Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

Space–time clustering and the permutation moments of quadratic forms

Yi-Hui Zhou[a]*, Gregory Mayhew[b], Zhibin Sun[c], Xiaolin Xu[d], Fei Zou[b] and Fred A. Wright[a]

Received 20 September 2013; Accepted 25 October 2013

The Mantel and Knox space–time clustering statistics are popular tools to establish transmissibility of a disease and detect outbreaks. The most commonly used null distributional approximations may provide poor fits, and researchers often resort to direct sampling from the permutation distribution. However, the exact first four moments for these statistics are available, and Pearson distributional approximations are often effective. Thus, our first goals are to clarify the literature and make these tools more widely available. In addition, by rewriting terms in the statistics, we obtain the exact first four permutation moments for the most commonly used quadratic form statistics, which need not be positive definite. The extension of this work to quadratic forms greatly expands the utility of density approximations for these problems, including for high-dimensional applications, where the statistics must be extreme in order to exceed stringent testing thresholds. We demonstrate the methods using examples from the investigation of disease transmission in cattle, the association of a gene expression pathway with breast cancer survival, regional genetic association with cystic fibrosis lung disease and hypothesis testing for smoothed local linear regression. © The Authors. *Stat* published by John Wiley & Sons Ltd.

Keywords: exact testing; resampling; statistical computing

1 Introduction

Mantel (1967) proposed an approach to detect clustering of location of events in space versus time of occurrence, by regressing a function of geographic distance on a function of distance in time. The prototypical application is to evaluate the evidence for communicable disease transmission, in contrast to sporadic occurrences that show no clustering. The approach has proven to be hugely popular, with 5200+ citations in the Science Citation Index as of 2013, with approximately 450 citations in each of recent years. Briefly, we let l_i and t_i represent the geographic location (space) and time of occurrence for the ith location–time sample, $i = 1, \ldots, n$. For samples i and j, we denote measures of location and time distances as $c_{ij} = f(l_i, l_j)$, $d_{ij} = g(t_i, t_j)$, and these elements populate the matrices **C** and **D**, respectively. For a final "regression" statistic

[a]Department of Statistics, North Carolina State University, Raleigh, NC, 27695, USA
[b]Department of Biostatistics, University of North Carolina, Chapel Hill, NC, 27599, USA
[c]Universities Space Research Association, Columbia, MD, 21044, USA
[d]School of Business, Nanjing University, Nanjing, Jiangsu, 210093, China
*Email: yihui_zhou@ncsu.edu

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

$$S = \sum_{i,j} c_{ij} d_{ij}, \tag{1}$$

high values are evidence of location–time clustering, and the author considered the power of various choices of f and g. He also noted that the Knox statistic, which records whether two locations or time points are less distant than predefined thresholds, is a special case. In addition, the paper solved for the mean and variance of S under permutation of sample labels for the location and time points. This permutation is equivalent to simultaneous permutation of rows and columns of one of the matrices **C** or **D**.

Much of Mantel's (1967) and subsequent work is concerned with finding powerful choices of f and g, but here, we assume that the statistic has been chosen, and our goal is to provide accurate testing. For numerous datasets, a normal approximation to S is inadequate, because of strong dependencies among the matrix elements. For the Knox statistic, p-values based on Poisson approximations (Knox, 1964) or a normal approximation (David & Barton, 1966) have been used. The improvements to Mantel and Knox tests for space–time interaction were proposed in several papers (Kulldorff & Hjalmars, 1999; Diggle et al., 1995; Jacquez, 1996; Baker, 1996) by not requiring as strong assumptions on the spatial and temporal scales of clustering. But in general, direct sampling from the permutation distribution had often been thought to be necessary, as enumeration of the $n!$ outcomes is of course infeasible for most datasets. An alternative approach is to use moment-based density approximations, but the skewness and kurtosis are important for tail accuracy. Siemiatycki (1978) provided the first four moments of S under permutation, for the most commonly encountered situation that **C** and **D** are symmetric with zero diagonals. The author described graphical patterns to aid in computing expectations of product terms, for example, in $c_{ij}c_{kl}c_{mn}c_{st}$, there are 23 distinct patterns of equality/inequality for the eight subscripts. In addition, moments of S were expressed in linear combinations of products of terms of varying order from **C** and **D**—the terms for the fourth moment involve nearly 150 non-zero coefficients. Although the bookkeeping is tedious, these operations reduce the complexity from a naive $O(n^8)$ to $O(n^3)$. With this reduction, density approximations become feasible for computing p-values, with reasonable accuracy even for stringent testing thresholds.

The space–time clustering statistic can easily be seen to resemble a quadratic form $\mathbf{y}^T\mathbf{A}\mathbf{y}$, where $\mathbf{y}$ is an $n\times 1$ vector with elements y_i, and $\mathbf{A}$ is a symmetric $n \times n$ matrix with elements a_{ij}. This can be seen by rewriting $\mathbf{y}^T\mathbf{A}\mathbf{y} = \sum_i \sum_j a_{ij}y_iy_j$, which is similar to (1), with a_{ij} and y_iy_j serving the roles of c_{ij} and d_{ij}. However, a key difference lies in the diagonals, that is, that a_{ii} and y_i^2 are not generally zero. Quadratic forms have been used for location–time clustering (Tango, 1984), but we are not aware that a direct equivalence has been described between the Mantel statistic and a quadratic form over permutations, and for the latter, to our knowledge, only the first two exact moments have been reported (Commenges, 2003). Quadratic forms arise in a number of disciplines, including epidemiology, genomic, economics, and other areas. The computation of exact moments enables robust analysis, while avoiding the additional computational cost of direct permutation.

Despite the popularity of the Knox–Mantel and related location–time clustering statistics, software has not been available to compute the four moments or subsequently obtain approximation p-values, despite a number of packages devoted to location–time surveillance (Robertson & Nelson, 2010). Similarly, quadratic forms are increasingly used, for example, in genomics problems (Tong et al., 2010). However, standard results for normal quadratic forms may not apply, such as for binary disease traits. The application of quadratic forms to non-normal data is often justified by appealing to asymptotics (Wu et al., 2011), but the use of exact methods may be preferred.

We have developed *R* code to compute the first four exact moments for the location–time statistic and for centered quadratic forms and to compute approximations to the exact permutation p-values using Pearson density approximations. We believe that the software and methods are useful additions to the statistician's toolkit.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

The ISI's Journal for the Rapid Dissemination of Statistics Research

2 Methods

2.1. The location–time statistic S

For symmetric **C** and **D** (with zero diagonals), we have implemented the Siemiatycki moment computation. The permutation approach involves simultaneous permutation of rows and columns of one of the matrices (say **D**), which is equivalent to permutation of the location versus time observations (Mantel, 1967). We use $\pi = 1, \ldots, n!$ as a subscript to represent a permutation of the n objects, with reordered indexes $\pi[1], \ldots, \pi[n]$. A random permutation is denoted as Π, and our task is to compute the first four moments of $S_\Pi = \sum_i \sum_j c_{ij} d_{\Pi[i],\Pi[j]}$. The key computations are shown in the Appendix, expressed in matrix form to exploit linear algebra routines in *R*. Approximate *p*-values are obtained by matching the exact moments to the Pearson family of distributions using the *PearsonDS* package, which automatically chooses the best-fitting type within the Pearson family.

2.2. Equivalence of the quadratic form statistic S

Here, the statistic is $S = \mathbf{y}^T\mathbf{A}\mathbf{y}$, for symmetric **A** with corresponding permutation random variable $S_\Pi = \mathbf{y}_\Pi^T\mathbf{A}\mathbf{y}_\Pi$. In many useful applications, **A** is centered, that is, the rows and columns sum to a constant μ. Here, we will assume $\mu = 0$, essentially without loss of generality, as non-zero μ values will offset S_Π by a constant $\mu\mathbf{y}^T\mathbf{y}$. Standard normal-theory results typically assume that **A** is positive definite, and the assumption is necessary for standard χ^2 distributional approximations. However, relaxing this assumption would considerably increase the variety of problems for which accurate *p*-values can be obtained. For example, in a genomic context, Zhou & Wright (2013) provided motivation for useful quadratic forms with eigenvalues summing to zero. Kuonen (1999) summarized a number of previous studies of quadratic form approximations, including those that are not positive definite, and described saddlepoint approximations applicable to normally distributed **y** only.

The moments computed by Siemiatycki were considerably simplified by assuming zero diagonals for **C** and **D**. Here, we describe a simple construction to map the quadratic form to the Mantel statistic. First, we define $\mathbf{C} = \mathbf{A} - \mathrm{diag}(\mathbf{A})$, that is, $c_{ij} = a_{ij}$ for $i \neq j$ and zero otherwise. Then we define **D** as the matrix with entries $d_{ij} = -\frac{1}{2}(y_i - y_j)^2$, and by this definition, each $d_{ii} = 0$. Our claim is that, for any π, $\sum_i \sum_j c_{ij} d_{\pi[i],\pi[j]} = \sum_i \sum_j a_{ij} y_{\pi[i]} y_{\pi[j]} = \mathbf{y}_\pi^T\mathbf{A}\mathbf{y}_\pi$.

Proof

By the constraint, $\sum_j a_{ij} = 0$, and therefore for any fixed π, we have $\sum_i \sum_j a_{ij} y_{\pi[i]}^2 = \sum_i y_{\pi[i]}^2 \sum_j a_{ij} = 0$, and by the same reasoning, $\sum_i \sum_j a_{ij} y_{\pi[j]}^2 = 0$. We have

$$\sum_i \sum_j c_{ij} d_{\pi[i],\pi[j]} = \sum_i \sum_j a_{ij} \left\{ -\frac{1}{2} \left(y_{\pi[i]} - y_{\pi[j]} \right)^2 \right\}$$

because each $d_{\pi[i],\pi[i]} = 0$. Expanding the right-hand side gives

$$\sum_i \sum_j a_{ij} y_{\pi[i]} y_{\pi[j]} - \frac{1}{2} \sum_i \sum_j a_{ij} y_{\pi[i]}^2 - \frac{1}{2} \sum_i \sum_j a_{ij} y_{\pi[j]}^2,$$

for which the last two entries are zero. Thus, $\sum_i \sum_j c_{ij} d_{\pi[i],\pi[j]} = \mathbf{y}_\pi^T\mathbf{A}\mathbf{y}_\pi$. □

As with the location–time statistic, we use Pearson family approximations to compute *p*-values. Because of the row/column constraint, several moment terms can be further simplified to lower order $O(n^2)$ (Appendix), which may be useful in applications for very large n.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

2.3. Permutation versus normal quadratic forms

Our motivation here is to perform approximations to exact inference, and our procedures only need the exchangeability assumption on the observed **y**, applying equally well to discrete or continuous data. For normal quadratic forms, where the elements of **y** are drawn randomly iid from a normal density, the null distribution may be computed as a weighted sum of independent χ^2_1 random variables, using the methods of Imhof (1961) or the saddlepoint approximation of Kuonen (1999), for example, as implemented in the *survey* package in *R*. A common technique used in genomics and other disciplines is to perform robust analysis by transforming data to be discrete-normal using rank-based inverse normal transformations. For example, if $r(y_i)$ is the rank of the ith observation, the transformed value is $y'_i = \Phi^{-1}(r(y_i)/(n+1))$. The use of normal scores in genetics was discussed and extensively critiqued by Beasley et al. (2009). An underlying theme in the application of normal scores appears to be a presumption that permutation of the scores is nearly equivalent to unconditional normal random sampling. For individual association tests, this assumption may be reasonable. For example, the permutation variance of the Pearson correlation coefficient between fixed vectors **x** and **y** is $1/(n-1)$, which is identical to the variance if **y** is randomly drawn as iid normal. However, permutation of **y** inherently creates negative correlation among the sampled elements. This dependence, which is slight for individual elements of **y** and decreases with n, remains highly consequential for S, because there are n^2 correlation terms among the elements. This effect of with-replacement sampling is especially strong if the eigenvalues of **A** do not contain a few dominant values (Zhou & Wright, 2013).

The permutation dependency phenomenon is illustrated in four panels in Supplementary Figure 1. For each panel, a single initial $m \times n$ matrix **X** was generated with elements drawn iid N(0,1) and row-centered, where $m = \{10, 1000\}$ and $n = \{50, 500\}$. Then we let $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ and compare the distribution of the unconditional normal quadratic form with that of permutation of normal scores. The figure illustrates that the variability under permutation is markedly less than for unconditional sampling, except when $n >> m$. Thus, even if an investigator transforms **y** to normal scores, the normal quadratic form null distribution cannot be used for permutation testing, and the methods described here remain relevant.

2.4. Example datasets

We illustrate our methods for four published examples, and for each of the first three examples, we use two different S statistics. The statistics are the same as proposed by the original authors or are otherwise well motivated within the context of the problem. For each example and choice of statistic, the analyst need only find **C** and **D**, or **y** and **A**, as appropriate to the problem. We note that these examples are useful not only for the observed statistics and p-values but also for the adequacy of the fit for the entire permutation distribution, and thus, the examples effectively illustrate the performance of our approximation in a variety of settings.

Example 1

In White et al. (1989), space–time clustering was used to investigate the evidence of transmissibility of dysentery in cattle for 37 outbreaks in farms in rural New York. Both the Mantel and Knox statistics were used, which we will denote S_{Mantel} *and* S_{Knox}*. Following the authors' implementation of the Mantel statistic, for f, we calculated the straight-line distance in kilometres between locations, and for g, we used the unsigned difference in days between outbreaks. The resulting matrices* **C** *and* **D** *were then used to calculate* S_{Mantel}*.*

The Knox statistic is the number of outbreak pairs that are close in space and time. Thresholds for defining closeness are required, and we used the thresholds of 5.5 km and 30 days chosen in White et al. (1989). In other words, $c_{ij} = 1$ *for* $f(l_i, l_j) < 5.5$ *km, and 0 otherwise. Similarly, for the Knox statistic,* d_{ij} *is an indicator for* $g(t_i, t_j)$ *less than 30 days, and* $c_{ii} = d_{ii} = 0$*. The resulting matrices* **C** *and* **D** *were then used to calculate* S_{Knox} *(which is twice the*

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

The ISI's Journal for the Rapid Dissemination of Statistics Research

statistic proposed by Kulldorff & Hjalmars (1999)). Although our moment calculations are exact, for an observed statistic s, density approximations to p-values tend to be closer to the mid p-value $p_{\text{mid}} = P(S > s) + \frac{1}{2}P(S = s)$ *than to the p-value* $P(S \geq s)$*. For most of the examples in this paper, the difference between the two is trivial and need not be considered. However, for this example,* S_{Knox} *statistic can assume only the 25 even values 0, 2, . . . , 48, and so we apply a continuity correction, by using the Pearson density approximation for* $s - 1$ *instead of s.*

Example 2

For pathway analysis of genetic expression data, the data are typically divided into $\mathbf{X}_{\text{path}}$*, which represents the* $m_{\text{path}} \times n$ *matrix of expression of* m_{path} *genes belonging to a pathway, and* $\mathbf{X}_{\text{comp}}$ *is the remaining* $m_{\text{comp}} \times n$ *complementary matrix of genes not in the pathway. We assume that both matrices are row centered and scaled. Expressions of genes are then compared to a clinical or experimental outcome* $\mathbf{y}$*, either by examining the association of* $\mathbf{y}$ *to only genes within the pathway (known as self-contained testing) or by contrasting the association with genes in the pathway versus that in the complement (competitive testing). Zhou & Wright (2013) proposed corresponding quadratic form statistics* $S_{\text{self}} = \mathbf{y}^T\mathbf{X}_{\text{path}}^T\mathbf{X}_{\text{path}}\mathbf{y}$*, and* $S_{\text{compet}} = \mathbf{y}^T\left(\frac{1}{m_{\text{path}}}\mathbf{X}_{\text{path}}^T\mathbf{X}_{\text{path}} - \frac{1}{m_{\text{comp}}}\mathbf{X}_{\text{comp}}^T\mathbf{X}_{\text{comp}}\right)\mathbf{y}$*, for which they obtained p-values using a weighted beta density approximation. However, for that approximation, only the first two moments are exact.* S_{compet} *has eigenvalues summing to zero, and for some, datasets can have a negative skew, making* χ^2 *density approximations ineffective. We use the breast cancer data of Miller et al. (2005), for which the pathway GO:0000184: "nuclear-transcribed mRNA catabolic process" (44 genes,* $n = 236$ *samples) was used in Zhou & Wright (2013) for an example in tests of association with survival. Here,* $\mathbf{y}$ *is the vector of martingale residuals for survival time,* $\mathbf{X}$ *is gene expression data, and both have been preresidualized for p53 mutation status.*

Example 3

Wright et al. (2011) described a genome-wide association analysis for lung function among 1978 cystic fibrosis (CF) patients, identifying the interval between the genes EHF and APIP on chromosome 11 as of interest. For an interval consisting of several genetics markers, we use an approach to perform regional genetic analysis, rather than testing individual markers. The approach compares similarities in the lung function phenotype between all pairs of individuals with a correlation-based measure in regional genotypes. The result (which we call S_{assoc1}*) is similar in spirit to a Mantel statistic, except that the individual elements represent similarity rather than distance. Specifically, we let* y_j *denote the phenotype for the jth individual, and the subsequent description is simplified by assuming* $\mathbf{y}$ *has been centered and scaled so that* $\sum_j y_j^2 = n - 1$*. We use* $d_{ij} = y_i y_j$ *for* $i \neq j$ *and* $d_{ii} = 0$*, following suggestions that the product* $y_i y_j$ *should be powerful in performing tests of phenotypic versus genotypic relatedness* (Elston et al., 2000)*. For m genetic markers in a region, with genotypes measured on the n individuals, we have an* $m \times n$ *genotype matrix* $\mathbf{G}$*, which has been centered and column scaled. For* $i \neq j$*, we use* $c_{ij} = \text{corr}(\mathbf{g}_{.i}, \mathbf{g}_{.j})$*, where* $\mathbf{g}_{.i}$ *is the ith column of* $\mathbf{G}$*, and "corr" is the Pearson correlation.*

A closely related quadratic form statistic (S_{assoc2}*) is the sum of squared score statistics across the markers, which is similar to* S_{assoc1} *but with slightly different genotype scaling, and with non-zero diagonals for the corresponding matrices. We use* $\mathbf{X}$ *to denote the matrix of genotypes, which have been row centered and scaled so that* $\sum_j x_{ij} = 0$ *and* $\sum_j x_{ij}^2 = 1$*. A single score statistic for the ith marker is* $\sum_j x_{ij}y_j$*, and* $S_{\text{assoc2}} = \sum_i \left(\sum_j x_{ij}y_j\right)^2$*, which can be shown to be* $S_{\text{assoc2}} = \mathbf{y}^T\mathbf{A}\mathbf{y}$*, where* $\mathbf{A} = \mathbf{X}^T\mathbf{X}$*.*

Example 4

Bowman & Azzalini (1997, pp. 86–90) described a dataset resulting from sampling aquatic life in a coral reef, with 42 observations of catch score, summarized as a log weight across numerous species, versus depth. The dataset has been used by these authors and others to demonstrate local linear regression, using a normal smoothing kernel.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

A standard test statistic for local linear regression can be expressed as a quadratic form, as follows. The derivation applies to the Nadaraya–Watson estimator (Nadaraya, 1964; Watson, 1964)

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} w_h(x_i - x) y_i}{w_h(x_i - x)}$$

with kernel function w_h, for the regression model $E(Y_i|x_i) = m(x_i)$. The fitted values $\hat{\mathbf{y}}$ can obtained using a smoothing matrix $\mathbf{M}$ (which depends on h) such that $\hat{\mathbf{y}} = \mathbf{My}$. As shown in Bowman & Azzalini (1997), an F-like statistic can be obtained using the ratio

$$F = \frac{\mathbf{y}^T \mathbf{U} \mathbf{y}}{\mathbf{y}^T \mathbf{V} \mathbf{y}},$$

with $\mathbf{U} = \mathbf{I} - \mathbf{1}_{n\times n}/n - (\mathbf{I} - \mathbf{M})^T(\mathbf{I} - \mathbf{M})$ and $\mathbf{V} = (\mathbf{I} - \mathbf{M})^T(\mathbf{I} - \mathbf{M})$. The p-value is $P(F > F_{\text{obs}})$, which can be rewritten as $P(\mathbf{y}^T(\mathbf{U} - F_{\text{obs}}\mathbf{V})\mathbf{y} > 0$, and so we use, finally, $\mathbf{A} = (\mathbf{U} - F_{\text{obs}}\mathbf{V})$ in the quadratic form. It is easy to show that $\mathbf{A}$ is symmetric with row/column sums of zero.

Bowman & Azzalini (1997) obtained *p*-values using moments from a normal quadratic form and a scaled chi-square density approximation, while acknowledging that the data included some non-normal features, such as truncation. They describe permutation analysis as an alternative approach, which they did not pursue further. For the same normal quadratic form, Kuonen (1999) reported *p*-values using a saddlepoint approximation. Here, we report *p*-values based for direct permutation and compare to results from our moment-based density approximation.

3 Results

Example 1

Figure 1 (left panel) shows a histogram of the S_{Mantel} statistic, overlaid with the normal density approximation. Although the normal approximation is based on exact moments, the presence of skew in the data creates a poor tail fit. In contrast, the approximation from our proposed method, which uses four moments and a Pearson type IV fit,

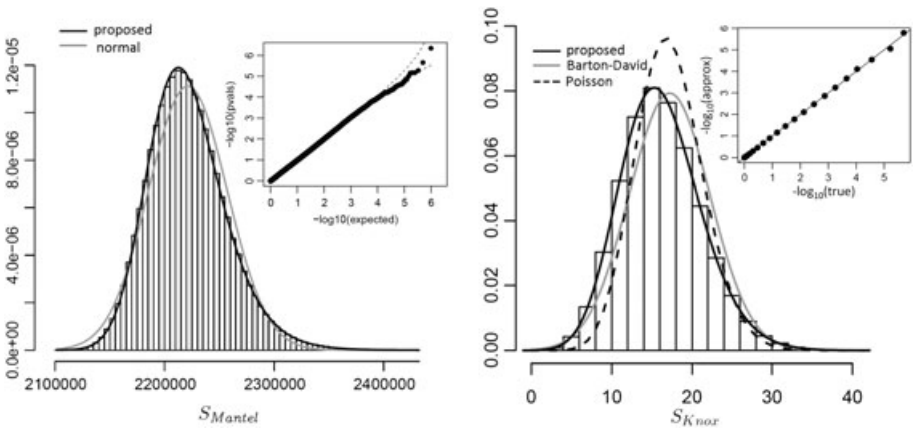

Figure 1. Performance of the proposed approach for space–time clustering analysis of the cattle data. The left panel shows a histogram of S_{Mantel} and a *q–q* plot of observed approximating *p*-values versus expected for 10^6 permutations.The right panel shows the analogous results for S_{Knox} for 10^6 permutations, along with density fits based on the Barton–David and Poisson approximations, as well as our proposed density fit. The inset shows the true permutation *p*-values for all possible outcomes, compared to that of the approximation.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

The ISI's Journal for the Rapid Dissemination of Statistics Research

is highly accurate. An observed versus expected q–q plot for p-values from our proposed procedure, applied to 10^6 *permutations, shows that the p-values are nearly uniform. The actual data show only marginal evidence of location–time clustering, with true* $p = 0.0703$*, with the density approximation of* $p = 0.0699$*. The results are similar for* S_{Knox} *(right panel). The proposed approximation (type IV) is accurate, while the two competing approximations in common use, based on David & Barton (1966), and a Poisson approximation are observably less accurate. For the actual data,* S_{Knox} *has a permutation-based* $p = 0.0681$ *and approximating* $p = 0.0696$*. Note that the tail probabilities do not degrade in accuracy, as shown by a comparison of true versus approximating p-values for the entire range of possible outcomes (inset of Figure 1, right panel).*

Example 2

Figure 2 shows histograms and q–q plots for S_{self} *and* S_{compet} *for the Miller breast cancer data for pathway GO:0000184. Here, again the fits (type VI for* S_{self} *and type IV for* S_{compet}*) are accurate, with a slight conservativeness of the approximating p-values in the extreme right tail for* S_{compet}*. For the observed data and* S_{self}*, the permutation based* $p = 0.080$ *and Pearson distributional approximation* $p = 0.081$*. For* S_{compet}*, the respective values are* $p = 0.822$ *and* $p = 0.817$*.*

Example 3

Figure 3 shows $-log_{10}(p)$ *for* S_{assoc1} *and* S_{assoc2} *for the CF data, where each statistic is plotted for the middle single-nucleotide polymorphism (SNP) in each 21-SNP window. q–q plots, produced for the interval showing the greatest evidence in the original data rs, again support the accuracy of the approximating p-values (type VI for all windows). The most highly significant region is in the interval between* EHF *and* APIP*, which is also supported by the single-SNP analysis. However, the evidence is much stronger for* S_{assoc1} *and* S_{assoc2} *than for single-SNP analysis and certainly significant in a genome-wide scan accounting for ~570,000 SNPs. We attribute the greater evidence from these statistics to the potential presence of multiple causal SNPs in the region, as proposed by Wright et al. (2011) in their analyses, because the moving window can capture the combined evidence from multiple SNPs. In fact, the relative genome-wide evidence may be even stronger for the regional methods, as they tend to have higher serial correlation than for individual SNPs and thus incur a smaller multiple-testing penalty. The use of* S_{assoc1} *and* S_{assoc2} *in this context is very similar to using sequence kernel association test* (Wu et al., 2011)*, which is designed for regional analysis and rare-variant testing of genetic association. However, these methods were formally designed for normal or binary phenotypes, and our use of exact moments adds considerable flexibility in handling the actual phenotype distribution.*

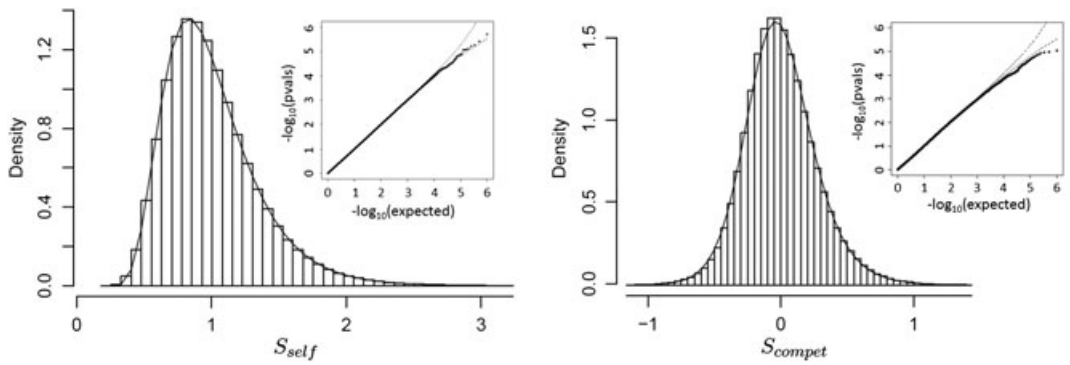

Figure 2. Example 2. Results for S_{self} (left panel) and S_{compet} (right panel) for the Miller breast cancer data, pathway GO:0000184 ($n = 236$, 44 genes in pathway).

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

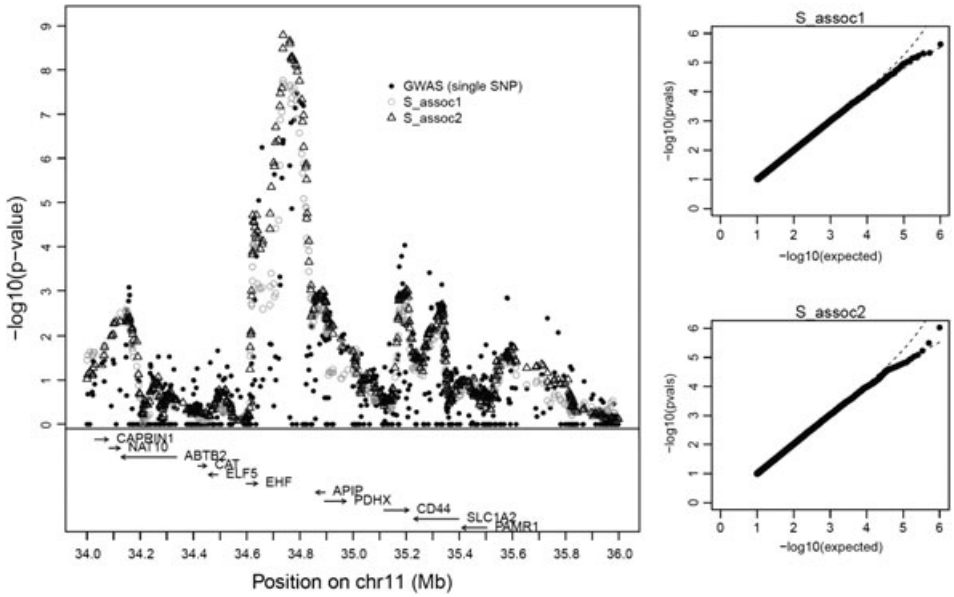

Figure 3. The left panel shows $-log_{10}$ p-values for S_{assoc1} and S_{assoc2} for the CF dataset. Each p-value is computed for a moving window of ± 10 SNPs around the center SNP. The two q–q plots for a fixed interval show that the proposed approximating p-values are approximately uniform under 10^6 permutations.

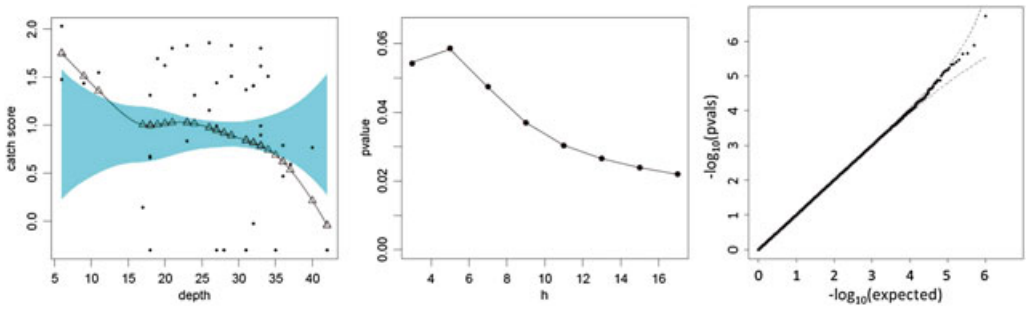

Figure 4. The application of the quadratic form approximation to the test statistics for local linear regression. Left panel: fitted curve and no-effected reference band. The triangles denote the fitted values for observed depth, obtained from the smoothing matrix as **My**. Middle panel: significance trace showing permutation p-values (dots) and the proposed approximation (line) as a function of h. Right panel: q–q plot for approximating p-values under permutation for $h = 5$.

Example 4

Figure 4 (left panel) plots the coral reef data, along with the smoothed local linear regression fit and confidence band from the R sm *package, for a kernel smoothing bandwidth of h=5, produced by the* sm *package in R. The fitted values at the extremes are clearly outside the reference band for the no-effect model. A "significance trace" (middle panel) shows p-values as a function of h, and for which the permutation-based p-values (dots) and type IV approximation (line) are nearly indistinguishable. The permutation-based p-values are generally lower than those obtained from the normal quadratic form, which were obtained for these data in Kuonen (1999). For example, for h=5, the permutation-based* $p = 0.058$ *but is 0.063 for the normal quadratic form.*

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

The ISI's Journal for the Rapid Dissemination of Statistics Research

4 Discussion

For standard space–time clustering statistics, our contribution has been to provide software for the moments and approximate p-values. Quadratic forms are used in a wide variety of settings, and the use of exact permutation moments has been often overlooked as an alternative to direct permutation. One setting where these approaches may be useful is SNP association pathway analysis, where the effect of sets of SNPs is aggregated and where direct permutation has been considered cumbersome, leading to alternative resampling proposals (Schaid et al., 2012). The use of our S_{compet} statistic, applied to genome-wide SNP association data, would enable true competitive testing for association pathway analysis. Such competitive testing had been viewed as infeasible, as a naive approach involves performing a full genome scan for each permutation.

Another point of consideration is whether direct sampling from the permutation distribution might be still preferable, as it provides an unbiased estimate of the permutation p-value. In high-throughput settings, however, extreme thresholds may be required to declare significance, and here, our approximation may be especially useful. In Zhou & Wright (2013), even the use of adaptive permutation (performing only as many permutations as necessary for high relative accuracy) was ∼250 times slower than the use of a moment-based analytic approximation. For the genome-scan setting of Example 3, p-values on the order of 10^{-8} are necessary in order to declare significance. Moreover, the *EHF/APIP* region for the CF example was initially identified using screening of individual SNPs, and so the investigator might compute the quadratic form p-value only in regions identified as of potential interest, avoiding the computational burden of genome-wide scans using the quadratic form statistic.

Appendix

We repeat here the notation from Siemiatycki (1978), modified to accord with our notation. The patterns of equivalence in the subscripts in $c_{ij}c_{kl}c_{mn}c_{st}$, for example, are represented in 23 graphical patterns in Appendix A of Siemiatycki (1978). We have r as the order of the moment, $q^{(r)}$ as the number of patterns associated with the rth moment, $\alpha^{(r)}$ as the pattern under consideration, $S_{c\alpha}^{(r)}$ as the sums derived from matrix $\mathbf{C}$, $P_{c\alpha}^{(r)}$ as the total of the products $\left(c_{i_1 i_2}\ldots c_{i_{2r-1}i_2}\right)$ under conditions of pattern α, $v_\alpha^{(r)}$ as the number of distinct subscripts in pattern α, $f_\alpha^{(r)}$ as the number of structurally equivalent forms of the αth pattern for the rth moment, and n as the sample size. All of the terms involving matrix $\mathbf{D}$ have analogous counterparts for $\mathbf{D}$. We have $q^{(1)}=1$, $q^{(2)}=3$, $q^{(3)}=8$, $q^{(1)}=23$, and $\alpha^{(r)}$; and $f_\alpha^{(r)}$ are shown in Appendix A of Siemiatycki (1978). The terms in $P_{c\alpha}^{(r)}$ are described subsequently. We have, finally, $E(S^r)=\sum_{\alpha=1}^{q^{(r)}}\frac{P_{d\alpha}^{(r)}P_{c\alpha}^{(r)}f_\alpha^{(r)}}{n(n-1)\cdots\left(n-v_\alpha^{(r)}+1\right)}$. All of the (central) moments follow from these non-central moments. Appendix B in Siemiatycki (1978) provides algebraic sums $Q_{c1}^{(1)},\ldots,Q_{c23}^{(4)}$ (using S instead of Q in the original reference), which are the building blocks for the $P_{c\alpha}^{(r)}$ terms. Most of the terms are $O(n^2)$ or less, but several of the terms are $O(n^3)$. Later, we describe how the terms reduce to $O(n^2)$ for the quadratic form when $\mathbf{A}$ is centered and $\mathbf{A}=\mathbf{X}^T\mathbf{X}$ for $m\times n$ matrix $\mathbf{X}$. Here, we consider m to be fixed and n increasing, so $O(mn^2)=O(n^2)$.

Many of the Siemiatycki Appendix B terms sum over i and j only for the matrices, and so are $O(n^2)$. The row sums of $\mathbf{A}$ are zero, and $\mathbf{C}=\mathbf{A}-\text{diag}(\mathbf{A})$, so $\sum_j c_{ij}=-a_{ii}$. Thus, the term

$$Q_{c3}^{(3)}=\sum_i\sum_j c_{ij}\left(\sum_k c_{ik}\right)\left(\sum_k c_{jk}\right)=\sum_i\sum_j c_{ij}a_{ii}a_{jj}.$$

Other terms of this type are $Q_{c2}^{(3)}$, $Q_{c2}^{(4)}$, $Q_{c5}^{(4)}$, $Q_{c6}^{(4)}$, $Q_{c7}^{(4)}$, $Q_{c10}^{(4)}$, and $Q_{c12}^{(4)}$. The term $Q_{c4}^{(3)}=\sum_i\sum_j\left(\sum_k c_{ik}c_{jk}\right)$ is not as obvious, but the term in parentheses can be reduced. The matrix with elements $\sum_k a_{ik}a_{jk}$, which is $\mathbf{A}^T\mathbf{A}$, can be

(wileyonlinelibrary.com) DOI: 10.1002/sta4.37

rewritten $(\mathbf{X}^T\mathbf{X})^T\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}$. Computing $\mathbf{G} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)$ can be performed as $O(m^2n)$, and so the final $\mathbf{GX}$ is $O(mn^2)$. Then expressing the desired result in terms of C instead of A is an $O(n)$ operation. This argument applies to $Q_{c4}^{(3)}$, $Q_{c4}^{(4)}$, $Q_{c8}^{(4)}$, and $Q_{c9}^{(4)}$. In some applications, the eigenvalues of $\mathbf{C}$ or of $\mathbf{C}^T\mathbf{C}$ may be available from parametric analysis of the data, and for this term, the following result may also be useful. $Q_{c8}^{(4)} = \sum_i \sum_j \left(\sum_k c_{ik}c_{jk}\right)^2$ has an additional simplification in terms of eigenvalues of $\mathbf{A}$. Let $\mathbf{1}_i$ denote the $1 \times n$ matrix of all zeros except the ith element, which is one. We have

$$\sum_i \sum_j \left(\sum_k c_{ik}c_{jk}\right)^2 = \sum_i \sum_j \left((\mathbf{1}_i\mathbf{C})(\mathbf{1}_i\mathbf{C})^T\right)^2 = \sum_i \sum_j \left(\mathbf{1}_i\mathbf{C}\mathbf{C}^T\mathbf{1}_i^T\right)^2 =$$

$$\cdots = \sum_i \left(\mathbf{1}_i\mathbf{C}\mathbf{C}^T\left[\sum_j \mathbf{1}_j^T\mathbf{1}_j\right]\mathbf{C}\mathbf{C}^T\mathbf{1}_i^T\right) = \sum_i \mathbf{1}_i\mathbf{C}\mathbf{C}^T\mathbf{C}\mathbf{C}^T\mathbf{1}_i^T,$$

noting that $\sum_j \mathbf{1}_j^T\mathbf{1}_j = \mathbf{I}$. Obtaining the singular value decomposition of $\mathbf{W} = \mathbf{C}\mathbf{C}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$ $Q_{c8}^{(4)} = \sum_i \left(\mathbf{1}_i\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}\mathbf{1}_i^T\right) = \sum_i \left(u_{i1}^2\lambda_1^2 + \ldots + u_{ip}^2\lambda_p^2\right)$, where $p = \min(m, n-1)$, and noting that $\sum_i u_{i1}^2 = 1$, we have $Q_{c8}^{(4)} = \lambda_1^2 + \cdots + \lambda_p^2$. Obtaining the singular value decomposition is $O(mn^2)$.

Acknowledgements

Supported in part by R01MH101819, P42ES005948, NSFC 71171105, and R01HL068890. We gratefully acknowledge the CF patients, the Cystic Fibrosis Foundation, the UNC Genetic Modifier Study, and the Canadian Consortium for Cystic Fibrosis Genetic Studies, funded in part by Cystic Fibrosis Canada and by Genome Canada through the Ontario Genomics Institute per research agreement 2004-OGI-3-05 (to P. R. D.), with the Ontario Research Fund—Research Excellence Program.

References

Baker, RD (1996), 'Testing for space–time clusters of unknown size', *Journal of Applied Statistics*, **23**(5), 543 –554.

Beasley, TM, Erickson, S & Allison, DB (2009), 'Rank-based inverse normal transformations are increasingly used, but are they merited?', *Behavior Genetics*, **39**(5), 580–595.

Bowman, AW & Azzalini, A (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford statistical science series, *Oxford University Press*, London.

Commenges, D (2003), 'Transformations which preserve exchangeability and application to permutation tests', *Journal of Nonparametric Statistics*, **15**(2), 171–185.

David, FN & Barton, DE (1966), 'Two space–time interaction tests for epidemicity', *British Journal of Preventive and Social Medicine*, **20**(1), 44–48.

Diggle, PJ, Chetwynd, AG, Haggkvist, R & Morris, S (1995), 'Second order analysis of space–time clustering', *Statistical Methods in Medical Research*, **4**, 124–136.

Elston, RC, Buxbaum, S, Jacobs, KB & Olson, JM (2000), 'Haseman and Elston revisited', *Genetic Epidemiology*, **19**(1), 1–17.

Imhof, JP (1961), 'Computing the distribution of quadratic forms in normal variables', *Biometrika*, **48**(3/4), 419–426.

Jacquez, GM (1996), 'A k nearest neighbour test for space–time interaction', *Statistics in Medicine*, **15**(17–18), 1935–1949.

Knox, G (1964), 'The detection of space–time interactions', *Applied Statistics*, **13**(1), 25–29.

Kulldorff, M & Hjalmars, U (1999), 'The Knox method and other tests for space–time interaction', *Biometrics*, **55**(2), 544–552.

Kuonen, D (1999), 'Saddlepoint approximations for distributions of quadratic forms in normal variables', *Biometrika*, **86**, 929–935.

Mantel, N (1967), 'The detection of disease clustering and a generalized regression approach', *Cancer Research*, **27**(2), 209–220.

Miller, LD, Smeds, J, George, J, Vega, VB, Vergara, L, Ploner, A, Pawitan, Y, Hall, P, Klaar, S, Liu, ET & Bergh, J (2005), 'An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival', *Proceedings of the National Academy of Sciences of the United States of America*, **102**(38), 13550–13555.

Nadaraya, EA (1964), 'On estimating regression', *Teoriya Veroyatnostei i ee Primeneniya*, **9**(1), 186–190.

Robertson, C & Nelson, AT (2010), 'Review of software for space–time disease surveillance', *International Journal of Health Geographics*, **9**(16). DOI: 10.1186/1476-072X-9-16.

Schaid, DJ, Sinnwell, JP, Jenkins, GD, McDonnell, SK, Ingle, JN, Kubo, M, Goss, PE, Costantino, JP, Wickerham, DL & Weinshilboum, RM (2012), 'Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies', *Genetic Epidemiology*, **36**1, 3–16.

Siemiatycki, J (1978), 'Mantel's space–time clustering statistic: computing higher moments and a comparison of various data transforms', *Journal of Statistical Computation and Simulation*, **7**(1), 13–31.

Tango, T (1984), 'The detection of disease clustering in time', *Biometrics*, **40**(1), 15–26.

Tong, L, Yang, J & Cooper, RS (2010), 'Efficient calculation of P-value and power for quadratic form statistics in multilocus association testing', *Annals of Human Genetics*, **74**(3), 275–285.

Watson, GS (1964), 'Smooth regression analysis', *Sankhya: The Indian Journal of Statistics, Series A*, **26**(4), 359–372.

White, ME, Schukken, YH & Tanksley, B (1989), 'Space–time clustering of, and risk factors for, farmer-diagnosed winter dysentery in dairy cattle', *Canadian Veterinary Journal*, **30**(12), 948–951.

Wright, FA, Strug, LJ, Doshi, VK, Commander, CW, Blackman, SM, Sun, L, Berthiaume, Y, Cutler, D, Cojocaru, A, Collaco, J M, Corey, M, Dorfman, R, Goddard, K, Green, D, Kent, JW, Lange, MM, Lee, S, Li, W, Luo, J, Mayhew, G & Naughton, KM (2011), 'Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2', *Nature Genetics*, **43**(6), 539–546.

Wu, MC, Lee, S, Cai, T, Li, Y, Boehnke, M & Lin, X (2011), 'Rare variant association testing for sequencing data with the sequence kernel association test (SKAT)', *The American Journal of Human Genetics*, **89**(1), 82–93.

Zhou, YH & Wright, FA (2013), 'Empirical pathway analysis, without permutation', *Biostatistics*, **14**(3), 573–585.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.