BIOMETRICS 000, 000–000 DOI: 000
000 0000

Web-based Supplementary materials for "A regression framework for assessing covariate effects on the reproducibility of high-throughput experiments" by Qunhua Li and Feipeng Zhang

This paper has been submitted for consideration for publication in *Biometrics*

1. Theoratical derivation

1.1 *Asymptotic properties*

To investigate the asymptotic properties of the proposed estimator in Section 3, we apply the standard theory of maximum likelihood estimators, similar to that of the ordinal regression. For ease of notation, we denote $\dot{f}(u) = \frac{df(u)}{du}$ and $\ddot{f}(u) = \frac{d^2f(u)}{du^2}$ as the first and second derivatives, respectively, $\Delta f(t_m) = f(t_m) - f(t_{m-1})$ as the difference for any function f, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for any vector $\mathbf{a}$. We rewrite the log of the likelihood function (8) as

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= n^{-1}\sum_{s=1}^{S}\sum_{i=1}^{n}\sum_{m=1}^{M} U_{im}^{s} \log\left(\Delta g^{-1}\left\{\boldsymbol{z}(t_m)^T\boldsymbol{\theta}\right\}\right)\\ &\equiv n^{-1}\sum_{s=1}^{S}\sum_{i=1}^{n}\sum_{m=1}^{M} U_{im}^{s} \log\left(\Delta g^{-1}\left\{\boldsymbol{h}(t_m)^T\boldsymbol{\alpha} + (\boldsymbol{x}^s)^T A(t_m)\boldsymbol{\beta}\right\}\right),\end{aligned}$$

where $\boldsymbol{h}(t) = (h_1(t), \ldots, h_m(t))^T$, and $A(t) = I_d \otimes \boldsymbol{h}(t)$, I_d is an identity matrix with d dimensions. The negative of the second derivative (or Hessian) matrix of the log-likelihood is given by

$$\widehat{V} = -n^{-1}\sum_{s=1}^{S}\sum_{i=1}^{n}\sum_{m=1}^{M} U_{im}^{s} \frac{\Delta \ddot{g}^{-1}\left\{\boldsymbol{z}(t_m)^T\boldsymbol{\theta}\right\}\boldsymbol{z}(t_m)^{\otimes 2} - \left[\Delta \dot{g}^{-1}\left\{\boldsymbol{z}(t_m)^T\boldsymbol{\theta}\right\}\boldsymbol{z}(t_m)\right]^{\otimes 2}}{\left[\Delta g^{-1}\left\{\boldsymbol{z}(t_m)^T\boldsymbol{\theta}\right\}\right]^{\otimes 2}}.$$

and the Fisher information matrix is $V = \mathrm{E}\widehat{V}$. Consistency and asymptotic normality follow straightforwardly under standard regularity conditions.

THEOREM 1: *Suppose that all the possible parameter values $\boldsymbol{\theta}$ are in a compact set, and V is a nonnegative definite matrix, then $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ in probability, as n goes to infinity. Furthermore, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normally distributed with mean zero and covariance matrix V^{-1}, as n goes to infinity.*

1.2 *Lower tail dependence*

Lower tail dependence measures the dependence betwen the variables in the lower-left quadrant of $[0,1]^2$ (See Nelsen (2006)). The coefficient of lower tail dependence is defined by

$$\lambda_L := \lim_{u\to 0+} P(F(X) < u | G(Y) < u) = \lim_{u\to 0+} \frac{C(u,u)}{u}.$$

For all of the Archimedean copulas in Table 4.1 in Nelsen (2006), their lower tail dependence are evaluated and summarized in Example 5.22 in Nelsen (2006). In particular, the relationship between the association parameter of the copula (θ) and the lower tail dependence is identical in the Nelson 4.2.12 and Clayton copulas, i.e., $\lambda_L = 2^{-1/\theta}$.

2. Appendix: Archimedean copulas in the homogeneous reproducibility class

In this appendix, we include a list of Archimedean copulas from Table 4.1 in Nelsen (2006) that can be written in the regression form of (2).

(1) Copula function: $C_\theta(t_1, t_2) = \max(1 - [(1-t_1)^\theta + (1-t_2)^\theta]^{1/\theta}, 0)$, where $\theta \in [1, \infty)$.

Name: (4.2.2) in Nelsen

Generator function: $\psi(t) = (1-t)^\theta$

Regression form: $g(C(t,t)) = \alpha_1 h(t)$, where $g(C(t,t)) = 1 - C(t,t)$, $h(t) = 1-t$, $\alpha_1 = 2^{1/\theta}$.

(2) Copula function: $C_\theta(t_1, t_2) = \frac{t_1 t_2}{1-\theta(1-t_1)(1-t_2)}$, where $\theta \in [-1, 1)$.

Name: Ali-Mikhail-Haq family, (4.2.3) in Nelsen

Generator function: $\log \frac{1-\theta(1-t)}{t}$

Regression form: $g(C(t,t)) = \alpha_1 h(t) + \alpha_2 h^2(t)$, where $g(C(t,t)) = \frac{1-C(t,t)}{C(t,t)}$, $h(t) = \frac{1-t}{t}$, $\alpha_1 = 2$, $\alpha_2 = (1-\theta)$.

(3) Copula function: $C_\theta(t_1, t_2) = \exp(-[(-\log t_1)^\theta + (-\log t_2)^\theta]^{1/\theta})$, where $\theta \in [1, \infty)$.

Name: Gumbel-Hougaard family, (4.2.4) in Nelsen

Generator function: $(-\log t)^\theta$

Regression form: $g(C(t,t)) = \alpha_1 h(t)$, where $g(C(t,t)) = \log(C(t,t))$, $h(t) = \log(t)$, $\alpha_1 = 2^{1/\theta}$.

(4) Copula function: $C_\theta(t_1, t_2) = \max(\theta t_1 t_2 + (1-\theta)(t_1 + t_2 - 1), 0)$, where $\theta \in (0, 1]$

Name: (4.2.7) in Nelsen

Generator function: $\psi(t) = -\log[\theta t + (1-\theta)]$

Regression form: $g(C(t,t)) = \alpha_1 h(t) + \alpha_2 h(t)^2$, where $g(C(t,t)) = 1 - C(t,t)$, $h(t) = 1 - t$, $\alpha_1 = 2$, $\alpha_2 = -\theta$

(5) Copula function: $C_\theta(t_1, t_2) = t_1 t_2 \exp(-\theta \log t_1 \log t_2)$, where $\theta \in (0, 1]$

Name: Gumbel-Barnett family, (4.2.9) in Nelsen

Generator function: $\psi(t) = \log(1 - \theta \log t)$

Regression form: $g(C(t,t)) = \alpha_1 h(t) + \alpha_2 h^2(t)$, where $g(C(t,t)) = \log C(t,t)$, $h(t) = \log t$, $\alpha_1 = 2$, $\alpha_2 = -\theta$.

(6) Copula function: $C_\theta(t_1, t_2) = (1 + [(t_1^{-1} - 1)^\theta + (t_2^{-1} - 1)^\theta]^{1/\theta})^{-1}$, where $\theta \in [1, \infty)$

Name: (4.2.12) in Nelsen

Generator function: $\psi(t) = (\frac{1}{t} - 1)^\theta$

Regression form: $g(C(t,t)) = \alpha_1 h_1(t) + \alpha_2 h_2(t)$, where $g(C(t,t)) = \log \frac{C(t,t)}{1-C(t,t)}$, $h_1(t) = 1$, $h_2(t) = \log \frac{t}{1-t}$, $\alpha_1 = -\frac{\log 2}{\theta}$, $\alpha_2 = 1$.

(7) Copula function: $C_\theta(t_1, t_2) = \max(1 + \frac{\theta}{\log[e^{\theta/(t_1-1)} + e^{\theta/t_2 - 1}]}, 0)$, where $\theta \in [2, \infty)$

Name: (4.2.18) in Nelsen

Generator function: $\psi(t) = \exp(\frac{\theta}{t-1})$

Regression form: $g(C(t,t)) = \alpha_1 h_1(t) + \alpha_2 h_2(t)$, where $g(C(t,t)) = \frac{1}{1-C(t,t)}$, $h_1(t) = 1$, $h_2(t) = \frac{1}{1-t}$, $\alpha_1 = -\frac{\log 2}{\theta}$, $\alpha_2 = 1$.

3. Supplementary figures

[Figure 1 about here.]

4. Supplementary tables

[Table 1 about here.]

[Table 2 about here.]

References

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G., et al. (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods* **2,** 345–350.

Li, Q., Brown, J. B., Huang, H., Bickel, P. J., et al. (2011). Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5,** 1752–1779.

Nelsen, R. B. (2006). *An Introduction to Copula, 2nd.* Springer Verlag, New York.

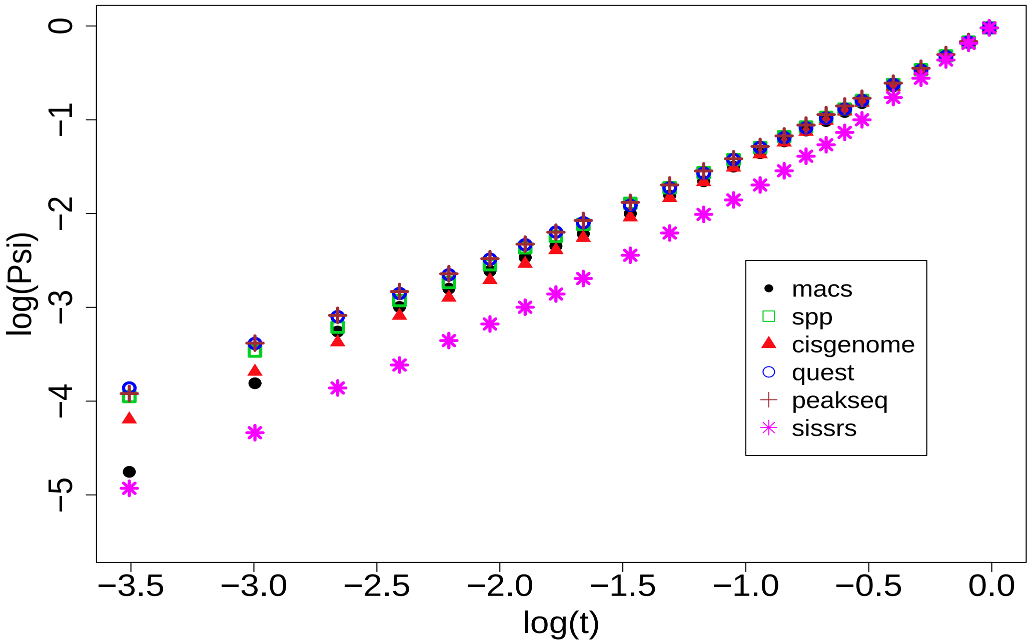

Figure 1: Verification of the functional form for the regression model for the ChIP-seq data in Li et al. (2011). Empirical data shows that there is an approximated linear trend between $\log(\Psi(t))$ and $\log(t)$ for most peak callers.

Table 1: Accuracy of estimation at different sample sizes and spacing of cutoffs under canonical model specification. Data are simulated from a Gumbel-Hougaard copula with $\theta_G = 1.0, 2.0$ and 3.0. The table shows the mean and the standard deviation of the estimated parameters α_1 over n data sets with $M =$ 20, 50 and 100.

		n=500			n=1,000			n=10,000		
θ_G	$\alpha_1 = 2^{\frac{1}{\theta_G}}$	$M = 20$	$M = 50$	$M = 100$	$M = 20$	$M = 50$	$M = 100$	$M = 20$	$M = 50$	$M = 100$
1	2.000	1.978	2.001	2.007	1.989	1.997	1.992	1.997	1.996	1.996
		0.096	0.090	0.094	0.068	0.066	0.073	0.019	0.021	0.019
2	1.414	1.402	1.404	1.416	1.416	1.415	1.418	1.410	1.413	1.414
		0.130	0.152	0.136	0.092	0.075	0.082	0.027	0.028	0.026
3	1.260	1.242	1.254	1.240	1.253	1.255	1.259	1.257	1.258	1.257
		0.170	0.199	0.192	0.119	0.117	0.116	0.037	0.044	0.045

Table 2: Lab and platform effects on the reproducibility of differentially expressed genes in a microarray study (Irizarry et al., 2005). The data is fitted using the regression model: $\mathrm{logit}(\Psi(t)) = \alpha_1 + \alpha_2\mathrm{logit}(t) + \beta_L X_L + \beta_P X_P + \beta_{PL} X_P X_L + \beta_{Lt} X_L \mathrm{logit}(t) + \beta_{Pt} X_P \mathrm{logit}(t) + \beta_{PLt} X_P X_L \mathrm{logit}(t)$.

		Estimate	95% confidence interval
Baseline	α_1	-0.520	[-0.768, -0.272]
	α_2	0.917	[0.809, 1.026]
Two-color oligo	β_P	-0.377	[-0.738, -0.015]
Lab 2	β_L	0.033	[-0.317, 0.384]
Two-color*Lab 2	β_{LP}	0.050	[-0.458, 0.558]
Two-color*logit(t)	β_{Pt}	0.142	[-0.024, 0.309]
Lab 2*logit(t)	β_{Lt}	0.006	[-0.148, 0.159]
Two-color*Lab 2*logit(t)	β_{PLt}	-0.067	[-0.299, 0.165]