The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1863-19

Advanced School and Conference on Statistics and Applied Probability in Life Sciences

24 September - 12 October, 2007

Nonparametric Modelling for Classification and Multiple Hypothesis Testing

Peter Hall

University of Melbourne
Dept. of Mathematics and Statistics
3010 Melbourne, Australia

NONPARAMETRIC MODELLING FOR CLASSIFICATION AND MULTIPLE HYPOTHESIS TESTING

Peter Hall, University of Melbourne & University of California Davis

work with Yao-ban Chan, Sandy Clarke and Aurore Delaigle

Statistical problems involving high-dimensional data vectors

Multiple hypothesis testing problems not infrequently take the form:

We have a vector $X = (X_1, \ldots, X_p)$ of test statistics, of which the ith permits us to test the hypothesis H_{0i} that the mean, μ_i, of a population equals zero, against the alternative H_{1i} that $\mu_i > 0$, or H_{2i} that $\mu_i \neq 0$.

Classification problems can take the form:

Given vectors $X_i = (X_{i1}, \ldots, X_{ip})$, for $1 \leq i \leq m$, and $Y_j = (Y_{j1}, \ldots, Y_{jp})$, for $1 \leq j \leq n$, from populations Π_X and Π_Y, respectively, construct a classifier that assigns a new data vector Z to either Π_X or Π_Y.

Some problems of signal detection are similar.

Questions we would like to answer

How does dependence among $X_1, \ldots, X_p$ affect the level accuracy and power of simultaneous hypothesis tests?

In particular, if the dependence is short range, how does it impact on the performance of classifiers or methods for signal detection? What if the dependence is long range?

We shall discuss theoretical models that enable us to answer these questions.

Our focus will be on multiple hypothesis testing, but similar models can be used to respond to the same questions in the contexts of classification and signal detection.

Models for high-dimensional data vectors

Generally, different data vectors can fairly be assumed to be independent, but of course their components cannot.

Therefore we wish to model the strength of dependence among vector components.

We shall use simple time-series models, noting that more complex models often lead to similar conclusions.

Linear models

Arguably, the simplest models are linear in structure although nonparametric in character; we do not make parametric assumptions about the distribution of the disturbances.

For example, considering the components X_i of a single data vector $X = (X_1, \ldots, X_p)$, we may write

$$X_i = \sum_k \theta_k \, \epsilon_{i+k} \, ,$$

where the θ_k's are constants and the random variables ϵ_i, for $-\infty < i < \infty$, are independent and identically distributed.

Linear models (cont.)

A linear model for X_i is often appropriate when testing hypotheses about a mean, in cases where the variance is known. More generally, "toy" linear models provide insight into more sophisticated settings where the actual test statistic is relatively complex.

In many of these more complex cases the main conclusions are identical to those under the simple linear model.

For example, this is typically the case when the test statistic is of Student's t type, in particular when it incorporates an empirical scale correction.

Nonlinear, inhomogeneous models

Moreover, it is often straightforward to remove both homogeneity and linearity, for example:

$$X_i = r_i \sum_k \theta_k \, \epsilon_{i+k}^{s_i} \, ,$$

where the ϵ_i's are now assumed to be nonnegative, and the constants r_i and s_i are also nonnegative.

All the results that we shall discuss have analogues in this setting.

Multiple hypothesis testing: mechanism

Suppose we conduct p tests, based on the respective values of $X_1, \ldots, X_p$. Here, X_i represents a test statistic computed from the ith of a sequence of samples.

In particular, for $1 \leq i \leq p$ we reject the ith null hypothesis, $H_{0i} : \mu_i = 0$, if $X_i > t$, say. If $X_i \leq t$ then we do not reject H_{0i}.

Multiple hypothesis testing: literature

The literature on multiple testing procedures is vast, and only a part of it is confined to statistics journals. Review-type contributions include those of Hochberg and Tamhane (1987), Pigeot (2000), Dudoit, Shaffer and Boldrick (2003), Bernhard, Klein and Hommel (2004) and Lehmann and Romano (2005, Chap. 9).

Benjamini and Hochberg (1995) introduced an approach, which has become very popular, to controlling false discovery rate. See also Simes (1986), Hommell (1988), Hochberg (1988), Sarkar (1998) and Sen (1999).

Multiple hypothesis testing: literature (cont.)

Benjamini and Yekuteli (2001) specified conditions under which simultaneous, dependent hypothesis tests, conducted as though they were independent, give conservative results. Efron (2007) suggested correlation corrections for large-scale simultaneous hypothesis testing.

Blair, Troendle and Beck (1996) proposed methods for controlling family-wise error rates in multiple procedures, and Holland and Cheung (2002) discussed robustness of family-wise error rate.

Multiple hypothesis testing: family-wise error rate

Let N, a random variable, denote the number of rejected hypotheses:

$$N = \sum_{i=1}^{p} I(X_i > t)\,. \tag{1}$$

If each of $H_{01}, \ldots, H_{0p}$ is correct, and if we view the sequence of p tests as a test of the "simultaneous hypothesis" H_0 that each of the component hypotheses H_{0i} is true, then the significance level of the simultaneous test equals the probability that $N \geq 1$.

This is the family-wise error rate (FWER) of the procedure. The generalised family-wise error rate (GFWER) is the probability $P_0(N \geq k)$ of at least k false discoveries.

Multiple hypothesis testing: formula for error rate (cont.)

For example, if $0 < \alpha < 1$ and we define $\beta = -\log(1-\alpha)$; if we choose t, in (1), to satisfy

$$P_0(X > t) = \frac{\beta}{p} + o\big(p^{-1}\big)\,; \tag{2}$$

and if

the test statistics X_i are independent and identically distributed as X;

then the generalised family-wise error rate (GFWER) converges in Poisson fashion:

$$P_0(N \geq k) \to \sum_{j=k}^{\infty} \frac{\beta^j}{j!}\, e^{-\beta}\,.$$

Multiple hypothesis testing: formula for error rate (cont.)

In particular, if $\beta = -\log(1-\alpha)$ then the family-wise error rate converges to α as p increases: $P_0(N \geq 1) \to \alpha$.

Here and in (2), P_0 denotes probability computed under H_0.

For the sake of simplicity we shall retain the assumption that, under the null hypothesis, the test statistics X_i are (asymptotically) identically distributed, and discuss the effect that lack of independence has on the FWER.

Subsequently we shall address the effects on false discovery rate (FDR); the impact there is very similar to the effects on FWER.

Multiple hypothesis testing: formula for error rate (cont.)

The case of non-identical null distributions of the test statistics can also be treated. It differs from the identical-distribution context principally in terms of complexity.

Main conclusions

1. Light-tailed test statistic distributions.

If the test statistic distribution is light tailed, decreasing like $\exp(-x^\gamma)$ where $\gamma \geq 1$, then the difficulties caused by dependence decrease as p, the number of simultaneous tests, increases.

In particular, the number of clusters of false discoveries declines, and the distribution of critical-point exceedences closely resembles its counterpart for independent data.

In this setting, methods that would normally be recommended only for independent data can give very good control of family-wise error rate and false-discovery rate.

Main conclusions (cont.)

2. Heavy-tailed test statistic distributions.

Only for relatively heavy-tailed data, where the tail of the marginal distribution decreases more slowly than e^{-x}, is this property violated.

In such cases, clusters of exceedences can occur, and methods based on the assumption that test statistics are independent are not adequate for controlling error rates.

The difficulty can be overcome by employing conservative methods (e.g. based on Bonferroni bounds) or by modelling the distributions of clusters.

Intuitive arguments

1. Light-tailed test statistic distributions.

In the case of light-tailed marginal distributions, exceedences above a high level (appropriate when the number of simultaneous tests, p, is very large) occur only because neighbouring disturbances ϵ_i are fortuitously aligned.

Indeed, since the tail is so light then it is highly unlikely that a single disturbance is so great as to carry the process,

$$X_i = \sum_k \theta_k \, \epsilon_{i+k}$$

close to, or over, the level for several different, neighbouring values of i. Instead different, moderately large disturbances reinforce one another, by chance, at a particular i.

Intuitive arguments (cont.)

1. Light-tailed test statistic distributions. (cont.)

However, at adjacent indices i the circumstances that led to alignment change, and the propensity for level exceedence quickly diminishes or even disappears.

Consequently, clusters of exceedences seldom arise. That is, the pattern of exceedences appears as though it was produced by a sequence of independent tests, and as a result, both generalised family-wise error rate, and false-discovery rate, can be controlled by appealling to standard arguments for independent tests.

Intuitive arguments (cont.)

2. Heavy-tailed test statistic distributions.

When test statistics have relatively heavy-tailed distributions, the probability that a single disturbance is so great that it carries the value of a test statistic over a high level, not just for one but for several consecutive indices i, is relatively high.

In such cases clusters of exceedences can occur, and methods based on independent data are not adequate for controlling error rates.

Intuitive arguments (cont.)

3. Small numbers of simultaneous tests.

The arguments and properties above, especially those in the light-tailed setting, are applicable only to exceedences of very high levels.

Very high levels are relevant only when the number of simultaneous tests is particularly large. Therefore the properties discussed above tend not to be noticed in conventional multiple testing problems, where the number of tests is relatively small.

Theoretical results: Assumptions

Assume that the test statistics X_i are generated as,

$$X_i = \sum_k \theta_k \, \epsilon_{i+k} \, ,$$

where the ϵ_i's are independent and identically distributed. Suppose too that,

> $\theta_k \geq 0$ for each k, $\theta_k = 0$ for all but a finite number of values of k, and $\theta_k \neq 0$ for some k.

The constraint that $\theta_k \geq 0$ is imposed here only to remove the need to impose conditions on the lower tail of the common distribution of the errors ϵ_i. It does not materially influence the results.

Given $\beta > 0$, let $t = t(p, \beta)$ be such that

$$P_0(X > t) = \frac{\beta}{p} + o\big(p^{-1}\big) \, .$$

Theoretical results: Assumptions (cont.)

We consider first the case where the upper tail of the marginal distribution of ϵ is light, and in particular decreases like $\exp(-x^\gamma)$ for some $\gamma \geq 1$.

Specifically, we assume that the density f of the distribution of ϵ satisfies, as $x \to \infty$, either

$$f(x) = \exp\left\{o\left(x^\gamma\right)\right\} \exp\left(-C\,x^\gamma\right),$$

where $\gamma > 1$; or, in the case $\gamma = 1$,

$$f(x) \sim C_1\,x^{C_2} \exp(-C\,x),$$

for constants $C, C_1 > 0$ and $C_2 \geq 0$.

Theoretical results (1): Light-tailed test statistic distributions

Let $1 \leq I_1 \leq I_2 \leq \ldots$ denote the indices i for which $X_i > t$.

Theorem. *Under the above conditions, and for each $C > 0$, the point process $I_1\, p^{-1}, I_2\, p^{-1}, \ldots$, restricted to the interval $[0, C]$, converges weakly, as $p \to \infty$, to a homogeneous Poisson process on $[0, C]$, with intensity β.*

Implications of theorem for GFWER

Recall that the generalised family-wise error rate (GFWER) is given by $P_0(N \geq k)$, where

$$N = \sum_{i=1}^{p} I(X_i > t)$$

is the number of false discoveries; and that, under the assumption that the hypothesis tests are independent,

$$P_0(N \geq k) \to \sum_{j=k}^{\infty} \frac{\beta^j}{j!} e^{-\beta}, \tag{3}$$

where P_0 denotes probability computed under the null hypothesis.

It follows from the theorem that, provided $\gamma \geq 1$ and the upper tail of the marginal distribution decays like $\exp(-x^\gamma)$ where $\gamma \geq 1$, result (3) does not require independence; it also holds under dependence.

False discovery rate (FDR)

The false-discovery rate approach (e.g. Benjamini and Hochberg, 1995) involves a step-down procedure but can be framed in a similar way to GFWER.

In particular, for $i \geq 1$ let $t_1 > t_2 > \ldots$ denote a sequence depending on p and with the property that

$$P_0(X > t_i) = \frac{i\beta}{p} + o\big(p^{-1}\big) \, .$$

(Thus, the t defined earlier is here denoted by t_1.)

Write N_i for the number of values X_i that lie in the interval $(t_i, t_{i-1}]$, where we take $t_0 = \infty$.

False discovery rate (cont. 1)

The event that the step-down method of Benjamini and Hochberg (1995) does not reject any of the hypotheses H_{0i}, for $1 \leq i \leq k$, is equivalent to the event that, for each i in the latter range, $X_i = X_{(p-j+1)} \leq t_j$, where $X_{(1)} \leq \ldots \leq X_{(p)}$ represent the order statistics of the sequence $X_1, \ldots, X_p$.

For example, if k denotes the largest j for which $X_{(p-j)} \leq t_{j-1}$, then H_{0i} is rejected for each i such that $X_i = X_{(p-j+1)}$, where $1 \leq j \leq k$.

False discovery rate (cont. 2)

Therefore, to describe properties of the false-discovery rate approach, we need to understand not just the distribution of N but more generally the distribution of

$$N^{(k)} = \sum_{i=1}^{p} I(X_i > t_k)\,.$$

Note that $N^{(k)} = N_1 + \ldots + N_k$, where

$$N_i = \sum_{j=1}^{p} I(t_i \leq X_j < t_{i-1})\,.$$

False discovery rate (cont. 3)

Under the assumption that the hypothesis tests are conducted independently the variables $N_1, \ldots, N_k$ are asymptotically independent and Poisson-distributed with mean β.

False discovery rate (cont. 4)

Therefore, the probability that the null hypotheses corresponding to the k largest values of X_i are all rejected under the FDR approach, when they are in fact all correct, is given by,

$$P_0\left(N^{(i)} \geq i \quad \text{for} \quad 1 \leq i \leq k\right)$$
$$\rightarrow P\left(Q_1 + \ldots + Q_i \geq i \quad \text{for} \quad 1 \leq i \leq k\right),$$

where $Q_1, \ldots, Q_k$ are independent and identically Poisson-distributed with mean β.

(The probability on the right-hand side here is dominated by β, for each $k \geq 1$, although this is useful only if $\beta < 1$.)

False discovery rate (cont. 5)

From the previous transparency: Under the assumption that the hypothesis tests are conducted independently,

$$P_0\left(N^{(i)} \geq i \quad \text{for} \quad 1 \leq i \leq k\right)$$

$$\rightarrow P\left(Q_1 + \ldots + Q_i \geq i \quad \text{for} \quad 1 \leq i \leq k\right).$$

The theorem, valid for light-tailed marginal distributions, implies that this result continues to hold under weak dependence.

Heavy-tailed marginals

We continue to assume that the test statistics X_i are generated as,

$$X_i = \sum_k \theta_k \, \epsilon_{i+k} \, ,$$

where the θ_k's are nonnegative and the ϵ_i's are independent and identically distributed.

If the tails of the distribution of ϵ decay like $\exp(-x^\gamma)$ for some $\gamma < 1$, or if they decay at a polynomial rate, then clustering of false discoveries can occur.

Heavy-tailed marginals (cont. 1)

1. Heavy, but nevertheless exponential, tails

If the density f of the distribution of ϵ satisfies, as $x \to \infty$,

$$f(x) \sim C_1 \, x^{C_2} \, \exp\left(-C \, x^{\gamma}\right) \, ,$$

for constants $\gamma < 1$, $C, C_1 > 0$ and $C_2 \geq 0$, then asymptotic clustering of false discoveries occurs only if there is a tie for the largest value of θ_k; not otherwise.

Heavy-tailed marginals (cont. 2)

To provide intuition we treat the case where $\theta_1 = \ldots = \theta_r$ and each other θ_k vanishes.

In this setting, having $\epsilon_1 + \ldots + \epsilon_r > x$ implies that, with high probability, one of the values of $\epsilon_1, \ldots, \epsilon_r$ is very close to x, or greater than x, and the other values are all significantly smaller than x. (Here and below we assume that x is large.)

That is, just one of the ϵ_i's is responsible for the level exceedence, and its influence can persist, through weights in the moving average, to ensure that $\epsilon_{j+1} + \ldots + \epsilon_{j+r} > x$ for values of j other than simply $j = 0$.

Heavy-tailed marginals (cont. 3)

2. Marginal distributions with polynomially heavy tails

Here a suitable model is,

$$P(\epsilon > x) \sim C\, x^{-\rho}$$

as $x \to \infty$, where $C, \rho > 0$ are constants.

For each i such that $\theta_i > 0$, let $\theta_i = \theta_{i1} \geq \ldots \geq \theta_{ir_i}$ be a ranking of the r_i, say, nonzero values of θ_{i+j} that are not strictly greater than θ_i. Define $\theta = 0$ for $j \geq r_i + 1$, $p_{iq} = \theta_{iq}^{\rho} - \theta_{i,q+1}^{\rho}$.

Heavy-tailed marginals (cont. 4)

2. Marginal distributions with polynomially heavy tails (cont.)

For each integer $q \geq 1$, put $p_q = (\sum_i p_{iq})/(\sum_i \theta_i^\rho)$. Let M_0 denote a random variable for which $P(M_0 \leq q) = p_q$.

Then the distribution of M_0 is the limiting distribution of cluster size for false discoveries.