Sensitivity analysis for the cross-match test, with applications in genomics

Ruth Heller, Shane T. Jensen, Paul R. Rosenbaum, Dylan S. Small[1]

Technion and the University of Pennsylvania

Abstract. The cross-match test is an exact, distribution free test of no treatment effect on a high dimensional outcome in a randomized experiment. The test uses optimal nonbipartite matching to pair $2I$ subjects into I pairs based on similar outcomes, and the cross-match statistic A is the number of times a treated subject was paired with a control, rejecting for small values of A. If the test is applied in an observational study in which treatments are not randomly assigned, it may be comparing treated and control subjects who are not comparable, and may therefore falsely reject a true null hypothesis of no treatment effect. We develop a sensitivity analysis for the cross-match test, and apply it in an observational study of the effects of smoking on gene expression levels. In addition, we develop a sensitivity analysis for several multiple testing procedures using the cross-match test and apply it to 1627 molecular function categories in Gene Ontology.

Keywords: Cross-match test; multiple testing; nonbipartite matching; observational study; sensitivity analysis.

[1] *Address for correspondence:* Dr. Ruth Heller, Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa, Israel. This work was supported by a grant SES-0849370 from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation and grant BSF 2008049 from the U.S.-Israel Binational Science Foundation. E-mail: ruheller@techunix.technion.ac.il. 14 January 2010

1 The cross-match test for a randomly assigned treatment

1.1 An observational study of the effects of smoking on gene expression levels

Does smoking cause changes in gene expression? If it does, what specific changes does it cause? Spira et al. [41] compared expression levels in human airway epithelial cells of 9968 genes in 34 current smokers and 23 never smokers. Analyses of data of this sort typically emphasize the dimensionality of the response and the associated problems of multiple testing; these are two important problems, but there are others. The treatment, here smoking, is not assigned at random to some individuals and denied to others, so smokers and nonsmokers may differ systematically in unmeasured ways that affect gene expression, so differing expression levels may not be effects caused by smoking. To what extent are conclusions sensitive to small or moderate departures from random treatment assignment? Would a high dimensional test or multiple comparison procedure reach very different conclusions if the analysis allowed for moderate departures from random assignment? We investigate this by developing a sensitivity analysis for a multivariate permutation test, the cross-match test, and for associated multiple-test procedures. In the study by Spira et al. [41], some of the changes in expression levels turn out to quite insensitive to bias from nonrandom assignment to smoking or control, but other changes are fairly sensitive.

In a randomized experiment, the cross-match test is a randomization test, and §1 applies the test to the data from Spira et al. [41] to test the null hypothesis that smoking does not affect the 9968 gene expression levels, ignoring for a moment the fact that people were not randomly assigned to smoke or not smoke. In §2, issues of

multiple testing are addressed and the cross-match test is applied to 1627 hypotheses about subsets of genes defined by Gene Ontology, continuing to ignore the absence of random assignment. Then §3 introduces a sensitivity analysis for the cross-match tests, asking about the magnitude of bias from nonrandom assignment that would need to be present to alter the conclusions reached by the randomization test. The sensitivity analysis is combined with corrections for testing many hypotheses in §4. Uses, limitations and practicalities of the cross-match test are discussed in §5.

1.2 Definition of the cross-match statistic

There are $2I$ subjects, $\ell = 1, 2, \ldots, 2I$, where subject ℓ is treated if $Z_\ell = 1$ and is a control if $Z_\ell = 0$, and there are $n = \sum_{\ell=1}^{2I} Z_\ell$ treated subjects and $2I - n$ controls in total. If subject ℓ receives the treatment, then this subject exhibits an M-dimensional response $\mathbf{y}_{T\ell}$ whereas if subject ℓ receives the control, then response $\mathbf{y}_{C\ell}$ is observed instead, so the response actually observed from subject ℓ is $\mathbf{Y}_\ell = Z_\ell\, \mathbf{y}_{T\ell} + (1 - Z_\ell)\, \mathbf{y}_{C\ell}$ and the effect of the treatment on ℓ, namely $\mathbf{y}_{T\ell} - \mathbf{y}_{C\ell}$, is not observed for any subject ℓ; see Neyman [26] and Rubin [39]. Write $\mathcal{F} = \{(\mathbf{y}_{T\ell}, \mathbf{y}_{C\ell}),\ \ell = 1, 2, \ldots, 2I\}$. Fisher's [10] sharp null hypothesis H_0 of no treatment effect says $H_0 : \mathbf{y}_{T\ell} = \mathbf{y}_{C\ell}$ for $\ell = 1, 2, \ldots, 2I$.

The cross-match test [35] is performed as follows. A $2I \times 2I$ symmetric distance matrix is defined, with row k and column ℓ giving a 'distance' between $\mathbf{Y}_k$ and $\mathbf{Y}_\ell$. The $2I$ subjects are then paired into I non-overlapping pairs to minimize the total of the I distances within pairs. For notational convenience, the subjects are renumbered, $j = 1, \ldots, 2I$ so that subject $2i - 1$ and $2i$ are paired for $i = 1, \ldots, I$.

The cross-match statistic A is the number of pairs containing a treated subject and a control, that is:

$$A = \sum_{i=1}^{I} Z_{2i-1}\ (1 - Z_{2i}) + (1 - Z_{2i-1})\ Z_{2i}. \tag{1}$$

A small value of A suggests that the distribution of $\mathbf{Y}_\ell$ is different for treated and control subjects [35].

The optimal pairing of $2I$ subjects into I pairs to minimize the total distance inside pairs is an 'optimal nonbipartite matching;' see [4, 27] for a textbook discussion, [8] for an algorithm with Fortran code, [5] for a literature review and C code, and [14, 20, 21, 22] for several applications of nonbipartite matching in statistics. In particular, Lu, et al. [22, 23] have made Derigs' [8] Fortran code available from inside R.

If there is an odd number, $2I + 1$, of subjects, then a pseudo-subject is added to the distance matrix at zero distance from everyone else, $I + 1$ pairs are formed as above, and the pair containing the pseudo-subject is discarded. In this way, the least matchable subject is the discarded subject.

1.3 Example of computing the cross-match statistic

In the study by Spira et al. [41], $\mathbf{Y}_\ell$ is the 9968-dimensional vector of logarithms of expression levels. The distance matrix is the 57×57 matrix of Euclidean distances among the $\mathbf{Y}_\ell$. Because $34 + 23 = 57$ is odd, a pseudo-subject is added at zero distance from all 57 subjects, as discussed in §1.2, making a 58×58 matrix. The 58 subjects are paired to minimize the total distance within the 58 pairs, and the

pair containing a pseudo-subject is discarded; in this case, the discarded subject is a smoker. Then there are $2I = 56$ subjects, $n = 33$ of whom are smokers, in $I = 28$ pairs.

Figure 1 depicts the calculations with the aid of a multidimensional scaling that plays no role in the test itself but is helpful in seeing what is happening. For the $2I = 56$ paired individuals, the 56×56 distance matrix was used in Kruskal's non-metric multidimensional scaling algorithm (isoMDS in the MASS package in R with two dimensions and the default settings). Paired points are connected by a line. The left-most pair is a cross-match, pairing a smoker with a nonsmoker. There are $A = 5$ cross-matches and $I - A = 28 - 5 = 23$ matches that are not cross-matches.

With expression levels, $\mathbf{Y}_\ell$ has numeric coordinates, but this is not an essential feature of the cross-match test. Instead, $\mathbf{Y}_\ell$ might be a 'word' consisting of a sequence of 'letters,' such as a DNA base sequence, with a suitable distance defined between different 'words'. Alternatively, $\mathbf{Y}_\ell$ might record both numeric intensities and geometric locations of those intensities, as in fMRI brain imaging, where two individuals i and ℓ are close if they have similar intensities at neighboring locations. Instead, $\mathbf{Y}_\ell$ might record the dates and locations of the international travel of person ℓ, where two people i and ℓ are close if they were often in the same locations on the same dates.

1.4 Null distribution of the cross-match statistic

Write $\mathbf{Z} = (Z_1, \ldots, Z_{2I})^T$ where subject $2i-1$ is paired with subject $2i$, $i = 1, \ldots, I$. Write $|S|$ for the number of elements in a finite set S. In a randomized experiment,

n of the $2I$ subjects would be picked at random for treatment, so there are $\binom{2I}{n}$ possible values, $\mathbf{z}$, of $\mathbf{Z}$, namely the values in the set $\mathcal{Z}$,

$$\mathcal{Z} = \left\{ \mathbf{z} = (z_1, \ldots, z_{2I})^T : \sum_{j=1}^{2I} z_j = n,\ z_j \in \{0, 1\},\ j = 1, \ldots, 2I \right\},$$

so $|\mathcal{Z}| = \binom{2I}{n}$. To say that $\mathbf{Z}$ is picked at random from $\mathcal{Z}$ is to say that

$$\Pr\left(\mathbf{Z} = \mathbf{z} \,\middle|\, \sum_{j=1}^{2I} Z_j = n,\ \mathcal{F} \right) = \frac{1}{|\mathcal{Z}|} = \frac{1}{\binom{2I}{n}} \text{ for each } \mathbf{z} \in \mathcal{Z}. \tag{2}$$

If Fisher's sharp null hypothesis of no treatment effect, $H_0 : \mathbf{y}_{T\ell} = \mathbf{y}_{C\ell}$ for $\ell = 1$, 2, ..., $2I$, were true, then $\mathbf{Y}_\ell = Z_\ell \mathbf{y}_{T\ell} + (1 - Z_\ell)\, \mathbf{y}_{C\ell} = \mathbf{y}_{C\ell}$ is a function of $\mathcal{F}$, so the matching is a function of $\mathcal{F}$, and the randomization (2) determines the exact null distribution of the cross-match statistic, A in (1). Alternatively, the same null distribution of A may be obtained from the null hypothesis that the $\mathbf{Y}_\ell$ are independent and identically distributed independent of $\mathbf{Z}$; see [35].

The null distribution $\Pr(A = a \mid \mathcal{F})$ has a simple form. We must first determine the support of this distribution. Write $\mathcal{A}_{n,I}$ for the possible values of A with n treated subjects and $2I - n$ controls. Clearly $A \leq \min(n, 2I - n)$, and $A = \min(n, 2I - n)$ is possible. If there were $a < \min(n, 2I - n)$ cross-matches, then there must be a pair i with $Z_{2i-1} + Z_{2i} = 2$ and a pair i' with $Z_{2i'-1} + Z_{2i'} = 0$; swapping Z_{2i} and $Z_{2i'}$ increases the number of cross-matches by 2. If n is odd, then there must be at least one cross-match, but if n is even, there can be 0 cross-matches. If n is even and $n \leq I$, then $\mathcal{A}_{n,I} = \{0, 2, 4, \ldots, n\}$, whereas if n is odd and $n \leq I$, then $\mathcal{A}_{n,I} = \{1, 3, 5, \ldots, n\}$. If $n > I$ and n is even, then $\mathcal{A}_{n,I} = \{0, 2, 4, \ldots, 2I - n\}$,

Table 1: Exact null randomization distribution of the cross-match statistic, A, for $2I = 56$ subjects in $I = 28$ pairs with $n = 33$ treated subjects.

a	$\Pr(A = a)$	$Pr(A \leq a)$
1	0.00000023	0.00000023
3	0.00002705	0.00002728
5	0.00081143	0.00083871
7	0.00973713	0.01057583
9	0.05625895	0.06683478
11	0.17184552	0.23868030
13	0.29081550	0.52949580
15	0.27696714	0.80646294
17	0.14662966	0.95309261
19	0.04115920	0.99425181
21	0.00548789	0.99973970
23	0.00026030	1.00000000

whereas if $n > I$ and n is odd then $\mathcal{A}_{n,I} = \{1, 3, 5, \ldots, 2I - n\}$.

If there are $a \in \mathcal{A}_{n,I}$ cross-matched pairs with $Z_{2i-1} + Z_{2i} = 1$, then there are $(n-a)/2$ pairs with $Z_{2i-1} + Z_{2i} = 2$, and $I - a - (n-a)/2 = I - (n+a)/2$ pairs with $Z_{2i-1} + Z_{2i} = 0$, making a total of $a + (n-a)/2 + I - (n+a)/2 = I$ pairs with $\sum_{j=1}^{2I} Z_j = a + 2(n-a)/2 = n$ treated subjects. Under the null hypothesis, the $\binom{2I}{n}$ values of $\mathbf{z} \in \mathcal{Z}$ are equally probable, so

$$\Pr(A = a \mid \mathcal{F}) = \kappa(a, n, I) = \begin{cases} \frac{2^a\, I!}{\binom{2I}{n} a! \left(\frac{n-a}{2}\right)! \left(I - \frac{n+a}{2}\right)!} & \text{for } a \in \mathcal{A}_{n,I} \\ 0 & \text{for } a \notin \mathcal{A}_{n,I} \end{cases} . \qquad (3)$$

Table 1 gives the randomization distribution of A for $2I = 56$ subjects in $I = 28$ pairs with $n = 33$ treated subjects and $2I - n = 23$ controls. If the study by Spria et al. [41] had been a randomized experiment, with individuals randomly assigned to their roles as smokers or never smokers, and if smoking did not affect expression

levels, the chance of $A = 5$ or fewer cross-matches is 0.000839, so the null hypothesis would be rejected at the conventional 0.05 level.

2 Testing multiple hypotheses of no treatment effect

When Fisher's sharp null hypothesis of no treatment effect, $H_0 : \mathbf{y}_{T\ell} = \mathbf{y}_{C\ell}$ for $\ell = 1$, 2, ..., $2I$, is rejected we will often wish to ask which coordinates of $\mathbf{Y}_\ell$ are affected. Let $\mathbf{s}$ be an M-dimensional vector of 0's and 1's with at least one 1, and let $\mathbf{Y}_\ell(\mathbf{s})$, $\mathbf{y}_{T\ell}(\mathbf{s})$, and $\mathbf{y}_{C\ell}(\mathbf{s})$ be the sub-vectors of, respectively, $\mathbf{Y}_\ell$, $\mathbf{y}_{T\ell}$, and $\mathbf{y}_{C\ell}$ of dimension $s_+ = \sum_{m=1}^{M} s_m$ containing the coordinates for which $s_m = 1$. The hypothesis $H_\mathbf{s}$ asserts that the treatment does not affect these s_+ coordinates, $H_\mathbf{s} : \mathbf{y}_{T\ell}(\mathbf{s}) = \mathbf{y}_{C\ell}(\mathbf{s})$ for $\ell = 1, 2, \ldots, 2I$. Apply the cross-match test to $\mathbf{Y}_\ell(\mathbf{s})$, count the number of cross-matches, $a(\mathbf{s})$, and let $p(\mathbf{s})$ be the resulting P-value computed as in §1.4. In a randomized experiment, each such P-value is a valid test of its null hypothesis, so $\Pr\{p(\mathbf{s}) \le \alpha\} \le \alpha$ if $H_\mathbf{s}$ is true.

There are $2^M - 1$ hypotheses $H_\mathbf{s}$, so one cannot test them all and reject whenever $p(\mathbf{s}) \le 0.05$, because this would lead to a large number of false rejections. There are many possible strategies; e.g., [9].

Bonferroni inequality. A simple familiar strategy is to test all $2^M - 1$ hypotheses, rejecting all hypotheses $H_\mathbf{s}$ with $p(\mathbf{s}) \le \alpha / \left(2^M - 1\right)$. Under this strategy, the probability of falsely rejecting at least one true hypothesis (i.e., the family wise error rate or FWER) is at most α, and the expected number of false rejections is α. In many contexts, this strategy will be quite conservative.

Holm's procedure. Holm's [17] procedure involves a few more steps, but it also

falsely rejects at least one true hypothesis with probability at most α, thereby controlling the FWER. It is less conservative than the Bonferroni procedure.

Closed testing. In closed testing [24], one would follow the approach in [19], rejecting $H_{\mathbf{s}}$ at level α if $p(\mathbf{s}') \leq \alpha$ for all $\mathbf{s}'$ such that $s_m = 1$ implies $s'_m = 1$ for all m. An advantage of this procedure is that all tests are done at level α, and yet the probability of falsely rejecting at least one true hypothesis is at most α. The procedure tends to be impractical for large M, but it is practical when M is small, or when M itself is large but a suitably restricted subset of hypotheses $H_{\mathbf{s}}$ is tested.

Benjamini-Hochberg procedure. The method of Benjamini and Hochberg [1] has been shown to control the false discovery rate (FDR), or the proportion of rejections that are false rejections, when the $p(\mathbf{s})$'s are independent and under certain other conditions. In these circumstances, the Benjamini-Hochberg procedure's more lenient standard typically rejects many more hypotheses than the Holm procedure. The Benjamini-Hochberg procedure appears to control the false discovery rate in most circumstances that are not highly artificial [30, 44], but artificial exceptions are known to exist [13]; see also [40]. The $p(\mathbf{s})$'s produced by the cross-match test are not independent, so use of the Benjamini-Hochberg procedure may be reasonable but is not formally known to control the false discovery rate.

Complementary partitions. Suppose that the M coordinates of $\mathbf{Y}_\ell$ can be partitioned into $\widetilde{M} \leq M$ mutually exclusive sets of coordinates, ordered by priority, where hypothesis $\widetilde{H}^{(1)}$ asserts that set 1 is unaffected, $\overline{H}^{(1)}$ asserts that the union of the remaining $\widetilde{M} - 1$ sets, 2, 3, ..., $\widetilde{M}$ is unaffected, $\widetilde{H}^{(2)}$ asserts that set 2 is unaffected, $\overline{H}^{(2)}$ asserts that the union of the remaining $\widetilde{M} - 2$ sets, 3, 4, ..., $\widetilde{M}$

is unaffected, and so on. Notice that, for the last hypothesis, $\overline{H}^{\left(\widetilde{M}-1\right)} = \widetilde{H}^{\left(\widetilde{M}\right)}$. For instance, with $\widetilde{M} = 2$, $\widetilde{H}^{(1)}$ might refer to the expression levels of all known oncogenes, and $\widetilde{H}^{(2)}$ might refer to all other genes. Let $p^{(0)}$ be the P-value from the test of no effect on $\mathbf{Y}_\ell$ from §1.4, and let $\widetilde{p}^{(k)}$ and $\overline{p}^{(k)}$ be the P-values when the cross-match test is used to test $\widetilde{H}^{(k)}$ and $\overline{H}^{(k)}$, respectively. Test the hypothesis of no effect, H_0, as done in §1.4 rejecting if $p^{(0)} \leq \alpha$; if H_0 is rejected, test both $\widetilde{H}^{(1)}$ and $\overline{H}^{(1)}$ rejecting $\widetilde{H}^{(1)}$ if $\widetilde{p}^{(1)} \leq \alpha$, rejecting $\overline{H}^{(1)}$ if $\overline{p}^{(1)} \leq \alpha$; ... if both $\widetilde{p}^{(k)} \leq \alpha$ and $\overline{p}^{(k)} \leq \alpha$, then test both $\widetilde{H}^{(k+1)}$ and $\overline{H}^{(k+1)}$, rejecting $\widetilde{H}^{(k+1)}$ if $\widetilde{p}^{(k+1)} \leq \alpha$, rejecting $\overline{H}^{(k+1)}$ if $\overline{p}^{(k+1)} \leq \alpha$; .... As discussed in [36, Proposition 3], the chance that this procedure tests and rejects at least one true hypothesis is at most α because the hypotheses

$$\left\langle H_0, \left\{\widetilde{H}^{(1)}, \overline{H}^{(1)}\right\}, \left\{\widetilde{H}^{(2)}, \overline{H}^{(2)}\right\}, \ldots, \left\{\widetilde{H}^{\left(\widetilde{M}-1\right)}, \overline{H}^{\left(\widetilde{M}-1\right)}\right\}\right\rangle$$

form a sequentially exclusive sequence of hypotheses.

As this incomplete list of multiple testing procedures suggests, there is often an advantage in lending some priority or structure to the $2^M - 1$ possible hypotheses. For instance, in genomics, the molecular function categories within Gene Ontology [11] provide one possible approach to (i) limiting the number of hypotheses, or (ii) organizing the hypotheses.

2.1 Application to the genomics study of effects of smoking

We used the 1627 molecular function categories within Gene Ontology [11] that contain at least 2 probe sets, to identify the functional categories where the smoking has an effect on the expression profile. That is, we did not use all $2^{9968}-1$ hypotheses, but rather the 1627 hypotheses $H_{\mathbf{s}}$ where the binary vector $\mathbf{s}$ picked out the genes in a function category.

We applied the Holm and Benjamini-Hochberg procedures with $\alpha = 0.05$ to the 1627 P-values, $p(\mathbf{s})$, from the cross-match test. Using the Holm procedure, 30 hypotheses were rejected, corresponding to the functional categories where at most 3 cross-matches were observed. Using the Benjamini-Hochberg procedure, 83 hypotheses were rejected, corresponding to the functional categories where at most 5 cross-matches were observed. Figure 2 displays the sorted P-values, as well as the adjusted P-values from the Holm and the Benjamini-Hochberg procedures. The appearance of Figure 2 reflects the discrete nature of the statistic A in (1). Here, the adjusted P-values for a hypothesis H_i is the smallest nominal level of the multiple testing procedure at which H_i would be rejected, given the value of all test statistics involved; see [43].

The analyses just presented acted as if the study by Spira et al. [41] had been a randomized experiment, with individuals randomly assigned to their roles as smokers or never smokers. Of course, individuals are not randomly assigned to smoke or not; indeed, smokers and nonsmokers differ in various ways. Could the significant differences in gene expression found above be due to small biases from nonrandom treatment assignment? Or would it take very large departures from random assign-

ment to produce these differences?

3 Sensitivity analysis for the cross-match test

3.1 Sensitivity to nonrandom treatment assignment

In point of fact, subjects were not randomly assigned to their roles as smokers and never smokers, so the randomization distribution in (2) that would be applicable in a randomized experiment is not applicable in the study by Spira, et al. [41]. What magnitude of departure from random assignment in (2) would need to be present to alter the conclusion that smoking causes changes in expression levels in human airway epithelial cells?

The sensitivity model [31, 33, 34] builds a family of distributions on $\mathcal{Z}$ in two steps: first, the treatment assignments, Z_j, given $\mathcal{F}$ are independent with unknown probabilities,

$$\Pr\left(Z_j = 1 \mid \mathcal{F}\right) = \pi_j;$$

then, the distribution of $\mathbf{Z}$ is returned to $\mathcal{Z}$ by conditioning on $\sum_{j=1}^{2I} Z_j = n$,

$$\Pr\left(\mathbf{Z} = \mathbf{z} \,\middle|\, \mathcal{F},\ \sum_{j=1}^{2I} Z_j = n\right) = \frac{\prod_{j=1}^{2I} \pi_j^{z_j} \left(1-\pi_j\right)^{1-z_j}}{\sum_{\mathbf{b}\in\mathcal{Z}} \prod_{j=1}^{2I} \pi_j^{b_j} \left(1-\pi_j\right)^{1-b_j}} \text{ for } \mathbf{z} \in \mathcal{Z}. \qquad (4)$$

Following in the spirit of [6], the magnitude of the departure from random assignment is measured by a parameter, $\Gamma \geq 1$, such that two subjects may differ in their odds

of treatment by at most a factor of Γ:

$$\frac{1}{\Gamma} \leq \frac{\pi_j (1-\pi_k)}{\pi_k (1-\pi_j)} \leq \Gamma, \quad \forall j,k. \tag{5}$$

If $\Gamma = 1$, then $\pi_j = \pi_k \; \forall j,k$, and (4) equals the randomization distribution (2). For fixed $\Gamma > 1$, the distribution (4) is unknown but deviates from random assignment by a bounded magnitude. A sensitivity analysis considers, for several values of $\Gamma \geq 1$, the range of possible inferences, say the interval of possible significance levels.

The model (4) may be rewritten in terms of a logit model involving an unmeasured covariate u_j with $u_j \in [0,1] \; \forall j$; specifically, set $\gamma = \log(\Gamma) \geq 0$,

$$\pi_j = \frac{\exp(\alpha + \gamma u_j)}{1 + \exp(\alpha + \gamma u_j)}$$

so that

$$\Pr\left(\mathbf{Z} = \mathbf{z} \;\middle|\; \mathcal{F}, \sum_{j=1}^{2I} Z_j = n\right) = \frac{\exp\left(\gamma \mathbf{z}^T \mathbf{u}\right)}{\sum_{\mathbf{b} \in \mathcal{Z}} \exp\left(\gamma \mathbf{b}^T \mathbf{u}\right)}, \quad \mathbf{u} \in [0,1]^{2I}. \tag{6}$$

(To see that this representation is always possible, set $\alpha = \min_j \log\{\pi_j/(1-\pi_j)\}$ and $u_j = [\log\{\pi_j/(1-\pi_j)\} - \alpha]/\gamma$ for $\gamma > 0$ or $u_j = 0$ for $\gamma = 0$; then the odds ratio in (5), namely $\pi_j(1-\pi_k)/\{\pi_k(1-\pi_j)\}$, becomes $e^{-\gamma} \leq \exp\{\gamma(u_j - u_k)\} \leq e^{\gamma}$ for $\forall j,k$ implying $u_j \in [0,1]$.)

3.2 Bounds on the significance level for fixed Γ

For fixed $\mathbf{u} \in [0,1]^{2I}$, the distribution of the cross-match statistic under model (6) and the null hypothesis H_0 of no effect is

$$\Pr\left(h(\mathbf{Z}) \leq a \;\middle|\; \mathcal{F},\ \sum_{j=1}^{2I} Z_j = n\right) = \frac{\sum_{\mathbf{z}\in\mathcal{Z}} \chi\left\{h(\mathbf{z}) \leq a\right\} \exp\left(\gamma \mathbf{z}^T \mathbf{u}\right)}{\sum_{\mathbf{b}\in\mathcal{Z}} \exp\left(\gamma \mathbf{b}^T \mathbf{u}\right)} \tag{7}$$

where

$$h(\mathbf{z}) = \sum_{i=1}^{I} z_{2i-1}\ (1 - z_{2i}) + (1 - z_{2i-1})\ z_{2i} \tag{8}$$

and $\chi(E) = 1$ if event E occurs and $\chi(E) = 0$ otherwise. Of course, (7) is unknown because $\mathbf{u}$ is unknown. For each fixed $\Gamma \geq 1$, the following proposition places an upper bound on (7) and hence an upper bound on the P-value from the cross-match statistic. Proposition 1 is proved in the appendix.

Proposition 1 *For fixed $\gamma = \log(\Gamma) \geq 0$, the probability (7) is maximized for $\mathbf{u} \in [0,1]^{2I}$ by a vector $\mathbf{u}$ with $u_j = 0$ or $u_j = 1$ for every j, and with $u_{2i-1} = u_{2i}$ for at least $I - 1$ pairs.*

In Proposition 1, the fewest cross-matches occur for a $\mathbf{u}$ such that at least $I - 1$ pairs have $u_{2i-1} = u_{2i}$, that is, paired subjects have the same u_j. Because $h(\mathbf{z})$ in (8) is symmetrical in the I pairs, the bound on (7) may be obtained at a $\mathbf{u}$ with $u_j = 0$ or $u_j = 1$ for all j and $u_1 \leq u_2 \leq \cdots \leq u_{2I}$, so the number of candidate $\mathbf{u}$'s is of order $O(I)$. Proposition 2 in the next section gives a practical method for computing the probability (7).

3.3 Sensitivity Distribution of the Cross-match Statistic

In light of Proposition 1, we evaluate (7) with $u_{2i-1} = u_{2i}$ for all I pairs; a single pair has negligible effect on (7) for moderate I. The following proposition gives an explicit form for the bounding distribution.

Proposition 2 *Suppose that*

$$\pi_j = \frac{e^{\alpha+\gamma}}{1+e^{\alpha+\gamma}},\ j = 1, \ldots, 2m, \quad \pi_j = \frac{e^{\alpha}}{1+e^{\alpha}} \text{ for } j = 2m+1, \ldots 2I.$$

Then for $a \in \mathcal{A}_{n,I}$

$$\begin{aligned} & \Pr\left(h\left(\mathbf{Z}\right) = a \,\middle|\, \mathcal{F},\ \sum_{j=1}^{2I} Z_j = n \right) \\ = & \sum_{k=\max(0,n+2m-2I)}^{\min(2m,n)} \frac{\binom{2m}{k}\binom{2I-2m}{n-k}\exp\left(\gamma k\right)}{\sum_{\ell=\max(0,n+2m-2I)}^{\min(2m,n)}\binom{2m}{\ell}\binom{2I-2m}{n-\ell}\exp\left(\gamma \ell\right)} \\ & \times \sum_{b \in \mathcal{A}_{k,m}} \kappa\left(b,k,m\right)\ \kappa\left(a-b, n-k, I-m\right) \end{aligned}$$

where $\kappa\left(\cdot,\cdot,\cdot\right)$ *is defined in (3).*

Proof. Before conditioning on $\sum_{j=1}^{2I} Z_j = n$, the quantity $\sum_{j=1}^{2m} Z_j$ is the number of treated subjects among the m pairs with $\pi_j = e^{\alpha+\gamma}/\left(1+e^{\alpha+\gamma}\right)$, so $\sum_{j=1}^{2m} Z_j$ is binomial with $2m$ trials and probability of success $e^{\alpha+\gamma}/\left(1+e^{\alpha+\gamma}\right)$; similarly, $\sum_{j=2m+1}^{2I} Z_j$ is an independent binomial with $2I-2m$ trials and probability of success $e^{\alpha}/\left(1+e^{\alpha}\right)$. Then the conditional probability is given by the extended hypergeometric distribu-

tion,

$$\Pr\left(\sum_{j=1}^{2m} Z_j = k \,\middle|\, \sum_{j=1}^{2I} Z_j = n\right) = \frac{\binom{2m}{k}\binom{2I-2m}{n-k}\exp(\gamma k)}{\sum_{\ell=\max(0,n+2m-2I)}^{\min(2m,n)}\binom{2m}{\ell}\binom{2I-2m}{n-\ell}\exp(\gamma \ell)}.$$

Conditionally, given $(\sum_{j=1}^{2m} Z_j = k,\ \sum_{j=1}^{2I} Z_j = n)$, the $\binom{2m}{k}$ possible values of $(Z_1, \ldots, Z_{2m})$ are equally probable, so the conditional probability of b cross-matches in the first m pairs is $\kappa(b,k,m)$ for $b \in \mathcal{A}_{k,m}$. In parallel, conditional on $(\sum_{j=1}^{2m} Z_j = k,\ \sum_{j=1}^{2I} Z_j = n)$, the $\binom{2I-2m}{n-k}$ possible values of $(Z_{2m+1}, \ldots, Z_{2I})$ are equally probable, so the chance of $a-b$ cross-matches is $\kappa(a-b, n-k, I-m)$. Moreover, these two events are conditionally independent. Therefore, conditional on $(\sum_{j=1}^{2m} Z_j = k,\ \sum_{j=1}^{2I} Z_j = n)$, the chance of $a \in \mathcal{A}_{n,I}$ cross-matches is

$$\sum_{b \in \mathcal{A}_{k,m}} \kappa(b,k,m)\ \kappa(a-b, n-k, I-m),$$

proving the proposition. ■

3.4 Application to the genomics study of effects of smoking

Table 2 presents the sensitivity analysis. The table gives the upper bound on the P-value for a bias of size Γ when, as in §1.3, there are $A = 5$ cross-matches in a study of this size. Again, the parameter Γ measures the magnitude of the departure from random assignment. A bias of magnitude $\Gamma = 10$ is enormous: two subjects may differ in their odds of smoking by a factor of 10 — one may be ten times more likely to smoke than the other because of an unmeasured covariate with very strong association with gene expression levels. At the conventional 0.05 level, the null

Table 2: Sensitivity analysis for the cross-match test when applied to all 9968 expression levels. The table shows the maximum possible P-value from the cross-match test for departures from random assignment of various magnitudes, Γ.

Γ	1	2	5	8	10
$\max_{\mathbf{u}} \Pr(A_1 \leq 5)$	0.00084	0.00142	0.00931	0.02877	0.04799

hypothesis would be rejected even if the bias Γ was of size 10.

For comparison, one of the least sensitive conclusions from an observational study is that heavy cigarette smoking is a cause of lung cancer. Hammond's [15] study of smoking and lung cancer, for instance, becomes sensitive at about $\Gamma = 6$; see [34, §4]. Moreover, this is true despite the smaller sample size and many outcomes in the study by Spira, et al. [41]. Table 2 exhibits far less sensitivity to unmeasured bias: much more bias would be needed to explain the results found by Spira, et al. [41] than the results found by Hammond [15], even though Hammond's study is insensitive to large unmeasured biases. In thinking about this, one should keep in mind that Hammond [15] matched for many covariates, while Table 2 compares unmatched groups, so larger biases may be plausible in Table 2; see Heller et al. [16] for discussion of matching in genomics.

In a nonrandomized study of treatment effects, if a conclusion is sensitive to small departures from random assignment, for instance $\Gamma = 1.1$, then the conclusion should not be dismissed but should be viewed with greater caution. See Rosenbaum (2002, 2010) for discussion with numerous examples.

4 Sensitivity analysis for testing multiple hypotheses

The method illustrated in §3.4 suffices to examine sensitivity to bias in testing one hypothesis. We now turn to the issues that arise when, as in §2, multiple hypotheses are tested. Many of these issues are discussed in [37], and so are only sketched here.

For each hypotheses $H_{\mathbf{s}}$ in §2, for each specific value of $\Gamma \geq 1$, and for each value of the unobserved covariate $\mathbf{u} \in [0,1]^{2I}$, there is a P-value, say $p_{\Gamma,\mathbf{u}}(\mathbf{s})$, from the cross-match test, and the computations in §3 provide a sharp upper bound, say $\overline{p}_{\Gamma}(\mathbf{s})$, on $p_{\Gamma,\mathbf{u}}(\mathbf{s})$, so $p_{\Gamma,\mathbf{u}}(\mathbf{s}) \leq \overline{p}_{\Gamma}(\mathbf{s})$ for all $\mathbf{u} \in [0,1]^{2I}$.

In principle, there is one true value of the unobserved covariate, $\mathbf{u}$, and we would like to use the corresponding $p_{\Gamma,\mathbf{u}}(\mathbf{s})$ in a multiple testing procedure, perhaps one of the procedures in §2. We cannot do this because we do not know $\mathbf{u}$.

All of the procedures in §2 are monotone in the $2^M - 1$ possible P-values: if $H_{\mathbf{s}}$ is not rejected by a given set of P-values, then making some of the P-values larger while making none of them smaller will not lead to rejection of $H_{\mathbf{s}}$. It follows that if $H_{\mathbf{s}}$ is rejected by using $\overline{p}_{\Gamma}(\mathbf{s})$ in place of $p_{\Gamma,\mathbf{u}}(\mathbf{s})$, then it would also be rejected by the correct but unknown $p_{\Gamma,\mathbf{u}}(\mathbf{s})$'s.

In other words, it is safe to assume that the multiple testing procedure would reject $H_{\mathbf{s}}$ at the true $\mathbf{u}$ if it does reject $H_{\mathbf{s}}$ with the upper bounds, with $\overline{p}_{\Gamma}(\mathbf{s})$, used in place of the unknown $p_{\Gamma,\mathbf{u}}(\mathbf{s})$. Is the converse true as well? Is it safe to assume that the multiple testing procedure would accept $H_{\mathbf{s}}$ for some $\mathbf{u} \in [0,1]^{2I}$ if it accepts $H_{\mathbf{s}}$ with the upper bounds, $\overline{p}_{\Gamma}(\mathbf{s})$ used in place of the unknown $p_{\Gamma,\mathbf{u}}(\mathbf{s})$? The answer depends upon the multiple testing procedure. The issue is developed precisely and in detail in [37], so it will only be sketched briefly here.

Although each bound $p_{\Gamma,\mathbf{u}}(\mathbf{s}) \leq \overline{p}_{\Gamma}(\mathbf{s})$ is sharp, being attained for some $\mathbf{u} \in [0,1]^{2I}$, there may be no one $\mathbf{u} \in [0,1]^{2I}$ such that $p_{\Gamma,\mathbf{u}}(\mathbf{s}) = \overline{p}_{\Gamma}(\mathbf{s})$ for all $\mathbf{s}$. The unobserved covariate $\mathbf{u}$ that most disrupts the inference about $H_{\mathbf{s}}$ is unlikely to be the same as the unobserved covariate $\mathbf{u}'$ that most disrupts the inference about $H_{\mathbf{s}'}$. For instance, with just two hypotheses, $\mathbf{s}$ and $\mathbf{s}'$, one might have $p_{\Gamma,\mathbf{u}}(\mathbf{s}) = 0.025$ and $p_{\Gamma,\mathbf{u}}(\mathbf{s}') = 0.05$, so Holm's procedure would reject both hypotheses at this $\mathbf{u}$, and one might have $p_{\Gamma,\mathbf{u}'}(\mathbf{s}) = 0.05$ and $p_{\Gamma,\mathbf{u}'}(\mathbf{s}') = 0.025$, so Holm's procedure would also reject both hypotheses at this $\mathbf{u}'$, yet $\overline{p}_{\Gamma}(\mathbf{s}) \geq 0.05 = p_{\Gamma,\mathbf{u}'}(\mathbf{s})$ and $\overline{p}_{\Gamma}(\mathbf{s}') \geq 0.05 = p_{\Gamma,\mathbf{u}}(\mathbf{s}')$, so Holm's procedure rejects neither hypothesis with the upper bounds, $\overline{p}_{\Gamma}(\mathbf{s})$, used in place of the unknown $p_{\Gamma,\mathbf{u}}(\mathbf{s})$. In other words, Holm's procedure might reject $H_{\mathbf{s}}$ for a given Γ for all $\mathbf{u} \in [0,1]^{2I}$, but Holm's procedure applied to the upper bounds, $\overline{p}_{\Gamma}(\mathbf{s})$ might accept $H_{\mathbf{s}}$. Applying Holm's procedure to the bounds $\overline{p}_{\Gamma}(\mathbf{s})$ is valid but conservative: the family-wise error rate is controlled, but some hypotheses that would be rejected by checking the $\mathbf{u}$'s one at a time may not be rejected by the bounds, $\overline{p}_{\Gamma}(\mathbf{s})$.

In [37], it is shown that the situation is different for the method of complementary partitions in §2: that procedure and other instances of testing in order [36] are not conservative. That is, if $H_{\mathbf{s}}$ is rejected by the upper bounds, the $\overline{p}_{\Gamma}(\mathbf{s}')$'s, then it is rejected for every $\mathbf{u} \in [0,1]^{2I}$ and if $H_{\mathbf{s}}$ is not rejected by the upper bounds, the $\overline{p}_{\Gamma}(\mathbf{s}')$'s, then there exists a $\mathbf{u} \in [0,1]^{2I}$ for which $H_{\mathbf{s}}$ is not rejected. Certain procedures, including the ones mentioned in this paragraph, are stopped by one large P-value, and these are the procedures for which the sensitivity analysis is not conservative; see [37] for specifics.

Table 3: Sensitivity analysis with $\Gamma = 1$, $\Gamma = 5$, or $\Gamma = 10$. The case $\Gamma = 1$ is the usual randomization inference. The left side of the table indicates the number of hypotheses that were rejected at the 0.05 level by the three methods of multiple testing. The right side of the table gives the value of the cross-match statistic, A, required for rejection.

	Number of rejected null hypotheses			Value of A required for rejection		
Γ	Bonferroni	Holm	Benjamini Hochberg	Bonferroni	Holm	Benjamini Hochberg
1	30	30	83	3	3	5
5	6	6	30	1	1	3
10	0	0	6	Not possible	Not possible	1

4.1 Example of sensitivity analysis for multiple testing

Continuing the analysis from §3.4 of Spira et al.'s [41] data, we performed the sensitivity analysis for multiple testing with $\Gamma = 5$ and $\Gamma = 10$. Table 3 shows the results, including results considered previously using the randomization test for which $\Gamma = 1$. As in §3.4, the results for several molecular function categories are remarkably insensitive to unmeasured biases, comparable to the studies linking heavy smoking with lung cancer.

Table 4 displays the six least sensitive molecular function categories, with rejected null hypotheses by the Holm procedure at $\Gamma = 5$ and by the Benjamini-Hochberg procedure at $\Gamma = 10$. As indicated in Table 3, the six rejected sets are those where the observed number of cross-matches was 1, its smallest possible value in a data set with n odd. Figure 3 parallels Figure 1, but refers only to the 92-dimensional cross-match test for molecular function category GO:0016616; here, there is $A = 1$ cross-match.

How does this analysis compare to the analysis performed by Spira et al.[41]?

Table 4: The six molecular function categories identified in Table 3 with $\Gamma = 10$.

Gene Set ID	Description
GO:0004033	Aldo-keto reductase activity.
GO:0004601	Peroxidase activity.
GO:0016614	Oxidoreductase activity, acting on CH-OH group of donors.
GO:0016616	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor.
GO:0016903	Oxidoreductase activity, acting on the aldehyde or oxo group of donors.
GO:0016684	Oxidoreductase activity, acting on peroxide as acceptor.

They found 97 genes to be differentially expressed between never smokers and current smokers. A significant molecular function category in the GO ontology was then determined by overrepresentation in that category of the 97 significant genes, where the judgement of overrepresentation depended on an assumption that the genes are independent. There are several differences between the analyses. Of course, our paper has emphasized a sensitivity analysis, addressing the possibility that the division of people into smokers and nonsmokers is not random, but rather is related to unmeasured attributes of these people. In addition to this, when performing a cross-match test in a GO category, we do not assume these genes are independent. Assuming independent expression levels for genes that share a GO category is, perhaps, not the most comfortable of assumptions.

Three of the six least sensitive functional categories found by our analysis, namely GO:0004033, GO:0004601, and GO:0016616, were also determined to be significantly

over-represented by the Spira et al. analysis. Our analysis strengthens their conclusion about these three categories by adding the observation that only large biases from nonrandom treatment assignment could explain this pattern of expression levels. In agreement with Spira et al., we found that an additional category, "glucuronosyltransferase activity" category (GO:0015020), was over-expressed when judged as if from a randomized experiment ($\Gamma = 1$), but with $A = 5$ cross-matches, this finding is sensitive to biases of moderate size from nonrandom treatment assignments. Another category, "transferase activity, transferring hexosyl groups" category (GO:00016758) was found to be significantly overrepresented by Spira et al. (they quote a P-value of 0), but was not significant by our analysis, even in a randomization test ($\Gamma = 1$) because the number of cross-matches was 7. Obviously, the discrepancy here is not due to the sensitivity analysis, because it is present even in the randomization test ($\Gamma = 1$), so it reflects some difference in the judgements of the two testing procedures, possibly the reliance on independent genes in their analysis.

5 Discussion

The cross-match test judges whether treated and control groups differ on a high dimensional response $\mathbf{Y}_\ell$ by pairing individuals with similar values of $\mathbf{Y}_\ell$ and counting the number of times, A, that treated individuals are paired with controls. If A is small, then the hypothesis of no effect of the treatment on $\mathbf{Y}_\ell$ is rejected. Previous work [35] considered the behavior of the cross-match statistic A in a randomized experiment, but many applications, for instance in genomics, are not experiments, so the behavior of A may be affected by some unmeasured way that treated and

control subjects are not comparable. Spira et al.'s [41] study of gene expression levels in smokers and nonsmokers is not an experiment: people are not randomly assigned to smoke or not, and they may differ in ways that have not been recorded. Here, we have proposed a sensitivity analysis for the cross-match test, which asks about the magnitude of bias from unmeasured covariates that would need to be present to alter the rejection of the null hypothesis. In Spira et al.'s [41] study, the magnitude Γ of the departure from randomized assignment needed to alter certain conclusions is quite large, greater even than the magnitude required to alter the conclusion in [15] that heavy smoking causes lung cancer, one of the least sensitive conclusions found in an observational study. We also showed how the statistic may be used in conjunction with multiple testing procedures to isolate affected parts of $\mathbf{Y}_\ell$.

The cross-match test is an omnibus test. It is one appropriate test when the investigator does not know the nature of the effect of the treatment on the coordinates of $\mathbf{Y}_\ell$. So far as we know, it is currently the only omnibus nonparametric test for which a sensitivity analysis is available. An omnibus test should not be used if one is interested only in focused alternatives to the hypothesis of no effect, such as shifts in location. If the investigator knew, for instance, the direction of the effect for every coordinate of $\mathbf{Y}_\ell$, then multivariate tests that exploit this knowledge would have much greater power. One such test would orient the M coordinates of $\mathbf{Y}_\ell$ in the anticipated direction, calculate the M separate Wilcoxon rank sum statistics, and take the sum of these M statistics as the test statistic [32]. This is actually a univariate rank test with scores summed over the M coordinates of $\mathbf{Y}_\ell$, so the

method of sensitivity analysis for univariate rank tests in [31] may be used. Also, this test may be applied to test each subhypothesis $H_{\mathbf{s}}$ involving a subset of the coordinates of $\mathbf{Y}_\ell$, so it may be combined with multiple testing procedures along the lines illustrated here for the cross match test.

The behavior of the cross-match test is affected by the choice of distance function used to judge whether $\mathbf{Y}_i$ is close $\mathbf{Y}_\ell$. We used the Euclidean distance applied to the log of expression levels. An advantage of the Euclidean distance is that it is not estimated from the data, so the distance between $\mathbf{Y}_i$ and $\mathbf{Y}_j$ is not affected if $\mathbf{Y}_\ell$ is an outlier. Some further properties of the Euclidean distance are: (i) the distance between $\mathbf{Y}_i$ and $\mathbf{Y}_j$ may be strongly affected by a single coordinate of $\mathbf{Y}_i$ or $\mathbf{Y}_j$, (ii) the coordinates of $\mathbf{Y}_\ell$ must be in commensurate units, because they are combined without further standardization, and (iii) no account is taken of covariances among the coordinates of $\mathbf{Y}_\ell$. These properties may be judged to be advantages or disadvantages depending upon the context. The Mahalanobis distance would address (ii) and (iii), but can be strongly distorted by a single outlier and, at the least, it requires care when $2I \leq M$.

6 Appendix I: Proof of Proposition 1

In Proposition 1, the proof that (7) is maximized with $u_j = 0$ or $u_j = 1$ for every j, is exactly parallel to the proof of Proposition 2 in [31, page 495] and is omitted. So for the remainder of the proof, we assume $u_j \in \{0, 1\}$. To complete the proof, it must additionally be shown that (7) is maximized with $u_{2i-1} = u_{2i}$ for at least $I - 1$ pairs. If $\gamma = 0$, there is nothing to prove; therefore, as $\Gamma \geq 1$ and $\gamma = \log(\Gamma)$, we

may restrict attention to $\gamma > 0$.

For $i = 1, \ldots, I$, let $V_i = Z_{2i-1} + Z_{2i}$, so that $V_i \in \{0, 1, 2\}$, the V_i are independent, $V_i = 1$ for a cross match, A is the number of 1's among the V_i, and $\sum_{i=1}^{I} V_i = \sum_{j=1}^{2I} Z_j$. So $\Pr\left(A \leq a \left| \sum_{j=1}^{2I} Z_j = n \right.\right)$ in (7) equals the probability of a or fewer 1's among the V_i given $\sum_{i=1}^{I} V_i = n$.

Because conditioning on $\sum_{j=1}^{2I} Z_j = n$ eliminates α in (6), we may set α to any arbitrary number without changing the distribution on $\mathcal{Z}$. Write $\lambda = \gamma/2$. It is tidy to set $\alpha = -\gamma/2 = -\lambda$, as the interval of π_j's is then symmetric about $\frac{1}{2}$,

$$\pi_j \in \left[\frac{e^{-\gamma/2}}{1+e^{-\gamma/2}},\ \frac{e^{\gamma/2}}{1+e^{\gamma/2}}\right] = \left[\frac{e^{-\lambda}}{1+e^{-\lambda}},\ \frac{e^{\lambda}}{1+e^{\lambda}}\right] = \left[\frac{1}{1+e^{\lambda}},\ \frac{e^{\lambda}}{1+e^{\lambda}}\right].$$

In light of this and using $u_j \in \{0, 1\}$ we have

$$\pi_j \in \left\{\frac{e^{\lambda}}{1+e^{\lambda}},\ \frac{1}{1+e^{\lambda}}\right\} \text{ for every } j \tag{9}$$

with the consequence that

$$\begin{aligned}
\Pr(V_i = 1) &= \pi_{2i-1}(1-\pi_{2i}) + \pi_{2i}(1-\pi_{2i-1}) \\
&= \frac{2e^{\lambda}}{(1+e^{\lambda})^2} \text{ if } \pi_{2i} = \pi_{2i-1} \\
&= \frac{e^{2\lambda}+1}{(1+e^{\lambda})^2} \text{ if } \pi_{2i} \neq \pi_{2i-1}
\end{aligned}$$

so the unconditional probability of a cross-match, $\Pr(V_i = 1)$, is larger for $\pi_{2i} \neq \pi_{2i-1}$ than for $\pi_{2i} = \pi_{2i-1}$.

Now every π_j satisfies (9). Suppose there are two pairs, say i and k, such that

$\pi_{2i} \neq \pi_{2i-1}$ and $\pi_{2k} \neq \pi_{2k-1}$. To simplify notation without loss of generality, suppose the pairs are $i = 1$ and $k = 2$ and

$$\pi_1 = \pi_3 = \frac{e^\lambda}{1+e^\lambda}, \;\; \pi_2 = \pi_4 = \frac{1}{1+e^\lambda} \tag{10}$$

We will show that swapping π_2 and π_3 does not decrease $\Pr\left(A \leq a \left| \sum_{j=1}^{2I} Z_j = n \right.\right)$. If such swaps are pursued for as many pairs, i and k, as possible, one obtains the bounding $\mathbf{u}$ described in the statement of Proposition 1, thereby proving the result. So to complete the proof, it suffices to show that swapping π_2 and π_3 does not decrease $\Pr\left(A \leq a \left| \sum_{j=1}^{2I} Z_j = n \right.\right)$.

Because $(V_1, V_2) \underline{\;||\;} (V_3, \ldots V_I)$ it follows that

$$(V_1, V_2) \underline{\;||\;} (V_3, \ldots V_I) \;\Big|\; \left(V_1 + V_2, \sum_{i=3}^{I} V_i \right);$$

see [7]. In particular, if $A_{g,h} = \sum_{i=g}^{h} Z_{2i-1}\,(1 - Z_{2i}) + (1 - Z_{2i-1})\,Z_{2i}$, then $A = A_{1,2} + A_{3,I}$ and

$$A_{1,2} \underline{\;||\;} A_{3,I} \;\Big|\; \left(V_1 + V_2, \sum_{i=3}^{I} V_i \right), \tag{11}$$

so that, continuing to use $\sum_{i=1}^{I} V_i = \sum_{j=1}^{2I} Z_j$,

$$\begin{aligned} \Pr\left(A \leq a \left| \sum_{j=1}^{2I} Z_j = n \right. \right) &= \Pr\left(A_{1,2} + A_{3,I} \leq a \left| \sum_{i=1}^{I} V_i = n \right. \right) \\ &= E\left\{ \Pr\left(A_{1,2} + A_{3,I} \leq a \left| \sum_{i=1}^{2} V_i, \sum_{i=3}^{I} V_i \right. \right) \;\Big|\; \sum_{i=1}^{I} V_i = n \right\} \end{aligned} \tag{12}$$

Combining (V_1, V_2) $\underline{||}$ $(V_3, \ldots V_I)$, (11) and (12), $\Pr\left(A \leq a \left| \sum_{j=1}^{2I} Z_j = n \right.\right)$ would be made larger (i.e., not smaller) for all a if $\Pr\left(A_{1,2} \leq c \left| \sum_{i=1}^{2} V_i = m \right.\right)$ were made larger (i.e., not smaller) for all (c, m).

Now given $V_1 + V_2 = m$,

$$\begin{aligned} A_{1,2} &= Z_1\,(1 - Z_2) + (1 - Z_1)\,Z_2 + Z_3\,(1 - Z_4) + (1 - Z_3)\,Z_4 \\ &= \begin{cases} 0 \text{ if } m = 0 \\ 1 \text{ if } m = 1 \\ 1 \text{ if } m = 3 \\ 0 \text{ if } m = 4 \end{cases} \end{aligned}$$

whereas if $m = 2$ then $\Pr\left(A_{1,2} = 0 \middle| V_1 + V_2 = 2\right) = \Psi / (\psi + \Psi)$ and

$$\Pr\left(A_{1,2} = 2 \middle| V_1 + V_2 = 2\right) = 1 - \Pr\left(A_{1,2} = 0 \middle| V_1 + V_2 = 2\right) = \psi / (\psi + \Psi)$$

where

$$\begin{aligned} \Psi &= \Pr\left\{(V_1, V_2) = (2, 0) \text{ or } (V_1, V_2) = (0, 2)\right\} \\ &= \pi_1 \pi_2 (1 - \pi_3)(1 - \pi_4) + \pi_3 \pi_4 (1 - \pi_1)(1 - \pi_2) \end{aligned}$$

and

$$\begin{aligned} \psi &= \Pr\left\{(V_1, V_2) = (1, 1)\right\} \\ &= \left\{\pi_1 (1 - \pi_2) + \pi_2 (1 - \pi_1)\right\} \left\{\pi_3 (1 - \pi_4) + \pi_4 (1 - \pi_3)\right\}. \end{aligned}$$

If (10) is true then

$$\Pr\left(A_{1,2}=0|\,V_1+V_2=2\right)=\frac{2e^{2\lambda}}{\left\{e^{2\lambda}+1\right\}\left\{e^{2\lambda}+1\right\}+2e^{2\lambda}}$$

but if π_2 and π_3 are interchanged, then this probability increases to

$$\Pr\left(A_{1,2}=0|\,V_1+V_2=2\right)=\frac{e^{4\lambda}+1}{\left\{e^{2\lambda}+1\right\}\left\{e^{2\lambda}+1\right\}+2e^{2\lambda}}.$$

It follows that the swap of π_2 and π_3 (or of Z_2 and Z_3) does not decrease $\Pr\left(A\leq a\left|\sum_{j=1}^{2I}Z_j=n\right.\right)$, proving Proposition 1.

References

[1] Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate - a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, 57, 289-300.

[2] Benjamini, Y. and Yekutieli, Y. (2001), "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, 29, 1165-1188.

[3] Benjamini, Y., Krieger, A. & Yekutieli, D. (2006), "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 93, 491-507.

[4] Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. (1988), *Combinatorial Optimization*, New York: Wiley

[5] Cook, W., Rohe, A. (1999), "Computing minimum-weight perfect matchings," *INFORMS Journal of Computation*, 11, 138-148. Software: http://www2.isye.gatech.edu/~wcook/

[6] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E. (1959), "Smoking and lung cancer," *Journal of the National Cancer Institute*, 22, 173-203.

[7] Dawid, A. P. (1979), "Conditional independence in statistical theory (with Discussion)," *Journal of the Royal Statistical Society*, B, 41, 1-31.

[8] Derigs, U. (1988), "Solving nonbipartite matching problems by shortest path techniques," *Annals of Operations Reseach*, 13, 225-261.

[9] Dudoit, S., and van der Laan, M. (2007), *Multiple Testing Procedures with Application in Genomics*, New York: Springer.

[10] Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

[11] Gene Ontology Consortium (2000), "Gene Ontology: tool for the unification of biology," *Nature Genetics*, 25, 25-29.

[12] Grant, G. , Manduchi, E. and Stoeckert, C. (2007), "Analysis and Management of Microarray Gene Expression Data," *Current Protocols in Molecular Biology*, supplement 77 (unit 19.6).

[13] Guo, W. and Rao, M. B. (2008), "On control of the false discovery rate under no assumption of dependency," *Journal of Statistical Planning and Inference*, **138**, 3176–3188.

[14] Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. R. (2004), "Optimal matching before randomization," *Biostatistics*, 5, 263-275.

[15] Hammond, E. C. (1964), "Smoking in relation to mortality and morbidity," *Journal of the National Cancer Institute*, 32, 1161–1188.

[16] Heller, R., Manduchi, E., and Small, D. (2009), Matching methods for observational microarray studies. *Bioinformatics*, 25, 904-909.

[17] Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6, 65-70.

[18] Kropf, S., Lauter, J., Eszlinger, M. , Krohn, K. and Paschke, R. (2004), "Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses," *Journal of Statistical Planning and Inference*, 125, 31-47.

[19] Lehmacher, W., Wassmer, G. and Reitmeir, P. (1991), "Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate, "*Biometrics*, 47, 511-21.

[20] Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001), "Matching with doses in an observational study of a media campaign against drug abuse," *Journal of the American Statistical Association*, 96, 1245-1253.

[21] Lu, B., Rosenbaum, P.R. (2004), "Optimal matching with two control groups," *Journal of Computational and Graphical Statististics*, 13, 422-434.

[22] Lu, B.: Propensity score matching with time-dependent covariates," *Biometrics*, 61, 721-728.

[23] Lu, B., Greevy, R., Xu, X., and Beck, C. (2010), "Optimal nonbipartite matching and its statistical applications," Submitted manuscript. (See also their nonbimatch package in R.)

[24] Marcus, R., Peritz, E. and Gabriel, K. R. (1976), "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, 63, 655-60.

[25] Nam, D. and Kim, S. (2008), "Gene-set approach for expression pattern analysis," *Briefings in Bioinformatics*, 9 (3), 189-197.

[26] Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments: Essay on principles, Section 9," In Polish, but reprinted in English with Discussion by T. Speed and D. B. Rubin in *Statistical Science*, 5, 463-480.

[27] Papadimitriou, C.H., Steiglitz, K. (1982), *Combinatorial Optimization: Algorithms and Complexity*, Englewood Cliffs, NJ: Prentice Hall.

[28] Reiner, A. (2007), "FDR Control by the BH Procedure for Two-Sided Correlated Tests with Implications to Gene Expression Data Analysis," *Biometrical Journal*, 49, 107-126.

[29] Reiner,A. and Yekutieli,D. and Benjamini,Y. (2003), "Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures," *Bioinformatics*, 19, 368-375.

[30] Romano, J. P., Shaikh, A. M., and Wolf, M. (2008), "Rejoinder: Control of the false discovery rate under dependence using the bootstrap and subsampling," *Test*, 17, 461-471,

[31] Rosenbaum, P. R. and Krieger, A. M. (1990), "Sensitivity analysis for two-sample permutation inferences in observational studies," *Journal of the American Statistical Association*, 85, 493-498.

[32] Rosenbaum, P. R. (1991), "Some poset statistics," *Annals of Statistics*, 19, 1091-1097.

[33] Rosenbaum, P. R. (1995), "Quantiles in nonrandom samples and observational studies," *Journal of the American Statistical Association*, 90, 1424-1431.

[34] Rosenbaum, P. R. (2002), *Observational Studies*. New York: Springer.

[35] Rosenbaum, P. R. (2005), "An exact, distribution free test comparing two multivariate distributions based on adjacency," *Journal of the Royal Statistical Society*, B, 67, 515-530.

[36] Rosenbaum, P. R. (2008), "Testing hypotheses in order," *Biometrika*, 95, 248-252.

[37] Rosenbaum, P. R. and Silber, J. H. (2009), "Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units," *Journal of the American Statistical Association*, 104, 501-511.

[38] Rosenbaum, P. R. (2010), *Design of Observational Studies*, New York: Springer.

[39] Rubin, D.B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688-701.

[40] Samuel-Cahn, E. (1996), "Is the Simes improved Bonferroni procedure conservative?" *Biometrika*, 83, 928-933.

[41] Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., Palma, J., and Brody, J. S. (2004), "Effects of cigarette smoke on the human airway epithelial cell transcriptome," *PNAS*, 101, 10143-10148.

[42] Storey, J.D. (2002), "A direct approach to the false discovery rate," *Journal of the Royal Statistical Society* B, 64, 479-498.

[43] Westfall, P. H. and Young, S. S. (1993), "On adjusting P-values for multiplicity. *Biometrics*, 49, 941-945.

[44] Yekutieli, D. (2008), "Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling," *Test*, **17**, 458–460.

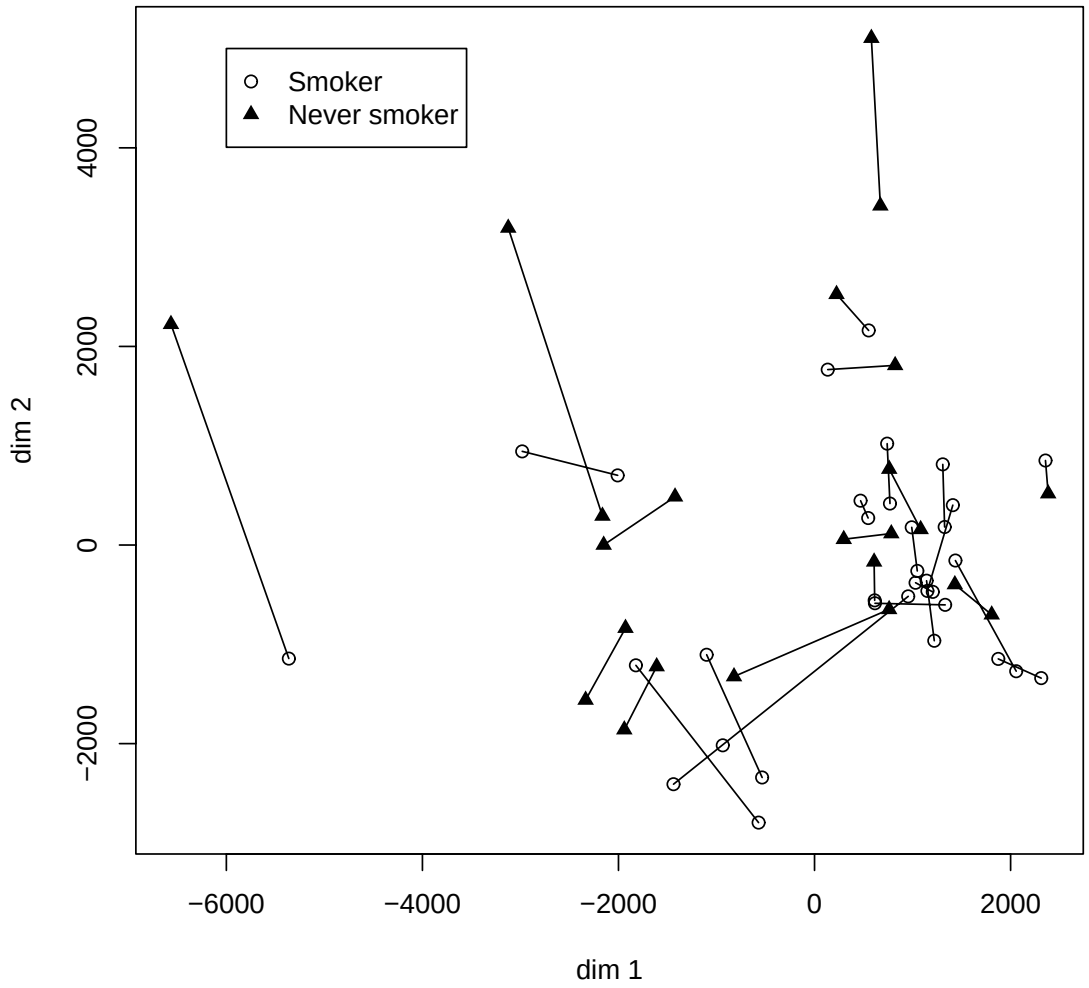

Figure 1: A two-dimensional representation of the 9968-dimensional cross-match test. Paired subjects are connected by a line. The two dimensions are from a multidimensional scaling of the 56×56 distance matrix for the 56 subjects who were paired. The multidimensional scaling is for graphical purposes only; it plays no role in the test. Because there are five instances in which a circle is connected to a triangle, the cross-match statistic is $A=5$.

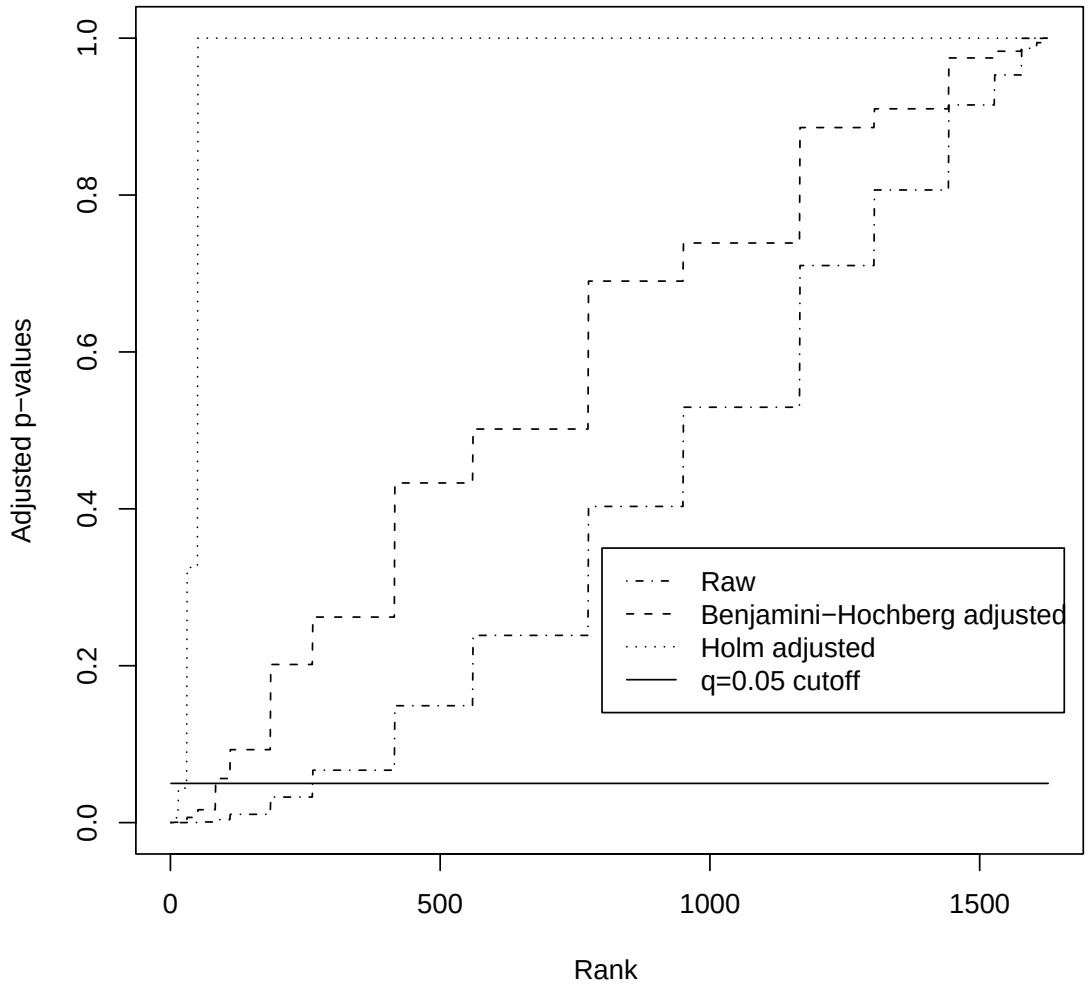

Figure 2: The raw p-values, as well as the adjusted p-values from the Holm and the Benjamini-Hochberg procedures. At the 0.05 level, 30 and 83 hypotheses are rejected using the Holm and the Benjamini-Hochberg procedure respectively.

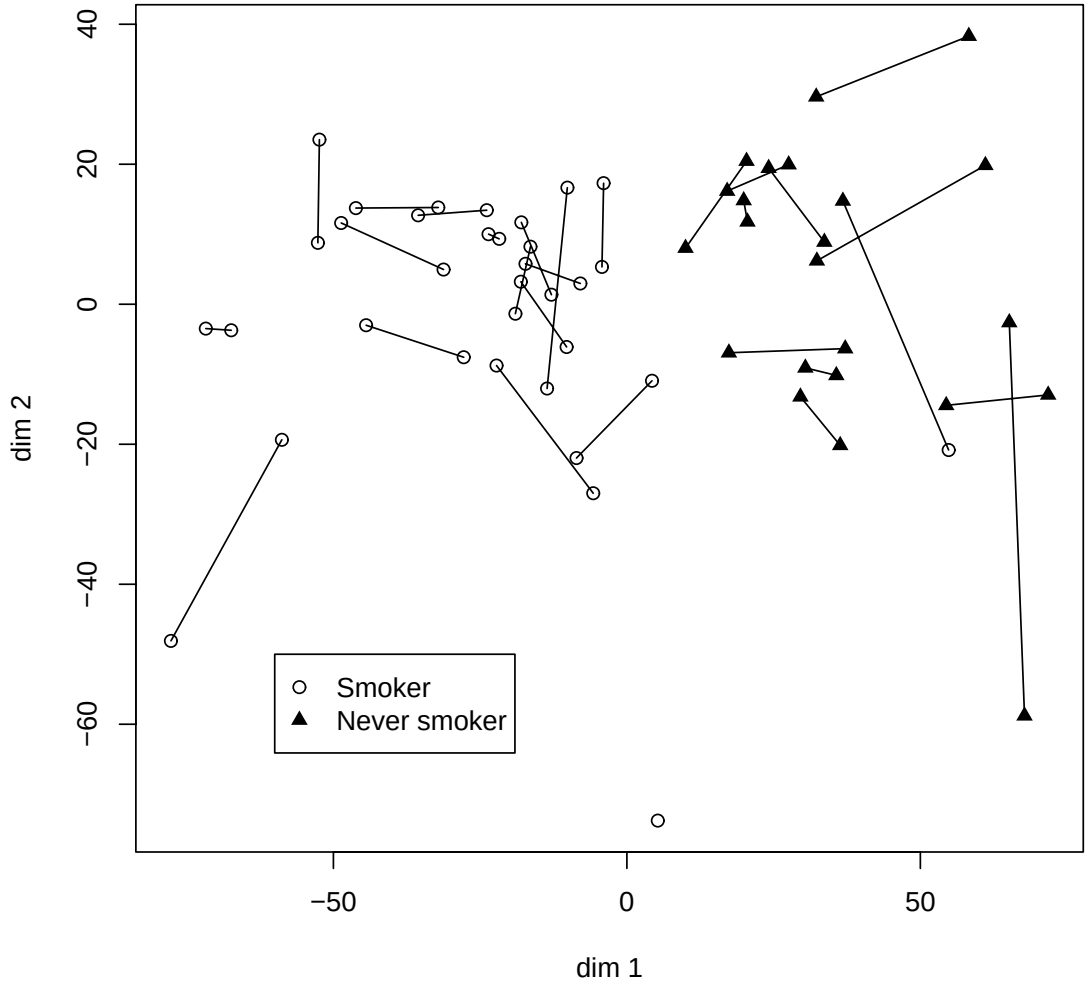

Figure 3: A two-dimensional representation of the 92-dimensional cross-match test of the molecular function GO:0016616. Because there is only 1 instance in which a circle is connected to a triangle, the cross-match statistic is *A=1*.