

Convex Hierarchical Testing of Interactions

JACOB BIEN*, NOAH SIMON † and ROBERT TIBSHIRANI‡

March 31, 2014

Abstract

We consider the testing of all pairwise interactions in a two-class problem with many features. We devise a hierarchical testing framework that considers an interaction only when one or more of its constituent features has a nonzero main effect. The test is based on a convex optimization framework that seamlessly considers main effects and interactions together. We show—both in simulation and on a genomic dataset from the SAPPHIRe study—a potential gain in power and interpretability over a standard (non-hierarchical) interaction test.

1 Introduction

We consider the standard two-class problem with $y_i \in \{1, 2\}$ and p features $\{x_{i1}, x_{i2}, \dots, x_{ip}\}$ measured on each of $i = 1, 2, \dots, n$ observations. Large-scale hypothesis testing for the effects of individual features (such as gene expressions; see Section 5) is a challenging problem and has received much attention in recent years (e.g. Efron 2010, Dudoit & van der Laan 2008). The problem of testing for interactions between pairs of features is even more difficult, as there are $\binom{p}{2}$ interactions. Buzkov, Lumley & Rice (2011) show that standard permutation tests cannot be used for interaction testing (because the correct null hypothesis is difficult to enforce), and propose instead a parametric bootstrap-based approach. Simon & Tibshirani (2012) devise a permutation approach that exploits the close relationship between the “forward” logistic model (based on $Y|X$) and a “backward” discriminant analysis (Gaussian) model (based on $X|Y$).

When p is large, the large number of potential pairwise interactions can result in low power for detecting the true effects. One strategy used by data analysts is to first screen the data for significant main effects, and then test for interactions only among those features that are themselves significant. This approach can be effective, but it has some drawbacks. Specifically, at what threshold does one stop entering main effects? And should this threshold depend on the strength of the interactions?

*Departments of Biological Statistics and Computational Biology and Statistical Science, Cornell University, jbien@cornell.edu

†Department of Biostatistics, University of Washington, nrsimon@uw.edu

‡Departments of Health, Research & Policy, and Statistics, Stanford University, tibs@stanford.edu

The above two-stage strategy can be viewed as “hierarchical”: interactions are considered only if both constituent main effects are significant. In this paper we propose a convex formulation that models main effects and interactions together, in a hierarchical fashion. It provides a testing framework that seamlessly combines main effects and interactions. We call the method *convex hierarchical testing* (CHT). The method is closely related to the recently proposed hierarchical lasso regression method (“hierNet”) of Bien, Taylor & Tibshirani (2013). A difference is that CHT seeks marginal interactions while hierNet looks for conditional interactions. We focus exclusively on pairwise interactions in the paper, but discuss possible extensions to higher order interactions in Section 7.

2 Testing interactions using a convex formulation

Our proposal has three main components. The first is to define interactions and main effects in terms of a “backward model” (Simon & Tibshirani 2012). The second component is to relate the testing problem to the solution path of a convex optimization problem. The third component is to introduce hierarchical sparsity within the convex optimization framework. We present these components in the next three subsections.

2.1 Defining interaction and main effects via the backward model

A common approach to testing interactions is to consider $\binom{p}{2}$ logistic regression models of the form

$$\log \left[\frac{p(Y|X_j, X_k)}{1 - p(Y|X_j, X_k)} \right] = \alpha_0 + \alpha_j X_j + \alpha_k X_k + \alpha_{jk} X_j X_k,$$

and then to test whether $\alpha_{jk} = 0$. However, Simon & Tibshirani (2012) argue that this definition of interaction is less natural than one coming from considering a “backward model” in which the feature vector $X \in \mathbb{R}^p$ is modeled conditional on the class label $Y \in \{1, 2\}$:

$$X | Y = \ell \sim N_p(\mu^{(\ell)}, \Sigma^{(\ell)}).$$

In particular, they redefine an interaction between X_j and X_k to mean that $\text{Cor}(X_j, X_k | Y = \ell) = \rho_{jk}^{(\ell)} = (\Sigma_{jj}^{(\ell)} \Sigma_{kk}^{(\ell)})^{-1/2} \Sigma_{jk}^{(\ell)}$ depends on ℓ . Their main criticism of defining interactions based on the forward model is that if the *marginal* quantity $\text{Var}(X_j | Y = \ell) = \Sigma_{jj}^{(\ell)}$ depends on ℓ , then $\alpha_{jk} \neq 0$ for all k . This does not correspond to what a biologist, say, would consider an interesting interaction because it is not a property of the *pair* of variables (X_j, X_k) . Likewise a natural definition for main effects in the backward model is a difference in class mean for that variable. Hence, we work in the backward model and test hypotheses of two kinds, which we will refer to as “main effects” and “interactions”:

$$\begin{aligned} H_{0,j} : \mu_j^{(1)} &= \mu_j^{(2)} && \text{for } 1 \leq j \leq p \\ H_{0,jk} : \rho_{jk}^{(1)} &= \rho_{jk}^{(2)} && \text{for } 1 \leq j < k \leq p. \end{aligned}$$

For testing $H_{0,j}$, a common choice would be the standard t -statistic,

$$w_j = \frac{\bar{x}_j^{(1)} - \bar{x}_j^{(2)}}{\sqrt{s_j^{(1)2}/n_1 + s_j^{(2)2}/n_2}},$$

where $\bar{x}_j^{(\ell)} = n_\ell^{-1} \sum_{i:y_i=\ell} x_{ij}$ and $s_j^{(1)2} = (n_\ell - 1)^{-1} \sum_{i:y_i=\ell} (x_{ij} - \bar{x}_j^{(\ell)})^2$ are the sample mean and variances within class ℓ . For testing $H_{0,jk}$, a common choice would be the difference of the Fisher transformed sample correlations between the two classes:

$$z_{jk} = \left(\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \right)^{-1/2} \left[\operatorname{arctanh}(\hat{\rho}_{jk}^{(1)}) - \operatorname{arctanh}(\hat{\rho}_{jk}^{(2)}) \right].$$

Here, $\hat{\rho}_{jk}^{(\ell)} = (n_\ell - 1)^{-1} \sum_{i:y_i=\ell} (x_{ij}^{(\ell)} - \bar{x}_j^{(\ell)})(x_{ik}^{(\ell)} - \bar{x}_k^{(\ell)}) / (s_j^{(\ell)} s_k^{(\ell)})$ is the sample correlation within class ℓ . Both w_j and z_{jk} are scaled so that they are approximately standard normal (for large n_1 and n_2).

2.2 Test statistics through convex optimization

We would like to select interactions based on the size of $|z_{jk}|$, but also somehow give a “boost” to interactions whose main effects are large. One could try to achieve this through a two-stage procedure in which one first screens the individual features and then tests for interactions only among those features selected at the first phase. This kind of method is explored for example in Kooperberg & LeBlanc (2008), Hsu, Jiao, Dai, Hutter, Peters & Kooperberg (2012), and Wu & Zhao (2009). However, such an approach to the hierarchy requirement can lead to algorithmic shortsightedness. In particular, a very strong interaction will be ignored if the corresponding main effects fail to make the threshold in the first phase. We seek a method that enforces the hierarchy constraint but *jointly* considers which interactions and main effects to include in the model.

Suppose that we define a testing procedure through a convex optimization problem involving both w_j and z_{jk} . Let $\beta^+, \beta^- \in \mathbb{R}^p$ and $\theta \in \mathbb{R}^{p(p-1)}$ be optimization variables. Given the objective function

$$\begin{aligned} L_\lambda(\beta^+, \beta^-, \theta) &= \frac{1}{2} \sum_{j=1}^p (w_j - (\beta_j^+ - \beta_j^-))^2 + \frac{1}{2} \sum_{j=1}^p \sum_{j \neq k} (z_{jk} - \theta_{jk})^2 \\ &\quad + \lambda \sum_{j=1}^p [\beta_j^+ + \beta_j^-] + \lambda \sum_{j=1}^p \sum_{k \neq j} |\theta_{jk}|, \end{aligned} \tag{1}$$

we may define the problem

$$\min_{\beta^+, \beta^-, \theta} L_\lambda(\beta^+, \beta^-, \theta) \text{ s.t. } \beta_j^+ \geq 0, \beta_j^- \geq 0 \text{ for } 1 \leq j \leq p,$$

where λ is a tuning parameter. For each fixed $\lambda \geq 0$, the pair $(\hat{\beta}^+(\lambda) - \hat{\beta}^-(\lambda), \hat{\theta}(\lambda))$ is unique. Consider the path of solutions formed by varying λ from ∞ to 0. The solution path goes from $(0, 0) \in \mathbb{R}^{p+p(p-1)}$ to (w, z) , and is piecewise-linear with knots at the values of λ for which individual coordinates of $\hat{\theta}_{jk}(\lambda)$ or $\hat{\beta}_j^+(\lambda) - \hat{\beta}_j^-(\lambda)$ become nonzero. It is straightforward to show that these knots occur precisely at the values of the standard test statistics introduced in the previous section:

$$|w_1|, \dots, |w_p|, |z_{12}|, \dots, |z_{p-1,p}|. \tag{2}$$

This observation suggests how a regularized regression problem can be viewed as producing test statistics: one defines the test statistic associated with a variable to be the λ value at which the corresponding parameter becomes nonzero.

Now in this setup, the k th knot is just equal to the k th largest value among those in (2), so our test for interactions is just the usual one, based on the size of $|z_{jk}|$. We have not obtained anything new. To exploit hierarchy, we will modify the optimization problem as described in the next section.

2.3 Convex hierarchical testing

The procedure described above does not share information between main effects and interactions. Our proposal in this paper is to add a convex hierarchy constraint to the problem, which will lead to *main-effect “informed” thresholds* for testing the interactions (and likewise interaction “informed” thresholds for testing main effects).

Bien et al. (2013) develop a hierarchical interactions lasso method in the forward model. The hierarchical sparsity is achieved by adding a set of convex constraints to the lasso problem. We may similarly impose hierarchy in the backward model by modifying (1) to get a hierarchical interactions test in the backward model:

$$(\hat{\beta}^+, \hat{\beta}^-, \hat{\theta}) = \arg \min L_\lambda(\beta^+, \beta^-, \theta) \text{ subject to } \beta_j^+, \beta_j^- \geq 0, \sum_{k:k \neq j} |\theta_{jk}| \leq \beta_j^+ + \beta_j^-. \quad (3)$$

We solve this problem for all λ , and define the test statistic associated with an interaction to be the λ value at which the corresponding parameter becomes nonzero. This is the main proposal of this paper.

The addition of the constraint imposes a “budget” $\beta_j^+ + \beta_j^-$ on the total interactions that involve feature j . In particular, the constraint $\sum_k |\theta_{jk}| \leq \beta_j^+ + \beta_j^-$ implies that at least one of β_j^+ and β_j^- must be non-zero, in order for θ_{jk} to be non-zero. Although in theory we could have $\hat{\beta}_j^+ = \hat{\beta}_j^-$ with both values positive, this happens with probability zero under reasonable assumptions (Bien et al. 2013). As a result, $\hat{\theta}_{jk} \neq 0$ implies $\hat{\beta}_j \neq 0$, and similarly for $\hat{\theta}_{kj}$. Thus the jk th interaction is nonzero if at least one of $\hat{\beta}_j$ or $\hat{\beta}_k$ is nonzero. This property has been called *weak hierarchy* (see, e.g., Bien et al. (2013)), in contrast to *strong hierarchy*, which requires both $\hat{\beta}_j$ and $\hat{\beta}_k$ to be nonzero, in order for $\hat{\theta}_{jk}$ to be nonzero. Problem (3) is convex, due to the fact that we have represented each main effect β_j as the difference of two non-negative quantities β_j^+, β_j^- . It would not be convex if we had used $|\beta_j|$ in place of $\beta_j^+ + \beta_j^-$ in the constraint above. This is because the set $\{(x, t) : \|x\|_1 \leq |t|\}$ is not convex.

Working in the optimization-based testing framework of the previous section, we consider the solution path (parameterized by λ) of this problem and then define the test statistics for interactions and main effects to be the λ values at which these values become nonzero (i.e., the knots of the path). In particular, for testing the jk th interaction, we take the largest λ for which either $\hat{\theta}_{jk}$ or $\hat{\theta}_{kj}$ is nonzero, and for testing the j th main effect we compute the largest λ for which either $\hat{\beta}_j^+$ or $\hat{\beta}_j^-$ is non-zero. That is, letting $\hat{\beta}(\lambda) = \hat{\beta}^+(\lambda) - \hat{\beta}^-(\lambda)$ and

$\hat{\theta}(\lambda)$ denote the solution as a function of λ , our proposed test statistics are

$$\begin{aligned}\hat{\lambda}_j &= \sup\{\lambda \geq 0 : \hat{\beta}_j(\lambda) \neq 0\} \\ \hat{\lambda}'_{jk} &= \max\{\hat{\lambda}_{jk}, \hat{\lambda}_{kj}\}\end{aligned}$$

where

$$\hat{\lambda}_{jk} = \sup\{\lambda \geq 0 : \hat{\theta}_{jk}(\lambda) \neq 0\}. \quad (4)$$

In Lemma 2 of the online supplement, we prove that (3) has a unique solution for each $\lambda > 0$, so $\hat{\lambda}_j$ and $\hat{\lambda}_{jk}$ are well-defined. Furthermore, we prove in Proposition 2 of the online supplement that $|\hat{\theta}_{jk}(\lambda)|$ is non-increasing in λ , which means that $\hat{\lambda}_{jk}$ is in fact the unique point in the path where $\hat{\theta}_{jk}(\lambda)$ becomes nonzero.

Without the hierarchy constraints in (3), we would have $\hat{\lambda}_{jk} = |z_{jk}|$ and $\hat{\lambda}_j = |w_j|$ as in Section 2.2. The weak hierarchy property of the solution to (3) implies that

$$\hat{\lambda}'_{jk} \leq \max\{\hat{\lambda}_j, \hat{\lambda}_k\}.$$

While one might assume that finding the knots of (3) would be computationally intensive, requiring one to solve the problem at many values of λ , it turns out that there is an explicit analytical form for the knots of this path, meaning that computing the test statistics is in fact computationally simple.

Theorem 1. *The knots of the solution path of (3) have the following closed-form expressions:*

$$\begin{aligned}\hat{\lambda}_j &= \max\left\{|w_j|, \frac{|w_j| + \|z_{j\cdot}\|_\infty}{2}\right\} \\ \hat{\lambda}_{jk} &= \min\left\{|z_{jk}|, \frac{|z_{jk}|}{2} + \frac{[|w_j| - \|\mathcal{S}(z_{j\cdot}, |z_{jk}|)\|_1]_+}{2}\right\},\end{aligned} \quad (5)$$

where $z_{j\cdot} = \{z_{jk} : k \neq j\} \in \mathbb{R}^{p-1}$ is the vector of interaction contrasts involving the j th variable and \mathcal{S} is the soft-thresholding function so that $\|\mathcal{S}(z_{j\cdot}, |z_{jk}|)\|_1 = \sum_{\ell: |z_{j\ell}| > |z_{jk}|} (|z_{j\ell}| - |z_{jk}|)$.

Proof. See Proposition 4 in the online supplement. \square

These formulae are somewhat complex, but we can interpret them loosely as follows. Each main effect is “boosted” by the size of the largest interaction in its row, due to the hierarchy constraint. In contrast, each interaction is shrunk by as much as half of its size, with the shrinkage amount less when the main effect is large or the interaction is large relative to the other interactions in that row. Interestingly, $\hat{\lambda}_{jk}$ depends only on w_j and those interactions in the j th row that are at least as large (in absolute value) as z_{jk} .

At one extreme, suppose $w_j = 0$. In this case, $\hat{\lambda}_j = \|z_{j\cdot}\|_\infty/2$ and $\hat{\lambda}_{jk} = |z_{jk}|/2$ (compared to the non-hierarchical statistics, which would be 0 and $|z_{jk}|$). On the other extreme, $|w_j| \gg$

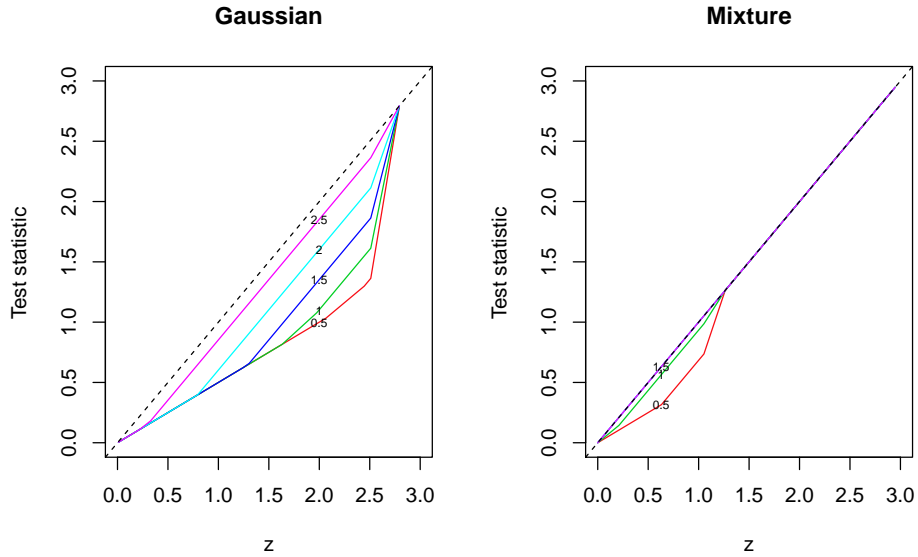


Figure 1: Graphical illustration of formula (5) for two different distributions of interactions (two panels) and different size of main effects w (colored lines). Broken line is the 45° line. Full details in text.

$\|z_{j\cdot}\|_1$, in which case $\hat{\lambda}_j = |w_j|$ and $\hat{\lambda}_{jk} = |z_{jk}|$ (which are identical to the non-hierarchical statistics).

Figure 1 gives a graphical illustration of the formula in (5). We set the number of interactions to 50. In the left panel the interaction contrasts z_{1j} are generated as $N(0, 1)$. The plot shows the test statistic $\hat{\lambda}_j$ as a function of z_{1j} and the main effect w_1 (different colored curves with main effect indicated), along with the 45° line. We see that the interaction effect is shrunk substantially until it reaches about 2.75, and the amount of shrinkage is less when the main effect is larger. In the right panel there are $p - 1 = 49$ small interactions distributed as $N(0, .5^2)$ and one large interaction whose value varies along the horizontal axis. Now we see that there is shrinkage only about until a value of 1.5, and a main effect of 1.5 is sufficient to ensure no shrinkage at all.

The knot-based test statistics produce a single ranking of all interactions and main effects. Our test rejects any null hypotheses whose corresponding knots are greater than a threshold. This threshold is chosen to meet a desired false discovery rate (FDR). In Section 6, we give a method for estimating the FDR. In this way, a practitioner can choose a cutoff with an acceptable FDR. As mentioned above, we call our method *convex hierarchical testing* (CHT). Algorithm 1 spells out the full procedure, which consists of computing the test statistics and then estimating the FDR at a series of cutoffs. The corresponding version of this proposal given in (1) that does not have the hierarchy constants we call the *all pairs* method.

Algorithm 1: Algorithm for convex hierarchical testing

Input: Main effect and interaction contrasts, w_1, \dots, w_p and z_{jk} for $1 \leq j < k \leq p$, as defined in Section 2.1.

- 1 Compute $\hat{\lambda}_j$ for $1 \leq j \leq p$ and $\hat{\lambda}'_{jk}$ for $1 \leq j < k \leq p$ using (5).
- 2 Sort these.
- 3 Reject all hypotheses $H_{0,jk}$ for which $\hat{\lambda}'_{jk} \leq \bar{\lambda}$ (and, if main effects are of interest, all $H_{0,j}$ for which $\hat{\lambda}_j \leq \bar{\lambda}$).
- 4 Repeat B times: Do steps 1-3 on data permuted as described in Section 6.
- 5 Use (7) to compute $\widehat{FDR}(\bar{\lambda})$.

3 Some insight into the objective function (3)

Although the ranking of interactions from the above procedure comes from a seemingly complicated optimization problem, the solutions actually have a simple form. In particular, we prove in the online supplement (see Lemma 1) that

$$\begin{aligned}\hat{\theta}_{jk}(\lambda) &= \mathcal{S}(z_{jk}, \lambda + \hat{\alpha}_j(\lambda)) \\ \hat{\beta}_j^+(\lambda) - \hat{\beta}_j^-(\lambda) &= \mathcal{S}(w_j, \lambda - \hat{\alpha}_j(\lambda))\end{aligned}\tag{6}$$

Here $\mathcal{S}(x, t) = \text{sign}(x) \cdot (|x| - t)_+$ is the soft-thresholding function, and the value $\hat{\alpha}_j(\lambda) \in [0, \lambda]$ emerges from the solution to problem (3), with $\hat{\alpha}_j(\lambda) = 0$ if the hierarchy constraint $\sum_k |\hat{\theta}_{jk}| \leq \hat{\beta}_j^+ + \hat{\beta}_j^-$ is loose (i.e., a strict inequality) and greater than zero otherwise.

For the all-pairs problem following (1), $\hat{\alpha}(\lambda) = 0$ gives the solution. Thus, we can think of $\hat{\alpha}_j(\lambda)$ as the bridge between the main effects and interactions that ensures hierarchy. Its value depends on both the interactions and the main effects. It is easy to see that the j th main effect becomes nonzero at the knot $\hat{\lambda}_j = |w_j| + \hat{\alpha}_j(\hat{\lambda}_j)$ and the jk th interaction becomes nonzero at $\hat{\lambda}_{jk} = |z_{jk}| - \hat{\alpha}_j(\hat{\lambda}_{jk})$. Thus, the solution path $\hat{\alpha}_j(\lambda) \geq 0$ is responsible for the hierarchy-related “boost” we described in the introductory section.

When $|w_j|$ is large enough relative to the $|z_{jk}|$'s, $\hat{\alpha}_j(\lambda) = 0$, i.e. hierarchy holds automatically. When $|z_{jk}|$ is large relative to $|w_j|$, then we may have $\hat{\alpha}_j(\lambda) > 0$, and this can be as large as λ . From (6), we see that $\hat{\alpha}_j(\lambda) > 0$ means that $|w_j|$ are shrunk by less (or even not at all if $\hat{\alpha}_j(\lambda) = \lambda$) and that the interactions are shrunk by more (up to twice as much as in the all-pairs approach). This gives some intuition for Theorem 1.

4 A simulation study

We simulate Gaussian data from the backward model with $n = 200$ observations and $p = 50$ features in two classes $y = 1, 2$. In all cases, we take $\mu^{(1)} = 0$ and $\Sigma^{(1)} = I_p$. We consider six scenarios, each of which has 10 nonzero interactions:

- **WEAK HIERARCHICAL TRUTH (SMALL INTERACTIONS):** We take $\mu_j^{(2)} = 2$ for $j = 1, \dots, 5$, and then select a random 10 interactions $(j_i, k_i) \in \{1, \dots, 5\} \times \{6, \dots, p\}$ to

be nonzero:

$$\Sigma_{jk}^{(2)} = \begin{cases} 0.3 & \text{if } (j, k) = (j_i, k_i) \text{ for some } i = 1, \dots, 10 \\ \Sigma_{jk}^{(1)} & \text{otherwise.} \end{cases}$$

- **WEAK HIERARCHICAL TRUTH:** Same as above, but with 0.5 instead of 0.3.
- **STRONG HIERARCHICAL TRUTH:** We take $\mu_j^{(2)} = 2$ for $j = 1, \dots, 5$, and then take

$$\Sigma_{jk}^{(2)} = \begin{cases} 0.5 & \text{if } 1 \leq j, k \leq 5, j \neq k \\ \Sigma_{jk}^{(1)} & \text{otherwise.} \end{cases}$$

- **NO MAIN EFFECTS TRUTH:** Same as Strong Hierarchical Truth except $\mu^{(2)} = 0$.
- **NO MAIN EFFECTS TRUTH (LARGE INTERACTIONS):** Same as above, but with 0.9 instead of 0.5.
- **ANTI-HIERARCHICAL TRUTH:** We take $\mu_j^{(2)} = 2$ for $j = 1, \dots, 5$, and then take

$$\Sigma_{jk}^{(2)} = \begin{cases} 0.5 & \text{if } 6 \leq j, k \leq 10, j \neq k \\ \Sigma_{jk}^{(1)} & \text{otherwise.} \end{cases}$$

We compare CHT with the all-pairs testing procedure, along with two different two-stage screening methods: in the “strong” version we retain all main effects with z scores above the 75th percentile, and then in the second stage test for interactions only among the retained variables. In the weak version, we consider all interactions among pairs of variables where at least one variable has a z score above the 75th percentile.

Figure 2 shows the true (as opposed to the estimated) false discovery rate for testing interactions, averaged over 100 simulations. In the weak hierarchical scenario with small interactions, we see that CHT shows a substantial improvement in FDR over all pairs, with the weak screen method performing a little worse. In the weak hierarchical scenario with larger interactions, the same ordering of methods holds, though the differences are less pronounced. In the strong hierarchical truth scenario, the strong screening rule does best (by a small amount). We see that in all other scenarios, the strong screening rule does unacceptably poorly. In the three scenarios where hierarchy does not hold, all pairs does best. When no main effects are present and the interactions are large (middle right), CHT does as well as all pairs. This behavior can be explained by (5): when all main effects are small enough, we have $\hat{\lambda}_{jk} \approx |z_{jk}|/2$, which has the same ordering as all pairs. For the screening methods, on the other hand, if a main effect is small, large interactions can go completely undetected. In the anti-hierarchical setting, we construct a scenario in which the hierarchy assumption is explicitly violated. Not surprisingly, CHT and the screening methods do poorly compared to all pairs. Figure 3 shows the performance of the methods in a scenario identical to the “Weak Hierarchical Truth” but with $p = 100$ and $n = 50$. There

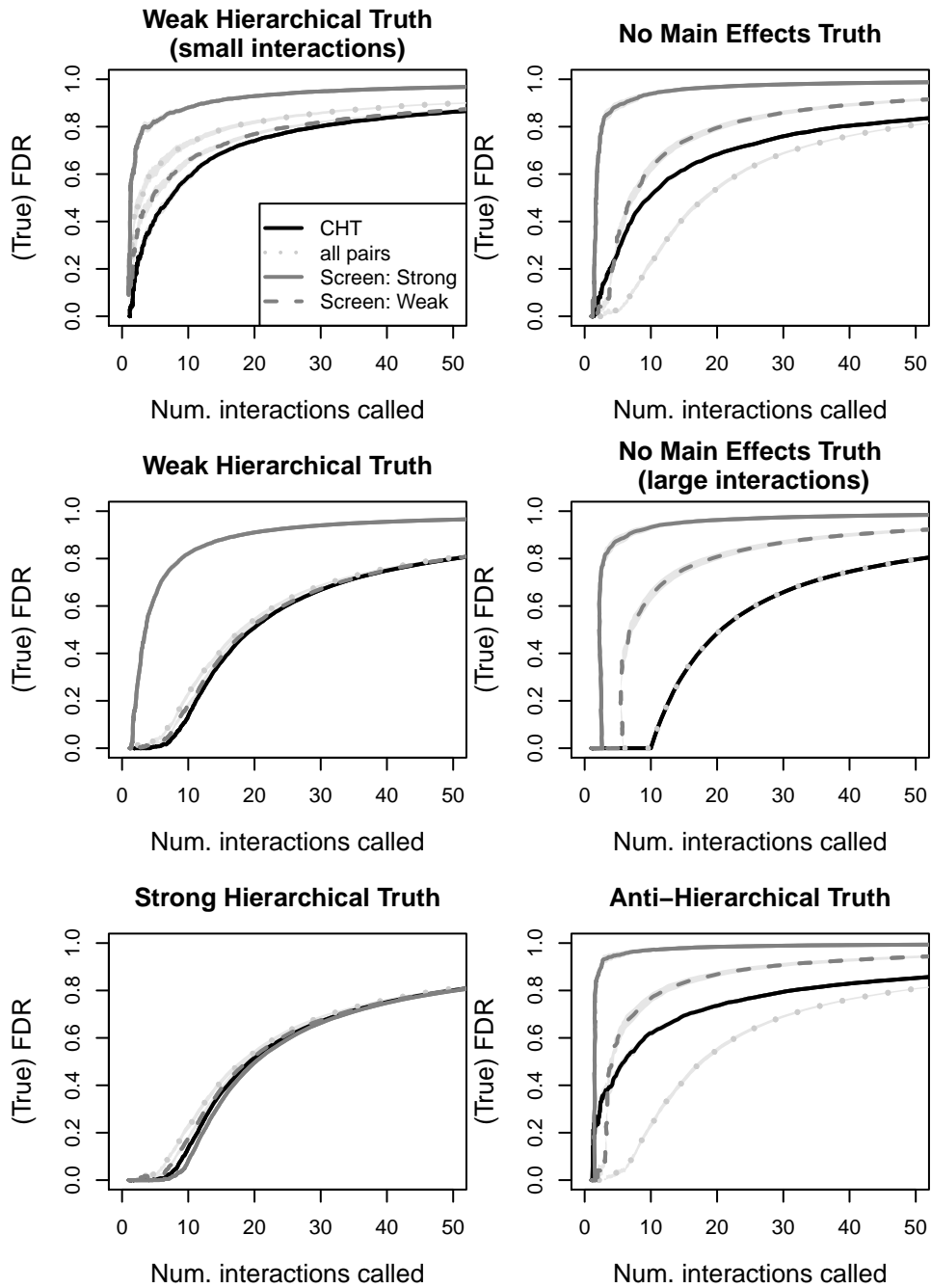


Figure 2: (True) false discovery rates of four different procedures over six different settings. Error bars are shown in light gray. CHT and the screening methods do better than all pairs when the truth is hierarchical (left column). When there are no main effects but there are large interactions (middle right), CHT and all pairs are able to perfectly identify all interactions, whereas the screening methods are not.

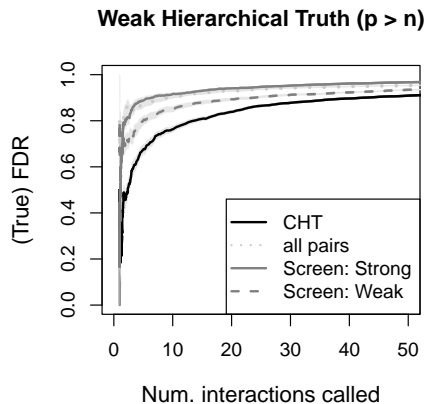


Figure 3: (*True*) false discovery rates of four different procedures when the truth is weak hierarchical with $p = 100$, $n = 50$. Only ten of the 4950 interactions are actually nonzero.

are still only ten nonzero interactions, but now there are 4950 interactions to choose among. The high FDR values show that this is a more challenging scenario; however, CHT performs well compared to the other methods.

In Figure 4, we vary the strength of the main effects in a weak hierarchical scenario with 40 nonzero interactions. We compare all-pairs, CHT, and screen-weak in their ability to correctly detect interactions while controlling FDR at a given sample size. We estimate the average number of non-null interactions called significant (over 50 replications) with $FDR < .2$, for varying sample sizes (horizontal axis) and size of the main effect (varying across panels). When no main effects are present, the all-pairs method does best, and CHT does much better than screen-weak (which is unable to detect over half of the interactions regardless of increasing sample size because these interactions have main effects that are too small). In the other two scenarios CHT does best.

5 Real data example- SAPPHiRe study data

This dataset was analyzed in Park & Hastie (2008), following the study of Huang, Lin, Narasimhan, Quertermous, Hsiung, Ho, Grove, Olivier, Ranade, Risch & Olshen (2004). The study sought to find genes associated with hypertension. A sample of 580 Chinese women, 216 hypotensive and 364 hypertensive, were studied. The predictors (see Table 1) are menopausal and insulin resistant status as well as genotypes on 21 distinct loci.

The first four predictors (all non-genetic) have the strongest effects individually, although none was significantly different across the two groups (details not shown). Table 2 shows the first ten interactions found by the all-pairs and convex hierarchical test methods. Five interactions are shared across these lists. It is interesting to observe how these lists are similar and how they are different. Every gene-gene interaction found by CHT is also in the all-pairs list. Every interaction found *only* by all-pairs and not by CHT is a gene-gene interaction, while every interaction found by CHT but not by all-pairs involves at least one

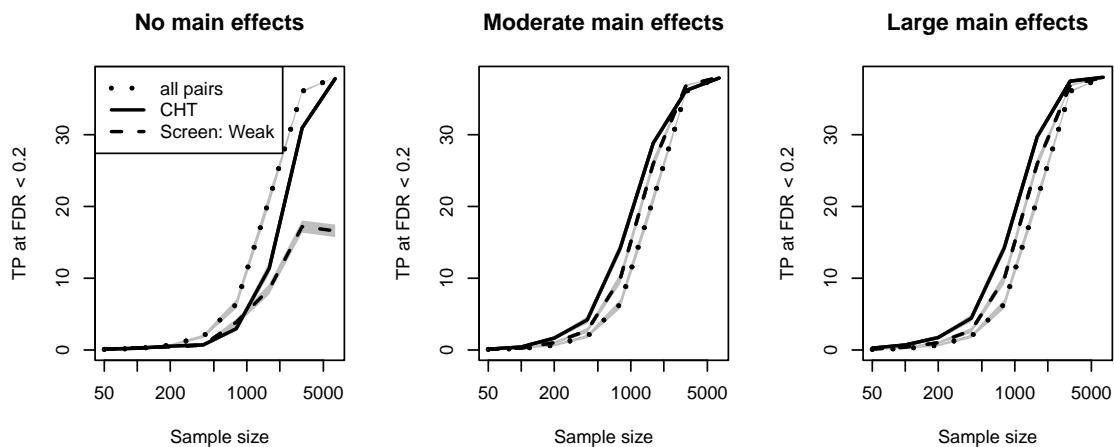


Figure 4: Average number of true positives (*i.e.*, non-null interactions called significant) with $FDR < .2$ (over 50 replications), for varying sample sizes (horizontal axis, logarithmic scale) and size of the main effect. True number of non-null interactions is 40.

Predictor number	Name	Predictor number	Name
1	Reached menopause?	14	PTPN1i4INV
2	insulin t=-10	15	Cyp11B2x1INV
3	insulin t=60	16	PTPN1x9INV
4	insulin t=120	17	ADRB3W1R
5	HUT2SNP5	18	KLKQ3E
6	HUT2SNP7	19	AGT2R1A1166C
7	BADG16R	20	AVPR2G12E
8	AVPR2A1629G	21	MLRI2V
9	AGT2R2C1333T	22	AGTG6A
10	PPARG12	23	Cyp11B2-5paINV
11	CD36x2aINV	24	PTPN1i1
12	MLRi6INV	25	PTPN1i4
13	Cyp11B2i4INV		

Table 1: List of predictors in the Sapphire dataset.

<i>All pairs</i>
PTPN1x9INV : Cyp11B2-5paINV CD36x2aINV : MLRi6INV Cyp11B2-5paINV : PTPN1i4 PTPN1i4INV : Cyp11B2-5paINV Cyp11B2i4INV : PTPN1x9INV CD36x2aINV : KLKQ3E PTPN1x9INV : MLRI2V Reached menopause? : AGT2R1A1166C Cyp11B2i4INV : PTPN1i4 AGT2R2C1333T : CD36x2aINV
<i>Convex hierarchical test</i>
PTPN1x9INV : Cyp11B2-5paINV Reached menopause? : AGT2R1A1166C CD36x2aINV : MLRi6INV insulin t=60 : KLKQ3E insulin t=-10 : HUT2SNP7 CD36x2aINV : KLKQ3E insulin t=60 : Cyp11B2i4INV PTPN1i4INV : Cyp11B2-5paINV insulin t=-10 : ADRB3W1R Reached menopause? : insulin t=120

Table 2: *Top ten interactions found by all-pairs and convex hierarchical test methods.*

All pairs	
Predictors	Bootstrap frequency
CD36x2aINV : KLKQ3E	0.46
Cyp11B2-5paINV : PTPN1x9INV	0.43
Cyp11B2-5paINV : PTPN1i4	0.32
MLRI2V : PTPN1x9INV	0.31
MLRi6INV : CD36x2aINV	0.27
Reached menopause? : AGT2R1A1166C	0.26
PTPN1x9INV : Cyp11B2i4INV	0.25
CD36x2aINV : AGT2R2C1333T	0.21
insulin t=60 : Cyp11B2i4INV	0.20
PTPN1i4INV : BADG16R	0.20
Convex hierarchical test	
PTPN1x9INV : Cyp11B2-5paINV	0.64
Reached menopause? : AGT2R1A1166C	0.47
CD36x2aINV : MLRi6INV	0.33
insulin t=60 : KLKQ3E	0.33
insulin t=-10 : ADRB3W1R	0.32
insulin t=-10 : HUT2SNP7	0.30
insulin t=60 : Cyp11B2i4INV	0.29
Cyp11B2-5paINV : PTPN1i4	0.26
insulin t=-10 : Cyp11B2i4INV	0.25
CD36x2aINV : KLKQ3E	0.22

Table 3: *Ten most frequent interactions found by all-pairs and convex hierarchical test methods over 100 bootstrap replications.*

non-genotype predictor. Figure 5 depicts the main effects and interactions found by CHT for different values of the regularization parameter λ .

To examine the stability of the two analyses, we carried out a bootstrap analysis, recording the top ten pairs appearing in the analysis from each of 100 bootstrap samples. The ten most frequently occurring interactions are shown in Table 3. We see that the CHT test shows a higher reproducibility among the top ten selected interactions. Interestingly, six of the interactions are shared across the lists, and the identical observations hold about the similarities and differences in the lists as in Table 2.

We note that the top interactions found by our procedure, were not found in Park & Hastie (2008). However this may not be surprising, as their paper focused on multivariate modelling and conditional effects.

6 Estimation of the false discovery rate

Permutations provide a convenient and robust way to estimate false discovery rates in large-scale hypothesis testing. For example Simon & Tibshirani (2012) devise a permutation scheme for the all-pairs interaction test. In this scheme, one randomly assigns a component

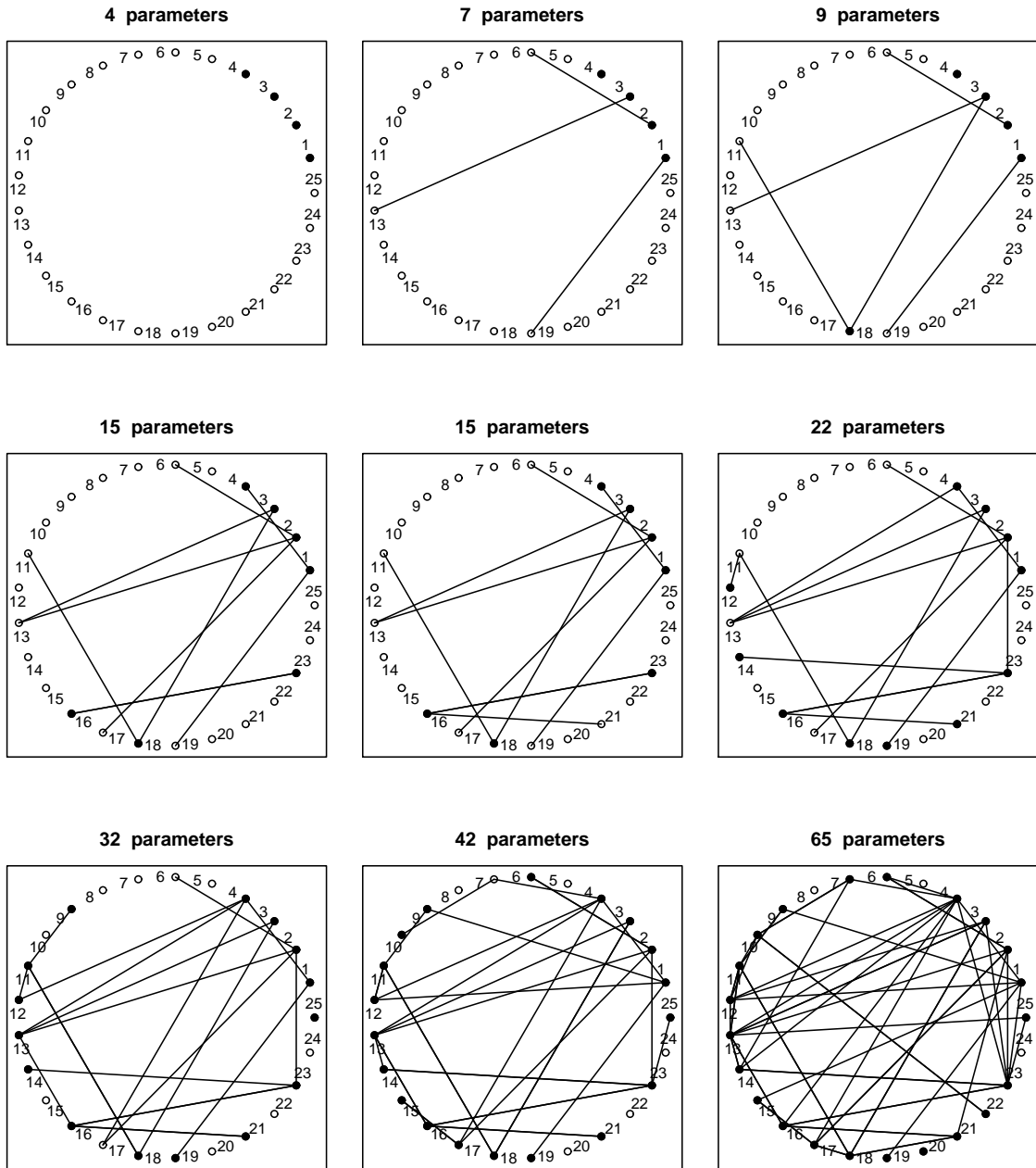


Figure 5: *Convex hierarchical testing: main effects (black dots) and interactions (edges), for 9 different decreasing values of λ . Weak hierarchy ensures that each edge is incident to at least one black dot.*

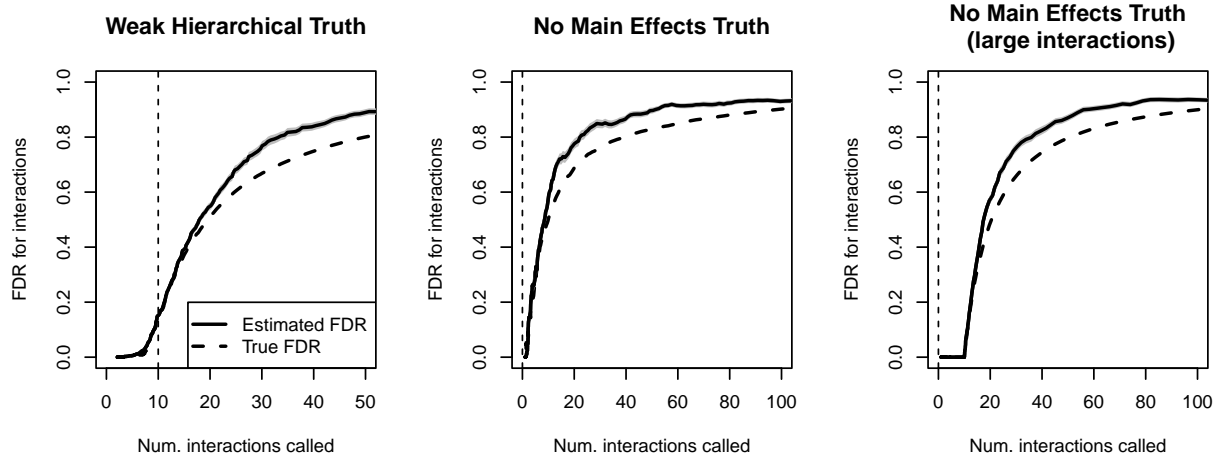


Figure 6: *Estimation of FDR for convex hierarchical testing using permutations. Result is an average over 50 simulations (with one standard error bars shown as well). Vertical line is drawn at true number of non-zero interactions.*

of the interaction contrast to group 1 or group 2, by flipping the sign of the component at random.

This scheme can be easily adapted to the present setting. Here are the details. The idea is to retain the main effect contrasts w_j from the original fit and to create randomized versions of the interactions. In particular, let $f : \mathbb{R}^n \rightarrow \mathbb{R}^{p^2-p}$ represent the function of the class labels such that $z_{jk} = f(y)_{jk}$. For $b = 1, \dots, B$, we get random permutations $y^{*(b)} \in \mathbb{R}^n$ of y and compute $z^{*(b)} = f(y^{*(b)})$.

Using (5), we get $\hat{\lambda}_{jk}^{*'(b)}$ based on $(w, z^{*(b)})$. Finally, we estimate the FDR as

$$\widehat{FDR}(\lambda) = \min \left\{ \frac{\frac{1}{B} \sum_{j,k,b} I(\hat{\lambda}_{jk}^{*'(b)} > \lambda)}{\sum_{jk} I(\hat{\lambda}_{jk}' > \lambda)}, 1 \right\} \quad (7)$$

Note that this estimate of FDR pools the null distributions from all jk pairs. This kind of pooled null distribution is commonly used, for example in the SAM procedure (Tusher, Tibshirani & Chu 2001) and the aforementioned interaction test of Simon & Tibshirani (2012). Its accuracy is quite high in simulation studies, although we know of no rigorous results on its asymptotic properties.

Figure 6 shows the estimated FDR from this method for three of the scenarios described earlier. We observe that the estimate is fairly accurate, especially when the number of interactions called is small, but tends to over-estimate the true FDR by a moderate amount for larger numbers of interactions called. This may be due to the interdependence of the test statistics $\hat{\lambda}_{jk}$ for each j . Over-estimation of the FDR corresponds to being conservative, which is less of a concern than under-estimation.

In future work, it would be important to study the theoretical properties of this permutation estimate.

7 Discussion

We have proposed a hierarchical method for large-scale interaction testing that biases its search towards interactions exhibiting at least one moderate main effect. Our testing procedure is defined in terms of a convex optimization problem but can be expressed in closed-form. Examination of the form of the statistic shows that it incorporates hierarchy in a gentler way than two-step procedures that screen out interactions based on main effects. This distinction allows it to include large interactions even when hierarchy is violated (as seen in simulation).

This work could be generalized in several ways. We have focused exclusively on pairwise interactions: extensions to k -way interactions, for $k > 2$ would bound the sum of such interactions by the size of the $k - 1$ order effect. With appropriate definitions for the interaction components, z_{jk} , one could also apply this procedure to interaction testing for the proportional hazards model in survival analysis. More generally, the idea of formulating a test statistic based on the knots of a convex optimization problems solution path may be of interest in contexts beyond testing interactions.

Acknowledgments. Robert Tibshirani was supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

References

- Bien, J., Taylor, J. & Tibshirani, R. (2013), ‘A lasso for hierarchical interactions’, *Annals of Statistics* **42**(3), 1111–1141.
- Buzkov, P., Lumley, T. & Rice, K. (2011), ‘Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions’, *Ann. Hum. Genet.* **1**, 36–45.
- Dudoit, S. & van der Laan, M. J. (2008), *Multiple Testing Procedures with Applications to Genomics*, Springer.
- Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Institute of Mathematical Statistics Monographs, Cambridge University Press.
- Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U. & Kooperberg, C. (2012), ‘Powerful cocktail methods for detecting genome-wide gene-environment interaction’, *Genetic Epidemiology* **36**(3), 183–194.
URL: <http://dx.doi.org/10.1002/gepi.21610>
- Huang, J., Lin, A., Narasimhan, B., Quertermous, T., Hsiung, C. A., Ho, L.-T., Grove, J. S., Olivier, M., Ranade, K., Risch, N. J. & Olshen, R. A. (2004), ‘Tree-structured supervised learning and the genetics of hypertension’, *Proceedings of the National Academy of Sciences* **101**, 10529–10534.

- Kooperberg, C. & LeBlanc, M. (2008), ‘Increasing the power of identifying gene-gene interactions in genome-wide association studies’, *Genetic Epidemiology* **32**(3), 255–263.
URL: <http://dx.doi.org/10.1002/gepi.20300>
- Park, M. Y. & Hastie, T. (2008), ‘Penalized logistic regression for detecting gene interactions’, *Biostatistics* **9**(1), 30–50.
URL: <http://biostatistics.oxfordjournals.org/content/9/1/30.abstract>
- Simon, N. & Tibshirani, R. (2012), A permutation approach to testing interactions in many dimensions, Technical report, Stanford University.
- Tusher, V., Tibshirani, R. & Chu, G. (2001), ‘Significance analysis of microarrays applied to transcriptional responses to ionizing radiation’, *Proc. Natl. Acad. Sci. USA*. **98**, 5116–5121.
- Wu, Z. & Zhao, H. (2009), ‘Statistical power of model selection strategies for genome-wide association studies’, *PLOS Genetics* **5**, 1–14.