SUPPLEMENTARY MATERIALS

(A): A Bivariate Normal Copula

A bivariate normal copula is expressed as

$$C(u_1,u_2)=\Phi_\theta(\Phi^{-1}(u_1),\Phi^{-1}(u_2)), \tag{1}$$

where

$$\Phi_\theta(u_1,u_2)=\int_{-\infty}^{\Phi^{-1}(u_1)}\int_{-\infty}^{\Phi^{-1}(u_2)}\frac{1}{2\pi\sqrt{1-\theta^2}}\,exp\left[-\frac{x^2-2\theta xy+y^2}{2(1-\theta^2)}\right]dxdy \tag{2}$$

is the bivariate standard normal distribution function with the correlation parameter $\theta\in[-1,1]$ and

$$\Phi(u_i)=\int_{-\infty}^{\Phi^{-1}(u_i)}\frac{1}{\sqrt{2\pi}}\,exp\left[-\frac{1}{2}x^2\right]dx, i=1,2 \tag{3}$$

denotes the univariate standard normal distribution function. We find the probability density function of copula, $c(u_1,u_2)$ by differentiating the $C(u_1,u_2)$ with respect to u_1 and u_2. i.e.

$$c(u_1,u_2)=\frac{\partial C(u_1,u_2)}{\partial u_1\,\partial u_2}. \tag{4}$$

Let $q_i=\Phi^{-1}(u_i)$. Therefore,

$$c(u_1,u_2)=\frac{\partial\Phi(q_1,q_2)}{\partial q_1\,\partial q_2}\frac{\partial q_1}{\partial u_1}\frac{\partial q_2}{\partial u_2}. \tag{5}$$

But

$$\frac{\partial q_i}{\partial u_i}=\frac{\partial\Phi^{-1}(u_i)}{\partial u_i}=\left(\frac{\partial\Phi(q_i)}{\partial q_i}\right)^{-1}. \tag{6}$$

Equation (5) therefore becomes

$$c(u_1,u_2)=\frac{\partial\Phi}{\partial q_1\,\partial q_2}\left(\frac{\partial\Phi(q_1)}{\partial q_1}\right)^{-1}\left(\frac{\partial\Phi(q_2)}{\partial q_2}\right)^{-1}. \tag{7}$$

The copula density function thus becomes

$$\begin{aligned}c(u_1,u_2)=\ &\frac{1}{\sqrt{1-\theta^2}}exp[\frac{1}{2}(\Phi^{-1}(u_1))^2+(\Phi^{-1}(u_2))^2\\&-\frac{(\Phi^{-1}(u_1))^2+2\theta\Phi^{-1}(u_1)\Phi^{-1}(u_2)+(\Phi^{-1}(u_2))^2}{2(1-\theta^2)}],\end{aligned} \tag{8}$$

and the likelihood function in terms of the normal copula is

$$f(x_1,x_2)=c(u_1,u_2)\prod_{i=1}^{2}f_i(x_i). \tag{9}$$

$f(x_1,x_2)$ reduces to a bivariate normal if $f_i(x_i)$ is normal.

The log-likelihood function becomes

$$\ell_n(\theta) = \sum_{j=1}^{n} log\left[c(u_1, u_2)\right] + \sum_{j=1}^{n}\sum_{i=1}^{2} log(f_i(x_i)) \,. \tag{10}$$

The dependence parameter θ is then estimated as

$$\hat{\theta} = \arg\, \max_{\theta \in \Theta} \ell_n(\theta) \,. \tag{11}$$

The Kendall's τ and Spearman's ρ for the normal copula are given as

$$\tau = \frac{2}{\pi} \text{arc}\, \sin(\theta) \tag{12}$$

and

$$\rho = \frac{6}{\pi} \text{arc}\, \sin(\frac{\theta}{2}) \,, \tag{13}$$

respectively. For proofs of (12) and (13), see [1].

(B): Canonical Maximum Likelihood Estimation (CMLE) Method

In this approach, no parametric assumptions are made on the marginals and therefore, it relies on the concept of empirical marginal transformation. The transformation approximates the unknown parametric marginal $F_i(x_i)$ with empirical distribution function $\hat{F}_i(x_i)$ known as a pseudo-sample (u_i) . $\hat{F}_i(x_i)$ is given by

$$\hat{F}_i(x_i) = \frac{n}{n+1}\frac{1}{n}\sum_{j=1}^{n} I\left(X_i \le x_i\right) \tag{14}$$

where I is the indicator function. Rescaling the empirical distribution by $\frac{n}{n+1}$ avoids the the boundary values. The log-likelihood function becomes

$$\ell(\theta) = \sum_{j=1}^{n} log\, c(\hat{F}_1(x_1), ..., \hat{F}_m(x_m); \theta) + \sum_{j=1}^{n}\sum_{i=1}^{m} log(f_i(x_i)) \tag{15}$$

The dependence parameter θ_i is then estimated as

$$\hat{\theta} \sim \arg\, \max \sum_{j=1}^{n} log\, c(\hat{F}_1(x_1), ..., \hat{F}_m(x_m); \theta), \tag{16}$$

since the last summand does not depend on θ. Genest et al. [2] showed that the resulting pseudo-likelihood estimator is consistent and asymptotically normal under the condition that F_i is continuous.

(C): Permutation-Based Resampling Method

Permutation approach provides an efficient method to testing when data do not conform to the distribution assumptions. To compute unadjusted *p*-value for each gene, we follow the procedure below.

1. Permutate the quantitative outcome column B times as you hold the gene expressions matrix fixed.

2. For the b^{th} permutation, $b = 1, ..., B$, compute test statistics $\hat{\theta}_{1b}, ..., \hat{\theta}_{Gb}$ for each hypothesis using equation (16).

3. After the B permutations are done, for two-sided alternative hypotheses, the permutation *p*-value for hypothesis H_i is

$$p_i = \frac{\#\{b : |\hat{\theta}_{ib}| \geq |\hat{\theta}_i|\}}{B}, \tag{17}$$

where $\hat{\theta}_i$ is the original $\hat{\theta}$ for the i^{th} gene before the permutation.

(D): Estimation of the False Discovery Rate (FDR)

The procedure below outlines the steps followed in the estimation of FDR for a given p-value [3].

1. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(G)}$ be the ordered p-values. This also denotes the ordering of the features in terms of their evidence against the null hypothesis.

2. For a range of λ, say λ = 0, 0.01, 0.02, ... , 0.95, calculate

$$\hat{\pi}_0(\lambda) = \frac{\#(p_j > \lambda)}{G(1-\lambda)}.$$

3. Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ.

4. Set the estimate of π_0 to be $\hat{\pi}_0 = \hat{f}(1)$.

5. Calculate $\hat{q}(p_{(G)}) = \hat{\pi}_0 p_{(G)}$.

6. For $i = G-1, G-2, \ldots, 1$, calculate

$$\hat{q}(p_{(i)}) = \min\left\{\frac{\hat{\pi}_0 G p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right\}.$$

7. The estimated *q*-value for the i^{th} most significant feature is $\hat{q}(p_{(i)})$.

REFERENCES

[1] McNeil AJ, Frey R, Embrechts P. Quantitative risk management: concepts, techniques and tools. Princeton series in finance. Princeton, N.J: Princeton University Press; 2005. OCLC: ocm60796246.

[2] Genest C, Ghoudi K, Rvest LP. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 1995; 82(3): 543-552. https://doi.org/10.1093/biomet/82.3.543

[3] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003; 100: 9440-9445. https://doi.org/10.1073/pnas.1530509100