Published in final edited form as:
Biometrics. 2009 June ; 65(2): 463–469. doi:10.1111/j.1541-0420.2008.01075.x.

Estimating a Multivariate Familial Correlation Using Joint Models for Canonical Correlations: Application to Memory Score Analysis from Familial Hispanic Alzheimer's Disease Study

Hye-Seung Lee[1,*], **Myunghee Cho Paik**[2,**], and **Joseph H. Lee**[3,***]

[1]Pediatrics Epidemiology Center, University of South Florida, Tampa, Florida 33612, U.S.A.

[2]Department of Biostatistics, Columbia University, New York, New York 10032, U.S.A.

[3]Sergievsky Center, Columbia University, New York, New York 10032, U.S.A.

Summary

Analysis of multiple traits can provide additional information beyond analysis of a single trait, allowing better understanding of the underlying genetic mechanism of a common disease. To accommodate multiple traits in familial correlation analysis adjusting for confounders, we develop a regression model for canonical correlation parameters and propose joint modeling along with mean and scale parameters. The proposed method is more powerful than the regression method modeling pairwise correlations because it captures familial aggregation manifested in multiple traits through maximum canonical correlation.

Keywords

Canonical correlation; Estimating equation; Multivariate familial correlation analysis; Regression model

1. Introduction

Memory scores are measured with other cognitive functions to assist in the diagnostic process of Alzheimer's disease (AD). AD is a common disease, whose underlying causes include multiple genetic as well as environmental factors (St. George-Hyslop and Petit, 2005). Because of this multifactorial nature of the disease, gene identification studies of AD have achieved limited success. To overcome this problem, recent attention has been paid to analyzing memory scores (McClearn et al., 1997; Lee et al., 2004).

Memory impairment is considered to be on the pathway to AD. It is generally agreed that accumulation of amyloid plaques and neurofibrillary tangles in the brain leads to neuronal cell deaths in certain brain regions such as the hippocampus, which affects memory performance prior to onset of AD. Hence, if a gene has any influence on AD, the influence would be more direct on memory than on AD. Further, memory scores can provide information for both affected and unaffected members, thereby enhancing power in family studies.

**email*: E-mail: leeh@epi.usf.edu. ***email*: E-mail: mp9@columbia.edu. ****email*: E-mail: joelee@sergievsky.cpmc.columbia.edu.

8. Supplementary Materials Web Appendices and Tables referenced in Section 3 and 4 are available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

Familial correlation is a fundamental analytic tool to investigate genetic influences on a continuous trait. Familial correlation is the correlation of a trait among family members. For a defined relationship (e.g., parent-offspring, siblings, or cousins, etc.), familial correlation greater than zero implies that genetic factors influence the trait. But nongenetic factors complicate the interpretation. In analyzing memory scores from AD families, memory declines with age and tends to be lower in individuals with lower level of education, regardless of having a susceptibility gene. In addition, familial correlation will be lower if there is a large difference in age or education level between two family members, even if they share the same susceptibility gene. The literature on the application of normal theory in the estimation of Pearson's correlation is vast as reviewed by Rao and Province (2000). However, these approaches do not fully address the issues of adjustment of nongenetic differences among family members in the analysis, where normality assumption is most likely to be violated. Moreover, because multiple memory scores are involved in the diagnosis of AD, which are correlated with each other, it would be better to utilize multiple memory scores to extract condensed information for the memory domain.

In modeling of familial correlations for a vector outcome, a straightforward extension of existing works will be challenged by the complicated structure. Table 1 illustrates a case of two memory scores (Y_1 and Y_2) and two family members (Rel_1 and Rel_2). Even for this simple case, a random vector with four elements (Y_1 of Rel_1, Y_2 of Rel_1, Y_1 of Rel_2, and Y_2 of Rel_2) generates six correlation coefficients: two (ρ_1 and ρ_2)describing correlation between two memory scores within each member, which could be assumed to be the same, and four correlation coefficients (φ_{11}, φ_{12}, φ_{21}, and φ_{22}) describing correlation between the two family members. The first two are not of our primary interest and we call them nuisance correlation. The remaining four correlation coefficients, which characterize familial correlation, are of interest. With three scores and two family members, the number of correlation coefficients reflecting familial correlation increases to 9. It is clear that modeling all pairwise correlation coefficients separately will complicate the analysis and interpretation. Thus, we need a better way to characterize familial correlation for a vector outcome. In this article, we propose to use maximum canonical correlation to effectively summarize familial correlations for multiple outcomes. In our example, maximum canonical correlation is the maximum correlation between a linear combination of memory scores from one family member and that from another member over all possible linear combinations. This setup allows us to investigate a genetic influence on memory domain as a whole without focusing on specific memory scores.

For a single outcome, pair-specific adjustment for correlation model was first addressed by Ziegler et al. (2000), implementing the generalized estimating equation (GEE) as proposed by Prentice and Zhao (1991), which extends the GEEs by Liang and Zeger (1986). Later, Yan and Fine (2004) added a variance model and proposed a joint modeling of mean, variance, and correlation. These works allow us to estimate familial correlations as a function of pair-specific covariates such as age differences of two family members, without assuming the normality. As in the model by Yan and Fine (2004), we will express maximum canonical correlation as a function of pair-specific covariates through regression modeling. To our knowledge, none of the existing work addresses pair-specific regression modeling for canonical correlations.

In Section 1, we developed a regression model for canonical correlations and proposed an estimating procedure. Adopting the joint GEEs by Yan and Fine (2004) for each outcome, we then estimated trait-specific means, variances, and nuisance correlations, along with canonical correlation. Section 2 describes memory data from Hispanic AD families and analyzes familial correlations in each memory score using the method by Yan and Fine (2004). In Section 3, we develop the regression model for canonical correlation and present the estimating procedure and asymptotic properties of resulting estimators. In Section 4, finite sample properties are evaluated through simulations. Section 5 revisits our motivating data analysis by implementing

the proposed approach, and Section 6 introduces a weighting method through sensitivity analysis. Lastly, Section 7 summarizes our work.

2. Data and Familial Correlation Analysis for a Single Trait

In this section, we describe our motivating data and show a single trait analysis using the method by Yan and Fine (2004).

2.1 Memory Data from AD Families

The recruitment of familial AD in Caribbean Hispanics began in 1998. The main goal of the study is to identify susceptibility genes for AD and to characterize those genes and other environmental factors that influence the expression of AD. Families were selected from multiple sources, including clinics in the Dominican Republic and in Puerto Rico, the Alzheimer's Disease Research Center-Memory Disorders Center, and doctors' private offices in the Department of Neurology and the General Medical Services at Columbia University. Once a potential proband with AD was identified, a structured family history interview was conducted with available family members to determine if the patient had living siblings or relatives with AD. If the family history interview revealed additional affected family members, we interviewed and examined all other living relatives. All family members completed the same medical and neuropsychological examinations, and their diagnoses were required to meet National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association research criteria for probable or possible AD (Ramas et al., 2002; Lee et al., 2004).

Memory was a part of the neuropsychological tests, which employed selective reminding test (SRT) and Benton visual retention test (BVRT). The SRT is used to assess verbal memory and dementia, while the BVRT examines nonverbal memory. The tests were administered to all family members. In the SRT, subjects were administered six trials in which they were given a list of 12 unrelated words to memorize. After each attempt at recalling the list, the subject was reminded only of the words that were not recalled and then asked to recall the entire list. Verbal memory was measured from the total recall score (TR; max score = 72, failure < 25) or the number of words recalled with and without reminding over the six trials (LTR; max score = 72, failure < 15), along with delayed recall, which was measured by asking the subject to recall the original 12-word list 15 minutes after completing the SRT (DR; max score = 12, failure < 4). Using those three verbal scores, we aim to examine familial correlation for sib pairs in these AD families.

A total of 855 subjects from 181 families yielded 1459 sib pairs with the three verbal memory scores. As shown in Table 2, the number of family members varied from 2 to 19. The mean scores for TR, LTR, and DR were 20.10 (SD = 17.22), 13.45 (SD = 14.27), and 2.70 (SD = 2.90), respectively; 55% were affected with AD, and 65% were women. The mean age was 72 years (SD = 13), and the mean age difference between two siblings was 7 years (SD = 5, lower quartile[Q1] = 3, median[Q2] = 6, and upper quartile [Q3] = 11). The mean time of education was 6 years (SD = 5), and the mean difference of years of education was 3 (SD = 3, Q1 = 1, Q2 = 2, and Q3 = 4). We first examined simple familial correlation. The Pearson's correlations (95% CI) for TR, LTR, and DR scores were 0.302 (0.254, 0.348), 0.307 (0.260, 0.353), and 0.251 (0.202, 0.298), respectively. Because these correlations significantly differ from zero, one may hypothesize that genetic factors influence TR, LTR, and DR scores. However, familial correlations were reduced as the differences in age or years of education increased, and the familial correlation for DR was no longer significant when age difference was greater than the third quartile. Table 3 illustrates the phenomenon.

2.2 Single Trait Familial Correlation Analysis

Because both subject-specific levels of age, education, or sex and pair-specific differences in those levels can confound estimation of familial correlation, we applied the method by Yan and Fine (2004) to adjust this feature, which uses three GEEs for the mean, scale, and Pearson's correlation parameters. For the mean model, identity link function was used, and subject-specific covariates include sex, age, and years of education. The scale model was an intercept model. For the correlation model, using the rescaled Fisher's *z*-transformation, we first fitted an intercept model and then a model with pair-specific covariates including the absolute values of differences in years of education and age, and the indicator of discordance for sex. We analyzed one memory score at a time.

Table 4 shows regression coefficient estimates and standard errors for the mean, scale, and correlation models. The p-values are for the null hypothesis that parameters from each model are equal to zero. The familial correlation estimates are presented from the correlation models. TR, LTR, and DR scores decreased with age but increased with years of education. There was no significant sex difference in TR. After adjusting for subject-specific confounders in the mean model (model 1), the estimated correlation for siblings from the intercept model was 0.068 (p-value = 0.023) in TR. When additional adjustment was made for the differences between two siblings in their education, age, and sex (model 2), the estimated correlation increased to 0.209 (p-value = 0.005). In other words, if two siblings had no differences in those pair-specific confounders, the familial correlation would be 0.209, instead of 0.068. Analogously, for LTR, the estimated correlation would be 0.164 (p-value = 0.030) rather than 0.064 (p-value = 0.035); and for DR, the estimated correlation would be 0.113 (p-value = 0.132) rather than 0.036 (p-value = 0.214).

In this single-trait analysis, we found that familial correlation estimates for sib pairs decreased after adjusting for subject-specific confounders in the mean model; however, they increased after additionally adjusting for pair-specific differences. This phenomenon was consistently observed for all three outcomes. Because these pair-specific differences are known confounders, we presume that estimates from the adjusted model 2 are likely to be closer to the true values. Although correlations for DR were not significant at the level of 0.05, we observed a similar tendency in these three verbal memory scores. Thus, it is of interest to determine a method to summarize these three familial correlations as a multivariate familial correlation for this verbal memory domain.

3. Proposed Method for Multivariate Familial Correlation Analysis

In this section, we propose a regression model for maximum canonical correlations to implement that for multivariate familial correlation analysis. For the *j*th member from the *i*th family, the outcome $Y_{ij} = (Y_{1ij}, Y_{2ij}, \cdots, Y_{mij})$ is an $m \times 1$ vector carrying m multiple traits, and X^*_{ij} is a matrix with m rows and columns including all potential covariates, $j = 1, \cdots, n_i$. We denote the conditional mean of Y_{ij} given X^*_{ij} by $\mu_{ij} = (\mu_{1ij}, \mu_{2ij}, \cdots, \mu_{mij})$.

3.1 Regression Model for Canonical Correlation

Canonical correlations are often defined as the eigenvalues of a certain function of covariances. For a pair from the *i*th family, let $Y_i = (Y_{i1}, Y_{i2})$ be the $2m \times 1$ outcome vector, and the variance of *Yi* be

$$\Sigma_i = \begin{pmatrix} \Sigma_{11i} & \Sigma_{12i} \\ \sigma_{21i} & \Sigma_{22i} \end{pmatrix},$$

where $\Sigma_{11i} = \mathrm{Var}(Y_{i1})$, $\Sigma_{12i} = \mathrm{Cov}(Y_{i1})\ \mathrm{Cov}(Y_{i2})$, $\Sigma_{21i} = \Sigma_{21i}{}^{T}$, and $\Sigma_{22i} = \mathrm{Var}\ (Y_{i2})$ When a and b are $m \times 1$ vectors, standard approach states that maximum canonical correlation between $a^{T} Yi1$ and $b^{T} Y_{i2}$, say γ_{1i}, over all possible choices of a and b is a positive square root of the maximum eigenvalue of $\Sigma_{11i}^{-1}\Sigma_{12i}\Sigma_{22i}^{-1}\Sigma_{21i}$. Denoting eigenvalues of $\Sigma_{11i}^{-1}\Sigma_{12i}\Sigma_{22i}^{-1}\Sigma_{21i}$ as $\left(\gamma_{1i}^{2}\gamma_{2i}^{2}, \cdots, \gamma_{mi}^{2}\right)$ Borga (1995) showed that ordered eigenvalues of $B_i^{-1}A_i$ are (γ_{1i}, γ_{2i}, $\cdots$ γ_{mi}, $-\gamma_{mi}$, $\cdots$, $-\gamma_{2i}$, $-\gamma_{1i}$) when

$$A_i = \begin{pmatrix} 0 & \Sigma_{12i} \\ \Sigma_{21i} & 0 \end{pmatrix}$$

and

$$B_i = \begin{pmatrix} \Sigma_{11i} & 0 \\ 0 & \Sigma_{22i} \end{pmatrix},$$

where **0** denotes a corresponding zero matrix. Note that the eigenvectors of $B_i^{-1}A_i$ contain a linear combination of eigenvectors of $\Sigma_{11i}^{-1}\Sigma_{12i}\Sigma_{22i}^{-1}\Sigma_{21i}$ and a linear combination of eigenvectors of $\Sigma_{11i}^{-1}\Sigma_{12i}\Sigma_{22i}^{-1}\Sigma_{21i}$. Using the covariance function $B_i^{-1}A_i$, we set up a regression model for maximum canonical correlation.

For the pth pair from the ith family, $p = 1, \cdots, P_i$, we first consider a standardized variable for a moment so that μ_i^p is a zero vector and Σ_i^p is a correlation matrix. Because we are not interested in correlations among traits within each member, we assume that nuisance correlation matrices are identical for all pairs ($\Sigma_{11i} = \Sigma_{22i} = \Sigma$) and drop subscript i for B. As for the eigenvector corresponding to the largest eigenvalue, we assume common eigenvector for all pairs. Because our interest is to model the pair-specific largest eigenvalue; e_1 denotes the common eigenvector. Hence, we have the following eigenvalue equation:

$$B^{-1}A_{ip}e_1 = \gamma_{1i}^{p}e_1, \tag{1}$$

where γ_{1i}^{p} is the pair-specific largest eigenvalue of $B^{-1}A_{ip}$.

As for the systematic part of the model, we assume that $g\left(\gamma_{1i}^{p}\right) = X_i^{p}\alpha$ where $g(\cdot)$ is twice the differentiable link function, X_i^{p} is a $1 \times r$ pair-specific covariate vector and α is $r \times 1$ regression parameter that relates the covariate to the maximum canonical correlation. One useful link function is $g\left(\gamma_{1i}^{p}\right) = \tanh^{-1}\left(\gamma_{1i}^{p}\right)$, which guarantees $\gamma_{1i}^{p}(\alpha) \in [-1, 1]$. That is, $\gamma_{1i}^{p}(\alpha) = e^{X_i^{p}\alpha} - 1/e^{X_i^{p}\alpha} + 1$. We will use notation $\gamma(\alpha)$ to emphasize the dependence of γ on α.

As the correlation regression model in the single trait case uses a product of single trait between two members as an "outcome," we need a similar random quantity to construct a regression model for canonical correlations. Consider a random matrix

$$A_{ip}^{S} = \begin{pmatrix} 0 & Y_{i1}Y_{i2}{}^{T} \\ Y_{i2}Y_{i1}{}^{T} & 0 \end{pmatrix},$$

then we have $E\left\{A_{ip}^{S}\right\}=A_{ip}$. To construct an unbiased estimating equation for α, we rearrange equation (1) to $\left\{A_{ip}-\gamma_{1i}^{p}(\alpha)B\right\}e_1=0$ and subsequently $e_{1^T}\left\{A_{ip}-\gamma_{1i}^{p}(\alpha)B\right\}e_1=0$. Note that we have $E\left[e_{1^T}\left\{A_{ip}^{S}-\gamma_{1i}^{p}(\alpha)B\right\}e_1\right]=0$ From this, assuming B is known, an unbiased estimating equation for α can be constructed as follows:

$$\begin{aligned} U(\alpha) &= \sum_{i=1}^{n} U_i \\ &= \sum_{i=1}^{n} \frac{\partial\gamma_{1i}(\alpha)}{\partial\alpha} V_i^{-1}\left[\widehat{e}_1^{T}\left\{A_i^{S}-\gamma_{1i}(\alpha)B\right\}\widehat{e}_1\right]=0, \end{aligned} \tag{2}$$

where $A_i^{S}=\left(A_{i1}^{S},\cdots,A_{iP_i}^{S}\right)$, $\gamma_{1i}(\alpha)=\left[\gamma_{1i}^{1}(\alpha),\cdots,\gamma_{1i}^{P_i}(\alpha)\right]$, V_i is the variace of $A^{s}{}_i$ which is a $P_i \times P_i$ matrix, and $\hat{e}_1$ is an eigenvector corresponding to the largest eigenvalue of $B^{-1}\bar{A}^{S}$ and $\bar{A}^{S}=\Sigma_{i=1}^{n}\Sigma_{p=1}^{P_i}A_{ip}^{S}/N$ where N is the total number of pairs. Let $\widehat{\alpha}$ be the solution of equation (2). Theorem 1 states that $\widehat{\alpha}$ is consistent and asymptotically normally distributed. A proof is given in Web Appendix A.

Theorem 1 (Asymptotic property of $\widehat{\alpha}$ for known B). *Under regularity conditions,* $\widehat{\alpha}$ *is consistent for the true* α, α_0, *and* $\sqrt{n}(\widehat{\alpha}-\alpha_0)$ *is asymptotically normally distributed with mean* 0 *and variance* $n\left\{W_0^{-1}\right\}W_1\left\{W_0^{-1}\right\}^{T}$ *where* $W_0=\Sigma_{i=1}^{n}\widehat{e}_1^{T}B\widehat{e}_1E\left[\frac{\partial\gamma_{1i}(\alpha)}{\partial\alpha}V_i^{-1}\frac{\partial\gamma_{1i}(\alpha)^{T}}{\partial\alpha}\right]$ *and* $W_1=\Sigma_{i=1}^{n}E\left[U(\alpha)U(\alpha)^{T}\right]$.

3.2 Joint Modeling with Trait-Specific Parameters

We drop the assumption that Y_{ij} is a standardized variable; that is, μ_{ij} is no longer a vector of zero, and Σ_i^{p} is not a correlation matrix. Further, nuisance correlation B is unknown. We estimate trait-specific means, variances, nuisance correlations, along with canonical correlation, through a joint regression method in the framework of GEEs.

As in the single-trait case, for the kth trait, we define $Y_{ki}=(Y_{ki1},\ldots,Y_{kin_i})$ and $\mathrm{Var}\left(Y_{kij}|X_{ij}^{*}\right)=\phi_{kij}v_{kij}\left(\mu_{kij}\right)$, where φ_{kij} is a part of variance that does not depend on μ_{kij}. Denote $s_{ij}=(s_{1ij}, s_{2ij}, \ldots$ where $S_{kij}=\frac{\left(Y_{kij}-\mu_{kij}\right)^2}{v_{kij}}$. For the scale parameter model, the $n_i \times 1$ vector $s_{ki}=(s_{ki1},\ldots,s_{kini})$ serves as an "outcome." Let X_{1i} and X_{2i} be $n_i \times p$ and $n_i \times q$ covariate matrices for mean and scale factor whose columns consist of a subset of columns of $X^{*}{}_i$. We assume that $g_{1k}(\mu_{ki})=X_{1i}\beta_k$ and $g_{2k}(\varphi_{ki})=X_{2i}\zeta_k$, where g_{1k} and g_{2k} are link functions, respectively. As for nuisance correlations, we assume those to be common in all subjects. That is, $\rho_{ijkl}=\mathrm{Corr}(Y_{ijk}, Y_{ijl} \mid X_i^{*})=\rho_{kl}$, where $1 \le k < l \le m$. We define an $\frac{m(m-1)}{2}\times 1$ vector $\varphi=(\rho_{12}, \rho_{13}, \ldots, \rho(m-1)m)$, and use $z_i=(z_{i1}+z_{i2})/2$ for a pair as an "outcome," where $z_{ij}=(z_{ij12}, z_{ij13}, \ldots, z_{ij(m-1)m})$, and $Z_{ijkl}=\left(Y_{ijk}-\mu_{ijk}\right)\left(Y_{ijl}-\mu_{ijl}\right)/\sqrt{v_{ijk}\phi_{ijk}v_{ijl}\phi_{ijl}}$. Finally, for canonical correlation, we use equation (2), but note that Y_{kij} will need to be standardized as $Y_{kij}^{*}=\frac{Y_{kij}-\mu_{kij}}{\sqrt{\mathrm{Var}\left(Y_{kij}|X_{ij}^{*}\right)}}$ for A_i^{S}; A_i^{*S} denotes the standardized A_i^{S}.

Denoting covariance matrices of Y_{ki} and s_{ki} by V_{1ik} and V_{2ik}, respectively, we construct the following estimating equation for the parameter $\theta = (\beta_1, \ldots, \beta_m, \zeta 1, \ldots, \zeta m, \varphi, \alpha)$:

$$U^*(\theta) = \sum_{i=1}^{n} U_i^*(\theta) = \sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \mu_{ki}}{\partial \beta_k} V_{1ik}^{-1} (Y_{ki} - \mu_{ki}) \\ \frac{\partial \phi_{ki}}{\partial \zeta_k} V_{2ik}^{-1} (S_{ki} - \phi_{ki}) \\ \sum_{p=1}^{P_i} \left(z_i^p - \varphi \right) \\ \frac{\partial \gamma_{1i}(\alpha)}{\partial \alpha} V_i^{-1} \left[\widehat{e}_i^T \left\{ A_i^{*S} - \gamma_{1i}(\alpha) B \right\} \widehat{e}_1 \right] \end{pmatrix} = 0, \quad (3)$$

where $k = 1, \ldots, m$. Let $\widehat{\theta}$ be the solution of equation (3). Theorem 2 states that $\widehat{\theta}$ is consistent and asymptotically normally distributed. A similar proof to Theorem 1 can be obtained but is omitted for brevity. Note that V_{1ik}, V_{2ik}, and V_i can be replaced by working covariance matrices without loss of asymptotic property from the framework of GEEs.

Theorem 2. *Under regularity conditions, the r* 1 *vector* $\widehat{\theta}$ *is consistent for the true* θ, θ_0, *and* $\sqrt{n}\left(\widehat{\theta} - \theta_0\right)$ is asymptotically normally distributed with mean of 0 *and variance of* $n\left\{W_0^{*-1}\right\} W_1^* \left\{W_0^{*-1}\right\}'$, *where* $W_0^* = E\left[-\frac{\partial U^*(\theta)}{\partial \theta} \right]$ *and* $W_1^* = E\left[U^*(\theta)\, U^* \times (\theta)' \right]$. *Details of* W_0^* *and* W_1^* *are given in Web Appendix B.*

4. Simulations

Simulation studies were conducted to evaluate the finite sample performance of regression canonical correlation estimator from the proposed model. Considering the cases of $m = 2$ and $m = 3$, we replicated 500 times when the number of pairs varied 100, 200, and 500. In standard canonical correlation analysis, it is known that weaker canonical correlations require a larger number of samples (Stevens, 1986). Through simulations, Lee (2007) showed that jackknife estimator via deletion of the *i*th pair reduces the bias in the estimation of canonical correlations, but the variance of the jackknife estimator is much larger than the variance of the biased estimator. Therefore, for the proposed model, we applied the jackknife technique (Quenouille, 1949; Tukey, 1958) in the estimation and employed bootstrap variance estimates of the bias-corrected estimator when the number of pairs is less than 500.

Outcome was generated from the standardized multivariate normal of length 4 ($m = 2$) and 6 ($m = 3$), recognizing that our joint regression model estimates canonical correlation while standardizing for each outcome. We set 0.5 for the components of nuisance correlation matrix Σ. In computing Σ_{12i}, we used the eigen decomposition $B\text{-}1A_i = P_i \Lambda_{1i} P\text{-}1_i$,where Λ_i is the diagonal matrix whose diagonal components are eigenvalues of $B\text{-}1A_i$ and P_i is the matrix whose columns are eigenvectors corresponding to each eigenvalue. In the equation, because we want to compute Σ_{12i} depending only on γ_{1i}, common eigenvector P was assumed and specified by implementing an initial Σ_{12} similar to the dataset. For Λ_i, at $\alpha_0 = (0.3, 0.5, 1.0, 1.5, 1.8)$, we computed $\gamma_{1i}(\alpha_0) = \frac{e^X i^{\alpha} 0 - 1}{e^X i^{\alpha} 0 + 1}$ considering (i) $X_i = 1$ and (ii) $X_i \sim U(0, 1)$, which translated to γ_{1i} ranging from 0.12 to 0.72. For $m = 2$, we set the second-ordered canonical correlation $\gamma_{2i} = 0.5\gamma_{1i}$ and for $m = 3$, we additionally set $\gamma_{3i} = 0.3\gamma_{1i}$. The choice of Σ or P does not affect the evaluation of α estimators.

Tables C.1-C.4 in Supplementary Materials summarize the simulation results for $\widehat{\alpha}$ obtained by solving equation (2) and the jackknife estimates $(\widehat{\alpha})$, reporting the bias, simulation variance

(EMP), and coverage rate of 95% confidence interval (95% CR). For the bootstrap jackknife variance estimator in 100 and 200 pairs, we obtained jackknife estimates for 1000 bootstrap samples from each run of the simulation in each configuration. Bias of $\widehat{\alpha}$ was negligible when $n = 500$. With 200 pairs, bias was not negligible when $\alpha_0 < 0.5$; with 100 pairs, bias was not negligible when $\alpha_0 < 1.0$. In both cases, we note that all finite sample biases were corrected with jackknife estimates $\tilde{\alpha}$. However, the coverage rates of $\tilde{\alpha}$ from the standard approach were lower than the expected lower bound when biases were nonnegligible. After applying the bootstrap, we observed the bias for variance estimates of $\tilde{\alpha}$ to be corrected and the coverage rates from that to be within the expected 95% confidence limits. Trends were similar, but biases were generally larger under uniform X_i and $m = 3$. With the number of pairs less than 500, we recommend using jackknife bias correction along with the bootstrap jackknife variance estimation, especially for weaker canonical correlations.

5. Multivariate Familial Correlation Analysis

We revisit the analysis of three memory scores in Section 2 and implement the proposed method to present a multivariate familial correlation for the verbal memory domain. Table 5 (unadjusted) presents the results. The mean model included sex, age, and education as subject-specific covariates. Scale and nuisance correlation models were unadjusted. Two canonical correlation models were fitted: Model 1 for intercept model and Model 2 adjusted for the differences in years of education and age, and the indicator of discordance for sex. The results for the mean and scale models were the same as those in Table 4, so we omitted them in this presentation. In addition, we did not consider the bootstrap jackknife variance because our samples included 1459 pairs. The familial canonical correlations for sib pairs were 0.084 (p-value = 0.005) for model 1 and 0.154 (p-value = 0.004) for model 2, respectively. The familial canonical correlation from the model 2 represents a multivariate familial correlation when there were no differences in age, education level, and sex between two siblings. Hence, we conclude that there is a significant genetic effect on verbal memory domain measured by TR, LTR, and DR in AD families, after controlling for both subject and pair-specific nongenetic differences between two family members.

6. Sensitivity Analysis Using Weighted Estimating Equations

Until now, we have not taken into account the fact that families were selected because the families had at least two members affected with AD. Thus, the interpretation of results should be conditioned on the sampling scheme. To draw an unconditional inference, the proposed method can be extended using a weighting method where the weight is the inverse of a selection probability, when the selection probability for each family can be estimated. Employing the weighted GEEs by Robins, Rotnitzky, and Zhao (1995), which applies the Horvitz-Thompson (1952) inverse probability, we consider $U^{**}(\theta) = \sum_{i=1}^{n} \frac{1}{\pi_i} U_i^*(\theta)$ where $U_i^*(\theta)$ is equation (3) and π_i denotes a selection probability for the ith family.

Because we did not have data to estimate selection probability, we conducted a sensitivity analysis by varying selection probability. It is generally agreed that the selection probability for a given family will increase as the number of affected family members increases. Assuming selection probability is proportional to the number of affected members or the number of total members for the ith family, we considered two situations: (i) $\pi_i = ca_i$ and (ii) $\pi_i = cn_i$, where c is a constant, and a_i and n_i are the number of AD affected members and the number of total family members, respectively. Note that the choice of c does not affect the results. Table 5 (adjusted) includes the results. In model 1, the familial correlation unadjusted for selection was greater than that adjusted for $\pi_i = ca_i$, while it was smaller than the one adjusted for $\pi_i = cn_i$. In model 2, the familial correlation unadjusted for selection was smaller than those adjusted for two types of selection probabilities. This observation suggests that the support for genetic

influence on verbal memory domain would have been greater if families were randomly selected. Our sensitivity analysis shows that the qualitative conclusion drawn in Section 5 remains the same even after some plausible selection bias was adjusted.

7. Summary and Further Study

In this article, we propose a regression approach for a multivariate familial correlation analysis implementing maximum canonical correlation. Multivariate quantitative trait analysis is a recent topic of interest in the genetics of common diseases. Because most common diseases are characterized by a combination of multiple subclinical phenotypes, multiple quantitative traits used to help the diagnosis of a disease can enhance our understanding of the role of underlying genetic factors toward the disease of interest. To support, we developed a regression model for canonical correlation parameter to adjust for pair-specific confounders, and then jointly modeled the canonical correlation parameters with trait-specific mean, scale, and nuisance correlation parameters. This regressionadjusted maximum canonical correlation is to interpret a multivariate familial correlation as if two family members had no differences in nongenetic confounders.

We further propose a means to take into account selection in our multivariate familial correlation. As in our example, families are often selected only when they include affected members. Hence, as the number of affected family members increases, the possibility of selection for a given family increases. Unlike the analysis of dichotomized outcomes, there is no standard way of conditioning out a selection or ascertainment bias in familial correlation analysis. We extended our proposed model to implement a weighting method where the weight is the inverse of a selection probability. However, this approach can be difficult in practice because it is often unavailable to estimate reliable selection probabilities for each family. We also did not have data available for the estimation of selection probabilities, so sensitivity analysis was included using different selection probabilities.

The proposed approach can have a great impact in various areas using canonical correlations, such as social-science areas that use greater dimension of outcomes, as well as other types of genetic analyses implementing familial correlations, such as heritability estimation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by the National Institute on Aging, National Institutes of Health R37 AG15473 and National Institute of Neurological Disorders and Stroke R01 NS036928. The first author is grateful to Dr Jeffrey Krischer for his support. We thank the reviewers and associate editor, whose comments substantially improved our manuscript.

References

Borga, M. Thesis No. 507. 1995. Reinforcement learning using local adaptive models. ISBN 91-7871-590-3

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 1952;47:32–685.

Lee H-S. Canonical correlation analysis using small number of samples. Communications in Statistics: Simulation and Computation 2007;36:32–985.

Lee JH, Flaquer A, Stern Y, Tycko B, Mayeux R. Genetic influences on memory performance in familial Alzheimer disease. Neurology 2004;62:32–421.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:32–22.

McClearn GE, Johansson B, Berg S, Pedersen NL, Ahern F, Petrill SA, Plomin R. Substantial genetic influence on cognitive abilities in twins 80 or more years old. Science 1997;276:32–1563. [PubMed: 9122704]

Prentice RL, Zhao LP. Estimating equations for parameters in mean and covariances of multivariate discrete and continuous responses. Biometrics 1991;47:32–839.

Quenouille M. Approximation tests of correlation in time series. Journal of the Royal Statistical Society 1949;11:32–84.Series B

Ramas SN, Santana V, Williamson J, Ciappa A, Lee JH, Rondon HZ, Estevez P, Medrano M, Torres M, Stern Y, Tycko B, Mayeux R. Familial Alzheimer disease among Caribbean Hispanics. Archives of Neurology 2002;59:32–91.

Rao DC, Province MA. The future of path analysis, segregation analysis, and combined models for genetic dissection of complex traits. Human Heredity 2000;41:32–42.

Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 1995;90:32–121.

Stevens, J. Applied Multivariate Statistics for the Social Sciences. Lawrence Erlbaum Associates, Inc; Mahwah, New Jersey: 1986.

St. George-Hyslop PH, Petit A. Molecular biology and genetics of Alzheimer's disease. Comptes Rendus Biologies 2005;328:32–130.

Tukey J. Bias and confidence in not quite large samples. Annals of Mathematical Statistics 1958;29:614.

Yan J, Fine J. Estimating equations for association structures. Statistics in Medicine 2004;23:32–874.

Ziegler A, Kastner C, Brunner D, Blettner M. Familial association of lipid profile: A generalized estimating equations approach. Statistics in Medicine 2000;19:32–3357.

Table 1

Correlation structure for two outcomes from two family members

	Y_1 of Rel_1	Y_2 of Rel_1	Y_1 of Rel_2	Y_2 of Rel_2
Y_1 of Rel_1	1	ρ_1	ϕ_{11}	ϕ_{12}
Y_2 of Rel_1	ρ_1	1	ϕ_{21}	ϕ_{22}
Y_1 of Rel_2	ϕ_{11}	ϕ_{21}	1	ρ_2
Y_2 of Rel_2	ϕ_{12}	ϕ_{22}	ρ_2	1

Table 2

Distribution of the number of family members for sib pairs

No. of members	2	3	4	5	6	7	8	9	10	11	12	15	16	17	18	19
No. of families	48	29	28	25	19	12	3	3	2	3	4	1	1	1	1	1

Table 3

Familial correlation estimates (95% CI) using standard calculation

		No. of families (No. of pairs)	TR	LTR	DR
Overall		181 (1459)	0.302 (0.254, 0.348)	0.307 (0.260, 0.353)	0.251 (0.202, 0.298)
Age	≤3	138 (409)	0.441 (0.359, 0.516)	0.435 (0.353, 0.510)	0.396 (0.310, 0.474)
difference	>3 and ≤6	112 (326)	0.294 (0.191, 0.390)	0.278 (0.174, 0.375)	0.216 (0.109, 0.316)
(years)	>6 and ≤11	123 (427)	0.267 (0.176, 0.352)	0.242 (0.151, 0.329)	0.245 (0.153, 0.332)
	>11	85 (297)	0.162 (0.049, 0.271)	0.241 (0.131, 0.345)	0.077 (-0.037, 0.189)
Education	≤1	141 (553)	0.323 (0.246, 0.395)	0.324 (0.248, 0.397)	0.209 (0.127, 0.287)
difference	>1 and ≤2	84 (260)	0.413 (0.306, 0.508)	0.390 (0.281, 0.488)	0.370 (0.259, 0.470)
(years)	>2 and ≤4	93 (310)	0.279 (0.173, 0.378)	0.361 (0.259, 0.454)	0.296 (0.191, 0.394)
	>4	84 (336)	0.162 (0.056, 0.264)	0.145 (0.039, 0.248)	0.139 (0.033, 0.243)

Note: 95% CI: 95% confidence interval using Fisher's *z*-transformation.

Table 4

Familial correlation estimates using the method by Yan and Fine (2004)

			TR			LTR			DR		
Model			**PE**	**SE**	**P**	**PE**	**SE**	**P**	**PE**	**SE**	**P**
Mean		Int	70.131	4.072	0.000	52.128	3.436	0.000	10.928	0.669	0.000
		Male	-0.711	0.941	0.450	-1.638	0.755	0.030	-0.361	0.152	0.018
		Age	-0.717	0.049	0.000	-0.577	0.041	0.000	-0.120	0.008	0.000
		Educ	0.605	0.114	0.000	0.536	0.098	0.000	0.088	0.017	0.000
Scale		Int	181.173	10.621	0.000	125.063	8.130	0.000	5.288	0.309	0.000
Corr	Model 1	Sib	0.136	0.059	0.023	0.129	0.061	0.035	0.072	0.058	0.214
		ρ	0.068			0.064			0.036		
	Model 2	Sib	0.424	0.150	0.005	0.331	0.152	0.030	0.227	0.151	0.132
		ρ	0.209			0.164			0.113		
		dedu	-0.049	0.021	0.022	-0.051	0.019	0.007	-0.020	0.017	0.235
		dage	-0.016	0.011	0.134	-0.002	0.014	0.876	-0.013	0.013	0.307
		dsex	-0.060	0.102	0.554	-0.081	0.111	0.464	0.009	0.109	0.931

Note: PE: parameter estimate; SE: standard error from the sandwich variance estimate; p: p-value; Corr: Pearson's correlation model; model 1: unadjusted for pair-specific confounders; model 2: adjusted for pair-specific confounders; ρ: Pearson's correlation corresponding to the parameter estimate from Corr; dage: age difference; dedu: education difference; dsex: sex difference.

Table 5

Familial canonical correlation estimates using the proposed joint modeling

			TR, LTR, and DR								
						Adjusted					
			Unadjusted			$\pi_i = Ca_i$			$\pi_i = cn_i$		
Model			**PE**	**SE**	**P**	**PE**	**SE**	**P**	**PE**	**SE**	**P**
Corr	Model 1	Sib	0.169	0.060	0.005	0.130	0.062	0.038	0.253	0.093	0.007
		γ_1	0.084			0.065			0.126		
	Model 2	Sib	0.310	0.107	0.004	0.541	0.167	0.001	0.429	0.164	0.009
		γ_1	0.154			0.264			0.211		
		dedu	-0.017	0.015	0.267	-0.082	0.034	0.016	-0.013	0.018	0.463
		dage	-0.012	0.008	0.143	-0.021	0.012	0.070	-0.019	0.011	0.086
		dsex	-0.005	0.074	0.950	0.020	0.131	0.879	-0.003	0.101	0.973
Corr(TR, LTR)			0.846	0.031	0.000	0.772	0.033	0.000	0.874	0.032	0.000
Corr(TR, DR)			0.753	0.033	0.000	0.695	0.039	0.000	0.771	0.031	0.000
Corr(LTR, DR)			0.770	0.036	0.000	0.698	0.042	0.000	0.811	0.036	0.000

Note: Unadjusted: unadjusted model for selection probability; Adjusted: adjusted model for selection probability; PE: canonical correlation estimate; Corr: maximum canonical correlation model; $\gamma 1$: canonical correlation corresponding to the parameter estimate from Corr; π_i: selection probability; c: constant; a_i: number of AD affected family members; n_i: number of family members; the rest of configurations are as in Table 4.