1 Methods

Suppose for each subject $i = 1, \ldots, n$, we have k longitudinal measurements $y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})'$, p SNPs of interest as a row vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ with x_{id} coded as 0,1 or 2 for the count of the minor allel for SNP $j = 1, \ldots, p$, and $z_i = (z_{i1}, z_{i2}, \ldots, z_{iq})$ is a row vector for q variates. We assume common effect sizes of the SNPs and covariates on the longitudinal phenotype/trait measurements. Thus, we construct the design matrix for the SNPs and covariates as:

$$X_i = \begin{pmatrix} x_i \\ x_i \\ \vdots \\ x_i \end{pmatrix}, Z_i = \begin{pmatrix} 1 & z_i \\ 1 & z_i \\ \vdots & \vdots \\ 1 & z_i \end{pmatrix}$$

where x_i and z_i are row vectors of length p and q respectively. X_i is a $k \times p$ matrix, and Z_i is a $k \times (q+1)$ matrix. Denote the regression coefficient $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_{q+1})'$ for X_i and Z_i respectively. The marginal mean of each measurement, $E(y_{im}|x_i, z_i) = \mu_{im}$, relates to the SNPs and covariates through a generalized linear model (GLM):

$$g(\mu_i) = \eta_i = Z_i\varphi + X_i\beta = H_i\theta$$

with $H_i = (Z_i, X_i), \theta = (\varphi', \beta')'$ and $g(.)$ as a suitable link function.

The estimates of β and φ can be obtained by solving the GEE [Liang and Zeger, 1986]:

$$U(\beta) = \sum_{i=}^{n} U_i(\beta) = \sum_{i=1}^{n} (\frac{\partial \mu_i}{\partial \theta'}) A_i^{-\frac{1}{2}} R_w(\alpha)^{-1} A_i^{-\frac{1}{2}} (Y_i - \mu_i) = 0,$$

with

$$\frac{\partial \mu_i}{\partial \theta'} = \frac{\partial g^{-1}(H_i\theta)}{\partial \theta'}, \quad A_i = \operatorname{diag}\left(v(\mu_{i1}), v(\mu_{i2}), \ldots, v(\mu_{ik})\right).$$

$R_w(\alpha)$ is a working correlation matrix depending on some unknown parameter α. For convenience, a working independence model with $R_w = I$ is often used. With a canonical link function and a working independence model, we have a closed form of the U vector with two parts corresponding to SNPs and covariates, and its covariance estimator:

$$U = (U_{.1}, U_{.2})' = \sum_i (Z_i, X_i)' (Y_i - \mu_i)$$

$$\hat{\Sigma} = \sum_i (Z_i, X_i)' \widehat{\operatorname{var}}(Y_i) (Z_i, X_i) = \sum_i (Z_i, X_i)' (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' (Z_i, X_i) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \tag{1}$$

The SPU test is defined as

$$T_{SPU(\gamma)} = \sum_{j=1}^{p} U_{.2,j}^{\gamma-1} U_{.2,j}$$

with weight

$$W_j = U_{.2,j}^{\gamma-1}$$

for a series of integer value $\gamma = 1, 2, \ldots, \infty$, leading to the sum of powered score (U) tests abbreviated as SPU tests. We often use $\gamma \in \Gamma = \{1, 2, \ldots, 8, \infty\}$, as a γ greater than 8 was demonstrated to perform similarly to $\gamma = \infty$ [Pan et al., 2014]. In particular, when $\gamma \to \infty$ we have: $T_{SPU(\gamma)} \propto \underset{j}{max}|U_{.2,j}|$, which behaves similarly to the **UminP** test.

We will use a simulation method to calculate the p-value from each $T_{SPU(\gamma)}$ [Lin, 2005, Seaman and Müller-Myhsok, 2005]. Specifically, suppose T is short notation of $T_{SPU(\gamma)}$ for a specific γ and $\hat{\Sigma}_{.2}$ is the covariance matrix of the score vector $U_{.2}$ based on original data (see Equation 1). We draw B samples of the score vector from its null distribution: $U_{.2}^{(b)} \sim MVN\left(0, \hat{\Sigma}_{.2}\right)$, with $b = 1, 2, \ldots, B$, and thus obtain a statistics under null hypothesis: $T^{(b)} = \sum_{j=1}^{p} U_{.2,j}^{(b)\gamma}$. We then can calculate the p-value of $T_{SPU(\gamma)}$ as $P_{SPU(\gamma)} = \sum_{b=1}^{B} \frac{I(T^{(b)} \geq T^{obs})+1}{B+1}$.

aSPU family tests

The aSPU test is designed with the aim of data adaptively combining the results of multiple SPU tests. Specifically, it takes the minimum p-value from all SPU(γ) tests: $T_{aSPU} = \underset{\gamma\in\Gamma}{min} P_{SPU(\gamma)}$. The p-value of aSPU test can be obtained through similar simulation based method, or permutation/bootstrap given the asymptotic normality of the score vector (U) may not hold [Pan et al., 2014].

The aSPUw test is a diagonal-variance-weighted version of the SPU test, defined through:

$$\begin{aligned} T_{SPUw(\gamma)} &= \sum_{j=1}^{p} \left(\frac{U_{.2,j}}{\sqrt{\hat{\Sigma}_{.2,jj}}} \right)^{\gamma} \\ T_{aSPUw} &= \underset{\gamma\in\Gamma}{min} P_{SPUw(\gamma)} \end{aligned}$$

The aSPUw test is designed to complement the performance of aSPU test. As the standard deviations of SNVs in a region may vary a lot, there is possibility that a *non-informative* SNV has *larger* standard deviations than other associated SNVs, and the SPU test statistic will be dominated by the noise coming from the null but with larger standard deviation SNV, thus leads to concealing association signals and eventually reduce the test power. Another advantage **aSPUw** brings about is it makes jointly analyze the effect of RVs and CVs possible by giving them an inverse-standard-deviation weight closely related to MAF.

The aSPU(w).score test adds the GEE score test statistics into the aSPU(w) test. Specifically,

$$T_{aSPU.Score} = \min\left\{ \underset{\gamma\in\Gamma}{\min} P_{SPU(\gamma)}, P_{Score} \right\},$$

We expect to see this aSPU variant assimilates the power complemented by the GEE score test under some situations where the correlation among SNPs matters.

The aSPU.aSPUw.score test is a more comprehensive test which is designed to combine the complementing powers coming from aSPU, aSPU weighted and the Score tests. In most scenarios, it achieves a quasi-best performance. Specifically,

$$T_{aSPU.aSPUw.Score} = \min\Big\{\min_{\gamma\in\Gamma} P_{SPU(\gamma)}, \min_{\gamma\in\Gamma} P_{SPUw(\gamma)}, P_{Score}\Big\},$$

Permutation correction for aSPU family tests on rare variants

While MAF of RVs are usually low, e.g. between 0.001 to 0.01, the asymptotically Normal distribution of either *beta* coefficient or score vector may or may not hold. The **simulation-based** p-value calculating method is thus not sufficient for RV case and need modification. As a remedy, we developed a permutation algorithm that generates the empirical null distribution of $U_{.2}^{(b)}$ and in the same time maintains the relationship between longitudinal traits and possible covariates such as age, gender, etc, for subject i. The algorithm is also robust to missing data as this is a usual case in longitudinal data settings. The **permutation** algorithm can be implemented as follows:

1. identify the max k across all n subjects, which is the number of longitudinal measurements, e.g. $k = 4$.

2. detect if the data has missing values, if yes, fill the missing value with NA to complement the data dimension (for example, subject i with $Y_i = (y_{i,1},,,y_{i,4})'$ has two missing measurements at time 2 and time 3. After missing value complementing, it becomes $Y_i = (y_{i,1}, \text{NA}, \text{NA}, y_{i,4})'$). Now we should have all the subjects with each Y_i of dimension equal to $k \times 1$.

3. complement H_i to be of full dimension, i.e. $k \times (p + q + 1)$, for covariates and SNVs. Now we should have $\begin{pmatrix} Y_i & H_i \end{pmatrix}$ as an augmented matrix of dimension $k \times (p + q + 2)$ for each subject i, where $H_i = (Z_i, X_i)$. For total n subjects, we have row-wise binded matrix
$$M = \begin{pmatrix} Y_1 & H_1 \\ Y_2 & H_2 \\ \vdots & \vdots \\ Y_n & H_n \end{pmatrix}$$
of dimension $nk \times (p + q + 2)$.

4. permute the SNV chunk among different individuals, i.e. the X_i in $\begin{pmatrix} Y_i & Z_i, X_i \end{pmatrix}$ with the X_j in $\begin{pmatrix} Y_j & Z_j, X_j \end{pmatrix}$, where $i \neq j$.

5. with permuted
$$M^{*(b)} = \begin{pmatrix} Y_1 & Z_1, X_1^{*(b)} \\ Y_2 & Z_1, X_2^{*(b)} \\ \vdots & \vdots \\ Y_n & Z_1, X_n^{*(b)} \end{pmatrix}$$

we refit the GEE model and get the $U_{.2}^{*(b)}$

6. repeat step 4 - 5 B times to produce $U_{.2}^{*(b)}$ with $b = 1, 2, \ldots, B$.

After we get enough $U_{.2}^{*(b)}$ to form an empirical null distribution, the left work of aSPU test for RVs will be exactly the same as for CVs.

References

[Liang and Zeger, 1986] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

[Lin, 2005] Lin, D. (2005). An efficient monte carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787.

[Pan et al., 2014] Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics*, pages genetics–114.

[Seaman and Müller-Myhsok, 2005] Seaman, S. and Müller-Myhsok, B. (2005). Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *The American Journal of Human Genetics*, 76(3):399–408.