_0$ and $p_\pi$ is the marginal density of $\tilde{Y}$ under $\pi$. Having observed $Y = y$, the corresponding FAB p-value could be obtained by computing the probability that $p_\pi(Y)/p_\pi(\tilde{Y}) > p_\pi(y)/p_\pi(\tilde{Y})$, where $\tilde{Y} \sim P_0$. However, while simple in theory, such a testing procedure can be computationally tedious. In many cases (including the case of t-distributions described below), $p_\pi$ might not have a closed form expression and would need to be approximated numerically. Additionally, the null distribution of $p_\pi(Y)/p_\pi(\tilde{Y})$ is likely not to be a standard distribution and would have to be approximated as well, possibly using a Monte Carlo method that would require a numerical approximation to $p_\pi$ at each iteration.

However, in many cases there is an alternative, easy-to-implement p-value procedure that allows for incorporation of prior information, obtained by generalizing the procedure for

---

**Figure 2.** Left and center panel: Densities of p-values for $b \in \{0, 0.25, 0.5, 0.75, 1\}$, and $\theta = 1$ (left) and $\theta = -1$ (center), given by lines of increasing lightness. The density corresponding to $b = 0$ (the UMPU p-value) is displayed with a thick black line. Right panel: Power of the FAB test as a function of $\theta$ and $b$ for $\alpha = 0.05$ (lowest set of lines) $\alpha = 0.10$ (middle set of lines), and on-average across $\alpha$-values (top set of lines), or equivalently, one minus the expected p-value (EPV).
The proof is the same as that of Theorem 6 but with \( \Phi \) replaced by \( F_0 \). This theorem says that for evaluating the hypothesis \( T \sim F_0 \), the \( p \)-value function \( p(T, b) = 1 - |F_0(T + b) - F_0(-T)| \) is uniformly distributed under the null hypothesis, for any choice of \( b \), and corresponds to a hypothesis test that rejects when \( |T + b/2| \) is large.

For example, suppose \( F_0 \) is symmetric around zero and consider the location family \( \{ F_0 : \theta \in \mathbb{R} \} \) where \( F_0(T) = F_0(T - \theta) \). Then \( p(T, b) \) can be used as a \( p \)-value for evaluating \( H : \theta = 0 \). Additionally, the following generalization of Theorem 3 indicates that we test a best based on \( p(T, b) \) will be more powerful than one based on \( p(T, 0) \) if \( \theta \) and \( b \) have the same sign:

**Theorem 8.** Let \( T \sim F_0 \). Then \( Pr(p(T, b) < p(T, 0)|\theta) > F_0(\text{sign}(b) \times \theta) \).

The proof is the same as that of Theorem 3, but with \( \Phi \) replaced by \( F_0 \). Additional results, analogous to those in the previous subsection for normal models, hold for such location families where \( F_0 \) is symmetric about zero.

What remains is to select a value of \( b \) based on prior information. Suppose \( F_0 \) has variance 1 and that \( Y = \theta + \sigma \epsilon \), where \( \epsilon \sim F_0 \). Then under the null hypothesis \( H : \theta = 0 \), \( Y/\sigma \sim F_0 \) and the \( p \)-value \( p(Y|\sigma, b) \) is uniformly distributed for any value of \( b \). Suppose the prior mean and variance of \( \theta \) are \( \mu \) and \( \tau^2 \), respectively. The choice of \( b = 2\mu\sigma/\tau^2 \), the same value as in the normal case, can be justified as follows: First, note that via Bayes’ rule, the marginal density of \( Y \) can be expressed as \( p_\sigma(Y) = \pi(0)p_\sigma(0)/\pi(0)y \), where \( \pi(0) \) and \( \pi(0)y \) are the prior and conditional densities of \( \theta \) at the point zero. The likelihood ratio \( p_\sigma(Y)/p_\sigma(0) \) can therefore be expressed as \( \pi(0)/\pi(0)y \), and so a value of \( y \) that makes the likelihood ratio small is one that makes \( \pi(0)y \) large.

Intuitively, it makes sense to reject \( H \) if \( Y \) is far from a value \( \tilde{y} \) for which \( p_\sigma(\tilde{y})/p_\sigma(0) \) is small, or equivalently \( \pi(0)y \) is large. To find such a value, let \( \hat{\theta}(\tilde{y}) \) be the linear estimator \( \hat{\theta}(\tilde{y}) = (\tau^2\tilde{y} + \sigma^2\mu)/\tau^2 + \sigma^2 \). While not in general the Bayes estimator under nonnormal models, it does minimize the Bayes risk among all linear estimators of the form \( cY + d \). As such, \( \pi(0)y \) should be large if \( \tilde{y}(\hat{\theta}) = 0 \), or equivalently, if \( \tilde{y} = -\mu\sigma^2/\tau^2 \). Rejecting \( H \) when \( Y \) is far from \( \tilde{y} \) thus leads to rejecting when \( |Y + \mu\sigma^2/\tau^2| \) is large. Following the same argument as in the proof of Theorem 2, the resulting \( p \)-value is \( p(Y/\sigma, 2\mu\sigma/\tau^2) = 1 - |F_0(Y/\sigma + 2\mu\sigma/\tau^2) - F_0(0)| \).

A FAB \( p \)-value can also be constructed in cases where the scale parameter is unknown. A particularly useful case is a FAB \( p \)-value based on a \( t \)-statistic. Specifically, suppose \( Y_1, \ldots, Y_n \sim \text{iid } N(\theta, \sigma^2) \) with \( \sigma^2 \) unknown and let \( S \) be the sample standard deviation. If \( \theta = 0 \) and \( n \) is large then \( T = \sqrt{n}\tilde{Y}/S \) will be approximately standard normal and \( 1 - |\Phi(T + b) - \Phi(T)| \) will be approximately uniform. If the sample size is insufficient to justify this approximation, then a \( t_{n-1} \) null distribution should be used for construction of the \( p \)-value. Like the standard normal null distribution, \( t \) distributions are continuous and symmetric about zero, and a \( p \)-value function may be defined as described by Theorem 7. For example, suppose that we have a normal estimator \( \hat{\theta} \) of \( \theta \), so that \( \hat{\theta} \sim N(\theta, \sigma^2) \), where \( \theta \) and \( \sigma^2 \) are both unknown. Suppose additionally that we have an estimator \( \hat{\sigma}^2 \) of \( \sigma^2 \) for which \( \hat{\sigma}^2/\hat{\sigma}^2 \sim \chi^2_{\nu} \), and that is independent of \( \hat{\theta} \). Then \( \hat{\theta}/\hat{\sigma} \sim t_\nu \) if \( \theta = 0 \). Considering the form of the FAB \( p \)-value in the known-variance case, it seems reasonable to use as a \( p \)-value

\[
1 - |F_\nu(\hat{\theta}/\hat{\sigma} + 2\mu\hat{\sigma}/\tau^2) - F_\nu(-\hat{\theta}/\hat{\sigma})|, \tag{4}
\]

where \( F_\nu \) is the CDF of the \( t_\nu \) distribution, \( \mu \) and \( \tau^2 \) are the mean and variance of a distribution describing the indirect information about \( \theta \), and \( \hat{\sigma} \) is a prior guess as to the value of \( \sigma \). This \( p \)-value is uniformly distributed under the null distribution \( \hat{\theta}/\hat{\sigma} \sim t_\nu \), regardless of the values of \( \mu, \tau^2 \), and \( \hat{\sigma} \). Furthermore, this \( p \)-value is easy to calculate and is numerically stable insofar as the CDF of the \( t \)-distribution may be calculated.

However, this \( p \)-value function does not correspond exactly to a Bayes-optimal test. Letting \( \nu = \hat{\theta}/\hat{\sigma} \), a Bayes-optimal level-\( \alpha \) test of \( H : \theta = 0 \) is one that rejects when the ratio \( p_\nu(T)/p_\nu(0) \) exceeds its \( 1 - \alpha \) quantile under the null distribution, where \( p_\nu \) is the \( t \)-density function and \( p_\nu \) is the marginal density of \( T \) under a prior distribution for \( \theta \) and \( \sigma^2 \). To see the difficulty in using a formally Bayes-optimal test to construct a \( p \)-value function in this case, consider a \( N(\mu, \tau^2) \) prior distribution for \( \theta \) and an arbitrary prior distribution for \( \sigma^2 \). Marginally over the prior distribution for \( \theta \) but conditional on \( \sigma^2 \), we have \( \nu \sim N(\mu, \sigma^2 + \tau^2) \), and the resulting marginal distribution for \( T = \hat{\theta}/\hat{\sigma} \) is a noncentral \( t \)-distribution with \( \nu \) degrees of freedom and noncentrality parameter \( \mu/\sqrt{\sigma^2 + \tau^2} \), where \( c = \sqrt{\sigma^2 + \tau^2} \). Let \( p_\nu(t|\sigma^2) \) denote the corresponding density for \( T \). The marginal density \( p_\nu(T) \), needed to construct the Bayes-optimal test, is obtained by integrating \( p_\nu(t|\sigma^2) \) with respect to the prior distribution for \( \sigma^2 \). To obtain the corresponding \( p \)-value, we would first have to compute \( p_\nu(t)/p_\nu(0) \) where \( t \) is the observed value of \( \hat{\theta}/\hat{\sigma} \), and then compute the probability of obtaining a more extreme value under the null hypothesis, that is, \( Pr(p_\nu(T)/p_\nu(0) > p_\nu(t)/p_\nu(0)) \). Additional details are required to evaluate the noncentral \( t \)-distribution, numerical approximation of which is often poor. For these reasons, the \( p \)-value function (4) is recommended unless there is a strong desire for a more formal incorporation of specific prior information about \( \sigma^2 \).

## 3. Adaptive \( p \)-Values From Independent Normal Data

### 3.1. FAB \( p \)-Values via Indirect Information

In this section the utility of the FAB \( p \)-value is broadened by the development of statistical methods with which indirect information may be summarized and incorporated into
the p-value. Generalizing the approach taken by Yu and Hoff (2018) and Burris and Hoff (2019) in the context of confidence interval construction, the proposed method is to use a statistical model to relate a parameter \( \theta \) of interest to other data that are available, in which case the indirect information itself may be stochastic. This does not affect the uniformity of the FAB p-value, as long as the summary of the indirect information is statistically independent of the direct information.

**Corollary 2.** Let \( Z \) and \( b \) be independent random variables with \( Z \sim N(0,1) \). Then \( \text{Pr}(1 - |\Phi(Z + b) - \Phi(-Z)| < u) = u \).

**Proof.** By independence, the conditional probability given \( b \) that the p-value is less than \( u \) is equal to \( u \) for each \( b \), and so it equal to \( u \) marginally over values of \( b \) as well. \( \square \)

Specifically, we develop adaptive FAB p-values in multiparameter settings, where indirect information about a given parameter is derived from the direct information about other parameters. In this way, the FAB p-value procedure can adapt to patterns among the parameters. To start, suppose a data vector \( Y \) with uncorrelated elements will be sampled from a multivariate normal population, so that the **sampling model** for the data is

\[
Y \sim N_p(\theta, \Sigma),
\]

where \( \theta \in \mathbb{R}^p \) is unknown and for now assume \( \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2) \) is known. For each \( j = 1, \ldots, p \) we will construct a FAB p-value \( p(Y_j/\sigma_j, b_j) = 1 - |\Phi(Y_j/\sigma_j + b_j) - \Phi(-Y_j/\sigma_j)| \), where \( Y_j \) is the \( j \)th element of \( Y \) and \( b_j \) is independent of \( Y_j \), so that \( p(Y_j/\sigma_j, b_j) \) is uniformly distributed under the null hypothesis \( H_j : \theta_j = 0 \).

Indirect information about \( \theta_j \) that is independent of \( Y_j \) may be derived from the other entries of \( Y \). To facilitate generalization to the correlated case considered in the next section, we write these elements as \( G_j^\top Y \), where \( G_j \) is the \( p \times (p - 1) \) matrix obtained by deleting the \( j \)th column from the \( p \times p \) identity matrix. Then \( G_j^\top Y \sim N_{p-1}(G_j^\top \theta, G_j^\top \Sigma G_j) \), and \( G_j^\top Y \) is statistically independent of \( Y_j \). By **Corollary 2**, a p-value \( p(Y_j/\sigma_j, b_j) \) is uniformly distributed under \( H_j \) for any statistic \( b_j \) that is a function of \( G_j^\top Y \).

While \( G_j^\top Y \) does not contain direct information about \( \theta_j \), it does contain information about \( G_j^\top \theta \), and so \( G_j^\top Y \) can be used to provide information about \( \theta_j \) if there are relationships among the entries of \( \theta \). Consider a Gaussian **linking model** to describe these relationships, of the form

\[
\theta \sim N_p(\mu, \Psi).
\]

Straightforward linear algebra shows that the conditional distribution of \( \theta \) given \( G_j^\top Y \) is \( N_p(m, V) \), where

\[
V = [\Psi^{-1} + G_j^\top (\Sigma G_j)^{-1} G_j^\top]^{-1},
\]

\[
m = V[\Psi^{-1} \mu + G_j^\top (\Sigma G_j)^{-1} G_j^\top Y].
\]

In some cases it may be easier to compute these as \( V = \Psi - \Psi G_j [\Psi^2 - \Psi^2]^{-1} \Psi^2 \Sigma G_j \) and \( m = \mu + \Psi G_j [\Psi^2 - \Psi^2]^{-1} \Psi^2 G_j (Y - \mu) \).

To the extent that the linking model is believed, the conditional distribution of \( \theta_j \) given \( G_j^\top Y \) is therefore \( \theta_j | G_j^\top Y \sim N(m_j, V_{jj}) \), where \( m_j \) and \( V_{jj} \) denote elements \( j \) and \( [j, j] \) of \( m \) and \( V \), respectively. This distribution quantifies the indirect information about \( \theta_j \): It is the distribution of \( \theta_j \) conditional on the data \( G_j^\top Y \) that is independent of \( Y_j \) under the sampling model (5), which in turn provides indirect information about \( \theta_j \) via the relationships among the elements of \( \theta \) based on the linking model (6).

Even if the linking model is not believed, it may still be used as a way to construct a data-adaptive p-value that will be uniformly distributed if \( H_j \) is true, as long as \( m_j \) and \( V_{jj} \) are statistically independent of \( Y_j \) under the sampling model (5). The performance of such a p-value under alternatives to \( H_j \) will depend on the extent to which the linking model is a good representation of the heterogeneity of the elements of \( \theta \). The accuracy of this representation may be improved by estimating \( \mu \) and \( \Psi \) from the data itself. In the applications we consider, \( \mu \) and \( \Psi \) will depend on some lower dimensional parameter, say \( \gamma \). This parameter may be estimated using hierarchical modeling methods, as the sampling model for the indirect information, \( G_j^\top Y \sim N_{p-1}(G_j^\top \theta, G_j^\top \Sigma G_j) \), together with the induced linking model \( G_j^\top \theta \sim N_{p-1}(G_j^\top \mu, G_j^\top \Psi G_j) \) constitute a Gaussian mixed effects model. For example, the maximum likelihood estimator of \( \gamma \) based on data independent of \( Y_j \) is the maximizer in \( \gamma \) of the density of the marginal distribution of \( G_j^\top Y \) if \( G_j^\top Y \sim N_{p-1}(G_j^\top \mu, G_j^\top \Psi G_j) \).

Now we are in a position to apply the results of the previous section. By **Theorem 1** the Bayes-optimal test of \( H_j \) given the indirect information \( G_j^\top Y \) rejects when \( |Y_j/\sigma_j + m_j/\sigma_j| \) is large, and by **Theorem 2** the Bayes-optimal p-value is given by \( p(Y_j/\sigma_j, b_j) \) where \( b_j = 2m_j/\sigma_j \). However, as \( m_j \) and \( \sigma_j \) depend on the unknown parameter \( \gamma \) via \( \mu_j \) and \( \Psi_j \), we instead use estimates \( \tilde{m}_j \) and \( \tilde{\sigma}_j \) based on \( \mu_j \) and \( \Psi_j \), where \( \tilde{\gamma} \) is obtained from \( G_j^\top Y \). Since \( G_j^\top Y \) is independent of \( Y_j \), \( \tilde{b}_j = 2\tilde{m}_j/\tilde{\sigma}_j \) is also independent of \( Y_j \) and so \( 1 - |\Phi(Y_j/\sigma_j + b_j) - \Phi(-Y_j/\sigma_j)| \) is uniformly distributed if \( \theta_j = 0 \) by **Corollary 2**.

If \( \sigma^2_j \) is not known but a high-precision estimator is available, then the value of the estimator can be plugged into the formula for the FAB p-value. Alternatively, if an estimator \( \tilde{\sigma}_j^2 \) is available that is independent of \( Y_j \) and for which \( v \tilde{\sigma}_j^2/\sigma_j^2 \sim \chi^2_v \), then by **Theorem 7** the p-value \( 1 - F_v(Y_j/\tilde{\sigma}_j + \tilde{b}_j) - F_v(-Y_j/\tilde{\sigma}_j) \) is uniformly distributed if \( \theta_j = 0 \) and \( \tilde{b}_j \) is independent of \( Y_j/\tilde{\sigma}_j \), where \( F_v \) is the CDF of the t-distribution with \( v \) degrees of freedom. Based on the discussion at the end of **Section 2**, we use \( \tilde{b}_j = 2\tilde{m}_j/\tilde{\sigma}_j \) if \( \tilde{\sigma}_j \) is estimated from \( G_j^\top Y \) as described in the previous paragraph. Note that an additional estimator \( \tilde{\sigma}_j \) of \( \sigma_j \) is also required, which must be independent of \( Y_j/\tilde{\sigma}_j \) for exact uniformity of the p-value to be maintained when \( \theta_j = 0 \). Availability of such an estimator will depend on the particular application. For example, if estimates of the \( \sigma^2_j \)'s for which \( k \neq j \) are also available, then \( \tilde{\sigma}_j^2 \) may be based on them.
### 3.2. Example: Small Area Inference With the Fay–Herriot Model

The 2002 Educational Longitudinal Study (ELS) gathered data on a sample of U.S. high schools and their students. From each participating school, a small sample of 10th grade students were selected and given a survey and a standardized reading exam. For \( p = 684 \) schools the sample size was 2 students or more, and among these the median school-level sample size was 21 and the maximum sample size was 50. In this section, we use adaptive FAB \( p \)-values to evaluate for each school the evidence that their school-specific mean score on the reading exam deviates from a particular national average value.

Let \( \bar{Y}_j \) and \( \sigma_j^2 \) be the sample mean and variance of the \( n_j \) reading test scores of students sampled from school \( j \). A Gaussian within-school sampling model implies that the vector \( \bar{Y} \) of school-specific sample means is normally distributed, with

\[
\bar{Y} \sim N_p(\theta, \text{diag}(\sigma_1^2/n_1, \ldots, \sigma_p^2/n_p)),
\]

where \( \theta \in \mathbb{R}^p \) is the vector of “true” school-specific means, meaning that \( \theta_j \) is the average exam score had all students in school \( j \) participated in the study. The sampling model also implies that \( (n_j - 1)\hat{\sigma}_j^2/\sigma_j^2 \sim \chi^2_{n_j - 1} \), independently for \( j = 1, \ldots, p \) and independent of \( \bar{Y} \).

In this numerical illustration, we construct \( p \)-values for evaluating the hypothesis that \( \theta_j = 50 \), a score that corresponds to the intended national average on the exam. Assuming the Gaussian within-school sampling model, the statistic \( T_j = \bar{Y}_j - 50/\hat{\sigma}_j \) has a \( \chi^2_{n_j - 1} \) distribution under the null hypothesis \( H_j: \theta_j = 50 \). Letting \( v_j = n_j - 1 \), the \( p \)-value \( 1 - |F_{\chi^2_{n_j - 1}}(T_j + \hat{\beta}_j) - F_{\chi^2_{n_j - 1}}(-T_j)| \) will be uniformly distributed under the null hypothesis as long as \( \hat{\beta}_j \) is independent of \( T_j \). The usual UMPU \( p \)-value based on \( T_j \) is obtained by setting \( \hat{\beta}_j = 0 \). An adaptive FAB \( p \)-value for each school \( j \) can be constructed by setting \( \hat{\beta}_j \) to a value that utilizes indirect information from schools other than \( j \). We do this via a Gaussian linking model for the \( \theta_j \)’s of the form

\[
\theta \sim N_p(X\beta, \tau^2 I),
\]

where \( X \) is a \( p \times q \) matrix of observed school-level characteristics, and \( \beta \) and \( \tau^2 \) are unknown. For these data, the characteristics \( x_j \) of school \( j \) includes a numeric measure of the number of students in school \( j \) who participated in a free lunch program, the total enrollment of school \( j \), and seven indicator variables that encode three categorical variables: school type (public, Catholic, or other private), region of the country (West, Midwest, South and East), and urbanicity (urban, suburban, rural). Together, a sampling model such as \( (9) \) and a linking model such as \( (10) \) are sometimes referred to as a Fay–Herriot model (Fay and Herriot 1979), a linear mixed-effects model that is often used in the small area estimation literature to obtain stable estimates for each of many groups or “areas” by sharing information across groups (Ghosh and Rao 1994). Here we are only using the linking model to obtain a value of \( \hat{\beta}_j \) with which the FAB \( p \)-value for group \( j \) is constructed. Validity of the linking model is not necessary for the FAB \( p \)-value to be uniformly distributed under the null hypothesis \( H_j \).

Based on the discussion in Section 3.1, we set \( \hat{\beta}_j = 2(\hat{\beta}^\top x_j - 50)/(\hat{\sigma}^2/\sqrt{n_j})/\tau^2 \), where \( \hat{\beta} \), \( \tau^2 \) and \( \hat{\sigma}^2 \) are the maximum likelihood estimators (MLEs) obtained by fitting the Fay–Herriot hierarchical model (with a common within-school variance) to the data from schools other than \( j \). Since \( \hat{\beta}_j \) is statistically independent of the sample from school \( j \), the FAB \( p \)-value \( 1 - |F_{\chi^2_{n_j - 1}}(T_j + \hat{\beta}_j) - F_{\chi^2_{n_j - 1}}(-T_j)| \) is uniformly distributed if \( \theta_j = 50 \), even if the linking model is incorrect. Figure 3 describes some aspects of this data analysis. The left-most panel plots the direct estimate \( \bar{Y}_j \) of each \( \theta_j \) versus the indirect estimate \( \hat{m}_j \) obtained from the Fay–Herriot model. The positive correlation between these two estimates suggests that the indirect information can be of use for school-level inference. The middle panel of the figure plots the FAB \( p \)-values versus the UMPU \( p \)-values corresponding to UMPU \( t \)-tests applied to each school individually. The FAB \( p \)-value is smaller than the UMPU \( p \)-value for 529 out of 684 schools (77%). Additionally, the fraction of the FAB \( p \)-values less than \( \alpha \) is more than that for the UMPU \( p \)-values for all values of \( \alpha \) up to about 0.76, which means that a test based on the FAB \( p \)-values will reject more null hypotheses than one based on the UMPU \( p \)-values while maintaining the same Type I error rate of \( \alpha \), for every \( \alpha \) less than 0.76. Of more interest might be the small

---

**Figure 3.** Results from the ELS analysis. The left panel plots direct estimates of school-level means (minus the national average of 50) versus the indirect estimates. The middle and right panels plot the FAB and UMPU \( p \)-values.
p-values, displayed in the right-most panel of the figure. There were 316 FAB p-values and 295 UMPU p-values less than 0.05, for a difference of 21 schools.

3.3. Example: Spatial Linking Models

In this subsection, we perform a small numerical simulation study that is meant to mimic the search for signals along a one-dimensional lattice, such as a chromosome. Let \( Y \sim N_p(\theta, I) \) and suppose the true \( \theta \) is a realization of a discrete Markov chain taking values in \( \{-1, 0, 1\} \) and with transition probability matrix

\[
P = \begin{pmatrix}
0.975 & 0.025 & 0.000 \\
0.010 & 0.990 & 0.010 \\
0.000 & 0.025 & 0.975
\end{pmatrix}.
\]

Figure 4 displays one realization of such a process for \( p = 1000 \).

For each \( j \in \{1, \ldots, p\} \), we construct a FAB p-value from \( Y \) to evaluate \( H_j : \theta_j = 0 \) using the (incorrect) linking model \( \theta \sim N_p(\mu, \Psi) \) where \( \mu \) is a scalar, \( \mathbf{1} \) is a \( p \)-dimensional vector of ones, and \( \Psi \) is the covariance matrix of a conditional autoregressive process, so that the conditional distribution of \( \theta_j \) given the other elements of \( \theta \) is \( N(\beta_0 + \beta_1 (\theta_{j-1} + \theta_{j+1}), \tau^2) \) for some values of \( \beta_0, \beta_1, \) and \( \tau^2 \). To construct a FAB p-value for \( H_j \), the procedure is to

1. obtain estimates \( \tilde{\theta}_j = (\tilde{\mu}, \tilde{\beta}_0, \tilde{\beta}_1, \tilde{\tau}^2) \) of the parameters \( \theta = (\mu, \beta_0, \beta_1, \tau^2) \) of the Gaussian autoregressive linking model, using \( G_j^T Y \) as data;
2. compute plug-in estimates \( \tilde{m}_{ij}, \tilde{v}_{ij} \) of the conditional mean and variance \( m_{ij}, v_{ij} \) of \( \theta_j \) given \( G_j^T Y \), using \( \tilde{\theta} \) and Equations (7) and (8);
3. compute the FAB p-value \( p(Y_j, \tilde{b}_j) = 1 - \Phi(Y_j + \tilde{b}_j) - \Phi(-Y_j) \), where \( \tilde{b}_j = 2 \tilde{m}_{ij}/\tilde{v}_{ij} \).

Further computational details are available in the supplementary materials for this article.

Computing the FAB p-value for each of the 1000 hypotheses is somewhat time consuming, as the autoregressive model in Step 1 is fit separately for each \( j \in \{1, \ldots, 1000\} \), each time without \( Y_j \) to ensure that \( \tilde{b}_j \) is independent of \( Y_j \). One faster alternative is to break the vector \( Y \) into contiguous subvectors, and then construct the FAB p-values in one subvector using linking model parameters estimated from the remainder of the subvector. Such an approach will maintain the uniformity of the FAB p-values under \( H_j \), but will not make optimal use of the information in \( G_j^T Y \). Another option would be to fit the autoregressive model once using the entire \( Y \)-vector, and use the resulting linking model parameter estimate \( \tilde{\theta} \) to construct each p-value. However, since in this case \( Y_j \) is used to obtain the common \( \tilde{\theta} \), the value of \( \tilde{b}_j \) will depend slightly on \( Y_j \), violating the sufficient condition for \( p(Y_j, \tilde{b}_j) \) to be uniformly distributed under \( H_j \). However, as this linking model has a small number of parameters relative to \( p \), we might expect the influence of \( Y_j \) on \( \tilde{b}_j \) to be slight. We investigate this claim numerically in the simulation study that follows.

One-hundred \( \theta \)-vectors were simulated from the discrete Markov chain described above, and from each a single observed data vector \( Y \sim N_p(\theta, I) \) was simulated. For each \( Y \)-vector and each index \( j \), three p-values for evaluating \( H_j : \theta_j = 0 \) were computed: the adaptive FAB p-value described in Steps 1–3 above; an “approximate” adaptive FAB p-value that estimates the parameters of the linking model only once for each simulated \( Y \)-vector; and the usual p-value \( p(Y_j, 0) \) based on the UMPU test of \( H_j : \theta_j = 0 \). For each simulated data vector and each type of p-value, we computed the number of discoveries (rejected null hypotheses) and the false discovery proportion (FDP) using the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995) with a false discovery rate (FDR) controlled to be less than 0.2.

We first compare each of the three p-value procedures in terms of FDR control. On average across datasets, the number of “true nulls” (\( \theta_j \)’s equal to zero) was about 574 out of 1000. Therefore, the Benjamini–Hochberg (BH) procedure using the UMPU p-values should attain an actual FDR of about \( 0.574 \times 0.2 \approx 0.115 \), since the UMPU p-values are independent of one another. Modulo Monte Carlo error, this is what was observed—the FDP based on the UMPU p-values was 0.106 on average across the 100 datasets. The average FDP using the FAB and approximate FAB procedures were both about 0.108, very similar to that of the UMPU procedure and well below the target FDR (the across-simulation standard deviations of FDP were 0.074 and 0.144 for the FAB and UMPU procedures, respectively). We note that control using the BH procedure is guaranteed for the UMPU p-values since they are statistically independent. In contrast, the FAB p-values are dependent since, for example, the p-values for \( \theta_j \) and \( \theta_{j+1} \) are both functions of \( Y_j \) and \( Y_{j+1} \) (the indirect information for \( Y_j \) includes \( Y_{j+1} \) and vice versa). This leads to positive dependence among spatially proximal FAB p-values. However, several theoretical results have
been obtained showing that for some types of positive dependence the BH procedure maintains the target FDR (Benjamini and Yekutieli 2001; Sarkar 2002), or does so asymptotically.

Figure 5. Some results from the simulation study. The panel on the left displays the fraction of true positives discovered by the FAB procedure relative to the UMPU procedure for each of the 100 simulated datasets. The panel in the middle displays the CDFs of nonnull p-values for each simulated dataset in dashed blue lines (FAB) and pink lines (UMPU), with average CDFs given in thick blue (FAB) and red (UMPU) lines. The panel on the right displays the distribution of the approximate FAB p-values corresponding to true nulls.

Some of the results relating to power are displayed in Figure 5. The first panel plots the fraction of true positives discovered using the exact FAB procedure, versus the fraction discovered by the UMPU procedure. The proportion discovered by the FAB approach was as high or higher for all 1000 simulated datasets. On average across datasets, the mean proportion of true positives that were discovered was 0.02 for the UMPU procedure. The proportion discovered using the exact FAB procedure, versus the fraction corresponding to true nulls. and we would therefore expect the p-values to be very similar. In particular, since the FAB p-values are uniformly distributed for js such that \( \theta_j = 0 \), we expect the corresponding approximate FAB p-values to be nearly uniformly distributed as well. This is confirmed in the third panel of Figure 5, which plots a histogram of the approximate FAB p-values for the true null hypotheses, across all iterations of the simulation study.

4. Adaptive p-Values From Dependent Data

4.1. Data With Known Covariance

In many applications, the elements of the data vector \( \mathbf{Y} \) will be correlated. Examples include the case that \( \mathbf{Y} \) is a vector of estimates of linear regression coefficients from a nonorthogonal design matrix, and situations where the elements of \( \mathbf{Y} \) correspond to measurements that are spatially or temporally related. In such cases, FAB p-values that are marginally uniform under their null hypotheses may be obtained as in the previous section by generalizing the form of the indirect information used to construct the p-value for each hypothesis.

Specifically, consider constructing a p-value for \( H_j : \theta_j = 0 \) based on the model \( \mathbf{Y} \sim N_p(\mathbf{\theta}, \mathbf{\Sigma}) \) where \( \mathbf{\Sigma} \) is not necessarily diagonal. In this subsection, we consider the case that \( \mathbf{\Sigma} \) is known—cases where it is unknown are considered in the following two subsections. As before, we refer to \( Y_j/\sigma_j \) as the direct information for \( \theta_j \). Indirect information about \( \theta_j \) that is independent of \( Y_j \) may be based upon an appropriate linear transformation of \( \mathbf{Y} \). Let \( \mathbf{G}_j \) be any \( p \times (p - 1) \) matrix whose columns form a basis for the null space of the \( j \)th column of \( \mathbf{\Sigma} \). Then \( \mathbf{G}_j^\top \mathbf{Y} \sim N_{p-1}(\mathbf{G}_j^\top \mathbf{\theta}, \mathbf{G}_j^\top \mathbf{\Sigma} \mathbf{G}_j) \), and is statistically independent of \( Y_j \). By Corollary 2, a p-value \( p(Y_j/\sigma_j, \tilde{b}_j) = 1 - \Phi(Y_j/\sigma_j + \tilde{b}_j) - \Phi(-Y_j/\sigma_j) \) is uniformly distributed under \( H_j \) for any statistic \( \tilde{b}_j \) that is a function of \( \mathbf{G}_j^\top \mathbf{Y} \).

As before, while \( \mathbf{G}_j^\top \mathbf{Y} \) does not provide direct information about \( \theta_j \), it does provide information about \( \mathbf{G}_j^\top \mathbf{\theta} \) and so it can be used to provide information about \( \theta_j \) if there are relationships...
among the entries of $\theta$. If such relationships can be encoded in a linking model $\theta \sim N_p(\mu, \Psi)$, then a FAB $p$-value for $H_1$ may be constructed exactly as in the previous section, modulo the more general form of the indirect information $G_j^T Y$ where $G_j$ depends on $\Sigma$. In particular, the conditional expectation and variance of $\beta_j$ given $\theta$ using $Y$ under the linking model may still be obtained from Equations (7) and (8). Additionally, if reasonable values of $\mu$ and $\Sigma$ are not known in advance, then they may be estimated from the indirect information $G_j^T Y$ and the resulting FAB $p$-value will still be uniform under $H_1$.

In practice, it is rare that the covariance of $Y$ will be completely known. However, if it is known sufficiently to decorrelate $Y$ then FAB $p$-values may be obtained that are exactly uniform under their null hypotheses. Otherwise, if $\Sigma$ may be consistently estimated, then FAB $p$-values may be constructed that are asymptotically uniform. In this case, we can also replace the normality assumption for $Y$ with asymptotic normality. We illustrate procedures for these two scenarios in the following subsections.

### 4.2. Data With Known Correlation

Here we show how to construct FAB $p$-values for regression coefficients of a normal linear regression model, a case where the correlations among the direct estimates are known. Construction of FAB $p$-values in other cases of known correlation is similar.

We use essentially the same method of data splitting as was used in Hoff and Yu (2019) in the context of confidence interval construction. Suppose we wish to construct $p$-values for the elements of $\beta \in \mathbb{R}^p$ in the linear regression model $Y \sim N_p(X \beta, \sigma^2 I)$, where $X \in \mathbb{R}^{n \times p}$ is an observed design matrix and $\beta \in \mathbb{R}^p$ and $\sigma^2 \in \mathbb{R}^+$ are unknown. Our sampling model is based on the probability distribution of the ordinary least-squares (OLS) regression estimator, $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$, with $\hat{\beta}$ being independent of the usual unbiased estimator $\hat{\sigma}^2$ of $\sigma^2$, for which $\hat{\sigma}^2 / \sigma^2 \sim \chi^2_n$ with $n = n - p$. The direct estimate for a particular coefficient $\beta_j$ is $\hat{\beta}_j \sim N(\beta_j, h_{jj} \sigma^2)$ where $h_{jj} = (X^T X)^{-1}_{jj}$, and so $T_j = \hat{\beta}_j / [h_{jj}^{1/2} \hat{\sigma}] \sim t_{n-p}$ under $H_j : \beta_j = 0$. To find the indirect information, let $G_j$ be a $p \times (p - 1)$ matrix whose columns span the null space of the $j$th column of $X$, so that $G_j^T \hat{\beta}$ is independent of $\hat{\beta}_j$.

Letting $F_v$ denote the CDF of a t-distribution with $v$ degrees of freedom, the distribution of $1 - [F_v(T_j + \hat{b}_j) - F_v(-T_j)]$ will then be uniform under $H_j : \beta_j = 0$ for any scalar function $b_j$ of $G_j^T \hat{\beta}$.

A value of $\hat{b}_j$ may be obtained from a linking model for $\beta$. The marginal distribution of $G_j^T \hat{\beta}$ under the linking model $\beta \sim N_p(\mu, \Sigma)$ is $G_j^T \hat{\beta} \sim N_p(-G_j^T \mu, G_j^T \Sigma G_j + \sigma^2 (X^T X)^{-1})$. If $\mu$ and $\Sigma$ are sufficiently structured then marginal maximum likelihood estimates $\hat{\mu}, \hat{\Sigma}, \hat{\sigma}^2$ may be obtained from this marginal model for $G_j^T \hat{\beta}$. These estimates may then be used to obtain estimates $\hat{m}_j$ and $\hat{\nu}_j$ of the conditional mean $m_j$ and variance $\nu_j$ of $\beta_j$ given the indirect information. Since $\hat{m}_j, \hat{\nu}_j, \hat{\sigma}^2$ are obtained from $G_j^T \hat{\beta}$, they are independent of $\hat{\beta}_j$ (and $\hat{\sigma}^2$), and so the FAB $p$-value $1 - |F(T_j + 2\hat{m}_j \hat{\sigma} / \hat{\nu}_j) - F(-T_j)|$ is uniformly distributed under $H_j$.

This methodology can also be applied in cases where there is only a linking model for a subset of parameters. For example, consider the linear regression model

$$ Y \sim N_p(W \alpha + X \beta, \sigma^2 I), $$

where $W$ and $X$ are observed design matrices. If interest is primarily in $\beta$ and a reasonable linking model relating $\alpha$ to $\beta$ is not available, then FAB $p$-values for $\beta$ alone may be obtained as just described, except noting that $\hat{\beta} \sim N_p(\beta, \sigma^2 \Omega)$ where $\Omega$ is the appropriate submatrix of $[(WX)^T (WX)]^{-1}$. Letting $G_j$ be an orthogonal basis for the null space of the $j$th column of $\Omega$, we can obtain an estimated conditional distribution for $\beta_j$ using a linking model $\beta \sim N_p(\mu, \Psi)$ and the indirect information $G_j^T \hat{\beta}$. Specifically, $G_j^T \hat{\beta}$ is independent of $\hat{\beta}_j$, and so the indirect information about $\beta_j$ is then given by $G_j^T \hat{\beta} \sim N_{p-1}(G_j^T \hat{\mu}, G_j^T \Sigma G_j)$.

As a numerical illustration we perform an analysis of another aspect of the 2002 Educational Longitudinal Study dataset. Let $Y_i$ be the test score of student $i$, $s_i$ be a numerical measure of their socioeconomic status (SES) and let $v_i$ be a three-dimensional vector indicating the sex, parents’ education, and status as a native English speaker of student $i$. Also, let $g_i \in \{0, 1\}$ be the binary vector indicating in which of $p = 684$ schools student $i$ is enrolled. Let $w_i = (g_i, v_i) \in \mathbb{R}^{p+3}$ and let $x_i = s_i g_i \in \mathbb{R}^p$. The linear regression model

$$ y_i = \alpha^T w_i + \beta^T x_i + \epsilon_i $$

can be used to evaluate the evidence for each school that SES is related to test score, controlling for effects of $v_i$, by obtaining $p$-values for each element of $\beta$. Note that the entries of $\hat{\beta}$, the OLS estimator of $\beta$, are correlated with each other because of the presence of the covariate $v_i$.

We compute indirect information for testing each $\beta_j$ using $G_j^T \hat{\beta}$ and an exchangeable linking model $\beta \sim N_p(\mu, \tau^2 I)$. The model for the indirect information is thus $G_j^T \hat{\beta} \sim N_{p-1}(\mu G_j^T 1, \sigma^2 G_j^T \Omega G_j + \tau^2 I)$. From this model, we obtain marginal maximum likelihood estimates $(\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2)$ of $(\mu, \tau^2, \sigma^2)$ that are statistically independent of $T_j$. We then compute the FAB $p$-value $1 - |F(T_j + 2\hat{\mu} \hat{\tau} / \hat{\sigma}) - F(-T_j)|$, which is uniformly distributed if $\beta_j = 0$.

Some results of this analysis are plotted in Figure 6. The first panel plots the direct OLS estimate $\hat{\beta}_j$ for each school $j$ versus its estimated precision $(\hat{\sigma}^2 w_{jj})^{-1/2}$ (the reciprocal of the standard error). While the variability of the $\hat{\beta}_j$’s around their average value (given by a horizontal black line) is large, the variability is highest among estimates corresponding to schools with low sample sizes (and hence low precision). This suggests that much of the observed variability among the elements of $\hat{\beta}$ is due to within-school sampling variability, and that the heterogeneity of the actual $\beta_j$’s is much lower. An informal assessment of this is made with the two gray curves in the plot, which are at the across-school average of the $\hat{\beta}_j$’s plus and minus 1.96 times the standard error (the reciprocal of the horizontal coordinate of the plot). Indeed, most of the OLS estimates fall within the
standard errors of the overall average. Nevertheless, an $F$-test of the global null hypotheses of no across-school variation in $\beta_j$’s has a $p$-value of less than $10^{-4}$, suggesting nonzero across-school variation.

The middle panel of Figure 6 plots the UMPU $p$-values versus the FAB $p$-values for evaluating $H_j : \beta_j = 0$ for each school $j$. The unfamiliar pattern is due to the fact that, while there is evidence of across-school variability in the $\beta_j$’s, the variation is centered around a positive value that is relatively far from zero. As a result, the FAB $p$-values have adaptively become nearly one-sided, so that $p_j^{\text{FAB}} \approx p_j^{\text{UMPU}} / 2$ if $\beta_j$ is positive, and $p_j^{\text{FAB}} \approx 1 - p_j^{\text{UMPU}} / 2$ if $\beta_j$ is negative, where $p_j^{\text{FAB}}$ and $p_j^{\text{UMPU}}$ are the FAB and UMPU $p$-values for school $j$, respectively. Since most $\beta_j$’s are positive (about 85%) this leads to many more “small” FAB $p$-values than small UMPU $p$-values, as shown in the right-most plot in the figure. For example, there are 245 FAB $p$-values less than 0.05, but only 188 UMPU $p$-values below this level. Additionally, the empirical CDF of the FAB $p$-values is larger than that of the UMPU $p$-values over the interval (0, 1), which means that a test based on the FAB $p$-values will reject more null hypotheses than one based on the UMPU $p$-values while maintaining the same Type I error rate of $\alpha$, for every $\alpha \in (0, 1)$.

There may be some concern that since the FAB $p$-values in this case are nearly the same as one-sided $p$-values, their power to detect evidence that a given parameter is less than zero is drastically reduced. While this is true, there is not much evidence in the data that any of the $\beta_j$’s are actually less than zero. For example, there was only one $p$-value smaller than 0.05 that corresponded to a negative value of $\beta_j$ (a $p$-value of 0.04). The data suggest that the $\beta_j$’s have a small amount of variation around a positive value, a data feature to which the FAB $p$-value procedure has adapted.

### 4.3. Asymptotically Uniform Null $p$-Values

In the preceding linear regression example, adaptive FAB $p$-values with exactly uniform null distributions are available because the covariance of the observed data vector is known up to a scalar. For each element of the data vector, this permits construction of an independent vector of indirect information. In cases where the covariance is not sufficiently known to permit this decorrelation, an independent vector of indirect information may not be available. However, if a consistent estimator of the covariance is available then it is possible to construct approximate adaptive FAB $p$-values with null distributions that are asymptotically uniform. Asymptotically uniform null $p$-values are also available in cases where the direct information is only asymptotically normally distributed. This may be the case if, for example, the direct estimates consist of regression coefficient estimates from a generalized linear model.

Consider the general case where the data available include a consistent and asymptotically normal estimator $\hat{\theta}$ of $\theta$, so that $\sqrt{n}(\hat{\theta} - \theta) \overset{\mathcal{d}}{\rightarrow} N_p(0, \Sigma)$ as $n \rightarrow \infty$, and a consistent estimator $\hat{\Sigma}$ of $\Sigma$. The standard Wald $p$-value for $\theta_j$ that is asymptotically uniform under $H_j : \theta_j = 0$ can be written as $1 - \Phi(Z_j) - \Phi(-Z_j)$, where $Z_j = \sqrt{n}\hat{\theta}_j / \hat{\sigma}_j$. The proposed approximate FAB $p$-value for $\theta_j$ is $1 - \Phi(Z_j + \tilde{b}_j) - \Phi(-Z_j)$, where $\tilde{b}_j$ is obtained from indirect information about $\theta_j$ that is asymptotically independent of the direct information $\theta_j$. This indirect information about $\theta_j$ can be constructed as follows:

Define the function $\Gamma : \mathbb{R}^p \rightarrow \mathbb{R}^{p \times (p - 1)}$ so that $\Gamma(s)$ is a $p \times (p - 1)$ matrix whose columns form an orthonormal basis for the null space of $s$. Specifically, define $\Gamma(s)$ so that the basis is obtained by Gram–Schmidt orthogonalization of the collection of vectors in $\mathbb{R}^p$ consisting of $s$ and the standard basis vectors $\{e_k : k \in \{1, \ldots, p\} \setminus \{j\}\}$. This construction implies that $\Gamma$ is a continuous function of $s$. Now let $\tilde{G}_j = \Gamma(\hat{\sigma}_j)$, where $\hat{\sigma}_j$ is the $j$th column of the true covariance matrix $\Sigma$. The indirect information about $\theta_j$ can be derived from $\tilde{G}_j^\top \hat{\theta}$. This random vector will not be exactly independent of the direct estimate $\hat{\theta}_j$ because $\tilde{G}_j$ is not exactly equal to $G_j$ and $\hat{\theta}_j$ might not be exactly normally distributed, but $\tilde{G}_j^\top \hat{\theta}$ should be approximately independent of $\hat{\theta}_j$ for large $n$, since $\tilde{G}_j$ is converging to $G_j$, $\tilde{G}_j^\top \hat{\theta}$ is asymptotically uncorrelated with $\hat{\theta}_j$ and both are asymptotically normal. The following result

![Figure 6](image-url). Evaluation of school-specific associations between SES and standardized test performance. The left panel plots OLS estimates versus their precision. The middle and right-most panels plot UMPU and FAB $p$-values.
shows that indeed $\hat{G}_j^\top \hat{\theta}$ is uncorrelated with and independent of $\hat{\theta}_j$ in an asymptotic sense.

**Theorem 9.** Let $\tilde{\Sigma} \overset{P}{\to} \Sigma$ and $\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\to} E$, where $E \sim N_p(0, \Sigma)$. Then as $n \to \infty$,

1. $\text{Cor}(\hat{G}_j^\top \hat{\theta}, \hat{\theta}_j) \to 0$;
2. $\text{Pr}(\{\sqrt{n}(\hat{G}_j^\top \hat{\theta} - \theta) \in A\} \cap \{\sqrt{n}(\hat{\theta}_j - \theta_j) \in B\}) \to \text{Pr}(G^\top E \in A) \times \text{Pr}(E_j \in B)$,

where $G_j = \Gamma(\sigma_j)$, $\hat{G}_j = \Gamma(\hat{\sigma}_j)$, and $A \subset \mathbb{R}^{p-1}$ and $B \subset \mathbb{R}$ are measurable sets.

**Proof.** Let $\hat{H}_j$ be the $p \times p$ matrix obtained by binding the $j$th standard basis vector to $G_j$, and define $H_j$ analogously. Since $\Gamma$ is continuous we have that $\hat{G}_j \overset{P}{\to} G_j$ and also $\hat{H}_j \overset{P}{\to} H_j$. Both results then follow by Slutsky’s theorem, since $\hat{H}_j^\top \sqrt{n}(\hat{\theta} - \theta) \overset{d}{\to} H_j^\top E = (G_j^\top E_1 E_j)$, and $G_j^\top E$ is independent of $E_j$. \(\square\)

Now we construct a class of $p$-value functions that make use of the indirect information and are asymptotically uniform under each null hypothesis. Consider the $p$-value function $1 - |\Phi(\hat{Z}_j + \hat{\sigma}f(\hat{G}_j^\top \hat{\theta}, \hat{G}_j)/\sqrt{n}) - \Phi(-\hat{Z}_j)|$ where $f : \mathbb{R}^{p-1} \times \mathbb{R}^p \to \mathbb{R}$ is a continuous function. For example, $f$ could be a function that, when input $(G_j^\top \theta, G_j)$, returns $\mathbb{E}[\hat{\sigma}j(\hat{G}_j^\top \theta)/V[\theta_j][G_j^\top \theta]]$ based on a linking model $G_j^\top \theta \sim N_{p-1}(G_j^\top \theta, G_j^\top \Sigma G_j)$. Since neither $G_j$ nor $\theta$ are known, we use plug-in estimates $\hat{G}_j^\top \hat{\theta}$ and $\hat{G}_j$.

An asymptotic analysis of this FAB $p$-value where $\theta$ is fixed at a single nonzero value for all sample sizes is not particularly interesting, since in this case both the FAB and UMPU $p$-values converge in probability to 0. To see this, note that in this case $f(G_j^\top \hat{\theta}, G_j)$ converges to $f(G_j^\top \theta, G_j)$ by the continuous mapping theorem, and so the FAB adjustment $\hat{b}_j = \hat{\sigma}f(G_j^\top \theta, G_j)/\sqrt{n}$ will converge to zero and so too the difference between the FAB and UMPU $p$-values, that latter of which converges in probability to 0 since the UMPU test is consistent against all nonzero alternatives. More interesting is the behavior of the FAB $p$-value under a sequence of local alternatives that decrease in magnitude with $n$, say $\theta = \theta_0/\sqrt{n}$ when the sample size is $n$. Now recall that $f(G_j^\top \theta, G_j)$ returns a plug-in estimate of $\mathbb{E}[\hat{\sigma}j(G_j^\top \theta)/V[\theta_j][G_j^\top \theta]]$ under the linking model $G_j^\top \theta \sim N_{p-1}(G_j^\top \theta, G_j^\top \Sigma G_j)$. It is natural to use a function $f$ that is scale equivariant, which implies that $f$ satisfies $f(cG_j^\top \theta, G_j) = f(G_j^\top \theta, G_j)/c$ for positive scalars $c$. In this case, we have $f(G_j^\top \hat{\theta}, G_j)/\sqrt{n} = f(G_j^\top \sqrt{n \hat{\theta}}, G_j)$. Now $\sqrt{n \hat{\theta}}$ is approximately distributed as $N_p(\theta_0, \Sigma)$, while $\hat{G}_j$ converges in probability to $G_j$, so multiplication of $\sqrt{n \hat{\theta}}$ by $\hat{G}_j$ gives an indirect information vector that is asymptotically independent of the direct estimate $\hat{\theta}_j$, and so the resulting $p$-value will be asymptotically uniformly distributed if $\theta_0j = 0$. This is summarized as follows:

**Theorem 10.** Let $\tilde{\Sigma} \overset{P}{\to} \Sigma$ and $\hat{\theta} \overset{d}{\to} Y/\sqrt{n} + o_p(1/\sqrt{n})$, where $Y \sim N_p(\theta_0, \Sigma)$. Then as $n \to \infty$,

\[
1 - |\Phi(\hat{Z}_j + \hat{\sigma}f(\hat{G}_j^\top \hat{\theta}, \hat{G}_j)/\sqrt{n}) - \Phi(-\hat{Z}_j)| \overset{d}{\to} 1 - |\Phi(Z_j + \sigma f(\hat{G}_j^\top Y, G_j) - \Phi(-Z_j)),
\]

where $\hat{Z}_j = \sqrt{n \hat{\theta}}/\hat{\sigma}_j, Z_j = Y_j/\sigma_j$, and $f : \mathbb{R}^{p-1} \times \mathbb{R}^p \to \mathbb{R}$ is a continuous function that satisfies $f(x, b) = f(a, b)/c$ for $c > 0$. The right side of (11) is uniformly distributed if $\theta_0j = 0$.

**Proof.** The $p$-value function $p(z, b) = 1 - |\Phi(z + b) - \Phi(-z)|$ is continuous in $z$ and $b$, and so by the continuous mapping theorem it suffices to show that $\hat{Z}_j \overset{d}{\to} Z_j$ and $\hat{\sigma}f(\hat{G}_j^\top \hat{\theta}, \hat{G}_j)/\sqrt{n} \overset{d}{\to} \sigma f(\hat{G}_j^\top Y, G_j)$, which converges to $f(\hat{G}_j^\top Y, G_j)$ by the continuous mapping theorem and that $\hat{G}_j^\top \sqrt{n \hat{\theta}} \overset{d}{\to} G_j^\top Y$ (which follows from the consistency of $\hat{G}_j$ for $G_j$ and Slutsky’s theorem). \(\square\)

We evaluate this result empirically in a simulation study with $p = 30$ and four sample sizes $n \in \{200, 400, 800, 1600\}$. For each sample size $n$, 5000 binary vectors $Y \in \{0, 1\}^n$ were constructed with elements simulated as $Y_i \sim \text{binary}(\theta^\top x_i/(1 + e^\theta^\top x_i))$, independently for $i = 1, \ldots, n$. The $p$-dimensional vector of regression coefficients $\theta$ had 15 elements equal to $3/\sqrt{n}$ and 15 elements equal to zero, and the elements of each $x_i \in \mathbb{R}^p$ were simulated independently from a standard normal distribution. For each simulated $Y$ vector, the MLE $\hat{\theta}$ was obtained, along with an estimate $\hat{\sigma}/n$ of its variance based on the observed information matrix. For each estimated coefficient $\hat{\theta}_j$, a Z-statistic $\hat{Z}_j = \sqrt{n \hat{\theta}_j}/\hat{\sigma}_j$ was computed and used to construct both a Wald and FAB $p$-value. The FAB $p$-value was constructed using a linking model where $\theta$ is an iid sample from a mixture of a $N(\mu, \tau^2)$ distribution and a point-mass at zero. For each $j$, the parameters $\mu$, $\tau^2$ and the mixture proportions were estimated using marginal maximum likelihood and the indirect information $\hat{G}_j^\top \hat{\theta}$. The FAB $p$-value under this linking model is $1 - |\Phi(\hat{Z}_j + 2 \hat{\sigma}_j \bar{\mu}_j/\sqrt{n \tau_j^2}) - \Phi(-\hat{Z}_j)|$, where $\bar{\mu}_j$ and $\tau_j^2$ are estimated from the linking model using $\hat{G}_j^\top \hat{\theta}$.

Some results are shown in Figure 7. The first column of the figure shows histograms of the Wald and FAB $p$-values corresponding to nonnull coefficient values. As can be seen, the FAB $p$-values tend to be smaller, consistently across the four sample sizes. For example, the fraction of nonnull FAB $p$-values below 0.05 was 0.4, 0.42, 0.43, and 0.44 for $n \in \{200, 400, 800, 1600\}$ as compared to 0.29, 0.31, 0.32, and 0.32 for the Wald $p$-values. More generally, the CDFs of the nonnull FAB $p$-values are larger than those of the Wald $p$-values, meaning that, on average across simulations and hypotheses, tests based on the FAB $p$-values are more powerful than those based on the Wald $p$-values, at every level. The second and third columns of the figure plot histograms of the $p$-values corresponding to zero coefficient values. Neither the Wald nor the FAB null $p$-values are exactly
Figure 7. Distributions of approximate $p$-values from the simulation study. The first column displays CDFs of Wald and FAB $p$-values for nonnull parameters. The second and third columns display distributions of $p$-values for null parameters.
uniformly distributed, however the uniformity of these distributions increases with $n$. Interestingly, the shape of the null $p$-value distributions is slightly different for the smaller sample sizes, although near zero the shapes are similar. For example, the fraction of null FAB $p$-values below 0.05 was 0.06, 0.06, 0.05, and 0.05 for $n \in \{200, 400, 800, 1600\}$ as compared to 0.07, 0.06, 0.05, and 0.05 for the Wald $p$-values.

5. Discussion

The FAB $p$-value is derived from a test statistic that has optimal average power with respect to a probability distribution over the possible values of the mean of a normal distribution. Ideally, this probability distribution over the normal mean should represent prior or indirect information. However, regardless of the distribution used to construct it, the FAB $p$-value is uniformly distributed under the null hypotheses.

In multiparameter settings, indirect information about one parameter may be derived from data on the other parameters, and then used to construct an adaptive FAB $p$-value. One way to do this is with a linking model that relates the parameters to each other. A Gaussian linking model is convenient, in terms of both the estimation of the parameters in the linking model, and the simplicity of the FAB $p$-value when the indirect information is in the form of a normal distribution. However, other linking models could certainly be used, which would yield different optimal test statistics and different $p$-value functions.

Related to this, an interesting question is the extent to which a FAB $p$-value based on a Gaussian linking model has good performance on average with respect to a non-Gaussian linking model. It seems plausible that in many scenarios, even if the true values of the parameters are better represented by a non-Gaussian linking model, adaptive $p$-values constructed using a Gaussian linking model will at least perform better than the UMPU $p$-values, on average across the parameters (this was the case for the example in Section 3.3, which used a misspecified linking model). To see why, consider the simplest case where the direct data for parameter $\theta_j$ is $Z_j \sim N(\theta_j, 1)$. By Theorem 3, the probability that the FAB $p$-value is smaller than the UMPU $p$-value is at least $\Phi(\text{sign}(\tilde{b}_j) \times \tilde{\theta}_j)$, where $\tilde{b}_j$ is two times the ratio of the expected value of $\tilde{\theta}_j$ to its variance, as estimated from the linking model. So roughly speaking, as long as the linking model is good enough so that $\Phi(\text{sign}(\tilde{b}_j) \times \tilde{\theta}_j)$ is greater than 1/2 on average across $\tilde{\theta}_j$'s, then a majority of the adaptive FAB $p$-values should be lower than the corresponding UMPU $p$-values. The most challenging case for the FAB procedure is perhaps when the $\theta_j$'s are unrelated and centered around zero. However, if the linking model contains the mean-zero normal distributions with covariance proportional to the identity matrix (as it did in each example presented in this article), then the lack of structure among the $\theta_j$'s should be reflected in the estimates of the parameters in the linking model. In this case, the values of the $\tilde{b}_j$'s should be close to zero and the FAB $p$-values will be approximately the same as the UMPU $p$-values.

This article has focused on constructing adaptive FAB $p$-values in a variety of multiparameter settings, rather than recommending how they might be used. In some settings, such as the analysis of the ELS data presented in Sections 3.2 and 4.2, it is the school-specific $p$-values themselves that may be of interest, as the faculty of a given school are presumably primarily concerned with, for example, the relationship between SES and test scores in their own school rather than what this relationship is on average across schools. In other settings, it may be of more interest to use the $p$-values as inputs into other procedures that maintain global error rates. In such cases, it must be remembered that the adaptive FAB $p$-values discussed in this article are dependent by construction, and so procedures that rely on independent $p$-values must be used with caution. However, for most scenarios we expect that this dependence among FAB $p$-values will be positive dependence of some sort, in which case results of Clarke and Hall (2009) suggest that target error rates will still be approximately maintained for some popular procedures designed for independent $p$-values. Alternatively, FAB $p$-values could be used as inputs into procedures that accommodate, and even make use of, the dependence among $p$-values (Efroin 2007; Romano, Shaikh, and Wolf 2008; Sun and Cai 2009; Fan, Han, and Gu 2012).

However, there are some sequential testing scenarios where the dependence structure of FAB $p$-values simplifies so that null $p$-values are independent, permitting their use in a variety of global testing procedures. For example, Grazier G’sell et al. (2016) and Ramdas et al. (2017) considered sequential multiple testing scenarios where hypotheses are ordered and inferences are made sequentially. In these cases, if the FAB $p$-values are constructed so that $\tilde{b}_j$ depends only on data from populations that precede population $j$ in the sequence, then a simple conditioning argument shows that all $p$-values corresponding to true null hypotheses are independent of each other. As a result, such FAB $p$-values can be combined in a variety of ways to evaluate the global null hypothesis (see, e.g., Heard and Rubin-Delanchy 2018). The author is currently investigating additional ways that this independence property of sequentially computed FAB $p$-values can be exploited for different sequential testing objectives.

Supplementary Materials

Replication materials, additional numerical examples, and software to compute FAB $p$-values for parameters in linear and generalized linear models are available at https://phoff.github.io/FABInference.

References

Barber, R. F., and Ramdas, A. (2017), “The $p$-Filter: Multilayer False Discovery Rate Control for Grouped Hypotheses,” Journal of the Royal Statistical Society, Series B, 79, 1247–1268. [1255]

Benjamin, D. J., and Berger, J. O. (2019), “Three Recommendations for Improving the Use of $p$-Values,” The American Statistician, 73, 186–191. [1254]

Benjamini, Y., and Hochberg, Y. (1995), “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” Journal of the Royal Statistical Society, Series B, 57, 289–300. [1255,1262]

Benjamini, Y., and Yekutieli, D. (2001), “The Control of the False Discovery Rate in Multiple Testing Under Dependency,” The Annals of Statistics, 29, 1165–1188. [1255,1263]

Birnbaum, A. (1954), “Combining Independent Tests of Significance,” Journal of the American Statistical Association, 49, 559–574. [1255]

Burrus, K., and Hoff, P. (2019), “Exact Adaptive Confidence Intervals for Small Areas,” Journal of Survey Statistics and Methodology, 8, 206–230. [1255,1260]
Cai, T. T., Sun, W., and Wang, W. (2019), "Covariate-Assisted Ranking and Screening for Large-Scale Two-Sample Inference" (includes discussions and a rejoinder), *Journal of the Royal Statistical Society, Series B*, 81, 187–234. [1255]

Clarke, S., and Hall, P. (2009), "Robustness of Multiple Testing Procedures Against Dependence," *The Annals of Statistics*, 37, 332–358. [1263,1268]

Dickhaus, T. (2014), *Simultaneous Statistical Inference: With Applications in the Life Sciences*, Heidelberg: Springer. [1254,1256,1257]

Efron, B. (2007), "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, 102, 93–103. [1268]

Fan, J., Han, X., and Gu, W. (2012), "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence," *Journal of the American Statistical Association*, 107, 1019–1035. [1268]

Ghosh, M., and Rao, J. N. K. (1994), "Small Area Estimation: An Appraisal" (with comments and a rejoinder by the authors), *Statistical Science*, 9, 55–93. [1255,1261]

Good, I. J. (1963), "Shorter Confidence Intervals for the Mean of a Normal Distribution With Known Variance," *The Annals of Mathematical Statistics*, 34, 574–586. [1254]

Held, L. (2020), "A New Standard for the Analysis and Design of Replication Studies," *Journal of the Royal Statistical Society, Series A*, 183, 431–448. [1255,1258]

Hoff, P., and Yu, C. (2019), "Exact Adaptive Confidence Intervals for Linear Regression Coefficients," *Electronic Journal of Statistics*, 13, 94–119. [1255,1264]

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008), "Control of the False Discovery Rate Under Dependence Using the Bootstrap and Subsampling," *TEST*, 17, 417–442. [1268]