Statistical methods for microarray data

Goeman, Jelle Jurjen

CHAPTER 5

Testing against a high-dimensional alternative

Abstract

As the dimensionality of the alternative increases, the power of classical tests tends to diminish quite rapidly. This is especially true for high-dimensional data in which there are more parameters than observations. In this paper we discuss a score test on a hyperparameter in an empirical Bayesian model as an alternative to classical tests. It gives a general test statistic which can be used to test a point null hypothesis against a high-dimensional alternative, even when the number of parameters exceeds the number of samples. This test will be shown to have optimal power on average in a neighbourhood of the null, which makes it a proper generalization of the locally most powerful test to multiple dimensions. To illustrate this new locally most powerful test we investigate the case of testing the global null hypothesis in a linear regression model in more detail. The score test is shown to have significantly more power than the F-test whenever under the alternative the large-variance principal components of the design matrix explain substantially more of the variance of the outcome than the low-variance principal components. The score test is also useful for detecting sparse alternatives in truly high-dimensional data, where its power is comparable to the test based on the maximum absolute t-statistic.

5.1 Introduction

In a linear regression model one traditionally uses the F-test to test the global null hypothesis that all regression coefficients are zero. However, it is well known that the F-test has low power when the number of covariates in the

This chapter will appear as: J. J. Goeman, S. A. van de Geer and J. C. van Houwelingen (2006) Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society, Series B* **68**, in press. The definitive version will be available at `http://www.blackwell-synergy.com`.

model is close to the number of samples. The F-test even breaks down completely when the number of covariates exceeds the number of samples. Similar behaviour is known for the likelihood ratio test in generalized linear models. In general, classical tests tend to perform badly when used against high dimensional alternatives.

This paper explores testing of a simple null hypothesis against a high-dimensional alternative. We shall formulate a simple test which can be used in high-dimensional models regardless of the number of parameters. This test is constructed as a locally most powerful test (score test) on the hyperparameter in an empirical Bayesian model. The same type of test has been introduced for specific models in the context of microarray gene expression data, where it is used to generalize a test for association between a clinical variable and a single gene to a test for association between a clinical variable and a group of genes. Goeman et al. (2004) have applied this methodology in generalized linear models with a canonical link function and Goeman et al. (2005) in the Cox proportional hazards model. For examples of real data applications we refer to these papers.

In the present paper we explore the general power properties of this type of test in more detail, adopting a purely frequentist point of view. The test will be shown to have optimal average power in a neighbourhood of the null hypothesis, a property which follows as a corollary to the Neyman-Pearson lemma. This property makes the test a natural generalization of the locally most powerful test to higher dimensions, and motivates us to refer this high-dimensional version of the locally most powerful test simply as the locally most powerful test.

We shall also look more closely into the relatively simple case of a high-dimensional alternative in a linear model. In this model there are few distracting details and many quantities can be explicitly calculated. We investigate the regions of the parameter space where the empirical Bayes score test has most and least power and situations where we may expect good power.

In the linear model it is also relatively easy to investigate links with other tests, most notably the F-test. It turns out that the F-test can be formulated as an empirical Bayesian score test with a different prior distribution, a fact which gives insight into the power properties of the F-test. We also investigate relationships between our empirical Bayes procedure with principal components tests and with a typical multiple testing procedure from microarray data analysis which uses the maximum of all absolute univariate T-statistics as a test statistic for the global null. All these comparisons will be illustrated with simulations based on real microarray data (taken from Van de Vijver et al., 2002).

5.2 Empirical Bayes testing

Suppose we have observations $\mathbf{y}$ (typically an n-vector), the distribution of which is assumed to depend on a p-vector of parameters $\boldsymbol{\beta}$. In this model we want to test

$$H_0 : \boldsymbol{\beta} = \mathbf{0}$$

against $H_A : \boldsymbol{\beta} \neq \mathbf{0}$. There may also be some nuisance parameters, but we assume them known for the moment.

If the dimension p of the alternative is large, the alternatives can range over a huge space and H_A typically allows many widely different distributions of $\mathbf{y}$. Some of the alternatives may even induce the same distribution of $\mathbf{y}$ as H_0, especially if $p > n$. In a generalized linear model, for example, the distribution of $\mathbf{y}$ depends on $\boldsymbol{\beta}$ only through $X\boldsymbol{\beta}$, where X is an $n \times p$ design matrix. If $p > n$, there are many alternatives which have $\boldsymbol{\beta} \neq \mathbf{0}$ but $X\boldsymbol{\beta} = \mathbf{0}$. These alternatives give rise to the same distribution of $\mathbf{y}$ as the null hypothesis, which means we can never hope to have any power against these alternatives. This is typical for high-dimensional alternatives: a minimax type approach which tries to have power against all alternatives is bound to fail.

Therefore it seems a sensible approach to focus the power of the test on what we choose to be the most interesting alternatives. This can be done in a Bayesian fashion by assigning the vector $\boldsymbol{\beta}$ a distribution. This distribution should give most probability mass to the alternatives which are perceived as more likely (as in a prior distribution) or simply as more 'interesting' to detect.

What this distribution should be depends very much on the model and the purpose of the test. However, a good choice for such a distribution is usually one that is 'unbiased', i.e. it is symmetric around the null hypothesis and therefore has $\mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$. This is sensible, because we are usually equally interested in detecting the alternative that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ as in detecting $\boldsymbol{\beta} = -\boldsymbol{\beta}_0$ for every $\boldsymbol{\beta}_0$. The covariance matrix of $\boldsymbol{\beta}$ may then be chosen in general as $\mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}') = \tau^2\Sigma$ for some well-chosen positive (semi-)definite $p \times p$ matrix Σ. The choice $\Sigma = I$ deserves special attention, because it follows from an exchangeability assumption: the density of all permutations of the vector $\boldsymbol{\beta}$ is equal (Bernardo and Smith, 1994, p. 180). Under this exchangeable assumption one is not prejudiced as to which elements of $\boldsymbol{\beta}$ are expected to be large or which elements of $\boldsymbol{\beta}$ are expected to be similar. This assumption is useful when there is no structure or ordering in the parameters that can be readily exploited and when the typical range of the parameter values is similar.

One can complete the specification of the distribution of $\boldsymbol{\beta}$ by choosing a value τ_0^2 for τ^2 and a distributional shape. In the generalized linear model setting, taking the maximum likelihood estimate of $\boldsymbol{\beta}$ will then result in one of

many familiar penalized regression methods, depending on the choice of the distribution of $\boldsymbol{\beta}$. Choosing $\boldsymbol{\beta}$ to have i.i.d. normal entries results in a (generalized) ridge regression (Hoerl and Kennard, 1970). Choosing the regression coefficients $\boldsymbol{\beta}$ i.i.d. double exponential results in the LASSO method (Tibshirani, 1996). These methods are frequently used in estimation and prediction problems in high-dimensional regression models.

We can also use the chosen distribution of $\boldsymbol{\beta}$ as a tool to rephrase our testing problem, rewriting it in terms of the marginal distribution of $\mathbf{y}$. Let $f(\boldsymbol{\beta}; \mathbf{y})$ be the likelihood of $\boldsymbol{\beta}$ for given $\mathbf{y}$. The marginal density of $\mathbf{y}$ is

$$\bar{f}(\tau^2; \mathbf{y}) = \mathrm{E}_{\boldsymbol{\beta}|\tau^2}[f(\boldsymbol{\beta}; \mathbf{y})],$$

which can be interpreted as the likelihood of τ^2 in a new marginal model of $\mathbf{y}$. In this new model, rejecting the new null hypothesis $\bar{H}_0 : \tau^2 = 0$ implies rejecting the old $H_0 : \boldsymbol{\beta} = \mathbf{0}$, as the two imply the same distribution of $\mathbf{y}$.

The testing procedure based on testing $\bar{H}_0 : \tau^2 = 0$ against $\bar{H}_A : \tau^2 = \tau_1^2$ can be called "empirical Bayes testing", because we have put a prior on the parameter vector $\boldsymbol{\beta}$ of the model, which depends on an unknown hyperparameter τ^2, and our inference on $\boldsymbol{\beta}$ proceeds through inference on τ^2. On the other hand it can also simply be called "Bayesian testing", because once the shape of the distribution and the value of τ_1^2 are chosen, the model H_A is fully Bayesian.

One important use of testing $\bar{H}_0$ in the marginal model of $\mathbf{y}$ lies in Lemma 1, a corollary to the Neyman-Pearson Lemma. It says that if we take a specific distribution of $\boldsymbol{\beta}$ and construct a likelihood ratio test in the marginal model, the resulting test has optimal power on average over the chosen distribution of alternatives.

Lemma 1 (Empirical Bayes version of Neyman-Pearson) *Let A_1 be the critical region of a likelihood ratio test of $\bar{H}_0 : \tau^2 = 0$ against $\bar{H}_A : \tau^2 = \tau_1^2$ in the marginal model $\bar{f}$, with associated power function $\bar{w}_{\tau_1^2}(\boldsymbol{\beta}) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A_1]$; and let A be the critical region of any test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$, with power function $w(\boldsymbol{\beta}) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A]$. Then*

$$w(\mathbf{0}) \leq w_{\tau_1^2}(\mathbf{0})$$

implies

$$\mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}[w(\boldsymbol{\beta})] \leq \mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}[w_{\tau_1^2}(\boldsymbol{\beta})].$$

This is a well-known result. The proof is immediate from the Neyman-Pearson Lemma when it is observed that $\mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}[w(\boldsymbol{\beta})] = \mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}\{\mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A]\} = \mathrm{P}_{\mathbf{y}|\tau_1^2}[A]$.

The result of Lemma 1 could immediately be used in practice if we were willing to completely specify the distribution of $\boldsymbol{\beta}$, or at least to specify the

shape of the distribution up to a number of parameters which can be estimated. In most cases, however, we should be reluctant to do this, for two reasons. In the first place, the marginal likelihood is a complicated p-dimensional integral, which often makes it difficult to estimate hyperparameters and usually almost impossible to find the distribution of the test statistic, except in very special cases. Secondly, specifying the distributional shape of $\boldsymbol{\beta}$ means specifying whether the interesting alternatives have a $\boldsymbol{\beta}$ with a few large entries or many small ones. This is a kind of judgement which is typically very difficult to make in high-dimensional data. In a high-dimensional regression model, for example, it is usually not known whether there are few large or many small regression coefficients. A wrong choice of the distribution of $\boldsymbol{\beta}$ could mean low power. How can we avoid specifying the distributional shape of $\boldsymbol{\beta}$?

5.3 The locally most powerful test

It turns out that we can design a test for $\bar{H}_0$ in the marginal model which manages to avoid full specification of the distribution of $\boldsymbol{\beta}$ and avoids evaluation of the complicated marginal likelihood as well. This can be done by constructing the test as a score test.

The traditional score test is a one-sided test of $H_0^* : \theta = \theta_0$ against $H_A^* : \theta > \theta_0$ in a one parameter model with likelihood $f^*(\theta; \mathbf{y})$. It rejects when the score test statistic $S^*(\mathbf{y}) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f^*(\theta_0; \mathbf{y}) \geq k$ for some constant k. If θ_0 is on the edge of the parameter space, $S^*(\mathbf{y})$ should be taken as the right-sided derivative. For typical values of the test size α the critical value k is almost invariably positive, because, by the properties of the score function, $S^*(\mathbf{y})$ has zero expectation under the null hypothesis.

The score test is known as the "locally most powerful test" as a consequence of Lemma 2. This lemma says that the score test has optimal slope of the power function among all tests of at most the same size, so that it has optimal power against local alternatives close to the null.

Lemma 2 (Score test property) *Suppose that the derivative $\frac{\mathrm{d}}{\mathrm{d}\theta} f^*(\theta; \mathbf{y})$ exists a.e. and is bounded in a (right-)neighbourhood of θ_0. Then for any test of H_0^* with critical region A and power function $w(\theta) = \mathrm{P}_{\mathbf{y}|\theta}[A]$, the derivative $\frac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0)$ exists. Moreover, if $w^*(\theta) = \mathrm{P}_{\mathbf{y}|\theta}[S^* \geq k]$ is the power function of the score test, then either of*

(i) $w(\theta_0) = w^*(\theta_0)$

(ii) $w(\theta_0) \leq w^*(\theta_0)$ *and* $k \geq 0$

implies

$$\frac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0) \leq \frac{\mathrm{d}}{\mathrm{d}\theta} w^*(\theta_0).$$

The proof of this lemma is given in Section 5.12.

A more extensive treatment of locally most powerful tests in one dimension is given in Cox and Hinkley (1974). They show that the score test can be interpreted as the limit for $\theta_1 \downarrow \theta_0$ of the likelihood ratio test of H_0^* against the point alternative $H_1^* : \theta = \theta_1$. Score tests are typically useful when testing an 'easy' null hypothesis against a 'complicated' alternative, because score testing does not require estimation of θ. Our high-dimensional alternative is a good example of such a complicated alternative.

We shall apply score testing in the empirical Bayesian setting by testing $\bar{H}_0 : \tau^2 = 0$ against $\bar{H}_A : \tau^2 > 0$ in the marginal model using the score test statistic

$$S = \frac{\mathrm{d}}{\mathrm{d}\tau^2} \log \bar{f}(0; \mathbf{y}),$$

which is automatically a right-sided derivative as $\bar{f}$ is only defined for $\tau^2 \geq 0$. This test has two very useful properties, which we have formulated as Lemma 3 and Lemma 4.

The first property is important both for computation and for modelling. Lemma 3 says that the test statistic S can be found with simple matrix operations from the conditional likelihood $f(\boldsymbol{\beta}; \mathbf{y})$ and the covariance matrix of $\boldsymbol{\beta}$. This implies that we do not need numerical integration to find the value of the test statistic and that we do not have to specify the distributional shape of the distribution of $\boldsymbol{\beta}$.

Lemma 3 (Score test statistic) *Suppose $\boldsymbol{\beta} = \tau \mathbf{b}$, where $\mathrm{E}\mathbf{b} = \mathbf{0}$ and $\mathrm{E}(\mathbf{b}\mathbf{b}') = \Sigma$ and the distribution of $\mathbf{b}$ does not depend on τ. Suppose also that loglikelihood $\log f(\boldsymbol{\beta}; \mathbf{y})$ and its first two derivatives exist a.e. and are bounded in a neighbourhood of $\boldsymbol{\beta} = \mathbf{0}$. Then the empirical Bayes score test statistic $S = \frac{\mathrm{d}}{\mathrm{d}\tau^2} \log \bar{f}(0; \mathbf{y})$ exists and is given by*

$$S = \frac{1}{2}\mathbf{s}'\Sigma\mathbf{s} - \frac{1}{2}\mathrm{trace}[\Sigma\mathbf{I}]$$

where $\mathbf{s} = \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{0}; \mathbf{y})$ is the score function and $\mathbf{I} = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log f(\mathbf{0}; \mathbf{y})$ the observed Fisher information of $\boldsymbol{\beta}$ in H_0.

The proof of this lemma is a simple calculation, which is given in Section 5.12.

The second and most important property of the score test based on S is given in Lemma 4. It is again an optimality property, which effectively combines the statements of Lemmas 1 and 2. Lemma 4 says that the empirical Bayes score test, which has optimal slope of the power function in the marginal model $\bar{f}$, has optimal expected slope of the power function in the conditional model f.

This lemma only holds for the exchangeable version of the test with $\Sigma = I$, although a more general version can also be formulated.

Lemma 4 (Locally Optimal Power) *Suppose the conditions of Lemma 3 hold with $\Sigma = I$. Let $\bar{w}(\boldsymbol{\beta}) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[S \geq k]$ be the power function of the exchangeable score test of H_0. Let $w(\theta) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A]$ be the power function of any test of H_0. Then either of*

(i) $w(\mathbf{0}) = \bar{w}(\mathbf{0})$

(ii) $w(\mathbf{0}) \leq \bar{w}(\mathbf{0})$ *and* $k \geq 0$

implies

$$\mathrm{E}_{\boldsymbol{\xi}}[\tfrac{\mathrm{d}}{\mathrm{d}\tau^2} w_{\boldsymbol{\xi}}(0)] \leq \mathrm{E}_{\boldsymbol{\xi}}[\tfrac{\mathrm{d}}{\mathrm{d}\tau^2} \bar{w}_{\boldsymbol{\xi}}(0)]$$

where $w_{\boldsymbol{\xi}}(\tau) = w(\tau\boldsymbol{\xi})$, $\bar{w}_{\boldsymbol{\xi}}(\tau) = \bar{w}(\tau\boldsymbol{\xi})$ and $\boldsymbol{\xi}$ has a uniform distribution on the unit p-ball ($p = \dim(\boldsymbol{\beta})$). The same result holds when $\boldsymbol{\xi}$ has any other distribution on the unit p-ball such that $\mathrm{E}(\boldsymbol{\xi}) = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{\xi}\boldsymbol{\xi}') \propto I$.

The proof of the lemma is given in Section 5.12. In fact, Lemma 4 follows from Lemma 2 in more or less the same way as Lemma 1 follows from the Neyman-Pearson Lemma.

By Lemma 4 we see that the score test in the exchangeable empirical Bayesian model has optimal expected slope of the power function, where the expectation is with respect to taking a random direction in p-space. This is the property that motivates its name of locally most powerful test. It is an interesting sidenote that even if $p = 1$, by Lemma 3 the high-dimensional score test based on S is not the same as the ordinary one-dimensional score test based on S^*, because the test based on S is a two-sided test, whereas the test based on S^* is one-sided. By Lemmas 3 and 4 the test based on S is the proper generalization of the one-dimensional score test from one-sided to two-sided alternatives.

5.4 Nuisance parameters

The presence of nuisance parameters complicates some of the issues described above. When nuisance parameters are present, the null hypothesis is not simple anymore but composite. In that case strict optimality in the sense of Lemma 4 is impossible.

The issue of nuisance parameters is usually tackled by switching to the profile likelihood (Pawitan, 2001). When using a score test, switching to the profile likelihood is very easy: one can simply plug in the maximum likelihood estimate of the nuisance parameter under the null hypothesis. This can be easily

seen in a simple two parameter model with loglikelihood $g(\theta, \eta)$ and profile likelihood $\hat{g}(\theta) = g\{\theta, \hat{\eta}(\theta)\}$. In this situation

$$\frac{\partial \hat{g}}{\partial \theta} = \frac{\partial g}{\partial \theta} + \frac{\partial g}{\partial \eta}\frac{\partial \eta}{\partial \theta}. \tag{5.1}$$

The second term on the right hand side is zero, because $\partial g/\partial \eta$ is always zero in $\hat{\eta}$.

This simple plugging in of the null estimate of the nuisance parameters can also be understood by viewing the score test again as a (profile) likelihood ratio test of $\theta = \theta_0$ versus $\theta = \theta_1$ for $\theta_1 \downarrow \theta_0$. In the limit the maximum likelihood estimate of η is the same under the alternative as under the null.

In the empirical Bayes model of this paper the situation is basically the same. A similar argument to (5.1) can be used to check in the proof of Lemma 3 that plugging in the estimate under the null is equivalent to using the profile likelihood. For this derivation it makes no difference whether one uses the conditional profile likelihood, starting with likelihood f and the maximum likelihood estimate $\hat{\eta}(\boldsymbol{\beta}; \mathbf{y})$ of the nuisance parameter η as a function of $\boldsymbol{\beta}$, or whether one uses the marginal likelihood $\bar{f}$ and the maximum likelihood estimate $\bar{\eta}(\tau^2; \mathbf{y})$ from the marginal model for given τ^2. Both profile likelihoods lead to the same test.

See section 5.6 for an example of a model with nuisance parameters.

5.5 Distribution of the test statistic

The specification of the locally most powerful test in the previous sections is not fully complete, as it only provides us with the test statistic to be used. To be able to use the test in practice, we must also know the distribution of the test statistic under the null, so as to be able to find the cutoff for significance and/or the p-value. There is no general method for finding the null distribution, and this may require some extra work when the concept of the locally most powerful test is to be applied in the context of a specific model. We only give some general comments here. See section 5.6 and Goeman et al. (2004) and Goeman et al. (2005) for concrete examples.

First, we look at the null distribution of S. It should be noted that, aside from having zero expectation under the null, the test statistic S is not yet standardized and, in general, should not be expected to follow any standard textbook distribution. It is usually not easy to directly apply asymptotic results on the distribution of the score statistic, because the marginal likelihood $\bar{f}$, from which the score statistic was derived, is not generally a product of n contributions of the individuals. Asymptotic arguments may be used in specific models (as in Goeman et al., 2005), but we have no general theory yet.

In many cases, however, one can find a reasonably good approximation to the distribution of S because the expression for S, as given in Lemma 3, is relatively easy. The mean of S and its variance can often be explicitly calculated. This allows approximation of the null distribution by moment matching to a tabulated distribution (this strategy was used in Goeman et al., 2004). Other practical options for finding the distribution of S include numerical integration or permutation methods. Exact calculation of the distribution function of S is possible in special cases, such as testing the global null hypothesis in the linear model with normal errors, which is the case we shall turn to now.

5.6 The linear model

The optimality property implied in Lemma 4 is very appealing, but it has its limitations. Good power is guaranteed, but only locally near the null and on average over many possible alternatives. To investigate more closely what Lemma 4 is worth for specific alternatives, we shall examine the simplest case of the linear model in detail.

Assume that $\mathbf{y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I)$, where X is an $n \times p$ design matrix of full rank $\min(n, p)$. For simplicity we ignore the intercept parameter α which would normally be included (See Goeman et al., 2004, on how to deal with the nuisance parameter α). The score vector for this model is $\mathbf{s} = \sigma^{-2} X'\mathbf{y}$ and the observed Fisher information is $\mathbf{I} = \sigma^{-2} X'X$, so the general empirical Bayes score test statistic is

$$\tilde{S}_\Sigma = \tfrac{1}{2\sigma^4} \mathbf{y}' X\Sigma X'\mathbf{y} - \tfrac{1}{2\sigma^2} \operatorname{trace}(X\Sigma X').$$

It is more convenient to work with the equivalent test statistic $\sigma^{-2}\mathbf{y}' X\Sigma X'\mathbf{y}$, whose distribution does not depend on σ^2. Because σ^2 is not known, we plug in its maximum likelihood estimate $\hat{\sigma}_0^2 \propto \mathbf{y}'\mathbf{y}$ under the null hypothesis. The resulting test statistic is

$$S_\Sigma = \frac{\mathbf{y}' X\Sigma X'\mathbf{y}}{\mathbf{y}'\mathbf{y}}, \tag{5.2}$$

whose distribution also does not depend on the nuisance parameter σ^2. We study the exchangeable case $\Sigma = I$, as 'the' locally most powerful test statistic

$$S = \frac{\mathbf{y}' XX'\mathbf{y}}{\mathbf{y}'\mathbf{y}}.$$

To find the distribution function of S, we can use the following identity (Azzalini and Bowman, 1993):

$$\mathrm{P}\{S > t\} = \mathrm{P}\{\mathbf{y}'(XX' - tI)\mathbf{y} > 0\}.$$

The distribution function of the quotient S can therefore be found through the distribution function of a quadratic form in normal variables. We use numeric methods developed by Imhof (1961) to calculate the latter distribution function. Reasonably good approximations to the 5% and 1% cutoff values can also be found by equating the moments of S to those of a gamma distribution, a strategy which was used in Goeman et al. (2004).

It is interesting to note a connection between the test statistic S and the method of partial least squares (PLS), which is often used for high-dimensional data in chemometrics (Brown, 1993). The first component of a partial least squares regression is $XX'\mathbf{y}$, so the test statistic S can be viewed as a test for correlation between the first PLS component and $\mathbf{y}$.

5.7 Power of the score test

We want to gain insight in the power of the locally most powerful test in practice. It has already been said that when the alternatives are high-dimensional, it is impossible to have power against all alternatives. To see which are the alternatives that our score test cannot detect, we check which alternatives have an expected test statistic that is smaller than expected under the null. These alternatives have power below the size α of the test.

Under the null hypothesis, the test statistic S has expectation

$$\mathrm{E}_{\mathbf{y}|\mathbf{0}}[S] = \frac{1}{n}\mathrm{trace}(XX').$$

Under the alternative the expectation of S can be well approximated by taking the expectations of the numerator and the denominator separately

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] \approx \frac{\boldsymbol{\beta}'X'XX'X\boldsymbol{\beta} + \sigma^2\mathrm{trace}(XX')}{\boldsymbol{\beta}'X'X\boldsymbol{\beta} + n\sigma^2}.$$

This approximation is not only asymptotically exact, but also for small sample size if $\mathbf{y}$ is either dominated by $X\boldsymbol{\beta}$ or by σ^2 (i.e. in any of the limits $n \to \infty$, $\sigma^2 \to 0$, $\sigma^2 \to \infty$ or $\boldsymbol{\beta} \to \mathbf{0}$).

The difference between the expectations is

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] - \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S] \approx \frac{\boldsymbol{\beta}'X'XX'X\boldsymbol{\beta} - \frac{1}{n}\boldsymbol{\beta}'X'X\boldsymbol{\beta}\cdot\mathrm{trace}(XX')}{\boldsymbol{\beta}'X'X\boldsymbol{\beta} + n\sigma^2}.$$

To interpret this expression we must look at the principal components of X and the amount of variance of $\mathbf{y}$ that each principal component explains. Call

$$r^2 = \frac{\boldsymbol{\beta}'X'X\boldsymbol{\beta}}{\boldsymbol{\beta}'X'X\boldsymbol{\beta} + n\sigma^2},$$

the fraction of the variance of $\mathbf{y}$ explained by the alternative. We use the spectral decomposition. Write $X'X = \sum_{i=1}^{n} \lambda_i Q_i$, where $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ are eigenvalues of $X'X$ and Q_i is the $p \times p$ projection matrix that projects onto the eigenvector of $X'X$ corresponding to the eigenvalue λ_i. Note that we can stop the decomposition at the n-th component because the rank of $X'X$ is $\min(n, p) \leq n$. Use of the spectral decomposition gives $r^2 = \sum_{i=1}^{n} r_i^2$, with $r_i^2 = \lambda_i \boldsymbol{\beta}' Q_i \boldsymbol{\beta} / (\boldsymbol{\beta}' X'X \boldsymbol{\beta} + n\sigma^2)$, and

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] - \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S] = \sum_{i=1}^{n} \lambda_i r_i^2 - \frac{1}{n} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} r_j^2.$$

This can be recognized as proportional to the covariance of the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ of variances of the principal components of X and the vector $\mathbf{r} = (r_1^2, \ldots, r_n^2)'$, which gives the fraction of the variance of $\mathbf{y}$ explained by these components.

This small exercise has a few interesting conclusions. In the first place there are many alternatives, especially in the $p \geq n$ case, for which the locally most powerful test has negligible power. These are the alternatives for which the low-variance principal components of X explain most of the variance of $\mathbf{y}$. These undetectable alternatives may have any value of r^2, even $r^2 = 1$: an alternative with $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] \leq \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$ and $r^2 = 1$ will even have zero power.

Fortunately for the score test, a negative covariance of $\boldsymbol{\lambda}$ and $\mathbf{r}$ occurs only seldomly in real data, because the measurements in X are often noisy or inaccurate. The uninformative noise tends be dominant in the small-variance principal components of X.

How can a test be most powerful on average if it has such low power against many alternatives? The reason for this lies in the assumption of exchangeability that underlies the test. By Lemma 4 the power is optimal on a small p-ball with $\boldsymbol{\beta}'\boldsymbol{\beta} = c$. The alternatives on this ball have very diverse values of r^2: alternatives which have $\boldsymbol{\beta}$ in directions corresponding to the eigenvectors of the large eigenvalues of $X'X$ have large r^2, others have small r^2. It is very difficult to have much power against alternatives with small r^2. Even an 'oracle' which knows the direction of $\boldsymbol{\beta}$ and only tests whether $\|\boldsymbol{\beta}\| = 0$ will have low power if the true $\boldsymbol{\beta}$ has low r^2. Average power will increase, therefore, if some power on the low-potential alternatives is sacrificed in exchange for a gain in power for the high-potential alternatives. This is the advantageous trade-off that the exchangeable empirical Bayes score test makes.

If negative covariance of $\boldsymbol{\lambda}$ and $\mathbf{r}$ leads to $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] < \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$, conversely a positive covariance of the same $\boldsymbol{\lambda}$ and $\mathbf{r}$ leads to $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] > \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$ and potentially good power. Against some of these alternatives the score test must even have very good power, as the test is locally most powerful on average by Lemma 4.

We come back to this in Sections 5.8 and 5.10, where we compare the locally most powerful test with the F-test.

It has to be remarked that the problems of lower expectation of the test statistic S under the alternative than under the null typically disappear when n is large. If we let n grow to kn by observing k samples from each covariate pattern, $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S]$ will eventually become larger than $\mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$, because letting n grow in this setup means augmenting both $\boldsymbol{\lambda}$ and $\mathbf{r}$ with zeros, so that the correlation between the two increases. Similarly, if we have $p < n$ to begin with, there are at least $n - p$ zero elements of $\boldsymbol{\lambda}$ with corresponding zero elements of $\mathbf{r}$, so that the smallest elements of $\boldsymbol{\lambda}$ and $\mathbf{r}$ automatically coincide and there are few alternatives with $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] \leq \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$.

5.8 A new look at the F-test

In the $p < n$ situation it is possible to apply both the locally most powerful test and the F-test, which makes it interesting to compare the two. The F-test statistic in our linear model is a constant times

$$\tilde{F} = \frac{\mathbf{y}'X(X'X)^{-1}X'\mathbf{y}}{\mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y}}.$$

We find it convenient to transform $\tilde{F}$ by the strictly increasing function $g(x) = (x^{-1} + 1)^{-1}$ to the equivalent test statistic $F = g(\tilde{F})$, which is given by

$$F = \frac{\mathbf{y}'X(X'X)^{-1}X'\mathbf{y}}{\mathbf{y}'\mathbf{y}}.$$

Under the null the transformed F has a beta distribution with parameters $\frac{1}{2}p$ and $\frac{1}{2}(n-p)$.

It is now easy to compare F with the locally most powerful test statistic $S = (\mathbf{y}'XX'\mathbf{y})/(\mathbf{y}'\mathbf{y})$. We can immediately notice that, if the design is orthogonal (i.e. $X'X \propto I$) both tests are equivalent. Note that the design is always orthogonal if $p = 1$, so the locally most powerful test for $p = 1$ is equivalent to the F-test and hence to the two-sided t-test.

More fundamental insights follow when comparing F with the general expression for the empirical Bayesian score test statistic given in (5.2): as $S_\Sigma = \mathbf{y}'X\Sigma X'\mathbf{y}/(\mathbf{y}'\mathbf{y})$, we have $F = S_{(X'X)^{-1}}$. It follows that we can look at the F-test as the empirical Bayes score test based on the prior covariance $\mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}') = \tau^2(X'X)^{-1}$ for τ^2 very small. By Lemma 1, the F-test therefore optimizes the power on average over this distribution of $\boldsymbol{\beta}$. The F-test is therefore especially directed against alternatives in directions where the variance of the distribution of $\boldsymbol{\beta}$ is large. These directions are the directions of the eigenvectors

of small eigenvalues of $X'X$. These are also the directions where a large r^2 requires a very large $\|\boldsymbol{\beta}\|$. Vice versa, the directions of the eigenvectors of large eigenvalues of $X'X$ get a small prior variance of $\boldsymbol{\beta}$. These are, therefore, of small importance to the F-test: $\boldsymbol{\beta}$ is a priori not expected to lie in these directions. The directions of the eigenvectors of large eigenvalues of $X'X$ are the directions in which a small investment of $\|\boldsymbol{\beta}\|$ results in a large r^2.

We get a similar look at the power properties of the F-test if we orthogonalize the design by taking $\tilde{\boldsymbol{\beta}} = (X'X)^{1/2}\boldsymbol{\beta}$ and $\tilde{X} = X(X'X)^{-1/2}$. This results in $X\boldsymbol{\beta} = \tilde{X}\tilde{\boldsymbol{\beta}}$ for all $\boldsymbol{\beta}$ so the distribution of $\mathbf{y}$ is unchanged. Unlike the F-test, the locally most powerful test is not invariant under change of parametrization: under the assumption of exchangeability $\mathrm{E}[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}'] = \tau^2 I$ on $\tilde{\boldsymbol{\beta}}$ we now get the F-test as the locally most powerful test for the new parametrization. Applying the reasoning of Section 5.6 to the new parametrization, we see that the F-test optimizes power not over small balls with $\boldsymbol{\beta}'\boldsymbol{\beta} = c$, but on small ellipsoids with $\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}'X'X\boldsymbol{\beta} = c$, which are ellipsoids of alternatives that have the same r^2. All alternatives with the same r^2 have the same potential power, so there is no trade-off and all alternatives in the ellipsoid are given equal power. The expected test statistic under the alternative minus the expected test statistic under the null for the F-test is

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[F] - \mathrm{E}_{\mathbf{y}|\mathbf{0}}[F] = r^2(1 - \frac{p}{n}),$$

which only depends on $\boldsymbol{\beta}$ through r^2. It is positive whenever $r^2 > 0$ and $p < n$.

The main difference between the empirical Bayes score test and the F-test is therefore that, while for the F-test all alternatives with the same r^2 are as credible and interesting to detect, the score test is explicitly directed at finding parsimonious alternatives, which can explain $\mathbf{y}$ with minimal expenditure of $\|\boldsymbol{\beta}\|$.

There is no easy analytic expression which shows for which alternatives in the $p \leq n$ situation the F-test has more power than the score test and vice versa. However, it can be convincingly argued that for those alternatives in which the large variance principal components of X explain most of the variance of $\mathbf{y}$, the score test has more power, while for the alternatives in which the small-variance principal components explain most of the variance of $\mathbf{y}$, the F-test is more powerful. This can be seen by writing XX' in a spectral decomposition as $XX' = \sum_{i=1}^{n} \lambda_i P_i$, where P_i is the $n \times n$ projection matrix for projection on the i-th principal component. Then

$$S = \sum_{i=1}^{n} \lambda_i \frac{\mathbf{y}'P_i\mathbf{y}}{\mathbf{y}'\mathbf{y}},$$

so the test statistic S is a weighted sum of the test statistics $\mathbf{y}'P_i\mathbf{y}$ which test whether the i-th principal component is associated with $\mathbf{y}$. The weights are proportional to the variance of the principal components. In the same way

$$F = \sum_{i=1}^{n} \frac{\mathbf{y}'P_i\mathbf{y}}{\mathbf{y}'\mathbf{y}},$$

the statistic F is the unweighted sum of the same test statistics. Comparing the two composite tests, one can argue that one has more power than the other if it puts heavier weights on the terms with most power. We'll illustrate this point with simulations in Section 5.10.

An interesting type of alternative against which the locally most powerful test can be expected to have more power than the F-test is a factor-analysis type setup, in which a limited number of latent variables linearly determines both the covariates X and the outcome variable $\mathbf{y}$, but both are measured with error (Bartholomew and Knott, 1999). In this case the latent variables tend to show up in the large-variance principal components of X, while the uninformative noise tends to dominate the small-variance principal components. This setup is not unrealistic for many practical problems, especially in high-dimensional data, as the covariates can often be seen as noisy measurements of more or less the same underlying mechanisms. In this kind of alternative one would normally apply principal components testing: reducing the matrix X to its first few principal components and then applying the F-test. An important advantage of the locally most powerful test over principal components testing is that there is no need to choose the number of principal components. We come back to principal components testing in Section 5.10.

5.9 Sparse alternatives

In the previous sections we have established that the locally most powerful test is especially directed against parsimonious alternatives with small $\|\boldsymbol{\beta}\|$. A different type of parsimonious alternative is the sparse alternative, in which only a few entries of $\boldsymbol{\beta}$ are non-zero. This type of alternative is especially of interest in regression modelling.

A test which specifically aims to detect this type of sparse alternative in a regression model is a multiple testing procedure. This type of testing procedure is often used in microarray data analysis. There are many variants, but the most basic form is the following: for $i = 1, \ldots, p$ a t-test statistic t_i is calculated to test for association of each covariate with the outcome $\mathbf{y}$. The test statistic $\tilde{T}_{\max} = \max(|t_1|, \ldots, |t_p|)$ is used to test whether there is an association between any covariate and $\mathbf{y}$. The critical value of $T_{\max}$ can be found either conservatively using the Bonferroni adjustment, or using numerical methods.

Different though this test may seem from the locally most powerful test, there is still a connection. First, we can transform each $|t_i|$ to $g(t_i^2)$, using the function $g(x) = (x^{-1}+1)^{-1}$ also used in section 5.8. This results in test statistics with a beta distribution with parameters $\frac{1}{2}$ and $\frac{1}{2}(n-1)$. As $g(x^2)$ is increasing in $|x|$, the test statistic

$$T_{\max} = \max\{g(t_1^2), \ldots, g(t_p^2)\}$$

is equivalent to $\tilde{T}_{\max}$. Next, we write $\mathbf{x}_i$ for the i-th column of X, then $g(t_i^2) = \mathbf{y}'\mathbf{x}_i\mathbf{x}_i'\mathbf{y}/(\mathbf{y}'\mathbf{y} \cdot \mathbf{x}_i'\mathbf{x}_i)$. However, as we can write $XX' = \sum_{i=1}^p \mathbf{x}_i\mathbf{x}_i'$, we can say that

$$S = \sum_{i=1}^{p} \mathbf{x}_i'\mathbf{x}_i g(t_i^2),$$

so the locally most powerful test statistic is a weighted sum of the same (transformed) t-test statistics over which $T_{\max}$ is the maximum. The weights are proportional to the variance of $\mathbf{x}_i$.

Perhaps surprisingly, by Lemma 4 the score test is more powerful than the test based on $T_{\max}$ in the situation where p is large and r^2 is very small, even when only a single regression coefficient is non-zero. Suppose $\boldsymbol{\beta}$ is given a single non-zero entry at random, of fixed size, but with random sign. This distribution of $\boldsymbol{\beta}$ has $\mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$ and, if p is large $\mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}') \approx \tau^2 I$ for some τ^2. By Lemma 4, the score test has optimal power on average to detect these alternatives if τ^2 is small.

This optimality can again be understood in terms of the principal components. If there are few principal components with large variance, it is probable that the $\mathbf{x}_i$ with the positive regression coefficient also has a major part of its variance in the direction of these large-variance principal components. If $\mathbf{y}$ is correlated with $\mathbf{x}_i$, it is therefore automatically correlated with these principal components, and therefore with many other covariates $\mathbf{x}_j$, which also tend to have a large part of their variance in the direction of the large-variance principal components. A single regression coefficient may therefore lead to many significant t-statistics. In this situation there may be more information in the sum of the t-statistics than in the maximum.

Simulations in section 5.10 illustrate these points.

5.10 Simulations

Many of the points raised in the previous sections require some illustration. We'll do this using simulations in the linear model. The simulations are based on real data in the sense that the design matrix X is a taken as a real biological

data set: a microarray data set of gene expression measurements of $p = 4911$ genes, measured for $n = 294$ breast cancer patients (obtained from Van de Vijver et al., 2002, after removing some genes and patients due to missing values). The matrix X was normalized to have both row and column means zero. After this normalization X has rank $n - 1$ and a ratio of the largest to the smallest non-zero singular value of 26.6. Using this design matrix X, values of $\mathbf{y}$ are simulated based on the models chosen below.

First we compare the locally most powerful test with the F-test, to illustrate the statements from Section 5.8 that the score test has more power when the large variance principal components of X explain most of the variance of $\mathbf{y}$. As we cannot use the F-test when $p > n$, we reduce the matrix X to X^* by selecting as covariates only the $p^* = 52$ genes belonging to the apoptosis pathway.

The simulation setup is as follows. We write X^* in a singular value decomposition as $X^* = U\Lambda^{1/2}V'$, with U an $n \times p^*$ semi-orthogonal matrix, V a $p^* \times p^*$ orthogonal matrix and Λ a $p^* \times p^*$ diagonal matrix with diagonal elements $\boldsymbol{\lambda}^* = (\lambda_1, \ldots, \lambda_{p^*})'$, where each λ_i is the variance of the i-th principal component. To vary the amount of variance explained by the principal components, we choose the regression coefficients as $\boldsymbol{\beta} = V\Lambda^{(s/2-1)}\boldsymbol{\lambda}$ for various values of s. In this setup the i-th principal component has regression coefficient $\lambda_i^{s/2}$ and explains a fraction r_i^2 of the variance of $\mathbf{y}$ proportional to λ_1^{s+1}. Hence, if $s > 0$, the large variance principal components have larger regression coefficients and therefore explain more of the variance of $\mathbf{y}$; if $-1 < s < 0$, the large variance principal components have smaller regression coefficients, but still explain more of the variance than the small-variance principal components, while if $s < -1$, the small-variance principal components dominate $\mathbf{y}$. By varying σ^2 as a function of s we can obtain all values of r^2 for every s.

To estimate the power for these alternatives, we generated 10000 $\mathbf{y}$ vectors each from alternatives with various values of s and r^2. The cutoff at level α for the S statistic was found using the methods of Imhof (1961). The results are given in table 5.1. They show that the power of the score test and the F-test is comparable for $s = -1/2$, although the F-test still has a slight advantage here. The score test is substantially more powerful for larger values of s, the F-test is more powerful for smaller values. This is in line with the theoretical discussion in section 5.8.

It is also interesting to compare the locally most powerful test with the test P_1, which is the F-test that tests whether the first principal component of X^* is correlated with $\mathbf{y}$. The results are also given in table 5.1. We can see that the locally most powerful test is comparable in power to the test P_1 for high values of s, but it is consistently better for all the alternatives considered.

In a second simulation experiment we look at sparse alternatives in high-

TABLE 5.1: *Monte Carlo power comparison between the locally most powerful test, the F-test and the test P_1, which uses only the first principal component for testing. The tests use $\alpha = 0.05$. The various alternatives are given by their r^2 and a coefficient s: $s > 0$ means that large-variance principal components get larger regression coefficients, $s < 0$ vice versa.*

	$r^2 = 0.02$			$r^2 = 0.05$		
alternative	F	S	P_1	F	S	P_1
$s = 1.5$	0.14	0.52	0.52	0.35	0.92	0.90
$s = 1$	0.14	0.46	0.44	0.35	0.88	0.82
$s = 0.5$	0.14	0.36	0.31	0.34	0.79	0.66
$s = 0$	0.13	0.24	0.19	0.34	0.58	0.39
$s = -0.5$	0.14	0.13	0.10	0.35	0.32	0.18
$s = -1$	0.14	0.08	0.06	0.34	0.14	0.08
$s = -1.5$	0.14	0.06	0.05	0.35	0.07	0.05

	$r^2 = 0.10$			$r^2 = 0.15$		
alternative	F	S	P_1	F	S	P_1
$s = 1.5$	0.76	1.00	1.00	0.96	1.00	1.00
$s = 1$	0.76	1.00	0.99	0.96	1.00	1.00
$s = 0.5$	0.76	0.99	0.92	0.96	1.00	1.00
$s = 0$	0.75	0.92	0.67	0.96	0.99	0.86
$s = -0.5$	0.76	0.65	0.31	0.96	0.89	0.43
$s = -1$	0.76	0.27	0.10	0.96	0.44	0.13
$s = -1.5$	0.75	0.10	0.05	0.96	0.13	0.05

dimensional data. We compare the power of the locally most powerful test with the power of the test based on $T_{\max}$, the maximum absolute t-statistic, as discussed in Section 5.9.

For this we reverted back to the original high-dimensional data set with $p = 4911$ genes. We generated alternatives $\boldsymbol{\beta}_{m,j}$ for $j = 1, \ldots, p$ and $m = 1, 3, 10, 30$, such that each alternative $\boldsymbol{\beta}_{m,j}$ has the m regression coefficients $\beta_j, \ldots, \beta_{j+m-1}$ equal to 1 and all others equal to zero (taking $\beta_i = \beta_{i-p}$ if $i > p$). Table 5.2 shows the power of the tests based on S and $T_{\max}$ on average against the alternatives $\boldsymbol{\beta}_{m,1}, \ldots, \boldsymbol{\beta}_{m,p}$ with m non-zero regression coefficients. In the simulations the value of σ^2 was taken to be equal for all alternatives $\boldsymbol{\beta}_{m,1}, \ldots, \boldsymbol{\beta}_{m,p}$ and was chosen to get a certain average r^2 over these alternatives. We generated 2 replicates for each of the alternatives, so that each power calculation is based on $2p \approx 10000$ Monte Carlo samples of $\mathbf{y}$.

A complicating factor in this simulation is the lack of a simple and accurate method to find the distribution function of the statistic $T_{\max}$, because of the dependence of the t-statistics. We used simulation to find the α cutoff of

$T_{\max}$ for the design matrix X. The 0.05-cutoff was found at 0.062, using 20000 simulations of $\mathbf{y}$ under the null. Note that this is only slightly below the crude Bonferroni corrected cutoff for p beta$(\frac{1}{2}, \frac{1}{2}(n-1))$ variables, which is at 0.064.

TABLE 5.2: *Monte Carlo power comparison between the locally most powerful test and the test based the maximum of p absolute t-statistics using $\alpha = 0.05$. The reported power values are on average over p different sparse alternatives with m non-zero regression coefficients.*

alter-	$r^2 = 0.01$		$r^2 = 0.02$		$r^2 = 0.05$		$r^2 = 0.10$		$r^2 = 0.20$	
native	S	$T_{\max}$	S	$T_{\max}$	S	$T_{\max}$	S	$T_{\max}$	S	$T_{\max}$
$m = 1$	0.12	0.10	0.17	0.16	0.33	0.40	0.54	0.74	0.76	0.97
$m = 3$	0.11	0.09	0.17	0.14	0.34	0.32	0.55	0.61	0.80	0.90
$m = 10$	0.11	0.09	0.17	0.14	0.35	0.29	0.58	0.54	0.83	0.84
$m = 30$	0.11	0.09	0.17	0.13	0.34	0.28	0.55	0.51	0.80	0.79

The table confirms the theoretical result of Section 5.9 that for sparse alternatives close to the null the score test is slightly superior to the test based on $T_{\max}$. This superiority disappears quite quickly, however, when the single covariate explains a large portion of the variance of $\mathbf{y}$. Looking at decreasingly sparse alternatives, the $T_{\max}$ statistic loses power, as can be expected, but the score test remains more or less stable. What is perhaps most surprising about table 5.2, is that even though the tests are constructed in a very dissimilar way, the average power is still quite similar. The $T_{\max}$ still has good power against not-so-sparse alternatives, while the locally most powerful test has good power against sparse alternatives far from the null.

5.11 Discussion

For testing against a multi-dimensional alternative there are no uniformly most powerful tests. Tests may only be optimal locally for some alternatives, or optimal on average over a region of alternatives. When choosing a test against multi-dimensional alternatives, it is therefore important to consider against which alternatives the chosen test has good power. When constructing such a test, one can use empirical Bayes modelling to design a test which has optimal power on average against a chosen region of alternatives. Thinking about these issues is especially relevant when the data are high-dimensional, because the power of often-used classical tests tends to diminish rapidly when the dimensionality increases.

A drawback of empirical Bayes design of hypothesis tests, is that the construction of the test requires integration over complicated distributions in possibly high-dimensional space. In this paper we have shown in general how to

avoid this problem by using a score test. This test has the property that it is locally most powerful: it has optimal average power in a well-defined neighbourhood of the null hypothesis.

In the linear model, we have shown that this test has good power for many important alternatives, even in the classical low-dimensional situation. The empirical Bayes score test often has better power than the F-test in the situations where there are errors in variables in the design matrix X, when a small set of latent variables influences both the covariates in X and the outcome variable $\mathbf{y}$, or more generally when the large-variance principal components of X explain more of the variance of $\mathbf{y}$ than the small-variance ones. We have also shown that the empirical Bayes score test has good power in truly high-dimensional situations, even against sparse alternatives. If the fraction of variance of $\mathbf{y}$ explained by the covariates is low, the test even outperforms the test based on the maximum absolute t-statistics of all covariates, a test which is designed to find sparse alternatives.

As high-dimensional data become more and more common, so will the need for testing against high-dimensional alternatives. This paper has given a general theoretical outline and presented a concrete example of a model in which the test has good power. But locally most powerful testing in high dimensions has many more potential applications, both in generalized linear models and more generally.

5.12 Proofs of the lemmas

Proof of Lemma 2: To prove Lemma 2, we have to adopt a slightly more formal notation. Shorthand f_θ for the density of $\mathbf{y}$ and let μ be a dominating measure, so that we can write $\mathrm{P}_{\mathbf{y}|\theta}(\mathbf{y} \in A) = \int_A f_\theta \,\mathrm{d}\mu$. Also, let $\mathbf{1}_{\{\cdot\}}$ denote an indicator function.

To prove the existence, we write $w(\theta) = \int_A f_\theta \,\mathrm{d}\mu$, so by the dominated convergence theorem $\frac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0) = \int_A \frac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} \mathrm{d}\mu < \infty$.

Furthermore, noting that $\frac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} = S^* f_{\theta_0}$, and using $\mathbf{1}_A - \mathbf{1}_B = \mathbf{1}_{A\setminus B} - \mathbf{1}_{B\setminus A}$ twice, we can calculate

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0) - \tfrac{\mathrm{d}}{\mathrm{d}\theta} w^*(\theta_0) &= \int_A \tfrac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} \,\mathrm{d}\mu - \int_{S^* \geq k} \tfrac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} \,\mathrm{d}\mu \\
&= \int_{A, S^* < k} S^* f_{\theta_0} \,\mathrm{d}\mu - \int_{A^c, S^* \geq k} S^* f_{\theta_0} \,\mathrm{d}\mu \\
&\leq k \int_{A, S^* < k} f_{\theta_0} \,\mathrm{d}\mu - k \int_{A^c, S^* \geq k} f_{\theta_0} \,\mathrm{d}\mu \\
&= k \int_A f_{\theta_0} \,\mathrm{d}\mu - k \int_{S^* \geq k} f_{\theta_0} \,\mathrm{d}\mu \\
&= k\{w(\theta_0) - w^*(\theta_0)\}.
\end{aligned}$$

The latter term is at most zero whenever condition (i) or (ii) holds. □

Proof of Lemma 3: The assumptions of bounded derivatives combined with the assumption that the distribution of $\mathbf{b}$ is free of τ allows us to interchange limits and integrals in the following calculations. For simplicity we suppress the dependence on $\mathbf{y}$ in the notation.

$$S = \lim_{\tau^2 \downarrow 0} \frac{\frac{\mathrm{d}}{\mathrm{d}\tau^2} \mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]}{\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]} = \lim_{\tau^2 \downarrow 0} \frac{\mathrm{E}_{\mathbf{b}}[\frac{\mathrm{d}}{\mathrm{d}\tau^2} f(\tau\mathbf{b})]}{\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]} = \lim_{\tau^2 \downarrow 0} \frac{\mathrm{E}_{\mathbf{b}}[(\frac{\mathrm{d}f(\tau\mathbf{b})}{\mathrm{d}\boldsymbol{\beta}})'\mathbf{b}]}{2\tau \mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]}.$$

The limit evaluates to $0/0$, so we use l'Hôpital's rule to get

$$S = \lim_{\tau^2 \downarrow 0} \frac{\mathrm{E}_{\mathbf{b}}[\mathbf{b}' \frac{\partial^2 f(\tau\mathbf{b})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\mathbf{b}]}{2\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})] + 2\tau \frac{\partial}{\partial\tau^2}\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]} = \frac{\mathrm{E}_{\mathbf{b}}[\mathbf{b}' \frac{\partial^2 f(\mathbf{0})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\mathbf{b}]}{2\mathrm{E}_{\mathbf{b}}[f(\mathbf{0})]}.$$

Now it only remains to rewrite $\frac{\partial^2 f(\mathbf{0})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = f(\mathbf{0}) \cdot \{\mathbf{s}\mathbf{s}' - \mathbf{I}\}$. □

Proof of Lemma 4: Assume $w(\mathbf{0}) = \bar{w}(\mathbf{0})$. By Lemma 3 every distribution of $\boldsymbol{\beta}$ with $\mathrm{E}(\boldsymbol{\beta}) = 0$ and $\mathrm{E}(\boldsymbol{\beta}'\boldsymbol{\beta}) \propto \tau^2 I$ leads to the same test statistic and therefore to the same power function. Without loss of generality we can therefore assume that $\bar{w}$ is the power function of the score test in the empirical Bayesian model in which $\boldsymbol{\beta}$ is distributed as $\tau\boldsymbol{\xi}$. By Lemma 2 we have $\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\beta}|\tau^2}[w(\boldsymbol{\beta})] \leq \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\beta}|\tau^2}[\bar{w}(\boldsymbol{\beta})]$ in $\tau^2 = 0$. The boundedness assumptions of Lemma 3 allow interchanging differentiation and integration, so we get

$$\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\beta}|\tau^2}[w(\boldsymbol{\beta})] = \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\xi}}[w(\tau\boldsymbol{\xi})] = \mathrm{E}_{\boldsymbol{\xi}}\Big[\frac{\mathrm{d}}{\mathrm{d}\tau^2} w(\tau\boldsymbol{\xi})\Big],$$

both for w and for $\bar{w}$, from which the result follows. □