Tests of linear hypotheses using indirect information

Andrew MCCORMACK* and Peter D. HOFF

Department of Statistical Science, Duke University, Durham, North Carolina, U.S.A.

Key words and phrases: empirical Bayes; $F$-test; frequentist testing; hierarchical model; invariant test; multilevel data; small area estimation.

MSC 2020: Primary 62F03; Secondary 62J05.

Abstract: In multigroup data settings with small within-group sample sizes, standard $F$-tests of group-specific linear hypotheses can have low power, particularly if the within-group sample sizes are not large relative to the number of explanatory variables. To remedy this situation, in this article we derive alternative test statistics based on information sharing across groups. Each group-specific test has potentially much larger power than the standard $F$-test, while still exactly maintaining a target type I error rate if the null hypothesis for the group is true. The proposed test for a given group uses a statistic that has optimal marginal power under a prior distribution derived from the data of the other groups. This statistic approaches the usual $F$-statistic as the prior distribution becomes more diffuse, but approaches a limiting “cone” test statistic as the prior distribution becomes extremely concentrated. We compare the power and $P$-values of the cone test to that of the $F$-test in some high-dimensional asymptotic scenarios. An analysis of educational outcome data is provided, demonstrating empirically that the proposed test is more powerful than the $F$-test.

Résumé: Dans les contextes de données multigroupes avec de petites tailles d’échantillon au sein d’un groupe, les tests $F$ standards d’hypothèses linéaires propres à un groupe peuvent avoir une faible puissance, en particulier si les tailles d’échantillon au sein d’un groupe ne sont pas importantes par rapport au nombre de variables explicatives. Pour remédier à cette situation, les auteurs de cet article construisent des statistiques de test alternatives basées sur le partage d’informations entre les groupes. Chaque test spécifique au groupe a potentiellement une puissance beaucoup plus grande que le test $F$ standard, tout en maintenant exactement un taux d’erreur de type I cible si l’hypothèse nulle pour le groupe est vraie. Le test proposé pour un groupe donné utilise une statistique qui a un pouvoir marginal optimal selon une distribution antérieure dérivée des données des autres groupes. Cette statistique s’approche de la statistique $F$ habituelle à mesure que la distribution antérieure devient plus diffuse, mais s’approche d’une statistique de test limitative à mesure que la distribution antérieure devient extrêmement concentrée. Les auteurs comparent la puissance et les seuils observés du test du cône avec ceux du test-$F$ dans quelques scénarios asymptotiques en haute dimension. Une analyse des données sur le rendement scolaire est présentée, montrant empiriquement que le test proposé est plus puissant que le test $F$.

1. INTRODUCTION

Multigroup data analysis often occurs through the lens of separate but related linear regression models for each of several groups. For example, letting $y_{i,j}$ be a real-valued outcome and $x_{i,j} \in \mathbb{R}^p$ be a vector of features for the $i$th subject in group $j$, the relationship between $y_{i,j}$ and
The vector of outcomes in group $j$, and $\mathbf{X}_j \in \mathbb{R}^{n_j \times p}$ be the matrix of explanatory variables, this model can be expressed as

$$y_j \sim N_{n_j}(\mathbf{X}_j \beta_j, \sigma_j^2 \mathbf{I})$$

(1)

independently across groups $j = 1, \ldots, m$.

Estimators of $\beta_1, \ldots, \beta_p$ can be broadly categorized as being either “direct” or “indirect”. A direct estimator of $\beta_j$ is one that makes use of data only from group $j$, such as the ordinary least squares (OLS) estimator $\hat{\beta}_j = (\mathbf{X}_j^\top \mathbf{X}_j)^{-1} \mathbf{X}_j^\top y_j$. While $\hat{\beta}_j$ has minimum variance among unbiased estimators, if $n_j$ is small, then it may be preferable to reduce variance further by introducing bias in the form of “indirect” information from the other groups. Such estimators are often derived by imagining a normal model for across group variation, i.e.,

$$\beta_1, \ldots, \beta_m \sim \text{i.i.d. } N_p(\mathbf{0}, \mathbf{V}).$$

(2)

This across-group model is sometimes referred to as a “linking model” in the small-area-estimation literature, as it “links” the group-specific parameters $\beta_1, \ldots, \beta_m$ together through the parameters $(\mathbf{0}, \mathbf{V})$. For given values of $(\beta_0, \mathbf{V}, \sigma_1^2, \ldots, \sigma_m^2)$, the mean squared estimation error for each $\beta_j$, on average with respect to (2), is minimized by the conditional expectation

$$\tilde{\beta}_j = E(\beta_j|y_j) = \left(\mathbf{V}^{-1} + \mathbf{X}_j^\top \mathbf{X}_j / \sigma_j^2\right)^{-1} \left(\mathbf{X}_j^\top y_j / \sigma_j^2 + \mathbf{V}^{-1} \mathbf{0}\right).$$

This estimator can be interpreted as a Bayes estimator when (2) is thought of as a prior distribution (Lindley & Smith, 1972) and is sometimes referred to as the “best linear unbiased predictor” (BLUP) when (2) is thought of as a sampling model for the groups (Henderson, 1974), despite the fact that $\tilde{\beta}_j$ is biased as an estimator of $\beta_j$ (the “U” in “BLUP” refers to average bias with respect to (2), which is essentially the across-group average bias of $\tilde{\beta}_1, \ldots, \tilde{\beta}_m$). In practice, values of $(\beta_0, \mathbf{V}, \sigma_1^2, \ldots, \sigma_m^2)$ are estimated from $y_1, \ldots, y_m$ and then plugged into the equation for each $\tilde{\beta}_j$, yielding a so-called empirical Bayes estimator that is an “indirect estimator” in the sense that it combines direct information from group $j$ with indirect information from other groups, via the estimates of $(\beta_0, \mathbf{V}, \sigma_1^2, \ldots, \sigma_m^2)$. While the empirical Bayes estimator for a given group $j$ has a potentially higher mean squared error than the corresponding OLS estimator, the across-group average mean squared error of the empirical Bayes estimators is typically lower than that of the OLS estimators, even if the linking model (2) is incorrect or only a conceptual device, for example, if the groups are not randomly selected.

Analogously, we may consider direct and indirect methods for hypothesis testing. The most widely used frequentist level-$\alpha$ direct test of a linear hypothesis about $\beta_j$ is the standard $F$-test, whose test statistic is a function of data only from group $j$. In contrast, a frequentist level-$\alpha$ indirect test is one where the test statistic for group $j$ is allowed to depend on the data from the other groups. In particular, data from groups other than $j$ might suggest that the vector $\beta_j$ lies in a particular direction. We can use this indirect information to select a level-$\alpha$ test of $\beta_j$ that has more power than the $F$-test in this particular direction, at a cost of having less power in other directions. As long as the data are independent across groups, such a procedure will maintain a type I error rate of $\alpha$ if the hypothesis is true, while having increased power as compared to the $F$-test if the hypothesis is false and the indirect information from the other groups is reasonably accurate.

In comparison to methods for indirect estimation, methods for indirect hypothesis testing are relatively undeveloped. Notable work in this area by O’Gorman (2002, 2006) examines adaptive
procedures for testing subsets of regression coefficients in a linear model. For some non-normal error distributions, such procedures have higher power than the $F$-test. Recently, Hoff (2022) proposed an indirect analogue to the standard $t$-test and corresponding $P$-value for a univariate parameter $\theta_j$ based on a normally distributed estimator $\hat{\theta}_j$ and indirect data from groups other than group $j$. The indirect test of $H_j : \theta_j = 0$ has a rejection region that is asymmetric around zero and is chosen to maximize expected power with respect to a “prior distribution” which is derived from indirect data that is independent of $\hat{\theta}_j$. Such a test is “frequentist”, as it maintains an exact level-\(\alpha\) type I error rate, but is also Bayesian in that it minimizes a Bayes risk (one minus the prior expected power) and so it is referred to as being “frequentist and Bayesian”, or FAB. Inversions of such tests were used by Yu & Hoff (2018) to construct indirect confidence intervals for means in multiple normal populations. Their confidence intervals are essentially a multigroup extension of the interval proposed by Pratt (1963), who obtained the confidence interval for the mean of a normal population that has minimum prior expected width among those having $1 - \alpha$ frequentist coverage. These Bayes-optimal frequentist procedures can be derived from various likelihood ratios, and so in this sense, they are related to methods that use Bayes factors as statistics in frequentist hypothesis tests (Good, 1992; Chacon et al., 2007). Such an approach has been applied to testing hypotheses about multinomial probabilities (Good & Crook, 1974) and for evaluating nonparametric goodness of fit (Aerts, Claeskens & Hart, 2004).

In this article we develop FAB alternatives to the $F$-test for evaluating group-specific linear hypotheses in multigroup regression settings. Essentially, the proposed FAB test for $\beta_j$ is a level-\(\alpha\) test that has maximum expected power with respect to the “prior” distribution $\beta_j \sim N_p(\hat{\beta}_0, \Psi)$, where $(\hat{\beta}_0, \Psi)$ are estimated from models (1), (2), and data from groups other than $j$. In the next section, we derive the form of the optimal FAB statistic and obtain a numerical approximation to facilitate its calculation. A theoretical power comparison of the $F$-test and a simplified version of the FAB test, which we call the cone test, appears in Section 3. We show that the $F$-statistic is a special case of the FAB statistic. We also show that the ratio of the $F$-test $P$-value to the cone test $P$-value may range from zero to infinity. Asymptotic power comparisons between the $F$-test and the cone test are also provided, where it is shown that the cone test can have higher power than the $F$-test in certain high-dimensional scenarios. In Section 4 we describe in greater detail how the FAB statistic can be used in multigroup settings, and review some methods for obtaining estimates of the linking model parameters. A data analysis example considering educational test scores from multiple schools appears in Section 5. The FAB tests that share information across schools lead to a substantially greater number of null hypotheses being rejected. A discussion follows in Section 6. All of the proofs of theoretical results presented in this article can be found in the Appendix.

2. TESTING LINEAR HYPOTHESES

2.1. The FAB Test

Consider testing a linear hypothesis for the parameter $\beta \in \mathbb{R}^p$ based on an observation of $y$ from the linear model $y \sim N_n(X\beta, \sigma^2 I)$, where $X \in \mathbb{R}^{n \times p}$ is a known matrix of predictors, and $\beta$ and $\sigma^2$ are unknown. For the moment, we assume that $X$ has full rank. Recall that linear hypotheses of the form $H : A\beta = c$ may be expressed as $H : \theta = 0$ for a transformed linear model with mean $X\theta$ (Seber & Lee, 2003), so without loss of generality, we consider the null hypothesis $H : \beta = 0$. Note that the null and alternative models are invariant under data rescalings of the form $y \rightarrow \epsilon y$ for scalars $\epsilon \neq 0$. For this reason, we restrict attention to tests based on statistics that are invariant with respect to the group of rescalings of $y$. Any such statistic must be a function of a maximal invariant statistic, such as the unit vector $u = y/||y||$. The norm $|| \cdot ||$ denotes the Euclidean norm throughout the remainder of this article. A test statistic based on $u$ has the advantage that its null distribution does not depend on any unknown parameters, as $u$ is uniformly distributed on the sphere $S^{p-1}$ if $\beta = 0$. 

\[ \text{DOI: 10.1002/cjs.11760} \]
One scale-invariant statistic is the usual $F$-statistic, $F(y) = \{(SST - SSR)/p\}/\{SSR/(n - p)\}$, where $SST = y\top y$ is the total sum of squares and $SSR = ||y - X\hat{\beta}||^2$ is the residual sum of squares, with $\hat{\beta}$ being the OLS estimator. To see that this depends on $y$ only through $u$, let $P$ be the projection matrix onto the space spanned by the columns of $X$. Then we can write

$$F(y) = \frac{n - p}{p} \frac{y\top y - y\top (I - P)y}{y\top (I - P)y} = \frac{n - p}{p} \frac{y\top Py/y\top y}{1 - y\top Py/y\top y} = \frac{n - p}{p} \frac{u\top Pu}{1 - u\top Pu}.$$ 

Note that if $n \leq p$, then $P = I$ and $u\top Pu = 1$ is constant in $y$, and so in this case any test based on $u\top Pu$ (such as the $F$-test) has power equal to its level. Even if $n > p$, since the distribution of the $F$-statistic depends on $\beta$ only through $||X\beta||^2$, its power is constant on level-sets of $||X\beta||^2$. In this sense, the $F$-test is “looking” in all directions equally for evidence against the null hypothesis.

If prior information about the direction of $\beta$ is available, it may be preferable to use a test that has more power in this direction, at the cost of having lower power in the opposite direction. Specifically, suppose that prior information about $(\beta, \sigma^2)$ is available in the form of a prior density $\pi(\beta, \sigma^2)$. Let $p(u|\beta, \sigma^2)$ be the density of $u$ implied by the normal model (1). The prior expected power of a test function $\phi : \mathbb{S}^{n-1} \rightarrow [0, 1]$ is given by

$$E(\phi) = \int \int \int \phi(u) \ p(u|\beta, \sigma^2) \pi(\beta, \sigma^2) \ d\beta \ d\sigma^2$$

$$= \phi(u) \ p_\pi(u) \ du,$$

where $p_\pi(u) = \int p(u|\beta, \sigma^2) \pi(\beta, \sigma^2) \ d\beta \ d\sigma^2$ is the marginal density of $u$ induced by $\pi$, with respect to the uniform measure on $\mathbb{S}^{n-1}$. The Bayes-optimal level-$\alpha$ test $\phi_\alpha$ is the test that maximizes $E(\phi)$ among all level-$\alpha$ tests. Since the null distribution of $u$ is uniform on $\mathbb{S}^{n-1}$, by the Neyman–Pearson lemma, the optimal test is given by

$$\phi_\alpha(u) = \begin{cases} 1 & \text{if } p_\pi(u) > c_\alpha, \\ 0 & \text{if } p_\pi(u) \leq c_\alpha, \end{cases}$$

where $c_\alpha$ is chosen such that $\phi_\alpha$ has a type I error rate equal to $\alpha$.

Primarily for computational reasons, we consider the case where $\pi$ corresponds to a normal distribution $\beta \sim N_p(\beta_0, \Psi)$ for $\beta$ and a point-mass distribution on $\sigma_0^2$ for $\sigma^2$. The marginal distribution of $y$ under this prior is $y \sim N_n\left( X\beta_0, X\Sigma X\top + \sigma_0^2 I \right)$, and the corresponding distribution for $u$ is the angular Gaussian distribution $u \sim AG(\mu, \Sigma)$ with $\mu = X\beta_0$ and $\Sigma = X\Sigma X\top + \sigma_0^2 I$. The density of $u \sim AG(\mu, \Sigma)$ with respect to the uniform probability distribution on the sphere is derived in Pukkila & Rao (1988), and is given by $p(u|\mu, \Sigma) \propto x^{-n} e^{r^2/2} I_n(r)$, where $x = u\top \Sigma^{-1} u$, $r = u\top \Sigma^{-1} \mu/x$, and $I_n(r) = \int_0^\infty z^{n-1} e^{-(z-r)^2/2} \ dz$. The approximation and computation of this integral is described below. Therefore, the Bayes-optimal test of $H : \beta = 0$ based on $u$ rejects the hypothesis when $x^{-n} e^{r^2/2} I_n(r)$ is large, or equivalently, for large values of the FAB test statistic

$$T_{FAB}(u) = r^2/2 + \log I_n(r) - n \log x.$$  

DOI: 10.1002/cjs.11760
The function $I_n(r)$ can be computed recursively as

\[ I_1(r) = 2\pi \Phi(r) \]
\[ I_2(r) = e^{-r^2/2} + rI_1(r) \]
\[ I_n(r) = (n-2)I_{n-2}(r) + rI_{n-1}(r), \]

where $\Phi$ is the standard normal CDF (Pukkila & Rao, 1988). While the recursion can be performed quite quickly, it can be numerically unstable if $r$ is negative and $n$ is large. Alternatively, for large $n$ we have the following approximation:

\[ \log I_n(r) \approx (n/2 - 1) \log 2 + \log \Gamma(n/2) + \sqrt{nr - r^2}/4. \]

This is based on a Taylor series expansion and an application of Rocktaschel’s (1922) approximation to the gamma function. Based on this approximation, an approximately optimal test statistic is

\[ T_{AFAB}(u) = r^2/4 + \sqrt{nr - n \log x}. \] (4)

The null distributions of $T_{FAB}$ and $T_{AFAB}$ may be easily obtained via Monte Carlo simulation, because under the null hypothesis $H : \beta = 0$, the distribution of $u$ is uniform on $\mathbb{S}^{n-1}$ and so it is free of any unknown parameters. Furthermore, an approximation to the $P$-value corresponding to any test statistic $T$ based on $u = y/||y||$ may be obtained as follows:

1. Simulate $y^{(1)}, \ldots, y^{(S)} \sim \text{i.i.d. } N_n(0, I)$;
2. Compute $u^{(s)} = y^{(s)}/||y^{(s)}||$ for $s = 1, \ldots, S$;
3. Compute $\hat{p} = 1/S \sum_{s=1}^S I(T(u^{(s)}) \geq T(u)).$

This Monte Carlo approximation $\hat{p}$ to the actual $P$-value can be made arbitrarily accurate by increasing the Monte Carlo sample size $S$.

One desirable feature of the FAB test is that, unlike the $F$-test, when $n \leq p$ the FAB test has nontrivial power against certain alternative hypotheses, meaning that the power of the test is strictly greater than the level of the test. In contrast, the $F$-test has power equal to the level of the test when $n \leq p$. If the null hypothesis is false and the FAB prior distribution is concentrated around the true parameter values, the FAB test will have nontrivial power. This is especially useful in settings where the sample size is small or when there are a large number of predictor variables. Conceptually, when $n \leq p$, the FAB test is comparable to ridge regression, where a prior distribution is used to add additional structure to an otherwise degenerate inference problem. Regardless of the rank of $X$, both the $F$ and FAB tests maintain the correct level. It should be noted that if $X$ is not full-rank, there does not exist an unbiased test of the hypothesis $H : \beta = 0$ because of the lack of identifiability of the regression model. In Section 3 we provide some additional insight on the performance of the FAB test relative to the $F$-test as a function of the values of $n$ and $p$.

2.2. Extensions of the FAB Test for Non-Normal Errors

The frequentist validity of the tests based on the statistics $T_{FAB}$ and $T_{AFAB}$ depends only on the distribution of $u$ being uniform over the unit sphere under the null hypothesis. This means that the tests described above are also valid for testing $H : \beta = 0$ in any linear regression model $y = X\beta + \sigma \epsilon$, where the distribution of $\epsilon$ is spherically symmetric about the origin. Spherically symmetric distributions include heavy-tailed error distributions, such as the multivariate $t$ and...
Cauchy distributions. The test given above is usually not the marginally most powerful test based on \( u \) under the prior distribution \( \beta \sim N_p(\beta_0, \Psi) \), as the normality of \( \epsilon \) is used in the Neyman–Pearson lemma when constructing \( T_{FAB} \). If the error distribution of \( \epsilon \) is known, an analogous likelihood ratio statistic can be used to construct a marginally most powerful test based on \( y \). Typically, this likelihood ratio statistic must be computed numerically, since the marginal density of \((X\beta + \sigma \epsilon)/||X\beta + \sigma \epsilon|| \) generally does not have a simple expression when \( \epsilon \) has a non-normal distribution.

More generally, it might be the case that the linear regression model has the form \( y = X\beta + \Sigma^{1/2} \epsilon \), where the components of the error term are correlated. If \( \Sigma \) is known and \( \epsilon \sim N_p(0, I) \), the FAB test can be directly applied to the transformed data \( \Sigma^{-1/2} y \sim N_p(\Sigma^{-1/2} X\beta, I) \). However, if \( \Sigma \) is known, then there is no longer any need to require that the test statistic be a function of the maximal invariant statistic \( u \), since the distribution of \( \Sigma^{-1/2} y \) under the null hypothesis is fixed. A marginally most powerful likelihood ratio test statistic based on \( y \) under the prior distribution \( \beta \sim N_p(\beta_0, \Psi) \) is given by

\[
T_{FAB,y}(y) = y^\top \Sigma^{-1} y - (y - X\beta_0)^\top (\Sigma^{-1} + \Sigma^{-1/2} \Sigma^{-1/2})^{-1} (y - X\beta_0). \tag{5}
\]

It is possible to extend the above test to an approximate level-\( \alpha \) test when \( \Sigma \) is unknown and repeated measurements are available, although we do not pursue this extension any further in this article.

If spherical symmetry under the null distribution is suspect, randomization tests using \( T_{FAB} \) and \( T_{AFAB} \) may be implemented instead, assuming the elements of \( \epsilon \) are exchangeable. In this case, under the null hypothesis, we have \( y = \epsilon \), and so the elements of \( y \) are exchangeable. Therefore, the null distribution of any statistic \( t(y) \), conditional on the unique elements of \( y \), is the distribution of \( t(y^\sigma) \), where \( y^\sigma \) is obtained from \( y \) by randomly permuting its elements uniformly. The \( P \)-value for the randomization test based on a statistic \( t(y) \) is obtained by comparing \( t(y) \) to the distribution of \( t(y^\sigma) \). Specifically, a Monte Carlo approximation to the \( P \)-value of such a randomization test may be obtained by replacing \( y^{(1)}, \ldots, y^{(S)} \) in step 1 above with \( S \) independent random permutations of the elements of the observed data vector \( y \).

### 3. THEORETICAL POWER COMPARISONS

In this section we first examine the FAB test statistic under the prior distribution \( \beta \sim N_p(\beta_0, \gamma (X^\top X)^{-1}) \). In the extreme cases where either \( \beta_0 = 0 \) or \( \gamma \rightarrow \infty \), the FAB test is equivalent to the \( F \)-test. In another extreme case where \( \gamma = 0 \), the FAB test is equivalent to a test which we call the cone test. When both \( \beta_0 \) and \( \gamma \) are nonzero, the FAB test can be seen as an interpolation between the simpler \( F \) and cone tests. As such, in the remainder of this section we compare the asymptotic performance of the \( F \) and cone tests. In particular, we show that the cone test, if correctly specified, can substantially outperform the \( F \)-test when the dimension \( p \) of the regression subspace is large. This suggests that the FAB test is especially useful in settings with a large number of predictor variables, in which the \( F \)-test has low power.

#### 3.1. The FAB, \( F \), and Cone Tests

Suppose that the prior distribution \( \beta \sim N_p(\beta_0, \gamma (X^\top X)^{-1}) \) is used in the FAB test statistic (3). Under this prior distribution, the marginal distribution of \( X\beta \) is \( N(X\beta_0, \gamma P) \), where \( P \) is the orthogonal projection matrix onto \( \text{col}(X) \). This marginal distribution has the property that \( E(X\beta) = X\beta_0 \) and the distribution of \( X\beta - X\beta_0 \) is rotationally invariant for all orthogonal transformations that fix the subspace \( \text{col}(X) \). This class of prior distributions with covariance matrices of the form \( \gamma (X^\top X)^{-1} \) is useful in situations where it is not feasible to model the entire prior covariance matrix of \( \beta_0 \). Moreover, intuition for the behaviour of the FAB test statistic can be
obtained by using this class of prior distributions. We have that
\[ \Sigma^{-1} = \left( \gamma P + \sigma_0^2 I \right)^{-1} = \sigma_0^{-2} \left( I - \frac{\gamma}{\gamma + \sigma_0^2} P \right) = \sigma_0^{-2} (1 - w) (I - P), \]
where \( w = \gamma / (\gamma + \sigma_0^2) \). Expanding the expression (3), the
FAB test statistic in this case takes the form
\[ T_{FAB}(u) = n \log (\sigma_0) - \frac{n}{2} \log \left( 1 - w + w \| (I - P)u \|^2 \right) + \log \left( \int_0^\infty z^{n-1} \exp \left( -z^2/2 + z \frac{(1 - w) u^\top X \beta_0}{(1 - w + w \| (I - P)u \|^2)^{1/2}} \right) dz \right). \] (6)

The last term in (6) is simply an expansion of the term \( \log I_n(r) \) in (3). This expression for
the FAB test statistic shows that \( T_{FAB} \) is a function of \( \| (I - P)u \|^2 \) and \( u^\top X \beta_0 \). For a fixed value of \( \| (I - P)u \|^2 \), \( T_{FAB} \) is a strictly increasing function of \( u^\top X \beta_0 \), while for a fixed positive value of \( u^\top X \beta_0 \), \( T_{FAB} \) is a strictly decreasing function of \( \| (I - P)u \|^2 \). Consequently, the FAB
test statistic is large when \( u \) is simultaneously close to the vector \( X \beta_0 / \| X \beta_0 \| \) and close to the subspace \( \text{col}(X) \). From (6), it is immediately apparent that when \( \beta_0 = 0 \), \( T_{FAB} \) is a strictly
decreasing function of \( \| (I - P)u \|^2 \) and so it results in a test that is equivalent to the \( F \)-test.
Similarly, if \( \gamma \to \infty \), the prior distribution becomes diffuse and the FAB test statistic converges
pointwise to a limiting test statistic that is a strictly decreasing function of \( \| (I - P)u \|^2 \). It is shown in
the Appendix that the FAB test is asymptotically equivalent to the \( F \)-test as \( w \to 1 \).

Another extreme case occurs when \( \gamma = 0 \) so that \( w = 0 \) and \( T_{FAB} \) is a strictly increasing
function of \( u^\top X \beta_0 \). We call the resulting test with \( \gamma = 0 \) the cone test with test direction
\( \mu = X \beta_0 / \| X \beta_0 \| \in \mathbb{S}^{n-1} \). When the level of the cone test is less than 1/2, the test has the rejection region
\[ R_\alpha = \left\{ y : \frac{\mu^\top y}{\| y \|} > c_{1-\alpha} \right\}, \quad \mu \in \mathbb{S}^{n-1}, \] (7)
where the number \( c_{1-\alpha} > 0 \) is chosen to make this a level-\( \alpha \) test. We call the test with rejection region \( R_\alpha \) a cone test because the set \( R_\alpha \) forms a cone in \( \mathbb{R}^n \) that is rotationally symmetric about the ray extending in the direction \( \mu \) from the origin. By construction, the cone test is identical to
the likelihood ratio test of \( H : \beta = 0 \) against the simple alternative hypothesis \( H_\alpha : \beta = \beta_0 \).

3.2. Power and \( P \)-value Comparisons of the \( F \) and Cone Tests

The feature of the test statistic (6) that differentiates the FAB test from the \( F \)-test is the dependence of the FAB test statistic on \( u^\top X \beta_0 \). The quantity \( u^\top X \beta_0 \) is the cosine of the angle between the scaled data \( y / \| y \| \) and the direction of the cone test. As the prior distribution becomes more concentrated about \( \beta_0 \), the FAB test can be approximated by the cone test. Below we compare the asymptotic properties of the \( F \)-test and the cone test with rejection region given by (7).

Our first lemma compares the \( P \)-value functions of the \( F \)-test and cone test. In the univariate
setting, it is shown in Hoff (2022) that for a given observation \( y \), the \( P \)-value function of the \( F \)-test \( p_F(y) \) (or equivalently the \( P \)-value function of the two-sided \( t \)-test) at \( y \) can be at
most twice as large as the \( P \)-value function of the FAB test \( p_{FAB}(y) \). However, the log-ratio
\( \log \left( p_F(y) / p_{FAB}(y) \right) \) can be arbitrarily close to \( -\infty \), meaning that it is possible for the FAB
\( P \)-value to be substantially larger than the \( F \)-test \( P \)-value if the true \( \beta \) points in a direction opposite to that of \( \beta_0 \). The following lemma shows that in a multivariate setting the logarithm of the \( P \)-value ratio \( \log \left( p_F(y) / p_{FAB}(y) \right) \), instead of being bounded above by \( \log(2) \), is also unbounded. In particular, for observations \( y \) where \( Py = \lambda \mu, \ \lambda > 0 \), the \( P \)-value ratio of such
an observation can be expressed as a ratio of probabilities of Dirichlet random variables.
Lemma 1. Let $p_C(y)$ be the $P$-value function for the cone test with rejection region (7) for testing the null hypothesis $H_0 : \beta = 0$, where $\mu \in \text{col}(X)$ with $\|\mu\| = 1$. If the observation $y$ is of the form $y = a\mu + bv$ with $v \in \text{col}(X)^\perp$, and $a, b > 0$, then

$$\frac{p_F(y)}{p_C(y)} = \frac{P \left( \sum_{i=1}^p s_i^2 > c(y) \right)}{P \left( s_1^2 > c(y) \right)},$$

where $\left( s_1, \ldots, s_p \right) \sim \text{Dirichlet}(\frac{1}{2}, \ldots, \frac{1}{2})$ and $c(y) = a^2/(a^2 + b^2)$. In particular, for such a $y$, the logarithm of the $P$-value ratio can be bounded below by

$$\log \left( \frac{p_F(y)}{p_C(y)} \right) \geq \log \left( 2 \frac{n-1}{n-p} \right) + \frac{p-1}{2} \log \left( \frac{c(y)}{1-c(y)} \right),$$

which for $1 < p < n$ tends to $\infty$ as $c(y) \to 1$.

From this lemma it is seen that for any observation $y$ with $Py = \lambda \mu$, $\lambda > 0$, the $P$-value of the $F$-test is larger than the $P$-value of the cone test. Moreover, for such an observation, the $P$-value ratio converges to infinity as $\|I - P\|y \to 0$ when $p > 1$. Therefore, unlike in the univariate setting, the $P$-value of the cone test can be orders of magnitude smaller than the corresponding $F$-test $P$-value when $y$ is close to $\mu$. In practice, it is not realistic to observe a $y$ with $Py = \lambda \mu$, $\lambda > 0$. However, since the $P$-value ratio $p_F(y)/p_C(y)$ is continuous at $y$ as long as $p_C(y) \neq 0$, the conclusion of Lemma 1 can be extended to observations $y$ with $Py \approx \lambda \mu$, $\lambda > 0$. For instance, if $p_F(y)/p_C(y) > M$, then there is a neighbourhood of $y$ where this inequality holds for $y$ within this neighbourhood.

As the cone test depends on the regression subspace $\text{col}(X)$ only through $\mu \in \text{col}(X)$, the performance of the cone test is independent of the dimension $p$ of the regression subspace. This contrasts with the power of the $F$-test, which deteriorates as $p$ grows. Formalizing this, consider the sequence of models

$$M_n : y_n \sim N_n(X_n\beta_n, \sigma^2 I), \quad \beta_n \in \mathbb{R}^{p_n}, n \in \mathbb{N}.$$  

Define $\rho_n(c_n, p_n, \sigma^2)$ to be the power of a level-$\alpha$ test ($\rho_n$ will be the power of either the $F$-test or the cone test) of the null hypothesis $H_{0,n} : \beta_n = 0$ under the alternative hypothesis that has $\|X_n\beta_n\| = c_n$. It is of interest to assess the impact of $p_n$ and $n$ on the power functions of the $F$ and cone tests. Table 1 summarizes the limiting power $\lim_{n \to \infty} \rho_n(c_n, p_n, \sigma^2)$ of both the $F$-test and the cone test under various asymptotic regimes, when the test direction of the cone test is correctly specified. The cone test direction is correctly specified when the null hypothesis does not hold and $X\beta/\|X\beta\| = \mu$. These asymptotic regimes differ depending on whether $p_n$ is taken to be fixed as $n$ increases or whether $p_n/n \to \gamma \in (0, 1)$ as $n$ increases. Summarizing Table 1, the correctly specified cone test will have higher limiting power than the $F$-test in settings with a large number of predictor variables. A more complete description of the asymptotics of the $F$-test is provided in the following lemma. This lemma shows that under the regime where $p_n/n \to \gamma$ as $n \to \infty$, the value of $\|X_n\beta_n\|$ needs to diverge from 0 at a rate of $\|X_n\beta_n\| = n^{1/4}$ if the power of the $F$-test is to be greater than its level.

Lemma 2. Let $\rho_n(c_n, p_n, \sigma^2)$ denote the power of the level-$\alpha$ $F$-test in the sequence of models (8). If $p_n = p_0$ and $c_n = c_0$ are constants, then $\alpha < \lim_{n \to \infty} \rho_n(c_0, p, \sigma^2) < 1$. If $\gamma \in (0, 1)$, then $\lim_{n \to \infty} \rho_n(c_0, \gamma n^\frac{1}{2}, \sigma^2) = \alpha$, and if $c_n = n^{1/4}$, then the $F$-test has limiting power $\lim_{n \to \infty} \rho_n(n^{1/4}, \gamma n^\frac{1}{2}, \sigma^2) \in (\alpha, 1)$. 

DOI: 10.1002/cjs.11760  

The Canadian Journal of Statistics / La revue canadienne de statistique
Lemma 3. Let \( \rho_n(\mu_n, v_n, p_n, \sigma^2) \) denote the power of the level-\( \alpha \) cone test with rejection region \( \{ y : \langle y, \mu_n \rangle > q_{n,1-\alpha} \} \) in the sequence of models (8), where \( q_{n,1-\alpha} \) is an appropriate level-\( \alpha \) quantile, \( \mu_n \in \mathbb{S}^{n-1} \), and \( v_n = X_n\hat{\beta}_n \). If \( ||v_n||^2 = c_0 \) is constant and the mean direction of the cone test is nearly correctly specified so that \( ||\mu_n - v_n/c_0|| = o(n^{-1/2}) \), then \( \liminf_{n \to \infty} \rho_n(\mu_n, v_n, p_n, \sigma^2) \in (\alpha, 1) \), where the power function does not depend on \( p_n \). If \( ||v_n|| = n^{1/4} \) and if \( ||\mu_n - (v_n/||v_n||)|| = (n^{-1/4} - an^{-1/2}) \) for some \( a > 0 \), then \( \liminf_{n \to \infty} \rho_n(\mu_n, v_n, p_n, \sigma^2) \in (\alpha, 1) \), and if \( ||\mu_n - (v_n/||v_n||)|| = o(n^{-1/4}) \), then \( \liminf_{n \to \infty} \rho_n(\mu_n, v_n, p_n, \sigma^2) = 1 \).

This lemma implies that the performance of a misspecified cone test depends on the dimension \( n \), where the test direction \( \mu_n \) must converge to the direction of the mean vector under the alternative \( X_n\hat{\beta}_n/||X_n\hat{\beta}_n|| \) at a rate of \( o(n^{-1/2}) \) if the test is to have limiting power greater than \( \alpha \). If \( \theta_n = \angle(0, v_n, X_n\hat{\beta}_n) \) is the angle between \( v_n \) and \( X_n\hat{\beta}_n \), then this condition is equivalent to \( \theta_n = o(n^{-1/2}) \). Essentially, if \( \theta_n = \theta_0 > 0 \) is constant, the power of the cone test diminishes as \( n \) increases but does not depend on \( p_n \), while the power of the \( F \)-test diminishes as \( p_n \) increases but it is not significantly altered by \( n \).

In both Lemmas 2 and 3, it is not unrealistic to assume that \( ||X_n\beta_n|| \to \infty \) when the null hypotheses \( H_{0,n} \) do not hold. One common such scenario is where \( m \) independent replications are observed, each with design matrix \( X, \in \mathbb{R}^{k \times p_0} \), so that \( p_n = p_0 \) and \( \beta_n = \beta_0 \) are constant, while \( X_n = (1_m \otimes \hat{X}) \), \( n = km \). In this case, \( c_n = ||X_n\beta_n|| = \sqrt{n/k}||\hat{X}\beta_0|| \) and the \( F \)– and correctly specified cone tests will have limiting power 1 by the above lemmas. A related case is when \( p_n = p_0 \) and \( \beta_n = \beta_0 \) for all \( n \) and the entries of \( X_n \) are independent standard normal random variables. Then, \( E(||X_n\beta_n||^2) = n||\beta_0||^2 \) and \( E(||X_n\beta_0||) \asymp \sqrt{n} \). Therefore, the rate \( c_n = n^{1/4} \) appearing in Lemmas 2 and 3 has the alternative hypothesis diverging away from the null hypothesis at a rate that is slower than the \( \sqrt{n} \)-rate that occurs in the above replication scenarios.

It is not realistic to assume that the cone test direction is correctly specified in practice. In the multigroup setting to be considered in the following section, even as the amount of auxiliary information from other groups increases, it will generally not be true that the test direction of the cone test can be estimated consistently from the auxiliary information. Therefore, we do not recommend using the cone test on its own. Rather, we recommend using the FAB test statistic (3) or (4) with a nonzero \( \Psi \), which provides a principled compromise between the \( F \) and cone tests. If the prior information about \( \beta \) is precise and accurate, the resulting FAB test will look similar to the cone test. If, instead, only weak prior information is available, the FAB test will behave similarly to the \( F \)-test.
4. FAB TESTING IN MULTIGROUP SETTINGS

4.1. The Multigroup FAB Test

Thus far, the prior distribution \( \beta \sim N_p(\beta_0, \Psi) \) used in the FAB test was assumed to be known at the outset. In this section we demonstrate how to choose the prior distribution in a data-dependent manner in a multigroup setting. Specifically, we consider the model (1), where \( y_j \in \mathbb{R}^{n_j} \) and \( \beta_j \in \mathbb{R}^p \). We construct \( m \) different FAB tests for each of the separate hypotheses \( H_j : \beta_j = 0, j = 1, \ldots, m \) in this model. To ease notation, we focus on testing the hypothesis \( H : \beta_1 = 0 \), where any of the other hypotheses can be tested in a similar manner by relabelling the groups.

Given the multigroup regression model (1), the prior distribution in (2), which assumes that the regression coefficients are drawn from a common multivariate normal distribution, is used as a device to share information across the different groups in our multigroup FAB test. Such a prior distribution will also be referred to as a linking model. If the parameters \( \beta \) are known, the FAB test introduced in Section 2 could be directly applied. As this is generally not the case, the data from groups 2 through \( m \) will be used to obtain estimates of \( \beta_0 \) and \( \Psi \). A level-\( \alpha \) multigroup FAB procedure can be constructed as follows:

1. Obtain estimates \( \hat{\beta}_0 = \hat{\beta}_0(y_2, \ldots, y_m) \) and \( \hat{\Psi} = \hat{\Psi}(y_2, \ldots, y_m) \) of the linking model parameters using observations from every group but the first group.
2. Plug in the values of \( \hat{\beta}_0 \) and \( \hat{\Psi} \) into the FAB test statistic in (3), where \( \mu = X_1 \hat{\beta}_0 \) and \( \Sigma = X_1 \hat{\Psi} X_1^T + \sigma_0^2 I \). Denote the observed FAB test statistic by \( T_{FAB}(\frac{y_1}{||y_1||}, \hat{\beta}_0, \hat{\Psi}) \).
3. Reject the null hypothesis if \( T_{FAB}(\frac{y_1}{||y_1||}, \hat{\beta}_0, \hat{\Psi}) > q_{1-\alpha}(\hat{\beta}_0, \hat{\Psi}) \), where \( q_{1-\alpha}(\hat{\beta}_0, \hat{\Psi}) \) is the \( 1 - \alpha \) quantile of \( T_{FAB}(u, \hat{\beta}_0, \hat{\Psi}) \), where \( u \sim \text{Unif}(\mathbb{S}_{p-1}) \). This quantile can be found by Monte Carlo simulation.

This procedure results in a level-\( \alpha \) test since the independence of \( y_1 \) and \( (\hat{\beta}_0, \hat{\Psi}) \) implies that when \( \beta_1 = 0 \),

\[
P \left( T_{FAB}(\frac{y_1}{||y_1||}, \hat{\beta}_0, \hat{\Psi}) > q_{1-\alpha}(\hat{\beta}_0, \hat{\Psi}) \right) = E \left( P \left( T_{FAB}(\frac{y_1}{||y_1||}, \beta^*_0, \Psi^*) > q_{1-\alpha}(\beta^*_0, \Psi^*) \right) | \hat{\beta}_0 = \beta^*_0, \hat{\Psi} = \Psi^* \right) = E(\alpha) = \alpha.
\]

We emphasize that this FAB test will be a level-\( \alpha \) test, regardless of the validity of the linking model. All that is required for this test to have level-\( \alpha \) is that \( y_1/||y_1|| \) must be independent of \( (\hat{\beta}_0, \hat{\Psi}) \) and it must be uniformly distributed over the unit sphere.

Another desirable feature of the multigroup FAB test is that it approximately has the Bayes-optimal power if the linking model holds and if the parameter estimates \( (\hat{\beta}_0, \hat{\Psi}) \) are close to \( (\beta_0, \Psi) \). This follows by construction, since the likelihood ratio test that the multigroup FAB test approximates is the most powerful test marginally over the linking model. The following lemma shows that if consistent estimates of the prior parameter for the multigroup FAB test are available, then asymptotically this test has the Bayes-optimal power.

**Lemma 4.** Let \( \rho(\beta_m, \Psi_m) \) denote the marginal power of the FAB test, where the test statistic is constructed using the prior distribution \( \beta \sim N_p(\beta_m, \Psi_m) \) and the power is evaluated under the
marginal distribution $\beta \sim N_p(\beta_0, \Psi_0)$. If $(\beta_m, \Psi_m) \rightarrow (\beta_0, \Psi_0)$, then $\rho(\beta_m, \Psi_m) \rightarrow \rho(\beta_0, \Psi_0)$. If $(\hat{\beta}_m, \hat{\Psi}_m)$ converges to $(\beta_0, \Psi_0)$ in probability, then $E\left(\rho(\hat{\beta}_m, \hat{\Psi}_m)\right) \rightarrow \rho(\beta_0, \Psi_0)$. In particular, if $(\hat{\beta}_m, \hat{\Psi}_m)$ are consistent estimates of the prior parameters $(\beta_0, \Psi_0)$ in a multigroup FAB test with $m$ groups, then asymptotically as the number of groups grows, the marginal power of the multigroup FAB test converges to the marginal power of the Bayes-optimal test $\rho(\beta_0, \Psi_0)$.

The simulations in Table 2 empirically demonstrate how the marginal power of the multigroup FAB test converges to the marginal power of the Bayes-optimal test. The powers in this table are computed by drawing observations from the model

$$y_i \sim N_{12}(X \beta_i, \mathbf{I}), \quad \beta_i \sim N_p(1, \tau^2 \mathbf{I}), \quad i = 1, \ldots, m,$$

where the design matrix $X$ has orthonormal columns equal to the first four standard basis vectors in $\mathbb{R}^{12}$. Moment-based estimates of $(\beta_0, \tau^2)$, as described in the subsequent section, are used in the multigroup FAB test. From Table 2 it is evident that when 20 groups are present, the power of the multigroup FAB test is nearly identical to that of the Bayes-optimal test. Additionally, it is seen that as the between-group variance $\tau^2$ increases relative to $\sigma^2$, the marginal power of every test decreases. Note that the marginal power of the Bayes-optimal test and of the $F$-test does not depend on the number of groups.

4.2. Constructing a Data-Dependent Prior Distribution

In this section we review some standard methods for obtaining estimates of $\beta_0$ and $\Psi$ from the random effects model

$$y_j | \beta_j \sim N_{n_j}(X_j \beta_j, \sigma_j^2 \mathbf{I}), \quad \beta_j \sim N_p(\beta_0, \Psi), \quad j = 2, \ldots, m.$$ (9)

For simplicity, assume that $\sigma_2^2 = \cdots = \sigma_m^2$, where we denote the common value by $\sigma^2$. The marginal distribution of $y_j$ is $N(X_j \beta_0, \Sigma_j)$ with $\Sigma_j = X_j \Psi X_j^\top + \sigma^2 \mathbf{I}$. Under this marginal model for the $y_j$'s, the maximum likelihood estimator of $\beta_0$ is

$$\hat{\beta}_0 = \left(\sum_{j=2}^m X_j^\top \Sigma_j^{-1} X_j\right)^{-1}\left(\sum_{j=2}^m X_j^\top \Sigma_j^{-1} y_j\right),$$ (10)

where $\hat{\Sigma}_j$ is the maximum likelihood estimator of $\Sigma_j$. The maximum likelihood estimators of $\sigma^2$ and $\Psi$ have to be found numerically using, for example, the R package lme4 (Bates et al., 2015). As a simpler alternative, moment-based estimates of $\sigma^2$ and $\Psi$ can be found, which then can be

| $\tau^2$ | 1/3 | 1 | 3 |
|---|---|---|---|
| Number of groups | 3 | 8 | 20 |
| Multigroup FAB test | 0.55 | 0.61 | 0.64 |
| Bayes-optimal test | 0.65 | 0.57 | 0.57 |
| $F$-test | 0.58 | 0.36 | 0.25 |

Table 2: Marginal power ($\pm 0.01$) for differing values of the between-group variance.
substituted into the values of $\sum_j$ appearing in (10). If $P_j$ is the projection matrix onto $\text{col}(X_j)$, the residual maximum likelihood estimate (REML) of $\sigma^2$ is given by

$$\hat{\sigma}_{REML}^2 = \left( \sum_{j=2}^m (n_j - p) \right)^{-1} \sum_{j=2}^m \| (I - P_j)y_j \|^2. \tag{11}$$

A simple moment-based estimate of $\Psi$ is

$$\hat{\Psi} = \frac{1}{m - 1} \sum_{j=2}^m \left( (X_j^T X_j)^{-1} X_j^T (y_j - X_j \hat{\beta}_0) (y_j - X_j \hat{\beta}_0)^T X_j (X_j^T X_j)^{-1} - \hat{\sigma}_{REML}^2 (X_j^T X_j)^{-1} \right). \tag{12}$$

As $\hat{\Psi}$ depends on $\hat{\beta}_0$, which in turn depends on $\hat{\Psi}$, an iterative procedure is needed to find suitable estimates. Such an iterative procedure can be initialized by taking $\hat{\Psi} = I$ in (10).

In summary, there is flexibility as to what estimates of $\beta_0, \Psi$, and $\sigma^2$ are used in the multigroup FAB procedure, as long as such estimates are independent of $y_1$. If the values of the $n_j$'s and $p$ are large, the estimates in (11) and (12) may be easier to compute than the maximum likelihood estimates. The restriction that $\sigma^2_2 = \cdots = \sigma^2_m$ can also be lifted at the expense of additional computational effort. In this case, either the MLE or direct analogues of the estimators in (11) and (12) could be used. Lifting this restriction may result in better estimates of the linking model parameters $\beta_0$ and $\Psi$ if the error variances across groups are different. However, the FAB test remains a level-$\alpha$ test regardless of the particular linking model parameter estimates chosen.

4.3. Modelling the Error Variances

As discussed in Section 3, the FAB test can approximately be viewed as a combination of the cone and $F$-tests. Roughly, if $\beta_1$ is given the prior distribution $N(\beta_0, \Psi)$, the prior mean $\beta_0$ determines the test direction of the cone test, while the relative magnitudes of $\Psi$ and $\sigma^2_0$ determine how similar the FAB test is to either the $F$-test or the cone test. Recall that $\sigma^2_0$ was the location of the point mass prior distribution placed on $\sigma^2_1$. In this section we pursue a more sophisticated FAB test where a linking model for the error variances $\sigma^2_j$ is added to the model (9). For illustrative purposes, we consider the inverse-gamma model

$$\sigma^2_1, \ldots, \sigma^2_m \sim \text{i.i.d. Inverse-Gamma}(\alpha, \beta). \tag{13}$$

Other linking models over the error variances can be handled analogously.

Two additional steps are needed to incorporate the inverse-gamma linking model into the multigroup FAB test previously discussed. First, the FAB test statistic (3) is altered to account for the new linking model (13). The hyperparameters $(\alpha, \beta)$ of the error variance linking model appear in this modified test statistic. Second, the observations in groups 2 through $m$ are used to obtain estimates of these hyperparameters to be used in the modified test statistic.

As before, by the Neyman–Pearson lemma, the FAB test statistic is the likelihood ratio test of the densities of $y_1/\|y_1\|$ under the new linking model and under the null hypothesis. Under the null hypothesis, $y_1/\|y_1\|$ remains uniformly distributed over the sphere, while under the linking model we have $y_1/\|y_1\| \sim AG(\mu, \Sigma)$ with $\mu = X_1 \beta_0$, $\Sigma = X_1 \Psi X_1^T + \sigma^2_1 I$ and $\sigma^2_1 \sim IG(\alpha, \beta)$. Therefore, the FAB test statistic is

$$T_{IG-FAB}(u) = \int_0^\infty |\Sigma|^{-1/2} x^{-n} I_{n}(r) \exp((r^2 - \mu^T \Sigma^{-1} \mu)/2)\pi_{\alpha,\beta}(\sigma^2_1) d\sigma^2_1, \tag{14}$$

DOI: 10.1002/cjs.11760 The Canadian Journal of Statistics / La revue canadienne de statistique
where \( x \) and \( r \) are defined as in Section 2 and \( \pi_{u, \beta} \) is the density of an \( IG(\alpha, \beta) \) distribution. This statistic can be found via Monte Carlo approximation or numerical integration over the inverse gamma distribution.

Parameter estimates of \( \beta_0, \Psi, \alpha, \) and \( \beta \) based on the data in groups 2 through \( m \) can be substituted into (14) to obtain a test statistic that shares information across groups. One strategy for obtaining such estimates is to estimate \( \beta_0 \) and \( \Psi \) via equations (10) and (12) as described in the previous section. Estimates of \( \alpha \) and \( \beta \) can be obtained through the method of moments.

### 4.4. Testing Other Linear Hypotheses

In this section we describe how to test linear hypotheses in the multigroup regression model that are more general than the hypothesis \( H : \beta_1 = 0 \). For instance, in (1) it may be of interest to test if a subset of components of \( \beta_1 \) are 0, or to test the hypothesis \( H : \beta_1 = \beta_2 \) that the regression coefficients of two different groups are equal.

Define \( \beta_{1;l}^* = \left( \beta_{11}, \ldots, \beta_{1l} \right) \) for \( l < m \), and let \( A \in \mathbb{R}^{p \times l} \), \( v \in \mathbb{R}^l \) with \( v \in \text{col}(A) \). A FAB procedure for testing the linear hypothesis \( H : A \beta_{1;l} = v \) on the regression coefficients of the first \( l \) groups in the model (1) is described below. We make the extra assumption in (1) that the error variances \( \sigma^2 \) are all equal to \( \sigma^2 \). Define \( y_{l:k}^* = \left( y_{1:k}^T, \ldots, y_{l:k}^T \right) \), and \( X_{l:k}^* = \left[ X_{1:k}^T, \ldots, X_{l:k}^T \right] \). The multigroup regression model (1) under the homoskedasticity assumption can be rewritten as

\[
y_{1:l} \sim N(X_{1:l} \beta_{1:l}, \Sigma^2 I), \quad y_j \sim N_{n_j}(X_j \beta_j, \Sigma^2 I), \quad j = l + 1, \ldots, m.
\]

Let \( S = \{ X_{1:l} \beta_{1:l} : A \beta_{1:l} = 0 \} \), and take \( W \) to be a full-rank orthonormal matrix whose rows span the subspace \( S^\perp \). Also take \( \beta_{1;l}^* \) to be a solution to the equation \( A \beta_{1;l}^* = v \), where we define \( \mu^* = WX_{1:l} \beta_{1;l}^* \). By the definition of \( W \), the value of \( WX_{1:l} \beta_{1;l}^* \) is independent of the particular solution \( \beta_{1;l}^* \) chosen. Moreover, if the hypothesis \( H \) holds, this implies that the hypothesis \( H^* : WX_{1:l} \beta_{1;l} - \mu^* = 0 \) also holds. When \( X_{1:l} \) is full-rank, or more generally when \( WX_{1:l} \) is full-rank, these hypotheses are equivalent, meaning that \( H^* \) is true if and only if \( H \) is true.

As \( W y_{1:l} - \mu^* \sim N(WX_{1:l} \beta_{1:l} - \mu^*, \Sigma^2 I) \), the hypothesis \( H^* \) is identical to the hypothesis considered in Section 2, namely testing that the mean of the isotropic, multivariate normal random vector \( W y_{1:l} - \mu^* \) is 0. Applying the results from Section 2, if \( \beta_{1:l} \) is given the prior distribution \( \beta_{1:l} \sim N_p(1 \otimes \beta_0, I \otimes \Psi) \), the FAB test of \( H^* \) is exactly the likelihood ratio test with test statistic (3), where \( u = (W y_{1:l} - \mu^*)/||(W y_{1:l} - \mu^*)|| \) and \( \mu \) and \( \Sigma \) are the marginal mean and variance of \( W y_{1:l} - \mu^* \). We note that the common variance assumption is crucial in this setting to ensure that the distribution of \( u \) under the null hypothesis is pivotal. If this assumption is violated, the resulting test might not be a level-\( \alpha \) test.

A data-dependent prior distribution over \( \beta_{1:l} \) is obtained by finding estimates for \( \beta_0 \) and \( \Psi \) in the linking model (2). Such estimates are found exactly as in Section 4.2, using the observations \( y_{1+1}, \ldots, y_m \) that are not from the first \( l \) groups. This procedure does not utilize any information from the portion \( (I - W^TW)y_{1:l} \) of \( y_{1:l} \) that lies in \( S \). If \( \tilde{W} \) is a full-rank matrix with rows that span \( S \), then marginally under the prior distribution (2), \( \tilde{W} y_{1:l} \sim N(WX_{1:l}(1 \otimes \beta_0), \Sigma^2 I + \tilde{W}(I \otimes \Psi)\tilde{W}^T) \). The vector \( \tilde{W} y_{1:l} \) therefore does provide some useful information about the covariance structure of \( \beta_{1:l} \) This information is most easily incorporated into a maximum likelihood approach for estimating \( \beta_0 \) and \( \Psi \). However, in hypotheses where \( l \ll m \), such as the hypothesis that a subset of the components of \( \beta_1 \) are 0 when a large number of groups are present, we recommend using the simpler prior parameter estimates based on the observations \( y_{l+1}, \ldots, y_m \).
5. EXAMPLE: EVALUATING STANDARDIZED TEST SCORES

5.1. Overview of the ELS Data

In this section we demonstrate the efficacy of the multigroup FAB test on educational outcome data. The 2002 educational longitudinal study (ELS) dataset includes demographic information of 15,362 students from a collection of 751 schools across the United States in an effort to inform educational policy. We identify schools with ethnic disparities in educational outcomes by testing for mean differences in test scores by ethnicity after accounting for other variables, such as socioeconomic status and parental education level. Darling-Hammond (1998) and Camara & Schmidt (1999) describe several studies that have detected test differences between ethnicities after accounting for such variables, and also describe potential causes. To further evaluate the cause of such disparities and find potential remedies, it is useful to first identify schools in which disparities likely exist.

As some schools had only a small number of students who were surveyed, sharing information between the schools can help to improve the sensitivity of within-school testing procedures. On average, 20 students were surveyed per school; however, 34 of the schools had fewer than 10 students who were surveyed. The response variable that we analyze is a nationally standardized composite math and reading score that is recorded for each student in the study. For each student, we model the relationship between their test score and the following dependent variables: ethnicity, native language, sex, parental education, and a composite index of the student’s socioeconomic status. Ethnicity is aggregated into four broad categories: Asian, Black, Hispanic, and White. A separate linear regression model with the aforementioned predictor variables is used for each school. Independently across schools \( j = 1, \ldots, 751 \), we take

\[
y_j \sim N_{nj}\left(Z_j\alpha_j + X_j\beta_j, \sigma_j^2 I\right), \quad j = 1, \ldots, 751,
\]

where \( y_j \) is the vector of test scores for the students in school \( j \), \( \beta_j \) represents the regression coefficients for the ethnicity variables, and \( \alpha_j \) represents the regression coefficients for the nonethnicity variables in school \( j \). For every school \( j \), we test the hypotheses \( H_j : \beta_j = 0 \) by projecting out the nonethnicity variables in (16), a process that was described in Section 4.4.

Figure A1 in the Appendix illustrates normal QQ-plots of the residuals in the projected models for nine different schools. These plots suggest that the projected regression model is a reasonable model for the ELS data. Figure 1 shows the distribution of the least-square estimates of the \( \beta_j \) coefficients for schools with full-rank projected design matrices. From this figure, we conclude that it is not unrealistic to assume that the \( \beta_j \)'s follow the multivariate normal linking model in (2). The 95% confidence ellipses for \( \beta_j \) in Figure 1 are based on a method-of-moments estimate of \( \Psi \).

Lastly, we examine the empirical distribution of the scaled squared-residuals \( e_j^2 = \| (I - P_j) y_j \|^2 \), \( j = 1, \ldots, 751 \) to determine a suitable linking model for the error variances. A kernel density estimate of the marginal density of \( e_j^2/(n_j - p) \), \( j = 1, \ldots, 751 \) is shown in Figure 2. Under a linking model that assumes homoskedasticity, \( \sigma^2 = \sigma_1^2 = \cdots = \sigma_{751}^2 \), and \( e_j^2/(n_j - p) \) has a chi-squared marginal distribution scaled by the constant \( \sigma^2/(n_j - p) \). The second plot in Figure 2 sets \( \hat{\sigma}^2 = \frac{1}{751} \sum_{j=1}^{751} e_j^2/(n_j - p) \) and displays a kernel density estimate of the marginal distribution of \( \{ \hat{\sigma}^2 w_{11}/(n_1 - p), \ldots, \hat{\sigma}^2 w_{751}/(n_{751} - p) \} \), where this distribution is found by simulating \( w_j \sim \chi_{n_j-p}^2 \). It is apparent that the kernel density estimate of this marginal distribution does not match the kernel density estimate of the observed marginal distribution. Two other possible linking models are the inverse-gamma linking model \( \sigma_j^2 \sim IG(\alpha, \beta) \) and the truncated normal linking model \( \sigma_j^2 = \sigma_0^2 |z_j|, \quad z_j \sim N(\mu, \tau^2) \). Fitting both of these models, it is seen

DOI: 10.1002/cjs.11760
FIGURE 1: Pair plots of the least-squares estimates of the ethnicity regression coefficients along with 95% confidence ellipses for $\beta_j$.

FIGURE 2: Kernel density estimates of the marginal distribution of the scaled squared residuals under different linking models.

in Figure 2 that the marginal density of $\sigma_j^2 w_j / (n_j - p)$, $w_j \sim \chi^2_{n_j - p}$, $j = 1, \ldots, 751$ under the truncated normal linking model with $z_j \sim N(0.2, 1.3)$ matches the observed marginal density more closely than the marginal density under the inverse-gamma linking model.

5.2. Methodology

We test the hypothesis $H_j : \beta_j = 0$ in the model (16) by projecting out the nonethnicity variables as described in Section 4.4. That is, if $W_j$ is a full-rank orthonormal matrix whose rows span $\text{col}(Z_j)^\perp$, the problem of testing $\beta_j = 0$ is reduced to testing this same hypothesis in the model $W_j y_j \sim N(W_j X_j \beta_j, \sigma_j^2 I)$. The FAB or $F$-tests in the reduced model will have power greater than the level against some alternatives as long as $W_j X_j \neq 0$, a condition which holds for 634 of the 751 schools. The remaining 117 schools are not considered in our analysis. Out of
these 634 schools, 169 have full-rank projected design matrices $W_jX_j$. In the FAB tests we construct, the multivariate normal linking model (2) is used to model $\hat{\beta}_j$. Parameter estimates of $\beta_0$ and $\Psi$ are obtained using only the 169 schools that have full-rank design matrices. The method-of-moments estimates of $(\beta_0, \Psi)$ described in Section 4.2 are used as estimates of these linking model parameters. As there are a sufficiently large number of observations from which to obtain parameter estimates, other estimation procedures will produce similar estimates.

Two different linking model assumptions on the error variances $\sigma_j^2$ are used to illustrate how sensitive the FAB testing procedure is to the error variance linking model. In practice, to avoid $p$-hacking, only a single linking model should be used. The first is simply the homoskedastic linking model that uses the FAB test statistic (3), where the parameter $\sigma_0^2$ is estimated by the mean squared error of the reduced model, pooled across schools. The second truncated normal linking model assumption assumes that $\sigma_j^2 = \sigma^2 z_j$, $z_j \sim N(0.2, 1.3)$, $j = 1, \ldots, 751$. We refer to these tests as FAB-HS and FAB-TN, respectively. Again, the pooled mean squared error is used to estimate $\sigma_0^2$ in the truncated normal linking model. Denoting the marginal distribution of $\sigma_j^2$ by $\pi_{TN}(\sigma_j^2)$, analogous to Equation (14), the FAB test statistic for school $j$ is given by

$$T_{\text{FAB-TN}}(u) = \int |\tilde{\Sigma}|^{-1/2} x^{-n} I_n(r) \exp \left((r^2 - \hat{\mu}^\top \tilde{\Sigma}^{-1} \hat{\mu})/2\right) \pi_{TN}(\sigma_j^2) d\sigma_j^2, \quad (17)$$

where $\tilde{\Sigma} = W_jX_j\Psi X_j^\top W_j^\top + \sigma_0^2 I$, and $x$ and $r$ are defined as in Section 2 based on the data from school $j$. The integral in (17) is approximated via Monte Carlo by drawing 100 observations of $\sigma_j^2$ from the truncated normal distribution.

To obtain the correct level when testing the hypothesis $H_j : \beta_j = 0$ for school $j$, theoretically, all of the estimates $\left(\hat{\beta}_0, \hat{\Psi}, \hat{\sigma}_0^2\right)$ should be computed by leaving out the data from school $j$. For instance, when testing the hypothesis for school 2, $\hat{\beta}_0$ should be of the form $\hat{\beta}_0(y_1, y_3, \ldots, y_{751})$. However, as these parameter estimates are not significantly altered by leaving out school $j$, it is not necessary in practice to recomputed these estimates for each school. Figure A2 in the Appendix displays the $P$-values for the homoskedastic FAB test, where the linking model parameter estimates are either recomputed when testing each hypothesis or the same parameter estimates are used for testing every hypothesis. It is seen that the $P$-values are nearly identical in these two cases, showing that at least when a large number of groups are present, it is not necessary to recompute the linking model parameter estimates for each hypothesis under consideration.

5.3. Results

We compare the empirical performance of three tests: the $F$-test, the homoskedastic FAB test, and the truncated normal FAB test. In the second plot in Figure 3, it is shown that the logarithm of the $P$-values for the homoskedastic and truncated normal FAB tests are nearly identical. This provides some evidence that the choice of the linking model over the $\sigma_j^2$s is not of critical importance. In the first plot, the magnitudes of the logarithm of the $P$-values of the $F$-test are seen to generally be larger than the logarithm of the $P$-values for the homoskedastic FAB test.

Table 3 summarizes the percentage of null hypotheses rejected for the FAB tests and the $F$-test. Across all of the 634 schools with nonzero projected design matrices, at a level of 0.05, the FAB test rejects nearly twice as many hypotheses as the $F$-test. At a level of 0.01, the relative improvement is even greater, with the FAB tests rejecting approximately 3 times as many hypotheses. The average number of hypotheses rejected is higher for the 169 schools that
have full-rank projected design matrices than for schools that do not. If $W_jX_j$ is not full-rank, then the full vector $\beta_j$ is not identifiable in (16); rather only certain linear functions of $\beta_j$ are identifiable. Therefore, if a school does not have a full-rank projected design matrix, not as much prior information can be leveraged in the FAB test. Effectively, prior information can only be used to directly inform the identifiable “portion” of $\beta_j$. For instance, if the first column of $W_jX_j$ happened to be 0, the FAB test would not directly utilize the information from the other schools about the marginal distribution of the first component of $\beta_j$.

Lastly, we examine the performance of the FAB tests relative to the $F$-test after controlling the false discovery rate (FDR). Controlling the FDR is an alternative to controlling the family-wise error rate (FWER), where controlling for the latter can result in an overly restrictive testing procedure. The Benjamini–Hochberg procedure controls the FDR at a level $\alpha$ by sorting the observed $P$-values $p(1) < \cdots < p(m)$ and rejecting all hypotheses $H_i$, $i \leq i^*$ where $i^*$ is the largest $i$ for which $p(i) \leq \alpha i^*/m$ (Benjamini & Hochberg, 1995). This procedure is valid for $P$-values $\{p_1, \ldots, p_m\}$ that are independent, as in the $F$-test, or for $P$-values that satisfy certain types of positive dependence (Benjamini & Yekutieli, 2001). As the $P$-values obtained from the FAB test are constructed from estimates of the linking model parameters that are correlated, we expect the $P$-values to be positively dependent and the Benjamini–Hochberg procedure to be approximately valid for the FAB tests. However, further investigations are needed to verify whether the Benjamini–Hochberg procedure correctly controls the FDR for the multigroup FAB test. Table A1 in the Appendix compares the average number of hypotheses rejected at different levels, for schools with and without full-rank design matrices.
that are rejected for the $F$-test and the FAB-HS test, controlling the FDR at various levels by the Benjamini–Hochberg procedure. The plot of the sorted $P$-values for the $F$-test and the FAB-HS test in Figure 4 demonstrates that the empirical distribution of the FAB-HS $P$-values is stochastically smaller than the empirical distribution of the $F$-test $P$-values. In conclusion, whether controlling for size or for the FDR, the FAB tests in this setting are seen to be more powerful than the $F$-test.

6. DISCUSSION

In multigroup data analyses, inferences about one group can often be made more precise by using data from the other groups. In this article, we have shown how a FAB test for a linear hypothesis involving one group may be constructed with the aid of data from the other groups, via a linking model that describes relationships among group-specific parameters. If the data from one group is informative about another group’s parameters, then such a test will have higher power than a direct test that does not make use of this information. Additionally, even if the linking model is incorrect or noninformative, the FAB test maintains exact type I error rate control.

The FAB test statistic we have proposed is a function of the scale-invariant statistic $\frac{y}{\|y\|}$ which has a distribution that does not depend on $\sigma^2$. Effectively, by only considering scale-invariant test statistics, the null hypothesis has been reduced to a simple null hypothesis. This reduction to test statistics that are pivotal under the null hypothesis can be applied more broadly to test a wide variety of hypotheses. An interesting future direction is to develop FAB tests for nonparametric hypotheses, by restricting the test statistic to be a function of a pivotal quantity. One such example is to test the hypothesis that two distributions are equal using test statistics based on empirical distributions.

In nonparametric settings, it may also be necessary to consider more sophisticated prior distributions than the multivariate normal prior used in this article. In fact, the FAB prior distribution can itself be nonparametrically estimated. The possible utility of doing so is suggested by the ELS example where kernel density estimates of the scaled squared residuals are shown in Figure 2. Rather than using the parametric truncated normal or inverse-gamma linking models, a kernel density estimate of the density of the variance parameter could instead...
be used. However, if the model under consideration is parametric, for computational reasons it is preferable to keep the linking model as simple as possible.

This article is focused entirely on testing the values of regression coefficients. In theory, the FAB test presented could be inverted to provide a confidence region for each vector of regression coefficients. Hoff & Yu (2019) examine related confidence intervals for the elements of \( \beta \) in a regression model for a single group using shrinkage prior distributions. In the multigroup setting, properties such as the connectedness or convexity of such a confidence region that is found by inverting the FAB test warrant further study.

Some other aspects of the multigroup FAB test that warrant further study are the multiple testing properties of this test. The multigroup FAB test controls the type I error rate for each group and thus controls the per-comparison error rate, but it does not control the family-wise error rate (Dudoit & van der Laan, 2008). Standard methods, such as the Bonferroni correction, can be used to control the family-wise error rate, although such methods may produce tests with low power. Similarly, it is also of interest to study methods for controlling the false discovery rate of the multigroup FAB test.

ACKNOWLEDGEMENT

We thank the associate editor and referee for their helpful and constructive comments.

REFERENCES

Aerts, M., Claeskens, G., & Hart, J. D. (2004). Bayesian-motivated tests of function fit and their asymptotic frequentist properties. *The Annals of Statistics*, 32, 2580–2615.

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Benjamini, Y. & 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.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29, 1165–1188.

Camara, W. J. & Schmidt, A. E. (1999). *Group differences in standardized testing and social stratification*, College Entrance Examination Board, Report No. 99-5.

Chacon, J. E., Montanero, J., Nogales, A. G., & Perez, P. (2007). On the use of Bayes factor in frequentist testing of a precise hypothesis. *Communications in Statistics Theory and Methods*, 36, 2251–2261.

Darling-Hammond, L. (1998). Unequal opportunity: Race and education. *The Brookings Review*, 16(2), 28–32.

Dudoit, S. & van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*, Springer, New York.

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87, 597–606.

Good, I. J. & Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of the American Statistical Association*, 69, 711–720.

Henderson, C. R. (1974). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2), 423–447.

Hoff, P. D. (2022). Smaller \( p \)-values via indirect information. *Journal of the American Statistical Association*, 117, 1254–1269.

Hoff, P. D. & Yu, C. (2019). Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, 13, 94–119.

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B*, 24, 1–41.

O’Gorman, T. W. (2002). An adaptive test of significance for a subset of regression coefficients. *Statistics in Medicine*, 21, 3527–3542.
Lemma 1. Let $p_C(y)$ be the $P$-value function for the cone test with rejection region (7) for testing the null hypothesis $H_0 : \beta = 0$, where $\mu \in \text{col}(X)$ with $\|\mu\| = 1$. If the observation $y$ is of the form $y = a\mu + bv$ with $v \in \text{col}(X)^\perp$, and $a, b > 0$, then

$$
p_F(y) = \frac{P(\sum_{i=1}^p s_i^2 > c(y))}{P(\sum_{i=p+1}^n s_i^2 > c(y))},
$$

where $(s_1^2, \ldots, s_n^2) \sim \text{Dirichlet}_n\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$, and $c(y) = a^2/\left(a^2 + b^2\right)$. In particular, for such a $y$, the $P$-value ratio can be bounded below by

$$
p_F(y) \geq 2 \left(\frac{n-1}{n-p}\right)^{\frac{p-1}{2}} \left(\frac{c(y)}{1-c(y)}\right),
$$

which tends to $\infty$ as $c(y) \to 1$ if $1 < p < n$.

Proof. Without loss of generality, the regression subspace $\text{col}(X)$ can be assumed to be equal to $\text{span}\{e_1, \ldots, e_p\}$, where $e_i$ is the $i$th standard basis vector and $\mu$ can be taken to be $e_1$. Thus $y = ae_1 + bv$, and $v$ can be taken to be $e_{p+1}$ since both the $F$ and cone test statistics depend only on $y$ through the values of $a$ and $b||y||$. At $y = ae_1 + be_{p+1}$, the $F$ and cone test statistics are $((||P_Xy||^2/p)/(||I_n - P_Xy||^2/(n-p))) = (a^2/p)/(b^2/(n-p))$ and $\langle y/||y||, e_1 \rangle = a/(a^2 + b^2)^{1/2}$, respectively.

Under the null hypothesis, $y \sim N_n(0, \sigma^2 I_n)$ and the vector $s = y/||y||$ has $(s_1^2, \ldots, s_n^2) \sim \text{Dirichlet}_n(1/2, \ldots, 1/2)$. The $F$-test $P$-value at $ae_1 + be_{p+1}$ therefore has the form

$$
p_F(y) = P\left(\sum_{i=1}^p s_i^2 > \frac{a^2}{b^2}\right) = P\left(\sum_{i=p+1}^n s_i^2 > \frac{a^2}{a^2 + b^2}\right)
$$

since $||P_XY||^2/||(I_n - P_X)Y||^2 \overset{d}{=} \left(\sum_{i=1}^p s_i^2\right)/\left(\sum_{i=p+1}^n s_i^2\right)$. The cone test $P$-value is

$$
p_C(y) = P\left(s_1 > \frac{a}{\sqrt{a^2 + b^2}}\right) = \frac{1}{2} P\left(s_1^2 > \frac{a^2}{a^2 + b^2}\right).
$$
We bound this ratio of beta probabilities below by

$$\frac{p_F(y)}{p_C(y)} = \frac{2P(\sum_{i=1}^{p} s_i^2 > c)}{P(s_1^2 > c)} = \frac{2\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n-1}{2}\right) \int_x^1 x^{\frac{n}{2}-1} (1-x)^{\frac{n-p-1}{2}} dx}{\Gamma\left(\frac{p}{2}\right) \Gamma\left(\frac{n-p}{2}\right) \int_x^1 x^{\frac{1}{2}-1} (1-x)^{\frac{n-1}{2}} dx}$$

$$\geq \frac{2c_{p-1}^{-1} (1-x)^{\frac{n-p}{2}} dx}{c^{-2} \int_x^1 (1-x)^{\frac{n-1}{2}} dx} = 2 \left( \frac{n-1}{n-p} \right) c_{n-1}^{-\frac{1}{2}} (1-c)^{\frac{1}{2} - \frac{p}{2}}.$$  

The inequality above follows from $\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n-1}{2}\right) \geq \Gamma\left(\frac{p}{2}\right) \Gamma\left(\frac{n-p}{2}\right)$ and by applying $x^{n/2-1} \geq c_{n/2-1}, \ x^{-1/2} \leq c^{-1/2}$ for all $x \in [c, 1]$ in the numerator and denominator, respectively.  

\[\square\]

**Lemma 2.** Consider the sequence of models $\mathcal{M}_n: y_n \sim N_n(X_n \beta_n, \sigma^2 I)$, where $\beta_n \in \mathbb{R}^{p_n}$.

Define $\rho_n(c_n, p_n, \sigma^2)$ to be the power of the level-$\alpha$ $F$-test of the null hypothesis $H_{0,n}: \beta_n = 0$ under the alternative hypothesis that $||X_n \beta_n|| = c_n$. If $p_n = p_0$ and $c_n = c_0$ are constants, then $\alpha \leq \lim \inf_{n \to \infty} \rho_n(c_0, p_0, \sigma^2) < 1$. If $\gamma \in (0,1)$, then $\lim \inf_{n \to \infty} \rho_n(c_0, |\gamma n|, \sigma^2) = \alpha$, and if $c_n = n^{1/4}$, then the $F$-test has limiting power $\lim \inf_{n \to \infty} \rho_n(n^{1/4}, |\gamma n|, \sigma^2) \in (\alpha, 1)$.

**Proof.** The $F$ statistic is

$$F_n(y_n) = \frac{\| \mathbf{P}_{X_n} y_n \|^2 / p_n}{\| (\mathbf{I} - \mathbf{P}_{X_n}) y_n \|^2 / (n - p_n)}$$

In the first regime where $p_n = p_0$ and $c_n = c_0$, this statistic converges in probability to a $\chi^2_{p}$ distribution under the null hypothesis and a $\chi^2(c_0^2)$ distribution under the alternative hypothesis since the converges in probability to $\sigma^2$ and the numerator follows a chi-squared distribution. If $Z \sim \chi^2_{p}(c_0^2)$, then $\lim \inf_{n \to \infty} \rho_n(c_0, p_0, \sigma^2) = P\left( Z > \chi^2_{p_0,1-\alpha} \right) \in (\alpha, 1)$ as needed.

Next, consider the case where $p_n = |\gamma n|$. Under the alternative hypothesis, the numerator of the $F$-statistic has a $\sigma^2 \chi^2_{p_n}(c_n^2/\sigma^2)$ distribution with the stochastic representation $\sigma^2(z_1 + c_n/\sigma^2)^2 + \sigma^2 \sum_{i=2}^{p_n} z_i^2 = c_n^2 + 2z_1 c_n/\sigma^2 + \sigma^2 \sum_{i=1}^{p_n} z_i^2$, where $z_i \overset{i.i.d.}{\sim} N(0,1)$. Define $w_n = (n - p_n) \sigma^2 \sum_{i=1}^{p_n} (z_i^2 - 1) / \left( \sqrt{2p_n} \| (\mathbf{I} - \mathbf{P}_{X_n}) y_n \|^2 \right)$. As $n \to \infty$, $w_n \overset{d}{\to} N(0,1)$ and, consequently, $\sqrt{p_n/2} \left( F_{p_n,n-p_n,1-\alpha} - 1 \right) \to z_{1-\alpha}$, where $z_{1-\alpha}$ is the $1 - \alpha$ standard normal quantile. Consequently

$$\lim \inf_{n \to \infty} \rho_n(c_n, p_n, \sigma^2) = \lim \inf_{n \to \infty} P \left( w_n + \frac{c_n^2 / \sqrt{p_n} + 2z_1 c_n/\sqrt{p_n}}{\sqrt{2\| (\mathbf{I} - \mathbf{P}_{X_n}) y_n \|^2 / (n - p_n)}} > \sqrt{p_n/2} \left( F_{p_n,n-p_n,1-\alpha} - 1 \right) \right).$$

If $c_n = o\left(p_n^{1/4}\right)$, then $\lim \inf_{n \to \infty} \rho_n(c_n, p_n, \sigma^2) = P(N(0,1) > z_{1-\alpha}) = \alpha$, while if $c_n = c p_n^{1/4}$, then $\lim \inf_{n \to \infty} \rho_n(c_n, p_n, \sigma^2) = P \left( N(0,1) > z_{1-\alpha} - c^2/(\sqrt{2} \sigma^2) \right) \in (\alpha, 1)$, and the claims follow.  

\[\square\]
Lemma 3. Consider the sequence of models \( \mathcal{M}_n : y_n \sim N_n(X_n\beta_n, \sigma^2 I_n) \) where \( \beta_n \in \mathbb{R}^p \).

Define \( \rho_n(\mu_n, v_n, p_n, \sigma^2) \) to be the power of the level-\( \alpha < 1/2 \) cone test of the null hypothesis \( H_{0,n} : \beta_n = 0 \) with rejection region \( \{ y : \langle y/||y||, \mu_n \rangle > q_{n,1-\alpha} \} \), where \( q_{n,1-\alpha} \) is an appropriate level-\( \alpha \) quantile, \( \mu_n \in S^{n-1} \), and \( v_n = X_n\beta_n \) under the alternative hypothesis. If \( ||v_n||^2 = c_0 \) is constant and the mean direction of the cone test is nearly correctly specified so that \( \liminf_{n \to \infty} \rho_n(\mu_n, v_n, p_n, \sigma^2, \alpha) \in (\alpha, 1) \), where the power function does not depend on \( p_n \). If \( ||v_n|| = n^{1/4} \) and if \( ||\mu_n - (v_n/||v_n||)|| = (n^{-1/4} - an^{-1/2}) = O(n^{-1/4}) \) for some \( a > 0 \), then \( \liminf_{n \to \infty} \rho_n(\mu_n, v_n, p_n, \sigma^2, \alpha) \in (\alpha, 1) \) and if \( ||\mu_n - (v_n/||v_n||)|| = o(n^{-1/4}) \) then \( \liminf_{n \to \infty} \rho_n(\mu_n, v_n, p_n, \sigma^2, \alpha) = 1 \).

**Proof.** Define \( c_n = ||X_n\beta_n|| \) and, without loss of generality, we assume that \( X_n\beta_n = c_ne_1 \in \mathbb{R}^n \), where \( e_1 \) is the first standard basis vector. Under the alternative hypothesis, the cone test statistic has the stochastic representation

\[
\frac{y_n}{||y_n||} \cdot \mu_n = \frac{z_n + c_ne_1}{\sqrt{||z_n||^2 + 2c_n z_{n,1} + c_n^2}} \cdot e_1 + (\mu_n - e_1)
\]

where \( z_n \sim N_n(0, \sigma^2 I_n) \), and \( z_{n,j} \) is the \( j \)th component of \( z_n \). Thus

\[
\rho_n(\mu_n, c_n e_1, p_n, \sigma^2) \geq P \left( \frac{z_{n,1} + c_n}{\sqrt{||z_n||^2 + 2c_n z_{n,1} + c_n^2}} > q_{n,1-\alpha} + ||\mu_n - e_1|| \right).
\]

By the law of large numbers and Slutsky’s theorem, \( \sqrt{n}z_{n,1}/||z_n|| \overset{d}{\to} N(0, 1) \), and thus \( \sqrt{n}q_{n,1-\alpha} \to z_{1-\alpha} \), where \( z_{1-\alpha} \) is a standard normal quantile. Then assuming that \( ||\mu_n - e_1|| = o(n^{-1/2}) \) and \( c_n = c_0 \) is constant

\[
\liminf_{n \to \infty} \rho_n(\mu_n, c_n e_1, p_n, \sigma^2)
\]

\[
\geq \liminf_{n \to \infty} P \left( \frac{z_{n,1} + c_0}{\sqrt{\frac{1}{n}||z_n||^2 + \frac{2c_0}{n} z_{n,1} + \frac{c_0^2}{n}}} > \sqrt{n}q_{n,1-\alpha} + \sqrt{n}||\mu_n - e_1|| \right)
\]

\[
= P \left( z_{n,1} + c_0 > z_{1-\alpha} \right) > \alpha.
\]

Next, assume that \( c_n = n^{1/4} \) and \( ||\mu_n - e_1|| = (n^{-1/4} - an^{-1/2}) \), so that

\[
\liminf_{n \to \infty} \rho_n(\mu_n, c_n e_1, p_n, \sigma^2)
\]

\[
\geq \liminf_{n \to \infty} P \left( \frac{z_{n,1} + c_n}{\sqrt{\frac{1}{n}||z_n||^2 + \frac{2}{n^{3/2}} z_{n,1} + \frac{1}{\sqrt{n}}}} > \sqrt{n}q_{n,1-\alpha} + \sqrt{n}||\mu_n - e_1|| \right)
\]
Lemma 4. Under the prior distribution $\beta \sim N_p(\beta_0, \gamma(X^T X)^{-1})$, the FAB test is asymptotically equivalent to the $F$-test as $w \to 1$. That is, the probability that the $F$- and FAB tests lead to the same conclusion under the null hypothesis or any alternative hypothesis tends to 1 as $w \to 1$.

Proof. Let $R_{a,F} = \{y : F(y/\|y\|) > c_{a,F}\}$ and $R_{a,w} = \{y : T_{FAB,w}(y/\|y\|) > c_{a,w}\}$ be the rejection regions for the $F$- and FAB tests, respectively ($\alpha < 0.5$). It suffices to show that $\bigcap_{k>0} \bigcup_{y \in R_{a,F}} R_{a,F} \Delta R_{a,w}$ has Lebesgue measure 0, where $A \Delta B$ is the symmetric difference of sets.

As $w \to 1$, if $\|(I-P)u\|^2 > 0$, then

$$T_{FAB,w}(u) \to n \log(\sigma_0) - \frac{n}{2} \log(\|(I-P)u\|^2) + \log \left( \int_0^\infty \exp(-z^2)dz \right),$$

(A1)

where an appeal to the dominated convergence theorem with dominating function $z^{n-1}\exp(-z^2/2 + z[\bar{u}^T \bar{X} \beta_0])$ is used to take the limit $w \to 1$ inside the integral in (6). The statistic on the right-hand side of (A1) is denoted by $T_{FAB,1}(u)$. If $z \sim \text{Uniform}(\mathbb{S}^{n-1})$, then $T_{FAB,w}(z) \xrightarrow{a.s.} T_{FAB,1}(z)$ for any sequence $w_n \to w$ since $P(\|(I-P)z\|^2 = 0) = 0$. If $c_{a,1}$ is the $(1-\alpha)$-quantile of $T_{FAB,1}(z)$, then it is claimed that for any $w_n \to 1$, $c_{a,w_n} \to c_{a,1}$. If these were not the case, then there would exist a sequence $w_n \to 1$ with $|c_{a,w_n} - c_{a,1}| > \epsilon$ for all $n$. However, this is not possible since $T_{FAB,w}(z) \xrightarrow{P} T_{FAB,1}(z)$ and $T_{FAB,\infty}(z)$ has a continuous distribution with positive density on the interior of its support, implying that its quantiles are unique.

We have shown above that if $w_n \to 1$, then $\lim_n T_{FAB,w}(u) = c_{a,w_n} = T_{FAB,1}(u) - c_{FAB,1}$ for all $u$ with $\|(I-P)u\|^2 > 0$. As $T_{FAB,1}(u) - c_{a,1} > 0$ if and only if $F(u) - c_{a,F} > 0$, all $y \in \bigcap_{k>0} \bigcup_{y \in R_{a,F}} R_{a,F} \Delta R_{a,w}$ must satisfy $\|(I-P)y\|^2 = 0$. As this set has Lebesgue measure 0 (assuming $P \neq I$), this completes the proof.

Lemma 5. Let $\rho(\beta_m, \Psi_m)$ denote the marginal power of the FAB test, where the test statistic is constructed using the prior distribution $\bar{\beta} \sim N_p(\beta_m, \Psi_m)$ and the power is evaluated under the marginal distribution $\bar{\beta} \sim N_p(\beta_0, \Psi_0)$. If $(\beta_m, \Psi_m) \to (\beta_0, \Psi_0)$, then $\rho(\beta_m, \Psi_m) \to \rho(\beta_0, \Psi_0)$. If $(\beta_m, \Psi_m)$ converges to $(\beta_0, \Psi_0)$ in probability, then $E(\rho(\beta_m, \Psi_m)) \to \rho(\beta_0, \Psi_0)$. In particular, if $(\beta_m, \Psi_m)$ are consistent estimates of the prior parameters $(\beta_0, \Psi_0)$ in a multigroup FAB test with $m$ groups, then asymptotically as the number of groups grows, the marginal power of the multigroup FAB test converges to the marginal power of the Bayes-optimal test $\rho(\beta_0, \Psi_0)$.

Proof. Both $x = \sqrt{u(\Psi \Psi^T + \sigma_0^2 I)^{-1}u}$ and $r = u(\Psi \Psi^T + \sigma_0^2 I)^{-1}X \beta / x$ are continuous functions of $(u, \beta, \Psi)$. When $r \leq M$ and $z \geq 0$, the function $z^{n-1}e^{-(z-r)^2/2}$ is uniformly bounded above by the integrable function $z^{n-1}e^{-z^2/2+zM}$ for every $z \geq 0$. By the dominated convergence theorem, it follows that $I_n(r) = \int_0^\infty z^{n-1}e^{-(z-r)^2/2}dz$ is a continuous function of $r$ for $r \in (0, M)$, and as $M$ is arbitrary, $I_n(r)$ is continuous for $r \in (0, \infty)$. The FAB test statistic $r^2/2 + I_n(r) - n \log(x)$ is therefore a continuous function of $(u, \beta, \Psi)$.

Denote the FAB test statistic constructed using the prior distribution $\beta \sim N_p(\beta_m, \Psi_m)$ by $T_m(u)$ for $m = 0, 1, \ldots$, and let $c_m$ be the $1-\alpha$ quantile of $T_m(u)$ when $u$ is uniformly distributed.
over $\mathbb{S}^{n-1}$. By continuity, $T_m(u) \to T_0(u)$ almost surely. As the statistic $T_0(u)$ has a continuous distribution function and $T_m(u)$ converges to $T_0(u)$ in probability, the quantiles $c_m$ must converge to $c_0$ as $m \to \infty$. If $P_0$ is the marginal distribution of $u$ under the prior distribution $\beta \sim N_p(\beta_0, \Psi_0)$, then

$$\rho(\beta_m, \Psi_m) = P_0(T_m(u) > c_m) \to P_0(T_0(u) > c_0) = \rho(\beta_0, \Psi_0)$$

as $T_m(u) - c_m$ converges to $T_0(u) - c_0$ in probability and $T_0(u) - c_0$ has a continuous distribution function.

The above reasoning demonstrates that the power function is continuous in $(\beta, \Psi)$. Therefore, $\rho(\hat{\beta}_m, \hat{\Psi}_m)$ converges in probability to $\rho(\beta_0, \Psi_0)$. The random variables $\rho(\hat{\beta}_m, \hat{\Psi}_m)$ are bounded in the interval $[0, 1]$ and thus are uniformly integrable, implying that $E(\rho(\hat{\beta}_m, \hat{\Psi}_m)) \to \rho(\beta_0, \Psi_0)$. The marginal power of the multigroup FAB test is given by $E(\rho(\hat{\beta}_m, \hat{\Psi}_m))$.

Tables and Figures

| Table A1: Proportion of hypotheses rejected by the Benjamini–Hochberg procedure across either all schools or the schools with full-rank design matrices. |
|---------------------------------------------------------------|
| $\alpha$ | FAB-HS, All | FA-HS, Full-rank | $F$-test, All | $F$-test, Full-rank |
|----------|-------------|-----------------|-------------|-----------------|
| $\alpha = 0.1$ | 0.01 | 0.11 | 0.00 | 0.01 |
| $\alpha = 0.2$ | 0.09 | 0.24 | 0.00 | 0.02 |
| $\alpha = 0.5$ | 0.52 | 0.68 | 0.12 | 0.28 |

**Figure A1:** Normal $Q - \bar{Q}$ plots for nine different schools.
FIGURE A2: Comparison of the FAB-HS $P$-values when the prior parameter estimates of $\beta_0, \Psi, \sigma_0^2$ are either recomputed for each school or common estimates are used across all schools.

Received 25 March 2022
Accepted 19 October 2022