Web-based Supplementary Materials for "Marginal screening of 2×2 tables in large-scale case-control studies"

by Ian W. McKeague and Min Qian

July 23, 2018

1 Forward-stepwise screening

It is important in practice to carry out forward-stepwise screening in order to identify further associations from the collection of 2×2 tables. When the risk factors are almost uncorrelated, a straightforward approach is to apply the BST after the removal of each previously selected risk factor. In general, however, some risk factors may be highly correlated, so it is necessary to adjust for those previously selected when screening the remaining ones. The way we propose to do this is to use the BST procedure sequentially as follows. Given that the initial screening yields a significant result, the selected risk factor $\hat{k}_N$ is used to split the data on the remaining risk factors into two collections of $p-1$ tables: exposed or unexposed to $\hat{k}_N$. Mantel–Haenszel odds-ratio estimates from the pairs of these $2(p-1)$ tables corresponding to the same risk factor are then used in place of $\hat{\theta}_k$, and similarly for the standard errors $\hat{\tau}_k$, to form a new test statistic T_N. This statistic is calibrated in the same way as before,

except that the new correlation matrix $C_{\mathbf{X}}$ is estimated by the $(p-1)\times(p-1)$ submatrix of the original estimate $C_{\mathbf{X}}$ excluding the entries involving $\hat{k}_N$. The procedure is iterated until no further risk factors are found to be significant.

2 Proof of Theorem 1

We first derive the limiting distribution of $(\boldsymbol{X}-E\boldsymbol{X})/\sqrt{M_1}$ using the Lindeberg–Feller theorem, which applies here since under (C1) we have that $\boldsymbol{X}$ is the sum of the independent random vectors

$$L_s = \big(I(\eta_{s1}\le(1+\theta_1^{-1}\lambda_{s1})^{-1})\delta_{s1},\ldots,I(\eta_{sp}\le(1+\theta_p^{-1}\lambda_{sp})^{-1})\delta_{sp}\big)^{\mathsf{T}},\quad s=1,\ldots,M_1.$$

We need to check that $\mathrm{Cov}(\mathbf{X})/M_1$ converges to an invertible matrix. Note that the Lindeberg condition holds in our case since the summands are uniformly bounded.

Using Theorem 3.3 of Kou and Ying (1996) that gives $\mathrm{Var}(X_k)\tau_k^2\to 1$, by Remark 2 we have

$$\mathrm{Var}(X_k)/M_1 = \mathrm{Var}(X_k)\tau_k^2/(M_1\tau_k^2)\to(\pi\sigma_k^2)^{-1}$$

for each $k=1,\ldots,p$, under the stability condition (C3). In terms of the conditional correlation matrix of $\boldsymbol{X}$, denoted $C_{\boldsymbol{X},N}$, and using (C2), we then obtain the invertible limit

$$\begin{aligned}\mathrm{Cov}(\mathbf{X})/M_1 &= \mathrm{diag}\left(\sqrt{\mathrm{Var}(\boldsymbol{X})/M_1}\right)C_{\boldsymbol{X},N}\mathrm{diag}\left(\sqrt{\mathrm{Var}(\boldsymbol{X})/M_1}\right)\\ &\to D_{\boldsymbol{X}}C_{\boldsymbol{X}}D_{\boldsymbol{X}}/\pi,\end{aligned}$$

where $D_{\boldsymbol{X}} = \operatorname{diag}\left(\sigma_1^{-1}, \ldots, \sigma_p^{-1}\right)$.

The multivariate Lindeberg–Feller theorem, as stated e.g. in van der Vaart (1998), then gives

$$M_1^{-1/2}(\boldsymbol{X} - E\boldsymbol{X}) \xrightarrow{d} N(0, D_{\boldsymbol{X}} C_{\boldsymbol{X}} D_{\boldsymbol{X}}/\pi). \tag{S.1}$$

By Theorem 3.4 of Kou and Ying (1996),

$$\log \hat{\theta}_k - \log \theta_k = \tau_k^2 (X_k - EX_k) + O_P(\tau_k^2) \tag{S.2}$$

as $N \to \infty$, for each $k = 1, \ldots, p$. Denote $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^{\mathsf{T}}$, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^{\mathsf{T}}$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathsf{T}}$. Then by (S.1), (S.2), and Slutsky's lemma,

$$\begin{aligned} \sqrt{N}(\log \hat{\boldsymbol{\theta}} - \log \boldsymbol{\theta}) &= \sqrt{M_1/N}(N\boldsymbol{\tau}^2) \odot [M_1^{-1/2}(\boldsymbol{X} - E\boldsymbol{X})] + \sqrt{N} O_P(\boldsymbol{\tau}^2) \\ &\xrightarrow{d} (\sigma_1 Z_1, \ldots, \sigma_p Z_p)^{\mathsf{T}} \end{aligned} \tag{S.3}$$

where $\odot$ denotes the elementwise (Hadamard) product. The result then follows immediately from the following two lemmas. The first lemma gives the (regular) limiting behavior of $\log \hat{\theta}_N$ under the assumption that there is at least one active risk factor; the proof is included for completeness. The second lemma gives the (non-regular) limiting behavior of $\log \hat{\theta}_N$ when no risk factor is associated with the outcome.

Lemma 1. *If all conditions in Theorem 1 hold and $\boldsymbol{\theta}^{(0)} \neq \mathbf{1}$, then $\hat{k}_N \xrightarrow{P_N} k_0$ and $\sqrt{N}(\log \hat{\theta}_N - \log \theta_N) \xrightarrow{d} \sigma_{k_0} Z_{k_0}$, where Z_{k_0} is defined in Theorem 1.*

Proof. By Lemma 3.1 of Kou and Ying (1996), we have $\hat{\tau}_k/\tau_k \to 1$ in probability.

Thus

$$\begin{aligned}\hat{k}_N &= \arg\max_k |\log \hat{\theta}_k/\hat{\tau}_k| \\ &= \arg\max_k (\sqrt{N}\hat{\tau}_k)^{-1} \big| \log \theta_k^{(0)} + N^{-1/2} b_k + \tau_k^2 (X_k - EX_k) + O_{P_N}(\tau_k^2) \big| \\ &\xrightarrow{P_N} \arg\max_k \sigma_k^{-1} |\log \theta_k^{(0)}| = k_0,\end{aligned}$$

and $k_N = \arg\max_k |\log \theta_k/\tau_k| = \arg\max_k (\sqrt{N}\tau_k)^{-1} \big| \log \theta_k^{(0)} + N^{-1/2} b_k| \to k_0$. Hence

$$\begin{aligned}&\sqrt{N}\big(\log \hat{\theta}_N - \log \theta_N\big) \\ &\quad = \sqrt{N}\big(\log \hat{\theta}_N - \log \hat{\theta}_{k_0}\big) + \sqrt{N}\big(\log \hat{\theta}_{k_0} - \log \theta_0\big) + \sqrt{N}\big(\log \theta_{k_0} - \log \theta_N\big) \\ &\quad = \sqrt{N}\big(\log \hat{\theta}_{k_0} - \log \theta_0\big) + o_{P_N}(1) \\ &\quad \xrightarrow{d} \ \sigma_{k_0} Z_{k_0},\end{aligned}$$

where the second equality uses $\hat{k}_N \xrightarrow{P_N} k_0$ and $k_N \to k_0$ as $N \to \infty$. □

Lemma 2. *If all conditions in Theorem 1 hold and* $\boldsymbol{\theta}^{(0)} = \mathbf{1}$*, then* $\sqrt{N}(\log \hat{\theta}_N - \log \theta_N) \xrightarrow{d} \sigma_K Z_K + b_K - b_{\tilde{k}}$.

Proof. When $\boldsymbol{\theta}^{(0)} = \mathbf{1}$, $\log \boldsymbol{\theta} = N^{-1/2}\boldsymbol{b}$, and we have $k_N \in \arg\max_k |b_k(\sqrt{N}\tau_k)^{-1}| \to \arg\max_k |b_k \sigma_k^{-1}| = \tilde{k}$. Since $(Z_1, \ldots, Z_p)$ is multivariate normal with a non-singular covariance matrix, it is easy to see that

$$(Z_j + b_j/\sigma_j)^2 \neq (Z_k + b_k/\sigma_k)^2 \text{ for any } j \neq k \ \text{ a.s.}, \tag{S.4}$$

which implies that K is unique a.s. Also note that using (S.3) we have

$$\begin{aligned}\hat{\boldsymbol{\tau}}^{-1} \odot \log \hat{\boldsymbol{\theta}} &= \hat{\boldsymbol{\tau}}^{-1} \odot \boldsymbol{\tau} \odot \boldsymbol{\tau} \odot \mathrm{Var}(\boldsymbol{X})^{1/2} \odot \mathrm{Var}(\boldsymbol{X})^{-1/2} \odot (\boldsymbol{X} - E\boldsymbol{X}) \\ &\quad + O_{P_N}(\boldsymbol{\tau}) + (\sqrt{N}\hat{\boldsymbol{\tau}})^{-1} \odot \boldsymbol{b} \\ &\xrightarrow{d} (Z_1 + b_1/\sigma_1, \ldots, Z_p + b_p/\sigma_p)^{\mathsf{T}}.\end{aligned}$$

Define $h(\mathbf{t}) = (1_{\arg\max_k t_k = 1}, \ldots, 1_{\arg\max_k t_k = p})^{\mathsf{T}}$, where $\mathbf{t} = (t_1, \ldots, t_p)^{\mathsf{T}} \in \mathbb{R}^p$. Note that h is continuous at $\mathbf{t}$ if $\arg\max_k t_k$ is unique. Thus, using (S.4) and since $\sqrt{N} \log \hat{\theta}_N = (\sqrt{N}\hat{\boldsymbol{\tau}} \odot \hat{\boldsymbol{\tau}}^{-1} \odot \log \hat{\boldsymbol{\theta}})^{\mathsf{T}} h\big(|\hat{\boldsymbol{\tau}}^{-1} \odot \log \hat{\boldsymbol{\theta}}|\big)$, the result follows by applying Slutsky's lemma and the continuous mapping theorem in conjunction with the above display. □

3 HC and SGoF tests

Below we specify the procedures used to implement the higher criticism (Donoho and Jin, 2015) and exact binomial sequential goodness-of-fit metatest (Carvajal-Rodriguez et al., 2009) approaches that are compared with BST in our simulation study.

Higher Criticism (HC)

1. Carry out the standard Chi-squared test for a marginal association between the disease status D and each risk factor W_k, $k = 1, \ldots, p$. This yields a list of p-values $\pi_1, \pi_2, \ldots, \pi_p$.

2. Sort the p-values in ascending order: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$.

3. Calculate the Higher Criticism statistic

$$\mathrm{HC}^{+} = \max_{\{k:\, 1\le k\le p/2,\ \pi_{(k)}>1/p\}} \mathrm{HC}_k,$$

where $\mathrm{HC}_k = \sqrt{p}(k/p - \pi_{(k)})/\sqrt{\pi_{(k)}(1-\pi_{(k)})}$.

4. At significance level α, calculate the (Gumbel distribution) critical value

$$h_G(\alpha) = \frac{2\log\log(p) + \frac{1}{2}\big[\log\log\log(p) - \log(4\pi)\big] - \log\log[1/(1-\alpha)]}{\sqrt{2\log\log(p)}}.$$

5. Reject the global null hypothesis (1) if HC^{+} is larger than $h_G(\alpha)$.

Sequential goodness-of-fit metatest (SGoF)

1. Compile the same list of p-values $\pi_1, \pi_2, \ldots, \pi_p$ as in step 1 of the HC test.

2. At significance level α (or the false discovery rate in the full version of SGoF), calculate K, the number of p-values $\le \alpha$.

3. Calculate the critical value K_α as the upper α-quantile of the binomial distribution with number of trials p and probability of success α.

4. Reject the global null hypothesis (1) if $K \ge K_\alpha$.

It is important to note that the HC and SGoF tests are each calibrated under the assumption of *independent* p-values, which is of limited relevance in our setting, and which explains why they do not perform well in the highly-dependent p-value situation of Table 3 (they only respect the nominal level α in the independent case of Table 1, and in the weakly-dependent case of Table 2).

References

Carvajal-Rodriguez, A., de Una-Alvarez, J. and E. Rolan-Alvarez (2009). A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics*, 10:209.

Donoho, D. and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30, 1–25.

Kou, S. G. and Ying, Z. (1996). Asymptotics for a 2×2 table with fixed margins. *Statistica Sinica*, 6, 809–829.

van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press.