Published in final edited form as:
Comput Stat Data Anal. 2009 August 1; 53(10): 3640–3649. doi:10.1016/j.csda.2009.03.005.

The Choice of the Number of Bins for the *M* Statistic*

Laura Forsberg White[a,*], **Marco Bonetti**[b], and **Marcello Pagano**[c]

[a] Department of Biostatistics, Boston University School of Public Health, 715 Albany St, Boston, MA 02118

[b] Department of Decision Sciences, Univerisita' Bocconi, Milano, Italy

[c] Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, MA 02115

Abstract

Methods to monitor spatial patterns of disease in populations are of interest in public health practice. The *M* statistic uses interpoint distances between cases to detect abnormalities in the spatial patterns of diseases. This statistic compares the observed distribution of interpoint distances with that which is expected when no unusual spatial patterns exist. We show the relationship of *M* to Pearson's Chi Square statistic, x_n^2. Both statistics require the discretization of continuous data into bins and then are formed by creating a quadratic form, scaled by an appropriate variance covariance matrix. We seek to choose the number and type of these bins for the *M* statistic so as to maximize the power to detect spatial anomalies. By showing the relationship between *M* to x_n^2, we argue for the extension of the theory that has been developed for the selection of the number and type of bins for x_n^2 to *M*. We further show that spatial data provides a unique insight into the problem through examples with simulated data and spatial data from a health care provider. In the spatial setting, these indicate that the optimal number of bins depends on the size of the cluster. For large clusters, a smaller number of bins appears to be preferrable, however for small clusters having many bins increases the power. Further, results indicate that the number of bins does not appear to vary with *m*, the number of spatial locations. We discuss the implications of this result for further work.

Keywords

Dependent data; Distances; Pearson's Chi Square Statistic; Spatial Statistics; Surveillance

1 Introduction

Syndromic surveillance and a growing interest in real time monitoring the patterns of diseases in populations has led to increased research in novel methods for detecting spatial anomalies. Among the methods being developed is the *M* statistic [1]. This statistic monitors the distribution of the interpoint distances between cases and is designed to detect deviations from expected behavior in this distribution. The *M* statistic has been shown to be effective at detecting exogenous clusters when compared with other statistics [1,2,3] designed for this same

*Research supported in part by grants from the National Institutes of Health AI28076 and T32 AI007358, and NLM ILM007677.

* Corresponding author, Tel: +1-6174142833; Fax: +1-6176386484, email: lfwhite@bu.edu.

purpose. It claims some advantages, including the ability to detect a broad range of deviations, not just clusters. Additionally, it does not suffer from the curse of dimensionality and it does not make any parametric assumptions, or impose restrictions on the format of the spatial data (for example, exact locations, or aggregated over zip code, census tract, or some other region). A bivariate version of this statistic, incorporating temporal information with the spatial data, leads to greater power to detect aberrant behavior [4].

The M statistic is constructed by considering all the interpoint distances between m independent cases. Thus there are $N=\binom{m}{2}$ interpoint distances used in calculating the statistic. In order to check for goodness-of-fit, data is discretized into the $k \times 1$ vector of the frequency distribution of the observed (dependent) distances and the vector is compared to the corresponding $k \times 1$ frequency vector expected under the null hypothesis via the quadratic form,

$$M=(\mathrm{o}_N - \mathrm{e})^\top S^- (\mathrm{o}_N - \mathrm{e})$$

where o_N is the observed vector of the frequencies of the distances, e is the vector of expected frequencies, and S^- is the generalized inverse of the variance covariance matrix of o_N. This statistic can be described as a Mahalanobis distance between the observed and expected distribution of the distances, once discretized through the creation of the k classes.

It is also useful to see this statistic as a generalization of Pearson's Chi Square statistic, X_n^2 [5]. Pearson's statistic does exactly the same operation as M, but assumes that the underlying observations are independent. Thus, in that case, the observed $k \times 1$ vector of the observed relative frequencies is distributed as a multinomial random variable. That statistic can be written as

$$X_n^2=(\mathrm{x}_n - \mathrm{e})^\top \sum\nolimits_{MN}^{-} (\mathrm{x}_n - \mathrm{e})$$

where x_n is a multinomial random vector and $\sum_{MN}^{-}$ (here, MN indicates the multinomial distribution) is the generalized inverse of the corresponding variance covariance matrix of x_n. This variance covariance matrix has the standard form $\sum_{MN}=\{e\mathrm{I}_k - (1/n)\mathrm{ee}^\top\}$.

Both statistics have several appealing features. Among these are that they are intuitive and easy to interpret. It is also reasonable to assume that greater power is gained to detect aberrations from null behavior when the entire distribution is compared, as opposed to comparing a simple summary statistic, such as the mean or maximal value, as in the Kolmogorov-Smirnov test [6]. A challenge with X_n^2 and M arises when working with continuous, or nearly continuous data. In this case, one must aggregate the continuous data into k bins. The optimal method of doing this has been studied for X_n^2 (see references cited in the next section). This is an important consideration, as values for the power have been shown to vary dramatically, both for M [2] and X_n^2 [7] when k varies. Intuition might suggest that we select k to be large, so as to more closely represent the continuous distribution. However, there is a tradeoff for choosing large values for k, as this leads to an increase in the variance and the possibility of having bins with no observations. Further the particular anomaly to be detected in the data may dictate the optimal number of bins, as we show below.

In the rest of this paper, we present two main results. First, the *M* statistic is indeed essentially equivalent to X_n^2. This also implies that any appropriately constructed quadratic form is equivalent to X_n^2, regardless of the nature of the underlying observations. Second, we explore the implication of this result for applying the theory that has been developed for X_n^2 with regard to the selection of *k* for spatial data using the *M*. Specifically, in the cluster detection setting with *M*, we show that the optimal number of bins depends on the size of the cluster to be detected. We discuss the implications of this result. First, however, we provide a summary of relevant work on selecting *k* for X_n^2.

2 Bin Selection for X_n^2

Pearson's Chi Square statistic was introduced at the beginning of the 20th century and has been widely used since its introduction [5]. It is most suited for goodness-of-fit tests with discrete data, but is frequently extended for use with continuous data. We assume in what follows that the null hypothesis is of the form $H_0 : X \sim f(x; \theta_0)$, with both f and θ_0 fully specified. The null distribution of the X_n^2 statistic when testing for a specific distribution depends on the method of estimation of the parameters, if any are being estimated as part of the test, for example when testing the hypothesis that the observed data arises from a normal distribution without specifying the mean and variance [8]. When one uses the statistic with continuous data it then becomes necessary to categorize the data, and there are two main issues to consider in this process: how the bins should be partitioned (for instance equiprobable, equispaced or some other configuration) and how many bins, *k*, should be used. The focus of this paper addresses the latter question, however we begin by providing some comments on the former.

An analogous process to determining the type of bins to use is performed when creating a histogram of continuous data. In most of the popular statistical computing packages (for instance Splus/R, Stata and SAS), the default setting for this is to use equispaced bins when creating the histogram. One might consider this same process when using X_n^2 for continuous data. The benefit of this method is that when the aberration we wish to detect occurs in one of the cells with a small expected count, we are more likely to detect even a small absolute change. For instance, if only one observation is expected in each cell, then it would not take a large increase for a strong signal to be generated. However, a major problem of this method is that inevitable disparities in the cell counts will be created. If one cell contains a large portion of the data, while the remaining cells are sparse, then subtle changes between the null and alternative distributions could be masked within the large cell, and power may be compromised. Additionally, having cells with no observations is not only inefficient, but also leads to instability in the statistic, since X_n^2 is calculated by dividing by the expected number in each cell. If that number is approaching zero for at least one cell, then we can expect that the statistic will behave aberrantly. For instance, [9] cites research showing that the asymptotic behavior of X_n^2 is not as expected, giving some explanation for the well-known rule of thumb of having at least five as the expected count in each cell (though subsequent work has shown this particular cutoff to be conservative [10]).

As an alternative to equispaced cells, [7] suggest that equiprobable cells are probably better, unless there is a particular alternative that one is interested in detecting. In that case, it is optimal to use equiprobable cells, except in the area closest to the anticipated aberration; in that vicinity one should use more cells. For instance, if the type of alternative we are interested in detecting would lead to deviations in values on the extremes of the distribution, we should consider increasing the number of bins in the area at the extremes of the data. The use of equiprobable binning is also recommended when using the *M* statistic, as work with that statistic suggests that this is more powerful than the equispaced approach [2,3].

We now focus our attention on the selection of k, the number of bins. The goal is to maintain a balance between accurately portraying the distribution by having a sufficient number of bins, but without allowing that number to be so high as to create excessive variability. Table 1 provides a brief summary of some of the more conclusive papers addressing this topic for the X_n^2 test. Mann and Wald provide a rule that is easy to implement in this problem. They prescribe a formula that has k increase with $n^{2/5}$. However much of the subsequent work showed that their approach often prescribes too many bins [11]. We give particular attention to the work done by [7] and [12]. This work describes the asymptotic behavior of the power of X_n^2 as both k and n tend to infinity. We find this particularly useful for understanding how to select k with respect to n for scenarios that have not been previously characterized, as the results are general and can be applied to detect any alternative against a given null distribution.

In many of the cases for X_n^2, research indicates that k should increase with n, though $k = n$ is problematic as sparse cell counts will occur. Exceptions occur when testing for normality as shown in [13] and when the alternative has light tails compared to the hypothesized distribution [7,12]. In these cases, small k is optimal for all values of n. The work presented here indicates that when testing for spatial clusters with the M statistic the value of k does not appear to depend on m, the number of independent spatial locations, as much as it depends on the size of the cluster to be detected. In the following sections we give the theoretical underpinning for the extension of the X_n^2 results to the M statistic and then illustrate it through examples.

3 Dependent Data

The above discussion on bin selection deals with the case where the continuous observations on which the frequencies are calculated are independent and identically distributed. We now seek to show their applicability to the case of observations that are not independent. This is motivated by work done in spatial statistics. [1,14] have studied a statistic designed to detect spatial anomalies. This statistic is being used in syndromic surveillance [4], as well as genetics research [15,16]. The statistic considers the distribution of all of the interpoint distances between observations. For the case of syndromic surveillance, we consider the Euclidean distances between cases with a syndrome of interest. In the genetics setting, one uses genetic distances between, for example, viruses. In what follows, $N = \binom{m}{2}$ is the number of interpoint distances between the m independent spatial locations. The matrix S has been shown empirically to be of rank $k - 1$ in most cases. We use the generalized inverse described in [17, p. 27]. This generalized inverse is obtained by partitioning the matrix as

$$S = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where the rank of A is equal to that of S. The generalized inverse is computed as

$$S^- = \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

As we pointed out, the M statistic bears strong resemblance to X_n^2. In what follows we show that that these two statistics are essentially equivalent and therefore argue that the theory that

has been developed for X_n^2 can be applied to M. These results can be generalized to any quadratic form with an appropriate variance covariance matrix. Some consideration, however, must be given to the rate at which the asymptotic results apply. This is because with dependent data generated from m independent observations, we observe a much larger number of dependent observations (of order m^2). We can rewrite the statistics M and X_n^2 as

$$\begin{aligned} M &= (\mathrm{o}_N - \mathrm{e})^\top \begin{pmatrix} S_1 & \mathrm{s}_2 \\ \mathrm{s}_2 & s_3 \end{pmatrix}^- (\mathrm{o}_N - \mathrm{e}) = (\tilde{\mathrm{o}}_N - \tilde{\mathrm{e}})^\top S_1^{-1} (\tilde{\mathrm{o}}_N - \tilde{\mathrm{e}}) \\ &= \tilde{\mathrm{a}}_N^\top S_1^{-1} \tilde{\mathrm{a}}_N \end{aligned}$$

$$\begin{aligned} X_n^2 &= (\mathrm{x}_n - \mathrm{e})^\top \begin{pmatrix} \sum_1 & \sigma_2 \\ \sigma_2 & \sigma_3 \end{pmatrix}^- (\mathrm{x}_n - \mathrm{e}) = (\tilde{\mathrm{x}}_n - \tilde{\mathrm{e}})^\top \sum_1^{-1} (\tilde{\mathrm{x}}_n - \tilde{\mathrm{e}}) \\ &= \tilde{\mathrm{z}}_n^\top \sum_1^{-1} \tilde{\mathrm{z}}_n, \end{aligned}$$

where a tilde represents a $k - 1$ dimensional vector.

We use the Cholesky decomposition and let

$$\begin{aligned} S_1 &= L^\top L, \text{ so that } S_1^{-1} = (L^{-1})(L^{-1})^\top \\ \sum_1 &= T^\top T, \text{ so that } \sum_1^{-1} = (T^{-1})(T^{-1})^\top. \end{aligned}$$

Then we can rewrite M as

$$\begin{aligned} M &= \tilde{\mathrm{a}}_N^\top L^{-1} (L^{-1})^\top \tilde{\mathrm{a}}_N = \tilde{\mathrm{a}}_N^\top L^{-1} T T^{-1} (T^{-1})^\top T^\top (L^{-1})^{-\top} \tilde{\mathrm{a}}_N \\ &= \tilde{\mathrm{a}}_N^\top L^{-1} T \sum_1^{-1} T^\top (L^{-1})^{-\top} \tilde{\mathrm{a}}_N = \tilde{\mathrm{b}}_N^\top \sum_1^{-1} \tilde{\mathrm{b}}_N. \end{aligned}$$

Given the transformation $\tilde{\mathrm{b}}_N = T^{-1}(L^{-1})^\top \tilde{\mathrm{a}}_N$, Σ_1 is the variance covariance matrix of $\tilde{\mathrm{b}}_N$:

$$\begin{aligned} \mathrm{var}(\tilde{\mathrm{b}}_N) &= \mathrm{var}(T^\top (L^{-1})^{-\top} \tilde{\mathrm{a}}_N) = T^\top (L^{-1})^{-\top} \mathrm{var}(\tilde{\mathrm{a}}_N) L^{-1} T \\ &= T^\top (L^{-1})^{-\top} L^\top L L^{-1} T = T^\top T = \sum_1. \end{aligned}$$

Therefore we see that M and X_n^2 are essentially equivalent. Both are quadratic forms constructed from vector random variables that (when normalized) converge in distribution to multivariate normal distributions.

This equivalence motivates our use of the theory for the behavior of X_n^2 to M_N as $k(m)$ (or $k(N)$) $\to \infty$. Our objective is to better understand the optimal choice for k, given m. [7] and [12] provide useful information on the behavior of the power of X_n^2 as $k(n) \to \infty$. We can use this information to better understand how changing k with respect to m might impact the power for M_N. One should keep in mind however that because of the dependence among distances, the asymptotic behavior does not follow the usual iid rules (see also [1]).

The results with X_n^2 provided by [12] and [7] suggest that k should increase with m in many cases, however there are exceptions, as we show. Simulations indicate that in fact the optimal number of bins depends on the size of the cluster. We illustrate this point by example in the following section. We also observe that the optimal value for k does not appear to change as m increases.

4 Empirical Study

We provide empirical examples of the effect of varying k on the power to detect a spatial cluster. The first examples arise from data of simulated locations based on a parametric model. A final example comes from a real dataset composed of the locations, aggregated by census tract, of patients with upper respiratory tract infections seeking care at an eastern Massachusetts healthcare provider. In both examples, spatial clusters are superimposed on the data and the power to detect these clusters with the M statistic is estimated empirically. The question that we address is how to determine the value of k for a given value of m such that the power to detect the cluster will be maximized.

4.1 Simulated Data

The first examples that we present arise from simulated data. There are three scenarios considered here. First, locations are simulated on the unit line and distances are calculated as the difference between these values. Second, data is simulated from the uniform distribution on the unit square. The last model simulates cases according to a bivariate normal distribution, centered at the origin with variance covariance matrix given by the identity, $\mathbf{I}_2$. For each of these models, m assumes four fixed values, 50, 100, 150 and 200 and we vary k over 31 values between two and 3000.

The bin breaks, S matrix and null distribution of the M statistic are calculated using resampling methods. The datasets used to calculate power contain a cluster placed in a region that composes 0.1, 0.01, 0.001, and 0.0001 of the probability space (these are heretofore denoted as p_0). The number of points placed in the cluster region is $p_A m$, where p_A is the probability of being in the cluster under the cluster model and varies depending on the scenario in order to bound the power away from α and 1. In the case of multiple clusters, the total probability space used by the clusters is $2p_0$. Diffuse clusters were also considered. In this case 0.1 was the only value considered for p_0, so that there was a detectable difference between the expected and observed number of points in the region. The details of these clusters are provided below. An illustration of some of the unit square and bivariate normal configurations are shown in Figure 1. Results from all of the simulations are given in Figures 2 and 3.

Unit Line—In the first example, data are generated from a uniform distribution on the interval [0, 1]. The cluster comes from a uniform random variable with positive probability centered at 0.50. The probabilities of points being in the cluster region under the cluster model, p_A, were 0.2253, 0.0700, 0.0600, and 0.0500 to correspond with the four values of p_0. Additionally, diffuse clusters were simulated when $p_0 = 0.1$ with $p_A = 0.01$. Simulations were also run with the cluster location varying, though the results are not shown here since the location of the cluster did not seem to impact the results for the optimal value for k. Additionally a double cluster model was run with clusters centered at 0.1 and 0.6 and $p_A = 0.1510$, 0.0469, 0.0300, and 0.0335 for each of the four values of p_0. Figure 2 gives the estimated powers for the dense cluster simulations. Figure 3 shows the power for the diffuse cluster scenario. We see that as the size of the cluster decreases, the optimal number of bins increases. Additionally, as m increases the overall power increases, which is not surprising since we would expect the precision to increase.

Unit Square—The third and fourth rows of plots in Figure 2 illustrate the power to detect a dense cluster superimposed on spatial data uniformly distributed on a unit square. Four single cluster models were simulated with clusters centered at (0.5, 0.5), (0.25, 0.25), (0.16, 0.16) (on the corner of the region), and (0.5, 0.25). We report only those results from the cluster being centered at (0.5, 0.5) since the results were consistent, regardless of cluster location. The values of p_A considered were 0.2253, 0.1000, 0.0600, and 0.0500. Additionally we consider a double cluster model with clusters centered at (0.25, 0.25) and (0.75, 0.50) and the values of p_A being 0.67 the values of p_A for the single cluster model. Diffuse clusters were also generated when $p_0 = 0.1$, with $p_A = 0.01$ for one and two cluster models. In all scenarios, the optimal value of k increases as the size of the cluster decreases, regardless of m and power increases with m.

Bivariate Normal—The final simulated example arises from data that are generated from a bivariate normal distribution centered at the origin and with variance covariance matrix of $\mathbf{I}_2$. The values for p_A used in this scenario are 0.2253, 0.1000, 0.0600, and 0.0500. The points of the clusters are obtained by identifying a square region centered at the origin that contains approximately p_0 of the probability space. A large number of points are sampled from the bivariate normal distribution and $(1 - p_A)m$ of the points that fall outside the cluster region are randomly sampled without replacement. All points that are in the cluster region are used and additional points are generated from simulating data uniformly on the square to create $p_A m$ points in the cluster region. Again, the optimal value of k increases with a decreasing cluster size regardless of m and power increases with m.

From these examples the optimal value of k does not appear to be increasing with m. Also, the best performing value of k depends on the cluster size. Figure 4 illustrates the reason for this, which is that a key disturbance created in the distance distribution comes from the interpoint distances within the cluster. So, clusters that are very small create a bump in the distribution around very small distances. As the cluster size increases, the area of the distribution that is affected increases. Power is likely to be optimized when the disturbance is contained within a single bin. Figure 4 illustrates this phenomenon and Table 2 shows the length of the first bin compared to the size of the cluster. The dependence of the optimal value of k on the size of the cluster poses challenges for exploratory work and generally in surveillance settings where the size of the cluster that one is interested in is likely not known a priori. Additionally it is not clear how to best choose k if the spatial abnormality is not a cluster, but some other modification in the distribution.

4.2 Biosurveillance Data

This study is based on data collected over a period of 1399 days from a large healthcare provider in eastern Massachusetts. The data contains the locations, aggregated by census tract, of patients with upper respiratory tract infections (URI). This information is downloaded from a central database daily. The data has been previously analyzed in [4]. Days with case counts less than six were excluded from the analysis. The average daily case count for the days included in the analysis was 40.12 (sd = 22.30). We use this data to calculate the bin breaks, S matrix and null distribution of M.

In this example, m is varying from day to day. Additionally, the location data is aggregated. The impact of these two features on the choice of k is not clear. Therefore, these simulations are useful for observing how k behaves in this more realistic scenario.

We consider three alternative models. The first two contain single clusters of size six. The final model is a combination of the previous two, containing both clusters used in the previous example simultaneously, in other words, there are 12 points that are in clusters. For each of these three examples, k ranged over eight values between 2 and 50. The results of these simulations are given in Figure 5. The power curves demonstrate the same basic pattern

observed in the previous simulated examples. That is, we see a sharp initial increase in power as *k* increases and then a very gradual decline as *k* approaches and extends beyond *m*.

These results are consistent with those observed for the simulated data suggesting that the clusters comprise a relatively large part of the probability space. In fact, based on our observations here, it is reasonable to suggest a test that is performed for varying *k* and the final result chosen to be for the value of *k* that optimizes some criteria, analogous to the AIC criterion. For cluster models, this appears to be a reasonable approach. Additionally, we note that the differences in power from choosing *k* within a reasonably broad range of values are not dramatic.

5 Discussion

We noted the equivalence of the *M* and X_n^2 statistics, implying that the underlying distribution of the observations is unimportant, as long as appropriate variance covariance matrix is used when constructing the quadratic form. In other words, we can say that X_n^2 is a special case of a broader class of statistics composed of quadratic forms that can be constructed as

$$Q = \mathrm{a}_n^{\top} \sum^{-} \mathrm{a}_n$$

where Σ is the $k \times k$ variance covariance matrix, with rank $k - 1$ of the $k \times 1$ vector **a**. In the case of X_n^2, the vector of data follows a multinomial distribution and the variance covariance matrix takes on a known form. On the other hand, for the M statistic, in general, it difficult to specify the variance covariance matrix. In practice we use resampling methods to estimate the variance covariance matrix. In simple cases, such as the unit line, we can obtain an exact form of the variance covariance matrix.

Pearson's Goodness-of-Fit test and the *M* statistic compare an observed distribution with a hypothesized distribution, the null distribution, and quantify the differences between the two. While these statistics are intuitive and very useful, there can be some difficulty in implementing them with continuous data. As we have seen, the power of the test is impacted substantially by the choice of the type and number of classes.

We confirm the recommendation of using of equiprobable binning also with *M*, unless there is prior interest in detecting a particular alternative. In that case one should use a finer probability grid in the vicinity of the expected deviation and a coarser equiprobable grid elsewhere.

The power of the test is affected by the selection of the number of discrete categories, *k*. Our simulations indicate that in the case of spatial cluster models, the best performing number of bins depends on the size of the cluster to be detected. This would argue for a modification to the *M* statistic that would allow one to search over a range of values of *k*. The optimal choice of *k* will also be informative in indicating the size of the cluster that has been detected. Additionally, our simulations affirm that overall power increases as *m* increases, regardless of the choice of *k*.

We recognize that these guidelines have limitations. First, the results are based on asymptotic theory of the behavior of the power as *k* increases with respect to *n* or *m*. In application, the point where the asymptotic results provide an accurate guide may not be attained. Therefore, it is possible that the rule of thumb given may not provide the best value of *k* for a practical dataset. It is hoped that by examining simulations, we can shed some insight on finite sample

behaviors. It should also be noted that the results obtained for X_n^2 often contradict each other. For instance, [7] and [12] approach the same problem of describing the behavior of k as n goes to infinity and find results that mostly coincide, but do differ significantly in some cases. This, and the large body of literature that has attempted to address this problem, lead to the conclusion that this is clearly not a simple issue, and that to this point it has not yet been addressed in a definitive way. We contend, however, that the literature that does exist and the results shown in this paper are sufficient to provide guidelines that appear to be reasonable for this problem.

References

1. Bonetti M, Pagano M. The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. Statistics and Medicine 2005;24:753–773.
2. Ozonoff A, Bonetti M, Forsberg L, Pagano M. Power comparisons for an improved disease clustering test. Computational Statistics and Data Analyis 2005;48:679–684.
3. Kulldorff M, Tango T, Park P. Power comparisons for disease clustering tests. Computational Statistics and Data Analyis 2003;42:665–684.
4. Ozonoff A, Forsberg L, Bonetti M, Pagano M. A bivariate method for spatio-temporal syndromic surveillance. MMWR 2004;53(Suppl):61–66.
5. Pearson K. On the criterion that a given system of deviations from the probable case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag 1900:157–172.50
6. Hollander, M.; Wolfe, D. Nonparametric Statistical Methods. John Wiley and Sons; 1999.
7. Kallenberg W, Ooosterhoff J, Schriever B. The number of classes in chi-squared goodness-of-fit tests. Journal of the American Statistical Association 1985;80:959–968.
8. Kendall, M.; Stuart, A.; Ord, J.; Arnold, S.; O'Hagan, A. Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory. Oxford University Press, Inc.; 1994.
9. Cochran WG. The χ^2 test of goodness of fit. Annals of Mathematical Statistics 1952;23:315–345.
10. Koehler K, Larntz K. An empirical investigation of goodness-of-fit statistics for sparse multinomials. Journal of the American Statistical Association 1980;75:336–1278.
11. Williams CA Jr. On the choice of the number and width of classes for the chi-square test of goodness of fit. Journal of the American Statistical Association 1950;45:77–86.
12. Kallenberg W. On moderate and large deviations in multinomial distributions. The Annals of Statistics 1985;13:1554–1580.
13. Dahiya RC, Gurland J. How many classes in the pearson chi-square test? Journal of the American Statistical Association 1973;68:707–712.
14. Bonetti, M.; Pagano, M. On detecting clustering. Proceedings Biometrics Section, American Statistical Association; 2000. p. 37-44.
15. Graham D, Pagano M. Dimension reduction of hiv genotype with application to modeling the relationship between genotype and phenotype, in preparation.
16. Kowalski J, Pagano M, DeGruttola V. A non-parametrc test of gene region heterogeneity associated with phenotype. Journal of the American Statistical Society 2002;97:398–408.
17. Rao, C. Linear Statistical Inference and Its Applications. John Wiley and Sons; 1973.
18. Mann H, Wald A. On the choice of the number and width of classes for the chi-square test of goodness of fit. Annals of Mathematical Statistics 1942;13:306–317.
19. Oosterhoff J. The choice of cells in chi-square tests. Statistica Neerlandia 1985;39:115–128.
20. Koehler K, Gann F. Chi-squared goodness-of-fit tests: Cell selection and power. Communications in Statistics-Simulation 1990;19:1265–1278.

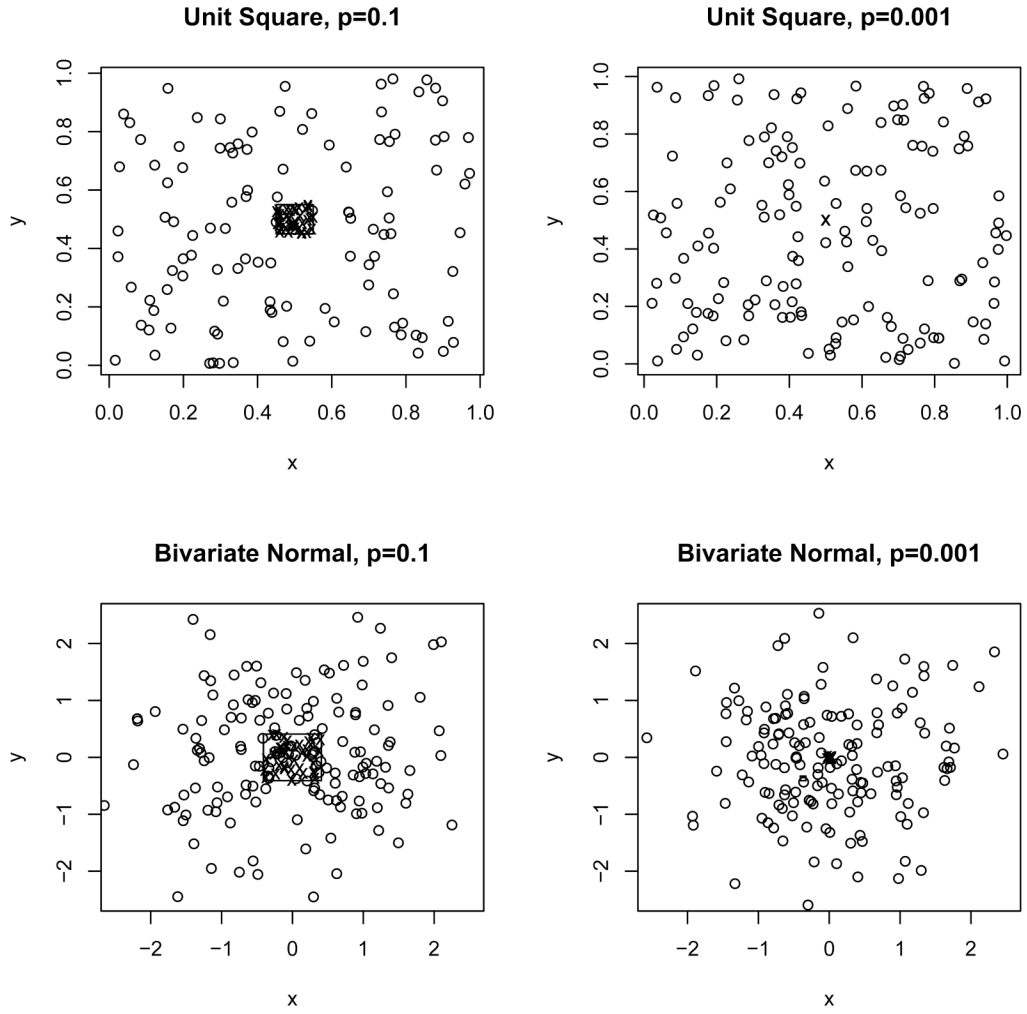

Fig. 1.
Configuration for the simulations for the Unit Square and Bivariate Normal scenarios when $p_A = 0.1$ or $p_A = 0.001$. The x's denote the points in the cluster and the o's denote non cluster points.

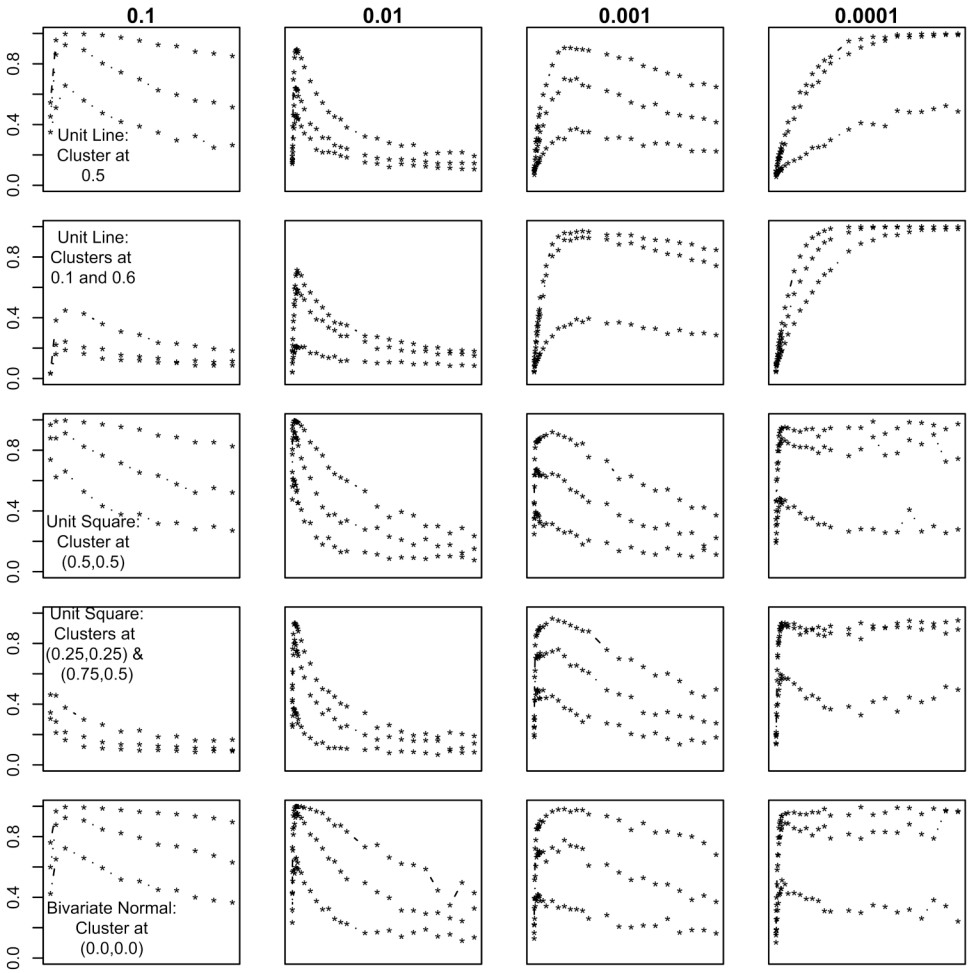

Fig. 2.

Power for the dense cluster model simulations. Within each plot the powers for m = 50, 100, 150, and 200 are shown. Each column represents a different value for p_0, indicated by the title. Power decreases with m. Each row of plots corresponds to a different clustering scenario. The x axis for each plot denotes values of k, where k ranges between 2 and 100 for p_0 = 0.1 and 2 and 3000 for p_0 = 0.01, 0.001, and 0.0001.

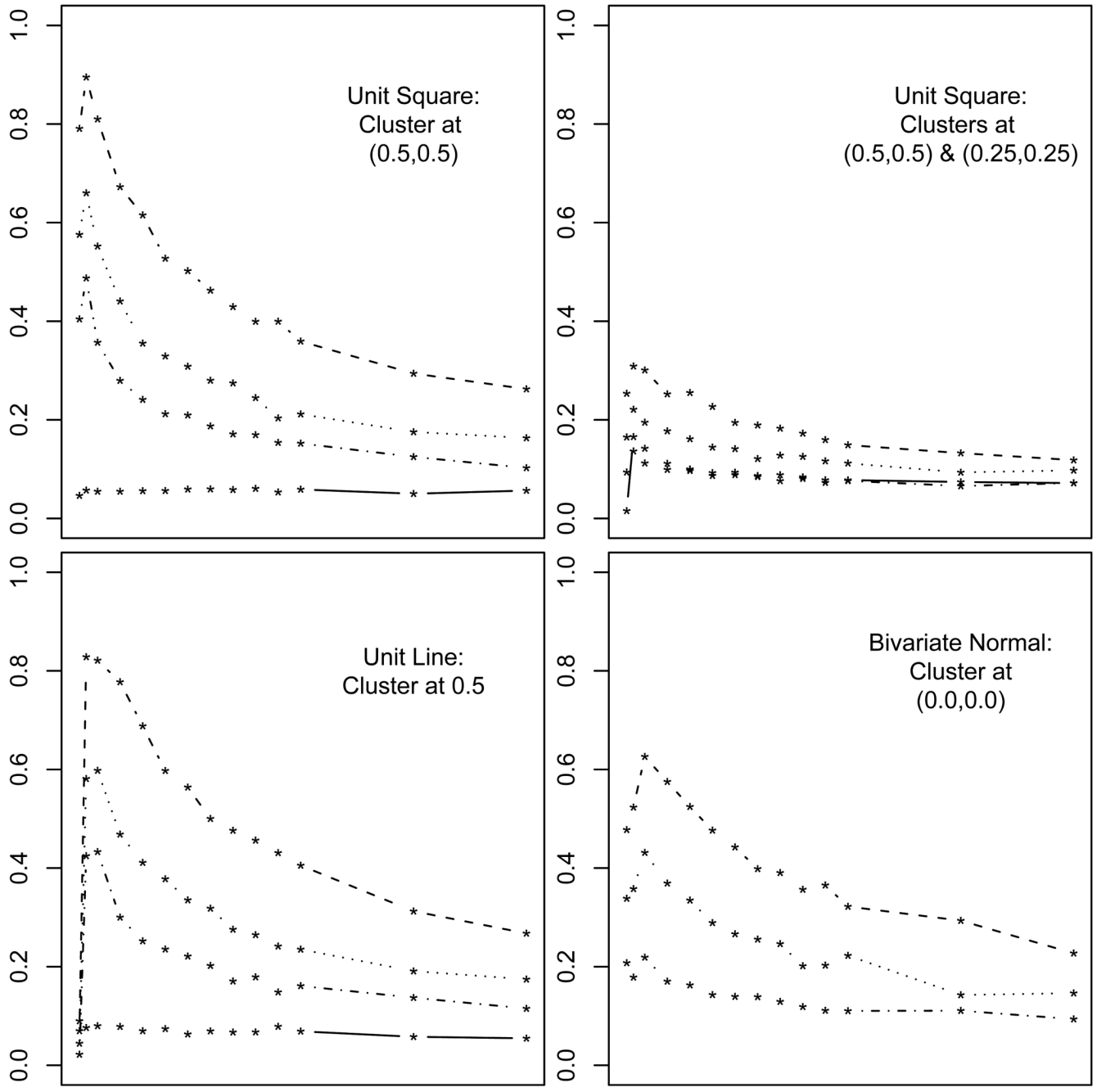

Fig. 3.

Power for the diffuse cluster model simulations. Within each plot the power for m = 50, 100, 150, and 200 are shown. Power is monotonically decreasing with m. The x axis denotes values of k, where k ranges between 2 and 3000.

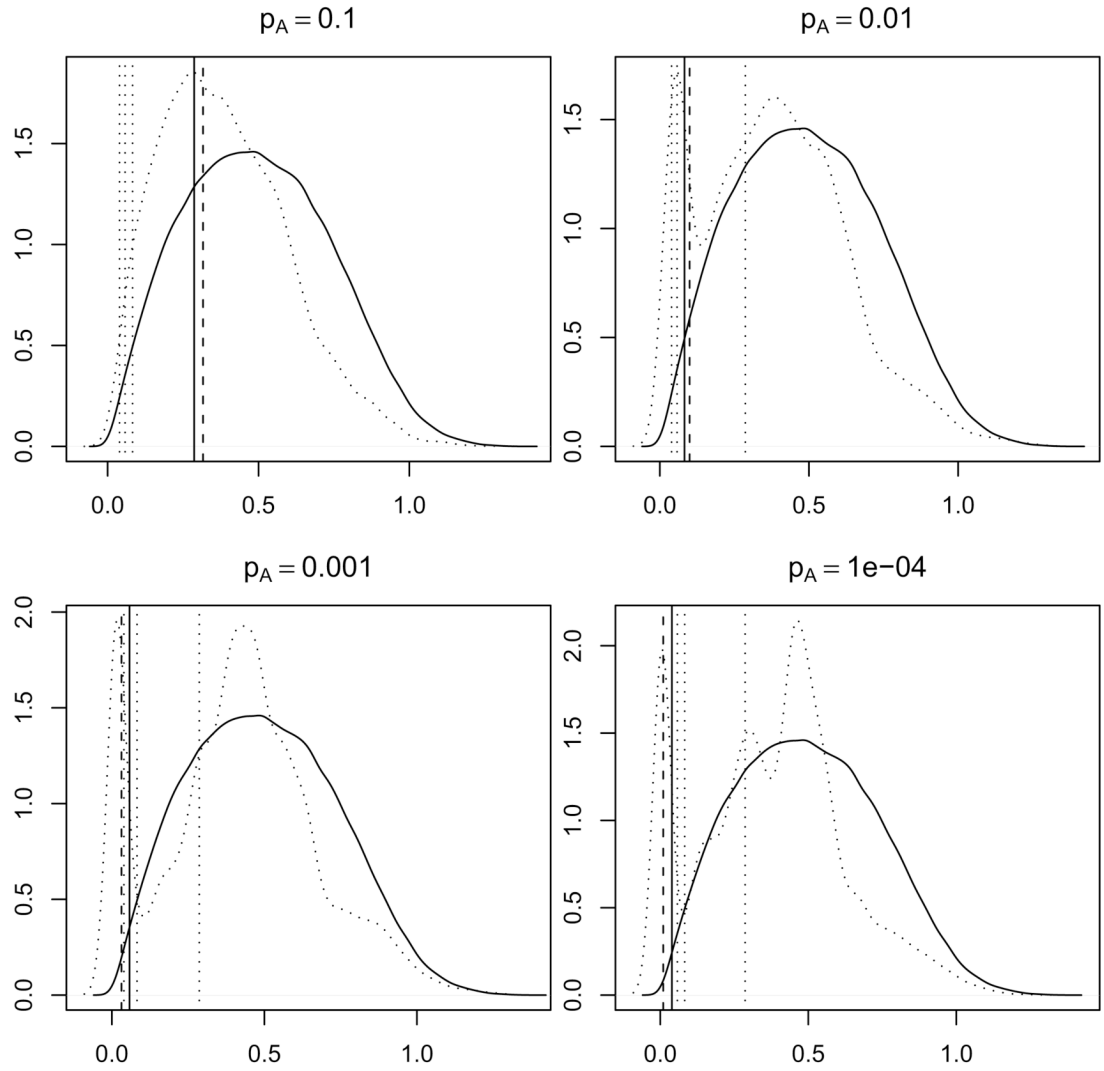

Fig. 4.

The null (black) and alternative densities (dotted) of the distances for the unit square simulations. The x axis represents distance. The width of the first bin for k = 5, 50, 100, and 200 are shown as vertical dotted lines with the optimal width being shown in black. The width of the cluster is shown with the dashed line.

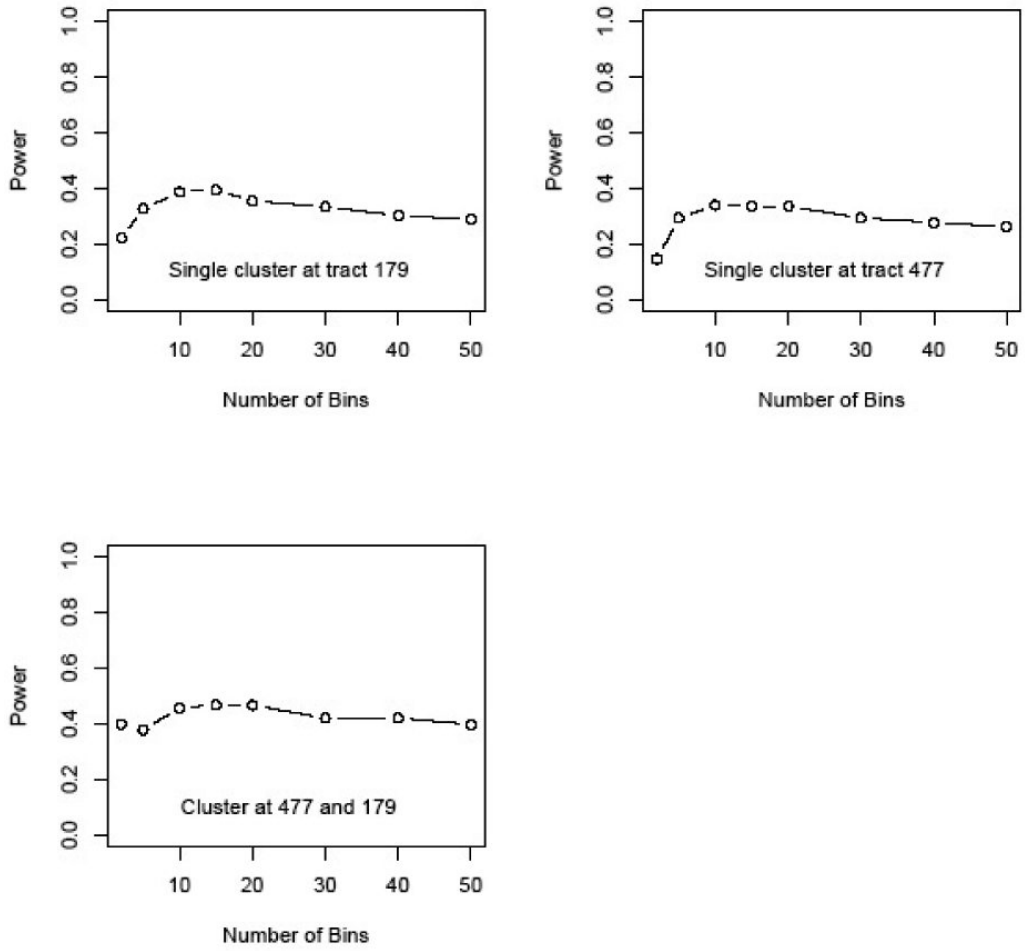

Fig. 5.
Power for three cluster models with data from health care provider, as *k* is varied.

Table 1

Summary of literature pertaining to cell selection for the X_n^2 statistic.

Author	Recommendation	Comments
[18]	$k = 4\left[\frac{2(n-1)^2}{c^2}\right]^{1/5}$	First proposal for value of k. Oversimplifies the problem.
[13]	k does not seem to vary with n. k does impact the power.	Only testing for normality. Did not consider many values of n.
[7]	Large number of bins best for alternative with heavy tails. If lighter tails or if tails are similar and there is a mean shift, then a smaller number number of bins is better. If have one light tail and one heavy tail, then a moderate number of bins is best.	Done by considering local alternatives and looking at the noncentrality parameter. Simulation results seem consistent with theoretical results.
[12]	k should always increase with n. When the non centrality parameter converges to a finite limit, k should increase more slowly with n.	Theory derived by considering nonlocal alternatives. Simulations are consistent with theory. Further simulations show that this result is more accurate than that of [7].
[19]	If not testing for a specific alternative, allow k to increase with n.	
[20]	k should increase slowly with n, so that expected cell counts increase with n. Recommend k of seven when n is 20, 10 – 15 for n of 50 and 20 when n is 100.	Testing for normality. Results based on simulations.

Table 2

Best performing number of bins, the length of the cluster, and the length of the first bin.

Model	p_0	Best k	Cluster Length	Length of first bin
Unit	0.1	10	0.1	0.0513
Line	0.01	80	0.01	0.0062
	0.001	700	0.001	0.0008
	0.0001	3000	0.0001	0.0002
Unit	0.1	5	0.3162	0.2898
Square	0.01	40	0.1000	0.0933
	0.001	200	0.0316	0.0390
	0.0001	1600	0.0100	0.0151
Bivariate	0.1	10	0.8180	0.1915
Normal	0.01	70	0.2500	0.0714
	0.001	500	0.0800	0.0472
	0.0001	1400	0.0300	0.0395

Note that the maximal value for k was 3000 in the simulations.