Communications of the Korean Statistical Society
2012, Vol. 19, No. 6, 899–904

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.6.899

Multiple Testing in Genomic Sequences Using Hamming Distance

Moonsu Kang[1,a]

[a]Department of Information Statistics, Gangneung-Wonju National University

Abstract

High-dimensional categorical data models with small sample sizes have not been used extensively in genomic sequences that involve count (or discrete) or purely qualitative responses. A basic task is to identify differentially expressed genes (or positions) among a number of genes. It requires an appropriate test statistics and a corresponding multiple testing procedure so that a multivariate analysis of variance should not be feasible. A family wise error rate(FWER) is not appropriate to test thousands of genes simultaneously in a multiple testing procedure. False discovery rate(FDR) is better than FWER in multiple testing problems. The data from the 2002–2003 SARS epidemic shows that a conventional FDR procedure and a proposed test statistic based on a pseudo-marginal approach with Hamming distance performs better.

Keywords: Pseudo-marginal approach, false discovery rate, Hamming distance, genomic sequence.

1. Introduction

High dimension and sample size data may appear in various areas of science (the dimension tends to ∞ while the sample size is small). This data models are abound in genomic studies, in particular, where the sample size n may be small and there are different epidemiologic strata G (> 2), so that the classical multivariate analysis of variance may not be pertinent (Sen, 2006, 2008; Kang and Sen, 2008, 2007; Ghosh, 2003). This study identifies the most significant genes (or positions) among a number of genes: Which positions are differentially expressed across the groups? The feature of this study is that the number of genes in a sequence(K) is significantly larger than the number of sequences(n). Control of the family wise error rate(FWER) is too conservative when there are many hypotheses such as microarray experiments (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). A false discovery rate(FDR) procedure is better than the FWER procedure to handle multiple testing problem in a large scale association study (Benjamini and Hochberg, 1995; Storey, 2002, 2003; Storey *et al.*, 2004). In multiple testing problems, the response variables are continuous, but may be count or discrete (or purely qualitative responses) of high-dimensional low sample size categorical data setups that complicate the multiplicity problems (Dudoit *et al.*, 2003). To motivate this problem, in SARS epidemic data (Dye and Gay, 2003), we have 900 genes (or positions) for each sequence and 14 samples downloaded from four locations, Guangdong, Beijing, Hong Kong and Taiwan. The response variables are a, c, g, and t, having even not ordered categories. Suppose we have a general model comprise G (> 2) groups of sequences. Each sequence has K positions, and in each position, there is a categorical response with C categories. n_{gkc} denotes the number of responses in category c at position k in the g^{th} group, $c = 1, \ldots, C; k = 1, \ldots, K$ and $g = 1, \ldots, G$. There is a set $\mathbf{C}$ of C^K

[1] Professor, Department of Information Statistics, Gangneung-Wonju National University, Gangneung 210-702, Korea. E-mail: moonsukang0223@gmail.com

joint labels $\mathbf{c} = (c_1, \ldots, c_K)$ in which each c_k takes on value $1, \ldots, C$. The number of observations in the g^{th} group with the label combination $\mathbf{c}$ is denoted by $n_g(\mathbf{c}), \mathbf{c} \in \mathbf{C}$. We also have $\sum_{\mathbf{c}\in\mathbf{C}} n_G(\mathbf{c}) = n_g$ $\sum_{\mathbf{c}\in\mathbf{C}} \pi_g(\mathbf{c}) = 1, \forall g = 1, \ldots, G$. The full multi-dimensional multinomial law is

$$\prod_{g=1}^{G} \left\{ \frac{n_g!}{\prod_{\mathbf{c}\in\mathbf{C}}(n_g(\mathbf{c}))!} \prod_{\mathbf{c}\in\mathbf{C}} [\pi_g(\mathbf{c})]^{n_g(\mathbf{c})} \right\}.$$

The total number of unknown parameters is $q^0 = G(CK - 1)$, but q^0 is too large compared to sample size n, where $n = \sum_{g=1}^{G} n_g$. This law may not be reasonable because of this problem. That is why the standard multivariate approach may be of limited utility (Kang and Sen, 2008; Sen, 2008, 2006). Thus, instead of modelling the multivariate approach to include many positions (or predictors), we approach each position separately and test thousands of positions using a FDR procedure with appropriate test statistics for each position. In this sense, the (pseudo) marginal diversity measures for each position may be combined into a composite measure that provide a less stringent way of categorical ANOVA. The categories are not ordered in the SNP model and a stochastic ordering may not be feasible. However, the Hamming distance may have ordering (Sen, 2005; Pinhero *et al.*, 2005). Even in that case, individual statistics (even coordinate-wise ones) do not have a known null hypothesis distribution. That is why we have to use their jackknife variance estimation and permutation distribution to construct some permutation tests (Sidak *et al.*, 1999; Krishnaiah and Sen, 1985). A pseudo-marginal approach based on Hamming distance provides some promising test statistics. Current FDR procedure along with the associated test statistic using this approach may be a useful tool for genomic studies. These perspectives are appraised in a nonstandard statistical analysis, using the 2002–2003 SARS epidemic data.

2. A Pseudo Marginal Model

The full multi-sample, multi-dimensional multinomial law may not be reasonable. For geographically separated sequences, the assumption of independence among G groups may be tenable but may not be independent within group sequences. For each sequence, the K positions may not be independent responses or identically distributed. Under the assumption of an inter-position stochastic dependence, we need to consider another measure of variation. The Gini-Simpson biodiversity index has found useful applications in genetics and in bioinformatics (Sen, 2005; Pinhero *et al.*, 2005). Mostly, categorical data models (without an ordering of the categories) appear, which preempts the use of measures of quantitative diversity analysis. Without much stringent structural regularity assumptions, the Hamming distance exploits the idea of Gini-Simpson diversity index in a variety of multidimensional setups (Sen, 2005; Pinhero *et al.*, 2005). Consider a set of n vectors $\mathbf{X}_{gi} = (X_{gi1}, \ldots, X_{gik})', i = 1, \ldots, n$, where X_{gik} stands for the particular label $(1, \ldots, C)$, for the i^{th} observation in the k^{th} position and g^{th} group, for $k = 1, \ldots, K$ and $g = 1, \ldots, G$. For two vectors $\mathbf{X}_{gi}, \mathbf{X}_{gj}, i \neq j$, the Hamming distance is defined as

$$d_{gij} = K^{-1} \sum_{k=1}^{K} I(X_{gik} \neq X_{gjk}). \tag{2.1}$$

For the entire sample of n vectors, we have the Hamming distance

$$D_{gn} = \binom{n}{2} \sum_{1\leq i<j\leq n} d_{gij}.$$

We exploit the following Gini-Simpson index: $\mathbf{I}(\pi) = 1 - \pi^t\pi = 1 - \sum_{c=1}^{C} \pi_c^2$, where $\pi = (\pi_1, \ldots, \pi_C)^t$ for a single multinomial population with C cells. Define $\mathbf{I}(\pi_{gk})$ for each $k = 1, \ldots, K$ and every $g = 1, \ldots, G$. For each $g\,(= 1, \ldots, G)$ and $k\,(= 1, \ldots, K)$, $\mathbf{I}(\pi_{\mathbf{gk}}) = 1-(\pi_{gk})^t\pi_{gk} = 1-\sum_{c=1}^{C}(\pi_{gkc})^2$. Also, define $\mathbf{I}(\pi_k)$ in the pooled sample, for each $k = 1, \ldots, K$. Define $H(\prod_g) = (1/K)\sum_{k=1}^{K} \mathbf{I}(\pi_{gk})$, $g = 1, \ldots, G$ as the Hamming distance based measure. In genomic studies, the following multiple hypotheses are represented in terms of the Gini-Simpson index.

$$H_0 : \mathbf{I}(\pi_{1k}) = \mathbf{I}(\pi_{2k}) = \cdots = \mathbf{I}(\pi_{Gk}), \quad k\,(= 1, \ldots, K) \quad vs.$$
$$H_1 : \text{There are at least one of } k's \text{ that } \mathbf{I}(\pi_{gk}) \neq \mathbf{I}(\pi_{g'k}), \quad 1 \leq g < g' \leq G.$$

3. Proposed Test Statistics and *P*-values

Let us consider the following asymptotic distribution of the test statistic. For each $k\,(= 1, \ldots, K)$ and each $g\,(= 1, \ldots, G)$, the estimate of $\mathbf{I}(\pi_{gk})$ using the Hamming distance is given by

$$\begin{aligned} U_{gk} &= \binom{n_g}{2} \sum_{1 \leq i < j \leq n_g} I(X_{gik} \neq X_{gjk}) \\ &= \sum_{c=1}^{C} \frac{n_{gkc}(n_g - n_{gkc})}{n_g(n_g - 1)}. \end{aligned}$$

This is a U-statistic based on a kernel of degree 2, an unbiased estimator of Gini-Simpson index. In the pooled sample,

$$\begin{aligned} U_k &= \binom{n}{2} \sum_{1 < i < j <= n} I(X_{ik} \neq X_{jk}) \\ &= \sum_{c=1}^{C} \frac{n_{kc}(n - n_{kc})}{n(n - 1)}, \end{aligned}$$

where $n_{kc} = \sum_{g=1}^{G} n_{gkc}$. Note that U_{gk} is the U-statistics with kernel $I(X_{gik} \neq X_{gjk})$ with degree of 2.

Suppose that the U-statistics is defined by

$$U = \binom{n}{2} \sum_{1 <= i < j <= n} g\,(X_{i1}, \ldots, X_{im})\,.$$

Then the asymptotic distribution of U is given by $\sqrt[2]{n}(U - \theta(F)) \sim N(0, m^2\zeta_1)$, where U is an unbiased estimator of $\theta(F)$ and $\zeta_1 = E(g(X_1, \ldots, X_m)g(X_m, \ldots, X_{2m-1}) - E(g(X_1, \ldots, X_m)g(X_{m+1}, \ldots, X_{2m})))$ by the projection method ((5.3.18) and (5.3.22) in Sen and Singer (1993)).

In these sense, the asymptotic distribution of this U-statistics ($m = 2$) is given by

$$\sqrt[2]{n_g}\left(U_{gk} - \mathbf{I}(\pi_{gk})\right) \sim N\left(0, 4\zeta_{1gk}\right),$$

where $\zeta_{1gk} = E\{I(X_{gik} \neq X_{gjk})I(X_{gi'k} \neq X_{gj'k}) - E(I(X_{gik} \neq X_{gjk})I(X_{gi'k} \neq X_{gj'k}))\}$. This ζ_{1gk} can be replaced by $\zeta_{1.k}$ from the pooled sample. $\zeta_{1.k}$ is estimated by the jackknife variance estimator of U_k, which is $\zeta_{1.k}\,(= 1/(n-1)\sum_{i=1}^{n}(U_{nk,i} - U_k)^2)$, because the jackknife variance estimator is more stable than other variance estimator (Sen and Singer, 1993; Sidak *et al.*, 1999; Krishnaiah and Sen, 1985;

Sen, 1977). For each $k(=1,\ldots,K)$, the test statistic L_k is defined as $\sum_{g=1}^{G} n_g[U_{gk} - U_k]^2/(4\hat{\zeta}_{1.k})$. By virtue of Cochran's theorem, it has χ^2 distribution with degree $G-1$ (Krishnaiah and Sen, 1985; Huber and Ronchetti, 1981). However, a conclusion based on this asymptotic distribution, whenever sample size is small, may give misleading results. Moreover, for n not adequately large, the p-values have discrete distribution without assuming uniform distribution for the associated p-values under the null hypothesis. Hence, it might be better to simulate the permutation distribution of the marginal test statistic L_k. At least for small to moderate values of the sample sizes, $n_1,\ldots,n_G$, the permutation distribution can be generated by considering all possible $n!$(equally likely) permutations of the combined sample observations among the G groups of (sizes $n_1,\ldots,n_G$). Hence, conditionally distribution-free tests may be constructed for the test statistic L_k. The corresponding p-value is defined as below. $\Pr(L_k > l_k|H_0)$, where L_k is a test statistic from the permuted distribution. Under the null hypothesis, the permutation distribution of L_k may be symmetric about 0, with mean $E_0(L_k) = 0$. Under the alternative hypothesis, the distribution is tilted to the right (Odeh, 1972; Silvapulle and Sen, 2004). and is why we use a right-sided test (Kang and Sen, 2008; Sen, 2008, 2006). However, though the distribution freeness holds under the null hypothesis, such distributions are more complex to evaluate.

4. Numerical Analysis

Following its origin in Southern China, the SARS epidemic resulted in 8422 infected people and 916 deaths (Dye and Gay, 2003). The SARS causative agent was identified as a corona virus 96 (SARSCoV) using the GenBank database. The SARS epidemic has an identified single-stranded and positive-sense RNA virus with large genome size and moderate mutation rate. For these sequences, an enormously high-dimensional purely qualitative categorical model is constructed. SARS complete sequences ($n = 14$) are downloaded and isolated from Guangdong, Beijing, Hong Kong and Taiwan: $n_1 (= 5)$ from Guangdong, $n_2 (= 4)$ from Beijing, $n_3 (= 3)$ from Hong Kong and $n_4 (= 2)$ from Taiwan. To simplify the measurement of variation, the sequences with no nucleotide changes are removed. The responses consist of not even ordered categories, a, c, g, and t and an ordering may not be feasible. The Hamming distance provides a stochastic ordering; however, individual statistics using Hamming distance do not have a known null hypothesis distribution in general. For these reasons, we use jackknife variance estimation $\zeta_{1.k}$ and permutation distribution to construct some permutation tests. There are K=900 genes (or positions) for each sequence and for each position, the test statistic described and corresponding p-value are computed. The test statistic L_k is defined as $\sum_{g=1}^{4} n_g[U_{gk} - U_k]^2/(4\hat{\zeta}_{1.k})$, $k = 1,\ldots,900$. P-value is also defined as $\Pr(L_K > l_k|H_0)$, where L_k is a test statistic from a permuted distribution. The permutation distribution can be generated by considering 14!(equally likely) permutations of the combined sample observations among four groups of (sizes 5, 4, 3 and 2). Based on 900 p-values from the positions, each p-value is computed from a permuted distribution. Based on the estimation of the proportion of true null hypothesis (Storey, 2003), we set it as 0.4. First, Storey's FDR procedure (Storey (normal)) (Storey, 2002), Benjamini and Hochberg FDR procedure (BH (normal)) (Benjamini and Hochberg, 1995) are computed with p-values using test statistics based on normal theory. Also, we calculate the same procedures (Storey (proposed) and BH(proposed)) with proposed test statistics with $\alpha = 0.1, 0.05$ in Table 1. Subsequently, false discovery rates with the proposed method work well in that those approaches control at any preassigned level α.

5. Concluding Remarks

We consider high dimension low sample size genomic sequences without ordering of response categories. When constructing an appropriate test statistics in this model, the classical ANOVA approach

Table 1: FDR procedures using the proposed test statistics compared with the test statistics based on normal theory

α	π_0	Storey (normal)	BH (normal)	Storey (proposed)	BH (proposed)
0.1	0.40	0.095	0.098	0.082	0.078
0.05		0.057	0.050	0.038	0.045
0.01		0.017	0.016	0.009	0.008

may not be tenable due to too large number of parameters and too small sample size. In this sense, a pseudo marginal model based on the Hamming distance was presented. The Hamming distance utilizes the idea of Gini-Simpson diversity index in a variety of multidimensional setups. For a small sample size, the permutation distribution was generated by considering all possible $n!$ (equally likely) permutations of the combined sample observations among the G groups of (sizes $n_1, \ldots, n_G$). FDR procedures along with the associated test statistics for each gene based on the proposed test statistics worked well in the set of p-values generated from the exact permutation theory and controls the FDR at any level α.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, *The Annals of Statistics*, **29**, 1165–1188.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, **18**, 71–103.

Dye, C. and Gay, N. (2003). Modeling the SARS epidemic, *Perspectives Epidemiology*, 300.

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from microarray experiments, *Bioinformatics*, **59**, 992–1000.

Huber, P. J. and Ronchetti, E. M. (1981). *Robust Statistics*, Wiley Series in Probability and Statistics, New York.

Kang, M. and Sen, P. K. (2007). *Multiple Testing in Genome-wide Studies*, University of North Carolina at Chapel Hill.

Kang, M. and Sen, P. K. (2008). Kendall tau type rank statistics in genomic data, *Applications of Mathematics*, **3**, 207–221.

Krishnaiah, P. R. and Sen, P. K. (1985). *Handbook of Statistics 4: Nonparametric Methods*, North-Holland, Netherlands.

Odeh, R. E. (1972). On the power of Jonckheere's k-sample test against ordered alternatives, *Biometrika*, **59**, 467–471.

Pinhero, H. P., Pinhero, A. D. S. and Sen, P. K. (2005). Comparison of genomic sequences using the hamming distance, *Journal of Statistical Planning and Inference*, **130**, 325–339.

Sen, P. K. (1977). Some invariance principles relating to jackknifing and their role in sequential analysis, *The Annals of Statistics*, **5**, 316–329.

Sen, P. K. (2005). Gini diversity index, hamming distance, and curse of dimensionality, *METRON - International Journal of Statistics*, **LXIII**, 329–349.

Sen, P. K. (2006). Robust statistical inference for high dimensional data models with application to genomics, *Austrian Journal of Statistics*, **35**, 197-214.

Sen, P. K. (2008). Kendall's tau in high-dimensional genomic parsimony, *Institute of mathematical Statistics, Collection Series*, **3**, 251–266.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*, Chapman and Hall/CRC, New York.

Sidak, Z., Sen, P. K. and Hajek, J. (1999). *Theory of Rank Tests*, Second Edition (Probability and Mathematical Statistics), San Diego, Academic Press, CA.

Silvapulle, M. J. and Sen, P. K. (2004). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*, Wiley-Interscience, New York.

Storey, J. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B*, **64**, 479–498.

Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value, *Annals of Statistics*, **3**, 2013–2035.

Storey, J., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach, *Journal of the Royal Statistical Society, Series B*, **66**, 187–205.

Received August 29, 2012; Revised November 14, 2012; Accepted November 15, 2012