Biometrics 000, 000–000 DOI: 000
000 0000

Asymptotic distribution of score statistics for spatial cluster detection with censored data

Daniel Commenges[1,2,*] and Benoit Liquet[1,2]

[1] INSERM, Epidemiology and Biostatistics Research Center, Bordeaux, F33076, France

[2] Université Victor Segalen Bordeaux 2, Bordeaux, F33076, France

email: daniel.commenges@isped.u-bordeaux2.fr

Summary: Cook, Gold and Li (2007) extended the Kulldorff (1997) scan statistic for spatial cluster detection to survival-type observations. Their approach was based on the score statistic and they proposed a permutation distribution for the maximum of score tests. The score statistic makes it possible to apply the scan statistic idea to models including explanatory variables. However, we show that the permutation distribution requires strong assumptions of independence between potential cluster and both censoring and explanatory variables. In contrast we present an approach using the asymptotic distribution of the maximum of score statistics in a manner not requiring these assumptions.

Key words: Asymptotic distribution; Cluster detection; Generalized linear model; Permutation test; Score test; Spatial scan statistic; Survival data.

1. Introduction

Cook, Gold and Li (2007) introduced a score statistic approach for spatial cluster detection with survival-type observations. This can be viewed as an extension of the scan statistic proposed by Kulldorff (1997) who proposed using the maximum over likelihood ratio statistics associated with potential clusters. Huang, Kulldorff and Gregorio (2007) recently extended the scan statistic to survival type data, but in a parametric framework and without adjusting on covariates. The inference was based on the permutation distribution of that statistic. The score statistic leads to simpler computations because the models under the alternative hypotheses do not need to be fitted. Cook, Gold and Li (2007) have also adopted the permutation approach for computing the p-value of the scan statistic based on the score. However, the permutation tests need an assumption of exchangeability which may not hold in this sort of application. Cook, Gold and Li (2007) proposed another statistic $\hat{W}_{loc}$ that we do not consider in this note.

The aim of this paper is to demonstrate the advantage of using the score statistics in this problem, to examine the exchangeability assumption, especially in the presence of censoring and covariates, and to contrast the permutation approach with an asymptotic approach proposed by Hashemi and Commenges (2002) in another context but which can be applied to the cluster detection problem.

2. Score test statistics

In the context of survival data, Huang, Kulldorff and Gregorio (2007) proposed a spatial scan statistic based on the likelihood ratio statistic. Nevertheless, their study was limited to a parametric assumption for the baseline risk function. To avoid this parametric hypothesis, Cook, Gold and Li (2007) defined a scan statistic based on normalized score statistics (or score test statistics), themselves based on Cox partial likelihood. Specically, consider the set

of Cox's proportional hazards models $\mathcal{M}_k$ for the possible effect associated with potential cluster k:

$$\lambda(t|Z_i(k), \mathbf{X_i}) = \lambda_0(t) \exp\left[\beta(k) Z_i(k) + \gamma^T \mathbf{X_i}\right] \quad i = 1, \ldots, n, \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, $Z_i(k)$ is an indicator with a value 1 if subject i belongs to a potential cluster area labeled k; we shall denote $Z_i = (Z_i(1), \ldots, Z_i(K))$; $\mathbf{X_i}$ is a vector of covariates. Note that if the potential clusters do not overlap there is only one $Z_i(k)$ equal to 1 for each i, while if the potential clusters overlap there may be several of them. We denote by Y_i the time of the event of interest and by C_i a censoring variable. For each i the observation is $(Z_i, Y_i^*, D_i, \mathbf{X_i})$, where $Y_i^* = \min(Y_i, C_i)$ and $D_i = I_{\{Y_i < C_i\}}$. Moreover we denote by n the size of the sample and by n_k the number of subjects in potential cluster k. In each model $\mathcal{M}_k$ the constraint $\beta(k) = 0$ defines a sub-model $\mathcal{M}_0$, specified by $\lambda(t|Z_i(k), \mathbf{X_i}) = \lambda_0(t) \exp(\gamma^T \mathbf{X_i})$, which is the same for all k. The hypothesis H_0 that we wish to test is that the data come from $\mathcal{M}_0$. The test problem considered by Cook, Gold and Li (2007) was testing H_0 against $H_{A.} = \cup H_{Ak}$, with $H_{Ak}: \ \beta(k) > 0$.

Note that another approach would be to define the alternative by the model $\mathcal{M}_f$ including the K variables of the vector Z_i (only $K-1$ if the potential clusters do not overlap). However, this approach is not very powerful if there are many potential clusters and only one has a higher risk. The reason is that the model $\mathcal{M}_f$ is much larger than the union of the models $\mathcal{M}_k$. It specifies a larger class of alternatives that does not incorporate the information that there is only one (or maybe a small number of) cluster(s).

For testing H_0 against $H_{A.} = \cup H_{Ak}$, Cook, Gold and Li (2007) have taken the following approach. Assume that for each k we have a statistic T_k and H_0 is rejected for high values of T_k, the test statistic is defined as the maximum of these K statistics, $T_{max} = \max T_k$.

The conventional choice for T_k in the scan statistic approach is the likelihood ratio statistic. Cook, Gold and Li (2007) proposed a score test statistic based on Cox partial likelihood.

The advantage of the score statistic is its simplicity, due in particular to the fact that the computation of both the statistic itself and its distribution involve only the null hypothesis. In particular, the score statistic for each k involves an estimator of γ, $\hat{\gamma}$, under the null hypothesis H_0, that is under model $\mathcal{M}_0$. Thus $\hat{\gamma}$ does not depend on k and can be computed once and for all. In the generalized linear model, including the Cox model, the score statistic can be expressed as a scalar product of the vector of values of the explanatory variable (the effect of which we wish to test) and a vector of residuals. These are ordinary residuals in generalized linear models and martingale residuals in the Cox model. Thus, in our problem, we can write the score as $U_k = Z^T(k)\underline{\mathrm{M}}$, where $\underline{\mathrm{M}}$ is the vector of so-called martingale residuals (Hashemi and Commenges, 2002) computed under $\mathcal{M}_0$. This very simple form for the score statistic may be used either for computing a permutation p-value or for deriving the asymptotic distribution. The main issue on which we focus in the remainder of this paper is the choice of the reference distribution for $T_{max} = \max T_k$, where T_k is the normalized score statistic (U_k divided by an estimator of its standard deviation) for testing H_0 against H_{Ak}.

3. Significance level

3.1 *Limitation of the permutation tests*

Let us take a close look at the assumptions needed for the permutation distribution to be valid under the null hypothesis.

Let us first consider the principle of permutation tests in a simple problem. Let Y denote an outcome of interest and Z a possible linked variable. If one is interested in testing the independence of Z and Y, one can consider a statistic $\phi(\underline{\mathrm{Z}}, \underline{\mathrm{Y}})$ based on an n-sample $(\underline{\mathrm{Z}}, \underline{\mathrm{Y}}) = ((Z_i, Y_i), i = 1, \ldots, n)$. We may construct the test by considering the distribution of this statistic under the null hypothesis conditional on $\underline{\mathrm{Z}}$; we thus only need the distribution of Y (here we have introduced an asymmetry between Z and Y in the spirit of regression

models). The permutation test makes it possible to get rid of distributional assumptions also for Y: by conditioning on the order statistic of $\underline{Y}$, the only values that $\phi(\underline{Z}, \underline{Y})$ can take are $\phi(\underline{Z}, P_j\underline{Y}), j = 1, \ldots, n!$, where P_j are permutation matrices (Kalbfleisch, 1978). With the exchangeability of $\underline{Y}$, the probability of each of these values is equal to $1/n!$. Mantel (1967) proposed an interesting permutation approach for detection of space-time clustering. This approach may be applied for instance to the scan statistic T_{max} in a simple problem where the outcome is uncensored and there is no covariate. The distribution of the statistic is obtained by permuting the vector $\underline{Y}$ and giving to each possible value the probability $n!$.

It is often believed that the permutation approach can be widely applied to build tests which respect the nominal type-one risk exactly. However, the exchangeability assumption is much more restrictive than appears at first sight. It generally does not hold in more complex problems, particularly when there are covariates or censoring. Cox and Oakes (1984) remarked that the exchangeability assumption would not hold when comparing the survival distributions in two groups if the censoring distributions were not the same in the two groups. Thus the permutation distribution applied to the log-rank statistic requires an assumption which may not hold. This is the main reason why the asymptotic distribution is used rather than the permutation approach for the log-rank statistic. Commenges (2003) noted that in the presence of explanatory variables, even the residuals were not exchangeable.

In the present context we consider a statistic $T_{max} = \phi(\underline{Z}, \underline{Y}^*, \underline{D}, \underline{X})$, where the four arguments in ϕ are the vectors of the potential cluster indicators, the observed follow-up time, the event indicator and the explanatory variable, the latter being a matrix if there are several explanatory variables. In general $(\underline{Y}^*, \underline{D}, \underline{X})$ are not exchangeable. Indeed, $(P\underline{Y}^*, P\underline{D}, P\underline{X})$, where P is a permutation matrix (different from the identity), does not have in general the same distribution as $(\underline{Y}^*, \underline{D}, \underline{X})$. A sufficient condition (not far from being necessary) is that C_i and X_i are independent of Z_i for all i. Let us take a very simple example with

$n = 2$; suppose that, conditionally on $\underline{Z}$, the expectation of X_1 is $g(Z_1)$ and that of X_2 is $g(Z_2)$. The vector $\underline{X}$ has expectation $(g(Z_1), g(Z_2))$; the permuted vector $P\underline{X} = (X_2, X_1)$ has expectation $(g(Z_2), g(Z_1))$ which is not the same, unless $g(x)$ is constant. The same problem occurs for the censoring variable, as has been already remarked by Cox and Oakes (1984) for the logrank test.

In conclusion, for exchangeability to hold we have to assume that both the censoring variable and the explanatory variables do not depend on the spatial location. This is a strong assumption which may not hold in practice. In fact, tests based on asymptotic distributions may be safer, although they may have difficulties of their own, the first one being to find the asymptotic distribution.

3.2 *Asymptotic distribution of the Score test*

Hashemi and Commenges (2002) tackled the problem of correcting the p-value when multiple cut-off values have been tried for dichotomizing an explanatory variable in a Cox model. They exploited the scalar product structure of the score statistic for computing the asymptotic covariance between T_k and $T_{k'}$ and used this to derive the asymptotic distribution of T_{max}. In fact, their approach can be directly applied to many multiplicity problems and in particular to the scan statistic. It is known that the score statistics T_k asymptotically (when the n_k tend towards infinity) have standard normal distributions under the null hypothesis. Under some regularity conditions, $\underline{T}$ has a multivariate normal distribution. Hashemi and Commenges (2002) gave formulas for estimating the covariance matrix of $\underline{T}$. The p_{value} associated with the observation $T_{max} = t_{max}$ is $p_{value} = P(T_{max} > t_{max}) = 1 - P(T_1 < t_{max}, \ldots, T_K < t_{max})$. The value of $P(T_1 < t_{max}, \ldots, T_K < t_{max})$ can be computed by numerical integration of the density of the multivariate normal distribution. For large K the integration may be performed by simulation.

What are the assumptions needed for the Hashemi-Commenges asymptotic distribution

to be valid? The censoring variable and the potential cluster indicator do not need to be independent, nor do the explanatory variables and the potential cluster indicator. However, if there are explanatory variables the proportional hazard model must be correct. Also this is an asymptotic result; for applying it each n_k must be relatively large.

4. Discussion

The permutation distribution is often believed to yield valid tests without assumptions but in fact it requires the exchangeability assumption which, in the problem at hand, takes the form of the independence of censoring variable and potential cluster, and also of explanatory variables and potential cluster. On the other hand the asymptotic distribution does not need these assumptions but needs the proportional hazard model to be correct. However, in the spirit of robust inference (Royall, 1986; Wei, Lin and Weissfeld, 1989), it is possible to estimate empirically the variances and covariances of score statistics. This is often used to construct the "sandwich estimator" for the variance of regression parameters but here it is simpler since we are dealing with score statistics. For instance, the covariance of $n^{-1/2}U_k$ and $n^{-1/2}U_k'$ can be estimated by $n^{-1}\sum_{i=1}^{n} U_{ik}U_{ik'}$. This version would be clearly more robust than the permutation approach. A limitation of using the asymptotic distribution is that there must be a sufficient number of subjects in each potential cluster. If the number of subjects is not large one may prefer the permutation distribution but in any case the assumptions needed for the permutation distribution to be valid should be critically examined.

Another question is the computation of the p-value. If the number of potential clusters is not large (say $n \leqslant 20$), the p-value using the asymptotic distribution can be obtained by numerical integration, for instance using the Genz algorithm (Genz, 1992). The precision obtained is better than that obtained by simulation in the permutation approach. If the number of clusters is large, the asymptotic p-value can still be computed by simulation (generating a large number of multivariate normal variables with the estimated covariance

matrix, computing the maximum and checking whether this maximum is above the observed t_{max}). Precision in that case should be approximately the same than that of the permutation p-value.

Cook, Gold and Li (2007) proposed another statistic $\hat{W}_{loc}$, which is different from the score test and thus should be less powerful. Moreover, the assumption of independence of censoring variable and potential cluster is also required in the argument leading to the computation of the p-value using this statistic. The discussion has been focused on survival data but the same arguments apply to other type of data. In particular, asymptotic distribution has been derived for generalized linear models (Liquet and Commenges, 2005).

References

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

Cox D.R. and Oakes D., (1984). *Analysis of survival Data*, Chapman & Hall, London.

Cook, A. J., Gold, D. R. and Li, Y. (2007). Spatial Cluster Detection for censored outcome data. *Biometrics* **63**, 540–549.

Genz, A.(1992). Numerical Computation of Multivariate Normal Probabilities *Journal of Computational and Graphical Statistics* **47**, 141-149.

Hashemi, R. and Commenges, D. (2002). Correction of p-values after multiple tests in a Cox proportional hazard model. *Lifetime Data Analysis* **8**, 335–348.

Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistics for survival data. *Biometrics* **63**, 109-118.

Kalbfleisch, J.D. (1978), Likelihood methods and nonparametric tests. *Journal of the American Statistical Assocociation* **73**, 167-170.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statitics* **26**, 1481-1496.

Liquet, B. and Commenges, D. (2005). Computation of the p-value of the maximum of

score tests in the generalized linear model; application to multiple coding *Statistics & Probability Letters* **71**, 33-38.

Mantel, N. (1967). The detection of disease clustering and a general regression approach. *Cancer Research* **27 Part 1**, 209-222.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). *Journal of the American Statistical Association* **84,** 1065-1073.

Received June 2007. *Accepted October* 2007.