Biometrika
2015, Vol. 102, No. 3, pp. 695-704.
Supplementary Materials

Efficient estimation of the number of false positives in high-throughput screening

Holger Rootzén and Dmitrii Zholud*

Abstract

This material contains introduction to the SmartTail software; proofs of Theorems 1, 2, and 3 in the paper; derivation of equation (10); additional results and discussion for dependent p-values; sandwich estimators for dependent p-values; additional plots for the yeast genome and salt stress screening data; and two additional examples: association mapping in *Arabidopsis Thaliana* and a fMRI brain scan experiment. MATLAB scripts to simulate from the tail mixture model and to compute the maximum likelihood estimates of the parameters of the mixture model (8) for the cases (i)-(iii) described in the paper are also available as supplementary online material.

1 SmartTail

SmartTail is a MATLAB implementation of the methods developed in the paper for handling false positives in high-throughput screening experiments. It may also be useful in more general tail estimation problems. SmartTail, and the data sets analyzed in this study, are available at www.smarttail.se.

The SmartTail graphical user interface consists of two panels. The left panel is used to fit the tail model (1) and to compare the distribution of the p-values from the true null distribution with the theoretical uniform distribution. The right panel can be used either for fitting the tail model (1) to analyze two set of p-values separately, or for fitting the extreme tail mixture model (2) with $\gamma_0 = \gamma_1$, and to compare it with the true null distribution in the left panel. Both panels are equipped with model checking tools including POT plots for the parameters of the model and for the estimator $\hat{F}(x)$; Kolmogorov-Smirnov test of model fit; test of the hypothesis $\gamma = \gamma_0$ for some fixed γ_0, and thus in particular for testing $\gamma = 1$ as suggested by Zholud (2014); test of equality of the γ-s in the two panels; and FDR and $pFDR$ plots. The left panel FDR and $pFDR$ plots are with respect to uniform distribution and the right panel FDR and $pFDR$ plots are with respect to the distribution in the left panel. Thus, if a sample from the true null distribution is loaded into the left panel and a sample from an experiment is loaded into the right panel, then the right panel FDR and $pFDR$ plots are with respect to the true null distribution. Estimates of the parameters c_i, γ_i, and of $F(x)$ and $FDR/pFDR$, with confidence intervals, are provided.

2 Proofs of the basic results

Throughout this supplementary material we use notation and equation numbers for the paper.

Proof of Theorem 1. We show that for $i = 0, 1$ there exist constants γ_i and functions $\alpha_i(y, u)$ such that

$$F_i(uy)/F_i(u) = y^{1/\gamma_i} \left\{1 + \alpha_i(y, u)\right\}, \qquad (1)$$

*Department of Mathematical Statistics
Chalmers University of Technology and University of Göteborg, Sweden.
E-mails: hrootzen@chalmers.se and dmitrii@chalmers.se

where $\alpha_i(y,u) \to 0$ as $u \to 0$, uniformly for $y \in [\varepsilon, 1]$ for any $0 < \varepsilon < 1$. Equation (3) in the paper then follows by replacing uy by x in (1) and taking $c_i(u) = F_i(u)/u^{1/\gamma_i}$.

(i) If $\xi_t, \xi_i > 0$, then by Beirlant et al. (2004), Equation (2.7), we have that $\bar{G}_t^{\leftarrow}(x) = \ell_t(x)x^{-\xi_t}$ and $\bar{G}_i(x) = \ell_i(x)x^{-1/\xi_i}$, where $\ell_t(x)$ is slowly varying as $x \to 0$ and $\ell_i(x)$ is slowly varying as $x \to \infty$. Hence

$$F_i(x) = \bar{G}_i\{\bar{G}_t^{\leftarrow}(x)\} = \ell_i\{\ell_t(x)x^{-\xi_t}\}\{\ell_t(x)\}^{-1/\xi_i}x^{\xi_t/\xi_i} = \ell(x)x^{1/\gamma}, \tag{2}$$

with $\gamma = \xi_i/\xi_t > 0$ and $\ell(x) = \ell_i\{\ell_t(x)x^{-\xi_t}\}\{\ell_t(x)\}^{-1/\xi_i}$. Let $u > 0$ be fixed. Since ℓ_t is slowly varying, $\ell_t(ux)(ux)^{-\xi_t} = \{1 + o(1)\}u^{-\xi_t}\ell_t(x)x^{-\xi_t}$, and it follows from Beirlant et al. (2004), Theorem (2.3) (i), that $\ell_i\{\ell_t(ux)(ux)^{-\xi_t}\}/\ell_i\{\ell_t(x)x^{-\xi_t}\} \to 1$ as $x \to 0$. Since also $\{\ell_t(ux)\}^{-1/\xi_i}/\{\ell_t(x)\}^{-1/\xi_i} \to 1$ it follows that $\ell(ux)/\ell(x) \to 1$ as $x \to 0$, so that $\ell(x)$ is slowly varying as $x \to 0$. Equation (1) hence follows from the Karamata uniform convergence theorem.

(ii) In this case, $\bar{G}_t^{\leftarrow}(x) = x^* - \ell_t(x)x^{\xi_t}$ with $\ell_t(x)$ slowly varying as $x \to 0$ and $\bar{G}_i(x) = \ell_i(x^* - x)(x^* - x)^{1/\xi_i}$, with $\ell_i(x^* - x)$ slowly varying as $(x^* - x) \to 0$, see Beirlant et al. (2004), equation (2.11). Thus $\bar{G}_i\{\bar{G}_t^{\leftarrow}(x)\} = \ell_i\{\ell_t(x)x^{\xi_t}\}\ell_t(x)^{1/\xi_i}x^{\xi_t/\xi_i}$, and the proof may be completed as in part (i). □

Proof of Theorem 2. The first assertion is a standard property of Poisson distributions. Further, if (i) holds then we have that the numbers of H_0 and H_1 p-values smaller that α_m are binomially distributed, and it follows from the Poisson limit theorem for binomial distributions that $||Q_i - Po\{\pi_i F_i(\alpha_m)\}|| \to 0$ for $i = 0, 1$. If (ii) is satisfied then $m_0/m \overset{P}{\to} \pi_0$ as $m \to \infty$. Thus, for any $\epsilon > 0$, the probability that the number of false positives is smaller than the number of positives in an extended sample of size $(\pi_0 + \epsilon)m$ tends to one as $m \to \infty$, and therefore the number of false positives is stochastically smaller than a binomial variable with $m_0 + \epsilon m$ trials and success probability $F_0(\alpha_m)$. Similarly one gets an upper bound for the number of true positives, and corresponding lower bounds. Using monotonicity we obtain $||Q_i - Po\{\pi_i F_i(\alpha_m)\}|| \to 0$ for $i = 0, 1$. □

Derivation of equation (10) in the paper. As in the paper, but omitting the 0-indexes to simplify the notation, let $p_1, p_2, \ldots p_m$ be a sample of mutually independent observations from a distribution F which satisfies (2), and hence, is such that $F(x) = F(u)(x/u)^{1/\gamma}$ for $x \leq u$. Let $N(x)$ denote the number of p_i-s which are less than or equal to x, so that $N(x)$ has a binomial distribution with parameters m and $F(x)$. The empirical distribution function estimator of $F(x)$ is given by

$$F_E(x) = N(x)/m,$$

and our estimator of $F(x)$ has the form

$$F(x) = \frac{N(u)}{m}\left(\frac{x}{u}\right)^{1/\hat{\gamma}} = F_E(u)\left(\frac{x}{u}\right)^{1/\hat{\gamma}},$$

for $\hat{\gamma}$ given by (7) of the paper, that is $\hat{\gamma} = N(u)^{-1}\sum_{p_i \leq u} -\log(p_i/u)$. We set $\hat{\gamma} = 0$ if $N(u) = 0$.

Since $E\{\hat{\gamma}|N(u)\} = \gamma$ for $N(u) > 0$, the law of total expectation gives

$$E(\hat{\gamma}) = E\left[E\left\{\hat{\gamma}|N(u)\right\}\right] = \gamma \mathrm{pr}\left\{N(u) > 0\right\},$$

and similarly

$$\mathrm{cov}\{F_E(u), \hat{\gamma}\} = \gamma F(u)\mathrm{pr}\left\{N(u) = 0\right\}.$$

Further, we have that

$$\begin{aligned} \mathrm{var}(\hat{\gamma}) &= E[\mathrm{var}\{\hat{\gamma}|N(u)\}] + \mathrm{var}[E\{\hat{\gamma}|N(u)\}] \\ &= \gamma^2\left[E\left\{\frac{1_{N(u)>0}}{N(u)}\right\} + \mathrm{var}\left\{1_{N(u)>0}\right\}\right]. \end{aligned}$$

Straightforward computation shows that if, say, $m > 1000$ and $mF(u) > 10$ then

$$E\left\{\frac{1_{N(u)>0}}{N(u)}\right\} \approx \frac{1}{mF(u)}\left\{1+\frac{1}{mF(u)}\right\}$$

with an error of at most one unit in the second decimal, see Johnson et al. (1992), Section 4.11, equation (3.92). Since $\mathrm{pr}\{N(u)=0\} \approx \exp\{-mF(u)\}$ we can with even smaller error replace $\mathrm{pr}\{N(u)=0\}$ and $\mathrm{var}\left\{1_{N(u)>0}\right\}$ by zero and $\mathrm{pr}\{N(u)>0\}$ by one in the expressions above. If we strengthen the assumptions further to, say, $mF(u) > 35$, then we get that $E\{1_{N(u)>0}/N(u)\} \approx 1/\{mF(u)\}$, with an error less that 3%.

Standard central limit theory ensures that $\hat{\gamma}$ and $F_E(x)$ are asymptotically normally distributed, and according to the computations above we can hence with small error assume that the following normal approximations hold,

$$\hat{\gamma} \approx N[\gamma, \gamma^2/\{mF(u)\}] \quad \text{and} \quad F_E(x) \approx N\left[F(x), \frac{F(x)\{1-F(x)\}}{m}\right].$$

The delta-method now gives that $\hat{F}(x) \approx N\left\{F(x), \sigma^2\right\}$, where

$$\sigma^2 = \frac{F(x)^2}{mF(u)}\left[1-F(u)+\frac{1}{\gamma^2}\left\{\log\left(\frac{x}{u}\right)\right\}^2\right].$$

Using $F(x) = F(u)(x/u)^{1/\gamma}$ we can then write

$$\begin{aligned}\frac{\mathrm{var}\{F_E(x)\}}{\mathrm{var}\left\{\hat{F}(x)\right\}} &= \left(\frac{u}{x}\right)^{1/\gamma}\left[\frac{1-F(u)}{1-F(x)}+\frac{1}{\gamma^2}\frac{\{\log\left(\frac{x}{u}\right)\}^2}{1-F(x)}\right]^{-1} \\ &\approx \left(\frac{u}{x}\right)^{1/\gamma}\left[1+\frac{1}{\gamma^2}\left\{\log\left(\frac{u}{x}\right)\right\}^2\right]^{-1},\end{aligned}$$

where the approximation is accurate for small values of $F(u)$. This concludes the derivation of (10).

3 Dependent p-values

This section provides the proof of Theorem 3 of the paper; presents sandwich estimators for dependent p-values; discusses asymptotic normality for more complex dependence structures than those in Theorem 3; and considers the conditional distribution of the number of false positives under dependence and the clustering of p-values which may occur when there is high local dependence.

3.1 Proof of Theorem 3

For the convenience of the reader, we first repeat the notation and statement of Theorem 3, with more details and explanations added, and then give the proof.

Omitting the subscript 0, suppose the p-values are observations of a stationary sequence $\{P_i\}$ with marginal distribution function $F(x)$ which satisfies Condition (i) or (ii) of Theorem 1, so that by (2)

$$F(x) = \ell(x)x^{1/\gamma}, \tag{3}$$

with $\gamma > 0$ a constant which does not depend on x and $\ell(x)$ slowly varying as $x \to 0$. For a sample of size m from $\{P_i\}$ and some small enough $u = u_m$ the estimator (6) of $F(x)$ for $0 \le x \le u_m$ is

$$\hat{F}(x) = F_E(u_m)\left(\frac{x}{u_m}\right)^{1/\hat{\gamma}(u_m)},$$

where, writing 1_i for the indicator function of the event $P_i \leq u_m$, $F_E(u_m) = m^{-1}\sum_{i=1}^m 1_i$ and $\hat{\gamma}(u_m) = \sum_{i=1}^m \{-\log(P_i/u_m)\,1_i\}/\sum_{i=1}^m 1_i$.

We use strong mixing to prove asymptotic normality of the estimators $F_E(u_m)$, $\hat{\gamma}(u_m)$ and $\hat{F}(x)$ as $m \to \infty$ and $x < u_m \to 0$. A major ingredient in the proof is blocking: the sequence is split up into $k_m = [m/r_m]$ big blocks $B_{m,i} = ((i-1)r_m, ir_m]$, $1 \leq i \leq k_m$ of length r_m, and asymptotically negligible separating small blocks of length ℓ_m are removed from the right ends of the big blocks, so that the remainder of the big blocks are asymptotically independent as $m \to \infty$.

Define a function f as follows,

$$\hat{F}(x) = f\left\{\frac{F_E(u_m)}{F(u_m)}, \tilde{\gamma}(u_m)\right\} = F(u_m)\frac{F_E(u_m)}{F(u_m)}\left(\frac{x}{u_m}\right)^{\frac{F_E(u_m)}{F(u_m)}\frac{1}{\tilde{\gamma}(u_m)}},$$

where

$$\tilde{\gamma}(u_m) = \frac{F_E(u_m)}{F(u_m)}\hat{\gamma} = \frac{1}{m}\sum_{i=1}^m \frac{-\log(P_i/u_m)\,1_i}{F(u_m)}.$$

Here $E\{F_E(u_m)/F(u_m)\} = 1$. Also, for $\gamma_m = E\{\tilde{\gamma}(u_m)\}$ and using partial integration, (3), and Karamata's theorem, we obtain that

$$\gamma_m = \frac{1}{F(u_m)}\int_0^{u_m} -\log(x/u_m)dF(x) = \frac{1}{\ell(u_m)u_m^{1/\gamma}}\int_0^{u_m} \ell(x)x^{1/\gamma-1}dx \to \gamma. \tag{4}$$

A first order Taylor expansion of f around $f(1,\gamma_m)$ leads to the approximation

$$\hat{F}(x) - F(u_m)\left(\frac{x}{u_m}\right)^{1/\gamma_m} \approx D\sum_{i=1}^m \{Z_{m,i} - E(Z_{m,i})\}, \tag{5}$$

where $D = D(x/u_m, u_m) = -m^{-1}\log(x/u_m)(x/u_m)^{1/\gamma}F(u_m)\gamma^{-2}$,

$$Z_{m,i} = \sum_{j\in B_i}\{-\log(P_j/u_m) + C\}F(u_m)^{-1}1_j,$$

and $C = C(x/u_m, u_m) = -\gamma_m\{1 + \gamma_m/\log(x/u_m)\}$.

To state the main theorem we need the some additional notation. Define

$$Z_{m,i}^{(1)} = \sum_{j\in B_i} -\log(P_j/u_m)F(u_m)^{-1}1_j, \quad Z_{m,i}^{(2)} = \sum_{j\in B_i} F(u_m)^{-1}1_j,$$

and set $\sigma_m^2 = k_m\mathrm{var}(Z_{m,1})$, $\sigma_{m,1}^2 = k_m\mathrm{var}(Z_{m,1}^{(1)})$, and $\sigma_{m,2}^2 = k_m\mathrm{var}(Z_{m,1}^{(2)})$. Introduce sample block sums

$$\hat{Z}_i = \hat{Z}_{m,i} = \hat{D}\sum_{j\in B_i}\{-\log(P_j/u_m) + \hat{C}\}F_E(u_m)^{-1}1_j,$$

where $\hat{D} = -m^{-1}\log(x/u_m)(x/u_m)^{1/\hat{\gamma}}F_E(u_m)\hat{\gamma}^{-2}$ and $\hat{C} = -\hat{\gamma}\{1 + \hat{\gamma}/\log(x/u_m)\}$. Further, set

$$s_m^2 = \sum_{i=1}^{k_m}(\hat{Z}_i - \bar{Z})^2, \tag{6}$$

for $\bar{Z} = k_m^{-1}\sum_{i=1}^{k_m}\hat{Z}_i$. Finally, let $\{\mathcal{B}_{i,j}\}$ be the σ-algebra generated by $P_i, \ldots, P_j$, and define the strong mixing coefficients

$$\alpha_{m,\ell} = \sup\{|\mathrm{pr}(AB) - \mathrm{pr}(A)\mathrm{pr}(B)| : A \in \mathcal{B}_{1,k}, B \in \mathcal{B}_{k+\ell,m}, 1 \leq k \leq m-\ell\}.$$

Now introduce the following conditions:

C1: There exist integers $\ell_m < r_m \to \infty$ with $r_m = o(m)$ such that, for $k_m = [m/r_m]$,

$$k_m(\alpha_{m,\ell_m} + \ell_m/m) \to 0 \quad \text{and} \quad k_m^{-1} m F(u_m) \to 0.$$

C2: There exist integers $w_m > 1$ such that

$$r_m\{F(u_m)\sigma_m\}^{-1} w_m\{mF(u_m)\sigma_m^{-1} e^{-w_m} + 1\} \to 0 \quad \text{and} \quad \sigma_m\{mF(u_m)\}^{-1} \to 0.$$

Theorem 3.1. *(i) Suppose C1, C2, and* (3) *hold, and that there exist constants* $0 < k < K$ *such that* $k \leq \sigma_{m,i}/\sigma_m \leq K$ *for* $i = 1, 2$. *Then for any fixed* $y \in (0, 1)$, *as* $m \to \infty$,

$$\frac{1}{D(y, u_m)\sigma_m}\left\{\hat{F}(yu_m) - F(u_m)y^{1/\gamma_m}\right\} \to_d N(0, 1), \tag{7}$$

and

$$\frac{m}{\sigma_{m,1}}(\hat{\gamma} - \gamma_m) \to_d N(0, 1), \quad \frac{m}{\sigma_{m,2}}\{F_E(u_m) - F(u_m)\} \to_d N(0, 1). \tag{8}$$

In particular $\hat{\gamma} \to_{pr} \gamma$ *and* $F_E(u_m)/F(u_m) \to_{pr} 1$.

(ii) If, in addition, $k_m var(Z_{m,1}^2) \to 0$, *then*

$$\frac{1}{s_m}\left\{\hat{F}(yu_m) - F(u_m)y^{1/\gamma_m}\right\} \to_d N(0, 1). \tag{9}$$

Proof. (i) We first show that, as $m \to \infty$,

$$\sigma_m^{-1}\left[\sum_{j=1}^{m}\{-\log(P_j/u_m) + C\}F(u_m)^{-1}1_j - m(\gamma_m + C)\right] \to_d N(0, 1), \tag{10}$$

$$\sigma_{m,1}^{-1}\left[\sum_{j=1}^{m}\{-\log(P_j/u_m)F(u_m)^{-1}1_j\} - m\gamma_m\right] \to_d N(0, 1), \tag{11}$$

and that

$$\sigma_{m,2}^{-1}\left\{\sum_{j=1}^{m}F(u_m)^{-1}1_j - m\right\} \to_d N(0, 1). \tag{12}$$

It then follows from (11), (12), and $k \leq \sigma_{m,i}/\sigma_m \leq K$ that the remainder terms in the first order Taylor expansion (5) of (7) tend to zero in probability, and (7) then is a consequence of (10).

To prove (10) we show that the conditions of Theorem 6.4 of Rootzén et al. (1998) are satisfied by $X_j = -\log P_j$, for n replaced by m and with u_n in the cited paper replaced by $v_m = -\log u_m$, and for $\alpha_1 = 1$, $\alpha_2 = C$, $\phi_{m,1}(x) = F(u_m)^{-1}xI_j(x)$ and $\phi_{m,2}(x) = F(u_m)^{-1}I_j(x)$, for $I_j(x) = 1$ if $x \geq v_m$ and 0 otherwise. Thus, in particular, with the notation above, $1_j = I_j\{-\log(P_j)\}$. For brevity, throughout this proof, equations numbers, basic conditions, theorems, and lemmas refer to Rootzén et al. (1998).

It follows from $F(x) = \mathrm{pr}(P_1 \leq x) = \ell(x)x^{1/\gamma}$ that

$$\mathrm{pr}(X_1 > x + t)/\mathrm{pr}(X_1 > t) = F(e^{-x}e^{-t})/F(e^{-t}) = e^{-x/\gamma}\{1 + o(1)\} \quad \text{as} \quad t \to \infty, \tag{13}$$

and hence (6.1) holds. By *C1* the basic assumptions hold, and, using Lemma 4.3, it follows that (4.7) and hence also (2.4) is satisfied. The first part of *C2* by straightforward calculation implies that (6.2) and (6.3) hold. Since the other conditions of Theorem 6.4 are satisfied by assumption, this proves

(10). Using the assumption $k \le \sigma_{m,i}/\sigma_m \le K$ for $i = 1, 2$, the results (11) and (12) are established in the same way.

The second part of *C2* includes that $m/\sigma_m \to \infty$, and then also $m/\sigma_{m,2} \to \infty$, and thus $\hat{F}(u_m)/F(u_m) \to_{pr} 1$ is implied by (12). By (4) we have that $\gamma_m \to \gamma$, and hence $\hat{\gamma} = \tilde{\gamma}F(u_m)/F_E(u_m) \to_{pr} \gamma$ is a consequence of $m/\sigma_{m,2} \to \infty$, the first part of (8), and $\hat{F}(u_m)/F(u_m) \to_{pr} 1$.

(ii) Let $\tilde{s}_m$ be defined in the same way as s_m but with Z instead of $\hat{Z}$. It follows from the last assertion of (i) that $D/\hat{D} \to_{pr} 1$ and hence, in the proof, we without loss of generality assume that $D = \hat{D} = 1$. According to Theorem 5.2

$$\tilde{s}_m^{-1}\left[\sum_{j=1}^{m}\{-\log(P_j/u_m) + C\}F(u_m)^{-1}1_j - m(\gamma_m + C)\right] \to_d N(0,1),$$

and $\tilde{s}_m/\sigma_m \to 1$. Hence the result follows from the first order Taylor expansion used in part (i) above if $s_m/\tilde{s}_m \to_{pr} 1$, or equivalently if $(s_m^2 - \tilde{s}_m^2)/\sigma_m^2 \to_{pr} 0$. Now, expanding squares and using Cauchy's inequality, that $(a+b)^2 \le 4(a^2+b^2)$, and that the expectation of a sample variance is bounded by the variance, this can be seen to hold if

$$\left\{(\hat{C} - C)F(u_m)F_E(u_m)^{-1}\right\}^2 \sigma_m^{-2}\sum_{i=1}^{k_m}\left\{\sum_{j\in J_i}F(u_m)^{-1}1_j - k_m^{-1}\sum_{j=1}^{r_m k_m}F(u_m)^{-1}1_j\right\}^2 \to_{pr} 0 \qquad (14)$$

and

$$\left\{F(u_m)F_E(u_m)^{-1} - 1\right\}^2 \sigma_m^{-2}\sum_{i=1}^{k_m}\left[\sum_{j\in J_i}\{-\log(P_j/u_m) + C\}F(u_m)^{-1}1_j \right. \qquad (15)$$

$$\left. -k_m^{-1}\sum_{j=1}^{r_m k_m}\{-\log(P_j/u_m) + C\}F(u_m)^{-1}1_j\right]^2 \to_{pr} 0.$$

The proofs of (14) and (15) are almost identical, so we only consider (14). It follows from the last part of (i) that $(\hat{C} - C)F(u_m)F_E(u_m)^{-1} \to_{pr} 0$. Further, again since the expectation of a sample variance is bounded by the variance, the expectation of the last part of (14) is bounded by $\sigma_m^{-2}k_m\mathrm{var}(Z_{m,2}) = \sigma_m^{-2}\sigma_{m,2}^{-2} \le K$, and hence it is tight, so that (14) holds. □

The conditions *C1* - *C2*, even though they contain rather many components, are simple, and very generally applicable, cf. the closely related papers Resnick and Starica (1997), Drees (2000), Drees (2003), and Rootzén (2009).

3.2 Sandwich estimators and more complex dependence structures

Under the conditions of Theorem 3, $\gamma_m \to \gamma$ and $F(yu_m) \sim F(u_m)y^{1/\gamma}$ and it then follows that $F(u_m)y^{1/\gamma_m} \sim F(yu_m)$. If further the bias $F(yu_m) - F(u_m)y^{1/\gamma}$ is of smaller order than σ_m, then the the distribution of the estimation error $\hat{F}(yu) - F(yu)$ may be approximated by a $N(0, s_m^2)$-distribution. If the P_i were independent, this would be the delta method, and in the present dependent case this kind of estimators are commonly referred to as sandwich estimators.

Dependence structures in high-throughput experiments may differ from stationary time series dependence. For example, in the yeast genome screening experiment, 100 strains of yeast were grown on each of two plates, and many such plates were grown sequentially in time. Clearly there is a risk of dependence between p-values coming from the same pair of plates, or between p-values coming from plates grown in, say, the same day, or from genes that are functionally related. We will, without proofs, indicate how the results above may be adapted to such complex situations where dependence is not just sequential in time. In particular, this provides sandwich estimators also for such situations.

The key is to divide the p-values up into asymptotically independent blocks, where the blocking may be determined by the structure of the experiment in a much more sophisticated way than in the stationary time series setting considered in Theorem 3.1. As before with $k_m = [m/r_m]$, assume that the p-values are split up into blocks B_i, where $i = 1, \ldots, k_m$, of r_m p-values, and that ℓ_m p-values are removed from each block to form new blocks B_i' consisting of the remaining parts of the blocks. Let $\{\mathcal{B}_i^-\}$ and $\{\mathcal{B}_i^+\}$ be the σ-algebras generated by $P_j, j \in B_k', k = 1, \ldots, i$ and by $P_j, j \in B_k', k = i, \ldots, k_m$, respectively, and define strong mixing coefficients by

$$\alpha_{m,\ell} = \sup\{|\mathrm{pr}(AB) - \mathrm{pr}(A)\mathrm{pr}(B)| : \; A \in \mathcal{B}_i^-, B \in \mathcal{B}_{i+1}^+, 1 \le i \le k_m - 1\}.$$

With $Z_{m,i}, Z_{m,i}^{(1)}$ and $Z_{m,i}^{(2)}$ defined as above, set

$$\sigma_m = \sum_{i=1}^{k_m} \mathrm{var}(Z_{m,i}), \quad \sigma_{m,i} = \sum_{i=1}^{k_m} \mathrm{var}(Z_{m,i}^{(i)}), \quad i = 1, 2,$$

and assume that $\sup\{\mathrm{var}(Z_{m,i}^2), i = 1, \ldots, k_m\} / \inf\{\mathrm{var}(Z_{m,i}^2), i = 1, \ldots, k_m\} \to 1$, and that $Z_{m,i}^{(1)}$ and $Z_{m,i}^{(2)}$ satisfy the corresponding conditions. Instead of $k_m \mathrm{var}(Z_m^2) \to 0$ assume that $k_m \sup\{\mathrm{var}(Z_{m,i}^2), i = 1, \ldots, k_m\} \to 0$, and further assume that all p-values have the same one-dimensional marginal distribution, with $F(x) = \ell(x) x^{1/\gamma}$.

With this new notation and new assumptions, Theorem 3.1 still holds. Thus the sandwich estimator can be used also for much more general dependence structures than time series dependence.

3.3 The conditional distribution of the number of false positives under dependence

In a stationary time series setting, the quite general Leadbetter (1974) conditions $\mathrm{D}(\alpha_m)$ and $\mathrm{D}'(\alpha_m)$ are commonly used to obtain Poisson convergence, as needed for Theorem 2. $\mathrm{D}(\alpha_m)$ is expected to hold for any reasonable experiment. However $\mathrm{D}'(\alpha_m)$ holds only in situations where small p-values do not cluster. If clustering does occur, then Theorem 3.1 of Rootzén et al. (1998) gives more general conditions which ensure that there still exist a limiting distribution, but this distribution is now a compound Poisson process. Estimation of the compounding probabilities is often uncertain, and in such situations one might not be able to estimate the full conditional distribution of the number of false positives, but only the expected number of false positives.

Clustering of p-values which could indicate dependence within blocks can be investigated informally by inspection of the samples, and there is also a large literature on formal estimation of the amount of clustering, as measured by the so-called Extremal Index, see e.g. Beirlant et al. (2004), Section 10.3.2. However, the issue is somewhat delicate: clustering caused by local dependence will violate the asymptotic Poisson distributions, but clusters of very small p-values may also be caused by non-null experiments occurring at neighboring locations, and this would then not contradict an asymptotic Poisson distribution. The latter situation, for example, is expected to occur in the brain scan experiment discussed below.

The Leadbetter conditions may also be extended to more general dependence structures. With the blocking B_i, B_i' as in the discussion of the effect on complex dependence structures on $\hat{F}(x)$, let $\bar{M}_i = \min\{P_j, j \in B_k', 1 \le k \le i\}$ and $M_i = \min\{P_j, j \in B_k'\}$ and introduce the following conditions.

D1: The blocking satisfies $\ell_m = o(m)$ and, for $m \to \infty$,

$$\sup\left\{|\mathrm{pr}(\bar{M}_i \le \alpha_m, M_{i+1} \le \alpha_m) - \mathrm{pr}(\bar{M}_i \le \alpha_m)\mathrm{pr}(M_{i+1} \le \alpha_m)| \; i = 1, \ldots, k_m - 1\right\} \to 0.$$

D2: The blocking satisfies that, as $m \to \infty$,

$$\sum_{i=1}^{k_m} \sum_{j \neq k,\, j,k \in B_i} \mathrm{pr}\left(P_j \leq \alpha_m, P_k \leq \alpha_m\right) \to 0.$$

If the P_j form a stationary time series, then *D1* is implied by $\mathrm{D}(\alpha_m)$ and *D2* by $\mathrm{D}'(\alpha_m)$.

Following the line of argument in Leadbetter (1974) it can be shown that if *D1* and *D2* hold and $mF(\alpha_m) \to \tau > 0$, then the number of P_j's which are smaller than α_m is asymptotically Poisson distributed.

4 Association mapping in Arabidopsis.

Zhao et al. (2007) study 95 *Arabidopsis Thaliana* samples, with measurements of flowering-related phenotypes together with genotypes in the form of over 900 short sequenced fragments, distributed throughout the genome. The goal was association mapping, that is identification of regions of the genome where individuals who are phenotypically similar are also unusually closely genetically related. A problem is that spurious correlations may arise if the population is structured so that members of a subgroup, say, samples obtained from a specific geographical area, tend to be closely related. One main thrust of the paper was to evaluate 9 different statistical methods to remove such spurious correlations. But of course the ultimate goal is to identify interesting genes.

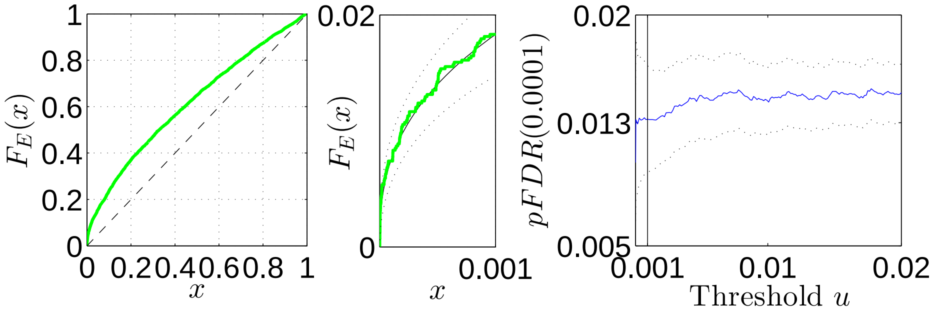

Figure 1. Goodness of fit and parameter stability plots for the KW analysis of the JIC4W data set *Left:* Empirical distribution function. Dashed line is the uniform distribution. *Middle:* Empirical distribution function for $p \leq 0.001$ (99 values). Solid line is (9) of the paper estimated using $u = 0.001$. Dotted lines are 95% pointwise confidence intervals. Note that x-axis scale is stretched 10 times. *Right:* pFDR at $\alpha = 0.0001$ as function of the threshold u, for $\pi_0 = 1$. Dotted lines are 95% pointwise confidence intervals.

Here we only consider the SNP (Single Nucleotide Polymorphism) data, essentially obtained from re-sequencing, and one phenotype, the one called JIC4W, which we choose since it was of special interest in the paper. Further, we only display results for two of the statistical methods, the KW method which just consisted in making Kruskal-Wallis tests without correction for population structure, and a method called Q+K which may have been the most successful of the 9 methods studied. The number of tests was 3745.

For the KW method, the maximum likelihood estimators using the mixture model (8) and method (iii) of the paper, with $u = 0.01$, were $\hat{p} = 0.4, \hat{\gamma}_0 = 1.2$, and $\hat{\gamma}_1 = 2.6$. However, the 95% confidence interval for p was the entire interval $(0, 1)$. Method (ii), that is fixing γ_0 to 1, for $u = 0.01$ gave $\hat{p} = 0.27$ and $\hat{\gamma}_1 = 2.40$, with 95% confidence intervals $(0.00, 0.54)$, and $(1.85, 2.96)$. Confidence intervals obtained by fitting (8) were very wide.

Instead, Figures 1 and 2 show that the model (9) fits both the Kruskal-Wallis and the Q+K p-values well (Kolmogorov-Smirnov p-values 0.43 and 0.38). The estimate of pFDR for $\alpha = 0.0001$, with 95% confidence intervals, for the Kruskal-Wallis test was 0.013 ± 0.0035, and for the Q+K the estimate was 0.15 ± 0.14, that is more than 11 times bigger. Both these numbers assume that the true null distribution is the uniform distribution. Zhao et al. (2007) argue that most of the Kruskal-Wallis p-values are spurious. We also did the same analysis for the other statistical tests studied by Zhao et al. For most, but not all of them, (11) gave a good fit. Of course the quality of the fit also depended on the choice of u

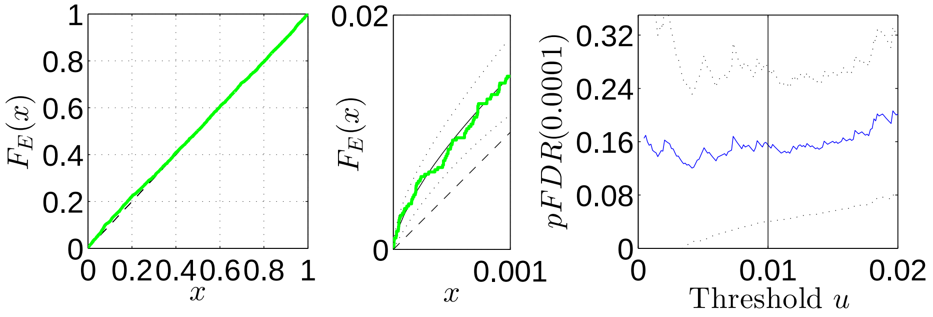

Figure 2. Goodness of fit and parameter stability plots for the Q+K analysis of the JIC4W data set *Left:* Empirical distribution function. Dashed line is the uniform distribution. *Middle:* Empirical distribution function for $p \leq 0.01$ (76 values). Solid line is (11) estimated using $u = 0.01$. Dotted lines are 95% pointwise confidence intervals. *Right:* p-FDR at $\alpha = 0.0001$ as function of the threshold u, for $\pi_0 = 1$. Dotted lines are 95% pointwise confidence intervals.

Again, to illustrate the gain in efficiency from using the estimates from Section 3, SmartTail estimated that $\text{var}\{F_E(0.0001)\}/\text{var}\{\hat{F}(0.0001)\}$ was 1.9 for the KW method with $u = 0.001$ and 1.2 for the Q+K method and $u = 0.01$ (the γ-values were 2.6 and 1.4, respectively).

5 fMRI brain scan experiment

Taylor and Worsley (2006) study the Functional Image Analysis Contest data set which contains results from an fMRI experiment aimed at understanding the language network in the human brain. Here we use the Block Experiment, Dehaene-Lambertz et al. (2006), in which 16 subjects were instructed to lie still in a scanner with eyes closed and to listen attentively to blocks of 6 sentences, either different ones or the same sentence, and either read by the same speaker or by different speakers. Each subject was asked to participate in two runs, with 16 blocks presented in each run. In Taylor and Worsley (2006), for each run and each voxel in the brain scans, the data was used to study the significance of two contrasts, different minus same sentence and different minus same speaker, and their interaction. Roughly 35,000 voxels per subject were used. For each voxel in each subject and each run sophisticated preprocessing was used to construct 3 t-test quantities. One subject dropped out, and one only completed one run, so the result was $(15 \times 2 + 1) \times 3 = 93$ sets of roughly 35,000 t-test quantities.

To study the fit of equation (9) we transformed these t-values to p-values using a t_{40}-distribution, see Taylor and Worsley (2006). For each of the 93 resulting data sets we performed a Kolmogorov-Smirnov test of the fit of the model (9) for the p-values which were smaller than the threshold $u = 0.01$. The smallest number of p-values less than 0.01 in any of these data sets was 117, and the largest number was 973.

Figure 3 shows that the distribution of the 93 goodness-of-fit p-values were somewhat skewed towards smaller values, as compared with the uniform distribution. However, the deviation from uniformity is small, and the overall impression is that (9) fits well.

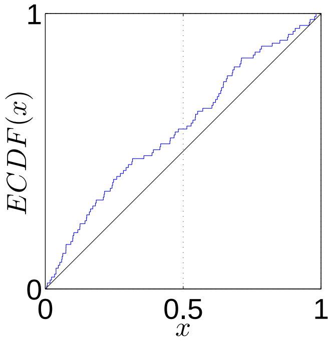

Figure 3. Empirical cumulative distribution function of the Kolmogorov-Smirnov p-values from 93 fMRI brain scan data sets.

In fact, even for the two data sets where (9) was clearly rejected, the Kolmogorov-Smirnov plots showed that the deviations from the model were quite moderate.

References

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes, Theory and Applications*. Wiley, Chichester, 2004.

G. Dehaene-Lambertz, S. Dehaene, J.-L. Anton, A. Campagne, A. Jobert, D. LeBihan, M. Sigman, C. Pallier, and J.-B. Poline. Functional segregation of cortical language areas by sentence repetition. *Human Brain Mapping*, 27(5):360–371, 2006.

H. Drees. Weighted approximations of tail processes for β-mixing random variables. *Ann. Appl. Probab.*, 10(4):1274–1301, 2000.

H. Drees. Extreme quantile extimation for dependent dat, with application to finance. *Bernoulli*, 9 (1):617–657, 2003.

N.L. Johnson, S. Kotz, and A.W. Kemp. *Univariate discrete distributions*. Wiley, New York, 1992.

M.R. Leadbetter. On extreme values in stationary sequences. *Probability Theory and Related Fields*, 28(4):289–303, 1974.

S. Resnick and C. Starica. Asymptotic behavior of hills estimator for autoregressive data. *Comm. Statist. Stochastic Models*, 13(4):703–721, 1997.

H. Rootzén. Weak convergence of the tail empirical process for dependent sequences. *Stochastic Process. Appl.*, 119(2):468–490, 2009.

H. Rootzén, M.R. Leadbetter, and L. de Haan. On the distribution of tail array sums for strongly mixing stationary sequence. *Ann. Appl. Probab.*, 8(3):868–885, 1998.

SmartTail, 2015. *url:* www.smarttail.se - Software for the analysis of false discovery rates in high-throughput screening experiments.

J.E. Taylor and K.J. Worsley. Inference for magnitudes and delays of response in the FIAC data using BRAINSTAT/FMRISTAT. *Human Brain Mapping*, 27:434–441, 2006.

K. Zhao, M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet*, 3(1):71–82, 2007.

D. Zholud. Tail approximations for the student t-, F-, and Welch statistics for non-normal and not necessarily i.i.d. random variables. *Bernoulli*, 20(4):2102–2130, 2014.