Supplementary Materials 1

Two dimensional Functional Principal Component Analysis

Consider a linear combination of functional values:

$$f = \int_S \int_T \beta(s,t)x(s,t)dsdt,$$

where $\beta(s,t)$ is a weight function and $x(s,t)$ is a centered random function. To capture the variations in the random functions, we chose weight function $\beta(s,t)$ to maximize the variance of f. By the formula for the variance of stochastic integral (1), we have

$$\text{var}(f) = \int_S \int_T \int_S \int_T \beta(s_1,t_1)R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_1dt_1ds_2t_2 \ , \qquad (1)$$

where $R(s_1,t_1,s_2,t_2) = \text{cov}(x(s_1,t_1),x(s_2,t_2))$ is the covariance function of the genetic variant function $x(s,t)$. Since multiplying $\beta(t)$ by a constant will not change the maximizer of the variance $Var(f)$, we impose a constraint to make the solution unique:

$$\int_T \int_T \beta^2(s,t)dsdt = 1. \qquad (2)$$

Therefore, to find the weight function, we seek to solve the following optimization problem:

$$\begin{aligned} &\max \quad \int_S \int_T \int_S \int_T \beta(s_1,t_1)R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_1dt_1ds_2t_2 \\ &\text{s.t.} \quad \int_T \int_T \beta^2(s,t)dsdt = 1. \end{aligned} \qquad (3)$$

By the Lagrange multiplier, we reformulate the constrained optimization problem (3) into the following non-constrained optimization problem:

$$\max_{\beta} \quad \frac{1}{2}\int_S\int_T\int_S\int_T \beta(s_1,t_1)R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_1dt_1ds_2t_2 + \frac{1}{2}\lambda(1-\int_T\int_T \beta^2(s_1,t_1)ds_1dt_1), \quad (4)$$

where λ is a parameter.

By variation calculus (2), we define the functional

$J[\beta] = \frac{1}{2}\int_S\int_T\int_S\int_T \beta(s_1,t_1)R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_1dt_1ds_2t_2 + \frac{1}{2}\lambda(1-\int_T\int_T \beta^2(s_1,t_1)ds_1dt_1)$. Its first

variation is given by

$$\begin{aligned}
\delta J[h] &= \frac{d}{d\varepsilon}J[\beta(s,t)+\varepsilon h(s,t)] \\
&= \frac{d}{d\varepsilon}\{\frac{1}{2}\int_S\int_T\int_S\int_T \beta(s_1,t_1)R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_1dt_1ds_2t_2 + \frac{1}{2}\lambda(1-\int_S\int_T \beta^2(s_1,t_1)ds_1dt_1)\}\,|_{\varepsilon=0} \\
&= \int_S\int_T\int_S\int_T h(s_1,t_1)R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_1dt_1ds_2t_2 - \lambda\int_S\int_T \beta(s_1,t_1)h(s_1,t_1)ds_1dt_1 \\
&= \int_S\int_T[\int_S\int_T[R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_2t_2 - \lambda\beta(s_1,t_1)]h(s_1,t_1)ds_1dt_1 \\
&= \int_S\int_T[\int_S\int_T R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_2t_2 - \lambda\beta(s_1,t_1)]^2 ds_1dt_1 = 0,
\end{aligned}$$

which implies the following integral equation

$$\int_S\int_T R(s_1,t_1,s_2,t_2)\beta(s_2.t_2)ds_2t_2 = \lambda\beta(s_1,t_1) \qquad (5)$$

for an appropriate eigenvalue λ. The left side of the integral equation (5) defines a two dimensional integral transform R of the weight function β. Therefore, the integral transform of the covariance function $R(s_1,t_1,s_2,t_2)$ is referred to as the covariance operator R. The integral equation (5) can be rewritten as

$$R\beta = \lambda\beta, \qquad (6)$$

where $\beta(s_1,t_1,s_2,t_2)$ is an eigenfunction and referred to as a principal component function. Equation (6) is also referred to as a two dimensional eigenequation. Clearly, the eigenequation (6)

looks the same as the eigenequation for the multivariate PCA if the covariance operator and eigenfunction are replaced by covariance matrix and eigenvector.

Since the number of function values is theoretically infinity, we may have an infinite number of eigenvalues. Provided the functions X_i and Y_i are not linearly dependent, there will be only $N-1$ nonzero eigenvalues, where N is the total number of sampled individuals ($N = n_A + n_G$). Eigenfunctions satisfying the eigenequation are orthonormal (Ramsay and Silverman, 2005). In other words, equation (6) generates a set of principal component functions

$$R\beta_k = \lambda_k \beta_k, \qquad \text{with } \lambda_1 \geq \lambda_2 \geq \cdots$$

These principal component functions satisfy

(1) $\int_S \int_T \beta_k^2(s,t)dsdt = 1$ and

(2) $\int_S \int_T \beta_k(s,t)\beta_m(s,t)dsdt = 0$, for all $m < k$.

The principal component function β_1 with the largest eigenvalue is referred to as the first principal component function, and the principal component function β_2 with the second largest eigenvalue is referred to as the second principal component function, and continues.

Computations for the Principal Component Function and the Principal Component Score

The eigenfunction is an integral function and difficult to solve in closed form. A general strategy for solving the eigenfunction problem in (5) is to convert the continuous eigen-analysis problem to an appropriate discrete eigen-analysis task (Ramsay and Silverman 2005). In this report, we use basis function expansion methods to achieve this conversion.

Let $\{\phi_j(t)\}$ be the series of Fourier functions. For each j, define $\omega_{2j-1} = \omega_{2j} = 2\pi j$. We expand each genetic variant profile $x_i(s,t)$ as a linear combination of the basis function ϕ_j:

$$x_i(s,t) = \sum_{k=1}^{K}\sum_{l=1}^{K} c_{kj}^{(i)}\phi_k(s)\phi_l(t). \tag{7}$$

Let $C_i = [c_{11}^{(i)},...,c_{1K}^{(i)},c_{21}^{(i)},...,c_{2K}^{(i)},...,c_{K1}^{(i)},...,c_{KK}^{(i)}]^T$ and $\phi(t) = [\phi_1(t),\cdots,\phi_K(t)]^T$. Then, equation (7) can be rewritten as

$$x_i(s,t) = C_i^T(\phi(s)\otimes\phi(t)),$$

where $\otimes$ denotes the Kronecker product of two matrices.

Define the vector-valued function $X(s,t) = [x_1(s,t),\cdots,x_N(s,t)]^T$. The joint expansion of all N random functions can be expressed as

$$X(s,t) = C(\phi(s)\otimes\phi(t)) \tag{8}$$

where the matrix C is given by

$$C = \begin{bmatrix} C_1^T \\ \vdots \\ C_N^T \end{bmatrix}.$$

In matrix form we can express the variance-covariance function of the genetic variant profiles as

$$\begin{aligned} R(s_1,t_1,s_2,t_2) &= \frac{1}{N}X^T(s_1,t_1)X(s_2,t_2) \\ &= \frac{1}{N}[\phi^T(s_1)\otimes\phi^T(t_1)C^TC[\phi(s_2)\otimes\phi(t_2)]. \end{aligned} \tag{9}$$

Similarly, the eigenfunction $\beta(s,t)$ can be expanded as

$$\beta(s,t) = \sum_{j=1}^{K}\sum_{k=1}^{K} b_{jk}\phi_j(s)\phi_k(t) \quad \text{or}$$

$$\beta(s,t) = [\phi^T(s)\otimes\phi^T(t)]b, \tag{10}$$

where $b = [b_{11},...,b_{1K},...,b_{K1},...,b_{KK}]^T$

Substituting expansions (9) and (10) of variance-covariance $R(s_1,t_1,s_2,t_2)$ and eigenfunction $\beta(s,t)$ into the functional eigenequation (5), we obtain

$$[\phi^T(s_1)\otimes\phi^T(t_1)]\frac{1}{N}C^TCb = \lambda[\phi^T(s_1)\otimes\phi^T(t_1)]b. \quad (11)$$

Since equation (11) must hold for all *t*, we obtain the following eigenequation:

$$\frac{1}{N}C^TCb = \lambda b. \quad (12)$$

Solving eigenequation (12), we obtain a set of orthonormal eigenvectors b_j. A set of orthonormal eigenfunctions is given by

$$\beta_j(s,t) = [\phi^T(s)\otimes\phi^T(t)]b_j, j = 1,...,J. \quad (13)$$

The random functions $x_i(s,t)$ can be expanded in terms of eigenfunctions as

$$x_i(t,s) = \sum_{j=1}^{J}\xi_{ij}\beta_j(s,t), i = 1,...,N, \quad (14)$$

where

$$\xi_{ij} = \int_S\int_T x_i(t,s)\beta_j(s,t)dsdt.$$

Max $\mathrm{Var}(\beta^T E[X-E(X)|Y]) = \beta^T \mathrm{cov}(E[X-E(X)|Y])\beta$
s.t. $\beta^T \mathrm{cov}(X,X)\beta = 1$

Supplementary Materials 2

Multivariate Functional Regression Models for Quantitative Trait Analysis

Assume that n individuals are sampled. Let $y_{ik}, k = 1,2,...,K$ be K trait values of the i-th individual. Consider a genomic region $[a, b]$. Let $x_i(t)$ be a RNA-seq profile of the *i*-th individual defined in the region $[a, b]$. Recall that a regression model for QTL analysis with the k-th trait and SNP data is defined as

$$y_{ik} = \mu_k + \sum_{j=1}^{J_1} x_{ij}\alpha_{kj} + \varepsilon_{ik} \qquad (1)$$

where μ_k is an overall mean of the k-th trait, α_{kj} is the main genetic additive effect of the j-th SNP in the genomic region for the k-th trait, x_{ij} is an indicator variable for the genotypes at the j-th SNP, $\varepsilon_{ik}, k = 1,..,K$ are independent and identically distributed normal variables with mean of zero and covariance matrix Σ.

Similar to the multiple regression models for QTL analysis with SNP data and multiple quantitative traits, the functional regression model for a quantitative trait can with RNA-seq data can be defined as

$$y_{ik} = \alpha_{0k} + \int_T \alpha_k(t)x_i(t)dt + \varepsilon_{ik}, \qquad (2)$$

where α_{0k} is an overall mean, $\alpha_k(t)$ are a genetic additive effect of a putative QTLs located at the genomic positions t for the k-th trait, $k = 1,...,K$, $x_i(t)$ is a genotype profile, ε_{ik} are independent and identically distributed normal variables with mean of zero and covariance matrix Σ.

Estimation of Additive Effects

We assume that both phenotypes and genotype profiles are centered. The genotype profiles $x_i(t)$ are expanded in terms of the orthonormal basis function as:

$$x_i(t) = \sum_{j=1}^{\infty} \xi_{ij} \phi_j(t) \qquad (3)$$

where $\phi_j(t)$ are sequences of the orthonormal basis functions. The expansion coefficients ξ_{ij} are estimated by

$$\xi_{ij} = \int_T x_i(t) \phi_j(t) dt \qquad (4)$$

In practice, numerical methods for the integral will be used to calculate the expansion coefficients.

Substituting equation (3) into equation (2), we obtain

$$\begin{aligned} y_{ik} &= \int_T \alpha_k(t) \sum_{j=1}^{\infty} \xi_{ij} \phi_j(t) dt + \varepsilon_i \\ &= \sum_{j=1}^{\infty} \xi_{ij} \int_T \alpha_k(t) \phi_j(t) dt + \varepsilon_{ik} \\ &= \sum_{j=1}^{\infty} \xi_{ij} \alpha_{kj} + \varepsilon_{ik}, i = 1,...,n, k = 1,...,K, \end{aligned} \qquad (5)$$

where $\alpha_{kj} = \int_T \alpha_k(t) \phi_j(t) dt$. The parameters α_{kj} are referred to as genetic additive effect scores for the k-th trait. These scores can also be viewed as the expansion coefficients of the genetic effect functions with respect to orthonormal basis functions:

$$\alpha_k(t) = \sum_j \alpha_{kj} \phi_j(t). \qquad (6)$$

Let

$$Y = [Y_1,...,Y_K] = \begin{bmatrix} Y_{11} & \cdots & Y_{1K} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{nK} \end{bmatrix},\ \xi = \begin{bmatrix} \xi_{11} & \cdots & \xi_{1J} \\ \vdots & \ddots & \vdots \\ \xi_{n1} & \cdots & \xi_{nJ} \end{bmatrix},,\quad \xi_i = \begin{bmatrix} \xi_{i1} \\ \vdots \\ \xi_{iJ} \end{bmatrix},,,$$

$$,\alpha_k = \begin{bmatrix} \alpha_{k1} \\ \vdots \\ \alpha_{kJ} \end{bmatrix}, \alpha = [\alpha_1,...,\alpha_K]\ , \varepsilon = \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1K} \\ \cdots & \cdots & \cdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nK} \end{bmatrix}.$$

Then, equation (5) can be approximated by

$$Y = \xi\alpha + \varepsilon \tag{7}$$

The standard least square estimators of α and the variance covariance matrix Σ are given by

$$\hat{\alpha} = (\xi^T\xi)^{-1}\xi^T(Y - \overline{Y}), \tag{8}$$

$$\hat{\Sigma} = \frac{1}{n}(Y - \xi\hat{\alpha})^T(Y - \xi\hat{\alpha}). \tag{9}$$

Denote the matrix $(\xi^T\xi)^{-1}\xi^T$ by A. Then, the estimator of the parameter α is given by

$$\hat{\alpha} = A(Y - \overline{Y}). \tag{10}$$

The vector of the matrix α can be written as

$$vec(\hat{\alpha}) = (A \otimes I)vec(Y - \overline{Y}). \tag{11}.$$

By the assumption of the variance matrix of Y, we obtain the variance matrix of $vec(Y)$:

$$\text{var}(vec(Y)) = \Sigma \otimes I. \tag{12}$$

Thus, it follows from equations (11) and (12) that

$$\Lambda = \text{var}(vec(\hat{\alpha})) = (I_k \otimes A)(\Sigma \otimes I_n)(I_k \otimes A^T) \\ = \Sigma \otimes (AA^T) \quad (13)$$

Test Statistics

An essential problem in QTL analysis or in integrative analysis of imaging and RNA-seq data is to test the **association of genomic region (or gene). Formally, we investigate the problem of** testing the following hypothesis:

$\alpha_k(t) = 0, \forall t \in [a,b], k = 1,...,K,$

which is equivalent to testing the hypothesis:

$H_0 : \alpha = 0$.

Define the test statistic for testing the association of a genomic region with K quantitative traits as

$T = \hat{\alpha}^T \Lambda^{-1} \hat{\alpha}$. (14)

Let $r = \text{rank}(\Lambda)$.

Then, under the null hypothesis $H_0 : \alpha = 0$, T is asymptotically distributed as a central $\chi^2_{(KJ)}$ or $\chi^2_{(r)}$ distribution if J components are taken in the expansion equation (3).

References

1. Henderson D and Plaschko P. (2006). Stochastic differential equations in science and engineering. World Scientific, New Jersey.
2. Sagan H. (1969). Introduction to the calculus of variation. McGraw-Hill, New York.